Collaborating with AI Agents:

Field Experiments on Teamwork, Productivity, and Performance

Harang Ju, Sinan Aral

Massachusetts Institute of Technology

harang@mit.edu, sinan@mit.edu

March 25, 2025

Abstract

To uncover how AI agents change productivity, performance, and work processes, we introduce MindMeld: 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, with randomized AI personality traits. 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 137% and allowed humans to focus 23% more on text and image content generation messaging and 20% less on direct text editing. Humans on Human-AI teams sent 23% fewer social messages, creating 60% greater productivity per worker and higher-quality ad copy. In contrast, human-human teams produced higher-quality images, suggesting that AI agents require fine-tuning for multimodal workflows. AI personality prompt randomization revealed that AI traits can complement human personalities to enhance collaboration. For example, conscientious humans paired with open AI agents improved image quality, while extroverted humans paired with conscientious AI agents reduced the quality of text, images, and clicks. In field tests of ad campaigns with ~5M impressions, ads with higher image quality produced by human collaborations and higher text quality produced by AI collaborations performed significantly better on click-through rate and cost per click metrics. Overall, ads created by human-AI teams performed similarly to those created by human-human teams. Together, these results suggest AI agents can improve teamwork and productivity, especially when tuned to complement human traits.

1 Introduction

Generative artificial intelligence (AI) tools have garnered attention for their potential to improve productivity and performance. Not only do they have far-reaching impacts across industries (Eloundou et al., 2024; Bick et al., 2024), but they also increase measurable productivity statistics, like output per unit time, as well as creativity and measurable quality. Researchers have found positive productivity effects of AI tools, with evidence of heterogeneity across tasks and baseline performance. For example, large language models (LLMs) decreased the average time taken for mid-level professional writing tasks by 40% and increased 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 5,000 papers, Vaccaro et al. (2024) shows that human-AI groups outperform humans alone in 85% of the studies.

This study addresses three critical gaps in understanding AI's impact on productivity and collaboration: 1) the lack of understanding of the role of AI agents, 2) the dearth of evidence concerning the mechanisms behind productivity improvements and how work processes change when humans collaborate with AI, and 3) whether and when model prompt engineering can improve the complementarity between AI agents and human labor. First, while the current literature, such as Dell'Acqua et al. (2023) and Chen and Chan (2024), reveal the productivity effects of AI by randomizing access to LLM chatbots, they are not multimodal, do not include context, do not allow the chatbots to take independent actions or use APIs to call outside of the platform, 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 have all these features, yet the existing scientific literature studies none of them. Moreover, while Liu et al.

(2023) explores levels of AI proactivity, which is key to understanding the effects of AI agents, it is a demonstration paper without a randomized controlled trial (RCT) measuring productivity effects. Similarly, Chen and Chan (2024) examines artificial workflows that do not reflect real-world settings.

Second, we lack insight into how work processes and collaboration patterns change when humans collaborate with AI agents compared to human-only teams and how these changes affect productivity and performance. Existing research has focused on the productivity effects of GPT chatbots (Noy and Zhang, 2023; Dell'Acqua et al., 2023). Separately, 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 collaboration settings, 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 in collaborative workspaces in addition to chat interfaces.

Third, prompts are a critical factor in LLM performance and have been studied extensively (Schulhoff et al., 2024). Indeed, the seminal paper introducing the GPT-3 model highlights that much of its power lies in its ability to learn and adapt to the context provided within the prompt, allowing it to generate coherent text without task-specific fine-tuning (Brown et al., 2020). Since then, other prompting methods, such as chain-of-thought prompting (Wei et al., 2023), have been shown to significantly improve task performance. However, little is known about how different prompting strategies interact with humans in collaborative workflows. In the human-robot interaction literature, studies have shown that factors like a robot's social communication and appearance influence the effectiveness of human-robot teams (Pamela J. Hinds and Jones, 2004; Jung et al., 2013; Jung and Hinds, 2018). For GPT-based agents, prompt engineering can be similarly critical; thus, it is essential to understand prompt effects on human-agent collaboration. Rigorous randomized controlled trials (RCTs), such as Jakesch et al. (2023), are needed to systematically evaluate the effects of prompts on human-AI collaboration and task outcomes.

To address these gaps, we developed MindMeld, a novel experimental platform designed to study human-AI collaboration in real-world, extensible tasks. MindMeld introduces several key innovations. It enables real-time collaboration between humans and AI agents, allowing participants to manipulate text, images, and workflows collaboratively in a chat-enabled workspace that mirrors existing online collaboration workflows. The platform supports randomized pairings of humans and AI (*i.e.*, Human-Human or Human-AI pairs) and allows for randomization of prompts. Critically, the AI can perform an equivalent set of actions that humans can on the platform. These include sending chat messages, editing the copy, selecting images, and generating new images using an external call to Dall-E 3. Thus, MindMeld captures every time-stamped keystroke, message, edit, API call, and intermediate output, providing a rich dataset that allows the detailed reconstruction of collaboration workflows. This represents a fundamental departure from the existing literature, which enables RCT-based evaluation of chatbots and co-pilots (LLMs) but does not enable randomized experiments analyzing human collaboration with AI agents.

We conducted a large-scale study using MindMeld to examine the effects of human-AI collaboration on ad design—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. As research on AI agents at this level of capability is relatively new, our randomization of human-AI teams serves as a bundle intervention, since AI agents exhibit many behaviors different from those of humans, with which we can use for further hypothesis generation. AI agents, built on GPT-4o, were further randomized by prompts designed to induce high or low levels of Big Five personality traits, enabling us to investigate how personality alignment influences collaboration, productivity, and quality. Teams worked collaboratively to create marketing campaigns for a real organization's year-end annual report, including generating ad images, writing text copy, and editing headlines. This process was fully recorded, resulting in 11,138 ads created, 183,691 messages exchanged, 1,960,095 text edits, 63,656 image edits, and 10,375 AI-generated images, offering an unprecedented level of detail to analyze teamwork dynamics.

Once the lab portion of the experiment was completed, we turned to the field evaluation. We first obtained human and AI quality ratings of the ads, including the overall quality of the ads, the quality of the images, the ad copy, the headlines, and the self-reported or AI-evaluated likelihood of consumer engagement with the ads (click-through rates). The ad copy was evaluated by human and AI raters on clarity, conciseness, grammar mistakes, persuasiveness, relevance to task, tone, and voice, the strength of the call to action, and the degree to which the ad was "attention-grabbing." We then ran the ads online, on X, generating over 4.9 million impressions, and evaluated click-through rates, cost-per-click, and view-through rates on the annual report, using DocSend, which allows us to see how much of the report consumers read, page by page, after clicking through on the ads.

Building on the capabilities of MindMeld, this study investigates how AI agents reshape collaboration dynamics, productivity, and performance compared to human-only teams. While prior work has shown that AI tools enhance productivity and reduce task completion times (Noy and Zhang, 2023; Brynjolfsson et al., 2023), these studies often treat AI as a passive tool rather than an active collaborator. As AI agents become integral to workflows, researchers are beginning to explore their roles as counterparts in collaborative systems rather than mere tools or mediums, emphasizing the importance of trust, transparency, and integration in human-AI partnerships (Makarius et al., 2020; Anthony et al., 2023; Collins et al., 2024). We hypothesize that AI agents, by reducing social coordination costs and enabling participants to focus more on content generation, will enhance individual productivity and lead to distinct communication patterns compared to human-human teams. Additionally, we investigate how the alignment between human personality traits and AI-induced traits influences collaboration outcomes. By randomizing AI prompts to induce high or low levels of Big Five personality traits, we examine how personality compatibility affects productivity, creativity, and collaboration quality. For instance, we hypothesize that high-openness pairings between humans and AI agents will lead to greater idea generation. These questions address critical gaps in understanding how AI agents interact with humans in multimodal workflows and how their design can be optimized for collaborative success.

Our findings reveal profound differences in how AI reshapes teamwork and productivity. We found that collaborating with AI agents significantly reshaped teamwork dynamics and enhanced productivity and performance compared to the human-only teams. In mining the rich collaboration data to understand how AI changes teamwork, we found Human-AI teams sent 45% more messages than Human-Human teams, revealing increased communication. The results were also consistent at the individual level: humans in Human-AI teams messaged more than those in Human-Human teams. Furthermore, both AI and humans in Human-AI teams sent more *content-* and *process*-oriented messages about their collaboration than those in Human-Human teams, especially messages containing suggestions, instructions, prioritization, judgment, and planning. Conversely, humans in Human-Human teams sent more messages that were *social* and *emotional*—including more messages that expressed rapport building, self-assessment, and concern—than humans in Human-AI teams. These results begin to unpack how work changes when humans collaborate with AI agents compared to when they collaborate with humans. We also tracked the work done on the collaborative workspace where humans and agents could both collaboratively edit the ad copy as well as generate images using external APIs (*i.e.*, Dall-E 3). We found that Human-AI teams made 84% fewer edits to the copy due to LLM's proficiency in writing high-quality ad copy. In contrast, LLMs were worse at predicting image quality, thus disadvantaging human-AI teams in the visual dimensions of ad creation and leading to human-AI teams producing 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 effort, including such coordinating activities like rapport building, than human participants in Human-Human teams.

The implications of these collaboration differences are then evident in the productivity and performance differences. Human-AI teams produced as many ads as Human-Human teams and 70% more ads on the individual level. First, we found in AI evaluations by `gpt-4o-mini` that ads created in Human-AI teams scored higher on the quality of the text but equally on image quality. Interestingly, human evaluations (using a separate set of 1,195 Prolific workers) also scored text quality higher but lower on image quality. Both human and AI evaluations are tied to the estimated likelihood of click-through rates. In addition to randomizing whether humans were paired with AI, we also orthogonally randomized a set of agent prompts designed to induce AI with either high or low levels of each of the Big Five personality traits. Through prompt randomization, we found heterogeneous prompt effects: a high or low personality prompt had differential effects on our outcomes depending on the personality of the human participant.

Overall, these findings suggest that AI's involvement in collaborative settings can drive productivity and performance by enabling participants to focus more on the content, possibly by reducing the social coordination costs of collaboration. Importantly, these effects vary across the personality of human collaborators, which means AI agents can be tuned to "fit" the personality of human collaborators to improve productivity and performance. By combining the precise manipulability of lab experiments with the rigor and real-world measurability of field studies, this study provides novel insights into the dynamics and potential of collaborating with AI agents.

2 Methods

Our study was preregistered and deemed minimal risk and exempt by the MIT Committee on the Use of Humans as Experimental Subjects (protocol E-5927).[1] Any non-pre-registered analyses are labeled *post hoc*.

2.1 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 1A). In the Human-Human condition, a participant joined a queue until another participant also

[1] See osf.io/jfzha and osf.io/95dhu.

joined the queue, 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.

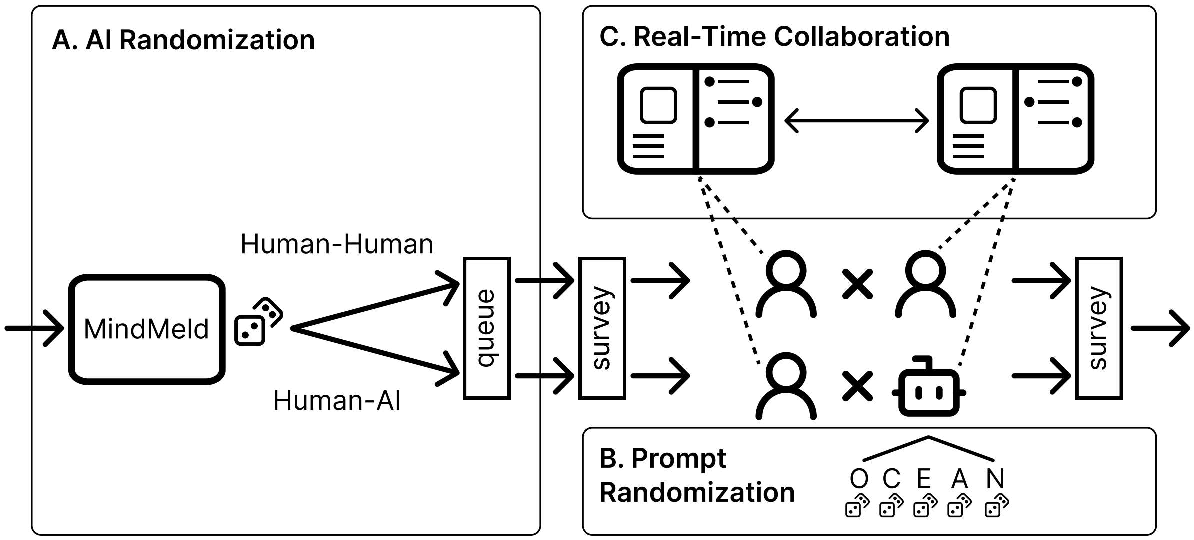

Figure 1: Overview of methods. (A) Participants are randomized into collaborating with another participant or an AI agent. (B) AI agents are assigned a personality profile based on Big Five traits, with each trait randomly set to either a high or low level. (C) Participants collaborate with another participant or an AI agent to produce ads in a real-time collaborative workspace.

Personality prompt randomization. In addition to the context of the task and chat, we prompt the model with personality prompts for the Big Five personality traits (McCrae and John, 1992). For each trait, we randomize whether the personality is high ($p = 0.5$) or low ($p = 0.5$; Figure 1B). To induce high or low personality traits, we use prompts generated by Jiang et al. (2023) using P^2 prompting, which uses a chain of prompts to generate a detailed description of individuals with the traits. The prompts are available on their GitHub repository.

Pre-task survey. After participants were paired with each other, they answered a 10-item survey, each 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 the pre-task survey, they entered the collaborative workspace and chat interface in which they could message each other and edit and submit ads (Figure 1C). The edits were synchronized and messages were transmitted between the participants in real-time using websockets (*i.e.*, tiptap.dev, pusher.com), as in commercial collaboration tools (*e.g.*, Google Docs) and chat applications (*e.g.*, Slack). The participants had 40 minutes to submit and could submit zero to 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 Gemuenden, 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. The excluded items include the following: "The team members communicated often in spontaneous meetings, phone conversations, etc.", "The team members communicated mostly directly and personally with each other.", and "There were mediators through whom much communication was conducted."

In addition, we asked four questions regarding the perception of AI, all on a 7-point Likert scale. They include two on 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."). The last two questions include one on their perception of their partner as an AI (*i.e.*, "I believe my partner was an AI during the task.") and the other in which we revealed the identity of their partner and 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?"). Only 1964 participants completed the post-task survey.

2.2 The MindMeld platform

To study the impact of collaborating with AI agents, we developed an online platform which we call *MindMeld* (Figure 2). Once a participant enters the platform, the participant is randomized into a queue for either a Human-Human or Human-AI condition. In the Human-Human condition, the participant collaborates with another participant; in the Human-AI condition, the participant collaborates with an AI agent. As far as we know, this randomization is a unique feature of our platform.

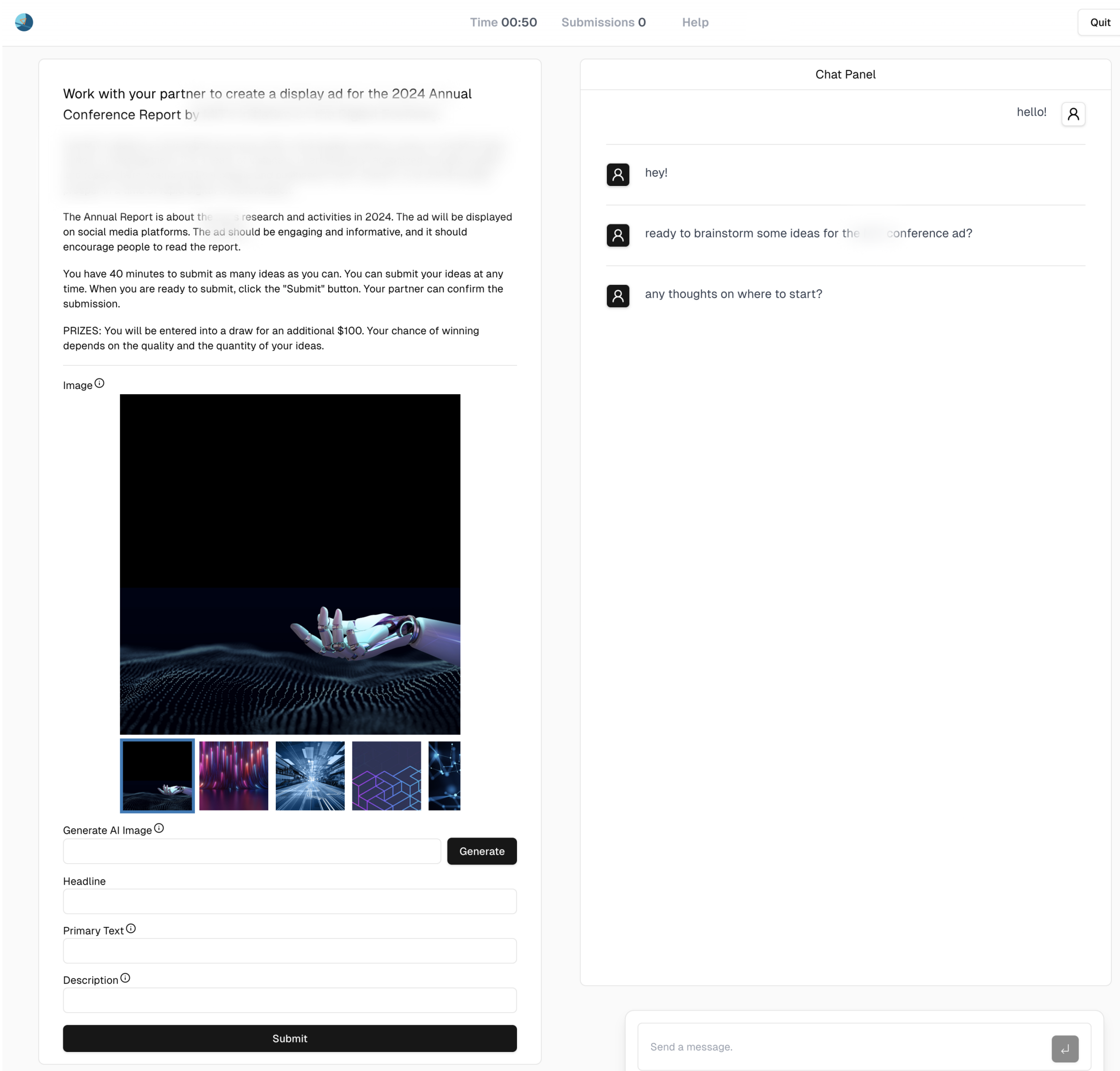

Figure 2: The MindMeld platform. On the left is the task panel, and on the right is the chat panel. In the Human-Human condition, chat messages and edits on the task panel, including text edits, image selections, and AI image generations, are synchronized in real-time. In the Human-AI condition, the participant chats with an AI agent with full context of the user interface (UI; see Section 2.3), and the AI can edit text, select images, and generate AI images.

On the left side of MindMeld (Figure 2) is the task panel where participants can create, edit, and submit the ads. The platform includes a selection of images participants can choose from, and the participants can also access the Dall-E 3 application programming interface (API) to generate new images. The copy includes a headline, primary text, and a description. All of the edits—including the image selection and image generation as well as the copy edits—are synchronized in real-time across the participants. On the right is the chat panel in which participants can chat in real-time with either a participant or an AI agent. The participants can submit the ads, and the interface will reset so they can submit the next ad. To the best of our knowledge, our platform is the first to simulate real-time collaboration between human-human or human-AI pairs with real-time chat and editing on synchronized text and image interfaces.

2.3 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 neural network. 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 following: the same text of the task given to participants, previous submissions, personality prompts (see Section 2.1), current copy, elapsed time, history of its own actions (see Section 2.3), history of its chain-of-thought (see Section 2.3), chat history, and general instructions. Moreover, a screenshot is taken of the image after every change and included as input so the AI can observe the image. The prompt template is shown in Appendix B.

Actions. To ensure that the Human-Human and Human-AI conditions are comparable, the AI agent can take the same actions a human participant can except for submitting ads. The actions include the following: sending messages, editing any of the ad copy (*i.e.*, 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 behaviors of the model, such as repeatedly sending the same message, and were determined through trial and error. The chain-of-thought prompts are shown in Appendix B.

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 2.1 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%. A balance check found that our sample was balanced on all measured covariates (Table 1). This study was run from October 15 to 18, 2024, and took a median of 46.32 minutes to complete. The participants were paid 9 US dollars, and two participants were each awarded a bonus payment of $100 who won a lottery based on productivity and performance.

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., gender, age) and psychological traits based on the Big Five personality dimensions. As shown, no significant differences were detected, indicating successful randomization.

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

Covariate	*All*	*Human-AI*	*Human-Human*
Individuals	2310	1258	1052
Teams	1834	1258	576
Gender (% Male)	50.8%	50.4%	51.3%
Age	42 ± 14	43 ± 15	42 ± 14
Openness	0.71 ± 0.20	0.71 ± 0.21	0.72 ± 0.19
Conscientiousness	0.79 ± 0.18	0.78 ± 0.18	0.79 ± 0.17
Extraversion	0.56 ± 0.21	0.55 ± 0.21	0.56 ± 0.21
Agreeableness	0.70 ± 0.22	0.70 ± 0.21	0.69 ± 0.24
Neuroticism	0.48 ± 0.23	0.49 ± 0.24	0.48 ± 0.22

Table 1: Randomization check. Personality traits are normalized from a 7-point Likert scale.

2.6 Incentives

To incentivize participants to create high-quality ads, we informed them that eligibility for an additional $100 prizes would be based on both the quality and quantity of their ads, as well as their performance on social media platforms. 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 best-performing ads.

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. See Appendix A for the prompts used in this analysis.

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. We built a web application that populates the image, ad copy, shortened link, call-to-action, and other user

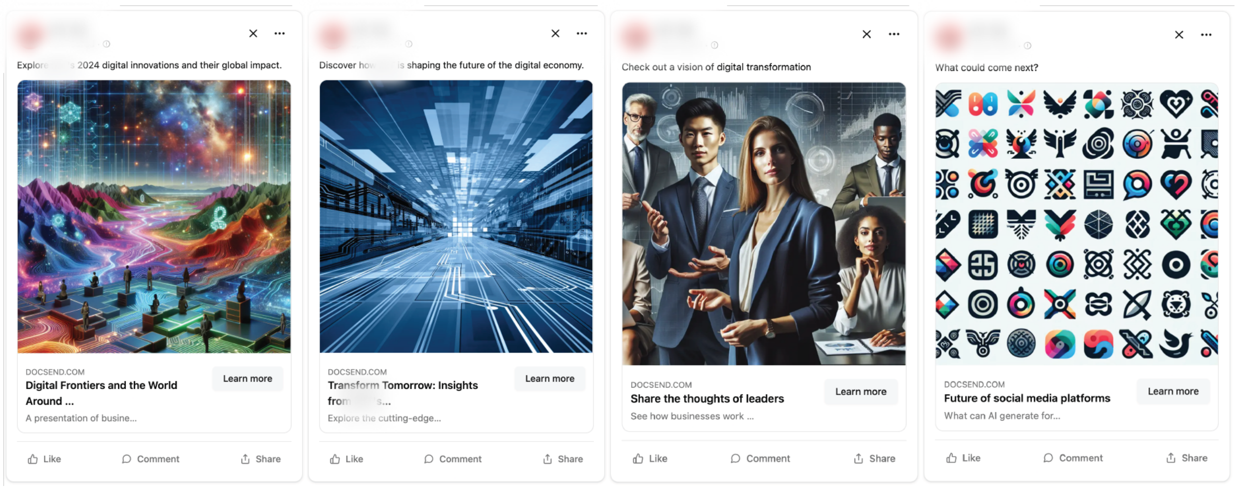

Figure 3: Examples of ad mockups.

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 ask the same questions as those given to human evaluations (see Section 2.9). Each item was on a 7-point Likert scale. The first question was "The text is present, clear, relevant, and engaging"; the second "The image is visually appealing"; and the third "I am likely to click on this ad." The exact prompts are shown in Appendix C.

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

[3]See OpenAI's documentation on vision and structured outputs.

submitting their surveys, resulting in a dropout rate of 0.42%. We used our custom platform for this survey, run on Google Cloud Platform's App Engine. The code is available on GitHub. This study was not a part of our pre-registration. This study was run 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 display ad, we asked the participants three questions regarding the quality of the text, the image, and the estimated clickthrough rate. Each item was on a 7-point Likert scale. The first question was "The text is present, clear, relevant, and engaging"; the second "The image is visually appealing"; and the third "I am likely to click on this ad." See Figure 4 for an example of an item on the survey.

2.10 Field evaluation of ad performance

To assess real-world advertising outcomes, we ran ad campaigns on X (formerly Twitter), 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 evidence of the impact of ads conditional on online algorithms. We do not use holdout in this study; each ad has

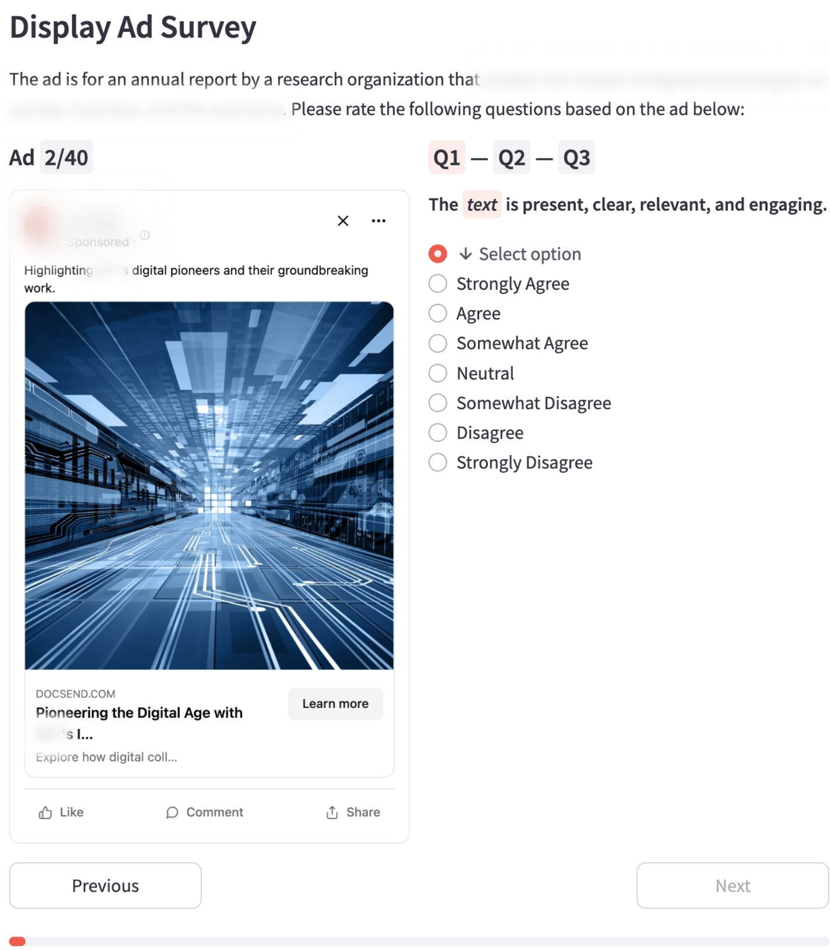

Figure 4: The user interface for the ad quality survey.

a unique DocSend link so we expect the outcome to be zero in a holdout group and will attribute all document visits to the ads.

To run as 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 click scores, ranging from low to high predicted click probability. The click scores were the human predictions of the likelihood of user clicks (see Section 2.9). We removed 8 ads that potentially violated moderation policies (*e.g.*, violence, sexual, drug) from the total set. We controlled for spend in our analyses to account for auto-bidding, which we had left on for our campaigns.

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).[4] 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. We measure the effects of human-AI collaboration using a standard regression model. For individual i, we use the following model:

$$Y_i = \delta \text{H-AI}_i + \sum_{p \in \text{Big Five}} \beta_p p_t + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i \tag{1}$$

where Y_t represents the outcome for team t, H-AI_i is 1 if individual i is working with an AI and 0 if individual i is working with another participant, p is a normalized score for each Big Five trait, and Gender_i is 1 if the participant is male and 0 if female. 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.

[4]The platform limit for split tests within each campaign is 5.
[5]The data is from https://data.census.gov/table/ACSDT5Y2020.B01003.

For team-level analysis for team t, we use the following model:

$$Y_t = \delta_t \text{H-AI}_t + \sum_{p \in \text{Big Five}} \beta_p \bar{p}_t + \sum_{i \in \text{Team}_t} \beta_{1i} \text{Age}_{ti} + + \sum_{g \in \{M,F\}} \text{Gender}_g, \tag{2}$$

where the outcome Y_t is the same as above, H-AI_t (Human-AI) is 1 if the team was Human-AI and 0 if the team was Human-Human, $\bar{p}_t$ is the average Big Five trait for both participants, and Gender_g is 1 if both are $g \in \{M,F\}$ and 0 if one participant is male and the other is female.

Prompt effects. In addition to the effects of human-AI collaboration, we measure the effects of prompts. For individual i in Human-AI teams (*i.e.*, $\text{H-AI}_i = 1$), we use the following regression model:

$$\begin{aligned} Y_i = & \sum_{P_{AI} \in \text{Big Five}} P_{AI} + \sum_{P_H \in \text{Big Five}} \beta_{P_H} P_{H,i} + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i \\ & + \sum_{P_{AI} \in \text{Big Five}} P_{AI} \times \left(\sum_{P_H \in \text{Big Five}} \beta_{P_{AI},P_H} P_{H,i} + \beta_{P_{AI},1} \text{Age}_i + \beta_{P_{AI},2} \text{Gender}_i \right) \end{aligned}$$

where the outcome Y_t is the same as above, P_{AI} is the orthogonally randomized, prompted high (*i.e.*, 1) or low (*i.e.*, 0) AI personality (Jiang et al., 2023).

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

$$Y_{tc} = \delta_t \text{H-AI}_{tc} + \text{Image}_{tc} + \text{Text}_{tc} + \text{Click}_{tc} + \text{Spend}_{tc} + u_c, \tag{3}$$

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 Teamwork

Collaborating with AI increases communication but decreases direct copy edits. To test how human-AI collaboration affects communication, we measured the number of chat messages sent by participants in both Human-Human and Human-AI groups. In the study, 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 as one does with chatbots. The MindMeld platform logs a timestamped record of each message sent by a participant or AI.

In our study, participants communicated significantly more in Human-AI teams than in Human-Human teams. Participants in Human-AI teams sent 45% more messages than those in Human-Human teams (Table 2). Interestingly, personality covariates were significantly associated with the number of messages sent in Human-AI teams: individuals scoring higher on openness and extraversion sent more messages, while those scoring higher on agreeableness sent fewer messages. These results indicate that collaborating with an AI partner encourages more frequent communication and that individual differences in personality traits, such as openness, extraversion, and agreeableness, further influence communication patterns in these interactions.

On the other hand, participants in Human-AI teams made significantly fewer direct copy edits compared to Human-Human teams, with a 60% decrease in edits (Table 3. The results on copy edits and messages suggest that collaboration with AI shifted participants' focus away from iterative textual refinements to edits through instructions and suggestions to the AI, as we will see further in the next 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 shift. 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), that artificially

	Messages	
Intercept	21.504***	9.977**
	(0.484)	(2.999)
Human-AI	13.669***	13.671***
	(0.834)	(0.834)
Openness		7.234**
		(2.175)
Conscientiousness		1.515
		(2.469)
Extraversion		4.721*
		(2.177)
Agreeableness		-4.144*
		(1.896)
Neuroticism		5.706**
		(1.902)
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.

constrained participants to specific modalities when interacting with AI, our results show that in more realistic collaborative settings with AI agents, participants naturally utilized both direct editing and interaction through messaging.

Human-AI teams focus on content and process over social and emotional communication. In addition to the number of messages sent across conditions, we investigate whether the types of messages sent vary across conditions. We used `gpt-4o-mini` to label each message independently with one of five *message categories*: Process, Content, Social, Emotional, Feedback, or Other (see Section 2.7 for more details). In our analysis, we found that both the AI and the individuals in Human-AI teams sent more content- and process-oriented messages, while Human-Human teams sent more social and emotional messages (Figure 5). This shift indicates that collaboration with AI emphasizes task-related communication over social interaction, possibly because participants can

	Copy Edits		*Image Selects*		*AI-Generated Images*	
Intercept	1367.225***	1630.869***	23.219***	23.447***	4.234***	5.921***
	(30.059)	(117.896)	(0.579)	(3.727)	(0.149)	(0.955)
Human-AI	-974.224***	-971.857***	7.432***	7.403***	0.401	0.420*
	(32.939)	(32.812)	(0.907)	(0.911)	(0.214)	(0.212)
Openness		185.814*		6.551**		1.877**
		(94.397)		(2.382)		(0.552)
Conscientiousness		-141.114		-1.662		-1.804*
		(96.456)		(2.667)		(0.717)
Extraversion		-21.398		-3.983		-0.065
		(83.181)		(2.523)		(0.552)
Agreeableness		-120.385		-2.113		-0.530
		(88.600)		(2.015)		(0.460)
Neuroticism		58.743		5.795*		0.496
		(75.416)		(2.636)		(0.558)
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.

focus more on the task without needing to navigate the social or emotional aspects of collaboration.

To further examine how communication differs across conditions, we labeled the messages with one of 36 *message intents* (see Appendix A for all labels). These intents 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 intents between the Human-Human and Human-AI groups in each message category (Table 4). See our model-free evidence in Figure 6. For example, in the Content category, Human-Human teams send more messages that show 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 messages about humor, concern, and apologies, while Human-AI teams send more messages about suggestions, appreciation, motivation, confirmation, and satisfaction. These findings further support the idea that Human-AI collaboration prioritizes

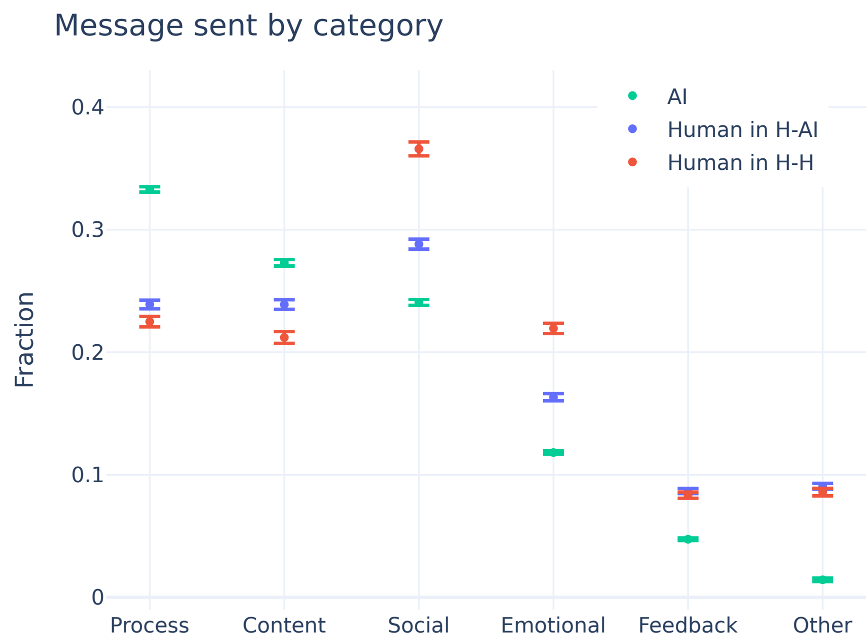

Figure 5: Participants in Human-AI teams send more process- and content-related messages while those in Human-Human teams send more social and emotional messages.

task-oriented communication, while Human-Human teams engage more in social and emotional exchanges.

Survey reports showed no differences in teamwork. 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.

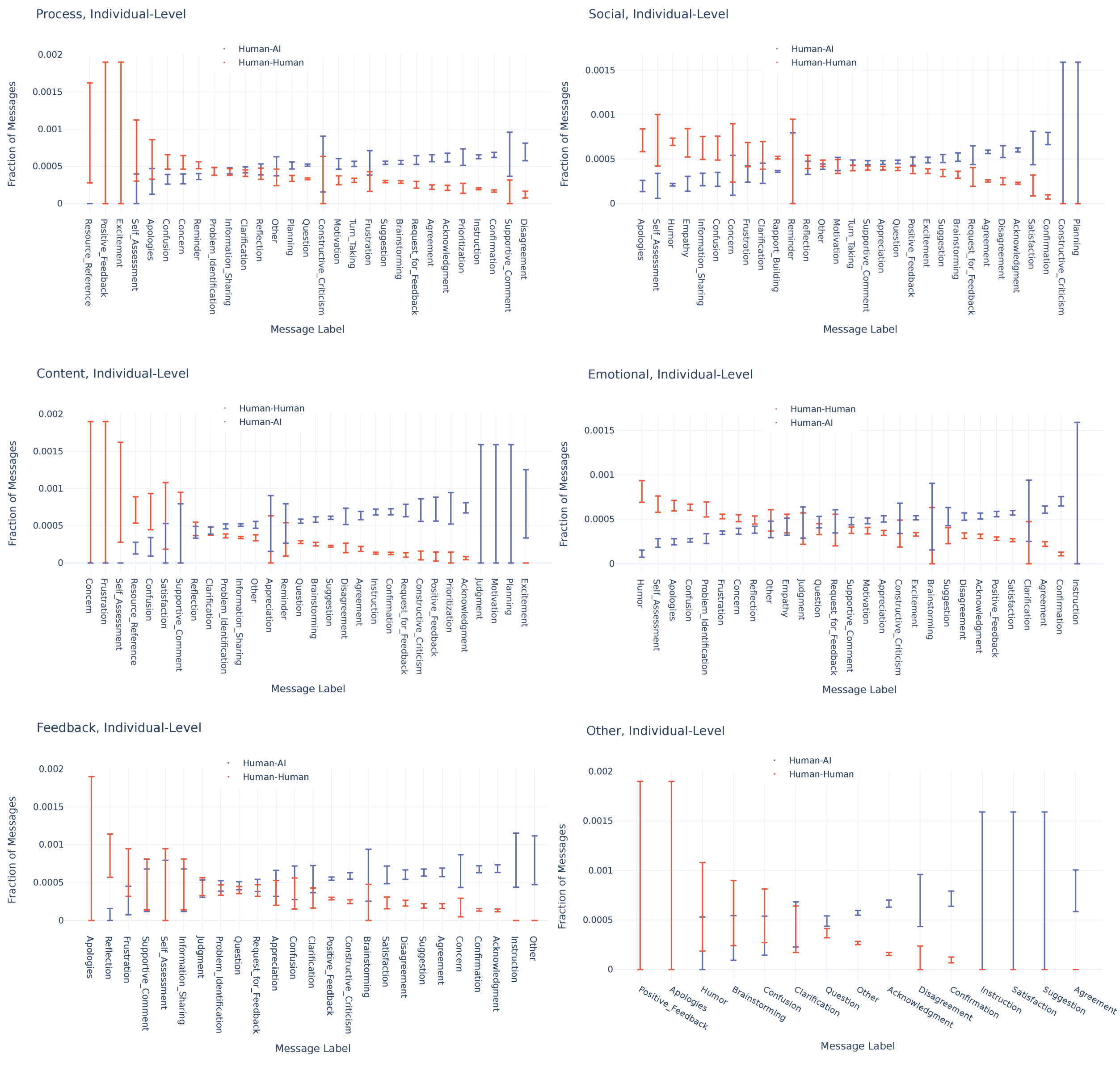

Figure 6: Message labeled by *intent* for each *category*.

	Content	*Process*	*Social*	*Emotional*	*Feedback*	*Other*
Intercept	0.156***	0.189***	0.415***	0.164***	0.053***	0.051***
	(0.025)	(0.021)	(0.028)	(0.017)	(0.010)	(0.011)
Human-AI	0.036***	0.025***	-0.085***	-0.045***	0.019***	0.018***
	(0.006)	(0.006)	(0.007)	(0.005)	(0.003)	(0.003)
Openness	0.024	-0.005	-0.021	0.016	-0.001	-0.018*
	(0.015)	(0.014)	(0.018)	(0.012)	(0.007)	(0.007)
Conscientiousness	-0.015	0.023	-0.010	-0.003	0.004	-0.006
	(0.018)	(0.016)	(0.022)	(0.013)	(0.007)	(0.008)
Extraversion	-0.006	-0.000	0.004	0.003	-0.012	0.006
	(0.016)	(0.015)	(0.017)	(0.013)	(0.007)	(0.007)
Agreeableness	0.028*	-0.014	0.008	-0.010	0.002	0.006
	(0.013)	(0.012)	(0.016)	(0.011)	(0.005)	(0.006)
Neuroticism	-0.007	-0.022	-0.020	0.033**	-0.007	-0.011
	(0.014)	(0.013)	(0.016)	(0.011)	(0.006)	(0.006)
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.

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 are near-substitutable with Human-Human teams with half as many individuals. On the MindMeld platform, participants were free to submit as many ads as they could produce within the time limit. On the team level, productivity was comparable between Human-Human and Human-AI teams (Table 5). However, on the individual level, participants in Human-AI teams individually submitted 60% to 73% as many ads as their counterparts in Human-Human teams. This suggests that AI collaboration supports individual productivity by enabling participants to generate more outputs compared to those in Human-Human teams.

	Submissions			
	Team-Level		*Individual-Level*	
Intercept	5.950***	7.483***	3.203***	3.935***
	(0.207)	(0.832)	(0.104)	(0.642)
Human-AI	-0.406	-0.395	2.341***	2.366***
	(0.246)	(0.246)	(0.169)	(0.169)
Personalities	No	Yes	No	Yes
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.

This focus on individual task completion is further reflected in ad copy completion rates. While completing tasks may seem like a minimal benchmark for productivity, it is particularly relevant for participants with lower performance levels, such as workers on platforms like Prolific. As shown in Table 6, with model-free evidence shown in Figure 7, participants in Human-AI teams had consistently higher completion rates for ad copy elements (e.g., headline, primary text, description) compared to participants in Human-Human teams. This finding suggests that AI collaboration provides support for individuals who might otherwise struggle to complete tasks, aligning with prior research on AI's role in enhancing performance for lower-performing participants.

3.3 Performance

We now evaluate the performance of Human-Human and Human-AI teams by assessing the quality of their outputs. Specifically, we examine differences in the quality of ad text and images and field evaluations of ad effectiveness to understand how collaboration dynamics influence overall effectiveness.

Ad text quality improves, but image quality declines in Human-AI teams. To determine the quality of ads, we use two *post hoc* measurements of ad quality: human evaluations and AI evaluations. For human evaluations, we recruited 1,195 participants from Prolific to rate the text

	Headline	*Primary Text*	*Description*
Intercept	0.728***	0.737***	0.651***
	(0.058)	(0.058)	(0.058)
Human-AI	0.189***	0.207***	0.199***
	(0.014)	(0.014)	(0.015)
Openness	0.006	0.009	0.019
	(0.034)	(0.035)	(0.036)
Conscientiousness	0.042	0.018	0.052
	(0.050)	(0.051)	(0.051)
Extraversion	0.003	-0.010	-0.017
	(0.036)	(0.036)	(0.038)
Agreeableness	0.017	-0.000	0.013
	(0.035)	(0.036)	(0.035)
Neuroticism	0.014	-0.009	-0.014
	(0.033)	(0.035)	(0.035)
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.

quality, the image quality, and the estimated likelihood of clicking on the ad (see Section 2.9 for details). The participants were shown a mockup of each ad as it would appear on a social media platform (Figure 3). In our AI evaluations, we used `gpt-4o-mini` to ask the AI model the same questions we ask the human evaluators (see Section 2.8 for details).

In our human evaluations of ad quality, we found that Human-AI teams had higher-quality text but lower-quality images than Human-Human teams (Table 7). We show model-free evidence in 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 the 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 the ads by Human-AI teams were created with the assistance of OpenAI's `gpt-4o`. However,

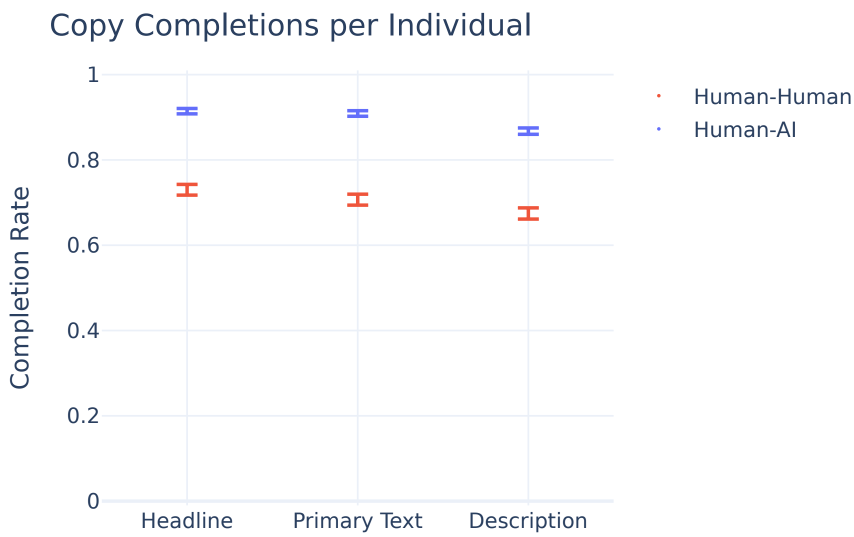

Figure 7: Copy completion rates.

human ratings show that AI introduces trade-offs: collaborating with GPT models improves text quality but reduces image quality. This trade-off is perhaps because GPT models were designed primarily for next-word prediction vs predicting image quality. These results suggest that while GPT models enhance text-focused tasks, their contributions to multimodal outputs like ads may require complementary tools or models designed specifically for image-related tasks. These findings also raise the question of whether and how these trade-offs in text and image quality meaningfully impact actual clickthrough rates.

3.4 Personality prompts

The second pre-registered set of randomization in our study focuses on AI personality traits. Specifically, we manipulate the Big Five personality traits for each AI, independently setting them to high or low levels using P^2 prompting (Jiang et al., 2023). This allows us to systematically investigate how AI personality traits influence collaborative work and whether there is heterogeneity in their effects based on the personality traits of the human collaborators, as measured through a pre-task survey. For more details on this approach, see Section 2.1.

	Human Evaluations			*AI Evaluations*		
	Text	*Image*	*Click*	*Text*	*Image*	*Click*
Intercept	4.535***	4.730***	3.592***	5.410***	6.438***	5.235***
	(0.070)	(0.070)	(0.072)	(0.056)	(0.054)	(0.045)
Human-AI	0.324***	-0.134***	-0.014	0.122***	-0.014	0.068***
	(0.018)	(0.018)	(0.019)	(0.015)	(0.014)	(0.013)
Openness	0.218***	0.111**	0.137**	0.153***	0.090**	0.111***
	(0.041)	(0.042)	(0.044)	(0.035)	(0.034)	(0.029)
Conscientiousness	-0.017	0.006	-0.015	0.018	0.100*	0.039
	(0.052)	(0.054)	(0.057)	(0.047)	(0.047)	(0.039)
Extraversion	-0.157***	-0.046	-0.034	-0.085*	-0.026	-0.027
	(0.044)	(0.044)	(0.047)	(0.038)	(0.035)	(0.032)
Agreeableness	0.032	-0.025	0.024	-0.008	-0.062*	-0.032
	(0.040)	(0.040)	(0.041)	(0.033)	(0.030)	(0.026)
Neuroticism	0.085*	0.034	0.047	0.033	0.028	0.016
	(0.041)	(0.041)	(0.042)	(0.033)	(0.031)	(0.026)
Sex[Male]	0.078***	0.000	0.026	0.016	0.003	-0.008
	(0.017)	(0.017)	(0.018)	(0.014)	(0.014)	(0.012)
Demographics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11138	11138	11138	11138	11138	11138

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

Table 7: AI and Human ratings of ads.

Teamwork. Examining the interaction effects of prompts with human personalities reveals significant heterogeneous prompt effects on collaboration outcomes (Table 8). First, on the level of communication measured by the number of messages, we see that a conscientious AI increases communication by a substantial amount, 62%, but only if the human counterpart is also high on conscientiousness. This intuitive result indicates a synergy between the conscientiousness of the AI and that of the human counterpart, where the alignment of traits, and not the traits individually, increases communication.

Table 8: Personality-heterogeneous prompt effects on collaboration outcomes on the individual level.

	Messages	*Copy Edits*	*Image Selects*	*AI-Generated Images*	*Submissions*
Intercept	28.033	-417.018	41.960*	5.430	5.356

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Table 8 continued

	Messages	*Copy Edits*	*Image Edits*	*AI-Image Generations*	*Submissions*
	(15.798)	(333.169)	(16.944)	(3.869)	(3.014)
Openness$_{AI}$	4.629	68.418	-1.154	-0.538	-0.329
	(12.011)	(264.775)	(13.691)	(3.166)	(2.096)
Openness$_{AI}\times$**Openness**$_{H}$	-5.612	-230.792	-9.978	-1.734	-1.488
	(7.145)	(165.372)	(7.789)	(1.575)	(1.224)
Openness$_{AI}\times$**Conscientiousness**$_{H}$	8.902	-21.331	5.713	0.679	-1.472
	(8.950)	(197.030)	(9.685)	(2.213)	(2.090)
Openness$_{AI}\times$**Extraversion**$_{H}$	-13.228	162.071	-3.790	0.882	-0.936
	(7.294)	(149.223)	(8.502)	(1.789)	(1.315)
Openness$_{AI}\times$**Agreeableness**$_{H}$	10.844	-108.768	3.001	0.104	3.273
	(7.951)	(179.521)	(8.084)	(1.701)	(1.676)
Openness$_{AI}\times$**Neuroticism**$_{H}$	-6.491	-3.857	2.923	1.468	0.418
	(6.918)	(164.836)	(9.042)	(1.940)	(1.264)
Conscientiousness$_{AI}$	-8.670	277.472	10.499	-0.912	0.149
	(11.251)	(260.988)	(14.708)	(3.102)	(2.433)
Conscientiousness$_{AI}\times$**Openness**$_{H}$	-0.992	-163.469	0.592	2.896	-0.367
	(6.861)	(165.094)	(7.629)	(1.571)	(1.265)
Conscientiousness$_{AI}\times$**Conscientiousness**$_{H}$	17.295*	-11.155	-7.208	-1.376	-0.625
	(8.408)	(194.647)	(9.875)	(2.177)	(2.302)
Conscientiousness$_{AI}\times$**Extraversion**$_{H}$	-8.598	156.372	-4.606	-2.429	1.338
	(7.318)	(159.277)	(8.211)	(1.764)	(1.314)
Conscientiousness$_{AI}\times$**Agreeableness**$_{H}$	-10.649	-94.053	1.922	0.059	0.257
	(7.475)	(198.311)	(8.107)	(1.676)	(1.442)
Conscientiousness$_{AI}\times$**Neuroticism**$_{H}$	2.633	-249.663	-9.956	0.221	0.257
	(6.636)	(159.721)	(9.050)	(1.854)	(1.343)
Extraversion$_{AI}$	-12.432	431.725	-18.305	-2.513	0.390
	(11.693)	(276.559)	(14.277)	(3.242)	(2.303)
Extraversion$_{AI}\times$**Openness**$_{H}$	-4.370	171.281	-0.090	-0.228	-0.156
	(7.059)	(157.808)	(7.788)	(1.627)	(1.265)
Extraversion$_{AI}\times$**Conscientiousness**$_{H}$	8.075	-262.434	10.594	2.489	-1.148
	(8.731)	(208.164)	(10.336)	(2.319)	(2.207)
Extraversion$_{AI}\times$**Extraversion**$_{H}$	3.410	-235.685	-0.867	-1.359	-0.371
	(7.330)	(155.943)	(7.779)	(1.725)	(1.298)
Extraversion$_{AI}\times$**Agreeableness**$_{H}$	3.789	-210.511	11.193	0.083	-0.572
	(8.016)	(181.480)	(8.002)	(1.645)	(1.444)
Extraversion$_{AI}\times$**Neuroticism**$_{H}$	7.986	-278.529	6.045	0.677	1.467
	(7.078)	(164.950)	(8.726)	(1.875)	(1.237)
Agreeableness$_{AI}$	0.741	270.288	0.796	2.037	-0.428
	(11.380)	(269.627)	(13.630)	(3.210)	(2.259)
Agreeableness$_{AI}\times$**Openness**$_{H}$	-2.466	-11.645	6.082	0.760	-1.105
	(7.018)	(162.741)	(7.579)	(1.618)	(1.286)
Agreeableness$_{AI}\times$**Conscientiousness**$_{H}$	-3.344	-40.849	-4.989	-2.589	-0.054
	(7.969)	(188.352)	(9.415)	(2.196)	(1.976)
Agreeableness$_{AI}\times$**Extraversion**$_{H}$	2.639	43.280	4.739	-0.681	3.115*
	(7.430)	(152.002)	(7.852)	(1.758)	(1.305)
Agreeableness$_{AI}\times$**Agreeableness**$_{H}$	5.759	-324.156	-2.213	0.479	-0.411
	(7.863)	(182.273)	(8.048)	(1.569)	(1.360)
Agreeableness$_{AI}\times$**Neuroticism**$_{H}$	-2.429	147.532	-1.327	-0.076	2.559*
	(6.710)	(163.766)	(9.180)	(1.886)	(1.228)
Neuroticism$_{AI}$	4.905	690.342**	-14.471	3.209	3.066
	(12.325)	(260.106)	(14.277)	(3.329)	(2.267)
Neuroticism$_{AI}\times$**Openness**$_{H}$	2.646	-106.095	22.124**	-0.818	-0.362

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Table 8 continued

	Messages	*Copy Edits*	*Image Edits*	*AI-Image Generations*	*Submissions*
	(7.055)	(159.805)	(7.701)	(1.651)	(1.302)
Neuroticism$_{AI}$ × **Conscientiousness**$_{H}$	-12.981	74.380	-4.622	-1.685	0.982
	(8.786)	(189.858)	(9.740)	(2.229)	(2.235)
Neuroticism$_{AI}$ × **Extraversion**$_{H}$	6.400	-136.372	11.654	2.640	0.746
	(7.252)	(154.798)	(8.068)	(1.736)	(1.305)
Neuroticism$_{AI}$ × **Agreeableness**$_{H}$	1.946	-435.239*	3.059	-1.308	-3.135
	(8.303)	(183.514)	(8.440)	(1.803)	(1.683)
Neuroticism$_{AI}$ × **Neuroticism**$_{H}$	1.862	-160.990	5.425	-1.292	-0.660
	(7.148)	(161.053)	(9.015)	(1.966)	(1.320)
Openness$_{H}$	14.867	393.999	-0.717	1.067	2.404
	(8.956)	(201.678)	(9.834)	(1.945)	(1.519)
Conscientiousness$_{H}$	-8.065	77.226	1.701	-0.158	1.052
	(11.145)	(263.025)	(12.535)	(2.768)	(2.722)
Extraversion$_{H}$	11.162	-41.462	-9.888	0.928	-1.894
	(9.525)	(189.557)	(9.505)	(2.163)	(1.680)
Agreeableness$_{H}$	-11.982	554.955*	-13.027	0.031	0.085
	(10.785)	(231.536)	(11.147)	(2.230)	(2.183)
Neuroticism$_{H}$	3.706	345.464	6.556	0.412	-2.405
	(9.506)	(229.219)	(10.524)	(2.197)	(1.810)
Demographics (+ interaction terms)	Yes	Yes	Yes	Yes	Yes
Observations	1258	1258	1258	1258	1258

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Next, we observed that AI neuroticism significantly influences edits on the display ad, including both copy and image selection. While a neurotic AI increased the number of copy edits, this effect reversed when the AI was paired with an agreeable human. For image selection, where participants choose from a set of images, a neurotic AI only increased the number of selections when collaborating with an open human. These findings underscore the nuanced ways in which AI and human personality traits interact to shape distinct facets of collaborative ad creation, such as editing and image selection.

Productivity. For productivity, we observed that AI agreeableness had significant heterogeneous effects on the number of submissions, on the individual level. While an agreeable AI by itself does not have any main effects, it increases the number of submissions by 2-3 ads when paired with either an extraverted or neurotic human. This highlights that distinct patterns also emerge for productivity outcomes.

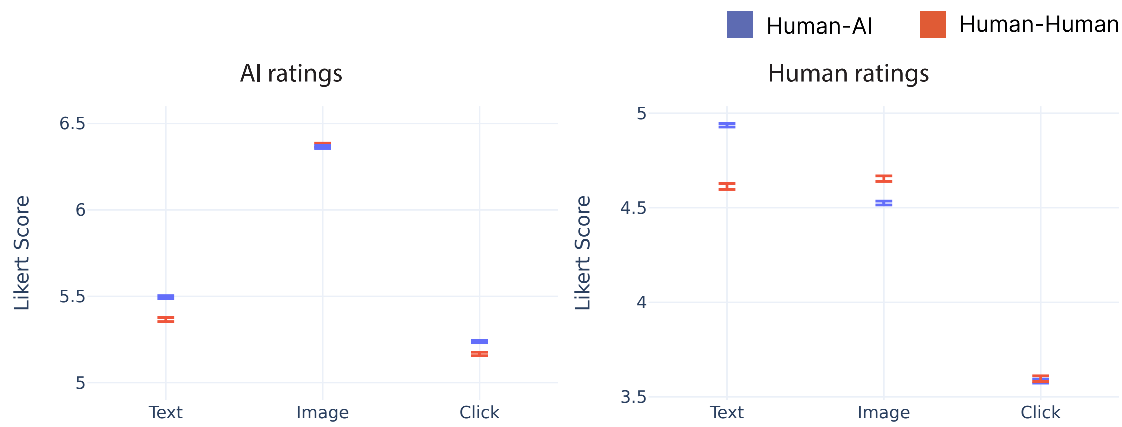

Figure 8: AI and Human ratings of ads.

Performance. To understand the effect of personality prompts on performance, we estimated heterogeneous prompt effects on the quality of completed ads. First, we examined the effects on our *post hoc* human evaluations of ad quality. In our analyses shown in Table 9, we observed numerous significant prompt effects, many of which have opposite heterogeneous effects depending on the personality of the participant. Interestingly, for text quality, we observed that a conscientious AI has a negative overall effect on text quality. However, this negative effect is reversed when the human is conscientious or open. AI extraversion has a negative effect when the human is agreeable but a positive effect when the human is extraverted. AI agreeableness has an overall positive effect, but this is reversed when the human is open or conscientious. AI neuroticism has a negative effect when the human is extraverted or neurotic but a positive effect when the human is agreeable.

Table 9: Personality-heterogeneous prompt effects on quality outcomes on the individual level.

	Human evaluations		
	Text	*Image*	*Click*
Intercept	4.856***	4.434***	3.342***
	(0.232)	(0.245)	(0.260)
Openness$_{AI}$	0.155	-0.014	0.214
	(0.174)	(0.184)	(0.193)
Openness$_{AI}$ $\times$ **Openness**$_{H}$	0.184	-0.216	-0.052
	(0.105)	(0.111)	(0.117)
Openness$_{AI}$ $\times$ **Conscientiousness**$_{H}$	-0.073	0.407**	0.060

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. Units are on a 7-point Likert scale. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Table 9 continued

	Human Evaluations		
	Text	*Image*	*Click*
	(0.130)	(0.142)	(0.149)
Openness$_{AI}$ × **Extraversion$_H$**	0.112	0.183	0.259*
	(0.111)	(0.118)	(0.123)
Openness$_{AI}$ × **Agreeableness$_H$**	0.039	-0.341**	-0.144
	(0.105)	(0.111)	(0.115)
Openness$_{AI}$ × **Neuroticism$_H$**	-0.069	0.008	-0.143
	(0.102)	(0.106)	(0.111)
Conscientiousness$_{AI}$	-0.628***	0.134	0.046
	(0.171)	(0.182)	(0.190)
Conscientiousness$_{AI}$ × **Conscientiousness$_H$**	0.392**	0.075	0.082
	(0.128)	(0.139)	(0.144)
Conscientiousness$_{AI}$ × **Extraversion$_H$**	-0.406***	-0.449***	-0.407**
	(0.110)	(0.118)	(0.123)
Conscientiousness$_{AI}$ × **Openness$_H$**	0.261*	0.104	0.145
	(0.104)	(0.111)	(0.116)
Conscientiousness$_{AI}$ × **Neuroticism$_H$**	-0.168	-0.167	-0.278*
	(0.099)	(0.104)	(0.108)
Conscientiousness$_{AI}$ × **Agreeableness$_H$**	0.047	-0.100	-0.097
	(0.105)	(0.109)	(0.113)
Extraversion$_{AI}$	0.232	0.139	0.103
	(0.173)	(0.185)	(0.193)
Extraversion$_{AI}$ × **Openness$_H$**	-0.072	-0.142	-0.157
	(0.103)	(0.111)	(0.116)
Extraversion$_{AI}$ × **Agreeableness$_H$**	-0.423***	-0.073	-0.210
	(0.106)	(0.111)	(0.115)
Extraversion$_{AI}$ × **Extraversion$_H$**	0.376**	-0.109	0.161
	(0.109)	(0.117)	(0.123)
Extraversion$_{AI}$ × **Conscientiousness$_H$**	-0.172	0.040	0.030
	(0.133)	(0.143)	(0.148)
Extraversion$_{AI}$ × **Neuroticism$_H$**	0.134	0.080	0.169
	(0.102)	(0.107)	(0.112)
Agreeableness$_{AI}$	0.491**	-0.158	-0.101
	(0.168)	(0.176)	(0.184)
Agreeableness$_{AI}$ × **Openness$_H$**	-0.437***	0.237*	0.123
	(0.102)	(0.108)	(0.113)
Agreeableness$_{AI}$ × **Conscientiousness$_H$**	-0.597***	-0.327*	-0.369*
	(0.133)	(0.141)	(0.148)
Agreeableness$_{AI}$ × **Extraversion$_H$**	0.173	-0.001	0.108
	(0.110)	(0.117)	(0.123)
Agreeableness$_{AI}$ × **Agreeableness$_H$**	-0.071	0.022	0.093
	(0.106)	(0.111)	(0.116)
Agreeableness$_{AI}$ × **Neuroticism$_H$**	0.122	-0.070	0.030
	(0.101)	(0.105)	(0.109)
Neuroticism$_{AI}$	0.068	0.210	0.168
	(0.175)	(0.185)	(0.195)
Neuroticism$_{AI}$ × **Openness$_H$**	0.008	0.046	0.022

Notes: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$. Heteroskedasticity-robust standard errors are reported. Units are on a 7-point Likert scale. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Table 9 continued

	Human Evaluations		
	Text	*Image*	*Click*
	(0.103)	(0.109)	(0.115)
Neuroticism$_{AI}$ × **Conscientiousness**$_{H}$	0.105	-0.004	0.111
	(0.130)	(0.139)	(0.145)
Neuroticism$_{AI}$ × **Extraversion**$_{H}$	-0.267*	0.094	-0.008
	(0.110)	(0.117)	(0.123)
Neuroticism$_{AI}$ × **Agreeableness**$_{H}$	0.356**	-0.055	-0.050
	(0.107)	(0.112)	(0.117)
Neuroticism$_{AI}$ × **Neuroticism**$_{H}$	-0.236*	-0.223*	-0.095
	(0.104)	(0.107)	(0.112)
Openness$_{H}$	0.161	0.088	0.039
	(0.127)	(0.136)	(0.145)
Conscientiousness$_{H}$	0.154	-0.055	0.111
	(0.173)	(0.188)	(0.197)
Extraversion$_{H}$	-0.210	0.063	-0.122
	(0.130)	(0.141)	(0.149)
Agreeableness$_{H}$	0.078	0.289	0.272
	(0.142)	(0.151)	(0.159)
Neuroticism$_{H}$	0.181	0.240	0.263
	(0.136)	(0.144)	(0.152)
Demographics (+ interaction terms)	Yes	Yes	Yes
Observations	29874	29874	29874

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. Units are on a 7-point Likert scale. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Next, we examined the heterogeneous effects of AI prompts on image quality. These effects are particularly important, given that ads created by Human-AI teams were rated as having lower image quality compared to those from Human-Human teams. For conscientious humans, AI openness improved image quality, while for agreeable humans, it had the opposite effect. AI conscientiousness negatively impacted image quality, but only when paired with extraverted humans. In contrast, AI agreeableness enhanced image quality for open humans but diminished it for conscientious humans. Finally, AI neuroticism reduced image quality specifically for neurotic humans.

Then, we analyzed the likelihood of clicking on the ads as an indicator of ad engagement. We only found that AI openness paired with human extraversion increased the click quality. Other combinations of personalities decreased click quality: AI conscientiousness paired with human extraversion or agreeableness and AI agreeableness paired with human conscientiousness, which is, interestingly, symmetrical between human and AI pairs.

3.5 Field study

To evaluate the real-world performance of ads created by Human-Human and Human-AI teams, we conducted a field experiment on a social media platform. 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: cost-per-click (CPC) and click-through rate (CTR), reported in Table 10, followed by view-through rate (VTR) and view-through duration (VTD) in Table 11. The field study extends lab findings and tests how the distinct productivity and quality profiles of Human-AI and Human-Human teams translate into advertising outcomes in a live setting. Broadly, we found that ads created by Human-AI teams performed similarly to those by Human-Human teams.

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 10. 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 reduce CPC, and this pattern holds when all predictors are combined. For CTR, expressed as a percentage, the analysis with collaboration indicates no direct Human-AI effect. Adding quality scores highlights text quality as a key driver of higher click-through rates, which remains consistent when all factors are included. These results suggest that Human-AI ads perform broadly equivalent to Human-Human ads in click metrics, with outcomes shaped by quality rather than team type.

View measures We assessed view metrics—view-through rate (VTR) and view-through duration (VTR; in log-seconds)—using similar regression models, as shown in Table 11. Our analysis of VTR shows no significant effects for collaboration type or quality. Our analysis of VTD reveals

[6] Higher 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).

	CPC ($)			CTR (%)		
Intercept	10.234***	11.636***	11.560***	0.000	-0.027*	-0.026*
	(0.323)	(0.686)	(0.698)	(0.006)	(0.013)	(0.013)
Human-AI	0.115		0.128	-0.000		-0.001
	(0.208)		(0.210)	(0.004)		(0.004)
Image		-0.268*	-0.264*		0.002	0.002
		(0.131)	(0.131)		(0.002)	(0.002)
Text		-0.154	-0.163		0.008***	0.008***
		(0.114)	(0.115)		(0.002)	(0.002)
Click		0.212	0.215		-0.002	-0.002
		(0.147)	(0.147)		(0.003)	(0.003)
Spend	-0.135***	-0.141***	-0.140***	0.006***	0.006***	0.006***
	(0.012)	(0.013)	(0.013)	(0.000)	(0.000)	(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 10: Effects on cost-per-click (CPC) and click-through rates (CTR) from the field study.

no Human-AI effect but uncovers text quality as a significant driver of longer viewing duration. These findings affirm the broad equivalence of Human-AI and Human-Human ads in view-through performance. While VTR remains largely unaffected by team type or quality, our analysis of VTD shows that the higher text quality of Human-AI teams can increase engagement time as well as CTR. Overall, these results confirm that Human-AI collaboration yields ads comparable to Human-Human efforts, with text and image quality differentially shaping real-world performance.

Personality prompts As before in Section 3.4, we examined the interaction effects of personality prompts on the outcomes of the field study. Table 12 shows the regression terms for the significant interaction effects, and we report the full regression in Section D. Among the significant interaction terms, we found that conscientious humans had 0.088% higher CTR but 0.072% lower CTR if the AI was prompted to be neurotic. An agreeable-prompted AI reduced CPC by $4.78, but this effect was reversed for an extraverted human by $3.11. Finally, a neurotic human paired with a neurotic AI saw CTR improved by 0.059%. These findings underscore that personality interactions between humans and AI agents significantly shape field performance and highlight the potential for tailored

	VTR			*VTD (log-sec)*		
Intercept	0.000	0.000	0.000	0.509***	0.000	0.000
	(0.006)	(0.011)	(0.012)	(0.052)	(0.116)	(0.118)
Human-AI	0.001		0.000	0.021		0.011
	(0.004)		(0.004)	(0.037)		(0.037)
Click		-0.000	-0.000		-0.003	-0.003
		(0.003)	(0.003)		(0.026)	(0.026)
Image		-0.003	-0.003		-0.001	-0.000
		(0.002)	(0.002)		(0.023)	(0.023)
Text		0.002	0.002		0.038*	0.037*
		(0.002)	(0.002)		(0.019)	(0.019)
Spend	0.001***	0.001***	0.001***	-0.001	0.005*	0.005**
	(0.000)	(0.000)	(0.000)	(0.002)	(0.002)	(0.002)
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.

Table 11: Effects on cost-per-click (CPC) and click-through rates (CTR) from the field study.

AI agents to improve advertising outcomes.

4 Discussion

4.1 Theoretical, practical, and methodological contributions

Theoretical contributions. This study advances theories of teamwork and collaboration by demonstrating how AI agents, particularly GPT-based models, reshape communication dynamics and workload distribution in human-AI teams (Schneider et al., 2021). We show that collaboration with GPT agents emphasizes task-oriented communication over social interactions, reducing social coordination costs and enabling participants to focus more on content generation. These findings extend theories of fit and complementarity in teamwork by illustrating how AI traits when aligned with human personality traits, enhance productivity and creativity (Ancona and Caldwell, 1992). For example, high-openness AI paired with open individuals fosters idea generation, while agreeable AI paired with extraverted individuals balances task-oriented and social communication. Tuning

	CPC ($)		*CTR (%)*	
Intercept	14.211***	16.533***	-0.067	-0.141*
	(2.920)	(3.041)	(0.057)	(0.059)
Spend	-0.144***	-0.150***	0.007***	0.008***
	(0.019)	(0.020)	(0.000)	(0.000)
Image		-0.461**		0.005
		(0.178)		(0.003)
Text		-0.314		0.012***
		(0.169)		(0.003)
Click		0.470*		-0.006
		(0.198)		(0.004)
Conscientiousness$_H$	-1.797	-1.913	0.088*	0.084
	(2.300)	(2.295)	(0.045)	(0.045)
Agreeableness$_{AI}$	-4.779*	-4.457	0.027	0.013
	(2.311)	(2.312)	(0.045)	(0.045)
Conscientiousness$_H$ $\times$ **Neuroticism**$_{AI}$	3.531*	3.537*	-0.072*	-0.068*
	(1.764)	(1.758)	(0.034)	(0.034)
Extraversion$_H$ $\times$ **Agreeableness**$_{AI}$	2.991	3.105*	-0.019	-0.020
	(1.542)	(1.537)	(0.030)	(0.029)
Neuroticism$_H$ $\times$ **Neuroticism**$_{AI}$	0.956	1.156	0.059*	0.056*
	(1.443)	(1.439)	(0.028)	(0.028)
Campaign RE	Yes	Yes	Yes	Yes
Observations	1859	1859	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 1-2.

Table 12: Significant interaction effects of human and prompted AI personalities on cost-per-click (CPC) and click-through rates (CTR) from the field study.

how AI agents interact and communicate is especially critical since communication patterns can predict team viability (Cao et al., 2021).

Additionally, this research contributes to our understanding of collaborative outcomes for GPT-based agents by highlighting their dual role as content generators and collaborators. Unlike prior work that frames AI tools as passive assistants (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Chen and Chan, 2024), our findings reveal that GPT-based agents actively influence the structure and quality of collaborative outputs, particularly in multimodal tasks. This study emphasizes the need to design GPT-based agents not only for efficiency but also for alignment with human workflows, paving the way for more nuanced and effective human-AI partnerships.

Practical contributions. This study provides actionable insights for organizations and designers deploying AI in collaborative environments. First, we demonstrate that GPT-based agents can significantly reduce social coordination costs, allowing individuals to focus more on task completion and content generation. By streamlining communication and reducing the need for social or emotional exchanges, AI agents enable greater individual productivity, particularly for low-performing participants. This finding is especially relevant for teams with varying skill levels or in high-pressure tasks where minimizing coordination overhead is crucial.

Second, our results highlight the importance of tailoring AI agents to align with the personality traits of their human collaborators. For instance, pairing high-openness AI with open individuals and agreeable AI with extraverted partners improves productivity and performance. These insights offer a practical roadmap for organizations to design and deploy AI agents that enhance team performance by optimizing personality fit. While studies have shown heterogeneity in the positive productivity effects of generative AI across role, function, and organization (Jaffe et al., 2024), it was unclear whether and how different versions of AI agents influence productivity and for what types of people. Customizing AI behavior based on human traits can drive better collaboration outcomes, particularly in creative or iterative workflows.

Third, the study underscores the strengths and limitations of GPT-based agents in multimodal workflows. While these agents excel at enhancing text quality, they underperform in tasks involving image-related outputs. This suggests that organizations should pair GPT-based agents with complementary tools specifically designed for image generation and evaluation. Such integrations can mitigate the trade-offs observed in multimodal tasks, ensuring consistent quality across all aspects of output.

Finally, this work offers guidance for structuring collaborative workflows in mixed human-AI teams. By identifying how GPT agents influence communication patterns and workload distribution, we provide a framework for assigning tasks that maximize individual and team productivity. For example, GPT agents can handle bulk or iterative edits while human collaborators focus on creative ideation and final quality checks. This division of labor can enhance efficiency while leveraging the unique strengths of both humans and AI, offering practical implications for industries ranging from marketing to product development and creative design.

Methodological contributions. This study introduces several methodological innovations that enhance our understanding of human-AI collaboration. First, the randomization of Human-Human and Human-AI teams represents a unique experimental design, allowing us to identify the distinct characteristics of human-human teamwork compared to collaborations involving AI agents. This approach provides causal evidence on how AI agents reshape communication patterns, workload distribution, and performance outcomes, filling a critical gap in prior research that often lacks such controls.

Second, we developed and employed a novel experimental ethnography approach, combining quantitative metrics with rich qualitative insights. The level of granularity in the data—capturing every time-stamped message, edit, and API call—enables a detailed reconstruction of collaboration workflows. This method bridges traditional experimental research with ethnographic approaches, offering a comprehensive view of team dynamics in real-time collaborative settings.

Third, this study introduces randomized personality prompts for AI agents, a methodological advance with significant implications for the burgeoning field of prompt engineering. Unlike traditional prompt engineering, which often focuses on one-shot outcomes, our approach explores how prompts influence interactions over time and within dynamic workflows (Schulhoff et al., 2024). By randomizing Big Five personality traits in AI agents, we provide a robust framework for studying the causal effects of personality alignment on collaboration outcomes. This methodology can be extended to test other dimensions of prompts, such as creativity, leadership styles, or domain expertise, across a range of interactive settings.

Finally, we built a state-of-the-art collaboration platform integrating the latest generative AI models and real-time collaborative tools, including APIs for ad editing and image generation. This platform not only advances the study of human-AI collaboration but also directly informs firms developing and deploying AI agents in workplace settings. By capturing how humans and

AI interact in realistic, task-driven environments, the platform provides actionable insights for organizations aiming to design effective and efficient collaborative workflows.

4.2 Limitations and future directions

Limitations. Several limitations in this study warrant further investigation. First, while we observe substantial differences in communication patterns between Human-Human and Human-AI teams, the underlying reasons remain unclear. These differences 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, reducing social or emotional exchanges. 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 purpose-built AI systems tailored to specific creative and evaluative tasks.

Finally, 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.

Future directions. This study opens several promising avenues for future research. First, while we randomized personality prompts to explore the interactive effects of human and AI personality alignment, future work could extend this approach to other dimensions of AI behavior. For example, prompts could be designed to vary in creativity, task orientation, or communication style, providing deeper insights into how different AI behaviors influence collaboration outcomes.

Second, 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.

Third, 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.

4.3 Conclusion

This study provides novel insights into how AI agents reshape teamwork, productivity, and performance in collaborative settings. By introducing randomized personality prompts and leveraging a state-of-the-art collaboration platform, we demonstrate that Human-AI teams communicate more, focus on task-related content, and achieve higher individual productivity compared to Human-Human teams. While collaboration with AI agents enhances text quality, it introduces trade-offs in multimodal outputs like images. Importantly, the forthcoming field evaluation of ad click-through and view-through rates will shed light on how these findings translate to real-world outcomes. Together, our results highlight the transformative potential of AI agents in collaborative workflows while underscoring the importance of aligning their design with human traits and task requirements.

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

D. G. Ancona and D. F. Caldwell. Demography and design: Predictors of new product team performance. *Organization science*, 3(3):321–341, 1992.

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.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *arXiv*, 2020.

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. Working Paper 31161, National Bureau of Economic Research, 2023.

H. Cao, V. Yang, V. Chen, Y. J. Lee, L. Stone, N. J. Diarrassouba, M. E. Whiting, and M. S. Bernstein. My team will go on: Differentiating high and low viability teams through team interaction. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), Jan. 2021. doi: 10.1145/3432929. URL https://doi.org/10.1145/3432929.

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. Candelon, 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.

S. Jaffe, N. P. Shah, J. Butler, A. Farach, A. Cambon, B. Hecht, M. Schwarz, and J. Teevan. Generative ai in real-world workplaces. Technical Report MSR-TR-2024-29, Microsoft, July 2024. URL https://www.microsoft.com/en-us/research/publication/generative-ai-in-real-world-workplaces/.

M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581196. URL https://doi.org/10.1145/3544548.3581196.

G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models, 2023.

M. Jung and P. Hinds. Robots in the wild: A time for more robust theories of human-robot interaction. *J. Hum.-Robot Interact.*, 7(1), May 2018. doi: 10.1145/3208975. URL https://doi.org/10.1145/3208975.

M. F. Jung, J. J. Lee, N. DePalma, S. O. Adalgeirsson, P. J. Hinds, and C. Breazeal. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, page 1555–1566, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450313315. doi: 10.1145/2441776.2441954. URL https://doi.org/10.1145/2441776.2441954.

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.

R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992. doi: 10.1111/j.1467-6494.1992.tb00970.x.

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.

T. L. R. Pamela J. Hinds and H. Jones. Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, 19(1-2):151–181, 2004. doi: 10.1080/07370024.2004.9667343. URL https://www.tandfonline.com/doi/abs/10.1080/07370024.2004.9667343.

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.

S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, and P. Resnik. The prompt report: A systematic survey of prompting techniques. *arXiv*, 2024.

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.

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
```

D Full interaction effects for human and prompted AI personalities for the field study

Table D.1: Personality-heterogeneous prompt effects on CPC and CTR.

	CPC		*CTR*	
Intercept	14.211***	16.533***	-0.067	-0.141*
	(2.920)	(3.041)	(0.057)	(0.059)
Spend	-0.144***	-0.150***	0.007***	0.008***
	(0.019)	(0.020)	(0.000)	(0.000)
Image		-0.461**		0.005
		(0.178)		(0.003)
Text		-0.314		0.012***
		(0.169)		(0.003)
Click		0.470*		-0.006
		(0.198)		(0.004)
Openness$_H$	-1.354	-1.425	0.048	0.051
	(1.722)	(1.718)	(0.034)	(0.033)
Conscientiousness$_H$	-1.797	-1.913	0.088*	0.084
	(2.300)	(2.295)	(0.045)	(0.045)
Extraversion$_H$	-3.003	-2.936	-0.023	-0.020
	(1.849)	(1.845)	(0.035)	(0.035)
Agreeableness$_H$	1.108	1.087	-0.038	-0.034
	(2.082)	(2.077)	(0.041)	(0.040)
Neuroticism$_H$	-1.295	-1.381	-0.002	-0.002
	(1.906)	(1.898)	(0.037)	(0.037)
Openness$_{AI}$	2.523	2.290	0.062	0.068
	(2.268)	(2.261)	(0.044)	(0.044)
Conscientiousness$_{AI}$	-2.175	-1.972	-0.002	0.004
	(2.315)	(2.313)	(0.046)	(0.045)
Extraversion$_{AI}$	-0.495	-0.583	-0.007	-0.004
	(2.313)	(2.304)	(0.045)	(0.045)
Agreeableness$_{AI}$	-4.779*	-4.457	0.027	0.013
	(2.311)	(2.312)	(0.045)	(0.045)
Neuroticism$_{AI}$	-0.911	-1.431	-0.017	-0.008
	(2.327)	(2.326)	(0.046)	(0.045)
Openness$_H$ $\times$ **Openness**$_{AI}$	-0.287	-0.082	-0.035	-0.044
	(1.424)	(1.428)	(0.028)	(0.028)
Openness$_H$ $\times$ **Conscientiousness**$_{AI}$	0.525	0.494	0.025	0.027
	(1.426)	(1.420)	(0.028)	(0.028)
Openness$_H$ $\times$ **Extraversion**$_{AI}$	2.151	2.095	-0.031	-0.033
	(1.420)	(1.416)	(0.028)	(0.027)
Openness$_H$ $\times$ **Agreeableness**$_{AI}$	0.688	0.458	-0.010	-0.005
	(1.388)	(1.385)	(0.027)	(0.027)
Openness$_H$ $\times$ **Neuroticism**$_{AI}$	-1.907	-1.645	-0.004	-0.008

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. *AI* refers to induced AI personality; *H* refers to human personality covariates.

Table 8 continued

	CPC		*CTR*	
	(1.421)	(1.420)	(0.028)	(0.028)
Conscientiousness$_H$ × **Openness**$_{AI}$	-0.695	-0.405	0.001	-0.002
	(1.737)	(1.734)	(0.034)	(0.034)
Conscientiousness$_H$ × **Conscientiousness**$_{AI}$	1.083	0.931	-0.040	-0.041
	(1.771)	(1.766)	(0.034)	(0.034)
Conscientiousness$_H$ × **Extraversion**$_{AI}$	1.315	1.477	-0.036	-0.035
	(1.747)	(1.743)	(0.034)	(0.034)
Conscientiousness$_H$ × **Agreeableness**$_{AI}$	0.247	0.231	-0.029	-0.024
	(1.728)	(1.725)	(0.033)	(0.033)
Conscientiousness$_H$ × **Neuroticism**$_{AI}$	3.531*	3.537*	-0.072*	-0.068*
	(1.764)	(1.758)	(0.034)	(0.034)
Extraversion$_H$ × **Openness**$_{AI}$	-2.166	-2.394	0.017	0.025
	(1.502)	(1.498)	(0.029)	(0.029)
Extraversion$_H$ × **Conscientiousness**$_{AI}$	1.178	1.116	-0.004	-0.009
	(1.502)	(1.503)	(0.029)	(0.029)
Extraversion$_H$ × **Extraversion**$_{AI}$	-0.836	-1.005	0.041	0.042
	(1.522)	(1.517)	(0.029)	(0.029)
Extraversion$_H$ × **Agreeableness**$_{AI}$	2.991	3.105*	-0.019	-0.020
	(1.542)	(1.537)	(0.030)	(0.029)
Extraversion$_H$ × **Neuroticism**$_{AI}$	0.588	0.483	0.055	0.056
	(1.485)	(1.480)	(0.029)	(0.028)
Agreeableness$_H$ × **Openness**$_{AI}$	-0.781	-0.788	-0.044	-0.046
	(1.584)	(1.578)	(0.031)	(0.031)
Agreeableness$_H$ × **Conscientiousness**$_{AI}$	0.624	0.512	0.021	0.022
	(1.572)	(1.568)	(0.031)	(0.031)
Agreeableness$_H$ × **Extraversion**$_{AI}$	-2.511	-2.314	0.054	0.050
	(1.619)	(1.614)	(0.032)	(0.031)
Agreeableness$_H$ × **Agreeableness**$_{AI}$	2.789	2.415	0.003	0.013
	(1.625)	(1.625)	(0.032)	(0.032)
Agreeableness$_H$ × **Neuroticism**$_{AI}$	-1.877	-1.479	0.031	0.020
	(1.590)	(1.589)	(0.031)	(0.031)
Neuroticism$_H$ × **Openness**$_{AI}$	-0.545	-0.655	-0.028	-0.025
	(1.423)	(1.418)	(0.028)	(0.027)
Neuroticism$_H$ × **Conscientiousness**$_{AI}$	-0.579	-0.568	0.011	0.006
	(1.450)	(1.448)	(0.028)	(0.028)
Neuroticism$_H$ × **Extraversion**$_{AI}$	1.466	1.361	-0.016	-0.017
	(1.486)	(1.482)	(0.029)	(0.029)
Neuroticism$_H$ × **Agreeableness**$_{AI}$	1.530	1.570	0.016	0.021
	(1.510)	(1.510)	(0.029)	(0.029)
Neuroticism$_H$ × **Neuroticism**$_{AI}$	0.956	1.156	0.059*	0.056*
	(1.443)	(1.439)	(0.028)	(0.028)
Campaign RE	Yes	Yes	Yes	Yes
Observations	1859	1859	2000	2000

Notes: *p<0.05; **p<0.01; ***p<0.001. Heteroskedasticity-robust standard errors are reported. *AI* refers to induced AI personality; *H* refers to human personality covariates.