Learning from examples: 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

It is widely believed that outsourcing cognitive work to AI boosts immediate productivity at the expense of long-term human capital development. An opposing possibility is that AI tools can support skill development by providing just-in-time, high-quality, personalized examples. This work explores whether using an AI writing tool undermines or supports performance on later unaided writing. In Study 1, forecasters predicted that practicing writing cover letters with an AI tool would impair learning compared to practicing alone. However, in Study 2, participants randomly assigned to practice writing with AI improved more on a subsequent writing test than those assigned to practice without AI (d = 0.40*)—despite exerting less effort, whether measured by time on task, keystrokes, or subjective ratings. In Study 3, participants who had practiced writing with AI again outperformed those who practiced without AI (d = 0.31***). Consistent with the positive impact of exposure to high-quality examples, these participants performed just as well as those who viewed–but could not edit–an AI-generated cover letter (d = 0.03, *ns*). In both Studies 2 and 3, the benefits of practicing with AI persisted in a one-day follow-up writing test. Collectively, these findings constitute an existence proof that, contrary to participants' intuition, using AI tools can improve, rather than undermine, learning.**

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, 6). For instance, in a 2024 poll, 62% of surveyed adults predicted that Generative AI will "lead to humans becoming less intelligent" (7). In January 2023, New York City public schools banned ChatGPT, citing "concerns about negative impacts on student learning (8)." When this ban was lifted three months later, it was not because of the potential of AI 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 (9)." The sentiment behind the initial ban aligns with teacher perceptions: 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%) (10).

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 (11)), make reasoning and arithmetic errors (12), and complete other tasks with varying degrees of accuracy.

Second, regardless of accuracy, 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 (13).

Third, technological tools reduce the need for the learner to be cognitively engaged with the task at hand. For instance, knowing that we will be able to 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 (14). And drivers who use GPS tend to have worse hippocampal-dependent spatial memory, both cross-sectionally and longitudinally (15, 16). To the extent that tools powered by generative AI 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.

There is a plausible, albeit less obvious, alternate hypothesis. Using AI tools to do our work alongside us could 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 and apply. 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 explanations alone (17, 18). Compared to textbooks, conventional computer tutoring programs (e.g., (19)), and even human teachers, today's AI tools are unprecedented in their ability to deliver hyper-personalized high-quality examples on command. Thus, AI tools may improve skill development if the benefits of exposure to just-in-time examples tailored to learners' needs outweigh the costs of diminished engagement.

There is little research on the effect of AI tools on subsequent skill development, as opposed to productivity during use. In working papers, results have been mixed. Some studies have found that interacting with AI tools improves skill on subsequent tests in which AI tools are not available (20, 21), while others have shown null or even negative effects (21–23). 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 focus exclusively on mathematics and computer programming.

In this investigation, we ask whether AI tools can increase intermediate-term skill development, above and beyond just 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 (24). Participants in our three pre-registered studies were American adults on the survey platform Prolific.

To differentiate the effects of AI use on learning versus performance, we developed a paradigm in which all participants were given a baseline writing test (i.e., revising a poorly written cover

DRAFT

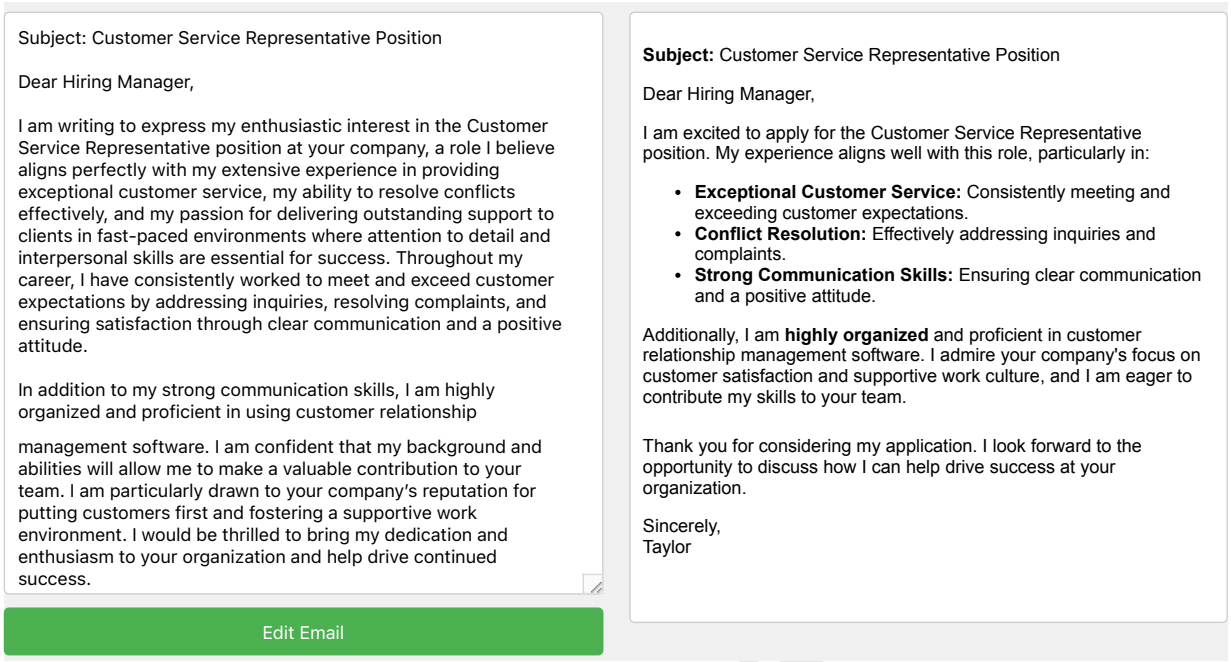

Fig. 1. The AI writing tool we created for this investigation takes inputted text (left panel) and generates a version that incorporates recommended writing principles (right panel). Users could edit the text in the left-hand box, which was prepopulated with the text they were instructed to rewrite. They could then copy and paste the revised output from the right-hand box.

letter). They were then introduced to evidence-based strategies for professional writing, with descriptions and examples for each one (25, 26). Next, participants were randomly assigned to one of three conditions: practicing rewriting a different cover letter with access to an AI tool based on the same writing principles that users had just learned (See Figure 1), practicing without this tool, or a no-practice control group. To assess gains in writing skill, participants completed an incentivized test in which they rewrote yet another cover letter without access to AI, with a cash bonus guaranteed for submissions ranked in the top 10 percent. To assess the persistence of skill improvement, all participants were invited to complete a similar incentivized test of writing skill one day later. See Figure 2. We used GPT4o to rate each cover letter for each of the five writing principles introduced in this experiment. We averaged these ratings to produce summary scores of writing skill and, in a random subsample ($n = 30$), validated these scores using trained human raters ($r = .83$, $p < .001$). 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. The relative likelihood of a cover letter securing an interview correlated positively with AI ratings of writing skill ($rs = .29$ and $.28$, $ps < .001$, for Study 2 and 3, respectively).

In Study 1, forecasters presented with this design were twice as likely to predict that practicing writing with the assistance of the AI tool would impair learning compared to practicing without the AI tool. In Study 2, however, participants who had practiced with the AI tool learned more (i.e., wrote better cover letters during the test phase) compared to either comparison group—an advantage that persisted in a one-day follow-up test. Finally, in Study 3, we explored the mechanism for these learning gains by introducing an example-only condition. Participants who had merely seen an AI-generated example (but did not have an opportunity to practice) improved in writing skill as much as participants who had practiced with an AI tool; benefits again persisted in a one-day follow-up test. Test-phase cover letters written by participants who had practiced with AI (Studies 2 and 3) or had seen an AI example (Study 3) were more likely to secure hypothetical job interviews compared to cover letters written by participants who had not practiced (Study 2) or had practiced without AI (Studies 2 and 3).

Study 1

Lay forecasters predicted that practicing with AI would hinder 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. 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 ($OR = 0.83$, $p = .239$) or any other measured demographic characteristic ($ps > .05$).

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 gives the answer for them.", "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 ways of learning, and AI can provide this.")

DRAFT

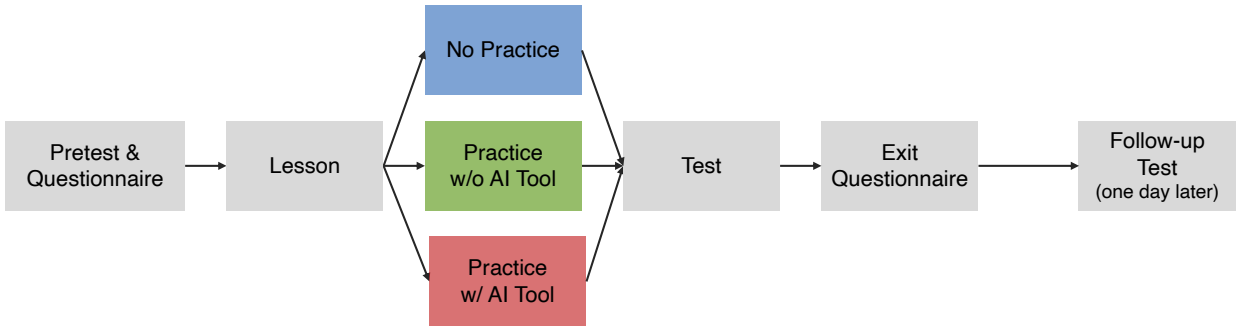

Fig. 2. Experimental design for Study 2. First, all participants completed a baseline questionnaire, a pretest (rewriting a poorly-written cover letter), and a lesson introducing five evidence-based principles of effective writing (25). Next, participants were randomly assigned to one of three conditions: 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.

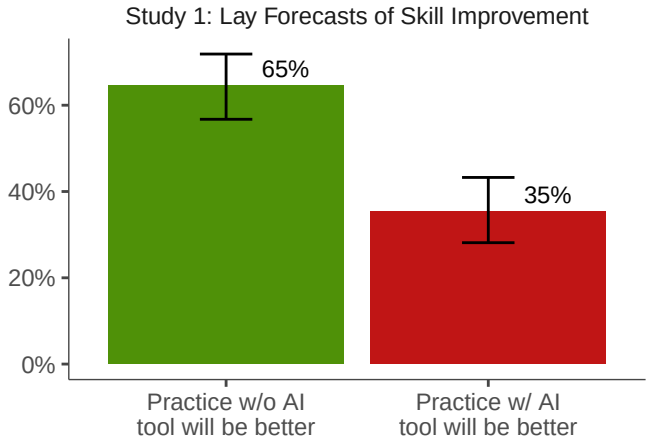

Fig. 3. Forecasters predicted that practicing without an AI tool would improve writing skill more than practicing with an AI tool. Error bars represent proportions ± 1 SE.

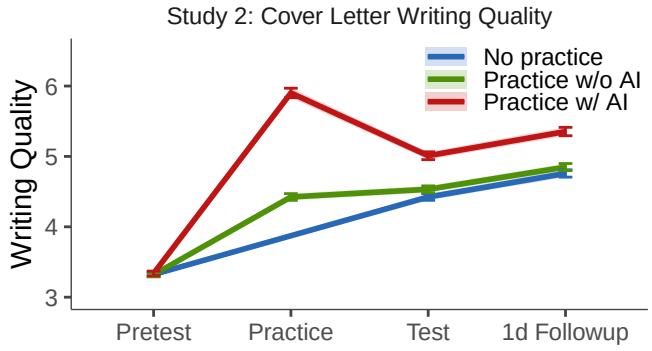

Fig. 4. In both the main test and the follow-up, participants who had practiced with the AI tool outperformed those who practiced without it and those who did not practice at all. Error bars represent means ± 1 SE. Means shown are for the subsample of participants (n = 1,294) who completed the one-day follow-up test. See Figure S4 for the equivalent figure in the full sample, excluding the one-day follow-up phase (N = 2,238).

Study 2

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 revised text instantly 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 yet another 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.

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 who were not (d = 1.01, p <.001).

The learning advantage of having practiced with AI was evident in the test phase: Consistent with our pre-registered hypothesis, participants who had practiced with the AI tool produced higher-quality writing than did participants who either had practiced without the AI tool (d = .38, p <.001) or who had not practiced at all (d = .47, p <.001). See Figure S4. 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 5.

AI practice was less effortful. The learning benefits of using an AI writing tool were evident despite reduced effort during the practice phase. 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), logged roughly a quarter as many keystrokes (26% d = -.44, p <.001) 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. Copying, pasting, and submitting the AI tool's output could be accomplished almost instantly. Yet, the majority of participants chose to interact

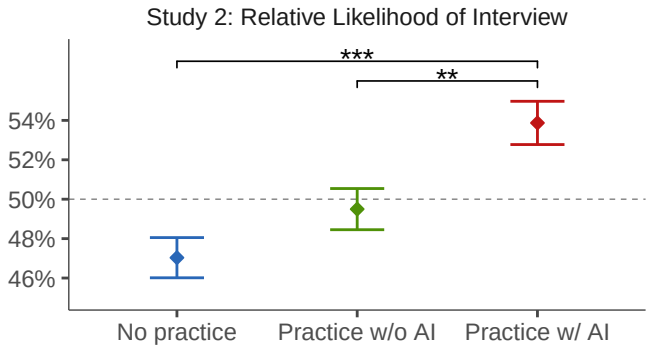

Fig. 5. During the test phase, cover letters written by participants who had practiced with AI were more likely to lead to interview offers than those from other conditions. Points depict the average proportion of times each cover letter was preferred in pairwise comparisons with cover letters from the other two conditions. Error bars represent proportions ± 1 SE. The dashed line at 50% represents no preference; values above this line indicate that cover letters were more likely to be preferred, while values below indicate they were less likely to be preferred.

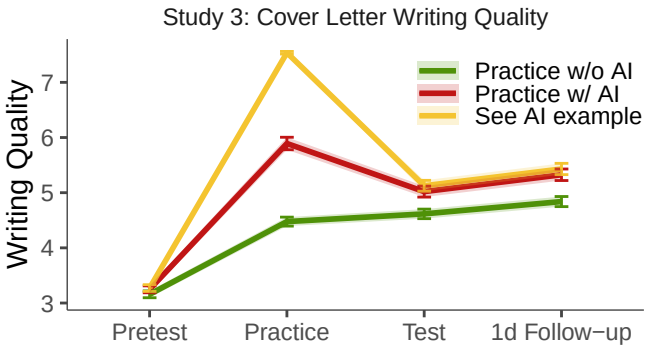

Fig. 6. In both the main test and the follow-up, participants who simply saw an AI-generated example improved just as much as those who practiced with AI and more than those who practiced without AI. Error bars represent means ± 1 SE. Means shown are for the subsample of participants (n = 608) who completed the one-day follow-up test. See Figure S6 for the equivalent figure in the full sample, excluding the one-day follow-up phase (N = 2,003).

with the task for over 3 minutes, and 95% made at least one edit to the AI-generated output before final submission. See Figure S5 in Supplementary Information for details.

Differential effort during practice raised the possibility that participants who had practiced with the AI tool outperformed those who had practiced on their own because they were less fatigued during the test phase. However, participants who had practiced with AI did not spend more time as those who had practiced without AI (d = .06, p = .235), but logged more keystrokes (d = .12, p = .026), and self-reported expending less effort (d = -.12, p = .023) during the test phase.

AI practice did not create the illusion of mastery. Following the test phase, there were no differences by condition on self-reported knowledge or motivation to improve. Despite improving more in objectively assessed writing skill, participants who had practiced with AI reported having learned no more than those who had either practiced alone or done no practice at all (ps >.05). Self-ratings of writing skill after the practice phase were also indistinguishable between participants who practiced with AI and those who practiced without the AI tool, or in the no-practice control group (ps >.05), but participants who had practiced without AI rated their skill more highly than those who did not practice (d = .14 p = .008). Finally, compared to participants who did not practice, participants who had practiced with AI were slightly less likely to request feedback after the test phase (.65 vs. .60, p = .039), but just as likely as participants who had practiced without AI (.64 vs. .60, p = .167). See Section B4 in the Supplementary Information for details.

The effects of practicing with AI persist. To examine whether the treatment effects persisted over time, we re-contacted all participants one day later. The majority of 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 to practice 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). See Figure 4 and Section B5 of Supplementary Information for details.

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 S12 in the Supplementary Information for details.

Study 3

To gain insight into what might be driving the benefit of practicing with AI, in Study 3 (N = 2003), 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, participants were shown 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 = 1.01, p <.001), and during the no-AI test phase (d = .28, p < .001). Their test-phase cover letters were also relatively more likely to secure them hypothetical job interviews (.51 vs. .47, p = .007).

Participants who had merely seen an AI-generated example also improved more in writing skill than those who had practiced without AI (d = .31, 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 S6 and 7.

Seeing an AI example was even less effortful than practicing with AI. During the practice phase, participants who saw the AI

DRAFT

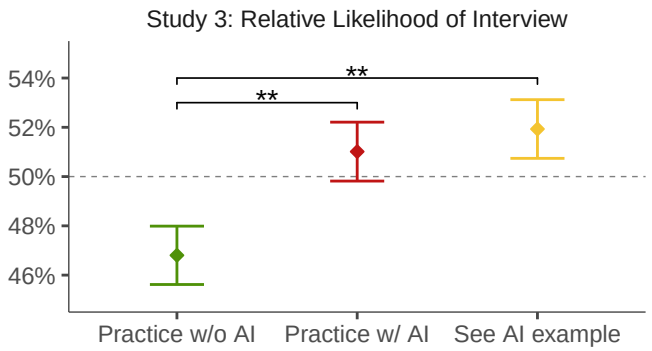

Fig. 7. During the test phase, cover letters written by participants who had seen an AI-generated example were about equally likely to lead to interview offers when compared to those assigned to practice with AI. Cover letters from both AI conditions outperformed those written by participants who were assigned to practice without AI. Points depict the average proportion of times each cover letter was preferred in pairwise comparisons with cover letters from the other two conditions. Error bars represent proportions ± 1 SE. The dashed line at 50% represents no preference; values above this line indicate that cover letters were more likely to be preferred, while values below indicate they were less likely to be preferred.

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 2.83 times fewer keystrokes than did participants who practiced without AI. As expected, participants exposed to the AI example logged 0 keystrokes.

During the test phase, participants who had seen the AI example worked for an additional 55 seconds more than participants who had practiced without AI ($d = 0.22$, $p < .001$). They also logged 30% more keystrokes than participants who had practiced with AI ($d = 1.86$ $p <.001$) and 48% more than those who practiced without AI ($d = 2.39$, $p < .001$). Across conditions, all participants self-reported similar levels of subjective effort ($ds < .08$, $ps > .05$).

Seeing an AI example did not create the illusion of mastery. As in Study 2, 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 C4 in the Supplementary Information for details.

The effects of seeing an AI example persist. When we recontacted a subsample of ($n = 800$) participants one day later, the majority responded ($n = 618$, 77%); 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 ($d = .32$, $p = .005$) and participants who had simply seen an AI example ($d = .39$, $p < .001$), 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 ($d = .07$, $p = .745$). See Figure 6 and Section C5 of Supplementary Information for details.

As in Study 2, the above findings were not moderated by individual differences. See Table S23 in the Supplementary Information for details.

Discussion

Contrary to the expectations of people shown our paradigm (Study 1), access to the AI writing tool led to improvements in writing skill (unassisted by AI) that persisted in a one-day follow-up test (Study 2). These learning gains cannot be attributed to greater effort, as participants who practiced with AI expended less effort during practice than those who practiced alone. Instead, these gains may be explained by exposure to a high-quality, just-in-time personalized example: participants who merely viewed an AI-generated example cover letter (without editing it) improved their writing skill just as much as those who given the option to practice editing the cover letter (Study 3).

While not obvious to forecasters (Study 1), teachers (10), and the general population (7), the benefits of seeing AI-generated examples are consistent with prior research on how people learn. In addition to the literature on worked examples mentioned earlier, research has shown that humans are especially adept at observing, imitating, and learning from others (27–29). 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 (30).

Future Directions. We want to highlight three promising directions for future research. First, we focused on the modal use case for AI tools in the workplace—professional writing (24). It is unclear whether learning from AI examples occurs in other domains. Indeed, in mathematics and computer programming, AI tools do not always support learning (22, 23). Indeed, writing may be particularly suited to learning by example. At a glance, a single AI-generated example visually communicates the elements of effective professional writing (e.g., shorter sentences and the strategic use of line breaks and bold-face formatting). In other domains, merely observing a solution may be less informative. For instance, the final answer to a math problem does not reveal the procedure that produced it.

Second, additional research is needed to explore moderators of learning from AI tools. Certain strategies for interacting with AI may bolster their effectiveness. 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 (20). Similarly, correlational evidence that asking AI for explanations as opposed to answers is associated with more learning in mathematics (22) and computer programming (21). On the other hand, other factors could 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

DRAFT

and learning, which we speculate would reduce the learning gains associated with practicing with AI.

Third, in our experimental paradigm, participants who interacted with the AI tool did so 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? Future research, ideally in field settings, is needed to establish the long-term benefits and costs of relying on 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 it reduced the effort users invested in practicing, the AI writing tool nevertheless accelerated skill development. Id accomplished this by providing high-quality, just-in-time, personalized examples of excellent writing. The underappreciated efficacy of timely and tailored examples has practical implications for the design of AI tools. Many AI tools designed to support learning are explicitly programmed not to "give away" answers. 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 baseball coach John Wooden declared that the four laws of learning are explanation, demonstration, imitation, and repetition (31). 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 to build 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 our studies were pre-registered. See pre-registrations for Study 1 at https://aspredicted.org/x9mm-7qwp.pdf.

Study 2 was run twice because of a technical issue that caused imbalanced missing data issue the first time it was run. The findings in this sample were consistent with the ones reported here. Pre-registration is available at https://aspredicted.org/4sw4-mpny.pdf. The version of Study 2 presented above was pre-registered https://aspredicted.org/xyyn-gmzc.pdf.

Pre-registartion for Study 3 is available at https://aspredicted.org/3mty-fcfy.pdf. The one-day follow-up for Study 3 was collected in three batches. We pre-registered batch 2 https://aspredicted.org/5jcx-bhg9.pdf, but report pooled results in the main text. Details on the separate batches are available in Supplementary Information Section C5.

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 ($N = 150$) were predominantly female (n = 93, 62%), and ranged in age 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%), and ranged in age from 18 to 82 (M = 36.0, SD = 12.5). Over half of the sample (58%) was white, with the rest being comprised by Black (33%), latino (6%), and asian (5%). 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).

Study 3 had similar prooportions 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 rest being comprised of 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. Participants first saw an introductory screen about what they were about to do. They then completed a brief questionnaire where they reported their demographics, their experience with AI writing tools, their motivation to improve their writing, and their perceived writing skill. They then completed a 2-minute pre-test in which they saw a poorly worded cover letter and were asked to improve it. After this, all participants completed a lesson about the five principles of effective writing. They were then randomized to the practice condition, or skipped ahead if assigned to the no-practice control. During the practice phase, participants rewrote a different cover letter, or (in the see example condition) generated an AI rewritten version of that letter. This example was not explicitly labeled as AI-generated. Then, participants saw the text they submitted (or the AI-generated example), and below it, AI-generated feedback highlighting one way in which this letter could be improved. See more information about the feedback procedure in Supplementary Information section A3. Immediately after seeignt this feedback screen, all participants then completed two questions, reporting how much they had learned and how hard they had worked on the task so far. They then edited a new cover letter in an incentivized, 7-minute test without the help of any AI tools. To minimize the possibility of cheating, we used custom JavaScript to restrict copy-pasting functionality. Finally, participants were invited to see optional feedback and were asked if they had used any outside resources during the test.

A small percentage of participants (2.88%) admitted to cheating. As per our pre-registration, these participants are included in our analyses, but see Tables S6 and S16 in Supplementary Information to see results excluding them, which are consistent with our main interpretation.

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

DRAFT

judgments, offering reliability and consistency across various evaluation contexts (32, 33). 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 Table S6 and S16.

To validate these ratings, the first author and a trained research assistant took a sample of $n = 30$ cover letters from Study 2, and rated them on the 5 principles. The average of these two ratings correlated more highly with the computer ratings ($r = .83$, $p < .001$), than the average interrater reliability ($r = .69$, $p < .001$).

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 highly with Claude ratings ($r = .71$, $p < .001$), and that the effects are not attenuated by using different models (See Tables S6 and S16), suggesting that our effects are not explainable by same-model bias.

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 feedback for their test 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 use the Benjamini-Hochberg correction (34).

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. 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.
8. K. Roose, Don't Ban ChatGPT in Schools. Teach With It. *New York Times* (2023).
9. 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).
10. L. Lin, A quarter of U.S. teachers say AI tools do more harm than good in K-12 education (2024).
11. 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).
12. 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).
13. 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).
14. B. Sparrow, J. Liu, D. M. Wegner, Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science* **333**, 776–778 (2011).
15. 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).
16. 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).
17. 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).
18. 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).
19. 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).
20. H. Kumar, D. M. Rothschild, D. G. Goldstein, J. Hofman, Math Education with Large Language Models: Peril or Promise? *SSRN Electronic Journal* (2023).
21. M. Lehmann, P. B. Cornelius, F. J. Sting, Ai meets the classroom: When does chatgpt harm learning? *arXiv preprint arXiv:2409.09047* (2024).
22. H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, R. Mariman, Generative AI Can Harm Learning (2024).
23. 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).
24. A. Bick, A. Blandin, D. J. Deming, The rapid adoption of generative ai, *Tech. rep.*, National Bureau of Economic Research (2024).
25. T. Rogers, J. Lasky-Fink, *Writing for Busy Readers: Communicate More Effectively in the Real World* (Penguin, 2023).
26. H. C. Shulman, D. M. Markowitz, T. Rogers, Reading dies in complexity: Online news consumers prefer simple writing. *Science Advances* **10**, eadn2555 (2024).
27. A. Bandura, *Social Learning Theory* (General Learning Press, New York, 1971).
28. A. N. Meltzoff, Imitation and other minds: The like me hypothesis. (2005).
29. 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).
30. A. K. Ericsson, Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine* **15**, 988–994 (2008).
31. Charlie Rose : WHUT : June 14, 2010 9:00am-10:00am EDT (2010).
32. 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).
33. 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).
34. 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).

DRAFT

Supplementary Information for

Learning from examples: AI assistance can enhance rather than hinder skill development

Contents

A Additional methods **2**
A1 AI Ratings . . . 2
A2 Pairwise Comparisons . . . 2
A3 Feedback . . . 2

B Results Study 2 **5**
B1 Randomization, Balance, and Missingness . . . 5
B2 AI practice improved writing skill . . . 6
B3 AI practice was less effortful . . . 7
B4 Seeing an AI example did not discourage motivation for future learning . . . 9
B5 The effects of practicing with AI persist . . . 9
B6 AI practice was equally effective across subgroups . . . 9

C Results Study 3 **11**
C1 Randomization, Balance, and Missingness . . . 11
C2 AI examples improve writing skill . . . 12
C3 Seeing AI examples was less effortful . . . 15
C4 Seeing an AI example did not discourage motivation for future learning . . . 16
C5 The effects of seeing an AI example persist . . . 16
C6 Seeing AI examples was equally effective across subgroups . . . 18

DRAFT

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.

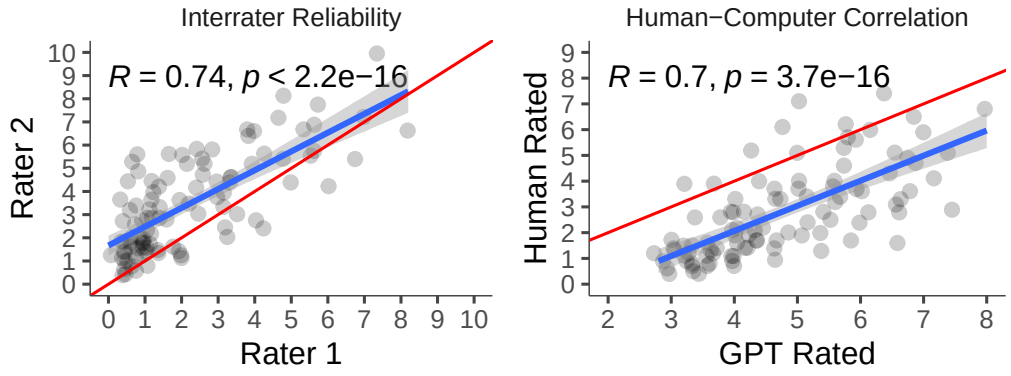

Fig. S1. Interrater correlations and correlations between AI and human ratings.

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 followup, 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, min, max, med (plot?). 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.

As shown in Figure S2, the relative likelihood of an interview correlated positively with the AI-generated measure of writing skill.

Table S2 shows the beta regressions for the relative likelihood by condition.

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 participants submission verbatim, then read "Here is one way in which it could be made better." The feedback was personalized and created by GPT-4o model.

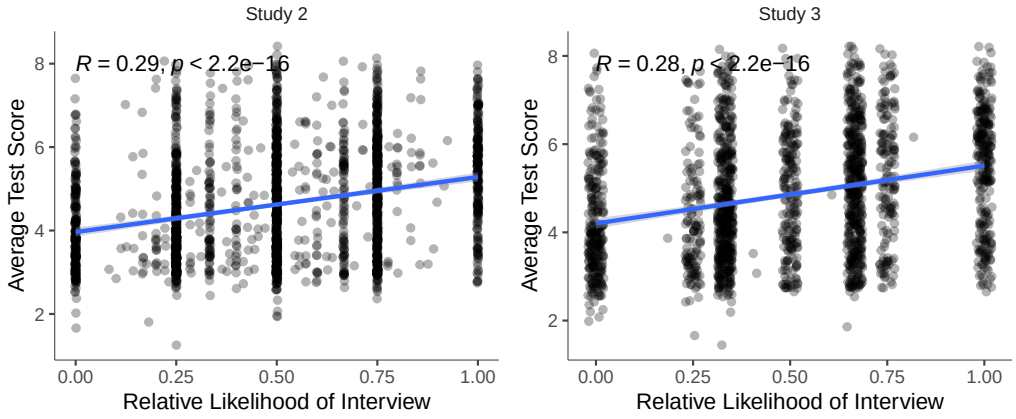

Fig. S2. Correlations between the relative likelihood of interview and AI-generated writing quality scores.

Table S2. Results from beta regressions predicting the relative likelihood of an interview from test phase cover letters. The reference category is practice with AI for Study 2, and practice without AI for Study 3.

	(1)	(2)	(3)	(4)
(Intercept)	0.085***	-4.816	-0.061**	-0.449
	(0.022)	(5.017)	(0.021)	(0.677)
No practice	-0.147***	-0.250***		
	(0.031)	(0.073)		
Practice wo AI	-0.096**	-0.110		-0.103
	(0.031)	(0.073)		(0.079)
Practice w AI			0.080**	
			(0.030)	
See AI example			0.097***	0.055
			(0.030)	(0.078)
Precision (ϕ)	12.899***		16.174***	
	(1.192)		(1.471)	
Symmertry (Log(ν))	-0.117		0.350***	
	(0.086)		(0.080)	
Demographic and baseline performance covariates	No	Yes	No	Yes
Num.Obs.	2188	2153	1934	1917
AIC	2204.7	-10885.8	2447.1	-14615.7
BIC	2233.2	-10721.3	2474.9	-14454.5
Log.Lik.	-1097.349		-1218.535	
RMSE		0.28		0.30

DRAFT

The feedack prompt is shown in Figure S3.

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 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. S3. Feedback prompt

B. Results Study 2

B1. 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 followup samples, and did not differ by condition.

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

Condition	Main Sample	Followup Sample
No practice	2.25%	13.38%
Practice wo 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.

DRAFT

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 wo 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

DRAFT

B2. 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).

Table S5. Practice effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	LM	ER	EN	F	ER
Means — (SE)										
Practice wo 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 wo 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. Followup is the subsample of non-attriting participants who returned to the one-day followup. LM to ER are disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

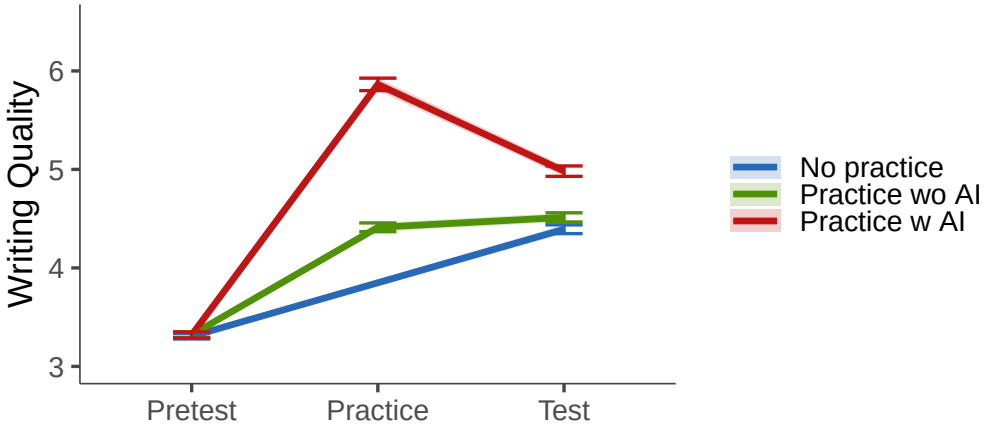

Fig. S4. Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means ± 1 SE. (*N* = 2,238).

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, $r = .06$, $p = .25$.

DRAFT

Table S6. Test effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	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 wo 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 wo 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 wo 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. Followup is the subsample of non-attriting participants who returned to the one-day followup. LM to ER are disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

B3. 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 wo 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 wo 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 wo 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 wo 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 wo 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 wo 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 wo 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 S5, 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.

DRAFT

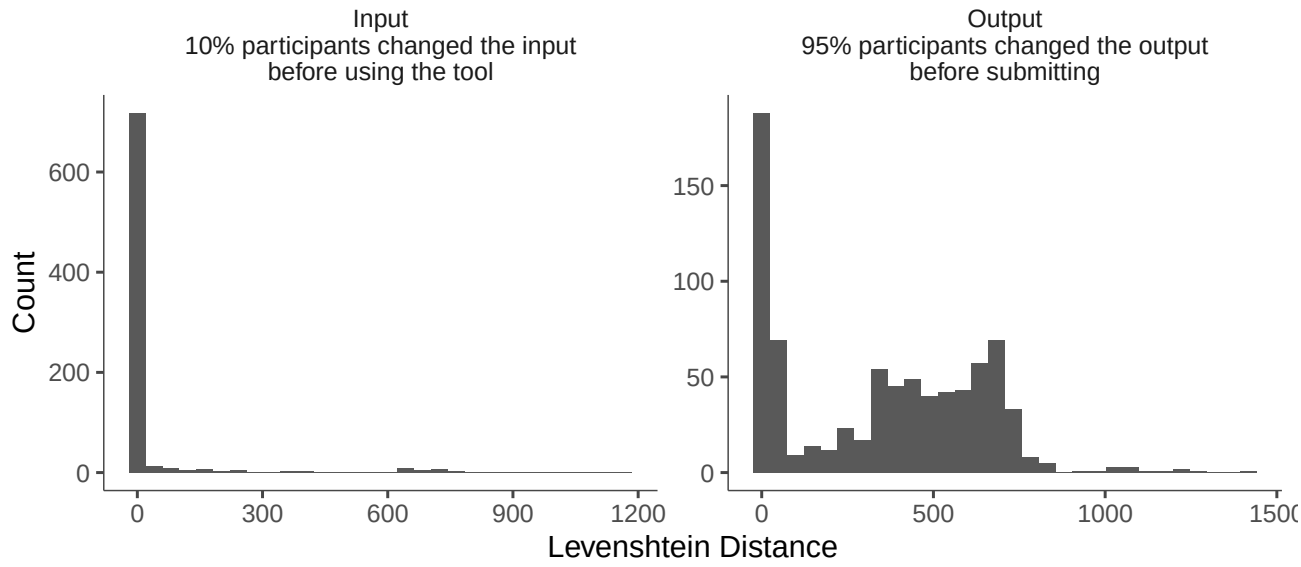

Fig. S5. Levenshtein distance (number of additions, modifications or deletions) between the original text and the text passed along to the AI tool (Input); and between the AI's output text and what users submitted as their final work (Output).

B4. Seeing an AI example 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 wo 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 wo AI	-.00	.13*	1.10
	(.053)	(.053)	(.128)
No practice vs. Practice w AI	.08	.10	1.30*
	(.053)	(.053)	(.149)
Practice wo AI vs. Practice w AI	.08	-.03	1.17
	(.053)	(.053)	(.134)

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

B5. The effects of practicing with AI persist. 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.

B6. 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

DRAFT

Table S11. Followup effects

	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 wo 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 wo 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 wo 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 disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

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.

DRAFT

Table S12. 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, and thus control for all other pre-treatment variables. Per. = Perceived

DRAFT

C. Results Study 3

C1. 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 S13, missingness and attrition rates were low for the main and follow-up samples and did not differ by condition.

Table S13. Missingness and attrition proportions and test in Study 3.

Condition	Main Sample	Followup Sample
Practice wo 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 S14.

Table S14. Randomization checks for pre-treatment variables. *p*-values are BH corrected. SMD = Standardized Mean Difference.

	Overall	Practice wo 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

DRAFT

C2. AI examples improve writing skill. The AI tool improved performance while participants used it. Table S15 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).

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 S16. For participants assigned to practice with the AI tool, The quality of AI rewrites did not correlate with participants' final submissions, $r = .06$, $p = .25$.

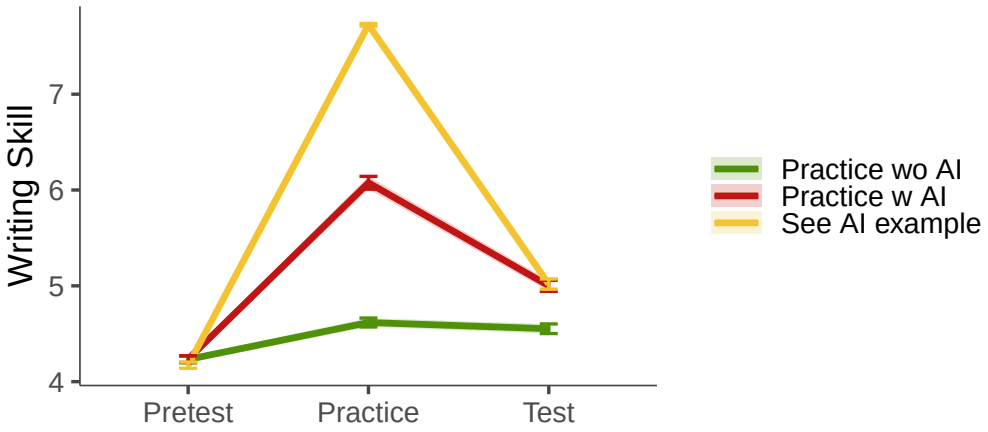

Fig. S6. Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means ± 1 SE. (*N* = 2,003).

DRAFT

Table S15. Practice effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	LM	ER	EN	F	ER
Means — (SE)										
Practice wo 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 wo 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 wo 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 disagregated 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 S16. Test effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	LM	ER	EN	F	ER
Means — (SE)										
Practice wo 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 wo 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 wo 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 disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

DRAFT

C3. Seeing AI examples was less effortful. Table S17 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 S17. Practice effort differences

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

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

Table S18 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 S18. Test effort differences

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

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

Table S19 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.

DRAFT

Table S19. Learning rate differences

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

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

C4. Seeing an AI example did not discourage motivation for future learning. Table S20 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 S20. Motivation

	Perceived learning	Perceived writing skill	Asked for feedback
Means — (SE)			
Practice wo 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)
extbfEffect Sizes (d)			
Practice wo AI vs. Practice w AI	-.01	.02	1.19
	(.057)	(.057)	(.122)
Practice wo 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. *** $p < .001$, ** $p < .01$, * $p < .05$.

C5. The effects of seeing an AI example persist. Table S21 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.

The follow-up analyses pool three separate follow-up samples collected on consecutive days. Table S22 are the results for each of these samples separately.

DRAFT

Table S21. Followup effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	LM	ER	EN	F	ER
Means — (SE)									
Practice wo 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 wo AI vs. Practice w AI	.29**	.40***	.32**	.34**	.18	.11	.24*	.28**	.23*
	(.106)	(.107)	(.104)	(.106)	(.106)	(.106)	(.106)	(.106)	(.106)
Practice wo 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 disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

DRAFT

Table S22. Follow-up effects by data collection batch

	Overall	Batch 1	Batch 2	Batch 3
Means — (SE)				
Practice wo 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 wo AI vs. Practice w AI	.29**	.43	.07	.40**
	(.106)	(.387)	(.187)	(.147)
Practice wo 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 disagregated scores for each of the five principles. LM = Less is More, ER = Easy Reading, EN = Easy Navigation, F = Formatting, ER = Easy Responding. *** $p < .001$, ** $p < .01$, * $p < .05$.

C6. 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 S23. Metrics 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.

DRAFT