EMPIRICAL EVIDENCE THAT USING AI TOOLS CAN ENHANCE HUMAN COGNITION

Benjamin Lira Luttges

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2025

Supervisor of Dissertation

Angela L. Duckworth, Rosa Lee and Egbert Chang Professor

Graduate Group Chairperson

Sara Jaffee, Professor of Psychology, University of Pennsylvania

Dissertation Committee

Lyle Ungar, Professor of Computer and Information Science, University of Pennsylvania
Martin Seligman, Zellerbach Family Professor of Psychology, University of Pennsylvania

ABSTRACT

EMPIRICAL EVIDENCE THAT USING AI TOOLS CAN ENHANCE HUMAN COGNITION

Benjamin Lira Luttges

Angela L. Duckworth

Many worry that AI tools boost productivity in the short-term, at a long-term cost in the development of human capability. Using AI will rob us of the opportunity to learn, will destroy our motivation for hard work, and will perpetuate the biases in its training data. I propose two theoretically distinct mechanisms through which AI influences human cognition–engagement and information–and provide three empirical demonstrations of AI tools proving information (Chapters 1 and 3) and motivation (Chapter 2).

In Chapter 1, I show that superior information—in the form of just-in-time, personalized examples—can compensate for decreased engagement, yielding net positive effects on skill development. Participants writing cover letters with AI assistance improved more on writing tests than those practicing without AI, despite exerting less effort. A second experiment showed participants learned more from exposure to AI-generated examples even when they could not edit them. These gains were unique to AI, outperforming static online examples and matching feedback from professional editors.

In Chapter 2, I show that an AI chatbot can increase motivation among Khan Academy users. In this year-long field investigation, I built two AI-based chatbots that delivered situation modification and emotional reframing interventions. Relative to the week before exposure to the chatbots, students increased their time on task by 10-11% and worked on more challenging problems after completing interventions.

Finally, in Chapter 3, I demonstrate the potential for AI to enhance human decision making in the high-stakes setting of college admissions. I fine-tuned a language model to assess personal qualities like leadership, training it on professional admissions officers' evaluations of 3,131 applicant

essays. The AI successfully reproduced expert judgment, and did so equally well across demographic subgroups, showing no evidence of bias. Scaling its application to a national sample of 309,594 applicants revealed that these scores incrementally predicted graduation rates without encoding demographic information. This demonstrates that AI need not be a biased 'black box'—careful training can produce interpretable systems that enhance rather than distort human judgment.

Collectively, the chapters in this dissertation show that smarter machines do not inevitably produce stupider humans. AI tools, when carefully designed and deployed, can improve motivation and information processing, making us better writers, more motivated learners, and wiser decision makers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Preface

War to the death should be instantly proclaimed against them. Every machine of every sort should be destroyed... Let there be no exceptions made, no quarter shown.

Samuel Butler, 1863

Society has long feared intelligent machines. Well before the Terminator and HAL-9000, the novelist Samuel Butler called for machine genocide. Perhaps because of all this cultural conditioning, I was wary of AI when I started my doctoral studies. In what follows, I'll describe how empirical evidence shifted me from skepticism to cautious optimism that AI tools, properly designed, can improve human capabilities.

My first real contact with AI came in 2012, during a semester abroad at the University of Leuven in Belgium. I took a course called Fundamentals of Artificial Intelligence, where we learned the "classical" toolkit—search algorithms, constraint satisfaction, and hand-crafted heuristics—designed to let computers solve well-defined problems. It was a far cry from the Hollywood image of AI. Seeing how these systems worked under the hood stripped away the mystique: the techniques were clever, but understandable.

A few years later, the field shifted almost overnight. Deep neural networks, long a theoretical curiosity that was impractical to implement, started outperforming older methods on speech recognition, image classification, and language processing. The gains were spectacular—but opaque. The very systems that were setting performance records were also becoming harder to interpret. At the same time, troubling examples emerged: medical triage algorithms that under-treated Black patients, risk-assessment tools that gave harsher scores to minorities, hiring systems that filtered out women. Accuracy had improved, but fairness and explainability had suffered.

This was the state of affairs when I started working on one of my first projects during the doctoral program, now Chapter 3 in this dissertation. I was analyzing the possibility of using predictive AI to evaluate personal qualities at scale in college admissions. If the tool, like others before it, traded efficiency for bias, the bargain was unacceptable. In fact, the best evidence we had so far pointed in this direction. AI extracted measures of application essay content and style had been shown to reveal applicants' social class.

My collaborators and I were surprised to see that with careful design, AI could accurately measure personal qualities in admissions, which predicted college graduation six years later. Moreover, this performance did not come at the expense of reduced equity, and modern techniques allowed us to "open the black box" and see where the model was looking at to make predictions about applicants' personal qualities. I learned that, contrary to what many critics argued, bias wasn't baked into AI by design; it was contingent on the recipe. Design choices—not algorithms in the abstract—decide fairness.

Just months after our paper was published, ChatGPT launched publicly. Anyone now had access to a system that could generate text about anything that read as if it was written by a human. I had tinkered with earlier AI models that generated text, but none of these were as open to the public nor as powerful as ChatGPT. Like many, I experienced what Ethan Mollick has called "the two sleepless nights". As I experimented with AI, I realized that now any applicant could draft an essay at the press of a button.

Soon enough, it dawned on me that the stakes were higher than just college admissions. If a readily accessible tool could produce an essay about any topic, would that make writing obsolete? And if writing is how we learn to think, would AI lead to people outsourcing all of their thinking? If so, AI could reduce motivation for effortful but meaningful thought. We often settle for the path of least resistance, even if it is bad for us in the long run. And reduced exposure would then lead to the atrophy of our mental capacities. So I began losing sleep over the question of whether AI would make us lazy and dumb.

I was not alone in worrying about this. New York City public schools banned ChatGPT within a month of its release, citing fears of "cognitive laziness." Two-thirds of Americans told pollsters that large language models will make society less intelligent. Was this just another moral panic? After all—from Socrates' critique of writing to Nicholas Carr's essay 'Is Google making us stupid'—many before me have worried about the technologies of their time. In hindsight, many of these concerns turned out to be far too pessimistic. The only way to know was to resort to experimentation.

In work not mature enough for this dissertation, we asked people how much of GDP should the US allocate for foreign aid and explain their recommendation. Then, we randomly assigned them to get more information, either through an AI chatbot assistant or through a traditional web search. After revising their original judgments, we evaluated them for critical thinking. The chatbot group outperformed the Google group, showing sharper reasoning despite not experiencing any extra effort and, in fact, spending about four times longer engaging with the material. Far from inducing laziness, the AI prompted deeper—and more sustained—thinking, directly contradicting my initial fear that it would make users passive or shallow.

Building on these findings about thinking quality, we next turned to measuring learning directly. We hypothesised that AI would be a coach or a crutch depending on how people used it. So we built two AI conditions to compare to a no-AI control. The AI-crutch condition simply gave away the answer, allowing users to copy and paste without thinking. The AI-coach condition forced the user to try first, and then compare their attempt with the AI's answer. Much to our surprise, we found that both AI conditions—even the AI-crutch—improved writing skill, tested without access to AI.

We focused on this counterintuitive finding, and collected the data now reported in Chapter 1. We found that despite people's predictions, even when users were allowed to copy and paste AI text, they learned more, even though they spent considerably less time practicing. A follow-up study gave us some clues as to why. Even if in this case AI reduced engagement, the access to higher quality information in the shape of precisely timed, personalized, high quality examples resulted in improved performance.

A year's worth of AI bots in the wild showed us that AI can also target motivation directly. About the time when we started researching AI and thinking, Angela got a call from Sal Khan, and we started working on AI assistants for motivational support in learning. After careful development with highschoolers and a year of deployment, we found that chatbot use was associated with students attempting more math problems, working for longer, and perhaps most surprisingly, attempting harder questions (Chapter 2).

Reflecting on these findings highlighted an apparent contradiction: Sometimes AI dramatically increased engagement (like in the foreign aid study), and sometimes it reduced it (like in the writing study). And yet in both cases, AI used resulted in improved performance. Engagement alone could not explain the pattern of results we observed. Looking back at our findings we realized the other lever is better information. Both matter. AI tools can supply higher-signal examples, explanations, or feedback (information). They can also raise or lower the user's willingness to expend cognitive effort (engagement). Whether people get "smarter or stupider" depends on the net effect of those two levers. See Figure 1.

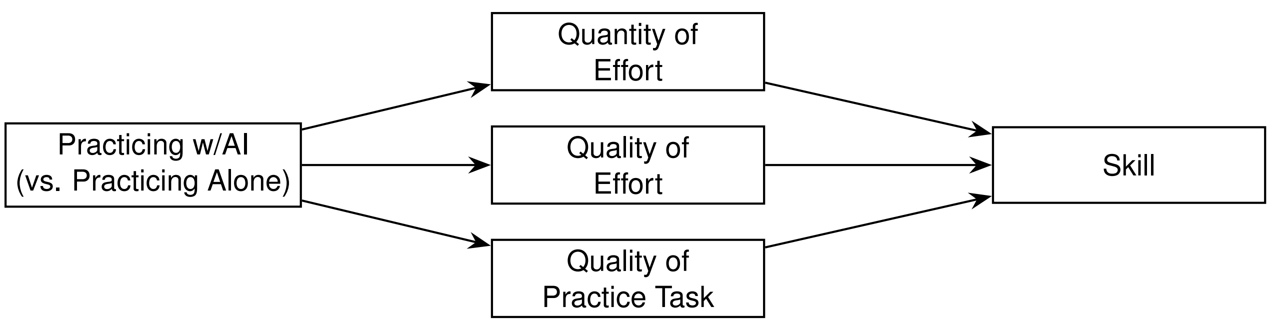

Figure 1.1: Using AI tools can improve human skill by improving task efficiency, quantity, and/or quality of engagement.

For me, the question is no longer whether AI will make us stupid or smart, but rather, when will it do so, and how can we carefully design AI (and the contexts around it) to ensure it makes us smarter. Taken together, the studies in this dissertation show that AI can make us smarter—better writers, more motivated learners, and more consistent decision makers—when designers balance the two levers of engagement and information wisely. Where information quality soars, lower engagement may be an acceptable trade-off; where engagement rises, information need not be as important.

This insight inspires future work. Can we design for both better information and higher engagement? How can AI help us achieve flow? How can AI boost performance while also deepening agency and meaning? Which design choices sharpen thinking, and which contexts tip the balance toward superficiality? Finally, can we raise objective engagement while dulling the aversive feeling of effort?

As these technologies continue to evolve, my initial skepticism has turned into cautious optimism. I hope that in the future I can continue to explore these questions, so that rather than a machine genocide, we get to harness them to be our best selves.

CHAPTER 2

Learning, not cheating: AI assistance can enhance rather than hinder skill development

2.1. Abstract

Does using AI make you stupid? Contrary to the predictions of lay forecasters, participants randomly assigned to practice writing with AI 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 tools would be less helpful, and less worth paying for, than personalized feedback from experienced human editors, we found the converse: practicing with an AI tool 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 tools can improve, rather than undermine, learning. [1]

2.2. Introduction

Generative AI (henceforth AI) tools are increasingly powerful and prevalent (2), 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) (3–5).

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 (6–8). For instance, in a 2024 poll, 62% of surveyed adults predicted that Generative AI will "lead to humans becoming less intelligent" (9). In January 2023, New York City public schools banned ChatGPT, citing "concerns about negative impacts on student learning (10)." 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

[1] This work is currently under review at Science Advances

that students are participating in and will work in a world where understanding Generative AI is crucial (11)." 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%) (12).

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 (13)), make reasoning and arithmetic errors (14), 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 (15).

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 (16). And drivers who use GPS tend to have worse hippocampal-dependent spatial memory, both cross-sectionally and longitudinally (17, 18). 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.

And yet there is a plausible, albeit less obvious, alternate hypothesis: Using AI tools to help do our work 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 (19, 20). Compared to textbooks, conventional

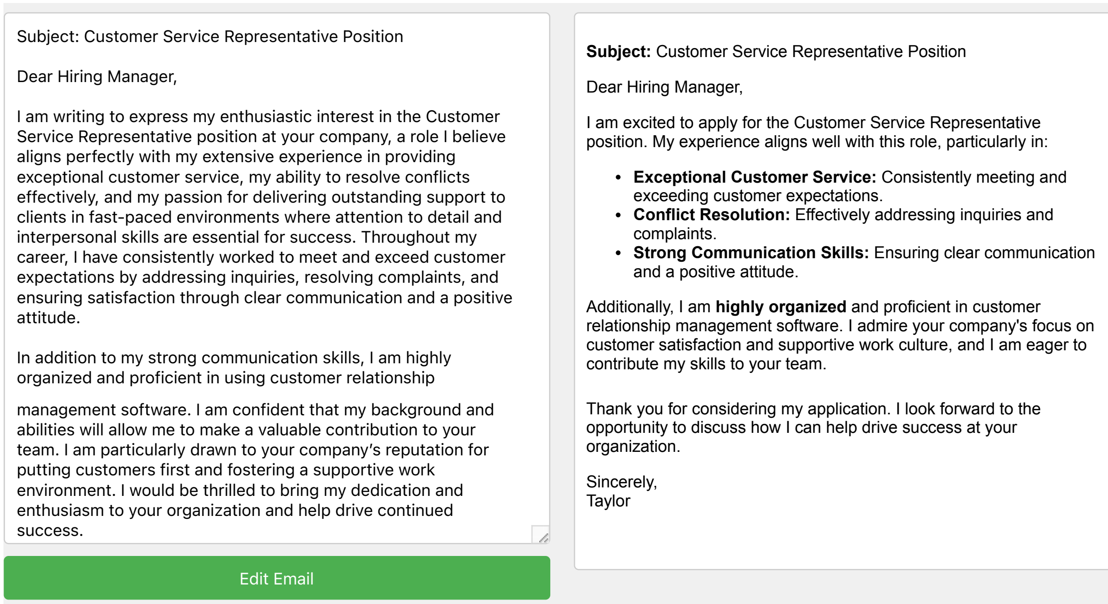

Figure 2.1: The AI writing tool we created for this investigation takes inputted text (left panel) and generates a version that incorporates recommended writing principles (right panel). Users could edit the text in the left-hand box, which was prepopulated with the text they were instructed to rewrite. They could then copy and paste the revised output from the right-hand box.

computer tutoring programs (e.g., (21)), and even human teachers, AI tools may be able to provide higher-quality, just-in-time examples exquisitely tailored to a learner's specific needs. Indeed, it may be 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, AI tools may improve skill development if the upside of exposure to excellent examples tailored to the learners' needs outweighs the downside of diminished engagement.

There is little research on how AI tools influence skill development. 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 (22, 23), while others have shown null or even negative effects (23–25). 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,

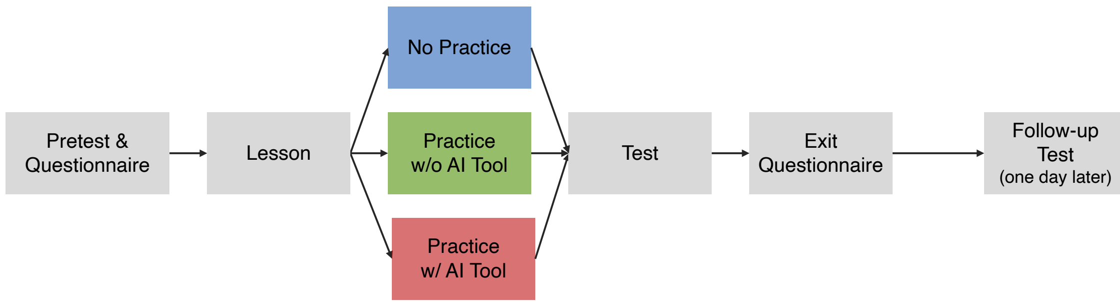

Figure 2.2: Experimental design for Study 2. First, all participants completed a baseline questionnaire, a pretest (rewriting a poorly-written cover letter), and a lesson introducing five evidence-based principles of effective writing (1). Next, participants were randomly assigned to one of three conditions: 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.

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 (26). 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 letter). They were then introduced to evidence-based strategies for professional writing, with descriptions and examples for each one (1, 27). 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 2.1), 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 2.2. 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. 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).

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. If participants wrongly believe that AI is detrimental to learning, they might underinvest in AI as a learning tool. Study 3 suggests this is the case, as forecasters predicted that AI would be less effective at providing feedback than experienced human editors, and accordingly, were willing to pay less for it. Study 4 explores whether participants learn more from AI than they would by accessing other ecologically valid human sources of feedback: a low-cost Google Search, and higher-cost personalized feedback from experienced human editors. Contrary to forecasters' predictions, participants who had practiced with AI improved their writing more than those who had access to Google Search, and even than those who received personalized feedback from professional editors. This finding is particularly striking given that our editors averaged 25 years of experience and included professors of journalism, news editors, and even one Pulitzer Prize winner. Finally, in Study 5, 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), had practiced without AI (Studies 2 and 5); and as likely as those written by participants who received feedback from experienced editors or used google (Study 4).

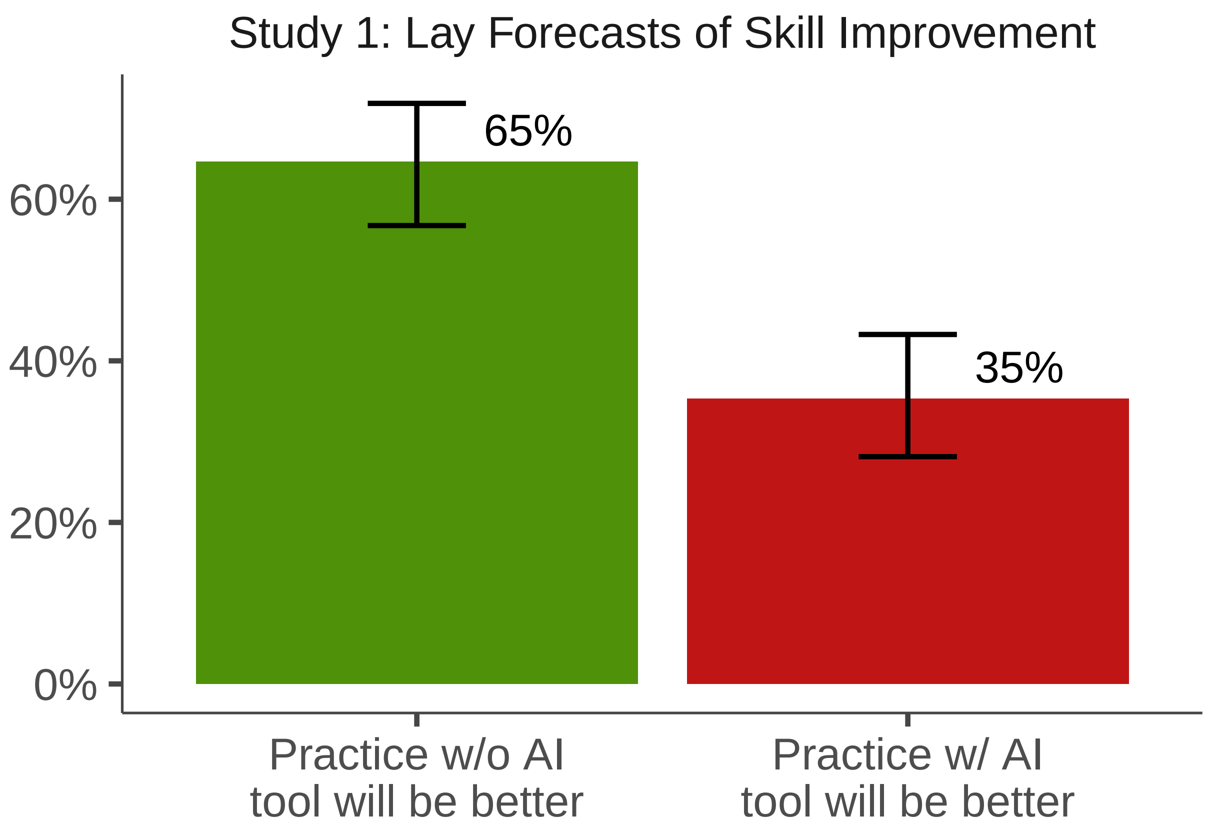

Figure 2.3: Forecasters in Study 1 predicted that practicing without an AI tool would improve writing skill more than practicing with an AI tool. Error bars represent proportions ± 1 SE.

2.3. Study 1

2.3.1. 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 2.3. Participants made this prediction regardless of self-reported experience with AI ($OR = 0.83$, $p = .239$) or any other measured demographic characteristic (ps $> .05$).

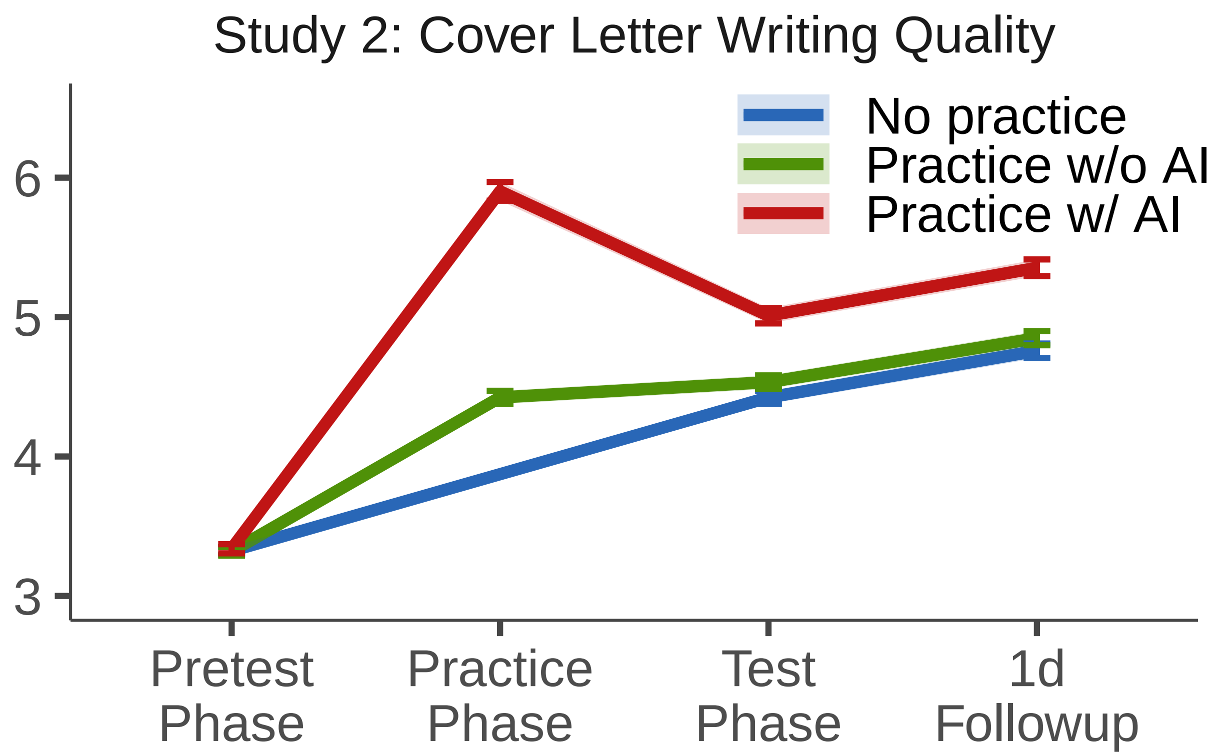

Figure 2.4: In both the main test and the follow-up of Study 2, participants who had practiced with the AI tool outperformed those who practiced without it and those who did not practice at all. Error bars represent means ± 1 SE. Means shown are for the subsample of participants (n = 1,294) who completed the one-day follow-up test. See Figure A.3 for the equivalent figure in the full sample, excluding the one-day follow-up phase (N = 2,238).

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.")

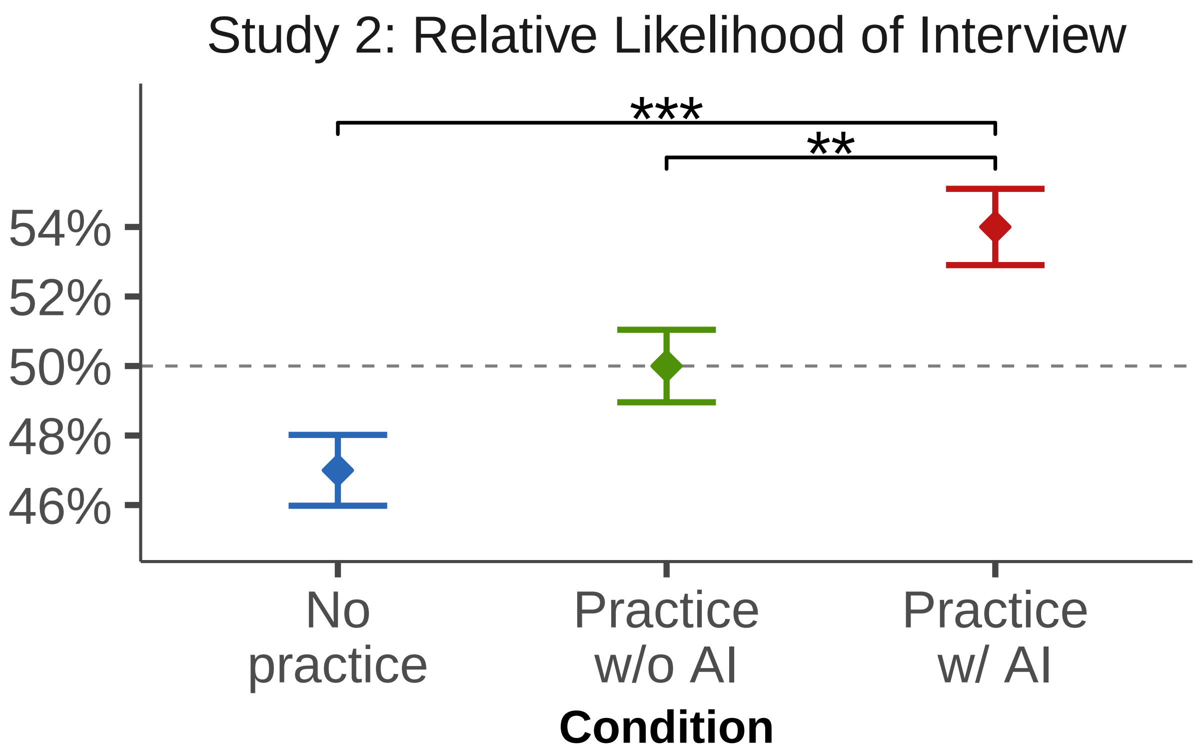

Figure 2.5: During the test phase of Study 2, cover letters written by participants who had practiced with AI were more likely to lead to interview offers than those from other conditions. Points depict the average proportion of times each cover letter was preferred in pairwise comparisons with cover letters from the other two conditions. Error bars represent proportions $\pm$ 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.

2.4. 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) (1). 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 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.

2.4.1. AI practice improved writing skill

Consistent with other studies demonstrating the productivity benefits of AI tools (3, 4), 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 preregistered 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 A.3. 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 2.5.

2.4.2. AI practice was less effortful

The learning benefits of using an AI writing tool were evident despite reduced quantity and quality of 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, quantity of effort), logged roughly a quarter as many total keystrokes (26% d = -.44, p <.001) and 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. Copying, pasting, and submitting the AI tool's output could be accomplished almost instantly. Yet, the majority of participants 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 A.4 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.

2.4.3. 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 A.2.4 in the Supplementary Information for details.

2.4.4. 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. 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). The magnitude of the follow-up effect size was not significantly different from the immediate test phase effect size (condition × time interactions were non-significant (ps > .69). See Figure 2.4 and Section A.2.5 of Supplementary Information for details.

2.4.5. 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 A.11 in the Supplementary Information for details.

2.5. Study 3

Given unwarranted pessimism about the pedagogical power of AI, in Study 3 we asked 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. To rule out whether these differences in willingness to pay were fully explainable by social preferences (e.g., a preference for paying individuals rather than AI corporations), we also asked them to rate the degree of confidence they had that AI feedback would help them improve their writing.

2.5.1. Lay writers were willing to pay more for feedback from expert human editors than for AI feedback

Out of a maximum 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, $d = 0.39$).

2.5.2. Lay writers predicted that editorial feedback would be more beneficial

Participants had more confidence that experienced human editors ($M = 6.68$ on a 0-10 likert scale) would help them improve their writing, compared to feedback from an AI tool ($M = 5.64$; $t(148) = -4.77$, $p <$.001, $d = 0.39$). This suggests that underinvestment in AI as a learning resource was not uniquely explained by social preferences.

2.5.3. Predicted effectiveness mediated underinvestment in AI feedback

Perceived effectiveness 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

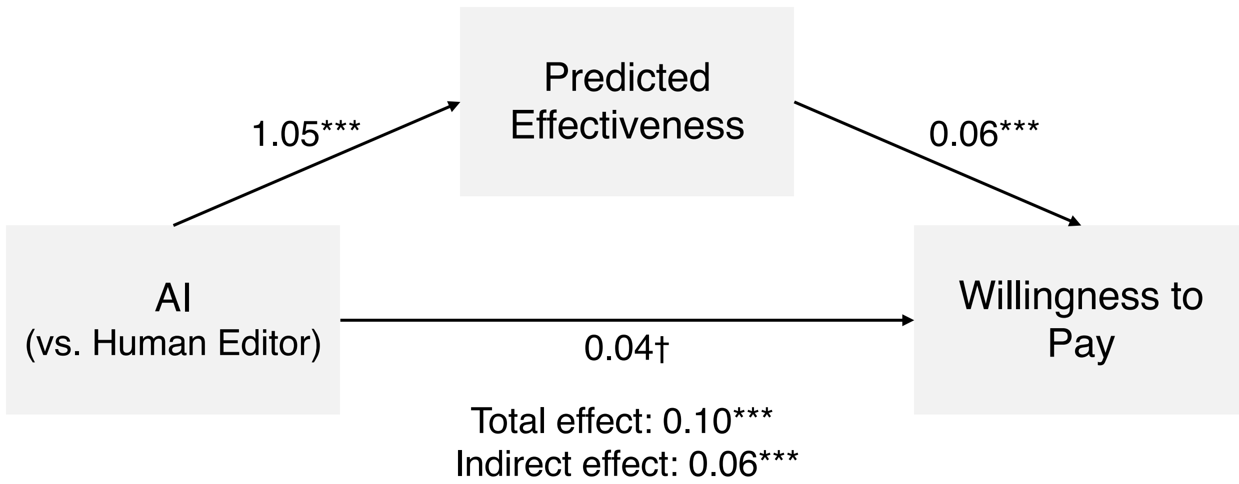

Figure 2.6: Predicted effectiveness mediates underinvestment in AI feedback

were accounted for, the remaining preference for human feedback was small and only marginally significant (c' = 0.04, SE = 0.02, p = .054). See Figure 2.6 and Section A.3.1 of Supplementary Information for details.

2.6. Study 4

In Study 4, we compared practicing with AI to practicing with feedback from human experts, and, in addition, to practicing with access to examples and tips from Google search. Specifically, N = 2,997 participants 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) search online for cover letter examples, or (3) receive feedback from professional human editors. Letters written by participants assigned to the third condition were given to professional editors, who provided written constructive feedback and a rewritten version of the cover letter. The next day, participants were asked to improve the cover letter they had submitted further with the resources afforded by their assigned condition. At the end of the session, all participants completed a test of writing skill (rewriting yet another cover letter without access to the AI writing tool) and an exit questionnaire.

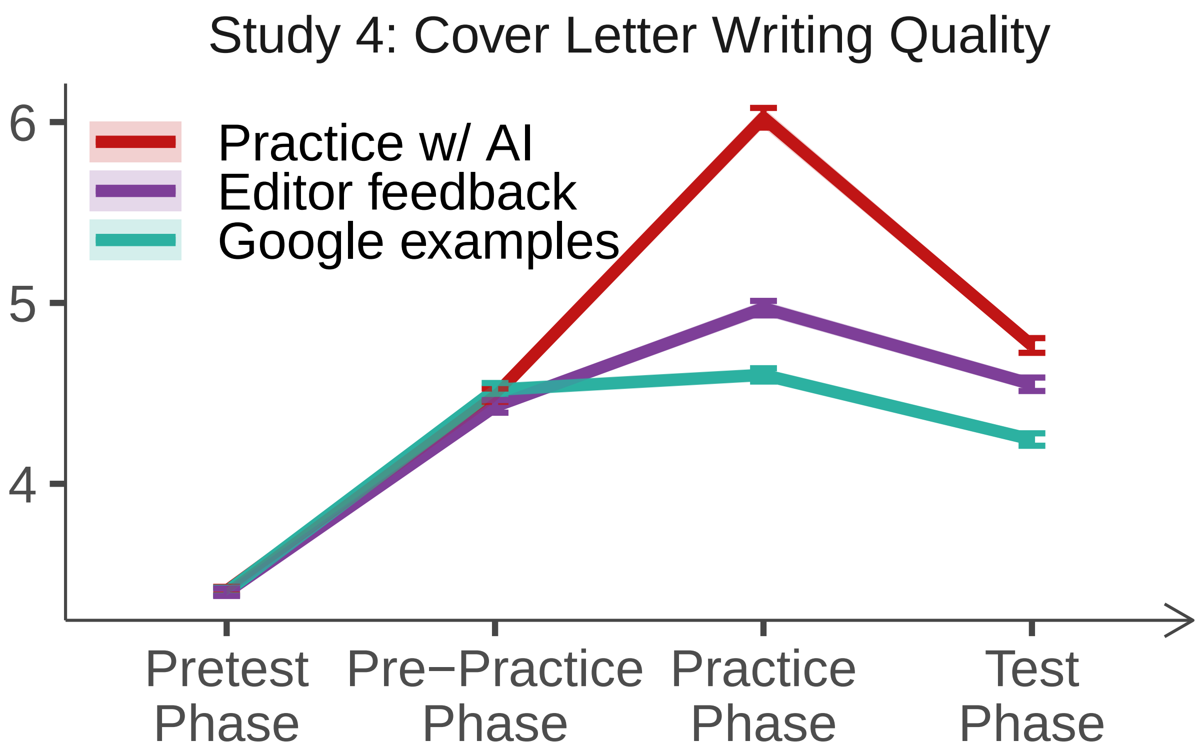

Figure 2.7: Study 4 participants who had practiced with the AI tool outperformed those who practiced with access to Google examples and feedback from professional editors. Error bars represent means ± 1 SE.

2.6.1. AI practice improved writing skill more than getting feedback from professional editors and looking for examples online

During the practice phase, participants working with AI outperformed participants who were able to Google cover letter examples and tips ($d = 1.03$, $p <.001$) or who received personalized feedback from professional editors ($d = 0.76$ $p <.001$).

During the test phase, when all 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 in the Google condition ($d = 0.46$, $p <.001$, $p <.001$), and event than those who had received personalized feedback from professional editors ($d = 0.20$, $p <.001$). See Figure 2.7.

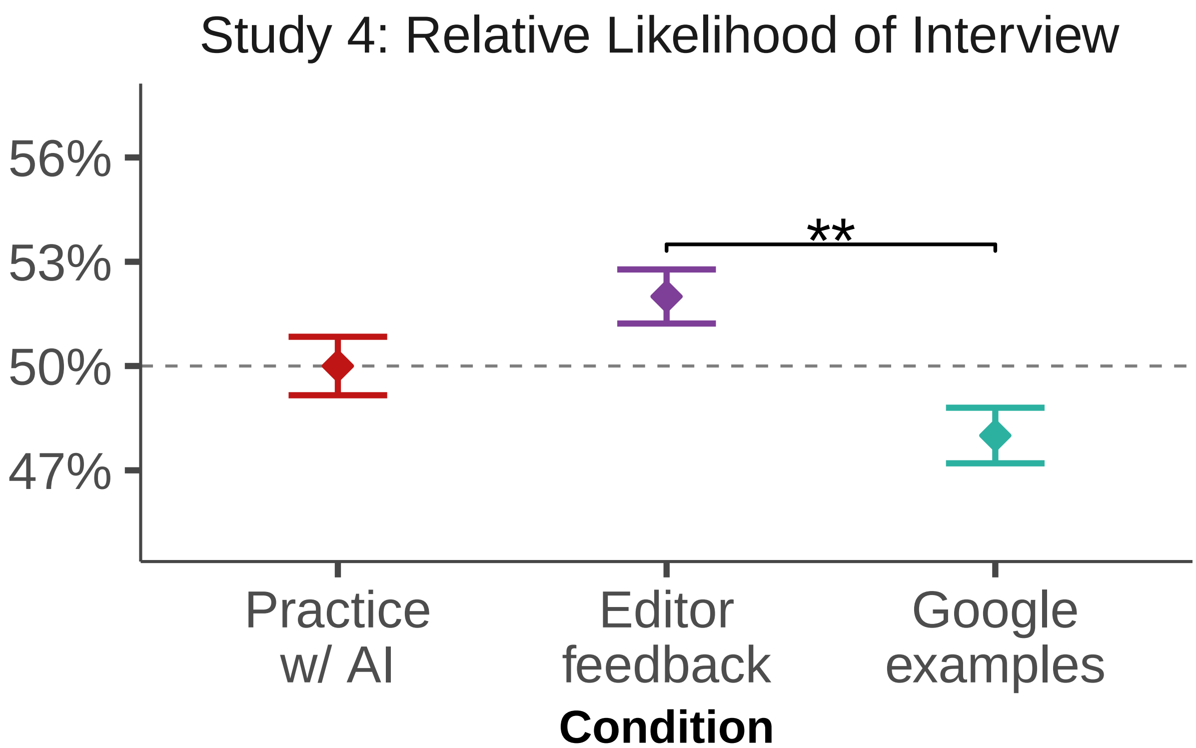

Figure 2.8: During the test phase of Study 4, cover letters written by participants who had practiced with AI were about equally likely to lead to interview offers when compared to those who had received feedback from professional editors. Cover letters from both of these conditions outperformed those written by participants who were assigned to browse through Google examples. 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.

However, cover letters written by participants who had practiced with AI were not significantly more likely to secure a hypothetical job interview than cover letters by participants (.50 vs. .52, p = .102), or participants who had who had access to Google Search (.50 vs. .48, p = .141). See Figure 2.8.

A possible concern is that reviewing many letters consecutively caused editors to get tired and give worse feedback over time. This could suggest our experiment compared AI against a disadvantaged "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 A.4.2 of Supplementary Information for details.

2.6.2. AI practice wasn't any more effortful than getting feedback from professional editors and was less effortful than looking for examples online

During the practice phase, participants who used AI spent 1.18 fewer minutes than Google users ($d = 0.32$, $p < .001$) and about the same time as those receiving editor feedback (difference = 0.43 minutes, $d = 0.09$, ns). Nevertheless, participants across all three groups reported similar levels of subjective effort (ps > .05). Objective measures told a similar story: AI users logged 76% fewer keystrokes than Google users ($d = 0.32$, $p < .001$), but not fewer than those working with editor feedback (68% fewer, $d = 0.08$, ns). They also typed at a slower pace than both Google users ($d = 0.34$, $p < .001$) and those who received editor feedback ($d = 0.41$, $p < .001$).

Once participants were tested without external support, differences in effort narrowed. Participants who had practiced with AI spent slightly more time writing than Google users (median = 5.72 vs. 5.16 minutes, $d = 0.18$, $p < .001$) and about the same as those given editor feedback (median = 5.72 vs. 5.52 minutes, $d = 0.07$, ns). Self-reported effort did not differ between AI users and participants who received editor feedback ($d = .02$, $p > .05$), though Google users rated the task as easier than AI users (d = -0.22, $p < .001$) and participants who received editor feedback (d = -0.24, $p < .001$). Participants across the three conditions accrued approximately the same amount of keystrokes (ps > .05).

2.6.3. AI practice did not create more of an illusion of mastery than getting feedback from professional editors or looking for examples online

Participants who had practiced with AI did not report learning more than those who received feedback from expert human editors ($d = 0.07$, ns), nor did they consider their writing skills to be better (d = -0.02, ns). Both of these groups, however, correctly recognized that they had learned more than those assigned to search for online cover letter examples and tips (ds = .28 and

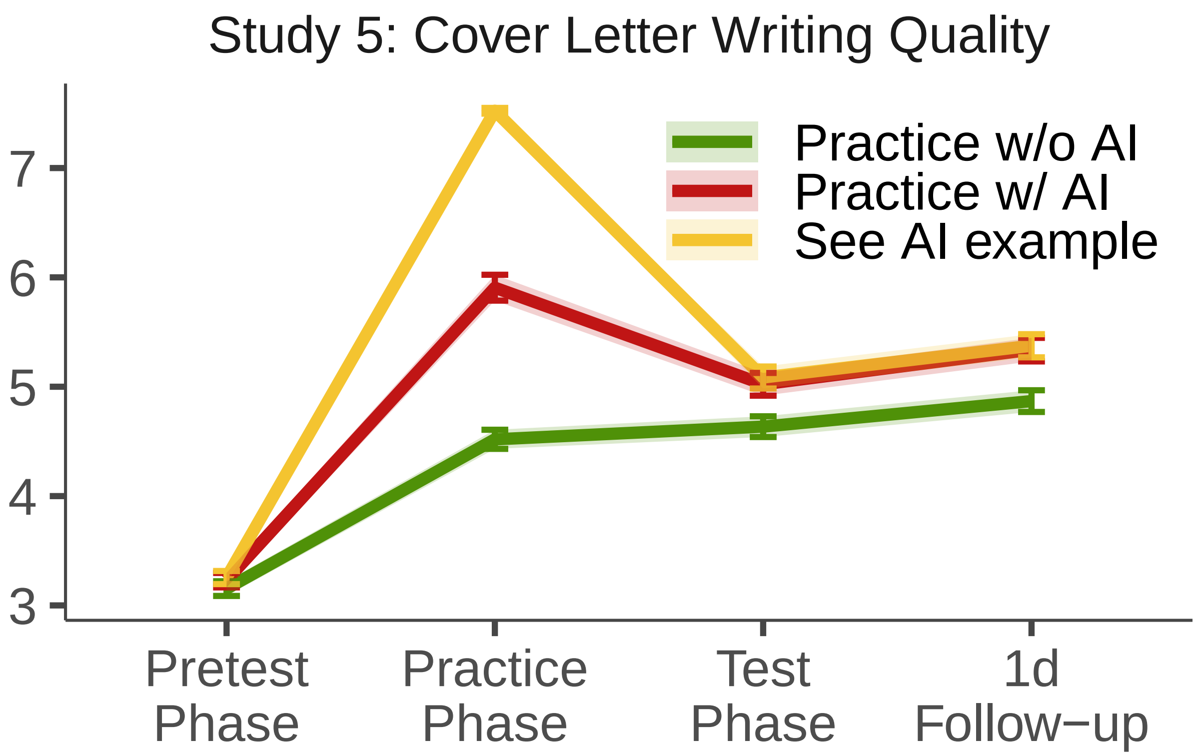

Figure 2.9: In both the main test and the follow-up of Study 5, participants who simply saw an AI-generated example improved just as much as those who practiced with AI and more than those who practiced without AI. Error bars represent means ± 1 SE. Means shown are for the subsample of participants (n = 608) who completed the one-day follow-up test. See Figure A.10 for the equivalent figure in the full sample, excluding the one-day follow-up phase (N = 2,003).

.20, respectively), and evaluated their writing skills accordingly (ds = .11 and .12, respectively). AI users and those who received expert editor feedback did not request feedback on their test submissions at different rates (71% and 70%, respectively). Both groups were more likely to request feedback than those who had practiced with access to Google search (64%, p = .003 and .007, respectively). See Section A.4.6 in the Supplementary Information for details.

2.6.4. Results were not moderated by individual differences

As in Study 2, the above findings were consistent across all measured individual differences. See Table A.20 in the Supplementary Information for details.

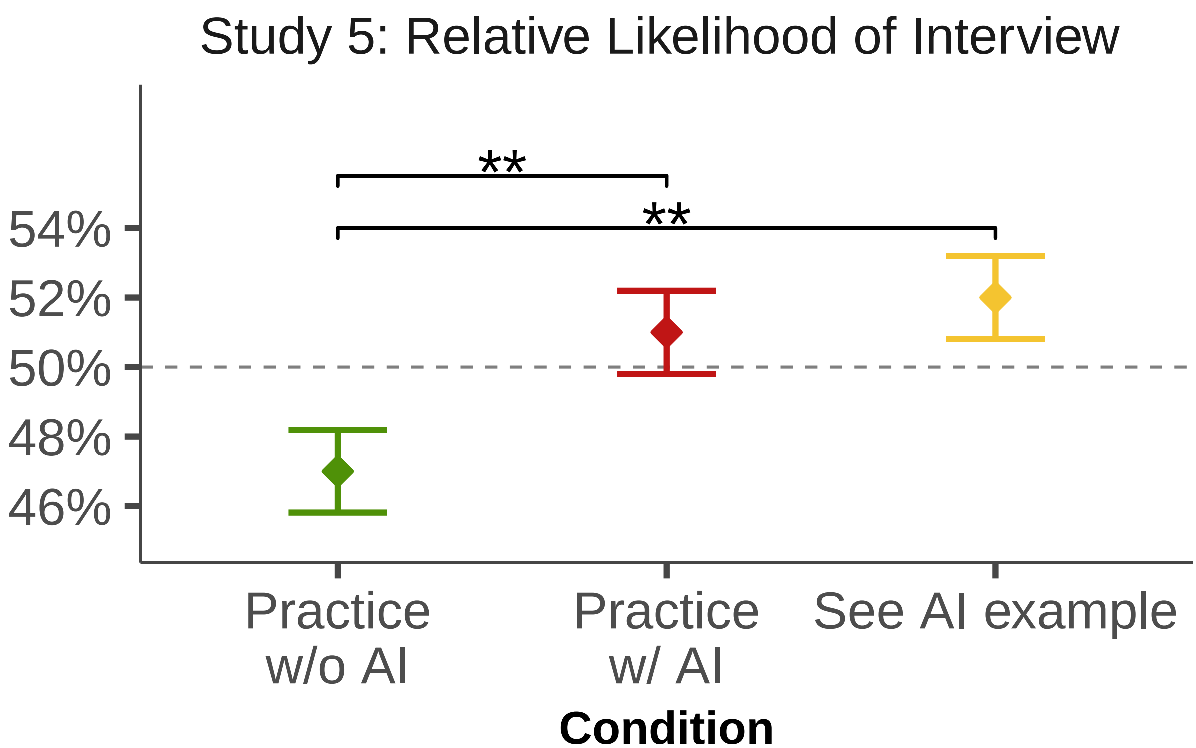

Figure 2.10: During the test phase of Study 5, cover letters written by participants who had seen an AI-generated example were about equally likely to lead to interview offers when compared to those assigned to practice with AI. Cover letters from both AI conditions outperformed those written by participants 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.

2.7. Study 5

To gain insight into what might be driving the benefit of practicing with AI, 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, 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.

2.7.1. 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 A.10 and 2.10.

2.7.2. 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 2.83 times fewer keystrokes and 2.10 fewer keystrokes per minute (d = 0.15, p = .007) 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 56 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 60% 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 $> .05$).

2.7.3. 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 A.5.4 in the Supplementary Information for details.

2.7.4. The benefits of seeing an AI example were just as large a day later

When we recontacted a subsample of ($n = 800$) participants 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 2.9 and Section A.5.5 of Supplementary Information for details.

2.7.5. Results were not moderated by individual differences

As in the previous studies, the above findings were consistent across all measured individual differences. See Table A.32 in the Supplementary Information for details.

2.8. 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, later 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 can 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 as much as those given the option to practice editing the cover letter (Study 5).

While not immediately apparent to forecasters, teachers (12), and the general population (9), the benefits of viewing 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 (28–30). 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 (31).

Study 1 forecasters who believed AI would harm learning, cited reduced effort as the mechanism, and prior research supports this concern (**? ?**)—all things being equal, the more effort a learner invests in a given practice task, the more they learn. So why, in Studies 2 and 5, did we observe the opposite? That is, why did practicing with AI improve writing skill more than practicing alone while reducing both the quantity and quality of effort? Likewise, why in Study 4, did writers who googled how to compose a cover letter work longer and harder than writers who practiced with AI, yet learn less? We posit a third mechanism that offsets decrements in human effort: the quality of the practice task itself. See Figure 2.11. Perhaps the excellent personalized examples provided by AI are analogous to demonstrations of an Olympic coach. For every moment of focused effort invested, the learning dividends are greater.

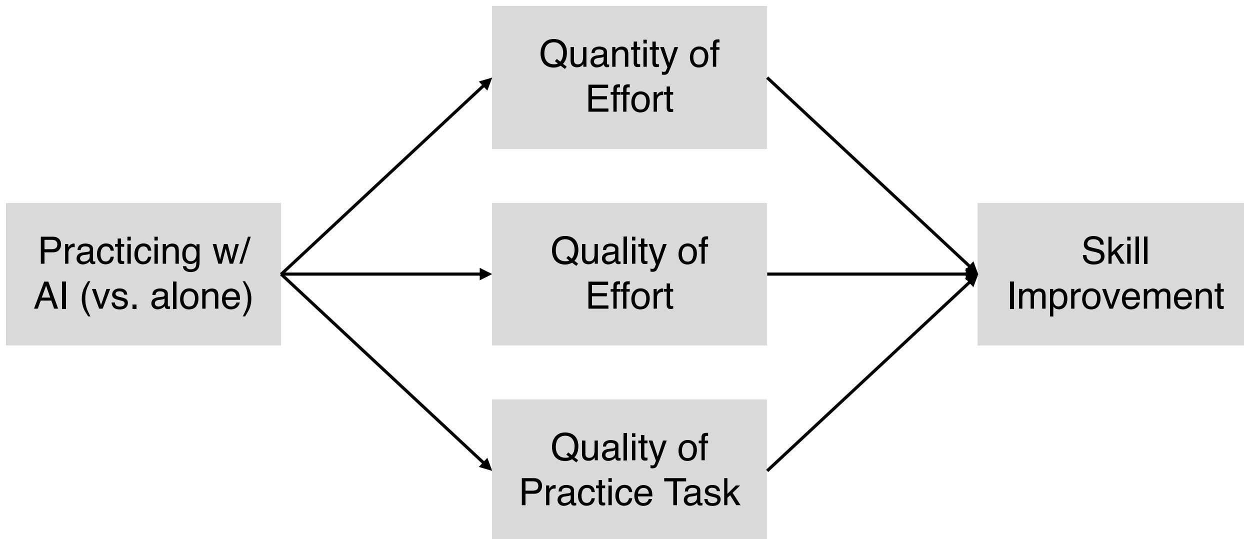

Figure 2.11: **Conceptual mechanism.** Access to AI reduces practice effort but raises the *quality of the practice task* by providing high-quality, personalized examples. The quality boost offsets lower effort, yielding greater post-practice writing skill.

2.8.1. Future Directions

Three promising directions for future research are worth highlighting. First, it is not yet clear whether learning from AI examples extends beyond writing–the most common workplace use case for AI tools (26). Writing may be particularly suited to learning by example because 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, 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. Studies of AI tutors in these domains have, perhaps not surprisingly, found null or negative effects (24, 25).

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 (22). Similarly, correlational evidence suggests that asking AI for explanations as opposed to answers is associated with more learning in mathematics (24) and computer programming (23). 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 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? 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 (32). Future research, ideally in field settings, is needed to establish the long-term benefits and costs of relying on AI tools.

2.8.2. 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. It 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 basketball coach John Wooden declared that the four laws of learning are explanation, demonstration, imitation, and repetition (33). 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.

2.9. Methods

2.9.1. 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.

2.9.2. Pre-registration

All studies were preregistered. Accordingly, the analyses presented in the main text were also pre-registered, with exceptions noted below.

The preregistration for Study 1 is at https://aspredicted.org/x9mm-7qwp.pdf.

The preregistration for Study 2 is at https://aspredicted.org/xyyn-gmzc.pdf. The preregistration for the pairwise comparisons analysis is available at https://aspredicted.org/4sw4-mpny.pdf. After preregistration, we considered a keystrokes-per-minute metric of effort intensity (rather than duration). Thus, the keystrokes-per-minute analysis was not preregistered.

The preregistration for Study 3 is at https://aspredicted.org/xyyn-gmzc.pdf. Analyses on confidence on AI vs. human feedback, and the mediation analysis were not preregistered.

The preregistration for Study 4 is at https://aspredicted.org/xyyn-gmzc.pdf. The preregistration for the pairwise comparisons analysis is available at https://aspredicted.org/4sw4-mpny.pdf. The keystrokes-per-minute metric of effort was not pre-registered.

The preregistration for Study 5 is at https://aspredicted.org/3mty-fcfy.pdf. The one-day follow-up for Study 5 was collected in three batches: a pilot, followed by a second and third batch. We preregistered batch 2 https://aspredicted.org/5jcx-bhg9.pdf, but report pooled results in the main text. Details on the separate batches are available in the Supplementary Information Section A.5.5. The preregistration for the pairwise comparisons analysis is available at https://aspredicted.org/4sw4-mpny.pdf. The keystrokes-per-minute metric of effort was not pre-registered.

2.9.3. 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 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 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).

2.9.4. 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 A.1.3. Immediately after seeing 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 A.5 and A.25 in Supplementary Information to see results excluding them, which are consistent with our main interpretation.

2.9.5. 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 (34, 35). See our prompts in Table A.1. 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 A.5 and A.25.

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 strongly with Claude ratings ($r = .71$, $p < .001$), and that the effects are not attenuated by using different models (See Tables A.5 and A.25), suggesting that our effects are not explainable by same-model bias.

2.9.6. 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 (36).

CHAPTER 3

Help Me Focus: AI-delivered motivational interventions enhance student persistence in online learning environments

3.1. Abstract

Can generative AI provide motivational support at scale? While psychological interventions have proven to be effective in controlled settings, their scalability in natural contexts remains understudied. This year-long field investigation shows that AI-delivered interventions benefit learners in an ecologically valid setting. Relative to the week before completing the intervention, students increased their time on task by 10% after Situation Modification and 11% after Emotional Reframing. Students also worked on more challenging items following exposure to the interventions. Effects on accuracy were less robust: average accuracy did not improve, but under Emotional Reframing higher-skill students showed modest gains. Exploratory text analyses reveal that AI-facilitated interactions are associated with reduced negative affect. These findings suggest that scalable emotion regulation interventions can significantly enhance engagement in digital learning environments, highlighting the potential for AI-supported emotional scaffolding to address barriers to educational persistence and achievement.

3.2. Introduction

Motivation drives behavior across domains including consumer decisions, workplace performance, health, and education. People often know what they should do—stick to fitness plans, manage finances, complete online courses—yet struggle to persist when faced with frustration, effort, or competing demands. The core challenge in behavior change is not delivering information but sustaining motivation over time.

Traditionally, human mentors, managers, and coaches provide motivational support by recognizing when individuals struggle and intervening in real-time (37). However, as interactions shift to digital environments, this personalized, adaptive support is often missing. The consequences are well-documented: employees disengage from workplace training, consumers abandon financial plans, and learners drop out of online courses (38, 39). Without timely intervention, engagement and persistence decline, limiting the effectiveness of digital interventions.

3.2.1. The Process Model of Behavior

Psychological research offers a promising framework for addressing these motivational failures. The process model of emotion regulation (40) describes how individuals respond to challenges that trigger frustration, anxiety, or disengagement from tasks. The model identifies two key strategies that proactively shape emotional responses.

First, situation modification involves structuring the environment to reduce the need for willpower. In workplace settings, this might involve removing distractions to enhance productivity. In consumer finance, it could mean automating savings to prevent impulsive spending. Second, emotional reframing involves interpreting negative emotions such as frustration as signals of progress rather than failure. For instance, framing difficulty as a necessary step in skill development can increase persistence in learning, just as reframing discomfort in exercise can encourage continued effort.

Behavioral interventions based on these strategies have successfully improved consumer persistence, workplace performance, and learning outcomes (41, 42). However, scaling these interventions remains a challenge, particularly in digital settings where human coaches are unavailable.

3.2.2. AI as a Scalable Motivational Coach

The proliferation of artificial intelligence-powered conversational agents presents a promising avenue for addressing the scalability challenges inherent in traditional motivational coaching. These systems offer the potential to deliver real-time, personalized motivational support to large populations at substantially reduced costs compared to human mentors. The theoretical foundation for AI-mediated motivational support rests on several key advantages that distinguish these systems from conventional interventions.

Contemporary research suggests that individuals often perceive AI-generated communications as demonstrating comparable or superior empathy relative to human-authored content, particularly in digital environments where concerns about social evaluation are minimized (43). This phenomenon may stem from users' reduced anxiety about being judged by an artificial agent, creating conditions more conducive to authentic self-disclosure and receptivity to motivational messages. The absence of perceived social judgment in AI interactions may paradoxically enhance the perceived authenticity of the motivational support provided.

Unlike static behavioral interventions that rely on predetermined content delivery schedules, contemporary generative AI systems possess the capability to dynamically adapt motivational strategies based on real-time user interactions and contextual factors (44). This adaptive capacity enables the customization of motivational messaging to align with individual learning styles, emotional states, and progress trajectories, potentially enhancing intervention effectiveness through increased relevance and timeliness. The ability to process and respond to nuanced user inputs represents a significant advancement over traditional one-size-fits-all approaches to motivational support.

Furthermore, AI-driven motivational systems can provide instantaneous, on-demand support to virtually unlimited numbers of users simultaneously, representing a fundamental shift from the resource-intensive, one-to-one model of human coaching (45). This scalability characteristic positions AI as a viable solution for addressing motivational support needs across diverse populations and educational contexts where human resources are limited or prohibitively expensive.

Despite these theoretical advantages, several fundamental limitations may constrain the effectiveness of AI-based motivational interventions, raising important questions about the sufficiency of artificial agents in facilitating sustained behavioral change. Social learning theory posits that meaningful motivation emerges through authentic interpersonal relationships characterized by mutual understanding, trust, and emotional resonance (46). Research in developmental psychology provides compelling evidence for this position: studies of early childhood language acquisition demonstrate that children cannot acquire language skills through passive exposure to digital media alone, requiring instead the social scaffolding provided by human interaction (47). If motivational processes operate through similar mechanisms—as suggested by theories of social cognitive development—then AI-based interventions may prove fundamentally insufficient for sustaining long-term behavioral change.

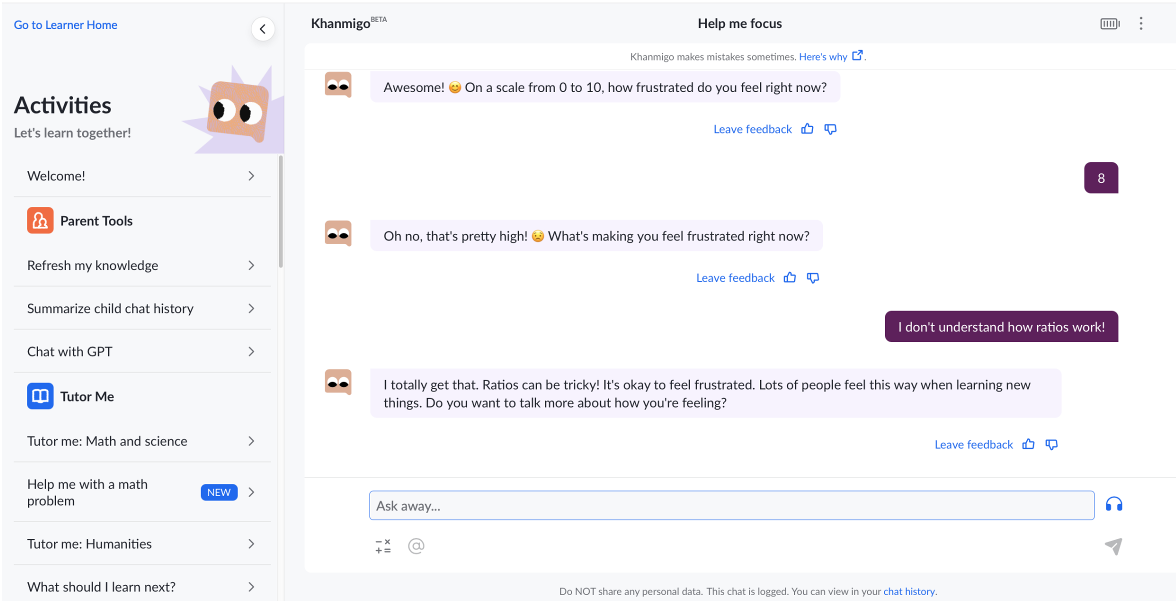

Figure 3.1: The interface that the users saw when they decided to engage with the bot.

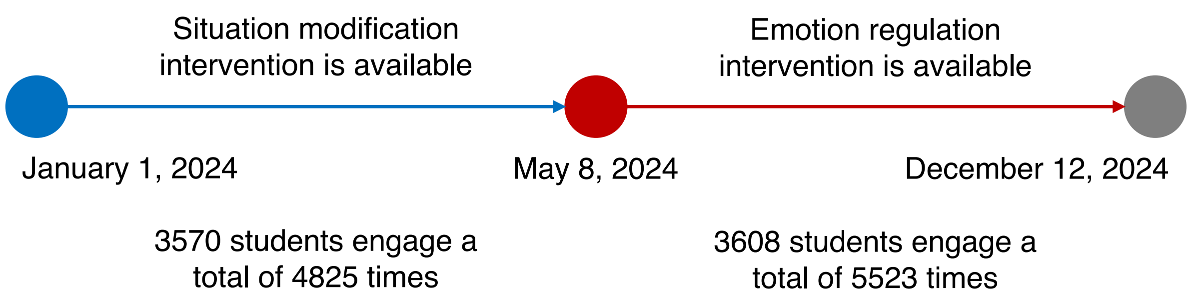

Figure 3.2: Data collection times and sample sizes.

Additionally, emerging research has identified significant concerns regarding the reliability and appropriateness of AI-generated advice, particularly in domains requiring sensitivity to individual psychological states and contexts (48). Poorly calibrated or contextually inappropriate motivational interventions may not merely fail to support users but could actively undermine motivation through frustration, misalignment with user needs, or provision of counterproductive guidance. The potential for such adverse outcomes underscores the need for careful evaluation of AI-mediated motivational systems before widespread implementation.

3.2.3. Current Investigation

This study presents the first large-scale, year-long observational analysis of AI-powered motivational coaching in a digital performance setting. Using behavioral data from Khan Academy, we evaluate two AI-driven motivational strategies:

First, the situation modification intervention encouraged users to structure their environment and habits in ways that reduce reliance on willpower. The interaction began by asking the participant to rate their distraction on a scale of 0 - 10. Then, it explained to students that modifying their environments can reduce distractions. It proceeded by helping them identify potential distractors in their immediate environment and assisting them in modifying their environment accordingly.

Second, the emotional reframing intervention helped users reinterpret frustration as an expected and productive part of progress. It started by asking students to rate their frustration on a scale of 0 - 10. Then, it validated these negative emotions and guided students in reinterpreting the emotion in a more positive way.

Since AI usage was self-selected, this study does not constitute a traditional experiment. Instead, we employ a fixed-effects methodology to account for timing and user heterogeneity, controlling for time-varying confounds through Bayesian item response theory (IRT) estimates of skill and motivation. This approach allows us to isolate the impact of AI-driven motivational interventions from broader shifts in user performance over time.

By evaluating AI-driven motivation at scale, this study makes three key contributions:

AI in behavioral interventions: Can AI effectively sustain engagement and persistence across digital performance environments? The role of human interaction in motivation: Does AI-based motivation substitute for or complement human-delivered support? Scalability of motivation science: How can emotion regulation strategies be embedded into AI systems to drive persistence in learning, work, and decision-making?

Beyond education, these findings have broad implications for AI-driven behavioral interventions in domains such as employee training, consumer engagement, and digital health. As AI systems increasingly mediate human decision-making, understanding when and how AI can sustain motivation is a critical question for marketing, management, and behavioral science.

3.3. Results

3.3.1. When do users decide to engage with the prompts

Because students could choose when to engage with the interventions, our first step was to examine the timing of that choice. If students tended to use a prompt when they were already disengaged or performing poorly, any apparent post-intervention improvement could simply reflect regression to the mean. Our data suggest this was not the case.

We compared user activity in the week before prompt use to a one-month baseline and found no evidence of systematic dips in engagement or performance preceding the intervention (Figure 3.3). In fact, students were more likely to use the situation modification prompt during periods of above-average activity, and to use the emotional reframing prompt when they had recently been attempting more difficult items.

3.3.2. Chatbot use is related to increases in motivation but not performance

As preregistered, students spent significantly more time working during the following week. Specifically, they worked 10% longer after Situation Modification ($\beta = 0.426,\ p < 0.001$) and 11% longer after Emotional Reframing ($\beta = 0.558,\ p < 0.001$).

This increased engagement did not translate into higher performance, contrary to our pre-registered hypothesis. In the week following either prompt, students did not show a significant improvement in accuracy (Situation Modification: $\beta = 0.425,\ p > 0.05$; Emotional Reframing: $\beta = 0.586,\ p > 0.05$).

Exploratory pre-registered analyses suggest that after engaging with the chatbot, users attempted more difficult items. This suggests that chatbot users may challenge themselves more, even if this does not immediately translate to higher accuracy. See Table 3.1 and Figure 3.4.

Our exploratory analyses provide additional context for these findings. The positive post-intervention

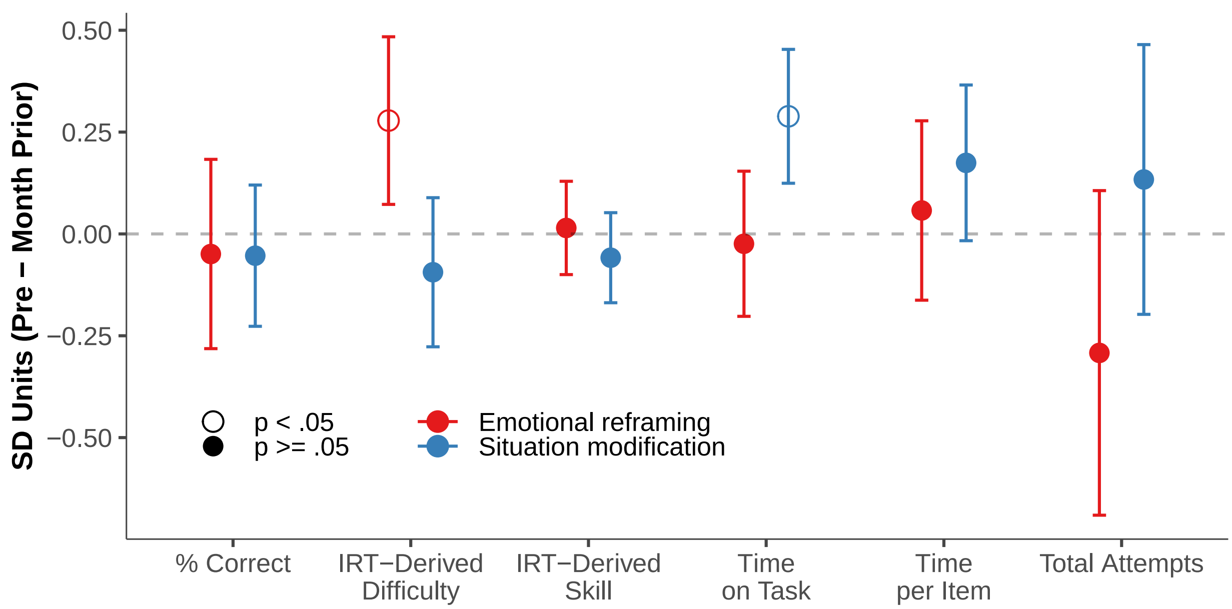

Figure 3.3: Users chose to engage with the prompts at times that were, for the most part, comparable to baseline. However, they were more likely to click on (1) the Emotional Reframing bot when they had been working on harder problems, and (2) the Situation Modification prompt when they had accumulated more time on task than usual. Baseline = One month before the pre-intervention period.

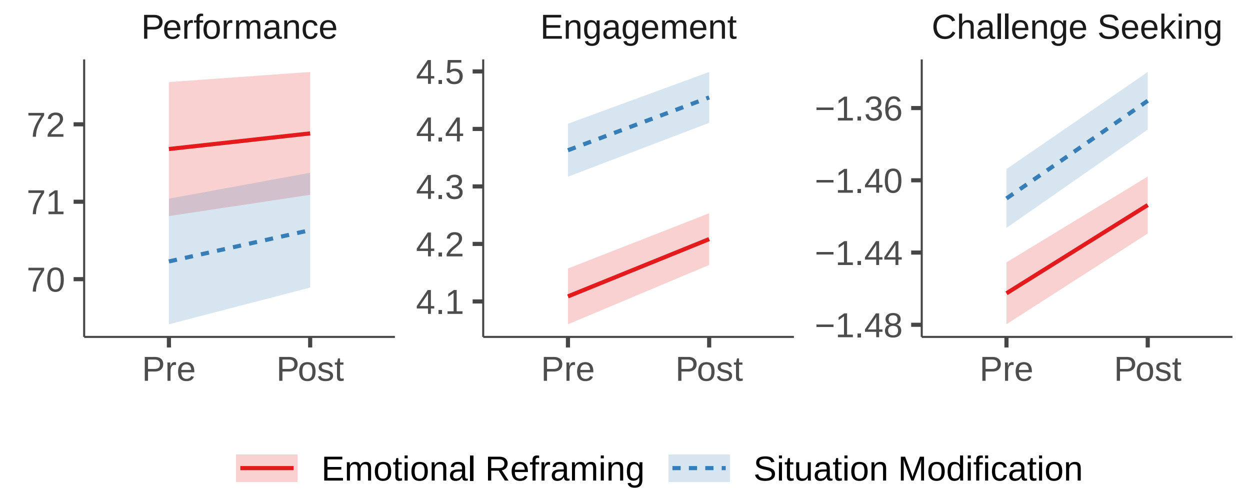

Figure 3.4: After students engaged with the situation modification and emotion regulation prompts, they saw increases in their motivation (operationalized as total time on task), and the difficulty of the items they attempted. There were no detectable effects on skill as measured by the proportion of correct responses. Results adjust for general time trends, all time-invariant student characteristics, item difficulty, and student's time-varying skill levels.

coefficients for difficulty ($\beta = 0.129,\ p < 0.001$ for Situation Modification; $\beta = 0.119,\ p < 0.001$ for Emotional Reframing) indicate that users attempted more challenging items after interacting with the chatbot. This increased willingness to tackle difficult problems aligns with the observed increase in engagement duration, suggesting that chatbot use may enhance user confidence or motivation.

As expected, user skill was strongly associated with higher accuracy ($\beta = 61.608$, $p < 0.001$ for situation modification; $\beta = 68.997$, $p < 0.001$ for emotional reframing), longer engagement durations, and willingness to attempt more difficult items. Similarly, item difficulty was negatively associated with accuracy and engagement duration.

3.3.3. Heterogeneity by baseline skill and gender

We find no moderation by gender. Under Emotional Reframing, higher-skill students show higher post-intervention accuracy; under Situation Modification, accuracy does not differ by baseline skill.

Table 3.1: Regression Results

	Engagement: Log(Time on Task)		Challenge-Seeking: Item Difficulty		
	Situation	Emotion	Situation	Emotion	
	(1)	(2)	(3)	(4)	
Post-Intervention Effect	0.426^{***}	0.558^{***}	0.129^{***}	0.119^{***}	
	(0.088)	(0.083)	(0.037)	(0.035)	
Time-Varying User Skill	1.057^{***}	0.737^{**}	-0.363^{***}	-0.537^{***}	
	(0.225)	(0.265)	(0.096)	(0.112)	
Time-Varying Item Difficulty	-0.294^{**}	-0.433^{***}			
	(0.099)	(0.099)			
Observations	1,782	1,834	1,782	1,834	
R^2	0.877	0.885	0.832	0.837	
Adjusted R^2	0.603	0.616	0.457	0.456	
Residual Std. Error	0.906 (df = 551)	0.932 (df = 547)	0.389 (df = 552)	0.402 (df = 548)	1

Note:

3.3.4. Analysis of conversations: Both prompts reduced users' negative emotion

Over the course of the conversations, it looks like the bot was effective in reducing the negative affect experienced by students. As shown in Figure 3.6, as users chatted with the bot, their language reflected reductions in negative sentiment and increases in positive sentiment.

3.4. Discussion

In this investigation, we deployed two motivational support chatbots—based on situation modification and emotional reframing—for a year on Khan Academy. To measure their impact, we compared user behavior before and after exposure to the chatbot intervention. Because users chose when to use the chatbots, we used Bayesian IRT to account for potential time-varying confounds. After these adjustments, students who engaged with either prompt worked longer, attempted more items, and chose harder items. However, these immediate behavioral changes did not translate into performance changes immediately. Students did not perform significantly better after engaging with the chatbots.

Why did these AI-mediated interventions improve persistence? Prior research suggests that these process model-based strategies help. In one experiment, students asked to study with their phones

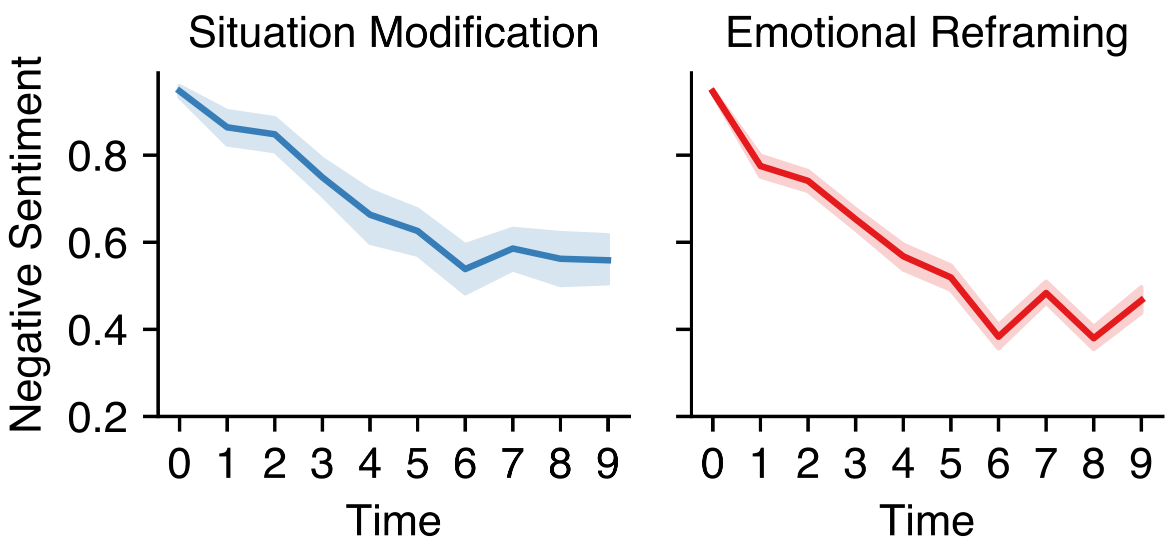

Figure 3.5: Negative sentiment decreased over the span of the conversation. Shaded area represents bootstrapped 95% confidence intervals. Ns = 405 and 2052 conversations for situation modification and emotional reframing. The correlations between turn number and time bins is -.33 and .39 for Situation Modification and Emotional reframing. If we do this within person (as I think we should), the correlations are weaker: -.07 and -.36, respectively.

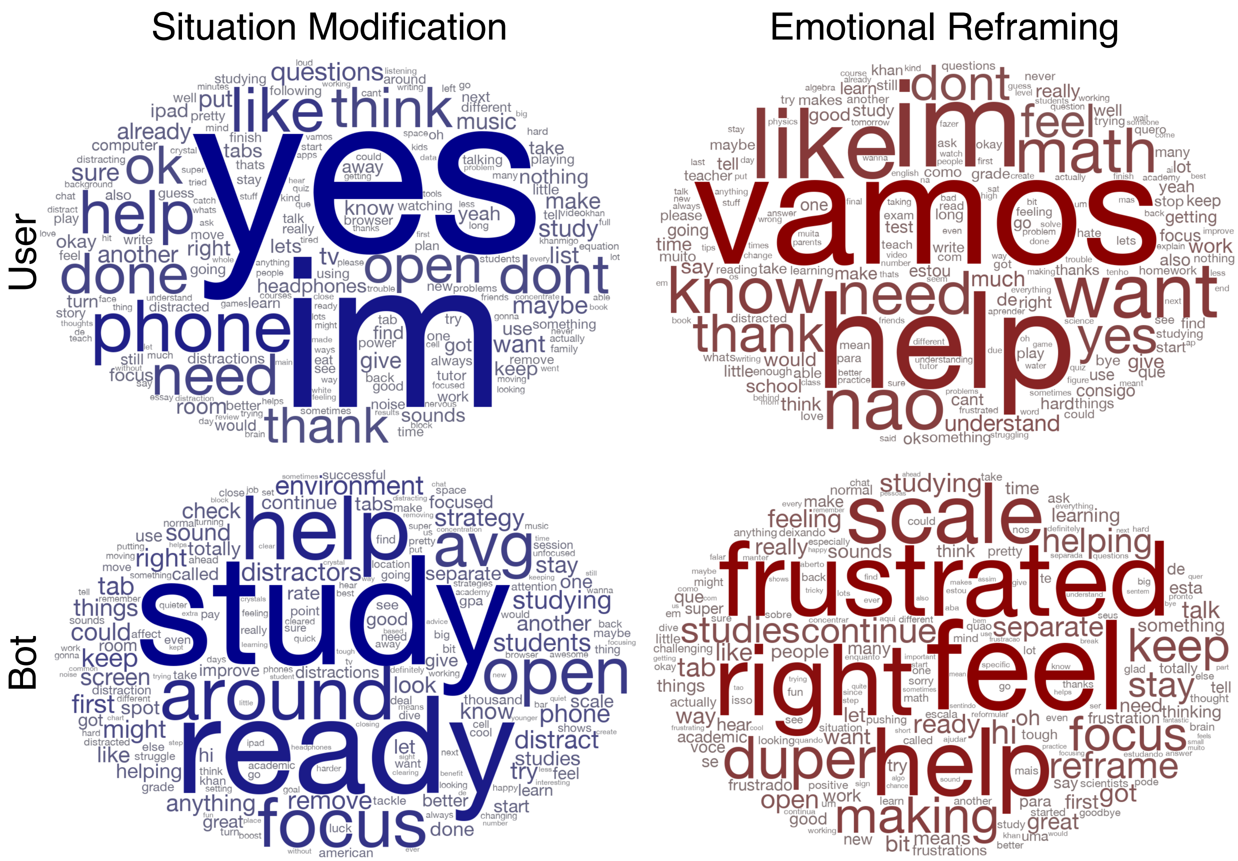

Figure 3.6: Word clouds from TF-IDF vectorization of user and bot text from situation modification and emotional reframing interventions.

further away from them received higher grades in subsequent tests. In another, providing students with the reframe that mistakes are a part of learning and not indicative of their worth also increased persistence. While expected for emotional reframing, our results suggest that situation modification may also be effective due to reductions in negative affect. Future research should confirm the relative increases in persistence of situation modification strategies due to reductions in negative affect.

If these results hold, they offer important implications for the deployment of positive online behaviors. First, is the idea that AI-chatbots, when adequately prompted, can work like motivational coaches. Just like tutors and mentors help people with the challenges that occur when things are difficult, complicated, or frustrating, AI-chatbots can help users deal with these difficulties. This can result in increased persistence in positive behaviors that would otherwise be ended.

Lower-skilled learners and users, however, may require supports beyond motivational interventions alone. Exploratory analyses of our data revealed that low-skilled users did not benefit from chatbot exposure. This pattern suggests that these users may require content-based instructional supports rather than purely motivational ones. Increased effort and persistence are unlikely to improve performance when learners lack fundamental knowledge or skills necessary for task completion. For such users, motivational interventions may be insufficient without concurrent provision of scaffolded learning content or skill-building resources.

3.4.1. Limitations and Future Directions

Several limitations of this study are worth highlighting. First and foremost, the highly naturalistic setting in which we collected the data meant that we were unable to randomly assign users to condition, weakening the causal interpretation of these findings. That said, the pattern of behavior prior to exposure to the chatbots does not suggest that regression to the mean is a viable explanation for these findings. We capitalized on the temporal resolution of the data to control for two crucial time-varying confounds: user motivation and skill, as well as item difficulty. If these findings are replicated in controlled experiments, we can have greater confidence that AI-based motivational coaches can increase persistence in challenging tasks.

A second limitation is generalizability. We tested these motivational coaches in the context of online learning on the Khan Academy platform. For the most part, this means school-aged children working on math. It is unclear how well these kinds of AI motivational coaches would translate to other important but tedious tasks and users. For instance, could these kinds of motivational AI agents help users navigate adult tasks like filing taxes or investing for retirement? Future studies could address these possibilities.

Finally, measuring changes in skill is difficult. Better measures and a more intentional design might have allowed us to better understand how AI motivational coaches could improve performance. With the crude measurement of percent correct, we were unable to find robust effects on performance.

Future work can capitalize on these opportunities. First, testing motivational AI coaches with randomized experiments is crucial to establish causality and better test the mechanism of the reduction of negative emotion. Second, our setup allowed us to make only one intervention available at a time, and then the other. Future research could address the possibility of personalization and targeted deployment. For instance, user behavior and the content of conversations with the AI tutor (i.e., Khanmigo) could be mined to proactively detect when students might benefit from receiving this intervention. Likewise, the content of the intervention itself could be tailored, such that students who need a situation modification intervention get that, and students who at some other point need some other thing (e.g., a growth mindset intervention), get that. Finally, the longer term effects of using AI motivational coaches shoudl be studied.

3.4.2. Conclusion

Final paragraph. Finish with the contributions: large study, high ecological validity, real-life stakes (math ed matters), self-initiated intervention — no need to nudge the students, the students nudge themselves. Finish with a positive sentence about the contributions and what this means. Taken together, this study provides rare, ecologically valid evidence—at a real scale and with self-initiated use—that AI-delivered, theory-based motivational prompts reliably increase persistence and encourage learners to tackle more challenging material. Performance gains appear to be contingent on baseline skill, highlighting a design principle: combining motivational coaching with instructional scaffolding. As digital learning and training increasingly rely on AI, embedding emotion-regulation supports offers a practical, scalable path to improving user engagement without requiring heavy-handed nudges or human coaches.

3.5. Methods

3.5.1. Setting and data

We analyzed de-identified platform logs from Khan Academy over a one-year period (January–December 2024). Two motivational support chatbots grounded in the process model of emotion regulation were available within the product: *Situation Modification* (SM) and *Emotional Reframing* (ER). Each exposure was user-initiated via an on-screen prompt (Figure 3.1); no pop-up or forced assignment was used. Logs include time-stamped item attempts (start/end times, correctness) and chatbot conversation text for users who chose to interact.

3.5.2. Participants

The analytic sample comprised 3,570 learners who engaged with the situation modification intervention and 3,608 students who wngaged with the emotional reframing intervention during the study window. Additionally, to be included, participants had to had had activity in both the pre- and post-exposure windows defined below. We included the first qualifying exposure per user per condition (i.e., each prompt was anlayzed separately). Users without activity in either the pre- or post-window were excluded from the corresponding analysis.

3.5.3. Interventions

The situation modification intervention encouraged users to structure their environment or routines to reduce reliance on willpower (e.g., minimizing distractions, planning short focused sessions). The emotional reframing intervention helped users reinterpret frustration and mistakes as expected, informative signals of learning progress. Both prompts were delivered via a short, text-based exchange that adapted to user inputs; no content tutoring was provided.

3.5.4. Design and exposure definition

Because use of the chatbots was self-initiated, we adopted a within-learner pre/post design with fixed effects and calendar-time controls. For each user's first exposure in a condition, we defined a pre-exposure window as the 7 days preceding the exposure (days $[-7, -1]$) and a post-exposure window as the 7 days following the exposure (days $[+1, +7]$). To avoid contaminating post measures with the triggering session, we excluded activity from the exposure session itself (day 0) in primary analyses. We computed outcomes within each window and compared post versus pre, controlling as described below.

3.5.5. Measures

Engagement was operationalized as the total time on task during the window, log-transformed for analysis. Time on task was computed from item start/end timestamps.

We operationalized challenge-seeking behavior as the average IRT-derived difficulty of items attempted during the window. Difficulty parameters were obtained from an item-response model described below and scaled so that higher values indicate more difficult items.

Performance was operationalized as the percentage of correct responses during the window. As a complementary specification, we also analyzed time-varying IRT ability ("skill") estimates.

3.5.6. Item and learner latent traits

We estimated time-varying learner skill and motivation and item difficulty using a Bayesian model at the attempt level (equations in the main text). Let s index learners, i items, and t ordered attempts. We modeled log-duration and correctness jointly:

$$\begin{aligned}
\text{Log-durations}_{s,i,t} &\sim \mathcal{N}(\mu_{\text{duration}}, \sigma_{\text{duration}}), \\
\mu_{\text{duration}} &= \text{difficulty}_i + \text{motivation}_{s,t} - \text{skill}_{s,t}, \\
p(\text{Correct}_{s,i,t} = 1) &= \sigma(\text{skill}_{s,t} - \text{difficulty}_i + \text{motivation}_{s,t}).
\end{aligned}$$

Latent skill and motivation followed separate Gaussian random walks with exponential priors on their step sizes; item difficulties had hierarchical Gaussian priors. Posterior summaries (posterior means) for skill, motivation, and item difficulty were carried forward into the outcome models.

3.5.7. Statistical analysis

Primary analyses were estimated separately for SM and ER. We fit linear models at the user $\times$ window level with user fixed effects and calendar-week fixed effects. For engagement we modeled log time on task; for challenge-seeking, average item difficulty; and for performance, percent correct. All models included time-varying skill and (when relevant) the average item difficulty in the window as covariates. Standard errors were clustered at the user level.

As robustness checks, we (i) excluded the 24 hours before and/or after exposure; (ii) varied the window length (3, 7, 14 days); (iii) re-estimated effects at the attempt level with user fixed effects; (iv) examined the second exposure to assess attenuation; and (v) repeated the performance analysis using IRT skill as the outcome. Heterogeneity analyses tested moderation by baseline skill (pre-exposure skill) and gender when available.

3.5.8. Ethical considerations

Analyses were conducted on de-identified data provided under a data-use agreement with Khan Academy. No interventions were randomized by the researchers; chatbot use was voluntary and user-initiated within the platform.

3.5.9. Pre-registration

Analyses, primary/secondary outcomes, and exclusion rules were preregistered prior to data analysis at https://aspredicted.org/g2p8-z8jg.pdf. Any deviations from the preregistered plan are noted in the Appendix.

CHAPTER 4

Using Artificial Intelligence to Assess Personal Qualities in College Admissions

Personal qualities like prosocial purpose and leadership predict important life outcomes, including college success. Unfortunately, the holistic assessment of personal qualities in college admissions is opaque and resource-intensive. Can artificial intelligence (AI) advance the goals of holistic admissions? While cost-effective, AI has been criticized as a "black box" that may inadvertently penalize already disadvantaged subgroups when used in high-stakes settings. Here we consider an AI approach to assessing personal qualities that aims to overcome these limitations. Research assistants and admissions officers first identified the presence/absence of seven personal qualities in n = 3,131 applicant essays describing extracurricular and work experiences. Next, we fine-tuned pretrained language models with these ratings, which successfully reproduced human codes across demographic subgroups. Finally, in a national sample (N = 309,594), computer-generated scores collectively demonstrated incremental validity for predicting six-year college graduation. We discuss challenges and opportunities of AI for assessing personal qualities.[2]

4.1. Introduction

Many colleges embrace the ideals of holistic review. In a recent survey by the National Association for College Admissions Counseling, 70% of admissions officers said they consider personal qualities to be an important factor when selecting applicants (49). This aim is justified by longitudinal research affirming that personal qualities—whether referred to as "non-cognitive skills," "social-emotional competencies," "personality," or "character"—predict positive life outcomes in general and success in college in particular (50–53). Moreover, a holistic admissions process can advance equity, some argue, as applicants are able to demonstrate qualifications not reflected in their standardized test scores, which tend to be highly correlated with socioeconomic advantage (54).

However, history shows that equity is certainly not guaranteed by holistic review. A century ago, Columbia University first began requiring applicants to write a personal essay, which admissions

[2]This work was published at *Science Advances*. Supplementary Online Materials for this chapter are available at https://www.science.org/doi/suppl/10.1126/sciadv.adg9405/suppl_file/sciadv.adg9405_sm.pdf

officers evaluated for evidence of "good character" (55). Previously, the university's admissions decisions had been based primarily on standardized test scores. The result was a growing proportion of Jewish students in each entering class, which in turn led to concerns that, as Columbia's dean at the time put it, the campus was no longer welcoming to "students who come from homes of refinement" (p. 87). It has been argued that for Columbia and other Ivy League colleges in that era, not requiring the justification, explanation, or even disclosure of these summary character judgments enabled the unfair exclusion of qualified Jewish applicants.

Although its aims may be nobler today, the holistic review process itself remains much the same. Admissions officers still rely heavily on the personal essay to evaluate an applicant's personal qualities (49). The particulars of how, or even which, personal qualities are assessed, remain undisclosed to either applicants or the public, and even the "admissions officers themselves simply do not have a common definition of holistic review beyond 'reading the entire file'" (56). As one admissions officer put it, the status quo of holistic review is both "opaque and secretive (57)."

Recently, a more transparent and systematic process has been recommended for the holistic review of personal qualities in college admissions. Specifically, admissions officers have been urged to assess individual personal qualities separately (as opposed to making a summary judgment of "good character"), to use structured rubrics (as opposed to intuition), and to carry out multiple, independent evaluations (as opposed to relying on a single officer's judgment) (54, 58). Such recommendations represent the application of basic psychometric principles and, in research contexts, have long been used to increase the reliability, validity, and interpretability of human ratings (59, 60). Moreover, the transparency of this systematic approach should limit bias—whether accidental or intentional.

In college admissions, however, this ideal is hardly ever achieved. The soaring number of applications that admissions officers must review—which for the majority of colleges has more than doubled in the last two decades—affords extraordinarily limited time to review each one (61, 62). These logistical and budgetary constraints are likely to continue to prohibit the implementation of best practices that, were resources unlimited, could optimize reliability, validity, interpretability—and in turn, equity.

Can artificial intelligence (AI) advance the aims of holistic review? With stunning efficiency, AI systems identify patterns in data and, with stunning fidelity, apply learned models to new cases. For example, a computer algorithm could be trained to generate personal quality scores from student writing instantaneously, reliably, and at near-zero marginal cost. However, there are concerns that the "black box" of an AI algorithm may inadvertently perpetuate, or even exacerbate, bias against disadvantaged subgroups (63, 64). Such bias has been shown in the domains of hiring, criminal justice, and medical diagnosis (65–67). In college admissions, AI-quantified essay content and style have been shown to correlate more strongly with household income than do SAT scores (68). Opaque AI algorithms that provide fertile ground for bias recall the anti-Semitic holistic review practices of a century ago.

Efforts within the AI community to address these issues have given rise to concepts such as Human-Centered AI (69, 70) and Explainable AI (71). These frameworks emphasize alignment with stakeholder objectives, interpretability, and equity—while promoting the idea of automation as a complement rather than a substitute for human control (70). Rather than simply maximizing predictive accuracy, these approaches prioritize alignment with stakeholder goals (e.g., admitting students who demonstrate prosocial purpose, (72)), interpretability (e.g., providing separate, face-valid scores for separate personal qualities rather than a single summary score of character with no evidence of face validity), and rigorously auditing model outputs for unintended bias. By prioritizing these aspects, digital technology can facilitate the identification of discrimination and contribute to rectifying historical exclusion (73).

In this investigation, we developed an artificial intelligence (AI) approach to assessing personal qualities with these priorities in mind. We began with a de-identified sample of 309,594 college applications (see **Figure 1**). Each included a 150-word essay describing an extracurricular or work activity of the applicant's choice. Next, in a *Development Sample* of 3,131 essays, research assistants (RA) and admissions officers (AO) identified the presence or absence of seven different personal qualities commonly valued by universities and shown in prior research to predict college success (51). See **Table 4.1**. Research assistant and admissions officer ratings were used to fine-tune separate RoBERTa language models (74) for each personal quality. We then confirmed each model's interpretability as well as evidence of convergent, discriminant, and predictive validity by demographic subgroup. Finally, we applied these fine-tuned models to the *Holdout Sample* of 306,463 essays, examining associations between computer-generated personal quality scores, demographic characteristics, and six-year college graduation.

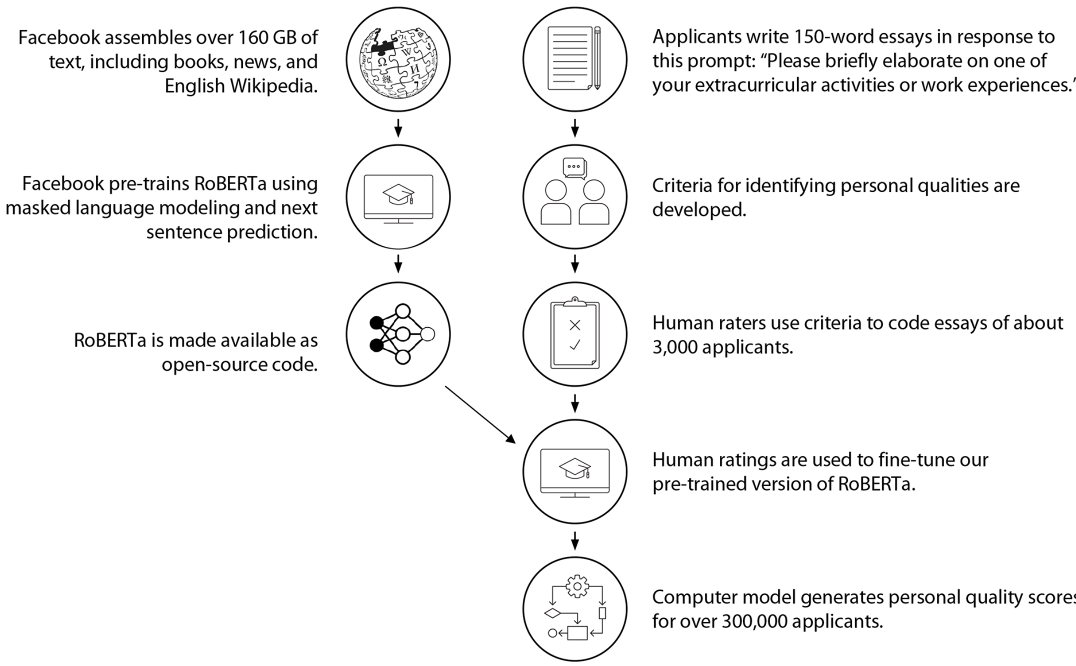

Figure 4.1: An artificial intelligence approach to assessing personal qualities in college admissions.

Table 4.1: Personal qualities and example essay excerpts

Personal quality	Fictionalized excerpts
Prosocial purpose Helping others, wanting to help others, considering the benefits to others, mentioning reasons for helping others, or reflecting on how enjoyable or rewarding it is to help others.	Every summer for the last three years, I worked as camp counselor at a camp for young children from underprivileged families. Helping children realize their hidden talents is one of the most rewarding experiences I have ever had. I've been so fulfilled by watching these children develop confidence in their abilities. This experience has been so important to me, and it showed me that a career in education is where I belong.
Leadership Serving in a leadership role, commenting on what he or she did in his or her capacity as a leader, or discussing the value, meaning, or importance of leadership.	I was chosen to be cheerleading captain during my senior year. My freshman year captain had a huge impact on my life, and I felt like it was my time to pay it forward. I am so proud of everything I did for the girls: creating a mentorship system, organizing events and fundraisers, and encouraging everyone to work as hard as they could. At the end of the year, a few girls thanked me. I was completely overcome with emotion. I've never felt so gratified in my life.
Learning Improving, learning, or developing knowledge, skills, or abilities.	I played softball in high school. When I started, I was not a very strong player. When I finally made the varsity team my senior year, I was determined to have a better season. I worked constantly to improve my game – during practice and on my own time. My skills grew so much. Because of my hard work, I finished the year with the best record on my team!
Goal pursuit Having a goal and/or a plan.	I have been playing soccer since I was six years old. Unfortunately, last year I injured my knee, and it has been a struggle to get back to the level I was playing at before my injury. It has been really challenging, but I've been doing physical therapy and practicing everyday so that I can be a varsity starter this year.
Intrinsic motivation Describing the activity as enjoyable or interesting. Liking the activity or identifying with it.	Running track is so much more than a sport to me. It's a challenge and an adventure, and I put everything I have into it. I love every aspect of it, even the afternoons I spend drenched in sweat in the scorching heat.
Teamwork Working with or learning from others. Valuing what fellow participants bring to the activity.	I've been on my school's debate team since my freshman year, and was elected co-captain because of my commitment to the team's success. My fellow co-captains and I worked together to get our team ready for competitions. We knew that a strong team performance was more important than the successes of a few individuals. We stressed teamwork and cooperation between our teammates. Because we focused on team effort, we earned first place at the state meet.
Perseverance Persisting in the face of challenge.	I've learned to become a gracious victor and to grow from defeat. Track has helped me overcome my fear of losing, and even helped me put my life in perspective. I've learned to keep working and fighting even when the odds seem impossible to beat. There were many times that I found myself lagging, but I pulled ahead at the end because I never gave up. The most important thing I've learned is to never let anything stand in my way.

Note. Our data use agreement with Common App does not allow us to publish real excerpts to protect student identity.

Table 4.2: Descriptive statistics and correlations between human ratings and computer-generated likelihoods of personal qualities in the Development Sample

	Research Assistant Ratings							Admissions Officer Ratings						
Personal Quality	PP	LD	TW	LR	PS	IM	GP	PP	LD	TW	LR	PS	IM	GP
Computer-Generated Likelihoods														
1. Prosocial purpose (PP)	.86***	−.01	−.04*	−.09***	−.12***	−.05**	.04*	.80***	−.13***	.13***	−.22***	−.24***	−.09**	−.19***
2. Leadership (LD)	−.01	.81***	.15***	−.01	.00	−.09***	.05**	.13***	.73***	.16***	−.15***	−.06***	−.16***	.01
3. Teamwork (TW)	−.07***	.18***	.62***	.06**	.07***	−.02	.06**	−.18***	.16***	.62***	.07***	.10***	−.03	.10***
4. Learning (LR)	−.10***	−.05**	.04*	.77***	.11***	−.01	−.03	−.28***	−.15***	−.07***	.65***	.07***	.01	.09***
5. Perseverance (PS)	−.16***	−.01	.06**	.10***	.67***	.03	.05**	−.35***	−.10***	.11***	.08***	.48***	.09***	.26***
6. Intrinsic motivation (IM)	−.05**	−.09***	.00	−.03	.04*	.73***	.03	.08***	−.24***	−.08***	−.01	.08***	.45***	−.05***
7. Goal pursuit (GP)	.06***	.06**	.06***	−.01	.02	.02	.59***	−.31***	−.05**	.12***	.15***	.27***	.06***	.45***
Descriptive Statistics														
Human Inter-Rater Reliability	.83	.78	.61	.73	.66	.63	.57	.60	.49	.30	.31	.24	.23	.15
Frequency of Human Rating	.34	.18	.26	.42	.19	.42	.31	.28	.25	.22	.44	.21	.41	.25
Mean of Computer-Generated Likelihood	.36	.19	.26	.45	.19	.45	.32	.30	.25	.22	.46	.24	.42	.25

Note. Inter-rater reliability for human raters was measured with Krippendorf's alpha. Correlations between human ratings and computer-generated likelihoods for the same personal qualities are shown along the diagonals. All correlations are point-biserial correlation coefficients between binary human ratings and continuous computer-generated likelihoods. * $p < .05$. ** $p < .01$. *** $p < .001$. n = 3,131. n for inter-rater reliability = 206 essays coded by multiple research assistants, n = 3131 essays coded by two admission officers.

4.2. Results

On average, research assistants and admissions officers found evidence for two out of seven personal qualities in each essay. As shown in **Table 4.2**, some personal qualities were more commonly observed than others. For instance, research assistants and admissions officers identified leadership in 42% and 44% of essays, respectively; in contrast, they identified perseverance in only 19% and 21% of essays, respectively. Correlations between research assistant and admission officer ratings ranged between $\phi = .193$ and $.703$, $ps < .001$.

Using these binary human ratings, we fine-tuned separate RoBERTa models to produce continuous likelihood scores for each personal quality, and each kind of rater. See **Section 2** in **Supplementary Materials** for details on model pretraining and fine-tuning.

4.2.1. Model interpretability

We used the *transformers-interpret* package (75, 76) to identify the words (or fractions of words) that these fine-tuned RoBERTa models relied on most to generate personal quality scores. As shown in **Figure 4.2,** there was reasonable evidence of face validity. For instance, RoBERTa assigned higher scores for leadership when essays mentioned "president," "leader," and "captain." Models trained on admission officer ratings produced similar attribution scores: average word-level attribution scores correlated between .392 and .983, $ps < .001$. See **Section 7** in **Supplementary Materials** for details.

4.2.2. Convergent and discriminant validity of computer-generated likelihoods in the Development Sample

Computer-generated likelihoods for each personal quality converged with human ratings of the same personal quality (rs ranged from .59 to .86, average $r = .74$, for research assistants; rs ranged from .45 to .80, average $r = .62$, for admission officers). In contrast, computer-generated likelihoods for a particular personal quality did not correlate with human ratings of other personal qualities (rs from -.16 to .18, average $r = .01$, for research assistants; rs from -.35 to .27, average $r = -.03$, for admission officers). See **Table 4.2**. Unsurprisingly, the more reliably human raters were able

Figure 4.2: Complete or partial words on which RoBERTa models finetuned on research assistants relied most for generating personal quality scores. Font size is proportional to word importance. Darker words are more common. Token "gru" is a fraction of the word "grueling", Token "unte" is a fraction of the word "volunteer". Words importance is not invariant across essays, it depends on word context. Word importance and frequency were largely independent (r = -.03, p $<$ .001). For instance, for intrinsic motivation, the model relied more on the word "pleasure" then the word "fun," but essays were more likely to contain the word "fun" then the word "pleasure."

to code each personal quality, the better the computer-generated likelihoods of personal qualities matched these ratings (r = .95, p = .001, for research assistants; r = .94, p = .001, for admission officers). In the sub-sample of essays that were coded by multiple raters, model scores correlated more strongly with human ratings than human ratings correlated with each other ($M_{human-computer}$ = .74, $M_{human-human}$ = .69, t = 4.16, p = .006, for research assistants; $M_{human-computer}$ = .60, $M_{human-human}$ = 0.28, t = 19.40, p < .001, for admissions officers). There were positive correlations between computer-generated likelihoods for personal qualities from models trained on research assistants and admissions officers (rs ranged from .394 to .869, ps < .001).

4.2.3. Convergent validity does not vary by demographic subgroup in the Development Sample

Correlations between human ratings and computer-generated likelihoods of personal qualities were similar across subgroups. For example, the average correlation between human-rated and computer-generated personal quality scores was .74 for female applicants and .73 for male applicants, for research assistants. The pattern of results was equivalent for admission officers. As shown in **Tables S11 and S12**, after correcting for multiple comparisons (77), 13% and 9% of the correlations differed by subgroup for research assistants and admissions officers, respectively. In about half of these comparisons, the models were more accurate for the marginalized group, while in the other half, the majority subgroup was favored. In most of these cases the difference between the correlations was not very large (mean $|\Delta r_{RA}| = .054$, range of Δr_{RA} = -.121 – .056), mean $|\Delta r_{AO}| = .10$, range of Δr_{AO} = -.206 – .123).

4.2.4. Human ratings and computer-generated likelihoods were largely unrelated to demographics in the Development Sample

Demographic characteristics were largely unrelated to personal qualities, whether assessed by human raters (mean $|\phi_{RA}| = 0.02$, mean $|\phi_{AO}| = 0.03$) or by computer algorithm (mean $|d_{RA}| = 0.06$, mean $|d_{AO}| = 0.08$). One exception is that female applicants were rated as more prosocial than male applicants ($\phi_{RA} = .13$, $\phi_{AO} = .12$ $p < .001$ for human ratings, $d_{RA} = 0.26$, $d_{AO} = .28$, $p < .001$ for computer-generated likelihoods, p-values adjusted for multiple comparisons (77))—in line with other research showing gender differences in prosocial motivation and behavior favoring women (78). See **Table S5** in **Supplementary Materials** for details.

4.2.5. Computer-generated likelihoods were as predictive of college graduation as human raters in the Development Sample

To compare the predictive validity of the computer-generated likelihoods with human ratings, we ran two logistic regression models in which personal qualities predicted college graduation. The computer-generated likelihoods were slightly more predictive than the human ratings, but the difference in the AUCs was not significant ($AUC_{human} = .565$, $AUC_{computer} = .574$, $\Delta AUC = .009$, $p = .274$, for research assistants; $AUC_{human} = .587$, $AUC_{computer} = .603$, $\Delta AUC = .017$, $p = .120$, for admission officers). Coefficients were slightly larger for computer-generated likelihoods as compared to human ratings ($t = 2.33$, $d = 0.882$, $p = .059$, for research assistants; $t = 3.89$, $d = 1.469$, $p = .008$, for admissions officers).

Table 4.3: Odds ratios from binary logistic regression models predicting six-year college graduation in the $N = 306{,}463$ Holdout Sample

	Research Assistant		Admission Officer	
	(1)	(2)	(1)	(2)
Computer-generated likelihoods of personal qualities				
Prosocial purpose	1.132***	1.075***	1.252***	1.116***
	(0.005)	(0.005)	(0.006)	(0.006)
Leadership	1.133***	1.065***	1.214***	1.084***
	(0.005)	(0.005)	(0.005)	(0.005)
Teamwork	1.080***	1.031***	1.135***	1.062***
	(0.005)	(0.005)	(0.005)	(0.005)
Learning	1.065***	1.045***	1.146***	1.034***
	(0.004)	(0.005)	(0.005)	(0.005)
Perseverance	1.071***	1.012**	1.089***	1.047***
	(0.005)	(0.005)	(0.005)	(0.006)
Intrinsic motivation	1.068***	1.007	1.142***	1.009
	(0.004)	(0.005)	(0.005)	(0.005)
Goal pursuit	1.041***	1.005	1.048***	1.030***
	(0.004)	(0.005)	(0.005)	(0.005)
Race/ethnicity (vs. white)				
Black		0.774***		0.775***
		(0.019)		(0.019)
Latino		0.871***		0.868***
		(0.019)		(0.019)
Asian		0.735***		0.739***
		(0.017)		(0.017)
Other		0.749***		0.750***
		(0.017)		(0.017)
No race reported		0.849***		0.853***
		(0.013)		(0.013)
Parental education (vs. no parent w/ college degree)				
One parent w/ college degree		1.199***		1.198***
		(0.012)		(0.012)
Two parents w/ college degree		1.335***		1.334***
		(0.012)		(0.012)
Female		1.435***		1.430***
		(0.010)		(0.010)
Married parents		1.311***		1.308***
		(0.011)		(0.011)
English language learner		0.769***		0.774***
		(0.015)		(0.016)
Title 1 high school		0.951***		0.947***
		(0.013)		(0.013)
Out-of-school activities (OSA)				
Number of OSA		1.250***		1.241***
		(0.005)		(0.005)
Time per OSA		1.088***		1.083***
		(0.004)		(0.004)
Proportion sports		1.042***		1.035***
		(0.005)		(0.005)
Standardized test scores		1.489***		1.482***
		(0.006)		(0.006)
Constant	3.555***	2.533***	3.585***	2.543***
	(0.004)	(0.014)	(0.004)	(0.014)
AUC	.560	.689	.576	.690

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

4.2.6. Computer-generated likelihoods were largely independent of demographics but, in support of criterion validity, predicted graduation in the Holdout Sample

Next, we applied the models fine-tuned on research assistants and admissions officers to the *Holdout Sample* of 306,463 essays. For both categories of models, reliability across models trained on different subsets of the data was high (range of Cronbach's $\alpha = .990$ to .997, for research assistants; range of Cronbach's $\alpha = .988$ to .998, for admission officers). Even when considering any two models, they were likely to produce similar results (average inter-model correlation ranged from .910 to .967 for research assistants, .896 to .978 for admission officers). Correlations between computer-generated likelihoods for personal qualities from models trained on research assistants and admissions officers ranged from .418 to .896, $ps < .001$.

As in the development sample, computer-generated likelihoods for personal qualities were similar across demographic subgroups (mean $|d_{RA}| = 0.05$, mean $|d_{AO}| = 0.06$). In contrast, and as expected, demographics were more strongly related to standardized test scores (mean $|d| = 0.38$) and degree of participation in out-of-school activities (mean $|d| = 0.17$). See **Tables S7** and **S10** in **Supplementary Materials** for details.

About 78% of students in the *Holdout Sample* graduated from college within 6 years. As shown in Model 1 in **Table 4**, computer-generated likelihoods for personal qualities were each modestly predictive of college graduation when controlling for each other (ORs from 1.041 to 1.132, $ps < .001$, $AUC = .560$, for research assistants; ORs from 1.048 to 1.252, $ps < .001$, $AUC = .576$, for admission officers). To estimate a ceiling on how much the essays can predict subsequent college graduation, we trained a RoBERTa model to predict college graduation from students' responses. This model achieved an out-of-sample AUC of .626, suggesting that consistent with previous research (68) essays do encode information predictive of graduation outside of personal qualities. The same procedure using personal qualities results in smaller out-of-sample AUCs ($AUC_{RA} = .557$, $AUC_{AO} = .568$). See **Section 8** in **Supplementary Materials** for details.

As shown in Model 2 in **Table 4**, in the models trained on research assistants, five of seven personal qualities remained predictive of college graduation when controlling for each other, demographics, standardized test scores, and out-of-school activities (ORs from 1.012 to 1.075, ps $<$ .01). In the models trained on admissions officers, six of seven personal qualities remained predictive (ORs from 1.030 to 1.116, ps $<$ .01). See **Figure S2** in **Supplementary Materials** for details on imputation.

As a further test for fairness, we tested whether the predictive power of computer-generated likelihoods of personal qualities was equivalent across subgroups. We added interaction terms between each personal quality and standardized test scores and each demographic characteristic. After controlling for multiple comparisons (77), we confirmed that the predictive effect of personal qualities was equal across demographic subgroups. Comparatively, the predictive accuracy of standardized tests differed across subgroups (mean $|\beta|$ = -0.053). We also tested for differences in predictive validity in intersections of two demographic subgroups (e.g., Black English language learners, Women in title 1 high schools). There were no consistent or theoretically interpretable patterns in these intersectional analyses. See **Section 9** in **Supplementary Materials** for details.

4.3. Discussion

In a national dataset of over 300,000 college applications, we evaluated an artificial intelligence approach to measuring personal qualities from student writing. Specifically, we fine-tuned RoBERTa language models using expert ratings of prosocial purpose, leadership, teamwork, learning, perseverance, intrinsic motivation, and goal pursuit, respectively, in applicants' essays about their out-of-school activities. We found that these models demonstrated convergent, discriminant, and predictive validity—and this evidence was consistent across demographic subgroups. In addition, computer-generated scores were largely independent of demographics.

In contrast, two prior studies found that AI-extracted admission essay content and style correlate with socioeconomic status. Alvero and colleagues (68) found that students from wealthier families tend to write about certain essay topics (e.g., human nature), whereas disadvantaged students tend to write about others (e.g., tutoring groups). Likewise, Pennebaker and colleagues (79) found that categorical words (e.g., articles, prepositions) versus dynamic words (e.g., pronouns, adverbs) in college essays correlate with parental education at r = .22. Why do our results differ? It seems likely that personal qualities are distributed more evenly across demographic subgroups than the topics students choose to write about or the words they use to do so. However, we cannot rule out methodological differences. Alvero et al. (68) used essays from the University of California system, and Pennebaker et al. (79) used essays from a large state university. In contrast, our sample included a larger and more diverse set of public and private four-year colleges from across the United States. In addition, both of these prior studies used personal statements totaling several hundred words, whereas the essays to which we had access were a maximum of 150 words and focused specifically on extracurricular activities and work experiences. Finally, rather than using unsupervised topic modeling or dictionary approaches, we fine-tuned a language representation model using human ratings that themselves were shown to be unbiased.

Several limitations of this investigation suggest promising directions for future research. First, while our national data set was unusually large and diverse, it did not include the 650-word personal essay now required by the Common Application. Unfortunately, applicants in 2008-09 submitted their personal essays as attached PDF files that were not feasible to de-identify. A replication and extension of our study using a more recent cohort of applicants should not face this limitation.

Second, and relatedly, because the majority of applicants in our sample submitted their high school transcripts as attached PDF files that could not be de-identified, our data set included high school GPAs for only a subsample of 43,592 applicants whose school counselors entered grades directly into the Common Application online portal. While our robustness check using this subsample (see **Supplementary Materials Table S52**) affirm the conclusions of our main analyses, future research should not face this limitation.

Third, the observed effect sizes for personal qualities predicting college graduation were modest, both in absolute terms and relative to the predictive validity of standardized test scores. They were, however, somewhat larger than predictive validities of questionnaire measures of personal qualities like growth mindset (80). As context, a growing literature suggests that long-term life outcomes are extremely difficult to predict with precision (81, 82), in part because the greater the number of factors that determine an outcome, the smaller the influence of any single one (83, 84).

Relatedly, it is worth noting that myriad factors unmeasured in this investigation have been shown to influence college graduation, including the ability to afford tuition payments (85), academic preparation and support (86, 87), and sense of belonging (80, 88).

Fourth, college graduation was the only outcome available in our dataset. We therefore could not evaluate the impact of personal qualities on other aspects of college success—such as GPA, extracurricular involvement, contributions to the campus community—nor on social or emotional well-being (89). This limitation, while not atypical, illuminates a more general concern with research on college admissions, namely the lack of explicit, consensual priorities for what college admissions decisions are aimed at optimizing and how such outcomes are operationalized. One unexpected benefit of evaluating AI approaches, therefore, is the critical perspective brought to the current status of holistic review and selective admissions. Thus, future research and practice should focus on clarifying the goals of holistic review (56) before automating parts of the process.

Finally, inter-rater reliability estimates and human-computer correlations were lower for admissions officers than for research assistants. These disparities may reflect differences in methodology (e.g., research assistants received more training on the coding instructions) or in rater perspective (e.g., heterogeneity in admission officers' ratings may reflect differences in the priorities of their universities). Our data do not distinguish between these possibilities. Regardless, it seems likely that the more reliable ratings of research assistants provided a more consistent signal for the models to learn from, resulting in higher human-computer correlations for research assistants compared to admissions officers. Notably, computer-generated scores for personal qualities were at least as, if not more, predictive of college graduation when the algorithm was trained by admissions officers as when it was trained by research assistants. While surprising, this pattern of results underscores the fact that increasing reliability does not always increase validity. By analogy, a questionnaire can achieve nearly perfect internal reliability when items are practically synonymous, but only at the cost of content and predictive validity (90).

In sum, this investigation suggests that an artificial intelligence approach to measuring personal qualities warrants both optimism and caution. On one hand, our findings demonstrate that AI models trained on human ratings are not only efficient (yielding millions of personal quality scores in a matter of minutes, replicating human ratings with uncanny precision) but also interpretable (as opposed to an inscrutable "black box") and auditable for fairness to demographic subgroups. On the other hand, Campbell's Law (91) states that the more weight given to an assessment in high-stakes decisions (as opposed to low-stakes research), the greater the incentive for distortion. It is not hard to imagine how applicants might try to mold their essays—perhaps using AI tools such as ChatGPT—to match what admissions officers, and the algorithms they train, are looking for. We can only assume that applicants from more advantaged backgrounds would be better positioned to do so. What's more, algorithms make mistakes, particularly insofar as they look for patterns and thus, by design, are blind to exceptions. For instance, our fine-tuned RoBERTa model gives the sentence "I donated heroin to the children's shelter" an extremely high score for prosocial purpose. Thus, we recommend artificial intelligence be used to augment, not replace, human judgment. No algorithm can decide what the goals of a university's admissions process should be nor what personal qualities matter most for that community. Seeing algorithms as complements rather than replacements for human judgment may also counter algorithm aversion—the tendency to trust human decision makers over algorithms, even in the face of contradictory evidence (92). With these caveats in mind, we conclude with the observation that progress in any field depends on dissatisfaction with the status quo; there is no doubt that when it comes to the assessment of personal qualities in college admissions, we can do better.

4.4. Materials and Methods

4.4.1. Participants

After exclusions, our sample consisted of 309,594 students who applied to universities in 2008-09. To provide labeled data for the machine learning algorithm, we set aside a *Development Sample* consisting of 3,131 applications for manual coding. We used stratified random sampling to ensure representation across demographic groups and levels of involvement in extracurricular activities. The *Holdout Sample* was composed of the remaining 306,463 essays. We applied the fine-tuned algorithm to these essays and tested the relationship between the computer-generated likelihoods of personal qualities and demographics as well as college graduation. See **Section 1** in **Supplementary Materials** for details on missing data and exclusion criteria.

4.4.2. Measures

Extracurriculars essay

In up to 150 words, applicants who completed the Common Application were asked to respond to the following prompt: "Please briefly elaborate on one of your activities or work experiences." We excluded all essays shorter than 50 characters, most of which were mentions to attachments (e.g., "See attached"). The critical role of extracurricular commitments (i.e., structured pursuits outside of the classroom) in the expression and development of personal qualities in youth has been documented in the literature on positive youth development (93, 94).

Standardized test scores

Over half (55%) of the Holdout Sample reported SAT scores, 14% reported ACT scores, 25% reported both, and 6% reported neither. Using published guidelines (95), we converted ACT scores to SAT scores. For students who reported both test scores, we selected the higher score, and for students who reported neither, data were considered missing.

Extracurricular activities

Applicants listed up to seven extracurricular activities and for each, indicated the years they had participated. For each applicant, we computed the total number of extracurricular activities, mean years per activity, and the proportion of activities that were sports.

Demographics

We obtained the following demographic information from the Common Application: race/ethnicity, parental education, gender, parents' marital status, English language learner status, and type of high school (i.e., Title 1 public school vs. other kinds of schools).

College graduation

We obtained data from the 2015 National Student Clearinghouse (NSC) database (www.studentclearinghouse.org) to create a binary six-year graduation measure (0 = did not earn a bachelor's degree from a four-year institution within six years of initial enrollment; 1 = earned a bachelor's within six years). We obtained institutional rates of graduation within six years from the National Center for Educational Statistics (NCES). We control for any potential effects of baseline institutional effects on the odds of graduation in the **Supplementary Materials Table S53**.

4.4.3. Analytic Strategy

To handle missing data, we used multiple imputation ($m = 25$), employing the mice package in R (96). We used predictive mean matching for graduation rates and college admissions test scores. For school type, we used polytomous regression. In the *Holdout Sample,* 5.7%, 12.2%, and 7.1% of students were missing data on admissions test scores, six-year institutional graduation rates, and high school Title 1 status, respectively.

In binary logistic regression models, we standardized all continuous variables to facilitate interpretation of odds ratios. Factor variables were dummy-coded and, along with binary variables, were not standardized, such that the effects shown indicate the expected change in the odds of each variable relative to the comparison group.

When averaging correlations together, we transformed the correlation coefficients to z-scores using Fisher's transformation, averaged them, and transformed them back to correlation coefficients.

Following convention, we report p-values for our analyses. It is important to note that p-values do not directly indicate practical importance, especially in the context of large sample sizes. With larger samples, even small effects can yield statistically significant results, potentially misleading interpretations of the findings. Therefore, we emphasize the importance of focusing on effect sizes, which provide a more meaningful measure of the magnitude of associations or differences.

4.4.4. RoBERTa fine-tuning procedure

Robustly Optimized BERT Pretraining Approach (RoBERTa) (74) is an advanced language representation model considered a meaningful innovation that improves on prior algorithms in the field of natural language processing. It is a deep neural network that has been pretrained by having it predict masked words in extremely large volumes of generic text (i.e., books and English Wikipedia).

The fine-tuning process consists of adjusting the parameters of the final layers in order to maximize predictive accuracy in particular tasks (e.g., text classification) and in a particular corpus of text (e.g., admissions essays).

We used a subset of essays that were not manually coded to do a round of pretraining to optimize the RoBERTa model to our admission essay corpus. To do this, we trained RoBERTa to predict a masked word given the surrounding words. This process resulted in a RoBERTa model optimized for the particular prompt the essays in our corpus were answering. See **Section 2** in **Supplementary Materials** for technical details on the pretraining process.

To begin the fine-tuning procedure, the second and third authors read random batches of 50 applicant essays to identify salient personal qualities commonly identified by colleges as desirable and/or shown in prior research to be related to positive life outcomes. After reading and discussing nine batches of 450 essays each, they developed criteria for seven personal qualities: prosocial purpose, leadership, teamwork, learning, perseverance, intrinsic motivation, and goal pursuit.

Next, we trained five research assistants to apply these criteria until each coder achieved adequate inter-rater reliability with either the second or third author across all seven attributes (Krippendorff's alpha $>$ .80). Raters then coded all 3,131 essays in the *Development Sample.* Most of the essays were coded by a single rater (n = 2,925; 93% of the *Development Sample*). To assess inter-rater reliability, pairs of raters independently coded a subset of essays (n = 206; 7% of the *Development Sample*).

Additionally, we recruited 36 admissions officers to provide expert ratings of personal qualities. We recruited them through Character Collaborative, a mailing list sent by NACAC, and the College Guidance Network. Admissions officers completed a short training which consisted on reading definitions, examples, and rating an example essay, and then were able to rate as many essays as they desired. Each admissions officer rated an average of 86 essays. Each essay in the *Development Sample* was rated by two different admissions officers.

We used these manually annotated data sets to fine-tune two sets of separate RoBERTa models to estimate the probability of each personal quality: one set on the ratings by research assistants and one set on the ratings by admission officers. After fine-tuning these models, we evaluated the performance of the models and applied it to the holdout sample of 306,463 essays, yielding more than two million continuous codes.

4.4.5. Acknowledgements

We thank Natalie Yee, Xena Wang, The Character Collaborative, NACAC, and Damien Crone for their help in this research. We thank Sarah Walter Kotlinski, Lisa Mortini, Zoey Stenson, Cigus Vanni, Janey Stephens, Susan Kastner Tree, Robert Luo, Elizabeth Lecroy, Megan Baryenbruch, Tiffany Tzeng, Brenda E. Bolden, Ky Putnam, Kate Kindbom, Jill Medina, Jenny Saluti, Matt Ogawa, Matt Ogawa, Matthew K Ogawa, Risa Sang-urai Harms, Heather Fomin, Sarah C Murphy, Jonathan Rice, Joe Johnson, Stephanie Metruk, Holly Buttrey, Lauren Kawakami, Faithe L.A. Beadle, Allison Jacobsmeier, and Florence Hines, who provided expert ratings of personal qualities. We also thank Donald Kamentz for assistance in acquiring the data and for general advice, and to Parker Goyer for her assistance in coding the NSC data.

Funding

This research was supported by the Charles and Lynn Schusterman Family Philanthropies, the Walton Family Foundation, the Mindset Scholars Network, the Bill & Melinda Gates Foundation, the Joyce Foundation, the Overdeck Family Foundation, and the Raikes Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Author contributions

BL: conceptualization, data curation, formal analysis, methodology, project administration, resources, software, supervision, validation, writing (original draft), and writing (review & editing). MG: data curation, investigation. AQ data curation, investigation. CS: software. AR: methodology, software, validation. LU: conceptualization methodology, supervision, validation, and writing (review, editing). SH: data curation, methodology, resources. LH: methodology, writing (review & editing). SDK: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, software, visualization, writing (original draft), and writing (review & editing). ALD: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, writing (original draft), and writing (review & editing).

Competing interests

The authors declare no competing interests.

Ethics statement

This research was approved by the University of Pennsylvannia IRB.

Data and Materials Availability

Analysis files are available at https://zenodo.org/record/8250087. The raw data for this study is not available in order to protect the privacy and anonymity of the applicants, per our data use agreement with the Common Application. Please contact Brian Kim at the Common Application (bkim@commonapp.org) for questions pertaining to student application data, and Joshua Leake at the National Student Clearinghouse (leake@studentclearinghouse.org) for questions pertaining to student graduation data.

CHAPTER 5

Conclusion

People worry that a world in which artificial intelligence tools get smarter and smarter is a world in which the users of such tools get stupider and stupider. These fears are widespread. Poll after poll, op-ed after op-ed worries about how in a world where AI can do all our thinking for us, we become cognitively enfeebled, stupid, lazy, and biased. These worries are not unfounded: In reputable studies, students learning with AI tools perform worse than their counterparts learning alone (24), workers who use AI perform better, but see reductions in their intrinsic motivation (97), and when AI tools are in high stakes settings like medical diagnoses and college admissions they can set in stone the biases present in their training data (66, 68).

However, the chapters of this dissertation show that this is not inevitable, and with careful design, the opposite is possible. AI can help us learn, become more persistent when learning challenging material, and assist us in making fairer decisions at scale. In Chapter 1, participants randomly assigned to use AI to help them write cover letters improved their writing more than those who practiced on their own, those who googled examples online, and even those who received personalized feedback from professional editors. They improved more despite exerting less effort. In chapter 2, AI motivational coaches, when carefully designed helped students improve persistence and challenge seeking when learning difficult material. Finally in chapter 3, AI systems fine tuned on college admissions officers' judgments were able to reliably measure personal qualities from students essays, without introducing demographic bias, while retaining interpretability and incremental predictive validity for six-year college graduation.

Taken together, these results challenge the knee-jerk that AI tools invariably erode human learning, motivation, and decision-making. With the right design, AI tools can improve rather than hinder these capabilities.

Moreover, the studies suggest a unifying framework that helps us understand the mechanisms by which AI can enhance human capabilities. Namely, AI tools can influence the quality and quantity of engagement, and the utility of the learning task itself (e.g., by providing better information). All else being equal, more and higher quality engagement, and a higher task utility result in improved capabilities.

However, all else is not always equal. AI tools can affect these levers in opposite directions. For example, in chapter 1, the AI writing tool reduced the amount of time and the intensity of the work during that time (i.e., quantity and quality of engagement). In chapter 2, AI led to increases in learning time, but did not change the usefulness of the task itself. Whether AI functions as a coach or a crutch depends on the profile of engagement (how long and how deeply learners invest) and the yield of the task (how much that engagement builds transferable skill).

These studies prompt us to reconsider the relationship between effort and learning. The common assumption that "no pain, no gain" holds for cognitive tasks may be incomplete. While desirable difficulties are important, our findings suggest that superior information can make learning more efficient, achieving greater gains with less effort. AI's ability to reduce extraneous cognitive load allows users to focus their effort on what matters most.

5.1. Future Directions

5.1.1. Learning

While my data thus far show that AI can catalyze learning, other studies have found that AI tools are either neutral or harmful.(23, 24) In future research, I will explore the boundary conditions and moderators surrounding the benefits of AI on learning. Does AI also support the acquisition of declarative knowledge, as opposed to procedural skills? Can it be helpful in other domains, such as mathematics or computer programming? In what situations are AI tools less helpful or detrimental (e.g., learners facing severe time pressure or competing demands on their attention)? What metacognitive strategies would enhance the learning benefits of AI tools?

5.1.2. Motivation

I have found that the effects of AI use on engagement can be nuanced. Practicing writing with an AI tool reduced both objective and subjective measures of effort. In contrast, compared to using google to do research, using an AI chatbot increased engagement *without* increasing the subjective experience of effort. These findings complement prior work showing that the phenomenology of effort is dissociable from the actual effort invested in a task. In future research, I would like to explore how AI can increase objective engagement while reducing the typically aversive subjective experience of effort. Can AI tools help people enter the flow state (i.e., maximal attention to the task at hand minus the sensation that this engagement is effortful)? On the other hand, might reducing subjective effort lead to overreliance on AI, crowding out learning from other people? And to the extent that subjective effort enhances the value of our work (i.e., the Ikea effect), will AI undermine a sense of meaning and purpose?

5.1.3. Judgment and Decision Making.

In future research, I will investigate scenarios in which AI tools inadvertently cause shallow thinking, overconfidence, or overreliance–and how they might be optimized to foster even deeper reflection. For instance, what types of prompts, feedback, or conversational strategies encourage more deliberative reasoning? And can AI systems be designed to systematically challenge cognitive biases, promote actively open-minded thinking, or scaffold counterfactual reasoning?

5.2. Conclusion

The evidence suggests that smarter machines do not doom us to become stupider humans. Instead, using AI tools can enhance rather than diminish our capabilities—but only when we understand the psychological mechanisms at play. The stakes are considerable: get this right, and AI becomes a catalyst for human flourishing; get it wrong, and we risk the cognitive atrophy that critics rightly fear. The challenge, then, is to harness these remarkable technologies not as a replacement for human intellect, but as a catalyst for it. My hope is that we will be able to thread that needle, and use AI in ways that help us become our best selves.

APPENDIX A

SUPPLEMENTARY ONLINE MATERIALS FOR CHAPTER 2

A.1. Additional methods

A.1.1. AI Ratings

After participants completed the procedure outlined in Figure 2.2, 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 A.1 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 A.1: 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 A.1.

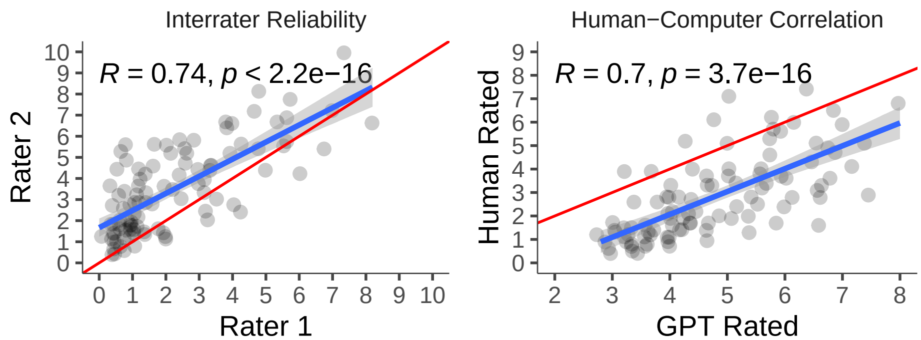

Figure A.1: Interrater correlations and correlations between AI and human ratings.

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.

A.1.2. 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.

A.1.3. 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.

The feedack prompt is shown in Figure A.2.

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.

Figure A.2: Feedback prompt

A.2. Results Study 2

A.2.1. 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 A.2, missingness and attrition rates were low for the main and followup samples, and did not differ by condition.

Table A.2: Missingness and attrition proportions and test in Study 2.

Condition	Main Sample	Followup 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 A.3.

Table A.3: 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
n	2238	755	752	731
Age (mean (SD))	36.22 (12.71)	35.99 (13.01)	36.39 (12.70)	36.29 (12.42)
Gender (%)				
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)
Black (%)	745 (33.3)	262 (34.7)	223 (29.7)	260 (35.6)
Asian (%)	134 (6.0)	49 (6.5)	42 (5.6)	43 (5.9)
Latino (%)	155 (6.9)	48 (6.4)	65 (8.6)	42 (5.7)
Other (%)	62 (2.8)	23 (3.0)	22 (2.9)	17 (2.3)
Education Level (%)				
Less than high school degree	14 (0.6)	5 (0.7)	5 (0.7)	4 (0.5)
High school graduate	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)
Motivation to improve writing (%)				
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 (%)				
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)

A.2.2. AI practice improved writing skill

The AI tool improved performance while participants used it. Table A.4 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).

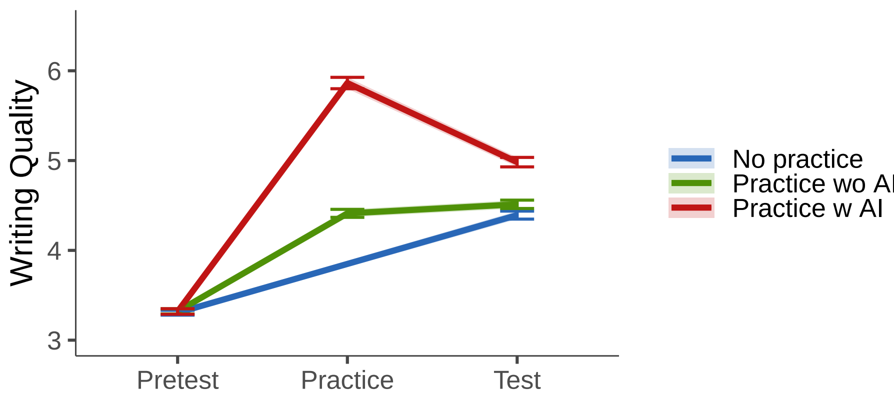

Figure A.3: Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means ± 1 SE. (N = 2,238).

Table A.4: Practice effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	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. 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. *** $p < .001$, ** $p < .01$, * $p < .05$.

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 A.5. 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 A.5: 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 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.	.09	.09	.09	.10	.09	.10*	.11*	.01	.09	.01
Practice w/o AI	(.053)	(.053)	(.052)	(.053)	(.056)	(.053)	(.053)	(.053)	(.053)	(.053)
No practice vs.	.47***	.36***	.46***	.46***	.48***	.34***	.28***	.42***	.46***	.14**
Practice w/ AI	(.054)	(.053)	(.053)	(.054)	(.057)	(.053)	(.053)	(.053)	(.054)	(.053)
Practice w/o AI vs.	.38***	.28***	.36***	.36***	.39***	.23***	.17**	.41***	.36***	.13*
Practice w/ AI	(.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. *** $p < .001$, ** $p < .01$, * $p < .05$.

A.2.3. AI practice was less effortful

Table A.6 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 A.6: 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 A.7 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 A.7: 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 A.8 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.

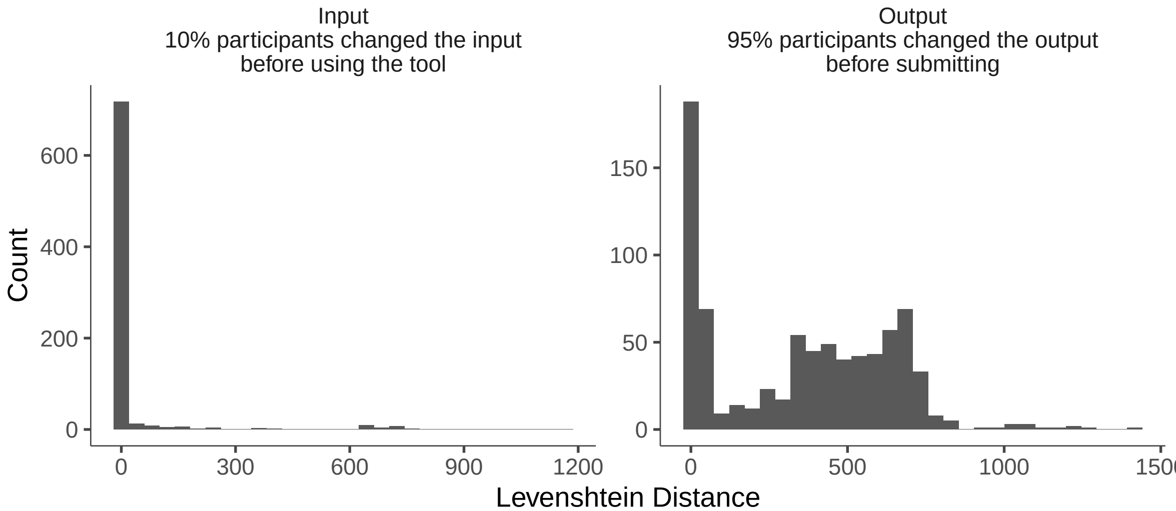

Figure A.4: Levenshtein distance (number of additions, modifications or deletions) between the original text and the text passed along to the AI tool (Input); and between the AI's output text and what users submitted as their final work (Output).

Table A.8: 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 A.4, 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.

A.2.4. AI practice did not discourage motivation for future learning

Table A.9 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 A.9: 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. *** $p < .001$, ** $p < .01$, * $p < .05$.

A.2.5. The effects of practicing with AI persist

Table A.10 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 A.10: 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 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.	.05	.06	.07	.08	.11	.08	-.02	.05	-.03
Practice w/o AI	(.056)	(.054)	(.055)	(.056)	(.056)	(.056)	(.056)	(.056)	(.056)
No practice vs.	.46***	.33***	.44***	.45***	.32***	.25***	.40***	.45***	.17**
Practice w/ AI	(.057)	(.055)	(.056)	(.057)	(.056)	(.056)	(.057)	(.057)	(.056)
Practice w/o AI vs.	.41***	.28***	.37***	.37***	.22***	.17**	.42***	.40***	.20***
Practice w/ AI	(.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. *** $p < .001$, ** $p < .01$, * $p < .05$.

A.2.6. 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 $\times$ 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 A.11: 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 F**
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**
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

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

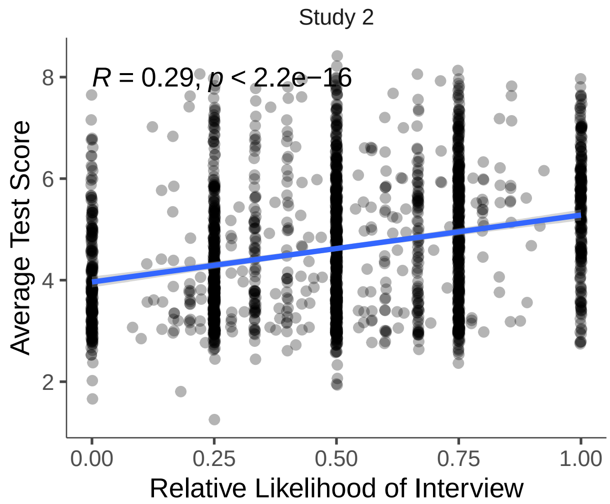

Figure A.5: Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

A.2.7. 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 A.5.

As shown in Table A.12, participants who had practiced writing cover letters with AI were more likely to be invited to a hypothetical job interview.

Table A.12: Results from beta regressions predicting the relative likelihood of an interview from test phase cover letters. The reference category is practice with AI.

	(1)	(2)
(Intercept)	0.085***	−4.816
	(0.022)	(5.017)
No practice	−0.147***	−0.250***
	(0.031)	(0.073)
Practice w/o AI	−0.096**	−0.110
	(0.031)	(0.073)
Precision (ϕ)	12.899***	
	(1.192)	
Symmetry (Log(ν))	−0.117	
	(0.086)	
Demographic and baseline performance covariates	No	Yes
Num.Obs.	2188	2153
AIC	2204.7	−10 885.8
BIC	2233.2	−10 721.3
Log.Lik.	−1097.349	
RMSE		0.28

A.3. Results Study 3

A.3.1. 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 confidence 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 A.6 and Section A.3.1 of Supplementary Information for details.

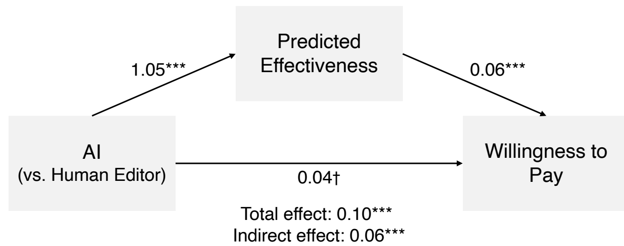

Figure A.6: Predicted effectiveness mediates underinvestment in AI feedback

A.4. Results Study 4

A.4.1. Randomization, Balance, and Missingness

Participants in Study 4 were recontacted to return two days after the initial intake section. While most people responded, X% of recontacted participants did not respond. As shown in Table A.13, missingness and attrition rates were low and did not differ by condition.

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 A.14.

Table A.13: 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

Table A.14: 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	S
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.
Gender (%)					.666	0.
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.
Black (%)	473 (15.8)	164 (16.4)	145 (14.5)	164 (16.5)	.666	0.
Asian (%)	204 (6.8)	61 (6.1)	68 (6.8)	75 (7.5)	.666	0.
Latino (%)	253 (8.4)	79 (7.9)	95 (9.5)	79 (7.9)	.666	0.
Other (%)	102 (3.4)	35 (3.5)	40 (4.0)	27 (2.7)	.666	0.
Education Level (%)					.964	0.
< high school	10 (0.3)	4 (0.4)	2 (0.2)	4 (0.4)		
High school	271 (9.0)	80 (8.0)	94 (9.4)	97 (9.7)		
Some college	508 (17.0)	178 (17.8)	168 (16.8)	162 (16.3)		
Associate	283 (9.4)	91 (9.1)	98 (9.8)	94 (9.4)		
Bachelor's	1165 (38.9)	398 (39.8)	394 (39.4)	373 (37.4)		
Master's	596 (19.9)	199 (19.9)	193 (19.3)	204 (20.5)		
Doctoral/Professional	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.
Motivation to Improve Writing (%)					.748	0.
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.
Never	516 (17.2)	188 (18.8)	173 (17.3)	155 (15.6)		
Hardly ever use them	1211 (40.4)	383 (38.3)	411 (41.1)	417 (41.9)		
A few times per week	557 (18.6)	197 (19.7)	173 (17.3)	187 (18.8)		
About once a week	503 (16.8)	158 (15.8)	178 (17.8)	167 (16.8)		
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.

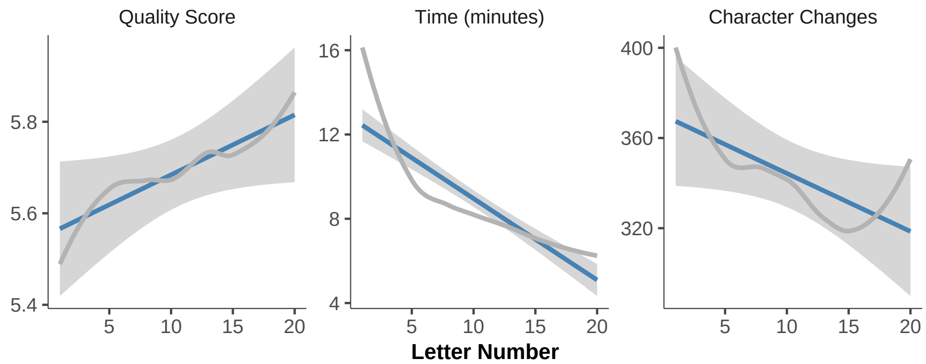

Figure A.7: Editor performance improved over the course of the task, with participants becoming more efficient at editing emails while reducing both editing time and the number of character-level changes (measured by Levenshtein distance). The line and shaded area show a linear model fit, while the gray line represents the best-fit loess curve.

A.4.2. 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.

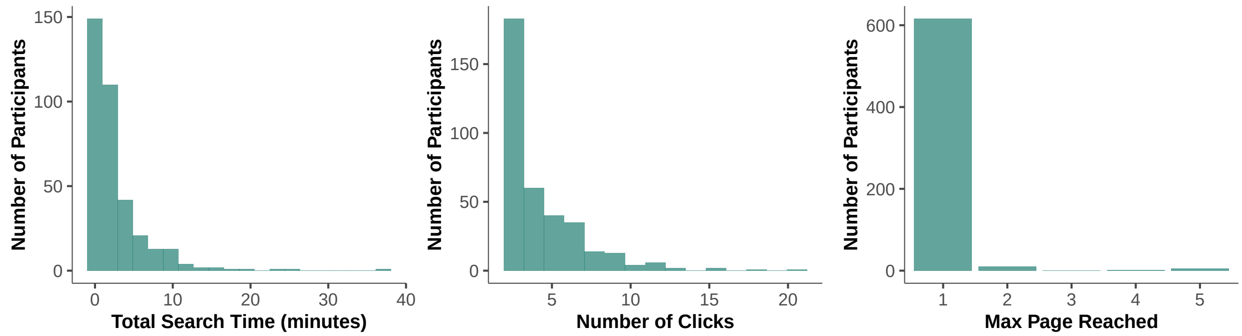

Figure A.8: Web browsing descriptive statistics.

A.4.3. 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 [CITE]. 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.

A.4.4. 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 **??** 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).

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

Table A.15: Practice effects

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

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 A.16.

Table A.16: Test effects

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

A.4.5. AI practice wasn't any more effortful than getting feedback from professional editors and was less effortful than looking for examples online

Table A.17 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 A.17: Practice effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/sec
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 A.18 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 A.18: Test effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes	Keystrokes/sec
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. *** $p < .001$, ** $p < .01$, * $p < .05$.

Table A.19: 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. *** $p < .001$, ** $p < .01$, * $p < .05$.

A.4.6. 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 A.19

A.4.7. 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 $\times$ 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 A.20

Table A.20: 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	0.99	0.990	0.805	0.997	0.919	0.930	0.990	0.99	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	0.99	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	0.99	0.923	0.314	0.402
Other	0.790	0.889	0.885	0.990	0.916	0.99	0.899	0.798	0.95	0.861	0.877	0.990	0.798	0.419
Race/Etnicity														
White	0.267	0.402	0.402	0.542	0.95	0.877	0.878	0.887	0.862	0.877	0.805	0.99	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	0.99	0.99
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	0.99	0.99
Other	0.894	0.798	0.877	0.990	0.99	0.805	0.959	0.66	0.805	0.701	0.877	0.427	0.99	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	0.14	0.166	0.990	0.99
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	0.99
Associate	0.990	0.990	0.899	0.690	0.99	0.798	0.990	0.542	0.990	0.14	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	0.14	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	0.99	0.798	0.990
Somewhat motivated	0.996	0.957	0.605	0.805	0.732	0.854	0.278	0.542	0.838	0.99	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	0.69	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

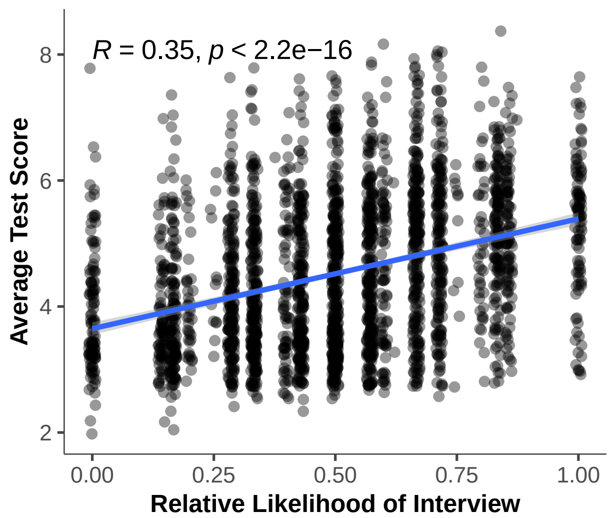

Figure A.9: Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

A.4.8. 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 A.9.

Table A.21 shows the beta regressions for the relative likelihood by condition.

Table A.21: Results from beta regressions predicting the relative likelihood of an interview from test phase cover letters. The reference category is practice with AI.

	(1)	(2)
(Intercept)	0.501***	−0.859
	(0.008)	(0.710)
Practice with Google Search	−0.020*	−0.023*
	(0.012)	(0.012)
Practice w/ Editor Feedback	0.015	0.012
	(0.012)	(0.012)
Observations	2,575	2,513
R^2	0.004	0.014
Adjusted R^2	0.003	0.010
Residual Std. Error	0.241 (df = 2572)	0.240 (df = 2501)
F Statistic	4.651*** (df = 2; 2572)	3.267*** (df = 11; 2501)
Note:		*p<0.1; **p<0.05; ***p<0.01

A.5. Results Study 5

A.5.1. 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 A.22, missingness and attrition rates were low for the main and follow-up samples and did not differ by condition.

Table A.22: Missingness and attrition proportions and test in Study 5.

Condition	Main Sample	Followup 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 A.23.

Table A.23: 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

A.5.2. AI examples improve writing skill

The AI tool improved performance while participants used it. Table A.24 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).

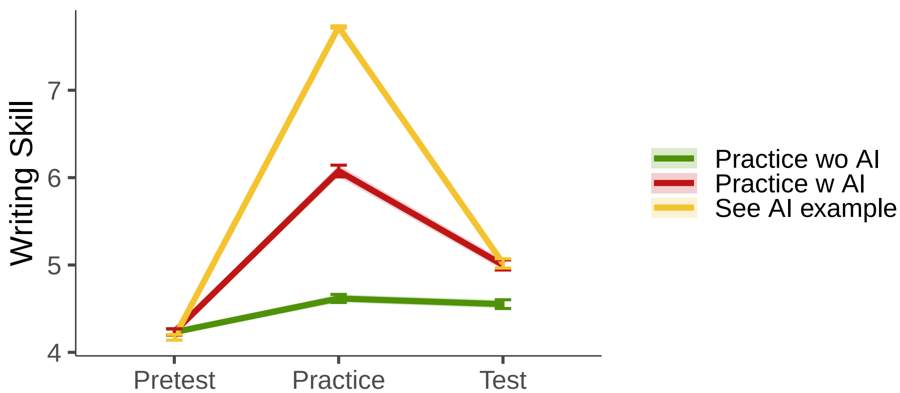

Figure A.10: Participants who had practiced with the AI tool outperformed those who had practiced without it and those who had not practiced at all. Error bars represent means ± 1 SE. (N = 2,003).

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 A.25. 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 A.24: Practice effects

	GPT-4o	Claude	Ex. Controls	Ex. Cheaters	Followup	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 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 A.25: Test effects

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

A.5.3. Seeing AI examples was less effortful

Table A.26 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 A.26: Practice effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
Practice w/o 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 w/o AI vs. Practice w/ AI	-.14*	-.55***	-.14*	-.07	-.10
	(.056)	(.056)	(.057)	(.056)	(.056)
Practice w/o 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 A.27 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 A.28 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 A.27: Test effort differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
Practice w/o 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 w/o AI vs. Practice w/ AI	.09	.06	.07	.10	.10
	(.057)	(.056)	(.057)	(.057)	(.056)
Practice w/o 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 A.28: Learning rate differences

	sqrt(Time)	log(Keystrokes)	Subjective Rating (0 - 10)	Time	Keystrokes
Means — (SE)					
Practice w/o 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 w/o AI vs. Practice w/ AI	.21***	.08	.25***	.15**	.27***
	(.057)	(.057)	(.057)	(.057)	(.057)
Practice w/o 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. *** $p < .001$, ** $p < .01$, * $p < .05$.

A.5.4. Seeing an AI example did not discourage motivation for future learning

Table A.29 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 A.29: Motivation

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

A.5.5. The effects of seeing an AI example persist

Table A.30 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 A.30: Followup effects

	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 A.31 are the results for each of these samples separately.

Table A.31: Follow-up effects 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 pre-treatment 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, ** $p < .01$, * $p < .05$.

A.5.6. 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 A.32: 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
Year of birth	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.

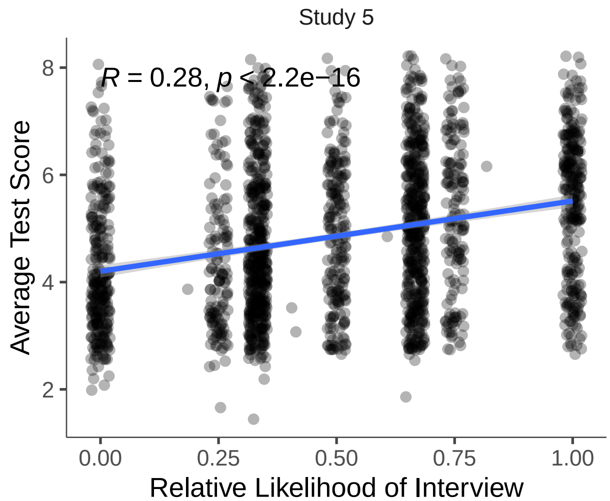

Figure A.11: Correlation between test score writing quality as rated by GPT-4o and relative likelihood of being offered a hypothetical interview.

A.5.7. Pairwise Compariosns

The relative likelihood of a cover letter receiving an invitation to an interview was correlated with the GPT-rated writing quality. See Figure A.11.

As shown in Table A.33, participants who had practiced writing cover letters with AI were more likely to be invited to a hypothetical job interview.

Table A.33: 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.

	(3)	(4)
(Intercept)	-0.061^{**}	-0.449
	(0.021)	(0.677)
Practice w/o AI		-0.103
		(0.079)
Practice w/ AI	0.080^{**}	
	(0.030)	
See AI example	0.097^{***}	0.055
	(0.030)	(0.078)
Precision (ϕ)	16.174^{***}	
	(1.471)	
Symmetry ($\log(\nu)$)	0.350^{***}	
	(0.080)	
Demographic and baseline covariates	No	Yes
Num.Obs.	1934	1917
AIC	2447.1	$-14\,615.7$
BIC	2474.9	$-14\,454.5$
Log.Lik.	-1218.535	
RMSE		0.30

[1] Todd Rogers and Jessica Lasky-Fink. *Writing for Busy Readers: Communicate More Effectively in the Real World*. Penguin, September 2023. ISBN 978-0-593-18748-7.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, April 2023.

[3] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381:187–192, 2023.

[4] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim 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. ISSN 1556-5068. doi: 10.2139/ssrn.4573321.

[5] Emma Wiles, Zanele T Munyikwa, and John J Horton. Algorithmic writing assistance on jobseekers' resumes increases hires. Technical report, National Bureau of Economic Research, 2023.

[6] Jake M Hofman, Daniel G Goldstein, and David M Rothschild. A sports analogy for understanding different ways to use ai. *Harvard Business Review*, 4, 2023.

[7] Stefano Puntoni, Rebecca Walker Reczek, Markus Giesler, and Simona Botti. Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, 85(1):131–151, 2021. doi: 10.1177/0022242920953847.

[8] Beatriz Borges, Negar Foroutan, Deniz Bayazit, Anna Sotnikova, Syrielle Montariol, Tanya Nazaretzky, Mohammadreza Banaei, Alireza Sakhaeirad, Philippe Servant, Seyed Parsa Neshaei, Jibril Frej, Angelika Romanou, Gail Weiss, Sepideh Mamooler, Zeming Chen, Simin Fan, Silin Gao, Mete Ismayilzada, Debjit Paul, Philippe Schwaller, Sacha Friedli, Patrick Jermann, Tanja Käser, Antoine Bosselut, EPFL Grader Consortium, and EPFL Data Consortium. Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences*, 121(49):e2414955121, December 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2414955121.

[9] Stephen Hawkins, Fred Duong, Ashley Fabrizio, and Daniel Yudkin. Between hesitation & hope: America's mixed feelings on generative artificial intelligence, October 2024. Available online: https://www.canva.com/design/DAGRVonfEGk/W6C6mAYwwJmRhS2dqjjfVg/view#1.

[10] Kevin Roose. Don't Ban ChatGPT in Schools. Teach With It. *New York Times*, January 2023.

[11] David 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/, May 2023.

[12] Luona Lin. A quarter of U.S. teachers say AI tools do more harm than good in K-12 education, May 2024.

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[14] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*, 2023.

[15] Matthew Fisher, Mariel K. Goddu, and Frank C. Keil. Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, 144(3):674–687, June 2015. ISSN 1939-2222, 0096-3445. doi: 10.1037/xge0000070.

[16] Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science*, 333(6043):776–778, August 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1207745.

[17] Eleanor A. Maguire, David G. Gadian, Ingrid S. Johnsrude, Catriona D. Good, John Ashburner, Richard S. J. Frackowiak, and Christopher D. Frith. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403, April 2000. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.070039597.

[18] Eva-Maria Griesbauer, Ed Manley, Jan M Wiener, and Hugo J Spiers. London taxi drivers: A review of neurocognitive studies and an exploration of how they build their cognitive map of london. *Hippocampus*, 32(1):3–20, 2022.

[19] John Sweller and Graham A Cooper. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction*, 2(1):59–89, 1985.

[20] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. Learning from examples: Instructional principles from the worked examples research. *Review of educational research*, 70(2):181–214, 2000.

[21] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.

[22] Harsh Kumar, David M. Rothschild, Daniel G. Goldstein, and Jake Hofman. Math Education with Large Language Models: Peril or Promise? *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4641653.

[23] Matthias Lehmann, Philipp B Cornelius, and Fabian J Sting. Ai meets the classroom: When does chatgpt harm learning? *arXiv preprint arXiv:2409.09047*, 2024.

[24] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI Can Harm Learning, 2024.

[25] Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. The gpt surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances. Technical report, Center for Open Science, 2024.

[26] Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.

[27] Hillary C. Shulman, David M. Markowitz, and Todd Rogers. Reading dies in complexity: Online news consumers prefer simple writing. *Science Advances*, 10(23):eadn2555, June 2024. ISSN 2375-2548. doi: 10.1126/sciadv.adn2555.

[28] Albert Bandura. *Social Learning Theory.* General Learning Press, New York, 1971.

[29] Andrew N. Meltzoff. Imitation and other minds: The like me hypothesis. In Susan Hurley and Nick Chater, editors, *Perspectives on Imitation: From Neuroscience to Social Science, Volume 2: Imitation, Human Development, and Culture*, pages 55–77. MIT Press, Cambridge, MA, 2005.

[30] Derek E Lyons, Andrew G Young, and Frank C Keil. The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, 104(50):19751–19756, 2007.

[31] A K Ericsson. Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine*, 15(11):988–994, 2008.

[32] Minkyu Shin, Jin Kim, Bas Van Opheusden, and Thomas L. Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, March 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2214840120.

[33] Charlie Rose : WHUT : June 14, 2010 9:00am-10:00am EDT, June 2010.

[34] Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121, August 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2308950121.

[35] Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, 8:

1272229, December 2023. ISSN 2504-284X. doi: 10.3389/feduc.2023.1272229.

[36] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[37] Mark R Lepper, Maria Woolverton, Donna L Mumme, and Jean-Luc Gurtner. Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. *Computers as cognitive tools*, pages 75–105, 1993.

[38] Justin Reich and José A Ruipérez-Valiente. The failure of online education. *Science*, 363(6423): 130–131, 2019.

[39] Katherine L Milkman, John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20):e2101165118, 2021.

[40] James J Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998.

[41] David S Yeager, Gregory M Walton, Shannon T Brady, Emirhan N Akcinar, David Paunesku, Laura Keane, Donald Kamentz, Gretchen Ritter, Angela L Duckworth, Rachel Urstein, et al. Teaching a lay theory before college narrows achievement gaps at scale. *Proceedings of the National Academy of Sciences*, 113(24):E3341–E3348, 2016.

[42] Michael D Mrazek, Eden D Ihm, Daniel C Molden, Alissa J Mrazek, Kimberly L Mrazek, and Jonathan W Schooler. Expanding minds: Growth mindsets of self-regulation and the influences on effort and perseverance. *Journal of Experimental Social Psychology*, 79:164–180, 2018.

[43] Dariya Ovsyannikova, Victoria Oldemburgo de Mello, and Michael Inzlicht. Third-party evaluators perceive ai as more compassionate than expert humans. *Communications Psychology*, 3 (1):4, 2025.

[44] Timothy W Bickmore and Rosalind W Picard. Relational agents: a model and implementation of building user trust. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403, 2005.

[45] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mHealth and uHealth*, 5 (6):e7785, 2017.

[46] Albert Bandura. *Social learning theory.* Prentice-Hall, Englewood Cliffs, NJ, 1977.

[47] Patricia K Kuhl, Feng-Ming Tsao, and Huei-Mei Liu. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 104(21):9096–9101, 2003.

[48] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. Chatbots and mental health: Insights into the safety of generative ai. *Journal of Consumer Psychology*, 34(3):481–491, 2024.

[49] National Research Council. *Assessing 21st Century Skills: Summary of a Workshop.* The National Academies Press, Washington DC, 2011. URL https://doi.org/10.17226/13215.

[50] T. E. Moffitt, L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, R. Houts, R. Poulton, B. W. Roberts, S. Ross, M. R. Sears, W. M. Thomson, and A. Caspi. A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7):2693–2698, February 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1010076108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1010076108.

[51] Mathilde Almlund, Angela Lee Duckworth, James Heckman, and Tim Kautz. Personality Psychology and Economics. In *Handbook of the Economics of Education*, volume 4, pages 1–181. Elsevier, 2011. ISBN 978-0-444-53444-6. doi: 10.1016/B978-0-444-53444-6.00001-8. URL https://linkinghub.elsevier.com/retrieve/pii/B9780444534446000018.

[52] Steven B. Robbins, Kristy Lauver, Huy Le, Daniel Davis, Ronelle Langley, and Aaron Carlstrom. Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin*, 130(2):261–288, 2004. ISSN 1939-1455, 0033-2909. doi: 10.1037/0033-2909.130.2.261. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.130.2.261.

[53] Patrick C. Kyllonen, Anastasiya A. Lipnevich, Jeremy Burrus, and Richard D. Roberts. Personality, Motivation, and College Readiness: A Prospectus for Assessment and Development: Personality, Motivation, and College Readiness. *ETS Research Report Series*, 2014(1):1–48, June 2014. ISSN 23308516. doi: 10.1002/ets2.12004. URL https://onlinelibrary.wiley.com/doi/10.1002/ets2.12004.

[54] Arthur L Coleman and Jamie Lewis Keith. Understanding Holistic Review in Higher Education Admissions. Technical report, College Board, New York, 2018.

[55] Jerome Karabel. *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton.* Houghton Mifflin Harcourt, 2005.

[56] Michael N. Bastedo, Nicholas A. Bowman, Kristen M. Glasener, and Jandi L. Kelly. What are We Talking About When We Talk About Holistic Review? Selective College Admissions and its Effects on Low-SES Students. *The Journal of Higher Education*, 89(5):782–805, September 2018. ISSN 0022-1546, 1538-4640. doi: 10.1080/00221546.2018.1442633. URL https://www.tandfonline.com/doi/full/10.1080/00221546.2018.1442633.

[57] Ruth Starkman. Confessions of an Application Reader. *The New York Times*, August 2013. ISSN 0362-4331. URL https://www.nytimes.com/2013/08/04/education/edlife/lifting-the-veil-on-the-holistic-process-at-the-university-of-california-berkeley.html.

[58] Trisha Ross Anderson and Richard Weissbourd. Character Assessment in College Admission. Technical report, Making Caring Common Project, Boston, 2020.

[59] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. *Noise: A Flaw in Human Judgment*. Harper Collins, 2021.

[60] J Philippe Rushton, Charles J Brainerd, and Michael Pressley. Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, 94(1):18, 1983.

[61] Eric Hoover. Working smarter, not harder, in admissions. *The Chronicle of Higher Education*, March 2017. URL https://www.chronicle.com/article/working-smarter-not-harder-in-admissions/.

[62] Melissa Korn. Some elite colleges review an application in 8 minutes (or less). *Wall Street Journal*, January 2018. ISSN 0099-9660. URL https://www.wsj.com/articles/some-elite-colleges-review-an-application-in-8-minutes-or-less-1517400001.

[63] Louis Tay, Sang Eun Woo, Louis Hickman, Brandon M Booth, and Sidney D'Mello. A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science*, 5 (1):1–30, 2022.

[64] Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107:1323–1351, 2022. ISSN 1939-1854. doi: 10.1037/apl0000695. Place: US Publisher: American Psychological Association.

[65] James Manyika, Jake Silberg, and Brittany Presten. What do we do about the biases in AI? *Harvard Business Review*, pages 1–5, 2019.

[66] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

[67] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research*, 81:1–12, 2018.

[68] Aj Alvero, Sonia Giebel, Ben Gebre-Medhin, a. l. antonio, M. L. Stevens, and Benjamin W. Domingue. Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science Advances*, 7(42):eabi9031, October 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abi9031. URL https://www.science.org/doi/10.

1126/sciadv.abi9031.

[69] Mark O. Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, January 2019. ISSN 2578-1863, 2578-1863. doi: 10.1002/hbe2.117. URL https://onlinelibrary.wiley.com/doi/10.1002/hbe2.117.

[70] Ben Shneiderman. Human-Centered Artificial Intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020. ISSN 19443900. doi: 10.17705/1thci.00131. URL http://aisel.aisnet.org/thci/vol12/iss3/1.

[71] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, December 2019. ISSN 2470-9476. doi: 10.1126/scirobotics.aay7120. URL https://www.science.org/doi/10.1126/scirobotics.aay7120.

[72] Our Selection Process, 2021. URL https://admission.stanford.edu/apply/selection/.

[73] Orly Lobel. *The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future*. Public Affairs, 2022.

[74] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

[75] Charles Pierse. Transformers Interpret, February 2021. URL https://github.com/cdpierse/transformers-interpret. original-date: 2020-05-27T20:32:08Z.

[76] Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks, June 2020. URL http://arxiv.org/abs/2002.04138.

[77] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, August 2001. ISSN 0090-5364. doi: 10.1214/aos/1013699998. URL https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-4/The-control-of-the-false-discovery-rate-in-multiple-testing/10.1214/aos/1013699998.full.

[78] Linda Kamas and Anne Preston. Empathy, gender, and prosocial behavior. *Journal of Behavioral and Experimental Economics*, 92:101654, June 2021. ISSN 22148043. doi: 10.1016/j.socec.2020.101654. URL https://linkinghub.elsevier.com/retrieve/pii/S2214804320306972.

[79] James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. When small words foretell academic auccess: The case of college admissions essays. *PLoS ONE*, 9(12):e115844, December 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0115844. URL https://dx.plos.org/10.1371/journal.pone.0115844.

[80] J. Parker Goyer, Gregory M. Walton, and David S. Yeager. The role of psychological factors and institutional channels in predicting the attainment of postsecondary goals. *Developmental Psychology*, 57:73–86, 2021. ISSN 1939-0599. doi: 10.1037/dev0001142. Place: US Publisher: American Psychological Association.

[81] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15): 8398–8403, April 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915006117. URL https://pnas.org/doi/full/10.1073/pnas.1915006117.

[82] Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. Exploring Limits to Prediction in Complex Social Systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694, Montréal Québec Canada, April 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883001. URL https://dl.acm.org/doi/10.1145/2872427.2883001.

[83] Stephan Ahadi and Edward Diener. Multiple determinants and effect size. *Journal of Personality and Social Psychology*, 56(3):398–406, 1989. ISSN 1939-1315. doi: 10.1037/0022-3514.56.3.398.

[84] Friedrich M Götz, Samuel D Gosling, and Peter J Rentfrow. Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*, 17 (1):205–215, 2022.

[85] Sara Goldrick-Rab. Following their every move: An investigation of social-class differences in

college pathways. *Sociology of Education*, 79(1):67–79, January 2006. ISSN 0038-0407, 1939-8573. doi: 10.1177/003804070607900104. URL http://journals.sagepub.com/doi/10.1177/003804070607900104.

[86] Daniel Hepworth, Ben Littlepage, and Katy Hancock. Factors influencing university student academic success. *Educational Research Quarterly*, 42(1):45–61, 2018.

[87] Sameano F. Porchea, Jeff Allen, Steve Robbins, and Richard P. Phelps. Predictors of long-term enrollment and degree outcomes for community college students: Integrating academic, psychosocial, socio-demographic, and situational factors. *The Journal of Higher Education*, 81(6):680–708, November 2010. ISSN 0022-1546, 1538-4640. doi: 10.1080/00221546.2010.11779077. URL https://www.tandfonline.com/doi/full/10.1080/00221546.2010.11779077.

[88] Mary C. Murphy, Maithreyi Gopalan, Evelyn R. Carter, Katherine T. U. Emerson, Bette L. Bottoms, and Gregory M. Walton. A customized belonging intervention improves retention of socially disadvantaged students at a broad-access university. *Science Advances*, 6(29):eaba4677, July 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aba4677. URL https://www.science.org/doi/10.1126/sciadv.aba4677.

[89] Warren W Willingham. *Success in College: The Role of Personal Qualities and Academic Ability*. College Board Publications, 1985.

[90] Jeremy D. W. Clifton. Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, 25(3):259, 2020. ISSN 1939-1463. doi: 10.1037/met0000236. URL https://psycnet.apa.org/fulltext/2019-47148-001.pdf. Publisher: US: American Psychological Association.

[91] Donald T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, January 1979. ISSN 01497189. doi: 10.1016/0149-7189(79)90048-X. URL https://linkinghub.elsevier.com/retrieve/pii/014971897990048X.

[92] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[93] Joseph L Mahoney, Reed W Larson, Jacquelynne S Eccles, and Heather Lord. Organized activities as developmental contexts for children and adolescents. *Organized activities as contexts of development: Extracurricular activities, after-school and community programs*, pages 3–22, 2005.

[94] Reed W. Larson. Toward a psychology of positive youth development. *American Psychologist*, 55(1):170, 2000. ISSN 1935-990X. doi: 10.1037/0003-066X.55.1.170. URL http://psycnet.apa.org/fulltext/2000-13324-016.pdf. Publisher: US: American Psychological Association.

[95] ACT. ACT–SAT concordance: A tool for comparing scores, 2013. URL https://www.act.org/

content/dam/act/unsecured/documents/reference.pdf.

[96] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL http://www.jstatsoft.org/v45/i03/.

[97] Suqing Wu, Yukun Liu, Mengqi Ruan, Siyu Chen, and Xiao-Yun Xie. Human-generative ai collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports*, 15(1):15105, 2025.