Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Prompt Adaptation as a Dynamic Complement in Generative AI Systems

Eaman Jahani, Benjamin S. Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, David Holtz

To cite this article:
Eaman Jahani, Benjamin S. Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, David Holtz (2026) Prompt Adaptation as a Dynamic Complement in Generative AI Systems. Information Systems Research

Published online in Articles in Advance 30 Apr 2026

. https://doi.org/10.1287/isre.2025.2029

https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–17
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Prompt Adaptation as a Dynamic Complement in Generative AI Systems

Eaman Jahani,[a,b] Benjamin S. Manning,[b,c] Joe Zhang,[d] Hong-Yi TuYe,[b,c] Mohammed Alsobay,[e] Christos Nicolaides,[b,f] Siddharth Suri,[g] David Holtz[b,h,*]

[a] Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742; [b] MIT Initiative on the Digital Economy, Cambridge, Massachusetts 02142; [c] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [d] Graduate School of Business, Stanford University, Stanford, California 94305; [e] Microsoft Research, New York, New York 10012; [f] School of Economics and Management, University of Cyprus, 2109 Aglantzia, Cyprus; [g] Microsoft Research, Redmond, Washington 98052; [h] Columbia Business School, Columbia University, New York, New York 10027

*Corresponding author

Contact: eaman@umd.edu, https://orcid.org/0000-0003-3879-4275 (EJ); benjamin.scott.manning@gmail.com, https://orcid.org/0009-0000-9977-2390 (BSM); joe.z.zhang@stanford.edu, https://orcid.org/0000-0003-4145-1126 (JZ); hytuye@mit.edu, https://orcid.org/0009-0000-2708-0467 (HYTY); malsobay@microsoft.com, https://orcid.org/0000-0001-5350-2061 (MA); nicolaides.christos@ucy.ac.cy, https://orcid.org/0000-0002-1485-2736 (CN); suri@microsoft.com, https://orcid.org/0000-0002-1318-8140 (SS); david.holtz@columbia.edu, https://orcid.org/0000-0002-0896-8628 (DH)

Received: April 9, 2025
Revised: December 16, 2025
Accepted: March 9, 2026
Published Online in Articles in Advance: April 30, 2026

https://doi.org/10.1287/isre.2025.2029

Abstract. As generative AI systems rapidly improve, a key question emerges: how do users adapt to these changes, and when does such adaptation matter for realizing performance gains? This paper studies *prompt adaptation*—how users adjust their inputs in response to evolving model behavior—using a common experimental design applied to two preregistered tasks with 3,750 total participants who submitted nearly 37,000 prompts. We show that the importance of prompt adaptation depends critically on task structure. In a task with fixed evaluation criteria and an unambiguous goal, user prompt adaptation accounts for roughly half of the performance gains from a model upgrade. In contrast, in an open-ended creative task where the space of acceptable outputs is effectively unbounded and quality is subjective, performance improvements are driven primarily by model capability; prompt adaptation plays a limited role. We further show that automated prompt rewriting cannot generally substitute for human adaptation: when aligned with task objectives, it can modestly improve performance, but when misaligned, it can actively undermine the gains from model improvements. Together, these findings position prompt adaptation as a dynamic complement whose importance depends on task structure and system design, and suggest that without it, a substantial share of the economic value created by advances in generative models may go unrealized.

History: Pei-Yu Chen, Senior Editor; Gordon Burtch, Associate Editor.

Conflict of Interest Statement: S. Suri is currently an employee of Microsoft. M. Alsobay was formerly a paid intern of and is currently an employee of Microsoft. D. Holtz was formerly a paid intern and visiting researcher at Microsoft, and is currently a visiting researcher at OpenAI.

Funding: E. Jahani was supported by National Science Foundation [Grant 1745640]. D. Holtz received an unrestricted gift from Microsoft, part of which was used to fund this research. This research was also supported by Microsoft's Accelerating Foundation Models Research Initiative (https://www.microsoft.com/en-us/research/collaboration/accelerating-foundation-models-research/).

Supplemental Material: The online appendix, as well as the data and replication code for this paper, is available at https://doi.org/10.1287/isre.2025.2029.

Keywords: generative AI • prompt engineering • prompt adaptation • human-AI interaction • complementary skills

1. Introduction

Generative AI is being integrated into work practices across the economy (Zhang and Kamel Boulos 2023, Bright et al. 2025), yielding notable productivity gains in tasks as diverse as software development, writing, and scientific research (Noy and Zhang 2023, Peng et al. 2023, Yu 2024, Brynjolfsson et al. 2025, Dell'Acqua et al. 2026). Recent research points to even greater potential ahead, demonstrating advances in automating core scientific processes (Manning et al. 2024), including tasks as complex as chemical research and proving mathematical theorems (Boiko et al. 2023, Romera-Paredes

et al. 2024). The adoption of generative AI is also occurring at an unprecedented pace, with recent research showing that approximately 28% of U.S. workers are already using generative AI in their jobs—a rate that significantly outpaces early adoption of personal computers and the internet at comparable points in their diffusion (Bright et al. 2025, Bick et al. 2026).

As with many other general-purpose technologies, the effectiveness of generative AI depends not only on the technology itself but on users' ability to craft inputs that produce high-quality results. To interact with generative AI systems, users provide written instructions—or *prompts*—that guide the model's behavior. These prompts can range from simple commands (e.g., "write a short story about a robot") to highly detailed specifications tailored to particular outputs (e.g., a series of paragraphs instructing an AI system to implement a complete piece of software). In this way, prompting serves as a complementary skill—one that, like spreadsheet modeling in the early PC era, can determine the productivity impact of a given tool (Brynjolfsson and Hitt 2000).

Prompting has quickly become an area of active research and practice. Scholars have developed taxonomies of prompt engineering techniques (Oppenlaender 2023), documented recurring patterns in prompt construction (Schulhoff et al. 2024), and examined how developers embed prompts into software systems (Liang et al. 2025). Other studies have explored prompting strategies for specific applications, including image generation (Don-Yehiya et al. 2023, Xie et al. 2023) and clinical documentation (Yao et al. 2024). In parallel, practitioners have built prompt libraries, shared tutorials, and developed tools to support prompt design. These developments signal a growing consensus that prompting plays a meaningful role in extracting value from generative AI systems.

Yet, despite this consensus, prompting remains understudied as a dynamic practice. Many prompt libraries and tutorials present effective prompts as reusable artifacts. But prompts that work well with one model version may underperform or break entirely with the next (Liang et al. 2025, Meincke et al. 2025). Whereas recent research increasingly views prompting as an adaptive process, empirical evidence remains limited on how these strategies evolve—both as users refine prompts for a single model and as they adjust to model updates—and on how these changes ultimately affect performance. This raises a broader question for individuals and organizations investing in prompting capabilities: are prompt strategies transferable across model versions, or must they be continually revised to match changing model behavior?

To begin exploring this question, we identify *prompt adaptation*[1] as a measurable behavioral mechanism through which user-side inputs evolve alongside technical advances. We conceptualize prompt adaptation as a *dynamic complement*—that is, a user capability that adapts in response to changes in a technological system and is critical to realizing the full economic value of system improvements. In contrast to static complements (e.g., fixed training, prompt templates), dynamic complements emerge through situated use with rapid feedback, respond to model-level change, and may be enabled or suppressed by system design.

Generative AI systems support a wide variety of task structures. Some tasks are steering oriented, where users aim to reach a clearly defined goal and success can be measured against an objective benchmark (e.g., write code to solve a specific problem). Others are creative, where goals are open-ended, evaluation is subjective, and improvement depends more on exploration and curation than on precision (e.g., make a beautiful picture). Recent work helps to establish these regimes: Vafa et al. (2025) investigate the steerability of models, that is, how effectively users can reach a specified target, whereas Orwig et al. (2024) and Zhou and Lee (2024) describe generative AI–assisted creation as a cocreative process of idea generation, filtering, and aesthetic evaluation. In organizational settings, steering tasks (or *bounded* tasks, herein) dominate operational workflows such as classification, brand-compliant content generation, or reproducing structured materials where quality can be defined ex ante. Creative tasks (or *unbounded* tasks, herein), by contrast, underpin innovation activities such as logo or campaign design, product ideation, and concept development, where success depends on novelty, appeal, and fit to evolving goals. Understanding how prompting and prompt adaptation operate across these two regimes is essential for organizations deploying generative AI at scale.

To assess the role of prompt adaptation in shaping outcomes across different types of tasks—and to separate its contribution from the direct effects of model improvement—we conducted two preregistered online experiments with an aggregate sample of 3,750 participants. The first, a bounded task, asks participants to iteratively prompt one of three randomly assigned text-to-image models—DALL-E 2, DALL-E 3, or DALL-E 3 with automated large language model (LLM)–based prompt rewriting—to reproduce a reference image as accurately as possible. It is explicitly designed to capture steering or target-matching behavior. Each participant submitted at least 10 prompts and was eligible for a substantial performance-based bonus, incentivizing careful refinement. The second, an unbounded task, captures creative or open-ended behavior: participants design a logo for a hypothetical organization based on a short textual vignette without a specified target image or objective benchmark. Apart from this difference in goal structure, the two tasks, conducted on the same custom interface, were identical to participants. By comparing outcomes across treatment arms, conducting

post hoc analyses that reevaluate prompts on alternative models, and contrasting results across the two experiments, we estimate how users adapted their prompts in response to model improvements and how those adaptations contributed to overall performance across both bounded and unbounded settings.

Across both tasks, we find that participants assigned to DALL-E 3 produced significantly better outputs than those assigned to DALL-E 2 but that the sources of these gains depend on task structure. In the bounded image replication task, about half of the gains came from participants *adapting* their prompts to exploit the new model's capabilities—replaying DALL-E 2 prompts on DALL-E 3 yields only about half the total improvement. By contrast, in the open-ended logo generation task, performance gains were driven primarily by improvements in model capability, with prompt adaptation accounting for a much smaller share of the overall effect (roughly 7%, compared with over 90% attributable to the model). Importantly, the prompt adaptation we observe does not appear to be limited to advanced "prompt engineers": benefits from prompt refinement after interacting with more capable models are observable across the outcome distribution. Finally, to assess whether automation can substitute for user-side prompt adaptation, we examine automated prompt revision via GPT-4-based rewriting. We find that such automation can either erode or modestly enhance performance depending on alignment with the end user's goals: automated rewriting substantially reduced gains in the replication task but slightly improved performance in the creative setting. Together, these findings position prompt adaptation as a dynamic complement whose importance depends on task structure and system design, shaping how advances in generative AI translate into realized value.

In terms of related literature and additional theory, our research builds on work in information systems, emphasizing the importance of dynamic, user-driven complements to digital technologies. Research on IT-enabled dynamic capabilities has shown that the value of new systems depends not only on technical infrastructure but on organizations' ability to reconfigure routines and user behaviors in response to ongoing change (Teece et al. 1997, Bharadwaj 2000, Joshi et al. 2010). Related work on post-adoptive IT use has demonstrated that users often engage only superficially with new systems and that meaningful performance gains tend to emerge only when users experiment with and refine their interaction strategies over time (Jasperson et al. 2005). Recent research on human-AI collaboration further underscores that interface design and task structure shape the degree to which users can learn from and adapt to model behavior (Fügener et al. 2022). And the concept of coevolution has been introduced to describe how humans and generative AI systems jointly adapt over time, forming interdependent capabilities that neither could realize alone (Böhm and Schedlberger 2023).

We also engage with work on general-purpose technologies, which has long emphasized that the productivity gains from technical advances depend on the development of new human and organizational complements (David 1990, Brynjolfsson 1993, Brynjolfsson and Hitt 2000, Brynjolfsson et al. 2021). We conceptualize prompt adaptation as a *dynamic* complement that coevolves with model capability, emerges through use rather than formal training, and whose importance depends on task structure and system design. In this way, prompt adaptation shapes how—and under what conditions—technical improvements translate into downstream economic value.

This paper makes three core contributions. First, we conceptualize prompt adaptation as a dynamic complement to improvements in generative AI models, one through which users actively shape how technical advances translate into realized performance. Leveraging a replay-based analysis enabled by our experimental design, we provide direct empirical evidence for this distinction by separating performance gains attributable to model capability from those arising through user-side adaptation and show that prompt adaptation can account for a substantial share of realized performance improvements. Second, we identify task structure as a key boundary condition for the importance of prompt adaptation: it is a first-order driver of gains in bounded, steerable tasks with fixed objectives, but plays a much more limited role in open-ended creative tasks where quality is unbounded in nature. Finally, we show that automation intended to simplify prompting is not a neutral substitute for user adaptation. We find that automated prompt rewriting can either undermine or complement performance, depending on its alignment with user goals. Together, these contributions clarify when and how user adaptation amplifies model improvements, positioning prompt adaptation as a dynamic complement whose economic importance depends on task structure and system design.

The remainder of the paper is organized as follows. We begin by presenting a simple conceptual framework that characterizes how output quality evolves with improvements in model capacity and with users' corresponding adjustments in prompting effort. We then describe our experimental design and the two task settings we study—a bounded, steerable image replication task and an open-ended logo generation task—along with the data and techniques used in our analyses. Next, we present our empirical findings, including a decomposition of the overall effect into components attributable to model improvements versus prompt adaptation, distributional effects across outcome quantiles, and the impact of automated prompt revision. We conclude by synthesizing these results and discussing their implications for organizations adopting generative AI.

2. Conceptual Framework

We first develop a stylized analytical framework to understand how overall output quality depends jointly on model capacity and on users' prompting behavior. Our goal is not to propose a fully normative model of user-AI interaction, but rather to distinguish improvements directly attributable to the model itself from those arising through user-side prompt adaptation. This distinction motivates our experimental design and yields testable predictions about how performance gains and the distributional impacts of those gains change as models improve.

Although our empirical analysis focuses on two tasks—image replication and logo generation—the same conceptual structure applies broadly to settings in which users interact with generative models, including text generation, code assistance, and scientific research. In all such environments, output quality reflects a combination of model capability, user skill, and prompting effort, and total performance improvements can be decomposed into direct (model-driven) and behavioral (prompting-driven) components. The two tasks we study correspond to different regimes of this framework: image replication is a bounded task with a clear ceiling (perfectly replicating the image pixel for pixel), whereas logo generation is a more open-ended creative task without an optimal outcome. We develop the core logic for the bounded case in the main text and provide complete derivations for both bounded and unbounded formulations in Online Appendix A. Nearly all of the main results carry over to the unbounded case, with minor differences arising from curvature differences across specifications.[2]

2.1. Notation and Problem Setting

Let $\theta \in (0,1]$ denote the model's capacity to translate prompts into high-fidelity outputs, $s \in (0,1]$ denote a user's baseline skill in prompt engineering, and $x \geq 0$ denote the effort the user expends on writing and refining prompts. Each unit of effort incurs a linear cost $c(x) = kx$ for some $k > 0$, reflecting the total cognitive and temporal costs of refining one's prompts. We assume k is small ($k \approx 0$), consistent with the low marginal cost of iterative prompting and using generative AI models more generally (Shahidi et al. 2025).

We begin with a bounded quality function,

$$Q(\theta, s, x) = 1 - e^{-\theta s x},$$

which is increasing in model capacity, user skill, and prompting effort, but exhibits diminishing marginal returns and has a natural upper bound on performance. We define the user's utility as

$$U(\theta, s, x) = Q(\theta, s, x) - k\,x.$$

The user chooses effort to maximize utility. We assume $\theta s > k$—cost is not prohibitive to effort—which implies there is a unique interior optimum $x^*(\theta, s)$:

$$x^*(\theta, s) = \frac{1}{\theta s} \ln\left(\frac{\theta s}{k}\right) > 0, \tag{1}$$

with an optimal quality

$$Q^* = Q(\theta, s, x^*(\theta, s)) = 1 - \frac{k}{\theta s} > 0. \tag{2}$$

It follows immediately from Equation (2) that optimal quality is increasing with both model capacity

$$\frac{\partial Q^*}{\partial \theta} = \frac{k}{\theta^2 s} > 0 \tag{3}$$

and user skill

$$\frac{\partial Q^*}{\partial s} = \frac{k}{\theta s^2} > 0. \tag{4}$$

2.2. Decomposition into Model and Prompting Effects

We now consider an upgrade from a model with capacity θ_1 to one with a higher capacity θ_2. As model capacity θ increases, Equation (1) implies that optimal effort $x^*(\theta, s)$ may change. Thus, an improved model affects performance through two channels: it directly increases output quality for a fixed prompt, and it indirectly changes users' optimal effort through prompt refining. We refer to these as the *model effect* and the *prompting effect*, respectively.

Formally, let $x^*(\theta_1, s)$ and $x^*(\theta_2, s)$ denote the user's optimal prompting effort before and after the upgrade. The total improvement in output quality is

$$\Delta Q = Q(\theta_2, s, x^*(\theta_2, s)) - Q(\theta_1, s, x^*(\theta_1, s)).$$

This total gain can be decomposed as

$$\begin{aligned} \Delta Q = &\underbrace{\left(Q(\theta_2, s, x^*(\theta_1, s)) - Q(\theta_1, s, x^*(\theta_1, s))\right)}_{\text{Model effect}} \\ &+ \underbrace{\left(Q(\theta_2, s, x^*(\theta_2, s)) - Q(\theta_2, s, x^*(\theta_1, s))\right)}_{\text{Prompting effect}}. \end{aligned} \tag{5}$$

The first term isolates the gain from upgrading the model while holding the user's prompt strategy fixed at its preupgrade optimum. The second term captures the additional improvement arising from user-side prompt adaptation. Put differently, even if model capacity improves, a person may forgo performance gains if his or her prompts are not updated accordingly.

2.3. Distributional Impacts of Model Improvements

A related question is how a transition from a model with capacity θ_1 to one with higher capacity θ_2 affects the distribution of realized output quality. In other words, does access to a more capable model level the playing field or widen performance gaps between

users? Under the bounded formulation of the framework analyzed here, $Q(\theta, s, x) = 1 - e^{-\theta s x}$, improvements in model capacity imply a specific pattern in how performance gains are distributed across users.

Proposition 1 (Equalizing Effect of Model Improvements). *Under the bounded formulation, improvements in model capacity compress the distribution of output quality across users.*

This aggregate result can be understood by examining the separate distributional implications of the model effect and the prompting effect.

Proposition 2 (Model Effect: Equalizing Channel). *Holding prompting effort fixed, an increase in model capacity compresses the outcome distribution: The spread of outcomes narrows as performance at the lower end of the distribution rises proportionally more.*

Proposition 3 (Prompting Effect: Expanding Channel). *Allowing users to reoptimize their prompting effort expands the outcome distribution: Performance differences across users widen as the upper portion of the distribution shifts further upward.*

The formal proofs and derivations for all three propositions are provided in Online Appendix A.

Taken together, these results imply that model improvements generate two opposing distributional forces. The direct mechanical improvement in capacity—the model effect—is equalizing, whereas the behavioral response—the prompting effect—is expanding. In aggregate, the equalizing channel dominates, producing the overall compression established in Proposition 1.[3] Intuitively, as model capacity increases, performance near the top of the distribution is already close to the performance ceiling and experiences smaller marginal gains, whereas adaptive prompting raises the performance of the upper tail more sharply.

3. Experiment Design and Methods

To empirically examine whether users do, in fact, adapt their prompts in response to model improvements—and how much this adaptation contributes to overall performance—we conducted two preregistered online experiments. The first experiment (hereinafter the *image replication experiment*) was conducted on a sample of 2,059 participants on Prolific between December 12 and December 19, 2023. In this experiment, participants were asked to replicate a target image as closely as possible using a generative AI model. The second experiment (hereinafter the *logo generation experiment*) was conducted on a sample of 2,067 participants on Prolific between June 30 and July 1, 2025. In this experiment, participants were asked to create a logo for a hypothetical organization described in a short textual vignette (e.g., a sign for a business or a crest for a university). In both experiments, the goal was to assess how both model capability and prompting behavior influence final outcomes.[4]

Together, these two experiments are designed to span distinct but empirically important regimes of generative AI use. The image replication experiment captures a bounded, steerable task in which success can be evaluated against a fixed target and output quality has a natural upper bound. This structure enables precise measurement of performance against an objective benchmark. By contrast, the logo generation experiment captures an open-ended, creative task in which goals are underspecified, quality is inherently comparative rather than absolute, and there is no clear performance ceiling. Despite these differences in task structure, both experiments share a common experimental architecture that allows us to decompose the performance gains into model-driven (model effect) and user-driven (prompting effect) components. Studying both settings allows us to assess whether prompt adaptation operates similarly when users are steering models toward an objective target versus when they are iterating within a creative design space and to examine how the relative contributions of model capability and user adaptation vary across these regimes.

3.1. Experimental Setting

In both experiments, each participant was randomly and blindly assigned to one of three model conditions: DALL-E 2, DALL-E 3 (hereafter "DALL-E 3 (Verbatim)" or simply "DALL-E 3"), or DALL-E 3 with automatic prompt revision ("DALL-E 3 (Revised)"). These models differ not only in technical capability but also in whether they apply hidden LLM-based modifications to user prompts. In the DALL-E 2 condition, participants interacted with an earlier-generation model, DALL-E 2, that interprets prompts directly without intermediate rewriting. In contrast, the DALL-E 3 API, by default, forwards user prompts to GPT-4 before image generation. This intermediate GPT-4 step rewrites the prompt (typically by adding detail or restructuring language) before passing it to the image model. This behavior is intended to improve image quality but is not observed by or known to the prompter. In the DALL-E 3 (Revised) condition, we allowed this default behavior to proceed unaltered. This condition allows us to test whether system-side prompt rewriting serves as a substitute for user-side prompt adaptation or instead changes users' ability to steer the model. In the DALL-E 3 (Verbatim) condition, we attempted to suppress GPT-4-based prompt rewriting by prepending a hidden system message instructing the model to leave the user's prompt unchanged. Although some rewriting still occurred, the rate and extent of modifications were substantially reduced. In all

conditions, participants were unaware that their prompt might be rewritten or modified before image generation.

Importantly, all generative models in the experiment were memoryless: each prompt was processed independently, with no carryover from prior attempts. Participants were explicitly informed of this property before beginning the task to ensure they understood that previous prompts would not influence subsequent generations. Although this setup departs from typical real-world interfaces that maintain conversational context, it enables us to isolate prompt-level adaptation and to measure users' iterative learning without confounding from model memory or accumulated dialogue.

We designed a custom experimental interface resembling ChatGPT, which was used for both tasks. The participant's "objective" (either the target image or the logo brief, depending on the task) was displayed on the right side of the screen, and a scrollable history of prompts and generated images was displayed on the left. For both experiments, participants had up to 25 minutes to submit at least 10 prompts. They were paid $4 for completing the task, plus an $8 bonus (a 200% increase) if their highest-scoring image (according to the relevant outcome metric below) was in the top 20% of participants. After they had finished submitting prompts, participants completed a demographic survey covering age, gender, education, occupational skills, and self-assessed proficiency in creative writing, programming, and generative AI. The median completion time in the image replication experiment (including the demographic survey) was 23.5 minutes, with an average hourly wage (including bonuses) of about $13. The median completion time in the logo generation experiment (including the demographic survey) was 26.1 minutes, with an average hourly wage (including bonuses) of about $11.50. We removed from our analyses any participants who did not submit at least 10 prompts, repeated the same prompt five or more times consecutively, or failed to complete the post task survey.[5] This resulted in a final sample of 1,893 participants and 18,560 prompts in the image replication experiment and a final sample of 1,857 participants and 18,425 prompts in the logo generation experiment.[6]

3.2. Image Replication Experiment

In the image replication experiment, participants were asked to replicate a "target image" as closely as possible using the generative model provided to them. Each participant was randomly assigned one of 15 target images in addition to the participant's model assignment. These images were drawn from three broad categories—business and marketing, graphic design, and architectural photography—to represent common use cases for text-to-image generation.[7] All images were sourced from platforms that permit free research use (e.g., Unsplash, Reshot, Shopify, Pixabay, Gratisography).

3.2.1. Outcome Definition: Expected CLIP Embedding Cosine Similarity.

The primary outcome in the image replication experiment is the similarity between each participant-generated image and the assigned target image, measured using the cosine similarity of contrastive language–image pretraining (CLIP) embeddings (Radford et al. 2021). CLIP is a neural network trained to jointly embed images and text into a shared latent space such that semantically or visually similar items lie close together. By embedding both the target image and each generated image into this space and computing the cosine similarity between them, we obtain a quantitative measure of how closely the generated image matches the target along both visual and conceptual dimensions.

Because the output of each generative model is stochastic, the same prompt can yield different images across attempts. To account for this variability, we generated 10 images for each prompt and computed their cosine similarity to the target image individually. We then averaged these 10 similarity scores to produce an expected quality score for each prompt—the primary outcome variable used in our analysis. As a robustness check, we replicated all analyses using DreamSim, a recently developed alternative to cosine similarity on CLIP embeddings that is based on perceptual similarity and aligns more closely with human judgments (Fu et al. 2023).[8] The two measures were highly correlated, and our findings were consistent across both.

3.3. Logo Generation Experiment

In the second experiment, each participant was asked to generate an original logo based on a short set of instructions, or "briefs." Each participant was randomly assigned one of 15 briefs in addition to the participant's model assignment. One example is as follows:

> The owners of Green Pine Brewery are seeking a logo for their family-owned craft brewery located in Vermont.
>
> - Required visual elements: pine tree, hops, barley
> - Context for use: beer labels, merchandise, digital marketing
>
> Please create the best possible design based on the client's description and the listed requirements.

All briefs followed this exact format: the name of the organization, a single sentence of background information, two to three required visual elements, and the anticipated context of use.

The briefs span a range of topics, from a community park sign to a soccer team crest to a university letterhead. We provide the full list of briefs in Online Appendix B. They were written with three goals in mind: (i) to capture a wide variety of realistic use cases, (ii) to include a small number of concrete required elements that provide shared reference

points for evaluators in an otherwise subjective task, and (iii) to allow for open-ended creative possibilities. This third feature contrasts with the image replication experiment, where the best possible performance is well defined (perfectly replicate the target image).

3.3.1. Outcome Definition: Bradley–Terry Strength.

Because each logo generation task was open-ended and lacked an objective reference image, we could not assess output quality using embedding-based similarity metrics (e.g., cosine similarity of CLIP representations). Instead, we employed the Bradley–Terry (BT) model (Bradley and Terry 1952), a standard approach for inferring latent quality scores when only relative judgments are available (Zermelo 1929, Ford 1957). At a high level, the BT model is a probabilistic model that estimates a strength parameter (which we refer to hereafter as "BT β") for each item based on the outcomes of many pairwise comparisons.

To operationalize this, within each logo generation task, we enumerated all possible pairwise matchups among the participant-generated images and their replayed counterparts. We then randomly sampled 20% of these comparisons and obtained outcomes, indicating which image better satisfied the design brief, using GPT-4.1 nano as an LLM judge. The resulting win-loss matrix was converted into continuous BT βs by fitting a maximum-likelihood model following a standard iterative procedure (Ford 1957, Hunter 2004). Random subsampling of the data and automated LLM-based evaluation were necessary, given the scale of our data set: even after subsampling, we conducted nearly 21 million pairwise comparisons across the 15 logo tasks. A detailed description of the procedure we used to generate BT βs is found in Online Appendix D.[9]

Because quality in the logo generation experiment was defined via pairwise comparisons rather than similarity to a fixed reference, we did not resample multiple generations per prompt prior to estimating BT βs. In our first experiment, generating multiple images per prompt increased image generation volume (and thus direct study costs) by an order of magnitude while delivering smaller-than-anticipated gains in statistical power. Moreover, because BT βs are estimated from pairwise comparisons, increasing the number of images by an order of magnitude would have required roughly two orders of magnitude more comparisons, further increasing evaluation costs. Given these considerations and consistent with our preregistration, we analyzed the logo generation experiment without prompt-level resampling.

3.4. Replay Analysis for Separating Model and Prompting Effects

A central goal of our experiment is to distinguish how much of the performance improvement in image replication stems from using a more capable model versus how much comes from users adapting their prompts. Recall that our conceptual framework shows that the total improvement in output quality from upgrading a generative AI model with capacity θ_1 to a model with higher capacity θ_2 can be decomposed into two parts: the model effect, that is, the gain from applying the same prompts to a better model,

$$M = Q(\theta_2, s, x^*(\theta_1, s)) - Q(\theta_1, s, x^*(\theta_1, s)),$$

and the prompting effect, that is, the additional improvement from adapting prompts to take advantage of the more capable model,

$$P = Q(\theta_2, s, x^*(\theta_2, s)) - Q(\theta_2, s, x^*(\theta_1, s)).$$

To estimate the model and prompting components, we conducted a replay-based analysis using prompts from participants in the DALL-E 2 and DALL-E 3 (Verbatim) conditions. In both experiments, we reevaluated prompts under the alternative model—replaying DALL-E 2 prompts on DALL-E 3 and DALL-E 3 prompts on DALL-E 2—thereby holding prompt content fixed while varying model capacity. These cross-model evaluations provide empirical analogues of $Q(\theta_2, s, x^*(\theta_1, s))$ and $Q(\theta_1, s, x^*(\theta_2, s))$, respectively and, together with the original prompt-model pairings, allow us to implement the decomposition into model and prompting effects shown in Equation (5).

In the image replication experiment, replay was conducted several weeks after the initial data collection. To guard against model drift (Chen et al. 2024), we also regenerated outputs for the original prompt-model combinations and used those for our analyses. In the logo generation experiment, replay occurred shortly after data collection, so this additional step was unnecessary.

4. Results

In this section, we present our empirical findings on how increases in model capacity affect user performance across the two tasks. For each task, we first examine whether access to a more capable model improves average output quality, how users adapt their prompts in response, and how much of the overall performance gain can be attributed to model improvements versus prompt adaptation using the replay analyses. Next, we study how these gains reshape the distribution of outcomes across users and interpret these distributional impacts through the lens of our conceptual framework. Finally, we assess the impact of automated prompt rewriting using the DALL-E 3 (Revised) condition.

4.1. Image Replication Task

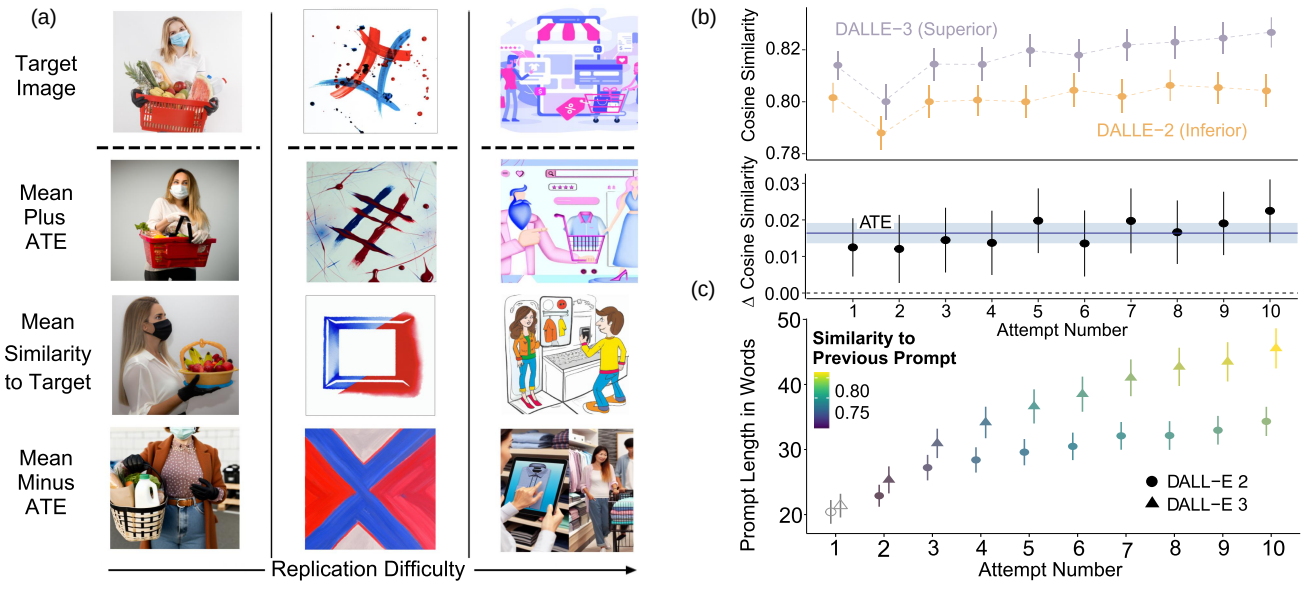
We begin with the image replication task. As discussed above, this task represents a bounded, steerable application of generative AI, in which users aim

to match an objective target, and output quality has a natural upper bound.

4.1.1. Overall Impact of Model Upgrades. We first examine whether participants using DALL-E 3 achieve higher performance than those using DALL-E 2, as implied by Equation (3). Figure 1 summarizes these findings. The top row of Figure 1(a) presents three of the target images provided to participants. Below each target, three generated images are shown, drawn from the full sample of participants across both model conditions. The middle row for each target shows the image whose cosine similarity to the target is closest to the mean similarity across all participants. The rows above (below) show images that are approximately one average treatment effect (ATE) more (less) similar to the target than the mean. This visualization provides qualitative intuition for the magnitude of the effect we estimate: the typical difference in fidelity between participants using DALL-E 2 and those using DALL-E 3 (Verbatim). Figure 1(b) shows that, across the 10 required prompt attempts, participants assigned to DALL-E 3 (Verbatim) produce images that are, on average, 0.0164 higher in cosine similarity to the target (95% confidence interval (CI): (0.0104, 0.0224), $p < 10^{-5}$). This improvement corresponds to roughly 0.19 standard deviations in performance. The gap persists across all attempts; participants using DALL-E 3 start off producing closer matches and maintain that edge through their 10th prompt.

Participants' dynamic prompting behavior also differs substantially between the two models. As shown in Figure 1(c), those assigned to DALL-E 3 write prompts that are, on average, 24% longer than those assigned to DALL-E 2, and this gap widens over successive attempts. Moreover, we observe that DALL-E 3 participants are more likely to reuse or refine their previous prompts (indicated by the shading intensity; see Online Appendix C for details), which suggests a more exploitative approach once they discover the model's capacity to handle detailed or complex instructions. Analyses of parts of speech confirm that these extra words likely provide additional descriptive information rather than mere filler: the proportion of nouns and adjectives—the two most descriptively informative parts of speech—is nearly identical across model conditions (48% for DALL-E 3 versus 49% for DALL-E 2; $p = 0.215$), suggesting that the increase in prompt length reflects the addition of semantically rich content rather than unnecessary verbosity. Together, these patterns illustrate prompt adaptation in practice: users progressively supply more information-dense prompts as they naturally learn the model's affordances, despite no explicit instructions to do so.

Figure 1. (Color online) Overall Performance and Prompting Behavior for the Image Replication Task



Notes. (a) For three example target images, the middle row shows participant-generated images closest to the mean similarity across all prompts. The rows above and below show images of approximately one average treatment effect (ATE) more or less similar to the target, illustrating the typical performance difference between model conditions. (b) Top: Average CLIP cosine similarity to the target image by attempt, separately for DALL-E 2 and DALL-E 3 participants. Bottom: The difference between these averages, with the solid horizontal line indicating the overall ATE and the shaded region showing the 95% confidence interval. (c) Average prompt length by attempt (y-axis), with error bars representing 95% confidence intervals. Shading intensity indicates the average textual embedding similarity between each prompt and the participant's previous prompt, capturing the extent of prompt reuse and refinement over time.

4.1.2. Replay Analysis and Decomposition of Effects. The differences we observe in prompting behavior suggest that users are actively adapting to the capabilities of the model to which they are assigned. But how much of the overall performance improvement for DALL-E 3 users is due to the model's enhanced technical capacity, and how much is due to users rewriting their prompts in response to that capacity? To answer this question, we turn to the replay analysis described in Section 3.4, which allows us to isolate these two effects empirically.

Figure 2(a) presents the results. To estimate the model effect, we hold prompts fixed and compare how the same DALL-E 2 prompts perform when evaluated on DALL-E 2 versus when replayed on DALL-E 3 (Verbatim). Because these prompts were written without knowledge of DALL-E 3's capabilities, any improvement reflects the gain from using a more capable model while holding the prompt fixed. We find that performance improves by 0.0084 in cosine similarity when these prompts are evaluated on DALL-E 3 ($p < 10^{-8}$; bootstrapped standard errors clustered at the participant level), which accounts for approximately 51% of the total difference in performance between the DALL-E 2 and DALL-E 3 arms.

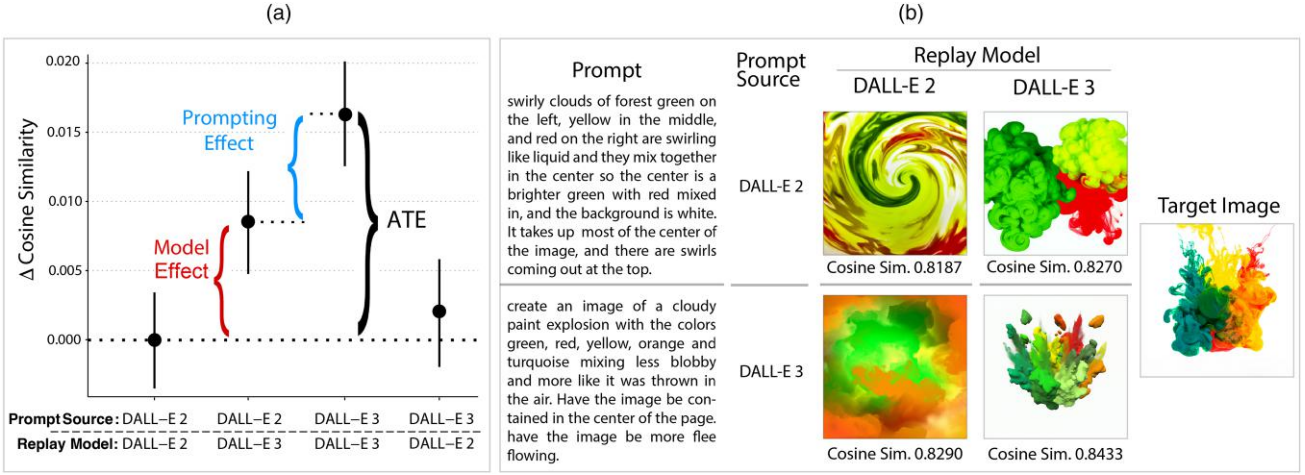
To estimate the prompting effect, we then compare the performance of these same DALL-E 2 prompts to the performance of prompts originally written by DALL-E 3 participants, both evaluated on DALL-E 3. Because both sets of prompts are evaluated on the same model, any difference reflects the effect of users adapting their prompts to the model's capabilities. We find that this prompting effect accounts for the remaining ≈ 48% of the total improvement,[10] corresponding to an increase of 0.0079 in cosine similarity ($p = 0.024$). Importantly, when we apply prompts written by DALL-E 3 users to DALL-E 2, we observe no performance benefit relative to the original DALL-E 2 prompts ($\Delta = 0.0020$; $p = 0.56$). This asymmetry reinforces the idea that the gains from prompt adaptation depend on the model's capacity to act on richer or more detailed prompt information.

Figure 2(b) illustrates these effects using a single target image. The two rows show different prompts submitted for that target, along with the images generated when evaluated by each model. In the top row, a prompt originally written by a DALL-E 2 participant yields a higher-fidelity image when replayed on DALL-E 3, demonstrating the improvement in output quality that comes from upgrading the model while holding the prompt fixed. In the bottom row, a prompt written by a DALL-E 3 participant produces a noticeably lower-quality image when rendered by DALL-E 2, underscoring the limits of prompt adaptation when the model lacks sufficient capacity to execute the instructions effectively.

Taken together, these findings offer empirical support for our theoretical claim: prompt adaptation operates as a dynamic complement that users deploy in response to improved model capabilities—and accounts for a substantial share of realized performance gains.

4.1.3. Distributional Effects. Next, we examine how model improvements and prompt adaptation reshape the distribution of performance outcomes in the image

Figure 2. (Color online) Replay Analysis and Effect Decomposition for the Image Replication Task



Notes. (a) Average performance of prompts evaluated across four prompt-model combinations. Comparing DALL-E 2 prompts evaluated on DALL-E 2 versus DALL-E 3 isolates the model effect. Comparing reused DALL-E 2 prompts on DALL-E 3 to original DALL-E 3 prompts on DALL-E 3 isolates the prompting effect. Error bars represent 95% confidence intervals based on bootstrapped standard errors clustered at the participant level. (b) A single target image with two submitted prompts: One written by a DALL-E 2 participant (top row) and one by a DALL-E 3 participant (bottom row). Images show how each prompt performs on both models, illustrating the model and prompting effects qualitatively.

Table 1. Distributional Effects of Model Upgrades in the Image Replication Task

Effect	Estimate (SE)	*p*-value
Total	0.022 (0.003)	$<10^{-5}$
Total × *Outcome Quantile (Percentile)*	−0.00012 (0.00005)	0.015
Model	0.011 (0.002)	$<10^{-5}$
Model × *Outcome Quantile (Percentile)*	−0.00006 (0.00003)	0.021
Prompting	0.011 (0.003)	$<10^{-5}$
Prompting × *Outcome Quantile (Percentile)*	−0.00006 (0.00005)	0.244

Notes. Negative interactions with outcome quantile indicate that gains from model upgrades are larger at lower points in the realized outcome distribution, implying compression of performance outcomes. The total and model effects exhibit statistically significant compression, whereas we do not detect a statistically significant distributional pattern for the prompting effect. The uninteracted total effect coefficient gives the QTE estimate at the zeroth percentile. SE, standard error.

replication task using a quantile treatment effects (QTE) perspective.

Table 1 reports estimates of the total, model, and prompting effects at different points of the realized outcome distribution, with outcome quantiles indexed by percentile. The total effect exhibits a negative and statistically significant interaction with outcome quantile (-0.000115, $p = 0.0152$), indicating that performance gains from upgrading from DALL-E 2 to DALL-E 3 are larger at lower points in the outcome distribution. This pattern is consistent with the framework's prediction that model improvements compress outcomes in bounded tasks. This compression is driven primarily by the model effect. Holding prompts fixed, the interaction between the model effect and outcome quantile is also negative and statistically significant (-0.000059, $p = 0.0210$), consistent with diminishing returns near the performance ceiling. Whereas the bounded formulation predicts that prompt adaptation exerts an expanding force on the outcome distribution, we do not detect a statistically significant interaction between the prompting effect and outcome quantile in this setting (-0.000056, $p = 0.2444$). This result does not rule out an expanding prompting channel but implies that any such effect is modest in magnitude relative to the dominant equalizing model effect.

4.1.4. Prompt Revision.

Finally, we examine whether automated prompt rewriting can serve as a substitute for human-driven prompt adaptation in the image replication task. In the DALL-E 3 (Revised) condition, user prompts were silently rewritten by GPT-4 before being submitted to the image model. Although participants in this condition still outperformed those using DALL-E 2 ($\Delta = 0.0069$; $p = 0.042$), they achieved substantially lower performance than participants using DALL-E 3 without rewriting. On average, automated prompt revision reduced the benefit of DALL-E 3 by 58% (95% CI: (40%, 76%)). Qualitative inspection of revised prompts suggests that the rewriting process often introduced additional details or altered phrasing in ways that were misaligned with participants' objective of replicating a specific target image as faithfully as possible. In this bounded, steerable task, such modifications reduced users' ability to precisely control the model's output and partially offset the performance gains achieved through direct user-side prompt adaptation.

4.2. Logo Generation Task

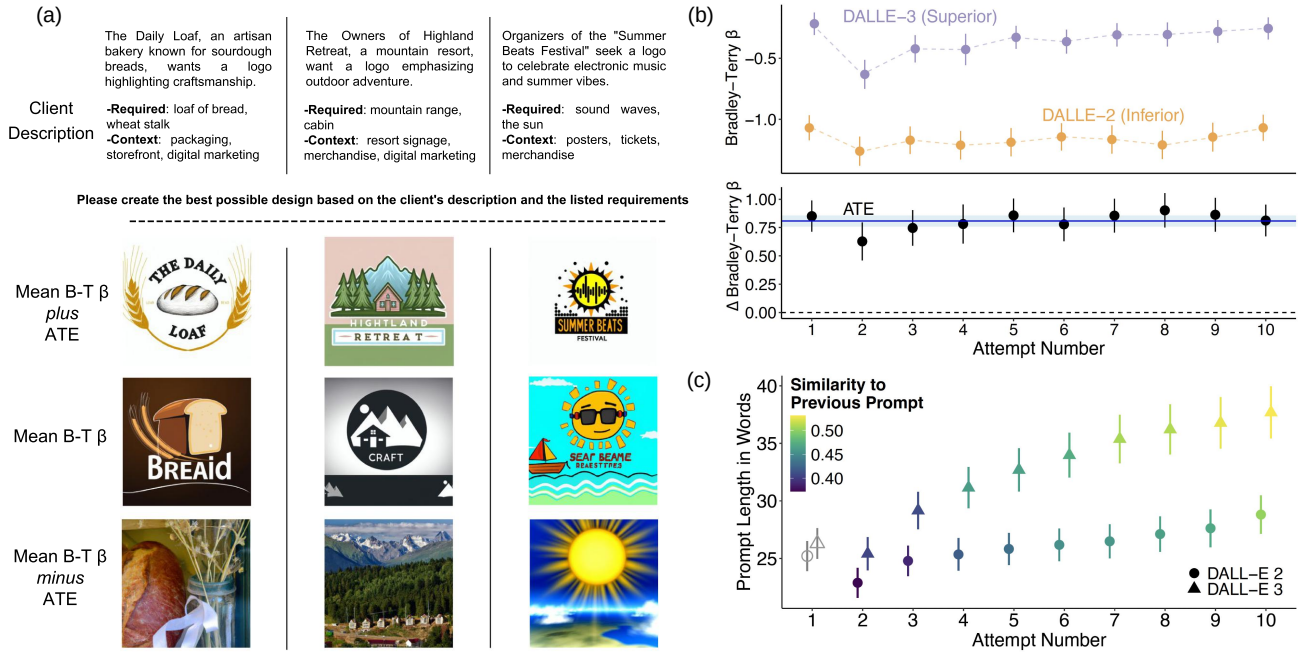
We next turn to the logo generation task. This task represents an open-ended, creative application of generative AI—such as design or ideation—in which success is evaluated comparatively rather than against an objective target, and output quality does not have a natural upper bound. Our empirical analyses mirror those of the image replication experiment.

4.2.1. Overall Impact of Model Upgrades.

We begin by examining whether access to a more capable model improves performance in the logo generation task. Participants assigned to DALL-E 3 produce substantially higher-quality logos than those assigned to DALL-E 2, with an estimated average treatment effect of 0.8075 in Bradley–Terry β (95% CI: (0.7120, 0.9030), $p < 10^{-10}$), corresponding to an improvement of approximately 0.57 standard deviations. Figure 3(a) provides qualitative intuition for the magnitude of this effect by showing representative logos at the mean, as well as approximately one average treatment effect above and below the mean. Figure 3(b) shows how this performance gap unfolds over successive prompting attempts. As in the image replication task, the advantage of the more capable model appears early and persists, although it does not appear to grow across attempts. Finally, Figure 3(c) shows prompting behavior across model conditions that closely mirrors the pattern observed in the image replication task. Participants using DALL-E 3 write prompts that are approximately 24.7% longer on average and exhibit greater similarity between successive prompts, consistent with more exploitative refinement. As in the image replication task, the share of nouns and adjectives is statistically indistinguishable across model conditions (43% for DALL-E 3 versus 42% for DALL-E 2; $p = 0.12$).

Together, these patterns show that in an open-ended, creative task without a fixed target or performance ceiling, increased model capability leads to large and persistent performance gains. These gains are accompanied by systematic changes in user prompting behavior as users adapt their prompts in response to

Figure 3. (Color online) Overall Performance and Prompting Behavior for the Logo Generation Task



Notes. (a) For three example logo instructions, the middle row shows participant-generated images closest to the mean Bradley–Terry β across all prompts. The rows above and below show images of approximately one average treatment effect (ATE) higher or lower quality, illustrating the typical performance difference between model conditions. (b) Top: Average Bradley–Terry β, separate for DALL-E 2 and DALL-E 3 participants. Bottom: The difference between these averages, with the solid horizontal line indicating the overall ATE and the shaded region showing the 95% confidence interval. (c) Average prompt length by attempt (y-axis), with error bars representing 95% confidence intervals. Shading intensity indicates the average textual similarity between each prompt and the participant's previous prompt, capturing the extent of prompt reuse and refinement over time.

the model's perceived capabilities. These results largely mirror those from the image replication task.

4.2.2. Replay Analysis and Decomposition of Effects. Next, we apply the same replay-based decomposition used in the image replication task to separate the overall performance gains in the logo generation experiment into components attributable to model improvements and to user prompt adaptation. Figure 4(a) presents the resulting decomposition.

The overall average treatment effect of upgrading from DALL-E 2 to DALL-E 3 is 0.8075 in Bradley–Terry β. In contrast to the image replication task, the vast majority of this gain is attributable to the model effect: Replaying prompts written by DALL-E 2 participants on DALL-E 3 increases performance by 0.748 ($p < 10^{-10}$), accounting for approximately 92.7% of the total effect. The corresponding prompting effect is small in magnitude (0.0577), statistically indistinguishable from zero ($p = 0.101$), and accounts for approximately 7.2% of the total effect. Consistent with this pattern, replaying prompts written for DALL-E 3 on DALL-E 2 yields no performance improvement ($\Delta\beta = -0.0125$, $p = 0.62$).

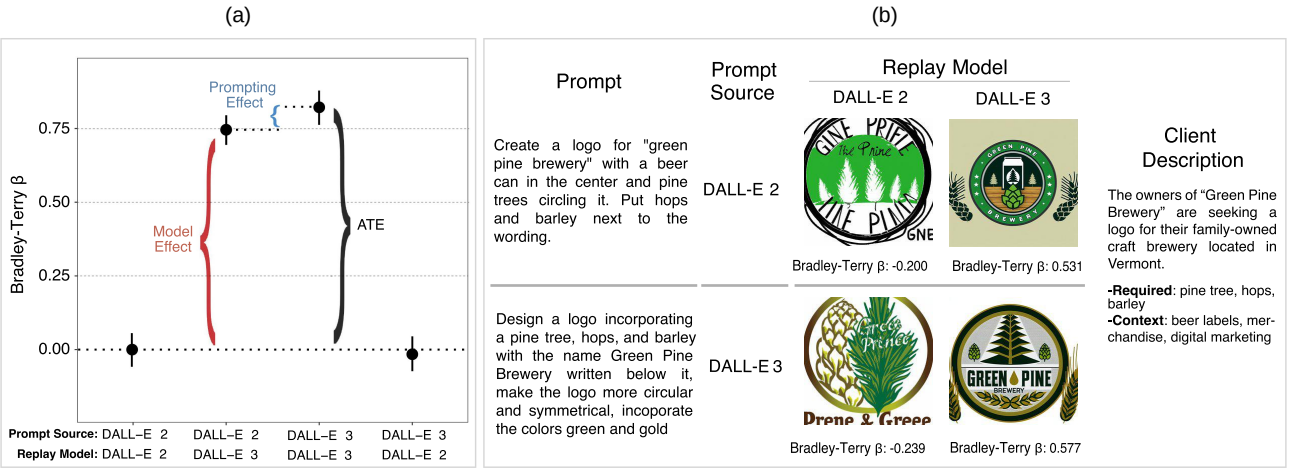
Figure 4(b) provides a qualitative illustration of this pattern for a representative logo brief. Although the prompt written by the DALL-E 3 participant is longer and more detailed, consistent with the prompt adaptation patterns shown in Figure 3(c), it does not perform substantially better on DALL-E 3 than the prompt written by the DALL-E 2 participant. Instead, the primary performance difference arises from evaluating either prompt on the more capable model. Although this comparison is illustrative rather than representative, it highlights how prompt adaptation in this task does not translate into large additional gains beyond those delivered by the model itself.

Taken together, these results indicate that in the logo generation task, realized performance gains are driven primarily by improvements in model capability, with prompt adaptation playing a comparatively limited role.

4.2.3. Distributional Effects. We also apply the same QTE approach used in the image replication task to examine how the model upgrade reshapes the distribution of performance outcomes in the logo generation experiment. Table 2 reports estimates of the total, model, and prompting effects at different points of the realized outcome distribution, with outcome quantiles indexed by percentile.

As was the case in the image replication task, the total effect of upgrading from DALL-E 2 to DALL-E 3

Figure 4. (Color online) Replay Analysis and Effect Decomposition for the Logo Generation Task



Notes. (a) Average performance of prompts evaluated across four prompt-model combinations. Comparing DALL-E 2 prompts evaluated on DALL-E 2 versus DALL-E 3 isolates the model effect. Comparing reused DALL-E 2 prompts on DALL-E 3 to original DALL-E 3 prompts on DALL-E 3 isolates the prompting effect. Error bars represent 95% confidence intervals based on bootstrapped standard errors clustered at the participant level. (b) A single client description with two submitted prompts: one written by a DALL-E 2 participant (top row) and one by a DALL-E 3 participant (bottom row). Images show how each prompt performs on both models, illustrating the model and prompting effects qualitatively.

exhibits a negative and statistically significant interaction with outcome quantile, indicating that performance gains are larger at lower points in the outcome distribution and that the upgrade compresses the distribution of outcomes. This pattern is again driven by the model effect: holding prompts fixed, the interaction between the model effect and outcome quantile is negative and statistically significant. By contrast, we do not detect a statistically significant interaction between the prompting effect and outcome quantile. Although this lack of statistical significance does not rule out the compressive impact of prompt adaptation predicted by the unbounded formulation of the model, it places an upper bound on the magnitude of any such effect in this setting.

Taken together, these results indicate that, as in the image replication task, distributional compression in the logo generation experiment arises primarily from improvements in model capability, with prompt adaptation playing a comparatively limited role in shaping the distribution of outcomes.

Table 2. Distributional Effects of Model Upgrades in the Logo Generation Task

Effect	Estimate (SE)	p-value
Total	1.135 (0.050)	$<10^{-5}$
Total × *Outcome Quantile (Percentile)*	−0.0065 (0.0008)	$<10^{-5}$
Model	1.012 (0.035)	$<10^{-5}$
Model × *Outcome Quantile (Percentile)*	−0.0053 (0.0005)	$<10^{-5}$
Prompting	0.122 (0.050)	0.014
Prompting × *Outcome Quantile (Percentile)*	−0.0013 (0.0008)	0.117

Notes. Negative interactions with outcome quantile indicate that gains from model upgrades are larger at lower points in the realized outcome distribution, implying compression of performance outcomes. As in the image replication task, distributional compression is driven by the model effect, whereas we do not detect a statistically significant distributional pattern for the prompting effect. The uninteracted total effect coefficient gives the QTE estimate at the zeroth percentile.

4.2.4. Prompt Revision. We conclude analyses of this experiment by examining the effects of automated prompt rewriting in the logo generation task. As in the image replication task, participants assigned to DALL-E 3 with automatic prompt revision outperform those using DALL-E 2, with an average treatment effect of 0.8337 in BT β ($p < 10^{-10}$). However, in contrast to the image replication task, automated prompt rewriting does not reduce performance relative to DALL-E 3 without rewriting; instead, it leads to a statistically non-significant 3.2% increase in BT β on average (95% CI: (−5.6%, 12.1%), $p = 0.44$).

One possible interpretation is that in an open-ended, creative task, the specific prompt rewriting used here, which primarily embellishes and elaborates user inputs, can enrich prompts in ways that facilitate exploration, introduce useful variation, or highlight design elements that improve perceived quality. Although we do not directly observe these mechanisms, the positive effect of rewriting in this task is consistent with the idea that automated prompt modification can complement user prompting when it is aligned with users' goals and the structure of the task.

5. Discussion

In this paper, we show that realized performance gains from generative AI upgrades reflect an interaction between model capability, user adaptation, and task structure. In a bounded, steerable task with an objective target (image replication), improvements in model quality translated into performance gains through two channels: roughly half of the improvement arose mechanically from the stronger model, whereas the other half reflected users adapting their prompts to better exploit the model's capabilities. In this setting, prompt adaptation is a substantive and economically meaningful complement to model improvement. By contrast, in an open-ended creative task without a natural performance ceiling (logo generation), performance gains were driven primarily by improvements in model capability. Although users again modified their prompting behavior when interacting with the stronger model, replay analyses indicate that prompt adaptation accounted for only a small share of the overall improvement. Thus, we argue that prompt adaptation functions as a dynamic complement to generative AI models, with task structure emerging as a key boundary condition: it plays a first-order role when success is defined against a fixed objective, but a more limited role when quality is open-ended and comparative.

A related question we address is whether prompt adaptation must occur through direct user experimentation or whether it can be partially automated by the system itself. In practice, many generative AI systems implement automated prompt rewriting to reduce user effort, implicitly positioning automation as a substitute for user-driven adaptation. Our findings suggest that the effectiveness of such automation depends critically on task structure and alignment with user goals. In the bounded image replication task, the DALL-E 3 endpoint's automated prompt rewriting reduced realized gains, consistent with misalignment between the rewriting logic and users' objective of faithfully reproducing a target image. By contrast, in the open-ended logo generation task, the same rewriting mechanism did not have a statistically significant impact on realized gains. Together, these findings underscore the importance of careful design in automated prompt rewriting systems. When system prompts are well aligned with user goals and task structure, automated rewriting has the potential to improve performance. But poorly aligned system prompts, especially when hidden, can degrade performance and limit users' ability to discover or exploit emergent but useful ways of interacting with a generative AI system. Such complexities are highlighted by related work, such as Yao et al. (2024), which studies automated prompt optimization in the context of clinical documentation.

More generally, our results both align with and help clarify themes in the literature on human-AI collaboration and the economics of technology, particularly the role of dynamic complements such as prompt adaptation that coevolve with technical change (Böhm and Schedlberger 2023). As generative AI models advance, often substantially from month to month, organizations that fail to adapt their prompting strategies in task settings where precision and steerability matter may forgo a meaningful share of the economic value these upgrades make possible. This coevolutionary pattern is not unique to AI. It echoes dynamics observed in earlier general-purpose technologies, where technical improvements often yielded modest returns until complementary skills and practices evolved to match (David 1990, Brynjolfsson 1993). However, the pace of generative AI advancement introduces a distinctive challenge: the window for adaptation is far shorter. Organizations that treat prompting as a one-time investment rather than an ongoing capability risk failing to capture the full value of model upgrades—echoing the productivity paradox observed when complementary assets lag behind technological potential (Brynjolfsson and Hitt 2000, Brynjolfsson et al. 2021). Our findings complement this stream of work by offering direct empirical evidence that a specific dynamic complement—user adaptation, specifically through prompt refinement—can account for a substantial share of realized performance gains from model upgrades in task settings where precision and steerability matter.

Seen from a broader organizational perspective, these findings also speak to long-standing work on IT-enabled dynamic capabilities (Teece et al. 1997, Bharadwaj 2000, Joshi et al. 2010) and post-adoptive IT use (Jasperson et al. 2005). Prompt adaptation, as we observe it, is not grounded in formal training or specialized skill. Participants in our study, drawn from a general population rather than a specialized group of prompt engineers, improved performance through trial and error within a single session. This contrasts with earlier IT transitions, where complementary capabilities often required extensive training (Attewell 1992, Von Hippel 2006). The relative accessibility of prompt adaptation highlights one potential pathway toward more broadly distributed productivity gains, provided that users are supported by scaffolds and interfaces that enable iterative refinement (Rogers 2003). At the same time, this accessibility introduces risk. Overoptimizing for a particular model version may reduce users' ability to adapt as systems evolve. This behavioral lock-in resembles challenges in architectural innovation, where tightly coupled routines inhibit flexibility when key system components change (Henderson and Clark 1990). Supporting long-term adaptation may require not just prompt training but workflows and learning mechanisms that encourage ongoing experimentation.

Our study has several limitations. First, our experimental interface was intentionally memoryless: each prompt was processed independently, and participants were informed that prior interactions would not influence subsequent generations. This simplification was necessary to isolate prompt-level adaptation in a controlled setting, but it departs from the conversational context that characterizes many real-world AI systems; our estimates therefore characterize performance and adaptation conditional on a memoryless interaction structure. Second, our focus was restricted to a single transition (from DALL-E 2 to DALL-E 3) and one type of generative AI (text-to-image generation). Although our conceptual framework suggests these mechanisms may generalize to other domains, further research is needed to assess how these dynamics play out in text generation, programming, scientific research, and other high-stakes settings. Third, we observe short-run adaptation behavior in a controlled setting, whereas longer-term learning dynamics, organizational feedback structures, or team-based workflows may shape prompting strategies differently in real-world environments. Finally, although our replay analysis helps isolate the contribution of prompt adaptation relative to model improvements, it does not fully disentangle the causal effects of specific prompt modifications (e.g., lengthening, lexical substitution, structural rephrasing), particularly given the path-dependent nature of user adaptation.

Building on these limitations, several promising research directions emerge. One important next step is to examine prompt adaptation in conversational, memory-enabled settings, where prior interactions persist and can be leveraged for cumulative learning by both users and models, and to study how this coevolution impacts performance. Relatedly, future studies might explore prompt adaptation longitudinally, observing how users develop durable heuristics that generalize across domains and respond to shifting model behavior over time. Another line of inquiry involves organizational-level complements to prompting, such as shared prompt repositories, collaborative refinement practices, or analytics dashboards that surface effective patterns. A third area concerns interface design—specifically, how features such as auto-complete, prompt scoring, and real-time feedback influence users' ability to experiment and adapt. Finally, researchers might explore how organizations balance standardization (which streamlines processes) with adaptability in prompting workflows in order to avoid creating rigid routines that lag behind technical advances.

For researchers and practitioners interested in the economics of AI, our findings reinforce the idea that complements play a central role in shaping performance. In some task settings, prompt adaptation can account for as much as half of the realized performance gains from model upgrades, underscoring that skill development and technological evolution must be treated as interdependent elements of innovation trajectories (Dosi 1982, Arthur 2009). As generative AI continues to advance, organizations that invest in *adaptive*, not static, complementary skills will be better positioned to realize its full value.

Acknowledgments

The authors thank Vivian Liu for early contributions to this project. The authors are also grateful to Greg Sun, Ethan Mollick, Nicholas Otis, Solene Delecourt, Rembrand Koning, Daniel Rock, Emma Wiles, Sonia Jaffe, Jake Hofman, and Benjamin Lira Luttges for their feedback. The authors have benefited from seminar and conference feedback at AFE 2025, the NYU-TAU Workshop on AI, MIT CODE, UC Berkeley, Microsoft, and the World Bank. E. Jahani, S. Suri, and D. Holtz led, directed, and oversaw the project; J. Zhang designed and built the experiment apparatus; B. S. Manning led the design of the online experiment flow and Qualtrics survey; J. Zhang led the prompt replay process; M. Alsobay led the analysis of prompt text, with contributions from HY. TuYe and E. Jahani; E. Jahani and HY. TuYe led all other data analysis and engineering, with contributions from J. Zhang, D. Holtz, and M. Alsobay; B. S. Manning, S. Suri, and D. Holtz led the writing of the manuscript; and all authors contributed to designing the research and writing the manuscript and supplementary information. Large language models were used in the drafting and editing of this paper. All ideas herein are those of the authors, who take full responsibility for the contents. All of the "target images" were collected from Unsplash, Reshot, Shopify, Pixabay, or Gratisography; all of these images have licenses for free use for commercial and noncommercial purposes. This study was reviewed by the UC Berkeley Committee for Protection of Human Subjects (Protocol 2023-06-16480).

Endnotes

[1] We use the phrase "prompt adaptation" in reference to changes in user prompting behavior in response to evolving model capabilities. This is distinct from the broader practice of prompt engineering, which includes static best practices, libraries, and templates, among other things.

[2] We note that both the bounded exponential and the unbounded logarithmic forms are illustrative rather than general and are not intended to be literal representations of the specific tasks we study. They adopt simple structures that allow us to decompose total performance gains into a "model effect" (improvements in capacity holding prompts fixed) and a "prompting effect" (improvements from user adaptation holding capacity fixed). This decomposition is a simplifying abstraction that isolates these mechanisms analytically, though in practice, they may interact. The comparative statics of $Q(\theta, s, x)$ also depend on its curvature and on how model capacity and skill interact (e.g., multiplicatively versus additively). We use these tractable forms to convey intuition, not to claim functional-form invariance.

[3] In the unbounded but concave logarithmic specification discussed in Online Appendix A, both the model and prompting channels are

equalizing, rather than opposing. The reversal stems from the additive interaction of capacity and skill, which makes improvements in θ more uniform across users rather than reinforcing skill differences.

[4] Full preregistration details for both experiments, including hypotheses and planned analyses, are available in an online repositories. Links to these online repositories can be found in the Online Appendix. All procedures were approved by an institutional review board (UC Berkeley Committee for the Protection of Human Subjects (CPHS), Protocol 2023-06-16480), and participants provided informed consent.

[5] This preregistered exclusion criterion was intended to filter out low-effort behavior in which participants repeatedly reused the same prompt in an attempt to complete the task as quickly as possible. Although repeating a prompt could also reflect legitimate exploration, given the stochastic nature of generative AI outputs, we note that this behavior occurred rarely in our data: the criterion affected 15 participants in the image replication experiment and 32 participants in the logo generation experiment (approximately 0.75%–1.75% of participants).

[6] Although participants were required to submit at least 10 prompts to be included in our analysis, the final prompt count in both experiments is slightly below $10 \times n_{participants}$. This is because we excluded some prompts due to technical issues (e.g., safety filter triggers, duplicate attempt numbers) and limited our main analysis to each participant's first 10 prompts to mitigate potential selection bias. Full details are provided in Online Appendix C.

[7] We selected the 15 target images with the goal of capturing diversity across common text-to-image use cases (business and marketing, graphic design, and architectural photography), while also taking into account practical considerations identified through pilot testing. Although a larger or systematically pretested image set would be desirable, there is no established standard for sampling text-to-image targets, and for a laboratory experiment of this scope, some pragmatic choices were necessary. Our aim was to ensure that tasks were neither trivially easy nor prohibitively difficult and to align broadly with categories commonly represented on repositories such as PromptBase. Target images were assigned using complete randomization across model-image cells, ensuring that image difficulty was orthogonal to model assignment. In addition, our analysis accounts for variation in difficulty by normalizing performance within each target image. Images were sourced from platforms that permit free research use for commercial and noncommercial purposes (e.g., Unsplash, Reshot, Shopify, Pixabay, Gratisography). A link to an online repository containing our target images can be found in Online Appendix B.4.2.

[8] DreamSim (Fu et al. 2023) is designed to better capture human perceptions of image similarity than traditional embedding-based methods. Our results are robust to using DreamSim in place of CLIP cosine similarity. Full DreamSim-based analyses are in Online Appendix H.

[9] A natural methodological question is whether BT βs derived from LLM-based pairwise comparisons should be calibrated against human judgments. We therefore attempted to benchmark LLM-based evaluations against human ratings. We recruited human evaluators to assess approximately 19,000 image pairs and estimated corresponding BT βs from those comparisons. Within individual logo briefs, however, we observed substantial disagreement among human raters, with brief-level Krippendorff's alpha values frequently below 0.6 and, in several cases, near zero or negative. This degree of disagreement indicates that human-derived BT β estimates are themselves unstable at the task level, reflecting the intrinsic ambiguity of the evaluation problem rather than noise in any particular rating source. Given this high level of subjectivity, we rely on LLM-based pairwise comparisons by GPT-4.1 nano to obtain a scalable measure of relative quality across designs while recognizing that quality in this task is inherently noisy and evaluative. Although LLM judgments do not constitute ground truth in any absolute sense, we treat them as a reasonable proxy for relative quality given the inherent instability of human raters in this setting.

[10] The total, model, and prompting effects are estimated using three separate two-way fixed-effects models, as explained in Online Appendix F. Because each effect is estimated from a different subset of counterfactual prompt-model comparisons with its own fixed effects, the resulting model and prompting effect estimates are not guaranteed to sum exactly to the total improvement.

References

Arthur WB (2009) *The Nature of Technology: What It Is and How It Evolves* (The Free Press, New York).

Attewell P (1992) Technology diffusion and organizational learning: The case of business computing. *Organ. Sci.* 3(1):1–19.

Bharadwaj A (2000) A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quart.* 24(1):169–196.

Bick A, Blandin A, Deming DJ (2026) The rapid adoption of generative AI. *Management Sci.*, ePub ahead of print January 20, https://doi.org/10.1287/mnsc.2025.02523.

Böhm K, Schedlberger L (2023) The use of generative AI in the domain of human creations—A case for co-evolution? *Proc. 9th Internat. Conf. Socio-Technical Perspect. IS (STPIS'23)* (CEUR Workshop Proceedings, Portsmouth, UK).

Boiko D, MacKnight R, Kline B, Gomes G (2023) Autonomous chemical research with large language models. *Nature* 624(7992):570–578.

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3/4): 324–345.

Bright J, Enock FE, Esnaashari S, Francis J, Hashem Y, Morgan D (2025) Generative AI is already widespread in the public sector: Evidence from a survey of UK public sector professionals. *Digit. Gov.: Res. Pract.* 6(1), Article 2, 13.

Brynjolfsson E (1993) The productivity paradox of information technology. *Comm. ACM* 36(12):66–77.

Brynjolfsson E, Hitt LM (2000) Beyond computation: Information technology, organizational transformation and business performance. *J. Econom. Perspect.* 14(4):23–48.

Brynjolfsson E, Li D, Raymond LR (2025) Generative AI at work. *Quart. J. Econom.* 140(2):889–942.

Brynjolfsson E, Rock D, Syverson C (2021) The productivity j-curve: How intangibles complement general purpose technologies. *Amer. Econom. J. Macroeconom.* 13(1):333–372.

Chen L, Zaharia M, Zou J (2024) How is ChatGPT's behavior changing over time? *Harvard Data Sci. Rev.* 6(2).

David PA (1990) The dynamo and the computer: An historical perspective on the modern productivity paradox. *Am. Econ. Rev.* 80(2):355–361.

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR (2026) Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organ. Sci.* 37(2):403–423.

Don-Yehiya S, Choshen L, Abend O (2023) Human learning by model feedback: The dynamics of iterative prompting with midjourney. Bouamor H, Pino J, Bali K, eds. *Proc. 2023 Conf. Empirical Methods Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), 4146–4161.

Dosi G (1982) Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Res. Policy* 11(3):147–162.

Ford LR (1957) Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* 64(8P2):28–33.

Fu S, Tamir N, Sundaram S, Chai L, Zhang R, Dekel T, Isola P (2023) DreamSim: Learning new dimensions of human visual similarity using synthetic data. Preprint, submitted June 15, https://arxiv.org/abs/2306.09344.

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. Systems Res.* 33(2):678–696.

Henderson RM, Clark KB (1990) Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Admin. Sci. Quart.* 35(1):9–30.

Hunter DR (2004) Mm algorithms for generalized Bradley-Terry models. *Ann. Statist.* 32(1):384–406.

Jasperson J, Carter PE, Zmud RW (2005) A comprehensive conceptualization of post-adoptive behaviors associated with IT-enabled work systems. *MIS Quart.* 29(3):525–557.

Joshi KD, Chi L, Datta A, Han S (2010) Changing the competitive landscape: Continuous innovation through IT-enabled knowledge capabilities. *Inform. Systems Res.* 21(3):472–495.

Liang JT, Lin M, Rao N, Myers BA (2025) Prompts are programs too! Understanding how developers build software containing prompts. *Proc. ACM Software Engrg.* 2(FSE):1591–1614.

Manning BS, Zhu K, Horton JJ (2024) Automated social science: Language models as scientist and subjects. NBER Working Paper No. 32381, National Bureau of Economic Research, Cambridge, MA.

Meincke L, Mollick E, Mollick L, Shapiro D (2025) Prompting science report 1: Prompt engineering is complicated and contingent. Preprint, submitted March 4, https://arxiv.org/abs/2503.04818.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

Oppenlaender J (2023) A taxonomy of prompt modifiers for text-to-image generation. *Behav. Inform. Tech.* 43(15):1–14.

Orwig W, Edenbaum ER, Greene JD, Schacter DL (2024) The language of creativity: Evidence from humans and large language models. *J. Creative Behav.* 58(1):128–136.

Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The impact of AI on developer productivity: Evidence from GitHub copilot. Preprint, submitted February 13, https://arxiv.org/abs/2302.06590.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, et al. (2021) Learning transferable visual models from natural language supervision. Meila M, Zhang T, eds. *Proc. 38th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 139 (PMLR, New York), 8748–8763.

Rogers EM (2003) *Diffusion of Innovations*, 5th ed. (The Free Press, New York).

Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar MP, Dupont E, Ruiz FJR, et al. (2024) Mathematical discoveries from program search with large language models. *Nature* 625(7995):468–475.

Schulhoff S, Ilie M, Balepur N, Kahadze K, Liu A, Si C, Li Y, et al. (2024) The prompt report: A systematic survey of prompting techniques. Preprint, submitted June 6, https://arxiv.org/abs/2406.06608.

Shahidi P, Rusak G, Manning BS, Fradkin A, Horton JJ (2025) The Coasean singularity: Demand, supply, and market design with AI agents. *The Economics of Transformative AI* (The University of Chicago Press, Chicago).

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strategic Management J.* 18(7):509–533.

Vafa K, Bentley S, Kleinberg J, Mullainathan S (2025) What's producible may not be reachable: Measuring the steerability of generative models. *Thirty-ninth Annual Conf. Neural Inform. Processing Systems (San Diego, CA)*.

Von Hippel E (2006) *Democratizing Innovation* (MIT Press, Cambridge, MA).

Xie Y, Pan Z, Ma J, Jie L, Mei Q (2023) A prompt log analysis of text-to-image generation systems. Ding Y, Tang J, Sequeda J, Aroyo L, Castillo C, Houben G-J, eds. *WWW'23: Proc. ACM Web Conf. 2023* (Association for Computing Machinery, New York), 3892–3902

Yao Z, Jaafar A, Wang B, Yang Z, Yu H (2024) Do clinicians know how to prompt? The need for automatic prompt optimization help in clinical note generation. *Proc. 23rd Workshop Biomedical Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), 182–201.

Yu Z (2024) The impacts of AI on scientific labor: Evidence from protein structure prediction. Preprint, submitted February 21, https://doi.org/10.2139/ssrn.4711334.

Zermelo E (1929) Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Math. Z.* 29(1):436–460.

Zhang P, Kamel Boulos MN (2023) Generative ai in medicine and healthcare: Promises, opportunities and challenges. *Future Internet* 15(9):286.

Zhou E, Lee D (2024) Generative artificial intelligence, human creativity, and art. *PNAS Nexu.* 3(3):pgae052.

Eaman Jahani is an assistant professor at the University of Maryland's Smith School of Business and a research affiliate at the World Bank and the MIT Initiative on the Digital Economy. He earned his PhD from MIT's Institute for Data, Systems, and Society and was a postdoctoral scholar in Statistics at UC Berkeley. His research uses network-based experimental design to study social norm interventions and how human-AI collaboration shapes creativity and collective decision-making.

Benjamin S. Manning is a PhD candidate at MIT Sloan. His research explores two complementary directions: (1) How can we use AI to better understand humans? (2) As AI systems increasingly act on people's behalf, how do markets change—and how can we improve the resulting outcomes? Prior to his doctoral studies, he was a high school math teacher. He holds a master's from Harvard and a BA from Washington University in St. Louis.

Joe Zhang is a PhD candidate in the Graduate School of Business at Stanford University. His research examines emerging human-AI systems and their implications for business and society. He uses computational methods and experiments to study how people interact with AI technologies and how these interactions shape outcomes in organizations and markets. Previously, he worked as a software engineer in the technology industry. He received his BSE in electrical engineering from Princeton University.

Hong-Yi TuYe is a PhD candidate at MIT Sloan's Information Technology group and a graduate student affiliate at both Stanford Digital Economy Lab and MIT Initiative on the Digital Economy. His research focuses on the labor market impacts of artificial intelligence, with emphasis on entrepreneurship, innovation, and digital experimentation. He has an MS in computational statistics and machine learning from University College London and a BA in economics from Columbia University.

Mohammed Alsobay studies how people—individually and collectively—interact with AI systems. He designs interactive experiments to address complementary challenges: building group-AI systems that improve collaboration, and developing human-centered evaluation methods that capture the ambiguity and interactivity of real-world AI use. He is a postdoctoral researcher in the Computational Social Science group at Microsoft Research and earned his PhD from the Information Technology group at MIT Sloan.

Christos Nicolaides is an associate professor at the University of Cyprus and a digital fellow at the MIT Initiative on the Digital

Economy. He received his PhD from the Massachusetts Institute of Technology in 2014. Christos' research focuses on statistical methods for large-scale empirical work, social influence mediated by technology, disease spreading and information diffusion through networks. His work has been published in *Nature Communications* and *PNAS* and covered by major media outlets.

Siddharth Suri is a senior principal researcher at Microsoft Research and a computational social scientist studying the interaction between technology and society. His early work examined how network structure shapes human behavior. He later helped pioneer large-scale "virtual lab" experiments on Amazon Mechanical Turk, leading to research on on-demand labor and the book *Ghost Work* (with Mary L. Gray). His recent work focuses on the future of work and the societal impacts of AI.

David Holtz is an assistant professor at Columbia Business School in the Decision, Risk and Operations Division, affiliated faculty at the Columbia University Data Science Institute, and a research affiliate at the MIT Initiative on the Digital Economy. He earned his PhD at MIT Sloan School of Management in the Information Technology group. Holtz's research focuses on online marketplaces and platform design, as well as the economic and societal impact of AI/ML systems.