When Everyone Writes Well: The Economics of Writing with AI*

Johan Fourie†

Abstract

Writing produces two things: the manuscript published now and the capability its author will hold next year. Artificial intelligence can raise the first while lowering the second, and which happens depends on the task the author delegates. The relevant comparison is between workflows rather than between using and refusing AI. This article separates writing into its tasks and applies a standard tool of economics to each, synthesising existing models into five predictions. Cheaper drafting lets an author test more versions of an argument, and the largest gains should go to researchers writing in a second language, whose language costs are unrelated to their evidence. Manuscript studies, including new evidence from economics working papers, fit this pattern, although they measure prose and output rather than argument quality. Polishing prose with a model while delegating the first formulation of a problem can leave a better text and a weaker author, and it makes the generated draft an anchor and a default. Cheap fluency makes polished prose a weaker guide to the underlying research. The evidence is uneven, and I note where it is thin. As models improve, the bottleneck moves, and so does the division of labour between author and machine.

Keywords: artificial intelligence; scholarly writing; human capital; language barriers; anchoring; judgement

JEL codes: D83; I23; J24; O33

*I thank Tyler Cowen, Di Kilpert, Leon-Ben Lambrechts, Kelsey Lemon, Deirdre McCloskey, Leonard Praeg, and Amy Rommelspacher for comments on earlier drafts. They are not responsible for the use I have made of them. I used Anthropic's Claude, OpenAI's Codex and Google's Gemini to map the literature, work through the accompanying derivation, classify the institutional affiliations in Section 8 by country, and draft and revise the manuscript; refine.ink, an AI review service, reviewed it. The claims, and the responsibility for them, are mine. Cite this paper as: Fourie, Johan. 2026. "When Everyone Writes Well: The Economics of Writing with AI." Working Paper, Department of Economics, Stellenbosch University.

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

1 Introduction

Imagine an economic historian working through an archive. She is considering two explanations for a document she has found. One is the standard account familiar to the literature. The second tries to make sense of an awkward inconsistency the standard account cannot explain. If she uploads the documents to a large language model requesting a draft, it returns the most likely explanation. She then revises that draft, accepting the main argument. Has the model merely saved her time or has it also shaped her judgement?

Now consider a researcher who has proved a theorem or completed an experiment. Her study was set up carefully, guided by disciplinary standards and years of experience. She has already reviewed the results, but she still needs to describe the method and report her conclusions. Here, a language model can help with the writing without changing the evidence or the argument; it simply reduces the cost of having to report work already done. Now imagine she is also not a native English speaker. She reads English fluently but writes it slowly, needing three hours to produce a page that a native speaker writes in one, and even when she takes the time, reviewers return her manuscripts with complaints about the language. A language model removes much of this burden without altering the quality of her evidence.

These examples illustrate the central distinction of this paper. Using a language model to write up a settled result or to translate an argument into fluent English is very different from using it to decide what the evidence means in the first place.

A common refrain raised against AI-assisted writing is that 'writing is thinking'. Critics warn that because students and scholars develop understanding through the act of writing, delegating it to artificial intelligence may leave them without the abilities that the finished text seems to show (Sparrow and Flenady, 2025; Corbin et al., 2026). In this paper, I argue that writing can contribute to thinking without being identical to it. We need to separate the text produced today from the capability its author retains tomorrow, because a polished manuscript does not necessarily mean its author has learnt more in the process of producing it.

Section 2 begins by separating the tasks involved in writing and the outcomes they produce. Section 3 looks beyond the current manuscript to what its author will be able to do on the next problem. There is also a parallel question about readers: the way an argument is expressed can shape what they believe and remember, even when the underlying evidence stays the same. A machine need not think, in any philosophical sense, for its output to influence human thought.

I use five standard economic frameworks to examine these effects. The analysis starts from an asymmetry in costs. Advocates of AI-assisted writing point to clear gains in output and assessed quality when prose becomes cheaper to produce (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2025). But while language models make it much cheaper to generate another draft, checking whether that draft is sound still requires effort. A model can assist with checking too, for instance by offering criticism, but the search results below lower the cost of drafting while holding the cost of checking fixed. Each section isolates one mechanism under explicit assumptions, and though the models share notation, they serve as linked benchmarks rather than a single unified model of how

authors choose what to delegate.

The first prediction comes from the economics of *search*. An author who spends less effort producing each draft can afford to consider more alternatives, and if she ranks them accurately, examining more candidates raises the expected quality of the argument she selects. Under the model's assumptions, the largest gains go to those for whom drafting is initially most expensive, particularly researchers writing in a second language. Section 4 sets out this result and its limits.

Judging an argument, however, is itself an acquired skill. This leads to the second prediction, which draws on the theory of *human capital*, or capabilities built through practice. If a model takes over the work through which an author learns, and nothing else replaces that practice, the current manuscript can improve even as her later unaided ability declines. Whether assistance narrows or widens gaps across authors then depends on why their abilities differ in the first place. Section 5 examines these possibilities.

The third prediction concerns how an initial draft affects subsequent evaluation and search. An author who reads a generated account before forming her own may favour alternatives that resemble that first account. The initial text thus becomes an *anchor* for her judgement and a *default* she may choose to accept rather than taking the time to work out her own argument. Section 6 looks at how an initial draft distorts the ranking of alternatives as well as how it encourages an author to stop searching early.

The fourth prediction turns from the writer to the reader, drawing on *signalling* theory. Fluent writing helps readers follow a paper, but when prose is costly to produce it also serves as a signal of the author's command of the subject. Once fluency becomes cheap for everyone, polished prose becomes a weaker guide to the quality of the research beneath it. Under the paper's assumptions, the association between fluency and research quality weakens, though it does not reverse: good writing becomes a weaker guide, not a sign of bad work. Section 9 develops this distinction.

A fifth claim, offered as a conjecture rather than a formal result, looks beyond the individual author to the field as a whole. If many researchers start from similar generated accounts, they may write more papers while exploring fewer distinct explanations. Choices that benefit each scholar in isolation could then narrow the range of ideas the discipline considers.

The empirical evidence across these five claims is uneven. Large-scale studies that screen millions of scientific manuscripts for markers of AI assistance find patterns consistent with the first prediction about who gains most (Kusumegi et al., 2025; He and Bu, 2026), and new evidence from 42,704 economics working papers in Section 8 shows the same pattern: since 2024, a detector of AI-like prose has flagged papers more often when their authors' institutions lie outside English-speaking countries. Neither source, however, shows whether authors selected stronger arguments. Classroom and adult learning experiments support the trade-off in the second prediction: assistance can raise performance while the tool is available yet lower unaided performance once it is removed unless paired with sustained human effort (Bastani et al., 2025; Cruces et al., 2026). By contrast, whether gaps between authors narrow or widen remains sensitive to the underlying assumptions and to whether performance is measured with or without the tool. The third prediction has not

yet been tested directly on scholars, whereas screening experiments in hiring and investment offer more direct support for the fourth, and there is no direct test of the fifth. Sections 8 and 9 weigh this evidence and its limits, bearing in mind that much of it comes from earlier models whose capabilities have since been surpassed.

This paper connects four strands of research. The first studies the economics of generative AI and knowledge production. Recent theoretical work shows that cheap machine answers can reduce the supply of new human-generated knowledge (Acemoglu, Kong and Ozdaglar, 2026) and weaken the quality signal in written communication (Gans, 2024). Reliance on a shared model can also lead decision makers to make correlated mistakes (Kleinberg and Raghavan, 2021). Empirical and practitioner studies, on the other hand, have catalogued how economists can delegate micro-tasks (Korinek, 2023) and measured AI adoption and its effects on output and quality across workers and scholars (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2025; Dell'Acqua et al., 2026; Kusumegi et al., 2025; He and Bu, 2026; Gross, Shvadron and Zhang, 2026). My contribution here is to disaggregate writing into its constituent tasks and trace how each affects both the finished manuscript and the author's future capability. Section 8 complements this analysis with descriptive evidence from economics working papers.

The second strand examines how writing shapes learning and skill acquisition. Cognitive research shows that writers often work out their ideas in the act of putting them into words (Flower and Hayes, 1981; Bereiter and Scardamalia, 1987). McCloskey (1985) argues on similar grounds that economic content cannot be separated from its expression. Where recent criticisms of AI-assisted writing in universities have focused on the philosophical objection that delegating prose bypasses thought (Sparrow and Flenady, 2025; Corbin et al., 2026), experimental and survey studies confirm that unguided AI assistance can boost immediate task performance and reduce cognitive effort while leaving learners less able to work unaided (Bastani et al., 2025; Lee et al., 2025; Cruces et al., 2026). These findings make the loss of practice a serious concern, though they do not tell us which writing tasks build human capital. This paper addresses the philosophical objection directly in Section 3 and separates initial argument formulation from prose execution in a model of human-capital formation in Section 5. Section 7 then places this risk of deskilling in an older economic debate stretching from Smith (1776) on the division of labour to modern task-based models of technology (Autor, 2015; Autor and Thompson, 2025): what happens when automating routine tasks removes the practice that builds expertise for the tasks that remain?

The third strand studies anchoring and default effects, drawing mainly on evidence from numerical estimation and savings or effort choices (Tversky and Kahneman, 1974; Madrian and Shea, 2001; Jachimowicz et al., 2019; Augenblick, Niederle and Sprenger, 2015). This paper extends those behavioural insights to sequential text generation, where a machine-supplied first draft combines both mechanisms: it shapes how an author evaluates rival accounts while offering an acceptable text at almost no immediate effort. Because stopping early is rational when a starting draft is already strong, Section 6 asks specifically when present bias leads an author to stop sooner than her own patient weighing of costs and benefits would justify, and Section 10 connects the practice

Table 1: Seven tasks in scholarly writing and their implications for delegation

Task	Function	Implications for delegation
Inscription and transcription	Express specified content in grammatical sentences	Often suitable for delegation
Externalisation	Put a partly formed idea into words so it can be examined	Can contribute to learning and discovery
Candidate generation	Produce alternative arguments or ways of expressing them	Requires a basis for comparison
Revision and criticism	Identify weaknesses and revise the account	May benefit from combining human and AI input
Selection and verification	Judge which claims are supported and important, and how to attribute them	The author remains accountable
Narrative design	Choose the causal account and how to present it	Requires judgement because these choices affect readers' beliefs
Deliberate practice	Develop the ability to perform these tasks in future	Its value cannot be judged from the current manuscript alone

of drafting an initial account oneself to the economics of active choice (Carroll et al., 2009).

The fourth strand concerns signalling in scientific communication and the governance of AI in academia. Screening experiments and a structural analysis of platform data find that cheap fluency can make written applications less useful to evaluators (Cowgill, Hernández-Lagos and Wright, 2026; Galdin and Silbert, 2025), textual studies document convergence in style and reasoning (Doshi and Hauser, 2024; Sourati, Ziabari and Dehghani, 2026), and recent policy studies examine automated detection and journal disclosure rules (Jabarian and Imas, 2025; He and Bu, 2026) as well as AI-generated peer review (Liang et al., 2024). Building on Spence (1973), Section 9 shows when cheap fluency weakens prose as a signal of research quality and why removing language barriers can instead make strong work by second-language authors easier to recognise. Section 10 then uses these results to evaluate journal disclosure policies and propose testable guidelines for researchers and journals.

The overall contribution is a synthesis of how economics can help guide our thinking on writing with large language models. Rather than asking whether AI helps or harms scholarly writing in the abstract, the paper asks which task the author delegated and which outcome a study measured. These questions help explain why seemingly contradictory findings can all hold at once.

2 What Writing Does

Because writing is both a way of working out ideas and a way of reporting them, evaluating delegation requires us to separate the tasks that a writer performs. Table 1 distinguishes seven such tasks.

In day-to-day work these tasks overlap. Planning a sentence can expose a missing step in a causal chain, and changing a word can clarify a distinction. Yet developing a causal argument and writing

it in grammatical prose are distinct tasks, even when an author performs them simultaneously.

Research in the psychology of writing supports this distinction. In the model of Flower and Hayes (1981), writers alternate between planning, translating ideas into text, and reviewing what they have written. Bereiter and Scardamalia (1987) similarly distinguish *knowledge telling* (retrieving existing ideas and writing them down) from *knowledge transforming*, in which formulating a text reshapes the writer's understanding. Coordinating substance and prose for an audience places heavy demands on working memory (Kellogg, 1994).

Philosophers of mind argue further that working with text on a page can form part of thinking itself, enabling people to follow chains of reasoning too complex to hold in working memory alone (Menary, 2007; Clark and Chalmers, 1998). Still, not every writing task contributes equally to thought, and expertise alters where writing aids reasoning. Oatley and Djikic (2008) observe that experienced writers put tentative ideas into words and use the draft to reconsider the argument itself, whereas novices tend to stop with the first version they express.

To track what delegation changes, the analysis distinguishes three outcomes: *manuscript quality*, *author capability*, and *field quality*. The first concerns the clarity and substantive accuracy of the finished paper. The second concerns what the researcher can do unaided, including on new problems; both are modelled formally below. The third outcome, field quality, is not modelled as an equilibrium choice; it concerns the diversity of explanations a discipline maintains and the effort readers must expend to identify reliable work. Whereas the economic case for delegation focuses on the manuscript (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2025; Korinek, 2023), the strongest criticisms focus on the author and the field (Sparrow and Flenady, 2025; Corbin et al., 2026; Doshi and Hauser, 2024), neither of which can be inferred from the finished text alone.

3 The Case for Human Prose

The main objection to delegating scholarly writing is that drafting is how scholars form judgement. Researchers rarely begin with a complete, internally consistent argument. Stating a claim in plain sentences forces them to specify what is happening and why, and revising those sentences exposes ambiguities that were easy to miss while the idea remained an intuition. Because repeated practice builds this capacity over time, a student or scholar who delegates the difficult parts of drafting may produce a passable paper without developing the analytical habits the text seems to reflect.

Recent critiques of AI in universities build on this point. Sparrow and Flenady (2025) distinguish knowing that a claim is true from knowing how to arrive at and evaluate it, arguing that university education teaches the first ability through practice in the second. Corbin et al. (2026) likewise contrast assignments graded solely as finished products with essays in which students work through ideas by drafting and revising. Instead of returning to pre-AI assessment, they call for teaching that evaluates the intellectual work students undertake and not merely whether they produced polished sentences unaided.

Two opposing newspaper columns illustrate how this debate has reached a wider public. Stephens

(2026) urges readers never to write with AI, arguing first that writing requires an author to state ideas in grammatical sentences open to scrutiny, and second that routine and serious writing draw on a single indivisible skill. The first claim restates the formation objection; the second is a testable hypothesis about skill transfer implying that delegating routine text lowers later unaided performance on demanding problems. Stephens cites a study in which a commercial detector flagged 56 of 100 education dissertations in one subfield as containing more than 10% AI-generated text, though widespread use does not by itself prove skill loss and the detection biases discussed in Section 8 apply to that figure. Taking the opposite view, Cruz (2026) argues that universities confuse archiving with thinking. After allowing students to use AI on final essays under written guidelines, he examined each student orally for 45 minutes and reported that at least half the class argued more clearly aloud than on paper. Yet oral defence is a distinct task that rewards extemporaneous speaking fluency, and Cruz's account includes neither a delayed test on new material nor a control group. Together, the two columns frame opposing hypotheses that empirical work can test.

Even for an experienced scholar, editing a machine-generated draft is not the same mental exercise as revising her own. A supplied text can direct attention to one explanation and away from details that do not fit (Section 6). Because generated prose is smooth and grammatical, it offers few surface flaws to prompt closer scrutiny; its weakness is more often an uncertain premise stated as settled fact. An author who could have built a stronger argument on her own may instead spend her time polishing a flawed one.

Having a human review the output does not solve this problem. An expert can check a generated claim against domain knowledge, but a novice may mistake fluency for rigour, and even capable evaluators scrutinise output less carefully when monitoring a generally reliable system. In a field experiment with 758 management consultants, Dell'Acqua et al. (2026) found that GPT-4 raised performance on tasks within its capabilities, but on a deceptive problem just outside its capabilities, consultants using the model were 19 percentage points less likely to reach the correct conclusion than unaided peers because they accepted plausible errors without checking the data. In a survey of 319 knowledge workers across 936 generative-AI workflows, Lee et al. (2025) similarly found that higher trust in the model was associated with lower self-reported critical-thinking effort. In a laboratory decision task, interfaces that forced participants to reason analytically before seeing AI advice reduced reliance on bad recommendations, though participants disliked the extra effort and gains were uneven (Buçinca, Malaya and Gajos, 2021). Retaining the final decision does not guarantee an independent judgement, nor can a model visit an archive or recognise unstated assumptions from data collection. What matters is whether the workflow requires the author to examine the evidence directly, regardless of who drafts the sentences.

A more fundamental objection questions the analytical framework of this paper directly. McCloskey (1985) rejects the premise that content and expression can be separated because economists discover the substance of an argument in the course of writing it out. Philosophy's linguistic turn (Rorty, 1967) and the extended-mind literature (Menary, 2007) similarly reject the view that language merely expresses thoughts formed in advance, and in linguistics Reddy (1979) criticises the

'conduit metaphor', the assumption that speakers transfer fixed meanings through words. Any model that separates the cost of generating text from the cost of evaluating its substance faces the same objection.

My response splits this objection into two parts. First, I accept that writing helps form both arguments and the capacity to judge them. In Section 5, that formative role enters through the model's second output, the author's future capability. Testing this mechanism does not require a philosophical definition of judgement, only delayed tests of what authors can do unaided on unfamiliar problems. Because unaided performance reflects several distinct capacities at once (domain knowledge, research design, analytical reasoning, and linguistic fluency) and a single essay score cannot separate them, Section 10 outlines an experimental design that distinguishes evaluating arguments from formulating them.

Second, as an objection to how the current manuscript is produced, McCloskey's critique is valid. Proposition 1 in Section 4 assigns separate costs to generating a candidate draft and checking it, and requires that lower drafting costs leave checking costs, the distribution of candidate quality, and the author's judgement unchanged. Writing down separate cost symbols does not make the two activities independent, so the proposition applies only to tasks where composing prose and judging substance interact weakly. As Table 1 suggests, that condition is most plausible when reporting a settled method and least plausible when forming an argument from unsettled evidence. McCloskey's objection is thus strongest precisely where this paper locates the greatest risk of delegation.

Reddy's critique highlights a related friction. Before an author can judge a machine-generated draft, she must reconstruct the argument its words express, and that interpretive effort is part of the checking cost. Cheap drafting therefore yields a smaller saving if making sense of machine prose takes nearly as much effort as drafting one's own argument.

Yet McCloskey's and Reddy's critiques also have the opposite implication. If putting an idea into words exposes its gaps, seeing where a model's summary diverges from what the author meant can expose gaps too, and increasingly so as models improve. Writing an empirical section, for instance, requires stating plainly what the computer code does. Because a language model can read code quickly, a mismatch between its description of the script and the intended specification can reveal a coding error. Prompted well, a model can also challenge unsupported claims before a paper reaches a referee (Section 5).

Human-generated prose also has intrinsic value beyond research productivity. Readers may value knowing that an essay was written in its author's own voice, though that value lies outside this paper's models of knowledge and skill formation. Many scholars also enjoy the craft of writing. Smith (1776) opened his analysis of wage differentials with the agreeableness of work itself, and labour economists consistently find that people accept lower pay for meaningful or enjoyable work (Maestas et al., 2023; Cassar and Meier, 2018). Experimental evidence on whether AI erodes intrinsic motivation is mixed. Across four studies with 3,562 participants, generative assistance raised immediate performance but left participants more bored and less intrinsically motivated on subsequent unaided work (Wu et al., 2025), whereas a smaller study found no significant difference

across support conditions (Fan et al., 2025).

Enjoyment also interacts with skill formation and task choice. A scholar who enjoys writing will practise it more, and mastering a demanding activity is often rewarding in itself (Loewenstein, 1999). In a model of workplace automation, Gans (2026) shows that workers may spend extra unrecorded time on tasks they enjoy, giving firms a reason to automate those tasks to redirect effort. An independent scholar faces no such contracting problem, yet enjoyment still affects which tasks writers delegate, and standard output measures miss the extra hours people invest in work they like. Even so, a researcher who needs three hours to write a page of English pays a much higher price for that amenity than a native speaker who needs one. Taking pleasure in writing is a sound reason to keep drafting oneself, but it is not grounds for barring others from assistance. Across all of these objections, the force of the concern depends on which task in Table 1 is delegated and on who the writer is.

4 Cheaper Search

Start with the case in which drafting and judging can be separated. Here, delegation can improve a manuscript by allowing the author to test more substantive versions of her argument, which I call 'candidates' to distinguish them from cosmetic rewordings. Suppose she cannot tell how well a candidate fits the evidence until she has produced and inspected it. Let c_v denote the effort needed to generate a candidate and c_s the effort needed to check and evaluate it, where the subscripts stand for variation and selection. With a total effort budget B, the number of candidates she can examine is

$$n = \frac{B}{c_v + c_s} \tag{1}$$

For example, with 40 hours to develop an argument, an author who needs seven hours to draft a reformulation and one hour to check it against the evidence can examine five versions. If a language model cuts generation time to one hour while checking still takes one hour, she can examine 20.[1]

Writing technologies shift these two costs differently. When word processors made it cheaper to alter text without retyping, writers revised more frequently, yet a larger share of their edits changed wording rather than substance (Ransdell and Levy, 1994). Cheaper line-editing did not automatically generate more substantive alternatives.

Suppose each candidate's true quality q_i is drawn independently from a fixed, non-degenerate distribution F with a finite mean. Assuming the author accurately identifies and keeps the best version inspected so far, her retained quality after n draws is $Q_n = \max\{q_1, \ldots, q_n\}$, defined on the same scale used in Section 6. Because the author can retain any earlier draft if later ones are worse, this is standard search with recall (Weitzman, 1979).

[1]Strictly, an author can examine only whole candidates, so the discrete count is $\lfloor B/(c_v + c_s) \rfloor$ and expected quality is weakly decreasing in cost between integer thresholds. The supplement proves both this discrete version (assuming $B, c_s > 0$ and that each compared author can afford at least one checked candidate) and a differentiable relaxation on $B/(c_v + c_s) > 1$.

Proposition 1 (Cheaper iteration, better argument)**.** *Expected final quality* $\mathbb{E}[Q_n]$ *is increasing and concave in the number of candidates* n *and decreasing in their cost. If writers initially differ only in production cost, a technology that lowers this cost to a common level, with checking costs unchanged, gives the largest gains to those whose initial production cost was highest.*

From basic order statistics, the expected gain from examining one additional candidate is

$$\mathbb{E}[Q_{n+1}] - \mathbb{E}[Q_n] = \int F(x)^n (1 - F(x))\, dx > 0. \tag{2}$$

At any threshold x, the next candidate improves on the current best when all n previous versions fell below x (with probability $F(x)^n$) and the new version exceeds x. Because each new candidate must improve on the best of a larger sample, this expected marginal gain diminishes as n rises, making expected quality concave in n.

The cross-sectional part of the proposition yields the first prediction. A scholar writing in a second language often faces high drafting costs c_v for reasons unrelated to the originality of her question or the quality of her evidence. By compressing differences in the cost of producing fluent English, language models reduce a friction that penalises researchers on linguistic grounds. Under the proposition's assumptions, the largest immediate gains from cheaper drafting go to authors whose initial writing costs are highest, above all second-language researchers (examined in Section 8).

Three scope conditions govern this result. First, applying the proposition to language barriers assumes that writing in a second language raises the cost of producing prose (c_v) without raising the cost or reducing the accuracy of evaluating it (c_s); if second-language scholars also find it harder to judge subtle differences in English drafts, their net gain is smaller. Second, the formula assumes independent draws from a fixed quality distribution, whereas multiple drafts from the same language model are correlated and may yield smaller gains (an issue that reappears at the field level in Section 9). Third, the search benchmark holds the author's judgement fixed and assumes accurate selection, whereas a model may either identify a missing comparison or bias the writer towards a conventional explanation (Section 6). Moreover, the checking cost c_s places an upper bound on search: as c_v approaches zero, n approaches B/c_s, not infinity. Only candidates the author actually checks contribute to the gain.

Early workplace experiments match this distributional pattern, providing tentative support for the first prediction, though low baseline scores need not reflect language costs. Measuring assisted output while the tool is in use, Noy and Zhang (2023) and Brynjolfsson, Li and Raymond (2025) find that in professional writing and customer-support tasks, generative assistance raised performance most among workers with the lowest baseline unassisted scores.

5 A Manuscript and a Writer

Drafting a paper produces two outputs, the current manuscript and the author's capacity to conduct future research. In economic terms, that capacity is human capital built through deliberate effort that depreciates when left unused (Becker, 1964). Even the stamina required for sustained cognitive work responds to training. In a field experiment with 1,636 primary-school pupils, structured practice at continuous mental effort significantly raised performance on subsequent demanding tasks (Brown et al., 2025). Removing the difficult parts of writing can also remove that training, so evaluating delegation requires considering both current output and future capability.

Following Table 1, the model distinguishes between delegating linguistic execution (expressing an argument the author has already worked out) and delegating representation formation (turning raw evidence into a structured account). Let $d_L \in [0,1]$ denote the delegated share of linguistic execution and $d_R \in [0,1]$ the delegated share of representation formation, and write $d = (d_L, d_R)$. An author who prompts a model for a full draft before formulating the argument herself chooses high d_L and high d_R; an author who writes a rough account first and uses the model to polish the prose combines low d_R with high d_L.

The quality of the current manuscript is given by

$$Q_t = G(n(d_L), J_t, b(d_R, J_t)) \,. \tag{3}$$

Here J_t denotes research-specific human capital, the capability accumulated within a discipline to evaluate candidate arguments and to formulate arguments from evidence, where 'judgement' refers to its evaluative component. Because the model takes the true scientific value of any candidate argument as given, it does not address what makes a research question worth asking. Separating evaluation from formulation matters directly below because those two components of J_t interact with cheaper candidate generation in opposite ways.

In this production function, delegating linguistic execution (d_L) allows the author to inspect more candidate versions $n(d_L)$, which weakly raises expected manuscript quality under accurate ranking and strictly raises it when additional draws have positive probability of improving on any retained draft (Proposition 1). The term b captures a ranking distortion when the author delegates initial formulation (d_R) and anchors on the machine's framing (Section 6). Accordingly, the model imposes strictly positive returns to generating more candidates only when ranking is unbiased, $b = 0$, on that non-degenerate domain (proved in Proposition 1 for accurate selection and extended in the supplement to non-degenerate Gaussian evaluation noise and to the formulation benchmark when machine draws can exceed the author's own account). When ranking is systematically biased, generating more candidates can lower the expected quality of the chosen paper.

The author's capability evolves from one project to the next according to

$$J_{t+1} = (1-\delta)J_t + L(g(d_R), r(d), v(d), f(d); J_t) \,. \tag{4}$$

Here δ represents skill depreciation in the absence of practice (Ericsson, Krampe and Tesch-Römer, 1993), and the learning function L depends on four activities: generating an initial account oneself (g), revising actively (r), verifying claims against primary sources and data (v), and absorbing critical feedback (f). Delegating initial formulation (d_R) directly reduces g, whereas delegating the polishing of an existing argument (d_L) need not. Used as a critic of the author's draft, a model can even increase time spent revising (r) and checking (v).

Suppose choices are evaluated by the weighted objective $W_t = Q_t + \omega J_{t+1}$, where ω is the weight placed on future capability relative to the current paper (a journal editor cares mainly about Q_t, whereas a scholar or doctoral supervisor also cares about J_{t+1}). For a small increase in delegation along coordinate $k \in \{L, R\}$, the marginal effect is

$$\frac{dW_t}{dd_k} = \underbrace{\frac{\partial Q_t}{\partial d_k}}_{\text{current manuscript}} + \omega \underbrace{\frac{dJ_{t+1}}{dd_k}}_{\text{future judgement}}. \tag{5}$$

This derivative identifies a marginal trade-off without ranking complete workflows or proving that a single sequence is globally optimal. Where delegating more of a task reduces learning, doing so is worthwhile only if the immediate gain to the manuscript outweighs the discounted loss in future capability. This trade-off yields the second prediction: delegating linguistic execution can raise assisted manuscript quality today through more checked candidates, but if the workflow also delegates initial formulation and other learning activities do not offset the loss, unaided capability tomorrow can fall.

Because skill acquisition is easier to test in classrooms than among faculty, the clearest empirical evidence on this trade-off comes from educational and experimental studies. Prior to generative AI, a meta-analysis of classroom experiments found that writing practice raised learning by roughly 0.30 standard deviations (Graham, Kiuhara and MacKay, 2020), and a meta-analysis of 86 studies on the 'generation effect' found that actively producing answers instead of reading them improved recall by about 0.40 standard deviations (Bertsch et al., 2007), though that advantage shrinks on more complex tasks. Remembering material is not the same as scholarly judgement, and neither literature directly measures the model's four learning activities, so while active writing builds understanding, an author need not write every sentence unaided.

Recent experiments with generative AI test what happens when the tool is removed. In a preregistered field experiment with nearly 1,000 secondary-school mathematics students, Bastani et al. (2025) compared unrestricted access to GPT-4 with a custom tutor prompted to give step-by-step hints instead of final answers. While the tools were available, practice scores were 48% higher with unrestricted access than in the control group, and 127% higher with the guided tutor. On a subsequent unaided exam in the same session, however, students who had used unrestricted assistance scored 17% below the unassisted control group, whereas the guided tutor largely prevented that drop. Although secondary-school mathematics and a same-session exam do not test delayed transfer in scholarly writing, the experiment cleanly isolates the two-output mechanism

in which higher assisted performance is accompanied by lower unaided capability. A randomised experiment with 1,174 adults solving workplace problems and writing business proposals adds an important qualification (Cruces et al., 2026). Access to generative AI closed roughly three-quarters of the initial performance gap between less- and more-educated participants while the tool was available, and treated participants did not score below the control group overall on a subsequent unaided task, yet lasting human-capital gains emerged only when intensive AI use was paired with sustained human effort.

A similar erosion could in principle occur across a research community. In Acemoglu, Kong and Ozdaglar (2026)'s model, individuals who solve hard problems for themselves generate a public good by adding new solutions to the public stock of knowledge; once AI supplies good answers cheaply, individuals have less private incentive to invest that effort, and under the model's assumptions the stock of human-generated knowledge declines towards zero ('knowledge collapse'). Extending that result to academic writing requires two strong additional premises: that an individual scholar's unaided drafting contributes to the field's shared knowledge base, and that delegating the draft removes that contribution with nothing replacing it.

The second prediction also asks how AI affects cross-sectional gaps between stronger and weaker authors. Whether cheaper candidate generation narrows or widens gaps in *assisted* performance depends on which component of J_t binds. Consider two benchmark cases derived in the supplement for authors who check the same number of candidates drawn from the same distribution. In the first, candidate quality and evaluation noise are normally distributed, and authors with higher J_t make smaller evaluation errors. Better judges gain more from each extra candidate because they are more reliable at selecting the best version from a larger sample, so search and judgement are complements and cheaper drafting *widens* the assisted gap. Experimental evidence on AI-assisted decision making points to a related mechanism: in an incentivised task with 732 participants, Caplin et al. (2026) show that holding baseline ability fixed, better *calibration* (knowing when one's own judgement is more reliable than the AI's) raises the benefit of AI assistance by roughly a fifth of the average effect, and because lower-ability participants are less calibrated, miscalibration limits how much assistance compresses performance gaps. In the second benchmark case, both authors evaluate arguments accurately, but the stronger author writes a high-quality baseline draft while the weaker author writes a poor one, and the language model draws alternatives from a common distribution. Because each author keeps the best draft available (her own or the machine's), the weaker author gains more from the machine's draws, so cheaper candidates weakly *narrow* the assisted gap. The workplace and adult-skill experiments of Noy and Zhang (2023), Brynjolfsson, Li and Raymond (2025), Dell'Acqua et al. (2026), and Cruces et al. (2026) fit this second scenario on tasks within the model's competence, though it remains an open empirical question which benchmark better describes frontier scholarship.

For later *unaided* capability, the gap across authors can move in the opposite direction. Suppose weaker authors learn at least as much as stronger ones from an hour of formulating an argument themselves, delegate at least as large a share of initial formulation d_R, and do not offset the

lost practice through extra revision, verification, or feedback.[2] Under those sufficient conditions, delegating initial formulation leaves the gap in future unaided capability at least as wide as without delegation. Each of these conditions is an explicit behavioural assumption, not a result derived from endogenous delegation choices. In Bastani et al. (2025), students with unrestricted access often requested direct answers instead of explanations, though that pattern does not prove that weaker students delegated more, and an interface that prompts heavy users to revise and verify more intensively could reverse the prediction. The value of an unaided attempt also depends on prior knowledge: instructional aids that help novices lose value and can reverse their effect as expertise rises (Kalyuga et al., 2003), so true beginners often learn more from worked examples than from unguided attempts (Kirschner, Sweller and Clark, 2006), whereas learners with some preparation benefit from attempting a problem before help arrives (Kapur, 2008).

The guided tutor in Bastani et al. (2025) illustrates that whether AI erodes skill or builds it depends on how the workflow is organised. A researcher who would otherwise submit a paper after a single draft might engage in more rigorous revision (r) and absorb more critical feedback (f) by using a language model to identify weaknesses in her argument; across more than 4,800 scientific manuscripts, Liang et al. (2024) found that GPT-4 raised many of the same substantive criticisms as human peer reviewers, and over half of surveyed authors judged its feedback helpful for improving early drafts. Assisted workflows can also create new forms of verification practice (v), such as the code-checking exercise described in Section 3, consistent with survey evidence that generative AI shifts knowledge workers' perceived effort from producing text towards verifying it (Lee et al., 2025). Whether such cross-checking improves later unaided verification has not been tested. Moreover, an accurate summary of flawed code will still miss a conceptual error if the author shares the flaw, and an overly agreeable model asked to critique a weak idea may simply restate it in formal language, leaving the author more confident in a bad argument.

Empirical evidence on how well researchers use AI to verify results is correspondingly mixed. In a 2024 replication study, teams relying exclusively on ChatGPT (GPT-4 and GPT-4o) reproduced published results in only 37% of cases, compared with 94% for unaided teams, whereas teams combining their own work with AI assistance achieved a 91% reproduction rate (statistically indistinguishable from the unaided benchmark) but caught fewer major coding errors (Brodeur et al., 2026). Because that study tested computational reproduction and model capabilities are advancing rapidly, findings from earlier models may understate what newer systems can do, a caveat that applies to both optimistic and pessimistic evidence across this literature. In a non-representative survey of 637 researchers, nearly nine in ten of those who reported saving time with AI spent more than a tenth of the saved time checking its output, and two in five reported a larger backlog of untested hypotheses than three years earlier (Codreanu et al., 2026). The evidence thus supports

[2]Formally, Assumption 2 in the supplement requires common learning and generation functions across authors with productive own generation ($L_g \geq 0$), a marginal learning return to own generation that is weakly higher at lower initial capability ($L_{gJ} \leq 0$), own generation weakly decreasing in formulation delegation ($g'(d_R) \leq 0$), weaker authors delegating at least as much ($d_R^L \geq d_R^H$), and revision, verification, and feedback equal across both authors and held constant along the comparison with zero formulation delegation.

neither uncritical adoption nor prohibition, pointing instead to evaluations of delayed unaided performance and workflows that retain practice in revision and verification.

6 The First Draft as an Anchor and a Default

Even an author with accurate judgement can be influenced by the first draft she sees. In psychology, when participants make numerical estimates, their answers are biased towards an initial starting value even when they know it was generated at random, a phenomenon known as *anchoring* (Tversky and Kahneman, 1974). In problem-solving and creative tasks, seeing an example solution similarly narrows the range of designs people subsequently produce, an effect called *fixation* (Jansson and Smith, 1991; Smith, Ward and Schumacher, 1993). When a researcher asks a language model for an opening draft before formulating her own account, that draft may shape both the alternatives she generates and the standard by which she judges them.

To formalise how this bias interacts with judgement, suppose the author assesses each candidate argument x_i according to

$$\widetilde{q}_i = q_i + \alpha(J_t, u_t)s(x_i, x_0) + \varepsilon_i, \qquad \alpha_J < 0, \quad \alpha_u > 0. \tag{6}$$

Here q_i is true scientific quality, $\widetilde{q}_i$ is perceived quality, ε_i is random evaluation noise, $s(x_i, x_0)$ measures how closely candidate x_i resembles the generated first draft x_0, and α is the weight placed on that resemblance (corresponding to the ranking distortion $b(d_R, J_t)$ in Section 5). The anchoring bias increases when the author's domain judgement J_t is weaker and when her uncertainty u_t about the problem is greater. If the model's opening draft sets out the standard explanation in the literature, the author assigns higher evaluations to subsequent framings that resemble that initial account.

Evidence from psychology and economics supports these assumptions. Across eleven preregistered estimation experiments, anchoring grew stronger when larger magnitudes widened the range of plausible answers (Lee and Morewedge, 2022), and in related studies anchoring grew weaker among participants with greater domain knowledge or relevant background facts (Wilson et al., 1996; Smith, Windschitl and Bruchmann, 2013). Domain expertise does not eliminate the bias, however. In mock-courtroom sentencing studies, experienced criminal-law judges and prosecutors were just as influenced as other legal professionals by a prosecutor's sentencing demand, including demands they were told had been generated at random, and their subjective confidence was unrelated to the size of the bias (Englich, Mussweiler and Strack, 2006). Uncertainty u_t therefore represents a wide range of plausible judgements rather than low subjective confidence. An initial reference point is not always harmful, either, because when absolute quality is difficult to judge in isolation, a benchmark can provide information and make local comparisons more precise (Dean et al., 2026).

Research on fixation and AI-assisted writing shows a similar pattern. In design experiments, engineers copied structural features of sample solutions, including flaws they had been warned to

avoid (Jansson and Smith, 1991), and brainstorming participants reproduced traits of examples they had just seen (Smith, Ward and Schumacher, 1993). In two recent experiments with 2,582 participants, Williams-Ceci et al. (2026) gave some writers an autocomplete assistant that subtly favoured one side of a policy issue. Compared with unaided writers, participants using the biased assistant shifted their stated attitudes towards the model's position by about 0.4 points on a five-point scale, a larger shift than occurred when the same arguments were presented as static reading material, and neither advance warnings nor post-task debriefings measurably reduced the shift. Although participants were writing a short opinion essay rather than ranking research hypotheses, the experiment shows that suggested prose can shift writers towards a conclusion without their noticing. Until tested directly on scholars, this mechanism remains suggestive.

In the formal derivation, systematic anchoring bias behaves very differently from random evaluation error. When candidate quality and evaluation errors are independent and normally distributed ($\widetilde{q}_i = q_i + \varepsilon_i$, with no preference for the first draft), inspecting more candidates still raises the expected *true* quality of the selected draft: the winning candidate tends to have a favourable error draw so the author overstates its quality (Smith and Winkler, 2006), yet expected true quality rises with n, and rises faster for more accurate judges (Section 5). Where the author holds a systematic preference for the initial framing, by contrast, more search can make the chosen argument worse. In an example where original accounts have bounded quality and conventional accounts receive a similarity bonus large enough to rank above any original framing despite fitting the evidence less well on average, generating more candidates raises the probability that a conventional account enters the pool and is selected, so expected true quality can rise over the first few draws and then fall. Because this loss comes entirely from the distorted selection rule, Section 5 restricts the claim that more candidates raise quality to unbiased ranking.

Work on narrative persuasion helps explain why a coherent first draft can have such a strong effect. In experiments by Barron and Fries (2025), advisers shifted decision makers' beliefs by supplying causal stories that connected data points both parties already knew; narratives that fitted the facts cleanly were more persuasive, and asking recipients to think through the data privately beforehand produced only a small, statistically insignificant reduction in persuasion, whereas explicit warnings about bias and incentive disclosures reduced persuasion substantially. A language model has no personal stake to disclose, yet its training favours common explanations in the literature, and the experiment shows how easily a coherent narrative can change beliefs without adding new facts.

Beyond biasing how an author evaluates alternatives, a generated first draft also changes what happens if she stops working. Before language models, producing an initial draft required effort, so a scholar had to write at least one rough version of her own to have a manuscript. Today, a single prompt produces a fluent draft at negligible cost, making that text the *default* outcome unless the author actively replaces it.

This shift operates through two stopping mechanisms, *satisficing* and *present bias*. A satisficing writer searches sequentially and stops at the first draft that meets an acceptable threshold (Simon,

1955), a rule that describes most participants in laboratory experiments with costly search (Caplin, Dean and Martin, 2011). When a language model produces a serviceable draft on the first try, a satisficing author may accept it instead of formulating an independent account. Broader evidence confirms the effect of defaults: automatic enrolment raised retirement-plan participation by tens of percentage points (Madrian and Shea, 2001), and a meta-analysis of 58 datasets found that defaults shifted choices by roughly two-thirds of a standard deviation on average (Jachimowicz et al., 2019), though the authors propose without testing that defaults matter less when decision makers already hold strong preferences.

Present bias supplies a second reason to stop early. People routinely overweight immediate effort relative to future rewards, especially when the immediate cost is cognitive or physical work (Augenblick, Niederle and Sprenger, 2015). Because revising a generated draft requires effort today while the payoff of a stronger contribution and sharper personal judgement arrives months later, a present-biased scholar may accept a draft she knows is imperfect. Stopping after fewer drafts is not automatically a mistake, since in a rational search model a stronger initial draft lowers the expected gain from further search and justifies stopping sooner (Weitzman, 1979), and having more alternatives under constant search costs does not by itself make a rational searcher accept a default (de Lara and Dean, 2025). What changes the structure of search here is that a language model places a passable draft at the start of the sequence before the author has invested any effort of her own.

The behavioural prediction is specifically about *excess* stopping relative to the author's own standard for patience. Suppose generating and checking the next candidate requires immediate effort $e = c_v + c_s$ and yields an expected future quality gain Δq_m. A patient author continues whenever $\Delta q_m > e$, whereas a present-biased author with $0 < \beta < 1$ continues only if $\beta\,\Delta q_m > e$. Whenever diminishing returns bring the marginal gain into the range $e < \Delta q_m \leq e/\beta$, the present-biased writer stops strictly earlier on that step than she herself would judge worthwhile from a distance. Anchoring and default effects can thus shape two distinct stages of the workflow: anchoring can distort *which* draft is selected among those examined, while satisficing and present bias can reduce *how many* independent drafts the author examines.

7 The Smithian Bargain

Moving from the individual writer's workflow to the institutions of scholarship raises a broader economic question: when a new division of labour removes learning from everyday work, how does society rebuild that skill? In Book I of *The Wealth of Nations*, Smith (1776) celebrated the division of labour for raising productivity by saving time between tasks and facilitating specialised machinery. Yet in Book V, he warned that a worker whose whole life is spent performing a few simple operations has no occasion to exercise his understanding and 'generally becomes as stupid and ignorant as it is possible for a human creature to become', losing the habit of sound judgement in ordinary life. Smith's remedy was public education.

I call the tension between those two outcomes the *Smithian bargain*. Specialisation raises material output while narrowing the formative scope of work itself, leaving society dependent on educational institutions to supply capabilities that daily labour no longer builds. Where West (1964) saw Book I and Book V as a contradiction, Rosenberg (1965) argued that they form a coherent whole because narrower tasks make formal education necessary and higher productivity makes it affordable. Rosenberg's reading captures the challenge of AI-assisted scholarship, which concerns both which skills are threatened by assistance and what institutional practices can cultivate them anew.

History offers both reassurance and a warning. Predictions of broad technological harm are common historically and have repeatedly proved overstated; as Mokyr, Vickers and Ziebarth (2015) document, predictions of permanent economy-wide technological unemployment have failed generation after generation. Losing part of the traditional craft of drafting will not end academic inquiry, and a discipline could publish more papers than ever even as its practitioners lose analytical habits that the older workflow forced them to develop. Useful historical comparisons must therefore look beyond aggregate output to ask which specific human capabilities are at risk and how institutions support adjustment.

Writing with a language model is a division of intellectual labour. By taking over the routine production of prose and offering rapid line-level criticism, a model can free the scholar to focus on primary sources and research design, the productivity gain Smith described. If that same delegation removes the formative practice through which researchers learn to build and test arguments, however, it carries the risk of deskilling he identified. Just as early twentieth-century electrification automated routine manual tasks while raising the demand for educated workers (Goldin and Katz, 1998), modern task-based models emphasise that machines typically replace rule-bound work while raising the value of human expertise on the tasks that remain (Autor, 2015; Autor and Thompson, 2025). Because a single technology can automate some tasks while making others more valuable, the analysis must proceed task by task.

Comparisons with earlier writing technologies clarify what is new. The printing press sharply reduced the cost of reproducing texts, and European cities that adopted it early grew faster than comparable neighbours (Dittmar, 2011). Centuries later, the word processor made it cheap to revise text and rearrange passages without retyping. Whether mechanical tools shaped the writing itself is unclear, though using a typewriter in 1882 led Nietzsche to observe that 'our writing tools are also working on our thoughts' (Kittler, 1999).

Yet neither print nor the word processor ever proposed the next paragraph or offered a causal explanation of a result. A language model does both. As a service, that is not wholly new, since a ghostwriter or a research assistant drafting part of a paper from brief instructions could perform similar work. Cost and attribution norms meant such assistance remained limited: ghostwriters provide a costly personal service, while academic colleagues and research assistants receive credit through co-authorship or acknowledgement. Crucially, when an assistant drafted a section, the practice of writing stayed inside the profession because the assistant learnt by writing it. Language

models alter the Smithian bargain by making that service cheap and accessible, while displacing the practice of drafting to no human successor at all.

If models come to draft and judge well, the labour-market return to acquiring those skills may also fall. When American telephone companies replaced manual switchboards with mechanical switching between 1920 and 1940, incumbent operators were significantly more likely a decade later to work in lower-paying occupations or to have left the labour force, whereas young women entering the labour market after the switch found other work without lower overall employment (Feigenbaum and Gross, 2024). Similarly, after the release of ChatGPT, freelancers in writing-related work on a large online platform reported fewer jobs and lower earnings, with no protection from their record of previous work (Hui, Reshef and Zhou, 2024). If machine capabilities continue to advance, an author who invests less time acquiring these skills may be making a privately sensible choice.

Two considerations qualify this conclusion, however. First, someone must still check the model's work, which requires judgement, and drafting is one of the ways authors acquire that judgement. Second, Smith's case for education never rested on wages alone; he wanted citizens to be able to judge for themselves in private life and public affairs, and that value survives any fall in market earnings.

8 Who Gains and Who Learns

The search model in Section 4 predicts that when a technology lowers the cost of producing fluent English, the largest immediate gains go to authors for whom writing in English is most costly, above all second-language researchers. Large bibliometric datasets and new evidence from economics working papers are consistent with that pattern, though measurement limits prevent them from showing whether authors selected stronger arguments in addition to fluent prose.

The baseline cost of writing in English varies sharply across the world. In an eight-country survey, Amano et al. (2023) find that non-native speakers spend substantially more time reading and writing in English and face more language-related rejections and revision requests than native speakers; among respondents with one English-language publication, those from countries with moderate or low English proficiency were about two and a half times as likely to report a rejection attributed to their English. Because professional human editing is priced in rich-country currencies, language barriers shape whose research is published regardless of scientific merit. Generative AI can reduce that editing burden, though subscription fees and limited digital infrastructure still restrict access in parts of the developing world, particularly across Africa (Okolo, Aruleba and Obaido, 2023).

Since the arrival of large language models, papers posted on preprint repositories have shifted rapidly. Using word patterns to identify probable AI assistance across more than two million preprints on arXiv, bioRxiv, and SSRN, Kusumegi et al. (2025) compare likely adopters with matched non-adopters and estimate that monthly output rose after adoption by 36.2% on arXiv,

52.9% on bioRxiv, and 59.8% on SSRN. The largest estimated gains appear among scholars classified by name as Asian and based at Asian institutions (up 89.3% on bioRxiv and 88.9% on SSRN, compared with 23.7% to 46.2% for those classified as Caucasian and based in English-speaking countries). Although these observational associations fit the search model's pattern, three measurement problems require care. First, in Liang et al. (2023), early open-source detectors misclassified a majority of essays by non-native English speakers as machine-generated because their wording was more statistically predictable; while commercial detectors have since achieved much lower false-positive rates on benchmark corpora (Jabarian and Imas, 2025), the classifiers used in large preprint studies have not been audited specifically for second-language bias, so part of the apparent gap across language groups could reflect differential detection error. Second, name-based classifications bundle language costs with cross-national differences in field mix and institutional resources. Third, adoption is dated to the first month containing a flagged paper, so months with more output are mechanically more likely to count as adopted; placebo flags produce a similar post-adoption rise (Renault, Bergeaud and Bosquet, 2026).

Economics working papers indicate a similar shift. I assembled 42,704 working papers issued between January 2015 and September 2026 across 32 series, comprising 31 series hosted by universities and policy institutions (including central banks) in North America, Europe, Japan, Australia, Chile, and South Africa, plus the economics collection of arXiv. Two textual indicators track AI-like prose style as a proxy for generative assistance in writing. The first applies Goldsmith-Pinkham (2026)'s detector, trained specifically on economics prose, to selected passages from each paper's introduction and one body section; in a comparison of 2026 NBER working papers, this simple detector flagged roughly a third of the papers identified by a commercial detector, so it likely misses much AI-assisted writing. The second indicator flags papers whose colon rate exceeds a fixed threshold of about 8.9 colons per 1,000 words, set at two standard deviations above the 2015–2021 average in an earlier calibration sample (exact threshold 8.9055).

Before 2023, the economics detector flagged fewer than one working paper in 200; by the first half of 2026, it flagged 15.6% (Figure 1, panel A). Over the same span, the share of papers with a high colon rate rose from around 4% to 15.2% in July–September 2026, whereas the detector flag dipped to 9.0% in that final partial period for reasons these data do not reveal. Neither stylistic indicator maps closely onto formal disclosures of AI use: across the 13,103 papers issued since 2023 on which all three measures are available, the correlation between a detector flag and a disclosure of use is just 0.07.

If authors adopt assistance more readily when it gains them more, and the detector treats groups alike, the first prediction suggests that AI-like prose should appear more often among authors facing higher English writing costs. Tracking authors' institutional affiliations across 6,939 working papers issued since January 2023 and assigning each institution to its country,[3] I use an affiliation outside

[3]International organisations and research networks that are not primary employers (such as the NBER and CEPR) were set aside, along with unidentifiable listings. Where an author's record matched known institutions, all were retained; otherwise the printed affiliation was coded by the first institution named. Two AI classifiers (Claude and Codex) agreed on 98.5% of the 1,838 distinct affiliation strings, and the remaining 27 were resolved from the printed

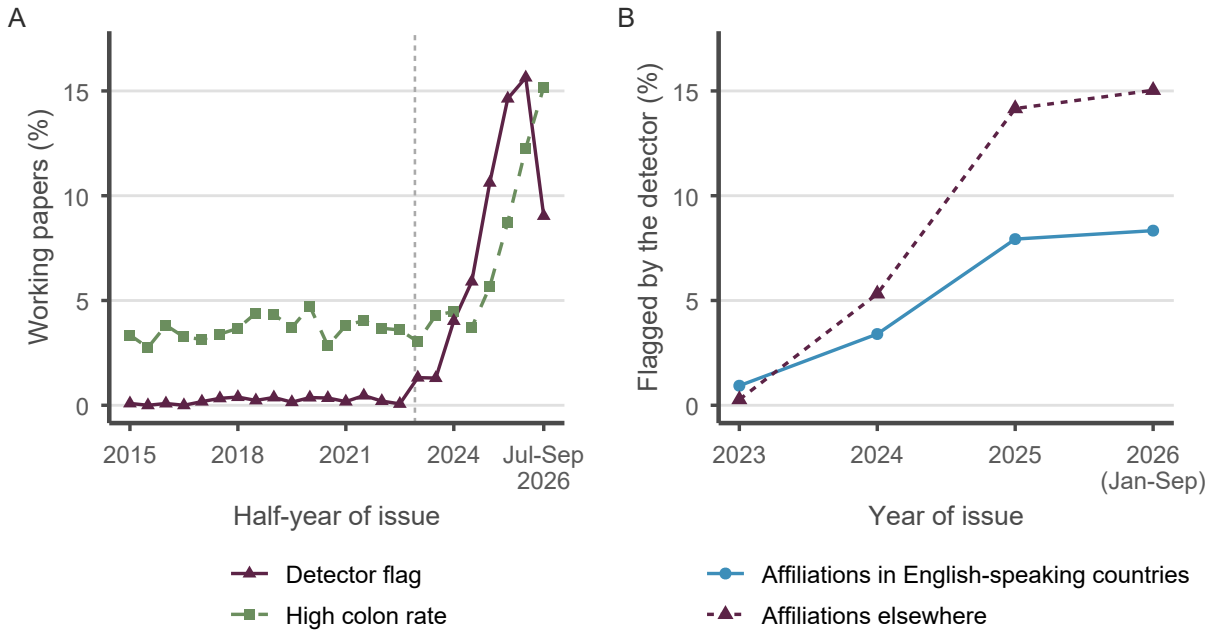


Figure 1: AI-like prose in economics working papers, 2015–2026. Panel A: shares of 42,704 working papers from 32 series flagged by the detector or with a high colon rate, by half-year of issue; each indicator uses the papers on which it can be measured, and the last period covers July–September 2026. The dotted line marks the release of ChatGPT. Panel B: shares flagged by the detector among papers issued since 2023 whose classified affiliations lie entirely in English-speaking countries (3,703) or entirely elsewhere (1,427); papers with affiliations on both sides (1,809) are omitted from the panel but included in the estimates in the text. 2026 covers January–September.

the United States, the United Kingdom, Ireland, Canada, Australia, and New Zealand as a rough proxy for writing in a second language. The detector flagged 8.3% of papers whose classified affiliations all lie outside those six countries, compared with 5.0% of papers whose affiliations all lie within them. Within the same series and half-year, the difference is 2.8 percentage points (95% confidence interval 1.5 to 4.1), and 2.0 points (0.5 to 3.6) excluding arXiv.

As panel B of Figure 1 shows, detector flags were rare in both groups in 2023, before a gap opened in 2024 and widened in 2025. With so few flags in 2023, however, that timing cannot prove that the detector treats second-language prose fairly (Liang et al., 2023). The colon indicator's gap between the two country groups was already present in 2023, consistent with pre-existing differences in writing style or paper mix, and because 2023 is not a pre-AI period, the comparison cannot separate baseline differences from early AI adoption. Overall, the working-paper pattern is consistent with the first prediction, though like Kusumegi et al. (2025) and Liang et al. (2023)

context. The within-series differences reported in the text come from linear probability models with series × half-year fixed effects and standard errors clustered by series, estimated on the 6,928 papers (5,393 excluding arXiv) that remain once singleton cells are dropped; papers with mixed affiliations enter with their own indicator (English-only omitted), and 95% confidence intervals use t_{G-1} critical values with 32 clusters (31 excluding arXiv).

it measures stylistic resemblance instead of argument quality, covers only papers with classifiable affiliations, and uses institutional country as a coarse proxy for English proficiency.

Published journal articles show a similar pattern. Analysing 5.2 million papers across 5,114 journals, He and Bu (2026) estimate that AI-assisted writing rose in every discipline after ChatGPT's release and grew fastest among authors based in non-English-speaking countries.

Equalising access to fluent prose need not equalise underlying capability. An experienced scholar may use a model to express an argument she has already worked out, whereas a less prepared writer may ask for the whole account and be less able to catch plausible errors, a contrast that fits the second prediction. Although untested in faculty research, the closest evidence comes from doctoral training. Applying a commercial detector to the dissertation abstracts of nearly all US STEM PhD graduates from 2019 to 2026, Gross, Shvadron and Zhang (2026) find that through May 2026 the classifier flags machine-written text in 41% of abstracts by students educated in non-English-speaking countries, against 23% for those educated in English-speaking countries. Because the detector flagged almost no abstracts from either group before 2023, baseline bias against second-language writers is unlikely to explain that gap.[4]

Lowering language barriers can also encourage stylistic conformity. In an experiment on stylistic convergence (separate from first-draft anchoring), Agarwal, Naaman and Vashistha (2025) find that suggestions from a Western-trained model shifted Indian participants' writing towards Western conventions and reduced culturally specific content. Because large bibliometric studies follow researchers who already publish, entry by previously excluded scholars remains unmeasured, leaving open how far cheaper English broadens participation versus standardises expression. In second-language writing pedagogy, Hyland (2026) similarly concludes from pedagogical principles that generative models supply valuable linguistic support while risking deskilling and deeper inequalities. Research policy must therefore weigh both who can produce an acceptable paper with assistance today and who builds the capability to judge the next one unaided tomorrow, lowering a native-speaker advantage unrelated to scientific merit without removing the practice through which judgement develops.

9 Readers and the Field

Writing continues to matter after a paper is finished because the way an argument is expressed shapes what readers believe and remember (McCloskey, 1983). In experiments by Graeber, Roth and Zimmermann (2024), statistical tables shifted beliefs more (and more quickly) than stories immediately, yet one day later statistics had lost about 73% of their initial effect, compared with 32% for stories, and narrative details that added no factual information still aided recall. Examining more than 6,900 spoken explanations of financial choices, Graeber, Roth and Schesch (2026)

[4] Among US-origin graduates of R1 universities in the 2023–2025 cohorts, those with flagged dissertation abstracts are 7.4 percentage points less likely to hold observed academic employment in the first half of 2026 (Gross, Shvadron and Zhang, 2026). This is an observational association with employment rather than a measure of unaided capability: it could reflect career plans affecting AI use, effects of AI use on training, or other differences between students, and the study does not separate them.

similarly found that explanations helped listeners learn on average because correct answers tended to come with richer reasoning, yet listeners were also more likely to follow richer explanations when independent assessments found little difference in the underlying arguments. Persuasive expression can thus make a claim more influential without its being more sound.

This distinction becomes critical when a language model can polish prose without improving the research beneath it. Good scholarly prose both helps readers follow an argument and, when costly to produce, signals the author's command of a subject (Spence, 1973). In Gans (2024)'s generative-communication model, making messages cheaper to write and read can improve direct comprehension while making high-quality messages harder to distinguish from weak ones. Once anyone can generate smooth sentences at negligible cost, fluency still helps readers follow a paper while telling them much less about the rigour of the work.

This mechanism yields the fourth prediction, concerning how readers infer quality as the signal degrades. Evaluators who are slow to adjust and continue treating fluent prose as a signal of strong evidence will overweight surface style. Yet their ability to separate strong papers from weak ones can fall even if they update their expectations optimally, simply because a formerly informative signal has lost precision; lower screening accuracy therefore says little on its own about whether readers overweight style. Weakening the association between fluency and quality is also very different from reversing it, which would mean that better-written papers contain worse evidence. In the model here, the cost of writing in English is unrelated to the quality of an author's evidence, so researchers who adopt assistance to overcome high language costs hold evidence of average quality. Where awkward English previously concealed strong qualifications among non-native speakers, language assistance makes their merit easier to recognise, and under the model's regularity conditions polishing their prose dilutes the overall link between fluency and research quality without turning it negative.[5]

Direct evidence on screening comes from human evaluators. In hiring and start-up investment experiments, Cowgill, Hernández-Lagos and Wright (2026) find that access to ChatGPT reduced screening accuracy by 4% to 9% overall as polished prose became a weaker signal of applicant quality, yet screening accuracy for applicants from non-English-speaking countries *improved*, suggesting that their merit had previously been masked by imperfect English. Galdin and Silbert (2025) find a parallel result in platform data on job applications, where cover letters lost much of their signalling value once they became cheap to generate, and the authors' structural model predicts that hiring would then depend less on applicant merit.

[5]Formally, Proposition 4 in the supplement establishes an attenuation result for pooled covariance: if mean research quality is equal across adoption groups, group-specific moments remain fixed as the assisted share rises, and the within-group covariance between fluency and quality is non-negative and no larger among assisted papers than among unassisted ones ($0 \leq \kappa_1 \leq \kappa_0$), then a higher share of assisted papers weakly reduces pooled covariance without turning it negative. Because the variance of fluency across papers also changes when assistance spreads, pooled covariance can fall even while the correlation coefficient holds steady. Reversing the sign of pooled covariance would require either a negative within-group covariance or a sufficiently negative between-group term (for instance, if adoption concentrated heavily among authors with the weakest underlying research rather than those facing the highest language costs). How much readers learn from prose, or how accurately editors screen submissions, then depends on how evaluators use the remaining clues.

Making writing artificially difficult again would be the wrong remedy, since clear prose helps readers understand and remember an argument. Instead, editors and referees should place less weight on surface fluency and more on the substance of the claims and evidence. Even within a manuscript, some choices are harder to delegate than others: producing grammatical sentences is very different from deciding which empirical results deserve primary emphasis. A paper that devotes pages to minor tables while barely explaining its main finding suggests that its author has not exercised scientific judgement. This aligns with recent economic work identifying the ability to judge which claims are true and which questions are worth pursuing as the binding constraint on AI-assisted science (Agrawal, McHale and Oettl, 2026).

Cheaper fluency may also affect the political economy of journal rules. Where academic institutions reward scarce writing skill, scholars who already write effortlessly lose part of their relative advantage when polished prose becomes cheap. Groups whose skills or rents are threatened by new technologies have historically resisted their adoption (Mokyr, 1992); a similar incentive could encourage resistance to AI writing tools quite apart from concerns about student learning. Whether this incentive influences current editorial policies is an untested hypothesis (predicting stronger support for restrictions where stylistic polish is most rewarded), attributes no motive to any individual critic, and leaves the human-capital objection to be judged on its empirical merits. It does clarify what rules should target: asking an author to state and defend her own formulation of a problem protects learning directly, whereas demanding prose that looks unassisted penalises surface style and disproportionately affects second-language writers (Section 8).

At the level of the discipline as a whole, generative assistance can raise the assessed creativity of an individual's work while making different people's outputs more alike (Doshi and Hauser, 2024). Field experiments with consultants (Dell'Acqua et al., 2026) and broader reviews of AI-assisted text (Sourati, Ziabari and Dehghani, 2026) document similar convergence in phrasing and reasoning, while textual comparisons find that ChatGPT essays are at least as clearly organised as British undergraduates' essays but contain far fewer markers of how strongly a claim is held (Jiang and Hyland, 2025). Whether scholarly papers share the same substantive omissions is a stronger claim requiring both that authors start from a common flawed draft and that subsequent revision fails to catch the flaw. Models of 'algorithmic monoculture' in automated hiring show how reliance on a shared evaluator can make each decision maker more accurate individually while worsening collective outcomes through correlated mistakes (Kleinberg and Raghavan, 2021). In science, where an unconvincing explanation today may be vindicated by later data, maintaining diverse rival explanations has real value, especially since large fields already concentrate attention on a small canon of established interpretations (Chu and Evans, 2021).

The fifth claim (that scholars could publish more papers while the field explores a narrower range of questions and explanations) therefore remains a conjecture for future research. A clean test would give researchers the same body of evidence and compare those assigned to start from a generated draft with those who formulate their own account first, measuring whether the assisted papers converge on fewer explanations and repeat the same omissions. Existing empirical studies

of AI and scientific novelty answer a different question because they identify papers using machine-learning research methods rather than generative writing tools: Hao et al. (2026) find that scientists using AI methods span a narrower range of topics across a field, whereas Bianchini et al. (2026) and Ding, Lawson and Shapira (2026) find more novel terms or unusual citation combinations within individual AI-assisted papers, findings that can hold simultaneously because within-paper novelty is compatible with narrower field-wide exploration.

Finally, testing more candidate drafts within one project must be distinguished from submitting more manuscripts for publication. Proposition 1 models an author who compares several versions privately before submitting her best paper; whether the field benefits from a surge in total submissions is a separate question. At *Organization Science*, submissions between January 2024 and November 2025 were 42% higher than in the same months two years earlier, the journal roughly doubled its roster of senior editors to screen the inflow, and a detector classified most abstracts submitted in early 2026 as partly AI-written (Gartenberg et al., 2026). Because referee and reader attention is scarce, a large increase in submissions can make strong work harder to find even if average quality holds steady, much as the expansion of print in early modern Europe required new indexes and reference works to filter material (Blair, 2010). Strict verification procedures and less editorial weight on surface fluency can help a field secure the search gains of cheap generation by publishing only carefully verified papers.

10 A Boundary for Use

The models and evidence above point to practical proposals for academic workflows and evaluation; because each rests on theoretical benchmarks and evidence from adjacent domains, each is a hypothesis worth testing.

For individual authors, the analysis suggests a clear sequence. Instead of prompting a model for a polished draft at the outset, the researcher first writes down her question, her main claim, the causal links connecting the evidence to that claim, and what remains uncertain. That rough sketch provides an independent baseline to compare against a generated draft. She can then prompt the model for rival explanations and distinguishing evidence, checking those possibilities directly against primary sources and data before seeing how the model ranks them.

Whether this self-imposed sequence changes researcher behaviour is an open empirical question. Writing an unaided rough account takes effort, and although people who recognise their own present bias sometimes welcome binding commitments (Augenblick, Niederle and Sprenger, 2015), a private workflow rule leaves the author free to abandon her initial framework at any time. Moreover, as Section 6 noted, experimental evidence on whether advance reasoning or warnings prevent anchoring and persuasion is mixed (Buçinca, Malaya and Gajos, 2021; Barron and Fries, 2025; Williams-Ceci et al., 2026). Structurally, writing an initial account oneself works like an 'active-choice' rule in retirement saving, where participants must state a decision of their own instead of passively accepting a default. Carroll et al. (2009) show that active choice outperforms a default when

procrastination is strong and decision makers understand their options, whereas a default can be superior when they lack the knowledge needed to choose well.

Where learning and training are the goal, assistance should therefore follow an independent attempt for prepared learners, while true novices may still need worked examples or structured guidance first (Section 5). Offering criticism of a student's draft after she has tried the problem herself retains far more of the formative practice than supplying a finished answer at the start. Curriculum assessment must then measure what learners can do *without* assistance after a delay: although retrieving knowledge from memory strengthens retention (Roediger and Karpicke, 2006), sound evaluation must separate repeating a familiar procedure from transferring understanding to a new problem (Barnett and Ceci, 2002), since a strong essay produced with AI does not show that its author could reason through the problem unaided.

Among journals, mandatory disclosure has been the primary policy response. By early 2025, roughly 70% of the 5,114 journals surveyed by He and Bu (2026) had adopted an AI policy, usually requiring authors to report their use, though existing rules often leave the threshold for reportable assistance unclear (Jabarian and Imas, 2025). In practice, these policies have had little visible effect: growth in estimated assistance was statistically indistinguishable between journals with and without them (He and Bu, 2026).Because estimated adoption varies widely across disciplines (growing fastest in the physical sciences), a single uniform rule may also fit poorly across fields. Moreover, explicit disclosures remain rare and barely correlate with detected AI-like prose. Across roughly 75,000 full-text papers published since 2023, only about 0.1% disclosed AI use in writing, or roughly one disclosure for every forty flagged papers (He and Bu, 2026). In the economics working papers of Section 8, 6.3% disclosed AI use against 8.4% flagged by the detector, and as noted earlier their correlation is 0.07. Neither measure counts AI use directly, and a disclosure can cover tasks other than writing, so these data cannot show what share of actual use authors report.

The signalling logic of this paper exposes the fundamental limit of disclosure mandates. Automated detection of hybrid human–AI text involves an unavoidable trade-off between false accusations and missed use (Jabarian and Imas, 2025), so an unverifiable disclosure rests on professional honesty rather than independent verification. By that standard, the disclosure attached to this paper is not a proof but a promise. If newer models no longer exhibit the stylistic patterns that current detectors flag and light AI editing becomes routine, a generic note that AI was used may tell readers as little as a note that the author ran a spell-checker. Instead of relying on unverifiable disclosures or treating fluent English as a measure of rigour, journals and editors can scrutinise the link between claims and evidence directly by insisting on explicit identification assumptions, precise citations, verifiable data, and reproducible code, for which the human author remains wholly accountable.

Empirical research on scholarly writing should now compare alternative workflows over time, measuring both what authors produce and what they learn. The two components of research-specific human capital J can be tested separately in a longitudinal experiment. To measure *evaluation*, for example, give participants a shared packet of evidence and ask them to rank candidate accounts of

known quality; to measure *formulation*, give them fresh evidence and ask them to write an account of their own. Administering both tasks without AI before the workflow intervention and again after a delay, on unfamiliar problems with language held constant and raters blind to treatment, can distinguish these two capacities far more cleanly than a single essay score. Although ranking arguments still draws on subject knowledge and formulating an account still requires language, together they test the working hypothesis that verification and revision train evaluative judgement while independent drafting trains the ability to form new arguments.

11 Conclusion

The worry that relying on an external tool might weaken human thought is older than language models. In Plato's *Phaedrus*, when the god Theuth presents the art of writing to King Thamus as an aid to memory and wisdom, Thamus replies that written records will breed forgetfulness and give people the appearance of wisdom without true understanding (Plato, ca. 370 BCE, 274c–275b). Thamus had a point: writing did reduce the need to memorise long narratives. Yet by freeing the mind from rote memorisation, written records also enabled scholars to build and test complex arguments that oral memory alone could never have sustained. Losing practice in one mental faculty was accompanied by gains in another.

In this paper, I have argued that writing can contribute to thinking without being identical to it. Much of the current debate treats writing as a single undifferentiated act, asking either whether AI improves the finished text (Noy and Zhang, 2023; Brynjolfsson, Li and Raymond, 2025) or whether the writer still learns by producing it (Sparrow and Flenady, 2025; Corbin et al., 2026). Both claims can hold at once because a better manuscript today does not guarantee that its author has learnt more for tomorrow. Separating linguistic execution from the initial formulation of an argument makes it possible to reconcile these findings across the paper's economic benchmarks. Cheaper drafting brings the largest immediate search gains to researchers facing the highest language costs, a pattern consistent with evidence from preprints, published articles, and economics working papers (Kusumegi et al., 2025; He and Bu, 2026), though none of these sources measures argument quality. Treating judgement as human capital, however, warns that combining delegated prose with delegated initial formulation can raise assisted output today while lowering unaided capability tomorrow, as classroom and adult learning experiments show (Bastani et al., 2025; Cruces et al., 2026). Economists since Smith (1776) have recognised that the division of labour can raise output while reducing what workers learn on the job. In scholarly writing, delegating the draft can remove the very practice that builds the judgement needed to evaluate the finished paper.

The same distinctions apply to how drafts are evaluated and how research is communicated. A machine-generated first draft can act as both an anchor and a default, biasing how an author ranks rival explanations while offering a passable text that present bias may tempt her to accept too soon. On the reader's side, cheap fluency makes polished prose a weaker signal of underlying research quality (Cowgill, Hernández-Lagos and Wright, 2026) without reversing the sign of

the association. Widespread reliance on shared generative drafts could also narrow the range of explanations a discipline explores. These predictions point to a testable sequence for experienced scholars: formulate a rough account of one's own first, then use the model for language assistance and critical challenge while checking every claim against primary evidence. Ideally, because the polish of an assisted manuscript cannot reveal what its author has learnt, longitudinal experiments that track workflows over time are needed to test these predictions.

Return one last time to the economic historian in the archive with her awkward source. If she writes down her own tentative explanation first, a language model can help her sharpen the prose and test the logic at a fraction of the old effort. If she reads the model's smooth, familiar account first, that anomalous record may never prompt her to question the conventional story. Looking at the finished manuscript alone, no reader could tell those two workflows apart.

Nor is the division of labour proposed here fixed. Language models continue to improve rapidly, and both formulating initial accounts from evidence and checking empirical claims may become cheaper to delegate in the future. The point of an economic framework is not to freeze today's boundary between human and machine, but to trace how that boundary shifts as capabilities improve. Making one task cheaper simply moves the bottleneck elsewhere. Verifying machine output still demands human judgement, and building that judgement requires practice. At the same time, readers still bear the cost of identifying reliable research amid abundant fluent prose. As models take over more of the work, the appropriate division of labour will shift with them. The opportunity is to delegate the routine work of expression while keeping the practice that builds human judgement.

Data availability

Section 8 reports indicators of AI-like prose and of disclosed AI use for 42,704 economics working papers issued between January 2015 and September 2026 in 32 publicly available series, and classified institutional affiliations for 6,939 of the papers issued since 2023. The working papers are available from the institutions that issue them and from arXiv, and are not redistributed. The paper-level indicators and affiliation groups, the classifiers' country codes without affiliation text, and the scripts that reproduce every number in Section 8 and Figure 1 are available at `https://github.com/johanfourieza/research/tree/main/2026/writing`, and contain no author names or email addresses. The formal claims are supported by a mathematical derivation and by an R script that verifies those results numerically; both accompany the submission as supplementary material and are deposited in the same repository.

References

Acemoglu, Daron, Dingwen Kong, and Asuman Ozdaglar. 2026. "AI, Human Cognition and Knowledge Collapse." National Bureau of Economic Research Working Paper 34910.

Agarwal, Dhruv, Mor Naaman, and Aditya Vashistha. 2025. "AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances."

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

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan D. González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel L. White, and Diogo Veríssimo. 2023. "The Manifold Costs of Being a Non-Native English Speaker in Science." *PLoS Biology*, 21(7): e3002184.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over Time: Dynamic Inconsistency in Real Effort Tasks." *Quarterly Journal of Economics*, 130(3): 1067–1115.

Autor, David, and Neil Thompson. 2025. "Expertise." *Journal of the European Economic Association*, 23(4): 1203–1271.

Autor, David H. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives*, 29(3): 3–30.

Barnett, Susan M., and Stephen J. Ceci. 2002. "When and where do we apply what we learn? A taxonomy for far transfer." *Psychological Bulletin*, 128(4): 612–637.

Barron, Kai, and Tilman Fries. 2025. "Narrative Persuasion." WZB Berlin Social Science Center Discussion Paper SP II 2023-301r. Current version: September 18, 2025.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. "Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics." *Proceedings of the National Academy of Sciences*, 122(26): e2422633122.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* New York:Columbia University Press for the National Bureau of Economic Research.

Bereiter, Carl, and Marlene Scardamalia. 1987. *The Psychology of Written Composition.* Hillsdale, NJ:Lawrence Erlbaum Associates.

Bertsch, Sharon, Bryan J. Pesta, Richard Wiscott, and Michael A. McDaniel. 2007. "The Generation Effect: A Meta-Analytic Review." *Memory & Cognition*, 35(2): 201–210.

Bianchini, Stefano, Valentina Di Girolamo, Julien Ravet, and David Arranz. 2026. "AI in science: When and where it makes a difference." *Research Policy*, 55(6): 105478.

Blair, Ann M. 2010. *Too Much to Know: Managing Scholarly Information before the Modern Age.* New Haven:Yale University Press.

Brodeur, Abel, David Valenta, Alexandru Marcoci, Juan P. Aparicio, Derek Mikola, Bruno Barbarioli, Rohan Alexander, Lachlan Deer, Tom Stafford, Lars Vilhuber, et al. 2026. "AI-assisted teams outperform AI-led teams but not human-only teams in assessing research reproducibility in quantitative social science." *Proceedings of the National Academy of Sciences*, 123(22): e2524747123.

Brown, Christina, Supreet Kaur, Geeta Kingdon, and Heather Schofield. 2025. "Cognitive Endurance as Human Capital." *Quarterly Journal of Economics*, 140(2): 943–1002.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *Quarterly Journal of Economics*, 140(2): 889–942.

Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making." *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Caplin, Andrew, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye. 2026. "The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration." *Management Science*, 72(7): 5843–5852.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. "Search and Satisficing." *American Economic Review*, 101(7): 2899–2922.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics*, 124(4): 1639–1674.

Cassar, Lea, and Stephan Meier. 2018. "Nonmonetary Incentives and the Implications of Work as a Source of Meaning." *Journal of Economic Perspectives*, 32(3): 215–238.

Chu, Johan S. G., and James A. Evans. 2021. "Slowed Canonical Progress in Large Fields of Science." *Proceedings of the National Academy of Sciences*, 118(41): e2021636118.

Clark, Andy, and David J. Chalmers. 1998. "The Extended Mind." *Analysis*, 58(1): 7–19.

Codreanu, Mihai, Alex Imas, Juan Mateos-Garcia, Joseph Emmens, Evalyne Muiruri, Arthur Turrell, Julian Jacobs, Atoosa Kasirzadeh, Ana Trišović, Yiyuan Chen, Tanya Rodchenko, Catherine Pollard, Scott Strand, Daniel Rock, Zanna Iscenko, Fabien Curto Millet, Neil Thompson, and James Manyika. 2026. "AI in Science: Early Insights." *arXiv preprint 2609.28504*.

Corbin, Thomas, Jack Walton, Peter Bannister, and Jean-Philippe Deranty. 2026. "On the essay in a time of GenAI." *Educational Philosophy and Theory*, 58(3): 198–210.

Cowgill, Bo, Pablo Hernández-Lagos, and Nataliya Langburd Wright. 2026. "Does AI Cheapen Talk? Theory and Evidence from Global Entrepreneurship and Hiring." *Management Science*. Articles in Advance.

Cruces, Guillermo, Diego Fernández Meijide, Sebastian Galiani, Ramiro H. Gálvez, and María Lombardi. 2026. "Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment." National Bureau of Economic Research Working Paper 34851.

Cruz, Joe. 2026. "I'm a College Professor. Writing Isn't as Important as We Think." *The New York Times, opinion, 29 September*.

Dean, Mark, Benjamin Enke, Thomas Graeber, and Pietro Ortoleva. 2026. "Reference Points as Information." Working paper, March 2026.

de Lara, Lucas, and Mark Dean. 2025. "Rational Choice Overload." Working paper, July 2, 2025.

Dell'Acqua, Fabrizio, Edward McFowland, III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2026. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." *Organization Science*, 37(2): 403–423.

Ding, Liangping, Cornelia Lawson, and Philip Shapira. 2026. "Does Artificial Intelligence Advance Science?" *arXiv preprint 2606.05118*.

Dittmar, Jeremiah E. 2011. "Information Technology and Economic Change: The Impact of the Printing Press." *Quarterly Journal of Economics*, 126(3): 1133–1172.

Doshi, Anil R., and Oliver P. Hauser. 2024. "Generative AI enhances individual creativity but reduces the collective diversity of novel content." *Science Advances*, 10(28): eadn5290.

Englich, Birte, Thomas Mussweiler, and Fritz Strack. 2006. "Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." *Personality and Social Psychology Bulletin*, 32(2): 188–200.

Ericsson, K. Anders, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. "The role of deliberate practice in the acquisition of expert performance." *Psychological Review*, 100(3): 363–406.

Fan, Yizhou, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology*, 56: 489–530.

Feigenbaum, James, and Daniel P. Gross. 2024. "Answering the Call of Automation: How the Labor Market Adjusted to Mechanizing Telephone Operation." *Quarterly Journal of Economics*, 139(3): 1879–1939.

Flower, Linda, and John R. Hayes. 1981. "A Cognitive Process Theory of Writing." *College Composition and Communication*, 32(4): 365–387.

Galdin, Anaïs, and Jesse Silbert. 2025. "Making Talk Cheap: Generative AI and Labor Market Signaling." Working paper. arXiv:2511.08785.

Gans, Joshua S. 2024. "How Will Generative AI Impact Communication?" *Economics Letters*, 242: 111872.

Gans, Joshua S. 2026. "But I Like Doing This! Enjoyable Tasks, Contracting, and Automation." National Bureau of Economic Research Working Paper 35309.

Gartenberg, Claudine, Sharique Hasan, Alex Murray, and Lamar Pierce. 2026. "More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review." *Organization Science*, 37(3): 795–812.

Goldin, Claudia, and Lawrence F. Katz. 1998. "The Origins of Technology-Skill Complementarity." *Quarterly Journal of Economics*, 113(3): 693–732.

Goldsmith-Pinkham, Paul. 2026. "econ-ai-detector." *Software repository.*

Graeber, Thomas, Christopher Roth, and Constantin Schesch. 2026. "Explanations." Department of Economics, University of Zurich ECON Working Paper 490.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann. 2024. "Stories, Statistics, and Memory." *Quarterly Journal of Economics*, 139(4): 2181–2225.

Graham, Steve, Sharlene A. Kiuhara, and Meade MacKay. 2020. "The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis." *Review of Educational Research*, 90(2): 179–226.

Gross, Daniel P., Dror Shvadron, and Hansen Zhang. 2026. "How Heavily Are PhD Students Offloading Their Scientific Writing to AI? Evidence from US STEM PhD Dissertations." National Bureau of Economic Research Working Paper 35859.

Hao, Qianyue, Fengli Xu, Yong Li, and James Evans. 2026. "Artificial Intelligence Tools Expand Scientists' Impact but Contract Science's Focus." *Nature*, 649: 1237–1243.

He, Yongyuan, and Yi Bu. 2026. "Academic Journals' AI Policies Fail to Curb the Surge in AI-Assisted Academic Writing." *Proceedings of the National Academy of Sciences*, 123: e2526734123.

Hui, Xiang, Oren Reshef, and Luofeng Zhou. 2024. "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market." *Organization Science*, 35(6): 1977–1989.

Hyland, Ken. 2026. "Writing in the AI Era: Rethinking Writing, Research and Teaching." *Journal of Second Language Writing*, 72: 101302.

Jabarian, Brian, and Alex Imas. 2025. "Artificial Writing and Automated Detection." National Bureau of Economic Research Working Paper 34223.

Jachimowicz, Jon M., Shannon Duncan, Elke U. Weber, and Eric J. Johnson. 2019. "When and why defaults influence decisions: a meta-analysis of default effects." *Behavioural Public Policy*, 3(2): 159–186.

Jansson, David G., and Steven M. Smith. 1991. "Design fixation." *Design Studies*, 12(1): 3–11.

Jiang, Feng (Kevin), and Ken Hyland. 2025. "Rhetorical Distinctions: Comparing Metadiscourse in Essays by ChatGPT and Students." *English for Specific Purposes*, 79: 17–29.

Kalyuga, Slava, Paul Ayres, Paul Chandler, and John Sweller. 2003. "The expertise reversal effect." *Educational Psychologist*, 38(1): 23–31.

Kapur, Manu. 2008. "Productive failure." *Cognition and Instruction*, 26(3): 379–424.

Kellogg, Ronald T. 1994. *The Psychology of Writing.* New York:Oxford University Press.

Kirschner, Paul A., John Sweller, and Richard E. Clark. 2006. "Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching." *Educational Psychologist*, 41(2): 75–86.

Kittler, Friedrich A. 1999. *Gramophone, Film, Typewriter.* Stanford:Stanford University Press.

Kleinberg, Jon, and Manish Raghavan. 2021. "Algorithmic Monoculture and Social Welfare." *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

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

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.

Lee, Chang-Yuan, and Carey K. Morewedge. 2022. "Noise Increases Anchoring Effects." *Psychological Science*, 33(1): 60–75.

Lee, Hao-Ping (Hank), Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers."

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. "GPT detectors are biased against non-native English writers." *Patterns*, 4(7): 100779.

Liang, Weixin, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel S. Smith, Yian Yin, Daniel A. McFarland, and James Y. Zou. 2024. "Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis." *NEJM AI*, 1(8): AIoa2400196.

Loewenstein, George. 1999. "Because It Is There: The Challenge of Mountaineering... for Utility Theory." *Kyklos*, 52(3): 315–343.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics*, 116(4): 1149–1187.

Maestas, Nicole, Kathleen J. Mullen, David Powell, Till von Wachter, and Jeffrey B. Wenger. 2023. "The Value of Working Conditions in the United States and the Implications for the Structure of Wages." *American Economic Review*, 113(7): 2007–2047.

McCloskey, Donald N. 1983. "The Rhetoric of Economics." *Journal of Economic Literature*, 21(2): 481–517.

McCloskey, Donald N. 1985. "Economical Writing." *Economic Inquiry*, 23(2): 187–222.

Menary, Richard. 2007. "Writing as thinking." *Language Sciences*, 29(5): 621–632.

Mokyr, Joel. 1992. "Technological Inertia in Economic History." *Journal of Economic History*, 52(2): 325–338.

Mokyr, Joel, Chris Vickers, and Nicolas L. Ziebarth. 2015. "The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?" *Journal of Economic Perspectives*, 29(3): 31–50.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the 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.

Okolo, Chinasa T., Kehinde Aruleba, and George Obaido. 2023. "Responsible AI in Africa—Challenges and Opportunities." In *Responsible AI in Africa: Challenges and Opportunities.* , ed. Damian Okaibedi Eke, Kutoma Wakunuma and Simisola Akintoye, 35–64. Cham:Palgrave Macmillan.

Plato. ca. 370 BCE. "Phaedrus." Translated with introduction and commentary by R. Hackforth, Cambridge University Press, 1952.

Ransdell, Sarah E., and C. Michael Levy. 1994. "Writing as Process and Product: The Impact of Tool, Genre, Audience Knowledge, and Writer Expertise." *Computers in Human Behavior*, 10(4): 511–527.

Reddy, Michael J. 1979. "The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language." In *Metaphor and Thought.* , ed. Andrew Ortony, 284–324. Cambridge:Cambridge University Press.

Renault, Thomas, Antonin Bergeaud, and Clément Bosquet. 2026. "Comment on Scientific Production in the Era of Large Language Models: Outcome-Triggered Treatment Timing and Spurious Event-Study Dynamics." arXiv:2605.17979.

Roediger, Henry L., and Jeffrey D. Karpicke. 2006. "Test-enhanced learning: Taking memory tests improves long-term retention." *Psychological Science*, 17(3): 249–255.

Rorty, Richard M., ed. 1967. *The Linguistic Turn: Recent Essays in Philosophical Method.* Chicago:University of Chicago Press.

Rosenberg, Nathan. 1965. "Adam Smith on the division of labour: two views or one?" *Economica*, 32(126): 127–139.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics*, 69(1): 99–118.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations.* London:W. Strahan and T. Cadell.

Smith, Andrew R., Paul D. Windschitl, and Kathryn Bruchmann. 2013. "Knowledge Matters: Anchoring Effects Are Moderated by Knowledge Level." *European Journal of Social Psychology*, 43(1): 97–108.

Smith, James E., and Robert L. Winkler. 2006. "The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis." *Management Science*, 52(3): 311–322.

Smith, Steven M., Thomas B. Ward, and Jay S. Schumacher. 1993. "Constraining effects of examples in a creative generation task." *Memory & Cognition*, 21(6): 837–845.

Sourati, Zhivar, Alireza S. Ziabari, and Morteza Dehghani. 2026. "The Homogenizing Effect of Large Language Models on Human Expression and Thought." *Trends in Cognitive Sciences*. Online ahead of print.

Sparrow, Robert, and Gene Flenady. 2025. "Bullshit universities: the future of automated education." *AI & Society*, 40: 5285–5296.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics*, 87(3): 355–374.

Stephens, Bret. 2026. "I'm Begging You: Never Write With A.I." *The New York Times, opinion, 4 August.*

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under uncertainty: Heuristics and biases." *Science*, 185(4157): 1124–1131.

Weitzman, Martin L. 1979. "Optimal Search for the Best Alternative." *Econometrica*, 47(3): 641–654.

West, Edwin George. 1964. "Adam Smith's two views on the division of labour." *Economica*, 31(121): 23–32.

Williams-Ceci, Sterling, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2026. "Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues." *Science Advances*, 12(11): eadw5578.

Wilson, Timothy D., Christopher E. Houston, Kathryn M. Etling, and Nancy Brekke. 1996. "A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents." *Journal of Experimental Psychology: General*, 125(4): 387–402.

Wu, Suqing, Yukun Liu, Mengqi Ruan, Siyu Chen, and Xiao-Yun Xie. 2025. "Human-Generative AI Collaboration Enhances Task Performance but Undermines Human's Intrinsic Motivation." *Scientific Reports*, 15: 15105.