The First Draft Problem: Productivity, Quality, and Physiological Strain in LLM-Assisted Writing

Xufei Liu
The Wharton School, xufei@upenn.edu

Hummy Song
The Wharton School, hummy@wharton.upenn.edu

Christian Terwiesch
The Wharton School, terwiesch@wharton.upenn.edu

Abstract. The use of large language models (LLMs) has grown rapidly over the past few years, especially in health care, where documentation burden remains a persistent problem for practicing clinicians. However, despite the deployment of many ambient scribes for documentation, it is still unclear how LLMs impact those who work with them. Across two preregistered studies, we examine how this shift—from writing documentation to editing an LLM-generated draft—impacts workers. We study this "first draft problem": eliminating the blank page does not remove the work of documentation but relocates it into revising and editing. In Study 1, an online experiment with 16 clinicians producing documentation following a teletherapy patient encounter, we establish the puzzle: counter to our hypotheses, editing an LLM-generated draft of the documentation as opposed to writing it from scratch does *not* result in time savings or quality improvements on average. Instead, while some clinicians experience no time savings but large improvements in quality after making substantial revisions, others experience meaningful time savings and no change in quality after making limited revisions. In other words, we observe significant heterogeneity in how clinicians engage with the same draft, which leads to variation in productivity and quality. In Study 2, a lab experiment with 100 university students summarizing video interviews, we unpack the mechanisms underlying this heterogeneity using granular keystroke logs, validated self-reports of workload, and micro-level physiological and eye-tracking measurements. In this setting, we find the expected productivity improvements: editing reduces task completion time by roughly 70%, raises quality, and lowers self-reported workload. Yet the physiological evidence tells a more nuanced story. Physiological markers indicate that fatigue accumulates faster when editing an LLM-generated draft than when writing from scratch. Thus, we find that (a) time savings can be real, but are setting dependent; (b) quality improvements are possible, but the relationship between how heavily a draft is revised and the quality of the result differs sharply across our two settings; and (c) while editing an LLM draft lowers reported cognitive load, the body still tires faster over time.

Funding: This work was supported by the Wharton Behavioral Lab, the Mack Institute for Innovation Management, and the Wharton AI & Analytics Initiative.

Acknowledgments: The authors thank Maurice Schweitzer and seminar participants at the 2026 INFORMS Healthcare Conference for helpful comments; Dr. Muniya Khanna for assistance with Study 1 materials and clinician recruitment; and Robert Botto and the team at the Wharton Behavioral Lab for assistance in setting up Study 2.

Key words: Human-AI Collaboration, Large Language Models, Health Care Operations, Behavioral Operations, Cognitive Load

History: Updated August 19, 2026

1. Introduction

From lawyers to scientists and from journalists to clinicians, many professions in today's knowledge economy spend a large part of their work writing documents. Given this importance of writing and the large opportunity costs of the writers, the past decades have produced a stream of technologies aiming to improve the productivity of the writing process. Word processors, real-time spell and grammar checking, and voice dictation are examples of such technologies, yet they all pale in comparison to the latest innovation in this area—large language models (LLMs). With the widespread adoption of ChatGPT and similar chatbots built on LLM technology, millions of lay and professional writers face the same question: "Should I write the document from scratch (a blank page) or should I ask an LLM to propose a first draft and then focus on editing and improving it?"

We empirically study this choice, writing versus editing, in the domain of health care. Although LLMs have many potential applications in health care operations, including diagnostics and treatment (Meincke et al. 2026, Su et al. 2025), the task of drafting clinical documentation is widely seen as the least controversial and lowest-hanging fruit (Roberts 2024). Rather than summarizing a patient encounter by starting with a blank page, clinicians—in our case, mental health professionals providing teletherapy—use an audio recorder to capture what was said during the encounter (a practice referred to as ambient listening). They then rely on an LLM to turn this audio into a first draft of the notes, which they can decide to edit or not. Such an LLM-assisted workflow (Patwardhan et al. 2025) holds the promise of saving time, improving productivity, and thereby helping to address the ongoing burnout and attrition problem of the clinical workforce (Apathy et al. 2023). Given this potential, many such applications are currently being piloted in practice (Duggan et al. 2025, Olson et al. 2025, Sanmark et al. 2025).

From an operations management perspective, the documentation task of preparing the visit summary note consists of two subtasks: (1) writing an initial note and (2) revising and editing the note until a certain quality target is met. The promise of the LLM technology as applied to this type of clinical documentation is simple: eliminate the first subtask in the hope that the time spent on the second does not grow enough to offset the savings. We call this tension the "first draft problem": eliminating the blank page by using an LLM does not remove the work of documentation, but rather relocates it into the second subtask of revising and editing. The chief aim of our work is to compare (a) the workflow resulting from clinicians writing the documentation from scratch (henceforth referred to as the "writing condition") with (b) the workflow of using an LLM to generate a first

draft and then editing and revising the document until the clinician is willing to accept the quality of the document (henceforth referred to as the "editing condition").

Whereas prior studies focus primarily on task completion time, we compare the writing condition and the editing condition across four dimensions:

1. *Productivity*: Following the argument of several prior studies investigating the impact of LLMs on productivity, including customer support (Ni et al. 2025, Brynjolfsson et al. 2025), idea generation (Meincke et al. 2024), and management consulting (Dell'Acqua et al. 2026), we compare the total time taken to complete and submit the summary note across the two conditions.

2. *Quality*: In addition to measuring the time to complete the note, we compare the actual summary notes that are submitted under the two conditions. Specifically, we assess the length of the summary note and evaluate its content as our measure of documentation quality.

3. *Workflow*: With the introduction of the new LLM technology, workers are likely to adjust the way in which they organize and sequence their work. For example, they may work faster throughout, increase the idle time between subsequent cases, or spend more time revising than adding content. In other words, as we move from the writing to the editing condition, we expect more aspects of the workflow to change beyond just eliminating the subtask of writing the initial note. We track these shifts using granular keystroke and editing logs.

4. *Cognitive load and fatigue*: Engaging with the new LLM technology is also likely to impact the cognitive load and fatigue experienced by the workers. For example, the cognitively easy parts of the task may be accomplished by the LLM, leaving the more effortful work for the worker. This would raise concerns about provider well-being and long-term sustainability, with direct implications for how managers set case loads after adopting LLM technologies. Furthermore, a worker's physiological response may diverge from their perceived cognitive load in important ways. We assess each of these aspects across the two conditions.

The empirical results reported in this paper come from two preregistered experiments. In Study 1, we recruited 16 experienced clinicians providing teletherapy care. Each clinician completed two documentation cases—one writing from scratch and one editing an LLM-generated draft—with condition order and case assignment randomized. We then observed their productivity and the quality of the documentation they produced.

While Study 1 focuses on replicating an experimental setting as close to the real work environment as possible (with real clinicians, real patient cases, and a clinically relevant task), the focus of Study 2 is on obtaining data at a more granular level to assess changes in workflow and in cognitive

load and fatigue. For that, we created an artificial documentation task of preparing a video interview summary that was completed by 100 students in a laboratory environment. In this controlled setting, we measure and track keystrokes to capture workflow. We also utilize biosensors (including video-based eye trackers, ear-clip photoplethysmography (PPG) sensors, and electrodermal activity electrodes) to measure physiological responses such as heart rate, pupil dilation, and fixation length to assess cognitive load and fatigue.

Applying this four-pronged framework, with its novel perspective on the micro-level execution of the documentation task, we establish four findings.

First, on productivity, the time savings promised by LLM assistance are real but highly setting-dependent. In Study 1, editing an LLM-generated draft yields no reduction in average completion time relative to writing from scratch; instead, we find large heterogeneity across clinicians, with some completing the note faster by writing and others faster by editing. In Study 2, the gain predicted by prior work does materialize: students working from an LLM draft complete tasks in roughly 70% less time on average.

Second, on quality, the gains are possible but not uniform. In Study 1, there is no significant difference in documentation quality on average, yet among clinicians who edit, those who revise the draft more heavily produce higher-quality notes. In Study 2 this relationship reverses: documentation quality is higher on average when editing an LLM-generated summary compared to writing from scratch, and students who revise more heavily produce lower-quality summaries.

Third, on workflow, the shift from author to editor changes not only how much time the task takes but how that time is spent. Keystroke logs from Study 2 show that editing reallocates effort away from composing text and toward reviewing and selectively revising the draft. When writing from scratch, students spend 55% of task time actively producing text, 38% reviewing the text without altering, and 7% deleting text; when editing, these shares change to 18%, 76%, and 6% respectively, and students navigate the document at more than double the click rate. The text work that remains tilts from production toward correction: students delete 81 characters for every 100 characters they add when editing, compared with 14 when writing. Yet these revisions are selective rather than extensive—the median submission removes only 1.4% of the draft's characters and retains a cosine similarity of 0.98 to the LLM draft—confirming that the transition reorganizes the task rather than simply removing its first subtask.

Fourth, on cognitive load and fatigue, perception and physiology diverge. Students self-report substantially lower cognitive load when editing compared to writing. Our biosensor data, however,

tell a more cautionary story: fatigue accumulates faster when editing than when writing. In other words, editing feels easier even as several physiological markers suggest it is more taxing over sustained use.

Altogether, our findings show that LLM assistance does not simply speed up an unchanged task; it reconstitutes it, shifting the worker from author to editor and redistributing the costs and benefits of documentation across different margins. This is the first draft problem in practice: the blank page disappears, but the burden it represented does not. Rather, it is redistributed from writing into editing and from perceived effort into physiological strain. We do not dispute the claim that LLMs can help with clinical documentation and writing. Rather, we locate the uneven gains and hidden costs that determine where, and for whom, that help appears. In doing so, this paper makes three contributions: it introduces a methodology that pairs physiological and eye-tracking measures with validated self-reports to assess the cognitive load of LLM-assisted work; it shows that the productivity gains documented in controlled settings need not carry over to professional, high-stakes documentation; and it documents that the relationship between revision intensity and quality can reverse across settings, so that more editing is not reliably better editing.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature and develops our hypotheses. Section 3 presents Study 1, the online experiment with clinicians, and Section 4 presents Study 2, the lab experiment with students. Section 5 concludes.

2. Related Literature and Hypotheses

Our work sits at the intersection of three streams of literature. The first studies human-AI collaboration and task allocation in operations, examining how humans and algorithms divide, delegate, and monitor work (Fugener et al. 2022, 2025, Boyacı et al. 2024, Dell'Acqua 2022). We contribute micro-level evidence on what happens after the delegation: how workers reorganize their effort when the machine takes over the first draft. The second stream measures the productivity effects of LLMs in lab and field settings (Noy and Zhang 2023, Brynjolfsson et al. 2025, Ni et al. 2025, Dell'Acqua et al. 2026). Relative to this work, we show that average effects mask substantial heterogeneity across settings and workers, and that realized gains hinge on how users revise the draft. The third stream studies workload, fatigue, and their operational consequences in service settings (Corbett 2024, Bavafa and Jonasson 2023, Liu et al. 2026). We extend its measurement toolkit by pairing validated self-reports with continuous physiological and eye-tracking measures—and show that the two can diverge.

2.1. Time savings from LLM-generated drafts

The case for large time savings rests on a simple assumption: creating new text is the most time-consuming part of documentation, and the draft removes the need to do so. Noy and Zhang (2023) provide clean results in their lab experiment, where professionals given LLM drafts completed writing tasks much faster and with higher quality. Documentation settings should be especially favorable: such tasks are formulaic by nature, and frontier models approach expert parity (Patwardhan et al. 2025). Thus, if revisions are minimal, the time saved by not needing to write from scratch should be the dominant effect. Field evidence is consistent with this prediction—early deployments of ambient scribes reported large time savings and decreased burden (Duggan et al. 2025, Olson et al. 2025, Pearlman et al. 2025, Sanmark et al. 2025). We therefore predict:

HYPOTHESIS 1. *Editing an LLM-generated draft versus writing from scratch reduces the time needed to complete a documentation task.*

At the same time, when deployed at scale, productivity gains are heterogeneous across workers, tasks, and settings (Brynjolfsson et al. 2025, Ni et al. 2025, Dell'Acqua et al. 2026). This holds in documentation settings as well; models that write fluently still fall short in real clinical workflows (Kupferschmidt et al. 2025). The reason lies in what happens after the draft is handed over: it may still contain hallucinations, errors, and biases (Maynez et al. 2020, Shah 2024, Chen et al. 2025), and when the stakes of a missed error are high, revision absorbs more of the time the LLM should have saved. This tension is not new—speech recognition also promised time savings for documentation, but delivered mixed efficiency results while introducing errors for clinicians to catch (Hodgson and Coiera 2016). Our two settings differ along exactly this dimension of stakes: professional documentation by clinicians carries high stakes for errors while lab summaries by students do not. This contrast lets us examine whether productivity gains are consistent across settings and task difficulty.

2.2. Quality of the resulting documentation

When working with an LLM-generated draft, the average quality of documentation should increase. First, it should establish a floor for quality by providing the base documentation note; as a result, workers who would otherwise perform below that floor are raised to it (Noy and Zhang 2023, Bayer and Renou 2024, Cui et al. 2026). This is especially true as LLM-generated drafts approach expert parity on documentation tasks (Patwardhan et al. 2025). Second, a human-in-the-loop should make further edits only when they judge that those edits improve the draft. We therefore predict:

HYPOTHESIS 2. *Editing an LLM-generated draft versus writing from scratch improves the quality of the completed documentation.*

This prediction rests on the assumption that human revisions add value, which may not necessarily be true. In some prior work, research has shown that inexperienced workers using LLMs produce lower-quality revisions and worse outcomes (Otis et al. 2024, van Buchem et al. 2024).

2.3. Composition of work

When provided with an LLM-generated draft, the worker's role shifts from author to editor. How workers balance the resulting allocation of effort between writing and editing is a primary driver of their productivity with the draft (Fugener et al. 2022, 2025): when the model provides the first draft, the only work remaining is revision. In clinical documentation specifically, the volume of text written and revised is already tied to documentation burden, with more writing corresponding to greater burden and lower efficiency (Apathy et al. 2023).

While the first two hypotheses test aggregate outcomes, they do not test the mechanism that creates them. Two workers can produce similar quality and time savings through very different workflows. We therefore focus on the workflow itself as part of the impact of working with LLM-generated drafts and predict:

HYPOTHESIS 3. *Editing an LLM-generated draft versus writing from scratch shifts effort away from creating new text toward revising and reviewing existing text.*

2.4. Cognitive load and fatigue

Finally, if editing changes the workflow of documentation, it also changes the cognitive load that workflow creates. Cognitive load is an operational outcome in its own right: workers under higher load take longer on tasks and produce lower and more variable quality (Corbett 2024, Bavafa and Jonasson 2023), and sustained load is a key driver of attrition over time (Liu et al. 2026). The case that a draft lowers load is direct: LLM ambient documentation tools reduce the amount clinicians need to write (Duggan et al. 2025, Olson et al. 2025), and workers given an additional LLM teammate report more positive emotions (Dell'Acqua et al. 2025).

That case has two distinct implications, corresponding to two related but separable constructs. *Cognitive load* is the demand a task places on the worker at a given moment: if the draft removes the work of composing text, each task should feel less demanding while it is performed. *Fatigue* is the state that develops with time on task, as sustained effort erodes the worker's capacity to mobilize attention. Documentation is sustained work—clinicians complete case after case—so if a

draft genuinely lightens the work, it should not only lower the load of each task but also slow the rate at which fatigue builds across them. We therefore predict:

HYPOTHESIS 4A. *Editing an LLM-generated draft versus writing from scratch lowers the cognitive load perceived by the worker.*

HYPOTHESIS 4B. *Editing an LLM-generated draft versus writing from scratch slows the rate at which fatigue accumulates over sustained work.*

Testing these predictions requires measuring both constructs. Existing evidence on LLMs and worker well-being relies primarily on self-reports, so we measure perceived cognitive load using the NASA-TLX and Paas mental effort scales (Spiliotopoulou et al. 2026, Hart and Staveland 1988, Paas 1992). Fatigue, by contrast, is a trajectory rather than a level, and we track it through physiological signals recorded continuously as the work unfolds. Appendix A describes these measures and their interpretation. Importantly, the two types of measures need not agree: perceived effort and physiological strain can diverge, and we treat them as complementary rather than interchangeable.

The counterargument to both hypotheses is that editing carries cognitive load costs of its own, as workers still need to revise a draft they did not write. Revising work that one did not write creates higher cognitive load than revising one's own work (Boyacı et al. 2024). Sustained LLM collaboration is also associated with reduced well-being through a lack of human connection, which itself feeds back into cognitive load (Tang et al. 2023, Kim and Lee 2024, Spiliotopoulou et al. 2026). Finally, reviewing a draft that looks "good enough" can invite workers to "fall asleep at the wheel" (Dell'Acqua 2022); supervision may look passive, but it requires sustained vigilance that is difficult to maintain over time (Shaw and Nave 2026).

On balance, however, we expect the provision of an LLM draft both to lower cognitive load and to slow the accumulation of fatigue: eliminating the need to write should outweigh the added load of supervision and revision.

3. Study 1: Online Study with Clinicians

Study 1 with clinicians and Study 2 with students share the same within-participant design: each participant completes one task writing from scratch and another editing an LLM-generated draft. Each hypothesis is thus tested through within-participant contrasts between the two conditions.

3.1. Study Design and Sample Population

Our first experiment is an online study with mental health clinicians providing teletherapy. We recruit 16 licensed and practicing clinicians and ask them to complete a two-hour study remotely on their own computers. They are compensated at $200 per hour ($400 total), which is delivered as an Amazon gift card after the completion of the study. Clinicians are also told that all responses will be reviewed for quality and that the top two participants will receive a $100 bonus. The study took place from January 2025 to March 2025, and the design, sample size, and analysis plan were preregistered on AsPredicted (#207660).[1]

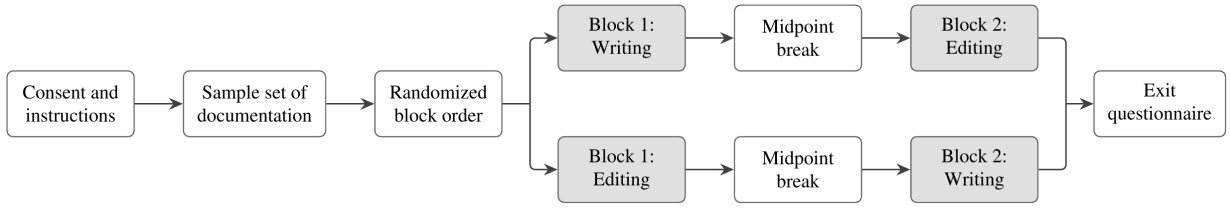
All recruited clinicians actively practice cognitive behavioral therapy (CBT), an evidence-based modality primarily used for treating depression and anxiety. In practice, these clinicians perform a teletherapy patient intake interview on their first encounter. Each interview is an hour long and consists of a conversation between a patient and a clinician, following CBT guidelines. Specifically, the teletherapy intake interview follows the Anxiety and Related Disorders Interview Schedule (ADIS), a semi-structured clinical interview designed to diagnose anxiety, depression, and other related disorders. After the intake interview is conducted, clinicians are responsible for two pieces of documentation: (i) a patient interview summary, which we refer to as the case summary, and (ii) a list of patient top problems, which we refer to as top problems. The case summary is a 300 to 600 word summary of the teletherapy interview covering the main takeaways and concerns. The top problems are a ranked list of up to three problems (such as social anxiety or depression), each given a severity score from 1 to 10, with 10 being the highest severity.

Clinicians are asked to complete two sets of documentation for two different teletherapy patient cases. As described above, this includes a case summary and top problems. Each patient case consists of a complete transcript of a real, de-identified teletherapy intake interview between an adolescent and the practicing clinician.

The entire study flow can be seen in Figure 1. At the very beginning of the study, we show clinicians a sample set of documentation of the case summary and top problems. They are told that each case will take approximately 60 minutes to complete and that they are not able to step away from their device once they begin a patient case. For one of the two cases, clinicians must create the set of documentation by writing from scratch in the provided text box. We will refer to this as the writing condition. For the other case, clinicians are asked to edit an LLM-generated draft

[1] The preregistration specified that all collected sets of documentation would be retained for a future assessment of quality, but did not explicitly hypothesize about quality.

Figure 1 Study 1: Experimental flow



Note. Each gray block depicts a segment of the study in which clinicians complete one documentation case. The order of the writing and editing conditions, and the assignment of patient cases to conditions, are randomized across clinicians.

which pre-populates the text box. We will refer to this as the editing condition. For each case, when viewing the teletherapy interview transcript, clinicians are able to take notes that will then appear on subsequent pages. The text box for submitting the case summary is on the next page, and the text box for submitting top problems is on the page after that. Each clinician completes both cases; the order of conditions and the assignment of patient cases to conditions are both randomized. Between the two cases, clinicians arrive at a midpoint break page where they are able to take a break prior to proceeding with the second case. At the end of the study, clinicians are asked to complete an exit questionnaire where we ask how useful the provision of a draft was in their documentation process.

To create the LLM-generated draft for clinicians to edit, we created a custom GPT bot built on GPT-4o in January 2025, the newest available model at the time. The bot was provided ADIS interview guidelines, a sample teletherapy interview transcript, and a sample set of documentation for that transcript. Then, it was given each of the two new teletherapy interview transcripts separately and asked to generate a set of documentation for each according to the guidelines and the sample documentation set.

3.2. Outcome Measures

The survey is deployed using Qualtrics, with JavaScript to record timestamps, completion times, and time on task. We can observe when clinicians leave the tab or open another window. We also inform them that copy and paste is disabled, so all work should be done in the Qualtrics window where time is tracked, rather than on another tab or window.

Completion time is the number of seconds a clinician is actively on the Qualtrics tab while producing a given piece of documentation, measured separately for the case summary and for the top problems.

Quality is scored after the study by two independent LLM graders, Claude and Gemini. To grade the patient interview summaries, we apply a fixed rubric with five dimensions (factual accuracy, coverage of core ideas, compression quality, structure, and insight) for a maximum possible score

of 100, and average the grades from Claude and Gemini.[2] As ground truth for grading, we use the documentation actually submitted by the clinician who saw the patient; this documentation was also reviewed by a CBT expert for accuracy. To grade top problems, we use the true top problems for the patient case, assigning one point for each of the three correctly listed top problems, for a maximum possible score of 3.

3.3. Empirical Strategy

We analyze within-participant differences between the editing and writing conditions for completion time (to test Hypothesis 1) and for quality (to test Hypothesis 2). To account for the order in which conditions were completed and for which patient case was assigned to each condition, we use fixed effects for patient case and block order. For clinician i producing documentation for patient case c, we estimate:

$$\mathrm{Y}_{ic} = \beta_0 + \beta_1 \,\mathrm{Edit}_{ic} + \gamma_c + \lambda_{o(i)} + \varepsilon_{ic}, \tag{1}$$

where Y_{ic} is the outcome of interest for clinician i on case c. Edit_{ic} is an indicator that equals 1 if clinician i completes case c in the editing condition and 0 in the writing condition, γ_c is the patient-case fixed effect, $\lambda_{o(i)}$ is the block-order fixed effect, and ε_{ic} is the error term, which is clustered at the participant level. We estimate Equation 1 separately with completion time and with the quality score as the outcome, each measured for the case summary and for the top problems. In all estimations, the parameter of interest is β_1. Under Hypothesis 1, we predict $\beta_1 < 0$ for completion time. Under Hypothesis 2, we predict $\beta_1 > 0$.

3.4. Results

3.4.1. Participants and Analysis Sample We preregistered two exclusions: (i) participants who leave the browser tab for more than 30 minutes while on the webpage for drafting either the case summary or top problems, and (ii) participants who spend less than five minutes total on writing or editing across the entire study. No participant left the tab for more than 30 minutes or fell below the five-minute threshold, and one participant was excluded for a browser glitch that locked them out of the study prematurely. We collected data until we had 16 participants total who were not excluded.

[2] The full rubric for quality is provided in Appendix B.

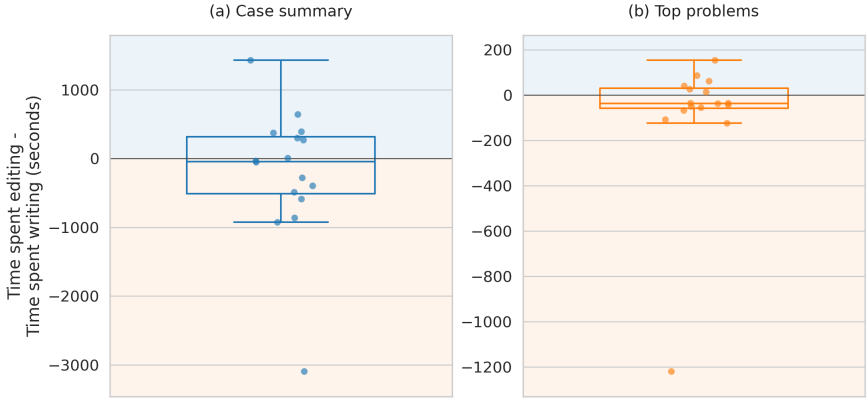
3.4.2. Impact on Productivity We do not find a significant difference in completion time between the editing and writing conditions. Table 1 reports the corresponding estimates of β_1 in Equation 1. On average, clinicians in the editing condition spend 205.38 fewer seconds drafting summaries and 87.00 fewer seconds drafting top problems, but both estimates carry large standard errors and neither is statistically significant. Thus, we fail to reject the null for Hypothesis 1.

Table 1 Study 1: Effect on completion time and output length

	Completion time (seconds)		Output length (words)		
	(1) Case summary	(2) Top problems	(3) Notes	(4) Case summary	(5) Top problems
Editing	−205.38	−87.00	−60.38	178.56***	−2.81
	(249.84)	(78.92)	(68.41)	(35.66)	(3.97)
Writing condition mean	1,532.39	218.63	172.56	420.69	35.12
Editing condition mean	1,327.01	131.63	112.19	599.25	32.31
Number of observations	32	32	32	32	32
Number of participants	16	16	16	16	16
Patient case fixed effects	Yes	Yes	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Coefficient reports the effect of being in the editing condition (β_1 in Equation 1); the writing condition is the reference condition. $+\ p<0.1$, $*\ p<0.05$, $**\ p<0.01$, $***\ p<0.001$.

Figure 2 Study 1: Within-participant differences in completion time between conditions



Note. Each point is the within-participant difference in completion time (editing condition minus writing condition) for one clinician, and each boxplot reports the 25th, 50th, and 75th percentiles of the 16 participants. Points below zero indicate that the clinician was faster in the editing condition.

The null average masks substantial heterogeneity. For case summaries, 9 of 16 clinicians are faster in the editing condition, while 7 are faster in the writing condition. For top problems, 10 clinicians are faster in the editing condition, while 6 clinicians are faster in the writing condition. Figure 2 plots the distribution of the differences between the editing condition and the writing

condition: clinicians who fall below the zero line spend less time in the editing condition, while those who fall above the zero line spend less time in the writing condition. The median is close to zero for both parts of documentation, and a single outlier clinician spends nearly one hour longer writing than editing. Further exploratory analysis does not suggest that this heterogeneity can be explained by the clinician's level of experience. The distribution makes clear that the effect of an LLM-generated draft depends on how the clinician chooses to work with it, and that choice is visible in the output: summaries in the editing condition are 178.56 words longer than those written from scratch (Table 1, column (4)), suggesting that clinicians retain much of the comprehensive draft rather than rewriting it to their own length.

3.4.3. Impact on Quality Hypothesis 2 predicts that editing improves the quality of what clinicians produce. We find that, on average, it does not. Table 2 reports the differences in quality in the editing condition versus the writing condition, with patient interview summaries from the editing condition scoring only 0.97 points lower than those in the writing condition. There is also no significant average change in the quality scores for top problems. Similar to what we observed with completion time, the null average masks heterogeneity: for case summaries, 9 of 16 clinicians score higher in the editing condition and 7 score higher in the writing condition (Figure 3). Thus, we also do not find support for Hypothesis 2.

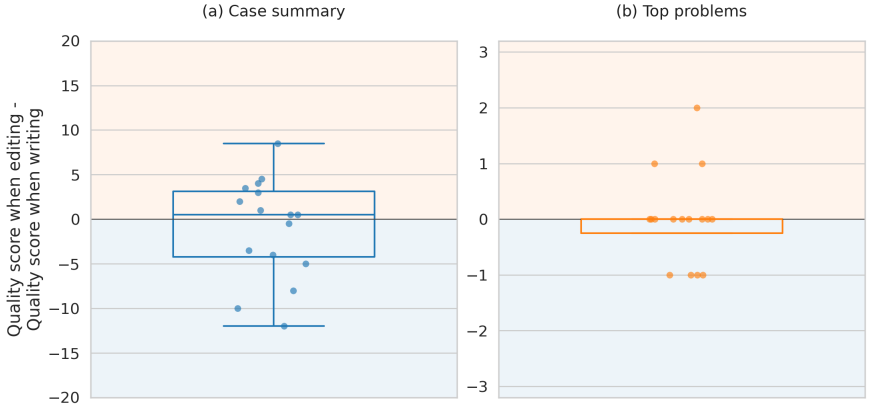
Table 2 Study 1: Effect on quality

	(1) Case summary (0–100)	(2) Top problems (0–3)
Editing	−0.97	0.00
	(1.18)	(0.21)
Writing condition mean	85.22	2.25
Editing condition mean	84.25	2.25
Number of observations	32	32
Number of participants	16	16
Patient case fixed effects	Yes	Yes
Block order fixed effects	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Coefficient reports the effect of being in the editing condition (β_1 in Equation 1); the writing condition is the reference condition. Case summary quality is assessed on a 0 (lowest) to 100 (highest) scale, and top problems quality is assessed on a 0 (lowest) to 3 (highest) scale. $+\, p < 0.1$, $*\, p < 0.05$, $**\, p < 0.01$, $***\, p < 0.001$.

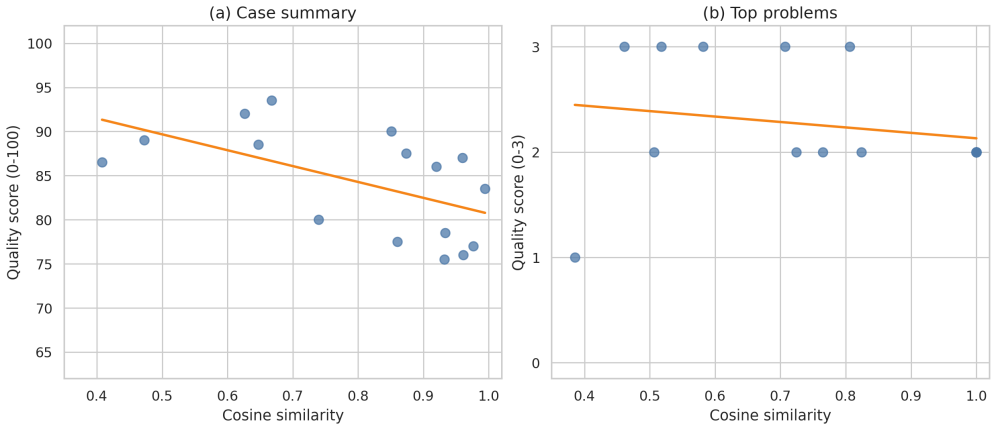
That said, for those in the editing condition, we find that case summary quality increases significantly when more revisions are made (captured by lower cosine similarity to the draft; $p < 0.05$).

Figure 3 Study 1: Within-participant differences in quality between conditions



Note. Each point is the within-participant difference in the quality score (editing condition minus writing condition) for one clinician, and each boxplot reports the 25th, 50th, and 75th percentiles of the 16 participants. Points above zero indicate higher quality in the editing condition.

Figure 4 Study 1: Quality versus cosine similarity to the LLM-generated draft, editing condition only



Note. Each point is one submitted case summary from the editing condition ($N = 16$), with the line showing the fitted regression. Higher cosine similarity indicates lighter editing.

Figure 4 plots quality against cosine similarity against the LLM-generated draft for case summaries created in the editing condition. We find that quality falls by 1.8 points (out of 100) for every 0.1 increase in cosine similarity. Summaries above the median quality score are 0.23 lower in cosine similarity than those below it. We do not find the same pattern for top problems in a statistically significant manner. In other words, quality in the editing condition depends on how much the clinician is willing to change the draft of the case summary.

3.5. Clinician Perceptions of the LLM-Generated Draft

At the conclusion of Study 1, clinicians responded to exit questionnaires asking about their perceptions of working with LLM-generated drafts. The most common positive themes are that the LLM-generated draft serves as an effective foundation (11 mentions) and speeds up the workflow (10 mentions). Some also mentioned that the provided draft was more comprehensive than one they

would have written from scratch (4 mentions), in line with the significantly longer case summaries in the editing condition. There were negative themes as well: concerns about accuracy and confirmation bias (8 mentions), a mismatch in writing style (7 mentions), and feeling constrained by the template the draft creates (6 mentions).

In addition, when asked directly, just over half of clinicians (56%) said they preferred having the draft, while 31% reported mixed feelings and 13% preferred working without it. These mixed perceptions are consistent with the heterogeneity in time and quality in Study 1: clinicians who scrutinize the draft's accuracy need more time to revise it.

Thus, Study 1 leaves us with a puzzle: some clinicians reaped large benefits from the LLM-generated draft, while others saw small or even negative returns. This variability produced null average effects for both time and quality. Clinicians' own assessments split the same way, with just over half preferring to have the draft. When provided a draft, the clinician's role inherently changes from writer to editor, and the way clinicians respond to this change drives the large differences in impact. Some perform very light revisions, while others rewrite multiple sections. The relevant question, then, is not whether an LLM can save time overall, but how the human cost of verification factors into the changed work.

Working with the draft clearly imposes a real cognitive load, but without physiological measurements we cannot observe how large that load is. This, combined with the large heterogeneity in time and quality differences, motivates Study 2, which examines the micro-level details of how people work with LLM-generated drafts.

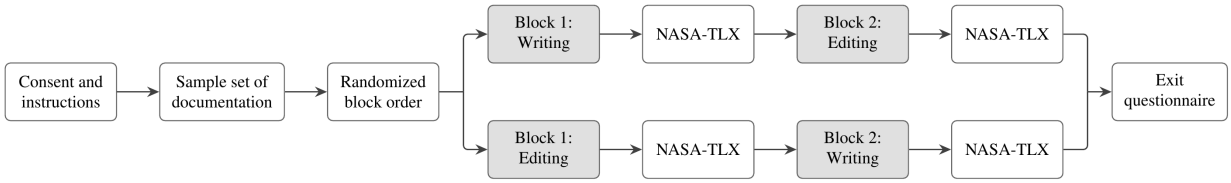
4. Study 2: Lab Study with Students

4.1. Study Design and Sample Population

Our second experiment is an in-person lab study with university students, designed to capture the micro-level interactions underlying changes in the composition of work, in cognitive load, and in the fatigue that accumulates over sustained work. We recruit 100 students through the university lab subject pool, and each student completes a single in-lab session consisting of two back-to-back 45-minute blocks of continuous work in which they summarize video interviews.

Each student receives a base payment of $30, with a $0.50 bonus for each completed summary and a bonus of $100 if their summaries are later rated to be in the top 10% quality of all summaries. The study took place from January to March 2026, and the design, sample size, and analysis plan were preregistered on AsPredicted (#267782).

Figure 5 Study 2: Experimental flow



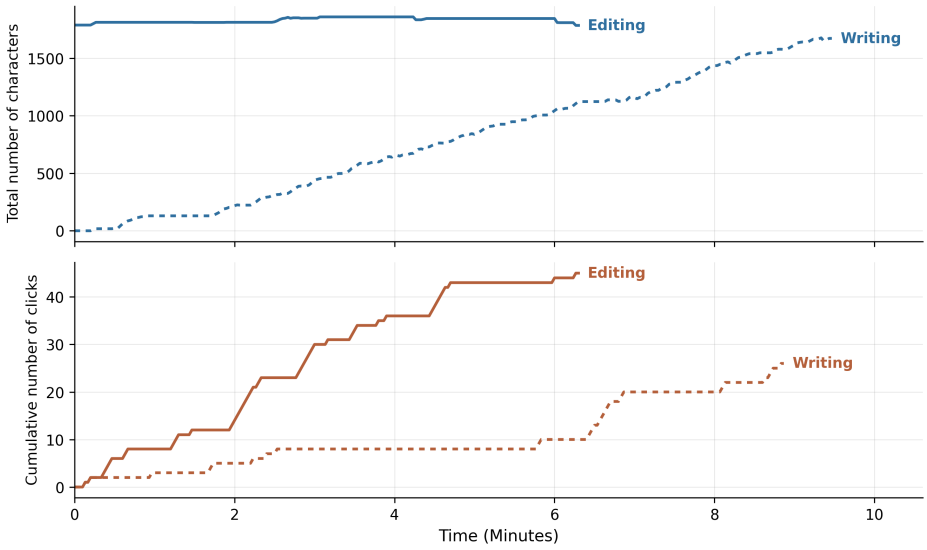
Note. Each gray block depicts a segment of the study in which students summarize a series of five-minute video interviews. The order of the writing and editing blocks is randomized across students.

The full study flow can be seen in Figure 5. At the very beginning of the study, we calibrate the eye-tracking device to the student's height and seating distance. We also apply a clip on their earlobe to capture heart rate. Next, students receive instructions for the tasks they will complete and view an example five-minute video interview and read the corresponding example summary. Each interview is a pre-recorded five-minute video interview in which a student research assistant (RA) answers questions about student life (Appendix C), and the video interviews are stored in an interview bank of 18. Students then complete the summarization tasks in two blocks, one per condition, in a randomized block order. In one block, students are provided a blank text box and asked to write the summary from scratch. In the other block, students are provided an LLM-generated draft and asked to revise the draft until they are satisfied with the summary. To create the LLM-generated draft, we used GPT-5.2 Pro (December 2025), the newest available model at the time. Video interviews shown in each block are drawn from the interview bank. In both conditions, students may take notes while watching the interview.

While students work, a timer is visible at the top of the page, counting down the 45 minutes second by second. Once the timer reaches zero, students are automatically moved to the next block. Students continuously receive new summarization tasks from the video interview bank until the timer ends, so students always have a task in front of them rather than racing to finish a fixed set. In addition, after each task, they are asked to rate their mental effort on the Paas scale (Paas 1992). Students also complete a NASA Task Load Index (NASA-TLX) workload questionnaire, a more in-depth measure of cognitive load, at the conclusion of each block. The session ends with an exit questionnaire.[3] Students work in a controlled lab environment that is identical across the session and across students, which should remove exogenous shocks to their physiological state, and the task pages are identical across conditions and tasks so that the interface itself does not affect physiological responses. This design allows us to study within-student contrasts with the source material held constant across conditions.

[3] See Appendix D for the questions and the corresponding scales.

Figure 6 Study 2: Sample draft interaction record, by condition



Note. Each panel overlays one summarization task from each condition for the same student. Solid lines show the editing condition and dotted lines show the writing condition. The top panel plots the total character count of the summary field. The bottom panel plots the number of clicks.

4.2. Outcome Measures

The survey interface is deployed on Qualtrics with JavaScript for keyboard and mouse tracking. This allows us to record timestamps, completion times, keystrokes, clicks, and other interactions for both the summary field and the notes field sampled at two-second intervals. The results are granular enough to measure the work process, such as when students are writing versus editing. A sample draft interaction record can be seen in Figure 6. In general, writing from scratch produces a steady, monotonic accumulation of characters in the summary field, whereas editing produces plateaus of little activity interspersed with short bursts of writing and revising. Both the rate of clicking and amount of clicking are significantly higher when editing compared to writing. We also use an iMotions clip on the ear for heart rate and a GazeScanner eye-tracking camera operated through iMotions to measure eye movements and pupil dynamics.[4]

The preregistered primary outcomes capture productivity, workflow, cognitive load, and fatigue. Productivity is captured by the overall time it takes to complete the summarization tasks. Workflow has multiple components, including creation intensity (the number of seconds during which new text is created), revision intensity (the number of seconds during which text is deleted), character production rate (how quickly character count increases, in characters per minute), and click rate

[4] Following the preregistration, we also collected electrodermal (skin-conductance) and blink data. Neither could be used: electrode adherence varies with skin type, and blinks are frequently missed for students wearing glasses or contact lenses, so both measures had systematically missing data. Thus, we do not report results from them in the paper, but the raw data are available from the authors upon request.

(the number of clicks per minute on the page). We also record the length of the notes and of the submitted summary in characters, along with the number of sentences in each, captured through sentence-ending periods.

Cognitive load and fatigue are measured separately, following the distinction drawn in Section 2.4. Cognitive load is self-reported: students rate their mental effort on the Paas 9-point scale after each task and complete the NASA-TLX scales at the end of each 45-minute block. Fatigue is captured by the sensor measures, which are recorded continuously as students work (Boksem and Tops 2008, Hopstaken et al. 2015). Pupil diameter and mean fixation duration index fatigue through their trajectories: pupil diameter declines and fixations lengthen as sustained effort accumulates (Beatty 1982, van der Wel and van Steenbergen 2018, Hopstaken et al. 2015, Rayner 1998, Schleicher et al. 2008). Heart rate, recorded from the ear clip, serves as a general index of physiological arousal (Jorna 1992, Malik et al. 1996) and is reported alongside the eye-tracking measures for completeness. Because fixation duration also responds to momentary difficulty, and because any single signal is ambiguous on its own, we interpret the sensor measures jointly and rely on trajectories rather than levels. Appendix Table A.1 gives the interpretation and caveats for each measure.

Finally, output quality is scored after the session by Claude using the same five-dimension rubric as in Study 1 (Appendix B), for a maximum possible score of 100. Each unique submitted text is graded exactly once under a fixed deduction protocol that applies the rubric, so identical submissions receive identical scores. These quality scores provide the Study 2 tests of Hypothesis 2.[5]

4.3. Empirical Strategy

We analyze within-participant differences between the editing condition and writing condition for time (to test Hypothesis 1), for quality (to test Hypothesis 2), for workflow (to test Hypothesis 3), and for cognitive load and fatigue (to test Hypotheses 4a and 4b). To account for differences across video interviews and for the order of each block, we also use fixed effects for students, video interviews, and block order; the phase-level sensor models additionally include task fixed effects. Each table reports which fixed effects are included in its estimates. For student i summarizing video interview v on task t, we estimate:

$$Y_{ivt} = \beta_0 + \beta_1 Edit_{ivt} + \alpha_i + \delta_v + \kappa_{b(it)} + \varepsilon_{ivt}, \tag{2}$$

where Y_{ivt} is the outcome of interest for student i summarizing video interview v as task number t in sequence. For Hypothesis 1, Y_{ivt} is the total time in seconds the student takes to complete the

[5] As in Study 1, the quality-scoring procedure was not preregistered and was conducted *post hoc*.

summarization task. The same specification is also re-estimated with the quality, workflow, and task-level workload outcomes described below to test Hypotheses 2–4b. The following notation is used in all remaining equations in this section: $Edit_{ivt}$ is an indicator which equals 1 if student i completes task t in the editing condition and 0 in the writing condition, α_i is the student fixed effect, δ_v is the video-interview fixed effect for video v, $\kappa_{b(it)}$ is the block order fixed effect, and ε_{ivt} is the error term, which is clustered at the participant level. In each estimation, the parameter of interest is β_1. We predict $\beta_1 < 0$ under Hypothesis 1 and $\beta_1 > 0$ under Hypothesis 2. Under Hypothesis 3, we predict $\beta_1 < 0$ for creation intensity and the character production rate, and $\beta_1 > 0$ for revision intensity. Under Hypothesis 4a, we predict $\beta_1 < 0$ for the self-reported cognitive load scales (the task-level Paas mental effort rating and the block-level NASA-TLX scales).[6]

To measure how outcomes change as students repeat summarization tasks, we first estimate within-condition experience curves separately on the editing and writing samples:

$$Y_{ivt} = \lambda_0 + \lambda_1 CondTaskIndex_{it} + \alpha_i + \delta_v + \varepsilon_{ivt}, \tag{3}$$

where $CondTaskIndex_{it}$ is the within-condition task index, the count of summarization tasks student i has completed in the current condition. The parameter of interest is λ_1, the change in the outcome per consecutive task within a condition.

To test whether that change differs between the editing and writing conditions, we estimate:

$$\begin{aligned} Y_{ivt} = \gamma_0 + \gamma_1 CondTaskIndex_{it} + \gamma_2 Edit_{ivt} \\ +\gamma_3 \left(CondTaskIndex_{it} \times Edit_{ivt}\right) + \alpha_i + \delta_v + \kappa_{b(it)} + \varepsilon_{ivt}. \end{aligned} \tag{4}$$

The parameter of interest is γ_3, which tests whether the across-task trend differs between the editing and writing conditions. We estimate both specifications with task completion time, the workflow measures, and the task-level Paas mental effort rating as outcomes.

Finally, to test Hypothesis 4b, we estimate minute-level within-block trajectories for the sensor measures, which are continuously captured throughout the session:

$$Y_{ibm} = \theta_0 + \theta_1 TimeMin_{ibm} + \theta_2 Edit_{ib} + \theta_3 \left(TimeMin_{ibm} \times Edit_{ib}\right) + \alpha_i + \kappa_b + \varepsilon_{ibm}, \tag{5}$$

where Y_{ibm} is the outcome for student i in block $b \in \{1, 2\}$ at elapsed minute m of the 45-minute block. $TimeMin_{ibm}$ is the number of elapsed minutes in the block, and $Edit_{ib}$ is an indicator which

[6] The NASA-TLX scales are administered once per block as opposed to after each task. As such, we analyze the within-participant differences between the editing and writing condition by estimating Equation 2 at the block level rather than the task level.

equals 1 if block b is the editing condition and 0 if it is the writing condition. α_i is the student fixed effect, κ_b is the block order fixed effect, and ε_{ibm} is the error term, clustered at the participant level.

The parameter of interest, θ_3, tests whether the minute-by-minute trajectories differ between these two conditions. Under Hypothesis 4b, editing should slow the accumulation of fatigue, so we predict $\theta_3 > 0$ for pupil diameter (a flatter decline when editing) and $\theta_3 < 0$ for mean fixation duration (a flatter increase when editing). For the physiological estimates, we use the data as collected, with the exception of pupil diameter, which is recorded separately for the left and right eyes; we combine the two into an aggregate measure using the process in Appendix E.

We estimate Equation 5 over three time windows. The *full block* is the entire 45-minute task block for either the editing condition or the writing condition. Within it, we separate the two phases corresponding to the two activities that make up each task. The *video phase* is the first five minutes of each task, during which the student watches the video interview, and the *work phase* is the remainder of each task, during which the student writes or edits the summary. Phase-level models measure time as cumulative elapsed minutes within each phase across the block.

4.4. Results

4.4.1. Participants and Analysis Sample We preregistered the recruitment of 100 university students from the university lab subject pool.[7] The 100 students submitted summaries for 796 tasks. For our analyses, we exclude 35 tasks that were aborted by Qualtrics when the 45-minute block timer expired, leaving an analysis sample of 761 completed summaries (499 in the editing condition and 262 in the writing condition).

4.4.2. Impact on Productivity In this lab setting, we find strong support for Hypothesis 1. Table 3. Column (1) reports our estimate of β_1 in Equation 2. For students, being in the editing condition cuts task completion time by 418.45 seconds—a 69% reduction relative to writing from scratch. These time savings compound over the session. Because each block lasts a fixed 45 minutes, students complete almost twice as many summarization tasks in the editing condition as in the writing condition (Appendix Figure F.1). Speed also improves over consecutive tasks: each successive task in the editing condition takes 25.71 seconds fewer than the last, and each successive task in the writing condition takes 56.74 seconds fewer than the last (Appendix Table F.1).

[7] The lab excluded three students for whom more than 85% of the task-level heart rate readings were missing.

Table 3 Study 2: Effects on task completion time and quality

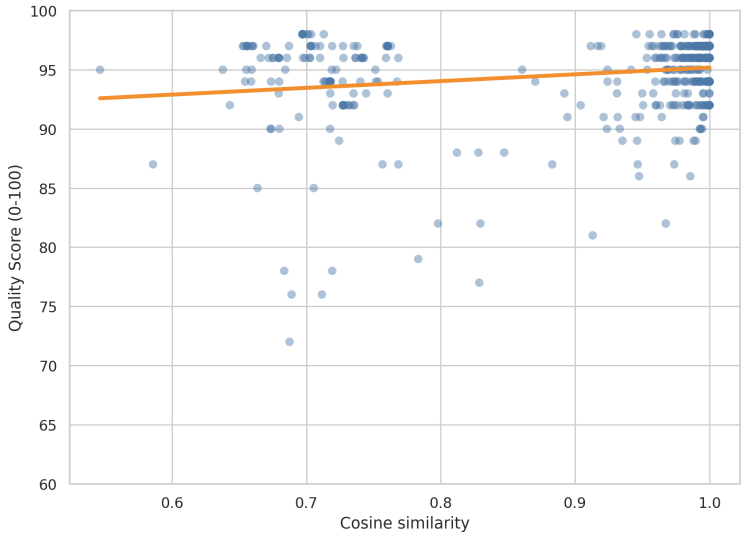
	Task completion time (seconds)	Quality score (0–100)	
	(1)	(2)	(3)
Editing	−418.45***	9.33***	
	(33.67)	(0.78)	
Cosine similarity to draft			39.29***
			(8.62)
Sample	All tasks	All tasks	Editing tasks only
Writing condition mean	605.03	85.10	—
Editing condition mean	193.00	94.69	—
Number of observations	761	761	499
Number of participants	100	100	100
Video interview fixed effects	Yes	Yes	Yes
Participant fixed effects	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Absorbed

Notes. Standard errors clustered at the participant level are in parentheses. Columns (1) and (2) report the effect of being in the editing condition (β_1 in Equation 2) on task completion time and summary quality; the writing condition is the reference condition. Column (3) regresses summary quality on cosine similarity to the LLM-generated draft using editing tasks only. Higher cosine similarity indicates lighter editing, and the coefficient is per unit of similarity. Block order is absorbed by participant fixed effects in column (3) because each student completes a single editing block. Quality is assessed on a 0 (lowest) to 100 (highest) scale using the five-dimension rubric in Appendix B. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

4.4.3. Impact on Quality We also find support for Hypothesis 2. We find that summaries in the editing condition score 9.33 points higher on quality (out of 100) compared to summaries in the writing condition (Table 3 Column (2)), with the editing condition summaries scoring higher along every dimension of the rubric. Unlike in Study 1, however, heavier revision of the LLM-generated draft is associated with *lower* quality. Figure 7 plots quality against cosine similarity to the LLM-generated draft in the editing condition, and Table 3 Column (3) reports the corresponding fixed-effects estimate. Quality rises with cosine similarity: every 0.1 increase in cosine similarity is associated with a 3.93-point increase in quality. Thus, the fewer revisions made, the better the final output. Unlike the clinicians in Study 1, whose heavier editing was associated with higher quality, additional revisions by students move the summary toward lower quality.

4.4.4. Impact on Workflow Table 4 reports Equation 2 estimates with the activity measures as outcomes in columns (1)–(4) and the final summary and notes-field measures as outcomes in columns (5)–(10). The changes in activity support the predicted shift. In the editing condition as compared to the writing condition, creation intensity is 342.55 seconds shorter and the character production rate is 388.59 characters per minute lower. This suggests that students who work with an LLM-generated draft create new content at a slower pace. Interestingly, revision intensity is

Figure 7 Study 2: Quality versus cosine similarity to the LLM-generated draft, editing condition only



Note. Each point is one submitted summary from the editing condition ($N = 499$), with the line showing the fitted regression. Higher cosine similarity indicates lighter editing.

Table 4 Study 2: Effect on workflow activity and final output measures

	Workflow activity				Summary			Notes		
	(1) Creation intensity (seconds)	(2) Revision intensity (seconds)	(3) Character production rate (chars/min)	(4) Click rate (click-s/min)	(5) Length (chars)	(6) Clicks	(7) Periods	(8) Length (chars)	(9) Clicks	(10) Periods
Editing	−342.55***	−39.44***	−388.59**	3.44***	133.23***	−14.48***	−1.27**	−155.89***	−3.29***	−0.99*
	(20.14)	(3.84)	(143.65)	(1.03)	(26.15)	(3.17)	(0.43)	(33.82)	(0.95)	(0.41)
Writing condition mean	586.21	61.34	455.25	2.94	1,523.76	31.24	13.06	1,084.75	7.86	6.50
Editing condition mean	246.90	22.71	9.33	6.63	1,665.88	16.22	11.76	889.34	4.42	4.94
Number of observations	761	761	761	761	761	761	761	761	761	761
Number of participants	100	100	100	100	100	100	100	100	100	100
Video interview fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Participant fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Coefficient reports the effect of being in the editing condition (β_1 in Equation 2); the writing condition is the reference condition. Creation (revision) intensity counts the active seconds in which the character count of the response fields increases (decreases). The Summary and Notes columns measure the final submitted text in the summary and notes fields: total character length, number of clicks in the field, and number of periods. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

also lower, which indicates students do less active typing overall, and in fact revise their own writing more than they revise the LLM-generated draft. Thus, editing does not substitute revision for creation one-for-one; there is less active typing of any kind. What does increase is navigation: the click rate more than doubles, rising by 3.44 clicks per minute.

The final submitted summaries show the same shift, with results in columns (5)–(10) of Table 4. On average, summaries in the editing condition are 133.23 characters longer than those in the

writing condition. Yet the work surrounding the summary still decreases. The total number of clicks in the summary field decreases by 14.48, and the amount of engagement with the notes field drops across all measures. There are 155.89 fewer characters in the notes, 3.29 fewer clicks in the notes field, and 0.99 fewer complete sentences ending with periods. Editing decreases preparatory work as students lean on the draft instead of on their own notes.

Altogether, Hypothesis 3 is partially supported. While creation intensity decreases as students write less, revision intensity also decreases, showing students edit less. Very little new text is produced in the editing condition; instead, the composition of the remaining work shifts away from production toward navigation and selective correction.

4.4.5. Impact on Cognitive Load and Fatigue With regards to perceived cognitive load, Hypothesis 4a is strongly supported. As shown in Table 5, Paas mental effort is 1.28 points lower in the editing condition compared to the writing condition, which is roughly a 21% reduction. Each NASA-TLX metric is also lower in the editing condition: mental demand by 18.54 points (28%), workload by 21.40 points (33%), frustration by 7.19 points (19%), and the overall score by 15.71 points (28%). Notably, the higher click rate documented in Table 4 does not translate into higher reported effort: task-level Paas ratings are unrelated to summary click rates.

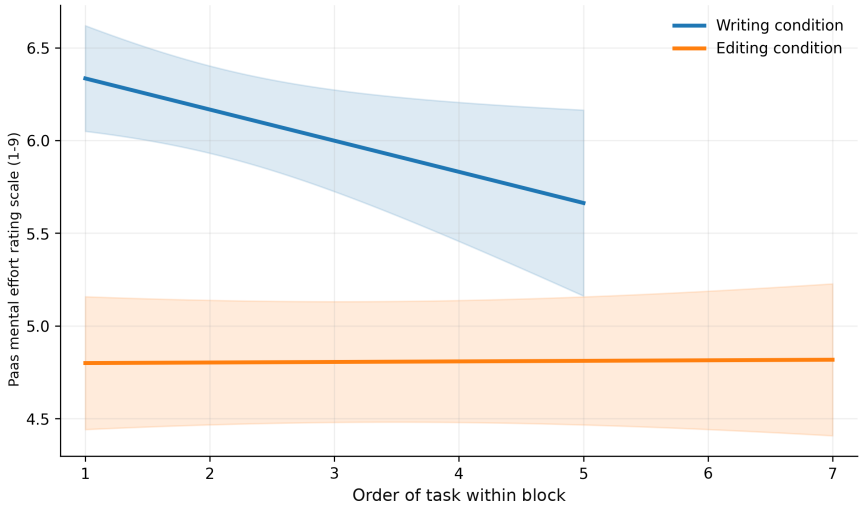
Table 5 Study 2: Effect on self-reported cognitive load

	(1) Paas mental effort rating (1–9)	(2) NASA-TLX mental demand (0–100)	(3) NASA-TLX workload (0–100)	(4) NASA-TLX frustration (0–100)	(5) NASA-TLX overall (0–100)
Editing	−1.28***	−18.54***	−21.40***	−7.19*	−15.71***
	(0.17)	(2.94)	(3.17)	(3.14)	(2.36)
Writing condition mean	6.16	65.30	65.10	37.90	56.10
Editing condition mean	4.81	46.85	43.85	30.65	40.45
Number of observations	761	200	200	200	200
Number of participants	100	100	100	100	100
Video interview fixed effects	Yes	No	No	No	No
Participant fixed effects	Yes	Yes	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Coefficient reports the effect of being in the editing condition (β_1 in Equation 2) for the task-level Paas outcome in column (1), and for the block-level NASA-TLX outcomes in columns (2)–(5); the writing condition is the reference condition. Paas mental effort rating scale is administered after each task on a 1 (lowest) to 9 (highest) scale. NASA Task Load Index (NASA-TLX) is administered at the end of each 45-minute block on a 0 (lowest) to 100 (highest) scale. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Before turning to the physiological evidence, we note that the self-reports themselves change as students work. Because editing shifts the type of work from creation to revision, the two conditions may produce different trajectories in reported effort. Using the interaction specification

Figure 8 Study 2: Fitted Paas mental effort trends over consecutive tasks, by condition



Note. Each line plots the task-level Paas mental effort rating (1–9) against the within-condition task index, with shaded 95% confidence bands. Estimated slopes are reported in Appendix Table F.2 column (3).

in Equation 4 with the task-level Paas rating as the outcome, Appendix Table F.2 column (3) reports the corresponding estimate of γ_3, interpreted as the difference in slopes across tasks with the writing condition as the baseline. Interestingly, the Paas interaction is positive (0.17 points per task). Consecutive writing tasks decrease reported load as students "get in the zone," whereas consecutive editing tasks show a slight yet not statistically significant increase. While this change in the editing condition is not statistically significant, the difference between the two slopes is significant, which is apparent in the fitted Paas trajectories plotted in Figure 8. Hence, although students do less active work per task when editing, their reported mental effort does not fall across tasks as it does in the writing condition.

Next, we turn to the physiological evidence using measures captured by the sensors. From this analysis, the pupil diameters and mean fixation duration tell an interesting story. Table 6 reports the difference in within-block slopes, Figure 9 plots the pupil diameter trajectory over time, and Appendix Figure F.2 plots fixation duration against time in the block. Pupil diameter declines over the block in both conditions, consistent with accumulating fatigue. It declines faster in the editing condition (a slope difference of -0.002 mm per minute, or roughly a 0.1 mm decline over the 45 minutes), indicating that students fatigue faster when editing than when writing. The same pattern holds when the two eyes are modeled separately rather than averaged (Appendix Table F.3). We see the same trend with mean fixation duration, which increases faster in the editing condition by 3.59 ms per minute, which also indicates that students fatigue faster when editing. The other continuously recorded measure, heart rate, shows no measurable difference over the block. The

Table 6 Study 2: Difference in within-block slopes of physiological measures over the full 45-minute block

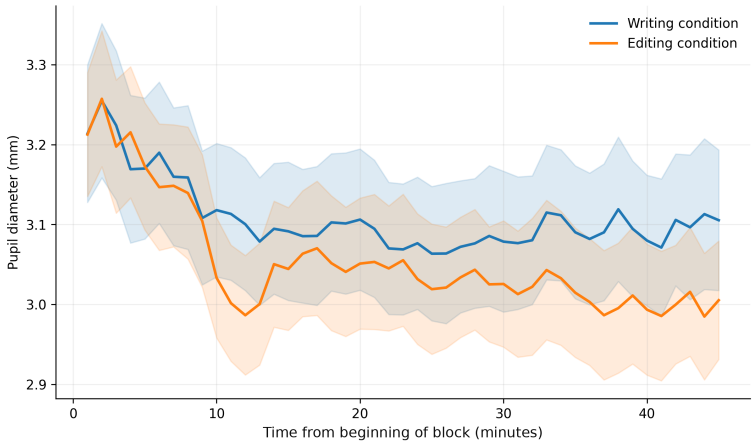
	(1) Pupil diameter (mm)	(2) Mean fixation duration (ms)	(3) Heart rate (bpm)
Editing − writing slope difference (per min)	−0.002***	3.59*	−0.01
	(0.0004)	(1.79)	(0.04)
Editing condition slope (per min)	−0.004***	3.64*	−0.03
	(0.0003)	(1.65)	(0.03)
Writing condition slope (per min)	−0.002***	0.05	−0.02
	(0.0003)	(0.69)	(0.03)
Number of observations	8,182	8,742	8,731
Number of participants	100	100	99
Video interview fixed effects	No	No	No
Participant fixed effects	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Each column reports, for the listed sensor outcome, the within-block slopes on elapsed minute from Equation 5: the first row is θ_3, the editing-minus-writing difference in slopes, and the rows below report the editing condition slope ($\theta_1 + \theta_3$) and the writing condition slope (θ_1). Sensor data are recorded continuously and aggregated to minute-level observations for estimation. Slope units are outcome units per minute, and the writing condition is the reference condition. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

overall pattern parallels the Paas scores: at the start of the block, pupil diameter is smaller in the writing condition, consistent with the higher load students report when writing. Over time, however, pupil diameter in the editing condition declines fast enough to catch up, mirroring the difference in Paas slopes across consecutive tasks. Editing may feel easier initially, but fatigue accumulates faster over the same period.

In Appendix Tables F.4, F.5, and Figure F.3, we further decompose the block into the video phase and the work phase. From this, we see that the difference in slope of pupil diameter is mostly in the work phase (−0.002 mm per minute), which is when students are actively writing or editing. During the video phase where students watch the video interview, the difference is not statistically

Figure 9 Study 2: Pupil diameter trajectory over the 45-minute block, by condition



Note. Each line plots the average pupil diameter at each elapsed minute of the block, with shaded 95% confidence bands, for the writing and editing conditions, respectively. Estimated slopes are reported on Table 6.

significant. Estimating the same phase-level models over consecutive tasks rather than elapsed minutes yields no significant condition differences in either phase (Appendix Table F.6), so the fatigue evidence is specific to the within-block timescale.

Together, we find strong support for Hypothesis 4a but reject Hypothesis 4b. Editing lowers the cognitive load students perceive on any given task, but it does not slow the accumulation of fatigue—it speeds it up. Pupil diameter declines faster and fixation duration lengthens faster when editing, and reported effort, which falls across consecutive writing tasks, stays flat when editing. Counter to Hypothesis 4b, students fatigue faster when editing an LLM-generated draft than when writing from scratch. In addition, in the exit surveys, 78% of students stated that writing from scratch is more challenging, and 77% stated it is more fatiguing (Appendix Figure F.4).

5. Discussion and Conclusion

Leveraging LLMs for documentation or writing-intensive work does not simply speed up an unchanged task. Rather, it reconstitutes the task, shifting the worker from serving as an author to serving as an editor. We study the impact associated with this shift across two studies. In Study 1, conducted with 16 clinicians producing documentation following a teletherapy patient encounter, we establish the puzzle: counter to our hypotheses (H1 and H2), editing an LLM-generated draft of the documentation as opposed to writing it from scratch does not result in time savings or quality improvements on average. The average conceals a split. Clinicians who revised heavily produced higher-quality documentation—a 1.8 point increase (out of 100) for each 10% decrease in cosine similarity to the draft—but those same revisions absorbed the time the draft was meant to save, while clinicians who revised lightly saved time without improving quality. Their stated preferences mirrored the split.

Study 2, conducted with 100 university students summarizing video interviews in a lab setting, unpacks the mechanisms underlying the heterogeneity observed in Study 1. Using granular keystroke logs, validated self-reports of workload, and micro-level physiological and eye-tracking measurements, we show that the author-to-editor shift is visible not only in productivity and text production but also in selected attentional trajectories. In the lab, the average gains predicted by prior literature (Noy and Zhang 2023, Peng et al. 2023) did materialize among students: students working with LLM drafts completed tasks in roughly 70% less time and produced higher-quality summaries. Notably, the effect of editing on quality reversed: among students, additional edits were associated with lower quality. Students also self-reported lower workload across all scales,

and approximately three-quarters of them named writing from scratch the more challenging and more fatiguing way to work. However, biosensors reveal a more nuanced story of cognitive load. Pupil-diameter and fixation-duration trajectories indicate that fatigue accumulates faster when editing than when writing. The Paas mental effort ratings show a similar pattern across consecutive tasks—reported effort falls when writing but not when editing. Writing grows easier but editing remains the same throughout. In addition, while students may not be conscious of the growing fatigue when editing, the body still registers the work.

Altogether, these studies suggest that assessing the true impact of revising LLM drafts requires more than a high-level view of productivity and time savings; it requires understanding how the workflow shifts and how workers respond to that shift under sustained use. The two settings differ in several ways at once—the training participants bring to the task, the stakes attached to errors, and how much room the draft leaves for improvement—and our design cannot separate them. We do not dispute that LLMs can help with documentation and writing; rather, we analyze where and how that help appears, and what it costs.

5.1. Theoretical Contributions

This paper makes several contributions to the operations management literature on human-AI collaboration. We introduce a novel methodology for measuring cognitive state through sensors and eye tracking alongside validated self-reports in a lab environment. Editing clearly lowers perceived workload, yet the pupil-diameter trajectory diverges by condition, and mean fixation duration lengthens faster during editing, even though heart-rate trajectories do not differ. This mixed pattern extends AI and worker well-being research that has largely relied on self-reports (Tang et al. 2023, Spiliotopoulou et al. 2026), while also showing why physiological measures should be reported separately rather than collapsed into a single claim about cognitive load.

We also demonstrate that the productivity gains documented in controlled lab settings may not carry over to less controlled settings and more difficult tasks. While editing an LLM draft is faster than writing from scratch for university students, the same result does not hold for licensed clinicians producing professional documentation. Clinicians, who remain accountable for accuracy and completeness, reallocate the time originally saved on writing to verifying and revising the provided draft.

Finally, we document that the association between revision intensity and quality reverses across our two settings. We cannot isolate why. Task expertise is one candidate—clinicians are trained against a documentation standard for which students have no counterpart—but the settings also

differ in the stakes attached to errors and in how much headroom the draft leaves, and separating these requires manipulating them within a single population. The same LLM output is a floor for users who are skilled in making edits, and a trap for users who are not, showing that more editing is not necessarily better editing. The heterogeneity in final output quality arises within the revision process itself. Granting access to LLMs does not guarantee a better end result, and more editing is not reliably better editing.

5.2. Limitations and Future Research

This study has limitations, and its results should be interpreted accordingly. First, Study 1 uses a sample of 16 clinicians who completed the tasks remotely, outside a fully integrated live workflow, and was powered to detect only large average effects. Thus, we use its central finding—that the same LLM draft generated widely heterogeneous impacts across clinicians—as motivation for the design of Study 2. Future work with larger clinician samples embedded in live clinical workflows can more precisely estimate average time savings and characterize which clinicians benefit most from LLM drafts.

Second, the participants in Study 2 are students, not clinicians. This population allowed us to collect dense measurements of interaction behavior and physiology in a lab setting. However, it limits generalizability: students have less accountability and less training in the documentation task than licensed clinicians. In addition, we simplified the task to suit a non-professional population, which could also affect the suitability of the LLM draft if the task moves outside the "jagged frontier" (Dell'Acqua et al. 2026). Future studies can test whether the pupil-trajectory differences replicate among clinicians and other professionals who (1) have tasks requiring greater domain knowledge and (2) feel a higher level of responsibility for the completed draft.

Third, our design keeps the LLM draft constant across participants by using pre-generated drafts. In doing so, we can isolate the revision process by clinicians and students without variation in prompting or in the drafts themselves. However, real deployments in professional settings would give users the ability to prompt the LLM themselves and iterate with it, which may change how they produce their final output. Future research can incorporate more flexibility with prompting and human-LLM interaction to study how the entire workflow changes, rather than the revision aspect alone.

Fourth, tasks within each study remain at similar difficulty, although difficulty varies between studies because the populations and domains differ. Future research can vary task difficulty within a single population to test whether the effects of LLM assistance on productivity and on sustained

attention depend on the task's location relative to the jagged technological frontier (Dell'Acqua et al. 2026). The simpler Study 2 task should lie within that frontier; more difficult tasks may change both the burden of verification and the physiological and attentional trajectories we observe.

5.3. Practical Implications and Conclusions

Our study has implications for managers who want to deploy LLMs in documentation or writing-intensive work. First, managers should not assume that having access to drafts will uniformly improve productivity across an organization. In settings involving high worker accountability, the time saved from creating documentation may shift to revising documentation, leading to a much smaller productivity gain. We also show that training matters for the same reason. The quality of the final output may depend on the worker's expertise, and workers who are not well trained may produce worse documentation with LLM assistance, even if they spend additional time revising and editing. Training should focus on what makes a good draft, what needs revision, and when to leave the draft as is.

In addition, managers should account for cognitive load as part of their decision to deploy LLMs. While editing might feel easier than writing from scratch, sustained editing still causes physiological fatigue and cognitive load to accumulate—and at a faster rate. Thus, if employees still need to work for the same amount of time, having them revise LLM-generated drafts can actually cause greater burnout and cognitive load, so ambient documentation tools need not reduce burden across the board (Duggan et al. 2025, Olson et al. 2025). Those deploying LLM assistance for documentation will need to consider more than just throughput, and instead consider whether it causes health care workers to burn out in the long run. The first draft problem does not disappear as models improve: deploying LLM drafts chooses where the work of documentation goes, and that choice deserves as much managerial attention as the time savings themselves.

References

Apathy NC, Rotenstein L, Bates DW, Holmgren AJ (2023) Documentation Dynamics: Note Composition, Burden, and Physician Efficiency. *Health Services Research* 58(3):674–685, URL http://dx.doi.org/10.1111/1475-6773.14097.

Bavafa H, Jonasson JO (2023) The Distributional Impact of Fatigue on Performance. *Management Science* 70(5):3319–3337, URL http://dx.doi.org/10.1287/mnsc.2023.4855.

Bayer RC, Renou L (2024) Interacting with Man or Machine: When Do Humans Reason Better? *Management Science* 72(1):594–608, URL http://dx.doi.org/10.1287/mnsc.2023.03315.

Beatty J (1982) Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* 91(2):276–292.

Boksem M, Tops M (2008) Mental Fatigue: Costs and Benefits. *Brain Research Reviews* 59(1):125–139, URL http://dx.doi.org/10.1016/j.brainresrev.2008.07.001.

Boyacı T, Canyakmaz C, de Véricourt F (2024) Human and Machine: The Impact of Machine Input on Decision Making Under Cognitive Limitations. *Management Science* 70(2):1258–1275, URL http://dx.doi.org/10.1287/mnsc.2023.4744.

Brynjolfsson E, Li D, Raymond L (2025) Generative AI at Work. *The Quarterly Journal of Economics* 140(2):889–942, URL https://arxiv.org/abs/2304.11771.

Chen Y, Kirshner SN, Ovchinnikov A, Andiappan M, Jenkin T (2025) A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *Manufacturing & Service Operations Management* 27(2):354–368, URL http://dx.doi.org/10.1287/msom.2023.0279.

Corbett CJ (2024) The Operations of Well-Being: An Operational Take on Happiness, Equity, and Sustainability. *Manufacturing & Service Operations Management* 26(2):409–430, URL http://dx.doi.org/10.1287/msom.2022.0521.

Cui KZ, Demirer M, Jaffe S, Musolff L, Peng S, Salz T (2026) The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers. *Management Science* URL http://dx.doi.org/10.1287/mnsc.2025.00535.

Dell'Acqua F (2022) Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper, Harvard Business School, URL https://www.fabriziodellacqua.com/, working paper.

Dell'Acqua F, Ayoubi C, Lifshitz H, Sadun R, Mollick E, Mollick L, Han Y, Goldman J, Nair H, Taub S, Lakhani K (2025) The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise. National Bureau of Economic Research (no. 33641), URL http://dx.doi.org/10.3386/w33641.

Dell'Acqua F, Edward McFowland I, Mollick E, Lifshitz H, Kellogg KC, Rajendran S, Krayer L, Candelon F, Lakhani KR (2026) Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Organization Science* 37(2):403–423, URL http://dx.doi.org/10.1287/orsc.2025.21838.

Duggan MJ, Gervase J, Schoenbaum A, Hanson W, Howell r J T, Sheinberg M, Johnson KB (2025) Clinician Experiences With Ambient Scribe Technology to Assist With Clinical Documentation. *JAMA Network Open* 8(2):e2460637, URL http://dx.doi.org/10.1001/jamanetworkopen.2024.60637.

Fugener A, Grahl J, Gupta A, Ketter W (2022) Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33(2):678–696, URL http://dx.doi.org/10.1287/isre.2021.1079.

Fugener A, Walzner DD, Gupta A (2025) Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work. *Management Science* 72(1):538–557, URL http://dx.doi.org/10.1287/mnsc.2024.05684.

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, volume 52, 139–183 (North-Holland), URL http://dx.doi.org/10.1016/S0166-4115(08)62386-9.

Hodgson T, Coiera E (2016) Risks and Benefits of Speech Recognition for Clinical Documentation: A Systematic Review. *Journal of the American Medical Informatics Association* 23(e1):e169–e179, URL http://dx.doi.org/10.1093/jamia/ocv043.

Hopstaken J, van der Linden D, Bakker A, Kompier M (2015) A Multifaceted Investigation of the Link Between Mental Fatigue and Task Disengagement. *Psychophysiology* 52(3):305–315, URL http://dx.doi.org/10.1111/psyp.12339.

Jorna P (1992) Spectral Analysis of Heart Rate and Psychological State: A Review of Its Validity as a Workload Index. *Biological Psychology* 34(2-3):237–257, URL http://dx.doi.org/10.1016/0301-0511(92)90017-o.

Kim BJ, Lee J (2024) The Mental Health Implications of Artificial Intelligence Adoption: The Crucial Role of Self-Efficacy. *Humanities and Social Sciences Communications* 11(1):1–15, URL http://dx.doi.org/10.1057/s41599-024-04018-w.

Kupferschmidt KL, O'Doherty KC, Skorburg JA (2025) Write on Paper, Wrong in Practice: Why LLMs Still Struggle with Writing Clinical Notes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1524–1534, URL http://dx.doi.org/10.1609/aies.v8i2.36651.

Liu BL, KC DS, Staats BR, Fundora MP (2026) Frontiers in Operations: Operational Overload: The Impact of Workload on High-Skilled Workforce Attrition. *Manufacturing & Service Operations Management* URL http://dx.doi.org/10.1287/msom.2025.0037, articles in Advance.

Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, Schwartz PJ (1996) Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal* 17(3):354–381, URL http://dx.doi.org/10.1093/oxfordjournals.eurheartj.a014868.

Maynez J, Narayan S, Bohnet B, McDonald RT (2020) On Faithfulness and Factuality in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, URL https://arxiv.org/abs/2005.00661.

Meincke L, Girotra K, Nave G, Terwiesch C, Ulrich KT (2024) Using Large Language Models for Idea Generation in Innovation. Working paper, The Wharton School, University of Pennsylvania, sSRN 4526071.

Meincke L, Terwiesch C, Huchzermeier A (2026) Evaluating LLMs for Dynamic, Multimodal Clinical Decision-Making. White paper, The Wharton School, University of Pennsylvania.

Ni X, Wang Y, Feng T, Lu LX, Wang Y, Zhou C (2025) Generative AI in action: Field experimental evidence from alibaba's customer service operations. Working Paper 5012601, SSRN.

Noy S, Zhang W (2023) Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381(6654):187–192, URL http://dx.doi.org/10.1126/science.adh2586.

Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, Stults CD, Jones VG, Shah SD, Shah T, Schwamm LH (2025) Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open* 8(10):e2534976.

Otis N, Clarke R, Delecourt S, Holtz D, Koning R (2024) The Uneven Impact of Generative AI on Entrepreneurial Performance. Harvard working paper, URL http://dx.doi.org/10.2139/ssrn.4671369.

Paas FGWC (1992) Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology* 84(4):429–434, URL http://dx.doi.org/10.1037/0022-0663.84.4.429.

Patwardhan T, Dias R, Proehl E, Kim G, Wang M, Watkins O, Posada Fishman S, Aljubeh M, Thacker P, et al. (2025) GDPval: Evaluating AI model performance on real-world economically valuable tasks. Working paper, OpenAI, arXiv:2510.04374.

Pearlman K, Wan W, Shah S, Laiteerapong N (2025) Use of an AI Scribe and Electronic Health Record Efficiency. *JAMA Network Open* 8(10):e2537000, URL http://dx.doi.org/10.1001/jamanetworkopen.2025.37000.

Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590, URL https://arxiv.org/abs/2302.06590.

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124(3):372–422, URL http://dx.doi.org/10.1037/0033-2909.124.3.372.

Roberts K (2024) Large Language Models for Reducing Clinicians' Documentation Burden. *Nature Medicine* 30(4):942–943, URL http://dx.doi.org/10.1038/s41591-024-02888-w.

Sanmark E, Vartiainen V, Sanmark J, Wettin K, Saari L, Entezarjou A (2025) Impact of an AI Medical Scribe After 375 000 Notes Generated Across Care Levels in a European Health System. medRxiv preprint, URL http://dx.doi.org/10.64898/2025.12.06.25341757.

Schleicher R, Galley N, Briest S, Galley L (2008) Blinks and Saccades as Indicators of Fatigue in Sleepiness Warnings: Looking Tired? *Ergonomics* 51(7):982–1010, URL http://dx.doi.org/10.1080/00140130701817062.

Shah SV (2024) Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. *JAMA Network Open* 7(8):e2425953, URL http://dx.doi.org/10.1001/jamanetworkopen.2024.25953.

Shaw SD, Nave G (2026) Thinking—fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. *The Wharton School Research Paper* URL http://dx.doi.org/10.2139/ssrn.6097646.

Spiliotopoulou E, Kraft T, Mankad S, Zheng Y (2026) Working with Generative AI vs Humans: The Impact on Well-being. Working paper, URL http://dx.doi.org/10.2139/ssrn.6197938.

Su H, Sun Y, Li R, Zhang A, Yang Y, Xiao F, Duan Z, Chen J, Hu Q, Yang T, Xu B, Zhang Q, Zhao J, Li Y, Li H (2025) Large Language Models in Medical Diagnostics: Scoping Review with Bibliometric Analysis. *Journal of Medical Internet Research* 27:e72062, URL http://dx.doi.org/10.2196/72062.

Tang P, Koopman J, Mai K, De Cremer D, Zhang J, Reynders P, Ng C, Chen I (2023) No Person Is an Island: Unpacking the Work and After-Work Consequences of Interacting with Artificial Intelligence. *Journal of Applied Psychology* 108(11):1766–1789, URL http://dx.doi.org/10.1037/apl0001103.

van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP (2024) Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study. *JMIR AI* 3:e60020, URL http://dx.doi.org/10.2196/60020.

van der Wel P, van Steenbergen H (2018) Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review* 25(6):2005–2015, URL http://dx.doi.org/10.3758/s13423-018-1432-y.

Appendix A: Interpretation of the Cognitive Load and Fatigue Measures

This appendix records what each measure indicates, in which direction, and with what caveats. Cognitive load is measured through the self-reported scales, and fatigue is measured through the continuously recorded sensor trajectories.

Table A.1 Study 2: Interpretation of the self-reported scales and physiological measures

Measure	Construct	Interpretation and direction	Caveats
Self-reported mental effort (NASA-TLX, Paas)	Cognitive load	Higher scores indicate greater perceived cognitive load (Hart and Staveland 1988, Paas 1992).	Retrospective; may not capture short-lived changes in load.
Pupil diameter	Fatigue	A *declining* diameter over time indicates accumulating fatigue and disengagement (Beatty 1982, van der Wel and van Steenbergen 2018, Hopstaken et al. 2015). We use the slope over the block, not the level.	Sensitive to screen luminance, gaze angle, and ambient lighting.
Mean fixation duration	Fatigue	*Lengthening* fixations indicate greater load or accumulating fatigue (Rayner 1998, Schleicher et al. 2008).	Responds to momentary difficulty as well as to sustained effort; the two cannot be separated, so we read it jointly with pupil diameter.
Heart rate	Fatigue; physiological arousal	A *higher* rate indicates greater arousal and effort (Jorna 1992, Malik et al. 1996).	Affected by posture, caffeine, and stress. Reported for completeness; null in every window we estimate.

Notes. No single measure is used on its own to argue that cognitive load or fatigue changes within a condition; we rely on the direction of multiple measures taken together, and on trajectories over time rather than levels. Self-reported scales are administered after each task (Paas) and at the end of each 45-minute block (NASA-TLX). Physiological signals are captured through wearable sensors and iMotions software. Eye-tracking signals are captured through a GazeScanner camera.

Appendix B: Quality Grading Rubric

Both studies score each submitted summary out of 100 across the same five dimensions: factual accuracy (30 points), coverage of core ideas (30 points), compression quality (20 points), structure and coherence (10 points), and insight and synthesis (10 points). The grading rubric below is provided to the LLM graders (Claude and Gemini in Study 1; Claude in Study 2) together with the source material and the instruction to grade each of the submitted responses.

1. **Factual Accuracy (0–30 points).** No incorrect claims; no misinterpretation of the speaker's intent; no invented details; precise terminology where the transcript is precise.

2. **Coverage of Core Ideas (0–30 points).** Captures all major arguments, steps, and conclusions; does not omit key framing or caveats; correct prioritization (main ideas > examples > anecdotes).

3. **Compression Quality (0–20 points).** Efficient information density; avoids copying phrasing verbatim; avoids redundancy; does not bloat minor points.

4. **Structure & Coherence (0–10 points).** Logical flow and clear progression of ideas; easy to reconstruct the original talk's logic.

5. **Insight & Synthesis (0–10 points).** Shows understanding, not transcription; connects ideas when appropriate; abstracts principles from examples.

In Study 1, the graded outputs are the case summaries and top problems. The primary quality score is the average of the Claude and Gemini grades. In Study 2, the graded outputs are the submitted video interview summaries with Claude as the sole grader.

Appendix C: Interview Questions for Video Interviews in Study 2

C.1. Practice Interview Question

1. Can you tell me about a week that felt especially hard this semester and what helped you get through it?

C.2. Health and Well-being

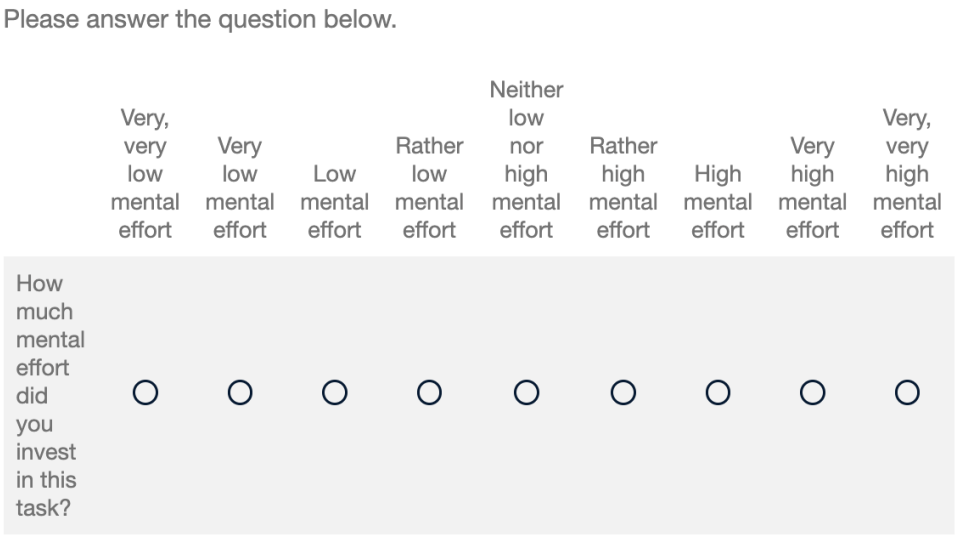
1. Walk me through a typical weekday. How do classes, studying, and sleep fit into your routine?
2. Are there daily and/or weekly habits that make you feel your best during the semester?
3. When things start to pile up, what stresses you out the most?
4. How do you usually recharge after a tough stretch?
5. Has being at [this university] changed how you think about balance or self-care?

C.3. Campus Belonging and Social Life

1. What kinds of clubs, teams, or communities are you involved in right now?
2. How did your expectations of [this university's] social life compare to the reality when you arrived?
3. Are there spaces or times on campus when you feel most comfortable or 'at home'?
4. How have your friendships or social life changed since you started at [this university]?
5. If you could give your first-year self one piece of advice, what would it be?

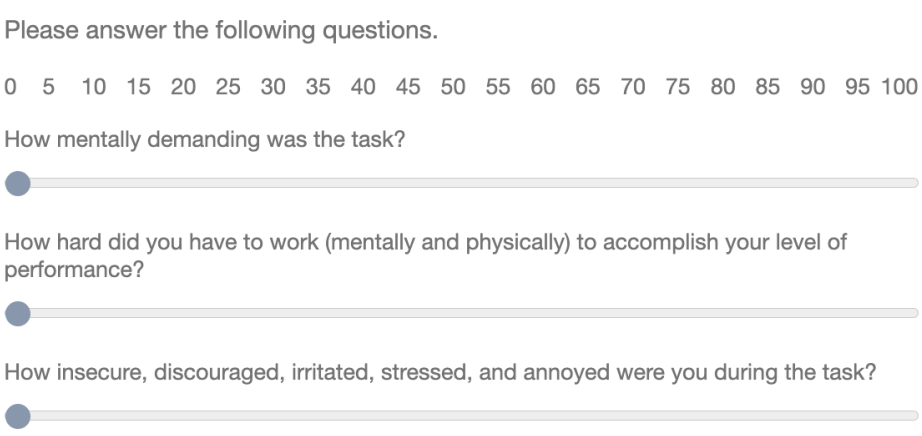
Appendix D: Self-Reported Mental Effort Scales

Figure D.1 Study 2: Paas mental effort scale after each task



Note. The Paas mental effort scale presented as radio buttons after each task.

Figure D.2 Study 2: NASA-TLX workload scale at the end of each block



Note. The NASA-TLX workload scale presented as three sliders at the end of each 45-minute block.

Appendix E: Physiological Measurement Cleaning and Preprocessing

This section documents the pupil-cleaning procedure, the construction of the video-versus-work phases, the model specification, and the full set of consecutive-task and eye-specific results that are summarized in the main text.

E.1. Pupil data cleaning

Pupil diameter is measured separately for the left and right eyes by the GazeScanner camera. Beginning from 9,003 candidate eye windows, we keep 8,422 matched windows. A window is kept only if both eyes are present and each eye contributes at least 10% within the range of valid samples. This removes 47 sparse windows where either eye had values outside the physically possible range. Windows in which only one eye is available are removed (9 windows). Both eyes are removed together when there is inter-eye disagreement that exceeds four robust standard deviations (0.24 mm). This removes 525 eye-discordant windows. The retained left- and right-eye series are then averaged into a matched average, which is the pupil-diameter measure used in the main text. Specific measurements for the left and right pupil diameters are retained separately for the supplementary models in Table F.3. Raw pupil values outside 1.5–8.0 mm are treated as invalid.

E.2. Phase construction and model specification

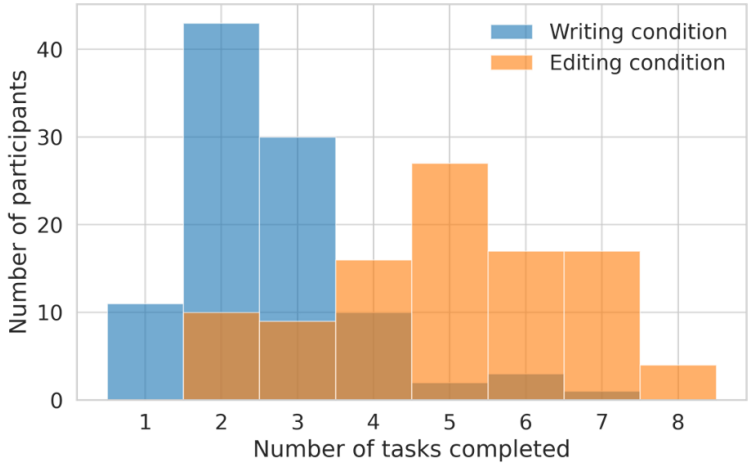
Each task contributes its first five minutes to the accumulated *video* timeline and its remaining minutes to the accumulated *work* timeline. All full-block models combine the video and work phases to measure impact over the entire 45 minutes. Full-block pupil values use the averaged pupil measurements. Each condition is retained when it has at least five participants so the editing and writing series may have different endpoints (since fewer students are able to write as much as they edit). Standard errors are clustered by participant. Within each phase we estimate slopes on two timescales: accumulated elapsed minutes within the phase, which is the specification reported in the main text, and the consecutive task index, which is reported in Appendix Table F.6.

E.3. Consecutive-task and eye-specific results

Table F.6 reports the sensor slopes estimated over the consecutive task index within each phase, and Table F.3 reports the eye-specific (left and right) pupil slopes. The consecutive-task editing-minus-writing pupil contrast is not significant in either the video phase (−0.01 mm per task, $p = 0.42$) or the work phase (−0.002 mm per task, $p = 0.89$). The eye-specific results mirror the average pupil diameter pattern: the editing-minus-writing difference in the accumulated work-phase slope is negative and significant for both eyes (−0.002 mm per minute, $p < 0.001$ left, $p < 0.004$ right), while the video-phase and consecutive-task eye-specific contrasts are not significant.

Appendix F: Additional Tables and Figures

Figure F.1 Study 2: Within-participant differences in number of tasks completed, by condition



Note. The histogram plots the distribution of the number of tasks completed per 45-minute block in the editing condition and the writing condition (499 versus 262 completed tasks in total).

Table F.1 Study 2: Within-condition experience curves for task completion time

	(1) Editing condition	(2) Writing condition
Within-condition task index	−25.71***	−56.74***
	(4.01)	(14.14)
Number of observations	499	262
Number of participants	100	100
Video interview fixed effects	Yes	Yes
Participant fixed effects	Yes	Yes
Block order fixed effects	No	No

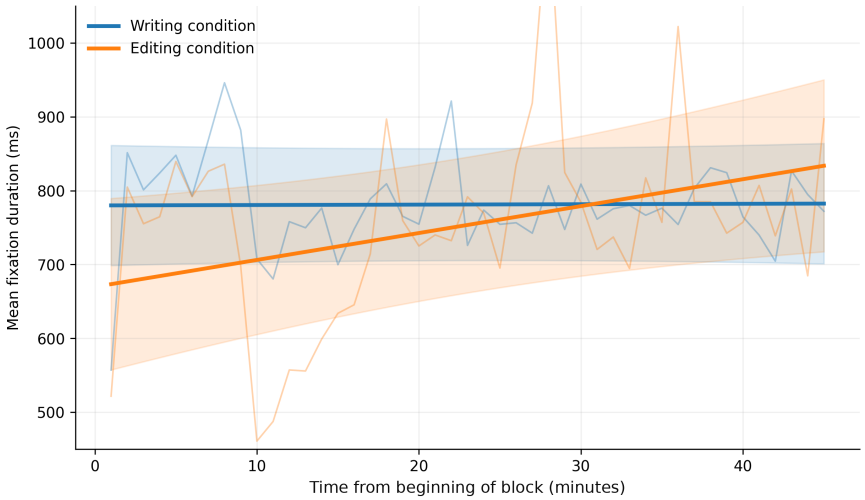
Notes. Standard errors clustered at the participant level are in parentheses. Each column reports the coefficient of the within-condition task index; the dependent variable is task completion time in seconds. $+\, p < 0.1$, $*\, p < 0.05$, $**\, p < 0.01$, $***\, p < 0.001$.

Table F.2 Study 2: Difference in across-task trends

	(1) Task completion time (seconds)	(2) Quality score (0–100)	(3) Paas mental effort rating (1–9)	(4) Creation intensity (seconds)	(5) Revision intensity (seconds)	(6) Summary character production rate (chars/min)	(7) Notes character production rate (chars/min)
Editing − writing slope difference (per task)	104.82***	0.28	0.17*	55.34***	8.91***	−437.59*	−14.97***
	(14.44)	(0.46)	(0.08)	(6.92)	(1.82)	(192.14)	(3.92)
Editing condition slope (per task)	−25.97***	0.34***	0.003	−16.00***	−2.60***	−14.07	−9.58***
	(4.27)	(0.10)	(0.03)	(2.66)	(0.56)	(8.98)	(2.34)
Writing condition slope (per task)	−130.80***	0.06	−0.17*	−71.33***	−11.51***	423.52*	5.39
	(14.46)	(0.40)	(0.08)	(6.76)	(1.82)	(184.50)	(3.34)
Number of observations	761	761	761	761	761	761	761
Number of participants	100	100	100	100	100	100	100
Video interview fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Participant fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. Each column reports, for the listed outcome, the across-task slopes from Equation 4: the first row is γ_3, the coefficient of the interaction between the within-condition task index and the editing condition indicator, and the rows below report the model-implied within-condition trends from the same estimation ($\gamma_1 + \gamma_3$ for the editing condition and γ_1 for the writing condition). Slope units are outcome units per consecutive task, and the writing condition is the reference condition. Summary quality is assessed on a 0 (lowest) to 100 (highest) scale, and the Paas mental effort rating is administered after each task on a 1 (lowest) to 9 (highest) scale. Creation (revision) intensity counts the active seconds in which the character count of the response fields increases (decreases), and the character production rates count the number of new characters written per minute in the summary and notes fields. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Figure F.2 Study 2: Mean-fixation-duration trajectory over the 45-minute block, by condition



Note. Each thin line plots the cross-student average mean fixation duration at each minute of the block. The thick lines show the fitted within-block trends from Equation 5 with shaded 95% confidence bands.

Table F.3 Study 2: Eye-specific (left and right) pupil-diameter trends during the video and work phases

Outcome	Writing slope (1)	Editing slope (2)	Editing−Writing (3)	*N*
Panel A. Accumulated minute within phase — video				
Left pupil diameter (mm)	−0.0062**	−0.0082***	−0.0020	2,378
	(0.0023)	(0.0010)	(0.0026)	
Right pupil diameter (mm)	−0.0048*	−0.0079***	−0.0031	2,378
	(0.0020)	(0.0011)	(0.0023)	
Accumulated minute within phase — work				
Left pupil diameter (mm)	−0.0024***	−0.0046***	−0.0021***	6,020
	(0.0003)	(0.0005)	(0.0006)	
Right pupil diameter (mm)	−0.0025***	−0.0042***	−0.0017**	6,020
	(0.0003)	(0.0005)	(0.0006)	
Panel B. Consecutive task within phase — video				
Left pupil diameter (mm)	−0.028+	−0.037***	−0.009	537
	(0.016)	(0.006)	(0.017)	
Right pupil diameter (mm)	−0.020	−0.035***	−0.015	537
	(0.014)	(0.006)	(0.014)	
Consecutive task within phase — work				
Left pupil diameter (mm)	−0.007	−0.011+	−0.004	476
	(0.012)	(0.006)	(0.015)	
Right pupil diameter (mm)	−0.012	−0.013+	−0.001	476
	(0.013)	(0.008)	(0.015)	
Participant fixed effects	Yes	Yes	Yes	
Block order fixed effects	Yes	Yes	Yes	
Task fixed effects	Yes	Yes	Yes	

Notes. Standard errors clustered at the participant level are in parentheses. Each row reports the slope of the eye-specific pupil diameter on accumulated minute (Panel A) or consecutive task index (Panel B) within the phase, in the writing condition (column 1), the editing condition (column 2), and their difference (column 3, editing minus writing). The writing condition is the reference condition. Left and right pupils are modeled separately (not averaged). The difference column is estimated directly from the interaction term and may deviate from the difference of the displayed slopes by one unit in the final digit due to rounding. All models are estimated on the video-versus-work pipeline, and N is the number of phase-level observations. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Table F.4 Study 2: Difference in within-phase slopes of physiological measures during the video phase

	(1) Pupil diameter (mm)	(2) Mean fixation duration (ms)	(3) Heart rate (bpm)
Editing − writing slope difference (per min)	−0.002	6.73	−0.15
	(0.002)	(6.59)	(0.18)
Editing condition slope (per min)	−0.008***	19.34**	0.02
	(0.001)	(5.92)	(0.10)
Writing condition slope (per min)	−0.006**	12.61**	0.17
	(0.002)	(3.99)	(0.16)
Number of observations	2,378	2,583	2,536
Number of participants	100	100	99
Video interview fixed effects	No	No	No
Participant fixed effects	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes
Task fixed effects	Yes	Yes	Yes

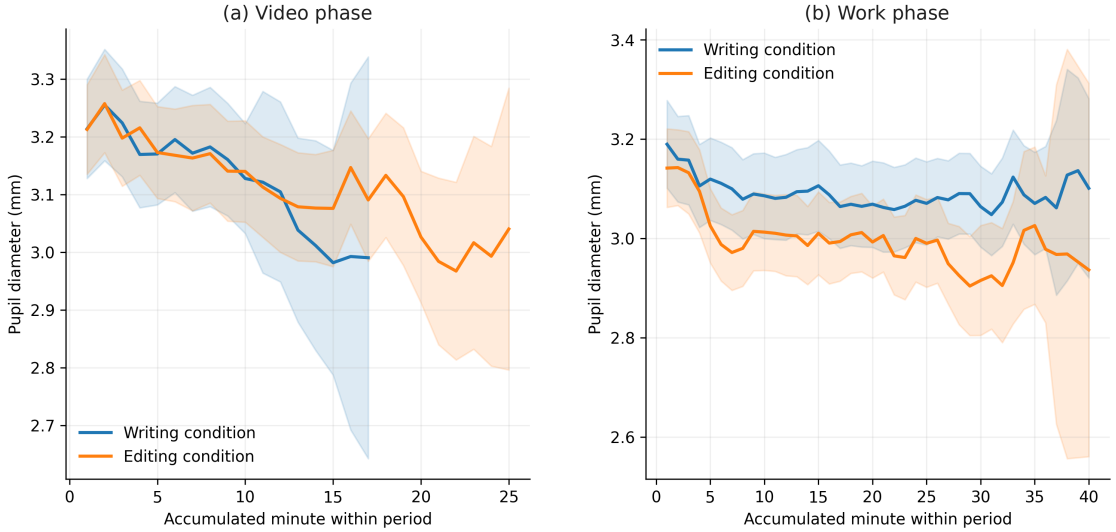
Notes. Standard errors clustered at the participant level are in parentheses. The video phase is the first five minutes of each task, during which the student watches the video interview. Each column reports, for the listed sensor outcome, the slopes on accumulated elapsed minute within the phase: the first row is the editing-minus-writing difference in slopes, and the rows below report the editing condition and writing condition slopes. Slope units are outcome units per minute, and the writing condition is the reference condition. The task fixed effects index the video-interview task being completed. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Table F.5 Study 2: Difference in within-phase slopes of physiological measures during the work phase

	(1) Pupil diameter (mm)	(2) Mean fixation duration (ms)	(3) Heart rate (bpm)
Editing − writing slope difference (per min)	−0.002**	0.94	0.04
	(0.001)	(1.85)	(0.06)
Editing condition slope (per min)	−0.004***	−0.29	−0.01
	(0.0005)	(1.63)	(0.04)
Writing condition slope (per min)	−0.002***	−1.23	−0.05
	(0.0003)	(1.09)	(0.04)
Number of observations	6,020	6,389	6,428
Number of participants	100	100	99
Video interview fixed effects	No	No	No
Participant fixed effects	Yes	Yes	Yes
Block order fixed effects	Yes	Yes	Yes
Task fixed effects	Yes	Yes	Yes

Notes. Standard errors clustered at the participant level are in parentheses. The work phase is the remainder of each task after the first five minutes, during which the student writes or edits the summary. Each column reports, for the listed sensor outcome, the slopes on accumulated elapsed minute within the phase: the first row is the editing-minus-writing difference in slopes, and the rows below report the editing condition and writing condition slopes. Slope units are outcome units per minute, and the writing condition is the reference condition. The task fixed effects index the video-interview task being completed. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Figure F.3 Study 2: Pupil-diameter trajectories during the video and work phases, by condition



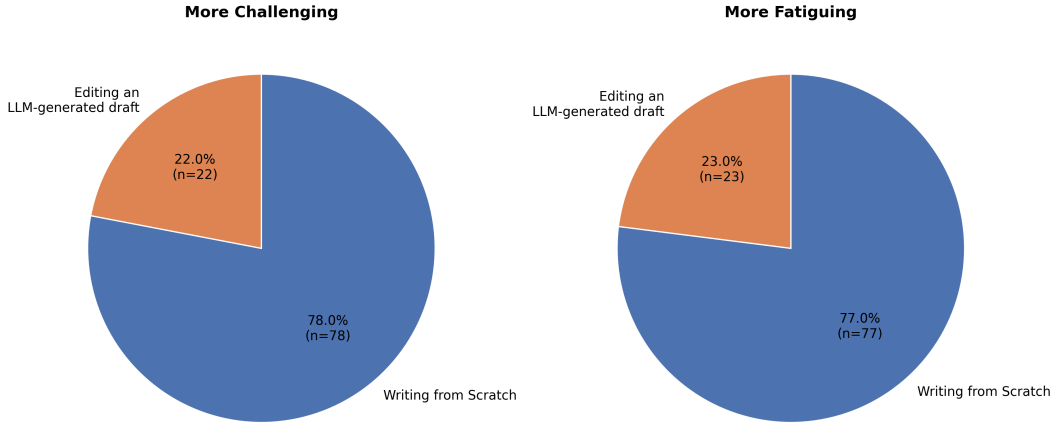
Note. Each panel plots the cross-student average pupil diameter against accumulated elapsed minute within the phase, with shaded 95% confidence bands. Left: video phase (watching the video interview). Right: work phase (writing or editing).

Table F.6 Study 2: Difference in across-task slopes of physiological measures during the video and work phases

Outcome	Writing slope (1)	Editing slope (2)	Editing−Writing (3)	*N*
Panel A. Video phase (consecutive task index)				
Pupil diameter (mm)	−0.02	−0.04***	−0.01	537
	(0.02)	(0.006)	(0.02)	
Mean fixation duration (ms)	14.54	83.05*	68.52+	544
	(24.45)	(35.61)	(36.86)	
Heart rate (bpm)	0.66	−0.19	−0.85	536
	(1.08)	(0.54)	(1.10)	
Panel B. Work phase (consecutive task index)				
Pupil diameter (mm)	−0.01	−0.01+	−0.002	476
	(0.01)	(0.007)	(0.02)	
Mean fixation duration (ms)	149.13*	81.46*	−67.67	481
	(71.07)	(31.84)	(74.59)	
Heart rate (bpm)	−0.10	0.37	0.47	476
	(1.02)	(0.44)	(1.00)	
Participant fixed effects	Yes	Yes	Yes	
Block order fixed effects	Yes	Yes	Yes	
Task fixed effects	Yes	Yes	Yes	

Notes. Standard errors clustered at the participant level are in parentheses. Each row reports the slope of the listed sensor outcome on the consecutive task index within the phase, in the writing condition (column 1), the editing condition (column 2), and their difference (column 3, editing minus writing). The writing condition is the reference condition. Pupil diameter is the matched average. All models are estimated on the video-versus-work pipeline, and *N* is the number of task-phase observations for that outcome. $+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

Figure F.4 Study 2: Student reports of which condition was more challenging and more fatiguing



Note. Exit-questionnaire responses ($N = 100$). The left panel plots the proportion of students who found writing from scratch versus editing the LLM-generated draft more challenging. The right panel plots the proportion of students who found writing versus editing more fatiguing.