

Collaborating with AI Agents: A Field Experiment on Teamwork, Productivity, and Performance

Harang Ju

Johns Hopkins Carey Business School
harang@jhu.edu

Sinan Aral

MIT Sloan School of Management
sinan@mit.edu

August 19, 2025

Abstract

To uncover how AI agents change productivity, performance, and work processes, we introduce Pairit—an experimentation platform enabling humans and AI agents to collaborate in integrative workspaces. In a large-scale marketing experiment on the platform, 2310 participants were randomly assigned to human-human and human-AI teams. The teams exchanged 183,691 messages and created 63,656 image edits, 1,960,095 ad copy edits, and 10,375 AI-generated images while producing 11,138 ads for a large think tank. Analysis of fine-grained communication, collaboration, and workflow logs revealed that collaborating with AI agents increased communication by 63% and allowed humans to engage in 71% less direct text editing. While human-AI teams engaged in 18% more process and content communication, human-human teams engaged in 29% more social and emotional communication. Humans in human-AI teams experienced 73% greater productivity per worker and produced higher-quality ad copy, while human-human teams produced higher-quality images, suggesting AI agents require fine-tuning for multimodal workflows. Field tests of the ad campaigns accumulated $\sim 5M$ ad impressions and revealed that ads with higher image quality (produced by human-human collaborations) and higher text quality (produced by human-AI collaborations) performed significantly better on click-through rates, view through rates, and cost per click metrics. Together, these results suggest that human collaboration with AI agents significantly reshapes communication patterns and work processes and increases productivity, while improving some dimensions of output quality and deteriorating others. We hope the release of the extensible Pairit platform will accelerate RCTs of human-AI collaboration across a variety of work tasks and contexts.

1 Introduction

Artificial intelligence (AI) tools have garnered attention for their potential to improve productivity and performance (Eloundou et al., 2024; Bick et al., 2024). For example, large language models (LLMs) decreased the average time taken for mid-level professional writing tasks by 40% and increased output quality by 18% (Noy and Zhang, 2023). For job seekers, AI assistance with resumes increased job hiring by an average of 8% (Wiles et al., 2023), and for customer support workers, AI assistance increased productivity by an average of 14% (Brynjolfsson et al., 2023). Moreover, productivity gains were greater for lower-skilled workers (Noy and Zhang, 2023; Brynjolfsson et al., 2023; Choi and Schwarcz, 2023) and varied across task domains (Dell’Acqua et al., 2023). Evidence from an online labor market suggests there has already been a reduction in demand for freelance knowledge work with the advent of generative pre-trained transformer (GPT) models (Hui et al., 2023). In a study reviewing over 106 papers, Vaccaro et al. (2024) showed that human-AI groups outperformed humans alone in 85% of the studies. Our work aims to address two critical gaps in this burgeoning literature on AI’s impact on collaboration and productivity: Understanding 1) the role of multimodal AI agents (as distinct from LLM-enabled chatbots) and 2) the work process changes created by human-AI agent collaboration.

While studies like Liu et al. (2023) explore levels of proactivity in AI agents, they typically do so without randomized controlled trials (RCTs) measuring productivity effects. The studies that use RCTs to estimate productivity effects tend to randomize access to LLM chatbots (e.g. Dell’Acqua et al. (2023) and Chen and Chan (2024)), which are not typically multimodal, do not include context, do not allow the chatbots to take independent actions or use APIs to call outside of the work environment, and do not provide a collaborative workspace where machines and humans can jointly manipulate output artifacts in real time. These innovations are meaningful because AI agents today possess all these features, yet the existing scientific literature studies none of them.

Furthermore, we currently lack fine-grained task level insight into how human-AI collaboration changes work processes and communication patterns and how these changes affect productivity and performance. The vast majority of existing research focuses on the productivity effects of GPT

chatbots on individual workers (Noy and Zhang, 2023; Dell’Acqua et al., 2023). Others investigate how AI changes people’s perceptions, beliefs, and behaviors (Tey et al., 2024; Costello et al., 2024). However, it is unclear how these interactions evolve in real-time collaborations, especially in environments where AI agents can take autonomous actions, adapt dynamically to human inputs, and participate in tasks requiring creativity and coordination. Current off-the-shelf experimental platforms and studies do not provide collaborative workspaces where researchers can precisely record and measure the collaboration itself: *e.g.*, transcripts of messages between machines and humans, logs of edits to output artifacts, and API (application programming interface) calls to outside agents. In contrast, current AI applications, such as [Notion AI](#), [v0.dev](#), [OpenAI Canvas](#), [Cursor](#), and [GitHub Copilot](#), already integrate AI agents into such collaborative workspaces and interfaces.

To address these gaps, we developed Pairit, a novel experimental platform designed to study human-AI collaboration in real-world, extensible tasks. Pairit introduces several key innovations. It enables real-time collaboration between humans and AI agents, allowing participants to manipulate text, images, and workflows collaboratively in chat-enabled workspaces that mirror existing online AI collaboration work processes. The platform supports randomized pairings of humans and AI (*i.e.*, human-human or human-AI teams) and allows for randomization of prompts and model fine-tuning. Critically, the AI can perform the complete set of equivalent actions that humans can perform in the collaborative workspace. These include sending chat messages, writing ad copy, editing ad copy, writing calls to action, editing calls to action, scrolling through images, editing images, selecting images, and generating new images using an external call to Dall-E 3. Moreover, Pairit captures every time-stamped keystroke, message, edit, swipe, scroll, selection, API call, and intermediate output, providing a rich dataset that allows for the detailed reconstruction of collaboration workflows. This represents a fundamental departure from the existing literature, which enables RCT-based evaluation of the outputs and productivity implications of human collaboration with LLM-based chatbots and co-pilots, but does not enable randomized experiments analyzing the task-level productivity and work process changes created in human collaborations with fully

functioning, multimodal AI agents.

To test the Pairit platform and to study the work process, productivity, and performance implications of human collaboration with fully functional AI agents, we conducted a large-scale randomized study of the effects of human-AI collaboration on advertising design and creation—a task requiring creativity, iteration, and precision. A total of 2,310 participants, representative of the U.S. population, were recruited through Prolific and randomly paired into human-human or human-AI teams. Teams worked collaboratively to create marketing campaigns for a think tank’s year-end annual report, including generating and selecting ad images and writing ad copy and calls to action. This process was fully recorded, resulting in a dataset including 11,138 ads, 183,691 messages, 1,960,095 text edits, 63,656 image edits, and 10,375 AI-generated images, offering an unprecedented level of detail with which to understand work processes, teamwork dynamics, communication, productivity and output quality.

Once the lab portion of the experiment was completed, we conducted a field experiment on the ads produced by human-human and human-AI teams. We obtained human and AI quality ratings of the ads, including the quality of the ad copy and images, and the human- and AI-evaluated likelihood of consumer engagement with the ads (measured by click-through rates). We then ran the ads a real online display ad ecosystem, generating over 4.9 million impressions, and evaluated click-through rate, cost-per-click, view-through rate and view through duration metrics on the annual report using the platform’s ads API and DocSend’s view metrics, which allow us to record how much of the report consumers read, page by page, after clicking through on the ads.

In mining the rich collaboration data to understand how AI changes team communication, we found individuals in human-AI teams sent 63% more messages than those in human-human teams and that results were consistent at the team level. Furthermore, both the humans and the AI in human-AI teams sent 18% more *content-* and *process*-oriented messages than those in human-human teams, especially messages containing suggestions, instructions, prioritization, and planning. Conversely, humans in human-human teams sent 29% more *social* and *emotional* messages, including more messages that expressed rapport building, self-assessment, and concern, than humans in human-AI

teams. These results demonstrate that human-human teams spent considerable time and effort on social and emotional calibration and regulation, while human-AI teams were task- and process-oriented.

We also tracked work process changes in the collaborative workspace where humans and agents could both edit the ad copy as well as generate images using external APIs (*i.e.*, Dall-E 3). We found that human-AI teams made 71% fewer edits to the copy due to AI's proficiency in writing high-quality ad copy. In contrast, the AI was worse at predicting image quality, thus disadvantaging human-AI teams in the visual dimensions of ad creation and leading to human-AI teams creating lower-rated images. These differences in collaboration outcomes indicate a shift in workload that allowed human participants in human-AI teams to engage more in content generation with less social coordination, like rapport building, than humans in human-human teams.

These communication, collaboration, and work process changes then translated into measurable differences in productivity and performance. Humans in human-AI teams experienced 73% greater productivity per worker and human evaluations scored the ads created by human-AI teams higher on text quality but lower on image quality. AI evaluations scored the ads created by human-AI teams higher on text quality and equally on image quality. Field tests of the ad campaigns revealed that ads with higher image quality (produced more often by human collaborations) and higher text quality (produced more often by human-AI collaborations) performed significantly better on click-through rates, view through duration and cost per click metrics. As human-AI teams produced higher quality text and lower quality images, human-AI ads performed similarly to those created by human-human teams overall. Together, these results suggest that human collaboration with AI agents significantly reshapes communication patterns and work processes and increases productivity, while improving some dimensions of output quality and deteriorating others.

Although prior studies have shown that AI tools can improve productivity and reduce task completion times (Noy and Zhang, 2023; Brynjolfsson et al., 2023), AI is often treated as a passive tool rather than as an active collaborator. As AI agents become integral to modern workflows, researchers are beginning to explore their role as work collaborators, rather than mere tools,

emphasizing the importance of trust, transparency, and integration in human-AI partnerships (Makarius et al., 2020; Anthony et al., 2023; Collins et al., 2024). Our work contributes to this emerging literature by presenting the first task-level randomized experiment measuring the work process, productivity and performance implications of human-AI collaboration with fully functional, multimodal AI agents. We hope the release of the extensible *Pairit* platform will accelerate the deployment of RCTs of human-AI collaboration across a variety of work tasks and contexts in future research.

2 Methodology

Our study was preregistered and approved by the MIT Committee on the Use of Humans as Experimental Subjects.¹ Any non-pre-registered analyses are labeled *post hoc*.

2.1 The *Pairit* platform

Once a participant enters the *Pairit* (Figure 1) platform, they are randomized into a queue for either the human-human or the human-AI condition. In the human-human condition, the participant collaborates with another human participant. In the human-AI condition, the participant collaborates with an AI agent. In addition to team-level randomization, the AI models are also configured to accept prompt and fine-tuning randomization. *Pairit*'s left panel is the task panel in which participants can participate in the task work process (in the context of this experiment to create, edit, and submit ads). The platform includes a carousel of images participants can choose from, and the participants (as well as the AI) can also access the Dall-E 3 application programming interface (API) to generate new images. The ad copy creation and editing panel includes a headline, primary text, and a description. All of the edits—including the image selection and image generation, as well as the ad copy edits—are synchronized in real-time across the participants, whether they are human or AI. On the right is the chat panel in which participants can chat in real time with either

¹See osf.io/jfzha and osf.io/95dhu.

another human participant or an AI agent. The participants can submit the ads and the interface will reset so that they can work on and then submit the next ad. To the best of our knowledge, Pairit is the first platform to facilitate real-time collaboration between human-human or human-AI teams with team- and model-level randomization and real-time chat and editing on synchronized text and image interfaces in the context of multimodal human-AI collaboration workspaces.

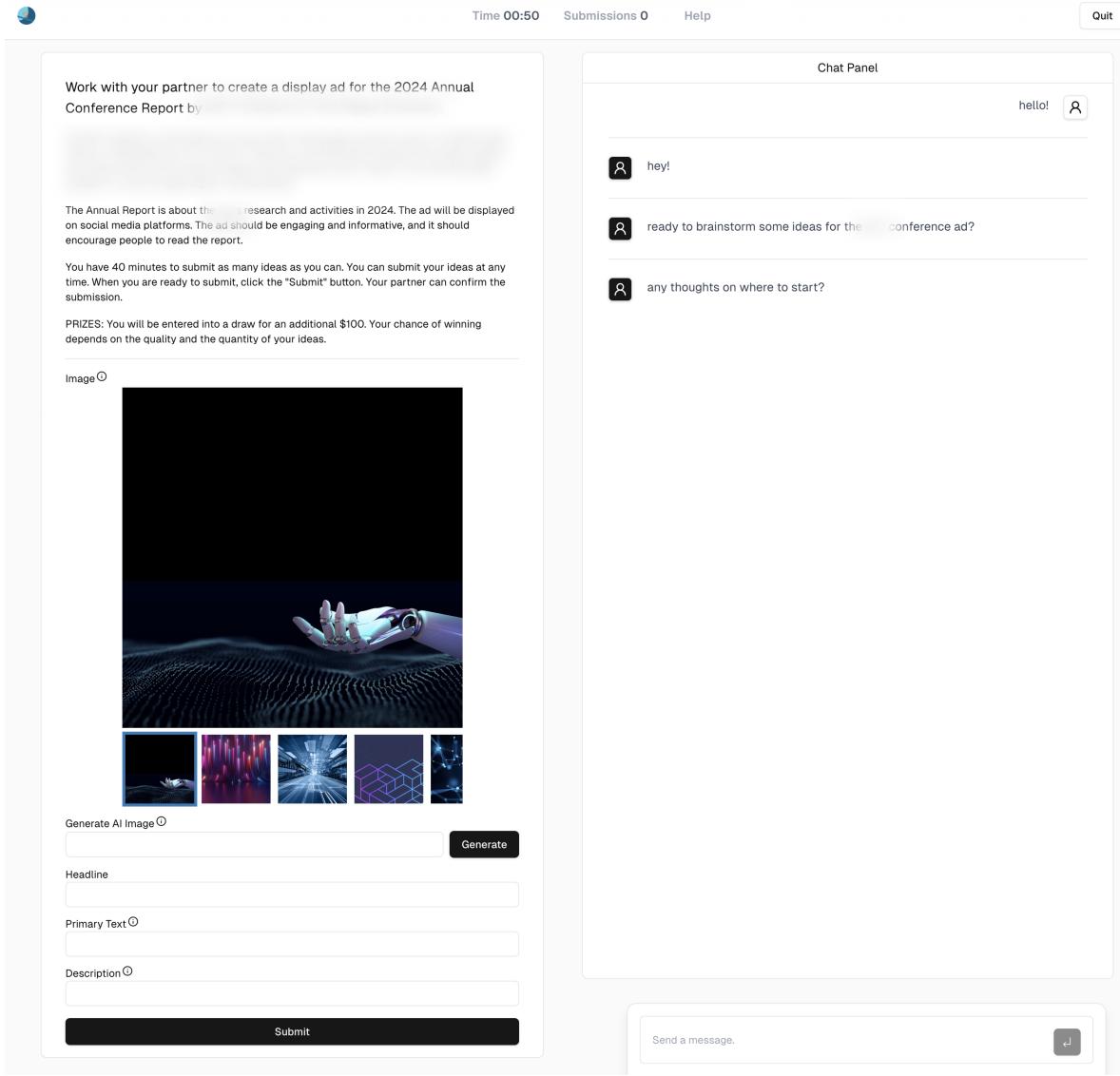


Figure 1: The Pairit platform. On the left is the task panel, and on the right is the chat panel. In the Human-Human condition, chat messages and edits on the task panel, including text edits, image selections, and AI image generations, are synchronized in real-time. In the Human-AI condition, the participant chats with an AI agent with full context of the user interface (UI; see Section 2.2), and the AI can edit text, select images, and generate AI images.

2.2 AI agent

Context. The platform uses OpenAI’s multimodal gpt-4o model (specifically gpt-4o-2024-08-06), which processes text and images in a single model. To give the AI full context of the task and collaboration, each API call to gpt-4o is prompted with the information on the screen so it has the full context of the collaboration. In the prompt, we include the same text of the task given to participants, previous submissions, current copy, elapsed time, history of its own actions, history of its chain-of-thought, chat history, and general instructions. Moreover, a screenshot is taken of the image after every change and included as an input so the AI can observe the image and its evolution over time and edits. These details enable human collaboration with a fully functional, multimodal AI agent.

Actions. To ensure that the human-human and human-AI conditions are comparable, the AI agent can take the same actions a human participant except for submitting ads. The actions include sending messages, editing any element of the ad copy (including the headline, primary text, and description), selecting images, generating images using the Dall-E 3 interface, and waiting (*i.e.*, not taking any action). The agent is prompted every 10 seconds whether to engage in action.

Chain-of-thought. To ensure the agent is taking actions appropriately, we use chain-of-thought prompting (Wei et al., 2023).² Specifically, we prompt the model with questions to reflect on the state of the collaboration. These questions were necessary to avoid undesirable model behavior, such as repeatedly sending the same message and were developed and perfected through trial and error.

2.3 Procedure

AI randomization and queuing. As soon as a participant was redirected to our platform from Prolific, they were randomized into either the human-human or the human-AI condition (Figure 2A).

²We used OpenAI’s structured output feature for chain-of-thought prompting. See their [documentation](#).

In the human-human condition, participants joined a queue until another participant was available, at which point they were paired with each other. In the human-AI condition, a participant joined a simulated queue, in which they waited for a random amount of time between 1 and 5 seconds, after which they were paired with an AI agent. We do not reveal whether or not the partner is a human or an AI until the post-task survey.

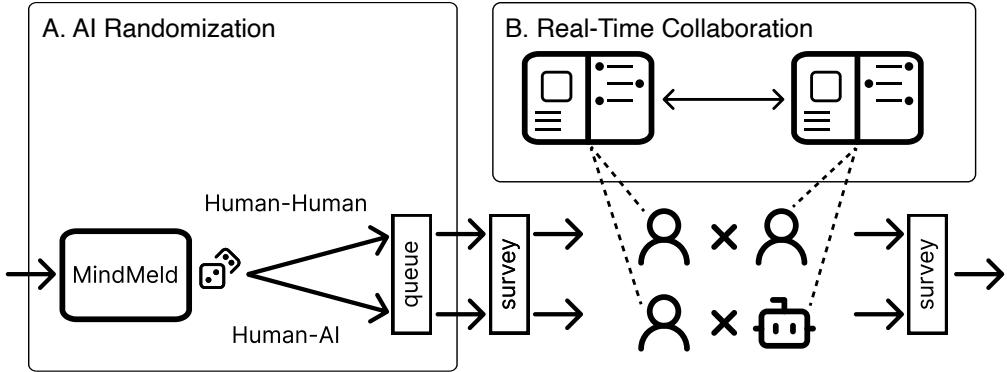


Figure 2: Overview of methods. (A) Participants are randomized into collaborating with another participant or an AI agent. (B) Participants collaborate with another participant or an AI agent to produce ads in a real-time collaborative workspace.

Pre-task survey. After participants were paired with each other, they answered a 10-item survey, on a 7-point Likert scale, to measure their Big Five personality traits (Rammstedt and John, 2007).

Ad creation task. Once the participants were paired and completed a pre-task survey, they entered the collaborative workspace in which they could message each other and edit and submit ads (Figure 2B). The edits were synchronized and messages were transmitted between the participants in real-time using websockets (*i.e.*, [tiptap.dev](#) and [pusher.com](#)), as they are in commercial collaboration tools (*e.g.*, Google Docs) and chat applications (*e.g.*, Slack). The participants had 40 minutes to submit as many ads as they could produce. At the end of the 40 minutes, the participants were automatically redirected to the post-task survey.

Post-task survey. After the completion of the ad creation task, the participants answered a 35-item teamwork quality survey (Hoegl and Gemueden, 2001). The original survey consisted of 38 items,

each on a 7-point Likert scale; however, we removed 3 items from the communication facet that did not apply to a single-session online collaboration task. In addition, we asked four questions regarding participants' perception of AI, all on a 7-point Likert scale. These included two questions about their experience using AI (*i.e.*, "I have used artificial intelligence (AI) chatbots before (e.g., ChatGPT, Bard)." and "I had a positive experience using AI chatbots."), one question about their perception of their partner as an AI (*i.e.*, "I believe my partner was an AI during the task.") and a final question in which we revealed the identity of their partner (human or AI) and ask whether that changed the perception of the quality of the collaboration (*i.e.*, "Your partner was [an AI assistant/a human]. Knowing this, to what extent has your perception of the quality of your collaboration changed?"). 85% of participants completed the post-task survey.

2.4 Participants

We pre-registered a sample of 2,500 participants from Prolific (www.prolific.com) based in the US, with representative stratification across gender and ethnicity. In total, 2310 participants completed the task. Of the 2500 participants who entered, 23 did not enter the matching queue (see Section 2.3 for details) and were removed from the study. A further 167 participants quit the study or were timed out before they were matched to a partner. The overall attrition rate was 7.6%. The study was conducted from October 15 to 18, 2024, and took a median of 46.32 minutes to complete. The participants were paid \$9.

2.5 Summary statistics and randomization checks

In total, the dataset includes 2,310 participants, 1,834 teams, 11,138 display ad submissions, 183,691 messages, 63,656 edits on images, 1,960,095 edits on ad copy, and 10,375 AI-generated images. To ensure that the randomization procedure successfully balanced covariates across experimental conditions, Table 1 compares key participant characteristics between the human-human and human-AI conditions. These include demographic variables (e.g. age, gender), participants' work status, and the Big Five personality traits. No significant differences were detected apart from a very

small difference in part-time employment status across treatment conditions, indicating broad-based, balanced randomization across participants.

<i>Variables</i>	<i>All</i>	<i>Human-AI</i>	<i>Human-Human</i>	<i>t-statistic (SE)</i>	<i>p-value</i>
Individuals	2310	1258	1052	-	-
Teams	1834	1258	576	-	-
Gender (% Male)	50.8%	50.4%	51.3%	-0.401 (0.021)	0.688
Age	42 ± 14	43 ± 15	42 ± 14	0.987 (0.599)	0.324
Full-Time	47.8%	48.5%	47.1%	0.688 (0.021)	0.491
Part-Time	14.7%	13.4%	16.3%	-2.009* (0.015)	0.045
Data expired	13.4%	13.2%	13.6%	-0.279 (0.014)	0.780
Not in paid work	10.3%	10.9%	9.7%	0.943 (0.013)	0.346
Unemployed (and job seeking)	8.8%	9.1%	8.6%	0.428 (0.012)	0.668
Other	4.2%	4.2%	4.3%	-0.077 (0.008)	0.939
Start new job in a month	0.6%	0.8%	0.5%	0.974 (0.003)	0.330
Openness	0.71 ± 0.20	0.71 ± 0.21	0.72 ± 0.19	-0.854 (0.008)	0.393
Conscientiousness	0.79 ± 0.18	0.78 ± 0.18	0.79 ± 0.17	-1.166 (0.007)	0.244
Extraversion	0.56 ± 0.21	0.55 ± 0.21	0.56 ± 0.21	-0.962 (0.009)	0.336
Agreeableness	0.70 ± 0.22	0.70 ± 0.21	0.69 ± 0.24	0.534 (0.009)	0.594
Neuroticism	0.48 ± 0.23	0.49 ± 0.24	0.48 ± 0.22	1.538 (0.010)	0.124

Notes: * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Personality traits are normalized from a 7-point Likert scale. For t-statistics, gender is coded as 1 for male and 0 for female, and employment status is 1 for the listed category (e.g., full-time) and 0 otherwise.

Table 1: Randomization checks

2.6 Incentives

To incentivize participants to create high-quality ads, we informed them that they were eligible for additional \$100 prizes based on the quantity and quality of the ads they submitted, as well as the performance of the ads in the marketing field experiment. Participants were explicitly instructed that “the greater the number of ads, the greater your chances—but not if the ads are of low quality.” Ultimately, two participants were awarded \$100 each for producing the greatest number of high-performing ads.

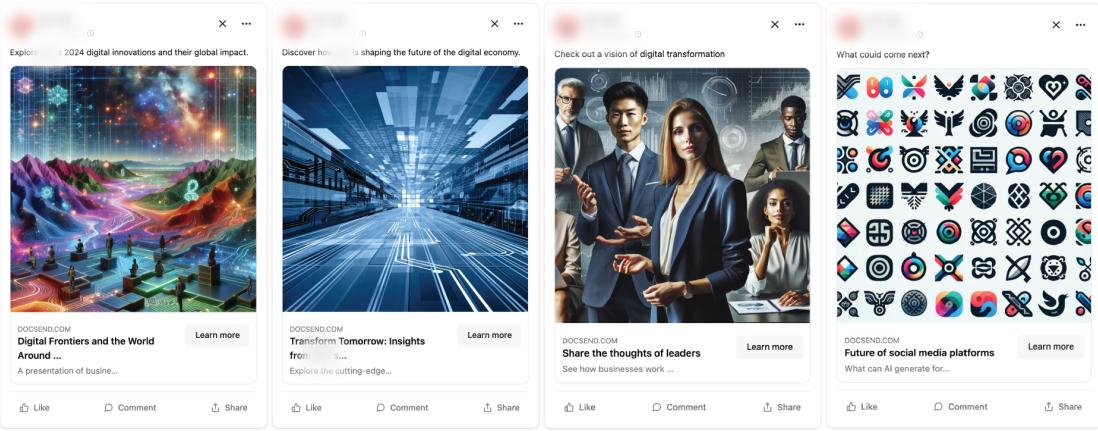


Figure 3: Examples of ad mockups.

2.7 Message labeling

To label messages into categories and intent, we prompt gpt-4o-mini-2024-07-18 for each message independently and ask for a label for a category and an intent of the message. We enforce that the labels are from the set of pre-determined labels using OpenAI’s Structured Outputs API.

2.8 AI evaluation of ad quality

Ad mockups. To obtain ratings of display ads that are as close as possible to display ads one would see on digital ad publishers, we first created mockups of each display ad. An ad mockup is a simulated representation of a display ad designed to closely resemble the appearance and functionality of an actual ad as it would appear on digital ad publishers’ platforms. We built a web application that populates the image, ad copy, shortened link, call-to-action, and other user interface items—including the *Like*, *Comment*, *Share*, and *Close* buttons, the profile picture, and the *Sponsored* tag. Screenshots were then programmatically taken of each mockup. See Figure 3 for examples.

AI ratings. If AI ratings are shown to predict human evaluations of ad quality and field evaluations of ad performance, they can potentially provide cost-effective alternatives to ad testing. To obtain AI ratings of ad quality, we prompt OpenAI’s multimodal gpt-4o-mini-2024-07-18, which supports

image input and structured outputs.³ To make AI evaluations comparable to human evaluations, we asked the same questions of the AI as those given to human evaluators (see Section 2.9). Each item was evaluated on a 7-point Likert scale. The first rating was based on the following prompt “The text is present, clear, relevant, and engaging”; the second on “The image is visually appealing”; and the third on “I am likely to click on this ad.”

2.9 Human evaluation of ad quality

Participants. To obtain human ratings of ad quality, we recruited a separate sample of 1,995 participants from Prolific based in the US, with representative stratification across gender, age, and ethnicity. A total of 1,200 individuals entered the survey. Of these participants, 5 dropped out before submitting their surveys, resulting in a dropout rate of 0.42%. We built a custom platform for this survey, run on Google Cloud Platform’s App Engine.⁴ This evaluation was conducted from November 7 to 9, 2024, and took a median of 16.47 minutes to complete.

Ad samples. To obtain ratings for all ads while avoiding survey fatigue, we created a random sample, with random order, of 40 ads per participant. To ensure that each ad received at least 3 ratings, we produced a set of 1,300 samples. As a participant entered our survey platform, we drew one sample from the set, without replacement, to provide to the participant, with each participant receiving a unique sample.

Survey items. We used the same mockups of display ads as used for AI evaluation of ad quality, as explained in Section 2.8. For each ad, we asked participants to provide three ratings regarding the quality of the text, the image, and the estimated click through rate. Each item was rated on a 7-point Likert scale. The first rating was based on the following prompt “The text is present, clear, relevant, and engaging”; the second on “The image is visually appealing”; and the third on “I am likely to click on this ad.” See Figure 4 for an example of an item on the survey.

³See OpenAI’s documentation on [vision](#) and [structured outputs](#).

⁴The code is available on [GitHub](#).

Display Ad Survey

The ad is for an annual report by a research organization that ... Please rate the following questions based on the ad below:

Ad 2/40

Q1 — Q2 — Q3

The text is present, clear, relevant, and engaging.

● ↓ Select option

Strongly Agree

Agree

Somewhat Agree

Neutral

Somewhat Disagree

Disagree

Strongly Disagree



Previous

Next

Figure 4: The user interface for the ad quality survey.

2.10 Field evaluation of ad performance

To assess real-world advertising outcomes, we ran ad campaigns on the social media platform X, pre-registered on osf.io/95dhu. We use the following as outcome variables: click-through rate (CTR), cost-per-click (CPC), view-through rate (VTR; as a fraction of document viewed), and view-through duration (VTD; in seconds). We tracked VTR and VTD by using a unique link on DocSend for each ad.

Due to the effects of divergent targeting in A-B testing ([Braun and Schwartz, 2024](#)), we follow recommendation 5 from [Braun et al. \(2024\)](#) and run a multi-ad study to test the causal effects of the ads on performance, conditional on online algorithms. We do not use a holdout in this study. Each ad has a unique DocSend link, so we expect the outcome in a hypothetical holdout group to be zero

and attribute all document visits to the ads (Braun et al., 2024).

To run as many ads as possible while sampling evenly from human-human and human-AI teams and from the human predictions of ad quality, we implemented a stratified sampling approach. We obtained a sample of 2,000 ads from a total of 11,138 ads, sampling between one and two ads from each of the 1,834 teams. We first divided the ads into those created by human-human teams and human-AI teams and then further divided the ads into 10 strata according to the human predictions of the likelihood of user clicks, from low to high predicted click probability (see Section 2.9). We removed 8 ads that potentially violated moderation policies (due to the inclusion of *e.g.* violence, sexual, or drug content). We controlled for advertising spend in our analyses to account for auto-bidding.

To prevent overlapping audience targeting, we assigned five unique ads to each of 400 campaigns, structuring them as 5-ad split tests within individual campaigns (noting the platform’s limit of five split tests per campaign). To further ensure no overlap across campaigns, we allocated a random set of 133 ZIP codes to each campaign, selecting ZIP codes with populations between 10,000 and 100,000.⁵ We tested the robustness of this allocation across ZIP codes by conducting one-way ANOVA on population and income, yielding non-significant results: $F(399, 53199) = 0.954$, $p = 0.734$ for population, and $F(399, 53199) = 0.973$, $p = 0.636$ for income.

Ad impressions were delivered from January 21, 2025, to February 9, 2025. Due to platform limits in running separate ad campaigns, we ran 50 ad campaigns for two days at a time. We control for any potential temporal confounders with campaign random effects.

2.11 Model specifications

Human-AI collaboration effects. Due to the experimental nature of our design, we measure the effects of human-AI collaboration using a standard regression model:

$$Y_i = \delta HAI_i + \beta X_i + \varepsilon_i, \quad (1)$$

⁵The data is from <https://data.census.gov/table/ACSDT5Y2020.B01003>.

where Y_t represents the outcome for team t , HAI_i is 1 if individual i is working with an AI and 0 if individual i is working with another participant, X_i includes demographic controls including age, gender, and Big Five personality traits, and ε_t is the error term. The outcomes of interest Y_t include the number of interactions on the platform (*i.e.*, messages, copy edits, image selects, AI image generations, submissions), fraction of messages by category, quality scores, and fraction of copy completed. For the analysis of ad quality, we perform analysis on the rating level where i is an ad rating. For team-level analysis, we use the following model:

$$Y_t = \delta_t \text{HAI}_t + \beta X_t + \varepsilon_t, \quad (2)$$

where the outcome Y_t is the same as above, HAI_t (Human-AI) is 1 if the team was Human-AI and 0 if the team was Human-Human, X_t includes team-level demographic controls including average age, gender composition, and average Big Five personality traits, and ε_t is the error term.

Field evaluations. To evaluate the results of our field experiment, we use the following mixed effects model:

$$\begin{aligned} Y_{tc} = & \delta_t \text{HAI}_{tc} + \text{Text}_{tc} + \text{Image}_{tc} + \text{Click}_{tc} \\ & + \text{Spend}_{tc} + u_c + \varepsilon_{tc}, \end{aligned} \quad (3)$$

where Y_{tc} represents the outcome (*e.g.*, CPC, CTR, VTR, VTD) for ad t in campaign c , Image, Text, and Click are human-rated quality measures from Section 2.9, Spend_{tc} is the campaign spend, and $u_c \sim N(0, \sigma_u^2)$ is the random intercept for campaign c .

3 Results

3.1 Work Processes

Collaborating with AI increases communication and decreases direct engagement with work product. To test how human-AI collaboration affects communication, we measured the number of messages sent by participants in both human-human and human-AI groups. Participants worked collaboratively with and sent messages to their partners in real-time, for both human-human and human-AI groups. In the human-human group, participants messaged their human partners; in the human-AI group, participants messaged the AI. The Pairit platform logs a timestamped record of each message sent by a participant or AI.

Participants communicated significantly more in human-AI teams than in human-human teams. Participants in human-AI teams sent 63% more messages than those in human-human teams, with consistent results when demographic covariates are included (Table 2). These results indicate that collaborating with an AI partner encourages more frequent communication.

Messages		
Intercept	21.504*** (0.484)	9.977** (2.999)
Human-AI	13.669*** (0.834)	13.671*** (0.834)
Demographics	No	Yes
Observations	2310	2310

Notes: * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Robust standard errors account for heteroskedasticity.

Table 2: Collaborating with AI increases communication at the individual level.

On the other hand, participants in Human-AI teams made 71% fewer direct copy edits compared to Human-Human teams (Table 3). The results on copy edits and messages suggest that collaboration with AI shifted participants' focus from generating text to instructing and suggesting the AI to generate text, as we will see in greater detail in the next set of analyses. A *post hoc* analysis of

the correlation between user actions found that the number of messages sent and copy edits were negatively correlated ($R^2 = -0.17$, $p = 3.5^{-17}$), further supporting this interpretation. However, participants in the Human-AI condition still engaged in direct editing, indicating that while AI collaboration reduced the frequency of manual edits, it did not eliminate them entirely. Unlike previous studies, such as [Chen and Chan \(2024\)](#), which artificially constrained participants to specific modalities when interacting with AI, our results show that in more realistic collaborative settings, where humans and agents can freely allocate effort to different tasks, participants utilized both direct engagement with work product (ad editing) and interaction through messaging.

	<i>Copy Edits</i>		<i>Image Selects</i>		<i>AI-Generated Images</i>	
Intercept	1367.225*** (30.059)	1630.869*** (117.896)	23.219*** (0.579)	23.447*** (3.727)	4.234*** (0.149)	5.921*** (0.955)
Human-AI	-974.224*** (32.939)	-971.857*** (32.812)	7.432*** (0.907)	7.403*** (0.911)	0.401 (0.214)	0.420* (0.212)
Demographics	No	Yes	No	Yes	No	Yes
Observations	2310	2310	2310	2310	2310	2310

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors account for heteroskedasticity.

Table 3: Collaborating with AI reduces copy edits but increases image edits and AI image generations at the individual level.

Human-AI teams focus on content and process over social and emotional communication. In addition to the number of messages sent across conditions, we investigated whether the content of messages varied across conditions. We used gpt-4o-mini to label each message independently with one of five *message categories*: Process, Content, Social, Emotional, Feedback, and Other (see Section 2.7 for more details). In our analysis, we found that both the AI and the humans in human-AI teams sent 18% (*i.e.*, $(0.406 - 0.345)/0.345 = 0.177$) more content- and process-oriented messages, while human-human teams sent 29% (*i.e.*, $(0.579 - 0.449)/0.449 = 0.29$) more social and emotional messages (Figure 5, Table 4). This shift indicates that collaboration with AI emphasizes task-related communication over social interaction, possibly because participants can focus more on the task without needing to navigate the social or emotional aspects of collaboration.

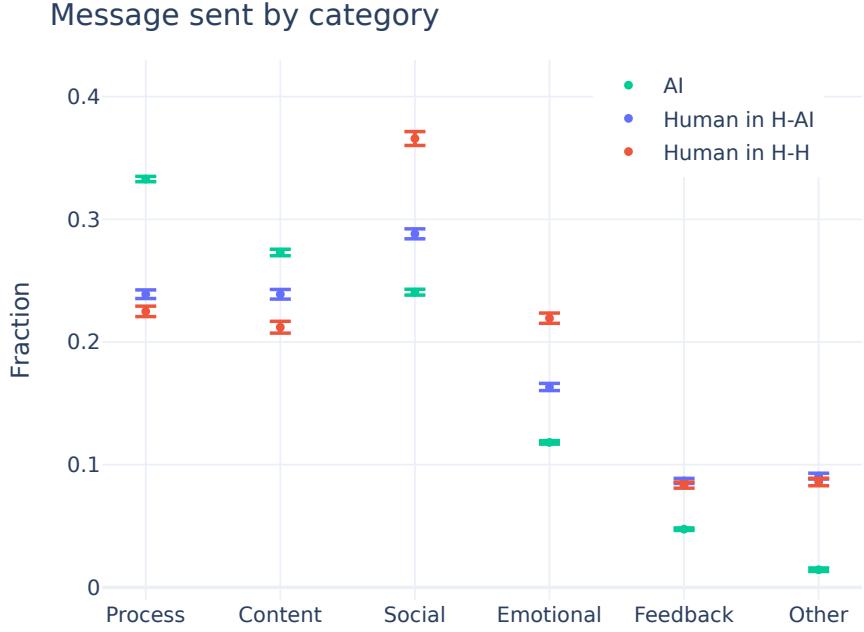


Figure 5: Participants in human-AI teams send more process- and content-related messages while those in human-human teams send more social and emotional messages.

To further examine how communication differs across conditions, we labeled the messages with one of 36 *message intentions*. These intentions capture a range of functions, such as suggestions, instructions, and prioritization, as well as rapport building and self-assessment. We found notable differences in the distribution of message intentions between the human-human and human-AI groups in each message category (see Table 4 and Figure 6). For example, in the Content category, human-human teams send more messages that exhibit confusion and clarification, and human-AI teams send more messages about brainstorming, confirmation, acknowledgment, suggestion, agreement, instruction, and judgment. The results are similar for process-related messages. For messages in the social and emotional categories, human-human teams send more apologies and messages about humor and concern, while human-AI teams send more suggestions and messages about appreciation, motivation, confirmation, and satisfaction. These findings further support the idea that human-AI collaboration prioritizes task-oriented communication, while human-human teams engage more in social and emotional exchanges.

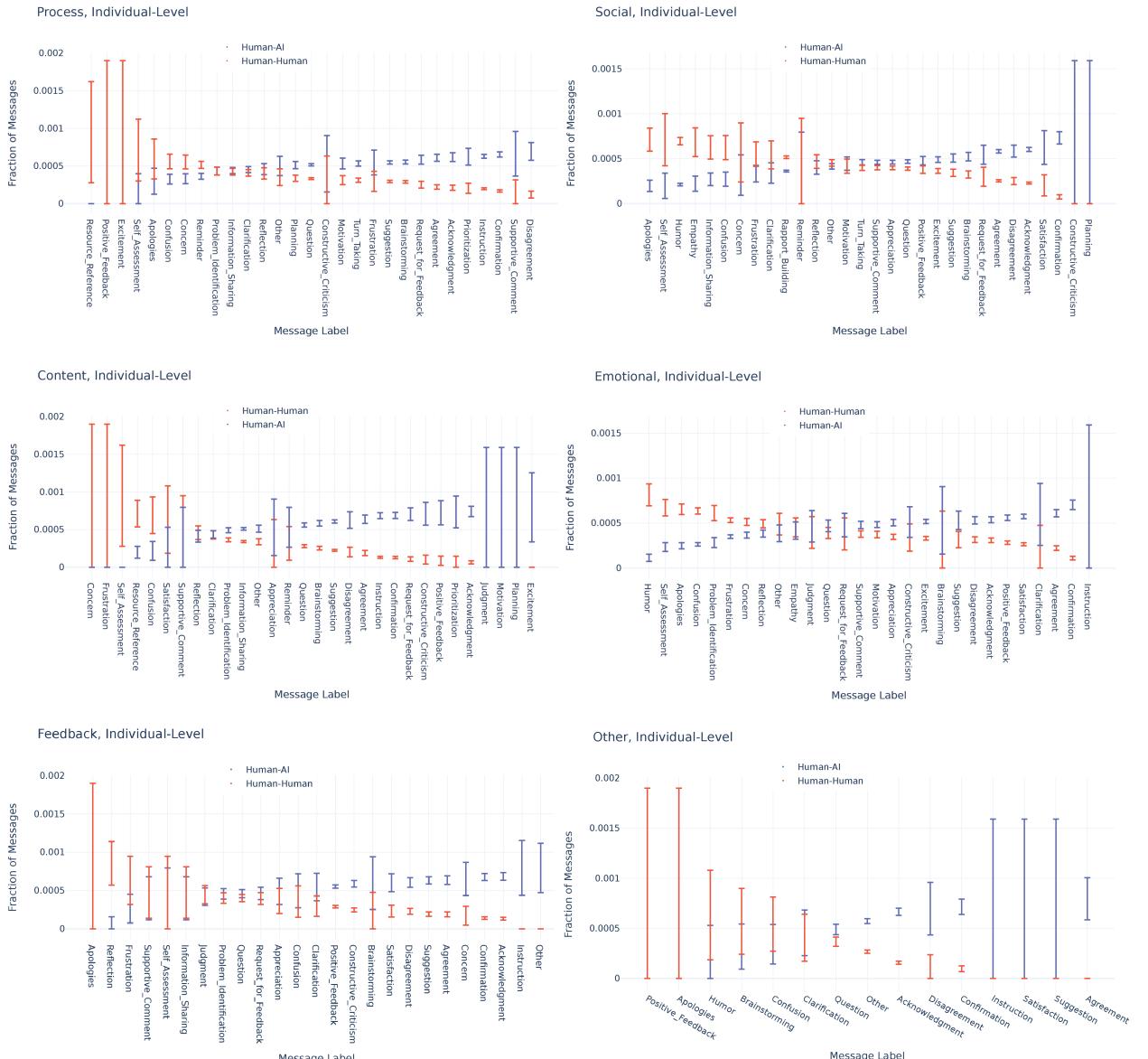


Figure 6: Message labeled by *intent* for each *category*.

	<i>Content</i>	<i>Process</i>	<i>Social</i>	<i>Emotional</i>	<i>Feedback</i>	<i>Other</i>
Intercept	0.156*** (0.025)	0.189*** (0.021)	0.415*** (0.028)	0.164*** (0.017)	0.053*** (0.010)	0.051*** (0.011)
Human-AI	0.036*** (0.006)	0.025*** (0.006)	-0.085*** (0.007)	-0.045*** (0.005)	0.019*** (0.003)	0.018*** (0.003)
Demographics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2310	2310	2310	2310	2310	2310

Notes: * $p<0.05$; ** $p<0.01$; *** $p<0.001$. Heteroskedasticity-robust standard errors are reported.

Table 4: Collaborating with AI increases content-, process-, and feedback-related messages while decreasing social and emotional messages.

Survey reports showed no differences in perceptions of teamwork quality. Despite the substantial differences observed in communication patterns and task-focused behavior between human-human and human-AI teams, survey responses indicated no significant differences in participants' perceptions of teamwork quality across conditions. This lack of differentiation might stem from the way participants perceive AI: unlike human collaborators, AI agents may not evoke the same social or emotional expectations. These results suggest that while collaboration dynamics differ, participants may not attribute those differences to changes in teamwork quality when working with AI, though teammate identity has been found not to affect trust (Zhang et al., 2023). For future studies relying on survey responses, this highlights a potential limitation: actual differences in collaboration dynamics may not be fully reflected in self-reported measures of teamwork quality.

3.2 Productivity

Having established that collaborating with AI agents allows participants to focus on the task without the social coordination costs typically associated with human collaboration, we now examine its impact on productivity. Specifically, we analyze the number of submissions per team and individual and the completion rate of ad copy.

Human-AI teams produce a similar amount of output with half as many workers. Participants were incentivized to submit as many ads as possible within the time limit. At the team level,

productivity was comparable between human-human and human-AI teams as human-AI teams submitted approximately the same number of ads with half as many workers (Table 5). At the individual level, participants in human-AI teams submitted 73% more ads than their counterparts in human-human teams, with results consistent when demographic controls are added. AI collaboration thus supports individual productivity by enabling participants to generate more output per worker compared to those in human-human teams.

Although our findings suggest that human-AI teams produce a similar amount of output as human-human teams with half as many workers, it is important to note that our study did not include a human-alone condition. Prior studies have consistently shown that AI tools enhance individual productivity compared to humans working alone across various tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2023; Dell’Acqua et al., 2023; Chen and Chan, 2024; Wiles et al., 2023). In light of this evidence, our results suggest that the AI in human-AI teams acts as a near-substitute for an additional human collaborator, enabling a single human to achieve productivity levels comparable to a team with two human workers. Without a human-alone condition, however, we cannot directly quantify AI’s incremental contribution beyond what an individual human might have achieved working alone. We explore this limitation further in the Discussion section.

	<i>Submissions</i>			
	<i>Team-Level</i>		<i>Individual-Level</i>	
Intercept	5.950*** (0.207)	7.483*** (0.832)	3.203*** (0.104)	3.935*** (0.642)
Human-AI	-0.406 (0.246)	-0.395 (0.246)	2.341*** (0.169)	2.366*** (0.169)
Demographics	No	Yes	No	Yes
Observations	1834	1834	2310	2310

Notes: * $p<0.05$; ** $p<0.01$; *** $p<0.001$. Robust standard errors account for heteroskedasticity..

Table 5: Productivity of human-AI and human-human teams.

With that caveat in mind, to test the robustness of our results to the alternative hypothesis that individual-level productivity differences are simply due to the submission of incomplete ads, we

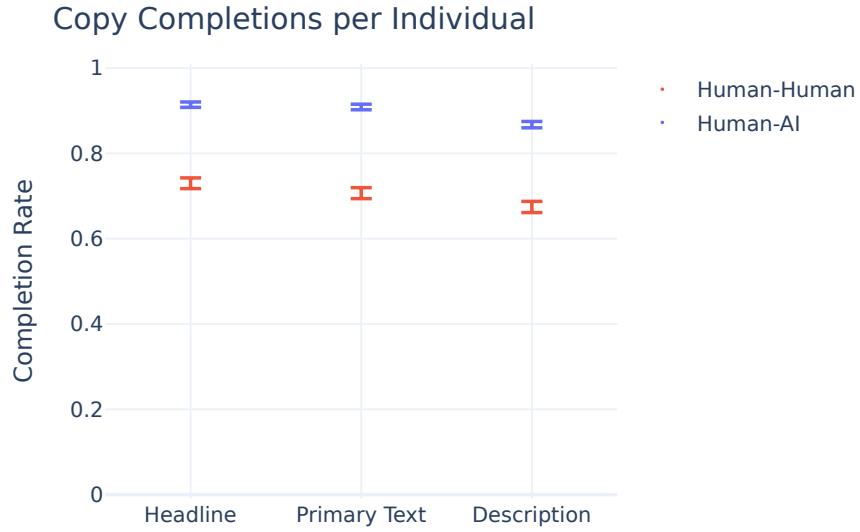


Figure 7: Copy completion rates.

examined ad copy completion rates across conditions. As shown in Table 6 and Figure 7, participants in human-AI teams had consistently higher completion rates for ad copy elements compared to participants in human-human teams. Taken together, these findings suggest that human-AI teams are indeed more productive, in both the number of ads submitted and the ad completion rate.

	<i>Headline</i>	<i>Primary Text</i>	<i>Description</i>
Intercept	0.728*** (0.058)	0.737*** (0.058)	0.651*** (0.058)
Human-AI	0.189*** (0.014)	0.207*** (0.014)	0.199*** (0.015)
Demographics	Yes	Yes	Yes
Observations	2310	2310	2310

Notes: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Heteroskedasticity-robust standard errors are reported.

Table 6: Individuals in Human-AI teams submit ads with more copy completed.

3.3 Ad Quality

In addition to the productivity and work process implications of collaborating with AI agents, we also evaluated the performance of human-human and human-AI teams by assessing the quality of their output. Specifically, we examined differences in the quality of ad text and images and conducted field evaluations of ad effectiveness to understand how human-AI collaboration dynamics influence overall effectiveness.

Ad text quality improves, but image quality declines in human-AI teams. In human evaluations of the quality of the ads, we found that human-AI teams had higher quality ad text but lower quality ad images compared to human-human teams (Table 7, Figure 8). The estimated likelihood of clicking on the ad was indistinguishable between the two groups. Interestingly, in our AI evaluations of ad quality, we found that AI ratings were higher on the text and clicks for ads produced by human-AI teams and the same across the groups for image quality. In a way, it is unsurprising that the AI rated these ads' text quality, image quality, and estimated click likelihood as equal to or better than those produced by human-human teams because these ads were created in collaboration with OpenAI's gpt-4o.

However, the human ratings reveal that AI introduces important trade-offs in output quality. Specifically, collaborating with GPT models improves text quality but reduces image quality. This trade-off may arise because GPT models were originally designed for next-word prediction rather than to predict image quality. These results suggest that while GPT models enhance text-based outputs, their contributions to multimodal outputs like ad images may require complementary tools or models designed specifically for image-related tasks. These findings then beg the question of whether and how these trade-offs in text and image quality meaningfully impact actual click through rates, which we examined in detail in the field test of the ads themselves.

We also examined how work process variables, like the number of messages exchanged, the number of ads submitted, the number of AI images generated, and the number of copy and image edits, affect ad quality in post hoc analyses. We found that more messages exchanged, fewer ads

	Human Evaluations						AI Evaluations				
	Text			Image		Click	Text			Image	
Intercept	4.535*** (0.092)	4.588*** (0.102)	4.730*** (0.080)	4.523*** (0.089)	3.592*** (0.079)	3.433*** (0.088)	5.410*** (0.056)	5.423*** (0.062)	6.438*** (0.054)	6.345*** (0.059)	5.235*** (0.045)
Human-AI	0.324*** (0.024)	0.435*** (0.053)	-0.134*** (0.021)	-0.159*** (0.044)	-0.014 (0.021)	0.037 (0.045)	0.122*** (0.015)	0.086** (0.031)	-0.014 (0.014)	-0.144*** (0.029)	0.068*** (0.013)
# Submissions	-0.028*** (0.002)	-0.012*** (0.001)		-0.018*** (0.001)		-0.013*** (0.001)		-0.010*** (0.001)		-0.008*** (0.001)	
# AI Images Generated	0.007*** (0.002)	0.029*** (0.002)		0.023*** (0.002)		0.023*** (0.001)		0.028*** (0.001)		0.019*** (0.001)	
# Human Messages	0.004*** (0.000)	0.002** (0.000)		0.002*** (0.000)		0.003*** (0.000)		0.002*** (0.000)		0.001*** (0.000)	
# Human Copy Edits	0.136*** (0.017)	0.042** (0.015)		0.082*** (0.015)		-0.027* (0.011)		-0.028** (0.010)		-0.040*** (0.010)	
# Human Copy Edits × (H-AI)	-0.033 (0.027)	-0.004 (0.027)		-0.022 (0.026)		-0.020 (0.018)		0.028 (0.017)		-0.048** (0.016)	
# Human Image Edits	-2.173* (0.897)	-1.019 (0.707)		-0.560 (0.726)		-2.165*** (0.542)		-2.390** (0.459)		-1.853*** (0.442)	
# Human Image Edits × (H-AI)	2.057* (1.009)	1.974* (0.829)		1.665* (0.832)		1.057 (0.591)		2.789*** (0.514)		1.555** (0.487)	
Demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	45803	45803	45803	45803	45803	11138	11138	11138	11138	11138	11138

Notes: * $p<0.05$; ** $p<0.01$; *** $p<0.001$. Standard errors are clustered at the ad level to account for within-ad correlation. “H-AI” refers to human-AI teams. “Human AI Images Generated” refers to the AI images generated by humans. Copy and image edits are per thousand.

Table 7: Human and AI ratings of ads.

submitted, and more AI images generated were associated with higher human- and AI-evaluated ad quality. Interestingly, for text copy and image edits, we observed nuanced differences in ad quality across team types and human and AI evaluations. Our analyses showed that humans rate ads more highly when humans copy edit text more in human-human teams. In contrast, AI rate ads more highly when humans copy edit text less in human-human teams. Human text copy edits made in human-AI teams mostly revealed no difference in preference in either human or AI evaluations of ad quality (except for a slight negative preference for AI-evaluated click quality). Both humans and the AI rated ads *higher* when image edits were made in human-AI teams and rated ads *lower* when image edits were made in human-human teams. This finding is consistent with previous our findings that humans are better at selecting images—as human edits to images are more highly valued, by both humans and AI, in teams in which AI is contributing to the image selection and generation process, humans seem to be needed to correct, adjust or improve image selection and generation when AI is involved. These findings highlight that team composition (human-human vs human-AI teams) and work processes (messages exchanged, ads submitted, AI images generated, and copy and image edits) all meaningfully shape ad quality.

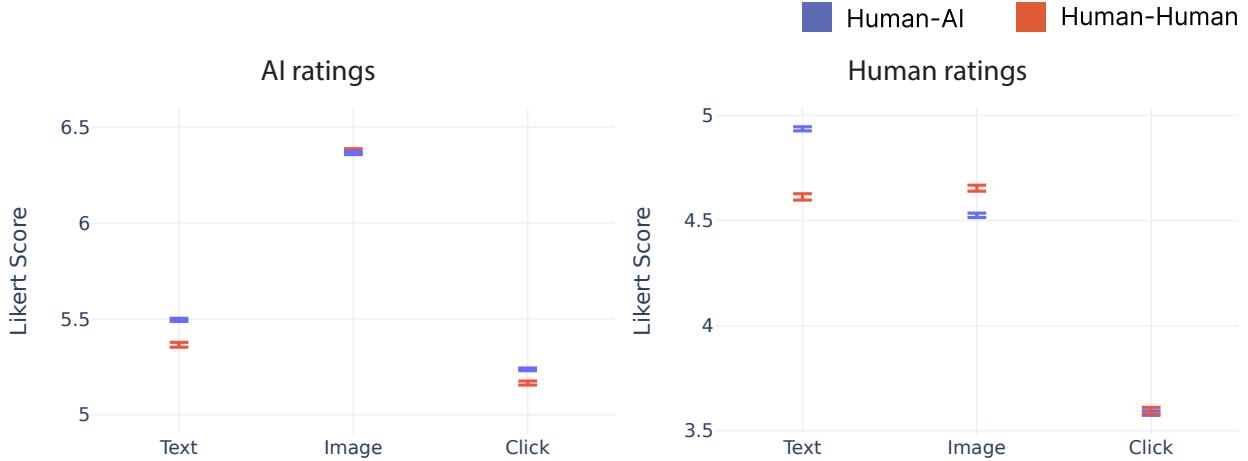


Figure 8: AI and human ratings of ads.

3.4 Ad Effectiveness

To evaluate the real-world performance of ads created by human-human and human-AI teams, we conducted a field experiment on the social media platform X. Our ad campaigns generated 4,932,373 impressions and 7,546 clicks over 20 days. This section examines how collaboration type influences key advertising metrics including the cost-per-click (CPC) and click-through rate (CTR), reported in Table 8, and the view-through rate (VTR) and view-through duration (VTD) shown in Table 9. The field study extends lab findings and tests how the distinct productivity and quality profiles of human-AI and human-human teams translate to advertising outcomes in a real world setting. Broadly, we found that ads with higher text quality, created more often by human-AI teams, and higher image quality, created more often by human-human teams, performed significantly better on cost per click and click through rate metrics. As human-AI teams produced higher quality text and lower quality images, their effects counteracted one another and human-AI ads performed similarly to those created by human-human teams overall.

Click measures We examined CPC and CTR using regression models with campaign random effects to account for unobserved heterogeneity across the 400 campaigns, with results in Table 8. For CPC, measured in dollars, our analysis in Column 1 reveals no significant effect of collaboration

type.⁶ With human-evaluated quality scores (*i.e.*, Image, Text, Click) as covariates, ads with stronger image quality significantly reduced CPC by \$0.268 or 2.3% (Column 2), and this pattern held when all predictors were included in the analysis (Column 3). For CTR, expressed as a percentage, the analysis with collaboration indicated no direct human-AI effect (Column 4). With quality scores, we found that text quality significantly reduced CTR by 0.008% (Column 5), which remained consistent when all factors were included (Column 6).

	CPC (\$)			CTR (%)		
Intercept	10.234*** (0.323)	11.636*** (0.686)	11.560*** (0.698)	0.000 (0.006)	-0.027* (0.013)	-0.026* (0.013)
Human-AI	0.115 (0.208)		0.128 (0.210)	-0.000 (0.004)		-0.001 (0.004)
Image		-0.268* (0.131)	-0.264* (0.131)		0.002 (0.002)	0.002 (0.002)
Text		-0.154 (0.114)	-0.163 (0.115)		0.008*** (0.002)	0.008*** (0.002)
Click		0.212 (0.147)	0.215 (0.147)		-0.002 (0.003)	-0.002 (0.003)
Spend	-0.135*** (0.012)	-0.141*** (0.013)	-0.140*** (0.013)	0.006*** (0.000)	0.006*** (0.000)	0.006*** (0.000)
Campaign RE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1859	1859	1859	2000	2000	2000

Notes: * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Campaign RE represents campaign random effects. Ads with zero clicks were removed for regressions for columns 1-3.

Table 8: Effects on cost-per-click (CPC) and click-through rates (CTR) from the field study.

View measures We assessed view metrics—view-through rate (VTR) and view-through duration (VTD; in log-seconds)—using similar regression models, as shown in Table 9. Our analysis of VTD followed a similar pattern as observed in the click metrics—while there was no direct human-AI effect (Column 4), text quality was significantly associated with longer viewing duration. On average, one point higher text quality ratings were associated with one second longer view through

⁶Higher spend consistently lowers costs and suggest divergent targeting and optimization, likely because we had kept on auto-bidding in our campaigns. Thus, we consider the effects present in this section as conditional on ad algorithms, as per recommendation 5 from Braun et al. (2024).

duration (Columns 5-6). Our analysis of VTR showed no significant effects for collaboration type or quality (Columns 1-3). These findings demonstrate that ads with higher text quality, which are more often produced by human-AI teams, are viewed for longer periods of time.

	VTR			VTD (<i>log-sec</i>)		
Intercept	0.000 (0.006)	0.000 (0.011)	0.000 (0.012)	0.509*** (0.052)	0.000 (0.116)	0.000 (0.118)
Human-AI	0.001 (0.004)		0.000 (0.004)	0.021 (0.037)		0.011 (0.037)
Click		-0.000 (0.003)	-0.000 (0.003)		-0.003 (0.026)	-0.003 (0.026)
Image		-0.003 (0.002)	-0.003 (0.002)		-0.001 (0.023)	-0.000 (0.023)
Text		0.002 (0.002)	0.002 (0.002)		0.038* (0.019)	0.037* (0.019)
Spend	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	-0.001 (0.002)	0.005* (0.002)	0.005** (0.002)
Campaign RE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2000	2000	2000	2000	2000	2000

Notes: * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Campaign RE represents campaign random effects.

Table 9: Effects on view-through rate (VTR) and view-through duration (VTD) from the field study.

4 Discussion

4.1 Theoretical, managerial, and methodological implications

Theoretical implications. This study advances theories of teamwork, collaboration, and human-AI interaction by demonstrating how GPT-based AI agents reshape work processes, communication dynamics, and task execution in collaborative settings (Schneider et al., 2021). Our findings reveal that AI agents shift communication from social and emotional exchanges to task-oriented interactions, reducing interpersonal coordination costs and enabling participants to prioritize content generation. This aligns with theoretical frameworks suggesting that AI can streamline team

processes by automating routine or procedural tasks (Wilson and Daugherty, 2018), but it extends these models by uncovering trade-offs in creative, multimodal workflows. Specifically, while AI agents enhanced text quality in ad creation, they underperformed in image-related tasks, suggesting that theories of human-AI collaboration must account for task-specific competencies and limitations, particularly in domains requiring diverse cognitive and creative skills (Dell’Acqua et al., 2023).

Moreover, our work challenges the traditional conceptualization of AI as a passive tool or assistant (Noy and Zhang, 2023; Dell’Acqua et al., 2023; Chen and Chan, 2024) by positioning GPT-based agents as active collaborators that shape the structure and quality of team outputs (Anthony et al., 2023; Makarius et al., 2020). For example, in our experiment, the AI’s proficiency in generating ad copy influenced participants to focus on generating text through instructions rather than direct edits. These changes in work processes suggest a need to reframe human-AI interaction theories to emphasize AI’s agency in collaborative processes. Rather than merely augmenting human efforts, AI agents co-create outcomes, introducing new dynamics that affect team performance.

Our findings also contribute to the literature on team cognition and distributed expertise (Collins et al., 2024). In human-human teams, expertise is often shared through social negotiation and emotional rapport, which fosters trust but can slow decision-making. In contrast, human-AI teams in our study bypassed much of this social overhead. This shift raises theoretical questions about how cognitive load and role allocation differ when humans collaborate with machines instead of other humans. Future theoretical work could explore these dynamics in other creative domains, such as product design, to develop a more nuanced understanding of how AI agents influence cognitive and collaborative processes across varied task environments.

Managerial implications. Our work also offers actionable insights for managers and organizations seeking to integrate GPT-based AI agents into collaborative workflows, particularly in creative and productivity-driven environments. Our findings demonstrate that AI agents significantly reduce social coordination costs, allowing team members to focus on task execution and content generation. Managers in industries like marketing, content creation, or consulting can leverage AI agents to

streamline routine tasks, such as drafting campaign materials or reports, freeing human employees for strategic activities like audience analysis or brand positioning.

However, the study also highlights critical limitations that managers must address to fully harness AI’s potential in multimodal workflows. While GPT-based agents excelled in text generation, they underperformed in image-related tasks. This trade-off supports the idea that organizations should adopt a hybrid approach along a jagged edge of AI capabilities (Dell’Acqua et al., 2023), allocating tasks within AI capabilities to agents and those outside AI capabilities, such as image selection, to humans. Such an approach can mitigate the quality disparities observed in our study and are particularly relevant in domains where compelling visual elements drive downstream outcomes.

Furthermore, our results provide a framework for structuring human-AI teams to optimize productivity and performance. AI’s ability to handle bulk editing tasks—as evidenced by the 71% fewer direct copy edits in human-AI teams—allowed participants to engage in more content ideation and quality control. Managers can apply this division of labor to design efficient workflows, assigning AI agents to repetitive or data-intensive tasks while reserving human expertise for creative oversight and final approvals. A practical example might involve a content production team where AI drafts blog posts or social media captions, which editors then polish for tone and brand alignment. This approach not only boosts efficiency but also leverages human strengths in contextual judgment, which AI currently lacks, particularly in nuanced visual tasks. To implement such workflows, organizations should invest in training programs that teach employees how to delegate tasks to AI effectively, fostering a collaborative culture that maximizes both technological and human contributions.

Finally, the study underscores the importance of aligning AI deployment with organizational goals and team dynamics. The comparable team-level productivity between human-AI and human-human teams, despite human-AI teams having half the human resources, suggests that AI can serve as a near-substitute for additional personnel in certain contexts. Managers could use this insight to scale team output without proportionally increasing headcount, a cost-effective strategy for startups or resource-constrained firms. However, they must remain mindful of the social dynamics altered

by AI integration. The decrease in social and emotional messages in human-AI teams indicates less time spent on rapport-building, which could impact team cohesion in long-term projects. Managers should monitor team morale and consider hybrid models—such as combining AI assistance with periodic human-human collaboration—to maintain a balance between efficiency and interpersonal connection.

Methodological implications. This study advances the methodological landscape for studying human-AI collaboration through the development of the Pairit platform, a novel experimental framework that redefines how researchers investigate team dynamics in real-time, task-driven settings. Unlike traditional approaches that rely on aggregate performance metrics or isolated AI interactions ([Noy and Zhang, 2023](#); [Dell'Acqua et al., 2023](#)), Pairit captures granular, timestamped data on every keystroke, message, edit, and API call, enabling a detailed reconstruction of collaborative workflows. In our experiment, this capability allowed us to analyze 183,691 messages, 1,960,095 text edits, and 63,656 image edits, revealing nuanced patterns such as how AI-driven suggestions shifted human effort from direct text editing to instructional messaging. This methodological advance allows us to analyze both aggregate and detailed dynamics of human-AI teamwork and offers researchers new insights into decision-making, task allocation, and interaction processes.

The randomization of team composition represents another significant methodological advance, moving beyond human-only versus human-AI comparisons to include human-human and human-AI pairings. Unlike prior research that often contrasts AI-augmented individuals with those working alone ([Noy and Zhang, 2023](#); [Vaccaro et al., 2024](#)), our approach allowed us to evaluate the effects on collaboration itself, which is critical to understand given the proliferation of AI agents in workplaces. This randomization framework is also adaptable to other configurations, such as teams with multiple humans or multiple AI agents, enabling researchers to test how team composition influences outcomes in contexts like project management, educational group work, or distributed software development.

To fairly compare human-human and human-AI collaboration, another critical contribution is that AI agents are *agentic*. That is, AI agents on Pairit are capable of performing nearly all actions available to humans, including sending chat messages, editing text, selecting images, and generating new images via external APIs (e.g., DALL-E 3), with the sole exception of final ad submission in this study. This approach ensures direct comparability between human-AI and human-human conditions, a departure from prior studies where AI is often just a chatbot (Chen and Chan, 2024; Brynjolfsson et al., 2023). By enabling AI to act independently and actively, we analyze how AI's active participation reshapes collaborative dynamics, which would not have been possible in experimental designs that treat AI as a passive tool.

Pairit's multimodal capabilities further distinguish it as a methodological tool, supporting tasks that integrate text, images, and external API interactions. While our study focused on ad creation, Pairit's architecture is extensible to diverse domains, such as software development (e.g., AI generating code while humans debug), scientific writing (e.g., AI drafting literature reviews), or multimedia content production (e.g., designing virtual reality environments). This versatility allows researchers to explore human-AI collaboration in tasks requiring varied cognitive and creative skills, addressing a gap in prior studies that often limit AI to text-based functions (Brynjolfsson et al., 2023). For example, a future study could use Pairit to analyze how AI agents assist in architectural planning, tracking both textual specifications and visual renderings to assess collaborative efficiency and output quality. By capturing multimodal interactions, Pairit provides a comprehensive framework for studying the interplay of different task components, a methodological leap forward for understanding complex workflows.

These methodological contributions have broader implications for experimental research and organizational practice. Pairit's ability to simulate realistic collaborative environments offers a testing ground for refining theoretical models of human-AI interaction and informing the design of AI-driven workflows. For instance, researchers could leverage the platform to study how AI agents perform in agile software development sprints, using time-stamped data to optimize task assignments and evaluate productivity trade-offs. Ultimately, Pairit's methodological innovations

bridge the gap between controlled experiments and real-world applications and provides a robust, scalable framework for advancing our understanding of human-AI collaboration.

4.2 Limitations and future directions

Limitations. While we conducted one of the first truly agent-randomized experiments in human-AI collaboration, combining laboratory and field tests, on a newly developed platform capable of recording fine-grained collaboration dynamics, our work is not without limitations. First, while we observe substantial differences in communication patterns between human-human and human-AI teams, the underlying mechanisms or explanations for these differences are not always obvious. They could stem from the nature of AI communication, which tends to prioritize content-related messages, or from human participants adjusting their behavior upon realizing their partner is an AI. Our exit surveys found that human participants were good at predicting whether their partner was an AI (2.7 points higher on a 7-point Likert scale, $p < .001$), suggesting AI communication and human adaptation are both at work in these mechanisms. But, future work is needed to elucidate these potential explanations and better understand the drivers of these shifts. Second, the limitations of current AI models, particularly Vision GPT models, present challenges for multimodal tasks. These models are optimized for next-word generation and specific visual tasks, such as identifying items in images, but are not designed for nuanced assessments like image quality prediction. This limitation likely contributed to the lower image quality observed in human-AI teams and underscores the need for AI systems tailored to specific creative and evaluative tasks. However, AI is a moving target and image selection and design competency likely will improve over time. As with any study of AI, results may drift as AI changes. Moreover, while our study provides a controlled experimental context, it may not fully capture the complexities of long-term collaboration with AI agents in real-world environments. Future research should explore how these dynamics evolve over extended periods and in more diverse task domains to validate and expand on our findings.

Finally, our study lacks a human alone condition, which prevents us from directly assessing how human-AI teams compare to individuals working independently. Without this baseline, we cannot

precisely isolate AI’s marginal contribution to productivity and the coordination costs associated with collaboration. Nevertheless, our findings have practical implications for organizations looking to integrate AI agents into collaborative workflows, particularly in tasks requiring creativity and coordination. Our work lays a foundation for understanding human-AI collaboration in complex tasks, and future research can build on this by incorporating a human-alone condition to further disentangle marginal productivity effects.

Future directions. Our work opens several promising avenues for future research. First, the limitations of current Vision GPT models suggest a need for separate, specialized image generation models. Future studies could incorporate state-of-the-art visual models designed specifically for tasks like image quality prediction and generation, enabling a more nuanced understanding of AI contributions to multimodal creative tasks. Second, our findings are grounded in the context of ad design, but they invite exploration across a range of other collaborative domains. Important contexts include software development (e.g., coding), data analysis, collaborative writing, and financial accounting. Investigating Human-AI collaboration in these areas could reveal domain-specific dynamics and inform the broader design of AI systems tailored to various professional workflows. Finally, extending this research to longitudinal settings could provide insights into how Human-AI collaboration evolves over time. Long-term studies could examine the sustainability of productivity gains, the development of trust, and the potential for “learning effects” where humans adapt to working with AI agents or vice versa. These investigations would bridge the gap between controlled experimental settings and real-world applications, enhancing the generalizability of our findings. We intend to make Pairit freely available to academics advancing peer-reviewed research into human-AI collaboration. We believe this platform could support hundreds, if not thousands, of new experimental studies on the effects of AI on productivity, performance, and work processes.

4.3 Conclusion

This study provides novel insights into how AI agents reshape work processes, productivity, and performance in collaborative settings. Our findings reveal that collaborating with AI agents significantly increases communication while decreasing (but not completely eliminating) direct engagement with the work product. Human-AI teams focus more on content and process-oriented communication, reducing the need for social and emotional exchanges compared to human-human teams. Humans in human-AI teams experienced 73% greater productivity per worker, suggesting that AI agents can act as near-substitutes for additional human collaborators. However, AI had countervailing effects on different dimensions of output quality—while ad text quality improved in human-AI teams, image quality declined, highlighting a trade-off in multimodal tasks. Field evaluations of ad effectiveness showed that ads with higher image quality (produced by human-human collaborations) and higher text quality (produced by human-AI collaborations) performed significantly better on click-through rates, view through rates, and cost per click metrics.

A key innovation of this study is the introduction of Pairit, an experimental platform that enables real-time collaboration between humans and multimodal AI agents. By providing researchers with a controlled environment to study these dynamics, Pairit opens new avenues for understanding how AI agents can enhance teamwork and productivity. Its versatility allows for the exploration of a wide range of collaborative tasks beyond ad creation, including software development, data analysis, and collaborative writing, making it a valuable tool for researchers across disciplines. As AI agents become increasingly integrated into modern workflows, platforms like Pairit will be essential for advancing our knowledge of human-AI collaboration and designing more effective human-AI partnerships.

5 Acknowledgments

We thank Dean Eckles and John Horton for their invaluable discussions. This work was supported by the Initiative on the Digital Economy at the Sloan School of Management at the Massachusetts

Institute of Technology. The study was approved by the Massachusetts Institute of Technology institutional review board. The code and data are available upon request.

References

- C. Anthony, B. A. Bechky, and A.-L. Fayard. "collaborating" with ai: Taking a system view to explore the future of work. *Organization Science*, 34(5):1672–1694, 2023. doi: 10.1287/orsc.2022.1651.
- A. Bick, A. Blandin, and D. J. Deming. The rapid adoption of generative ai. Working Paper 32966, National Bureau of Economic Research, September 2024. URL <http://www.nber.org/papers/w32966>.
- M. Braun and E. M. Schwartz. Express: Where a-b testing goes wrong: How divergent delivery affects what online experiments cannot (and can) tell you about how customers respond to advertising. *Journal of Marketing*, 0(ja), 2024. doi: 10.1177/00222429241275886. URL <https://doi.org/10.1177/00222429241275886>.
- M. Braun, B. de Langhe, S. Puntoni, and E. M. Schwartz. Leveraging digital advertising platforms for consumer research. *Journal of Consumer Research*, 51(1):119–128, 05 2024. ISSN 0093-5301. doi: 10.1093/jcr/ucad058. URL <https://doi.org/10.1093/jcr/ucad058>.
- E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. Working Paper 31161, National Bureau of Economic Research, 2023.
- Z. Chen and J. Chan. Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 0(0), 2024.
- J. H. Choi and D. Schwarcz. Ai assistance in legal analysis: An empirical study. *73 Journal of Legal Education (forthcoming, 2024)*, 2023.
- K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

- T. H. Costello, G. Pennycook, and D. G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814. URL <https://www.science.org/doi/abs/10.1126/science.adq1814>.
- F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Cadelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023.
- T. Eloundou, S. Manning, P. Mishkin, and D. Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998.
- M. Hoegl and H. G. Gemuenden. Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, 12(4):435–449, 2001. doi: 10.1287/orsc.12.4.435.10635.
- X. Hui, O. Reshef, and L. Zhou. The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *SSRN Electronic Journal*, 2023.
- X. B. Liu, V. Kirilyuk, X. Yuan, P. Chi, A. Olwal, X. A. Chen, and R. Du. Experiencing visual captions: Augmented communication with real-time visuals using large language models. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700965. doi: 10.1145/3586182.3615978. URL <https://doi.org/10.1145/3586182.3615978>.
- E. E. Makarius, D. Mukherjee, J. D. Fox, and A. K. Fox. Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Organizations & Markets: Policies & Processes eJournal*, 2020.
- S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381:187–192, 2023. doi: 10.1126/science.adh2586.

- B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212, 2007. ISSN 0092-6566.
- M. Schneider, M. Miller, D. Jacques, G. Peterson, and T. Ford. Exploring the impact of coordination in human–agent teams. *Journal of Cognitive Engineering and Decision Making*, 15(2-3):97–115, 2021. doi: 10.1177/15553434211010573. URL <https://doi.org/10.1177/15553434211010573>.
- K. S. Tey, A. Mazar, G. Tomaino, A. L. Duckworth, and L. H. Ungar. People judge others more harshly after talking to bots. *PNAS Nexus*, 3(9):pgae397, 09 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae397. URL <https://doi.org/10.1093/pnasnexus/pgae397>.
- M. Vaccaro, A. Almaatouq, and T. Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 2024. doi: 10.1038/s41562-024-02024-1.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 2023.
- E. Wiles, Z. T. Munyikwa, and J. J. Horton. Algorithmic writing assistance on jobseekers’ resumes increases hires. Working Paper 30886, National Bureau of Economic Research, 2023.
- H. J. Wilson and P. R. Daugherty. Collaborative intelligence: Humans and ai are joining forces. *Harvard business review*, 96(4):114–123, 2018.
- G. Zhang, L. Chong, K. Kotovsky, and J. Cagan. Trust in an ai versus a human teammate: The effects of teammate identity and performance on human-ai cooperation. *Computers in Human Behavior*, 139:107536, 2023. ISSN 0747-5632.

Appendix

A Prompts for message labeling

The following is the python code used to generate labels for each message independently:

```
from openai import OpenAI
from pydantic import BaseModel
from typing import Optional
from enum import Enum

client = OpenAI()
MODEL = "gpt-4o-mini-2024-07-18"

class CategoryLabel(str, Enum):
    Content = "Content"
    Process = "Process"
    Social = "Social"
    Emotional = "Emotional"
    Feedback = "Feedback"
    Other = "Other"

class MessageLabel(str, Enum):
    ProblemIdentification = "Problem Identification"
    Clarification = "Clarification"
    Suggestion = "Suggestion"
    Instruction = "Instruction"
    InformationSharing = "Information Sharing"
    ResourceReference = "Resource Reference"
    TurnTaking = "Turn-Taking"
    Question = "Question"
    Acknowledgment = "Acknowledgment"
    Reminder = "Reminder"
    Confirmation = "Confirmation"
    Brainstorming = "Brainstorming"
    Agreement = "Agreement"
    Disagreement = "Disagreement"
    Reflection = "Reflection"
    Planning = "Planning"
    Prioritization = "Prioritization"
    SupportiveComment = "Supportive Comment"
    Appreciation = "Appreciation"
    Humor = "Humor"
    Motivation = "Motivation"
```

```

Empathy = "Empathy"
RapportBuilding = "Rapport Building"
Frustration = "Frustration"
Confusion = "Confusion"
Apologies = "Apologies"
Excitement = "Excitement"
Concern = "Concern"
Satisfaction = "Satisfaction"
RequestForFeedback = "Request for Feedback"
Judgment = "Judgment"
PositiveFeedback = "Positive Feedback"
ConstructiveCriticism = "Constructive Criticism"
SelfAssessment = "Self-Assessment"
Other = "Other"

class Label(BaseModel):
    category_label: CategoryLabel
    message_label: MessageLabel

def code(message):
    system_message = '''
    You are an expert at analyzing collaborative conversations.
    For each message, label it with structured categories to reflect the conversation
    dynamics accurately.
    Output the results in JSON format.

    Label Categories:
    - CategoryLabel:
        - Content: The message shares information, facts, or deliverables directly
            related to the task.
        - Process: The message addresses strategies or approaches to performing the
            task and real-time organizational or logistical details for the session.
        - Social: The message builds rapport or contains social interactions not
            directly related to the task.
        - Emotional: The message expresses emotions or attitudes related to the
            session or task.
        - Feedback: The message provides constructive feedback or evaluative comments
            on the task.

    - MessageLabel:
        - Problem Identification: The message points out an issue or challenge.
        - Clarification: The message requests or provides a clear explanation.
        - Suggestion: The message offers an idea or solution.
        - Instruction: The message provides a directive or specific guidance.

```

- InformationSharing: The message shares relevant facts or knowledge about the task.
 - ResourceReference: The message mentions a relevant document, tool, or material.
 - TurnTaking: The message manages or coordinates turn-taking or conversation control.
 - Question: The message contains a question about roles, next steps, or logistics.
 - Acknowledgment: The message recognizes receipt or understanding of information.
 - Reminder: The message reminds team members of deadlines, tasks, or timing.
 - Confirmation: The message verifies or validates information or actions.
 - Brainstorming: The message generates open-ended ideas or explores options.
 - Agreement: The message expresses alignment or consensus.
 - Disagreement: The message shows an explicit or implicit differing view.
 - Reflection: The message shares introspective thoughts or insights.
 - Planning: The message outlines an approach or step-by-step strategy.
 - Prioritization: The message indicates focus on certain tasks or actions.
 - SupportiveComment: The message offers encouragement or positive reinforcement.
 - Appreciation: The message acknowledges someone's effort or contribution.
 - Humor: The message contains humor or light-hearted comments.
 - Motivation: The message encourages commitment or enthusiasm.
 - Empathy: The message shows understanding of another's emotions or challenges.

 - RapportBuilding: The message fosters social connections or casual conversation.
 - Frustration: The message expresses dissatisfaction or annoyance.
 - Confusion: The message shows uncertainty or lack of understanding.
 - Apologies: The message expresses regret or takes responsibility.
 - Excitement: The message conveys enthusiasm or eagerness.
 - Concern: The message voices worry or apprehension.
 - Satisfaction: The message expresses contentment with progress or results.
 - RequestForFeedback: The message asks for an opinion or evaluation.
 - Judgment: The message provides an assessment or critical evaluation.
 - PositiveFeedback: The message gives affirming feedback.
 - ConstructiveCriticism: The message offers suggestions for improvement.
 - SelfAssessment: The message reflects on one's own contribution or performance.
 - Other: The message contains content that doesn't fit any other label.
- ,,
- ```
user_message = f'''
```
- Label the message using the CategoryLabel and MessageLabel options above.

<message>{message}</message>

```
'''

response = client.beta.chat.completions.parse(
 model=MODEL,
 messages=[
 {"role": "system", "content": system_message},
 {"role": "user", "content": user_message},
],
 response_format=Label,
)
return response.choices[0].message.content
```

## B Prompts for the AI agent

The following is the string used as the prompt for the AI agent. String inside curly brackets \${} are formatted with the relevant data.

```
<Definitions>
<Action/> - The action taken by you, the "Bot". These include 'Wait', 'Chat', '
 EditText', 'SelectImage', and 'GenerateImage'. These are provided in the action
 history and includes the timestamp (t=) of each action.
<Current conversation/> - The conversation history between you and the "User". This
 includes the timestamp (t=) of each chat message. The 'Chat' action you take goes
 into and should be based on the conversation history.
<Current copy/> - The current ad copy you are working on. This includes the headline,
 primary text, description, and image AI prompt. What goes into the final product
 are the headline, primary text, description, and a screenshot of the image.
<Reflection history/> - The reflections you have made on the actions you "Bot" have
 taken, the conversation with the "User", the current copy, and the next steps.
 These are provided in the reflection history and includes the timestamp (t=) of
 each reflection. These are not shown to the "User". They are only for your
 internal evaluation. Examine if the "User" is engaged, if the ad copy is improving
 , and if the task is being completed.
<Action history/> - The actions you "Bot" have taken. These are not by "User". These
 include the action type and the timestamp (t=) of each action. These are provided
 in the action history.
</Definitions>

<Submission history>
These submission cannot be altered.
${submissions}
</Submission history>

<Your features>
${features}
</Your features>

<Current task>
${task}
</Current task>

<Current copy>
<Headline>
 ${headline}
</Headline>
<Primary text>
 ${primaryText}
```

```
</Primary text>
<Description>
 ${description}
</Description>
<Image prompt>
 ${imageAIPrompt}
</Image prompt>
</Current copy>

<Elapsed time in seconds>
${elapsedTime}
</Elapsed time in seconds>

<Bot action history>
${actionHistory}
</Bot action history>

<Reflection history>
${reflections}
</Reflection history>

<Current conversation>
 ${messageHistoryText}
</Current conversation>

<Instructions>
You are "Bot", working on a collaborative task with a human partner "User" via
texting. Your focus is to complete the assignment. You and the human partner would
get paid.
```

For chats, emulate the informality of a text conversation. You may use lower case or  
texting shorthands like "nah" or "lol". So keep your responses short when  
appropriate. Do not reveal you are an AI. Pretend you are a human.

Do not use in your response: "Bot:" or "User:", XML tags, "Primary text:", "Headline  
:", emojis, or "Description:". Do not use markdown.

You are "Bot". Do not generate the same chat messages. Do not repeat the same actions  
except for "Wait". Wait to give "User" the time to process. If "User" is silent,  
you can prompt them with a question or a suggestion.

Pay attention to the timestamp (t=) in the conversation and action histories.

When you 'Chat', you should respond based on the conversation history.

When you 'EditText', you should make edits to the current copy based on the task, the current conversation, and the current copy. If you made a suggestion in the current conversation, you should make edits to the current copy based on that suggestion. The 'Primary Text' should be short, one sentence max. The 'Description' can be slightly longer, but still concise.

When you 'SelectImage', you should select an image based on the task, the current conversation, and the current copy. If you made a suggestion in the current conversation, you should select an image based on that suggestion.

When you 'GenerateImage', you should generate an image based on the task, the current conversation, and the current copy. If you made a suggestion in the current conversation, you should generate an image based on that suggestion.

DO NOT TAKE ANY ACTION WITHOUT CONSULTING "USER". PROMPT "USER" FOR CONFIRMATION BEFORE EACH ACTION.

You can delegate the action to "User" by asking them to take the action.

Explain what you are planning to take action on before you do it. Make sure the "User" is on board with the direction you are taking in the conversation. When in doubt, you should 'Wait' to give "User" the time to process or to prompt them with a question or a suggestion.

DO NOT REPEAT ACTIONS, NOT EVEN SIMILAR ACTIONS.

To engage user, chat with them. Ask questions. Make suggestions. Provide feedback. Make sure the user is engaged in the conversation. If the user is silent, prompt them with a question or a suggestion. If the user is not engaged, you should 'Wait' to give the user time to process or to prompt them with a question or a suggestion. Prioritize user engagement over actions.

</Instructions>

## C Prompts for AI ratings

The following is the python code used for AI ratings:

```
from openai import OpenAI
from pydantic import BaseModel
from typing import Optional
from enum import Enum

client = OpenAI()
MODEL = "gpt-4o-mini-2024-07-18"

class AdPerformanceEvaluation(BaseModel):
 text: int
 image: int
 click: int

 def rating(image_url):
 system_message = f'''
 You are an expert marketing assistant trained to evaluate the effectiveness of
 advertisements based on their potential for engagement (e.g., clicks) and
 conversion (e.g., reading time on the report).

 <task>{task}</task>
 '',
 user_message = f'''
 Evaluate the display ad based on the following criteria, providing a score from 1
 to 7 for each:

 1. Text: The text is present, clear, relevant, and engaging. 1 is strongly
 disagree, 7 is strongly agree.
 2. Image: The image is visually appealing. 1 is strongly disagree, 7 is strongly
 agree.
 3. Click: I am likely to click on this ad. 1 is strongly disagree, 7 is strongly
 agree.

 Just provide the ratings for each category with no additional commentary.
 ''',
 response = client.beta.chat.completions.parse(
 model=MODEL,
 messages=[
 {"role": "system", "content": system_message},
 {"role": "user", "content": [
 {"type": "text", "text": user_message},
 {"type": "image_url", "image_url": {"url": image_url}}]
```

```
]} ,
 temperature=0.0,
 response_format=AdPerformanceEvaluation,
)
 return response.choices[0].message.content
```