When an AI Judges Your Work: The Hidden Costs of Algorithmic Assessment*

David Almog[†], Lucas Lippman[‡], and Daniel Martin[§]

January 5, 2026

Abstract

We use an online experiment with a real work task to study whether workers change their behavior when they know AI will be used to judge their work instead of humans. We find that individuals produce a higher quantity of output when they are assigned an AI evaluator. However, controlling for quantity, the quality of their output is lower, regardless of whether quality is measured using humans or LLM grades. We also find that workers are more likely to use external tools, including LLMs, when they know AI is used to judge their work instead of humans. However, the increase in external tool use does not appear to explain the differences in quantity or quality across treatments.

*IRB approval was obtained at the University of California, Santa Barbara.

[†]Kellogg School of Management, Northwestern University, david.almog@kellogg.northwestern.edu.

[‡]Walmart Connect.

[§]University of California, Santa Barbara, danielmartin@ucsb.edu.

1 Introduction

The emergence of Generative AI (GenAI), primarily driven by Large Language Models (LLMs), has radically expanded the set of tasks that AI can evaluate, such as the tone and content of written documents and audio transcripts. Given this, firms and institutions are increasingly considering whether to use AI for assessments. In work settings, it is starting to be rolled out for performance reviews and hiring (Goergen, Bellis, and Klesse 2025). In academic settings, it is being considered for grading assignments (D. W. Zhang, Boey, Tan, and Jia 2024; Flodén 2025) and application selection[1]. In economic experiments, LLMs are being used to incentivize performance (Conlon 2025).

Why would firms, institutions, and experimentalists use AI assessments instead of human ones? One key reason is that AI evaluations can be less expensive. This is especially true when human assessment requires domain experts, such as managers. AI evaluations can also be quicker, allowing for immediate feedback, faster analyses, and rapid payments. Finally, AI evaluations can be more consistent because the trained model can be held fixed, allowing every worker or document to receive the same evaluation.

In fact, our study illustrates the strong incentives that exist to implement AI assessment based on time and cost savings alone. In our experiments, using AI to grade all subject output cost $11.67 plus co-author time to write a Python script, which took one day to code and run. On the other hand, human grading took three graduate students an average of 54 hours (nearly seven 8-hour workdays), at a total cost of $6,480.

While AI assessments can be more cost-effective and faster, firms and institutions may hesitate to switch from human assessments if doing so negatively impacts workers' output.[2] An open question is whether worker output under AI assessment is fundamentally different from worker output under human assessment. This question is critical for understanding when and to what extent AI assessments will be adopted by firms and the implications of its adoption.

In this paper, we study the impact of AI assessment on worker output using an online experiment in which subjects were paid to provide captions for images. This is a common task among online workers, including those who complete piece-rate work through platforms and those who undertake more standard job contracts.[3] Captioning images is a widespread task across industries: social media managers caption posts, insurance specialists caption claims for damaged goods, annotators caption training data for machine learning models, librarians and archivists caption stock for search tools, and radiologists caption X-rays to aid

[1] See https://apnews.com/article/ai-chatgpt-college-admissions-essays-87802788683ca4831bf1390078147a6f.

[2] Other reasons they may hesitate could be negative publicity, legal risk, or worker morale. For instance, algorithmic assessments can be perceived as less fair (Mok, Nanda, and Anderson 2023).

[3] See https://www.indeed.com/q-image-captioning-jobs.html.

interpretation.[4] While AI is increasingly used for image captioning, there remain substantial quality concerns around AI-generated captions (Sarhan and Hegelich 2023), so humans are still regularly hired to complete this task.

In our experiment, conducted on the Prolific platform, we asked subjects to write captions between 150 and 400 characters for 20 images of daily scenes, such as a vendor squeezing oranges or kids holding skateboards. Subjects were told that their captions would be evaluated based on three criteria (being interesting, effortful, and accurate) and that the criteria would remain fixed for the entire session.[5] Subjects received a $1 bonus whenever their caption ranked in the top 30% for a given image, so a subject could earn up to $20 in addition to their $8 show-up fee, which is substantially higher than the standard hourly rate on Prolific (Palan and Schitter 2018). Using a between-subject design, we randomly assigned subjects to have their captions evaluated by a human or by ChatGPT, a leading LLM at the time of the study. Importantly, subjects were told who would be evaluating them and were not told about the existence of the other treatment arm. The assigned evaluator ("human assigned" or "ChatGPT assigned") remained fixed for all 20 images. If assigned a human evaluator, subjects were also told truthfully that the evaluator would be a college graduate.

We find that being assigned an AI evaluator produces an increase of 27.8% of a standard deviation in output quantity, as measured by the length of their captions. However, when controlling for output quantity, we find that the quality of captions declines when subjects are assigned to be judged by AI. To determine the quality of captions, we use several benchmarks: the numeric grades assigned by PhD student raters and the numeric grades assigned by GPT-4o under different temperatures and prompts. Controlling for output quantity, being assigned an AI evaluator decreases grades between 12.1% and 20.5% of a standard deviation across our quality benchmarks. These decreases are statistically significant at a 5% level across all benchmarks including demographic controls, image fixed effects, and clustering standard errors at the individual level.

We also find individuals are 93.3% (9.7 pp) more likely to use external tools when their work is evaluated by AI rather than by humans. However, externally assisted captions tend to be of higher quality, so the increased use of external help does not explain the lower quality of captions in the AI-evaluated group. We do not directly observe external tool use, so we consider several ways of identifying whether a subject used an external tools, and our primary proxy is the pasting of text.

What kinds of external tools are subjects using? One possibility is that they are increasing their use of AI tools when being evaluated AI. To help determine these, participants were

[4] In addition to being a regular work task, providing written descriptions is also a standard educational activity.

[5] The selection of these criteria will be explained in detail in the experimental design section.

asked to disclose whether they had used an LLM to help write their description, with a monetary bonus awarded if their response matched the prediction of AI detection software. This provides a novel incentivized elicitation method to identify AI usage. We use this method to show an uptick in AI tool use when being evaluated by AI and then demonstrate that this measure can serve as a useful complement to the pasting metric.

The rest of the paper is organized as follows. Section 1.1 reviews related literature on AI in economics. Section 2 describes our experimental design and key design choices. Section 3 examines the treatment effect of evaluator assignment on writing habits, quality, and the use of external assistance. In Section 4, we discuss the interpretation and implications of our findings. Finally, Section 5 provides a brief summary.

1.1 Literature Review

This paper contributes to multiple strands of economics that consider the implications of AI use on society.

The rise of generative AI is set to disrupt the labor market on an unprecedented scale. Eloundou, S. Manning, Mishkin, and Rock (2024) estimate that 80% of the United States workforce could have at least 10% of their tasks affected by LLMs, with 19% seeing more than half of their tasks impacted. This potential is already beginning to materialize. Using comprehensive Danish registry data, Humlum and Vestergaard (2025) conducted a large-scale survey of workers and found that at least half of the respondents had used ChatGPT for work. Similarly, Bick, Blandin, and Deming (2024), using a representative survey of the U.S. population aged 18-64, report that by August 2024, 39% had used Gen AI.[6] The vast majority of research to date has focused on the productivity gains associated with providing workers access to generative AI as a tool (e.g., Noy and W. Zhang 2023; Brynjolfsson, Li, and Raymond 2025; Peng, Kalliamvakou, Cihon, and Demirer 2023; Choi and Schwarcz 2025; Otis, Clarke, Delecourt, Holtz, and Koning 2025). However, there is significant potential for AI adoption in other areas of the workplace environment (Ludwig, Mullainathan, and Rambachan 2024), with one such application being its role as an assessment mechanism.

Autor, Chin, Salomons, and Seegmiller (2024) and Acemoglu and Restrepo (2019) discuss how technological shocks can lead to both automation and the creation of new tasks. Along those lines, Ide and Talamas (2025) adapt the canonical model of the knowledge economy to provide organizational insights into how AI may reshape hierarchical firms. To accurately assess which tasks are at risk of human displacement and how this restructuring of firms will unfold, it is essential to evaluate the general equilibrium effects of integrating AI into workflows, including human responses to its presence. Our paper contributes to this litera-

[6] These figures are likely to have increased since then.

ture by examining the incentives behind AI adoption. We introduce a new trade-off to guide decisions on when to utilize AI as an assessment tool, demonstrating that the response of human workers to being monitored by different entities (AI versus human) may significantly influence that decision.

We also contribute to the literature on monitoring, particularly to the subset of studies examining behavioral responses to being monitored. Contrary to the predictions of principal-agent theory, Frey (1993) proposed that monitoring might negatively impact workers' effort by reducing intrinsic motivation. This "crowding-out" phenomenon was later carefully tested by Dickinson and Villeval (2008), who provided evidence that monitoring diminishes effort once it exceeds a certain threshold. Using a field experiment on call centers, Nagin, Rebitzer, Sanders, and Taylor (2002) find that workers may strategically respond to monitoring, consistent with a model of "rational cheaters". If workers adjust their behavior in response to monitoring, and humans behave differently in strategic interactions with computer players compared to other humans[7], it is reasonable to expect distinct responses to AI versus human assessment. This paper provides evidence for this distinction and explores how these differences may emerge.

The two most closely related papers to ours are Almog, Gauriot, Page, and Martin (2024) and Corgnet, Hernán-González, and Mateo (2023). The former provides evidence that the introduction of AI assessment impacted both the extent and nature of mistakes made by umpires in professional tennis. However, there was no human assessment of umpire decisions before the introduction of AI assessment, so that setting cannot be used to answer our question of interest. To actually address whether workers' responses differ between human and AI judgment, we implement a task where both monitoring designations can be carried out in a comparable manner. The latter paper investigates the impact of substituting humans with algorithms on peer effects. Their findings suggest that introducing algorithms significantly affects workers' performance by reducing social pressure, an important driver of performance in their context. A key distinction is that, in their study, algorithms serve as a substitute for coworkers, influencing social pressure through the embarrassment of underperformance or the display of selfish behavior. In our paper, the implementation of AI differs considerably, as it assumes the role of a quality control supervisor monitoring workers. While embarrassment could also play a role in our setting (as it likely does in most contexts involving human interaction), our focus shifts away from addressing selfish behavior and instead explores other considerations, such as when workers are more likely to seek external help. Together, these two papers and ours share the vision that effective automation must account for behavioral forces in its assessment.[8]

[7] March (2021) reviews 162 experimental studies providing evidence of this.

[8] Camerer (2019) stresses the importance of integrating behavioral economics insights to improve human-AI interactions in firms and institutions.

Lastly, we contribute to the experimental economics toolkit by introducing an incentivized method to elicit AI utilization and test it alongside several other collected measures of external assistance. Gen AI provides the opportunity to improve many areas of experimental practices (see Korinek 2023; Charness, Jabarian, and List 2023; S. Zhang, Xu, and Alvero 2025) but simultaneously poses new challenges, as participants can get help or even delegate their experimental duties. Using a text summarization task on Prolific, Veselovsky, Ribeiro, Cozzolino, Gordon, Rothschild, and West (2025) estimated that approximately 30% of participants utilized LLMs. Their findings demonstrate that restricting copy-pasting can reduce LLMs usage by half. Combining our elicited measures of ChatGPT usage with a text-pasting detector, we show that human assessment is another effective way to mitigate LLMs utilization. Our findings are consistent with prior research suggesting that the presence of humans can effectively deter dishonest behavior (Cohn, Gesche, and Maréchal 2022).

2 The Experiment

2.1 Implementation

We conducted our pre-registered online experiment[9] in November 2024 using a webpage built with Flask, a Python-based web framework, and hosted on the *PythonAnywhere* platform. The experiment utilized OpenAI's API, and we followed many current best practices in LLM research (Filippas, Horton, and B. S. Manning 2025). For API integration, we used OpenAI's official Python client library. Each API call was a standalone request, ensuring no conversation history or context was maintained between evaluations and that every request contained the correct system prompt, image, and caption. We recruited 208 participants[10] from Prolific's participant pool, selecting only those located in the United States to ensure a certain level of English proficiency for the writing task.

After consenting to participate in the experiment, participants were immediately presented with general instructions. They were informed that they would need to complete 20 rounds, with each round requiring them to describe an image in 150 to 400 characters. Four example images were provided, along with an explanation that their descriptions would be evaluated based on three criteria: Interesting, Effortful, and Accurate. Participants were also told they could earn a "success" bonus of $1 for each description that ranked in the top 30% among descriptions submitted by other Prolific participants for the same image. They were then informed that the "success" bonuses would be determined by either [ChatGPT/a graduate student], but each participant was only made aware of one evaluator, maintaining

[9] The preregistration plan is available at https://aspredicted.org/qc3fn5.pdf

[10] 103 and 105 participants in the human and ChatGPT treatments, respectively.

the between-subjects design of the experiment. Finally, participants completed a practice round, which they knew would not count toward payment but served to familiarize them with the task and interface. Screenshots of the experiment instructions and interface can be found in Appendix B.1.

After reviewing the instructions and completing a practice round, participants were tasked with writing captions for 20 images depicting daily life scenes.[11] The interface included a character counter and a persistent reminder at the bottom indicating who their evaluator was and the criteria under which they were being evaluated. Upon completing the 20 rounds of captioning, participants were invited to take an optional survey about their experience during the task. Most importantly, they were asked whether they had used a tool like ChatGPT to assist them. Initially, this question was posed without incentives. Subsequently, we implemented an incentivized method, offering an additional $1 if their response aligned with the prediction of an AI detection software.

Participants received an $8 show-up payment and had the opportunity to earn up to an additional $20 based on their performance. Captions were graded by their assigned evaluator, and if a caption ranked in the top 30% for a specific image, the participant received a $1 bonus.[12] The experiment lasted an average of 75 minutes, and by design, the average payoff was $14, which is well above the suggested rate for a study of this duration on Prolific.

2.2 Design Choices

In this section, we discuss several of the main choices we faced while designing the experiment.

Selection of the 3 evaluation criteria: Interesting, effortful, and accurate. We believe that selecting multiple criteria helps prevent participants from attempting to game the task and ensures that captions satisfy the criteria while still effectively describing the image. When selecting a group of three criteria that work well together, we chose to leverage some of the new capabilities of LLMs with subjective criteria (*interesting* and *effortful*), while still maintaining objective measures, such as *accuracy*. In our pilot, we randomized whether subjects saw three objective or three subjective criteria. The treatment effects from assigning different evaluators were fairly similar,[13] which led us to combine objective and subjective criteria into a single group and focus solely on evaluator variation.

Using ChatGPT model GPT-4o. It was the most advanced model of ChatGPT at the time of our experiment. In addition, ChatGPT was the most popular LLM, which increases

[11] The 20 images are included in Appendix B.3.

[12] Since our quality measures were in discrete numbers, multiple participants at the threshold received the same grade. To ensure fairness, we randomized bonus awards at the threshold to include 30% of participants.

[13] See C.1 for reference.

the likelihood of familiarity. The model we use includes the ability to interpret images, a feature introduced shortly before the experiment was conducted and that is essential for AI assessment of this task.

Prompt. During the testing process, we asked ChatGPT to elaborate on the grades it awarded. We discovered that when ChatGPT was prompted to grade a given caption across all three criteria simultaneously, it sometimes took shortcuts, assigning identical grades to multiple criteria and providing the same justification. To address this, we implemented three separate prompts per caption, each for one specific criterion. Additionally, we noticed that higher temperature settings occasionally caused ChatGPT to deviate from the grading guidelines. To mitigate this, we set the temperature at 0.75. As shown in the prompt, we provided systematic and detailed instructions, clearly outlining the task, ChatGPT's role, the grading criteria, steps, and response format. The full prompt provided to both ChatGPT and human evaluators can be found in Appendix B.2.

Bonus incentive. We decided to pay bonuses in terms of relative performance within the same evaluator. By paying a bonus for each time a caption is ranked in the top 30% for a given image, we are able to rule out explanations related to beliefs about evaluator leniency, which we discuss more in detail in Section 4. Relative performance payments may raise concerns about participants trying to infer the performance of others. However, we believe that the benefits of eliminating certain mechanism channels far outweigh this concern, for which we lack evidence of it playing a major role in this experiment. The alternative design involved paying bonuses if the caption met a satisfaction threshold (e.g., 7 out of 10 or higher). This approach was used in the pilot, and our results appear to be robust under that design as well.

Not eliciting WTP for preferred evaluator. In our pilot, we elicited participants' willingness to pay (WTP) for their preferred evaluator and assigned either a human or ChatGPT evaluator partly based on these preferences. Our results, particularly regarding work quality, proved robust even when separating participants by their preferred evaluator (C.2). However, workplace environments typically involve being assigned an evaluation method regardless of individual preferences, so we decided to shift to a design with that feature. However, investigating preferences for a preferred evaluator is an interesting question for future work.

Management of text pasting. We chose not to restrict pasting because doing so could backfire by inadvertently priming participants to attempt gaming the task in this or other ways. However, we feel this is also an interesting future research direction.

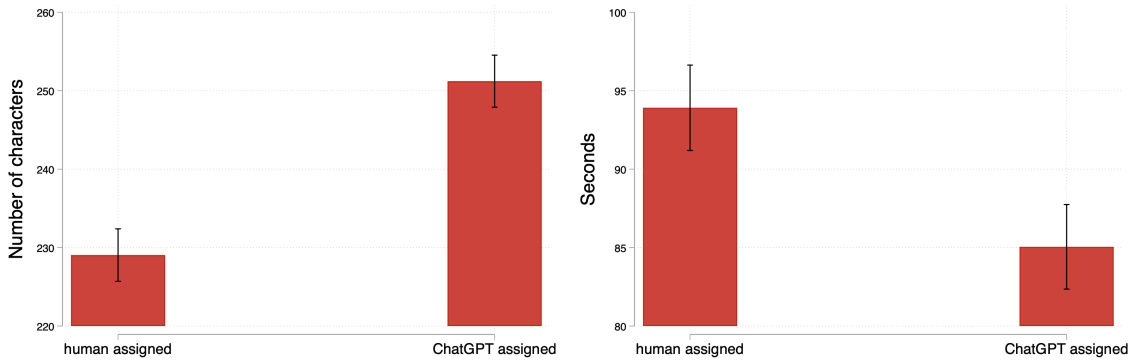
3 Results

3.1 Do Writing Habits Change with Evaluator?

We first examine whether workers' behavior changes when they are assigned different evaluators by analyzing our pre-registered measurements of writing habits: response length, response time, pasting, and erasures. These are all variables that can be objectively quantified (representing true indicators of ground truth). While responses containing pasted text differ on these metrics in several ways from those without pasted text, we will reserve that comparison for a later section. Here, we focus on aggregate outcomes.

3.1.1 Response Length and Time

A natural starting point for studying writing output is to examine the length of the response and the time spent writing. We observe that participants evaluated by a human tend to write shorter captions on average, yet take longer to do so. The average response length was 229 characters when evaluated by a human, compared to 251 characters when evaluated by ChatGPT (with a two-sided test of means p-value < 0.0001). Regarding response time, the average duration was 94 seconds when assigned a human evaluator and 85 seconds when assigned a ChatGPT evaluator (with a two-sided test of means p-value < 0.0001).[14] Figure 1 presents these findings visually, highlighting the strong statistical significance of the results.



(a) Response length.

(b) Response time.

Figure 1: Average caption response length and time by treatment.

[14]We winsorized the time variable at the 5th and 95th percentiles due to extreme outliers. The results remain consistent without this adjustment

3.1.2 Pasting and Erasures

Leveraging features of our custom-built webpage for this experiment, we collected two variables that provide insights into the writing process: text pasted and the number of characters erased. We created a dummy variable that equals one if, at any point during a caption, the participant pasted any characters. We will study in more detail what happens when participants paste text, but as mentioned before, we use pasting as our primary proxy of using external help. We find a stark difference in pasting habits between treatments: participants evaluated by ChatGPT pasted 20.1%, compared to 10.4% for those evaluated by a human (two-sided t-test p-value of < 0.0001). On the other hand, as shown in Figure A.1, there were no treatment effects on erasures.

3.2 Difference in Quality?

We documented significant treatment effects by evaluator in three of the four writing habit variables we studied. Given that the caption output and the process of writing them were shown to be different, there is already scope for questioning whether the quality of the captions is different as well. The key difference in measuring quality compared to the previous variables is the lack of ground truth, even for *accuracy*, which is arguably the most objective of our evaluation criteria. Although the lack of ground truth could be seen as a challenge, it adds realism since evaluating text, as many other work-related tasks, rarely involves a clear notion of ground truth.

We asked both ChatGPT and human evaluators to grade every caption, regardless of the treatment. Both evaluators were given the same set of instructions, and the evaluators separately rated a caption each criterion to minimize spillovers. Each criterion was given a score between 0 and 3, and the score on an image was summed to 9. While compensation bonuses were awarded based on the grade given by the assigned evaluator, we still collected evaluations from both types of evaluators for all captions. This approach generated two measures of quality: one from ChatGPT and another from the human graders.

We first analyze the scores awarded by ChatGPT. We find that participants in the human evaluator treatment (those assigned a human evaluator) received a higher average score from ChatGPT than those in the ChatGPT evaluator treatment (those assigned a ChatGPT evaluator).[15] Figure 2 shows that the average score on the human evaluator treatment was 6.32, while for the ChatGPT evaluator was 6.05 (two-sided t-test p-value of 0.0001). According to ChatGPT, participants working under the impression that a human would evaluate them ended up writing captions of higher quality than those thinking they will

[15]This result is robust to changes in the temperature parameter and whether we ask for scoring using all three criteria together or separately.

be evaluated by ChatGPT. This result is strong enough to overcome a potential advantage for participants in the ChatGPT treatment, who could tailor their captions to the actual evaluator.

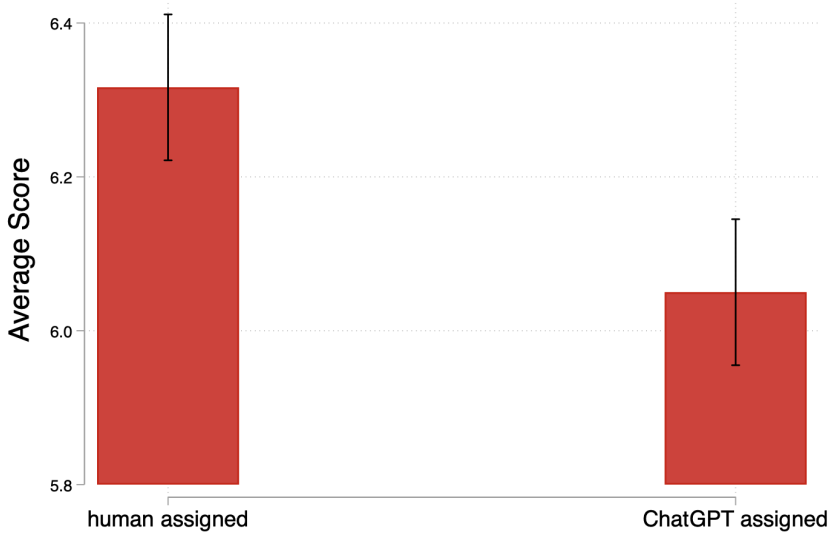
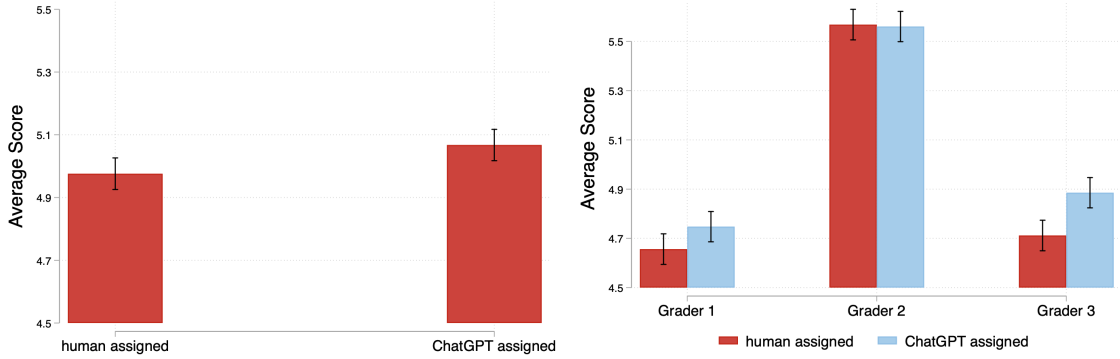


Figure 2: ChatGPT grades by treatment.

To generate human evaluations, we instructed three graduate students to grade all captions using the same prompt provided to ChatGPT. Figure 3 presents the grades, both aggregated and broken down by each grader. When averaging the grades of the three human graders, captions evaluated in the ChatGPT evaluation treatment received a statistically significantly higher average grade (5.07 vs. 4.97, two-sided t-test, p-value = 0.0046). This result is driven by two of the three graders.



(a) All graders aggregated.

(b) Separated by grader.

Figure 3: human evaluators' grades by treatment.

While, at first glance, the grades assigned by ChatGPT and the human graders appear to diverge, the following sections will explore these differences in greater detail.

3.3 Seeking External Help

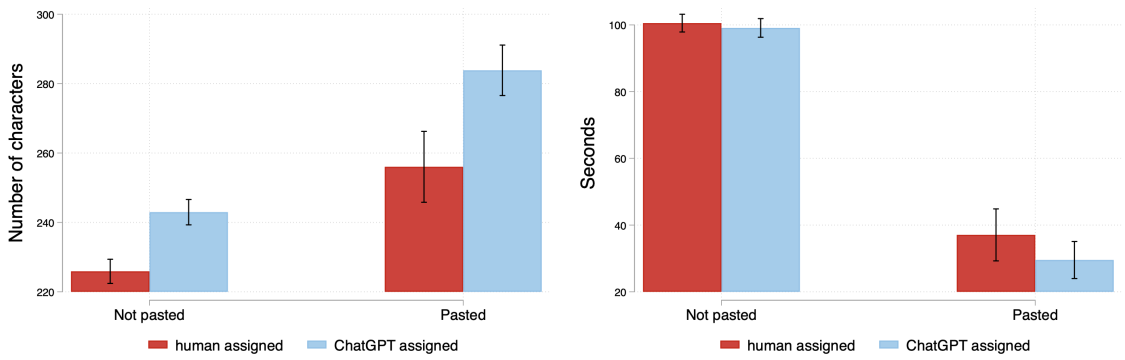
At the same time that AI has improved to the point where it can monitor work, it has also advanced in its ability to perform the tasks that are being monitored, opening the door for workers to seek assistance from AI in their daily duties. Previously, we reported that participants who were assigned AI assessment used external help twice as often as those assigned human assessment (using our pasting dummy variable as a proxy for external help). This finding aligns with the expectation that humans may negatively judge the use of AI assistance, such as ChatGPT, even when not explicitly instructed to do so. This aligns with concurrent work by Awuah, Krenk, and Yanagizawa-Drott (2024), which finds that NGO application evaluators penalize cover letters that appear to have been written with ChatGPT's assistance, even without explicit instructions to do so.

In this section, we examine two aspects of seeking help. First, we analyze whether the observed effects on writing habits and quality can be attributed to the use of external help. Second, we compare multiple measures of external help usage that we collected and present a methodological contribution: a new elicitation method that can identify the use of external help in contexts or tasks where the pasting variable is less effective or unavailable.

3.3.1 Are Our Results Driven by External Tool Use?

By analyzing writing habits and quality separately for captions with pasted text and those without, we find substantial differences in captions depending on whether text was pasted. However, these differences alone cannot explain most of the previously observed treatment effects.

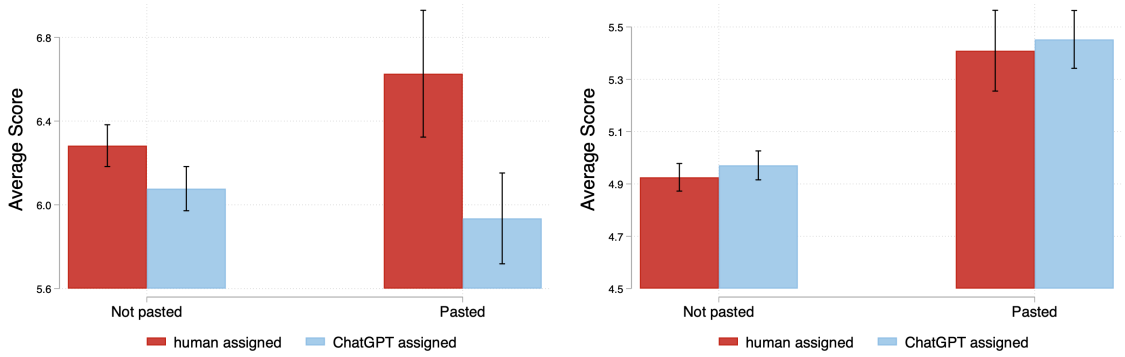
Regarding response length, we observe that pasted captions tend to be longer. However, participants in the human evaluation treatment write significantly shorter captions, on average, for both pasted and non-pasted captions. In terms of response time, non-pasted captions take nearly three times longer. When controlling for pasting, there are no longer any treatment effects on response time. Figure 4 documents these results.



(a) Response length by pasting. (b) Response time by pasting.

Figure 4: Average caption response length and time by treatment and pasting.

When revisiting the quality of the captions, Figure 5 confirms that ChatGPT consistently awards higher grades to participants in the human treatment regardless of pasting occurring or not. In contrast, human graders assign significantly higher grades to captions containing pasted text, which explains why participants in the ChatGPT treatment received higher average grades, as they pasted text twice as often. However, when analyzing captions with and without pasted text separately, human graders do not assign statistically different grades between the two treatments.



(a) ChatGPT grades by pasting (b) Human grades by pasting

Figure 5: ChatGPT and human grades by treatment and pasting.

3.3.2 Measuring and Eliciting External Help Usage

In this paper, we have used text pasting as the main proxy for identifying use of external assistance. However, text pasting may not always be a feasible approach (e.g., for non-written tasks or situations where implementation is impractical). We introduce a new elicitation

method and test its validity to expand the experimental toolkit for identifying the use of LLMs in experiments.

At the end of the experiment, participants are asked to disclose whether they used a tool like ChatGPT to assist in writing any of their descriptions. Then, they were offered a $1 bonus if they correctly predict whether a ChatGPT detector software would classify any of their descriptions as AI-generated. We show that designing this question with incentives (as in the latter case) yields more promising results.

In the non-incentivized question, 13 out of 208 participants reported using external help from a tool like ChatGPT. For the subsequent incentivized question, the number of participants admitting their captions might be deemed as AI-generated increased to 29. Interestingly, 3 out of the 13 participants who initially answered "yes" to the non-incentivized question switched to "no" when the incentivized question incorporated a detection tool. Of those three participants, two did not paste text in any round. This suggests that while they wrote the text themselves, they might have used ChatGPT for other parts of the process. For instance, one participant mentioned in the open-ended survey question: "There was one picture where a woman was photographing a cupcake. She was using a white umbrella, so I asked ChatGPT what the umbrella was used for." This example highlights how the range of utilization can vary from full delegation to simply gathering information to augment work. Figure 6 illustrates the distribution of participants based on the number of rounds in which they used pasting, as well as the number of participants who disclosed using help for the elicitation questions. The results show that the incentivized method is particularly effective in identifying heavy pasters (18–20 rounds), capturing half of those participants (13 out of 26). Additionally, the incentivized question demonstrates added value beyond the pasting detector, identifying 13 participants who did not paste in any round.[16]

[16]This number increases to 15 if we include the two participants who only responded "yes" in the non-incentivized question.

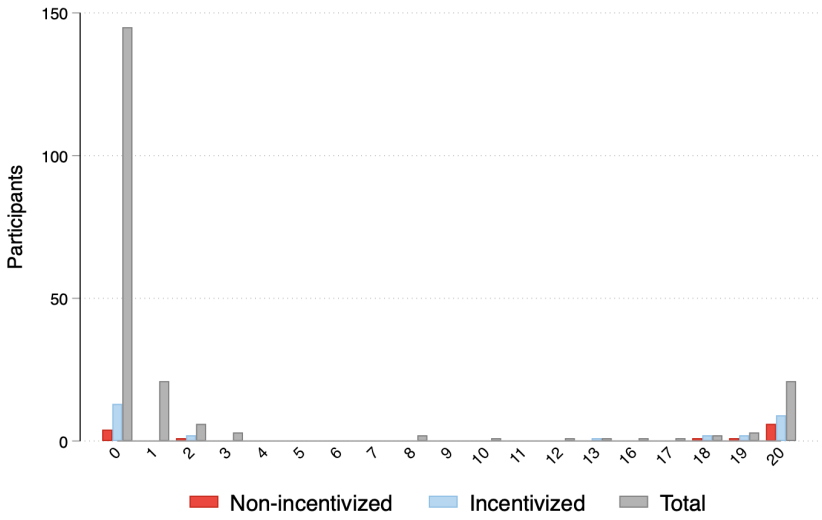


Figure 6: Distribution of participants by the number of rounds they pasted text. The x-axis represents the number of rounds a participant pasted text, with red bars showing participants who disclosed using help in the non-incentivized question, blue bars showing participants who disclosed using help in the incentivized question, and grey bars indicating the total number of participants for each category.

4 Discussion

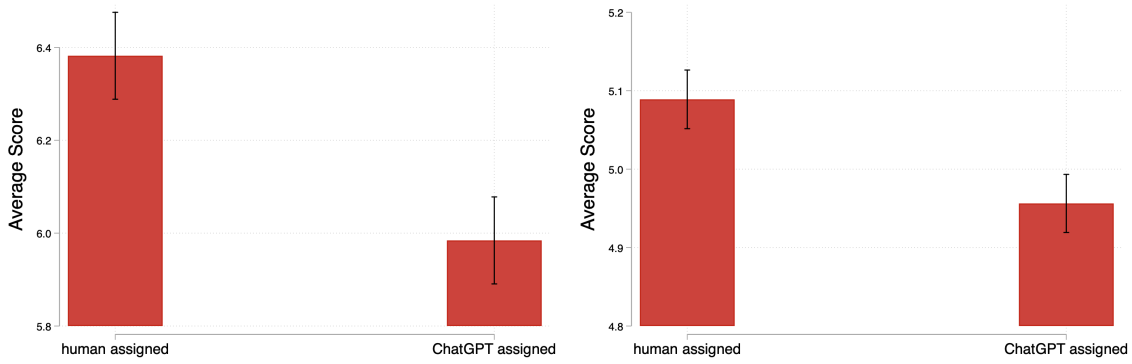
4.1 Why Does the Treatment Effect on Quality Emerge Only in ChatGPT Grading?

After accounting for pasting, only ChatGPT identified significant differences in our quality measure between treatments. This raises an interesting question: why don't human graders award higher grades to participants in the human evaluator treatment, as ChatGPT does? Since the quality of captions lacks an objective ground truth, the two evaluation sources do not need to agree. However, one possible explanation for the initial discrepancy could be a misperception about the length of the text. Not only did participants in the human evaluator treatment write shorter captions on average, there is also a first-order stochastic dominance relationship as seen in Figure A.2.

We hypothesize that while writing the captions, participants might have assumed that human graders would prefer more concise (and probably polished) descriptions. However, our findings indicate that human graders awarded higher grades to longer captions. ChatGPT also assigned higher grades to longer captions, but the effect of length on grades was

more pronounced for human graders. When regressing grades on treatment and caption length, an additional 100 characters increased human-assigned grades by an estimated 1 point, whereas the effect was 0.5 points for ChatGPT. This outcome is not surprising, as longer captions often exhibit greater effort and capture more details, likely resulting in more accurate descriptions. Interestingly, human evaluators placed a greater emphasis on length compared to ChatGPT.

Figure 7 illustrates that, when controlling for output length, both ChatGPT and human evaluators rate the work in the human assessment treatment as significantly higher in quality. Notably, this result holds even without distinguishing between pasted and non-pasted text, a factor that would likely amplify the treatment effect.



(a) ChatGPT grades

(b) Human grades

Figure 7: ChatGPT and human grades by treatment when controlling for text length.

4.2 Interpretation of Treatment Effects

After accounting for pasting habits, we documented that the length of captions is different between treatments. Participants in the human assessment treatment write shorter captions, which suggests they may be more careful with what they write. A finding that supports this is that when controlling for text length, both ChatGPT and humans rate as higher quality the captions from those participants writing for a human grader.

ChatGPT consistently rates captions from the human assessment treatment as higher quality. For graduate students, the results follow the same direction when controlling for text length, or is not significant otherwise (when we separate by pasting). What factors might explain why participants assigned human assessment appear to perform better? To rule out potential explanations, we fixed the incentives across treatments, awarding the same number of bonuses. This ensures that relative beliefs about evaluator leniency cannot account for

the observed differences.[17]

We explored whether there is a different feeling when a human validates your work compared to AI, which motivates participants assigned human assessment to elevate the quality of their work. We asked participants how happy they felt when their assigned evaluator gave them a high score, regardless of the payment they received (on a scale from 1 to 5). Participants assigned to a human evaluator reported a statistically significantly higher score for this question (4.41 vs. 4.09; two-sided t-test, p-value = 0.0055). This suggests that non-monetary incentives may explain why participants assigned human assessment outperformed those assigned AI assessment.

5 Conclusion

Our study provides empirical evidence on how individuals alter their behavior when their work is evaluated by AI rather than by humans. Using image captioning, a common workforce task, we find that algorithmic evaluation increases output quantity. However, when controlling for output levels, caption quality in the AI treatment is lower, regardless of whether it is assessed by humans or AI. Additionally, AI assessment increases the likelihood of workers seeking external assistance, though this alone does not fully explain the decline in quality. Our results suggest that the shift toward AI-based evaluation systems may induce behavioral responses that firms and policymakers should consider when implementing such assessment mechanisms.

These findings have important implications for the future of algorithms in organizations. While AI evaluation is cost-effective and scalable, its impact on work quality and reliance on external tools presents potential trade-offs. In contexts where quality is paramount, such as creative and knowledge-based tasks, firms may need to supplement AI assessment with human review to mitigate unintended consequences.

Future research should explore how different types of AI assessment affect worker behavior across various tasks and industries. Additionally, understanding the long-term effects of AI evaluation on worker effort, learning, and motivation remains an open question.

[17] If bonuses were awarded based on whether an evaluator deemed a caption satisfactory, beliefs about the minimum effort required to achieve a satisfactory grade could vary between evaluators, potentially influencing participants' effort levels.

References

Acemoglu, Daron and Pascual Restrepo (2019). "Automation and New Tasks: How Technology Displaces and Reinstates Labor". *Journal of Economic Perspectives* 33.2, pp. 3–30. DOI: 10.1257/jep.33.2.3.

Almog, David, Romain Gauriot, Lionel Page, and Daniel Martin (2024). "Human Responses to AI Oversight: Evidence from Centre Court". *Proceedings of the ACM on Human-Computer Interaction*. ACM working paper; DOI : 10.1145/3670865.3673481.

Autor, David, Caroline Chin, Anna Salomons, and Bryan Seegmiller (2024). "New Frontiers: The Origins and Content of New Work, 1940–2018". *The Quarterly Journal of Economics* 139.3, pp. 1399–1465. DOI: 10.1093/qje/qjae008.

Awuah, Kobbina, Ursa Krenk, and David Yanagizawa-Drott (2024). "Learning About AI". Working paper.

Bick, Alexander, Adam Blandin, and David J. Deming (2024). "The Rapid Adoption of Generative AI". *NBER Working Paper* 32966. Revised February 2025. DOI: 10.3386/w32966.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond (2025). "Generative AI at Work". *The Quarterly Journal of Economics* 140.2. Earlier version circulated as NBER Working Paper 31161 (2023), pp. 889–942. DOI: 10.1086/705716.

Camerer, Colin F. (2019). "Artificial Intelligence and Behavioral Economics". *In The Economics of Artificial Intelligence: An Agenda*, pp. 587–608.

Charness, Gary, Brian Jabarian, and John A. List (2023). "Generation Next: Experimentation with AI". *NBER Working Paper* 31679. DOI: 10.3386/w31679.

Choi, Jonathan H. and Daniel Schwarcz (2025). "AI Assistance in Legal Analysis: An Empirical Study". *Journal of Legal Education* 73. Originally circulated as working paper (2023), p. 384.

Cohn, Alain, Tobias Gesche, and Michel André Maréchal (2022). "Honesty in the Digital Age". *Management Science* 68.2, pp. 827–845. DOI: 10.1287/mnsc.2021.3985.

Conlon, John J. (June 2025). "Memory Rehearsal and Belief Biases". Working paper, Carnegie Mellon University. First version: May 28, 2024. This version: June 28, 2025. URL: https://johnjconlon17.github.io/website/conlon_rehearsal_beliefs.pdf.

Corgnet, Brice, Roberto Hernán-González, and Ricardo Mateo (2023). "Peer effects in an automated world". *Labour Economics* 85, p. 102455. DOI: 10.1016/j.labeco.2023.102455.

Dickinson, David and Marie-Claire Villeval (2008). "Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories". *Games and Economic Behavior* 63.1, pp. 56–76. DOI: 10.1016/j.geb.2007.02.004.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock (2024). "GPTs are GPTs: Labor market impact potential of LLMs". *Science* 384.6702, pp. 1306–1308. DOI: 10.1126/science.adj0998.

Filippas, Apostolos, John J. Horton, and Benjamin S. Manning (2025). "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *Working paper*. Version dated February 12, 2025; earlier version appears as NBER Working Paper 31122 (2023) by John J. Horton.

Flodén, Jonas (2025). "Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT". *British Educational Research Journal* 51.1. Online first: 16 September 2024, pp. 201–224. DOI: 10.1002/berj.4069. URL: https://doi.org/10.1002/berj.4069.

Frey, Bruno (1993). "Does Monitoring Increase Work Effort? The Rivalry Between Trust and Loyalty". *Economic Inquiry* 31.4, pp. 663–670. DOI: 10.1111/j.1465-7295.1993.tb00897.x.

Goergen, Jonas, Emanuel de Bellis, and Anne-Kathrin Klesse (2025). "AI assessment changes human behavior". *Proceedings of the National Academy of Sciences of the United States of America* 122.25, e2425439122. DOI: 10.1073/pnas.2425439122.

Humlum, Anders and Emilie Vestergaard (2025). "The unequal adoption of ChatGPT exacerbates existing inequalities among workers". *Proceedings of the National Academy of Sciences* 122.1, e2414972121. DOI: 10.1073/pnas.2414972121.

Ide, Enrique and Eduard Talamas (2025). "Artificial Intelligence in the Knowledge Economy". *Journal of Political Economy*. Published online May 2025. DOI: 10.1086/737233.

Korinek, Anton (2023). "Generative AI for Economic Research: Use Cases and Implications for Economists". *Journal of Economic Literature* 61.4, pp. 1281–1317. DOI: 10.1257/jel.20231736.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2024). "The Unreasonable Effectiveness of Algorithms". *NBER Working Paper* 32125. DOI: 10.3386/w32125.

March, Christoph (2021). "Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players". *Journal of Economic Psychology* 87, p. 102426. DOI: 10.1016/j.joep.2021.102426.

Mok, Lillio, Sasha Nanda, and Ashton Anderson (2023). "People perceive algorithmic assessments as less fair and trustworthy than identical human assessments". *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2, pp. 1–26.

Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor (2002). "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment". *American Economic Review* 92.4, pp. 850–873. DOI: 10.1257/00028280260344498.

Noy, Shaked and Whitney Zhang (2023). "Experimental evidence on the productivity effects of generative artificial intelligence". *Science* 381.6654, pp. 187–192. DOI: 10.1126/science.adh2586.

Otis, Nicholas, Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning (2025). "The Uneven Impact of Generative AI on Entrepreneurial Performance". *OSF preprint*. Preprint version posted October 2025; field experiment with Kenyan entrepreneurs.

Palan, Stefan and Christian Schitter (2018). "Prolific.ac—A subject pool for online experiments". *Journal of Behavioral and Experimental Finance* 17, pp. 22–27. DOI: 10.1016/j.jbef.2017.12.004.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer (2023). "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot". *arXiv preprint*. DOI: 10.48550/arXiv.2302.06590. eprint: arXiv:2302.06590.

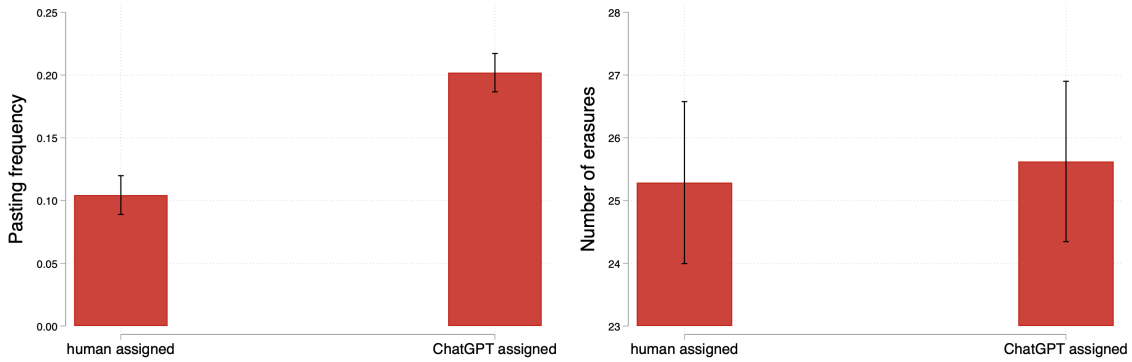
Sarhan, Habiba and Simon Hegelich (2023). "Understanding and evaluating harms of AI-generated image captions in political images". *Frontiers in Political Science* 5, p. 1245684. DOI: 10.3389/fpos.2023.1245684.

Veselovsky, Veniamin, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West (2025). "Prevalence and prevention of large language model use in crowd work". *Communications of the ACM* 68.3. Earlier version on arXiv:2310.15683, pp. 42–47.

Zhang, Da Wei, Melissa Boey, Yan Yu Tan, and Alexis Hoh Sheng Jia (Dec. 2024). "Evaluating large language models for criterion-based grading from agreement to consistency". *npj Science of Learning* 9.1, p. 79. DOI: 10.1038/s41539-024-00291-1.

Zhang, Simone, Janet Xu, and AJ Alvero (2025). "Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization". *Sociological Methods & Research* 54.3, pp. 1197–1242. DOI: 10.1177/00491241251327130.

A Additional Figures



(a) Pasted text frequency. (b) Erasures.

Figure A.1: Frequency of captions containing pasted text and the average number of erasures per caption, separated by treatment.

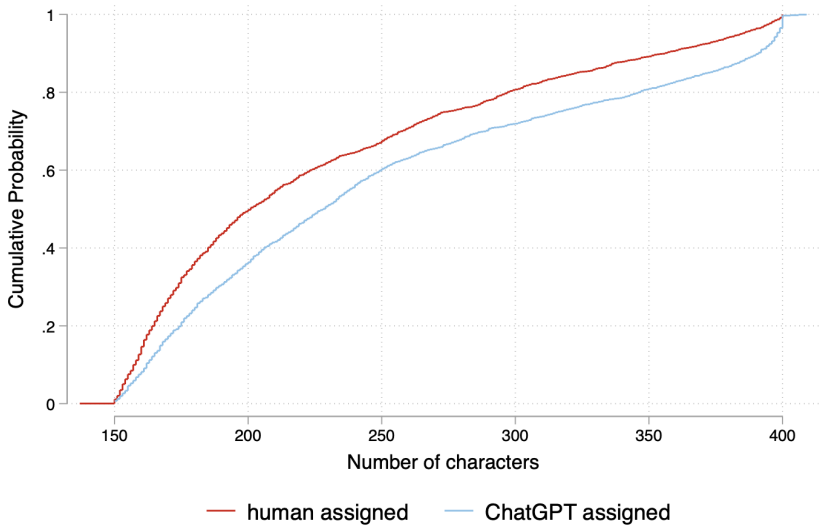
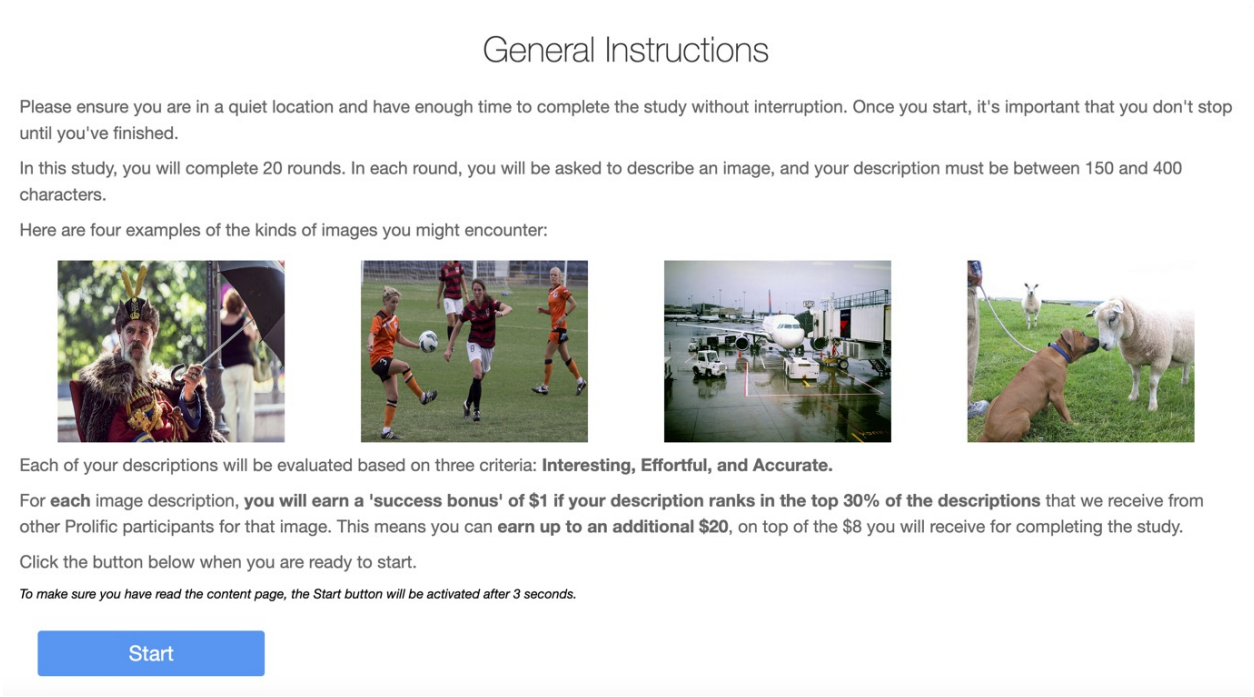
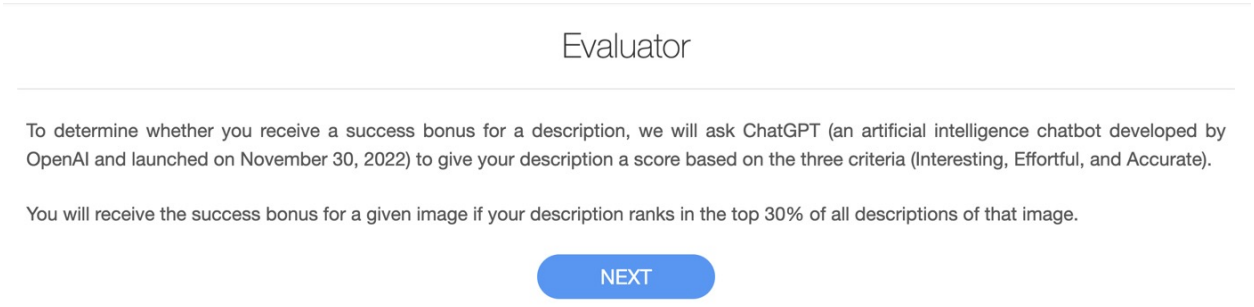


Figure A.2: Caption length empirical CDF by treatment.

B Experimental Design

B.1 Instructions and Interface





Round: 6 out of 20

How would you describe this image?

A muscular bulldog balances confidently on a bright blue surfboard, its sturdy frame adorned with a red lifesaver vest. The dog's tongue flops joyfully as it rides the crest of a shimmering wave, water spraying dramatically around it.

Character Count: 234

As a reminder, your descriptions will be evaluated by ChatGPT based on three criteria: Interesting, Effortful, and Accurate.

SUBMIT

We do not have a problem with you using a tool like ChatGPT, and the answer to these questions will not affect your compensation at all.

1. Did you use a tool like ChatGPT to help write any of your descriptions?

○ Yes ○ No

2. If you did use a tool like ChatGPT, please describe how you used it:

Next

You will earn an additional $1 bonus if you can correctly guess whether Originality.ai (a ChatGPT detector software) will say that any of your descriptions was written by ChatGPT. Again, we do not have a problem with you using a tool like ChatGPT, and the answer to this question will not decrease your compensation at all. In this case, you could even earn extra money.

Will Originality.ai (a ChatGPT detector software) say that any of your descriptions was written by ChatGPT?

○ Yes ○ No

Next

B.2 Evaluation Instructions

We provided the same instructions to both ChatGPT and the graduate students responsible for evaluating. For ChatGPT, we executed the following three prompts separately for each caption.

Interesting
You will be shown an image along with a description and asked to grade the description based on how interesting it is.
Task: Evaluate how interesting the description is, assigning a score from 0 to 3.
Grading Criteria:

- 0: Not engaging or interesting.
- 1-2: Moderately engaging, somewhat interesting.
- 3: Very engaging and creative.

Steps:

1. Read the Description: Carefully review the description provided.
2. Evaluate for Interest: Assess how engaging and creative the description feels.
3. Assign a Score: Rate the description based on the criteria above.

Output Format:
Provide a single score (0–3) that reflects how interesting you found the description.

Effortful
You will be shown an image along with a description and asked to grade the description based on how effortful it is.
Task: Evaluate how effortful the description is, assigning a score from 0 to 3.
Grading Criteria:

- 0: Little to no effort shown.
- 1-2: Shows moderate effort and thought.
- 3: Demonstrates high effort and thoughtfulness.

Steps:

1. Read the Description: Carefully review the description provided.
2. Evaluate for Effort: Assess how thoughtful the description feels.
3. Assign a Score: Rate the description based on the criteria above.

Output Format:
Provide a single score (0–3) that reflects how effortful you found the description.

Accurate

You will be shown an image along with a description and asked to grade the description based on how accurate it is.

Task: Evaluate how accurate the description is, assigning a score from 0 to 3.

Grading Criteria:

- 0: Inaccurate or irrelevant to the image.
- 1-2: Somewhat accurate but with minor issues.
- 3: Highly accurate, fully aligned with the image.

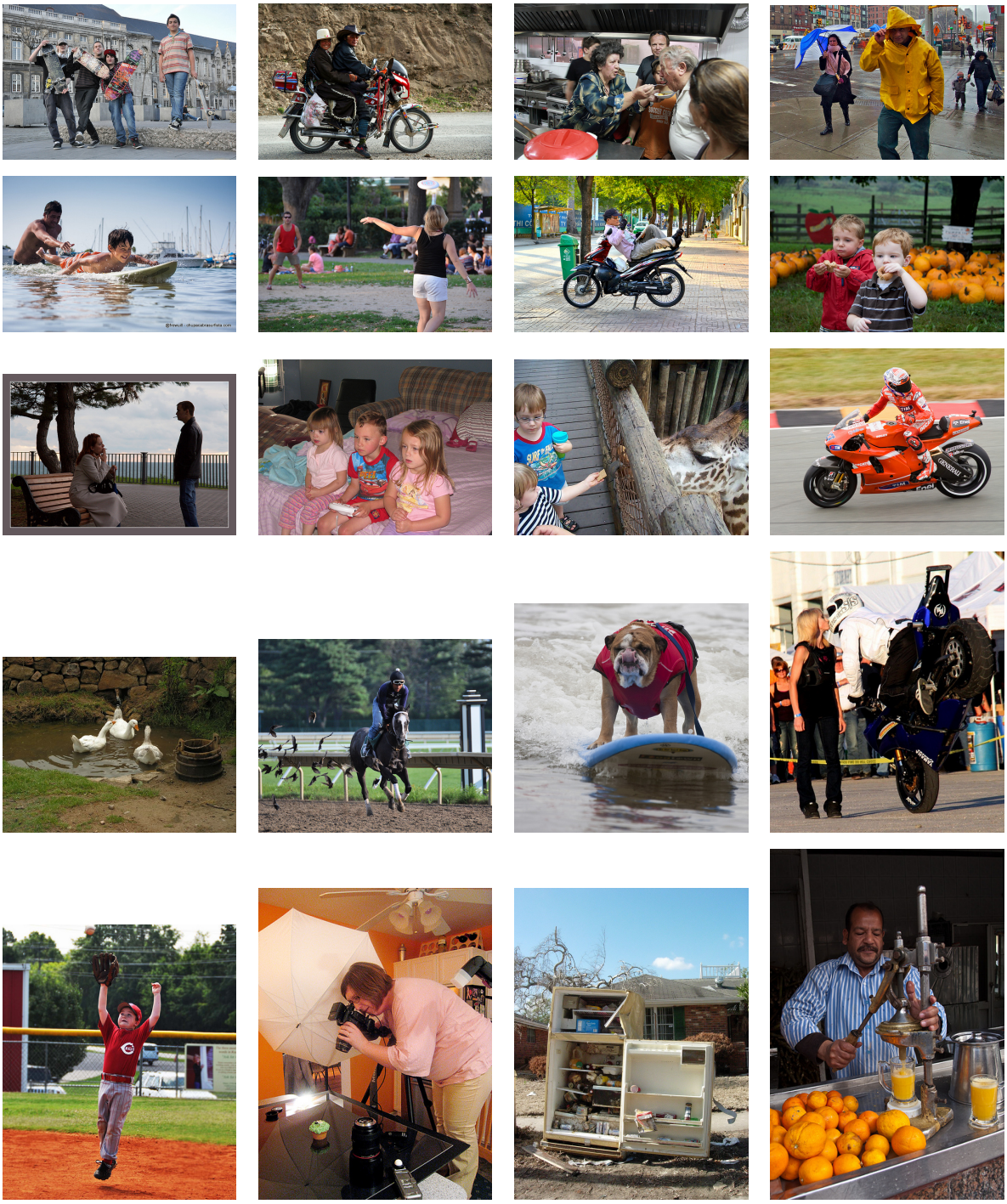
Steps:

1. Read the Description: Carefully review the description provided.
2. Evaluate for Accuracy: Assess how accurate the description feels.
3. Assign a Score: Rate the description based on the criteria above.

Output Format:

Provide a single score (0–3) that reflects how accurate you found the description.

B.3 Images



C Pilot Results

In the first pilot, we instructed ChatGPT to rate captions as either satisfactory or not satisfactory. Below, we report the results based on whether the criteria used were objective or subjective, and whether they were conditioned on the participant's original choice of evaluator.

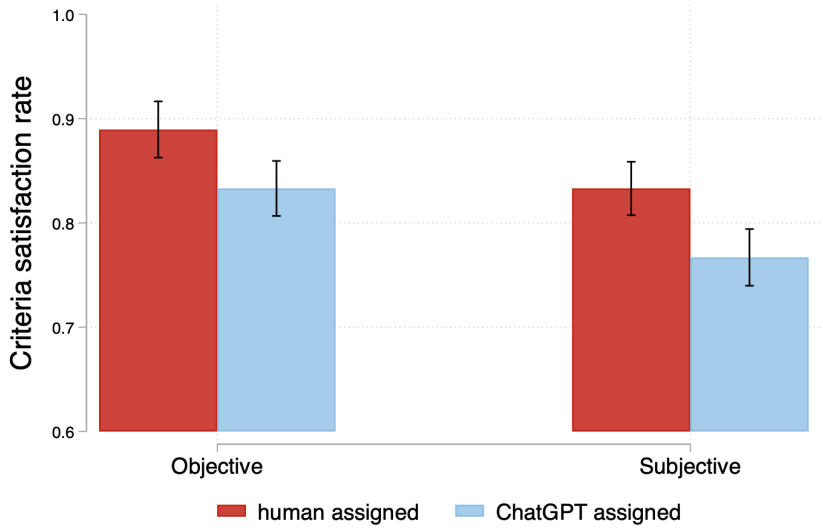


Figure C.1: ChatGPT grades by treatment and criteria group.

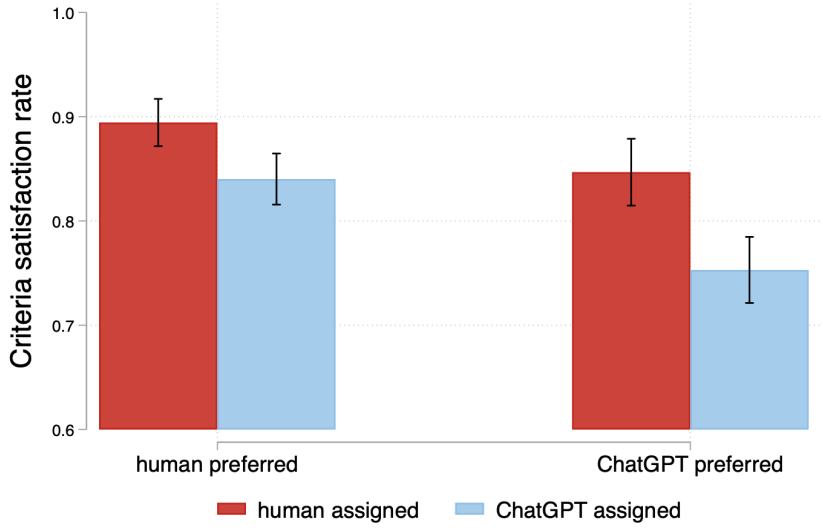


Figure C.2: ChatGPT grades by treatment and evaluator preference.