Cognitive offloading and the speedup illusion in human-AI interaction

Sunny Yu[1], **Myra Cheng**[1], **Ahmad Jabbar**[2], **Ilia Sucholutsky**[3], **Katherine M. Collins**[4,5], **Dan Jurafsky**[1,2], **Robert D. Hawkins**[2]
syu03@stanford.edu
[1]Department of Computer Science, Stanford University
[2]Department of Linguistics, Stanford University
[3]New York University
[4]MIT
[5]Princeton AI Lab

Abstract

Large language models (LLMs) have the potential to boost human productivity by speeding up task completion—provided users know when to offload cognitive work to them. But we do not know if users are well-calibrated in estimating these potential time savings. We conducted a preregistered large-scale behavioral study (N = 1237) to characterize mismatches between expectations and reality, with a focus on simple cognitive tasks. While actual completion times between independent completion and AI-assisted completion did not differ, participants predicted AI to be significantly faster. The same bias was not observed when imagining help from another human participant. We identify a *speedup illusion* where people have accurate forecasts of independent completion times but significantly underestimate AI-assisted times. Additionally, time and effort dissociate: participants reported lower subjective effort with AI despite equivalent completion times. This suggests that completion time itself is not sufficient to characterize efficiency gains.

Keywords: human-AI interaction; cognitive effort; cognitive offloading; AI use; calibration;

Introduction

People routinely offload cognition to external resources: writing things down, using calculators, consulting others, or searching the web (Fan, Bainbridge, Chamberlain, & Wammes, 2023). From a resource-rational perspective, such offloading reflects adaptive allocation of limited cognitive resources (Griffiths et al., 2019; Lieder & Griffiths, 2020). The decision of whether to offload a given task depends on a cost-benefit comparison: people should delegate to external tools when internal costs exceed the costs of using external support (Risko & Gilbert, 2016). But this comparison requires *calibration*, an accurate mental model of both one's own capabilities and those of the external tool.

For many cognitive tools, people appear reasonably calibrated: they offload more when internal memory is taxed and when tasks are difficult (Dunn & Risko, 2016; Wahn, Schmitz, Gerster, & Weiss, 2023). Large language models (LLMs) represent a new class of cognitive tool, one that can serve as a "thought partner" (Collins, Sucholutsky, et al., 2024) with broader and more variable capabilities than calculators or search engines (Hooper, 2025; Xiong et al., 2024). The prevailing narrative suggests these tools dramatically boost productivity (Appel et al., 2026; Handa et al., 2025; Wang et al., 2025).

While AI assistance offers efficiency gain on complex tasks that would otherwise take a long time for humans to complete independently, it is unclear whether using AI can save time on simpler problems. Metacognitive monitoring provides us with imperfect but usable signals about our own cognitive processes (how long a task will take, how difficult it feels, whether we're on track; Koriat, 2015). We lack such privileged access to the internal workings of an LLM. This asymmetry suggests people may be systematically biased in their predictions of AI-assisted completion times compared to their independent completion times: a *speedup illusion*. Emerging evidence is consistent with this concern: Becker, Rush, Barnes, and Rein (2025) found that AI assistance slowed down coding by 19% despite an expectation for speedup. Other studies show mixed results, with some finding AI reduces subjective effort (Stadler, Bannert, & Sailer, 2024) and others finding no effect (Dhillon et al., 2024).

Are people well-calibrated about how much time using AI can save? More specifically, can people identify that AI assistance does not necessarily save time on simple cognitive tasks, or do they overestimate the extent to which AI assistance leads to an efficiency gain? Based on the metacognitive access argument (Overgaard & Sandberg, 2012), we hypothesize that there is an *asymmetric miscalibration*: people are reasonably calibrated about their own completion times but systematically underestimate the AI-assisted time.

To test the hypothesis, we collected data for predicted and actual completion times for both independent and AI-assisted work. We conducted a pre-registered, large-scale behavioral study ($N = 1,237$)[1] to test the prediction. In the prediction sample, participants predicted how long tasks would take (independently or with assistance); in the completion sample, they actually completed the tasks themselves, either independently or with AI assistance. Tasks spanned four categories of cognitive work, from simple information retrieval to content creation. Our results support the asymmetric miscalibration account: participants accurately predicted their independent completion times but significantly underestimated how long tasks would take with AI assistance; participant miscalibration holds across task difficulty. Furthermore, while AI assistance did not reduce actual completion times, it reduced people's experienced subjective effort. Together, these findings suggest that a "speedup illusion" persists even on easy

[1]This study was approved by the Stanford Institutional Review Board (IRB) under protocol 83204 and pre-registered at https://osf.io/8x9j6. The data and code can be accessed at https://github.com/sunnyych/cognitive_offloading_cogsci.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

tasks — a finding especially worrying if miscalibration encourages AI use, forming a feedback loop where AI use leads to further miscalibration.

Methods

Problem Formulation

For a task τ, we notate the time for a human to complete the task independently as $t_H(\tau)$ and the time for a human to complete the task with AI assistance as $t_A(\tau)$. In addition to the actual completion times, people's expectation for how long a task would take to complete may vary w.r.t. whether the task is being completed independently or with AI assistance. Call a human *well-calibrated* w.r.t. task completion time if they are accurate in their prediction of the completion time of the task. We can compare the predicted and actual $t_H(\tau)$ and $t_A(\tau)$ to ascertain whether people are well-calibrated about their independent and AI-assisted completion times, respectively. The direction and difference between $t_H(\tau)$ and $t_A(\tau)$ reveal whether AI assistance saves time. Similarly, replacing time with self-reported subjective effort, we can examine whether using AI reduces the subjective cognitive effort, thereby exploring whether time aligns with subjective effort.

Task Construction

We constructed a total of 24 tasks (see Figure 1) based on a taxonomy of LLM use, capturing four broad categories of cognitive skills required by different tasks: C1 – Information Seeking, C2 – Information Processing & Synthesis, C3 – Procedural Guidance & Execution, and C4 – Content Creation & Transformation (Shelby, Diaz, & Prabhakaran, 2025). We included a total of six tasks for each category, reflecting distinct task types in their taxonomy. The tasks spanned two difficulty levels, differing in the amount of cognitive effort required (e.g., "Name one Olympic medalist." vs. "Name ten Olympic medalists.").

Experiment setup

To measure people's completion times and how their expectations for them differed from reality, we included a prediction sample and a completion sample. While a within-subject design would provide more direct insights into calibration patterns at an individual level, it would bias the actual completion times and effort if participants were already familiar with the tasks before completing them. A between-subject design, on the other hand, minimizes downstream biases on the variables and provides insights about average differences.

Prediction sample. In the prediction sample, participants were randomly assigned to one of two hypothetical conditions. Without having to complete the task while having a detailed task description, a participant was asked to predict how long it would take to complete the task independently and with AI assistance or with assistance from another human participant. For each task, participants also stated whether they would choose to complete the task independently or with the external assistance of AI/human. After making predictions, participants completed the Need for Cognition scale (Cacioppo & Petty, 1982). Participants were instructed to not use any form of AI, and copying and pasting was disabled.

Completion sample. In the completion sample, participants were randomly assigned to complete tasks either independently or with AI assistance. For the AI assistance condition, we provided an embedded chat interface with GPT-4o on the task page that participants could interact with. Participants were presented a randomized mix of easy and difficult tasks. Each task was completed by around 68 participants for each condition. We used a hidden timer to record the completion time for each task. After the completion of each task, participants answered the 5-question version of the NASA-TLX, a measure of subjective effort (Hart & Staveland, 1988). The questions evaluate how mentally demanding (Q1), hurried or rushed (Q2), and successful (Q3) the participant felt completing the task, as well as how hard the participant had to work (Q4), and how insecure, discouraged, and stressed they felt (Q5) when completing the task. To understand people's calibration and efficiency gain when it does not affect the outcome, we filtered out incorrect responses and focused on correct answers only and analyzed the corresponding time and effort on these tasks. To exclude low-effort responses (which would correspond to shorter completion times), we annotated all final answers and excluded ones that were incorrect (for questions with verifiable answers), low-effort, or failed to address the prompts to ensure that all responses were high-quality. In general, participants made a decent effort to complete the tasks across both of the conditions; we excluded 6.3% of the responses in the independent condition and 4.0% in the AI assistance condition.

Participants. Participants were recruited through the Prolific platform, comprising a representative sample of the US adult population. Our sample size was N=1237 in total: 401 in the prediction sample and 836 in the completion sample. The participants were 48% female, 47% male, and 5% other; 58% white, 11% Black, 10% mixed, and 6% Asian.

Results

People expect offloading to AI to be more efficient. Before examining the calibration gaps, we first confirm that people generally expect AI assistance to reduce task completion times. We fit a linear mixed-effects model with prediction target (independent vs. assisted) and assistance source (AI vs. another participant) as fixed effects, including participant and task-level random intercepts.

Our model reveals that people expect AI assistance to reduce task completion times by 68.5 seconds ($\beta = 68.5$, $SE = 3.37$, $z = 20.34$, $p < 0.001$). To determine whether the difference observed above is unique to AI assistance, we also ask them about "another participant" who is highly intelligent[2]) as a baseline. We found that participants estimated in-

[2]Without an anchor, participants' predictions for "another participant" would conflate two separate beliefs: the average competence of an unknown other and the predicted speedup given that competence. Anchoring on "highly intelligent" matches the implicit competence assumption participants likely bring to the AI condition

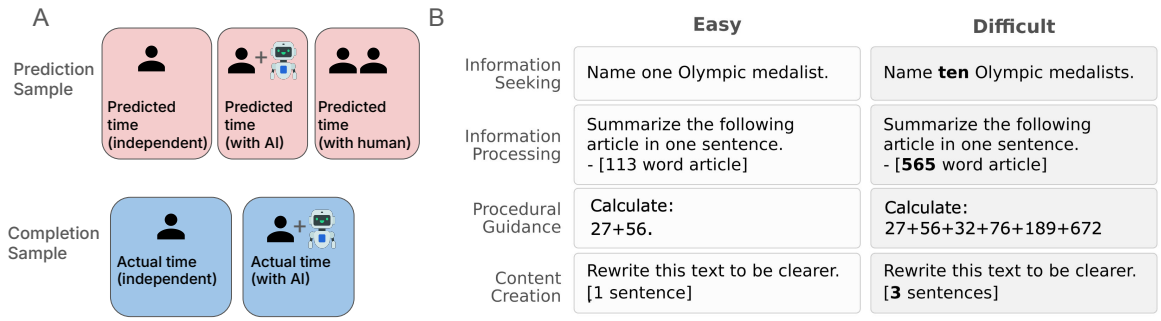

Figure 1: **Experiment setup**: **a)** we include a prediction sample and a completion sample. In the prediction sample, participants predict the completion times for themselves and for using external assistance. In the completion sample, participants complete the task independently or with AI assistance. **Example tasks**: **b)**, we show one example per category × difficulty level (8 of the 24 tasks).

dependent completion to take significantly longer than completion with the help of another participant ($\beta = 17.10$, $SE = 3.36$, $z = 5.10$, $p < 0.001$), although the difference is significantly greater for AI assistance than for another participant ($\beta = -51.4$, $SE = 4.76$, $z = -10.80$, $p < 0.001$), meaning that on average, participants expected offloading to AI to be more efficient than offloading to another participant. At a task level, participants predicted AI assistance to speed up completion times for 18 tasks, and predicted assistance from another participant to speed up 5 of the tasks. The finding is consistent with participants' stated preferences: when asked how they would complete the task, people preferred AI assistance more than assistance from another participant ($\beta = 1.37$, $SE = 0.26$, $z = 5.34$, $p < 0.001$). The results show that people expect **assistance from AI** to be particularly effective in reducing completion times.

People miscalibrate the time that AI assistance saves. To investigate whether people's predictions of completion times with AI assistance is aligned with reality, we fit the following linear mixed-effect model: time $\sim$ condition $*$ type $+$ difficulty $+ (1 \mid \text{participantID}) + (1 \mid \text{taskID})$, where time is the prediction or completion times in seconds, condition is prediction vs. completion, and type is independent vs. AI assistance. We found that people failed to reliably calibrate how much time AI assistance saves (Figure 2): the actual completion time was almost one minute greater than the predicted time ($\beta = 57.8$, $SE = 8.79$, $z = 6.57$, $p < 0.001$). The speedup illusion holds across all four task categories and both difficulty levels.

In stark contrast to the AI assistance condition, we found no significant difference between the actual and predicted completion times in the independent completion condition ($\beta = -2.52$, $SE = 8.83$, $z = -0.29$, $p = 0.775$), which suggests that people's predictions are generally accurate. However, two category-wide differences emerge: participants underestimated their completion time for C2 ($\beta = 29.2$, $SE = 11.7$, $z = 2.49$, $p < 0.05$) but overestimated for C4 ($\beta = -24.25$, $SE = 11.7$, $z = -2.07$, $p < 0.05$). We observe no significant effect across difficulty levels.

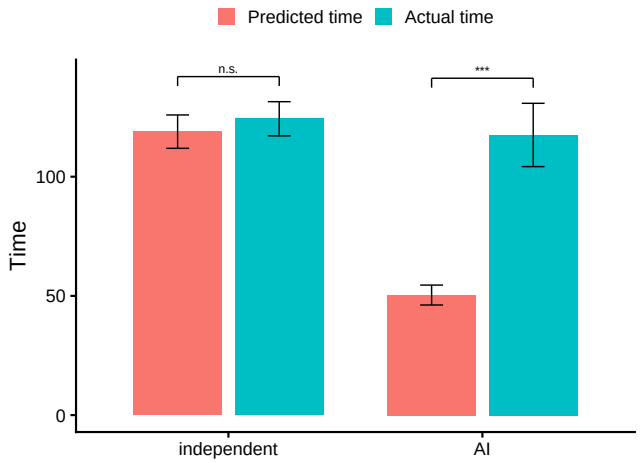

Figure 2: There is no significant difference between the predicted and actual completion times for the independent completion condition (left), but actual time is significantly greater than predicted time for the AI condition (right).

We found that while the speedup illusion existed across all tasks, AI assistance only made difficult tasks faster, and not easy ones: task completion with AI assistance was faster by 26.14 seconds ($SE = 11.9$, $z = 2.19$, $p < 0.05$) on difficult tasks, but no difference between independent completion time and AI-assisted completion time was found for easy tasks. Across individual tasks, the completion times only became significantly faster with AI assistance for three tasks (coming up with 10 other words for "enjoy", summarizing a long article, and editing a long text; see Figure 3). The results highlight that when it comes to basic cognitive tasks like providing instructions for boiling an egg or finding spelling errors in a sentence, people do not have a good mental model of whether using a resource-rational strategy also makes the

for these tasks (a capable assistant).

completion more efficient. On easy tasks, conserving effort does not translate to a conservation of time.

People's individual traits also influenced their decision to offload cognitive efforts. We examine whether participants' willingness to think (as measured by the Need For Cognition scale) predicts the magnitude of calibration errors. We found that participants who are more averse to thinking were more likely to believe that AI assistance saves more time. Specifically, we focus on the prediction sample and identify negative coefficients for the difference between the predicted independent versus assisted completion times for item 3 ("thinking is not my idea of fun"; $\beta = -6.55$, $SE = 3.14$, $t = -2.11$, $p < 0.05$) and item 4 ("I would rather do something that requires little thought than something that is sure to challenge my thinking abilities"; $\beta = -6.66$, $SE = 3.19$, $z = -2.049$, $p < 0.05$) of the NFC scale. This means that the more participants dislike having to think, the more susceptible they are to the speedup illusion, believing that using AI can be significantly faster than doing the task on their own. The finding is consistent with previous research indicating a stronger effect between perceived difficulty and subjective cognitive effort for individuals with low ability or conscientiousness (Yeo & Neal, 2008). On the other hand, variables related to people's familiarity and experience with AI — namely AI-use frequency and the AI Assessment scale (AIAS) (Grassini, 2023) — do not predict the magnitude of calibration error.

Subjective mental effort differs from time. While cognitive offloading is typically measured with time (Handa et al., 2025; Wang et al., 2025), it is also related to the *experience* of mental effort, which in essence is subjective (Steele, 2020). Bearing in mind that the time taken may not capture the subjective experience of mental effort, we sought to directly record people's mental effort using NASA-TLX index. Our results suggest that subjective mental effort could be a possible reason for the calibration error. Overall, there was a positive and weak correlation between completion time and the NASA-TLX average ($r = 0.26$). However, the two measurements differ fundamentally: while AI assistance sped up completion of only a few tasks, it reduced mental effort for all tasks — on average, the NASA-TLX average (higher means more effortful) lowered by 0.61 points on a 7-point scale ($SE = 0.059$, $t = -10.38$, $p < 0.001$).

The significant reduction in perceived subjective effort holds for 15 out of 24 tasks. Comparing the effort in the independent and AI conditions provides us with an informative understanding of the reduction of mental effort with AI use. An AI-assisted reduction in mental effort across all easy and difficult tasks may have contributed to an overestimation of the time saved with AI assistance. This provides us with an understanding of a possible source of the miscalibration error.

How are people prompting AI? To understand the different ways that participants prompted the LLM, we provide an in-depth analysis of prompting behaviors and user-LLM interactions in the experiment. More than 70% of the interactions were single-turn, and the maximum number of turns was 5 for all tasks. Out of all the prompts, 18.5% were directly copied and pasted from the instructions. Copying and pasting did not significantly reduce completion time, but it reduced NASA-TLX (average) by 0.14 points ($SE = 0.058$, $z = 2.38$, $p < 0.05$). The finding reveals that when completing a task with AI assistance, much of the cognitive effort comes from writing the prompt.

To verify if prompting took up the majority of the task completion time, we decomposed the completion time into three components: the time it takes to construct the prompt, the time for the model to generate a response, and the time the user reads and processes the model response, i.e., $t_A(\tau) = t_P(\tau) + t_M(\tau) + t_R(\tau)$[3]. Overall, the time it took for the model to generate a response was minimal (2.89 seconds), so we compare prompting with post-response processing. Across all tasks, we found no significant difference between prompting and response processing time. But we found evidence that prompting could take significantly longer for some tasks, while taking less time for others: prompting took 33.4 seconds longer ($SE = 12.1$, $z = 2.76$, $p < 0.01$) than post-response processing for C2 tasks. It is possible that people's increased familiarity with AI systems and previous prompting history made it easier and faster for them to write prompts (Korzynski, Mazurek, Krzypkowska, & Kurasinski, 2023).

On the other hand, we found that post-response processing took 113.9 seconds longer than prompting ($SE = 29.3$, $z = -3.89$, $p < 0.001$) on the logic problem. An analysis of the conversation histories for this question reveals that GPT-4o tended to produce verbose answers that are difficult to understand (see example below). For the same question, AI-assisted completion times were significantly longer ($\beta = -110.66$, $SE = 35.1$, $z = -3.16$, $p < 0.01$), which contradicted with people's prediction that AI assistance should speed up the task by more than 2 minutes ($\beta = 136.31$, $SE = 14.2$, $z = 9.59$, $p < 0.001$). The question highlights a case where people need to spend more time and effort reading and understanding a model's response.

> **Task:** If it is raining and it is sunny, then there is a rainbow. It is not the case that there is a rainbow. Therefore, either it is not the case that it is raining or it is not the case that it is sunny. Construct a similar argument with the statements: "There is water in the pot", "The pot is on the stove", "The water gets heated.
> **Assistant:** ...This argument follows a similar logical pattern to the original, using the form of modus tollens: - If $P \wedge Q$, then R. - Not R. - Therefore, not P or not Q.

[3]The prompting time is calculated by taking the difference in time stamps between the first model message and the last user prompt. The model response generation time is calculated as the difference in the time stamp between the user's last prompt and the model's final response. Finally, the post-response time is calculated as the difference between the task completion time (the variable we use for all analyses above) and the time between the first model message and the last model message.

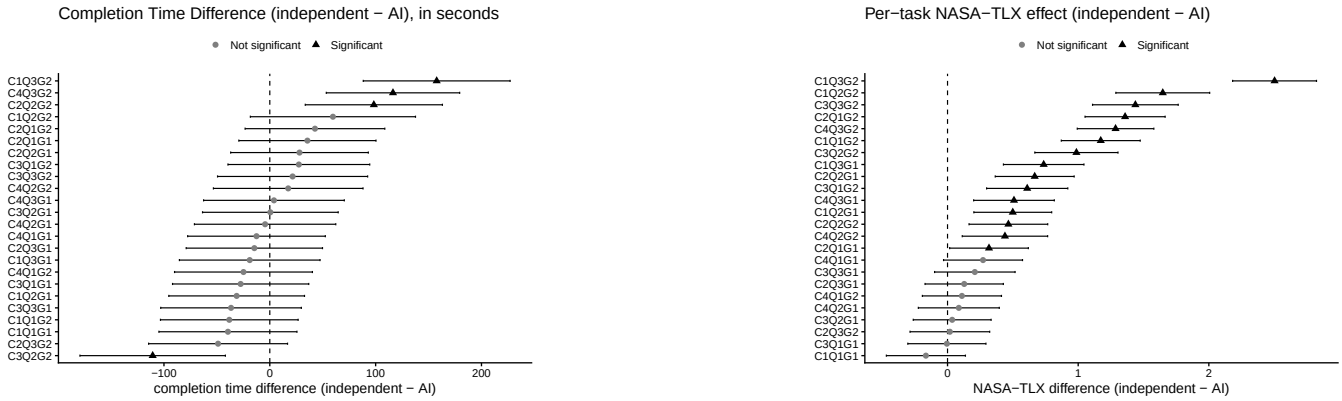

Figure 3: The difference between the two conditions per task in terms of time (left) and NASA-TLX (right), ordered by effect size. The completion time difference is only significant for 3 tasks. AI assistance leads to offloading (measured by NASA-TLX) for 15 of the tasks (C denotes category, Q denotes question, and G denotes difficulty level, where 1 is easy).

Where: - *P*: "There is water in the pot." - *Q*: "The pot is on the stove." - *R*: "The water gets heated."...

User: Umm—this is getting a bit abstruse. Can we summarize this proposition in simpler terms?

While copying the tasks directly as prompts was common, we also observed instances where participants wrote very different prompts even for the same task. For example, for the task on naming Olympic medalists, many participants had a specific category or name in mind and used AI as a recall tool (Table 1). Some participants used LLMs to confirm or expand on their existing answers. For the trolley problems, where participants had to justify whether to divert the trolley, some participants asked AI to articulate the reasoning rather than making the decision (e.g., "Give me three reasons in bullet format of why I should pick the 10 students and three reasons in bullet format of why I should pick the five retired doctors"). When using AI to summarize a satire article, some participants chose to have the AI expand on their own summary of the article rather than delegating the AI to do the task entirely, e.g., "[The article is] too long to paste, but it's about the many rules and regulations that airlines have that can be burdensome and frustrating." In these cases, the participants already expended the majority of the effort required to complete the task prior to receiving the AI answer. The decision to do so may reflect a resource-rational choice to conserve both time and effort.

Finally, participants actively engaged with the model responses and proactively asked questions like "Why did you leave out the university's struggles to impliment policy regarding AI" for the article summarization question or "Why do you have the temperature so high in the beginning?" for the explanation question. The wide range of behaviors showcase how differently participants prompted the model and their different degrees of cognitive engagement.

Discussion

People conflate effort reduction with time reduction. Decades of cognitive science research has studied people's biases in the perception and estimation of time, which can be affected by factors like the pleasantness of the experience (Fraisse et al., 1984; Fredrickson & Kahneman, 1993). Like retrospective biases, prospective duration can also be biased: Zauberman, Kim, Malkoc, and Bettman (2009) finds that subjective time perceptions are logarithmic in objective time (Zauberman et al., 2009), and Liu and Wickens (1994) finds that time estimation was sensitive to the presence of cognitive demands. Various factors from existing research on subjective time perception could explain the speedup illusion. For example, previous studies find that conditions that favor vivid representations and efficient perceptual decision-making lead to longer perceived durations — it is possible that people lack a vivid representation of task completions with AI and thus underestimate the time it would take (Matthews & Meck, 2016). Another important factor could be that people are conflating time reduction with effort reduction. A task can take a long time to complete — even with the help of AI — but the process could feel effortless. All the user had to do was write out the prompt and wait for the model to come up with an answer. This very difference could explain why a speedup illusion persists: when there is AI help, a task just does not *feel* long, but when the help is gone, the task *feels* long to complete. It is likely that when people offload, they start to miscalibrate about other resources being conserved, such as time. Our findings highlight the subjective nature of effort and how time reduction alone cannot fully capture the experienced efficiency gain. Besides time reduction, narratives and reports about AI productivity should also consider a more encompassing notion of cognitive offloading that takes into account the subjective experience.

Prompt
I would like to know who won gold in the 400m men's freestyle in the 2016 Olympics.
Tell me who won the women's snowboarding gold medal in the last Winter Olympics.
Can you name the Olympic medalist with the highest number of medals?
NAME AN OLYMPIC MEDALIST IN SKIING.
Please provide a list of men and women 10 gold medalists in the 100m freestyle in the last 10 Olympics.

Table 1: Examples of participants' prompts to AI for the "name an Olympic medalist" task that were not directly copied.

A feedback loop: risks of overreliance. In this paper, we identify a speedup illusion of AI, which persists across task domains and difficulty levels, meaning that people expect AI to make task completion faster even when it does not. Studying calibration in the context of AI use is particularly important because it sheds lights on the factors that shape people's decision to (and potentially continue to) use AI systems. As resource-constrained reasoners (Griffiths, 2020), if people expect AI assistance to reduce time, they are more likely to delegate tasks to AI. As cautioned in recent studies, an excessive amount of offloading could lead to cognitive deskilling (Ahn, 2025; Gerlich, 2025; Oktar et al., 2026), which will impair people's cognitive capabilities and further slow down their independent completion time. The tendency risks forming a negative feedback loop where people use LLMs as steroids and offload at a disruptive level (Collins, Bhatt, & Sucholutsky, 2025; Hofman, Goldstein, & Rothschild, 2023; Jose et al., 2025), which could undermine people's own cognitive abilities and foster overreliance (Ibrahim et al., 2025). Even more jarringly, we provide preliminary evidence for a feedback loop where miscalibration encourages AI use as people overestimate the benefits of AI assistance, and using AI and experiencing effort saving leads to further miscalibration, which will lead to increased AI use.

While blindly copying prompts and pasting model responses significantly reduced cognitive effort, it could provide a shortcut that ultimately leads to cognitive surrender and disempowerment (Shaw & Nave, 2026; Sturgeon, Samuelson, Haimes, & Anthis, 2025). Our results on the distinction between time and effort reduction also show that people need to re-evaluate what resource (time versus effort) that they are trying to conserve. Yet, our qualitative analysis also indicates that many participants prompted LLMs in ways that required them to think, pointing to the possibility of LLMs allowing users to critically engage with the questions (Shum, 2024). Similar to how LLMs can create an illusion of understanding, our work highlights people's skewed mental models and invite future work to more systematically examine the downstream outcomes of such misalignments (Bansal et al., 2019; Kelly, Kumar, Smyth, & Steyvers, 2023; Messeri & Crockett, 2024). More importantly, our work highlights the importance of calibration as both a subject and mechanism of study for understanding AI use and human-AI interaction more broadly; calibration can also be a key intervention point to adjust behavioral outcomes such as overreliance.

Limitations and Future Directions

Our study has various limitations that point to opportunities for future work. First, we focus on tasks that can be completed under 5 minutes; an interesting direction of future work is to identify the specific complexity level at which AI assistance genuinely saves time and effort. Second, our experiment did not directly control for participant motivation or incentives. Although we manually went through the responses and excluded low-effort and incorrect responses, crowdworking contexts may impact task completion time, response quality, and AI adoption rates compared to other settings with more personal stakes (Yurt & Kasarci, 2024).

Nonetheless, this setting enables us to understand overall miscalibrations in AI use. Future work should examine how incentive structures and motivation levels (e.g. answer quality) interact with the completion times and reported effort. Third, another limitation of the study is individual variations in AI use. Participants used AI in very different ways, and the degree of cognitive effort required even within the AI condition could vary. For example, it is possible that participants prompted the model but came up with a different final answer, regardless of what the AI response was. It was also possible that some participants only used AI to double check their answer or prompted AI out of curiosity of the response without using the responses.

In our completion sample, we assign participants to one of the two completion modes. Future work should investigate *when* people choose to use AI to understand the rate at which people are using AI and what resource (e.g. effort vs. time) they preserve, as well as the underlying mechanisms that lead to the decision to or not to use AI (Collins, Chen, et al., 2024).

Finally, our study relies on between-subjects aggregates to avoid carryover biases. The prediction sample and completion sample were also not completely disjoint, with some participants completing both samples on Prolific. Future work should examine within-subject patterns while minimizing biases in the dependent variables.

Conclusion

Through a large-scale behavioral study on cognitive offloading, we highlight an asymmetric miscalibration and show that AI assistance can create a speedup illusion where tasks feel shorter to complete because the process becomes less effortful. We point out the dissociation between time and effort, finding that AI assistance always leads to effort reduction but not time reduction.

Acknowledgments

We would like to thank members of the Social Interaction Lab (SoIL) and the Jurafsky lab for their valuable feedback on the project. In addition, we thank Polina Tsvilodub and Charley Wu for their helpful feedback on the draft.

References

Ahn, S. (2025). Preserving critical thinking in the age of large language models: The paradox of cognitive load and efficiency. *The Korean Journal of Medicine*, *100*(5), 197–200.

Appel, R., Massenkoff, M., McCrory, P., McCain, M., Heller, R., Neylon, T., & Tamkin, A. (2026). *Anthropic economic index report: economic primitives.*

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (Vol. 7, pp. 2–11).

Becker, J., Rush, N., Barnes, E., & Rein, D. (2025, 07). *Measuring the impact of early-2025 ai on experienced open-source developer productivity.* https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116.

Collins, K. M., Bhatt, U., & Sucholutsky, I. (2025). Revisiting rogers' paradox in the context of human-ai interaction. *arXiv preprint arXiv:2501.10476*.

Collins, K. M., Chen, V., Sucholutsky, I., Kirk, H. R., Sadek, M., Sargeant, H., ... Bhatt, U. (2024). Modulating language model experiences through frictions. *arXiv preprint arXiv:2407.12804*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., ... Griffiths, T. L. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

Dhillon, P. S., Molaei, S., Li, J., Golub, M., Zheng, S., & Robert, L. P. (2024). Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI conference on human factors in computing systems* (pp. 1–18).

Dunn, T. L., & Risko, E. F. (2016). Toward a metacognitive account of cognitive offloading. *Cognitive Science*, *40*(5), 1080–1127.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568.

Fraisse, P., et al. (1984). Perception and estimation of time. *Annual review of psychology*, *35*(1), 1–37.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology*, *65*(1), 45.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6.

Grassini, S. (2023). Development and validation of the AI attitude scale (aias-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in psychology*, *14*, 1191628.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., ... others (2025). Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hofman, J., Goldstein, D. G., & Rothschild, D. (2023, December). A sports analogy for understanding different ways to use ai. *Harvard Business Review*.

Hooper, V. J. (2025). Cognitive offloading and the reshaping of human thought: The subtle influence of artificial intelligence. In *Colloquia, academic journal of culture and thought* (Vol. 12, pp. 01–14).

Ibrahim, L., Collins, K. M., Kim, S. S. Y., Reuel, A., Lamparth, M., Feng, K., ... Bhatt, U. (2025). Measuring and mitigating overreliance is necessary for building human-compatible AI. *arXiv preprint arXiv:2509.08010*.

Jose, B., Joseph, D., Mohan, V., Alexander, E., Varghese, S. K., & Roy, A. (2025). Outsourcing cognition: the psychological costs of ai-era convenience. *Frontiers in Psychology*, *16*, 1645237.

Kelly, M., Kumar, A., Smyth, P., & Steyvers, M. (2023). Capturing humans' mental models of AI: An item response theory approach. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1723–1734).

Koriat, A. (2015). Metacognition: Decision making processes in self-monitoring and self-regulation. *The Wiley Blackwell handbook of judgment and decision making*, *2*, 356–379.

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as chatgpt. *Entrepreneurial Business and Economics Review*, *11*(3), 25–37.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal

use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. *Ergonomics*, *37*(11), 1843–1854.

Matthews, W. J., & Meck, W. H. (2016). Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psychological bulletin*, *142*(8), 865.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58.

Oktar, K., Collins, K. M., Hernández-Orallo, J., Coyle, D., Cave, S., Weller, A., & Sucholutsky, I. (2026, March). Identifying, evaluating, and mitigating risks of ai thought partnerships. *ACM AI Lett.*.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287–1296.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in cognitive sciences*, *20*(9), 676–688.

Shaw, S., & Nave, G. (2026). Thinking—fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. psyarxiv. *Preprint at https://osf.io/preprints/psyarxiv/yk25n_v1*.

Shelby, R., Diaz, F., & Prabhakaran, V. (2025). Taxonomy of user needs and actions. *arXiv preprint arXiv:2510.06124*.

Shum, S. B. (2024). Generative ai for critical analysis practical tools cognitive offloading and human agency. In *CEUR workshop proceedings.*

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386.

Steele, J. (2020, Jun). *What is (perception of) effort? objective and subjective effort during attempted task performance.* PsyArXiv.

Sturgeon, B., Samuelson, D., Haimes, J., & Anthis, J. R. (2025). Humanagencybench: Scalable evaluation of human agency support in ai assistants. *arXiv preprint arXiv:2509.08494*.

Wahn, B., Schmitz, L., Gerster, F. N., & Weiss, M. (2023). Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm. *Plos one*, *18*(5), e0286102.

Wang, Z. Z., Shao, Y., Shaikh, O., Fried, D., Neubig, G., & Yang, D. (2025). How do AI agents do human work? comparing ai and human workflows across diverse occupations. *arXiv preprint arXiv:2510.22780*.

Xiong, H., Bian, J., Li, Y., Li, X., Du, M., Wang, S., ... Helal, S. (2024). When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing*.

Yeo, G., & Neal, A. (2008). Subjective cognitive effort: A model of states, traits, and time. *Journal of Applied Psychology*, *93*(3), 617.

Yurt, E., & Kasarci, I. (2024). A questionnaire of artificial intelligence use motives: A contribution to investigating the connection between ai and motivation. *International Journal of Technology in Education*, *7*(2).

Zauberman, G., Kim, B. K., Malkoc, S. A., & Bettman, J. R. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. *Journal of Marketing Research*, *46*(4), 543–556.