Writing Is Not Thinking*

Johan Fourie†

Abstract

Writing produces two things: the manuscript published now, and the judgment its author will hold next year. Artificial intelligence can raise the first and lower the second. Which happens depends on which task the author delegates. This article separates writing into its tasks and prices each with a standard tool of economics. Cheaper drafting lets an author test more versions of an argument. The largest gains should go to researchers writing in a second language, whose cost of English is unrelated to the quality of their evidence. Manuscript studies fit that pattern but measure output, not argument quality. Delegating the work of first putting the problem into words can leave a better text and a weaker judge, and the generated draft becomes a starting point the author may not replace. Cheap fluency makes polished prose a weaker guide to the research beneath it; shared starting points can narrow what a field asks. The evidence is uneven, and I say where it is thin. The workflow the predictions make defensible is sequential: the author's own account first, delegated language after. Institutions that value judgment should protect the practice that builds it and remove language costs that exclude good evidence.

Keywords: artificial intelligence; scholarly writing; human capital; language barriers; reference points; judgment

JEL codes: D83; I23; J24; O33

*I thank Tyler Cowen, Di Kilpert, Leon-Ben Lambrechts, 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 Claude Opus 4.8, Claude Sonnet, Claude Fable 5 and OpenAI Codex for literature mapping, mathematical derivation, drafting and revision, and Claude and Codex for independent audits of the formal results. The manuscript was reviewed by refine.ink, an AI review service. I checked the cited claims against the underlying sources, verified the formal results analytically and numerically, and accept responsibility for the argument and any errors. This use exceeds AI-assisted copy editing and is disclosed because it is part of the practice the article examines. Cite this paper as: Fourie, Johan. 2026. "Writing Is Not Thinking." Working Paper, Department of Economics, Stellenbosch University.

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

'discursive writing is not thinking, but a direct verbal imitation of thought.'

Frye (1957)

An economic historian is working with an archive and two possible explanations. One is familiar in the literature. An unfamiliar rival explains an awkward document that the familiar account cannot. Before stating the problem in her own words, she asks a language model for a draft. The model will usually return a clear version of the familiar explanation. She can still edit it. But her task has changed: instead of constructing an account from the archive, she is now comparing revisions with a finished text. Has the model merely saved labour, or has it also changed her judgment?

Now take a researcher with good data and fluent reading English who writes English slowly. A page of publishable prose takes her three hours where it takes a native speaker one, and reviewers still return her manuscripts with complaints about the language. A language model can remove most of that cost at once. It has not improved her evidence. It has lowered the price of expressing it.

These are not opposite cases, one about risk and the other about benefit. The second researcher faces the historian's decision too. If she asks the model to account for her evidence before she has framed it herself, the same mechanism applies. Both researchers make the same pair of decisions; what differs is the price each pays for the technology. An honest account of AI in scholarly writing must therefore hold two margins together: a framing risk that every author runs, and a language cost that some authors bear much more heavily than others.

The bargain predates the language model. In Plato's *Phaedrus*, the god Theuth presents writing to King Thamus as an aid to memory and wisdom. Thamus replies that external marks will weaken trained memory and bestow the appearance of wisdom without its reality (Plato ca. 370 BCE, 274c–275b). His objection was not foolish. Writing reduced the return to memorising what could instead be recorded. It also allowed an argument to exceed the limits of unaided recall. One capacity declined; a larger one became possible. Thamus's warning survives because Plato wrote it down.

Writing can contribute to thinking without being identical to it. The question is whether the gains from delegating prose are worth the costs of delegation. Critics hold that producing prose is part of how a student or scholar comes to understand anything. Delegate it, and the author may never acquire the capacity that the finished text appears to demonstrate (Sparrow and Flenady 2025; Corbin et al. 2026). Advocates point instead to measured gains in output and assessed quality when prose becomes cheaper to produce (Noy and Zhang 2023; Brynjolfsson, Li and Raymond 2025). These claims need not contradict each other. Both can be true, though of different things.

The case against delegation is often compressed into a slogan: writing is thinking. Read as a claim about formation, this is the serious objection, and Section 2 presents it in that form. Yet the wisdom of delegation turns on three questions, each with its own answer. The first is about a product: does AI assistance improve the manuscript? The second is about formation: does

producing and revising prose build the author's later capacity to reason and judge? The third is about transmission: does prose form change what readers believe and remember? An answer to one does not settle the others. A machine need not think, in any philosophical sense, for its output to alter human thought.

This article uses five economic approaches to answer those three questions: search; human capital, the accumulation of skill through practice; behavioural models of reference points and defaults; signalling; and an accounting of effects on people other than the author. They begin from a common asymmetry. Writing technologies lower the cost of producing candidate text much more than they lower the cost of judging which candidate to keep. Text has become cheap. Judgment has not. A model that returns criticism may help with judgment, as Section 4 explains, but the asymmetry drives the results. Judgment is a stock, and practice builds it. The contribution is therefore not one more position in the debate. It is a set of predictions derived from models and stated so that evidence can confirm or reject them, together with the workflow those predictions imply.

1. *Search.* Cheaper drafting improves the argument ultimately selected. The largest immediate gains go to writers for whom producing English prose is most costly, above all researchers writing in a second language.

2. *Human capital.* When generated text replaces the practice that forms judgment, assisted output rises today while unaided capability falls tomorrow. Gaps in assisted performance narrow even as gaps in unaided capability widen relative to the counterfactual in which nobody delegates representation.

3. *Reference points and defaults.* An author who is uncertain about her argument and first reads a generated draft will evaluate alternatives against it and move toward its framing, despite retaining the final choice. The draft is also a plausible default, a ready-made option that stands unless she replaces it. When replacement requires immediate effort, she will stop earlier than her own reflective standard would choose. The benchmark is important: a good, free draft rationally shortens search. The prediction is therefore excess stopping against a patient standard, not merely fewer alternatives.

4. *Signalling.* When fluent prose becomes cheap to produce, fluency becomes a less reliable signal of quality. Readers and editors who continue to rely on its old meaning will make more mistakes. The relation weakens toward zero. Reversal requires a further condition, namely that adoption concentrates among authors with weaker evidence, which this article does not claim. This is a prediction about the pooled relation. Where poor English had concealed good evidence, removing it makes prose more informative and improves screening of that group.

5. *External effects.* If many authors draw on shared generative defaults, individual gains occur

alongside field narrowing: more material to screen, but fewer distinct questions and explanations under study.

The evidence is uneven, and its strength should be visible from the outset. Millions of scientific manuscripts have been screened for signs of AI assistance. Their patterns are consistent with the distributional part of the first prediction, but they do not measure the quality of the selected argument. The fifth prediction rests on a different and weaker source, a comparison of scientists classified by their use of AI tools, and has no direct test. The fourth has more direct support from screening experiments in hiring and investment. Sections 7 and 8 present both the evidence and its limits. Education supplies direct experimental support for the central contrast in the second prediction: assisted output rises while unaided capability falls. The cross-sectional claim about gaps between authors holds only under conditions given in Section 4; two receive empirical support, while a separate scope condition restricts the technology. Everything in the third prediction remains untested for scholars. The conclusion must therefore be conditional: what matters is where assistance enters the sequence of inquiry, and which cognitive acts a teacher or journal is trying to produce.

1 What Writing Does

Writing is both process and product, but that familiar distinction is still too coarse. Much of the disagreement becomes tractable once the process is opened up. At least seven tasks sit inside what scholars casually call writing.

Table 1: Tasks within scholarly writing

Task	Function	Initial allocation
Inscription and transcription	Put specified content into grammatical form	Often delegable
Externalisation	Make a vague representation explicit enough to inspect	Formative in learning and discovery
Candidate generation	Produce alternative claims, structures, examples and phrasings	Useful only with criteria for comparison
Revision and criticism	Locate failures and change the representation	Plausible site of human–AI complementarity
Selection and verification	Decide what is supported, important and attributable	Author remains accountable
Narrative design	Choose causal structure, emphasis, ordering and examples	Part of judgment because it changes belief
Deliberate practice	Build later capacity to perform these tasks	Cannot be valued by today's manuscript alone

Planning a sentence may reveal a missing causal link; changing a word may expose a change in

meaning. Yet grammatical realisation and causal structure remain different tasks, even when hand and mind perform them in a single movement.

Research on composition supports this division. Writing, in Flower and Hayes (1981), moves repeatedly among planning, translating and reviewing; it is not a one-way transfer from thought to page. Bereiter and Scardamalia (1987) separate knowledge telling, the retrieval and recording of existing ideas, from knowledge transforming, in which writing changes the writer's understanding. Meanwhile, Kellogg (1994) stresses the working-memory burden of coordinating content, language and audience. Menary (2007) makes the stronger claim that written marks can join the cognitive system: by manipulating them, writers complete tasks they could not complete internally. The broader philosophical case for treating a reliably integrated external resource as part of cognition comes from Clark and Chalmers (1998).

This literature establishes that writing can be part of thinking. It does not establish that every part of writing must remain manual. Oatley and Djikic (2008) describe expert writers as putting thought into external symbols and then working on those symbols. Expertise determines what happens next. A skilled writer treats a draft as material for substantive revision; a novice tends to preserve the first version of an idea she manages to write down.

The distinction among tasks yields three outcomes, of which the models below price the first two. The first is manuscript quality: the correctness, clarity and contribution of the finished text. The second is author capability: what the writer knows, what transfers to a new problem and what she can still do when the tool is removed. The third is field quality, which belongs to no single author: the number of distinct explanations a discipline keeps under study and the cost to readers of finding reliable work. The economic case for delegation usually counts only the first. Its strongest critics worry about the other two. A reader holds the manuscript; the other two outcomes leave no mark on it.

2 The Case for Human Prose

The strongest case against delegating scholarly prose begins with a rival account of production. Drafting does not merely express judgment; it helps to form it. The difficulty is part of the output. An author starts with impressions that are incomplete and partly inconsistent. A sentence forces her to choose agents and causes. Revision exposes choices that remained hidden while the thought was vague. Repetition makes that capacity more reliable. A student who avoids this process may deliver an acceptable text without acquiring the ability that such a text ordinarily demonstrates.

Sparrow and Flenady (2025) frame the objection as learning how before learning that. A university does more than transmit statements available elsewhere. It places students inside practices where they learn to frame an argument and weigh what another person tells them. Teachers matter as exemplars; education matters as a social relation. On this view, replacing an essay with a plausible product is one case of a larger error: confusing the delivery of content with the formation of a participant in a practice.

The argument of Corbin et al. (2026) is related but narrower. They distinguish the standardised, product-focused assignment from the essay as an exploratory engagement with uncertainty, divergence and recursive exchange between writer and text. Their response to generative AI is not a return to pre-AI assessment. It is a redesign of the surrounding pedagogy that makes process and the cultivation of judgment visible. This sharpens the objection here. The value at risk includes manual grammar and a form of activity that current assessment often measures poorly even without AI.

Formation is not the only issue. Editing a supplied draft differs from revising one's own attempt. The supplied text shapes what is omitted and which causal story seems natural, a mechanism formalised in Section 5 and treated there as untested. Fluency makes such a draft especially hard to reject: few visible mistakes force the reader to slow down. The central error may instead be a premise that the text never marks as uncertain. An author capable of constructing a better account may spend her available time improving the wrong one.

Nor does the word *oversight* settle the matter. The ability to detect a plausible error is unevenly distributed, and it must be acquired somewhere. An expert can compare a generated claim with a large stock of domain knowledge. A novice may confuse fluency with fit precisely because the assignment is meant to build that stock. Automation can alter attention as well: monitoring a generally reliable system may invite less scrutiny than producing the answer oneself. In one decision task, requiring participants to engage analytically with an AI recommendation reduced their reliance when it was wrong. Yet they disliked the interfaces that imposed this work, and the benefit was uneven (Buçinca, Malaya and Gajos 2021). Formal authority over the last choice does not guarantee an independent decision.

The relation between text and evidence poses a further problem. A language model can summarise material supplied to it and propose checks. It cannot independently inspect an archive, interview a witness or know which uncited premise comes from the researcher's experience. Fluent synthesis may therefore conceal a broken link between a claim and the world. The relevant comparison is not human prose against machine prose in isolation. It is between workflows with different access to evidence and different incentives to verify it.

The framework of this article then faces a deeper objection, first stated by an economist. McCloskey (1985) rejected the premise that content and expression are separable. In production terms, scholarship cannot be written as the sum of two functions. An author discovers the details of an argument only by writing them out, and in doing so often discovers a flaw in its foundations.

Philosophy reached the same territory by another route. The linguistic turn (Rorty 1967) denied that language merely transmits thoughts completed in advance; language helps form them. Menary (2007), cited above, offers one version of this claim. If it is right, producing a candidate and judging it may not be two activities in the first place.

Reddy (1979) makes the charge sharper still. English encodes a dominant picture of communication in which a speaker places meaning into words and a listener later takes it out. Reddy argues that this picture is false and distorts what people expect of one another. A model that prices prose

as a container, produced at one cost and inspected at another, invites exactly that criticism.

The objection cannot be answered in full. The useful response is to separate the part this article accepts from the part it does not.

McCloskey's claim has two parts. First, writing forms content: authors discover an argument's details and defects by writing them out. This article accepts that claim and is built around it. It is why the model has a second output. In the equation for J_{t+1}, delegating the first act of putting a problem into words reduces g, and g is an input to judgment. The claim thereby acquires a magnitude. How much judgment a particular delegation costs becomes a quantity to measure rather than a doctrine to affirm; Section 9 proposes such a measurement.

The second part concerns production of the manuscript itself, and here the article does not satisfy McCloskey. Proposition 1 prices the production and checking of a candidate as c_v and c_s. Its comparative static requires a fall in c_v to leave c_s, the distribution of candidate quality and the act of judgment unchanged. McCloskey doubts precisely those invariances; giving the costs different names does not establish them. The proposition therefore applies to tasks in which composing and judging interact weakly. Table 1 distinguishes tasks along this dimension. The interaction is weakest for a settled description of a method and strongest for an original argument under unsettled evidence, the same set of tasks in which this article says delegation is most dangerous. Her objection and the article's warning point to the same place.

Reddy's correction can also be translated into the model. A model supplies marks, not a candidate. Moving from those marks to a judgment requires reconstruction, productive work that the author performs whether the marks are hers or the machine's. This article places reconstruction inside c_s rather than pricing it separately, an acknowledged simplification. Two consequences follow. First, a model lowers the cost of producing marks, not the cost of building an argument from them. The number of candidates an author can check therefore remains bounded as drafting approaches free, as Section 3 shows. Second, the saving is smaller than cheap generation alone suggests: reconstructing an argument from someone else's marks is not obviously cheaper than reconstructing it from one's own.

There is also a point of agreement, though it is corroboration rather than reconciliation. Both accounts recommend nearly the same practice: write one's own account, receive criticism, revise. McCloskey derives the rule from the nature of writing; this article derives it from relative costs. Agreement about what to do leaves the disagreement about why intact. That disagreement is for the reader to weigh.

What survives the objection is the second output, provided J is read as the model reads it. To say that judgment cannot be separated from writing practice is to make a claim about what judgment *is*. The model requires something weaker: that an author's performance without the tool, on new problems and after a delay, can be measured. Section 9 proposes that measurement. A reader who denies even this has denied the existence of the article's second output, and none of the evidence assembled here can settle the disagreement. That is a genuine limit on what the article can establish.

Prose also acts on readers, an effect Section 8 takes up. One point belongs here: authenticity may matter independently of measured quality. A reader may reasonably value an essay as an encounter with the author's own expression. That value cannot be reduced to truth or efficiency, although the analysis below concentrates on epistemic and educational outcomes.

For many scholars, writing is also one of the pleasures of the work. Smith (1776) began his account of wage differences with the agreeableness or disagreeableness of employment itself. Workers still surrender substantial pay for work they find meaningful and enjoyable (Maestas et al. 2023; Cassar and Meier 2018). A paragraph that finally says what one means is consumption as well as production. Yet four experiments with 3,562 participants found that working with a generative model raised task performance while lowering intrinsic motivation and increasing boredom when participants returned to unassisted work (Wu et al. 2025). Enjoyment may subsidise formation because an author who likes writing practises it more, but that link is this article's conjecture, not a measured mechanism. Nor does a high cost of prose imply an absence of pleasure: effortful mastery can itself produce utility (Loewenstein 1999). Gans (2026) places the same feature inside an automation problem. When a worker supplies extra time to a task she enjoys, that time does not enter the payroll record. The firm may then automate the task in order to contain it. His setting is a firm choosing a contract; a scholar writing her own paper has no such counterparty. Two things still carry over. Enjoyment changes which tasks are delegated rather than only what they cost, and the time it adds is invisible in the standard measures. Even so, the researcher who pays three hours for each page of English pays far more for that pleasure than a native speaker. She may reasonably refuse the price. Valuing the activity supports a choice not to delegate; it does not justify requiring others to forgo the tool.

The force of these objections varies by task. A standard contract, a translation, a settled methods description and an original historical interpretation do not pose the same problem. Neither do an experienced scholar and a first-year student. The sections that follow treat the objections as economic mechanisms, not as sentiment about an old craft, and examine them in turn.

3 Cheaper Search

Begin with the modest positive case. An argument can be treated as a search among candidate versions. The quality of each candidate, meaning its fit to the evidence and the strength of its contribution, remains unknown until the author produces and examines it. Let production cost c_v units of effort and checking cost c_s, where the subscripts denote variation and selection. From a total effort budget B, the author can examine

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

candidates. Give her forty hours for the paper's argument. If a serious reformulation takes seven hours to produce and one hour to check against the evidence, she can examine five. If a language model lowers production from seven hours to one while checking still takes an hour, she can examine

twenty. The tool has quadrupled the number of arguments she can afford to test. It has checked none of them.

Writing technologies change these costs in different ways. The word processor made it cheap to return to an earlier draft, bringing composition closer to search with recall. But easier editing did not translate one-for-one into new ideas: writers revised more often, yet a larger share of revisions preserved the existing text instead of changing its meaning (Ransdell and Levy 1994). What matters economically is a meaningful candidate, not another keystroke or edit.

Suppose the quality of each meaningful revision is drawn from a fixed distribution F. Candidates vary in quality, and average quality is finite. The author retains the best draft she has checked, so $Q_n = \max\{q_1, \ldots, q_n\}$, where q_i denotes the quality of candidate i, as in Section 5. Earlier drafts are stored rather than lost; the author therefore searches with recall. In the classic framework, recall is an assumption, not a result (Weitzman 1979).

Proposition 1 (Cheaper iteration, better argument)**.** *Expected final quality $\mathbb{E}[Q_n]$ is increasing and concave in the number of revisions n, and decreasing in their cost. If a technology offers a common execution cost while judgment cost is fixed, and writers differ initially only in execution cost, the gain is largest for writers whose execution cost was highest.*

The gain from one more candidate is

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

and diminishes with n: each new draw must beat the best of a growing set. Lower production costs therefore improve the expected final draft only while candidates remain meaningful, earlier drafts remain available and the author's check can distinguish better candidates from worse ones.

The proposition's distributional clause gives the article's first prediction. Production cost c_v differs across writers. A researcher writing in a second language may face a high c_v for reasons unrelated to the quality of her data or question. Generative AI lowers precisely the cost that varies most across writers and least with the value of their science. The model therefore assigns the largest immediate gains to researchers writing in a second language.

That mapping requires two assumptions. First, language raises the cost of producing candidates but not the cost or accuracy of judging them. If second-language writers also select more noisily among English drafts, part of the predicted gain disappears. Second, candidates are independent draws from a fixed distribution. Outputs from one model are correlated, so their maximum improves more slowly. Section 8 returns to this problem at the level of the field.

A language model does more than cheapen a candidate. It can find a missing comparison, produce several versions of a difficult passage or direct revision toward an improvement rather than leaving the author to search blindly. It can also direct her toward whatever the model regards as the canonical account. The static proposition cannot distinguish helpful criticism from canonical pull: it holds the author's judgment fixed and assumes accurate selection.

The floor on cost is therefore important. As c_v approaches zero, n approaches B/c_s, not infinity.

Cheap generation cannot make careful checking free. An author who requests one hundred alternatives but reads only the first few is no longer following the process described by the proposition. Her capacity to generate has outrun her capacity to select.

Early experiments are consistent with the distributional part of the first prediction. In professional writing and customer support tasks, assistance raised measured performance most among participants whose unassisted performance was lowest (Noy and Zhang 2023; Brynjolfsson, Li and Raymond 2025). The outcome, however, is assisted output. These studies do not reveal what users can do after the tool is removed. The next section turns to that second output.

4 A Manuscript and a Writer

The formation objection requires the second output identified in Section 1: author capability. Economics already has a way to value an activity that produces an output today and an ability tomorrow. Human capital is the stock of skill embodied in a person. It is built through costly practice and depreciated through disuse (Becker 1964). Drafting, on this account, is not merely a cost of the present manuscript. It is an investment in the person who will judge the next one. Sustained, effortful thought is itself such a stock. In a field experiment with 1,600 primary-school students, practice at continuous cognitive work measurably increased it (Brown et al. 2025). Removing the effortful stretches of scholarship also removes practice at effort.

The model must now preserve the distinction in Table 1 between linguistic execution and externalisation. The workflow eventually recommended here delegates the former while retaining the latter. Let $d_L \in [0,1]$ be the share of linguistic execution delegated to a model, where $d_L = 0$ means the author produces every sentence and $d_L = 1$ means the model produces them all. Let $d_R \in [0,1]$ be the model's share of the separate task of first putting the problem into words. The two margins rise together in an answer-first workflow, where the author requests an account before forming one. They separate when she first writes her own account and later requests language help: high d_L, low d_R. Write $d = (d_L, d_R)$. Current manuscript quality is

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

increasing in the number of checked candidates $n(d_L)$, with $\partial n/\partial d_L > 0$, and in current judgment J_t and decreasing in b, a distortion in candidate selection. This distortion depends on d_R because the reference point in Section 5 comes from the generated first representation, not from the delegated sentence. Proposition 1 fixes J_t and sets $b = 0$.

Judgment changes through use:

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

Judgment depreciates at rate δ, and expert performance requires continued deliberate practice (Ericsson, Krampe and Tesch-Römer 1993). Four activities replenish it: the author's own generation

g, which turns a vague idea into explicit words; active revision r; verification v against sources and data; and feedback f from others. The separation of d_L and d_R now does real work. Own generation falls with d_R, because a model that forms the representation displaces that input. Its response to d_L is indeterminate: delegating the sentence that expresses an account already reached need not reduce the work of reaching it. Design determines the other three inputs. A system that supplies a finished answer first raises both delegation margins and can replace all four activities. A system that criticises the author's own attempt raises d_L, keeps d_R near zero and may increase revision and feedback while lowering language costs.

With $W_t = Q_t + \omega J_{t+1}$, the marginal effect of delegating a little more of either task, $k \in \{L, R\}$, is

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

This derivative describes a small change, not whether delegation is worthwhile in total; the latter requires comparing total changes. The parameter ω captures how much the author, or an institution acting for her, values future judgment relative to the present text. An editor assessing one submission may place ω near zero. A university teaching first-year students should not. The expression thus prices the difference between exercising judgment and acquiring it. If delegation crowds out formation, its gain to the manuscript must exceed the discounted loss of learning. If guided assistance adds enough revision, verification or feedback, both terms may instead be positive.

The second prediction follows. Under the search benchmark, assistance raises Q_t; when it displaces formative activity, it may lower J_{t+1}. The cross-sectional claim about gaps between authors requires more. It does not follow from G and L as written. The accompanying derivation supplies sufficient conditions. Extra checked candidates must be less valuable to an author who already judges well, allowing cheaper search to close rather than widen the assisted gap. An hour of the author's own generation must add less to a large stock of judgment than to a small one. Finally, the author with less judgment must ask the machine to frame the problem more often. Own generation must also fall as representation delegation rises and remain a productive input. Under these conditions, assisted quality gaps narrow in d_L, while unaided judgment gaps widen relative to the counterfactual in which nobody delegates.

A final condition defines the result's scope. Delegation must remove the author's generation without replacing it. If the interface instead supplies criticism, thereby raising revision and feedback most for the author who delegates most, divergence can reverse. This is the guardrail result of Bastani et al. (2025), expressed as a condition on the model rather than as a conclusion drawn from one experiment.

The third condition concerns behaviour and has direct support. Given unrestricted access, students overwhelmingly requested answers rather than explanations (Bastani et al. 2025). Fan et al. (2025) call this shortcut metacognitive laziness. Their experiment, unlike Wu et al. (2025), found no loss of intrinsic motivation; the motivational cost of assistance remains unsettled. The first condition is a restriction rather than a fact. If candidates and judgment were complements instead

of substitutes, cheaper search would widen the assisted gap. The evidence for the restriction comes from experiments in which assistance helped the weakest performers most. The contrast between assisted and unaided gaps should appear when the tool is withdrawn.

Evidence that generation contributes to learning long predates generative AI. A meta-analysis of 86 studies estimates a mean advantage of about 0.40 standard deviations for generation over reading (Bertsch et al. 2007). Most tasks in this literature test memory for simple verbal material, and the advantage shrinks as the material becomes more complex. Better recall is not yet better judgment; moving from one to the other is an inference, not a finding. A meta-analysis of 56 school experiments comes closer to the capacity of interest. Writing about content improved learning in science, social studies and mathematics by about 0.30 standard deviations on average (Graham, Kiuhara and MacKay 2020), although effects varied widely across designs. Neither literature separates generation, revision, verification and feedback. The four inputs to L are therefore a modelling choice, not an estimate. The evidence does not prove that every sentence must be produced manually. It does show that replacing generation with exposure can remove a productive learning activity.

The emerging AI evidence makes sequence visible. In a preregistered field experiment with almost one thousand secondary-school mathematics students, Bastani et al. (2025) compared unrestricted GPT-4 with a tutor designed to withhold direct answers and offer guided help. While available, the systems raised performance by 48 and 127 percent. Once access disappeared, students who had used the ordinary interface scored 17 percent below students who had never received it; the guarded tutor largely mitigated the loss. This is the one design that separates the two outputs within a single trial. Mathematics is not scholarly writing, and a short field experiment cannot identify long-run intellectual development. Even so, the experiment demonstrates the separation the argument requires: assisted output and acquired capacity can move in opposite directions, depending on the interface.

The same substitution appears at field level in the model of Acemoglu, Kong and Ozdaglar (2026). Good decisions in their economy require both shared general knowledge and knowledge of one's own problem. The costly effort that produces the second also enlarges the first. AI recommendations substitute for that effort, so each person rationally works less: a good answer is available without the cost. Yet everyone's effort feeds the common stock of knowledge, and no individual counts her own contribution to it. Once AI becomes sufficiently accurate, the economy can enter what the authors call knowledge collapse. Individual decisions remain well advised while general knowledge falls toward zero.

Applying this result to scholarship requires a premise absent from the two-period model above: drafting must contribute to a stock of judgment beyond the author's own and delegation must fail to replace what drafting removes. If so, individually sensible delegation cannot protect collective judgment, just as individually sensible fishing cannot protect a fishery.

The classroom evidence also shows why the student case cannot simply be inferred from the expert case. An experienced scholar may already possess the representation she asks the model

to express. For a student, forming that representation is often the assignment. The asymmetry has empirical support: as expertise rises, instructional aids that help novices lose value and may reverse their effect (Kalyuga et al. 2003). Yet this literature also warns against a universal attempt-first rule. True beginners often learn more from worked examples than from unguided attempts (Kirschner, Sweller and Clark 2006); learners with some preparation may benefit from struggling before help arrives (Kapur 2008). An unaided attempt can thus be a rational investment even when it lowers today's measured product. Which sequence protects formation depends on the learner's position along the expertise gradient. The point is to protect an input into future judgment, not to declare difficult prose virtuous.

The reverse possibility must remain open. Feedback is scarce. Many students submit once, receive a grade and rarely revise. A model that asks for a claim, identifies a contradiction and requires a repair may create more active practice than solitary work. The two-output model consequently supports neither unrestricted access nor a blanket ban. It asks what students can do without the tool, what they retain after a delay and which activities the interface removes or adds.

5 The First Draft as a Reference Point and a Default

Adequate judgment does not guarantee an independent evaluation of generated prose. People often value an option relative to a starting object rather than on its own terms. Economists call that starting object a reference point. Dean et al. (2026) model reference points as sources of comparative information when absolute value is uncertain. Such a reference can help: saying which of two objects is better is often easier than assigning either an absolute value. But the comparison also creates dependence on the starting point, especially when uncertainty is high and the reference is easy to understand.

A reduced-form application to writing is

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

The author's perceived quality $\widetilde{q}_i$ equals the true quality q_i, random error ε_i, and a distortion. The function $s(x_i, x_0)$ measures the similarity between candidate x_i and the first generated draft x_0; α gives that similarity its weight. The weight rises as the author's judgment J_t weakens and her uncertainty u_t increases. This expression adapts Dean et al. (2026); it is not their result. It states a testable mechanism: the less able an author is to value alternatives independently, the more the generated draft matters as a reference. If x_0 expresses the canonical account already familiar in the literature, later candidates receive a hidden premium for resembling it.

The two comparative statics do not have equal standing. Informational theory supplies $\alpha_u > 0$: uncertainty increases the reference's weight. This article conjectures $\alpha_J < 0$: weaker judgment does the same. Indeed, the informational theory could have the opposite welfare implication, because a reference may make nearby evaluations more precise (Dean et al. 2026). In the canonical model of reference-dependent preferences, moreover, the agent's expectation is the reference. Anticipated

use of a model would then supply a reference even before its draft appeared (Kőszegi and Rabin 2006).

Adjacent evidence comes from outside economics, and it concerns production after exposure rather than the ranking of alternatives. Engineers reproduced features of an example solution, including flaws against which they had been explicitly warned (Jansson and Smith 1991). Participants asked to generate new ideas likewise conformed to examples they had just seen (Smith, Ward and Schumacher 1993). The classic analogue in judgment research is anchoring (Tversky and Kahneman 1974). But no available evidence isolates the similarity premium in evaluation itself.

One recent experiment comes closer to the writing case. Williams-Ceci et al. (2026) gave 2,582 participants a writing assistant whose suggested continuations favoured one side of a public question. Participants who saw them reported attitudes nearer the model's position than participants who wrote unaided, by about 0.4 points on a five-point scale. The same arguments shown as fixed text moved them less, so the information alone does not account for the effect. Most participants judged the suggestions balanced. Those whose attitudes moved were the least likely to notice the bias. Warning participants beforehand and telling them afterwards did not measurably reduce the shift. The setting is opinion on public questions rather than the ranking of scholarly candidates, and the effect is small. The authors attribute the movement to having written in support of a position, a different mechanism from the similarity premium above. The experiment does show that text introduced during composition can move the writer's own position without her noticing.

Reference dependence dissolves the clean boundary between candidate production and selection. An author may retain the final choice yet judge every alternative against a model-generated starting point. More iteration then polishes the first framing rather than testing rivals to it. This is the ranking part of the third prediction. Additional candidates need not rescue the author because the same distortion ranks them all. In the manuscript-quality function of Section 4, that distortion is $b(d_R, J_t)$: delegation supplies the reference, and weaker judgment gives it more weight.

The Gaussian benchmark in the accompanying derivation sharpens the distinction between a reference point and random error. Suppose true quality and independent noise are normally distributed, and the author ranks candidates by $\widetilde{q}_i = q_i + \varepsilon_i$. As the candidate set grows, selecting the highest perceived quality still raises expected true quality. The increase is smaller than the winning score suggests because noise helped the winner; an author who fails to correct for this will be disappointed. Even so, more candidates do no harm in this benchmark (Smith and Winkler 2006).

A similarity premium differs because it pushes every ranking in the same direction. The derivation imposes two conditions: the bonus to a canonical account is large enough to make it outrank every original candidate, and that account is worse on average than the originals it displaces. It can therefore defeat better but less familiar alternatives. Adding candidates then increases the chance that a canonical candidate enters the set and wins. Expected quality may first rise and then fall as the set expands. The generator has not worsened; the selection rule has changed. Candidates are ranked differently, not drawn from a different distribution. The accompanying derivation states

the cases precisely.

Narrative persuasion provides adjacent evidence. In the experiment of Barron and Fries (2025), advisors gave causal explanations of identical objective data. Narratives moved recipients' beliefs even though sender and recipient possessed the same information. Empirical fit strongly predicted adoption, and advisors anticipated this by choosing components that made their preferred claims cohere with the facts. A private-reasoning stage produced only a small, statistically insignificant reduction in the persuasion gap. Warnings about bias and disclosure of the advisor's incentives reduced it much more.

A language model is not an advisor with a personal stake in the chosen interpretation. It does, however, have a training distribution and a tendency to produce narratives that fit familiar patterns. The narrower lesson survives the imperfect analogy: coherence can change belief without adding information, and forming a prior before exposure may offer insufficient protection. The interface study of Section 2, on nutrition decisions rather than writing, adds a related caution: an instruction to check an answer is not a design that causes checking to occur.

Participants' preference for the less demanding interface suggests a second mechanism. Before generation became cheap, no effort meant no manuscript. The blank page was a commitment device that nobody had to design. Unless an author could pay someone else, the only path to an acceptable text passed through her own attempt to construct one. A generated draft changes that zero-effort outcome from a blank page to a plausible manuscript. The draft becomes a default, the ready-made option that stands unless the author rejects it. This formalises a familiar intuition: being forced to write is being forced to think. Writing is no longer a decision to produce, but a decision to reject.

People often stop when an option is good enough rather than search for the best one (Simon 1955; Caplin, Dean and Martin 2011). Present bias strengthens that tendency: immediate effort weighs more heavily than later benefits, especially for real effort rather than money (Augenblick, Niederle and Sprenger 2015). Defaults persist even at high stakes. Automatic enrolment shifted participation in retirement plans by tens of percentage points (Madrian and Shea 2001); across 58 studies, defaults moved choices by about two-thirds of a standard deviation (Jachimowicz et al. 2019). Their influence weakens when preferences are strong. That heterogeneity matters: an author who knows what she wants to argue is precisely the chooser a default should move least. Rejecting generated prose requires hard thinking now, while the return, a better argument and stronger judgment, arrives later. An author may therefore accept a draft she recognises as not quite her own argument.

Earlier stopping does not by itself reveal bias; the benchmark must be rational search. An author who already holds a free draft of acceptable quality should continue only while the expected gain from another attempt exceeds its cost. A better incumbent rationally shortens search (Weitzman 1979). The boundary also holds when the menu grows. With constant search costs and a known quality distribution, a larger choice set does not by itself increase default choice. Learning about the distribution, rising search costs or choosing search depth in advance can reverse that result (de Lara

and Dean 2025). The behavioural prediction is therefore *excess* stopping. Suppose attempt $m+1$ imposes immediate effort $e = c_v + c_s$ and yields an expected later improvement Δq_m. A patient author continues if $\Delta q_m > e$. With present bias $\beta < 1$, she continues only if $\beta\,\Delta q_m > e$: one unit of immediate effort must return more than $1/\beta$ units of later value. If expected gains decline across attempts, she stops no later than her patient self, and strictly earlier whenever a reachable gain lies above e but no higher than e/β.

This stopping channel is distinct from reference dependence. The reference point distorts the ranking of candidates already examined; present-biased effort reduces the number examined below the rational benchmark. Together they form the two parts of the third prediction. Neither has been tested on scholars. The adjacent evidence makes them plausible, and no more.

6 The Smithian Bargain

The second prediction carries an institutional implication that the others inherit. Each identifies a capacity that some institution must either protect or replace. Economic history gives the trade-off a familiar form. In Book I of *The Wealth of Nations*, Smith (1776) explains how specialisation, less time lost between tasks, and task-simplifying machinery allow the division of labour to raise productivity. In Book V he returns to the same process with a warning. A worker confined to a few simple operations 'generally becomes as stupid and ignorant as it is possible for a human creature to become'; the 'torpor of his mind' leaves him unable to join rational conversation or form a just judgment concerning many even of the ordinary duties of private life. Smith's remedy is education. He counted the pins, and then counted what the pin-maker lost.

One theory can therefore produce more current output and less human capability. I call this exchange the Smithian bargain. Society accepts narrower learning within each job in return for higher output now, then relies on institutions such as schools to replace the learning the job no longer supplies.

The apparent tension between Smith's two books has generated its own debate. West (1964) called it a contradiction; Rosenberg (1965) denied that it was one. The narrowing of work makes compensatory education necessary, while the productivity released by specialisation makes that education affordable. Rosenberg's resolution applies this article's thesis to an older technology. The question is not merely which capacity disappears. It is which institution will reproduce it.

The history of technological anxiety supplies an important caution. Mokyr, Vickers and Ziebarth (2015) explain both its recurrence and the repeated failure of aggregate predictions of permanent technological unemployment. Task displacement alone does not imply the end of scholarship. Yet that history is not decisive here. A social function can survive and expand while a capability once formed by its old organisation declines. Historical comparison must identify both the skill at risk and the institution that supports transition; survival of the function is not a complete welfare test.

AI-assisted writing is a cognitive division of labour. A model can perform linguistic realisation and supply criticism, freeing the author to concentrate on evidence and causal judgment. This is

the gain of Book I. If the new allocation also removes the practice through which an author learns to externalise and repair an argument, it brings the loss of Book V. The lesson is neither to preserve every inherited task nor to assume that new expertise will form by itself.

Technology may also complement education. Goldin and Katz (1998) connect early twentieth-century electrification with rising demand for educated workers. In the modern task approach, machines replace activities that can be reduced to rules while increasing the value of expertise (Autor 2015; Autor and Thompson 2025). The proper unit of analysis is therefore the task, not the occupation or the person. One technology can remove some of a person's tasks and raise the return to others.

Writing, print and word processing remain useful comparisons, provided their differences are kept in view. Print sharply lowered reproduction costs, and European cities that adopted it in its first decades grew measurably faster than neighbouring cities (Dittmar 2011). The word processor lowered the cost of return and rearrangement. Writers have long suspected that the instrument reaches into the sentence: working on a typewriter in 1882, Nietzsche remarked that 'our writing tools are also working on our thoughts' (Kittler 1999).

Neither print nor the word processor, however, normally supplied an original semantic continuation or causal explanation. A language model does. Such continuations were always available through a ghostwriter or a secretary who could compose a letter from brief instructions, but they were scarce and costly personal services. The model makes them a nearly free commodity, the kind of price change analysed by the search model in Section 3. It can therefore enter selection through the content of the candidate it makes cheap. The discontinuity is one of price and scale, not kind; that is a reason for careful comparison, not for abandoning comparison.

7 Who Gains and Who Learns

The first prediction has a distributional component that can now be examined at scale. The pattern fits the prediction, but no more can be claimed: measurement problems are substantial, and none of the evidence observes whether authors selected better arguments.

Start with the unequal cost. In a survey spanning eight countries, Amano et al. (2023) find that researchers who are not native English speakers spend more time reading and writing in English and receive many more language-related rejections and revision requests. Among respondents with one English-language publication, those from countries of moderate or low English proficiency were about two and a half times as likely as native speakers to report a rejection attributed to their English. Professional editing, moreover, is priced in the currencies of rich countries. A burden unrelated to scientific value can therefore influence whose evidence and questions enter the record. The remedy is itself distributed unequally: weak infrastructure and high connectivity costs prevent many researchers across Africa from accessing AI tools (Okolo, Aruleba and Obaido 2023).

Kusumegi et al. (2025) examine what happened as the cost fell. They use word patterns to identify probable LLM assistance in more than two million preprints from arXiv, bioRxiv and SSRN,

then compare adopters with similar non-adopters. Estimated monthly output after adoption rose by 36.2 percent on arXiv, 52.9 percent on bioRxiv and 59.8 percent on SSRN. The largest gains appear in the groups singled out by the search model. Among scholars classified by name as Asian and affiliated with Asian institutions, the estimates reach 89.3 percent on bioRxiv and 88.9 percent on SSRN. Among researchers classified by name as Caucasian and based in English-speaking countries, they range from 23.7 to 46.2 percent. The authors expect scientific production to shift toward regions in which English is not the first language.

These estimates are associations, not experimental effects, and two measurement problems add to the ordinary caution. The first is classification. Errors in a text classifier need not be random. Detectors of machine-generated prose misclassify a majority of essays written by non-native English writers because unassisted second-language prose tends to have low perplexity, meaning more statistically predictable wording, which those classifiers flag (Liang et al. 2023). The particular classifiers used in these studies have not been audited for that pattern. Nevertheless, an error correlated with the very characteristic under study could create part of the measured language-group gap without any corresponding difference in tool use. Classifying authors by name adds another fallible proxy, and the resulting group labels combine language cost with field and institutional resources.

The second problem is timing. Adoption is dated to the first month containing a flagged paper, making the treatment date mechanically related to output: high-output months are more likely to contain at least one flag. Placebo flags produce a positive post-adoption pattern even without treatment (Renault, Bergeaud and Bosquet 2026). Authors who post more are also more visible in preprint repositories. This timing critique bears mainly on estimated levels; differential detector error threatens the comparison across language groups itself.

Published journals display the same broad movement. He and Bu (2026) analyse 5.2 million papers in 5,114 journals and estimate that AI-assisted writing rose in every discipline after ChatGPT's release, fastest among authors in non-English-speaking countries. By early 2025, about seventy percent of journals had introduced an AI policy, usually requiring disclosure. Yet growth in estimated AI assistance was statistically indistinguishable between journals with and without such policies. The two patterns are consistent with a cost falling most where it had been binding and with formal rules leaving incentives largely unchanged. Policies and enforcement were not randomly assigned, however, so the study cannot identify why the trends were similar.

Lower language costs also change what evaluators learn from prose. Cowgill, Hernández-Lagos and Wright (2026) model application essays and pitches as signals, then run hiring and start-up investment experiments. Giving applicants access to ChatGPT reduced screening accuracy by four to nine percent: polished prose became less informative about who was stronger. For applicants from non-English-speaking countries, however, access *increased* screening accuracy. Imperfect English had obscured real quality; removing that noise revealed information. In both directions, access changed the informativeness of prose. Section 8 develops the mechanism.

Equalising access is not the same as equalising capability. An experienced writer can use the

model to complement an account she already possesses. A less prepared user may be more likely to request a complete answer and less able to detect a plausible error. If so, the second prediction applies. In scholarly writing this remains a prediction, not an established fact.

A second distributional force pulls in the opposite direction. The system that removes a language barrier may reward conformity to its dominant style and stock of explanations. Suggestions from a Western-trained model shifted writing by participants in India toward Western styles and reduced culturally specific content (Agarwal, Naaman and Vashistha 2025). This is adjacent evidence of convergence, not a direct test of the first-draft mechanism in the third prediction. Lower language costs could broaden participation, although the available evidence measures output among existing authors rather than entry by new ones. Conformity could simultaneously narrow what that participation contributes.

The field that studies second-language writing has arrived at a similar two-sided position. Hyland (2026) holds that generative models neither end second-language writing nor repair its long-standing problems. They can supply linguistic resources and widen access to disciplinary genres. They can also deskill writers and deepen existing inequities. His argument is a statement about what the field should value rather than a finding, and it comes from the pedagogy of writing rather than from a comparison of costs.

Distribution must therefore be considered twice. Who can produce an acceptable manuscript with the tool? And who acquires the capacity to judge the next manuscript without it? A policy that asks only the first question may widen publication while weakening the formation of judgment. A policy that protects formation by demanding native-like manual prose preserves an exclusion that serves no purpose for researchers able to access the tool. The design problem is to remove a native-speaker advantage unrelated to scientific value without removing the practice through which judgment grows.

8 Narratives, Readers and the Field

Scholarly prose does more than state claims. Its form changes what readers believe. This is an old proposition in economics (McCloskey 1983) to which experiments now give measurable content. Graeber, Roth and Zimmermann (2024) compare stories and statistics in controlled settings. Statistics move beliefs more at first, but lose more of their effect after a one-day delay: about 73 percent, against 32 percent for stories. Qualitative material improves cue-dependent retrieval, the recovery of information when associated material prompts it, even when it contains no information about the state. This is not an argument for replacing statistical evidence with academic narrative. It is evidence that representational form changes what remains available to later judgment.

Graeber, Roth and Schesch (2026) examine more than 6,900 spoken explanations of financial choices. Explanations improve social learning on average because correct answers tend to come with richer reasoning, but they do not equally arrest the spread of false answers. Receivers imitate richer explanations more. When linguistic richness varies while independently coded arguments

remain approximately constant, form itself changes imitation. The setting is speech in bounded financial tasks, not AI-written scholarship. Its implication here is about decoupling rather than credulity.

Readers need not become any more sensitive to form. The relation between form and evidence need only change. Rich prose was once produced disproportionately by authors who also possessed strong evidence. If cheap generation breaks that association, the same reader response begins to reward explanations whose evidence has not improved. This hypothesis applies only if readers discount fluency more slowly than fluency becomes cheap. Rational adjustment is the counterforce in the fourth prediction, and its speed remains untested. Nor does lower screening accuracy prove that readers adjust too slowly: accuracy can fall even when evaluators correctly abandon a signal that has lost information.

Signalling theory states the mechanism. In Spence (1973), an action conveys information because it is costly enough that only some senders take it. What everyone can afford, no one can signal with. Clean prose could play two roles in science: helping readers understand and indicating an author's command of the subject. Once a model can polish anyone's prose, the second role becomes less reliable.

A model of generative communication gives the same conditional result. Lower production and reading costs can improve communication while making it harder for a receiver to distinguish high-quality messages (Gans 2024). The effect is not automatic: it depends on how the technology changes the costs and informativeness of the signal.

Attenuation follows; inversion does not. Adoption selected only by language cost cannot reverse the relation. Under the first prediction's assumption that the cost of English is unrelated to evidence quality, adopters are an average draw. Moving them into the fluent group pulls the pooled relation toward zero, not below it. The subgroup result in Section 7 weighs against inversion by another route. When poor English had concealed applicants' quality, access made their prose *more* informative, not less. The first prediction expects those writers to adopt most, so adoption does not concentrate among a group whose new fluency is misleading. Inversion requires something further: adoption must concentrate among authors with weaker evidence, not merely among authors for whom English costs more. This article does not claim that it does.

The accompanying derivation states this result as a condition on pooled covariance. A lower covariance does not by itself establish a lower correlation, a less informative posterior or worse optimal screening, because the variance of prose and the reader's updating can change too.

Experimental evidence can support the attenuation prediction. In the screening experiments of Cowgill, Hernández-Lagos and Wright (2026), discussed in Section 7, access changed how much evaluators learned from prose in both directions. Galdin and Silbert (2025) reach a parallel result for job applications: once generative AI makes cover letters cheap, written material loses its ability to separate candidates and matching becomes less meritocratic. Kusumegi et al. (2025) add a suggestive observational pattern. Complex writing predicts a higher probability of publication among manuscripts without detectable AI assistance, as it long has; conditional on detected assistance,

the relation reverses. But that subsample is endogenously selected, meaning that its composition is not random, and detector errors vary with the same textual features. Composition alone could produce the flip. The pattern is consistent with the experiments, but adds little independent support because the named biases can generate it unaided.

The response is not to make prose costly again. Prose remains the medium through which readers understand and remember an argument. Certification should move closer to claims and evidence because fluency is now a poorer proxy for them. Whether editors and readers actually reweight in this direction is a separate empirical question. A current economic analysis of AI-assisted science likewise treats human judgment about which claims are true and which questions are worth pursuing as the binding constraint (Agrawal, McHale and Oettl 2026).

A field-level effect follows. Generative AI can increase the measured creativity of one person's output while reducing diversity across people (Doshi and Hauser 2024). Sourati, Ziabari and Dehghani (2026) synthesise related evidence of convergence in language and reasoning; Section 7 described a similar movement in cultural style. The texts themselves differ in a related way. Jiang and Hyland (2025) find that essays produced by ChatGPT are organised at least as clearly as essays by British university students, but carry markedly fewer of the markers by which a writer signals how strongly a claim is held. Their corpus is student argumentative writing rather than research articles. These findings concern outcomes adjacent to scholarly narrative, not direct measures of it. They establish convergence in expression and framing. A stronger claim, that errors and omissions become correlated across papers, requires a defective shared starting point and verification that fails to catch the defect. Both links are testable; neither has been tested.

The distinction has a formal counterpart in models of algorithmic monoculture. Reliance on a common evaluator can reduce social welfare even when it improves each decision maker's isolated accuracy, because the resulting errors are shared rather than independent (Kleinberg and Raghavan 2021). The application here is to generated starting points, not automated hiring decisions.

In this setting, diversity provides insurance. When no one knows the correct framing, each rival account preserves an interpretation that later evidence may vindicate. Convergence is efficient right up to the moment it is wrong. Rapid convergence may improve average clarity while reducing the chance that anyone retains the account eventually shown to be right. Large fields already concentrate attention on a small set of canonical papers and framings (Chu and Evans 2021). Shared generative systems may intensify that tendency.

The fifth prediction is therefore that shared generative defaults narrow the questions and explanations examined by a field even as individual output rises. Hao et al. (2026) report an adjacent pattern among 41.3 million natural science papers. Scientists classified as users of AI tools publish 3.02 times more papers, receive 4.84 times more citations and become research leaders 1.37 years earlier than non-users. Yet their papers cover 4.63 percent less topical ground than comparable non-AI work, measured by the spread of paper embeddings, numerical representations of content whose distance measures similarity. Later papers building on the same AI-assisted article engage 22 percent less with one another, while AI-assisted work concentrates in data-rich areas.

Most observations in the study concern earlier machine learning rather than generative writing tools, so the result applies broadly to AI-augmented science. The association has the form the model predicts: higher measured output and citations alongside narrower topical coverage. But the comparisons lack a causal counterfactual, and selection by adopters into data-rich fields is composition, not congestion. The fifth prediction still lacks a direct test.

One final distinction matters. Proposition 1 concerns private iteration: many candidates within one research process, followed by selection of a final output. It does not imply that a field benefits from more submissions. Reader and referee attention is finite. If cheaper production increases submissions faster than it reduces the cost of screening and certification, valuable work becomes harder to find even if average paper quality does not decline. The problem is old. Faced with an abundance of print, early modern scholars developed indexes, compilations and reference genres as certification devices for cheap text (Blair 2010). The socially useful rule is to generate many, verify carefully and certify one, not to publish everything that cheap generation permits.

9 A Boundary for Use

The analysis does not issue a general permission or prohibition. It suggests a sequence of work. That sequence is a defensible heuristic, not the optimum of a welfare calculation. It carries its own costs: producing a first account takes effort, while delaying the model sacrifices information. Its elements should therefore be treated as testable practices, not settled rules.

The sequence begins with the author's own account. Before requesting finished prose, she records the research question, principal claim, evidence, causal structure and unresolved uncertainties in a form that can later be checked. Elegance is unnecessary. The purpose is to create a standard of comparison that does not depend on the generated draft.

This requirement resembles the active decisions used to counter procrastination in retirement saving (Carroll et al. 2009). It also acts as a soft commitment, raising the cost of accepting a draft without examination while leaving the final choice open. Whether such a non-binding device works is a conjecture. Sophisticated present-biased agents reveal demand for commitments that bind (Augenblick, Niederle and Sprenger 2015). The remedy is therefore conditional. In Carroll et al. (2009), active choice outperforms a default when procrastination is strong and the chooser knows her preferences; a default can dominate when she lacks the competence to choose. An own-account-first practice fits scholars better than beginners, who need support alongside it.

Two further practices concern the candidate set. First, generate alternatives before polishing one of them: request incompatible explanations and the evidence that would distinguish them. A single polished narrative is a stronger reference point than a visible set of rivals. Second, separate generation from evaluation. Check claims against primary sources and data, preferably before seeing the model's ranking. A link helps verification only if someone opens the source.

Where formation is the objective, assistance should follow an attempt. Criticism and post-solution explanation preserve more of the generation process than an answer-first interface. The

expertise gradient in Section 4 determines how much worked structure a true beginner first needs. Periodic unaided tasks should then measure what remains; retrieval also strengthens retention (Roediger and Karpicke 2006). Assessment should distinguish performance on nearby variants from transfer to genuinely new problems (Barnett and Ceci 2002). A student who succeeds only while the tool is available has demonstrated assisted production, not yet independent judgment.

Finally, disclosure should identify the delegated task. To say only that AI was used is too coarse. Readers need to know whether a system corrected grammar, generated prose, proposed claims, summarised evidence, selected a narrative or performed a check. Accountability becomes more meaningful when the delegated operation is visible. Present practice falls far short. In one full-text sample of roughly 75,000 papers published since 2023, about 0.1 percent disclosed any AI use in writing. By early 2025, there was roughly one disclosure for every forty papers showing statistical evidence of AI assistance (He and Bu 2026). This ratio inherits the detector error discussed in Section 7. Low disclosure in the presence of formal requirements is consistent with social or institutional costs of disclosure, but the study does not identify a mechanism.

The article's signalling argument reveals the standard's weakness. A disclosure that cannot be checked is a cheap message, and cheap messages are not informative. Detection is unreliable, and journal policies were not associated with measurably different trends. Task-level disclosure is therefore an aspirational norm resting on professional honesty, not a verification device. By its own test, the disclosure attached to this article is a promise, not a proof. What a journal can review is the chain from claim to evidence, and that is where its effort is better spent. Disclosure and verification should not be confused.

For journals, then, the reviewable object should include that evidentiary chain. Precise citations and available data become more important as prose becomes cheaper. Author responsibility remains essential. Editors should use neither AI detection as a proxy for quality nor fluency as a proxy for truth.

These practices remain hypotheses, supported to different degrees by adjacent evidence. In one narrative-persuasion experiment, private reasoning alone did little while warnings and incentive disclosure did more. Warnings did nothing measurable in the writing-assistant experiment of Section 5, so their value is unsettled. The proposed boundary therefore requires direct tests in scholarly writing: designs that compare sequences of work over time rather than isolated outputs.

10 Conclusion

Writing is not identical to thinking. It is a collection of practices through which people externalise thought and, in doing so, change it. Some practices perform linguistic execution. Others form the author's judgment or shape the reader's beliefs. Collapse them into a single activity and the resulting rule will be too permissive or too restrictive.

The five predictions show what can presently be said. Preprint evidence is consistent with the search model: the largest output gains appear among researchers writing in a second language and

estimated journal adoption rose fastest where English was most costly. Yet both results depend on detection of a kind that errs most for those same writers and neither measures argument quality. Screening experiments show fluent prose becoming a less reliable signal of quality. Reversal rather than attenuation requires a selection condition the article does not establish. Comparisons without a causal counterfactual place individual productivity gains alongside narrower topical coverage. In mathematics education, assisted output and unaided capability move in opposite directions; whether scholarly writing behaves the same way remains unknown. Neither part of the third prediction has been tested directly on scholars: that an initial generated draft draws uncertain authors toward its framing and that the immediate effort of replacing it makes them stop before their reflective standard would.

One conclusion nevertheless follows from all five predictions. Judgment is not a free resource distributed equally among users. It is human capital built through practice. Oversight therefore has a cost and a history; it is not a label placed on the final click.

Thamus was right that writing would weaken some forms of memory, but he did not foresee every capacity that external memory would support. The lesson is neither that technological fears are invariably mistaken nor that every displaced skill deserves preservation. An institution should ask which capacity is disappearing, which is being created and whether the new workflow reproduces the judgment on which it relies.

For scholarly writing, this reasoning makes one short sequence defensible. It does not establish the sequence as optimal: in the closest experiment, private reasoning before exposure offered little protection. The proposal remains a practice to test. The author first forms her own account of the problem. The model then reduces language costs and supplies criticism. Verification remains independent.

The appropriate boundary varies with the author and the stage of inquiry, but no single measure can locate it. The manuscript produced today is the visible outcome. Alongside it belongs the judgment its author retains next year. A third outcome is the range of explanations the field can still entertain, conditional on enough authors beginning from the same generated point for their framings to become correlated.

These claims can be disproved. If authors who receive a generated draft first later perform just as well without the tool as authors assisted only after their own attempt and if they do not converge more strongly on the initial framing, then the proposed formation and reference-point costs are small. If an own-account-first requirement provides no protection, it should be abandoned. The relevant test follows sequences of work over time. The quality of assisted prose alone cannot reveal what its author has learned.

The historian from the opening still has her awkward document. If she writes her own account first, the model can improve it and lower the cost of its prose. If she reads the model's account first, the document may never disturb the familiar story. No finished manuscript reveals the difference between those two mornings of work.

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.

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

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, 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.

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.

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.

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.

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.

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

Frye, Northrop. 1957. "The Realistic Oriole: A Study of Wallace Stevens." *The Hudson Review*, 10(3): 353–370. JSTOR 3848061.

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.

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

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.

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.

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

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.

Kőszegi, Botond, and Matthew Rabin. 2006. "A model of reference-dependent preferences." *Quarterly Journal of Economics*, 121(4): 1133–1165.

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.

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.

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, 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, 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.

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.

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.