NOT PEER REVIEWED

Coach not crutch: AI assistance can enhance rather than hinder skill development

Benjamin Lira[1], Todd Rogers[2], Daniel G. Goldstein[3], Lyle Ungar[1], and Angela L. Duckworth[1]

[1]University of Pennsylvania; [2]Harvard University; [3]Microsoft Research

Corresponding author: blira@upenn.edu

Does using AI make you stupid? Contrary to predictions by lay forecasters, participants randomly assigned to practice writing with access to an AI tool improved more on a writing test one day later compared to writers assigned to practice without AI—despite exerting less effort. Likewise, whereas forecasters predicted that AI would be less helpful, and less worth paying for, than personalized feedback from experienced human editors, we found the converse: practicing with AI increased skill more than feedback from human editors (or simply googling examples and tips). The results of a third experiment suggest that AI may teach by example: writers merely shown an AI-generated cover letter (without the opportunity to practice) performed as well as writers who practiced writing with the original AI tool. Collectively, the findings of these pre-registered studies constitute an existence proof that by providing personalized, high-quality examples, AI can improve, rather than undermine, learning.

Human Capital Development | Generative AI | Skill Acquisition

Generative AI (henceforth AI) tools are increasingly powerful and prevalent (1), and there is mounting evidence that they can dramatically boost performance. For example, working side-by-side with AI as a copilot has been shown to increase both quality and speed in a variety of professional writing tasks (e.g., emails, memos, short reports) (2–4).

Nevertheless, there is growing concern that AI tools will be used as a crutch, providing immediate gains in performance at the expense of long-run development of human capital (5–7). For instance, in a 2024 poll, 62% of surveyed adults predicted that Generative AI would "lead to humans becoming less intelligent" (8). In January 2023, New York City public schools banned ChatGPT, citing "concerns about negative impacts on student learning (9)." When this ban was lifted three months later, it was not because of AI's potential to scaffold learning, but instead because of the "reality that students are participating in and will work in a world where understanding Generative AI is crucial (10)." However, teachers remain worried: in a nationally representative poll in May 2024, four times as many K-12 educators judged the use of AI tools as net harmful (24%) than net beneficial (6%) (11).

Concerns that using AI tools hinders learning (while increasing short-term performance) are justified for at least three reasons. First, AI systems based on large language models like GPT-4 have been shown to confidently assert erroneous facts (i.e., hallucinations (12)), make reasoning and arithmetic errors (13), and complete other tasks with varying degrees of accuracy.

Second, the fluent and instantaneous solutions AI tools generate may contribute to an illusion of mastery. To the extent users conflate the skills of an AI tool with their own, they may be less likely to seek feedback and improve. Prior research has found that searching for information on the Internet, for example, creates an illusion whereby people conflate knowledge outside their heads with what they personally know themselves (14).

Third, technological tools reduce the learner's need to be cognitively engaged with the task at hand. For instance, knowing that one can search for a fact on a computer has been shown to reduce memory for that fact, instead encouraging recall of how to search for it (15). And drivers who use GPS tend to have worse hippocampal-dependent spatial memory, both cross-sectionally and longitudinally (16, 17). When AI-powered tools instantaneously produce turn-key solutions for complex cognitive tasks, they may be especially detrimental to learning. It is, after all, tempting to copy and paste the output of an AI tool without even laying eyes on it.

On the other hand, using AI tools to assist us in our work may, in fact, help us develop our own skills. In particular, the current generation of AI tools may teach by example, offering high-quality and personally tailored demonstrations of abstract principles that are otherwise difficult to grasp. Classic research shows that worked examples of math problems (i.e., not just answers but the step-by-step process by which problems are solved) scaffold learning more effectively than

Significance Statement

Does using AI make you stupid? It is widely believed that relying on AI tools inadvertently cheats us of the opportunity to develop our own capabilities. In this investigation, we examine the possibility that using AI—even while reducing effort invested in a task—may nevertheless enhance learning. Contrary to lay forecasts, participants who practiced writing cover letters with access to AI improved in writing skill more than participants who practiced without AI, participants who received personalized feedback from professional editors, or participants given access to a Google Search for cover letter examples and tips. Forecasters were not wrong about effort: participants worked longer and harder when practicing without access to AI. However, AI-related decrements in effort appear to have been offset by an increase in the utility of the practice task itself: writers merely shown an AI-generated cover letter (without any opportunity to practice) improved in skill as much as writers who practiced writing with the original AI tool. Taken together, these findings constitute an existence proof that, by providing personalized, high-quality examples, using AI tools can improve, rather than undermine, learning.

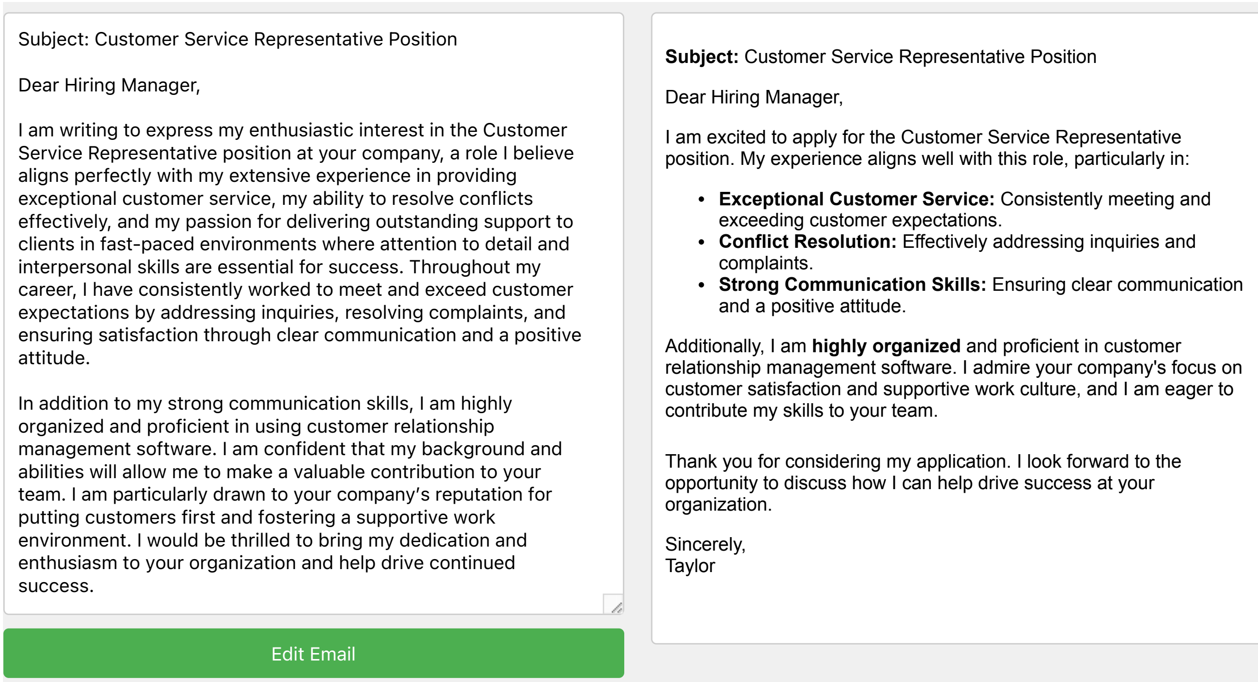

Fig. 1. The AI writing tool we created for this investigation takes inputted text (left panel) and generates a version that incorporates recommended writing principles (right panel).

explanations alone (18, 19). It seems likely that, compared to textbooks, conventional computer tutoring programs (e.g., (20)), and human teachers, AI tools may provide higher quality, just-in-time examples exquisitely tailored to a learner's specific needs. Indeed, it's possible that in both quality and didactic utility, AI examples (e.g., of an excellent cover letter) might surpass even those provided by domain experts (e.g., professional writers and editors). Thus, if the upside of exposure to excellent examples tailored to the learners' needs outweighs the downside of diminished engagement, AI tools may support skill development.

The scant research on how AI tools influence skill development has yielded mixed results. Some studies report that interacting with AI tools improves skill on subsequent tests in which AI tools are not available (21, 22), while others have shown null or even negative effects (22–24). Notably, these studies examine AI tutors, chatbots, or explanations explicitly designed to support learning, rather than simply providing solutions, as is typical in real-world use. Further, they focused exclusively on mathematics and computer programming.

In this investigation, we ask whether AI tools can support skill development, above and beyond merely improving performance while using the AI tool. We focus on writing—the most common use of AI at work, as ranked in a nationally representative survey of American adults in August 2024 (26). In our experimental paradigm all participants are given a baseline writing test (i.e., revising a poorly written cover letter), followed by a short lesson on five evidence-based strategies for professional writing (25, 27). Next, participants are randomly assigned to different types of practice (e.g., practice with access to an AI tool, practice without access to an AI tool, practice with feedback from professional editors, and practice with Google Search to find examples and tips). Finally, all participants complete an incentivized test in which they rewrite cover letters without access to AI.

In Study 1, forecasters presented with this design were twice as likely to predict that practicing writing with the assistance of AI would impair learning compared to practicing without it. In Study 2, however, participants who practiced with AI actually learned more (i.e., wrote better cover letters during the test phase) than participants who practiced without AI—an advantage that persisted in a one-day follow-up test. In Study 3, forecasters predicted that feedback from AI would be less helpful than feedback from professional editors and, accordingly, were less willing to pay for it. However, in Study 4, participants who practiced with AI learned more than participants who practiced with feedback from professional editors (or those who practiced with Google Search to find examples and tips). Finally, in Study 5, we explored the mechanism for these learning gains by introducing an example-only condition. Participants who saw an AI-generated example (but did not have an opportunity to practice) improved their writing as much as participants who practiced with AI; benefits again persisted in a one-day follow-up test.

Study 1: Forecasting the effect of AI on learning

We showed $N = 150$ participants screenshots of a random assignment study with three conditions. We asked them to rank-order these conditions according to how much they predicted future participants would learn in each.

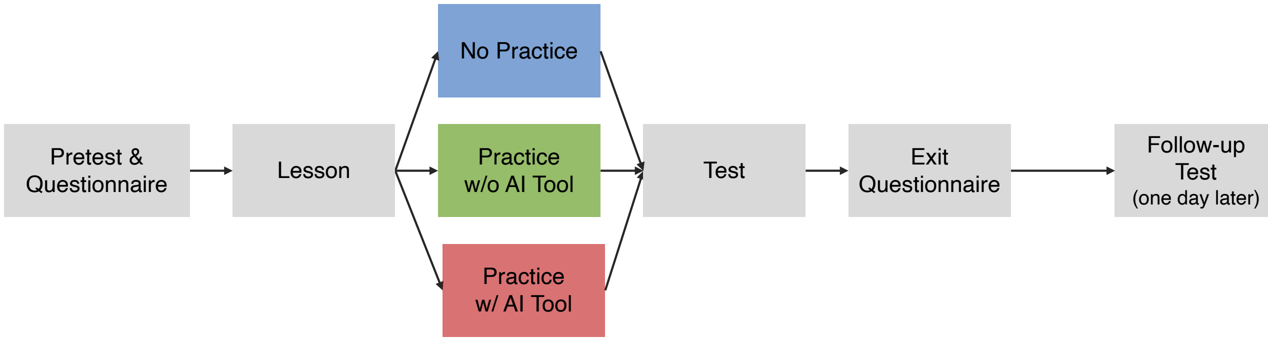

Fig. 2. Experimental design for Study 2. First, all participants completed a baseline questionnaire, a pretest (rewriting a poorly-written cover letter), and a lesson introducing five evidence-based principles of effective writing (25). Next, participants were randomly assigned to one of three conditions: practicing with an AI writing tool, practicing without an AI writing tool, or no practice. Then, all participants were tested on writing skill (rewriting a new cover letter without access to AI) and completed an exit questionnaire. Finally, to assess the persistence of skill improvement, participants were invited to complete a similar incentivized test of writing skill one day later.

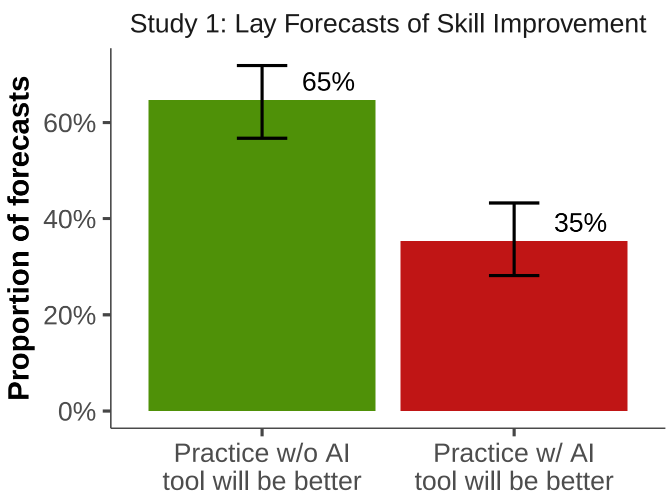

Fig. 3. Forecasters in Study 1 predicted that practicing without an AI tool would improve writing skill more than practicing with an AI tool. Error bars represent proportions $\pm$ 1 SE.

Lay forecasters predicted that practicing with AI would hinder learning. Confirming our pre-registered hypothesis, nearly twice as many forecasters (64.7%) ranked practicing alone above practicing writing with access to an AI tool as the converse (35.3%, χ^2 (1) $= 12.9$, $p < .001$), see Figure 3. Participants made this prediction regardless of self-reported experience with AI ($OR = 0.72$, $p = .096$) or any other measured demographic characteristic ($ps > .05$). See Table S2

In open-ended responses, forecasters who were pessimistic about the effect of the AI tool on learning speculated that it would crowd out effort (e.g., "Practicing alone would force more recall and problem-solving skills, while AI essentially just gives them the answer.", "I think oftentimes using AI impedes the learning process because it's the 'easy way."'). Those with positive views, on the other hand, cited the possibility of AI providing insights or examples that would be otherwise unavailable ("As much as I hate AI, I do not believe you can improve in any manner if you do not have examples or other ways of learning, and AI can provide this.")

Study 2: Effect of practicing with AI on learning

Study 2 tested whether the predictions of Study 1 forecasters were accurate. Specifically, $N = 2{,}238$ participants completed a baseline questionnaire and pretest (rewriting a poorly written cover letter), followed by a lesson introducing five principles of effective writing (i.e., Less is more, Make reading easy, Design for easy navigation, Use enough formatting but no more, Make responding easy) (25). Next, participants were randomly assigned to one of three practice conditions: (1) rewriting a new cover letter with an AI writing tool that instantly revised text based on these principles, (2) rewriting the new cover letter without the AI tool, or (3) a no-practice control. At the end of the session, all participants completed a test of writing skill (rewriting a new cover letter without access to the AI writing tool) and an exit questionnaire. One day later, all participants were invited to complete a similar incentivized test of writing skill. See Figure 2.

We assessed writing skill in two ways. First, we used GPT-4o to rate cover letters on each of the principles of effective writing and then averaged these ratings to produce a summary score (Cronbach's $\alpha = .81$). This operationalization demonstrated convergent validity with ratings by human research assistants who rated a random subsample of $n = 100$ cover letters on the same principles ($r = .70$, $p < .001$). Second, we recruited participants naive to this investigation to read two letters and select the one more likely to secure its writer a job interview. These two operationalizations converged: preferred cover letters obtained higher AI ratings of writing skill ($rs = .29$).

AI practice improved writing skill. Consistent with other studies demonstrating the productivity benefits of AI tools (2, 3), participants given access to the AI writing tool produced cover letters during the practice phase that were dramatically higher in quality than participants without access ($d = 1.01$, $p < .001$).

The learning advantage of having practiced with AI was evident in the test phase: Consistent with our preregistered hypothesis, participants who had practiced with the AI tool produced higher-quality writing than did participants who

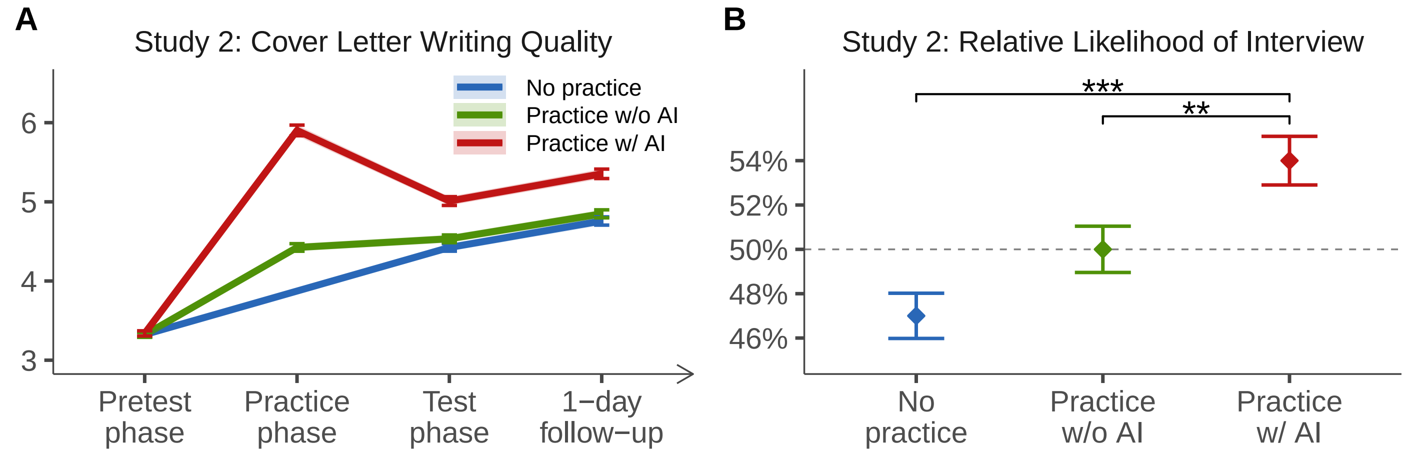

Fig. 4. A. Cover Letter Writing Quality, In both the main test and the follow-up of Study 2, participants who had practiced with AI outperformed those who practiced without it and those who did not practice at all. Error bars represent means $\pm$ 1 SE. Means shown are for the subsample of participants (n = 1,294) who completed the one-day follow-up test. See Figure S3 for the equivalent figure in the full sample. **B. Relative Likelihood of Interview.** During the test phase of Study 2, cover letters written by participants who had practiced with AI were judged more likely to generate hypothetical interview offers than those from other conditions. Error bars represent proportions $\pm$ 1 SE.

either had practiced without the AI tool (d = .38, p <.001) or who had not practiced at all (d = .47, p <.001). See Figure S3. Likewise, cover letters written by participants who had practiced with AI were more likely to secure a hypothetical job interview than cover letters by participants who had practiced without AI (.54 vs. .50, p = .002) or had not practiced at all (.54 vs. .47, p < .001). See Figure 4B.

AI practice was less effortful. Compared to participants who practiced alone, participants who practiced with the AI tool spent 0.44 fewer minutes during the practice phase (3.73 vs. 4.17; $d = -.12$, p = .025, quantity of effort), logged roughly a quarter as many total keystrokes (26% $d = -.44$, p <.001) and logged 65% fewer keystrokes per minute ($d = -.22$, p <.001, intensity or quality of effort), and self-reported expending less effort during practice ($d = -.31$, p <.001).

Nevertheless, it would be inaccurate to label writers practicing with AI as entirely disengaged. Rather than copying and pasting the AI tool's output, the majority of participants practicing with AI chose to interact with the task for over 3 minutes, and 95% made at least one edit to the AI-generated output before final submission. See Figure S4 in Supplementary Information for details.

AI practice did not create the illusion of mastery. In the exit questionnaire, participants who had practiced with AI reported learning about as much as participants who had practiced without it or not practiced. Likewise, they judged their writing skill similarly to those who had either practiced alone or done no practice at all ($|ds| \leq 0.10$, ps >.05). When asked if they wanted to see additional feedback, they were about as likely to say yes as participants who had practiced without AI (.64 vs. .60, p = .167), and slightly less likely than those who had not practice at all (.65 vs. .60, p = .039). See Section C4 in the Supplementary Information for details.

The benefits of practicing with AI were just as large a day later. To examine whether the treatment effects persisted over time, we re-contacted all participants one day later. Most participants responded (87%), and attrition rates did not differ by condition (13% to 14% χ^2 = .68, p = .710). Confirming our pre-registered hypothesis, participants who had practiced with the AI tool the previous day continued to outperform those who had practiced without the tool (d = .41, p <.001) as well as those who had not practiced at all (d = .46, p < .001). Participants who had practiced without the AI tool performed no better than those who did not practice (d = .05, p = .331). The magnitude of the follow-up effect size was not significantly different from the immediate test phase effect size (condition $\times$ time interactions were non-significant (ps > .69). See Figure 4A and Section C5 of Supplementary Information for details.

Results were not moderated by individual differences. None of the findings above were moderated by individual difference variables, including past experience with AI, age, gender, race, education, motivation to learn, and baseline writing skill, BH-corrected p-values >.05. See Table S13 in the Supplementary Information for details.

Study 3: Willingness to pay for AI vs. human editors

Given unwarranted pessimism about the pedagogical power of AI, in Study 3 we examined whether people underinvest in AI as a learning resource. Specifically, we asked N = 150 people to imagine they were participants in Study 2 and asked them how much they would be willing to pay for feedback from the AI tool vs. feedback from experienced human editors.

Participants were willing to pay more for feedback from expert human editors than for AI feedback. Out of a maxi-

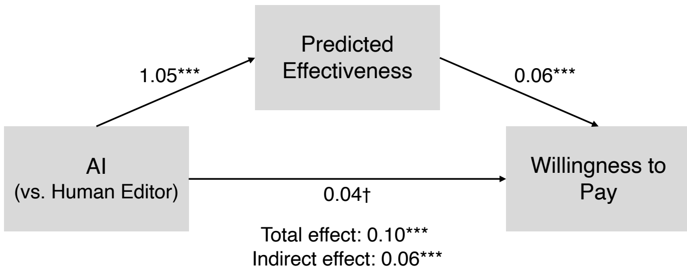

Fig. 5. Predicted effectiveness mediated underinvestment in AI feedback

mum of a dollar (the total amount of the hypothetical bonus participants could earn), participants on average were willing to pay 28 cents for the AI tool vs. 38 cents for feedback from an experienced human editor ($t(149) = -4.83$, $p <.001$, $d = 0.39$).

Participants predicted that editorial feedback would be more beneficial. Participants predicted that experienced human editors ($M = 6.68$ on a 0—10 Likert scale) would be more effective in helping them improve their writing, compared to feedback from an AI tool ($M = 5.64$; $t(148) = -4.77$, $p <.001$, $d = 0.39$).

Predicted effectiveness mediated underinvestment in AI feedback. People's beliefs about how effective AI and human editor feedback would be predicted willingness to pay (b = 0.06, SE = 0.01, $p < .001$) and accounted for most of the preference for human editors (indirect effect = 0.06, $p < .001$). Once these predictions were accounted for, the remaining preference for human feedback was reduced by 60% (c' = 0.04, SE = 0.02, $p = .054$). See Figure 5 and Section D1 of Supplementary Information for details.

Study 4: Practicing with AI vs. with expert feedback (vs. with Google Search)

Participants ($N = 2{,}997$) completed a baseline questionnaire and pretest, followed by the same lesson on effective writing as in Study 2, and then practiced rewriting a new cover letter on their own. Next, participants were randomly assigned to one of three conditions: (1) practice with AI, (2) practice with feedback from professional human editors, (3) practice with Google Search. Letters written by participants assigned to the second condition were given to professional editors averaging 25 years of professional editorial and writing experience. These 49 editors, recruited in partnership with the Journalists' Resource at Harvard Kennedy School, provided personalized revisions and constructive feedback. The following day, all participants attempted to improve the cover letter they had previously submitted, using the resources afforded by their assigned condition. Finally, all participants completed a writing test (rewriting a cover letter without access to AI) and an exit questionnaire.

Practicing with AI improved writing skill more than feedback from professional editors and Google Search. During the practice phase, participants working with AI outperformed participants who received personalized feedback from professional editors ($d = 0.76$, $p <.001$) and those assigned to Google cover letter examples and tips ($d = 1.03$, $p <.001$).

During the test phase, in which participants were asked to write a new cover letter without access to any outside resources, participants who had practiced with the AI tool produced higher-quality cover letters than those who had practiced with Google Search ($d = 0.46$, $p <.001$), and even outperformed those who had practiced with personalized feedback from professional editors ($d = 0.20$, $p <.001$). See Figure 6A.

Cover letters written by participants who had practiced with AI were as likely to secure a hypothetical job interview as cover letters by participants who had practiced with editor feedback (.51 vs. .51, $p = .843$), and more likely to than those written by participants who had practiced with Google Search (.51 vs. .47, $p = .024$). See Figure 6B.

A possible concern is that reviewing many letters consecutively caused editors to get tired and give worse feedback over time. This could suggest that our experiment compared AI against a "strawman" version of human expertise. However, analysis of the editors' performance revealed the opposite pattern: while editors became more efficient, spending less time per letter as the task progressed ($b = -0.387$, $p < .001$), the quality of their edits and feedback actually improved ($b = 0.013$, $p < 0.001$). Thus, the AI tool's advantage over professional editors cannot be attributed to a decline in editorial performance. See Section E2 of Supplementary Information for details.

Practicing with AI was not more effortful than getting feedback from professional editors and was less effortful than looking for examples online. During the practice phase, participants with access to AI spent about the same time as those who practiced with editor feedback (difference = 0.43 minutes, $d = 0.09$, $p > .05$), but about 1.18 fewer minutes than those who practiced with Google Search ($d = 0.32$, $p < .001$). Nevertheless, participants across all three groups reported similar levels of subjective effort (ps $> .05$). Likewise, participants who practiced with AI logged about as many keystroles as those who practiced with editor feedback (68% fewer, $d = 0.08$, $p > .05$) and 76% fewer keystrokes than those who practiced with Google Search ($d = 0.32$, $p < .001$). They also typed at a slower rate than participants who practiced with editor feedback ($d = -0.41$, $p < .001$) or those who practiced with Google Search ($d = -0.34$, $p < .001$).

No evidence of AI-induced illusion of mastery. Participants who had practiced with AI did not report learning more than those who had practiced with editor feedback ($d = 0.07$, $p > .05$), nor did they consider their writing skills to be better ($d = -0.02$, $p > .05$). Both groups, however, correctly recognized that they had learned more than those who had

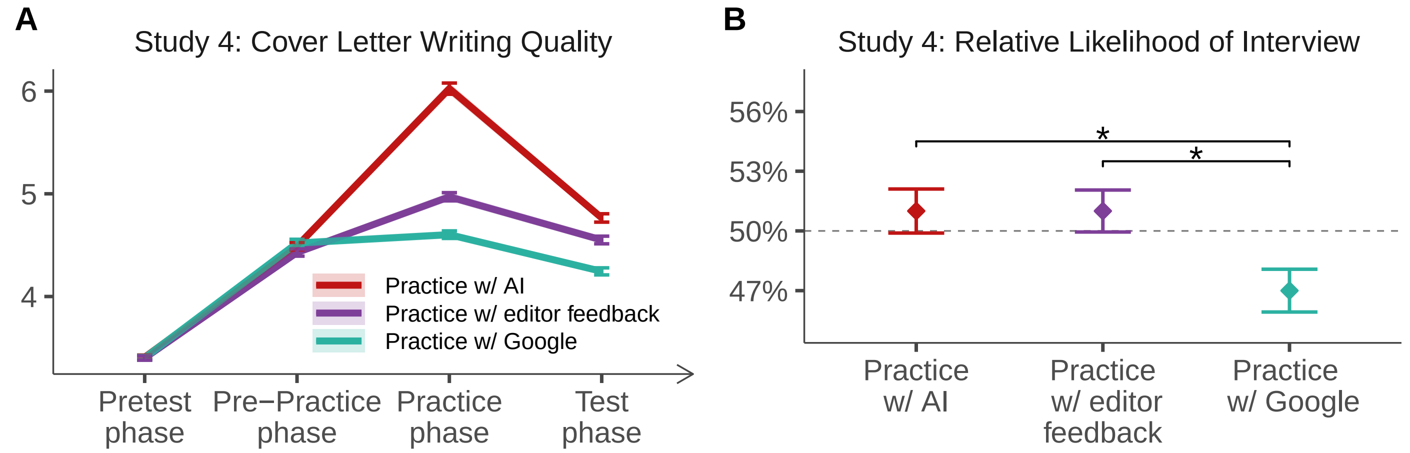

Fig. 6. A. Cover Letter Writing Quality. In the test phase, Study 4 participants who had practiced with the AI tool outperformed those who had practiced with either feedback from professional editors or access to Google Search. Error bars represent means $\pm$ 1 SE. **B. Relative Likelihood of Interview.** In the test phase, cover letters written by participants who had practiced with AI were about as likely to generate hypothetical interview offers as those written by participants who had practiced with feedback from professional editors. Cover letters from both of these conditions outperformed those written by participants who had practiced with Google Search. Error bars represent proportions $\pm$ 1 SE.

practiced with Google Search (ds = .28 and .20, respectively), and evaluated their writing skills accordingly (ds = .11 and .12, respectively). Participants who had practiced with AI and those who had practiced with editor feedback requested feedback on their test submissions at similar rates (71% and 70%, respectively). Both groups were more likely to request feedback than those who had practiced with Google search (64%, p = .003 and .007, respectively). See Section E6 in the Supplementary Information for details.

Results were not moderated by individual differences. As in Study 2, the above findings were consistent across all measured individual differences. See Table S22 in the Supplementary Information for details.

Study 5: Examples as the mechanism for AI learning gains

To better understand what drives the benefit of AI practice, in Study 5 (N = 2,003), we preregistered a replication and extension in which we replaced the no-practice condition of Study 2 with an example-only condition. In this condition, we showed participants an AI-generated writing example that they could not edit. To the extent that the benefit of practicing with AI was driven by exposure to examples, the example-only condition should improve performance in the test phase as much as the practice with AI condition.

Seeing an AI example was as effective as practicing with AI. As in Study 2, participants given access to the AI writing tool dramatically outperformed participants who did not get access to it, both while using it during the practice phase (d = 1.22, p <.001), and during the no-AI test phase (d = .34, p < .001). Their test phase cover letters were also relatively more likely to secure them hypothetical job interviews (.51 vs. .47, p = .008).

Participants who had merely seen an AI-generated example also improved more in writing skill than those who had practiced without AI (d = .37, p < .001), and produced letters that were relatively more likely to secure them interviews (.52 vs. .47, p = .007). Notably, they improved as much as participants who had practiced with the AI tool (and could edit its output, d = .03, p = .883), and were offered hypothetical interviews at similar rates as them (.51 vs. .52, p = .561). See Figures S10 and 7B.

Seeing an AI example was even less effortful than practicing with AI. During the practice phase, participants who saw the AI example spent 2.32 fewer minutes than participants practicing with AI (d = 0.99, p < .001) and 2.96 fewer minutes than participants practicing without AI (d = 1.13 p = < .001), and reported expending less effort than those practicing with an AI tool (d = 0.19, p = .003) and those practicing without AI (d = 0.32, p < .001). As in Study 2, participants who practiced with AI logged 74% fewer keystrokes and 68% fewer keystrokes per minute (d = 0.15, p = .007) compared to participants who practiced without AI. As expected, participants exposed to the AI example logged 0 keystrokes.

Seeing an AI example did not create the illusion of mastery. As in Studies 2 and 4, despite learning more, participants who had practiced with AI or had merely seen an AI-generated example reported learning similar amounts to those who practiced without AI (ps > .05) and rated their writing skill after practice at comparable levels (ps > .05). Moreover, all participants requested feedback at similar rates (proportions ranged from 63% to 67%). See Section F4 in the Supplementary Information for details.

The benefits of seeing an AI example were just as large a day later. When we recontacted a subsample of participants

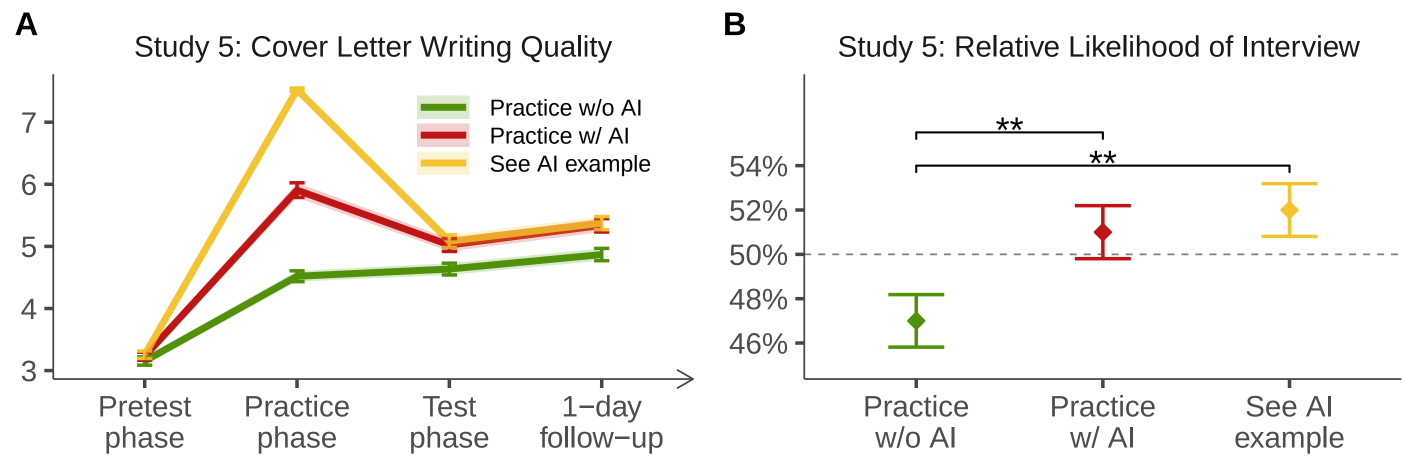

Fig. 7. A. Cover Letter Writing Quality. In both the main test and the follow-up of Study 5, participants who simply saw an AI-generated example improved just as much as those who practiced with AI and more than those who practiced without AI. Error bars represent means $\pm$ 1 SE. Means shown are for the subsample of participants (n = 608) who completed the one-day follow-up test. See Figure S10 for the equivalent figure in the full sample. **B. Relative Likelihood of Interview.** During the test phase, cover letters written by participants who had seen an AI-generated example were about equally likely to generate hypothetical interview offers when compared to those assigned to practice with AI. Cover letters from both AI conditions outperformed those written by participants assigned to practice without AI. Error bars represent proportions $\pm$ 1 SE.

($n = 800$) one day later, the majority responded ($n = 633$, 80%); the attrition rates ranged from 17% to 24% and did not differ by condition (χ^2 (2) $= 4.56$, $p = .102$). The effect remained robust after 24 hours. Participants who had practiced with the AI-tool ($d = .29$, $p = .006$) and participants who had simply seen an AI example ($d = .32$, $p = .003$), both continued to outperform those who had practiced without the tool. Participants who had merely seen an AI example performed as well as those who had practiced with AI ($d = .02$, $p = .830$). As in Study 2, the magnitude of the follow-up effect size was not significantly different from the immediate test phase effect size (condition $\times$ time interactions were non-significant ($ps > .96$). See Figure 7A and Section F5 of Supplementary Information for details.

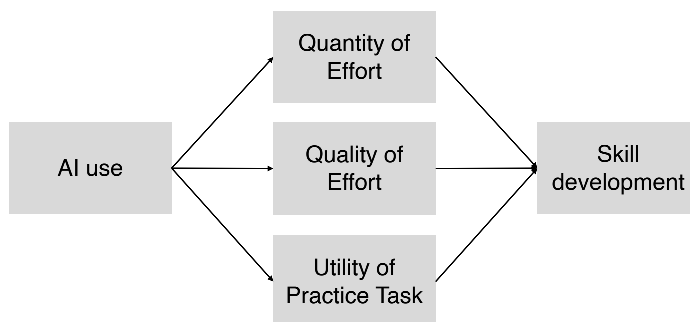

Fig. 8. Conceptual mechanism. Relying on AI may cause people to work less hard and for less time, but may nevertheless raise the utility of the practice task by providing high-quality, personalized examples. The increased task utility offsets lower effort, yielding greater post-practice writing skill.

Results were not moderated by individual differences. As in the previous studies, the above findings were consistent across all measured individual differences. See Table S35 in the Supplementary Information for details.

Discussion

Contrary to lay forecasters' expectations (Studies 1 and 3), participants who practiced writing cover letters with an AI tool learned more than those who practiced on their own (Studies 2 and 5), more than those who had access to Google cover letter examples, and even more than those who received personalized feedback from professional human editors (Study 4). Specifically, participants who had practiced with AI, subsequently wrote cover letters (without AI assistance) that were rated higher in writing quality and were more likely to secure a hypothetical interview—both immediately after practice and one day later. Learning gains were not the result of greater effort; in fact, participants who had practiced with AI expended less effort during practice than those who had practiced alone. Instead, these gains were at least in part explained by exposure to high-quality, just-in-time personalized examples: participants who merely viewed an AI-generated cover letter (without editing it) improved their writing skill as much as those given the option to practice editing the cover letter (Study 5).

Study 1 forecasters who believed AI would harm learning cited reduced effort as the mechanism—and prior research supports this concern (28, 29). All things equal, the more effort a learner invests in a given practice task, the more they learn. So why did practicing with AI improve writing skill more than practicing alone, even while reducing both the quantity and quality of effort? Likewise, why did writers who googled how to compose a cover letter work longer and harder than writers who practiced with AI, yet learn less? We propose a third mechanism that offsets decrements in human effort: the utility of the practice task itself. See Figure 8. Perhaps the excellent personalized examples provided by AI are analogous to demonstrations of an Olympic coach. The higher the utility of the practice task, the faster the learning.

While not immediately apparent to forecasters, teachers (11), and the general population (8), the pedagogical benefits of viewing AI-generated examples are consistent with prior research on learning. Research has shown that humans are especially adept at observing, imitating, and learning from others (30–32). Our findings also align with the expert performance literature: the most successful learners engage in deliberate practice, which (in addition to concentration, feedback, and repetition) depends upon detailed mental representations of excellent performance (33).

Future Directions. Three promising directions for future research are worth highlighting. First, it is not clear whether the observed benefits of AI examples would be as large in domains other than professional writing. At a glance, a single AI-generated example visually communicates the elements of effective professional writing (e.g., short sentences and boldface formatting). In other domains, however, merely observing a solution may be less informative. For instance, the final answer to a math or computer programming problem does not instantly reveal the procedure that produced it.

Second, additional research is needed to explore moderators of learning from AI tools. Certain metacognitive strategies may enhance the benefits of interacting with AI. For example, experimental research suggests that learners benefit more from AI explanations for math problems if they first try to solve them on their own (21). Similarly, correlational evidence suggests that asking AI for explanations as opposed to answers is associated with more learning in mathematics (23) and computer programming (22). Conversely, other factors may minimize learning benefits. In our experimental paradigm, participants practiced for as long as they wanted, with the foreknowledge that their skills would subsequently be tested (and rewarded monetarily) without access to AI. During practice, therefore, participants were incentivized to prioritize gains in acquired skill over performance in the moment. In real-world settings, there is often time pressure and competing incentives around performance and learning, which we speculate would reduce the learning gains associated with practicing with AI. More broadly, we believe concerns about AI as a crutch are justified when AI displaces effort without increasing task utility.

Third, in our experimental paradigm, participants interacted with the AI tool only once. It is common, however, to use AI tools repeatedly. When do repeated interactions lead to diminishing or even negative returns, and in what scenarios might skill development continue over time? Consider, for instance, the game of Go. The introduction of superhuman AI has been associated with an increase in the novelty and quality of decisions made by human Go players over time, with elite players reporting that they have been inspired by AI solutions they'd never seen before (34). Future research, ideally in field settings, is needed to establish the long-term benefits and costs of using AI tools.

Conclusion. Our findings should temper widespread concern that AI tools invariably boost momentary productivity at the expense of long-term skill development. Although the AI writing tool reduced the effort users invested in practicing, it nevertheless accelerated their skill development. The underappreciated efficacy of timely and tailored examples carries practical implications: Many AI tools designed to support learning are explicitly programmed not to "give away" answers. However, it may be that in addition to hints, leading questions, and explanations, learners benefit from demonstrations of the principles they are attempting to master.

Decades before the advent of generative AI, the legendary UCLA basketball coach John Wooden declared that the four laws of learning are explanation, demonstration, imitation, and repetition (35). Few learners have access to the best human teachers, coaches, and mentors, but generative AI now makes it possible to learn from personalized, just-in-time demonstrations tailored to any domain. In doing so, AI has the potential not only to boost productivity but also to democratize opportunities for building human capital at scale.

Methods

Ethical Considerations. The study was assessed by the University of Pennsylvania's IRB, and was approved before implementation (Protocol 853653). All participants completed informed consent.

Pre-registration. All studies were preregistered on https://aspredicted.org. Accordingly, the analyses presented in the main text were also pre-registered, with exceptions noted below.

Study 1 was preregistered (#191800). The moderation analyses were not pre-registered.

Study 2 was pre-registered (#201140). The pairwise comparisons analysis was pre-registered separately (#205316). After preregistration, we considered a keystrokes-per-minute metric of effort intensity rather than duration; thus, this analysis was not preregistered.

Study 3 was pre-registered (#236487). Analyses on predicted effectiveness of AI vs. human feedback and the mediation analysis were not preregistered.

Study 4 was pre-registered (#239157). The pairwise comparisons analysis was pre-registered separately (#243115). The keystrokes-per-minute metric of effort was not pre-registered.

Study 5 was pre-registered (#197704). The one-day follow-up was collected in three batches: a pilot, followed by second and third batches. We pre-registered batch 2 (#199451) but report pooled results in the main text, with separate batch details available in Supplementary Information Section F5. The pairwise comparisons analysis was pre-registered separately (#205315). The keystrokes-per-minute metric of effort was not pre-registered.

Participants. We sampled participants from Prolific from all our studies. We excluded all Prolific users who participated in one of the earlier studies from participation in subsequent studies.

Participants in Study 1 were predominantly female (n = 93, 62%), with ages ranging from 21 to 81 (M = 38.4, SD = 12.2). They were predominantly white (75%). A small proportion were students (13%), and most were employed (62%).

In Study 2, the sample was more evenly split between men (46%) and women (52%), ranging in age from 18 to 82 (M = 36.0, SD = 12.5). Over half of the sample (58%) was White, with the rest comprising Black (33%), Latino (6%), and Asian (5%) participants. Most participants (77%) had college degrees. Most participants (93%) were at least somewhat motivated to improve their writing, and had varying levels of experience with AI writing assistants (36% had tried them, but hardly ever used them, 47% used them at least a few times per week, and 17% had never used AI assistants before).

Participants in Study 3 were evenly split between men (52%) and women (48%), and their ages ranged from 18 to 75 (M = 42.4, SD = 14.4). They were predominantly white (67%). A small proportion were students (19%), and most were employed (56%).

In Study 4, the sample consisted of relatively fewer men (38%) than women (61%), and ranged in age from 19 to 92 (M = 41.5, SD = 13.8). Over half of the sample (74%) was White, with the rest comprising Black (16%), Latino (8%), and Asian (7%) participants. Most participants (74%) had college degrees. Most participants (94%) were at least somewhat motivated to improve their writing, and had varying levels of experience with AI writing assistants: 17% had never used AI assistants, 40% had tried them but hardly ever used them, and 43% used them at least weekly.

Study 5 had similar proportions of men (46%) and women (53%), and participants ranged in age from 18 to 95 (M = 37.9, SD = 12.6). Over half of the sample (64%) was White, with the remainder comprising Black (24%), Latino (8%), and Asian (6%) participants. Most participants (74%) had college degrees. Most participants (91%) were at least somewhat motivated to improve their writing, and had varying levels of experience with AI writing assistants (40% had tried them, but hardly ever used them, 42% used them at least a few times per week, and 18% had never used AI assistants before).

Procedure. All studies followed the same general structure with study-specific variations. Participants first saw an introductory screen and completed a brief questionnaire reporting demographics, prior experience with AI writing tools, motivation to improve their writing, and perceived writing skill. They then completed a short pre-test (2 minutes) in which they edited a poorly written cover letter, followed by a brief lesson on the five principles of effective writing.

In Studies 2 and 5, participants were then randomized to a practice condition (or skipped practice in the no-practice control). During practice, participants rewrote a new cover letter or, in the example-only condition, viewed an AI-rewritten version of the letter. This example was not explicitly labeled as AI-generated. After practice, participants saw either their own rewritten letter or the AI-generated example, along with AI-generated feedback highlighting one suggested improvement (see Supplementary Information, Section A3). Immediately after this feedback, participants rated how much they had learned and how hard they had worked so far. They then completed a 7-minute incentivized test of writing skill by editing a new cover letter without access to AI tools. To minimize the possibility of cheating, we used custom JavaScript to restrict copy-pasting functionality. Finally, participants could request optional feedback and were asked whether they had used outside resources during the test. A small percentage of participants (2.9%) admitted to doing so; as per our preregistration, these participants are included in analyses, though results excluding them are consistent with our main interpretation (see Tables S6, S18, and S27).

In Study 4, we adapted the procedure to compare AI practice with two ecologically valid human alternatives: feedback from professional editors and guidance from Google Search. Participants completed the same baseline questionnaire, pretest, and writing lesson. They then edited a new cover letter without any assistance and were randomly assigned to one of three conditions: (1) practice with AI, (2) practice with feedback from professional editors, (3) practice with a Google Search for cover letter examples and tips. In the second condition condition, participants' cover letters were forwarded to editors recruited in partnership with The Journalists' Resource at Harvard Kennedy School, who provided both written feedback and a rewritten version of the letter. The following day, participants were asked to revise their original cover letter using the resources from their assigned condition. All participants then completed a final writing test by rewriting a new cover letter without assistance, followed by an exit questionnaire.

Measurement. As per our pre-registration, we used OpenAI's GPT-4o to rate text samples for writing quality. To do this, we independently rated each cover letter and each writing principle. Research has demonstrated that large language models can provide ratings of writing quality that align closely with human judgments, offering reliability and consistency across various evaluation contexts (36, 37). See our prompts in Table S1. We then took the unweighted average of these 5 scores as our main dependent variable. See disaggregated analyses by each writing principle and additional outcomes on Tables S6, S18, and S27.

To validate these ratings, the first author and a trained research assistant took a sample of $n = 100$ cover letters from Study 2, and rated them on the 5 principles. The average of these two ratings correlated about as highly with the computer ratings (r = .70, p < .001) as they did with the average interrater reliability (r = .74, p < .001). These ratings had high internal consistency (Cronbach's αs = 0.81, 0.78, and 0.83 for studies 2, 4, 5, respectively.)

To address concerns that particular LLMs might be biased in favor of their own output, we also used Claude to rate the cover letters. We find that GPT ratings correlate strongly with Claude ratings ($r = .71$, $p < .001$), and that the effects are not attenuated by using different models (See Tables S6, S18, and S27), suggesting that our effects are not explainable by same-model bias.

Additionally, we asked a separate sample of participants to read pairs of test-phase cover letters randomly selected from different conditions, and to indicate which letter would be more likely to secure a job interview. Cover letters more likely to secure an interview obtained higher AI ratings of writing skill (rs = .29, .35, and .28, ps < .001, for Study 2, 4 and 5, respectively).

Statistical analysis. As per our pre-registrations, we fit ANCOVA models predicting outcomes from condition indicators, controlling for pretest score and baseline characteristics (age, gender, race/ethnicity, primary language, education level, motivation to improve writing skills, self-rated writing skill, experience with AI writing assistants, and baseline writing effectiveness). We used logistic regression to predict whether participants chose to see optional test feedback from condition, controlling for pretest score and baseline characteristics. Our analyses of the hypothetical hiring situation use beta regression, because the relative likelihood of a cover letter being preferred is bounded between 0 and 1. When correcting for multiple comparisons in exploratory moderation analyses, we used the Benjamini-Hochberg correction (38).

References

1. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4 (2023).
2. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
3. F. Dell'Acqua, E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, K. R. Lakhani, Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal* (2023).
4. E. Wiles, Z. T. Munyikwa, J. J. Horton, Algorithmic writing assistance on jobseekers' resumes increases hires, *Tech. rep.*, National Bureau of Economic Research (2023).
5. J. M. Hofman, D. G. Goldstein, D. M. Rothschild, A sports analogy for understanding different ways to use ai. *Harvard Business Review* **4** (2023).
6. S. Puntoni, R. W. Reczek, M. Giesler, S. Botti, Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing* **85**, 131–151 (2021).
7. B. Borges, N. Foroutan, D. Bayazit, A. Sotnikova, S. Montariol, T. Nazaretzky, M. Banaei, A. Sakhaeirad, P. Servant, S. P. Neshaei, J. Frej, A. Romanou, G. Weiss, S. Mamooler, Z. Chen, S. Fan, S. Gao, M. Ismayilzada, D. Paul, P. Schwaller, S. Friedli, P. Jermann, T. Käser, A. Bosselut, EPFL Grader Consortium, EPFL Data Consortium, Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences* **121**, e2414955121 (2024).
8. S. Hawkins, F. Duong, A. Fabrizio, D. Yudkin, Between hesitation & hope: America's mixed feelings on generative artificial intelligence (2024). Available online: https://www.canva.com/design/DAGRVonfEGk/W6C6mAYwwJmRhS2dqjjfVg/view#1.
9. K. Roose, Don't Ban ChatGPT in Schools. Teach With It. *New York Times* (2023).
10. D. C. Banks, ChatGPT caught NYC schools off guard. Now, we're determined to embrace its potential., https://www.chalkbeat.org/newyork/2023/5/18/23727942/chatgpt-nyc-schools-david-banks/ (2023).
11. L. Lin, A quarter of U.S. teachers say AI tools do more harm than good in K-12 education (2024).
12. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation. *ACM Computing Surveys* **55**, 1–38 (2023).
13. Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015* (2023).
14. M. Fisher, M. K. Goddu, F. C. Keil, Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General* **144**, 674–687 (2015).
15. B. Sparrow, J. Liu, D. M. Wegner, Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science* **333**, 776–778 (2011).
16. E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. J. Frackowiak, C. D. Frith, Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences* **97**, 4398–4403 (2000).
17. E.-M. Griesbauer, E. Manley, J. M. Wiener, H. J. Spiers, London taxi drivers: A review of neurocognitive studies and an exploration of how they build their cognitive map of london. *Hippocampus* **32**, 3–20 (2022).
18. J. Sweller, G. A. Cooper, The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction* **2**, 59–89 (1985).
19. R. K. Atkinson, S. J. Derry, A. Renkl, D. Wortham, Learning from examples: Instructional principles from the worked examples research. *Review of educational research* **70**, 181–214 (2000).
20. J. R. Anderson, A. T. Corbett, K. R. Koedinger, R. Pelletier, Cognitive tutors: Lessons learned. *The journal of the learning sciences* **4**, 167–207 (1995).
21. H. Kumar, D. M. Rothschild, D. G. Goldstein, J. Hofman, Math Education with Large Language Models: Peril or Promise? *SSRN Electronic Journal* (2023).
22. M. Lehmann, P. B. Cornelius, F. J. Sting, Ai meets the classroom: When does chatgpt harm learning? *arXiv preprint arXiv:2409.09047* (2024).
23. H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, R. Mariman, Generative AI Can Harm Learning (2024).
24. A. Nie, Y. Chandak, M. Suzara, A. Malik, J. Woodrow, M. Peng, M. Sahami, E. Brunskill, C. Piech, The gpt surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances, *Tech. rep.*, Center for Open Science (2024).
25. T. Rogers, J. Lasky-Fink, *Writing for Busy Readers: Communicate More Effectively in the Real World* (Penguin, 2023).
26. A. Bick, A. Blandin, D. J. Deming, The rapid adoption of generative ai, *Tech. rep.*, National Bureau of Economic Research (2024).
27. H. C. Shulman, D. M. Markowitz, T. Rogers, Reading dies in complexity: Online news consumers prefer simple writing. *Science Advances* **10**, eadn2555 (2024).
28. S. Melumad, J. H. Yun, Experimental evidence of the effects of large language models versus web search on depth of learning. *Available at SSRN 5104064* (2025).
29. N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, P. Maes, Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872* (2025).
30. A. Bandura, *Social Learning Theory* (General Learning Press, New York, 1971).
31. A. N. Meltzoff, Imitation and other minds: The "like me" hypothesis. *Perspectives on Imitation: Imitation, Human Development, and Culture*, S. Hurley, N. Chater, eds. (MIT Press, 2005), pp. 55–77.
32. D. E. Lyons, A. G. Young, F. C. Keil, The hidden structure of overimitation. *Proceedings of the National Academy of Sciences* **104**, 19751–19756 (2007).
33. A. K. Ericsson, Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine* **15**, 988–994 (2008).
34. M. Shin, J. Kim, B. Van Opheusden, T. L. Griffiths, Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences* **120**, e2214840120 (2023).
35. Charlie Rose : WHUT : June 14, 2010 9:00am-10:00am EDT (2010).
36. S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, J. J. Van Bavel, GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* **121**, e2308950121 (2024).
37. V. Hackl, A. E. Müller, M. Granitzer, M. Sailer, Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education* **8**, 1272229 (2023).
38. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
39. B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, L. Granka, In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication* **12**, 801–823 (2007).

ACKNOWLEDGMENTS. We thank the Wharton Behavioral Lab for their generous assistance with data collection and cost-sharing. Data, code, and materials are available here. We thank Hans Bode and Jacob P. for great research assistance. We thank Carmen Nobel and the Journalist Resource for assistance in recruiting professional editors, writers, and journalists. We extend special thanks to the many professional writers who generously contributed their time and expertise to this project, including Jim Davis, Theodore Fernando, Michael Ambrosino, Ronnie Willis, Jo Ann Livingston, Deborah Gdula, Mike Sherry, Jocelyn Dong, Kelly P. Kissel, Amabelle Ocampo, Katherine H. Donahue, Lindsay Pietenpol, Amanda Wolfe, Beth McLaughlin, Carol Katarsky, David Smith, Neil Reisner, Shawn Drury, Tanya Perez, Mary-Ann McHugh, Ashley Yeager, Pat Bellinghausen Zellar, Clarice Cardoso, Kathryn Marchocki, Simon C. Tung, Jones mark, Marion Davis, Christina Frangou, Nancy Peckenham, Stephanie Morrison, Mac McKerral, Mark Shwartz, Wacuka Mungai, Amanda Pearson, Elizabeth Blakey, Janet Raloff, Jim Johnson, Margaret Cadieux, Matt Wilson, Melanie Wachsman, Nettie Reynolds, Paul O'Donnell, Tim Wacker, Wendy Guild Swearngen, Nancy Lapid, Emily Matthews, Vladimire Herard, Steve Doig, and Lora Durance.

Supplementary Information for

Coach not crutch:
AI assistance can enhance rather than hinder skill development

Contents

A. Additional methods

A1. AI Ratings. After participants completed the procedure outlined in Figure 2, we had three writing samples for participants who practiced with or without AI (one for each phase of the experiment: pretest, practice, test), and two writing samples for participants who did not practice, or simply saw an AI-generated example. We used GPT-4o to rate these texts for five writing principles. Each text and rating was completed independently of each other (i.e., the model had no memory of seeing that text before or of having rated it for any of the other writing principles). For robustness checks, we also used Anthropic's Claude Haiku.

Table S1 shows the prompts used to have GPT-4o and Claude rate the rewritten cover letters on the five principles. Our pre-registered main outcome is the unweighted mean of these five principles.

Table S1. Prompt instructions given to GPT-4o and Claude for rating cover letters.

Writing principle	Rating prompt
Less is more	On a $0-10$ scale, how much does the text follow the Less is more principle? The text should use as few words as needed, as few ideas as needed, and make as few requests as needed.
Easy reading	On a $0-10$ scale, how much does the text make reading easy. The text should use short and common words, use straightforward sentences, and shorter sentences.
Easy navigation	On a $0-10$ scale, how much does the text make navigation easy. The text should make key information immediately visible, separate distinct ideas, place related ideas together, order ideas by priority, include headings when necessary, and use visuals if needed.
Formatting	On a $0-10$ scale, how much does the text use appropriate formatting. The text should follow readers expectations regarding formatting, use bolding, italics, underline, or highlight to draw attention to the most important ideas, and should not overdo formatting.
Easy responding	On a $0-10$ scale, how much does the text make responding easy. The text should simplify the steps required to act, organize the key information needed for action, and minimize the amount of attention required.

We used a randomly selected sample of 100 cover letters from Study 2 to validate the AI ratings. Two trained raters independently rated each of the 5 principles. The inter-rater correlation was $r = .74$, $p < .001$. The human raters correlated with the AI-generated ratings satisfactorily ($r = .70$, $p < .001$). See Figure S1.

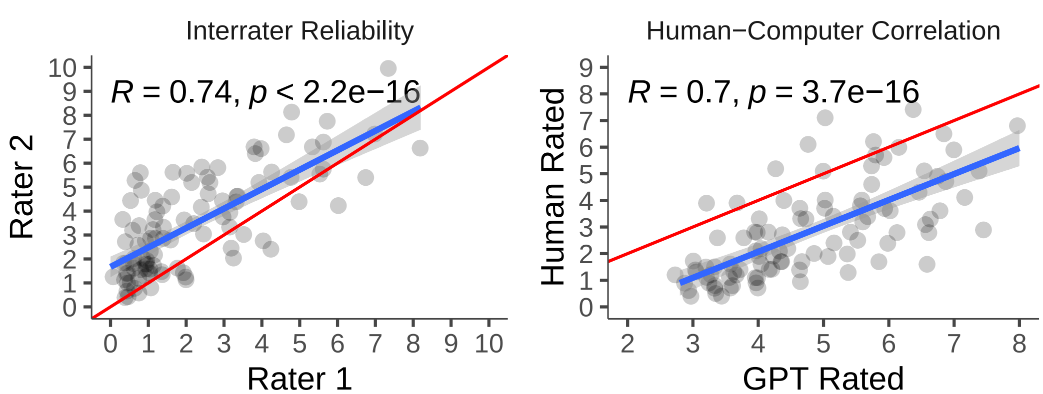

Fig. S1. Interrater correlations and correlations between AI and human ratings.

Claude and GPT-4o ratings were positively correlated, with pretest ratings being less so. In Study 2, the correlations were .38, .78, .65, and .67 for pretest, practice, test, and follow-up, respectively. All p-values were below .001.

A2. Pairwise Comparisons. Prolific participants were shown pairs of cover letters sampled from different conditions. They were asked to "Imagine you're hiring a social media manager for your company; which cover letter would make you more likely to offer an interview to the candidate? Choose one." Each cover letter was compared to at least three other letters, sampled uniformly at random from the other two conditions. Most letters were compared against 3 or 4 other letters. For each cover letter, we calculated the relative likelihood of it securing a hypothetical interview, defined as the total number of times that letter was preferred, divided by the total number of contests for that cover letter.

A3. Feedback. In Studies 2 and 3, participants received feedback for their submissions. The feedback was displayed immediately after the practice cover letter submission. The feedback page read: "Here is the email, then reproduced the participant's submission verbatim, then read "Here is one way in which it could be made better." The feedback was personalized and created by the GPT-4o model.

The feedback prompt is shown in Figure S2.

Feedback prompt

Take into account the following principles.

1. Less is more (use fewer words, include fewer ideas, make fewer requests).
2. Make reading easy (use short and common words, write straightforward sentences, write shorter sentences).
3. Design for easy navigation (make information immediately visible, group related ideas together, order ideas by priority, include headings).
4. Use enough formatting but no more (match formatting to readers' expectations, highlight, bold, or underline the most important ideas, limit your formatting).
5. Make responding easy (simplify the steps required to act, organize key information needed for action, minimize the amount of attention required).

I will show you a text, and I want you to act as a teacher providing feedback to the email, not the student. To do this, identify the principle that the text would benefit the most from implementing.
Your feedback:
- Should be clear, concise.
- Should reference the text wrote directly, Quote it and offer an alternative
- Start with something nice to say about the text

You can structure it as follows:
One sentence about what was good.
The email could be improved by focusing on **principle explained concretely in simple words**. For example:
- The email said: **example**
- Instead, it could have said: **rewritten example**
Make sure the feedback never addresses the person, but always focuses on the text. Never refer to you or your.
One sentence explanation, positive tone.

Fig. S2. Feedback prompt

B. Results Study 1

Table S2 reports results from a logistic regression predicting the probability of answering correctly as a function of demographic and background characteristics. Odds ratios (ORs) are displayed along with their standard errors, Wald z statistics, and *p*-values. The model includes age, sex, ethnicity, student status, employment, and prior AI experience as predictors.

Table S2. Demographic covariates and odds ratios for correct responses

Predictor	OR	SE	z	p-value
Prior AI experience	0.718	0.199	-1.664	0.096
Age (years)	1.014	0.021	0.675	0.500
Male (vs. Female)	1.035	0.467	0.073	0.942
White (vs. Other)	0.966	0.509	-0.068	0.946
Student (vs. Not)	1.120	0.618	0.183	0.855
Employment status				
Other	0.770	0.687	-0.382	0.703
Part-time	0.663	0.601	-0.684	0.494
Unemployed (seeking)	0.645	0.593	-0.738	0.461

C. Results Study 2

C1. Randomization, Balance, and Missingness. To allow users to format their responses flexibly, we used the TinyMCE rich text editor, which is interfaced with Qualtrics. While this allowed users to use bolding, lists, and italicizing, a small percentage of users experienced technical issues that resulted in their text data not being recorded (3.31%). These users did type in the box, as evidenced by their time and keystroke data, and completed the experiment.

There was also attrition in the follow-up sample. While most people responded, 13.45% of recontacted participants did not respond. This attrition was not selective by condition. As shown in Table S3, missingness and attrition rates were low for the main and follow-up samples, and did not differ by condition.

Table S3. Missingness and attrition proportions and test in Study 2.

Condition	Main Sample	Follow-up Sample
No practice	2.25%	13.38%
Practice w/o AI	3.86%	12.77%
Practice w/ AI	3.83%	14.23%
Overall	3.31%	13.45%
χ^2	3.966	0.685
p-value	0.138	0.710

Pre-treatment variables were balanced across experimental conditions, ensuring that random assignment was successful. To assess balance, we conducted a series of one-way ANOVAs for continuous variables and chi-square tests for categorical variables. Given the multiple comparisons, we applied the Benjamini-Hochberg (BH) procedure to control the false discovery rate. All statistical tests confirmed that none of the pre-treatment variables differed significantly across conditions. See Table S4.

Table S4. Randomization checks for pre-treatment variables in Study 2. *p*-values are BH multiple comparisons corrected. Continuous variables tested with ANOVA, binary and factor variables with χ^2 tests. SMD = Standardized Mean Difference.

	Overall	No practice	Practice w/o AI	Practice w/ AI	*p*	SMD
n	2238	755	752	731		
Age (mean (SD))	36.22 (12.71)	35.99 (13.01)	36.39 (12.70)	36.29 (12.42)	.923	0.021
Gender (%)					.636	0.077
Female	1189 (53.1)	421 (55.8)	394 (52.4)	374 (51.2)		
Male	1027 (45.9)	328 (43.4)	348 (46.3)	351 (48.0)		
Other	22 (1.0)	6 (0.8)	10 (1.3)	6 (0.8)		
Race/Ethnicity						
White (%)	1288 (57.6)	419 (55.5)	458 (60.9)	411 (56.2)	.213	0.073
Black (%)	745 (33.3)	262 (34.7)	223 (29.7)	260 (35.6)	.192	0.084
Asian (%)	134 (6.0)	49 (6.5)	42 (5.6)	43 (5.9)	.923	0.025
Latino (%)	155 (6.9)	48 (6.4)	65 (8.6)	42 (5.7)	.213	0.075
Other (%)	62 (2.8)	23 (3.0)	22 (2.9)	17 (2.3)	.923	0.030
Education Level (%)					.923	0.099
Less than high school degree	14 (0.6)	5 (0.7)	5 (0.7)	4 (0.5)		
High school graduate (high school diploma or equivalent including GED)	207 (9.2)	68 (9.0)	72 (9.6)	67 (9.2)		
Some college but no degree	321 (14.3)	117 (15.5)	109 (14.5)	95 (13.0)		
Associate degree in college (2-year)	168 (7.5)	61 (8.1)	60 (8.0)	47 (6.4)		
Bachelor's degree in college (4-year)	984 (44.0)	314 (41.6)	334 (44.4)	336 (46.0)		
Master's degree	474 (21.2)	165 (21.9)	153 (20.3)	156 (21.3)		
Doctoral degree (PhD)	44 (2.0)	16 (2.1)	11 (1.5)	17 (2.3)		
Non-PhD Professional degree (JD, MD)	26 (1.2)	9 (1.2)	8 (1.1)	9 (1.2)		
Perceived Writing Skill (mean (SD))	6.70 (1.70)	6.77 (1.67)	6.56 (1.72)	6.76 (1.69)	.192	0.083
Motivation to improve writing (%)					.410	0.126
Not at all motivated	33 (1.5)	6 (0.8)	15 (2.0)	12 (1.6)		
Hardly motivated	106 (4.7)	38 (5.0)	34 (4.5)	34 (4.7)		
Somewhat motivated	644 (28.8)	218 (28.9)	236 (31.4)	190 (26.0)		
Very motivated	932 (41.6)	311 (41.2)	311 (41.4)	310 (42.4)		
Extremely motivated	523 (23.4)	182 (24.1)	156 (20.7)	185 (25.3)		
Experience with AI writing assistants (%)					.701	0.094
I have never tried any AI writing assistant	354 (15.8)	109 (14.4)	139 (18.5)	106 (14.5)		
I have tried AI writing assistant(s) but hardly ever use them	859 (38.4)	291 (38.5)	288 (38.3)	280 (38.3)		
I use AI writing assistant(s) a few times per week	477 (21.3)	164 (21.7)	147 (19.5)	166 (22.7)		
I use AI writing assistant(s) about once a week	395 (17.6)	136 (18.0)	133 (17.7)	126 (17.2)		
I use AI writing assistant(s) every day	153 (6.8)	55 (7.3)	45 (6.0)	53 (7.3)		
Pretest Writing Skill (mean (SD))	3.32 (0.78)	3.31 (0.73)	3.32 (0.78)	3.32 (0.83)	.923	0.013

C2. AI practice improved writing skill. The AI tool improved performance while participants used it. Table S5 shows means and standardized differences for different measures of writing skill during the practice phase. The robustness checks included after the main specification, show that results are similar when using a different language model (Column 2), when not including control variables (Column 3), when excluding participants who admitted to cheating in the test phase (Column 4), for the subset of non-attriting participants to the follow-up phase (Column 5), and for each of the 5 principles separately (Columns 6 - 10).

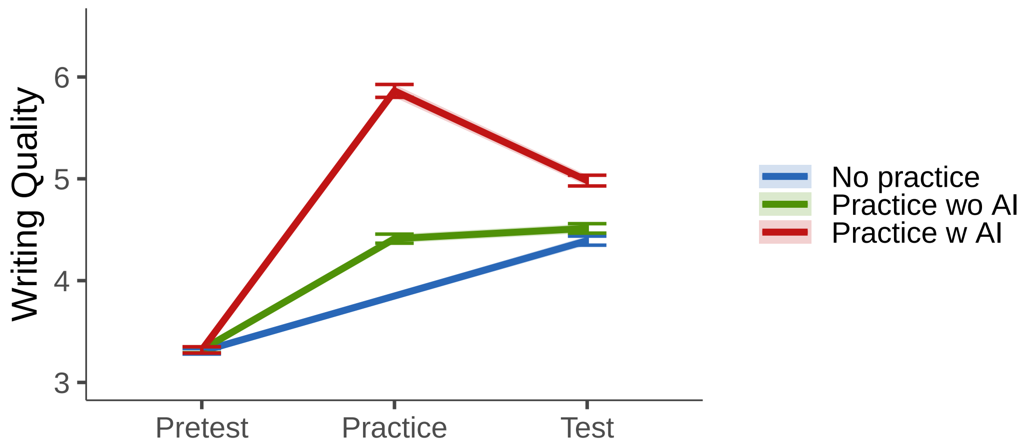

Fig. S3. Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means $\pm$ 1 SE. (N = 2,238).

Table S5. Differences in writing quality by condition in the practice phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Follow-up	LM	ER	EN	F	ER
Means — (SE)										
Practice w/o AI	4.58	6.52	4.41	4.42	4.27	4.29	6.31	5.65	3.38	4.26
	(.222)	(.091)	(.054)	(.055)	(.281)	(.230)	(.150)	(.258)	(.480)	(.268)
Practice w/ AI	6.05	7.01	5.86	5.89	5.78	5.56	7.11	6.75	6.30	5.54
	(.224)	(.092)	(.055)	(.055)	(.286)	(.232)	(.151)	(.260)	(.484)	(.271)
Effect Sizes (d) — (SE)										
Practice w/o AI vs. Practice w/ AI	1.01***	.81***	.98***	1.00***	1.05***	.84***	.82***	.65***	.93***	.73***
	(.056)	(.055)	(.055)	(.056)	(.061)	(.055)	(.055)	(.054)	(.056)	(.055)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. Follow-up is the subsample of non-attriting participants who returned to the one-day follow-up. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

During the test phase, when participants had to rewrite a cover letter without the help of the AI tool, participants who had practiced with AI outperformed participants who had not practiced, or had practiced without the AI tool. Again, the learning gains are robust to different specifications, subsamples, and measures or writing quality. See Table S6. For participants assigned to practice with the AI tool, the quality of AI rewrites did not correlate with participants' final submissions, $r = .06$, $p = .25$.

Table S6. Differences in writing quality by condition in the test phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Follow-up	LM	ER	EN	F	ER
Means — (SE)										
No practice	4.41	6.70	4.39	4.39	4.52	3.71	5.90	5.55	2.47	4.44
	(.161)	(.072)	(.047)	(.048)	(.200)	(.155)	(.143)	(.192)	(.394)	(.202)
Practice w/o AI	4.53	6.74	4.51	4.52	4.63	3.84	6.03	5.56	2.76	4.45
	(.160)	(.071)	(.048)	(.048)	(.199)	(.154)	(.142)	(.190)	(.392)	(.200)
Practice w/ AI	5.01	6.90	4.98	4.99	5.11	4.12	6.21	6.17	3.86	4.66
	(.161)	(.072)	(.049)	(.049)	(.202)	(.155)	(.143)	(.192)	(.394)	(.202)
Effect Sizes (d) — (SE)										
No practice vs. Practice w/o AI	.09	.09	.09	.10	.09	.10*	.11*	.01	.09	.01
	(.053)	(.053)	(.052)	(.053)	(.056)	(.053)	(.053)	(.053)	(.053)	(.053)
No practice vs. Practice w/ AI	.47***	.36***	.46***	.46***	.48***	.34***	.28***	.42***	.46***	.14**
	(.054)	(.053)	(.053)	(.054)	(.057)	(.053)	(.053)	(.053)	(.054)	(.053)
Practice w/o AI vs. Practice w/ AI	.38***	.28***	.36***	.36***	.39***	.23***	.17**	.41***	.36***	.13*
	(.054)	(.053)	(.053)	(.054)	(.057)	(.054)	(.053)	(.054)	(.054)	(.053)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. Follow-up is the subsample of non-attriting participants who returned to the one-day follow-up. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

C3. AI practice was less effortful. Table S7 shows OLS models predicting practice effort metrics from practice condition. Results show that participants practicing without AI expended more effort, measured subjectively or objectively, through keystrokes or practice time. As pre-registered, time is square-root-transformed, and keystrokes are log-transformed. Differences are slightly smaller when using untransformed variables.

Table S7. Practice effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
Practice w/o AI	2.37	4.31	6.52	6.76	430.95
	(.152)	(.322)	(.291)	(.913)	(57.349)
Practice w/ AI	2.30	3.36	5.93	6.62	383.38
	(.153)	(.325)	(.293)	(.919)	(57.887)
Effect Sizes (d) — (SE)					
Practice w/o AI vs. Practice w/ AI	-.07	-.45***	-.31***	-.02	-.13*
	(.053)	(.054)	(.053)	(.053)	(.053)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S8 shows OLS models predicting test effort metrics from practice condition. Results show some differences: participants who practiced with AI pressed more keys but reported less subjective effort.

Table S8. Test effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
No practice	2.32	5.01	6.55	5.62	399.10
	(.069)	(.213)	(.266)	(.257)	(41.391)
Practice w/o AI	2.15	4.87	6.91	4.98	409.09
	(.068)	(.211)	(.265)	(.255)	(41.143)
Practice w/ AI	2.19	5.05	6.69	5.14	446.59
	(.069)	(.213)	(.267)	(.257)	(41.408)
Effect Sizes (d) — (SE)					
No practice vs. Practice w/o AI	-.31***	-.09	.18***	-.32***	.03
	(.053)	(.053)	(.053)	(.053)	(.053)
No practice vs. Practice w/ AI	-.24***	.02	.07	-.24***	.15**
	(.053)	(.053)	(.053)	(.053)	(.053)
Practice w/o AI vs. Practice w/ AI	.07	.11*	-.11*	.08	.12*
	(.053)	(.053)	(.053)	(.053)	(.053)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S9 shows OLS models predicting learning rate metrics from practice condition. Learning rate is defined as the difference between test and pretest, divided by the effort metric. It shows how many points (10-point scale) the participant improved per unit effort (e.g., per minute spent practicing). Participants who practiced with AI improved their skill more efficiently.

Table S9. Learning rate differences. Means are the rate of improvement per unit sqrt(time (min)), log(keystrokes), subjective rating, raw time in minutes, and raw keystrokes.

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
Practice w/o AI	.28	.12	.27	.32	.43
	(.062)	(.094)	(.038)	(.096)	(.073)
Practice w/ AI	.43	.36	.37	.61	.62
	(.062)	(.094)	(.038)	(.097)	(.073)
Effect Sizes (d) — (SE)					
Practice w/o AI vs. Practice w/ AI	.39***	.40***	.41***	.47***	.40***
	(.054)	(.054)	(.054)	(.054)	(.054)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Most participants did not engage passively with the AI tool. As shown in Figure S4, an overwhelming majority of participants changed the AI tool's output text before submitting it as their answer. A smaller proportion of participants even edited the cover letter email *before* passing it along to the AI tool.

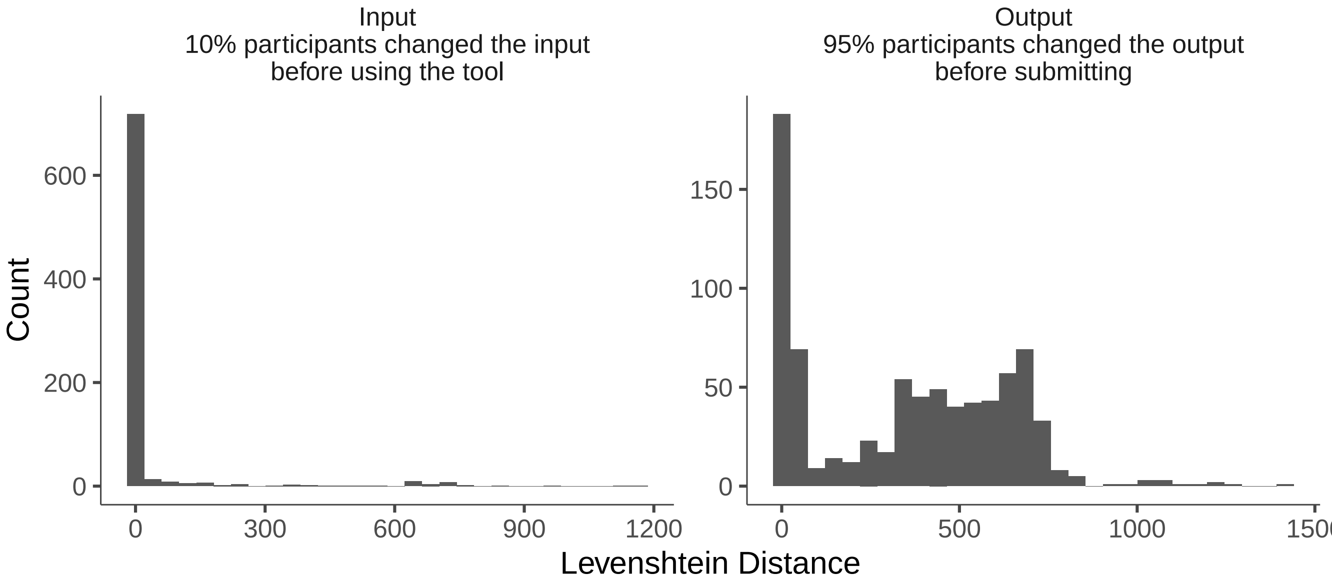

Fig. S4. Levenshtein distance (number of additions, modifications or deletions) between the original text and the text passed along to the AI tool (Input); and between the AI's output text and what users submitted as their final work (Output).

C4. AI practice did not discourage motivation for future learning. Table S10 presents differences in perceived learning, perceived writing skill, and the likelihood of asking for feedback across conditions, with effect sizes and means reported for each comparison. Despite objectively learning more, participants who practiced with AI perceived their learning and skill levels to be similar to those who practiced without AI and asked for feedback at comparable rates.

Table S10. Differences in motivational variables by condition.

	Perceived learning	Perceived writing skill	Asked for feedback
Means/Proportions			
No practice	5.91	6.40	.64
	(.209)	(.208)	(.064)
Practice w/o AI	5.90	6.61	.62
	(.207)	(.206)	(.066)
Practice w/ AI	6.03	6.56	.58
	(.209)	(.208)	(.068)
Effect Sizes (*d*s/odds ratios)			
No practice vs. Practice w/o AI	-.00	.13*	1.10
	(.053)	(.053)	(.128)
No practice vs. Practice w/ AI	.08	.10	1.30*
	(.053)	(.053)	(.149)
Practice w/o AI vs. Practice w/ AI	.08	-.03	1.17
	(.053)	(.053)	(.134)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

C5. The benefits of practicing with AI were just as large a day later. Table S11 shows means and standardized differences for measures of writing skill and related outcomes during the follow-up phase. The main specification demonstrates that participants who practiced with AI continued to outperform those who did not practice or practiced without AI. Robustness checks, including using a different language model (Column 2), excluding control variables (Column 3), and removing participants who admitted to cheating (Column 4) confirm the consistency of these effects. The results also hold when evaluating each of the five principles separately (Columns 5–9). These findings suggest that the benefits of practicing with AI are durable and persist even after participants stop using the tool.

Table S11. Differences in writing quality by condition in the follow-up phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	LM	ER	EN	F	ER
Means — (SE)									
No practice	4.73	6.75	4.75	4.75	4.31	6.36	5.56	2.43	5.01
	(.212)	(.094)	(.054)	(.054)	(.201)	(.206)	(.235)	(.510)	(.267)
Practice w/o AI	4.79	6.78	4.84	4.86	4.44	6.45	5.52	2.59	4.96
	(.211)	(.093)	(.053)	(.054)	(.200)	(.205)	(.234)	(.507)	(.266)
Practice w/ AI	5.34	6.95	5.35	5.37	4.72	6.67	6.14	3.85	5.30
	(.214)	(.094)	(.055)	(.055)	(.203)	(.208)	(.237)	(.515)	(.270)
Effect Sizes (d) — (SE)									
No practice vs. Practice w/o AI	.05	.06	.07	.08	.11	.08	-.02	.05	-.03
	(.056)	(.054)	(.055)	(.056)	(.056)	(.056)	(.056)	(.056)	(.056)
No practice vs. Practice w/ AI	.46***	.33***	.44***	.45***	.32***	.25***	.40***	.45***	.17**
	(.057)	(.055)	(.056)	(.057)	(.056)	(.056)	(.057)	(.057)	(.056)
Practice w/o AI vs. Practice w/ AI	.41***	.28***	.37***	.37***	.22***	.17**	.42***	.40***	.20***
	(.057)	(.055)	(.056)	(.056)	(.056)	(.056)	(.057)	(.057)	(.056)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

As mentioned in the main text the effect of practicing with AI and seeing an AI example did not become attenuated one day later. See Table S34.

Table S12. OLS model interacting condition with phase (Test vs. Follow-up) shows no attenuation of the effect of practicing with AI

Term	Estimate	SE	t	*p*-value
Intercept	-11.768	3.394	-3.467	0.001
Condition: No practice	-0.104	0.067	-1.541	0.123
Condition: Practice w/ AI	0.491	0.068	7.211	0.000
Condition: No practice × Follow-up	0.035	0.098	0.359	0.720
Condition: Practice w/ AI × Follow-up	0.039	0.099	0.399	0.690
Phase: Follow-up	0.330	0.069	4.787	0.000
Pretest score	0.394	0.026	15.220	0.000
Year of birth	0.007	0.002	4.284	0.000
Gender: Male	0.000	0.041	-0.012	0.990
Gender: Other	-0.272	0.210	-1.292	0.196
Race/ethnicity: White	0.108	0.092	1.167	0.243
Race/ethnicity: Black	-0.198	0.097	-2.048	0.041
Race/ethnicity: Asian	0.239	0.109	2.193	0.028
Race/ethnicity: Latino	0.070	0.095	0.736	0.462
Race/ethnicity: Other	0.093	0.128	0.724	0.469
Education: High school graduate (high school diploma or equivalent including GED)	0.775	0.274	2.824	0.005
Education: Some college but no degree	0.699	0.271	2.579	0.010
Education: Associate degree in college (2-year)	0.672	0.276	2.431	0.015
Education: Bachelor's degree in college (4-year)	0.682	0.269	2.536	0.011
Education: Master's degree	0.568	0.272	2.090	0.037
Education: Doctoral degree (PhD)	1.122	0.308	3.643	0.000
Education: Non-PhD Professional degree (JD, MD)	0.474	0.332	1.429	0.153
Writing skill	0.014	0.013	1.045	0.296
Motivation: Hardly motivated	-0.362	0.194	-1.865	0.062
Motivation: Somewhat motivated	-0.237	0.174	-1.360	0.174
Motivation: Very motivated	-0.338	0.174	-1.937	0.053
Motivation: Extremely motivated	-0.429	0.178	-2.410	0.016
Experience: I have tried AI writing assistant(s) but hardly ever use them	0.145	0.061	2.354	0.019
Experience: I use AI writing assistant(s) a few times per week	0.029	0.071	0.410	0.682
Experience: I use AI writing assistant(s) about once a week	0.040	0.072	0.549	0.583
Experience: I use AI writing assistant(s) every day	0.099	0.095	1.045	0.296

C6. AI practice was equally effective across subgroups. To test for moderation effects of pre-treatment demographic variables, we ran separate linear in which writing skill during the test phase was regressed on condition, the pre-treatment moderator of interest, writing skill at baseline, and an interaction term between the moderator × condition. After correcting the p-values for the interaction terms, none were significant at the .05 level, suggesting that practicing with AI was equally effective across groups.

Table S13. BH-corrected p-values for interaction terms from models predicting each outcome from condition interacted with pre-treatment variables.

	Test		**Follow-Up**		**Time Practice**	**Keys Practice**	**Effort Practice**	**Per. Learning**		**Per. Skill**		**Want Feedback**	
Level	**No AI**	**With AI**	**No AI**	**With AI**	**With AI**	**With AI**	**With AI**	**No AI**	**With AI**	**No AI**	**With AI**	**No AI**	**With AI**
Continuous Moderators													
Pretest	0.955	0.914	0.631	0.699	0.984	0.914	0.941	0.820	0.914	0.914	0.914	0.574	0.851
Year of Birth	0.533	0.931	0.868	0.955	0.914	0.955	0.914	0.914	0.955	0.955	0.914	0.914	0.618
Writing Skill	0.914	0.914	0.955	0.545	0.913	0.838	0.914	0.851	0.914	0.914	0.919	0.914	0.914
Gender (vs. Female)													
Male	0.955	0.914	0.970	0.955	0.699	0.914	0.955	0.699	0.979	0.914	0.851	0.699	0.533
Other	0.719	0.919	0.914	0.955	0.931	0.955	0.931	0.914	0.914	0.955	0.931	0.876	0.913
Race													
White	0.914	0.955	0.914	0.574	0.955	0.914	0.914	0.737	0.931	0.643	0.851	0.533	0.699
Black	0.914	0.913	0.914	0.851	0.914	0.914	0.955	0.574	0.919	0.295	0.699	0.574	0.533
Asian	0.973	0.973	0.955	0.737	0.964	0.931	0.955	0.931	0.890	0.919	0.955	0.914	0.699
Latino	0.919	0.914	0.970	0.914	0.914	0.868	0.533	0.868	0.944	0.533	0.914	0.970	0.574
Other	0.973	0.914	0.914	0.973	0.914	0.931	0.643	0.914	0.914	0.955	0.914	0.533	0.533
Motivation (vs. Not at all)													
Hardly	0.914	0.914	0.955	0.955	0.914	0.955	0.964	0.914	0.914	0.955	0.931	0.699	0.533
Somewhat	0.876	0.663	0.973	0.913	0.914	0.914	0.914	0.973	0.964	0.914	0.955	0.663	0.533
Very	0.851	0.566	0.980	0.914	0.851	0.968	0.964	0.970	0.982	0.914	0.973	0.699	0.533
Extremely	0.861	0.533	0.973	0.868	0.914	0.946	0.914	0.931	0.964	0.914	0.964	0.749	0.533
Experience with AI writing assistants (vs. None)													
Hardly ever	0.931	0.914	0.931	0.868	0.914	0.533	0.914	0.574	0.931	0.931	0.914	0.955	0.964
A few times per week	0.914	0.667	0.955	0.955	0.914	0.533	0.914	0.574	0.919	0.931	0.914	0.914	0.955
About once a week	0.876	0.214	0.914	0.955	0.955	0.699	0.914	0.574	0.914	0.931	0.919	0.914	0.970
Every day	0.914	0.914	0.914	0.964	0.919	0.914	0.984	0.914	0.931	0.955	0.914	0.955	0.931

Note. Models for test and follow-up performance, square-root practice time, log keystrokes, subjective effort, perceived learning and perceived writing skill or OLS models. Asking to see feedback was a binary Yes/No variable, and was modelled with logistic regression. Models match the pre-registered main specification, thereby controlling for all other pre-treatment variables. Per. = Perceived

C7. Pairwise comparisons. The relative likelihood of a cover letter receiving an invitation to an interview was correlated with the GPT-rated writing quality. See Figure S5.

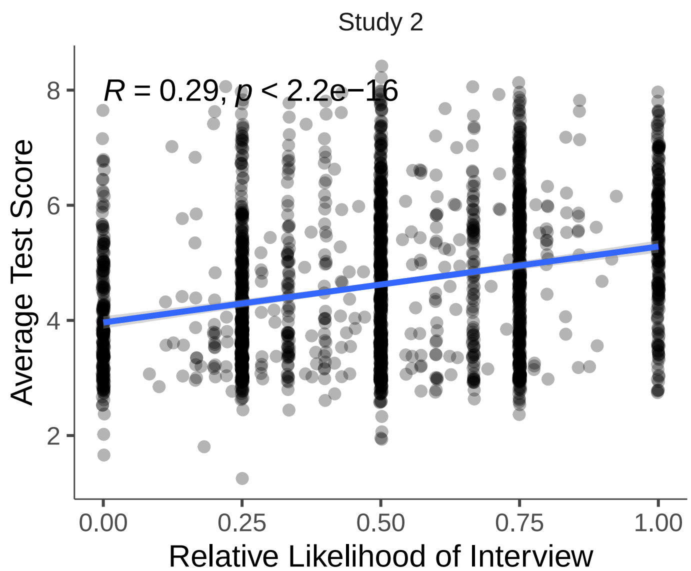

Fig. S5. Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

As shown in Table S14, participants who had practiced writing cover letters with AI were more likely to be invited to a hypothetical job interview.

Table S14. Beta regression models predicting relative likelihood of interview from condition.

	Without controls	With controls
Means (SE)		
Practice w AI	.50	.53
	(.01)	(.01)
Practice wo AI	.48	.48
	(.01)	(.01)
See AI example	.51	.49
	(.01)	(.01)
Model coefficients (SE)		
Precision (ϕ)	16.17***	5.36***
	(1.47)	(.30)
Symmetry (Log(ν))	.35***	-.91***
	(.08)	(.07)
Pairwise comparison (SE)		
Practice w AI - Practice wo AI	.08**	.21***
	(.03)	(.04)
Practice w AI - See AI example	-.02	.14**
	(.03)	(.04)
Practice wo AI - See AI example	-.10***	-.06
	(.03)	(.04)
Statistics		
N	1934.00	2153.00
AIC	2447.07	2329.29
BIC	2474.91	2499.52
log(Likelihood)	-1218.54	-1134.64

D. Results Study 3

D1. Mediation. Participants predicted they would learn more from the feedback of human editors than from AI (a path = 1.05, SE = 0.22, p <.001). In turn, higher predicted effectiveness predicted greater willingness to pay (b path = 0.06, SE = 0.01, p <.001). Participants' greater willingness to pay for human feedback was largely explained by their prediction that they would learn more from experienced human editors than from AI (indirect effect = 0.06, SE = 0.01, p < .001), accounting for most of the overall tendency to pay more for human editors (total effect = 0.10, SE = 0.02, p < .001). Once these predictions were taken into account, the remaining preference for human feedback was small and only marginally significant (c' path = 0.04, SE = 0.02, p = .054). See Figure S6 and Section D1 of Supplementary Information for details.

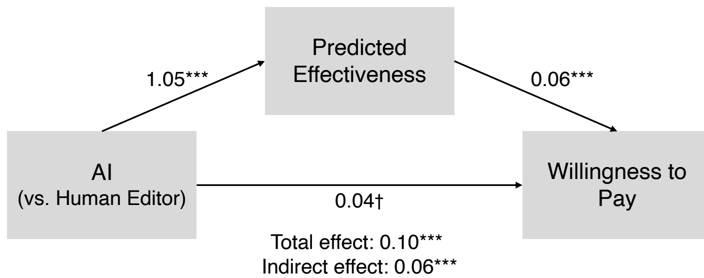

Fig. S6. Predicted effectiveness mediates underinvestment in AI feedback

E. Results Study 4

E1. Randomization, Balance, and Missingness. Participants in Study 4 were recontacted to return two days after the initial intake section. While most people responded, 14% of recontacted participants did not respond. As shown in Table S15, missingness and attrition rates were low and did not differ by condition.

Table S15. Missingness and attrition proportions and test in Study 4.

Condition	Retention Rate
Practice w/ AI	85.11%
Google examples	85.70%
Editor feedback	86.95%
Overall	85.92%
χ^2	1.446
p-value	.485

Pre-treatment variables were balanced across experimental conditions, ensuring that random assignment was successful. To assess balance, we conducted a series of one-way ANOVAs for continuous variables and chi-square tests for categorical variables. Given the multiple comparisons, we applied the Benjamini-Hochberg (BH) procedure to control the false discovery rate. All statistical tests confirmed that none of the pre-treatment variables differed significantly across conditions. See Table S16.

Table S16. Randomization checks for pre-treatment variables in Study 4. *p*-values are BH multiple comparisons corrected. Continuous variables tested with ANOVA, binary and factor variables with χ^2 tests. SMD = Standardized Mean Difference.

	Overall	Practice w/ AI	Google examples	Editor feedback	*p*	SMD
n	2997	1001	1000	996		
Age (mean (SD))	41.47 (13.83)	41.38 (13.78)	41.11 (13.76)	41.92 (13.95)	.666	0.039
Gender (%)					.666	0.059
Female	1824 (60.9)	596 (59.5)	628 (62.8)	600 (60.2)		
Male	1124 (37.5)	390 (39.0)	352 (35.2)	382 (38.4)		
Other	49 (1.6)	15 (1.5)	20 (2.0)	14 (1.4)		
Race/Ethnicity						
White (%)	2209 (73.7)	745 (74.4)	736 (73.6)	728 (73.1)	.950	0.020
Black (%)	473 (15.8)	164 (16.4)	145 (14.5)	164 (16.5)	.666	0.036
Asian (%)	204 (6.8)	61 (6.1)	68 (6.8)	75 (7.5)	.666	0.038
Latino (%)	253 (8.4)	79 (7.9)	95 (9.5)	79 (7.9)	.666	0.038
Other (%)	102 (3.4)	35 (3.5)	40 (4.0)	27 (2.7)	.666	0.048
Education Level (%)					.964	0.080
Less than high school degree	10 (0.3)	4 (0.4)	2 (0.2)	4 (0.4)		
High school graduate (high school diploma or equivalent including GED)	271 (9.0)	80 (8.0)	94 (9.4)	97 (9.7)		
Some college but no degree	508 (17.0)	178 (17.8)	168 (16.8)	162 (16.3)		
Associate degree in college (2-year)	283 (9.4)	91 (9.1)	98 (9.8)	94 (9.4)		
Bachelor's degree in college (4-year)	1165 (38.9)	398 (39.8)	394 (39.4)	373 (37.4)		
Master's degree	596 (19.9)	199 (19.9)	193 (19.3)	204 (20.5)		
Doctoral/Professional degree	164 (5.5)	51 (5.1)	51 (5.1)	62 (6.2)		
Perceived Writing Skill (mean (SD))	6.69 (1.67)	6.67 (1.68)	6.63 (1.69)	6.76 (1.63)	.666	0.051
Motivation to Improve Writing (%)					.748	0.083
Not at all motivated	24 (0.8)	9 (0.9)	6 (0.6)	9 (0.9)		
Hardly motivated	166 (5.5)	59 (5.9)	55 (5.5)	52 (5.2)		
Somewhat motivated	1036 (34.6)	354 (35.4)	330 (33.0)	352 (35.3)		
Very motivated	1135 (37.9)	387 (38.7)	390 (39.0)	358 (35.9)		
Extremely motivated	636 (21.2)	192 (19.2)	219 (21.9)	225 (22.6)		
Experience with AI writing assistants (%)					.666	0.090
I have never tried any AI writing assistant	516 (17.2)	188 (18.8)	173 (17.3)	155 (15.6)		
I have tried AI writing assistant(s) but hardly ever use them	1211 (40.4)	383 (38.3)	411 (41.1)	417 (41.9)		
I use AI writing assistant(s) a few times per week	557 (18.6)	197 (19.7)	173 (17.3)	187 (18.8)		
I use AI writing assistant(s) about once a week	503 (16.8)	158 (15.8)	178 (17.8)	167 (16.8)		
I use AI writing assistant(s) every day	210 (7.0)	75 (7.5)	65 (6.5)	70 (7.0)		
Pretest Writing Skill (mean (SD))	4.40 (0.67)	4.41 (0.67)	4.40 (0.68)	4.40 (0.65)	.964	0.008

E2. Details on professional editors. We recruited 44 professional writers and editors through Journalist's Resource, paying them $100 to edit 20 cover letters each. They were asked to spend about 7 minutes per cover letter. In total, they edited 1,227 cover letters.

To determine if editors changed their behavior over time, we fit a series of multilevel models, with cover letter position as a predictor, random intercept for editors, and text quality (as rated by GPT-4o), time spent per cover letter, and number of insertions, deletions, or modifications (i.e., Levenstein distance). These analyses include editors who were able to edit at least 20 letters, and focus on the first 20 letters edited. The quality of the edited letters improved over time (β = 0.013, SE = 0.005, p = 0.013), even though editors spent less time as the task progressed (β = -0.387, SE = 0.028, p < .001). Interestingly, editors also were able to improve the emails using fewer edits as the task progressed (β = -2.572, SE = 0.926, p = 0.00561). Despite spending less time per letter, editors produced higher quality edits with fewer but more effective changes, demonstrating clear learning effects in professional editing tasks.

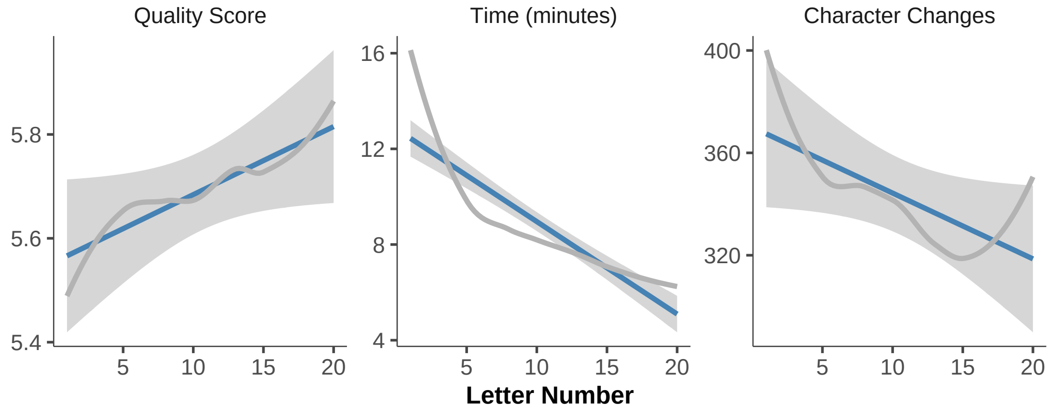

Fig. S7. Editor performance improved over the course of the task, with participants becoming more efficient at editing emails while reducing both editing time and the number of character-level changes (measured by Levenshtein distance). The line and shaded area show a linear model fit, while the gray line represents the best-fit loess curve.

E3. Details on participants' search behavior. We replicated the first five pages of a Google search for "cover letter examples" and made this available to participants through Qualtrics to monitor their search behavior.

Across 635 participants assigned to use Google to search for cover letter examples and tips, we found search behavior consistent with prior research on web-browsing behavior (39). Most participants (97%) stayed on page 1, while only 3% ventured beyond the first page of results. Participants spent an average of 2.8 minutes (median: 1.3 minutes) browsing search results, with an average of 0.9 minutes between clicks. On average, participants made 2.9 clicks during their search session.

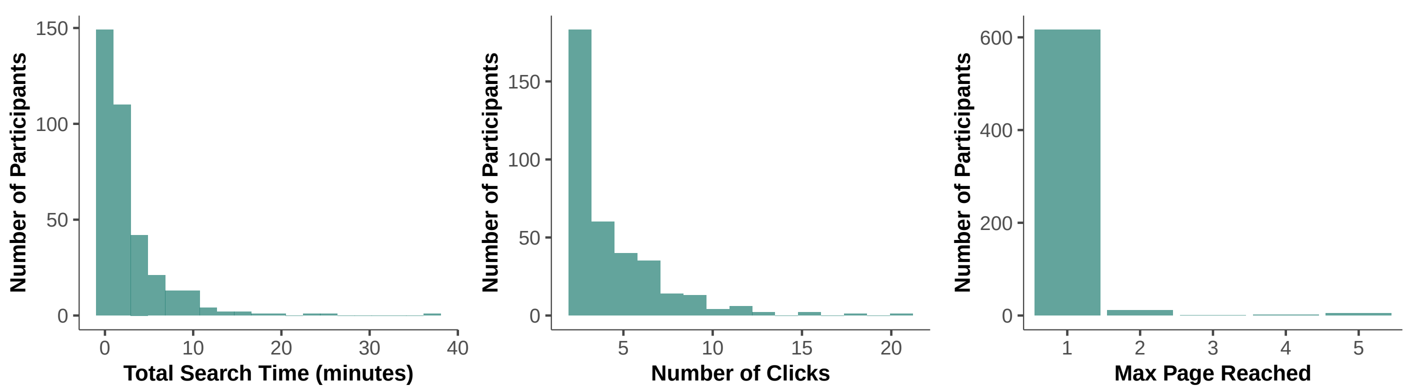

Fig. S8. Web browsing descriptive statistics.

E4. AI practice improved writing skill more than getting feedback from professional editors and looking for examples online. The AI tool improved performance while participants used it. Table S17 shows means and standardized differences for different measures of writing skill during the practice phase. The robustness checks included after the main specification, show that results are similar when using a different language model (Column 2), when not including control variables (Column 3), when excluding participants who admitted to cheating in the test phase (Column 4), and for each of the 5 principles separately (Columns 5 - 9).

Table S17. Differences in writing quality by condition in the practice phase

	GPT-4o	Ex. Controls	Ex. Cheaters	LM	ER	EN	F	ER
Means								
AI	6.24	6.22	6.28	5.61	7.26	6.88	5.41	6.04
	(.172)	(.048)	(.173)	(.167)	(.125)	(.190)	(.397)	(.214)
Google	4.81	4.80	4.84	4.23	6.33	5.85	2.71	4.91
	(.173)	(.048)	(.174)	(.167)	(.125)	(.191)	(.398)	(.215)
Editors	5.18	5.17	5.21	4.55	6.56	6.24	3.44	5.09
	(.172)	(.048)	(.173)	(.166)	(.124)	(.190)	(.396)	(.214)
Effect Sizes (d)								
AI vs. Google	-1.04***	-1.01***	-1.03***	-1.03***	-.93***	-.67***	-.85***	-.66***
	(.051)	(.050)	(.052)	(.051)	(.051)	(.050)	(.051)	(.050)
AI vs. Editors	-.77***	-.75***	-.77***	-.79***	-.70***	-.42***	-.62***	-.56***
	(.050)	(.049)	(.050)	(.050)	(.050)	(.049)	(.050)	(.050)
Google vs. Editors	.27***	.26***	.27***	.24***	.23***	.25***	.23***	.10*
	(.049)	(.048)	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

During the test phase, when participants had to rewrite a cover letter without the help of the AI tool, participants who had practiced with AI outperformed participants who searched for cover letter examples and tips and participants who received personalized feedback from professional editors. Again, the learning gains are robust to different specifications, subsamples, and measures or writing quality. See Table S18.

Table S18. Differences in writing quality by condition in the test phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	LM	ER	EN	F	ER
Means									
AI	4.98	5.14	4.77	4.99	3.98	6.34	6.00	3.63	4.93
	(.133)	(.103)	(.041)	(.134)	(.126)	(.133)	(.156)	(.333)	(.169)
Google	4.44	4.90	4.24	4.46	3.59	5.95	5.50	2.43	4.74
	(.133)	(.103)	(.040)	(.134)	(.126)	(.133)	(.156)	(.333)	(.169)
Editors	4.75	5.03	4.55	4.76	3.70	6.10	5.95	3.18	4.80
	(.133)	(.103)	(.040)	(.134)	(.126)	(.133)	(.155)	(.333)	(.169)
Effect Sizes (d)									
AI vs. Google	-.46***	-.27***	-.44***	-.46***	-.36***	-.34***	-.37***	-.41***	-.12*
	(.050)	(.050)	(.049)	(.050)	(.049)	(.049)	(.049)	(.049)	(.049)
AI vs. Editors	-.20***	-.12*	-.18***	-.20***	-.25***	-.21***	-.04	-.16**	-.09
	(.049)	(.049)	(.048)	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)
Google vs. Editors	.26***	.15**	.26***	.26***	.11*	.13**	.34***	.26***	.04
	(.049)	(.049)	(.048)	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

E5. AI practice wasn't any more effortful than getting feedback from professional editors and was less effortful than looking for examples online. Table S19 shows OLS models predicting practice effort metrics from practice condition. Results show that participants practicing without AI expended more effort, measured subjectively or objectively, through keystrokes or practice time. As pre-registered, time is square-root-transformed, and keystrokes are log-transformed. Differences are slightly smaller when using untransformed variables.

Table S19. Practice effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/min
Means						
AI	2.45	3.76	7.14	7.04	251.19	1.52
	(.112)	(.251)	(.217)	(.702)	(41.527)	(.089)
Google	2.76	4.68	7.11	8.96	369.45	1.78
	(.112)	(.251)	(.217)	(.702)	(41.464)	(.089)
Editors	2.54	4.50	7.27	7.50	281.33	1.83
	(.112)	(.251)	(.216)	(.702)	(41.482)	(.089)
Effect Sizes (d)						
AI vs. Google	.32***	.42***	-.02	.31***	.32***	.34***
	(.049)	(.049)	(.049)	(.049)	(.049)	(.050)
AI vs. Editors	.09	.34***	.07	.07	.08	.41***
	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)
Google vs. Editors	-.22***	-.08	.09	-.24***	-.24***	.06
	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S20 shows OLS models predicting test effort metrics from practice condition. Results show some differences: participants who practiced with AI pressed more keys but reported less subjective effort.

Table S20. Test effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/min
Means						
AI	2.32	5.44	7.95	5.55	415.78	1.92
	(.056)	(.177)	(.225)	(.218)	(39.760)	(.041)
Google	2.23	5.29	7.52	5.20	384.82	1.92
	(.056)	(.177)	(.225)	(.218)	(39.747)	(.041)
Editors	2.29	5.39	8.00	5.41	381.98	1.95
	(.056)	(.177)	(.225)	(.218)	(39.793)	(.041)
Effect Sizes (d)						
AI vs. Google	-.18***	-.09	-.22***	-.18***	-.09	-.02
	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)
AI vs. Editors	-.07	-.03	.02	-.08	-.10*	.07
	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)
Google vs. Editors	.11*	.06	.24***	.11*	-.01	.09
	(.049)	(.049)	(.049)	(.049)	(.049)	(.049)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

E6. AI practice did not create more of an illusion of mastery than getting feedback from professional editors or looking for examples online. As reported in the main text, there were no differences between participants who practiced with AI and participants who practiced with editor feedback on how much they thought they learned, their perceived skill, and their likelihood to ask for feedback after the test. See Table S21

Table S21. Differences in motivational variables by condition

	Perceived Learning	Perceived Writing Skill	Asked for Feedback
Means			
AI	6.56	6.74	.67
	(.227)	(.157)	(.054)
Google	6.01	6.59	.59
	(.227)	(.157)	(.058)
Editors	6.41	6.76	.66
	(.227)	(.156)	(.054)
Effect Sizes (d)			
AI vs. Google	-.28***	-.11*	1.38**
	(.049)	(.049)	(.147)
AI vs. Editors	-.07	.02	1.03
	(.049)	(.049)	(.112)
Google vs. Editors	.20***	.12*	.75**
	(.049)	(.049)	(.080)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

E7. Results were not moderated by individual differences. As in Study 2, we tested whether each of the pretreatment demographic variables moderated the effects of seeing an AI example. To do this, we ran separate linear in which writing skill during the test phase was regressed on condition, the pre-treatment moderator of interest, writing skill at baseline, and an interaction term between the moderator × condition. After correcting the p-values for the interaction terms, none were significant at the .05 level, suggesting that seeing AI examples was equally effective across groups. See Table S22

Table S22. BH-corrected *p*-values for interaction terms from models predicting each outcome from condition interacted with pre-treatment variables.

	Test		Time Practice		Keys Practice		Effort Practice		Per. Learning		Per. Skill		Want Feedback	
	GE	EF	GE	EF	GE	EF	GE	EF	GE	EF	GE	EF	GE	EF
Pretest	0.805	0.950	0.990	0.877	0.990	0.990	0.805	0.997	0.919	0.930	0.990	0.990	0.990	0.690
Year of Birth	0.996	0.402	0.990	0.748	0.877	0.894	0.798	0.852	0.765	0.708	0.936	0.695	0.990	0.789
Writing Skill	0.971	0.991	0.889	0.805	0.877	0.894	0.990	0.985	0.878	0.572	0.789	0.957	0.877	0.572
Gender vs. Female														
Male	0.695	0.957	0.877	0.805	0.899	0.894	0.891	0.950	0.981	0.889	0.990	0.923	0.314	0.402
Other	0.790	0.889	0.885	0.990	0.916	0.990	0.899	0.798	0.950	0.861	0.877	0.990	0.798	0.419
Race/Etnicity														
White	0.267	0.402	0.402	0.542	0.950	0.877	0.878	0.887	0.862	0.877	0.805	0.990	0.852	0.449
Black	0.748	0.922	0.748	0.990	0.891	0.877	0.442	0.211	0.189	0.533	0.140	0.427	0.990	0.990
Asian	0.150	0.432	0.990	0.789	0.894	0.605	0.519	0.805	0.937	0.899	0.990	0.805	0.805	0.378
Latino	0.189	0.789	0.402	0.889	0.318	0.894	0.981	0.605	0.899	0.805	0.852	0.789	0.990	0.990
Other	0.894	0.798	0.877	0.990	0.990	0.805	0.959	0.660	0.805	0.701	0.877	0.427	0.990	0.877
Education														
High school	0.990	0.990	0.882	0.877	0.990	0.874	0.990	0.419	0.998	0.189	0.140	0.166	0.990	0.990
Some college	0.990	0.990	0.877	0.899	0.959	0.877	0.990	0.333	0.990	0.140	0.143	0.186	0.990	0.990
Associate	0.990	0.990	0.899	0.690	0.990	0.798	0.990	0.542	0.990	0.140	0.144	0.166	0.990	0.990
Bachelor's	0.990	0.990	0.877	0.878	0.899	0.877	0.990	0.438	0.990	0.143	0.140	0.172	0.990	0.990
Master's	0.990	0.981	0.889	0.877	0.957	0.852	0.973	0.605	0.990	0.143	0.140	0.189	0.990	0.990
Doctoral/Professional	0.990	0.894	0.894	0.805	0.990	0.789	0.990	0.402	0.927	0.140	0.140	0.140	0.990	0.990
Motivation (vs. Not at all)														
Hardly motivated	0.990	0.990	0.611	0.768	0.877	0.899	0.690	0.889	0.901	0.885	0.577	0.990	0.798	0.990
Somewhat motivated	0.996	0.957	0.605	0.805	0.732	0.854	0.278	0.542	0.838	0.990	0.877	0.877	0.877	0.934
Very motivated	0.990	0.990	0.419	0.789	0.790	0.852	0.419	0.690	0.936	0.964	0.763	0.899	0.878	0.919
Extremely motivated	0.990	0.936	0.547	0.789	0.690	0.789	0.448	0.608	0.894	0.990	0.660	0.894	0.789	0.874
Experience with AI writing assistants (vs. None)														
Hardly ever	0.990	0.427	0.789	0.990	0.267	0.990	0.990	0.990	0.990	0.971	0.990	0.990	0.789	0.877
A few times per week	0.805	0.877	0.789	0.990	0.660	0.990	0.930	0.990	0.889	0.444	0.970	0.990	0.899	0.877
About once a week	0.805	0.798	0.278	0.790	0.189	0.990	0.990	0.990	0.981	0.990	0.990	0.985	0.981	0.990
Every day	0.899	0.150	0.957	0.894	0.990	0.832	0.790	0.695	0.790	0.990	0.852	0.790	0.790	0.852

Note. Models for test performance, square-root practice time, log keystrokes, subjective effort, perceived learning and perceived writing skill or OLS models. Asking to see feedback was a binary Yes/No variable, and was modelled with logistic regression. Models match the pre-registered main specification, thereby controlling for all other pre-treatment variables. Per. = Perceived

E8. Pairwise comparisons. The relative likelihood of a cover letter receiving an invitation to an interview was correlated with the GPT-rated writing quality. See Figure S9.

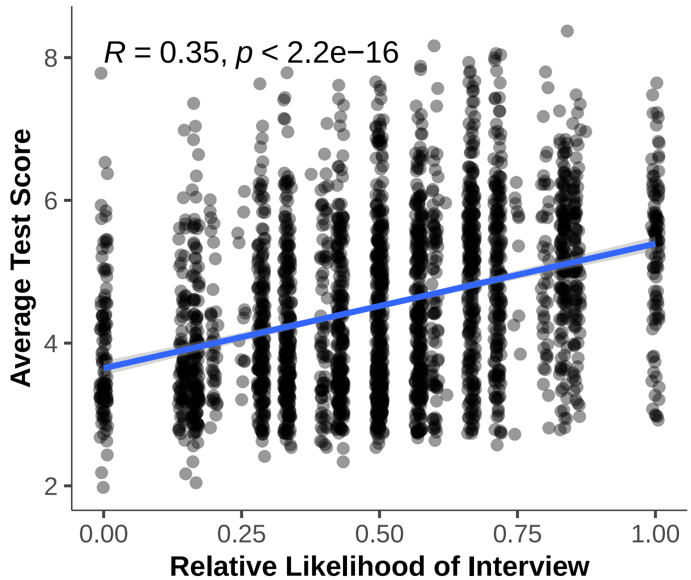

Fig. S9. Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

Table S23 shows the beta regressions for the relative likelihood by condition.

Table S23. Beta regression models predicting relative likelihood of interview from condition.

	As preregistered		All data	
	Without controls	With controls	Without controls	With controls
Means (SE)				
AI	.50	.51	.50	.50
	(.00)	(.01)	(.01)	(.01)
Google	.49	.49	.49	.49
	(.00)	(.01)	(.01)	(.01)
Editors	.51	.51	.51	.51
	(.00)	(.01)	(.01)	(.01)
Model coefficients (SE)				
Precision (ϕ)	16.78***	7.17***	9.94***	9.94***
	(1.28)	(.41)	(.74)	(.74)
Symmetry ($\log(\nu)$)	.60***	-.01	-1.05***	-1.05***
	(.07)	(.06)	(.09)	(.09)
Pairwise comparison (SE)				
AI - Google	.06*	.08*	.05	.05
	(.03)	(.04)	(.03)	(.03)
AI - Editors	-.01	-.00	-.04	-.04
	(.03)	(.04)	(.03)	(.03)
Google - Editors	-.06*	-.08*	-.09**	-.09**
	(.03)	(.04)	(.03)	(.03)
Statistics				
N	2575.00	2513.00	2575.00	2575.00
AIC	3697.77	3750.64	867.58	867.58
BIC	3727.04	3832.25	896.84	896.84
log(Likelihood)	-1843.89	-1861.32	-428.79	-428.79

F. Results Study 5

F1. Randomization, Balance, and Missingness. As in Study 2, technical issues caused small amounts of missing data. Overall, 5.64% of data was missing in for the test phase analysis, which was not differentially missing by condition. There was also attrition in the follow-up sample. While most people responded, 13.45% of recontacted participants did not respond. This attrition was not selective by condition. As shown in Table S24, missingness and attrition rates were low for the main and follow-up samples and did not differ by condition.

Table S24. Missingness and attrition proportions and test in Study 5.

Condition	Main Sample	Follow-up Sample
Practice w/o AI	4.61%	73.51%
Practice w/ AI	5.52%	70.40%
See AI example	6.77%	72.16%
Overall	5.64%	72.04%
χ^2	2.991	1.600
p-value	0.224	0.449

Pre-treatment variables were balanced across experimental conditions, ensuring that random assignment was successful. To assess balance, we conducted a series of one-way ANOVAs for continuous variables and chi-square tests for categorical variables. Given the multiple comparisons, we applied the Benjamini-Hochberg (BH) procedure to control the false discovery rate. All statistical tests confirmed that none of the pre-treatment variables differed significantly across conditions. See Table S25.

Table S25. Randomization checks for pre-treatment variables. *p*-values are BH corrected. SMD = Standardized Mean Difference.

	Overall	Practice w/o AI	Practice w/ AI	See AI example	*p*	SMD
n	2003	672	652	679		
Age (mean (SD))	37.89 (12.63)	37.77 (12.37)	37.87 (12.85)	38.03 (12.69)	.997	0.014
Gender (%)					.822	0.055
Female	1056 (52.7)	341 (50.7)	350 (53.7)	365 (53.8)		
Male	923 (46.1)	321 (47.8)	296 (45.4)	306 (45.1)		
Other	24 (1.2)	10 (1.5)	6 (0.9)	8 (1.2)		
Race/Ethnicity (%)						
White = 1	1287 (64.3)	419 (62.4)	430 (66.0)	438 (64.5)	.655	0.050
Black = 1	484 (24.2)	184 (27.4)	144 (22.1)	156 (23.0)	.324	0.082
Asian = 1	127 (6.3)	37 (5.5)	43 (6.6)	47 (6.9)	.715	0.039
Latino = 1	163 (8.1)	55 (8.2)	46 (7.1)	62 (9.1)	.655	0.051
Other = 1	3 (0.1)	1 (0.1)	2 (0.3)	0 (0.0)	.655	0.055
Education Level (%)					.655	0.152
Less than high school degree	10 (0.5)	3 (0.4)	3 (0.5)	4 (0.6)		
High school graduate	205 (10.2)	74 (11.0)	61 (9.4)	70 (10.3)		
Some college, no degree	305 (15.2)	104 (15.5)	110 (16.9)	91 (13.4)		
Associate degree	169 (8.4)	68 (10.1)	45 (6.9)	56 (8.2)		
Bachelor's degree	850 (42.4)	255 (37.9)	290 (44.5)	305 (44.9)		
Master's degree	401 (20.0)	144 (21.4)	126 (19.3)	131 (19.3)		
Doctoral degree (PhD)	36 (1.8)	14 (2.1)	11 (1.7)	11 (1.6)		
Professional degree (JD, MD)	27 (1.3)	10 (1.5)	6 (0.9)	11 (1.6)		
Writing Skill (mean (SD))	6.60 (1.70)	6.63 (1.67)	6.71 (1.69)	6.46 (1.73)	.228	0.100
Motivation (%)					.997	0.042
Not at all motivated	28 (1.4)	9 (1.3)	10 (1.5)	9 (1.3)		
Hardly motivated	154 (7.7)	50 (7.4)	53 (8.1)	51 (7.5)		
Somewhat motivated	639 (31.9)	221 (32.9)	202 (31.0)	216 (31.8)		
Very motivated	762 (38.0)	249 (37.1)	249 (38.2)	264 (38.9)		
Extremely motivated	420 (21.0)	143 (21.3)	138 (21.2)	139 (20.5)		
Experience with AI (%)					.655	0.103
Never used AI writing assistant	351 (17.5)	128 (19.0)	105 (16.1)	118 (17.4)		
Tried AI but hardly use	807 (40.3)	267 (39.7)	269 (41.3)	271 (39.9)		
Use AI a few times per week	375 (18.7)	108 (16.1)	133 (20.4)	134 (19.7)		
Use AI about once a week	343 (17.1)	127 (18.9)	102 (15.6)	114 (16.8)		
Use AI every day	127 (6.3)	42 (6.2)	43 (6.6)	42 (6.2)		
Pretest Writing Skill (mean (SD))	4.21 (0.88)	4.23 (0.90)	4.24 (0.90)	4.17 (0.84)	.655	0.051

F2. AI examples improve writing skill. The AI tool improved performance while participants used it. Table S26 shows means and standardized differences for different measures of writing skill during the practice phase. The robustness checks included after the main specification, show that results are similar when using a different language model (Column 2), when not including control variables (Column 3), when excluding participants who admitted to cheating in the test phase (Column 4), for the subset of non-attriting participants to the follow-up phase (Column 5), and for each of the 5 principles separately (Columns 6 - 10).

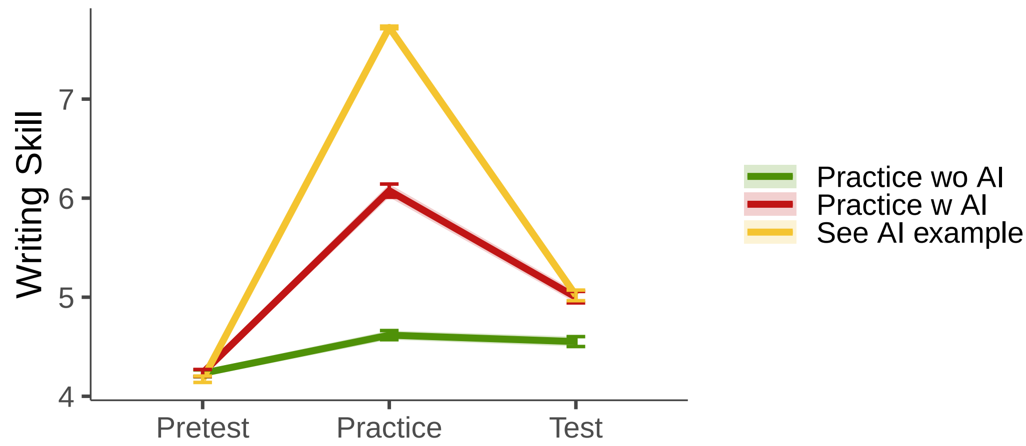

Fig. S10. Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means $\pm$ 1 SE. (*N* = 2,003).

During the test phase, when participants had to rewrite a cover letter without the help of the AI tool, participants who simply had seen an AI example outperformed participants who had practiced without the AI tool, and perfomed comparably to those who had practiced with the AI tool. Replicating Study 2, participants who had practiced with the AI tool performed better than those who had practiced without it. Again, the learning gains are robust to different specifications, subsamples, and measures or writing quality. See Table S27. For participants assigned to practice with the AI tool, The quality of AI rewrites did not correlate with participants' final submissions, $r = .06$, $p = .25$.

Table S26. Differences in writing quality by condition in the practice phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Follow-up	LM	ER	EN	F	ER
Means — (SE)										
Practice w/o AI	4.72	4.96	4.62	4.61	4.64	4.33	6.61	5.43	2.93	4.31
	(.374)	(.379)	(.048)	(.048)	(.651)	(.396)	(.265)	(.438)	(.795)	(.466)
Practice w/ AI	6.19	6.36	6.08	6.08	5.98	5.53	7.44	6.60	5.64	5.74
	(.373)	(.379)	(.048)	(.049)	(.656)	(.395)	(.265)	(.438)	(.794)	(.466)
See AI example	7.83	8.04	7.72	7.72	7.58	7.23	8.38	8.18	8.34	7.02
	(.375)	(.380)	(.048)	(.049)	(.656)	(.397)	(.266)	(.439)	(.797)	(.467)
Effect Sizes (d) — (SE)										
Practice w/o AI vs. Practice w/ AI	1.22***	1.15***	1.19***	1.21***	1.14***	.95***	.96***	.83***	1.06***	.95***
	(.060)	(.059)	(.059)	(.060)	(.112)	(.059)	(.059)	(.058)	(.059)	(.059)
Practice w/o AI vs. See AI example	2.58***	2.52***	2.54***	2.55***	2.50***	2.28***	2.07***	1.95***	2.12***	1.81***
	(.070)	(.070)	(.069)	(.070)	(.132)	(.067)	(.066)	(.065)	(.066)	(.063)
Practice w/ AI vs. See AI example	1.36***	1.37***	1.35***	1.35***	1.36***	1.33***	1.10***	1.12***	1.06***	.86***
	(.061)	(.061)	(.060)	(.061)	(.113)	(.060)	(.059)	(.059)	(.059)	(.058)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S27. Differences in writing quality by condition in the test phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Follow-up	LM	ER	EN	F	ER
Means — (SE)										
Practice w/o AI	5.39	5.28	4.55	4.55	4.71	4.39	7.01	6.19	4.26	5.13
	(.411)	(.426)	(.055)	(.055)	(.717)	(.391)	(.367)	(.479)	(.982)	(.496)
Practice w/ AI	5.82	5.84	5.00	5.00	5.00	4.69	7.11	6.61	5.29	5.38
	(.410)	(.426)	(.056)	(.056)	(.722)	(.390)	(.366)	(.478)	(.981)	(.495)
See AI example	5.87	5.95	5.02	5.03	5.08	4.66	7.08	6.78	5.47	5.36
	(.412)	(.427)	(.054)	(.054)	(.722)	(.392)	(.368)	(.480)	(.985)	(.497)
Effect Sizes (d) — (SE)										
Practice w/o AI vs. Practice w/ AI	.32***	.41***	.32***	.33***	.22*	.24***	.09	.28***	.33***	.16**
	(.057)	(.057)	(.056)	(.057)	(.106)	(.057)	(.057)	(.057)	(.057)	(.057)
Practice w/o AI vs. See AI example	.36***	.49***	.34***	.35***	.29**	.22***	.06	.39***	.38***	.14**
	(.056)	(.056)	(.056)	(.057)	(.106)	(.056)	(.056)	(.056)	(.056)	(.056)
Practice w/ AI vs. See AI example	.04	.08	.01	.02	.06	-.03	-.02	.11*	.06	-.01
	(.056)	(.056)	(.056)	(.057)	(.104)	(.056)	(.056)	(.056)	(.056)	(.056)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

F3. Seeing AI examples was less effortful. Table S28 shows OLS models predicting practice effort metrics from practice condition. Results show that participants seeing an AI example expended considerably less effort, measured subjectively or objectively, through keystrokes or practice time, when compared both to participants who practiced with AI and without it. As in Study 2, participants who practiced with AI still expended less effort than those who practiced without it. As pre-registered, time is square-root-transformed, and keystrokes are log-transformed. Differences are slightly smaller when using untransformed variables.

Table S28. Practice effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/min
Means						
Practice wo AI	2.83	5.01	6.17	9.00	259.34	27.62
	(.270)	(.541)	(.642)	(1.495)	(99.161)	(13.213)
Practice w AI	2.71	4.05	5.89	8.65	228.45	21.20
	(.270)	(.540)	(.641)	(1.493)	(99.050)	(13.200)
See AI example	1.85	.81	5.52	4.99	24.98	-8.17
	(.271)	(.542)	(.643)	(1.499)	(99.392)	(13.246)
Effect Sizes (d)						
Practice wo AI vs. Practice w AI	-.14*	-.55***	-.14*	-.07	-.10	-.15**
	(.056)	(.056)	(.057)	(.056)	(.056)	(.056)
Practice wo AI vs. See AI example	-1.13***	-2.41***	-.32***	-.83***	-.73***	-.84***
	(.059)	(.067)	(.056)	(.057)	(.056)	(.056)
Practice w AI vs. See AI example	-.99***	-1.86***	-.18**	-.76***	-.64***	-.69***
	(.058)	(.063)	(.056)	(.057)	(.056)	(.056)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S29 shows OLS models predicting test effort metrics from practice condition. Results show some differences: participants who had seen the AI example write for longer duing the test, and pressed more keys, however their subjective experience of effort was not different from those who practice with or without the AI tool.

Table S29. Test effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/min
Means						
Practice wo AI	2.45	5.40	7.05	6.14	432.01	68.46
	(.170)	(.562)	(.623)	(.655)	(109.856)	(19.117)
Practice w AI	2.50	5.52	7.19	6.35	466.90	72.19
	(.169)	(.561)	(.622)	(.654)	(109.733)	(19.090)
See AI example	2.57	5.86	7.21	6.63	517.26	77.84
	(.170)	(.563)	(.625)	(.656)	(110.112)	(19.159)
Effect Sizes (d)						
Practice wo AI vs. Practice w AI	.09	.06	.07	.10	.10	.06
	(.057)	(.056)	(.057)	(.057)	(.056)	(.057)
Practice wo AI vs. See AI example	.22***	.25***	.08	.23***	.24***	.15**
	(.056)	(.055)	(.056)	(.056)	(.055)	(.056)
Practice w AI vs. See AI example	.13*	.19***	.01	.13*	.14*	.09
	(.056)	(.056)	(.056)	(.056)	(.055)	(.056)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table S30 shows OLS models predicting learning rate metrics from practice condition. Learning rate is defined as the difference between test and pretest, divided by the effort metric. It shows how many points (10 point scale) the participant improved per unit effort (e.g., per minute spent practicing). Participants who had seen an AI example improved their skill more efficiently.

Table S30. Learning rate differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/min
Means						
Practice wo AI	.20	.15	.18	.31	.34	.20
	(.155)	(.272)	(.085)	(.278)	(.167)	(.277)
Practice w AI	.30	.22	.25	.45	.49	.30
	(.155)	(.272)	(.085)	(.277)	(.167)	(.277)
See AI example	.51	.97	.28	1.08	.62	1.00
	(.155)	(.273)	(.085)	(.278)	(.168)	(.278)
Effect Sizes (d)						
Practice wo AI vs. Practice w AI	.21***	.08	.25***	.15**	.27***	.12*
	(.057)	(.057)	(.057)	(.057)	(.057)	(.057)
Practice wo AI vs. See AI example	.62***	.94***	.38***	.86***	.53***	.90***
	(.057)	(.058)	(.056)	(.058)	(.057)	(.058)
Practice w AI vs. See AI example	.41***	.86***	.13*	.71***	.25***	.78***
	(.057)	(.058)	(.056)	(.057)	(.056)	(.058)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

F4. Seeing an AI example did not discourage motivation for future learning. Table S31 presents differences in perceived learning, perceived writing skill, and the likelihood of asking for feedback across conditions, with effect sizes and means reported for each comparison. Despite objectively learning more, participants who practiced with AI and saw an AI example perceived their learning and skill levels to be similar to those who practiced without AI and asked for feedback at comparable rates.

Table S31. Differences in motivational variables by condition

	Perceived learning	Perceived writing skill	Asked for feedback
Means — (SE)			
Practice w/o AI	5.26	6.33	.64
	(.549)	(.517)	(.670)
Practice w/ AI	5.25	6.36	.46
	(.549)	(.516)	(.669)
See AI example	5.42	6.23	.55
	(.551)	(.518)	(.671)
Effect Sizes (d)			
Practice w/o AI vs. Practice w/ AI	-.01	.02	1.19
	(.057)	(.057)	(.122)
Practice w/o AI vs. See AI example	.09	-.06	1.10
	(.056)	(.056)	(.121)
Practice w/ AI vs. See AI example	.09	-.08	0.921
	(.056)	(.056)	(.120)

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

F5. The benefits of seeing an AI example were just as large a day later. Table S32 shows means and standardized differences for measures of writing skill and related outcomes during the follow-up phase. The main specification demonstrates that participants who practiced with AI continued to outperform those who did not practice or practiced without AI. Robustness checks, including using a different language model (Column 2), excluding control variables (Column 3), and removing participants who admitted to cheating (Column 4) confirm the consistency of these effects. The results also hold when evaluating each of the five principles separately (Columns 5–9). These findings suggest that the benefits of practicing with AI are durable and persist even after participants stop using the tool.

Table S32. Differences in writing quality by condition in the follow-up phase

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	LM	ER	EN	F	ER
Means — (SE)									
Practice w/o AI	4.95	5.10	4.87	4.88	5.32	6.83	5.40	1.98	5.23
	(.776)	(.798)	(.109)	(.110)	(.750)	(.730)	(.847)	(1.829)	(.905)
Practice w/ AI	5.37	5.67	5.34	5.38	5.57	6.98	5.76	2.91	5.61
	(.781)	(.804)	(.103)	(.105)	(.756)	(.735)	(.853)	(1.842)	(.912)
See AI example	5.40	5.71	5.37	5.36	5.54	6.98	5.87	3.14	5.46
	(.781)	(.804)	(.103)	(.105)	(.756)	(.735)	(.854)	(1.843)	(.912)
Effect Sizes (d) — (SE)									
Practice w/o AI vs. Practice w/ AI	.29**	.40***	.32**	.34**	.18	.11	.24*	.28**	.23*
	(.106)	(.107)	(.104)	(.106)	(.106)	(.106)	(.106)	(.106)	(.106)
Practice w/o AI vs. See AI example	.32**	.43***	.35***	.33**	.16	.11	.31**	.35***	.14
	(.106)	(.107)	(.104)	(.106)	(.106)	(.106)	(.106)	(.107)	(.106)
Practice w/ AI vs. See AI example	.02	.03	.02	-.01	-.02	-.00	.07	.07	-.09
	(.104)	(.104)	(.101)	(.103)	(.104)	(.104)	(.104)	(.104)	(.104)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

The follow-up analyses pool three separate follow-up samples collected on consecutive days. Table S33 are the results for each of these samples separately.

Table S33. Differences in writing quality by condition in the follow-up phase by data collection batch

	Overall	Batch 1	Batch 2	Batch 3
Means — (SE)				
Practice w/o AI	4.95	6.02	5.47	4.85
	(.776)	(1.539)	(.492)	(.811)
Practice w/ AI	5.37	6.64	5.57	5.41
	(.781)	(1.446)	(.483)	(.817)
See AI example	5.40	6.51	5.88	5.32
	(.781)	(1.502)	(.472)	(.818)
Effect Sizes (d) — (SE)				
Practice w/o AI vs. Practice w/ AI	.29**	.43	.07	.40**
	(.106)	(.387)	(.187)	(.147)
Practice w/o AI vs. See AI example	.32**	.34	.29	.34*
	(.106)	(.354)	(.190)	(.149)
Practice w/ AI vs. See AI example	.02	-.09	.22	-.06
	(.104)	(.344)	(.177)	(.145)

Note. GPT-4o is the main specification. Ex. Controls is the main specification, unadjusted for demographic and pretreatment variables, Ex. Cheaters excludes the 3% of participants who admitted to cheating on the test phase. LM to ER are disaggregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

As mentioned in the main text the effect of practicing with AI and seeing an AI example did not become attenuated one day later. See Table S34.

Table S34. OLS model interacting condition with phase (Test vs. Follow-up) shows no attenuation of the effect of practicing with AI

Term	Estimate	SE	t	*p*-value
Intercept	-8.469	4.502	-1.881	0.060
Condition: Practice w AI	0.421	0.076	5.551	0.000
Condition: See AI example	0.467	0.075	6.253	0.000
Condition: Practice w/ AI $\times$ Follow-up	-0.007	0.158	-0.044	0.965
Condition: See AI example $\times$ Follow-up	-0.006	0.157	-0.036	0.971
Phase: Follow-Up	0.356	0.114	3.139	0.002
Pretest score	0.394	0.031	12.877	0.000
Year of birth	0.006	0.002	2.524	0.012
Gender: Male	-0.096	0.055	-1.724	0.085
Gender: Other	0.117	0.249	0.470	0.639
Race/ethnicity: White	0.215	0.117	1.846	0.065
Race/ethnicity: Black	-0.140	0.125	-1.125	0.261
Race/ethnicity: Asian	0.575	0.148	3.880	0.000
Race/ethnicity: Latino	0.216	0.123	1.756	0.079
Race/ethnicity: Other	-0.180	0.250	-0.719	0.472
Education: High school graduate (high school diploma or equivalent including GED)	-0.184	0.397	-0.464	0.642
Education: Some college but no degree	-0.187	0.394	-0.475	0.635
Education: Associate degree in college (2-year)	-0.207	0.399	-0.518	0.605
Education: Bachelor's degree in college (4-year)	-0.121	0.390	-0.310	0.756
Education: Master's degree	-0.033	0.393	-0.085	0.933
Education: Doctoral degree (PhD)	-0.227	0.439	-0.518	0.604
Education: Non-PhD Professional degree (JD, MD)	0.099	0.456	0.218	0.828
Writing skill	0.045	0.017	2.578	0.010
Motivation: Hardly motivated	-0.196	0.258	-0.760	0.447
Motivation: Somewhat motivated	-0.142	0.243	-0.582	0.561
Motivation: Very motivated	-0.338	0.244	-1.388	0.165
Motivation: Extremely motivated	-0.430	0.249	-1.731	0.084
Experience: I have tried AI writing assistant(s) but hardly ever use them	0.132	0.079	1.678	0.093
Experience: I use AI writing assistant(s) a few times per week	0.003	0.094	0.034	0.973
Experience: I use AI writing assistant(s) about once a week	0.070	0.095	0.730	0.465
Experience: I use AI writing assistant(s) every day	-0.063	0.131	-0.480	0.631

F6. Seeing AI examples was equally effective across subgroups. As in Study 2, we tested whether each of the pretreatment demographic variables moderated the effects of seeing an AI example. To do this, we ran separate linear in which writing skill during the test phase was regressed on condition, the pre-treatment moderator of interest, writing skill at baseline, and an interaction term between the moderator × condition. After correcting the p-values for the interaction terms, none were significant at the .05 level, suggesting that seeing AI examples was equally effective across groups.

Table S35. *p*-values for interaction terms predicting each outcome by condition and pre-treatment variables.

	Test		Follow-Up		Time Practice		Keys Practice		Effort Practice		Per. Learning		Per. Skill		Want Feedback	
Level	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**	**PAI**	**AIE**
Continuous Moderators																
Pretest	0.565	0.546	0.708	0.945	0.987	0.857	0.987	0.940	0.987	0.405	0.987	0.967	0.987	0.576	0.565	0.274
YOB	0.961	0.987	0.967	0.857	0.565	0.855	0.405	0.405	0.940	0.763	0.516	0.724	0.987	0.724	0.763	0.871
Writing Skill	0.943	1.000	0.405	0.987	0.878	0.987	0.967	0.871	0.763	0.707	0.986	0.405	0.900	0.987	0.434	0.405
Gender																
Male	0.816	0.987	0.987	0.987	0.987	0.532	0.793	0.535	0.565	0.405	0.450	0.703	0.724	0.535	0.707	0.565
Other	0.446	0.565	0.900	0.987	0.405	0.724	0.842	0.728	0.426	0.791	0.655	0.987	0.987	0.763	0.793	0.791
Race																
White	0.822	0.724	0.499	0.605	0.811	0.987	0.967	0.855	0.405	0.937	0.987	0.763	0.987	0.533	0.987	0.797
Black	0.718	0.405	0.791	0.718	0.987	0.901	0.987	0.899	0.274	0.565	0.760	0.945	0.987	0.499	0.793	0.446
Asian	0.940	0.766	0.734	0.532	0.703	0.405	0.934	0.605	0.341	0.565	0.937	0.987	0.989	0.280	0.535	0.565
Latino	0.987	0.899	0.405	0.405	0.763	0.987	0.405	0.816	0.405	0.940	1.000	0.987	0.857	0.624	0.987	0.991
Other	0.405		0.855		0.987		0.855		0.987		0.987		0.987		0.987	
Education Level																
High School Graduate	0.940	0.900	0.749	0.565	0.405	0.938	0.718	0.987	0.405	0.405	0.987	0.967	0.987	0.763	0.987	0.734
Some College	0.987	0.987	0.791	0.565	0.405	0.900	0.707	0.986	0.405	0.405	1.000	0.987	0.987	0.878	0.987	0.987
Associate Degree	0.900	0.987	0.791	0.565	0.447	0.855	0.707	0.940	0.446	0.520	0.987	0.987	0.989	0.811	0.987	0.814
Bachelor's Degree	0.986	0.987	0.793	0.499	0.405	0.899	0.707	0.967	0.405	0.405	1.000	0.987	0.989	0.791	0.987	0.964
Master's Degree	0.986	0.987	0.724	0.405	0.405	0.940	0.724	0.964	0.405	0.434	0.987	0.987	0.987	0.763	0.987	0.987
Doctoral Degree	0.987	0.987	0.524	0.724	0.516	0.832	0.987	0.940	0.406	0.516	0.987	0.987	0.832	0.760	0.987	0.964
Professional Degree	0.987	0.900	0.987	0.943	0.763	0.987	0.987	0.987	0.707	0.405	0.763	0.938	0.763	0.749	0.987	0.987
Motivation																
Hardly Motivated	0.987	0.900	0.624	0.899	0.768	0.763	0.832	0.987	0.763	0.900	0.685	0.707	0.760	0.535	0.405	0.565
Somewhat Motivated	0.977	0.535	0.763	0.763	0.871	0.585	0.937	0.987	0.763	0.791	0.791	0.763	0.797	0.707	0.536	0.763
Very Motivated	0.907	0.705	0.778	0.703	0.986	0.405	0.797	0.987	0.797	0.708	0.847	0.857	0.847	0.752	0.451	0.749
Extremely Motivated	0.855	0.536	0.724	0.707	0.763	0.724	0.763	0.987	0.987	0.763	0.987	0.900	0.763	0.763	0.535	0.535
Experience with AI Writing Assistants																
Hardly Ever Use Them	0.724	0.763	0.907	0.791	0.405	0.900	0.405	0.987	0.967	0.446	0.565	0.987	0.987	0.724	0.987	0.405
Use a Few Times Per Week	0.576	0.987	0.763	0.763	0.766	0.766	0.763	0.987	0.565	0.152	0.903	0.763	0.783	0.405	0.945	0.405
Use About Once a Week	0.846	0.707	0.623	0.786	0.406	0.855	0.405	0.987	0.763	0.718	0.967	0.987	0.899	0.405	0.707	0.797
Use Every Day	0.987	0.724	0.987	0.763	0.341	0.763	0.405	0.987	0.405	0.280	0.943	0.987	0.899	0.763	0.734	0.987

Note. Models for test and follow-up performance, square-root practice time, log keystrokes, subjective effort, perceived learning and perceived writing skill or OLS models. Asking to see feedback was a binary Yes/No variable, and was modelled with logistic regression. Models match the pre-registered main specification, and thus control for all other pre-treatment variables. Per. = Perceived, PAI = Practice with AI, AIE = See AI example.

F7. Pairwise Comparisons. The relative likelihood of a cover letter receiving an invitation to an interview was correlated with the GPT-rated writing quality. See Figure S11.

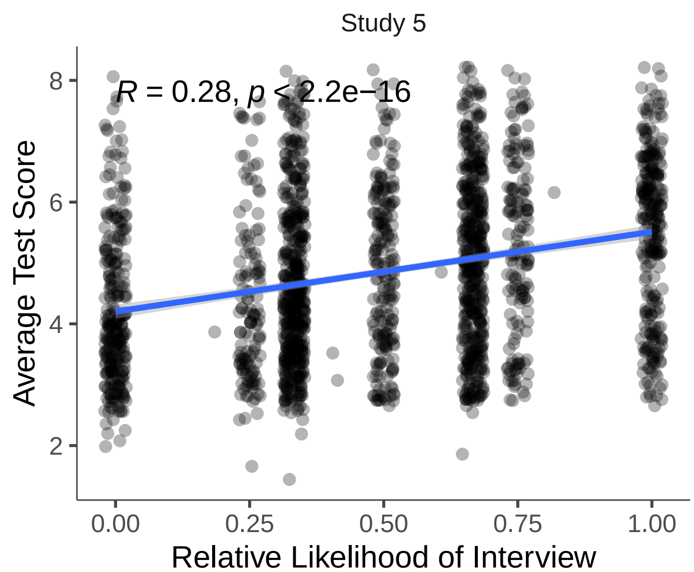

Fig. S11. Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

As shown in Table S36, participants who had practiced writing cover letters with AI were more likely to be invited to a hypothetical job interview.

Table S36. Beta regression models predicting relative likelihood of interview from condition.

	Without controls	With controls
Means (SE)		
Practice w AI	.50	.51
	(.01)	(.01)
Practice wo AI	.48	.48
	(.01)	(.01)
See AI example	.51	.51
	(.01)	(.01)
Model coefficients (SE)		
Precision (ϕ)	16.17***	7.77***
	(1.47)	(.51)
Symmetry (Log(ν))	.35***	-.23***
	(.08)	(.07)
Pairwise comparison (SE)		
Practice w AI - Practice wo AI	.08**	.09*
	(.03)	(.04)
Practice w AI - See AI example	-.02	-.03
	(.03)	(.04)
Practice wo AI - See AI example	-.10***	-.12**
	(.03)	(.04)
Statistics		
N	1934.00	1916.00
AIC	2447.07	2518.06
BIC	2474.91	2684.80
log(Likelihood)	-1218.54	-1229.03