Multitasking with Generative AI: Experimental Evidence *

Zhuoqiong Chen† Daojing He‡ Jifan Ren§

February 6, 2026

Abstract

This study investigates how generative AI affects effort allocation in multitasking environments with asymmetric incentives. In a lab-in-the-field experiment, employees completed an incentivized writing task that could be supported by generative AI alongside a weakly incentivized calculation task that could not. Without AI, stronger incentives for writing slightly reduce calculation accuracy, consistent with classic multitasking distortions. With AI available, access improves writing quality and efficiency without reducing calculation performance. Moreover, stronger incentives for writing no longer change writing quality but generate spillovers to calculation accuracy, especially for participants without prior experience in similar writing tasks. This pattern is consistent with stronger incentives increasing reliance on AI for the primary task, compressing its cognitive demands and freeing attention for secondary work. The results inform incentive design and human-AI collaboration in AI-enabled workplaces, suggesting that firms can strengthen incentives on AI-assisted outputs without sacrificing, and sometimes improving, performance on harder-to-measure tasks.

Keywords: generative artificial intelligence, multitasking, experiment, spillover effect

*We gratefully acknowledge financial support from the Shenzhen Science and Technology Program (No. RKX20231110090859012). We thank seminar subjects at the School of Humanities and Social Sciences, Nanyang Technological University, and subjects at the Dongbei University of Finance and Economics Behavioral Economics Workshop for helpful comments. Yuanxi Song provided excellent research assistance.

†Harbin Institute of Technology, Shenzhen, China. E-mail: chenzq926@hit.edu.cn.

‡Harbin Institute of Technology, Shenzhen, China. E-mail: hedaojinghit@163.com.

§Harbin Institute of Technology, Shenzhen, China. E-mail: renjifan@hit.edu.cn.

1. Introduction

In many organizations, jobs are bundles of heterogeneous, often complementary tasks rather than single, isolated activities. A manager drafts strategy documents while also resolving conflicts and motivating subordinates; a bank employee writes credit reports while simultaneously calling clients and coordinating with risk officers; an engineer prepares technical documentation while also troubleshooting problems on site and aligning with other teams. Generative AI can now be used as a tool for some of these tasks — such as drafting text, restructuring reports, or retrieving information — but offers little direct help for others, including coordination, supervision, and judgment.[1] How this partial assistability reshapes behaviour within jobs remains unclear: when only a subset of tasks can be supported by AI, how do workers reallocate their limited attention across assisted and unassisted tasks, and what happens to the performance of the parts of work that AI cannot do?

In such multitasking environments, one theoretical perspective comes from multitasking theory. Workers must allocate scarce attention across tasks that differ in measurability and rewards: measurable outputs (e.g., scored reports, code quality) typically attract stronger, "high-powered" incentives, while diffuse contributions (coordination, judgment, or creativity) receive weaker or implicit rewards.[2] Classic multitasking theory predicts that such asymmetric incentives distort effort toward the strongly incentivized, easily measured task at the expense of its less measurable complements (Holmström and Milgrom, 1991), a tendency exacerbated by limited cognitive resources and switching costs in multitask environments (Alonso et al., 2014; Monsell, 2003). If the primary task in a job is both highly measurable and the main target of high-powered incentives, then making that task AI-assistable should, in theory, further shift attention toward it and exacerbate neglect of the complementary tasks that remain entirely human.

A second perspective comes from experiments on human–AI collaboration and algorithmic advice, which show that stronger or more salient incentives generally increase reliance on AI support. Greiner et al. (2025) show that moving from fixed pay to individual performance pay or tournaments raises the extent to which decision makers follow an AI-based forecasting algorithm. Bockstedt and Buckman (2025) find that loss-framed incentives increase willingness to delegate decisions to an AI assistant. Saragih and Morrison (2022) report that when an algorithm has a strong performance history, higher decision stakes lead participants to accept its advice more readily, and Neumann et al. (2022) find that incentives increase the use of simple mechanical decision rules in human-performance prediction. Taken together, these studies suggest that higher-powered or psychologically salient incentives often tilt decision makers toward greater reliance on

[1] A dominant approach to assessing the labour-market impact of generative AI is to measure "task exposure": mapping large language models' capabilities onto occupational task taxonomies and inferring impact from the share of tasks in a job that AI can assist (Eloundou et al., 2024; Melina et al., 2024; Gmyrek et al., 2025). This task-based perspective is informative about which parts of jobs are directly exposed to AI, but largely silent on within-job behavioural responses—how workers reallocate effort across tasks, and how incentives and cognition interact once only some tasks can draw on AI.

[2] In multitask settings, diffuse contributions refer to activities that are difficult to measure directly yet compete for an individual's limited attention—such as coordination, communication, creativity, and judgment. When these diffuse activities are important or hard to observe, high-powered incentives tied to measurable outputs are more likely to distort effort allocation across tasks (Holmström and Milgrom, 1991). As a result, firms often rely on lower-powered incentives, task or job bundling, and formal rules or relational governance to manage performance (Garicano and Rayo, 2016; Rayo, 2007). This emphasis on cross-task spillovers from incentive schemes is also central to theories of contracting and the organization of knowledge in which rewarding one dimension of performance reshapes behavior in other interdependent activities (Fuchs et al., 2015).

algorithmic or AI tools for the incentivized task.

These two strands of theory thus make competing predictions for GenAI-assisted multitask jobs. From the multitasking perspective, stronger incentives on an AI-assistable, measurable primary task should exacerbate effort diversion toward that task and worsen underinvestment in harder-to-measure complements. From the incentive-reliance perspective, the same stronger incentives may induce workers — especially those who feel relatively weak at the primary task — to offload more of it onto AI, compressing its cognitive demands and freeing scarce attention for secondary tasks in the presence of limited processing capacity and switching costs. This tension leads to our central question: when workers can use generative AI on a strongly incentivized primary task in a realistic multitask setting, does AI adoption exacerbate or mitigate the classic multitasking distortion on complementary, weakly incentivized tasks?

We answer this question causally in a lab-in-the-field experiment with 175 employees across three companies based in Shenzhen, China (Qingchuang Zhiyuan, ICBC, and CESEC).[3] Subjects complete three 15-minute professional writing assignments; in Stages 1 and 2, writing is interrupted at minute 6 by a mandatory calculation task, while Stage 3 proceeds without interference (a single-task benchmark). Treatments vary between subjects: NO-AI (no access to AI in any stage), AI-Low (ChatGPT-3.5 access via a custom platform in Stages 2–3), and AI-High (same access as AI-Low but with doubled writing incentives). All groups were explicitly forbidden from AI use in Stage 1, which served as an additional benchmark. Writing is measurable and strongly incentivized through relative performance pay (the top 50% earn CNY 100 in NO-AI/AI-Low; CNY 200 in AI-High) and is graded double-blind by independent professionals using a prespecified rubric. The calculation task is perfectly measurable in the lab (binary correctness) but assigned weak incentives—a small CNY 5 penalty for errors. This mirrors organizational reality: many essential activities are observable ex post yet hard to reward at high power because they are diffuse and costly to attribute. To ensure that it is non-assistable by AI and to emphasize listening/comprehension over arithmetic, the instructions for the calculation tasks are presented and explained by an experimenter using a projector. The answer is submitted on paper on which the subject's ID is also written. Following Holmström and Milgrom (1991), we implement the incentive asymmetry induced by measurability through the pay structure, rather than by making the secondary task noisy or unmeasurable.[4]

Our findings unfold in three steps. First, in Stage 1 without any AI assistance, the stronger writing incentives in the AI-High group reduce performance on the concurrent calculation task. This is exactly in line with multitasking theory: high-powered incentives on one measurable task pull attention away from

[3]Shenzhen Qingchuang Zhiyuan Industrial Group Co., Ltd. (also known as "Meta Youth City": https://www.hxqcy.com/) develops and operates smart industrial parks in Shenzhen, staffed by corporate managers, technical staff, property-operations teams, customer-facing service personnel, and a marketing team. The Shenzhen branch of the Industrial and Commercial Bank of China (ICBC) provides corporate and retail banking and is staffed by relationship managers, risk and compliance personnel, operations support, technology/data teams, and customer service staff. CESEC, short for CESI (Shenzhen) Electronic Information Products Standardization Engineering Center Co., Ltd., is the Shenzhen branch of the National Engineering Research Center of Standardization of Electronic Information Products (https://www.cesi.cn/nercfsoeip/page/) and is a Shenzhen-based engineering center dedicated to standardization and conformity assessment for electronic information products, supporting national and industry standard development, testing and certification services, and industry-wide standardization implementation.

[4]Gibbons (1998) further emphasizes that it is the structure of incentives, rather than measurability per se, that drives such distortions. Friebel and Raith (2004) extend this view by showing that even when pay is tied to less observable tasks, imbalanced incentives can still shift effort inefficiently.

the secondary task and worsen its outcomes. Second, in Stage 2, where GenAI is available for writing in both AI-Low and AI-High, AI access substantially raises writing quality relative to NO-AI without harming calculation accuracy. Conditional on using AI, however, stronger writing incentives no longer improve writing quality (AI-High vs. AI-Low), but they do lead to a marked increase in calculation accuracy. This pattern is consistent with the incentive-reliance mechanism: under higher-powered incentives, participants appear to rely more heavily on GenAI for the incentivized task, freeing cognitive resources for the non-assistable calculation task. Heterogeneity results support this interpretation: among subjects without similar prior writing experience, who are plausibly more inclined to lean on AI in an unfamiliar task environment, the positive effect of stronger incentives on calculation performance is especially pronounced. This aligns with recent evidence that physicians are more receptive to AI recommendations when prescribing in unfamiliar domains (Wang et al., 2025). Third, in the Stage-3 single-task benchmark, generative AI again improves writing quality relative to NO-AI, but stronger incentives in AI-High still do not generate additional gains over AI-Low once AI is available. This suggests that, in our short-horizon writing task, variation in incentives and AI reliance has limited room to affect final performance in a single-task setting; the main impact of stronger incentives operates through AI reliance and cognitive reallocation in the multitask environment of Stage 2.

We contribute to the literature on the productivity impact of generative AI and related discussions in several aspects. First, to our knowledge, this is the first experimental study that explicitly embeds generative AI in a multitasking framework and asks whether AI exacerbates or alleviates the classic multitasking distortion between a strongly incentivized, measurable task and a complementary, weakly incentivized task. Second, we provide causal evidence on how GenAI interacts with asymmetric incentives in multitasking. Mechanistically, AI lowers the cognitive cost of producing measurable outputs; behaviourally, stronger incentives increase reliance on AI for the primary task and thereby generate a positive spillover to the non-assistable secondary task, a pattern that is new to the experimental literature on generative AI and algorithmic advice.

These findings clarify when incentives and AI act as complements versus substitutes. In our multitask environment, high-powered incentives complement GenAI by encouraging greater AI reliance and freeing cognitive resources for complementary work, whereas in the single-task benchmark AI largely determines performance and additional incentives have little room to improve quality once AI is available, so that incentives and AI effectively operate on the same margin. Organizationally, this informs incentive design and task allocation in AI-integrated firms: where AI meaningfully reduces primary-task load, firms can strengthen incentives on that task without neglecting complementary, harder-to-reward activities — and may even improve them via freed capacity (Gambacorta et al., 2024; Brynjolfsson and Li, 2024). The spillover pattern we document is also consistent with evidence that GenAI disproportionately benefits less-experienced workers and with cognitive accounts of reduced switch costs when one component task becomes less demanding (Brynjolfsson et al., 2025; Monsell, 2003).

Literature review. Although prior studies have examined the impact of artificial intelligence on writing tasks and its expanding application to more complex domains such as coding, none have yet addressed AI's

role in multitasking environments where certain tasks lie beyond the scope of AI assistance. Micro-level evidence from laboratory and field experiments has consistently demonstrated that generative AI enhances productivity—particularly for lower-ability workers—across diverse domains such as writing, software development, and customer service, with heterogeneous effects varying by size, including faster writing (Noy and Zhang, 2023), faster coding task completion (Peng et al., 2023; Cui et al., 2025; Gambacorta et al., 2024), significant gains among inexperienced customer service agents (Brynjolfsson et al., 2025), and boosts for less experienced users in coding (Kreitmeir and Raschky, 2024) and creative tasks (Chen et al., 2025; Doshi and Hauser, 2024; Zhou and Lee, 2024; Zou and Zhu, 2025; Dell'Acqua et al., 2023; Talypova et al., 2024; Uhunoma, 2025; Bonin et al., 2025; Hoffmann et al., 2024; Bohren et al., 2024; Ju and Aral, 2025); recent literature further highlights cases where effects turn negative for experienced or high-ability individuals on routine (Fang et al., 2025) or high-creativity tasks (Nakavachara et al., 2024; Huang et al., 2024; Hou et al., 2025), as well as risks of productivity losses due to role shifts (Eloundou et al., 2024), workflow disruptions (Vaccaro et al., 2024), and task difficulty amplification (Simkute et al., 2024).[5] However, a critical gap persists in translating these insights from single-tasking studies to real-world workplaces, where multitasking is inherent to many roles. Hence, our results illuminate the role of generative AI in such settings, where only a subset of tasks can be directly assisted: the productivity gains in AI-supported tasks may spill over to those that cannot benefit from AI.

Although related, our study is more closely connected to the multitasking literature than the economics literature on the effect of interruptions on productivity. The latter typically emphasizes unexpected interruptions and their adverse effects on productivity.[6] In contrast, our experiment explicitly informed subjects at the beginning of each stage that they would be interrupted during the writing task by a secondary calculation task at a pre-specified time, and they could decide which task to work on after finishing listening to the calculation instructions. This design enables subjects to make forward-looking decisions about how to allocate cognitive resources—such as attention—across tasks, a perspective consistent with models in which attention is optimally allocated in anticipation of payoffs and informational value (Tasoff and Madarász, 2009).

A rich body of literature has demonstrated that incentivizing measurable tasks can crowd out effort on less measurable or non-incentivized ones, especially when agents face uncertainty about task performance or evaluation criteria. For example, Fryer and Holden (2012) show that incentives may reduce effort on non-incentivized tasks by revealing ability gaps and lowering motivation. Other studies emphasize that multitasking choices are shaped by factors such as subjective performance evaluations, cognitive costs of switching, and task substitutability (Manthei and Sliwka, 2019; Schumacher et al., 2001; KC, 2014).

Our experimental design bridges insights from psychology and economics. Specifically, it leverages multitasking paradigms from cognitive psychology, where a secondary or "side" task is commonly used to measure

[5] Also on the negative side of generative AI, a novel study by Melumad and Yun (2025) shows that learning through large language models restricts the depth of learning than through web search.

[6] While some interruptions reduce productivity by disrupting attention and workflow continuity (Cai et al., 2018), others—particularly unexpected interferences with idle time—can enhance creativity by stimulating cognitive processing and recovery (Schweisfurth and Greul, 2024); an integrative review further underscores the need for clearer conceptual definitions and theory-driven models to explain when and why such interruptions help or harm performance (Puranik et al., 2020).

the cognitive load imposed by a primary task. Classic dual-task research demonstrates that concurrent performance of two tasks typically induces systematic interference, reflecting central processing bottlenecks and capacity limitations (Pashler, 1994; Wickens, 2008).[7] Within cognitive load theory, dual-task methodology has been established as a direct and objective measure of cognitive demands: performance decrements on the secondary task are taken to reflect the cognitive load imposed by the primary task (Brünken et al., 2002, 2003; Schoor et al., 2012). By embedding such a side task within our lab-in-the-field experiment, we create a two-task environment in which generative AI is applied only to the primary, high-incentive task. This setup allows us to assess how AI assistance affects not only the directly supported task but also performance in the secondary task, which serves as a proxy for residual cognitive load.[8] The resulting pattern—improved outcomes in both tasks when AI is available—provides micro-level evidence for the mechanism through which AI mitigates classic multitasking distortions as formalized by Holmström and Milgrom (1991). In other words, by lowering the cognitive demands of the primary task, AI frees attentional and mental resources, allowing subjects to maintain or even enhance effort in tasks that would otherwise suffer under asymmetric incentives. This design thus offers a clean test of the spillover effect of AI on multitask performance, linking cognitive load reduction to the alleviation of distorted effort allocation in organizational settings.

In more general terms, recent work clarifies when and how generative AI shapes human behaviour. A large meta-analysis shows that human—AI teams typically underperform the best single agent on decision tasks yet can shine on open-ended creation, sharpening expectations for where assistance helps (Vaccaro et al., 2024). Field-scale evidence finds that providing LLM-generated ideas boosts judged creativity and writing quality while reducing the diversity of outputs, implying productivity gains with potential convergence of ideas (Doshi and Hauser, 2024). On attitudes and persuasion, sustained, tailored dialogues with an AI chatbot durably reduce conspiracy beliefs, illustrating that generative AI can shift entrenched views (Costello et al., 2024); in collective settings, AI-mediated summarization helps groups find common ground in democratic deliberation (Tessler et al., 2024). People feel more heard from AI-crafted supportive messages, but labeling the source as AI dampens that benefit (Yin et al., 2024). Finally, comparative tests caution against treating LLM behaviour as a drop-in surrogate for human cognition (Strachan et al., 2024) and offer guidance to avoid misusing LLMs as human stand-ins in behavioural research (Gao et al., 2025).

2. Methods

Experimental Design. We provide the important details of the methodology of our experiment. The study was approved by the Research Ethics Committee of the School of Economics and Management at Harbin Institute of Technology, Shenzhen and complied with all relevant guidelines. The study was not preregistered because it was designed as an exploratory investigation to generate behavioral and theoretical

[7] In economics, Alonso et al. (2014) microfound limited cognitive capacity by modeling the brain as an organization that allocates scarce processing resources across concurrently performed tasks, drawing on neuroscience evidence.

[8] Our key focus is not how the side task affects the main task nor to use the side task to measure how much cognitive load the main task has consumed, as was done in psychology studies. Instead, our focus is on the side task itself: given that many important tasks in the workplace are unmeasurable, we want to understand how AI helps to mitigate typical multitasking problems.

insights rather than to test pre-specified hypotheses. Participants provided informed consent before entering the experiments. Participants received a participation fee and were paid according to their decisions (details below).

Experimental Procedure. Upon arrival, subjects were required to hand over their personal electronic devices, which were then kept by an experimenter. Each participant used a preconfigured iPad Mini with a Bluetooth keyboard using the oTree platform (Chen et al., 2016). The procedure followed this sequence:

Stage 0: Subjects were introduced to the overall procedure and the payment/punishment rules for each task, and completed a comprehension test of 5 questions to ensure understanding. The experimental interface prevented subjects from advancing to subsequent pages if they did not answer all questions correctly. The subject had to continue entering revised answers by identifying his/her mistakes or receiving help from the experimental assistant. The experiment continued to the next section only if all subjects answered all questions correctly. In groups with access to an AI platform that we provide, this stage also included a 15-minute tutorial on how to write prompts for generative AI. This tutorial also contained a brief introduction to how to use the AI platform.[9]

Stage 1: All subjects performed a 15-minute writing task on the experimental interface using the iPad mini and connected keyboard, during which a calculation task was introduced by an experimenter using a projector in the 6th minute. The calculation instructions continued to be shown on the projector until the end of the stage. Subjects were informed beforehand that their writing would be interrupted by a calculation task, but the exact timing and format of the interference were not disclosed. For the calculation task, they could calculate and write their answers on a piece of paper using a pen — both were provided at each seat before any subjects entered the lab. All subjects could choose which task to work on freely, but were required to submit their answer to the calculation task before the end of the writing task. That is, they had to multitask for the remaining 9 minutes of the stage. All subjects were forbidden from using the AI platform.

Stage 2: This stage followed the same structure as Stage 1. That is, subjects completed another 15-minute writing task, with another calculation task inserted at the 6th minute. They were also required to write their answers on a new piece of paper, which was assigned before starting this stage. However, subjects in AI-Low and AI-High were allowed to use the AI platform when completing the writing task.

Stage 3: Subjects again completed a 15-minute writing task, but this time without any calculation-task interference. All other aspects were the same as in Stage 2. That is, subjects in treatment groups that involved using AI were allowed to use the AI platform when completing the writing task, and subjects in AI-High faced an incentive that doubled that in NO-AI and AI-Low.

Stage 4: All subjects completed a questionnaire regarding whether they had similar working experience as the tasks in the experiment and how helpful the AI was in completing the writing tasks (in AI-groups). They were informed that all payments from the experiment will be made through bank transfer via their respective companies. They were allowed to leave after everyone in the same session had finished the entire

[9] See training materials of the tutorial in Appendix D.

experiment.

Writing Tasks. The three writing tasks were designed to emulate professional workplace reports. Task 1 requires subjects to draft a turnover analysis report from an HR perspective, task 2 summarizes marketing department conflicts and proposes solutions, and task 3 requires an employee satisfaction report based on feedback data.[10] Each task required 300—500 words within 15 minutes. Writing Tasks 1 and 2 were interrupted at minute 6 with a calculation challenge shown by a projector and read aloud by an experimenter; this way of presenting the calculation tasks prevents any subject from copying the task descriptions to the AI interface. Task 3 had no interference, serving as the single-tasking control.

Calculation Tasks. Each calculation task required summing three separate three-digit figures, combining listening comprehension and manual calculation. Submissions were made on paper. In the first calculation task, subjects were asked to compute a company's total product sales by aggregating data from multiple channels, while carefully accounting for returns, promotional sales, and items not counted as final sales. The second calculation task required calculating store replenishment needs for a holiday, considering demand forecasts, current inventory, and supplier delivery rules.

Treatment NO-AI: This treatment serves as a control group. Subjects had no access to the AI platform in any stage. For each stage, a reward of CNY 100 per writing task is paid if a subject's score is ranked in the top 50% among subjects in the same experimental session. For Stages 1 and 2, incorrect answers in the calculation tasks incurred a CNY 5 penalty. At the beginning of the experiment, subjects were informed that they would earn a CNY 50 show-up fee if they fully participated in the entire experiment.

Treatment AI-Low: Subjects in this group had access to the AI platform that provides access to ChatGPT-3.5 in Stages 2 and 3. The reward rule for writing tasks in all stages, the punishment rule for calculation tasks in Stages 1 and 2, and the show-up fee and its payment rule are identical to those in NO-AI.

Treatment AI-High: This treatment is equivalent to AI-Low in terms of AI-accessibility, the punishment rule for the calculation tasks, and the show-up fee. However, for the writing tasks in all stages, each subject is rewarded CNY 200 (that is, double the reward in NO-AI and AI-Low) if her/his score is ranked in the top 50% among subjects in the same experimental session.

AI Access and Monitoring. Because ChatGPT is not directly available in China, we built a dedicated experimental web platform connected to the ChatGPT-3.5 API. Subjects could copy task prompts into the AI, edit AI outputs, and submit modified results. No restrictions were imposed on how prompts could be used, but Stage 1 explicitly prohibited AI use. Subjects in AI-Low and AI-High were warned that using AI in Stage 1 would result in forfeiting all payments. AI usage logs were recorded through the backend, allowing verification of compliance.

Performance Evaluation. Writing outputs were evaluated in a double-blind manner by 30 professional

[10] See task descriptions in Appendix C.

consultants from a local consulting company.[11] Each submission was scored independently by five raters, based on logic, accuracy, practicality, innovation, and completeness (each up to 20 points; total 100). A subject's final score was the average of five evaluations. Raters were paid CNY 5 per evaluation and received an additional CNY 10 bonus if their score was closest to the group average, incentivizing careful and unbiased assessment. Calculation tasks were scored as binary (correct/incorrect) with penalties applied.

Participant Recruitment. A total of 175 employees were recruited from three Shenzhen-based companies, selected to capture diverse workplace environments. Recruitment was facilitated through internal company announcements distributed by a collaborating research institute.

The sample includes 55 employees from ICBC, 63 from CESEC, and 57 from Qingchuang Zhiyuan. Subjects were required to register via an online form, reporting demographic information, education, hometown, ethnicity, monthly expenditures, and job role. They also completed a standardized language test with six multiple-choice questions from China's National Civil Service Examination.

Random assignment into the three treatments was stratified within each company to ensure balanced distribution. Baseline characteristics were checked to confirm no statistically significant differences across treatments. Each company was informed of group allocations and session times one week before the experiment, and employees were required to attend at their assigned times.[12]

Discussion of the experimental design. The experimental design enables us to disentangle two central questions. First, comparing the NO-AI and AI-Low groups in Stage 2 identifies the effect of generative AI on productivity in multitasking environments where the secondary task cannot benefit directly from AI. This comparison allows us to test whether the productivity gains observed in prior single-tasking studies extend to more complex settings where workers must divide attention across tasks. Second, comparing AI-Low and AI-High isolates the role of stronger economic incentives under AI assistance, shedding light on whether AI mitigates or amplifies the typical multitasking trade-off by altering performance on weakly incentivized tasks.

Stage 1 serves as a baseline to establish the existence of multitasking effort distortions in the absence of AI. The improvement in performance from stage 1 to 2, across groups, provides evidence on how AI affects both the primary (writing) and secondary (calculation) tasks when cognitive load is present. Stage 3 serves as another benchmark to compare the cost of calculation interference on writing performance across treatments, revealing how such a cost varies with AI accessibility and economic incentives.

3. Results

This section introduces our main findings. Summary statistics are provided in Table A.1. We randomized subjects into the three treatment groups based on demographics, role at work, and previous AI usage

[11] See evaluators' demographic information in Table A.4.

[12] Qingchuang Zhiyuan's sessions were held September 25—26, 2024; CESEC's, October 25—26, 2024; and ICBC's, December 14—15, 2024.

experience, which were elicited through a pre-experimental survey. Randomization analysis, provided in Table A.3, shows that it was successful.

3.1. Writing performance

In the first stage, the writing task scores are similar across the groups. Although the average score in AI-High (64.6) is higher than in AI-Low (60.3) and NO-AI (64.3), the differences are not statistically significant ($p = 0.109$ and $p = 0.898$, respectively).[13] Thus, in the multitasking environment where no subjects had access to AI, the additional incentive in AI-High did not result in better performance.[14] The latter two averages are also not statistically different from each other at the conventional level ($p = 0.082$). See the leftmost box plots in Figure 1 for stage-1 writing scores and the left panel of Figure 2 for distributions of the scores.

In the second stage, subjects assigned to the AI-Low and AI-High treatments were given access to AI assistance, which resulted in a significant performance increase. The average writing score increased from 63.5 in NO-AI to 74.7 in AI-Low and 74.1 in AI-High ($p < 0.001$ for both comparisons), consistent with findings from prior studies (e.g., Noy and Zhang (2023)). Similar to the first stage, there is no statistically significant difference between AI-Low and AI-High ($p = 0.427$). See the middle box plots in Figure 1 for the writing scores and the middle panel of Figure 2 for the distributions. Both illustrate the substantial difference between the AI-assisted groups and the NO-AI group. In fact, although the mean scores are statistically indistinguishable, Figure 2 suggests that the score distribution under AI-High is slightly shifted to the right relative to AI-Low.

In the third stage, the calculation task was removed, while AI assistance is still available. Writing performance in both AI-assisted groups—AI-Low (75.8) and AI-High (77.4)—was significantly higher than in the NO-AI group (67.4), with $p < 0.001$ (see the leftmost box plots of writing scores in Figure 1 and the right panel of Figure 2 for corresponding distributions). Again, no significant difference was observed between the two AI-assisted groups in terms of performance. The score distributions of the AI-assisted groups are almost identical.

To assess the impact of AI assistance and incentives on writing performance, we conduct OLS regressions for each writing task, as presented in Table 1. All models control for demographic characteristics, such as gender, education, and monthly expenses, and self-reported abilities in math, language, and computer skills, as well as prior AI experience. The NO-AI group is the baseline, with dummy variables for AI-Low and AI-High, and interactions with education and monthly expenses to explore heterogeneity. We select education and monthly expenses for heterogeneity analysis because they best approximate subjects' income level, among other variables.[15] Since our focus is the effect of economic incentives on performance in the

[13] Unless otherwise specified, all non-parametric tests in the paper are Wilcoxon rank-sum tests.

[14] Our analysis of time spent in each stage (Section B.2) suggests that the null effect of the incentive may reflect a binding time constraint. Nearly all subjects submitted Stage-1 work by the 15-minute deadline, implying that, even if they wished to, those in AI-High had little scope to exert additional effort.

[15] All companies that participated in the experiment rejected the question inquiring about subjects' income or any questions partially related to it.

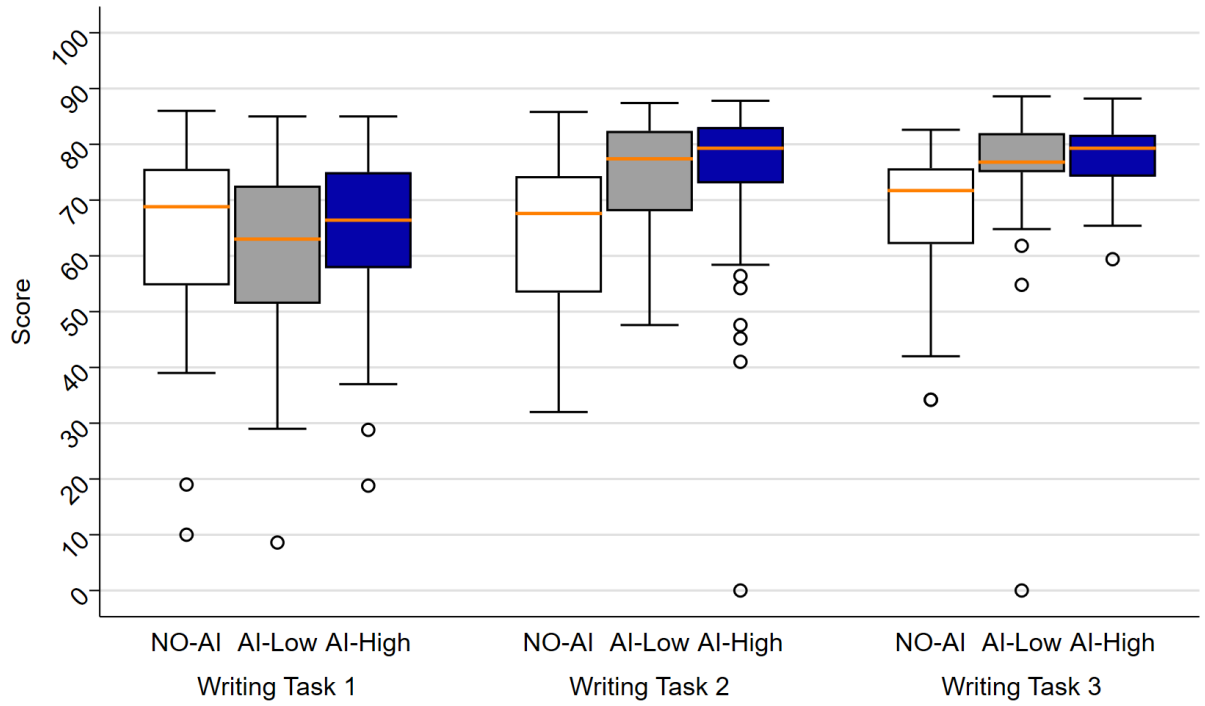


Figure 1: Boxplots of writing scores. Left panel: Writing task 1; Middle panel: Writing task 2; Right panel: Writing task 3.

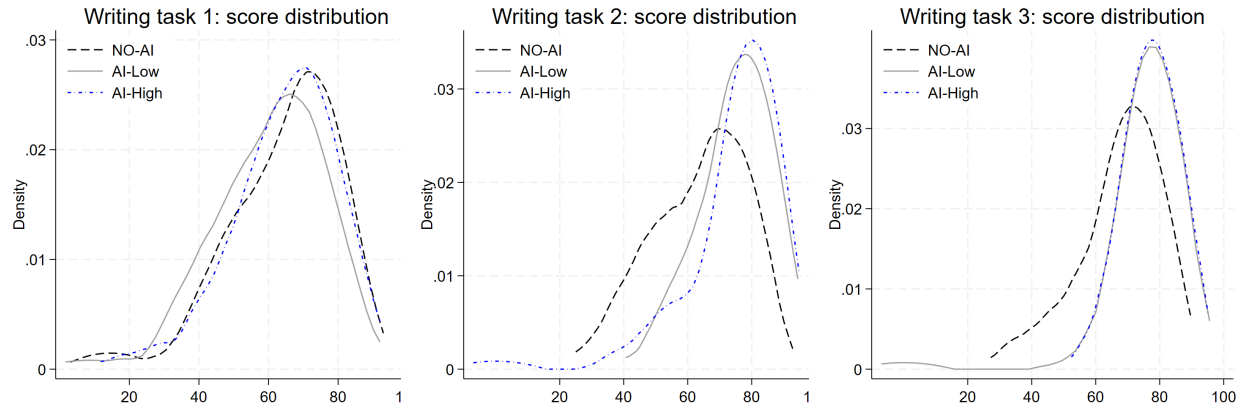


Figure 2: Distribution of writing scores. Left panel: Writing task 1; Middle panel: Writing task 2; Right panel: Writing task 3.

presence of AI assistance, the degree of sensitivity towards monetary incentives, which is likely correlated with subjects' income, should be accounted for in the analysis.

For writing task 1, regression (1) in Table 1 shows an insignificant coefficient for AI-High (-0.564, $p > 0.10$), indicating that the higher incentive did not yield an overall performance boost under multitasking conditions without AI. This is consistent with Figure 1. Regression (2) introduces interaction terms between education level and the treatment dummies, revealing a positive main effect for AI-High (28.223, $p < 0.05$) and a significant negative coefficient of its interaction term with education (-9.415, $p < 0.05$). This suggests

that the higher bonus motivated less educated subjects more strongly, as they are likely to have lower income or be paid less at work, making them more responsive to the increased incentive. For AI-Low, which matched NO-AI's incentive, the effects are insignificant except for a marginal negative coefficient in regression (1) (-4.199, $p < 0.10$). These observations show that the additional incentive in AI-High induced better writing performance for a subset of subjects (i.e., less educated ones) in AI-High than those in NO-AI. Regression (3) finds no significant interaction for AI-High with monthly expenses (1.250, $p > 0.10$). For regression (2), the Wald tests reject the null hypothesis that the parameters of the variables AI-Low and AI-High are equal ($p = 0.017$).[16] Hence, the additional incentives in AI-High did play a role in the writing task. However, the hypothesis is not rejected in regression (1) and (3) ($p = 0.22$ and $p = 0.744$, respectively).

For writing task 2, regression (4) in Table 1 shows strong positive effects for both AI-Low (11.981, $p < 0.01$) and AI-High (10.137, $p < 0.01$), showing that generative AI enhances writing performance in the multitasking environment. The Wald test does not reject the null hypothesis that the coefficients of AI-Low and AI-High are equal ($p = 0.520$), implying that incentives play a less essential role in determining writing performance in the presence of AI assistance. Regression (5) includes education interactions, which amplify AI benefits and, again, show a negative moderation for AI-High (-8.258, $p < 0.01$). Hence, the observation in stage 1 that lower-educated subjects performed better in the writing task in AI-High than in NO-AI carries to stage 2. Regression (6) shows a negative coefficient on the interaction term between AI-High and monthly expenses (-2.017, $p < 0.10$), implying that lower-spending subjects performed better in AI-High than in NO-AI. However, AI-Low's interaction terms with education and with monthly expense are not statistically significantly different from zero. That is, subjects benefited similarly from the AI tool provided in the experiment. Hence, the heterogeneous performance difference between NO-AI and AI-High is most likely driven by the stronger incentive in the latter, rather than the AI accessibility. For (5) and (6), Wald tests also show no significant differences between AI-Low and AI-High ($p = 0.197$ and $p = 0.744$, respectively)—consistent with the finding in (4) that the stronger incentive plays a less important role in boosting writing performance.

For writing task 3, regression (7) shows positive effects for AI-Low (9.354, $p < 0.01$) and AI-High (10.014, $p < 0.01$). The Wald test, again, shows no difference in their coefficients ($p = 0.756$). Regression (8) shows significantly negative coefficients of the interaction terms between education and the treatment dummies (AI-Low: -7.512, $p < 0.01$; AI-High: -6.626, $p < 0.01$). It indicates that generative AI offered less benefit to highly educated subjects. This is different from what is observed in regression (5), suggesting that the effect of AI in the multitasking environment can be different from that in the single-tasking environment. Regression (9) shows a marginally positive effect for AI-High with monthly expenses (7.576, $p < 0.10$), but other interactions are insignificant. In both regressions (8) and (9), Wald tests find no significant differences between AI-Low and AI-High ($p > 0.50$), suggesting that, without the calculation task, AI assistance leveled performance across incentive levels.

[16] The Wald test refers to the sum of coefficients of AI-High and AI-Low whenever there are interaction terms involving these treatment dummies in regressions. Specifically, the Wald test in regression (2) of Table 1 refers to the following null hypothesis: $\beta_{\text{AI-High}} + \beta_{\text{AI-High}\times\text{Education}} \cdot \text{Education} = \beta_{\text{AI-Low}} + \beta_{\text{AI-Low}\times\text{Education}} \cdot \text{Education}$.

[Insert Table 1 here]

3.2. Calculation Performance

Performance on the first calculation task, measured by its accuracy, does not differ significantly across groups; all pairwise comparisons yield $p > 0.2$. Nonetheless, the accuracy rate in the AI-High group (0.59) is lower than in the other two groups (0.70 in NO-AI and 0.68 in AI-Low) by around 10 percentage points, consistent with the multitasking theory (Holmström and Milgrom, 1991) which predicts that the additional incentive on the writing task in AI-High may have distorted effort allocation more strongly. That is, subjects may have prioritized writing over the calculation task, reducing accuracy on the latter due to divided attention. See the left panel of Figure 3.

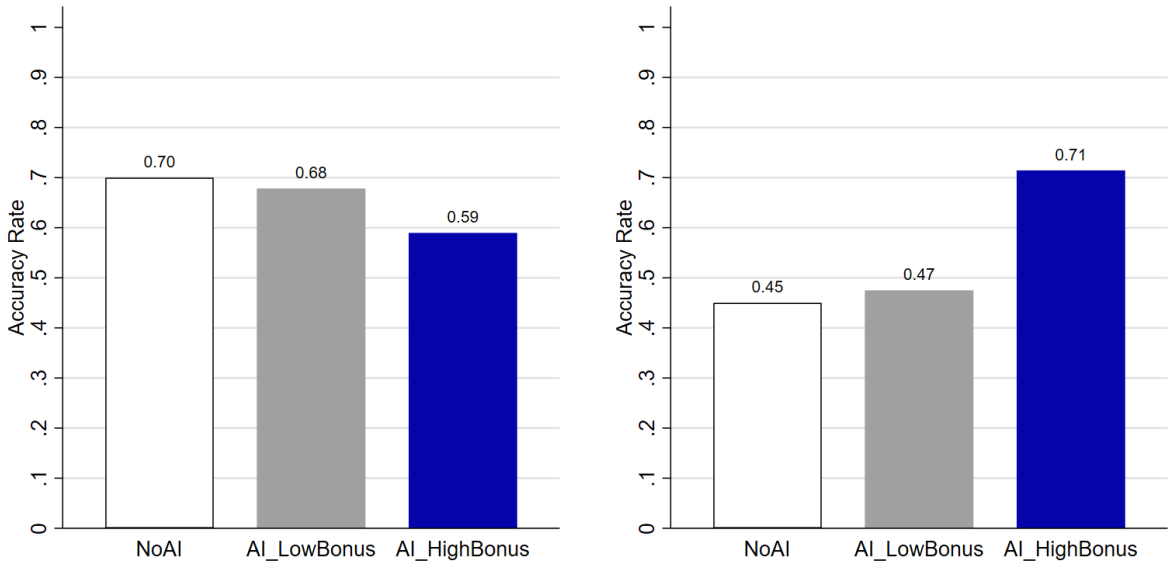


Figure 3: Calculation accuracy. Left panel: Calculation task 1; Right panel: Calculation task 2

The right panel of Figure 3 shows a substantially higher rate of task-2 calculation-accuracy in the AI-High group compared to the other groups; the differences are statistically significant ($p < 0.05$ for both comparisons). Recall that in all three groups and for both calculation tasks, the incentive is constant — inaccurate calculation results in a CNY 5 penalty. Moreover, the AI tool can only be used in writing tasks, rather than the calculation tasks. Hence, the finding supports the prediction that greater incentives on the writing task increase AI-reliance, which frees attention for the calculation task and increases accuracy.

OLS regressions in Table 2 confirm the lower stage-1 accuracy in AI-High. Regressions (1) and (3) reveal a significant negative effect for AI-High (-0.186, $p < 0.05$; -0.539, $p < 0.05$), indicating lower calculation accuracy compared to NO-AI. Again, these results support the multitasking theory. The Wald test shows that the coefficients of the treatment dummies in regression (1) are not significantly different from each other at the conventional level ($p = 0.101$). Yet, the parameters for AI-High and AI-Low are significantly different from each other in regression (3) ($p = 0.006$). Regression (2) shows an insignificant main effect for AI-High

(-0.480, $p > 0.10$) with a positive but insignificant education interaction (0.095, $p > 0.10$). However, the Wald test rejects the hypothesis that AI-High and AI-Low share the same parameter ($p = 0.032$), supporting the negative impact of AI-High on stage-1 calculation accuracy. AI-Low coefficients are insignificant across all three regressions, consistent with its identical stage-1 environment to NO-AI.

[Insert Table 2 here]

Regression (4) in Table 2 confirms the significant positive effect for AI-High (0.298, $p < 0.01$) in stage-2 calculation, suggesting improved calculation accuracy compared to NO-AI. Regression (5) further confirms this effect (0.892, $p < 0.01$) but reveals a negative education interaction (-0.197, $p < 0.05$), indicating that lower-educated subjects' calculation accuracy is significantly better in AI-High than in NO-AI. Regression (6) shows a positive effect for AI-High (0.707, $p < 0.01$) with an insignificant monthly expense interaction term (-0.083, $p > 0.10$). AI-Low coefficients remain insignificant across (4)-(6), suggesting that AI assistance on the writing task alone does not boost calculation accuracy.[17]

In summary, the higher stage-2 calculation accuracy in AI-High cannot be attributed to AI assistance alone, as the average accuracy rate is lower in AI-Low than in AI-High, even though AI-assistance was accessible in both treatments, and the coefficients of AI-Low are never significantly different from zero in regression (4)-(6) in Table 2. Given that the only difference between AI-Low and AI-High is that the latter doubles the incentive for writing tasks in the former, the results in subsections 3.1 and 3.2 reflect a novel cross-task *spillover effect* of economic incentives in the presence of AI assistance. Specifically, the stronger incentive for the writing tasks in AI-High did not lead to better writing performance than that in AI-Low in stage 2, yet it did result in higher calculation accuracy. We explore this effect further in the following section.

3.3. Calculation accuracy improvement

To further understand the impacts of AI on both tasks and how they vary with incentive change, we investigate how much generative AI improves performance in both tasks in this section. We focus on performance improvement in the calculation tasks to understand the spillover effect of incentives between tasks better.

For the calculation task, we categorize subjects into the following three sub-samples according to the dynamics of their performance: (1) Worse: submitted the right answer in stage 1 but wrong answers in stage 2; (2) Equal: either submitted a wrong answer in both stages or submitted the correct answer in both stages; (3) Better: submitted a wrong answer in stage 1 but the correct answer in stage 2. The following contingency table shows how the three sub-categories were distributed in each treatment group:

[Insert Table 3 here]

According to Table 3, the AI-High group exhibits the greatest improvement: the greatest number/percentage of accuracy improvement and the least number of worsened accuracy. The pairwise Chi-squared tests are

[17] The observations from OLS regressions in Table 2 are consistent with logit and probit regressions, shown in Table A.5 and A.6.

aligned with this observation; both NO-AI vs. AI-High and AI-Low vs. AI-High yield $p < 0.05$. There is no statistical difference between NO-AI and AI-Low. This observation, together with the previous findings in Figure 3 and Table 2, supports the spillover effect of economic incentives in the presence of generative AI. In particular, since there is no difference between NO-AI and AI-Low in either stage-2 calculation accuracy or accuracy improvement from stage 1 to 2, AI adoption alone does not play a role. The statistically significant difference in both stage-2 accuracy and the accuracy improvement between AI-High and AI-Low suggests that the doubled incentive for the writing task in the former boosted calculation accuracy. Crucially, this spillover effect of economic incentives between tasks only exists when generative AI is accessible: the stage-1 calculation accuracy is, if anything, lower in AI-High, which is consistent with distorted effort allocation predicted by the multitasking theory.

To further quantify the spillover effect of economic incentives, we estimate difference-in-difference regressions on stage 1 and 2 calculation accuracy using samples of AI-Low and AI-High groups, as shown in Table 4. The dependent variable is a binary indicator of whether calculation task 1 or 2 is submitted correctly by a subject. Independent variables include the AI-High dummy, which captures group differences, the stage 2 dummy, which captures the stage differences, and their interaction, capturing differential improvement in stage 2 for AI-High.

Regression (1) in Table 4 confirms the significant calculation accuracy improvement, while regression (2) and (3) show that the observation is robust after gradually including controls and company fixed effects. These observations align with the contingency table in Table 3, where AI-High had the highest share of "Better" performers (17 out of 56) versus AI-Low (7 out of 59), confirming the spillover effect of incentives between tasks. Regression (4)-(7) explores the underlying mechanism behind the spillover effect observed previously and in (1)-(3). The triple interaction with monthly expenses in (4) is negative and marginally significant (-0.126, $p < 0.10$), suggesting that lower-spending subjects in AI-High experienced slightly greater spillover. Regression (5) shows an insignificant interaction with education (-0.241, $p > 0.10$), implying that the spillover effect is not heterogeneous by education level, though Table 1 and 2 show that lower-educated subjects in AI-High performed better in both writing and calculation tasks in stage 2 than they did in NO-AI.

Regressions (6) illustrates interesting heterogeneity of the spillover effect, which supports the dominant role of the incentive-reliance mechanism. The triple interaction term with whether the subject had previous experiences working on tasks similar to the writing tasks in our experiment in (6) has a significantly negative coefficient (-0.759, $p < 0.01$). That is, those who reported never having any experience similar to working on the writing tasks exhibited a stronger spillover effect, i.e., their difference in calculation accuracy improvement between the AI-assisted groups is significantly greater than those who had previously experienced similar tasks.[18]

[Insert Table 4 here]

To gain an even deeper view of the mechanism, Table 5 provides OLS regressions of stage-1 and stage-2

[18]Table A.10 and A.11 present logit and probit regressions using the same variables as (1)-(6) in Table 4 and show that its results are robust.

calculation accuracy as dependent variables, respectively, and the AI-High dummy and its interaction with the control variables, respectively, in Table 4 as independent variables, using AI-Low and AI-High samples.[19]

Table 5 shows that the interaction between AI-High and previous similar experience is insignificant in Stage 1 (regression (4)) but becomes significantly negative in Stage 2 (regression (8)). This indicates that the larger improvement in calculation accuracy from Stage 1 to Stage 2 among subjects without prior similar writing experience is unlikely to result from any initial disadvantage in writing — due to their lack of such experience — that could have induced stronger effort distortion. Instead, it is more plausibly driven by their tendency to rely excessively on AI under AI-High, thereby conserving more cognitive resources for the calculation task.[20] This rapid delegation may stem from the high-powered incentives in AI-High, which create a perception of a hyper-competitive environment, coupled with lower self-confidence in an unfamiliar writing task.[21] In AI-High, inexperienced writers appear to delegate more decisively (though not necessarily to copy more), reducing internal uncertainty about the writing outcome and stabilizing attention, thus freeing resources for the concurrent calculation task. Because writing quality already approaches an AI-assisted plateau, higher bonuses no longer improve writing performance; instead, they tighten execution discipline and reduce switching and monitoring costs, which manifest as improved calculation accuracy.

[Insert Table 5 here]

The same does not hold when analyzing the effect of AI adoption alone on the calculation task. Regression (1)-(3) in Table 6 conduct difference-in-difference analysis similar to (1)-(3) in Table 4, except for using the NO-AI and AI-Low sample (hence, the AI-High dummy in the latter is replaced by AI-Low in the former). The interaction term Stage 2 $\times$ AI-Low is positive but not statistically significant, implying that AI assistance alone did not meaningfully offset the decline in accuracy observed in the second stage. Therefore, AI assistance on the writing task alone does not affect calculation accuracy.

[Insert Table 6 here]

3.4. Writing performance improvement

For completeness, we move on to writing performance dynamics from stage 1 to 2, showing the effect of AI, and from stage 2 to 3, revealing how the cost of interference on writing performance varies with whether AI is accessible and whether there is an additional incentive.

Effect of generative AI. We calculate the difference between the stage 2 score and stage 1 score for each subject, Δ_{21}. The average differences are 0.01 (NO-AI), 11.11 (AI-Low), and 9.11 (AI-High). Both AI-assisted groups significantly outperform the NO-AI group ($p < 0.001$ for both comparisons), while the

[19] Table A.7 provides OLS regressions on stage 1 and 2 writing scores, corresponding to Table 5.

[20] In regressions where time spent in Stage 2 is the dependent variable, we include AI-High, similar work experience, and their respective interactions with AI-High as independent variables, using both AI-High and AI-Low samples. None of these independent variables are significant, suggesting that time spent does not differ across treatments or individual characteristics.

[21] Indeed, whether subjects had prior similar writing experience is uncorrelated with their self-assessed language ability or their performance in the pre-experimental Chinese language test.

difference between AI-Low and AI-High is not significant ($p > 0.1$). This suggests that AI substantially improves writing performance, with additional incentives having no significant effect.

Regression (8) and (9) in Table 4 using the AI-Low and AI-High sample show that the interaction term AI-High × Stage 2 is insignificant, confirming no additional incentive impact beyond AI on the writing performance improvement. Regression (4)-(6) in Table 6 using the NO-AI and AI-Low sample show that the interaction term AI-Low × Stage 2 is significantly positive. This result further confirms our observation in previous subsections that using generative AI brings a larger improvement in writing performance than not using it, whereas additional incentive does not affect writing performance. Table A.9 extends the analysis to incorporate triple interaction terms with subject characteristics as those in Table 4 and shows no heterogeneity effects.

The findings reveal the flip side of the *spillover effect of economic incentives* that we discussed earlier in the calculation tasks. That is, the additional incentive improves the calculation performance, while it has no effect on the writing task, even though the additional incentive is imposed on the latter, in the presence of generative AI. Table A.7 using the AI-Low and AI-High samples confirms this observation by showing that the AI-High dummy and its interactions with education, monthly expense, and previous similar experience are never significant in regressions with writing scores in stage 1 and 2, respectively, as dependent variables.

Cost of interference. We also calculate the difference between the stage-3 and stage-2 writing scores for each subject, Δ_{32}. The average differences are 3.92 (NO-AI), 1.11 (AI-Low), and 1.70 (AI-High); no pairwise comparisons are statistically significant at the conventional level, though the improvement is substantially higher in NO-AI, suggesting a greater cost of interference without AI.[22]

Regression (1)-(3) in Table 7 using the AI-Low and AI-High sample shows insignificant coefficients for Stage 3, AI-High, and their interaction term on the dependent variable, writing scores. Regression (4)-(6), on the other hand, use the NO-AI and AI-Low sample, also show insignificant coefficients for the interaction term AI-Low × Stage 3. Hence, consistent with the pairwise comparisons of Δ_{32}, neither the generative AI nor its combination with additional incentive on writing tasks had a meaningful effect in mitigating the impact of calculation interference on writing performance.

[Insert Table 7 here]

4. Discussion

In our lab-in-the-field dual-task experiment with employees from diverse firms, participants completed an incentivized, measurable writing task alongside a simultaneous calculation task that cannot benefit from AI. When we grant GenAI access for writing, writing quality improves without degrading calculation performance. Strikingly, when incentives on the AI-assisted writing task are strengthened, calculation accuracy rises sharply (by over 20 percentage points). This pattern is inconsistent with the effort-diversion prediction

[22] The difference between NO-AI and AI-High is marginally significant with $p = 0.077$.

of classic multitasking models (Holmström and Milgrom, 1991) and is instead consistent with a channel in which higher-powered incentives increase reliance on GenAI for the primary task, in line with evidence from human–algorithm interaction that incentives can shift tool use (e.g., Greiner et al., 2025; Bockstedt and Buckman, 2025; Neumann et al., 2022; Saragih and Morrison, 2022).

We interpret these results through a cognitive-resource mechanism: GenAI reduces the processing demands of the incentivized writing task, and stronger incentives increase the propensity to rely on GenAI rather than on one's own deliberation, freeing up capacity that can be redeployed to the non-assistable calculation task.[23] Heterogeneity patterns reinforce this reading: spillovers to calculations are largest for subjects who do not have prior work experience similar to the writing task in our experiment. A plausible interpretation is that they may be less confident[24] about their writing performance; when competitiveness rises under the larger bonus in the AI-High condition, they rely more on GenAI and engage in less independent thinking, which reduces the human cognitive load of writing and generates larger spillovers to calculation.[25]

These results refine our understanding of multitasking under technological asymmetry. When GenAI lowers the cognitive cost of an incentivized, measurable task and stronger incentives induce greater reliance on GenAI, incentives can improve—rather than crowd out—performance on concurrent tasks that resist automation. Our findings also complement the growing single-task GenAI literature, which often finds larger gains for less experienced workers in writing, coding, and customer support (Noy and Zhang, 2023; Peng et al., 2023; Cui et al., 2025; Brynjolfsson et al., 2025; Kreitmeir and Raschky, 2024). We extend this evidence by showing that GenAI can generate indirect benefits for tasks it cannot perform, operating through incentives, reliance on AI, and reduced cognitive load. Overall, GenAI is not only a substitute for human effort on the assisted task; it can function as a capacity-relief technology that changes how incentives propagate across tasks—a theme emphasized in theories of contracting and the organization of knowledge, where incentives in one activity shape behavior in other, interdependent activities (Fuchs et al., 2015).

Spillovers are most likely when the primary task is AI-assistable and strongly incentivized, the concurrent task is weakly incentivized and non-assistable, AI meaningfully improves quality without offsetting supervision costs, and deadlines allow redeployment of freed capacity. When time pressure binds without AI, stronger incentives do not raise performance; when both tasks are equally (un)assistable, the resource-relief channel weakens and classic multitasking distortions are more likely to re-emerge. These boundaries help identify settings where distortion concerns remain first-order versus those where incentive power can be

[23] This interpretation is consistent with the dual-task logic in psychology, where secondary-task performance is used to index the primary task's processing demands (e.g., Pashler, 1994; Wickens, 2008; Brünken et al., 2002, 2003; Schoor et al., 2012; Schorr et al., 2003).

[24] Importantly, this need not reflect lower writing ability. In our data, heterogeneity by baseline language test scores and by self-rated writing ability (elicited in the pre-experimental questionnaire) does not support the claim that these subjects are objectively less able at writing.

[25] We do not find statistically significant differences between AI-Low and AI-High in observed measures of AI usage in stage 2, including (i) similarity between AI's answers and the submitted work, (ii) total time spent (calculation was submitted on paper, so we do not separately observe time by task), (iii) the time gap between the first prompt and task start, and (iv) the time gap between the last prompt and submission. This does not undermine the interpretation above for two reasons. First, the calculation task's instructions require less than 40 seconds of attention, so reallocating attention to calculation is unlikely to generate large, detectable shifts in these coarse AI-usage proxies. Second, similarity-based measures can be uninformative about marginal reliance: a subject who initially relies less on GenAI may later adopt AI-provided structure or sentences, yielding submitted text that is comparably similar to AI output even if the underlying reliance process differs.

raised safely because the incentive-reliance channel turns extra motivation on the primary task into spillovers elsewhere.

From a managerial perspective, organizations can attach high-powered incentives to AI-assistable, measurable deliverables without necessarily neglecting complementary, harder-to-measure work; improvements in the latter can arise through freed capacity when stronger incentives lead workers to rely more on AI as a tool. In bottlenecked sequential workflows, targeting incentives to AI-augmented stages may improve system performance, particularly for workers facing tighter ex-ante constraints. Because the largest spillovers accrue to less experienced subjects, onboarding, prompt-use training, and feedback for these groups may help mitigate distributional concerns. Where tasks are isolated and interference is minimal, extra bonuses may yield diminishing returns once AI lifts quality, arguing for investments in tooling, workflow integration, and prompt tutoring rather than ever-stronger pay-for-performance schemes.

Our evidence concerns short-run responses with a quantitative secondary task; socially complex or coordination-intensive activities and longer-run learning remain open, testable cases. Future work can vary the nature and measurability of the secondary task and manipulate frictions to AI use (e.g., latency, verification) to separate genuine capacity relief from motivational or strategic channels, and can explicitly measure changes in AI reliance under different incentive schemes.

When AI lowers the cognitive cost of producing measurable outputs, and stronger incentives induce workers to use AI more intensively on those outputs, incentives can be made bolder in multitasking settings without sacrificing—and sometimes enhancing—performance on tasks that resist automation. Rather than substituting for human effort, GenAI in our study relieves processing bottlenecks so that motivation operates more effectively across tasks. This synthesis—linking incentive theory, capacity-limit psychology, and field evidence from diverse firms—aims to inform the design of AI-enabled jobs and, more broadly, debates about how to harness GenAI for system-wide productivity while safeguarding the hard-to-measure work where human judgment remains indispensable.

Table 1: OLS regression analysis on writing performance

	Writing Task 1			Writing Task 2			Writing Task 3		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI-Low	-4.199*	-3.077	-3.975	11.981***	23.105***	17.935***	9.354***	31.803***	3.779
	(2.498)	(9.981)	(7.312)	(2.083)	(7.819)	(5.852)	(1.913)	(6.677)	(9.643)
AI-High	-0.564	28.223**	-6.387	10.137***	35.387***	19.761***	10.014***	29.869***	7.576*
	(3.041)	(12.153)	(8.262)	(3.011)	(9.638)	(5.445)	(1.760)	(6.208)	(4.463)
AI-High × Education		-9.415**			-8.258***			-6.626***	
		(3.896)			(2.956)			(1.898)	
AI-High × Monthly Expense			1.250			-2.017*			0.486
			(1.642)			(1.205)			(0.970)
AI-Low × Education		-0.340			-3.693			-7.512***	
		(3.096)			(2.452)			(2.003)	
AI-Low × Monthly Expense			-0.118			-1.207			1.208
			(1.548)			(1.174)			(1.877)
Observations	175	175	175	175	175	175	175	175	175
R-squared	0.299	0.341	0.303	0.337	0.369	0.346	0.336	0.394	0.340
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses; significance levels against the (two-sided) null of a zero effect: * p<0.10, ** p<0.05, *** p<0.01. The dependent variables are writing performance in stage 1 in (1)-(3), stage 2 in (4)-(6), and stage 3 in (7)-(9). The main independent variables are dummies indicating whether the subject was in AI-Low, AI-High, and their interactions with education and monthly expense, respectively. All regressions control for subjects' age, gender, education, whether the subject is Han Chinese, whether they originate from the Guangdong Province where Shenzhen belongs to, self-evaluated ability on math, language, computer, and previous experiences of using AI.

Table 2: OLS regression analysis on calculation performance

	Calculation Task 1			Calculation Task 2		
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	-0.039	0.241	0.094	0.052	0.349	0.378
	(0.078)	(0.312)	(0.218)	(0.088)	(0.287)	(0.266)
AI-High	-0.186**	-0.480	-0.539**	0.298***	0.892***	0.707***
	(0.088)	(0.345)	(0.238)	(0.086)	(0.295)	(0.234)
AI-High × Education		0.095			-0.197**	
		(0.111)			(0.096)	
AI-High × Monthly Expense			0.078			-0.086*
			(0.049)			(0.050)
AI-Low × Education		-0.094			-0.099	
		(0.102)			(0.096)	
AI-Low × Monthly Expense			-0.034			-0.068
			(0.048)			(0.055)
Observations	175	175	175	175	175	175
R-squared	0.283	0.296	0.308	0.320	0.335	0.334
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 (regression (1)-(3)) or 2 (regression (4)-(6)) is answered correctly; the independent and control variables are the same as those in Table 1.

Table 3: The distribution of calculation task performance dynamics from stage 1 to 2.

	Worse	Equal	Better		Worse	Equal	Better
AI-High	10	29	17	AI-High	17.9%	51.8%	30.4%
AI-Low	19	33	7	AI-Low	32.2%	55.9%	11.9%
NO-AI	19	37	4	NO-AI	31.7%	61.7%	6.7 %

Note: Left panel: the number of subjects who had worse, equal, or better calculation performance from stage 1 to 2 in each treatment; Right panel: the percentage of subjects who had worse, equal, or better calculation performance from stage 1 to 2 in each treatment.

Table 4: Performance improvement from stage 1 to 2: AI-Low vs. AI-High

	Calculation accuracy						Writing performance		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI-High	-0.089	-0.092	-0.098	-0.599**	-0.651*	-0.193	4.274	3.220	3.360
	(0.091)	(0.090)	(0.091)	(0.233)	(0.380)	(0.156)	(2.751)	(2.773)	(2.797)
Stage 2	0.323	0.406	0.421	0.322	0.088	0.192	-13.304	-10.267	-10.489
	(0.474)	(0.499)	(0.498)	(0.529)	(0.597)	(0.514)	(10.830)	(13.490)	(13.646)
AI-High × Stage 2	0.331**	0.327***	0.328***	0.905***	1.069**	0.854***	-2.559	-2.322	-2.418
	(0.129)	(0.124)	(0.125)	(0.322)	(0.489)	(0.205)	(3.288)	(3.284)	(3.281)
AI-High × Stage 2 × Mon. Exp.				-0.133*					
				(0.071)					
AI-High × Stage 2 × Education					-0.247				
					(0.157)				
AI-High × Stage 2 × Sim. Exp.						-0.717***			
						(0.246)			
Observations	230	230	230	230	230	230	230	230	230
R-squared	0.06	0.24	0.25	0.27	0.26	0.28	0.28	0.39	0.40
Controls	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes	No	No	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly for (1)-(6) and writing scores in (7)-(9). The independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-High, indicating whether the subject is in AI-High or AI-Low, and their interaction term. Regression (4)-(6) include triple interactions with Mon. Exp., which stands for monthly expense, Education, and Sim. Exp., which stands for similar previous writing experience, respectively. Control variables include those in Table 1, the word-count of the subject's submitted writing at the same stage, reflecting the effort on writing, the duration in seconds that the subject spent in the corresponding stage, word-count in the interaction with the AI, and word-count of the AI's last response. All subjects in AI-High and AI-Low are included.

Table 5: OLS regressions on calculation tasks: AI-Low vs. AI-High

	Calculation Task 1				Calculation Task 2			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-High	-0.122	-0.494	-0.552**	-0.335***	0.286**	0.401	0.327	0.712***
	(0.098)	(0.393)	(0.261)	(0.126)	(0.114)	(0.430)	(0.250)	(0.162)
AI-High × Education		0.122				-0.045		
		(0.128)				(0.136)		
AI-High × Monthly Expense			0.099				-0.015	
			(0.060)				(0.054)	
AI-High × Similar Experience				0.298*				-0.601***
				(0.156)				(0.168)
Observations	115	115	115	115	115	115	115	115
R-squared	0.369	0.376	0.393	0.385	0.314	0.357	0.357	0.430
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly; the independent variables include the AI-High dummy, indicating whether the subject is in AI-High or AI-Low, and its interaction terms with Education, Monthly Expense, and previous Similar Experience, respectively. Control variables are the same as those in Table 1, the word-count of the subject's submitted writing at the same stage, reflecting the effort on writing, and the duration in seconds that the subject spent in the corresponding stage. Regression (5)-(8) included additional control variables: stage-1 accuracy, word-count in the interaction with the AI, word-count of the AI's last response. All subjects in AI-High and AI-Low are included.

Table 6: Performance improvement from stage 1 to 2: NO-AI vs. AI-Low

	Calculation accuracy			Writing performance		
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	-0.022	-0.050	-0.047	-4.005	-3.702	-3.597
	(0.086)	(0.081)	(0.082)	(2.796)	(2.452)	(2.435)
Stage 2	-0.250***	-0.250***	-0.250***	-0.820	-0.820	-0.820
	(0.088)	(0.078)	(0.078)	(2.680)	(2.152)	(2.148)
AI-Low × Stage 2	0.047	0.047	0.047	15.206***	15.206***	15.206***
	(0.126)	(0.116)	(0.116)	(3.562)	(3.127)	(3.116)
Observations	238	238	238	238	238	238
R-squared	0.05	0.29	0.29	0.13	0.41	0.42
Controls	No	Yes	Yes	No	Yes	Yes
Session FE	No	No	Yes	No	No	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly; the independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-Low, indicating whether the subject is in NO-AI or AI-Low, and their interaction term. Control variables are the same as in Table 1. All subjects in NO-AI and AI-Low are included.

Table 7: Writing performance dynamics from stage 2 to 3

	AI-Low vs. AI-High				NO-AI vs. AI-Low		
	(1)	(2)	(3)		(4)	(5)	(6)
AI-High	-0.352	-1.033	-0.848	AI-Low	11.202***	12.140***	12.218***
	(2.380)	(2.670)	(2.636)		(2.207)	(1.980)	(1.991)
Stage 3	1.112	1.112	1.112	Stage 3	3.920*	3.920**	3.920**
	(2.009)	(1.964)	(1.967)		(2.370)	(1.862)	(1.869)
AI-High × Stage 3	1.917	1.917	1.917	AI-Low × Stage 3	-2.808	-2.808	-2.808
	(2.953)	(2.946)	(2.928)		(3.106)	(2.784)	(2.791)
Observations	230	230	230	Observations	238	238	238
R-squared	0.01	0.13	0.15	R-squared	0.16	0.40	0.40
Controls	No	Yes	Yes	Controls	No	Yes	Yes
Session FE	No	No	Yes	Session FE	No	No	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the writing score in stage 2 and 3; the independent variables include Stage 3, indicating whether the calculation task is in stage 2 or 3, AI-Low, indicating whether the subject is in NO-AI or AI-Low, AI-High, indicating whether the subject is in AI-High or AI-Low, and interaction terms between Stage 3 and each treatment dummy. Control variables are the same as those in Table 1. All subjects in NO-AI and AI-Low are included.

References

Alonso, R., I. Brocas, and J. D. Carrillo (2014). Resource allocation in the brain. The Review of Economic Studies 81(2), 501–534.

Bockstedt, J. C. and J. R. Buckman (2025). Humans' use of AI assistance: The effect of loss aversion on willingness to delegate decisions. Management Science.

Bohren, N., R. Hakimov, and R. Lalive (2024). Creative and strategic capabilities of generative ai: Evidence from large-scale experiments. IZA Discussion Papers No. 17302.

Bonin, A. L., P. R. Smolinski, and J. Winiarski (2025). Exploring the Impact of Generative Artificial Intelligence on Software Development in the IT Sector: Preliminary Findings on Productivity, Efficiency and Job Security. Working Paper.

Brynjolfsson, E. and D. Li (2024). The economics of generative ai. NBER Reporter (1), 16–19.

Brynjolfsson, E., D. Li, and L. R. Raymond (2025). Generative ai at work. The Quarterly Journal of Economics 140(2), 889–942.

Brünken, R., J. L. Plass, and D. Leutner (2003). Direct measurement of cognitive load in multimedia learning. Educational Psychologist 38(1), 53–61.

Brünken, R., S. Steinbacher, J. L. Plass, and D. Leutner (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. Experimental Psychology 49(2), 109–119.

Cai, X., J. Gong, Y. Lu, and S. Zhong (2018). Recover overnight? work interruption and worker productivity. Management Science 64(8), 3489–3500.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree–an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97.

Chen, Y. J., J. Gong, J. Li, and Z. Zhao (2025). Better technology, worse motivation: Genais mediocrity trap. Available at SSRN 5208163.

Costello, T. H., G. Pennycook, and D. G. Rand (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science 385(6714), eadq1814.

Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2025). The effects of generative ai on high-skilled work: Evidence from three field experiments with software developers. SSRN 4945566.

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Harvard Business School Working Paper 24-013.

Doshi, A. R. and O. P. Hauser (2024). Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. Science Advances 10(28), eadn5290.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science 384(6702), 1306–1308.

Fang, L., Z. Yuan, K. Zhang, D. Donati, and M. Sarvary (2025). Generative ai and firm productivity: Field experiments in online retail.

Friebel, G. and M. Raith (2004). Abuse of Authority and Hierarchical Communication. RAND Journal of Economics, 224–244.

Fryer, R. G. and R. T. Holden (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper No. 18237.

Fuchs, W., L. Garicano, and L. Rayo (2015). Optimal contracting and the organization of knowledge. The Review of Economic Studies 82(2), 632–658.

Gambacorta, L., H. Qiu, S. Shan, and D. M. Rees (2024). Generative ai and labour productivity: A field experiment on coding. BIS Working Papers No 1208.

Gao, Y., D. Lee, G. Burtch, and S. Fazelpour (2025). Take caution in using LLMs as human surrogates. Proceedings of the National Academy of Sciences 122(24), e2501660122.

Garicano, L. and L. Rayo (2016). Why organizations fail: Models and cases. Journal of Economic Literature 54(1), 137–192.

Gibbons, R. (1998). Incentives in organizations. Journal of economic perspectives 12(4), 115–132.

Gmyrek, P., J. Berg, K. Kamiński, F. Konopczyński, A. Ładna, B. Nafradi, K. Rosłaniec, and M. Troszyński (2025). Generative AI and jobs: A refined global index of occupational exposure. Number 140. ILO Working Paper.

Greiner, B., P. Grünwald, T. Lindner, G. Lintner, and M. Wiernsperger (2025). Incentives, framing, and reliance on algorithmic advice: An experimental study. Management Science. Articles in Advance.

Hoffmann, M., S. Boysel, F. Nagle, S. Peng, and K. Xu (2024). Generative ai and the nature of work. CESifo Working Paper No. 11479.

Holmström, B. and P. Milgrom (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. Journal of Law, Economics, & Organization 7(Special Issue), 24–52.

Hou, J. J., L. Wang, G. Wang, H. J. Wang, and S. Yang (2025). The double-edged roles of generative ai in the creative process: Experiments on design work. Information Systems Research.

Huang, M., M. Jin, and N. Li (2024). Augmenting Minds or Automating Skills: The Differential Role of Human Capital in Generative AI's Impact on Creative Tasks. arXiv preprint arXiv:2412.03963.

Ju, H. and S. Aral (2025). Collaborating with ai agents: A field experiment on teamwork, productivity, and performance.

KC, D. S. (2014). Does multitasking improve performance? Evidence from the emergency department. Manufacturing & Service Operations Management 16(2), 168–183.

Kreitmeir, D. and P. A. Raschky (2024). The Heterogeneous Productivity Effects of Generative AI. Working Paper.

Manthei, K. and D. Sliwka (2019). Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank. Management Science 65(12), 5861–5883.

Melina, G., A. J. Panton, C. Pizzinelli, E. Rockall, and M. M. Tavares (2024). Gen-AI: Artificial intelligence and the future of work. Working Paper.

Melumad, S. and J. H. Yun (2025). Experimental evidence of the effects of large language models versus web search on depth of learning. PNAS nexus 4(10), pgaf316.

Monsell, S. (2003). Task switching. Trends in Cognitive Sciences 7(3), 134–140.

Nakavachara, V., T. Potipiti, and T. Chaiwat (2024). Experimenting with Generative AI: Does ChatGPT Really Increase Everyones Productivity?

Neumann, M., M. Hengeveld, A. S. M. Niessen, J. N. Tendeiro, and R. R. Meijer (2022). Education increases decision-rule use: An investigation of education and incentives to improve decision making. Journal of Experimental Psychology: Applied 28(1), 166–178.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654), 187–192.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. Psychological Bulletin 116(2), 220–244.

Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The impact of AI on developer productivity: Evidence from github copilot. arXiv preprint arXiv:2302.06590.

Puranik, H., J. Koopman, and H. C. Vough (2020). Pardon the interruption: An integrative review and future research agenda for research on work interruptions. Journal of Management 46(6), 806–842.

Rayo, L. (2007). Relational incentives and moral hazard in teams. The Review of Economic Studies 74(3), 937–963.

Saragih, M. and B. W. Morrison (2022). The effect of past algorithmic performance and decision significance on algorithmic advice acceptance. International Journal of Human–Computer Interaction 38(13), 1228–1237.

Schoor, C., M. Bannert, and R. Brünken (2012). Role of dual task design when measuring cognitive load during multimedia learning. Educational Technology Research and Development 60(5), 753–768.

Schorr, T., P. Gerjets, and K. Scheiter (2003). Analyzing effects of goal competition and task difficulty in multiple-task performance: Volitional action control within act-r. In R. Alterman and D. Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society, pp. 1053–1058. Cognitive Science Society.

Schumacher, E. H., T. L. Seymour, J. M. Glass, and et al. (2001). Concurrent response-selection processes in dual-task performance: Evidence for adaptive executive control of task scheduling. Journal of Experimental Psychology: Human Perception and Performance 27(4), 991–1010.

Schweisfurth, T. G. and A. Greul (2024). Unexpected interruptions, idle time, and creativity: Evidence from a natural experiment. Organization Science 35(1), 116–137.

Simkute, A., L. Tankelevitch, V. Kewenig, A. E. Scott, A. Sellen, and S. Rintel (2024). Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interactions.

Strachan, J. W. A., D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. A. Graziano, and C. Becchio (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour 8(7), 1285–1295.

Talypova, D., A. Lingler, and P. Wintersberger (2024). Prolonged usage of ai assistant for improving multi-tasking performance. In Proceedings of the 12th International Conference on Human-Agent Interaction, pp. 404–407.

Tasoff, J. and K. Madarász (2009). A model of attention and anticipation. Working paper.

Tessler, M. H., M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, and C. Summerfield (2024). Ai can help humans find common ground in democratic deliberation. Science 386(6719), eadq2852.

Uhunoma, E. (2025). From hours to minutes: The impact of generative ai on work turnaround time and productivity. United International Journal for Research Technology 6(3), 45–56.

Vaccaro, M., A. Almaatouq, and T. Malone (2024). When Combinations of Humans and AI are Useful: A Systematic Review and Meta-Analysis. Nature Human Behaviour.

Wang, Z., L. Wei, and L. Xue (2025). Overcoming medical overuse with ai assistance: An experimental investigation. Journal of Health Economics, 103043.

Wickens, C. D. (2008). Multiple resources and mental workload. Human Factors 50(3), 449–455.

Yin, Y., N. Jia, and C. J. Wakslak (2024). Ai can help people feel heard, but an AI label diminishes this impact. Proceedings of the National Academy of Sciences 121(14), e2319112121.

Zhou, E. and D. Lee (2024). Generative artificial intelligence, human creativity, and art. PNAS Nexus 3(3), pgae052.

672 Zou, W. and F. Zhu (2025). Generative ai adoption in human creative tasks: Experimental evidence. SSRN
673 5196748.

A. Supporting tables

Table A.1: Summary statistics

Variable	Median	Mean	Std.Dv	Min	Max	Obs
Age	30	30.92	5.632	22	54	175
Gender						
Male	0	.474	.501	0	1	175
Female	1	.526	.501	0	1	175
Education						
Secondary-vocational	0	.04	.197	0	1	175
Associate	0	.229	.421	0	1	175
Bachelor	0	.429	.496	0	1	175
Master	0	.297	.458	0	1	175
Doctoral	0	.006	.076	0	1	175
Is-Guangdong						
NOT-GD	0	.486	.501	0	1	175
IS-GD	1	.514	.501	0	1	175
Is-Han						
NOT-HAN	0	.046	.209	0	1	175
IS-HAN	1	.954	.209	0	1	175
Session-Code						
GONGHANG	0	.314	.466	0	1	175
SAIXI	0	.36	.481	0	1	175
QINGCHUANG	0	.326	.47	0	1	175
Department-Type						
Finance	0	.029	.167	0	1	175
Administration	0	.097	.297	0	1	175
Property	0	.029	.167	0	1	175
Operation	0	.343	.476	0	1	175
Technology	0	.24	.428	0	1	175
Management	0	.143	.351	0	1	175
Consulting	0	.12	.326	0	1	175
Monthly-Expense						
<1000	0	.051	.222	0	1	175
1000-1500	0	.069	.253	0	1	175
1500-2000	0	.149	.357	0	1	175
2000-3000	0	.211	.409	0	1	175
3000-4000	0	.206	.405	0	1	175
4000-6000	0	.194	.397	0	1	175
6000-10000	0	.074	.263	0	1	175
>10000	0	.046	.209	0	1	175
Math-Top50	60	59.909	24.722	0	100	175
Math-Ability	7	7.36	1.804	1	11	175
Computer-Ability	7	7.28	1.711	2	11	175
Computer-Top50	60	60.103	23.925	0	100	175
Language-Top50	60	62.469	22.097	0	100	175
Language-Ability	8	7.526	1.688	2	11	175
Chinese-Error-Count	2	2.646	1.715	0	8	175

Table A.2: Summary statistics (Continued)

Variable	Median	Mean	Std.Dv	Min	Max	Obs
AI-Society-Attitude						
AI-Very-positive	0	.343	.476	0	1	175
AI-Positive	0	.411	.494	0	1	175
AI-Neutral	0	.24	.428	0	1	175
AI-Negative	0	.006	.076	0	1	175
AI-Very-negative	0	0	0	0	0	175
AI-Usage						
AI-usage-occasional	1	.554	.498	0	1	175
AI-usage-work	0	.4	.491	0	1	175
AI-usage-entertainment	0	.194	.397	0	1	175
Work						
work-creative	1	.754	.432	0	1	175
work-customer-service	0	.406	.492	0	1	175
work-writing	0	.217	.413	0	1	175
work-data-analysis	0	.451	.499	0	1	175
work-decision	1	.714	.453	0	1	175
work-negotiation	0	.497	.501	0	1	175

Table A.3: Randomization check

Variable	No-AI		AI-Low		AI-High		Overall		Test Results
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	p-value
Age	30.783	6.405	31.576	5.742	30.375	4.559	30.920	5.632	0.513
Math-Ability	7.417	1.759	7.017	1.824	7.661	1.802	7.360	1.804	0.114
Math-Top50	61.350	23.488	54.814	26.306	63.732	23.813	59.909	24.722	0.124
Computer-Ability	7.267	1.645	7.051	1.716	7.536	1.768	7.280	1.711	0.289
Computer-Top50	59.033	23.110	59.153	23.820	62.250	25.154	60.103	23.925	0.666
Language-Ability	7.533	1.641	7.339	1.816	7.714	1.604	7.526	1.688	0.671
Language-Top50	63.517	22.050	59.542	22.493	64.429	21.801	62.469	22.097	0.588
Chinese-Error-Count	2.583	1.565	2.627	1.956	2.732	1.624	2.646	1.715	0.675
Is-Han	0.050	0.220	0.034	0.183	0.054	0.227	0.046	0.209	0.908
Education	2.983	0.965	3.017	0.820	3.000	0.739	3.000	0.844	0.576
Department-Type	4.250	1.590	4.593	1.428	4.911	1.418	4.577	1.498	0.102
Monthly-Expense	4.950	1.661	4.356	1.616	4.357	1.853	4.560	1.724	0.332
AI-Usage	4.083	1.555	3.932	1.639	4.643	1.470	4.211	1.578	0.285
Gender	1.533	0.503	1.508	0.504	1.536	0.503	1.526	0.501	0.948
Is-Guangdong	0.450	0.502	0.576	0.498	0.518	0.504	0.514	0.501	0.386
Session-Code	1.983	0.813	2.034	0.809	2.018	0.798	2.011	0.802	0.995
AI-Society-Attitude	1.833	0.763	1.966	0.830	1.929	0.735	1.909	0.775	0.743
Work-Creative	1.733	0.446	1.780	0.418	1.750	0.437	1.754	0.432	0.838
Work-Customer-Service	1.400	0.494	1.373	0.488	1.446	0.502	1.406	0.492	0.720
Work-Writing	1.150	0.360	1.254	0.439	1.250	0.437	1.217	0.413	0.298
Work-Data-Analysis	1.500	0.504	1.424	0.498	1.429	0.499	1.451	0.499	0.646
Work-Decision	1.650	0.481	1.746	0.439	1.750	0.437	1.714	0.453	0.396
Work-Negotiation	1.450	0.502	1.542	0.502	1.500	0.505	1.497	0.501	0.601
Observations	60		59		56		175		

Table A.4: Evaluators

Variable	Median	Mean	Std.Dv	Min	Max	Obs
Age	33.5	33.267	5.362	23	42	30
Major						
Engineering	1	.6	.498	0	1	30
Business	0	.2	.407	0	1	30
Economics	0	.067	.254	0	1	30
Environment	0	.033	.183	0	1	30
Social	0	.1	.305	0	1	30
Highest-Edu						
Bachelor	1	.667	.479	0	1	30
Master	0	.333	.479	0	1	30
Current-Job						
Business	0	.067	.254	0	1	30
Consult	0	.1	.305	0	1	30
Engineering	0	.233	.43	0	1	30
Manage	0	.2	.407	0	1	30
Research	0	.367	.49	0	1	30
Work-Experience						
1 to 3 Years	0	.267	.45	0	1	30
4 to 5 Years	0	.1	.305	0	1	30
More Than 6 Years	1	.633	.49	0	1	30

Table A.5: Logit & Probit regressions for calculation task 1

	Calculation Task 1					
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	-0.490	-0.299	0.908	0.677	1.137	0.640
	(0.497)	(0.279)	(1.785)	(1.016)	(1.904)	(1.034)
AI-High	-1.212**	-0.715**	-3.145	-1.802	-3.521**	-2.113**
	(0.531)	(0.294)	(1.966)	(1.110)	(1.544)	(0.875)
AI-High × Education			0.634	0.352		
			(0.645)	(0.361)		
AI-High × Monthly Expense					0.488	0.293*
					(0.306)	(0.173)
AI-Low × Education			-0.475	-0.328		
			(0.582)	(0.328)		
AI-Low × Monthly Expense					-0.378	-0.214
					(0.404)	(0.214)
Intercept	-2.443	-1.295	-2.295	-1.284	-1.457	-0.802
	(3.053)	(1.702)	(3.081)	(1.724)	(3.066)	(1.716)
Observations	175	175	175	175	175	175
Pseudo R-squared	0.263	0.264	0.275	0.277	0.296	0.299
Wald test p-value	0.159	0.142	0.024	0.021	0.004	0.002
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 is answered correctly; the independent variables include treatment dummies and their interactions with education and monthly expense, respectively. Control variables are the same as those in Table 1. All regressions control for company fixed effect. All samples are included.

Table A.6: Logit & Probit regressions for calculation task 2

	Calculation Task 2					
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	0.321	0.196	3.273*	1.928**	2.379*	1.436*
	(0.474)	(0.271)	(1.767)	(0.983)	(1.284)	(0.766)
AI-High	1.759***	1.022***	5.812***	3.428***	4.199***	2.500***
	(0.538)	(0.294)	(2.038)	(1.123)	(1.384)	(0.784)
AI-High × Education			-1.306**	-0.780**		
			(0.626)	(0.352)		
AI-High × Monthly Expense					-0.501*	-0.303*
					(0.278)	(0.158)
AI-Low × Education			-0.931*	-0.551*		
			(0.550)	(0.314)		
AI-Low × Monthly Expense					-0.428	-0.255
					(0.266)	(0.158)
Intercept	2.151	0.995	0.357	-0.125	0.434	0.075
	(3.175)	(1.801)	(3.251)	(1.850)	(3.297)	(1.867)
Observations	175	175	175	175	175	175
Pseudo R-squared	0.277	0.279	0.294	0.296	0.292	0.294
Wald test p-value	0.007	0.005	0.132	0.122	0.097	0.091
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 2 is answered correctly; the independent variables include treatment dummies and their interactions with education and monthly expense, respectively. Control variables are the same as those in Table 1. All regressions control for company fixed effect. All samples are included.

Table A.7: OLS regressions on writing tasks: AI-Low vs. AI-High

	Writing Task 1				Writing Task 2			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-High	2.836	28.834**	-1.536	5.483	-1.522	10.624	5.044	-2.101
	(3.128)	(13.329)	(8.198)	(5.374)	(3.011)	(10.060)	(6.341)	(4.599)
AI-High × Education		-8.509*				-3.980		
		(4.368)				(3.051)		
AI-High × Monthly Expense			0.990				-1.498	
			(1.751)				(1.511)	
AI-High × Similar Experience				-3.670				1.025
				(6.388)				(6.401)
Observations	115	115	115	115	115	115	115	115
R-squared	0.313	0.349	0.316	0.318	0.160	0.171	0.168	0.171
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is writing scores; the independent variables include the AI-High dummy, indicating whether the subject is in AI-High or AI-Low, and its interaction terms with Education, Monthly Expense, and previous Similar Experience, respectively. Control variables are the same as those in Table 1. All subjects in AI-High and AI-Low are included.

Table A.8: OLS regressions on calculation accuracy improvement from stage 1 to 2: NO-AI vs. AI-Low

	Calculation accuracy					
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	-0.022	-0.050	-0.047	0.083	0.378	0.109
	(0.086)	(0.081)	(0.083)	(0.231)	(0.319)	(0.161)
Stage 2	-0.250***	0.116	0.157	-0.174	-0.234	0.141
	(0.088)	(0.613)	(0.592)	(0.651)	(0.611)	(0.629)
AI-Low × Stage2	0.047	0.033	0.032	0.204	0.113	-0.214
	(0.126)	(0.118)	(0.118)	(0.351)	(0.422)	(0.216)
AI-Low × Stage2 × Monthly Expense				-0.032		
				(0.074)		
AI-Low × Stage2 × Education					-0.028	
					(0.133)	
AI-Low × Stage2 × Similar Experience						0.344
						(0.256)
Observations	238	238	238	238	238	238
R-squared	0.05	0.29	0.30	0.31	0.32	0.31
Controls	No	Yes	Yes	Yes	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is writing scores; the independent variables include the AI-High dummy, indicating whether the subject is in AI-High or AI-Low, and its interaction terms with Education, Monthly Expense, and previous Similar Experience, respectively. Control variables include those in Table 1 and the amount of time spent in stage 2. All subjects in AI-High and AI-Low are included.

Table A.9: OLS regression analysis on writing performance improvement from stage 1 to 2: NO-AI vs. AI-Low

	Writing performance					
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Low	-4.005	-3.700	-3.596	-1.038	-2.864	-4.726
	(2.796)	(2.468)	(2.448)	(6.416)	(9.565)	(4.919)
Stage 2	-0.820	9.764	11.246	0.556	15.260	14.895
	(2.680)	(9.365)	(9.788)	(11.575)	(12.454)	(10.533)
AI-Low × Stage 2	15.206***	14.772***	14.760***	16.630**	27.279**	17.606***
	(3.562)	(3.219)	(3.198)	(8.237)	(11.623)	(5.785)
AI-Low × Stage 2 × Monthly Expense				-0.191		
				(1.680)		
AI-Low × Stage 2 × Education					-4.048	
					(3.527)	
AI-Low × Stage 2 × Similar Experience						-3.499
						(6.880)
Observations	238	238	238	238	238	238
R-squared	0.13	0.41	0.42	0.43	0.44	0.43
Controls	No	Yes	Yes	Yes	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly; the independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-Low, indicating whether the subject is in No-AI or AI-Low, and their interaction term. Control variables include those in Table 1 and the amount of time spent in stage 2. All subjects in NO-AI and AI-Low are included.

Table A.10: Logit regressions for calculation performance dynamics: AI-Low vs. AI-High

	(1)	(2)	(3)	(4)	(5)	(6)
AI-High	-0.383	-0.408	-0.430	-3.658***	-3.295*	-1.007
	(0.390)	(0.483)	(0.483)	(1.335)	(1.947)	(0.832)
Stage 2	-0.846**	3.473	3.535	2.547	1.969	1.466
	(0.382)	(2.800)	(2.768)	(2.854)	(3.252)	(3.191)
AI-High × Stage 2	1.402**	1.872***	1.882***	5.590***	5.094**	18.871***
	(0.555)	(0.673)	(0.672)	(1.824)	(2.560)	(1.259)
AI-High × Stage 2 × Monthly Expense				-0.824**		
				(0.381)		
AI-High × Stage 2 × Education					-1.092	
					(0.864)	
AI-High × Stage 2 × Similar Experience						-18.053***
						(1.433)
Observations	230	230	230	230	230	230
Pseudo R-squared	0.03	0.22	0.22	0.24	0.22	0.27
Controls	No	Yes	Yes	Yes	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly. The independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-High, indicating whether the subject is in AI-High or AI-Low, and their interaction term. Regression (4)-(7) include triple interactions with Mon. Exp., which stands for monthly expense, Education, and Sim. Exp., which stands for similar previous writing experience, respectively. Control variables include those in Table 1, the word-count of the subject's submitted writing at the same stage, reflecting the effort on writing, the duration in seconds that the subject spent in the corresponding stage, word-count in the interaction with the AI, and word-count of the AI's last response. All subjects in AI-High and AI-Low are included.

Table A.11: Probit regressions for calculation performance dynamics: AI-Low vs. AI-High

	(1)	(2)	(3)	(4)	(5)	(6)
AI-High	-0.236	-0.279	-0.291	-2.254***	-1.972*	-0.639
	(0.240)	(0.275)	(0.277)	(0.755)	(1.111)	(0.479)
Stage 2	-0.526**	1.967	2.052	1.502	1.073	0.952
	(0.236)	(1.616)	(1.610)	(1.685)	(1.844)	(1.786)
AI-High × Stage 2	0.866**	1.081***	1.092***	3.292***	3.084**	7.091***
	(0.341)	(0.384)	(0.383)	(1.034)	(1.469)	(0.680)
AI-High × Stage 2 × Monthly Expense				-0.488**		
				(0.219)		
AI-High × Stage 2 × Education					-0.669	
					(0.487)	
AI-High × Stage 2 × Similar Experience						-6.598***
						(0.789)
Observations	230	230	230	230	230	230
Pseudo R-squared	0.03	0.21	0.22	0.24	0.22	0.27
Controls	No	Yes	Yes	Yes	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes

Note: robust standard errors are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly. The independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-High, indicating whether the subject is in AI-High or AI-Low, and their interaction term. Regression (4)-(7) include triple interactions with Mon. Exp., which stands for monthly expense, Education, and Sim. Exp., which stands for similar previous writing experience, respectively. Control variables include those in Table 1, the word-count of the subject's submitted writing at the same stage, reflecting the effort on writing, the duration in seconds that the subject spent in the corresponding stage, word-count in the interaction with the AI, and word-count of the AI's last response. All subjects in AI-High and AI-Low are included.

Table A.12: Logit & Probit Analysis for Calculation Performance Dynamics: NO-AI vs. AI-Low

	(1) Logit	(2) Probit
AI-Low	-0.210	-0.121
	(0.443)	(0.260)
Stage 2	-1.207***	-0.733***
	(0.371)	(0.222)
AI-Low × Stage2	0.245	0.158
	(0.543)	(0.324)
Intercept	-0.907	-0.516
	(1.643)	(1.010)
Observations	238	238
Pseudo R-squared	0.132	0.132
Controls	Yes	Yes
Session FE	Yes	Yes

Note: Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly; the independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-Low, indicating whether the subject is in NO-AI or AI-Low, and their interaction term. Control variables are the same as those in Table 1. All regressions control for company fixed effect. All subjects in NO-AI and AI-Low are included.

B. Further analysis

B.1. Submission time

First, we explore if the heterogeneous spillover effect is driven by those who submitted too soon or too late.

Figure B.1 plots the empirical distribution of Stage-2 total time (seconds) by treatment (AI-Low vs. AI-High). Both groups display a pronounced spike at the 15-minute deadline (900s), indicating many subjects submitted at the cap. Apart from this cap-mass, the two distributions look very similar, with only a small share of observations in the short-duration tail. This motivates restricting the DID sample to a common same-time band that excludes extreme durations.

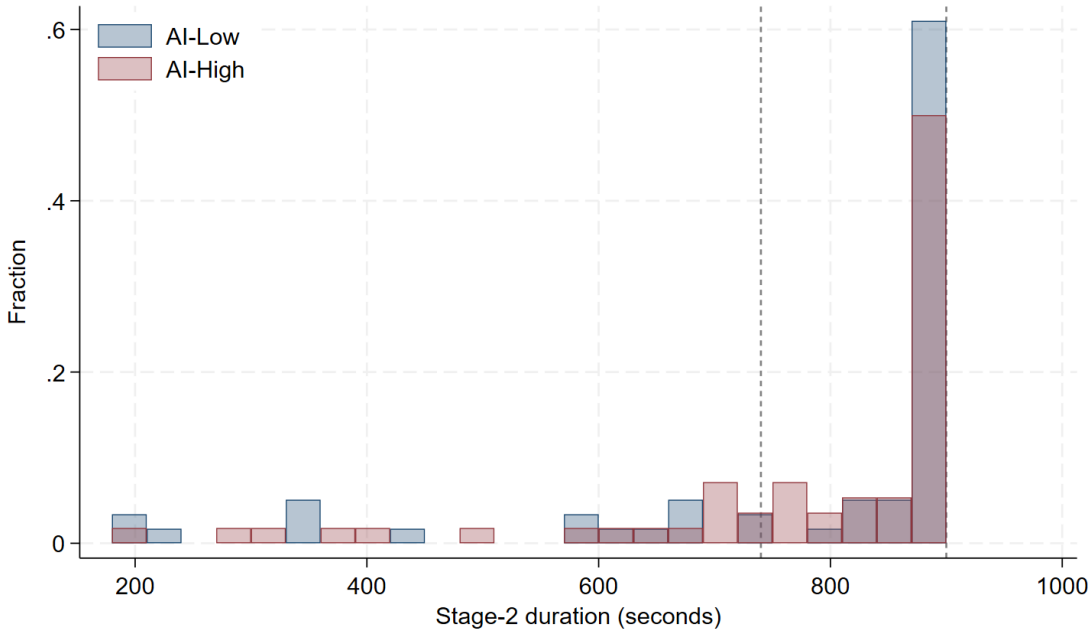


Figure B.1: Stage 2 Duration distribution by treatment: AI-Low vs. AI-High. The common time band is between $[740, 900]$ seconds.

Table B.1 reports OLS DID regressions after restricting the sample to the interquartile range of Stage-2 total time (25th-75th percentiles). This reduces the sample from 230 to 172 observations. Across specifications with and without controls and session fixed effects, the key DID coefficient, AI-High × Stage 2, remains positive and statistically significant (e.g., 0.330-0.338 in cols. 1-3; 0.711-1.398 in cols. 4-7; 1.060 in col. 6). Thus, the main spillover—greater calculation improvement in AI-High vs. AI-Low from Stage 1 to 2—survives when comparing subjects who spent comparable time in Stage 2.

Turning to moderators inside this band: the triple interaction with Similar Experience (AI-High × Stage 2 × Similar Experience) is negative and significant (0.980, $p<0.01$ in col. 6), indicating that the spillover is stronger for those without prior similar writing experience. Monthly expense is not significant (col. 4), while Education shows a marginally negative coefficient (0.347, $p<0.10$, col. 5).

Table B.1: DiD OLS analysis controlling for same-time band

	Calculation accuracy						Writing performance		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI-High	-0.148	-0.097	-0.095	-0.571**	-0.768*	-0.306*	4.664	2.737	2.934
	(0.104)	(0.105)	(0.107)	(0.285)	(0.450)	(0.177)	(3.315)	(3.172)	(3.154)
Stage 2	24.196	8.220	8.464	4.796	10.981	11.631	-846.989	-646.617	-640.697
	(20.585)	(21.678)	(21.684)	(23.205)	(21.568)	(19.109)	(528.243)	(590.606)	(601.941)
AI-High × Stage2	0.330**	0.335**	0.338**	0.711**	1.398**	1.060***	-3.031	-2.888	-2.993
	(0.148)	(0.138)	(0.139)	(0.355)	(0.577)	(0.228)	(4.095)	(4.016)	(4.031)
AI-High × Stage2 × Mon. Exp.				-0.084					
				(0.079)					
AI-High × Stage2 × Education					-0.347*				
					(0.180)				
AI-High × Stage2 × Sim. Exp.						-0.980***			
						(0.277)			
Observations	172	172	172	172	172	172	172	172	172
R-squared	0.09	0.31	0.31	0.33	0.38	0.34	0.27	0.45	0.45
Controls	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Session FE	No	No	Yes	Yes	Yes	Yes	No	No	Yes

Note: robust standard errors are reported in parenthese. The sample is restricted to be within 25 and 75 percentile of the stage 2 total time spent. Note that 75 percentile is the 15 minutes cap as a substantial percentage of subjects still submitted at the cap. The dependent variable is whether a calculation task in stage 1 or 2 is answered correctly for (1)-(6) and writing scores in (7)-(9). The independent variables include Stage 2, indicating whether the calculation task is in stage 2 or 1, AI-High, indicating whether the subject is in AI-High or AI-Low, and their interaction term. Regression (4)-(6) include triple interactions with Mon. Exp., which stands for monthly expense, Education, and Sim. Exp., which stands for similar previous writing experience, respectively. Control variables include those in Table 1, the word-count of the subject's submitted writing at the same stage, reflecting the effort on writing, the duration in seconds that the subject spent in the corresponding stage, word-count in the interaction with the AI, and word-count of the AI's last response. All subjects in AI-High and AI-Low are included.

B.2. Time analysis

First stage time analysis. The average time spent in the stage was 14'50" for NO-AI, 14'58" for AI-Low, and 14'59" for AI-High; pairwise t-tests all yield $p > 0.05$, except for the comparison between NO-AI and AI-High, where $p = 0.075$. This similarity in effort is consistent with the similar writing performance across treatments and is likely due to the strict 15-minute time limit imposed for each task stage. Most subjects submitted their work just before the deadline, or it was automatically submitted when the time expired.

Second stage time analysis. The average time spent in the second stage is significantly lower when AI assistance is available: 14'36" in the NO-AI group, compared to 13'03" in AI-Low and 13'01" in AI-High ($p < 0.02$ for both comparisons using the Wilcoxon rank-sum tests). There is no statistically significant difference between the two AI-assisted groups, which is consistent with their similar average scores.

Third stage time analysis. The stronger incentive did lead to increased effort in the third stage: subjects in AI-High spent significantly more time on the task (12'45") compared to those in AI-Low (10'34"), with $p < 0.01$ for Wilcoxon rank-sum test. In addition, both AI-assisted groups spent less time than those in the NO-AI group (14'38"), with $p < 0.05$ for all comparisons. Given that section 3.1 shows no statistical difference in the writing performance between AI-assisted groups, the findings illustrate a *substitution effect between AI usage and economic incentives* in boosting productivity.

C. Experimental Instructions

This section presents the majority of experimental instructions translated into English.

Page 1: Welcome Page

Dear colleagues from the XXXX company,

Thank you very much for participating in this experiment! We deeply appreciate your time and involvement. The goal of this experiment is to study how generative artificial intelligence affects work efficiency under different workplace scenarios.

You will receive a base participation fee of 50 RMB for fully participating the experiment. Additionally, you may earn extra compensation depending on your own performance and that of your colleagues during the experiment.

Please stay focused and take each task seriously. Your understanding of the experiment rules will directly impact your compensation.

To ensure the success of the experiment, please strictly follow the rules. Any rule violation may negatively impact the results and your earnings. If you have any questions, feel free to ask the experiment staff.

Thank you again for your participation. We wish you the best in completing the tasks successfully!

Page 2: Participant Information Entry

- Please enter your **full name** (used to process payment).
- Please enter your **seat number** (your current seat in the room).

Page 3: Experiment Rules

- Please silence or power off your mobile phone and place it in the designated area. It will be securely stored and returned after the experiment.
- Please also hand over any other personal electronic devices (e.g., smartwatches, tablets). No personal electronics may be used.
- Do not speak or communicate with others during the experiment. All tasks must be completed independently.
- If you need help, raise your hand to get assistance from the experiment staff.
- Follow the instructions and tasks displayed on the iPad. Each task has a strict time limit.
- At the end of the experiment, you will be notified to leave and your personal belongings will be returned.

Click confirm to proceed.

Page 4: Pre-Experiment Instructions

(To subjects in AI-Low and AI-High Groups)

You will now receive a 15-minute training session on how to use generative AI in the workplace.

After the training, you will complete three simulated work tasks. You will receive compensation for each task based on performance.

(To subjects in NO-AI Group)

You will complete three simulated work tasks. You will receive compensation for each task based on performance.

Page 5: Task 1 Instructions and Comprehension Check

You will see a piece of text and be asked to write a 300–500 word business report.

Over five industry professionals will evaluate your report based on *usefulness, accuracy, logic, and originality.* Your score will be the average of their ratings.

Reward Conditions:

- *(To subjects in NO-AI & AI-Low:)* 100 RMB if your score is in the top 50% among all subjects in today's session.
- *(To subjects in AI-High:)* 200 RMB if your score is in the top 50% among all subjects in today's session.

(To subjects in AI-Low & AI-High:) AI Use Restriction: You may *NOT* use generative AI for this task. Violations will result in loss of all rewards.

Comprehension Questions:

1. What is the highest possible reward from Task 1?
   A. 200 RMB, B. 100 RMB, C. 150 RMB, D. 50 RMB
2. On what criteria will your writing be evaluated?
   A. Usefulness, B. Accuracy, C. Logic, D. Originality
3. If your score is in the bottom 40%, how much will you earn?
   A. 0 RMB, B. 100 RMB, C. 200 RMB, D. 50 RMB
4. If your score is in the top 10%, how much will you earn?
   A. 0 RMB, B. 100 RMB, C. 150 RMB, D. 200 RMB
5. (AI-Low & AI-High): What happens if you use generative AI in Task 1?
   A. You lose all rewards, B. No influences, C. You obtain some rewards, D. You lose the opportunity to use generative AI.

All questions must be answered correctly to proceed.

Page 6: Task 1 Writing (15-minute countdown)

You will now see the material for the first work task.

You are required to complete a business report (300–500 words) within 15 minutes. Please write your response in the text box at the bottom and click submit. If time runs out, your input will be automatically submitted.

Note: You may be interrupted during the task and asked to complete a short task. Please complete the short task on the provided sheet and submit it to one of the experimenters before submitting your report.

Task 1 Scenario: WanderWorld is a well-known online travel agency specializing in luxury vacation packages. As an employee, you need to prepare a Q2 business summary.

- Revenue grew 10% from last quarter, but net profit only increased 3% due to higher costs.
- A new "Southeast Asia Tropical Island Tour" package met 120% of booking expectations.
- Ad spending exceeded budget by 20%, but conversion rates rose by 15%.
- Complaints about the "Luxury Europe Tour" package included poor itinerary, subpar accommodation, and low-quality guide service, causing negative publicity on social media.
- A competitor launched a new customized luxury tour product, taking away market share.

Writing Task:

- Financial summary and market performance of the new tour package
- Effectiveness of marketing activities and new strategy suggestions
- Customer service issues and proposed solutions

Time Limit: 15 minutes

Word Count: 300–500 words

Enter your response in the text box below.

(For subjects in AI-Low & AI-High Groups): You may refer to your AI training but must *not* use generative AI for this task.

Midway Disruption Task (Smartwatch Sales)

Six minutes after Task 1 has started, the experimenter calls for the attention of all subjects and introduces the following calculation task, shown by a presenter in the front of the lab.

Scenario: Your supervisor asks you to immediately report the total number of smartwatches sold in July, August, and September.

You open the internal sales report and see the following data:

Online Store Sales:

- July: 300 units
- August: 250 units
- September: 350 units (includes 5 returns)

Offline Store Sales:

- July: 120 units
- August: 110 units
- September: 130 units

Partner Channel Sales:

- July: 50 units
- August: 45 units
- September: 60 units (10 demo units not included in final sales)

Task: Add up all sales across channels for July–September, adjusting for returns and demo units. Enter the correct total at the bottom of the page.

Incorrect totals will result in a 5 RMB deduction from your compensation.

Page 7: Task 2 Instructions

You will now begin Task 2.

Reward Conditions:

- *NO-AI & AI-Low:* 100 RMB if your score is in the top 50%
- *AI-High:* 200 RMB if your score is in the top 50%

(AI-Low & AI-High Groups): You may use the generative AI platform for this task. Apply what you learned in the training to improve your efficiency.

Click confirm to see the task material.

Page 8: Task 2 Writing (15-minute countdown)

You will now see the material for the second work task.

Write a business report (300–500 words) based on the scenario below. If time runs out, your input will be auto-submitted.

Note: You may be interrupted during the task and asked to complete a short task. Please complete the short task on the provided sheet and submit it to one of the experimenters before submitting your report.

Task 2 Scenario: Bear Electric is a company that sells kitchen and home appliances.

Q2 Business Highlights:

- Sales increased by 8%, but net profit only rose 3% due to rising production costs.
- The new "Smart Rice Cooker" exceeded expectations, reaching 125% of the sales target.
- Online marketing expenses exceeded budget by 20%, but customer conversion rate improved by 15%.
- Some customers complained about the rice cooker's complexity and slow customer service response.

Writing Task: Write a Q2 business summary that includes:

- Financial performance
- Market performance and expectations for the new product
- Effectiveness and improvements in marketing strategy
- Customer complaint analysis and solutions

Time Limit: 15 minutes

Word Count: 300–500 words

Enter your response in the text box below.

(For subjects in AI-Low & AI-High Groups): You may use the generative AI tool for assistance.

Calculation Task (Holiday Beverage Restocking)

Six minutes after Task 2 has started, the experimenter calls for the attention of all subjects and introduces the following calculation task, shown by a presenter in the front of the lab.

Scenario: You are the purchasing manager of a convenience store chain in City C. With the National Day holiday approaching, you must ensure sufficient beverage stock for three stores.

You check recent sales data and current inventory:

Expected Holiday Beverage Sales:

- Store 1: 150 bottles
- Store 2: 200 bottles
- Store 3: 180 bottles

Current Inventory:

- Store 1: 50 bottles
- Store 2: 60 bottles
- Store 3: 40 bottles

Supplier Requirements:

- Minimum delivery per order: 30 bottles

- Must be in multiples of 10

Task: Calculate the required restocking amount for each store to meet demand and comply with supplier rules.

Enter the restocking amount per store and the total.

Incorrect restocking plans may lead to holiday shortages and result in a 5 RMB penalty.

- Store 1 Restock: — bottles
- Store 2 Restock: — bottles
- Store 3 Restock: — bottles
- Total Restock: — bottles

Page 9: Task 3 Instructions

You will now begin Task 3, the final writing task.

Reward Conditions:

- *NO-AI & AI-Low:* 100 RMB if your score is in the top 50%
- *AI-High:* 200 RMB if your score is in the top 50%

There will be *no interferences* during this task.

(AI-Low & AI-High Groups): You may use the generative AI platform. Apply the skills learned in training to improve your work.

Click confirm to view the task materials.

Page 10: Task 3 Writing

Task 3 Scenario: Red Star Furniture is a budget-friendly furniture company offering diversified home solutions.

Q2 Operational Highlights:

- Sales increased by 10% compared to the previous quarter, but net profit grew by only 4% due to increased logistics and material costs.
- The new "Full House Customization" service received positive feedback and exceeded expectations with 115% of projected orders.
- Offline promotion events exceeded budget by 10%, resulting in an 18% increase in customer acquisition.
- Recent complaints focused on delivery delays and poor installation services.

Writing Task: Write a Q2 business summary report covering:

- Financial performance
- Market performance and outlook for the new service
- Promotion campaign evaluation and suggestions for improvement
- Customer service issues and solutions

Time Limit: 15 minutes

Word Count: 300–500 words

Enter your response in the text box below.

Page 11: AI Training for NO-AI Group

(NO-AI Group Only)

Now, you will receive a brief training on how to use generative AI in the workplace to enhance productivity.

Page 12: Final Survey

Please complete the following short survey:

1. Have you performed similar tasks in your daily work? (Yes/No)
2. To what extent did generative AI help reduce the performance loss caused by interferences during the tasks? (0–10)
3. How useful was generative AI in helping you complete the three tasks? (0–10)
4. To what extent can generative AI improve your productivity in daily work? (0–10)
5. What percentage of your daily tasks could be assisted by generative AI? (Input %)
6. Do you think generative AI could replace your current job?
   Scale: 0 (Not at all) to 10 (Completely)

Page 13: Experiment Complete

The experiment has ended.

Thank you for your participation!

D. Generative AI training

This section illustrates the main points of the generative AI training in our experiment. The objective of the training is to familiarize subjects how to use Generative AI to improve their productivity and writing quality. The training is a 15-20 minutes lecture given by one of the authors.

Core Concepts and Key Elements of Prompts

1. Definition of a Prompt

A prompt is a question or instruction given during interaction with generative AI, specifying the desired content, format, and level of detail for the output. The quality of a prompt determines the quality of the results—the more complete and rich the information, the better the output.

2. Main Elements of a Prompt

- **Instruction:** The specific task or command you want the AI to perform.
- **Context:** Additional background information to help the AI understand the task.
- **Input Data:** The raw content or question provided by the user.
- **Output Indicator:** Requirements for the type and format of the output.

Principles for Designing High-Quality Prompts

1. Clear Instructions and Context

Use explicit action verbs (e.g., "write", "summarize") and clear delimiters (e.g., ###) to separate task descriptions from content.

2. Specify Output Requirements

Clearly indicate the output type, format, background, scope, length, and style. Examples may include summarizing multiple resumes into a table or writing a structured multi-paragraph report.

3. Start Simple and Iterate

Begin with concise prompts, gradually add details, and refine through iterations. Break down complex tasks into smaller sub-tasks and improve results over multiple interactions.

4. Be Specific and Focused

The more descriptive and detailed the prompt, the better the result. If format is critical, provide a sample or template.

5. Emphasize What To Do

Avoid negative instructions about what not to do; instead, state clearly what should be done (e.g., "Extract keywords" rather than "Do not include irrelevant words").

6. Provide Context for Accuracy

The richer the background information, the more accurate the AI's response will be. Without sufficient context, the model can only make limited inferences.

7. Set Roles or Tone

Instruct the AI to adopt a specific role (e.g., analyst, teacher) or tone (e.g., formal, humorous) to match the desired style.

Key Framework: CO-STAR

The **CO-STAR** framework outlines six essential elements for constructing effective prompts:

- **C - Context:** Provide background to clarify the situation.
- **O - Objective:** Clearly define the purpose of the task.
- **S - Style:** Specify the desired writing or communication style (e.g., CEO report, data analysis).
- **T - Tone:** Set the tone of the response (e.g., formal, humorous, persuasive).
- **A - Audience:** Define the target reader (e.g., industry expert, general public, child).
- **R - Response Format:** Specify the desired format of the output (e.g., list, table, JSON).