GENERATIVE AI AT WORK*

ERIK BRYNJOLFSSON
DANIELLE LI
LINDSEY RAYMOND

We study the staggered introduction of a generative AI–based conversational assistant using data from 5,172 customer-support agents. Access to AI assistance increases worker productivity, as measured by issues resolved per hour, by 15% on average, with substantial heterogeneity across workers. The effects vary significantly across different agents. Less experienced and lower-skilled workers improve both the speed and quality of their output, while the most experienced and highest-skilled workers see small gains in speed and small declines in quality. We also find evidence that AI assistance facilitates worker learning and improves English fluency, particularly among international agents. While AI systems improve with more training data, we find that the gains from AI adoption are largest for moderately rare problems, where human agents have less baseline experience but the system still has adequate training data. Finally, we provide evidence that AI assistance improves the experience of work along several dimensions: customers are more polite and less likely to ask to speak to a manager. *JEL codes:* D80, J24, M15, M51, O33.

I. INTRODUCTION

The emergence of generative artificial intelligence (AI) has attracted significant attention for its potential economic impact.

* We gratefully acknowledge the editors, Lawrence Katz and Andrei Shleifer, and anonymous referees for many valuable insights and suggestions. We are grateful to Daron Acemoglu, David Autor, Amittai Axelrod, Eleanor Dillon, Zayd Enam, Luis Garicano, Alex Frankel, Sam Manning, Sendhil Mullainathan, Emma Pierson, Scott Stern, Ashesh Rambachan, John Van Reenen, Raffaella Sadun, Kathryn Shaw, Christopher Stanton, Sebastian Thrun, and various seminar participants for helpful comments and suggestions. We thank Max Feng and Esther Plotnick for providing excellent research assistance and the Stanford Digital Economy Lab for financial support. The content is solely the responsibility of the authors and does not necessarily represent the official views of Stanford University, MIT, or the NBER.

The Quarterly Journal of Economics (2025), 889–942. https://doi.org/10.1093/qje/qjae044.
Advance Access publication on February 4, 2025.

Although various generative AI tools have performed well in laboratory settings, questions remain about their effectiveness in the real world, where they may encounter unfamiliar problems, face organizational resistance, or provide misleading information (Peng et al. 2023a; Roose 2023).

We provide early evidence on the effect of generative AI deployed at scale in the workplace. We study the adoption of a generative AI tool that provides conversational support to customer-service agents. We find that access to AI assistance increases the productivity of agents by 15%, as measured by the number of customer issues they are able to resolve per hour. The gains accrue disproportionately to less experienced and lower-skill customer-support workers indicating that generative AI systems may be capable of capturing and disseminating the behaviors of the most productive agents.

Computers and software have transformed the economy with their ability to perform certain tasks with far more precision, speed, and consistency than humans. To be effective, these systems typically require explicit and detailed instructions for how to transform inputs into outputs: a software engineer must program computers. Despite significant advances in traditional computing, many workplace activities, such as writing emails, analyzing data, or creating presentations, are difficult to codify and have therefore defied computerization.

Machine learning (ML) algorithms work differently from traditional computer programs. Instead of requiring explicit instructions, machine learning algorithms infer instructions from examples. Given a training set of images, for instance, ML systems can learn to recognize specific individuals without requiring a list of the physical features that identify a given person. As a result, ML systems can perform tasks even when no instructions exist (Polanyi 1966; Autor 2014; Brynjolfsson and Mitchell 2017). The data to train these ML systems is often generated by human workers, who naturally vary in their abilities. As a result, ML tools may differentially introduce lower-performing workers to new skills and techniques, causing varied productivity shifts even among workers engaged in the same task.

We study the effect of generative AI on productivity and worker experience in the customer-service sector, an industry with one of the highest surveyed rates of AI adoption (Chui et al. 2021). We examine the staggered deployment of a chat assistant, using data from 5,172 customer-support agents working for

a Fortune 500 firm that sells business-process software. The tool we study is built on Generative Pre-trained Transformer 3 (GPT-3), a member of the GPT family of large language models developed by OpenAI (OpenAI 2023). The AI system monitors customer chats and provides agents with real-time suggestions for how to respond. The AI system is designed to augment agents, who remain responsible for the conversation and are free to ignore or edit the AI's suggestions.

We have four sets of findings.

First, AI assistance increases worker productivity, resulting in a 15% increase in the number of chats that an agent successfully resolves per hour. These productivity increases are based on shifts in three components of overall productivity: a decline in the time it takes an agent to handle an individual chat, an increase in the number of chats that an agent handles per hour (agents may handle multiple chats at once), and a small increase in the share of chats that are successfully resolved.

Second, the impact of AI assistance varies widely among agents. Less skilled and less experienced workers improve significantly across all productivity measures, including a 30% increase in the number of issues resolved per hour. The AI tool also helps newer agents to move more quickly down the experience curve: treated agents with two months of tenure perform just as well as untreated agents with more than six months of tenure. In contrast, AI has little effect on the productivity of higher-skilled or more experienced workers. Indeed, we find evidence that AI assistance leads to a small decrease in the quality of conversations conducted by the most skilled agents. These results contrast, in spirit, with studies that find evidence of skill-biased technical change for earlier waves of computer technology and robotics (Bresnahan, Brynjolfsson, and Hitt 2002; Bartel, Ichniowski, and Shaw 2007; Dixon, Hong, and Wu 2020).

Third, we investigate the mechanism underlying our main findings. Agents who follow AI recommendations more closely see larger gains in productivity, and adherence rates increase over time. We also find that experience with AI recommendations can lead to durable learning. Using data on software outages—periods when the AI's output is unexpectedly interrupted due to technical issues—we find that workers see productivity gains relative to their pre-AI baseline even when AI recommendations are unavailable. The gains are most pronounced among workers who had greater exposure to AI and followed AI suggestions more

closely. In addition, we examine the heterogeneous impact of AI access by the technical support conversation topics that agents encounter. While AI systems improve with access to more training data, we find that the gains to AI adoption—when used to complement human workers—are largest for relatively rare problems, perhaps because agents are already capable of addressing the problems they encounter most frequently. When there is insufficient training data for a topic, the system may not provide suggestions at all. We further analyze the text of agents' chats and provide evidence that access to AI improves their English-language fluency, especially among international agents. Finally, we compare the text of conversations before and after AI and provide suggestive evidence that AI adoption drives convergence in communication patterns: low-skill agents begin communicating more like high-skill agents.

Fourth, we focus on the dynamics between agents and customers. Contact-center (formerly known as call center) work is often challenging, with agents frequently facing hostile interactions from anonymous, frustrated customers. Many agents also work overnight shifts to align with U.S. business hours due to the prevalence of offshoring. While AI assistance can help agents communicate more effectively, it may risk making agents seem mechanical or inauthentic. Our findings show that access to AI assistance significantly improves customer treatment of agents, as reflected in the tone of customer messages. Customers are also less likely to question agents' competence by asking to speak to a supervisor. Notably, these changes come alongside a decrease in worker attrition, which is driven by the retention of newer workers.

Our findings show that access to generative AI suggestions can increase the productivity of individual workers and improve their experience of work. We emphasize that these findings capture medium-run effects in a single firm. Our article is not designed to shed light on the aggregate employment or wage effects of generative AI tools. In the longer run, firms may respond to increasing productivity among novice workers by hiring more of them or by seeking to develop more powerful AI systems that replace labor altogether. While the introduction of generative AI may increase demand for lower-skill labor within an occupation, the equilibrium response to AI assistance may lead to across-occupation shifts in labor demand that instead benefit higher-skill workers (Autor, Levy, and Murnane 2003; Acemoglu and

Restrepo 2018; Acemoglu 2024). Unfortunately, our data do not allow us to observe changes in wages, overall labor demand, or the skill composition of workers hired at the firm.

The results also underscore the longer-term challenges raised by the adoption of AI systems. In our data, top workers increase their adherence to AI recommendations, even though those recommendations marginally decrease the quality of their conversations. Yet with fewer original contributions from the most skilled workers, future iterations of the AI model may be less effective in solving new problems. Our work therefore raises questions about how these dynamics play out over the long run.

This article is related to a large literature on the effect of technological adoption on worker productivity and the organization of work (Rosen 1981; Autor, Katz, and Krueger 1998; Brynjolfsson and Hitt 2000; Athey and Stern 2002; Bartel, Ichniowski, and Shaw 2007; Acemoglu et al. 2007; Bloom et al. 2014; Michaels, Natraj, and Van Reenen 2014; Garicano and Rossi-Hansberg 2015; Hoffman, Kahn, and Li 2018; Acemoglu and Restrepo 2020; Felten, Raj, and Seamans 2023). Many of these studies, particularly those focused on information technologies, find evidence that IT complements higher-skill or more educated workers (Bresnahan, Brynjolfsson, and Hitt 2002; Akerman, Gaarder, and Mogstad 2015; Taniguchi and Yamada 2022). For instance, Bartel, Ichniowski, and Shaw (2007) find that firms that adopt IT tend to use more skilled labor and adoption is associated with increased skill requirements for machine operators in valve manufacturing. Other research investigates how technology affects workers based on their educational attainment and occupation. Acemoglu and Restrepo (2020) find that the negative effects of robots on employment are most pronounced for workers in blue-collar occupations who lack a college education.

Fewer studies focus on AI-based technologies, generative or not. Zolas et al. (2020), Acemoglu et al. (2022), and Calvino and Fontanelli (2023) examine economy-wide data from the United States and the Organisation for Economic Co-operation and Development (OECD), finding that the adoption of AI tools is concentrated among larger and younger firms with relatively high productivity. So far, evidence is mixed on the effects of AI on productivity. Acemoglu et al. (2022) find no detectable relationship between investments in AI-specific tools and firm outcomes, and Babina et al. (2024) find evidence of a positive relationship between firms' AI investments and their subsequent growth and

valuations. All the studies stress that determining the productivity effects of AI technologies is difficult because AI-adopting firms differ substantially from non-adopters.

Other studies present mixed findings on the effects of AI tools on decision making, often revealing challenges in human-AI collaboration. As an example of a positive finding, Kanazawa et al. (2022) find that a non-generative AI tool which suggests customer-rich routes to taxi drivers reduces their search time by 5%, with the least experienced drivers benefiting the most. By contrast, several other studies find that humans assisted by AI make worse decisions than do either humans or AI alone (Hoffman, Kahn, and Li 2018; Poursabzi-Sangdeh et al. 2021; Angelova, Dobbie, and Yang 2023; Agarwal et al. 2023). In fact, a meta-analysis of more than 100 experimental studies concludes that on average, human-AI collaborations underperform both the AI alone and the best human decision makers (Vaccaro, Almaatouq, and Malone 2024). These results underscore the particular challenges introduced when using AI-based tools designed to augment human decision making.

We provide micro-level evidence on the adoption of a generative AI tool across thousands of workers employed by a specific firm and its subcontractors. Our work is closely related to several other studies examining the effects of generative AI in lab-like settings. Peng et al. (2023b) recruited software engineers for a specific coding task (writing an HTTP server in JavaScript) and show that those given access to the AI tool GitHub Copilot complete this task twice as quickly as a control group. Similarly, Noy and Zhang (2023) conduct an online experiment showing that subjects given access to ChatGPT complete professional writing tasks more quickly. In the legal domain, Choi and Schwarcz (2023) show that AI assistance helps law students on a law school exam, whereas in management consulting, Dell'Acqua et al. (2023) find that access to GPT-4 suggestions improves the quality of responses on some management consulting tasks, but can negatively affect performance on tasks outside its capabilities. Consistent with our findings, Noy and Zhang (2023), Choi and Schwarcz (2023), Peng et al. (2023a), and Dell'Acqua et al. (2023) find that generative AI assistance compresses the productivity distribution, with lower-skill workers driving the bulk of improvements.

This article provides new evidence of longer-term effects in a real-world workplace where we also track patterns of learning, customer-side effects, and changes in the experience of work.

II. Generative AI and Large Language Models

II.A. Technical Primer

The rapid pace of AI development and public release tools such as ChatGPT, GitHub Copilot, and DALL-E have attracted widespread attention, optimism, and alarm (White House 2022). Such tools are examples of "generative AI" a class of machine learning technologies that can generate new content—such as text, images, music, and video—by analyzing patterns in existing data. This paper focuses on an important class of generative AI: large language models (LLMs). LLMs are neural network models designed to process sequential data (Bubeck et al. 2023). An LLM is trained by learning to predict the next word in a sequence, given what has come before, drawing on a large corpus of text, such as Wikipedia, digitized books, and portions of the internet. From its knowledge base of the statistical co-occurrence of words, the LLM generates new text that is grammatically correct and semantically meaningful. Although LLM implies human language, these techniques can also be used to produce other forms of sequential data (text), such as computer code, protein sequences, audio, and chess moves (Eloundou et al. 2023).

Four factors are driving improvements in generative AI: computing scale, earlier innovations in model architecture, the ability to pretrain using large amounts of unlabeled data, and refinements in training techniques (Radford et al. 2018; Radford et al. 2019; Ouyang et al. 2022; Liu et al. 2023).

The quality of LLMs is strongly dependent on scale: the amount of computing power used for training, the number of model parameters, and the data set size (Kaplan et al. 2020). The GPT-3 model included 175 billion parameters, was trained on 300 billion tokens, and incurred approximately $5 million in training costs. The GPT-4 model includes 1.8 trillion parameters and was trained on 13 trillion tokens at an estimated computing-only cost of $65 million (Li 2020; Brown et al. 2020; Patel and Wong 2023).

Modern LLMs use two earlier key innovations in model architecture: positional encoding and self-attention. Positional encodings keep track of the order in which a word occurs in a given

input. Self-attention assigns importance weights to each word in the context of the entire input text. Together, these technological advances enable models to capture long-range semantic relationships in an input text, even when that text is broken up into smaller segments and processed in parallel (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017).

These innovations in model architecture enable LLMs to train on large amounts of unlabeled data from sources such as Reddit and Wikipedia. Unlabeled data are far more prevalent than labeled data, allowing LLMs to learn about natural language on a much larger training corpus (Brown et al. 2020). By seeing, for example, that the word "yellow" is more likely to be observed with "banana" or "sun" or "rubber duckie," the model can learn about semantic and grammatical relationships without explicit guidance (Radford et al. 2018). This approach enables LLMs to learn a foundational understanding of language patterns and relationships that can be adapted or fine-tuned for a specific task.

Fine-tuning can refine general-purpose LLMs's output to match the priorities of a specific setting (Ouyang et al. 2022; Liu et al. 2023). Fine-tuning can help eliminate factually incorrect or inappropriate responses or prioritize a particular tone of response. Such improvements make a general-purpose model better suited to its specific application (Ouyang et al. 2022). For example, a model trained to generate social media content can be further trained on labeled data that contain not just the content of a post but also information on the user engagement it attracts.

Together, these innovations in computing scale, model architecture, and training have generated meaningful improvements in model performance. The Generative Pre-trained Transformer (GPT) family of models, in particular, has attracted considerable attention for outperforming humans on tests such as the U.S. bar exam (Liu et al. 2023; Bubeck et al. 2023; OpenAI 2023).

II.B. The Economic Effects of Generative AI

Computers have historically excelled at executing preprogrammed instructions, making them particularly effective at tasks that can be described by explicit rules (Autor 2014). Consequently, computerization has disproportionately decreased the demand for workers performing routine and repetitive tasks such as data entry, bookkeeping, and assembly line work, reducing wages in these jobs (Acemoglu and Autor 2011). At the same

time, computerization has increased the demand for workers who have complementary skills such as programming, data analysis, and research. As a result, technology-related shifts in the labor market have contributed to increased wage inequality in the United States and have been linked to a variety of organizational changes (Katz and Murphy 1992; Brynjolfsson and Hitt 2000; Bresnahan, Brynjolfsson, and Hitt 2002; Autor, Levy, and Murnane 2003; Baker and Hubbard 2003; Michaels, Natraj, and Van Reenen 2014; OECD 2023).

In contrast, generative AI tools do not require explicit instructions to perform tasks. If asked to write an email denying an employee a raise, generative AI tools will likely respond with a professional and conciliatory note. The model will have seen many examples of workplace communication without the need for a programmer to explicitly define what professional writing looks like. These machine learning methods behind AI enable computers to perform nonroutine tasks that rely on tacit knowledge and experience. AI has shown promise on tasks traditionally dominated by highly skilled professionals insulated from prior waves of automation, including complex mathematics, scientific analysis, and financial modeling. For example, Github Copilot, an AI tool that generates code suggestions for programmers, has achieved impressive performance on technical coding questions and, if asked, can provide natural language explanations of how the code it produces works (Nguyen and Nadi 2022; Zhao 2023). In addition, beyond learning to predict good outcomes from human-generated data, ML models can implicitly identify characteristics or patterns of behavior that distinguish high and low performers. Generative AI could replace lower-skill workers with AI-based tools or they may use them to help lower-skill, less experienced workers get up to speed more quickly. Together, the rise of generative AI has the potential to significantly alter the established relationships among technology, labor productivity, and economic inequality (White House 2022).

Despite their potential, generative AI tools face significant challenges in real-world applications. At a technical level, popular LLM-based tools, such as ChatGPT, have been shown to produce false or misleading information unpredictably, raising concerns about their reliability in high-stakes situations. While LLM models often perform well on specific tasks in the lab (OpenAI 2023; Peng et al. 2023b; Noy and Zhang 2023), the types of problems that workers encounter in real-world settings are likely to

be broader and less predictable. Furthermore, generative AI tools often require prompts from human operators, but finding ways to effectively combine human and AI expertise is difficult. For instance, earlier research indicates that decision-support systems integrating AI with human judgment often perform worse than those that rely on humans or AI alone (Vaccaro, Almaatouq, and Malone 2024). These challenges raise concerns about the ability of AI systems to provide accurate assistance in every circumstance and—perhaps more important—workers' capacity to distinguish cases where AI suggestions are effective from those where they are not.

Finally, the efficacy of new technologies is likely to depend on how they interact with existing workplace structures. Promising technologies may have more limited effects in practice due to the need for complementary organizational investments, skill development, or business-process redesign (Brynjolfsson, Rock, and Syverson 2021). Because generative AI technologies are only beginning to be used in the workplace, little is currently known about their effects.

III. Our Setting: LLMs for Customer Support

III.A. Customer Support and Generative AI

We study the impact of generative AI in the customer-service industry, an industry at the forefront of AI adoption (Chui et al. 2021). Client interactions play a crucial role in building strong customer relationships and company reputation. However, as in many occupations, customer-service workers vary widely in productivity (Berg et al. 2018; Syverson 2011).

Newer workers require significant training and time to become more productive. Turnover is high: industry estimates suggest that 60% of agents in contact centers leave each year, costing firms $10,000 to $20,000 per agent in the United States (Gretz and Jacobson 2018; Buesing et al. 2020). Consequently, the average supervisor spends a large share of their time coaching new agents (Berg et al. 2018).

Customer service is also a setting where there is high variability in the abilities of individual agents. For example, top-performing customer-support agents are often more effective at diagnosing the underlying technical issue given a customer's problem description. They ask more questions before offering a

solution, spending more time initially to avoid wasting time later solving the wrong problem. Faced with variable productivity, high turnover, and high training costs, firms are increasingly turning to AI tools that might pick up some of these best practices of top performers (Chui et al. 2021).

At a technical level, customer support is well suited for current generative AI tools. Customer-agent conversations can be thought of as a series of pattern-matching problems in which one is looking for a superior sequence of actions. When confronted with an issue such as "I can't log in," an AI/agent must identify the likely problems and their solutions ("Can you check that caps lock is not on?"). At the same time, the agent must be attuned to the customer's emotional response, using reassuring rather than patronizing language ("That wasn't stupid of you at all! I always forget to check that too!"). Because customer-service conversations are widely recorded and digitized, pretrained LLMs can be fine-tuned with examples of successfully and unsuccessfully resolved conversations.

III.B. Data-Firm Background

We work with a company that provides AI-based customer-service support software (hereafter, the AI firm) to study the deployment of its tool at a client firm (hereafter, the data firm). Our data firm is a Fortune 500 company that specializes in business-process software for small and medium-sized businesses in the United States. It uses a variety of chat-based technical support agents, directly and through third-party outsourcing firms. The majority of agents in our sample work from offices in the Philippines, with smaller groups working in the United States and other countries. Across locations, agents are engaged in a fairly uniform job: answering technical support questions from U.S.-based small business owners.

Customer chats are assigned on the basis of agent availability, with no additional prescreening. The questions are often complex and support sessions average 40 minutes, with the majority of the chat spent trying to diagnose the underlying technical problem. The agent's job requires a combination of detailed product knowledge, problem-solving skills, and the ability to handle frustrated customers. While our data firm employed other groups of agents to provide chat-based support for different customer segments—such as self-employed people or larger businesses—

there was no additional sorting for queries related to U.S.-based small businesses.

Our firm measures productivity using three metrics that are standard in the customer-service industry: average handle time (AHT), the average time an agent takes to finish a chat; resolution rate (RR), the share of conversations successfully resolved; and net promoter score (NPS), based on a random post-chat survey that measures customer satisfaction by subtracting the percentage of clients who would not recommend an agent from the percentage who would. A productive agent fields customer chats quickly, while maintaining a high RR and NPS.

Across locations, agents are organized into teams with a manager who provides feedback and training to agents. Once a week, managers hold one-on-one feedback sessions with each agent. For example, a manager might share the solution to a new software bug, explain the implication of a tax change, or suggest how to better manage customer frustration with technical issues. Agents work individually, and the quality of their output does not directly affect others.

Agents employed by the data firm are generally paid a baseline hourly wage and receive bonuses for hitting specific performance targets, such as for chats per hour or RR. Although we lack data on individual pay, the managers we interviewed estimated that performance bonuses accounted for 20% to 40% of a typical agent's total take-home pay.

III.C. AI System Design

The AI system we study is designed to identify conversational patterns that predict efficient call resolution. The system builds on GPT-3 and is fine-tuned on a large data set of customer-agent conversations labeled with various outcomes, such as call resolution success and handling time. Our AI firm also up-weights the value of training chats if the chat was conducted by a top performer when training the AI. Many aspects of successful agent behavior are difficult to quantify, including when to ask clarifying questions, being attentive to customer concerns, deescalating tense situations, adapting communication styles, and explaining complex topics in simple terms. Explicitly training the AI model on text from top performers helps the AI system to pick up on these subtleties in behavior and tone. The AI firm further trains its model using a process similar in spirit to Ouyang et al.

(2022) to prioritize agent responses that express empathy, provide appropriate technical documentation, and limit unprofessional language. This additional training mitigates the potential for inappropriate responses while helping the LLM to distinguish successful behaviors of the top performers, including those they tacitly apply.

Once deployed, the AI system generates two main types of output: (i) real-time suggestions for how agents should respond to customers and (ii) links to the data firm's internal documentation for relevant technical issues. Online Appendix Figure A.I illustrates an example of AI assistance. In the chat window (Panel A), Alex, the customer, describes his problem to the agent. Here, the AI assistant generates two suggested responses (Panel B). The tool has learned that phrases like "I can definitely assist you with this!" and "Happy to help you get this fixed asap" are associated with positive outcomes. Panel C shows a technical recommendation from the AI system: a link to the data firm's internal technical documentation.

Importantly, the AI system we study is designed to augment (rather than replace) human agents. The output is shown only to the agent, who has full discretion over which, if any, AI suggestions to accept. Giving the agent final authority reduces the likelihood that customers receive off-topic or incorrect responses. The system does not provide suggestions when it has insufficient training data for a topic, leaving agents to respond on their own.

IV. Deployment, Data, and Empirical Strategy

IV.A. AI Rollout

AI assistance was introduced after a randomized control trial (RCT) involving a small number of agents. Online Appendix Figure A.II illustrates the timing of the rollout, which primarily took place during fall 2020 and winter 2021. Implementation varied among sites because of limited training resources and the firm's overall budget for AI assistance.

Agents gained access to the AI tool after completing a three-hour online session conducted by the AI firm. To maintain quality and consistency, training sessions were kept small and exclusively led by one of two employees from the AI firm, both of whom had prior contact-center experience and deep knowledge of the AI

system. Since they had other full-time responsibilities, the trainers had to limit the number of sessions they could conduct each week. The timing of sessions was adjusted to accommodate the time zones of the data firm's global workforce.

In addition, because generative AI was costly and relatively untested at that time, the data firm allocated a limited budget for AI deployment. Once the total number of on-boarded agents reached the predefined contractual limit, on-boarding ceased. However, when AI-enabled agents left, their slots were filled by new agents. The capacity and training session scheduling constraints created the variation in AI adoption that we analyze.

Managers at each office oversaw the selection and allocation of agents to training sessions. In interviews, employees of our AI firm reported that managers sought to minimize customer-service disruptions by assigning workers on the same team to different training sessions. After their initial onboarding session, workers received no additional training on using the AI software. At this time, the AI firm's small product management team did not have the capacity to provide ongoing support to the thousands of agents using the tool.

As a result of these considerations, our AI rollout effectively occurred at the individual level. In the same team and same office, individuals would be on-boarded to AI assistance at different times. In October 2020, in a team, the average share of active workers with access to AI assistance was only 5%, growing to 70% in January 2021. While our analysis primarily focuses on the individual adoption dates, we also provide results in Online Appendix Table A.II that instrument individual adoption dates with team-level adoption patterns.

IV.B. Summary Statistics

Table I provides details on the sample characteristics, divided into four groups: all agents (All); agents who never have access to the AI tool during our sample period (Never treated); pre-AI observations for those who eventually get access (Treated, pre); and post-AI observations (Treated, post). In total, we observe the conversation text and outcomes associated with 3 million chats by 5,172 agents. Within this, we observe 1.2 million chats by 1,636 agents in the post-AI period. Most of the agents in our sample (89%) are located outside the United States, mainly in the Philippines. For each agent, we observe their assigned man-

TABLE I
SUMMARY STATISTICS FOR THE SAMPLE OF CUSTOMER-SERVICE AGENTS

Variable	All	Never treated	Treated, pre	Treated, post
Chats	3,006,395	944,848	881,101	1,180,446
Agents	5,172	3,517	1,340	1,636
Number of teams	133	111	80	81
Share U.S. agents	0.11	0.15	0.081	0.072
Distinct locations	25	25	18	17
Average chats per month	128	83	147	188
Average handle time (min.)	41	43	43	35
St. average handle time (min.)	23	24	24	22
Resolution rate	0.82	0.78	0.82	0.84
Resolutions per hour	2.1	1.7	2	2.5
Customer satisfaction (NPS)	79	78	80	80

Notes. This table shows summary statistics of conversations, agent characteristics, and issue resolution rates, customer satisfaction and average call duration. The first column consists of all agents in our sample, the second column includes control agents who never receive AI access. The third column presents statistics for treated agents before they receive AI access, and the fourth column includes treated agents after AI model deployment.

ager, tenure, geographic location, and employer (the data firm or a subcontractor).

We rely on several key performance indicators, or outcome variables, all aggregated at the agent-month level, the most granular level with complete data. Our primary productivity measure is resolutions per hour (RPH), which reflects the number of chats a worker successfully handles per hour. RPH is influenced by several factors, which we also measure individually: AHT for an individual chat; the number of chats an agent handles per hour, which accounts for multitasking (CPH); and RR, the share of chats that are successfully resolved. We measure customer satisfaction using the NPS from post-call surveys. While our main outcome measures are at the agent-month level, some data, like chat duration, are available at a more granular chat level. We construct additional measures of sentiment, topics, and language fluency from chat text.

Our data set includes average handle time and CPH for all agents in our sample. However, subcontractors fail to consistently collect call-quality metrics for all agents. As a result, we only observe our omnibus productivity measure—resolutions per hour—for this smaller subset of agents with call-quality outcomes. Workers may work only for a portion of the year or part-time, so we calculate year-month observations based solely on the periods

that an agent is assigned to chats. Online Appendix J.B includes a more extensive discussion of our sample construction and key variables.

Figure I plots the raw distributions of our outcomes for each of the never-, pre-, and posttreatment subgroups. Several of our main results are readily visible in these raw data. In Panels A–D, we see that posttreatment agents do better along a range of outcomes, relative to both never-treated agents and pretreatment agents. In Panel E, we see no significant differences in surveyed customer satisfaction among treated and non-treated groups.

Focusing on our main productivity measure, Figure I, Panel A and Table I show never-treated agents resolve an average of 1.7 CPH, whereas posttreatment agents resolve 2.5 CPH. Some of this difference may be due to initial selection: treated agents already had higher resolutions per hour prior to AI model deployment (2.0 chats) relative to never-treated agents (1.7). This same pattern appears for CPH (Panel C) and RRs (Panel D): while ever-treated agents appear to be stronger performers at the outset than agents who are never treated, posttreatment agents perform substantially better. Looking instead at AHTs (Panel B), we see a starker pattern: pretreatment and never-treated agents have similar distributions of AHTs, centered at 40 minutes, but post-treatment agents have a lower average handle time of 35 minutes.

These figures, of course, reflect raw differences that do not account for potential confounding factors such as differences in agent experience or differences in selection into treatment. In the next section, we more precisely attribute these raw differences to the impact of AI model deployment.

IV.C. Empirical Strategy

We isolate the causal impact of access to AI recommendations using a standard difference-in-differences regression:

$$y_{it} = \delta_t + \alpha_i + \beta AI_{it} + \gamma X_{it} + \epsilon_{it}. \tag{1}$$

Our outcome variables, y_{it}, are performance measures for agent i in year-month t, with RPH as our main measure of productivity. We measure these outcomes in levels, and report percentage changes off the baseline pretreatment means. Our main variable of interest is AI_{it}, an indicator that equals one if AI assistance is activated for agent i at time t. All regressions include year-month fixed effects, δ_t, to control for common, time-varying factors such as tax season or the end of the business

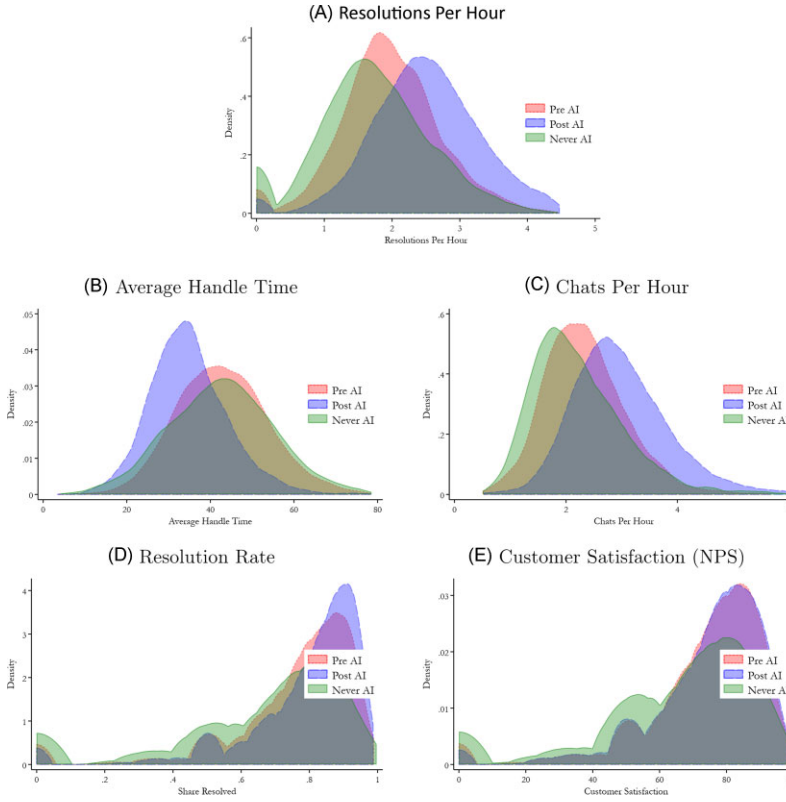

FIGURE I

Raw Productivity Distributions, by AI Treatment

This figure shows the distribution of various outcome measures. We split this sample into agent-month observations for agents who eventually receive access to the AI system before deployment (Pre AI, short dashed line), after deployment (Post AI, long dashed line), and for agent-months associated with agents who never receive access (Never AI, solid line). Our primary productivity measure is resolutions per hour, the number of customer issues the agent is able to successfully resolve per hour. We also provide descriptives for average handle time, the average length of time an agent takes to finish a chat; chats per hour, the number of chats completed per hour incorporating multitasking; resolution rate, the share of conversations that the agent is able to resolve successfully; and net promoter score (NPS), which is calculated by randomly surveying customers after a chat and calculating the percentage of customers who would recommend an agent minus the percentage who would not. All data come from the firm's software systems.

quarter. In our preferred specifications, we include agent fixed effects, $\alpha_{i,}$ to control for time-invariant differences in productivity across agents and time-varying tenure controls X_{it} (specifically, fixed effects for agent tenure in months). In our main specifications, we weight each agent-month equally and cluster standard errors at the agent level to reflect that AI access is rolled out individually, but Online Appendix Tables A.III and A.IV show that our results are robust to alternative weightings and clustering.

A rapidly growing literature has shown that two-way fixed effects regressions deliver consistent estimates only with strong assumptions about the homogeneity of treatment effects and may be biased when treatment effects vary over time or by adoption cohort (Cengiz et al. 2019; de Chaisemartin and D'Haultfœuille 2020; Sun and Abraham 2021; Goodman-Bacon 2021; Callaway and Sant'Anna 2021; Borusyak, Jaravel, and Spiess 2024). For example, workers may take time to adjust to using the AI system, in which case its effect in the first month may be smaller. Alternatively, the on-boarding of later cohorts of agents may be smoother, so that their treatment effects may be larger.

We study the dynamics of treatment effects using the interaction-weighted (IW) estimator proposed in Sun and Abraham (2021). Sun and Abraham (2021) show that this estimator is consistent assuming parallel trends, no anticipatory behavior, and cohort-specific treatment effects that follow the same dynamic profile. Our main difference-in-differences and event-study estimates are similar using robust estimators introduced in de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Borusyak, Jaravel, and Spiess (2024), as well as using traditional two-way fixed effects OLS. In fact, Online Appendix Figure A.VIII shows similar treatment effects across adoption cohorts (e.g., those that received AI access earlier or later, and were thus subject to potentially different versions of the model). We also show that our results are similar when clustering at different levels of granularity (Online Appendix Table A.III) and instrumenting agent adoption with the date on which the first worker in the agent's team received AI access (Online Appendix Table A.II).

TABLE II
MAIN EFFECTS: PRODUCTIVITY (RESOLUTIONS PER HOUR)

Variables	Resolutions/hour (1)	Resolutions/hour (2)	Resolutions/hour (3)
Post AI × Ever treated	0.469***	0.371***	0.301***
	(0.0325)	(0.0318)	(0.0329)
Ever treated	0.110**		
	(0.0440)		
Observations	13,192	12,295	12,295
R-squared	0.249	0.562	0.575
Year month FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
Agent FE	—	Yes	Yes
Agent tenure FE	—	—	Yes
DV mean	2.123	2.176	2.176

Notes. This table presents the results of difference-in-difference regressions estimating the effect of AI model deployment on our main measure of productivity, resolutions per hour, the number of technical support problems resolved by an agent per hour (resolutions/hour). Post AI × Ever treated captures the impact of AI model deployment on resolutions per hour. Column (1) includes agent geographic location and year-by-month fixed effects. Columns (2) and (3) include agent-level fixed effects, and column (3), our preferred specification described by equation (1), also includes fixed effects that control for months of agent tenure. Observations for this regression are at the agent-month level and all standard errors are clustered at the agent level. Section IV.A describes the AI rollout procedure. Robust standard errors are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

V. MAIN RESULTS

V.A. Overall Impacts

Table II examines the impact of the deployment of the AI model on our primary measure of productivity, RPH, using a standard two-way fixed effects model. In column (1), we show that, controlling for time and location fixed effects, access to AI recommendations increases RPH by 0.47 chats, up 23.9% from their pretreatment mean of 1.97. In column (2), we include fixed effects for individual agents to account for potential differences between treated and untreated agents. In column (3), we include additional fixed effects for agent tenure in months to account for time-varying experience levels. As we add controls, our effects fall slightly, so that with agent and tenure fixed effects, we find that the AI deployment increases RPH by 0.30 chats or 15.2%.

Figure II shows the accompanying interactive-weighted event-study estimates of Sun and Abraham (2021) for the impact of AI assistance on RPH. We find a substantial and immediate increase in productivity in the first month of deployment. This

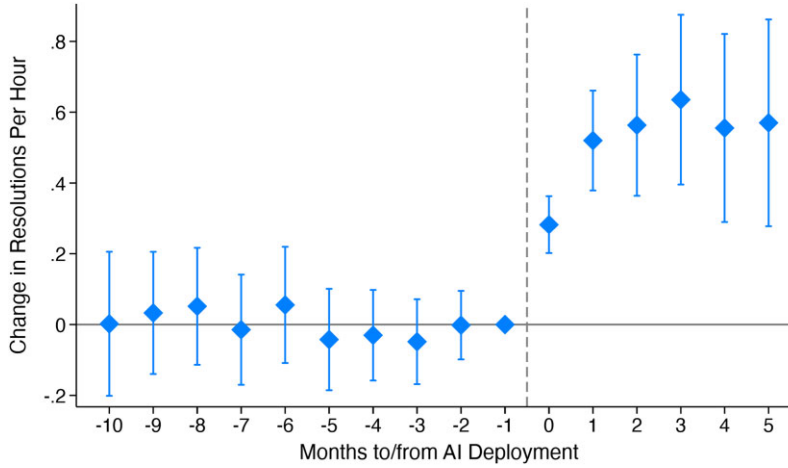

FIGURE II

Event Studies, Productivity

This figure plots the coefficients and 95% confidence intervals from event-study regressions of AI model deployment on our measure of productivity, resolutions per hour, using the Sun and Abraham (2021) interaction-weighted estimator. Our specification follows equation (1) and includes fixed effects for agent, chat year-month, and agent tenure in months. Observations are at the agent-month level, which is the most granular level at which resolutions per hour is available. Robust standard errors are clustered at the agent level. Section IV.A describes the rollout and Online Appendix J.C outlines the regression specification.

effect grows slightly in the second month and remains stable and persistent up to the end of our sample.

In Table III, we report additional results using our preferred specification with fixed effects for year-month, agent, location, and agent tenure. Column (1) documents a 3.7-minute decrease in the average duration of customer chats, an 8.5% decline from the baseline mean of 43 minutes (shorter handle times are considered better). Column (2) indicates a 0.37 unit increase in the number of chats that an agent can handle per hour. Relative to a baseline mean of 2.4, this represents an increase of roughly 15%. Unlike AHT, CPH accounts for the possibility that agents may handle multiple chats simultaneously. The fact that we find a stronger effect on this outcome suggests that AI enables agents to both speed up chats and multitask more effectively.

Table III, column (3) indicates a small, 1.3 percentage point increase in chat RRs. This effect is economically modest and insignificant given a high baseline RR of 82%. We interpret this

TABLE III
MAIN EFFECTS: ADDITIONAL OUTCOMES

Variables	AHT (1)	Chats/hour (2)	Res. rate (3)	NPS (4)
Post AI × Ever treated	−3.746***	0.365***	0.0132	−0.119
	(0.369)	(0.0345)	(0.00882)	(0.524)
Observations	21,839	21,839	12,295	12,541
R-squared	0.591	0.563	0.371	0.526
Year month FE	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes
Agent tenure FE	Yes	Yes	Yes	Yes
DV mean	40.64	2.559	0.822	79.59

Notes. This table presents the results of difference-in-difference regressions estimating the effect of AI model deployment on additional measures of productivity and agent performance. Post AI × Ever treated measures the impact of AI model deployment after deployment on treated agents for average handle time (AHT) in column (1); chats per hour, the number of chats an agent handles per hour in column (2); resolution rate, the share of technical support problems they can resolve in column (3); and net promoter score (NPS), an estimate of customer satisfaction in column (4). Our regression specification, equation (1), includes fixed effects for each agent, chat year-month, and agent months of tenure. Observations for this regression are at the agent-month level and all standard errors are clustered at the agent level. Section IV.A describes the AI rollout procedure. Robust standard errors are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

as evidence that improvements in chat handling do not come at the expense of problem solving on average. Finally, column (4) finds no economically significant change in customer satisfaction, as measured by net promoter scores: the coefficient is −0.12 percentage points and the mean is 80%.

Online Appendix Figure A.III presents the accompanying event studies for additional outcomes. We see immediate effects on AHT (Panel A) and CPH (Panel B), and relatively flat patterns for RR (Panel C) and customer satisfaction (Panel D). We interpret these findings as saying that on average, AI assistance increases productivity without negatively affecting resolution rates and surveyed customer satisfaction.

1. *RCT Analysis.* In August 2020, our data firm conducted a pilot analysis involving approximately 50 workers, with about half randomized into treatment. Unfortunately, we do not have information on the control-group workers that were part of the experiment, so we compare our treated group to all remaining untreated agents. Analysis of the randomized control trial, in Online Appendix Table A.I, shows similar effects on productivity as in our main sample. Online Appendix Figure A.V reports the accompanying event studies for our various outcomes, which are also similar to our main results.

2. *Robustness.* During the rollout process, managers decided which agents to on-board onto the AI system and scheduled when their training would occur. If managers allocated AI access to stronger workers first, our OLS results could overstate the effects of AI. To address this, in Online Appendix Table A.II, we instrument an individual agent's AI adoption date with the first adoption date of the worker's company, office location, and team. The effects on AHT and CPH are essentially identical to those under our main specification. However, resolutions increase by 0.55 CPH, compared with 0.30 in our main finding. We attribute the larger effect on RPH to the fact that this IV approach estimates a significant and larger impact on RRs.

We show similar results using alternative estimators and at different levels of clustering and weighting. Online Appendix Table A.V finds similar results using alternative difference-in-difference estimators introduced in de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Borusyak, Jaravel, and Spiess (2024). Similarly, Online Appendix Figure A.IV reports that our results are similar under alternative event-study estimators: de Chaisemartin and D'Haultfœuille (2020), Callaway and Sant'Anna (2021), Borusyak, Jaravel, and Spiess (2024), and traditional two-way fixed effects. In Online Appendix Table A.III, we show that our standard errors are similar whether clustering at the individual level, team level, or geographic location level. Finally, we explore robustness to alternative weighting. In Online Appendix Table A.IV, we weight agent-month observations by the number of customer chats that a worker conducts. Reweighting generates similar results to our main, equally weighted specifications.

V.B. Heterogeneity by Agent Skill and Tenure

There is substantial interest in the distributional consequences of AI-based technologies. An extensive literature suggests that earlier waves of information technology complemented high-skill workers, with effects on productivity, labor demand, and wage differentials. Together with important changes in relative supply and demand for skilled labor, these changes shaped patterns of wage inequality in the labor market (Goldin and Katz 1998, 2008). Unlike earlier waves of IT, generative AI does not simply execute routine tasks. Instead, as outlined in Section II.B,

AI models identify patterns in data that replicate the behaviors of many types of workers, including those engaged in nonroutine, creative, or knowledge-based tasks. These fundamental technical differences suggest that generative AI may impact different workers in different ways.

1. *Pretreatment Worker Skill.* We explore two components that are important for understanding the distributional consequences of AI adoption: its impacts by agent productivity and tenure. We measure an agent's "skill" using an index incorporating three key performance indicators: call-handling speed, issue RRs, and customer satisfaction. To construct this index, we compute an agent's ranking within its employer company-month for each component productivity measure and then average these rankings into a single index. Then we calculate the average index value over the three months for each agent, to smooth out month-to-month shocks in agent performance. An agent in the top quintile of this productivity index demonstrates excellence across all three metrics: efficient call handling, high issue RRs, and superior customer-satisfaction scores.

Figure III, Panel A shows how our productivity effects vary across workers in each quintile of our skill index, measured in the month before AI access. To isolate the effect of worker skill, our regression specification, available in Online Appendix J.C, includes a set of fixed effects for months of worker tenure. We find that the productivity effect of AI assistance is most pronounced for workers in the lowest skill quintile (leftmost side), who see an increase of 0.5 in RPH, or 36%. In contrast, AI assistance does not lead to any significant change in productivity for the most skilled workers (rightmost side).

In Figure III, Panels B–E we show that less skilled agents consistently see the largest gains across our other outcomes as well. For the highest-skilled workers, we find mixed results: a zero effect on AHT (Panel B); a small but positive effect for CPH (Panel C); and, interestingly, small but statistically significant decreases in RRs and customer satisfaction (Panels D and E).

These results are consistent with the idea that generative AI tools may function by exposing lower-skill workers to the best practices of higher-skill workers. Lower-skill workers benefit because AI assistance provides new solutions, whereas the best performers may see little benefit from being exposed to their own best practices. Indeed, the negative effects along measures of chat

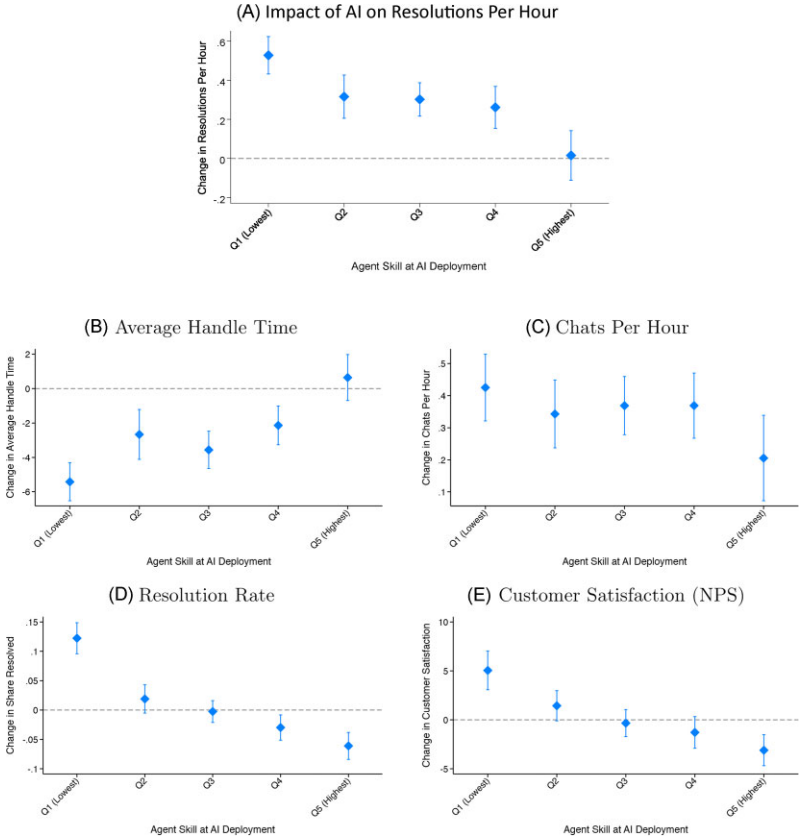

FIGURE III

Heterogeneity of AI Impact, by Skill at Deployment

These figures plot the effects of AI model deployment on five measures of productivity and performance, by pre-deployment worker skill controlling for agent tenure. Agent skill is calculated as the agent's trailing three-month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses for agent performance. In each month and company, agents are grouped into quintiles, with the most productive agents in each firm in quintile 5 and the least productive in quintile 1. Panel A plots the effects on resolutions per hour. Panel B plots the average handle time or the average duration of each technical support chat. Panel C graphs chats per hour, or the number of chats an agent can handle per hour. Panel D plots the resolution rate, and Panel E plots net promoter score, an average of surveyed customer satisfaction. All specifications include fixed effects for the agent, chat year-month, and months of tenure. Robust standard errors are clustered at the agent level. The regression specifications are available in Online Appendix J.C and results in Online Appendix Table A.VI.

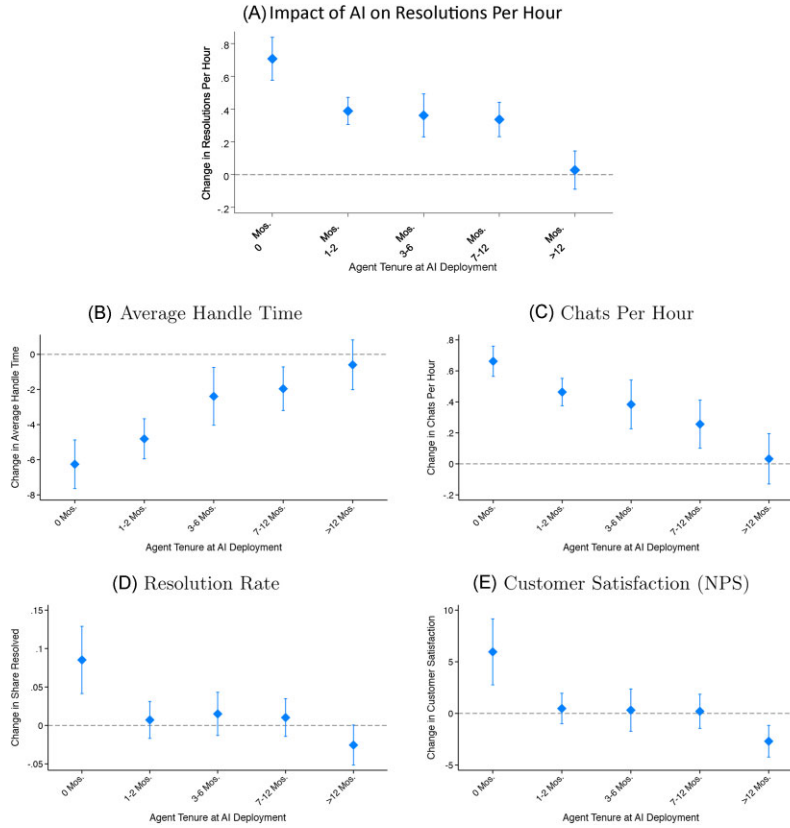

FIGURE IV

Heterogeneity of AI Impact, by Tenure at Deployment

These figures plot the effects of AI model deployment on five measures of productivity and performance by pre-AI worker tenure, defined as the number of months an agent has been employed when they receive access to the AI model. Panel A plots the effects on resolutions per hour. Panel B plots the average handle time or the average duration of each technical support chat. Panel C graphs chats per hour, or the number of chats an agent can handle per hour. Panel D plots the resolution rate, and Panel E plots net promoter score, an average of surveyed customer satisfaction. All specifications include fixed effects for the agent, chat year-month, and months of tenure. Robust standard errors are clustered at the agent level. The regression specifications are available in Online Appendix J.C and results in Online Appendix Table A.VI.

quality—RR and customer satisfaction—suggest that AI recommendations may distract top performers or lead them to choose the faster or less cognitively taxing option (following suggestions) rather than taking the time to come up with their own responses. Addressing this outcome is potentially important because the conversations of top agents are used for ongoing AI training.

Our results could be driven by mean reversion: agents who performed well just before AI adoption may see a natural decline in their productivity afterward, while lower-performing agents may rebound. To address this concern, we plot raw resolutions per hour in event time, graphed by skill tercile at AI treatment in Online Appendix Figure A.VII. If mean reversion were driving our observed effects, we would expect to see a convergence of productivity levels after treatment, with top tercile agents showing decreased performance and the least skilled agents demonstrating improved output. However, our analysis reveals a consistent linear increase in productivity across all skill levels after AI implementation, with no strong evidence of mean reversion, suggesting that productivity gains are attributable to AI assistance.

2. *Pretreatment Worker Experience.* We repeat our previous analysis for agent tenure to understand how the treatment effects of AI access vary by worker experience. To do so, we divide agents into five groups based on their months of tenure at the time the AI model is introduced. Some agents have less than a month of tenure when they receive AI access, whereas others have more than a year of experience. To isolate the effect of worker tenure, this analysis controls for the worker skill quintile at AI adoption, with the regression specification in Online Appendix J.C.

In Figure IV, Panel A, we see a clear monotonic pattern in which the least experienced agents see the greatest gains in RPH. Agents with less than one month of tenure improve by 0.7 RPH, with larger effects for less experienced workers. In contrast, we see no effect for agents with more than a year of tenure.

In Figure IV, Panels B–E, we report results for other outcomes. In Panels B and C, we see that AI assistance generates large gains in call-handling efficiency, measured by AHTs and CPH, respectively, among the newest workers. In Panels D and E, we find positive effects of AI assistance on chat quality, as measured by RRs and customer satisfaction, respectively. For the most experienced workers, we see modest positive effects for AHT (Panel B), positive but statistically insignificant effects on CPH

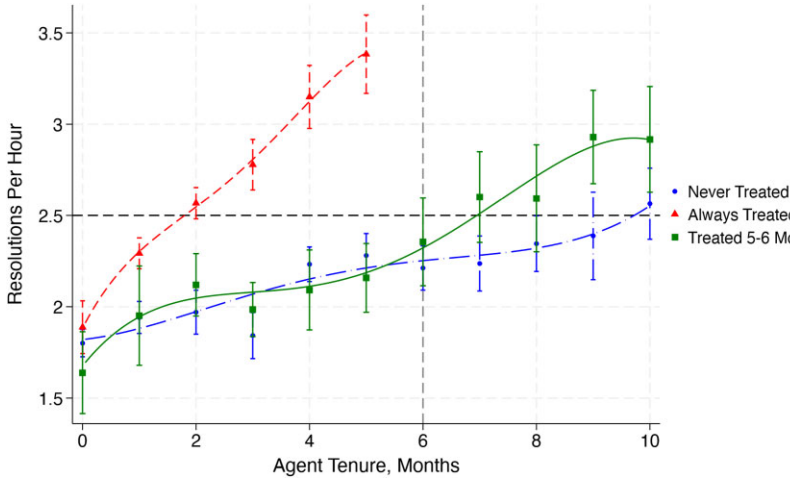

FIGURE V

Experience Curves by Deployment Cohort

This figure plots the relationship between productivity and job tenure. The short-dashed line plots the performance of always-treated agents, those who have access to AI assistance from their first month on the job. The long-dashed line plots agents who are never treated. The solid line plots agents who spend their first four months of work without the AI assistance, and gain access to the AI model during their fifth month on the job. Ninety-five percent confidence intervals are shown. Observations are at the agent-month level.

(Panel C), and small but statistically significant negative effects for measures of call quality and customer satisfaction (Panels D and E).

Overall, these patterns are very similar to our findings for agent skill, even though our regressions are designed to isolate the distinct roles of skill and experience—our skill regressions control for experience and vice versa. This suggests that even in the same task, access to AI systems disproportionately improves the performance of both novice and less skilled workers.

3. *Moving Down the Experience Curve.* To further explore how AI assistance affects newer workers, we examine how worker productivity evolves on the job. In Figure V, we plot productivity variables by agent tenure for three distinct groups: agents who never receive access to the AI model ("never treated"), those who have access from the time they join the firm ("always treated"),

and those who receive access in their fifth month with the firm ("treated 5 mo.").

We see that all agents begin at around 1.8 RPH. Never-treated workers (long-dashed blue line; color version available online) slowly improve their productivity with experience, reaching approximately 2.5 RPH 8–10 months later. In contrast, workers who always have access to AI assistance (short-dashed red line) increase their productivity to 2.5 RPH after only two months and continue to improve until they are resolving more than 3 CPH after five months of tenure. Comparing just these two groups suggests that access to AI recommendations helps workers move more quickly down the experience curve and reduces ramp-up time.

The final group in Panel A tracks workers who begin their tenure without access to AI assistance but who receive access after five months on the job (solid green line). These workers initially improve at the same rate as never-treated workers, but after gaining AI access in month 5, their productivity begins to increase more rapidly, following the same trajectory as the always-treated agents. These findings demonstrate that AI assistance not only accelerates ramp-up for new workers but also increases the rate at which even experienced workers improve in their roles.

In Online Appendix Figure A.VI, we plot these curves for other outcomes. We see clear evidence that the experience curve for always-treated agents is steeper for handle time, chats per hour, and resolution rates (Panels A–C). Panel D follows a similar but noisier pattern for customer satisfaction. Across many of the outcomes that we examine, agents with two months of tenure and access to AI assistance perform as well as or better than agents with more than six months of tenure who do not have access. AI assistance alters the relationship between on-the-job productivity and time, with potential implications for how firms might value prior experience, or approach training and worker development.

VI. Adherence, Learning, Topic Handling, and Conversational Change

We conduct the following analyses to understand better the mechanisms driving our main results. We examine patterns in how workers of varying skills engage AI recommendations. We look at how exposure to AI helps agents master their jobs, sharp-

ening their diagnostic skills and language fluency, and how AI assistance influences the communication patterns of higher- and lower-skill workers.

VI.A. Adherence to AI Recommendations

The AI tool we study makes suggestions, but agents are ultimately responsible for what they say to the customer. Thus far, our analysis evaluates the effect of AI assistance, irrespective of the frequency with which users adhere to its suggestions. Here we examine how closely agents adhere to AI recommendations and document the association between adherence and returns to adoption. We define adherence as the proportion of AI suggestions an agent typically adopts, when an AI suggestion is generated, omitting messages where the AI does not make suggestions. The AI company considers an agent to have adhered when they either directly copy the AI's proposed text or manually enter highly similar content. To gauge initial adherence, we classify each treated agent into a quintile based on their level of adherence during their first month using the AI tool.

Figure VI, Panel A shows the distribution of average agent-month-level adherence for our post-AI sample, weighted by the log number of AI recommendations provided to that agent in that month. The average adherence rate is 38%, with an interquartile range of 23%–50%: agents frequently disregard recommendations. In fact, the share of recommendations followed is similar to the share of other publicly reported numbers for generative AI tools; a study of GitHub Copilot reports that individual developers use 27%–46% of code recommendations (Zhao 2023). Such behavior may be appropriate, given that AI models may make incorrect or irrelevant suggestions. In Online Appendix Figure A.IX, we further show that the variation in adherence is similar within locations and teams, indicating that it is not driven by some organizational segments being systematically more supportive than others.

Figure VI, Panel B shows that returns to AI model deployment are higher when agents follow recommendations. We measure this by dividing agents into quintiles based on the share of AI recommendations they follow in the first month of AI access. Following equation (3) in the Online Appendix, we separately estimate the impact of AI assistance for each group, including year-month, agent, and agent-tenure fixed effects.

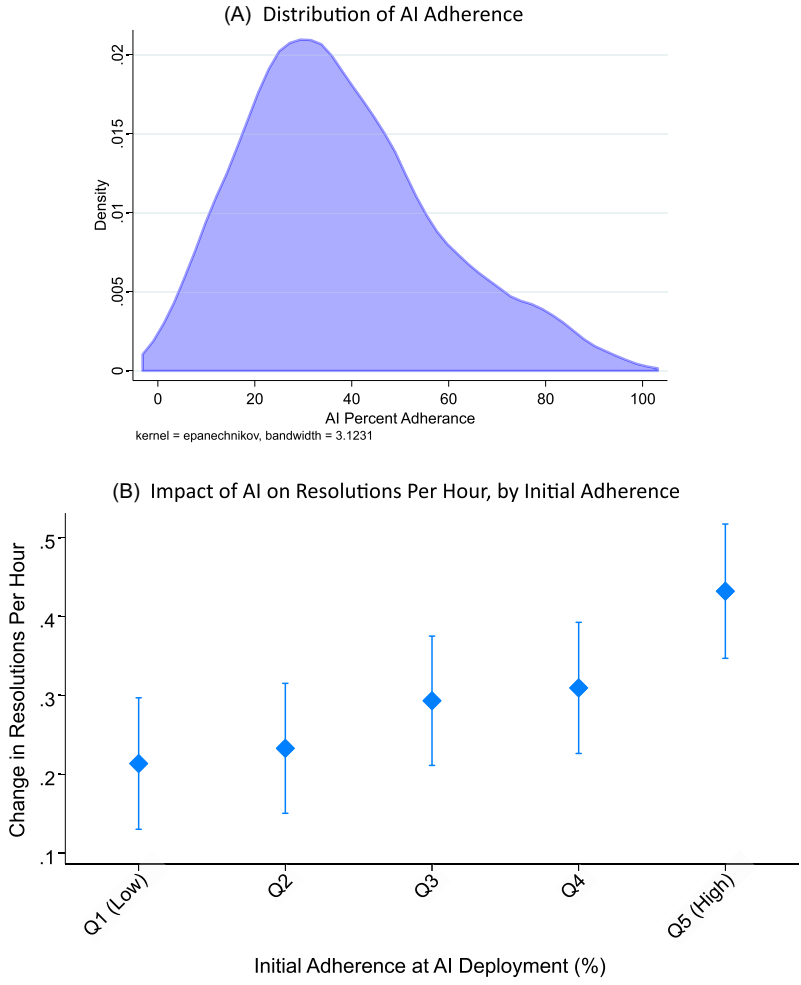

FIGURE VI

Heterogeneity of AI Impact, by AI Adherence

Panel A plots the distribution of AI adherence, averaged at the agent-month level, weighted by the log of the number of AI recommendations for that agent-month. Panel B shows the effect of AI assistance on resolutions per hour, by agents grouped by their initial adherence, defined as the share of AI recommendations they followed in the first month of treatment. The regression, outlined in Online Appendix J.C, is run at the agent-month level and includes fixed effects for agent, chat year-month, and agent tenure in months. Standard errors are clustered at the agent level. Results are in Online Appendix Table A.VII.

We find a steady and monotonic increase in returns by agent adherence. Among agents in the lowest quintile, we still see a 10% gain in productivity, but for agents in the highest quintile, the estimated impact is over twice as high, close to 25%. Online Appendix Figure A.X shows the results for our other four outcome measures. The positive correlation between adherence and returns holds most strongly for AHT (Panel A) and CPH (Panel B), and more noisily for RR (Panel C) and customer satisfaction (Panel D).

Our results are consistent with the idea that there is a treatment effect of following AI recommendations on productivity. We note that this relationship could also be driven by other factors: selection (agents who choose to adhere are more productive for other reasons) or selection on gains (agents who follow recommendations are those with the greatest returns). We consider the worker's revealed preference: do they continue to follow AI recommendations over time? If our results were driven purely by selection, we would expect workers with low adherence to continue having low adherence, since it was optimal for them to do so.

Online Appendix Figure A.XI plots the evolution of AI adherence over time, for various categories of agents. Panel A begins by considering agents who differ in their initial AI compliance, which we categorize based on terciles of AI adherence in the first month of model deployment (initial adherence). We see that compliance either stays stable or grows over time. The initially most compliant agents continue to comply at the same rates (just above 50%). The initially least compliant agents increase their compliance over time: those in the bottom tercile initially follow recommendations less than 20% of the time, but by month five their compliance rates have increased by over 50%.

Panel B divides workers up by tenure at the time of AI deployment. More senior workers are initially less likely to follow AI recommendations: 30% for those with more than a year of tenure compared with 37% for those with less than three months of tenure. Over time, all workers increase adherence, with more senior workers doing so faster, and the groups converge five months after deployment.

In Panel C, we show the same analysis by worker skill at AI deployment. We see that compliance rates are similar across skill groups, and all groups increase their compliance over time. In Online Appendix Figure A.XII we show that these patterns are robust to examining within-agent changes in adherence (that is,

adherence rates residualized by agent fixed effects), suggesting that increases in adherence over time are not driven exclusively by less adherent agents leaving.

The results in Online Appendix Figures A.XI and A.XII are consistent with agents, particularly those who are initially more skeptical, coming to value AI recommendations over time. We note that high-skill agents increase their adherence as quickly as their lower-skill peers, even though their productivity gains are smaller and—in the case of some quality measures—even negative. This suggests an alternate possibility: some agents may be overrelying on AI recommendations beyond what is optimal in the long run. Top agents, in particular, may see little additional value in taking the time to provide the highest-quality inputs when an adequate AI suggestion is readily available. High AI adherence in the present may then reduce the quality or diversity of solutions used for AI training in the future. However, in the short run, our analysis finds no evidence that the model is declining in quality over our sample period. In Online Appendix Figure A.VIII, we show that workers who received later access to the AI system—and therefore to a more recently updated version—had similar first-month treatment effects as those who received access to an earlier version of the model.

VI.B. Worker Learning

A key question raised by our findings so far is whether these improvements in productivity and changes in communication patterns reflect durable changes in the human capital of workers or simply their growing reliance on AI assistance. In the latter case, the introduction of AI assistance could actually lead to an erosion in human capital, and we would expect treated workers to be less able to address customer questions if they are no longer able to access AI assistance. For example, research in cognitive science has shown that individuals learn less about spatial navigation when they follow GPS directions, relative to using a traditional map (Brügger, Richter, and Fabrikant 2019).

We examine how workers perform during periods in which they are not able to access AI recommendations due to technical issues at the AI firm. Outages occur occasionally in our data and can last anywhere from a few minutes to a few hours. During an outage, the system fails to provide recommendations to some (but not necessarily all) workers. For example, outages may affect

agents who log into their computers after the system crashes, but not those working at the same time who had signed in earlier. Outages may also affect workers using one physical server but not another. Our AI firm tracks the most significant outages to perform technical reviews of what went wrong. We compile these system reports to identify periods in which a significant fraction of chats are affected by outages.

Online Appendix Figure A.XIII shows an example of such an outage, which occurred on September 10, 2020. The y-axis plots the share of posttreatment chats (e.g., those occurring after the AI system has been deployed for a given agent) for which the AI software does not provide any suggestions, aggregated to the hour level. The x-axis tracks hours in days leading up to and following the outage event (hours with fewer than 15 posttreatment chats are plotted as zeros for figure clarity). During non-outage periods, the share of chats without AI recommendations is typically 30% to 40%, reflecting that the AI system does not normally generate recommendations in response to all messages. On the morning of September 10, however, we see a notable spike in the number of chats without recommendations, increasing to almost 100%. Records from our AI firm indicate that this outage was caused by a software engineer running a load test that crashed the system.

Figure VII examines the impact of access to the AI system for chats that occur during and outside these outage periods. These regressions are estimated at the individual-chat level to precisely compare conversations that occurred during outage periods with those that did not. Because we do not have information on chat resolution at this level of granularity, our main outcome measure is chat duration. Panel A considers the effect of AI assistance using only post-adoption periods in which the AI system is not affected by a software outage. Consistent with our main results, we see an immediate decline in the duration of individual chats by approximately 10% to 15%.

In Panel B, we use the same pretreatment observations, but now restrict to post-adoption periods that are affected by large outages. We find that even during outage periods when the AI system is not working, AI-exposed agents continue to handle calls faster (equivalent to 15%–25% declines in chat duration). Because AI outages are rare, our estimates are noisy and could reflect differences in the types of chats that are seen during outage periods than during non-outage periods. However, when focusing

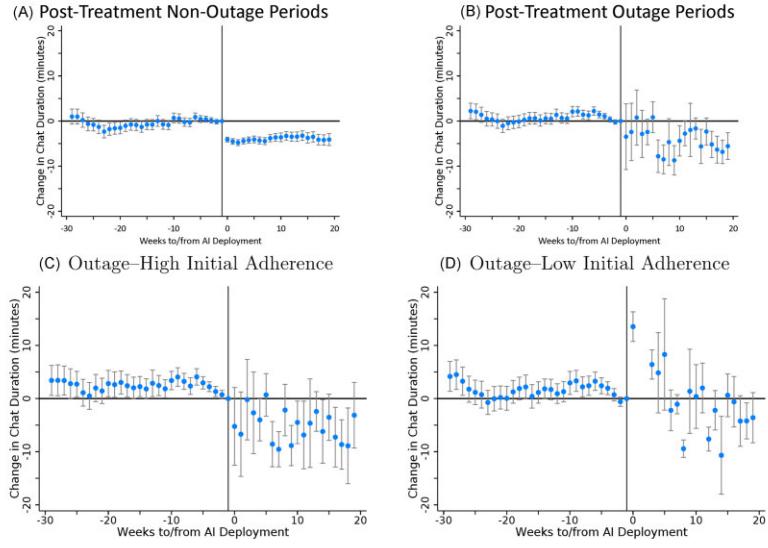

FIGURE VII

Chat Duration during AI System Outages

These figures plot event studies for the effect of AI system rollout of chat duration at the individual-chat level. Panel A restricts to posttreatment chats that do not occur during any period where there is a AI system outage. Panel B restricts to posttreatment chats that only occur during a large system outage. Panels C and D focus on outage only post-periods. Panel C restricts to only chats generated by ever-treated agents who with high initial AI adherence (top tercile), while Panel D restricts to agents with low initial adherence (bottom tercile). Agents who are never treated are excluded from this analysis. The regressions are run at the chat level with agent, year-month, and tenure fixed effects with standard errors clustered at the agent level. Regression results are reported in Online Appendix Table A.VIII.

on the size of estimated effects over time, an interesting pattern emerges. Rather than declining immediately post-adoption and staying largely stable as we see in Panel A for non-outage periods, Panel B shows that the benefit of exposure to AI assistance increases with time during outage periods. That is, if an outage occurs one month after AI adoption, workers do not handle the chat much more quickly than their pre-adoption baseline. If an outage occurs after three months of exposure to AI recommendations, workers handle the chat faster—even though they are not receiving direct AI assistance in either case.

Panel B highlights the potential scope for improving existing employee training practices. Prior to AI assistance, training was limited to brief weekly coaching sessions where managers reviewed select conversations and provided feedback. By necessity, managers can only provide feedback on a small fraction of the conversations an agent conducts. Moreover, because managers are often short on time and may lack training, they often simply point out weak metrics ("you need to reduce your handling time") rather than identifying strategies for how an agent could better approach a problem ("you need to ask more questions at the beginning to diagnose the issue better"). Such coaching can be ineffective and counterproductive to employee engagement (Berg et al. 2018). In contrast, AI assistance offers workers specific, real-time, actionable suggestions, potentially addressing a limitation of traditional coaching methods.

To better understand how learning might occur, in Figure VII, Panels C and D, we divide our main study of outage events by the initial adherence of the worker to AI, as described in Section VI.A. When a worker chooses not to follow a particular AI recommendation, they miss the opportunity to observe how the customer might respond. AI suggestions may prompt workers to communicate in ways that differ from their natural style, such as by expressing more enthusiasm or empathy or by frequently pausing to recap the conversation. Workers who do not try out these recommendations may never realize that customers could react positively to them.

Panel C reveals that workers with high initial adherence to AI recommendations experience significant and rapid declines in chat processing times, even during outages, relative to their pre-adoption baseline. In contrast, Panel D shows no such improvement for workers who frequently deviate from AI suggestions; they see no reduction in chat times during outage periods, even after prolonged AI access.

These findings suggest that workers learn more by actively engaging with AI suggestions and observing firsthand how customers respond. These findings are consistent with other evidence from education that higher adherence and engagement with LLM-generated responses positively affected learning (Kumar et al. 2023). In addition to directly improving productivity, exposure to AI assistance could supplement existing on-the-job training programs.

VI.C. Handling Routine and Nonroutine Topics

In addition to varying by the characteristics of the worker, the effect of AI assistance could depend on the types of problems it is asked to resolve. Agents encounter customer questions that range from common requests for help on-boarding an employee or changing a password to less common issues such as setting up wage garnishments in child support cases or ensuring compliance with international tax treaties. We examine how the effect of AI assistance varies between more and less routine customer problems. We use Gemini, an LLM developed by Google DeepMind, to classify the interactions into topic categories. The details of this process, along with our human validation of the LLM classification process, are described in Online Appendix J.B (Gemini Team 2023).

Online Appendix Figure A.XIV reports the distribution of conversation topics in our data set. Unsurprisingly, we observe a small number of frequent issues, accompanied by a long tail of less common problems. Specifically, the two most prevalent topics—payroll and taxes, and account access and management—make up half of all conversations, and the top 16 topics represent over 90% of all chats.

To evaluate the impact of AI assistance based on the frequency of customer inquiries, we categorize conversations into four distinct groups. The Payroll/Account category, comprising 50% of all chats, includes inquiries related to payroll, taxes, and account access and management. The next 25% of chats covers five categories, including those dealing with bank transfers or managing subscriptions. The following 15% of chats encompass nine topics, and the final 10% of chats consists of all the remaining topics. Our regression, in Online Appendix J.C, is conducted at the chat level, with a focus on chat duration.

Figure VIII, Panel A shows the average treatment effect of AI assistance based on how common the inquiry is. The pattern is non-monotonic and suggests that AI assistance has the greatest effect on workers' ability to handle problems that are moderately rare. Workers with access to AI assistance handle the most routine problems—payroll and account management—about four to five minutes faster, which corresponds to an approximately 10% decrease from the pretreatment mean duration for these topics. We see the largest decline, five to six minutes, for issues that are in the 75th–90th percentiles of topic rarity, corresponding to a

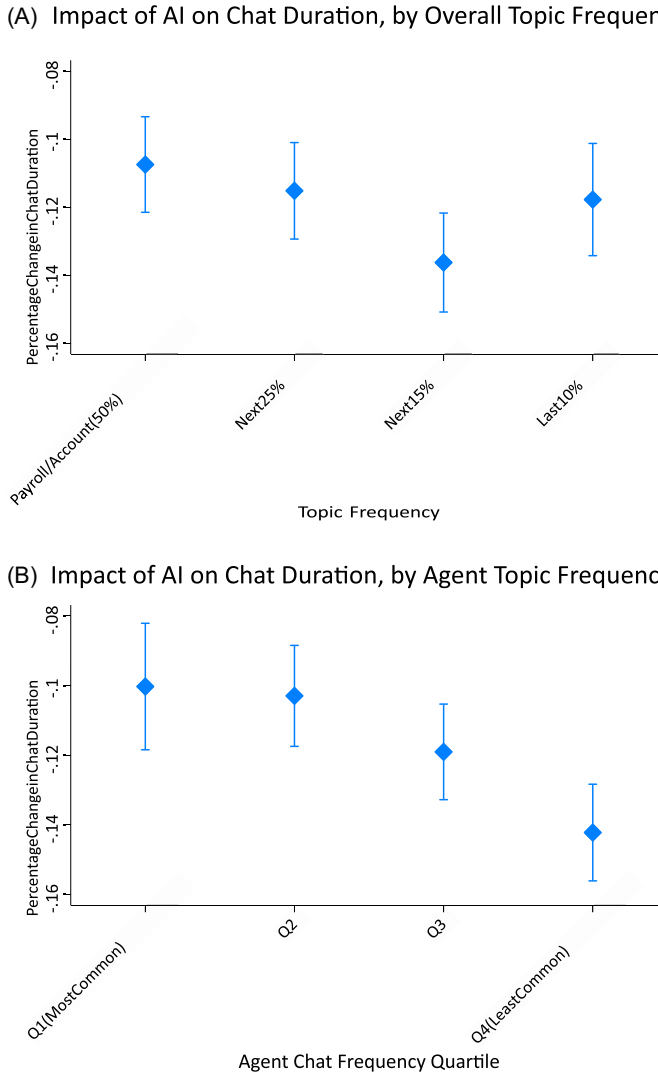

FIGURE VIII

AI Impact by Chat Topic

Panel A shows the impact of AI assistance on chat duration in minutes for each category of conversation-topic frequency relative to the category's pre-AI mean. Data is at the chat level and the regressions control for topic frequency, year-month fixed effects, agent fixed effects, and fixed effects in months of agent tenure. Panel B shows the impact of AI assistance grouped instead by topic frequency as encountered by the individual agent relative to the pre-AI mean. Online Appendix J.C details the regression specification and topic-category construction and results are in Online Appendix Table A.IX.

14% reduction from the pretreatment means for those topics. Finally, we see a smaller four-minute or 11% decrease for the most rare problems. It should be noted that the system does not provide suggestions at all when there was insufficient training data.

These results highlight the difference between the technical quality of an AI system and its potential productivity effects in real-world settings. AI models generally perform better when trained on large data sets, which provide diverse examples and richer contextual information. Such data sets enable the model to learn more robust and generalizable patterns while reducing the risk of overfitting (Halevy, Norvig, and Pereira 2009). Consequently, we might expect an AI system to function best when dealing with routine problems, where training data are abundant.

The value of AI systems is less straightforward when they are used to complement human workers. Customer-service agents, especially those dealing with common issues, are specifically trained to address these routine problems and become most experienced answering them. For example, even novice workers are likely to know how to reset a customer's password. In such cases, access to even high-quality AI assistance may not have a large complementary effect. Rather, as our findings suggest, the impact of an AI system on workplace productivity depends critically on its capabilities relative to workers' baseline skills. The greatest productivity gains may occur not where the AI system is most capable in absolute terms, but where its capabilities most effectively complement or exceed those of human workers.

In our setting, the heterogeneous effect of AI access appears to reflect both factors. AI access has the smallest reduction in handle time for problems where human agents are already well trained (very routine problems) or where its training data may be sparse (very rare problems). We see the largest improvements in the handle time for moderately uncommon problems. The AI system is likely to have enough training data to assess these problems, while individual agents are less likely to have had much firsthand experience. For example, the AI-recommended links to potentially relevant technical documentation may be particularly valuable for the types of cases where agents otherwise would need to search for an answer.

To examine the role of agent-specific experience, Figure VIII, Panel B plots the effect of AI assistance on chat duration by quartiles of topic frequency with respect to an individual agent, controlling for the overall frequency of a problem. AI assistance re-

duces conversation times by 15% for the least common problems compared with 10% for the most common. Once we control for a topic's overall frequency, we find a monotonic relationship between agent-specific exposure to a problem and the impact of AI. That is, holding constant the AI model's exposure to a problem, the effect of AI assistance is greatest for problems that a specific agent is least exposed to. Although AI in isolation may be most effective where training data is most plentiful, the marginal value of AI assistance appears to be highest where humans have a greater need for AI input.

VI.D. Conversational Style

1. *English Fluency.* The ability to communicate in clear, idiomatic English is crucial for customer satisfaction and the job performance of contact workers serving U.S. customers. In our data, 80% of the agents are based in the Philippines, where many residents are fluent English speakers for various cultural and historical reasons. However, cultural differences and language nuances occasionally lead to misunderstandings or a sense of disconnect, even when an agent's technical language skills are strong. We assess how AI assistance influences workers' ability to communicate clearly.

We measure the proficiency of text in two ways: its comprehensiblity and its *native fluency*. The comprehensibility score assesses whether the agent produces text that is cogent and easy to understand, using a scale of 1–5, where 1 indicates "very difficult to comprehend" and 5 signifies "very fluent and easily understandable, with no significant errors." In contrast, native fluency focuses on whether the text was likely to have been produced by a native speaker of American English. Native fluency is based on the Interagency Language Roundtable "functionally native" proficiency standard. The native-fluency score is also on a five-point scale where 1 indicates a writer is "definitely not a native American English speaker" and 5 indicates they definitely are. For instance, "I could care less" is grammatically incorrect but a common English-language expression. On the other hand, Filipino agents often use the greeting "to have a blessed day," which is grammatically correct, but not a common greeting in the United States. We use Gemini to score agents' text in each conversation along these

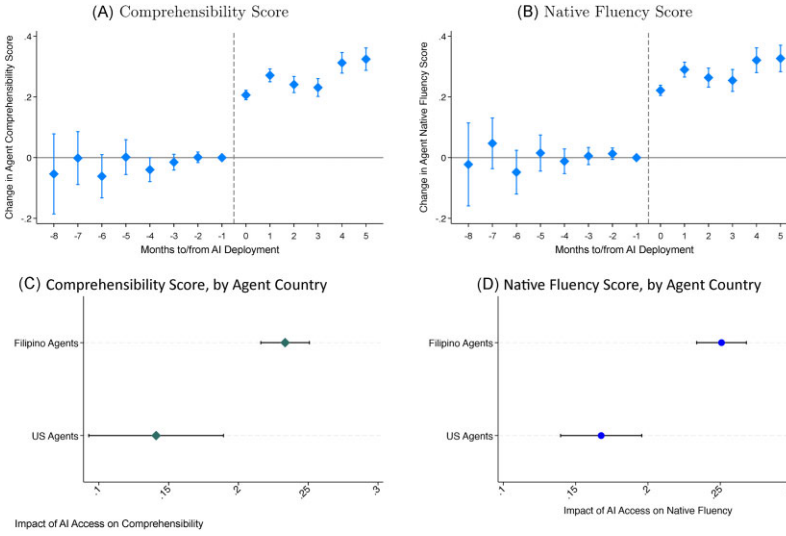

FIGURE IX

The Impact of AI on Language Skills

These figures show the impact of AI access on scores of agent comprehensibility in Panel A and native fluency in Panel B. Observations for this regression are at the agent-chat level, aggregated to the agent-month level. Regressions follow equation (1) and include agent, chat year-month, and months of agent-tenure fixed effects. Robust standard errors are clustered at the agent level in Panels A and B and by agent location in Panels C and D. For more details on the construction of the comprehensibility and native-fluency scores, refer to Online Appendix J.B. Regression results are in Online Appendix Table A.X.

two dimensions. For more information on our specific approach, prompts, and validation tests, see Online Appendix J.B.5.

The general level of both comprehensibility and native fluency is high. Before having AI access, the interquartile range of comprehensibility scores was 4.28–4.67; for native fluency it was 4.26–4.65. Despite this high baseline level, we find clear evidence that access to AI assistance increases proficiency scores. Online Appendix Figure A.XV presents the raw pre- and post-AI distribution of comprehensibility and native-fluency scores for never-treated workers, pretreatment workers, and posttreatment workers. The never-treated and pretreatment workers have identical distributions, but we see markedly higher scores for post-treatment workers. In Figure IX, Panels A and B we report the accompanying event studies, which show that AI access leads to large improvements in both comprehensibility and native fluency.

Finally, in Figure IX, Panels C and D, we report separate coefficients for U.S.- and Philippines-based workers. We see a positive effect for all workers, but a larger improvement for workers based in the Philippines.

2. *Textual Convergence.* The foregoing analysis focuses on an important but narrow aspect of how workers communicate. To gain a broader understanding of AI's influence on communication patterns, we examine how the text produced by workers evolves over time: do they change how they write relative to their pre-AI baseline, and does AI access affect the relative communication patterns of high- and low-skill workers? Because tacit knowledge is, by definition, not something that can be codified as a set of rules, we examine the overall textual similarity of conversations using textual embeddings, rather than looking for the presence of specific formulaic phrases (Hugging Face 2023).

Online Appendix Figure A.XVI, Panel A plots the evolution of agents' communication over time, before and after access to AI assistance. We compute the cosine similarity of agents' text in each given event-time week to their own chats from the month before AI deployment (week −4 to week −1). Cosine similarity runs from zero to one, with zero meaning two pieces of text are orthogonal (when represented as semantic vectors), and one indicating exact semantic similarity.

Before deploying AI, the similarity between a worker's own text from month to month is stable at 0.67, which reflects consistency in an individual agent's language use, while also capturing differences in the topics and customers that she faces. After AI deployment, the similarity of agents' text drops. The drop is equivalent to about half of a standard deviation of within-agent cosine similarity across the pre-period. This is consistent with the idea that AI assistance changes the content of agents' messages, rather than merely leading workers to type the same content but faster. Online Appendix Figure A.XVI, Panel B plots the average change in textual content separately by pre-AI worker skill. Lower-skill agents experience greater textual change after AI adoption, relative to top performers.

We find across-worker changes in communication changes with AI access. Online Appendix Figure A.XVI, Panel C plots the cosine similarity between high- and low-skill agents at specific moments in calendar time, separately for workers without (blue dots) and with (red diamonds) access to AI assistance. For

non-AI users, we define skill levels based on monthly quintiles of our skill index. For AI users, we use skill quintiles at the time of AI deployment. Without AI, high- and low-productivity workers show a moderate level of similarity in their language use, with an average cosine similarity between high and low workers of 0.55. This similarity remains stable over time, suggesting that there are no divergent trends between skill groups that do not have access to AI assistance. After AI adoption, text similarity between high- and low-skill workers begins increasing, from 0.55 to 0.61. Although this change may seem modest, it represents a substantial narrowing of language gaps, given that the average similarity of a high-skill worker's own pre- and post-AI text is only 0.67. The change is equivalent to half of a standard deviation of the average high- and low-worker textual similarity.

Taken together, the patterns in Online Appendix Figure A.XVI are consistent with AI assistance leading to more pronounced changes in how lower-skill workers communicate and ultimately to their communicating more like high-skill workers. We caution that changes in agent text can reflect many factors that are not directly related to a worker's style or tacit skills, such as changes in conversation topics driven by customers. As a result, this analysis is only suggestive.

VII. Effects on the Experience of Work

VII.A. Customer Sentiment

Qualitative studies suggest that working conditions for contact-center agents can be unpleasant. Customers often vent their frustrations to anonymous service agents; in our data, we see regular instances of swearing, verbal abuse, and "yelling" (typing in all caps). The stress associated with this type of emotional labor is often cited as a key cause of burnout and attrition among customer-service workers (Lee 2015).

Access to AI assistance may affect how customers treat agents, but in theory, the direction and magnitude of these effects are ambiguous. AI assistance may improve the tenor of conversations by helping agents set customer expectations or resolve their problems more quickly. Alternatively, customers may become more frustrated if AI-suggested language feels "corporate" or insincere.

TABLE IV
EXPERIENCE OF WORK

Variables	Mean customer sentiment (1)	Mean agent sentiment (2)	Share req. manager (3)
Post AI × Ever treated	0.177***	0.0198***	−0.00875***
	(0.0116)	(0.00599)	(0.00201)
Observations	21,218	21,218	21,839
R-squared	0.485	0.596	0.482
Year month FE	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes
Agent tenure FE	Yes	Yes	Yes
DV mean	0.141	0.896	0.0377

Notes. This table presents the results of difference-in-difference regressions estimating the effect of AI model deployment on measures of conversation sentiment and requests to speak to a manager (Share req. manager). Our regression specification, equation (1), includes fixed effects for each agent, chat year-month, and months of agent tenure. Observations for these regressions are at the agent-month level and all standard errors are clustered at the agent level. Measures of customer sentiment are created from conversation transcripts using SiBERT and are aggregated to the agent-month level. Online Appendix J.B elaborates on sentiment construction and Section IV.A describes the AI rollout procedure. Robust standard errors are in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

To assess this, we capture the affective nature of both agent and customer text, using sentiment analysis (Mejova 2009). For this analysis, we use SiEBERT, an LLM that is fine-tuned for sentiment analysis using a variety of data sets, including product reviews and social media posts (Hartmann et al. 2023). Sentiment is measured on a scale from −1 to 1, where −1 indicates negative sentiment and 1 indicates positive. In a given conversation, we compute separate sentiment scores for both agent and customer text. We aggregate these chat-level variables into a measure of average agent sentiment and average customer sentiment for each agent-year-month.

Figure X, Panels A and B consider how sentiment scores respond following the rollout of AI assistance. In Panel A, we see an immediate and persistent improvement in customer sentiment. This effect is large: according to Table IV, column (1), access to AI improves the mean customer sentiments (averaged over an agent-month) by 0.18 points, equivalent to half of a standard deviation. In Panel B, we see no detectable effect for agent sentiment, which is already very high at baseline. Table IV, column (2) indicates agent sentiments increase by only 0.02 points or about 1% of a standard deviation.

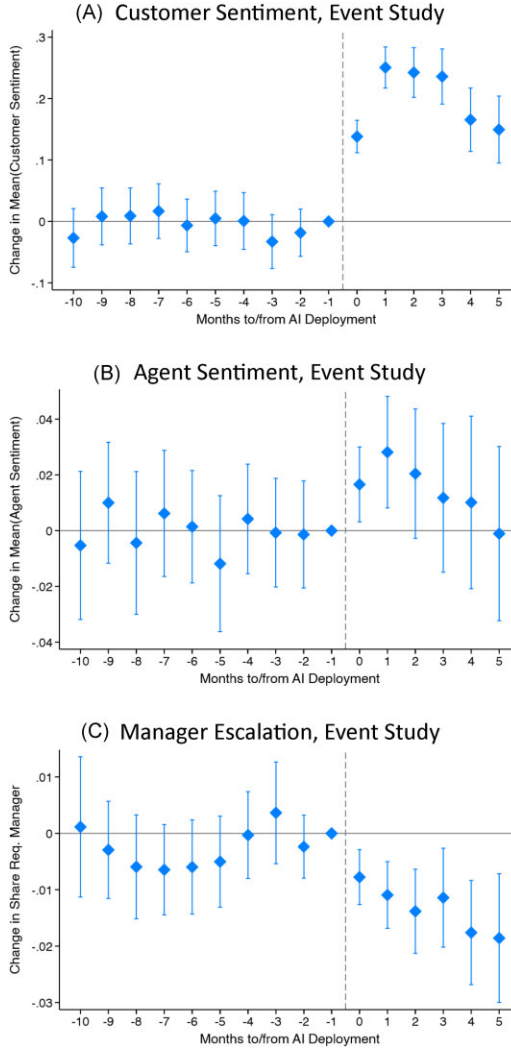

FIGURE X

Experience of Work

Each panel plots the effect of AI model deployment on the experience of work. Panel A plots the effect of AI model deployment on customer sentiment, Panel B plots the corresponding estimate for agent sentiment, and Panel C show the effects of AI assistance on customer requests for manager assistance. Sentiment is measured using SiEBERT, a fine-tuned checkpoint of RoBERTA, an English-language transformer model. Regressions follow equation (1) and include agent, chat year-month, and months of agent-tenure fixed effects. Observations are at the chat level, aggregated to the agent-month, and robust standard errors are clustered at the agent level.

Focusing on customer sentiment, Online Appendix Figure A.XVII, Panels C and D examine whether access to AI has different effects across agents. We find access to AI assistance significantly improves how customers treat agents of all skill and experience levels, with the largest effects for agents in the lower to lower-middle range of both the skill and tenure distributions. Consistent with our productivity results, the highest-performing and most experienced agents see the smallest benefits of AI access. These results suggest that AI recommendations, which were explicitly designed to prioritize more empathetic responses, may improve agents' demonstrated social skills and have a positive emotional effect on customers.

VII.B. Customer Confidence and Managerial Escalation

Changes in individual productivity may have broader implications for organizational workflows (Athey et al. 1994; Athey and Stern 1998; Garicano 2000). In most customer-service settings, frontline agents attempt to resolve customer problems but can seek the help of supervisors when they are unsure how to proceed. Customers, knowing this, will sometimes attempt to escalate a conversation by asking to speak to a manager. This type of request generally occurs when frustrated customers feel that the current agent is not adequately addressing their problem.

In Figure X, Panel C, we consider the effect of AI assistance on the frequency of chat escalation. The outcome variable we focus on is the share of an agent's chats in which a customer requests to speak to a manager or supervisor, aggregated to the year-month level. We focus on requests for escalation rather than actual escalations because we lack data on actual escalations and because requests are a better measure of customer confidence in an agent's competence or authority.

After the introduction of AI assistance, we see a gradual but substantial decline in requests for escalation. Relative to a baseline rate of approximately 6%, these coefficients suggest that AI assistance generates an almost 25% decline in customer requests to speak to a manager. In Online Appendix Figure A.XVII, Panels E and F, we consider how these patterns change depending on the skill and experience of the worker. Though these results are relatively noisy, our point estimates suggest that requests for

escalation are disproportionately reduced for agents who were less skilled or less experienced at the time of AI adoption.

VII.C. Attrition

The adoption of generative AI tools can affect workers in various ways, including their productivity, the level of stress they experience, how customers perceive them, and their overall job satisfaction. Although we cannot directly observe all these factors, we can analyze turnover patterns as a broad measure of how workers respond to AI implementation.

We compare attrition rates between AI-assisted agents and untreated agents with equal tenure. We drop observations for treated agents before treatment because they do not experience attrition by construction (they must survive to be treated in the future), and control for location and time fixed effects.

Consistent with our findings so far, Online Appendix Figure A.XVIII, Panel A shows that access to AI assistance is associated with the strongest reductions in attrition among newer agents, those with less than six months of experience. The magnitude of this coefficient, around 10 percentage points, translates into a 40% decrease relative to a baseline attrition rate in this group of 25%. In Panel B, we examine attrition by worker skill. We find a significant decrease in attrition for all skill groups, although without a clear gradient.

These results should be taken with more caution relative to our main results because attrition occurs once per worker, and therefore we are unable to include agent fixed effects. Our results may overstate the effect of AI access on attrition if, for example, the firm is more likely to give AI access to agents deemed more likely to stay.

VIII. Conclusion

Advances in AI technologies open up a broad set of economic possibilities. This article provides early empirical evidence on the effects of a generative AI tool in a real-world workplace. In our setting, we find that access to AI-generated recommendations increases overall worker productivity by 15%, with even larger effects for lower-skill and novice agents. These productivity gains in part reflect durable worker learning rather than rote reliance on AI suggestions. Furthermore, AI assistance appears to improve

worker on-the-job experiences, such as by improving customer sentiment and confidence, and is associated with reductions in turnover.

Our analysis is subject to some caveats and raises many unanswered questions. First, we note again that our findings apply for a particular AI tool, used in a single firm, in a single occupation, and should not be generalized across all occupations and AI systems. For example, our setting has a relatively stable product and set of technical support questions. In areas where the product or environment is changing rapidly, the relative value of AI recommendations may be different. For instance, AI may be better able to synthesize changing best practices, or could impede learning by promoting outdated practices observed in historical training data. Indeed, recent work by Perry et al. (2023) and Otis et al. (2023) have found cases in which AI adoption has limited or even negative effects.

Second, we report partial equilibrium short- to medium-run effects of AI deployment. Although we do not have access to compensation data, the managers we spoke to believed that workers may have received higher performance pay as a result of AI assistance, since these bonuses were typically tied to targets related to AHT and RRs. They caution that potential gains in bonus pay may not be long-lived because it is common practice to adjust performance targets if too many agents were reaching the goals. As a result, workers may eventually be subject to a ratchet effect if AI assistance leads performance targets to be readjusted upward.

More generally, we are not able to observe longer-run equilibrium responses. In principle, the increased productivity we observe could lead to either lower or higher demand for customer-service agents. If customer demand for assistance is inelastic, then the productivity gains we document will likely translate into less demand for human labor. A back-of-the-envelope calculation suggests the firm could field the same number of customer-support issues with 12% fewer worker-hours. Conversely, individuals may currently avoid contacting customer service because of the long wait times and low-quality service. AI assistance that improves this experience may boost consumer demand for product support, resulting in increased labor demand (Berg et al. 2018; Korinek 2022). In addition, the use of AI could create new jobs for customer-service agents, such as testing and training AI models (Autor et al. 2024). One manager

we spoke with reports that high-skill workers in some contact centers are already being tasked with reviewing AI suggestions and providing better alternatives. Other work shows that even low levels of AI adoption can affect market equilibrium prices and quantities, highlighting the need for more work on the equilibrium effects of AI on the labor market (Raymond 2023).

Finally, our findings also raise questions about the nature of worker productivity. Traditionally, a support agent's productivity refers to their ability to help the customers. Yet in a setting where customer-service conversations are fed into training data sets, a worker's productivity also includes the AI training data they produce. Top performers, in particular, contribute many of the examples used to train the AI system we study. This increases their value to the firm. At the same time, our results suggest that access to AI suggestions may lead them to put less effort into coming up with new solutions. Going forward, compensation policies that provide incentives for people to contribute to model training could be important. Given the early stage of generative AI, these and other questions deserve further scrutiny.

Stanford University and National Bureau of Economic Research, United States
Massachusetts Institute of Technology and National Bureau of Economic Research, United States
Massachusetts Institute of Technology, United States

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

Data Availability

Code replicating tables and figures from this article is available in the Harvard Dataverse, https://doi.org/10.7910/DVN/FSV1X7 (Brynjolfsson, Li, and Raymond 2024).

References

Acemoglu, Daron, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti, "Technology, Information, and the Decentralization of the Firm," *Quarterly Journal of Economics*, 122 (2007), 1759–1799. https://doi.org/10.1162/qjec.2007.122.4.1759

Acemoglu, Daron, and David Autor, "Skills, Tasks and Technologies: Implications for Employment and Earnings," In *Handbook of Labor Economics*, David Card and Orley Ashenfelter, eds. (Amsterdam: Elsevier, 2011), 1043–1171. https://doi.org/10.1016/S0169-7218(11)02410-5.

Acemoglu, Daron, and Pascual Restrepo, "Low-Skill and High-Skill Automation," *Journal of Human Capital*, 12 (2018), 204–232. https://doi.org/10.1086/697242

Acemoglu, Daron, Gary W. Anderson, David N. Beede, Cathy Buffington, Eric E. Childress, Emin Dinlersoz, Lucia S Foster, Nathan Goldschlag, John C. Haltiwanger, Zachary Kroff, Pascual Restrepo, and Nikolas Zolas, "Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey," NBER Working Paper no. 30659, 2022. https://doi.org/10.3386/w30659.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," NBER Working Paper no. 31422, 2023. https://doi.org/10.3386/w31422.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad, "The Skill Complementarity of Broadband Internet," *Quarterly Journal of Economics*, 130 (2015), 1781–1824. https://doi.org/10.1093/qje/qjv028

Angelova, Victoria, Will S. Dobbie, and Crystal Yang, "Algorithmic Recommendations and Human Discretion," NBER Working Paper no. 31747, 2023. https://doi.org/10.3386/w31747.

Athey, Susan, and Scott Stern, "The Impact of Information Technology on Emergency Health Care Outcomes," *RAND Journal of Economics*, 33 (2002), 399–432. https://doi.org/10.2307/3087465.

———. "An Empirical Framework for Testing Theories About Complimentarity in Organizational Design," NBER Working Paper no. 6600, 1998. https://doi.org/10.3386/w6600.

Athey, Susan, Joshua Gans, Scott Schaefer, and Scott Stern, "The Allocation of Decisions in Organizations," Working Paper no. 1322, Stanford Graduate School of Business, Stanford, CA, 1994. https://www.gsb.stanford.edu/faculty-research/working-papers/allocation-decisions-organizations.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger, "Computing Inequality: Have Computers Changed the Labor Market?" *Quarterly Journal of Economics*, 113 (1998), 1169–1213. https://doi.org/10.1162/003355398555874

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 118 (2003), 1279–1333. https://doi.org/10.1162/003355303322552801.

Autor, David, "Polanyi's Paradox and the Shape of Employment Growth," NBER Working Paper no. 20485, 2014. https://doi.org/10.3386/w20485.

Autor, David, Caroline Chin, Anna Salomons, and Bryan Seegmiller, "New Frontiers: The Origins and Content of New Work, 1940–2018," *Quarterly Journal of Economics*, 139 (2024), 1399–1465. https://doi.org/10.1093/qje/qjae008.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," *Journal of Financial Economics*, 151 (2024), 103745. https://doi.org/10.1016/j.jfineco.2023.103745

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, May 7–9, 2015, Conference Track Proceedings*, Bengio Yoshua and Yann LeCuneds. (2015). http://arxiv.org/abs/1409.0473.

Baker, George P., and Thomas N. Hubbard, "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information," *American Economic Review*, 93 (2003), 551–572. https://doi.org/10.1257/000282803322156981

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw, "How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills," *Quarterly Journal of Economics*, 122 (2007), 1721–1758. https://doi.org/10.1162/qjec.2007.122.4.1721

Berg, Jeff, Avinash Das, Vinay Gupta, and Paul Kline, "Smarter Call-Center Coaching for the Digital World," Technical Report, McKinsey & Company, New York, 2018.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen, "The Distinct Effects of Information Technology and Communication Technology on Firm Organization," *Management Science*, 60 (2014), 2859–2885. https://doi.org/10.1287/mnsc.2014.2013

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 91 (2024), 3253–3285. https://doi.org/10.1093/restud/rdae007

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *Quarterly Journal of Economics*, 117 (2002), 339–376. https://doi.org/10.1162/003355302753399526

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language Models Are Few-Shot Learners," In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, (Red Hook, NY: Curran Associates, 2020).

Brügger, Annina, Kai-Florian Richter, and Sara Irina Fabrikant, "How Does Navigation System Behavior Influence Human Behavior?" *Cognitive Research: Principles and Implications*, 4 (2019), 5.

Brynjolfsson, Erik, and Lorin M. Hitt, "Beyond Computation: Information Technology, Organizational Transformation and Business Performance," *Journal of Economic Perspectives*, 14 (2000), 23–48. https://doi.org/10.1257/jep.14.4.23

Brynjolfsson, Erik, and Tom Mitchell, "What Can Machine Learning Do? Workforce Implications," *Science*, 358 (2017), 1530–1534. https://doi.org/10.1126/science.aap8062

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Replication Data for: 'Generative AI at Work'," 2024, Harvard Dataverse, https://doi.org/10.7910/DVN/FSV1X7

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 13 (2021), 333–372. https://doi.org/10.1257/mac.20180386.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv working paper 2303.12712, 2023. https://arxiv.org/abs/2303.12712.

Buesing, Eric, Vinay Gupta, Sarah Higgins, and Raelyn Jacobson, "Customer Care: The Future Talent Factory," Technical Report, McKinsey & Company, New York, 2020. https://www.mckinsey.com/capabilities/operations/our-insights/customer-care-the-future-talent-factory.

Callaway, Brantly, and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225 (2021), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Calvino, Flavio, and Luca Fontanelli, "A Portrait of AI Adopters across Countries: Firm Characteristics, Assets' Complementarities and Productivity," Technical Report, Organisation for Economic Co-operation and Development, Paris, 2023.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," *Quarterly Journal of Economics*, 134 (2019), 1405–1454. https://doi.org/10.1093/qje/qjz014

Choi, Jonathan H., and Daniel Schwarcz, "AI Assistance in Legal Analysis: An Empirical Study," SSRN working paper, 2023. https://doi.org/10.2139/ssrn.4539836

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky, "Global Survey: The State of AI in 2021," Technical Report, McKinsey & Company, New York, 2021. https://www.mckinsey.com/businessfunctions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021.

Daron, Acemoglu, and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, 128 (2020), 2188–2244. https://doi.org/10.1086/705716

———, "The Simple Macroeconomics of AI," Working Paper no. 32487, National Bureau of Economic Research, Cambridge, MA, 2024.

de Chaisemartin, Clément, and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110 (2020), 2964–2996. https://doi.org/10.1257/aer.20181169

Dell'Acqua, Fabrizio, Edward McFowland, III, Ethan Mollick, Katherine Kellogg Lifshitz-Assaf, Saran Rajendran, Lisa Krayer, Francois Candelon, and Karim Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Working Paper no. 24-013, Harvard Business School, Cambridge, MA, 2023.

Dixon, Jay, Bryan Hong, and Lynn Wu, "The Robot Revolution: Managerial and Employment Consequences for Firms," SSRN working paper, 2020. https://dx.doi.org/10.2139/ssrn.3422581.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv working paper 2303.10130, 2023. http://arxiv.org/abs/2303.10130.

Felten, Edward W., Raj Manav, and Robert Seamans, "Occupational Heterogeneity in Exposure to Generative AI," SSRN working paper, 2023. https://dx.doi.org/10.2139/ssrn.4414065.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108 (2000), 874–904. https://doi.org/10.1086/317671

Garicano, Luis, and Esteban Rossi-Hansberg, "Knowledge-Based Hierarchies: Using Organizations to Understand the Economy," *Annual Review of Economics*, 7 (2015), 1–30. https://doi.org/10.1146/annurev-economics-080614-115748

Gemini Team, "Gemini: A Family of Highly Capable Multimodal Models," Technical Report, Google, Mountain View, CA, 2023. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Goldin, Claudia, and Lawrence F. Katz, "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics*, 113 (1998), 693–732. https://doi.org/10.1162/003355398555720

Goldin, Claudia, and Lawrence F. Katz, *The Race between Education and Technology*, (Cambridge, MA: Harvard University Press, 2008).

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 225 (2021), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Gretz, Whitney, and Raelyn Jacobson, "Boosting Contact-Center Performance through Employee Engagement," Technical Report, McKinsey & Company,

New York, 2018. https://www.mckinsey.com/capabilities/operations/our-insights/boosting-contact-center-performance-through-employee-engagement.
Halevy, Alon, Peter Norvig, and Fernando Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems*, 24 (2009), 8–12. https://doi.org/10.1109/MIS.2009.36
Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp, "More than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, 40 (2023), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005
Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li, "Discretion in Hiring," *Quarterly Journal of Economics*, 133 (2018), 765–800. https://doi.org/10.1093/qje/qjx042
Hugging Face, "sentence-transformers/all-MiniLM-L6-v2," 2023. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, Skill, and Productivity: The Case of Taxi Drivers," NBER Working Paper no. 30612, 2022. https://doi.org/10.3386/w30612.
Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," arXiv working paper 2001.08361, 2020. https://arxiv.org/abs/2001.08361.
Katz, Lawrence F., and Kevin M. Murphy, "Changes in Relative Wages, 1963–1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107 (1992), 35–78. https://doi.org/10.2307/2118323
Korinek, Anton, "How Innovation Affects Labor Markets: An Impact Assessment," Working Paper, Brookings Institution, Washington, DC, 2022. https://www.brookings.edu/wp-content/uploads/2022/06/How-innovation-affects-labor-markets-1.pdf.
Kumar, Harsh, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman, "Math Education with Large Language Models: Peril or Promise?," SSRN working paper, 2023. https://doi.org/10.2139/ssrn.4641653
Lee, Don, "The Philippines Has Become the Call-Center Capital of the World," *Los Angeles Times*, February 1, 2015. https://www.latimes.com/business/la-fi-philippines-economy-20150202-story.html.
Li, Chun, "OpenAI's GPT-3 Language Model: A Technical Overview," *Lambda*, June 3, 2020. https://lambdalabs.com/blog/demystifying-gpt-3.
Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge, "Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models," *Meta-Radiology*, 1 (2023), 100017. https://doi.org/10.1016/j.metrad.2023.100017
Mejova, Yelena, "Sentiment Analysis: An Overview," Survey Paper, University of Iowa, Computer Science Department, Iowa City, 2009.
Michaels, Guy, Ashwini Natraj, and John Van Reenen, "Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years," *Review of Economics and Statistics*, 96 (2014), 60–77. https://doi.org/10.1162/REST_a_00366.
Nguyen, Nhan, and Sarah Nadi, "An Empirical Evaluation of GitHub Copilot's Code Suggestions," in *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*, (2022).
Noy, Shakked, and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 381 (2023), 187–192. https://doi.org/10.1126/science.adh2586
OECD, *OECD Employment Outlook2023: Artificial Intelligence and the Labour Market*, (Paris: Organisation for Economic Co-operation and Development, 2023).

OpenAI, "GPT-4 Technical Report," Technical Report, OpenAI, 2023, https://cdn.openai.com/papers/gpt-4.pdf.

Otis, Nicholas G., Berkeley Haas, Rowan Clarke, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance," Working Paper no. 24-042, Harvard Business School, Cambridge, MA, 2023.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training Language Models to Follow Instructions with Human Feedback," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oheds. (Red Hook, NY: Curran Associates, 2022), 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Patel, Dylan, and Gerald Wong, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE," *Semianalysis*, July 10, 2023. https://www.semianalysis.com/p/gpt-4-architecture-infrastructure.

Peng, Baolin, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao, "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback," arXiv working paper 2302.12813, 2023a. https://arxiv.org/abs/2302.12813.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," arXiv working paper 2302.06590, 2023b. https://arxiv.org/abs/2302.06590.

Perry, Neil, Megha Srivastava, Deepak Kumar, and Dan Boneh, "Do Users Write More Insecure Code with AI Assistants?," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS '23*, (New York: Association for Computing Machinery, 2023), 2785–2799.

Polanyi, Michael, *The Tacit Dimension*, (Chicago: University of Chicago Press, 1966).

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach, "Manipulating and Measuring Model Interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, (New York: Association for Computing Machinery, 2021).

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training," preprint, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language Models Are Unsupervised Multitask Learners," preprint, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Raymond, Lindsey, "The Market Effects of Algorithms," Working Paper, 2023.

Roose, Kevin, "A Conversation with Bing's Chatbot Left Me Deeply Unsettled," *New York Times*, February 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Rosen, Sherwin, "The Economics of Superstars," *American Economic Review*, 71 (1981), 845–858. http://www.jstor.org/stable/1803469.

Sun, Liyang, and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225 (2021), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

Syverson, Chad, "What Determines Productivity?" *Journal of Economic Literature*, 49 (2011), 326–365. https://doi.org/10.1257/jel.49.2.326

Taniguchi, Hiroya, and Ken Yamada, "ICT Capital-Skill Complementarity and Wage Inequality: Evidence from OECD Countries," *Labour Economics*, 76 (2022), 102151. https://doi.org/10.1016/j.labeco.2022.102151

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone, "When Combinations of Humans and AI Are Useful: A Systematic Review and Meta Analysis," *Nature Human Behaviour*, 8 (2024), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, (Red Hook, NY: Curran Associates, 2017), 6000–6010.

White House, "The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America," Technical Report, White House, Washington, DC, 2022. https://www.whitehouse.gov/cea/written-materials/2022/12/05/the-impact-of-artificial-intelligence.

Zhao, Shuyin, "GitHub Copilot Now Has a Better AI Model and New Capabilities," *GitHub*, February 14, 2023. https://github.blog/ai-and-ml/github-copilot/github-copilot-now-has-a-better-ai-model-and-new-capabilities.

Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson, Kristina McElheran, David Beede, Catherine Buffington, Nathan Goldschlag, Lucia Foster, and Emin Dinlersoz, "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey," Working Paper no. 20-40, Center for Economic Studies, U.S. Census Bureau, Washington, DC, 2020. https://ideas.repec.org/p/cen/wpaper/20-40.html.