GENERATIVE AI AT WORK*

Erik Brynjolfsson

Danielle Li

Lindsey Raymond

November 18, 2024

First Draft: 23 April 2023

Abstract

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 percent 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 Classifications: D80, J24, M15, M51, O33

Keywords: Generative AI, Large Language Models, Technology Adoption, Worker Productivity, Worker Learning, Experience of Work, Organizational Design.

*Correspondence to erikb@stanford.edu, d_li@mit.edu, and lraymond@mit.edu. 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 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 emergence of generative artificial intelligence (AI) has attracted significant attention for its potential economic impact. Although various generative AI tools have performed well in laboratory settings, questions remain about their effectiveness in real-world settings, where they may encounter unfamiliar problems, face organizational resistance, or provide misleading information (Peng et al., 2023; Roose, 2023).

We provide early evidence on the impact 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 percent, 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. Yet, 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 impact 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 Generative Pre-trained Transformer 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 percent 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 percent increase in the number of issues resolved per hour. The AI tool helps also newer agents 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 impact 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).

Our third set of results investigates 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 adop-

tion drives convergence in communication patterns: low-skill agents begin communicating more like high-skill agents.

The fourth set of results focuses 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 access to generative AI suggestions can increase the productivity of individual workers and improve their experience of work. We emphasize, however, that these findings capture medium-run impacts within a single firm. Our paper 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 paper is related to a large literature on the impact of technological adoption on worker productivity and the organization of work (e.g. Rosen, 1981; Autor, Katz, and Krueger, 1998; Brynjolfsson and Hitt, 2000; Athey and Stern, 2002; Bartel, Ichniowski, and Shaw, 2007; Acemoglu et al., 2007; Hoffman, Kahn, and Li, 2017; Bloom et al., 2014; Michaels, Natraj, and Van Reenen, 2014; Garicano and Rossi-Hansberg, 2015; 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. For example, 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. Acemoglu et al. (2022); Zolas et al. (2020); Calvino and Fontanelli (2023) examine economy-wide data from the US and 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. For example, Acemoglu et al. (2022) find no detectable relationship between investments in AI-specific tools and firm outcomes, while Babina et al. (2022) 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 a non-generative AI tool that suggests customer-rich routes to taxi drivers reduces their search time by 5 percent, with the least experienced drivers benefiting the most. By contrast, several other studies find humans assisted by AI make worse decisions than do either humans or AI alone. (Hoffman, Kahn, and Li, 2017; Angelova, Dobbie, and Yang, 2023; Agarwal et al., 2023; Poursabzi-Sangdeh et al., 2021). 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.

In this paper, we provide micro-level evidence on the adoption of a generative AI tool across thousands of workers employed by a given firm and its subcontractors. Our work is closely related to several other studies examining the impacts of generative AI in lab-like settings. Peng et al. (2023) recruit software engineers for a specific coding task (writing an HTTP server in JavaScript) and show 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 AI assistance helps law students on a law school exam, while in management consulting, Dell'Acqua et al. (2023) find access to GPT-4 suggestions improves the quality of responses on some management consulting tasks, but can negatively impact performance on tasks outside its capabilities. Consistent with our findings, Noy and Zhang (2023), Choi and Schwarcz (2023), Peng et al. (2023) and Dell'Acqua et al. (2023) find generative AI assistance compresses the productivity distribution, with lower-skill workers driving the bulk of improvements. Our paper, however, 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.

1 Generative AI and Large Language Models

1.1 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 (The 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 "large language model" 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 "pre-train" using large amounts of unlabeled data, and refinements in training techniques (Radford and Narasimhan, 2018; Radford et al., 2019; Liu et al., 2023; Ouyang et al., 2022).

The quality of LLMs is strongly dependent on scale: the amount of computing power used for training, the number of model parameters, and the dataset size (Kaplan et al., 2020). The GPT-3 model included 175 billion parameters, was trained on 300 billion tokens, and incurred approximately $5 million dollars in training costs; the GPT-4 model, meanwhile, 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 within an input text, even when that text is broken up into smaller segments and processed in parallel (Vaswani et al., 2017; Bahdanau, Cho, and Bengio, 2015).

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 and Narasimhan, 2018). This approach enables LLMs to learn a foundational understanding of language patterns and relationships that can be adapted or further fine-tuned for a specific task.

Fine-tuning can refine general-purpose LLMs 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 or tweet, 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 US legal bar exam (Liu et al., 2023; Bubeck et al., 2023; OpenAI, 2023).

1.2 The Economic Impacts of Generative AI

Computers have historically excelled at executing pre-programmed 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 possess 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; Autor, Levy, and Murnane, 2003; Brynjolfsson and Hitt, 2000; Michaels, Natraj, and Van Reenen, 2014; Bresnahan, Brynjolfsson, and Hitt, 2002; Baker and Hubbard, 2003; 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 non-routine 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 (The 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., 2023; 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, yet 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 importantly—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 impacts.

2 Our Setting: LLMs for Customer Support

2.1 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 percent of agents in contact centers leave each year, costing firms $10,000 to $20,000 per agent in the United States (Buesing et al., 2020; Gretz and Jacobson, 2018). 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, pre-trained LLMs can be fine-tuned with examples of both successfully and unsuccessfully resolved conversations.

2.2 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 firms (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 employs 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 located 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 US-based small business owners.

Customer chats are assigned on the basis of agent availability, with no additional pre-screening. 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 individuals or larger businesses—there was no additional sorting for queries related to US-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-resolution rate and net promoter score.

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

2.3 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 dataset 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 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 hallucinations, and inappropriate responses while helping the LLM distinguish successful behaviors of the top performers, including those they tacitly apply.

Once deployed, the AI system generates two main types of output: 1) real-time suggestions for how agents should respond to customers and 2) links to the data firm's internal documentation for relevant technical issues. 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 of Appendix Figure A.I shows a technical recommendation from the AI system, namely 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.

3 Deployment, Data, and Empirical Strategy

3.1 AI Rollout

AI assistance was introduced after a randomized control trial (RCT) involving a small number of agents. Appendix Figure A.II illustrates the timing of the rollout, which primarily took place during

the fall of 2020 and winter of 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.

Additionally, 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 in our study.

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

3.2 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 percent, are located outside the United States, mainly in the

Philippines. For each agent, we observe their assigned manager, 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: the average time handling an individual chat (AHT); the number of chats an agent handles per hour, which accounts for multitasking (CPH); and resolution rate (RR), the share of chats that are successfully resolved. We measure customer satisfaction using the net promoter score 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 dataset includes average handle time and chats per hour 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. Appendix Section J.2 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 post-treatment subgroups. Several of our main results are readily visible in these raw data. In Panels A through D, we see that post-treatment agents do better along a range of outcomes, relative to both never-treated agents and pre-treatment 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, Panel A of Figure I and Table I show never-treated agents resolve an average of 1.7 chats per hour, whereas post-treatment agents resolve 2.5 chats per hour. 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 chats per hour (Panel C) and resolution rates (Panel D): while ever-treated agents appear to be stronger performers at the outset than agents who are never treated, post-treatment agents perform substantially better. Looking instead at average handle times (Panel B), we see a starker pattern: pre-treatment and never-treated agents have similar distributions of average handle times, 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 will more precisely attribute these raw differences to the impact of AI model deployment.

3.3 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}, is performance measures for agent i in year-month t, with resolutions per hour as our main measure of productivity. We measure these outcomes in levels, and report percentage changes off the baseline pre-treatment 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 quarter. In our preferred specifications, we also include agent fixed effects, α_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 Appendix Tables A.III and A.IV show 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, 2022). For example, workers may take time to adjust to using the AI system, in which case its impact 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 this estimator is consistent assuming parallel trends, no anticipatory behavior, and cohort-specific treatment effects that follow the same dynamic profile. Both our main differences-in-differences and event study estimates are similar using robust estimators introduced in de Chaisemartin and D'Haultfœuille (2020), Borusyak, Jaravel, and Spiess (2022), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), as well

as using traditional two-way fixed effects OLS. In fact, 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 our results are similar when clustering at different levels of granularity (Appendix Table A.III) and instrumenting agent adoption with the date the first worker within the agent's team received AI access (Appendix Table A.II).

4 Main Results

4.1 Overall Impacts

Table II examines the impact of the deployment of the AI model on our primary measure of productivity, resolutions per hour, using a standard two-way fixed effects model. In Column 1, we show, controlling for time and location fixed effects, access to AI recommendations increases resolutions per hour by 0.47 chats, up 23.9 percent from their pre-treatment 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 the deployment of AI increases RPH by 0.30 chats or 15.2 percent.

Figure II shows the accompanying IW 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 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 percent decline from the baseline mean of 43 minutes (shorter handle times are considered better). Next, 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 percent. Unlike average handle time, chats per hour account 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.

Column 3 of Table III indicates a small 1.3 percentage point increase in chat resolution rates. This effect is economically modest and insignificant given a high baseline resolution rate of 82

percent. We interpret this 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 percent.

Appendix Figure A.III presents the accompanying event studies for additional outcomes. We see immediate impacts on average handle time (Panel A) and chats per hour (Panel B), and relatively flat patterns for resolution rate (Panel C) and customer satisfaction (Panel D). We interpret these findings as saying that, on average, AI assistance increases productivity without negatively impacting resolution rates and surveyed customer satisfaction.

4.1.1 Randomized Control Trial 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 Appendix Table A.I, shows similar effects on productivity as in our main sample. Appendix Figure A.V reports the accompanying event studies for our various outcomes, which are also similar to our main results.

4.1.2 Robustness

During the roll-out process, managers decided which agents to onboard 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 impacts of AI. To address this, in 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 average handle time and chats per hour are essentially identical to those under our main specification. However, resolutions increase by 0.55 chats per hour, compared with 0.30 in our main finding. We attribute the larger impact on resolutions per hour to the fact that this IV approach estimates a significant and larger impact on resolution rates.

We also show similar results using alternative estimators and at different levels of clustering and weighting. Appendix Table A.IX finds similar results using alternative difference-in-difference estimators introduced in Callaway and Sant'Anna (2021), Borusyak, Jaravel, and Spiess (2022), de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2021). Similarly, Appendix Figure A.IV reports that our results are similar under alternative event study estimators: Callaway and Sant'Anna (2021), Borusyak, Jaravel, and Spiess (2022), de Chaisemartin and D'Haultfœuille

(2020), and traditional two-way fixed effects. Next, in Appendix Table A.III, we show 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 Appendix Table A.IV, we weight agent-month observations by the number of customer chats a worker conducts. Re-weighting generates similar results to our main, equally-weighted, specifications.

4.2 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, however, generative AI does not simply execute routine tasks. Instead, as outlined in Section 1.2, AI models identify patterns in data that replicate the behaviors of many types of workers, including those engaged in non-routine, creative, or knowledge-based tasks. These fundamental technical differences suggest that generative AI may impact different workers in different ways.

4.2.1 Pre-treatment 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 resolution rates, 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. Next, 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 resolution rates, and superior customer satisfaction scores.

Panel A of Figure III shows how our productivity effects vary across workers in each quintile of our skill index, measured in the month prior to AI access. To isolate the impact of worker skill, our regression specification, available in Appendix Section J.3, includes a set of fixed effects for months of worker tenure. We find the productivity impact of AI assistance is most pronounced for workers in the lowest skill quintile (leftmost side), who see an increase of 0.5 in resolutions per hour, or 36 percent. In contrast, AI assistance does not lead to any significant change in productivity for the most skilled workers (rightmost side).

In Panels B through E of Figure III 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 average handle time (Panel B); a small but positive effect for chats per hour (Panel C); and, interestingly, small but statistically significant *decreases* in resolution rates 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 them with 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 quality—resolution rate and customer satisfaction—suggests 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 prior to 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 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.

4.2.2 Pre-treatment Worker Experience

Next, 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, while others have more than a year of experience. To isolate the impact of worker tenure, this analysis controls for the worker skill quintile at AI adoption, with the regression specification located in Appendix Section J.3.

In Panel A of Figure IV, we see a clear monotonic pattern in which the least experienced agents see the greatest gains in resolutions per hour. Agents with less than 1 month of tenure improve

by 0.7 resolutions per hour, with larger effects for less-experienced workers. In contrast, we see no effect for agents with more than a year of tenure.

In Panels B through E of Figure IV, 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 average handle times and chats per hour, respectively, among the newest workers. In Panels D and E, we find positive impacts of AI assistance on chat quality, as measured by resolution rates and customer satisfaction, respectively. For the most experienced workers, we see modest positive effects for average handle time (Panel B), positive but statistically insignificant effects on chats per hour (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 within the same task, access to AI systems disproportionately improves the performance of both novice and less skilled workers.

4.2.3 Moving Down the Experience Curve

To further explore how AI assistance impacts 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 around at 1.8 resolutions per hour. Never-treated workers (long dashed blue line) slowly improve their productivity with experience, reaching approximately 2.5 resolutions per hour 8 to 10 months later. In contrast, workers who always have access to AI assistance (short dashed red line) increase their productivity to 2.5 resolutions per hour after only two months and continue to improve until they are resolving more than 3 chats per hour 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 more rapidly increase, following the same trajectory as the always-treated agents. These findings demonstrate AI assistance not only accelerates ramp-up for new workers, but also improves the rate at which even experienced workers improve in their roles.

In 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 through C). Panel D follows a similar but noisier pattern for customer satisfaction. Across many of the outcomes 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.

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

5.1 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.

Panel A of Figure VI 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 percent, with an interquartile range of 23 percent to 50 percent: Agents frequently ignore 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 percent to 46 percent of code recommendations (Zhao, 2023). Such behavior may be appropriate, given that AI models may make incorrect or

irrelevant suggestions. In 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

Panel B of Figure VI 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, we separately estimate the impact of AI assistance for each group, including year-month, agent, and agent tenure fixed effects.

We find a steady and monotonic increase in returns by agent adherence: Among agents in the lowest quintile, we still see a 10 percent gain in productivity, but for agents in the highest quintile, the estimated impact is over twice as high, close to 25 percent. Appendix Figure A.X shows the results for our other four outcome measures. The positive correlation between adherence and returns holds most strongly for average handle time (Panel A) and chats per hour (Panel B), and more noisily for resolution rate (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, however, 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 then 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.

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"). Here, we see that compliance either stays stable or grows over time. The most initially compliant agents continue to comply at the same rates (just above 50 percent). The least initially compliant agents increase their compliance over time: Those in the bottom tercile initially follow recommendations less than 20 percent of the time, but by month five their compliance rates have increased by over 50 percent.

Next, 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 percent for those with more than a year of tenure compared with 37 percent for those with less than three months of tenure. However, 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. Here, we see that compliance rates are similar across skill groups, and all groups increase their compliance over time. In Appendix Figure A.XII we show 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 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, however, 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 over-relying 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 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.

5.2 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 agents 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 in order 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.

Appendix Figure A.XIII shows an example of such an outage, which occurred on September 10, 2020. The y-axis plots the share of post-treatment 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 post-treatment chats are plotted as zeros for figure clarity). During non-outage periods, the share of chats without AI recommendations is typically 30 to 40 percent, reflecting that the AI system does not normally generate recommendations in response to all messages. On the morning of September 10th, however, we see a notable spike in the number of chats without recommendations, increasing to almost 100 percent. 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 in order 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 impact of AI assistance using only post-adoption periods in which the AI system is not impacted by a software outage. Consistent with our main results, we see an immediate decline in the duration of individual chats by approximately 10 percent to 15 percent.

In Panel B, we use the same pre-treatment observations, but now restrict to post-adoption periods that are impacted 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 percent to 25 percent 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, focusing 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. Yet, 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. However, 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 can be 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 Panels C and D of Figure VII, we divide our main study of outage events by the initial adherence of the worker to AI, as described in Section 5.1. 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 first-hand how customers respond. These findings are consistent with other evidence from education that higher adherence and engagement with LLM-generated responses positively impacted learning (Kumar et al., 2023). In addition to directly improving productivity, exposure to AI assistance could supplement existing on-the-job training programs.

5.3 Handling Routine and Non-Routine Topics

In addition to varying by the characteristics of the worker, the impact 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 onboarding 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 impact of AI-assistance varies between more and

less routine customer problems. We use Gemini, a large language model 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 Appendix Section J.2 (Gemini Team, 2024).

Appendix Figure A.XIV reports the distribution of conversation topics in our dataset. 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—comprise half of all conversations and the top 16 topics represent over 90 percent 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 percent of all chats, includes inquiries related to payroll, taxes, and account access and management. The next 25 percent of chats covers five additional categories, including those dealing with bank transfers or managing subscriptions. The following 15 percent of chats encompass nine additional topics, while the final 10 percent of chats consists of all the remaining topics. Our regression, in Appendix Section J.3, is conducted at the chat level, with a focus on chat duration.

Panel A of Figure VIII 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 impact 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 4 to 5 minutes faster, which corresponds to an approximately 10 percent decrease from the pre-treatment mean duration for these topics. We see the largest decline, 5 to 6 minutes, for issues that are in the 75th to 90th percentiles of topic rarity, corresponding to a 14 percent reduction from the pre-treatment means for those topics. Finally, we see a smaller 4 minute or 11 percent 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 impacts in real-world settings. AI models generally perform better when trained on large datasets, which provide diverse examples and richer contextual information. Such datasets 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.

However, 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 specifi-

cally 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 impact. 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 impact 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 first-hand 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, Panel B of Figure VIII plots the impact 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 reduces conversation times 15 percent for the least common problems compared with 10 percent 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 impact of AI assistance is greatest for problems that a specific agent is least exposed to. While 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.

5.4 Conversational Style

5.4.1 English Fluency

The ability to communicate in clear, idiomatic English is crucial to customer satisfaction and the job performance of contact workers serving US customers. In our data, 80 percent 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. In this section, 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 to 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 5 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, an LLM, to score agents' text in each conversation along these two dimensions. For more information on our specific approach, prompts, and validation tests, see Appendix J.2.5.

The general level of both comprehensibility and native fluency is high. Prior to having AI access, the interquartile range of comprehensibility scores was 4.28 to 4.67; for native fluency it was 4.26 to 4.65. Despite this high baseline level, we find clear evidence that access to AI assistance increases proficiency scores. Appendix Figure A.XV presents the raw pre- and post-AI distribution of comprehensibility and native fluency scores for never-treated workers, pre-treatment workers, and post-treatment workers. The never treated and pre-treatment workers have identical distributions, but we see markedly higher scores for post-treatment workers. In Panels A and B of Figure IX, we report the accompanying event studies which show AI access leads to large improvements in both comprehensibility and native fluency. Finally, in Panels C and D of Figure IX, we report separate coefficients for US- and Philippines-based workers. We see a positive impact for all workers, but a larger improvement for workers based in the Philippines.

5.4.2 Textual Convergence

The analysis above 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 impact 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).

Panel A of Appendix Figure A.XVI 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 0 to 1, with 0 meaning two pieces of text are orthogonal (when represented as semantic vectors), and 1 indicating exact semantic similarity.

Prior to the deployment of 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. However, following 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. Panel B of Appendix Figure A.XVI 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. Panel C of Appendix Figure A.XVI 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. Post-AI adoption, however, text similarity between high- and low-skill workers begins increasing, from 0.55 to 0.61. While 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 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, however, 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.

6 Effects on the Experience of Work

6.1 Customer Sentiment

Qualitative studies suggest that working conditions for contact center agents can be unpleasant. Customers often vent their frustrations to anonymous service agents and, 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 impact how customers treat agents, but, in theory, the direction and magnitude of these impacts 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.

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 datasets, including product reviews and tweets (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 then aggregate these chat-level variables into a measure of average agent sentiment and average customer sentiment for each agent-year-month.

Panels A and B of Figure X 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 Column 1 of Table IV, 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. Column 2 of Table IV indicates agent sentiments increase by only 0.02 points or about 1 percent of a standard deviation.

Focusing on customer sentiment, Panels C and D of Appendix Figure A.XVII examine whether access to AI has different impacts 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 impact on customers.

6.2 Customer Confidence and Managerial Escalation

Changes in individual productivity may have broader implications for organizational workflows (Garicano, 2000; Athey et al., 1994; Athey and Stern, 1998). In most customer service settings, front-line 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 Panel C of Figure X, we consider the impact 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 both because we lack data on actual escalations and because requests are a better measure of customer confidence in an agent's competence or authority.

Following the introduction of AI assistance, we see a gradual but substantial decline in requests for escalation. Relative to a baseline rate of approximately 6 percent, these coefficients suggest that AI assistance generates an almost 25 percent decline in customer requests to speak to a manager. In Panels E and F of Appendix Figure A.XVII, we consider how these patterns change depending on the skill and experience of the worker. While 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.

6.3 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. While 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, Panel A of Appendix Figure A.XVIII shows that access to AI assistance is associated with the strongest reductions in attrition among newer agents, those with less than 6 months of experience. The magnitude of this coefficient, around 10 percentage

points, translates into a 40 percent decrease relative to a baseline attrition rate in this group of 25 percent. In Panel B, we examine attrition by worker skill. Here, 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 impact of AI access on attrition if, for example, the firm is more likely to give AI access to agents deemed more likely to stay.

7 Conclusion

Advances in AI technologies open up a broad set of economic possibilities. Our paper provides early empirical evidence on the effects of a generative AI tool in a real-world workplace. In our setting, we find access to AI-generated recommendations increases overall worker productivity by 15 percent, with even larger impacts 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, within 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 Otis et al. (2023) and Perry et al. (2022) have found instances in which AI adoption has limited or even negative effects.

Second, we report partial equilibrium short- to medium-run impacts of AI deployment. While 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 average handle time and resolution rates. They caution, however, 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 hitting 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 percent 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., 2022). 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 impact 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 datasets, 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, United States
National Bureau of Economic Research, United States
Massachusetts Institute of Technology, United States

References

Acemoglu, Daron. 2024. "The Simple Macroeconomics of AI."

——— , **Philippe Aghion, Claire Lelarge et al.** 2007. "Technology, Information, and the Decentralization of the Firm*." *The Quarterly Journal of Economics* 122 (4): 1759–1799. 10.1162/qjec.2007.122.4.1759, _eprint: https://academic.oup.com/qje/article-pdf/122/4/1759/5234557/122-4-1759.pdf.

——— , **Gary Anderson, David Beede et al.** 2022. "Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey."

——— , **and David Autor.** 2011. "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*, Volume 4. 1043–1171, Elsevier.

——— , **and Pascual Restrepo.** 2018. "Low-Skill and High-Skill Automation." *Journal of Human Capital* 12 (2): 204–232. 10.1086/697242.

——— 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128 (6): 2188–2244. 10.1086/705716, _eprint: https://doi.org/10.1086/705716.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar et al. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." Working Paper 31422, National Bureau of Economic Research. 10.3386/w31422.

Akerman, Anders, Ingvil Gaarder, and Magne Mogstad. 2015. "The Skill Complementarity of Broadband Internet *." *The Quarterly Journal of Economics* 130 (4): 1781–1824. 10.1093/qje/qjv028, _eprint: https://academic.oup.com/qje/article-pdf/130/4/1781/30637431/qjv028.pdf.

Angelova, Victoria, Will S Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion." Working Paper 31747, National Bureau of Economic Research. 10.3386/w31747.

Athey, Susan, Joshua Gans, Scott Schaefer et al. 1994. "The Allocation of Decisions in Organizations." *Stanford Graduate School of Business*, https://www.gsb.stanford.edu/faculty-research/working-papers/allocation-decisions-organizations.

——— , **and Scott Stern.** 1998. "An Empirical Framework for Testing Theories About Complimentarity in Organizational Design." Working Paper 6600, National Bureau of Economic Research. 10.3386/w6600.

——— 2002. "The Impact of Information Technology on Emergency Health Care Outcomes." *RAND Journal of Economics* 33 (3): 399–432, https://ideas.repec.org/a/rje/randje/v33y2002iautumnp399-432.html.

Autor, David. 2014. "Polanyi's Paradox and the Shape of Employment Growth." Working Paper w20485, National Bureau of Economic Research. 10.3386/w20485.

——— , **Caroline Chin, Anna M Salomons et al.** 2022. "New Frontiers: The Origins and Content of New Work, 1940–2018." Working Paper 30389, National Bureau of Economic Research. 10.3386/w30389.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?*." *The Quarterly Journal of Economics* 113 (4): 1169–1213. 10.1162/003355398555874, _eprint: https://academic.oup.com/qje/article-pdf/113/4/1169/5406877/113-4-1169.pdf.

——— , **Frank Levy, and Richard J. Murnane.** 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4): 1279–1333, http://www.jstor.org/stable/25053940.

Babina, Tania, Anastassia Fedyk, Alex Xi He et al. 2022. "Artificial Intelligence, Firm Growth, and Product Innovation." May. 10.2139/ssrn.3651052.

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

Baker, George P., and Thomas N. Hubbard. 2003. "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information." *American Economic Review* 93 (3): 551–572. 10.1257/000282803322156981.

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw. 2007. "How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills*." *The Quarterly Journal of Economics* 122 (4): 1721–1758. 10.1162/qjec.2007.122.4.1721.

Berg, Jeff, Avinash Das, Vinay Gupta et al. 2018. "Smarter call-center coaching for the digital world."Technical report, McKinsey & Company.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun et al. 2014. "The Distinct Effects of Information Technology and Communication Technology on Firm Organization." *Management Science* 60 (12): 2859–2885. 10.1287/mnsc.2014.2013.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. "Revisiting Event Study Designs: Robust and Efficient Estimation."

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt. 2002. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." *The Quarterly Journal of Economics* 117 (1): 339–376. 10.1162/003355302753399526.

Brown, Tom B., Benjamin Mann, Nick Ryder et al. 2020. "Language Models are Few-Shot Learners." July. 10.48550/arXiv.2005.14165, arXiv:2005.14165 [cs].

Brynjolfsson, Erik, and Lorin Hitt. 2000. "Beyond Computation: Information Technology, Organizational Transformation and Business Performance." *Journal of Economic Perspectives* 14 (4): 23–48.

———— **, and Tom Mitchell.** 2017. "What Can Machine Learning, Do? Workforce Implications." *Science* 358 1530–1534. 10.1126/science.aap8062.

———— **, Daniel Rock, and Chad Syverson.** 2021. "The productivity J-curve: How intangibles complement general purpose technologies." *American Economic Journal: Macroeconomics* 13 (1): 333–72.

Brügger, Annina, Kai-Florian Richter, and Sara Irina Fabrikant. 2019. "How does navigation system behavior influence human behavior?" *Cognitive Research: Principles and Implications* 4 (1): 5. 10.1186/s41235-019-0156-5.

Bubeck, Sebastien, Varun Chandrasekaran, Ronen Eldan et al. 2023. "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712*.

Buesing, Eric, Vinay Gupta, Sarah Higgins et al. 2020. "Customer care: The future talent factory."Technical report, McKinsey & Company, https://www.mckinsey.com/business-functions/operations/our-insights/customer-care-the-future-talent-factory.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. 10.1016/j.jeconom.2020.12.001.

Calvino, Flavio, and Luca Fontanelli. 2023. "A Portrait of AI Adopters across Countries: Firm Characteristics, Assets' Complementarities and Productivity."Technical report, OECD, Paris. 10.1787/0fb79bb9-en.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner et al. 2019. "The Effect of Minimum Wages on Low-Wage Jobs*." *The Quarterly Journal of Economics* 134 (3): 1405–1454. 10.1093/qje/qjz014.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96. 10.1257/aer.20181169.

Choi, Jonathan H., and Daniel Schwarcz. 2023. "AI Assistance in Legal Analysis: An Empirical Study." August. 10.2139/ssrn.4539836.

Chui, Michael, Bryce Hall, Alex Singla et al. 2021. "Global survey: The state of AI in 2021."Technical report, McKinsey & Company, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick et al. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." September.

Dixon, Jay, Bryan Hong, and Lynn Wu. 2020. "The Robot Revolution: Managerial and Employment Consequences for Firms." *DecisionSciRN: Other Performance Management (Sub-Topic)*, https://api.semanticscholar.org/CorpusID:226197840.

Eloundou, Tyna, Sam Manning, Pamela Mishkin et al. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." March, http://arxiv.org/abs/2303.10130, arXiv:2303.10130 [cs, econ, q-fin].

Felten, Edward W., Manav Raj, and Robert Seamans. 2023. "Occupational Heterogeneity in Exposure to Generative AI." April. 10.2139/ssrn.4414065.

Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy* 108 (5): 874–904. 10.1086/317671, Publisher: The University of Chicago Press.

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

Gemini Team. 2024. "Gemini: A Family of Highly Capable Multimodal Models." working papers, Google, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Goldin, Claudia, and Lawrence F. Katz. 1998. "The Origins of Technology-Skill Complementarity*." *The Quarterly Journal of Economics* 113 (3): 693–732. 10.1162/003355398555720.

———— 2008. *The Race between Education and Technology*. Harvard University Press, , http://www.jstor.org/stable/j.ctvjf9x5x.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225 (2): 254–277. 10.1016/j.jeconom.2021.03.014.

Gretz, Whitney, and Raelyn Jacobson. 2018. "Boosting contact-center performance through employee engagement."Technical report, McKinsey & Company.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2): 8–12. 10.1109/MIS.2009.36.

Hartmann, Jochen, Mark Heitmann, Christian Siebert et al. 2023. "More than a Feeling: Accuracy and Application of Sentiment Analysis." *International Journal of Research in Marketing* 40 (1): 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li. 2017. "Discretion in Hiring*." *The Quarterly Journal of Economics* 133 (2): 765–800. 10.1093/qje/qjx042.

Hugging Face. 2023. "sentence-transformers/all-MiniLM-L6-v2." April, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka et al. 2022. "AI, Skill, and Productivity: The Case of Taxi Drivers." Working Paper 30612, National Bureau of Economic Research. 10.3386/w30612.

Kaplan, Jared, Sam McCandlish, Tom Henighan et al. 2020. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361*.

Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *The Quarterly Journal of Economics* 107 (1): 35–78. 10.2307/2118323.

Korinek, Anton. 2022. "How innovation affects labor markets: An impact assessment." working paper, Brookings Institution, https://www.brookings.edu/wp-content/uploads/2022/06/How-innovation-affects-labor-markets-1.pdf.

Kumar, Harsh, David Rothschild, Daniel Goldstein et al. 2023. "Math Education with Large Language Models: Peril or Promise?" *SSRN Electronic Journal*. 10.2139/ssrn.4641653.

Lee, Don. 2015. "The Philippines has become the call-center capital of the world." *Los Angeles Times*, https://www.latimes.com/business/la-fi-philippines-economy-20150202-story.html, Section: Business.

Li, Chun. 2020. "OpenAI's GPT-3 Language Model: A Technical Overview." June, https://lambdalabs.com/blog/demystifying-gpt-3.

Liu, Yiheng, Tianle Han, Siyuan Ma et al. 2023. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." April. 10.48550/arXiv.2304.01852, arXiv:2304.01852 [cs].

Mejova, Yelena. 2009. "Sentiment Analysis: An Overview." *University of Iowa, Computer Science Department*.

Michaels, Guy, Ashwini Natraj, and John Van Reenen. 2014. "Has ICT Polarized Skill Demand? Evidence from Eleven Countries Over Twenty-Five Years." *The Review of Economics and Statistics* 96 (1): 60–77, https://www.jstor.org/stable/43554913.

Nguyen, Nhan, and Sarah Nadi. 2022. "An Empirical Evaluation of GitHub Copilot's Code Suggestions." In *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*, 1–5, May. 10.1145/3524842.3528470, ISSN: 2574-3864.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Available at SSRN 4375283*. 10.2139/ssrn.4375283.

OECD. 2023. *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. Paris: Organisation for Economic Co-operation and Development.

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

Otis, Nicholas G, Berkeley Haas, Rowan Clarke et al. 2023. "The Uneven Impact of Generative AI on Entrepreneurial Performance." December.

Ouyang, Long, Jeff Wu, Xu Jiang et al. 2022. "Training language models to follow instructions with human feedback." March. 10.48550/arXiv.2203.02155, arXiv:2203.02155 [cs].

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

Peng, Baolin, Michel Galley, Pengcheng He et al. 2023. "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback."

Peng, Sida, Eirini Kalliamvakou, Peter Cihon et al. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot."

Perry, Neil, Megha Srivastava, Deepak Kumar et al. 2022. "Do Users Write More Insecure Code with AI Assistants?" *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, https://api.semanticscholar.org/CorpusID:253384456.

Polanyi, Michael. 1966. *The Tacit Dimension*. Chicago, IL: University of Chicago Press, , https://press.uchicago.edu/ucp/books/book/chicago/T/bo6035368.html.

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman et al. 2021. "Manipulating and Measuring Model Interpretability."

Radford, Alec, and Karthik Narasimhan. 2018. "Improving Language Understanding by Generative Pre-Training." https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

——— , **Jeff Wu, Rewon Child et al.** 2019. "Language Models are Unsupervised Multitask Learners." https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Raymond, Lindsey. 2023. "The Market Effects of Algorithms." working paper, https://www.dropbox.com/scl/fi/22p85oogcf67mour5q8y2/LRaymond_JMP.pdf?rlkey=3v7nt884tx78y4rxbwwgwhuu5&dl=0.

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

Rosen, Sherwin. 1981. "The Economics of Superstars." *The American Economic Review* 71 (5): 845–858, http://www.jstor.org/stable/1803469.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49 (2): 326–65. 10.1257/jel.49.2.326.

Taniguchi, Hiroya, and Ken Yamada. 2022. "ICT Capital-Skill Complementarity and Wage Inequality: Evidence from OECD Countries." *Labour Economics* 76 102151. 10.1016/j.labeco.2022.102151, arXiv:1904.09857 [econ, q-fin].

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

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. 2024. "When Are Combinations of Humans and AI Useful?." https://arxiv.org/abs/2405.06087.

Vaswani, Ashish, Noam Shazeer, Niki Parmar et al. 2017. "Attention Is All You Need." December. 10.48550/arXiv.1706.03762, arXiv:1706.03762 [cs].

Zhao, Shuyin. 2023. "GitHub Copilot now has a better AI model and new capabilities." February, https://github.blog/2023-02-14-github-copilot-now-has-a-better-ai-model-and-new-capabilities/.

Zolas, Nikolas, Zachary Kroff, Erik Brynjolfsson et al. 2020. "Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey." Working Papers 20-40, Center for Economic Studies, U.S. Census Bureau, https://ideas.repec.org/p/cen/wpaper/20-40.html.

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 US Agents	.11	.15	.081	.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	.82	.78	.82	.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. Column 1 consists of all agents in our sample, Column 2 includes control agents who were never receive AI access. Column 3 presents statistics for treated agents before they receive AI access and Column 4 includes treated agents after AI model deployment.

TABLE II
Main Effects: Productivity (Resolutions per Hour)

VARIABLES	(1) Resolutions/Hr	(2) Resolutions/Hr	(3) Resolutions/Hr
Post AI X 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

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table presents the results of difference-in-difference regressions estimating the impact 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 X 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 3.1 describes the AI rollout procedure.

TABLE III
Main Effects: Additional Outcomes

VARIABLES	(1) AHT	(2) Chats/Hr	(3) Res. Rate	(4) NPS
Post AI X 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

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on additional measures of productivity and agent performance. Post AI X Treated measures the impact of AI model deployment after deployment on treated agents for average handle time (AHT) in Column 1, chats per hour (Chats/Hr), the number of chats an agent handles per hour in Column 2, resolution rate (Res. 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 3.1 describes the AI rollout procedure.

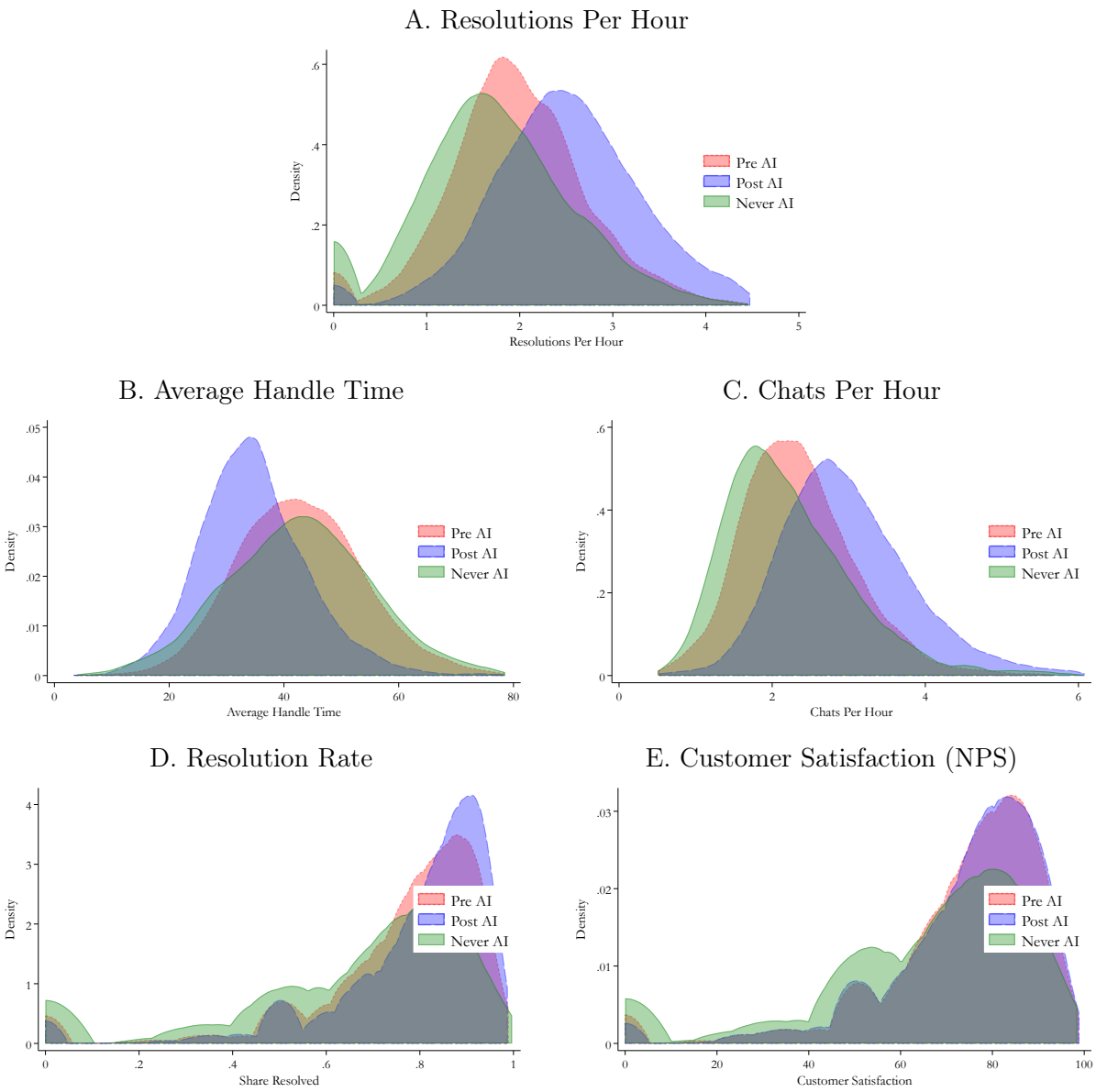
TABLE IV
Experience of Work

VARIABLES	(1) Mean(Customer Sentiment)	(2) Mean(Agent Sentiment)	(3) Share Req. Manager
Post AI X 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

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

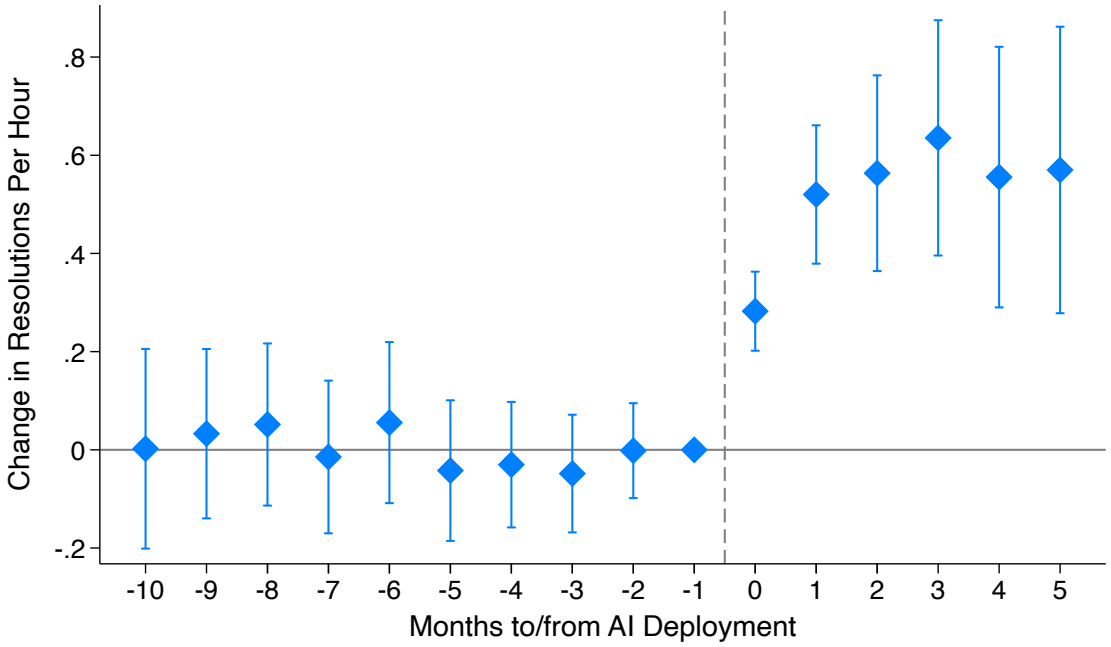
NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on measures of conversation sentiment and requests to speak to a manager ("Share Req. Manager"). Our regression specification, Equation 1, includes fixed effects for each agent, 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 aggregated to the agent-month level. Appendix Section J.2 elaborates on sentiment construction and Section 3.1 describes the AI rollout procedure.

FIGURE I
Raw Productivity Distributions, by AI Treatment



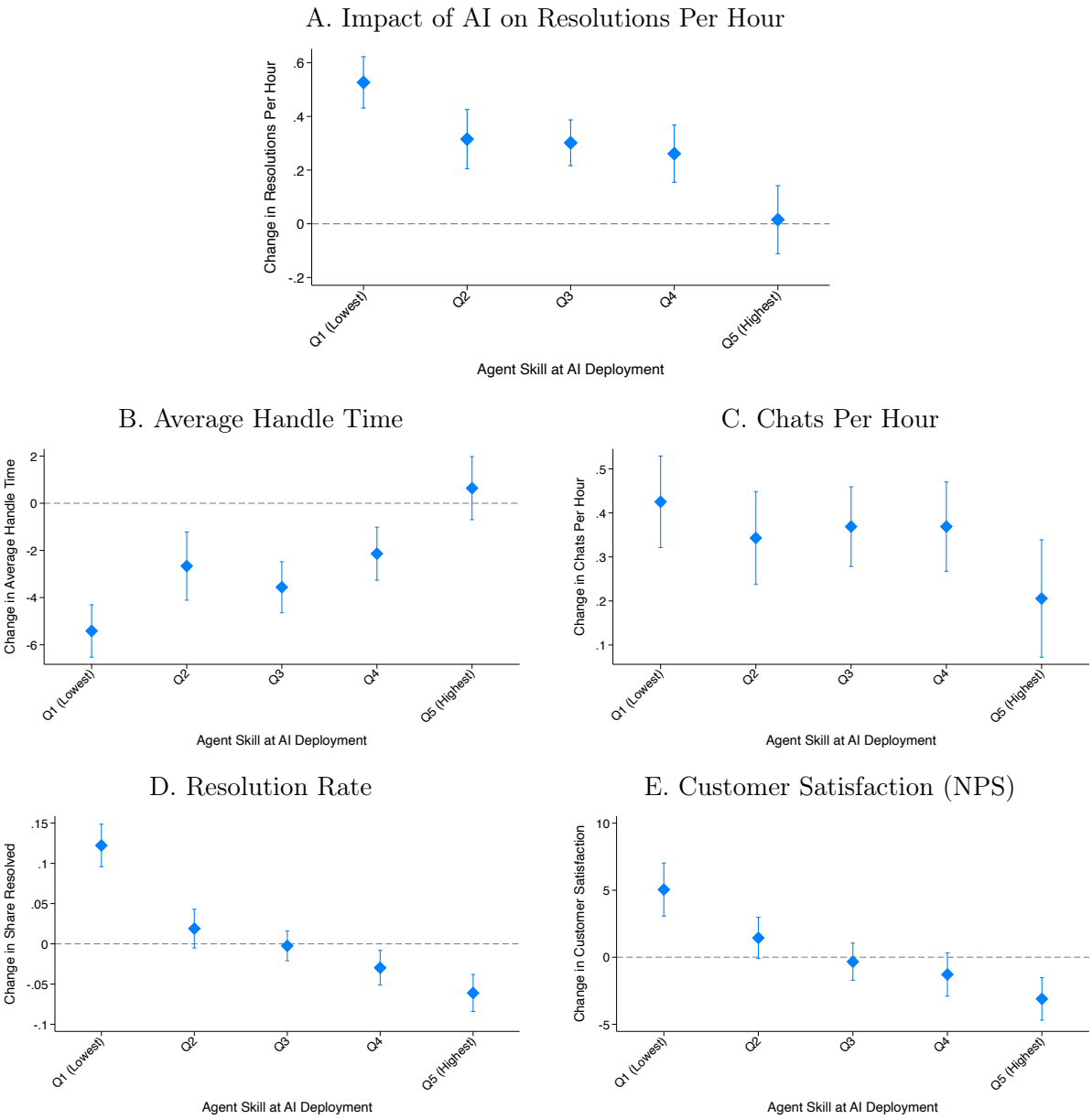
NOTES: This figure shows the distribution various outcome measures. We split this sample into agent-month observations for agents who eventually receive access to the AI system before deployment ("Pre AI", short dashed line), after deployment ("Post AI", long dashed line), and for agent-months associated 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 are 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 comes from the firm's software systems.

FIGURE II
Event Studies, Productivity



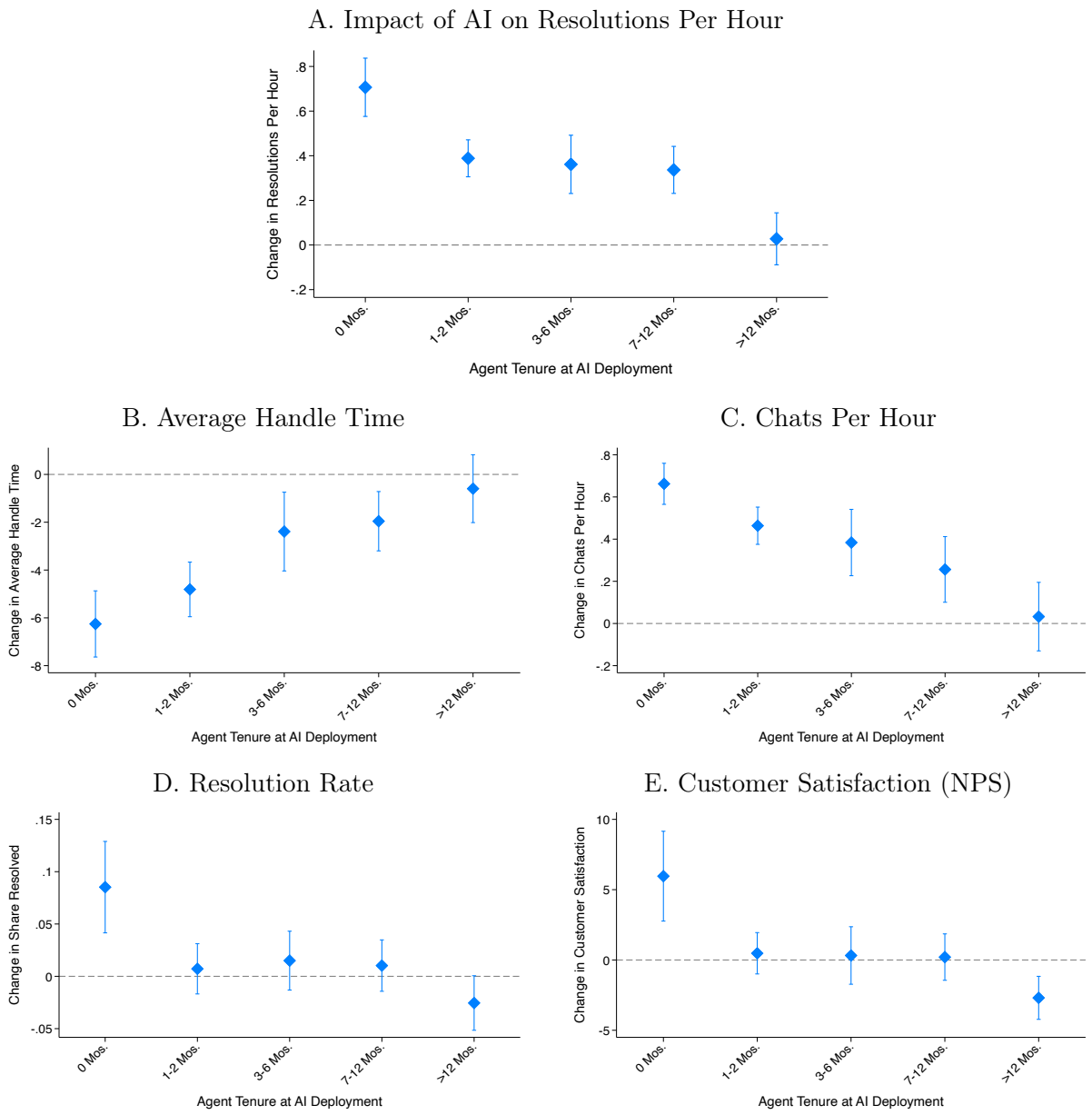
NOTES: This figure plots the coefficients and 95% confidence intervals from event study regressions of AI model deployment on our measure of productivity, resolutions per hour, using the Sun and Abraham (2021) interaction weighted estimator. Our specification follows Equation 1 and includes fixed effects 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 3.1 describes the rollout and Appendix Section J.3 outlines the regression specification.

FIGURE III: Heterogeneity of AI Impact, by Skill at Deployment



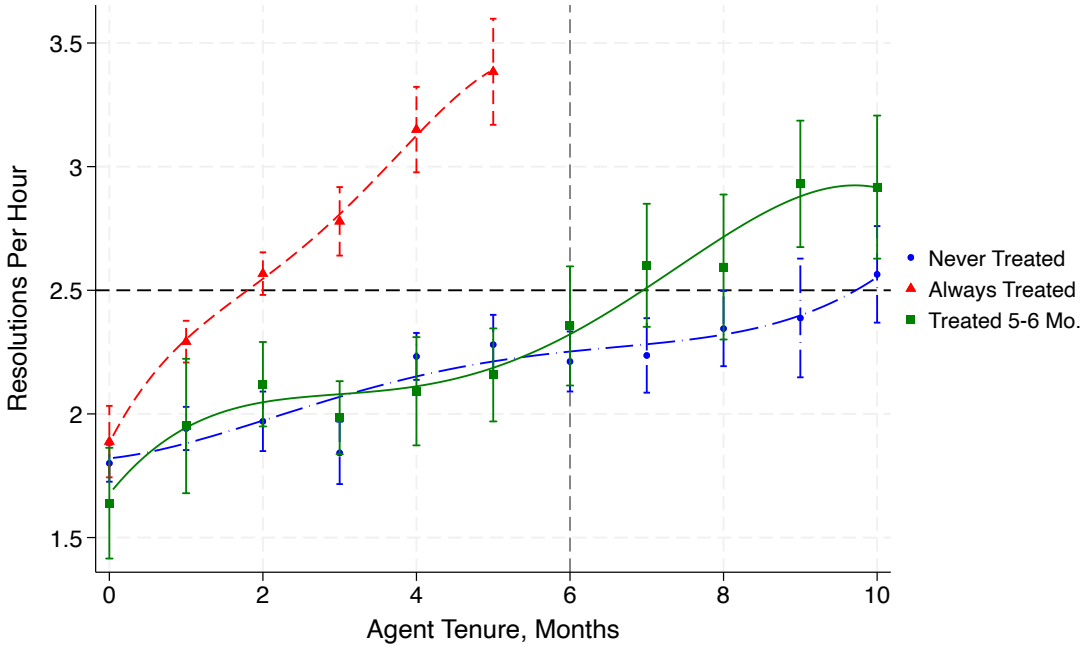
NOTES: These figures plot the impacts of AI model deployment on five measures of productivity and performance, by pre-deployment worker skill controlling for agent tenure. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses for agent performance. Within each month and company, agents are grouped into quintiles, with the most productive agents within each firm in quintile 5 and the least productive in quintile 1. Panel A plots the impacts 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 Appendix section J.3.

FIGURE IV: Heterogeneity of AI Impact, by Tenure at Deployment



NOTES: These figures plot the impacts 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 impacts 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 Appendix section J.3.

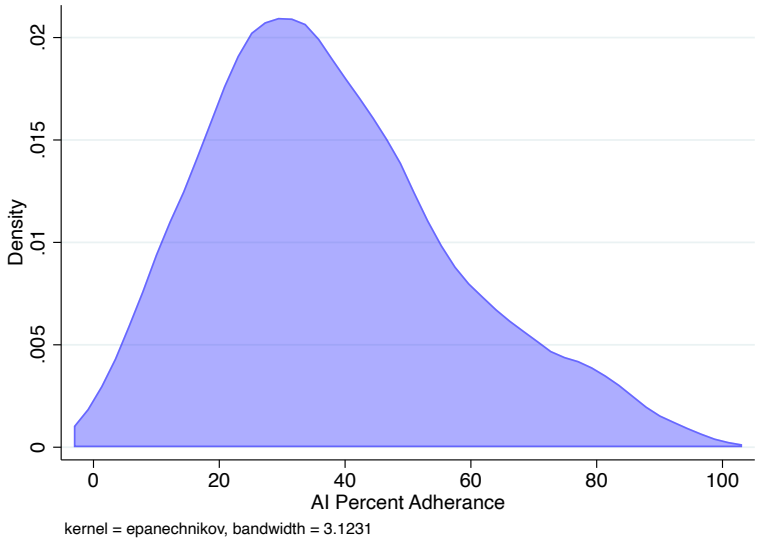
FIGURE V
Experience Curves by Deployment Cohort



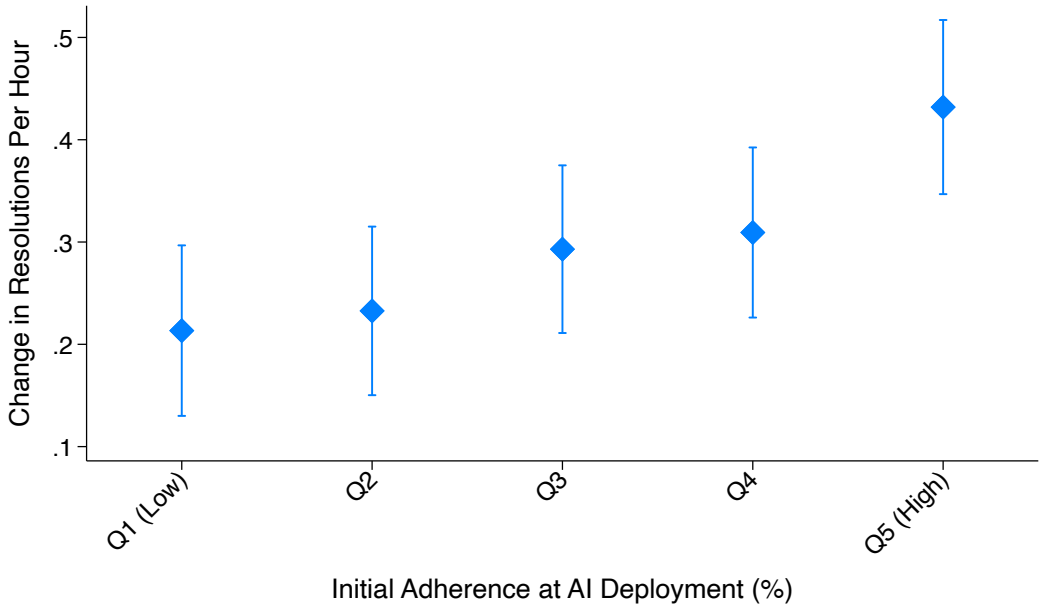
NOTES: This figure plots the relationship between productivity and job tenure. The short dashed line plots the performance of always-treated agents, those who have access to AI assistance from their first month on the job. The long dashed line plots agents who are never treated. The solid line plots 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. 95% confidence intervals are shown. Observations are at the agent-month level.

FIGURE VI
Heterogeneity of AI Impact, by AI Adherence

A. Distribution of AI Adherence

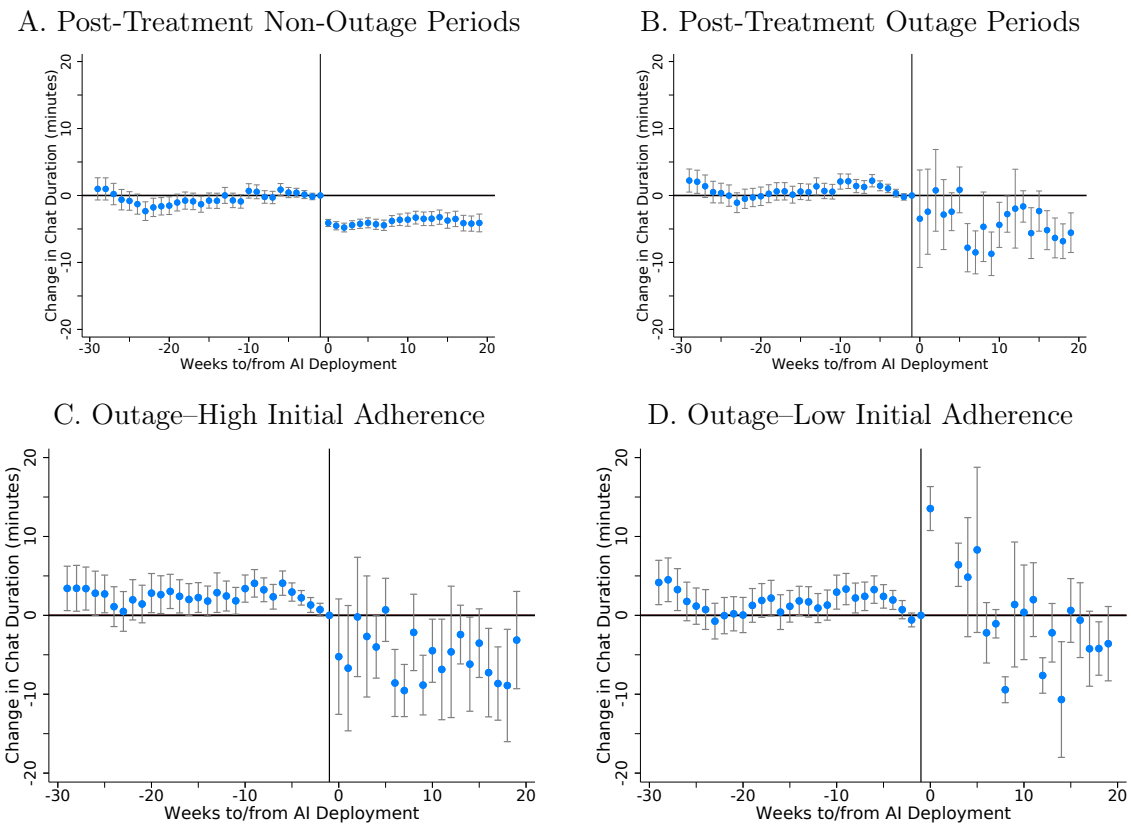


B. Impact of AI on Resolutions Per Hour, by Initial Adherence



NOTES: 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 impact of AI assistance on resolutions by 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 Appendix Section J.3, 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.

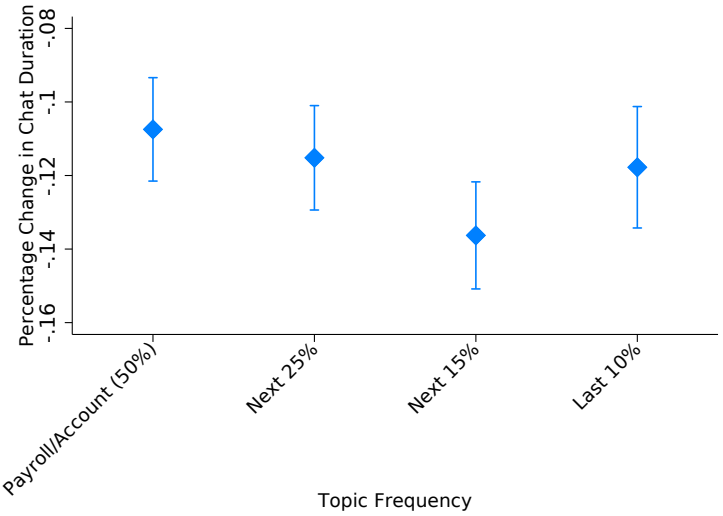
FIGURE VII
Chat Duration during AI System Outages



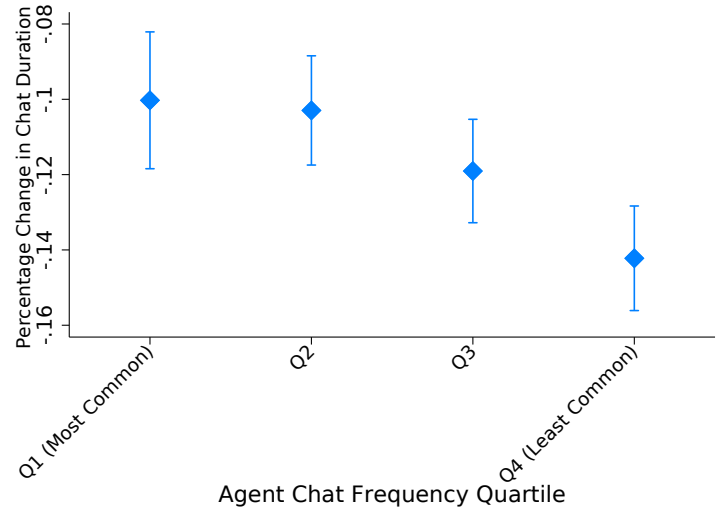
NOTES: These figures plot event studies for the impact of AI system rollout of chat duration at the individual chat level. Panel A restricts to post-treatment chats that do not occur during any period where there is a AI system outage. Panel B restricts to post-treatment chats that only occur during 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.

FIGURE VIII
AI Impact by Chat Topic

A. Impact of AI on Chat Duration, by Overall Topic Frequency

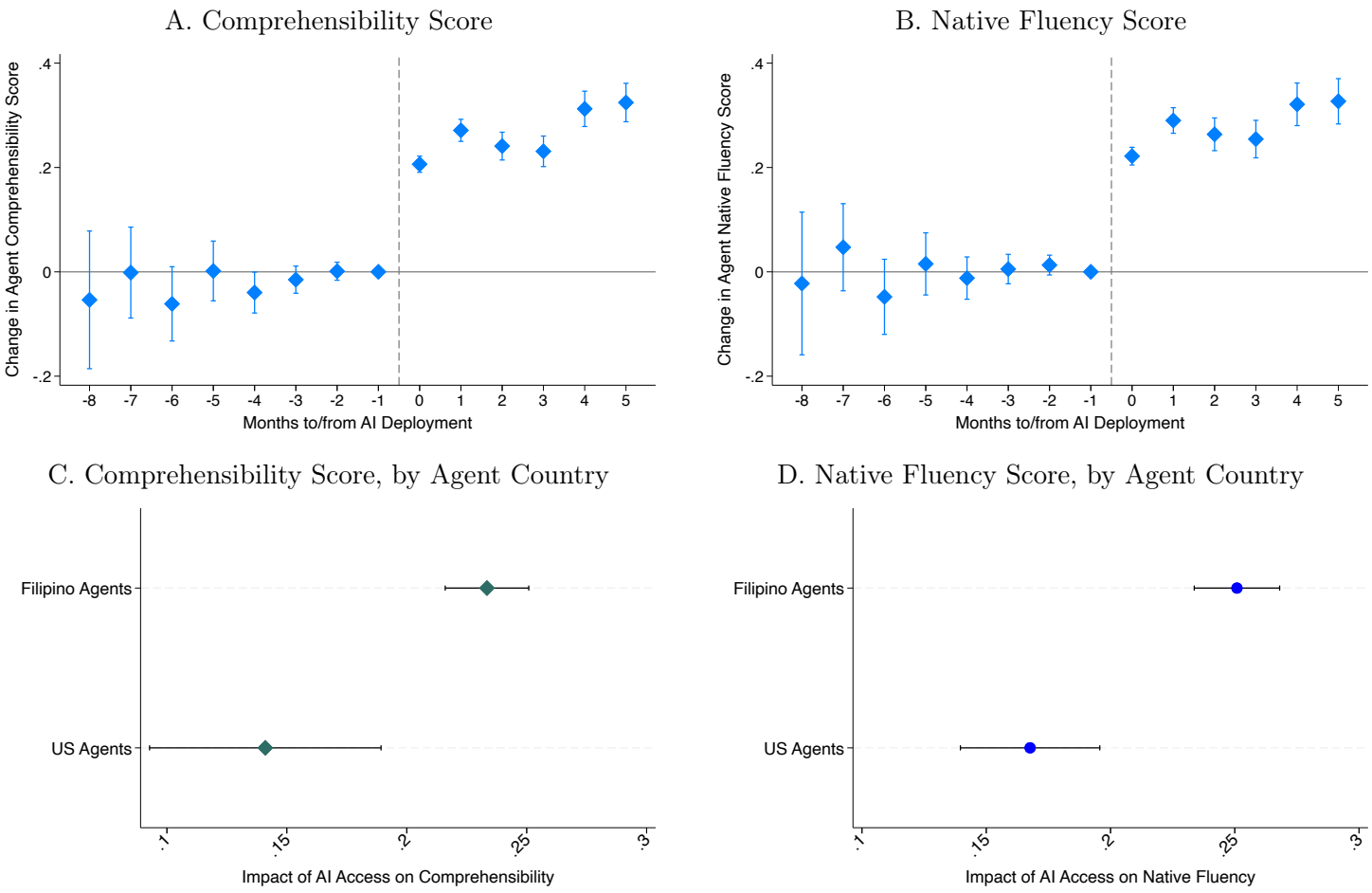


B. Impact of AI on Chat Duration, by Agent Topic Frequency



NOTES: 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. Appendix Section J.3 details the regression specification and topic category construction.

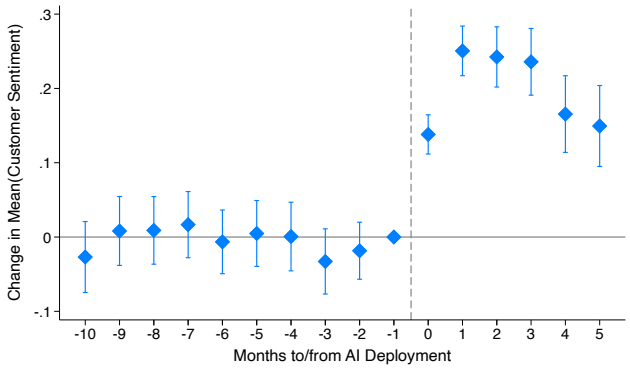
FIGURE IX
Impact of AI on Language Skills



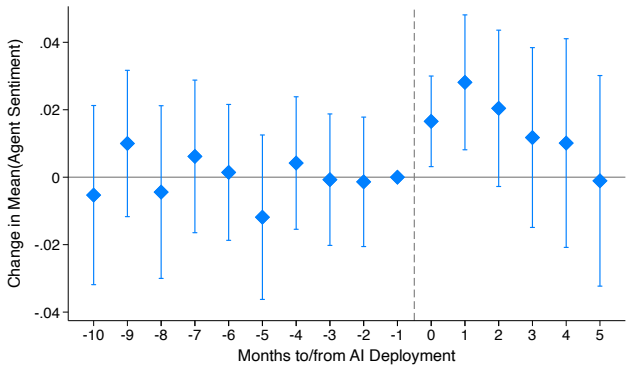
NOTES: These figures show the impact of AI access on scores of agent comprehensibility in Panel A and native fluency in Panel B. Observations for this regression are at the agent-chat level, aggregate to the agent-month level. Regressions follow Equation 1 and include agent, chat year-month and months of agent tenure fixed effects. Robust standard errors are clustered at the agent level in Panels A and B and agent location in Panels C and D. For more details on construction of the comprehensibility and native fluency scores, refer to Appendix Section J.2.

FIGURE X
Experience of Work

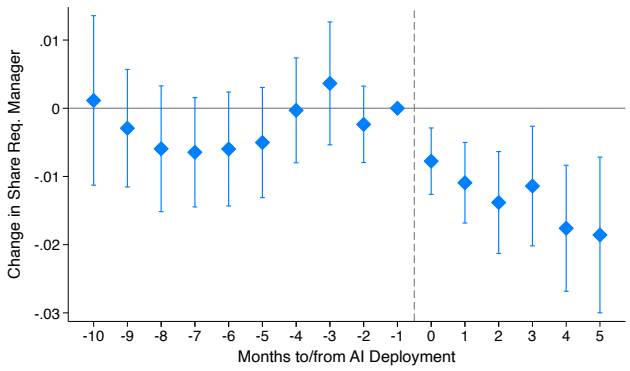
A. Customer Sentiment, Event Study



B. Agent Sentiment, Event Study



C. Manager Escalation, Event Study



NOTES: Each panel of this figure plots the impact of AI model deployment on the experience of work. Panel A plots the impact of AI model deployment on customer sentiment, Panel B plots the corresponding estimate for agent sentiment, and Panel C show the impacts of AI assistance on customer requests for manager assistance. Sentiment is measured using SiEBERT, a fine-tuned checkpoint of a 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.

Online Appendix:

GENERATIVE AI AT WORK

Erik Brynjolfsson

Danielle Li

Lindsey Raymond

November 18, 2024

JEL Classifications: D80, J24, M15, M51, O33

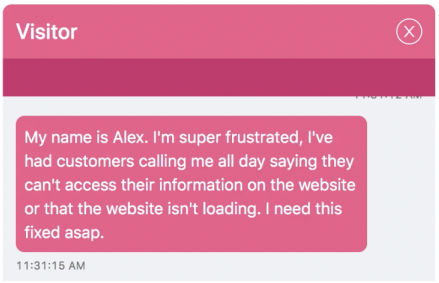
Keywords: Generative AI, Large Language Models, Technology Adoption, Worker Productivity, Worker Learning, Experience of Work, Organizational Design.

Contents

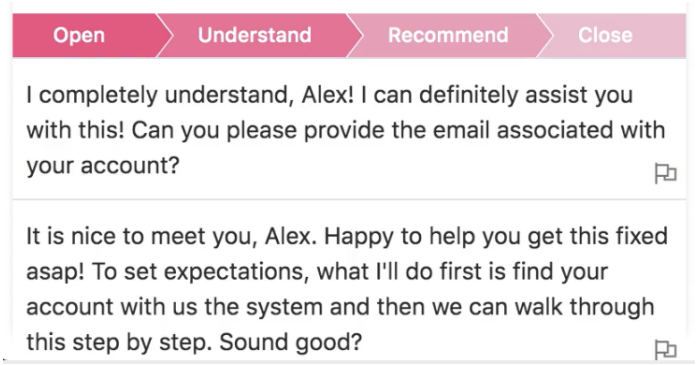
A AI System and Rollout

FIGURE A.I
Sample AI Output

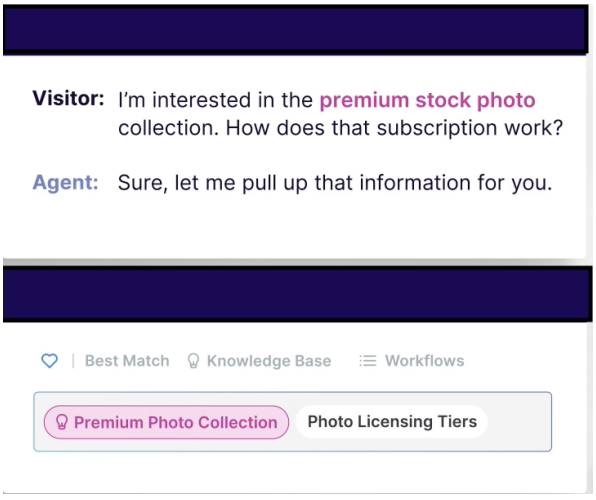
A. Sample Customer Issue



B. Sample AI-Generated Suggested Response

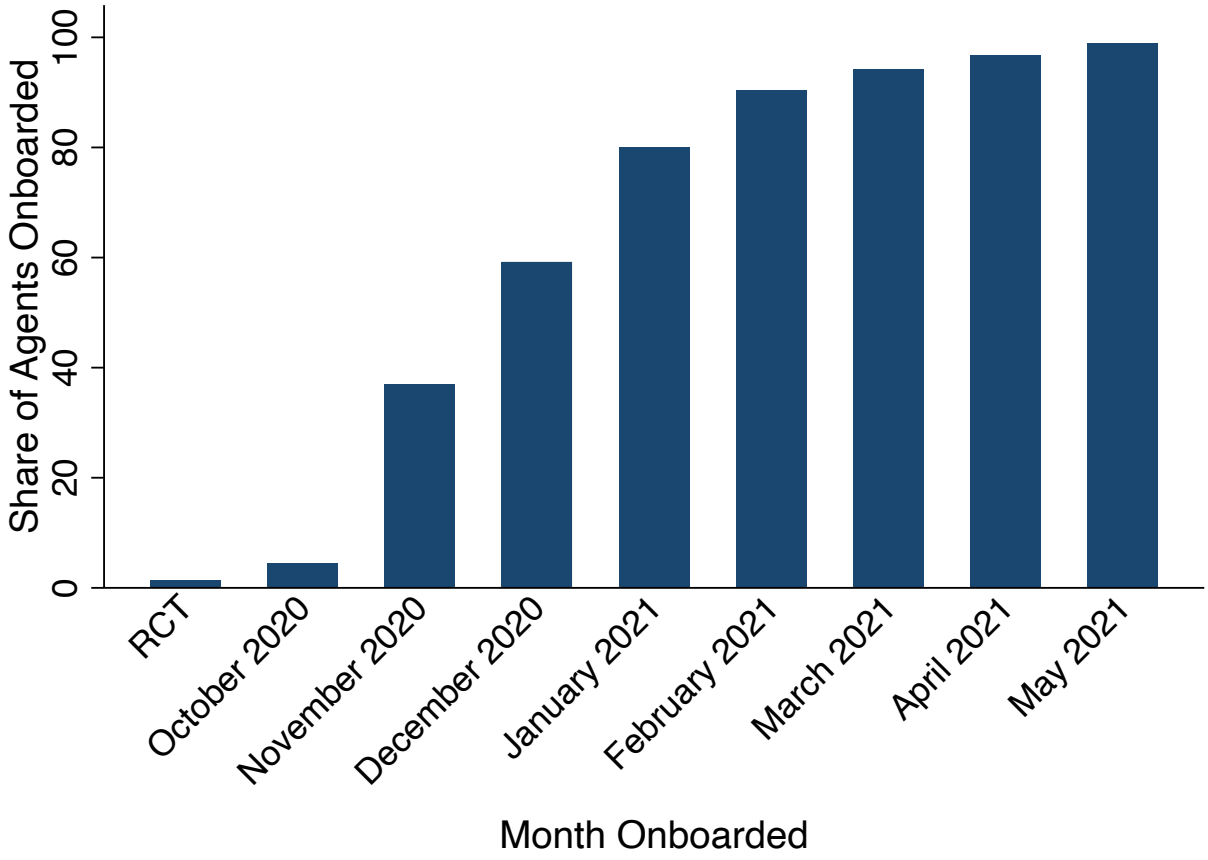


C. Sample AI-Generated Technical Link



NOTES: This figure illustrates AI-generated suggestions for customer service agents. Panel A shows a sample customer issue. Panel B displays AI-suggested responses for greeting and setting call expectations. Panel C presents a AI-recommended technical documentation excerpt from the company's internal knowledge base. Suggestions are only visible to agents, not customers, and agents can choose to use, modify, or ignore these suggestions when responding.

FIGURE A.II
Deployment Timeline



NOTES: This figure shows the share of agents deployed onto the AI system over the study period. Agents are deployed onto the AI system after a training session as described in Section 3.1. The small randomized control trial in August 2020 is analyzed in Section 4.1.1. All data are from the firm's internal software systems.

B Average Productivity Effects

TABLE A.I
Randomized Control Trial Analysis

VARIABLES	(1) Res./Hr	(2) AHT	(3) Chats/Hr	(4) Res. Rate	(5) NPS
Post AI X Ever Treated	0.202**	-3.713***	0.105	0.0169	-1.393
	(0.0850)	(1.045)	(0.0718)	(0.0246)	(1.529)
Observations	6,998	15,100	15,100	6,998	7,176
R-squared	0.568	0.574	0.566	0.383	0.517
Year Month FE	Yes	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes	Yes	Yes
DV Mean	1.956	43	2.378	0.808	79.68

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table focuses on the 22 agents who were onboarded as part of the pilot randomized control trial (RCT). Because we do not have information on the specific agents (there were around 25) selected to be part of the control group, we compare the 22 pilot-treated agents with observations for all pre-treatment agents, controlling for agent, chat year-month, and months of agent tenure fixed effects. Data are at the agent-month level and the standard errors are clustered at the agent level.

TABLE A.II
Main Effects: Company Team IV

VARIABLES	(1) Individual Treatment	(2) Res./Hr	(3) AHT	(4) Chats/Hr	(5) Res. Rate	(6) NPS
Earliest Team Treatment	0.311***					
	(0.0163)					
Post AI X Individual Treatment		0.550***	-3.622**	0.412***	0.0740***	-0.518
		(0.0943)	(1.444)	(0.0791)	(0.0184)	(0.946)
Observations	21,839	12,295	21,839	21,839	12,295	12,541
R-squared	0.813	0.177	0.102	0.115	0.007	0.007
Year Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	365.7					
Number of agent_id		2,163	3,633	3,633	2,163	2,214

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: In this table we instrument agent level date of AI adoption with the minimum date of AI deployment on that agent's team. Column 1 shows the first stage regression of earliest team adoption on individual adoption and Columns 2 through 6 shows the 2SLS estimates on measures of productivity and call quality and efficiency. Regressions follow Equation 1 and include agent level, chat-year, and months of agent tenure fixed effects. Data is available at the agent-month level and robust standard errors are clustered at the agent level. Information on team is missing for some workers. Workers without this information are grouped into a single "missing" team.

TABLE A.III
Main Effects: Robustness to Alternative Clustering

VARIABLES	(1) Res./Hr	(2) Res./Hr	(3) Res./Hr
Post AI X Ever Treated	0.301***	0.301***	0.301***
	(0.0329)	(0.0455)	(0.0498)
Observations	12,295	12,295	12,295
R-squared	0.575	0.575	0.575
Year Month FE	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes
DV Mean	2.176	2.176	2.176

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table shows robustness of our Column 3 in Table II to different levels of clustering. Column 1 clusters our robust standard errors at the agent level, Column 2 clusters at the company/team level and Column 3 clusters at the location (usually the agent's city of work). City of work may include call center locations for multiple companies (e.g. our data firm and an contracted outsourcing firm). Regressions follow Equation 1 and include agent level, chat-year, and months of agent tenure fixed effects. Data is available at the agent-month level and robust standard errors are clustered at the agent level.

TABLE A.IV
Main Effects: Chat-Weighted

VARIABLES	(1) Res./Hour	(2) AHT	(3) Chats/Hr	(4) Res. Rate	(5) NPS
Post AI X Ever Treated	0.252***	-3.100***	0.295***	0.00325	-0.119
	(0.0318)	(0.315)	(0.0263)	(0.00725)	(0.524)
Observations	12,295	21,839	21,839	12,295	12,541
R-squared	0.650	0.756	0.721	0.465	0.526
Year Month FE	Yes	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes	Yes	Yes
DV Mean	2.457	38.66	2.886	0.839	79.59

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on measures of productivity and agent performance. Observations are at the agent-month level, weighted by the number of chats associated with that agent-month. Post AI X Treated measures the impact of AI model deployment after deployment on treated agents for average handle time or average call duration, chats per hour, the number of chats an agent handles per hour, resolution rate, the share of technical support problems they can resolve and net promoter score (NPS), an estimate of customer satisfaction. Regressions follow Equation 1 and include agent level, chat-year, and months of agent tenure fixed effects. Data is available at the agent-month level and robust standard errors are clustered at the agent level.

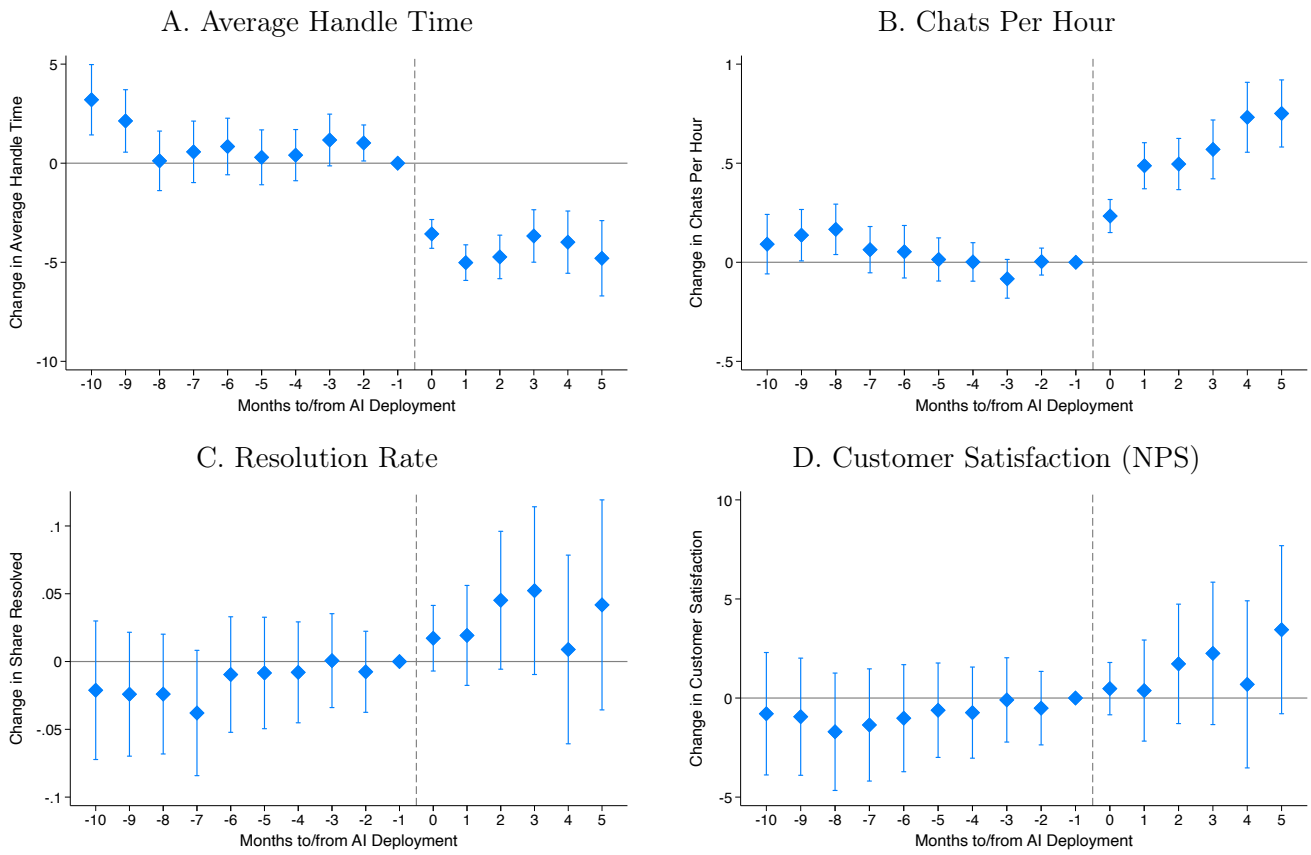
TABLE A.IX

Main Effects: Productivity (Resolutions per Hour), Alternative Difference-in-Difference Estimators

	Point Estimate	Standard Error	Lower Bound 95% Confidence Interval	Upper Bound 95% Confidence Interval
TWFE-OLS	0.296	0.032	0.233	0.360
Borusyak-Jaravel-Spiess	0.576	0.070	0.438	0.714
Callaway-Sant'Anna	0.489	0.059	0.374	0.605
DeChaisemartin-D'Haultfeuille	0.219	0.042	0.137	0.302
Sun-Abraham	0.521	0.094	0.337	0.705

NOTES: This table shows the impact of AI model deployment on our main productivity outcome, resolutions per hour, using robust difference-in-differences estimators introduced in Borusyak, Jaravel, and Spiess (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021). Regressions follow Equation 1 and include agent level, chat-year, and months of agent tenure fixed effects. Data is available at the agent-month level and standard errors are clustered at the agent level. Because of the number of post-treatment periods and high turnover of agents in our sample, we can only estimate five months of pre-period data using Borusyak, Jaravel, and Spiess (2022) and de Chaisemartin and D'Haultfœuille (2020).

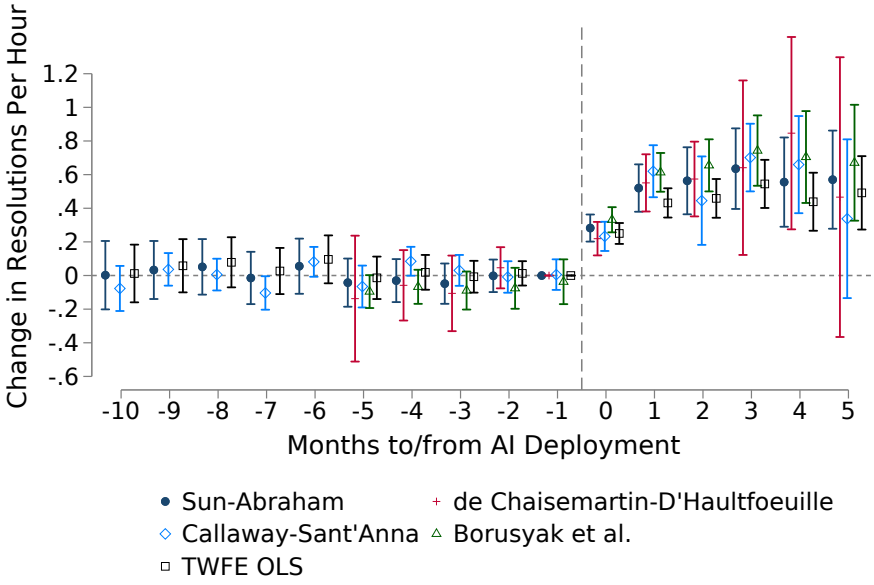
FIGURE A.III
Event Studies, Additional Outcomes



NOTES: These figures plot the coefficients and 95% confidence intervals from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. Panel A plots the average handle time or the average duration of each technical support chat. Panel B plots the number of chats an agent completes per hour, incorporating multitasking. Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots net promoter score, which is an average of surveyed customer satisfaction. All specifications follow Equation 1 and include agent and chat year-month and months of agent tenure fixed effects. Data is at the agent-month level and robust standard errors are clustered at the agent level.

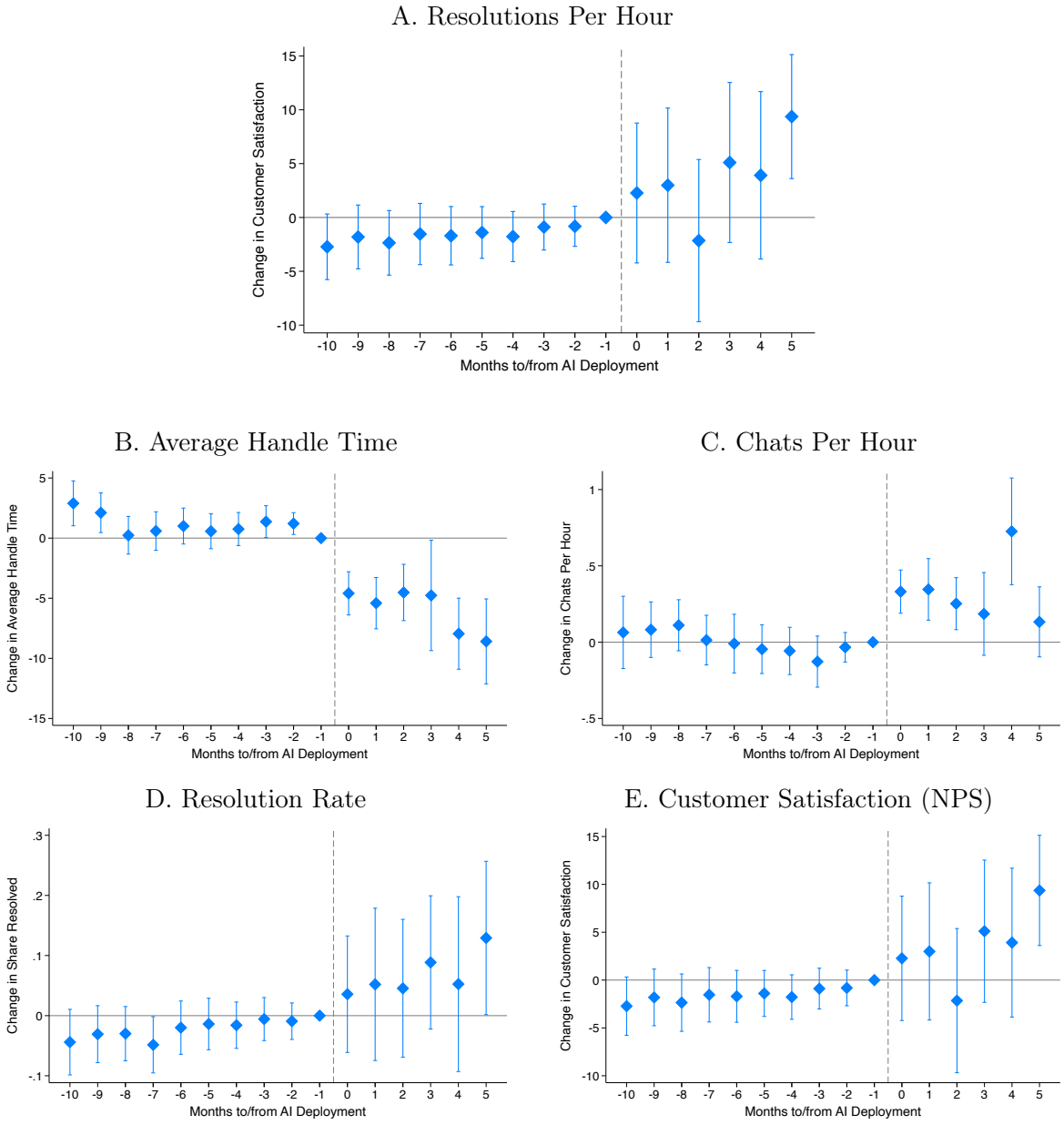
FIGURE A.IV
Event Studies, Resolutions Per Hour

A. Resolutions Per Hour



NOTES: This figure presents the effect of AI model deployment on our main productivity outcome, resolutions per hour, using a variety of robust dynamic difference-in-differences estimators introduced in Borusyak, Jaravel, and Spiess (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) and a standard two-way fixed effects regression model. Regressions follow Equation 1 and include agent level, chat-year, and months of agent tenure fixed effects. Data is available at the agent-month level and robust standard errors are clustered at the agent level. Because of the number of post-treatment periods and high turnover of agents in our sample, we can only estimate five months of pre-period data using Borusyak, Jaravel, and Spiess (2022) and de Chaisemartin and D'Haultfœuille (2020).

FIGURE A.V
Randomized Control Trial Analysis



NOTES: This figures plots event studies focusing on the 22 agents who were onboarded as part of the pilot randomized control trial (RCT). Because we do not have information on the specific agents (there were around 25) selected to be part of the control group, we compare the 22 RCT treated agents with observations for all pre-treatment agents, controlling for our usual agent, chat year-month, and months of agent tenure fixed effects in agent-month level regressions following Equation 1. Robust standard errors are clustered at the agent level.

C Impacts by Agent Skill and Tenure

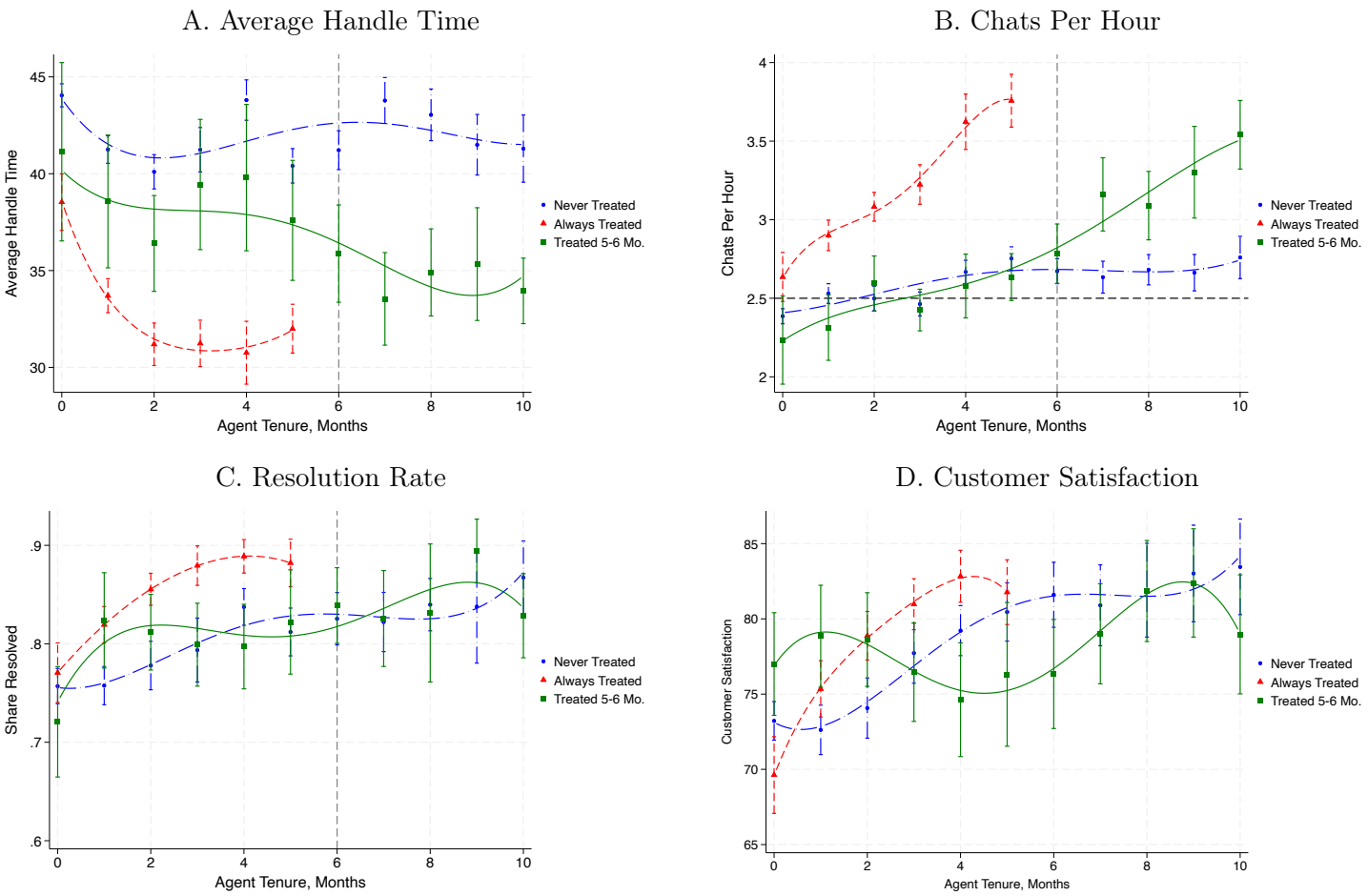
TABLE A.VI
Heterogeneity of AI Impact by Skill and Tenure, Resolutions per Hour

	(1) By Skill at AI	(2) By Tenure at AI
Q1 (Lowest Skill)	0.527***	
	(0.049)	
Q2	0.315***	
	(0.056)	
Q3	0.302***	
	(0.043)	
Q4	0.261***	
	(0.055)	
Q5 (Highest Skill)	0.015	
	(0.065)	
$<$ 1 Mos.)		0.707***
		(0.067)
1-2 Mos.		0.388***
		(0.042)
3-6 Mos.		0.361***
		(0.067)
7-12 Mos.		0.337***
		(0.054)
$>$ 12 Mos.)		0.028
		(0.059)
Year Month FE	Yes	Yes
Agent FE	Yes	Yes
Other FE	Tenure	Skill at AI
DV Mean	2.176	2.284
Observations	12,295	8,148

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

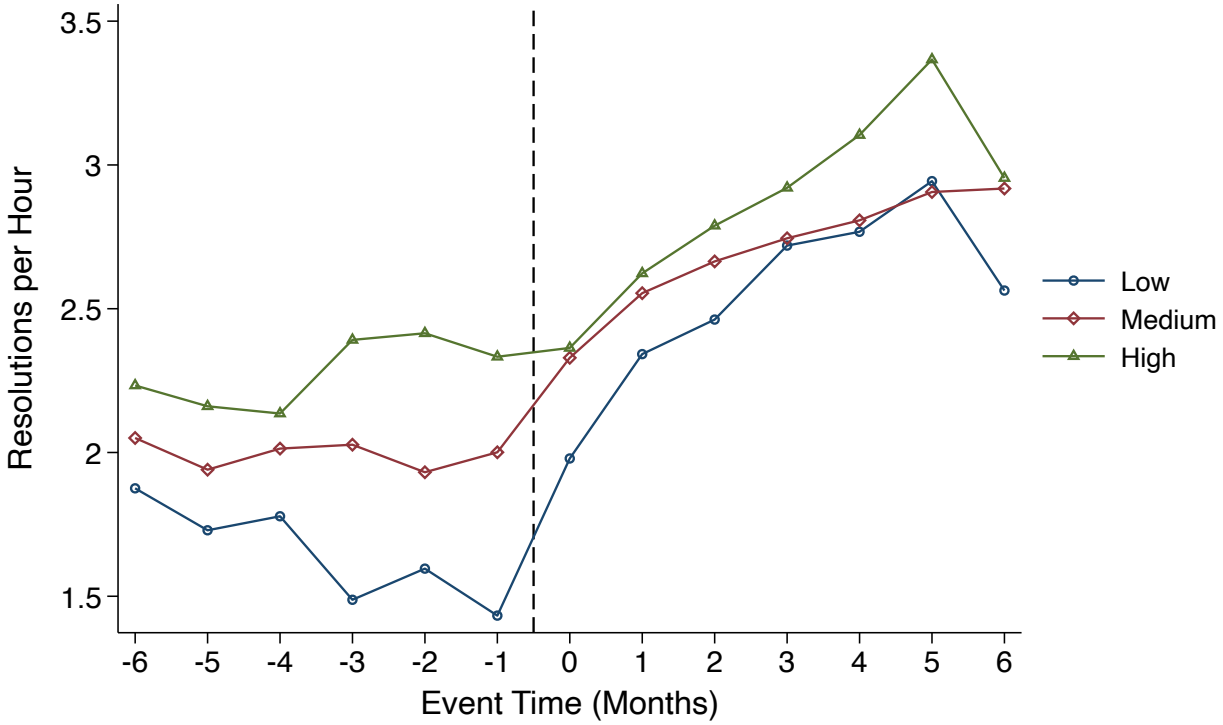
NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on resolutions per hour. Column 1 estimates the impact of AI access by worker skill at AI, including agent, year-month and months of agent tenure fixed effects. Column 2 estimates the effects by worker tenure at AI deployment, including agent, year-month and agent skill at AI deployment fixed effects. Observations are at the agent-month level and all standard errors are clustered at the agent level.

FIGURE A.VI
Experience Curves by Deployment Cohort, Additional Outcomes



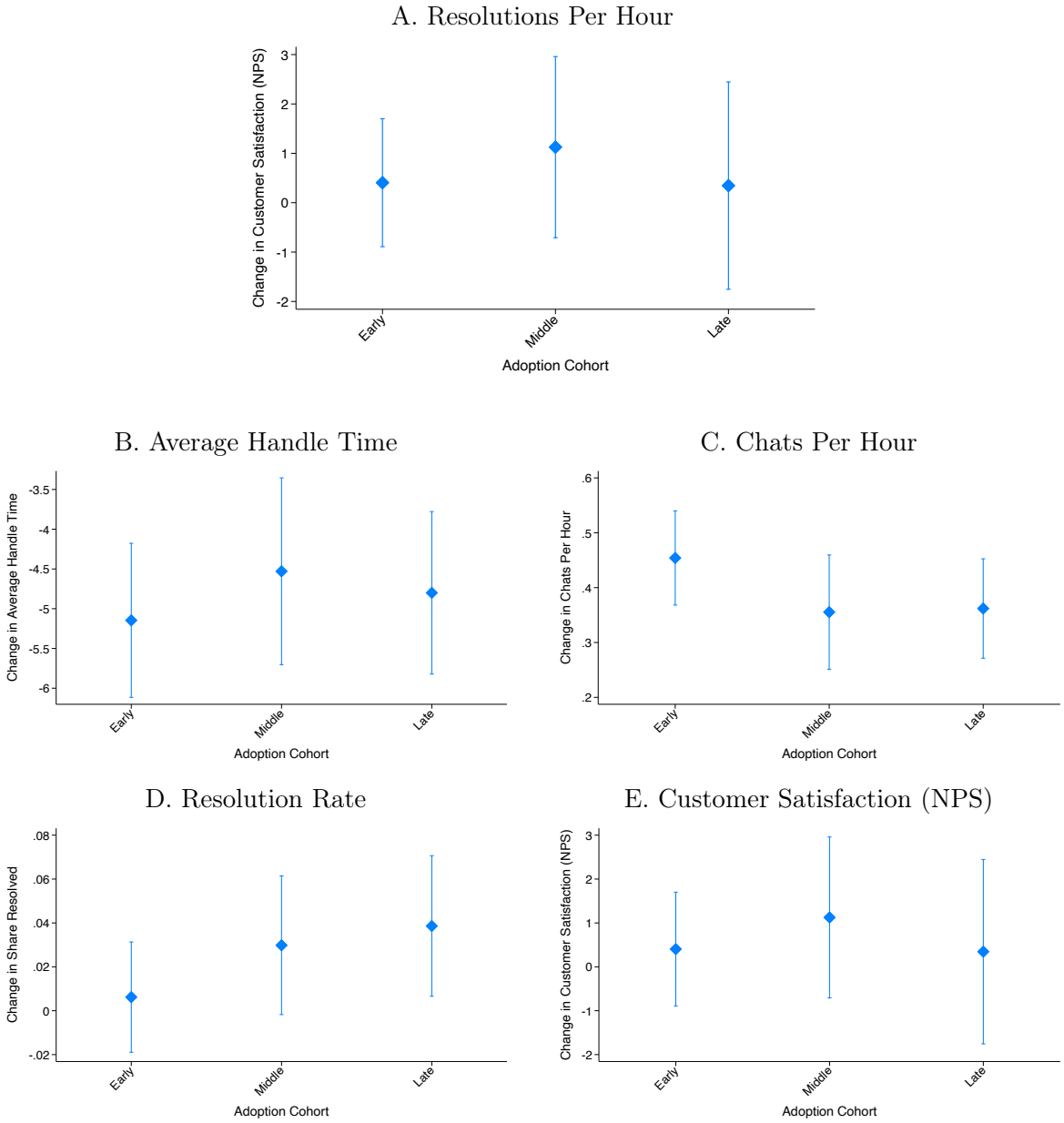
NOTES: These figures plot the experience curves of three groups of agents over their tenure, the x-axis, against five measures of productivity and performance. The short-dashed lines plot the performance of always-treated agents, those who are start work in their first month with the AI and always have access to the AI suggestions. 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 model, and gain access to the AI during their fifth month on the job. All panels include 95% confidence intervals and observations are at the agent-month level.

FIGURE A.VII
Resolutions per Hour over Time



NOTES: This graph depicts the evolution of average resolutions per hour for agents following the implementation of AI assistance. The graph segments agents into three groups based on their skill level at the time of AI deployment. The triangle-symbol line represents the highest-performing third of agents, those in the top tercile of the skill index. The diamond-symbol line illustrates the progress of agents in the middle tercile, while the circle-symbol line tracks those in the bottom tercile, representing the lowest-skilled third at the time of treatment. Agents are categorized based on their skill index at the time of AI implementation. For details on the skill index construction, refer to Appendix Section J.2.

FIGURE A.VIII
Productivity Impacts by Adoption Cohort



NOTES: These figures plot the treatment effect of AI assistance on various outcomes for workers who received access to the AI model in the early, middle, or later period of the rollout. Because the AI model is periodically updated with new data and pushed to all onboarded workers no more frequently than once a month, we focus on productivity outcomes in the first month after AI adoption in order to compare workers using earlier and later versions of the model. Data are at the agent-month, regression specification follows Equation 1 and we include chat year-month, agent, and months of tenure fixed effects. Robust standard errors are clustered at the agent level.

D Adherence to AI Suggestions

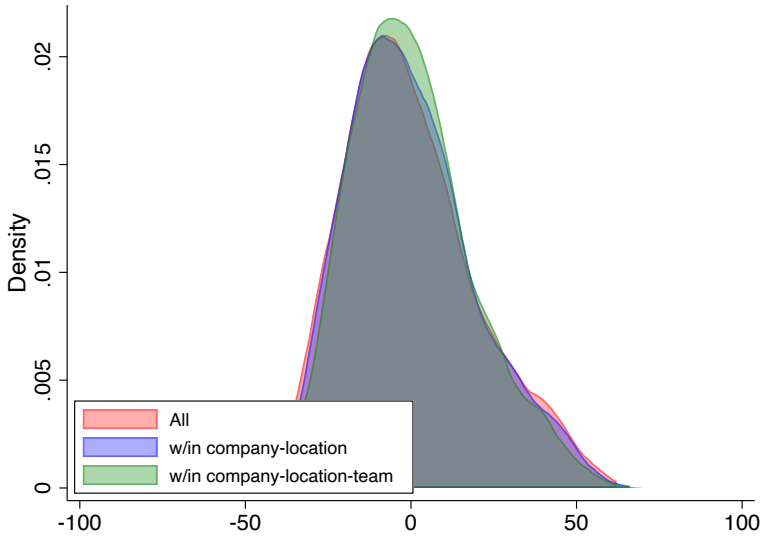
TABLE A.VII
Heterogeneity of AI Impact by Initial AI Adherence, Resolutions per Hour

	(1) By Adherence at AI
Q1 (Lowest Adherence)	0.213***
	(0.043)
Q2	0.233***
	(0.042)
Q3	0.293***
	(0.042)
Q4	0.309***
	(0.042)
Q5 (Highest Adherence)	0.432***
	(0.043)
Year Month FE	Yes
Agent FE	Yes
Agent Tenure FE	Yes
DV Mean	2.176
Observations	12,295

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

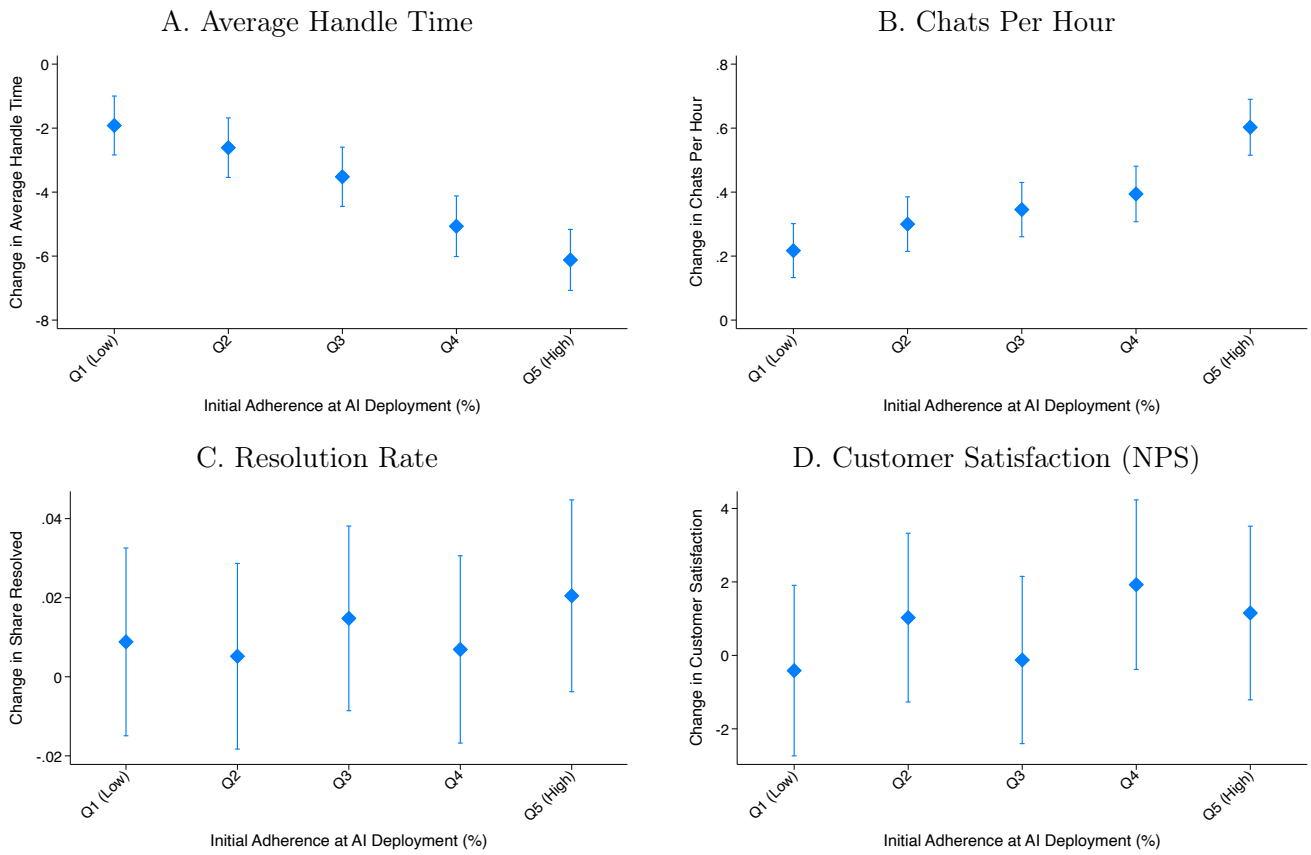
NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on resolutions per hour by initial adherence. Column 1 estimates the impact of AI access by initial adherence quintile, including agent, year-month and months of agent tenure fixed effects. Observations are at the agent-month level and all standard errors are clustered at the agent level. We include chat year-month, agent and months of agent tenure fixed effects.

FIGURE A.IX
Variation in AI Adherence



NOTES: This figure 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. "within company" refers to the overall distribution of adherence rates, adjusted by the company of employment. "within company-location" plots residual adherence after adjusting for company and location fixed effects, and "within company-location-team" plots residuals adjusted for company, location, and team fixed effects.

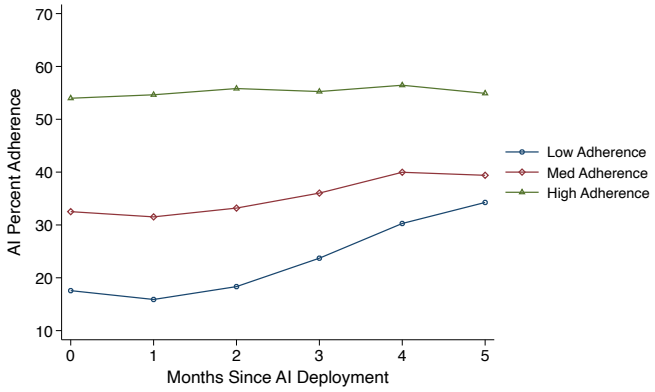
FIGURE A.X
Heterogeneity of AI Impact by Initial AI Adherence, Additional Outcomes



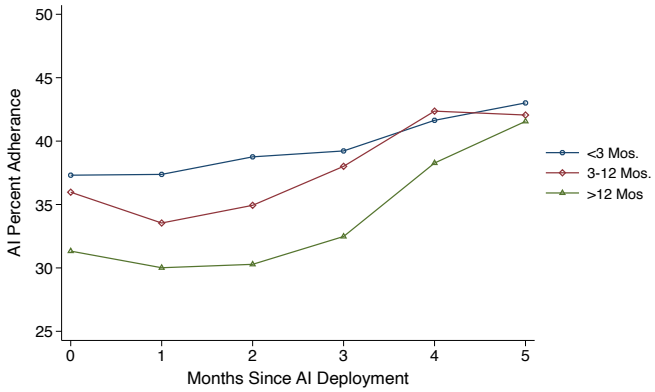
NOTES: These figures plot the impact of AI model deployment on additional measures of performance by quintile of initial adherence, the share of AI recommendations followed in the first month of treatment. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of chats an agent can handle per hour (including working on multiple chats simultaneously). Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots NPS, or net promoter score, is an average of surveyed customer satisfaction. Data is at the agent-level and all regressions include agent and chat year-month, months of agent tenure, with more details in Appendix section J.3.

FIGURE A.XI
AI Adherence over Time

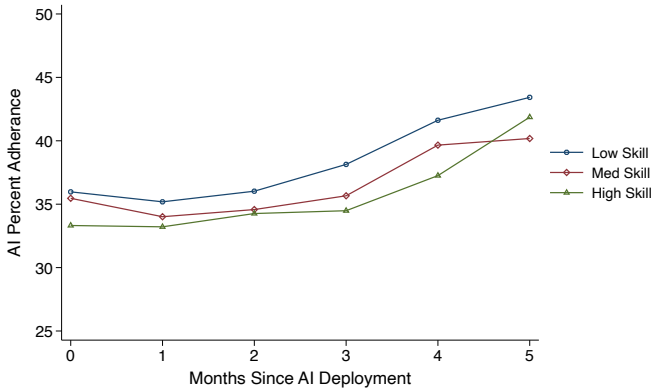
A. By Adherence at AI Model Deployment



B. By Agent Tenure at AI Model Deployment



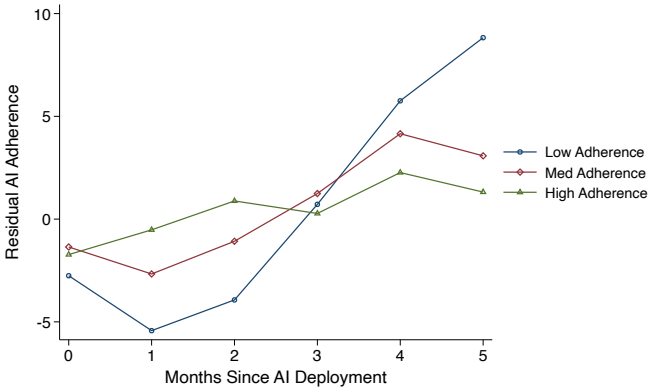
C. By Agent Skill at AI Model Deployment



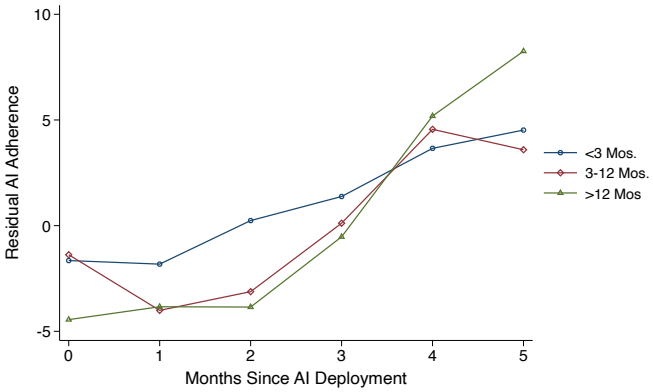
NOTES: This figure plots the share of AI suggestions followed by agents as a function of the number of months each agent has had access to the AI model. In Panel A, we divide agents into terciles based on their adherence to AI suggestions in the first month. In Panel B, we divide agents into groups based on their tenure at the firm at the time of AI model deployment. In Panel C, we divide workers into terciles of pre-deployment productivity as defined by our skill index.

FIGURE A.XII
Within-Agent AI Adherence over Time

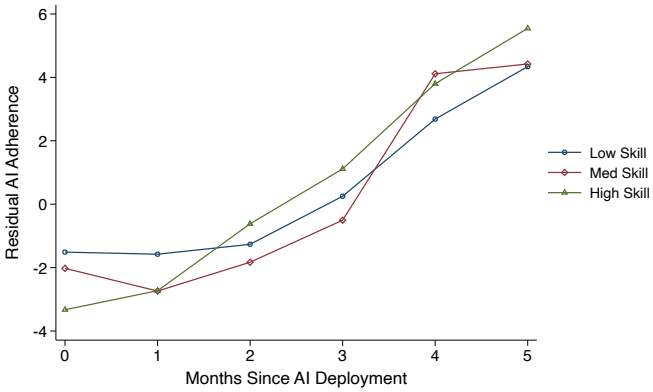
A. By Adherence at AI Model Deployment



B. By Agent Tenure at AI Model Deployment



C. By Agent Skill at AI Model Deployment



NOTES: This figure plots the residualized percentage of AI suggestions followed by agents as a function of the number of months each agent has had access to the AI model, after controlling for agent level fixed effects. In Panel A, we divide agents into terciles based on their adherence to AI suggestions in the first month. In Panel B, we divide agents into groups based on their tenure at the firm at the time of AI model deployment. In Panel C, we divide workers into terciles of pre-deployment productivity as defined by our skill index.

E Worker Learning

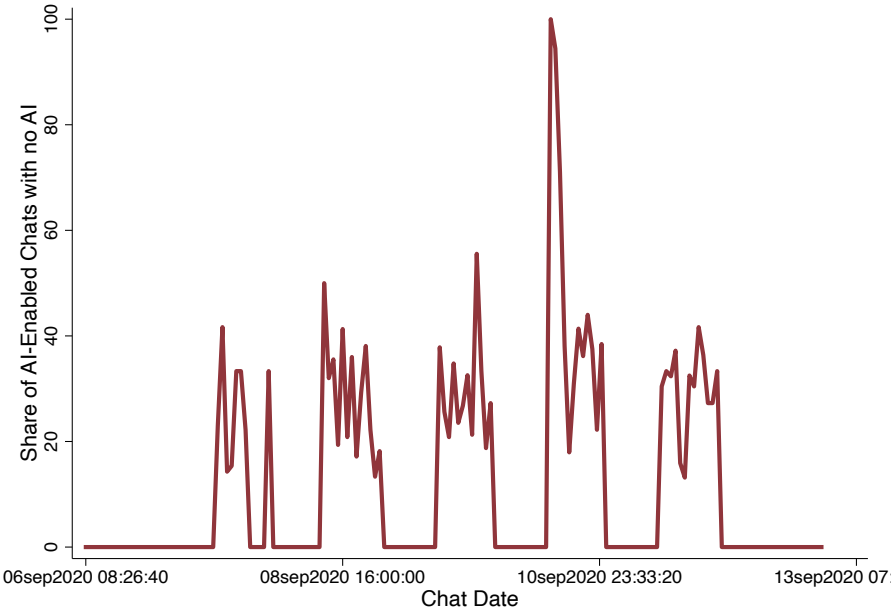
TABLE A.VIII
Chat Duration during AI System Outages

	(1)	(2)	(3) Outages	(4) Outages
	Non Outage	AI Outages	High Receptivity	Low Receptivity
lead10	0.684	2.085***	3.391***	0.284
	(0.550)	(0.581)	(0.880)	(1.590)
lead9	0.549	2.126***	4.056***	1.003
	(0.529)	(0.551)	(0.890)	(1.388)
lead8	-0.222	1.410***	3.239***	0.340
	(0.479)	(0.504)	(0.771)	(1.299)
lead7	-0.302	1.267***	2.354***	1.025
	(0.470)	(0.486)	(0.798)	(1.141)
lead6	0.874*	2.204***	4.081***	2.210**
	(0.463)	(0.468)	(0.786)	(1.079)
lead5	0.439	1.444***	2.953***	1.736*
	(0.389)	(0.378)	(0.591)	(0.885)
lead4	0.415	1.052***	2.222***	1.258*
	(0.336)	(0.321)	(0.471)	(0.703)
lead3	0.135	0.343	1.314***	0.218
	(0.305)	(0.297)	(0.482)	(0.627)
lead2	-0.146	-0.209	0.723*	-0.869*
	(0.248)	(0.247)	(0.411)	(0.469)
lag0	-4.105***	-3.471	-5.236	9.410***
	(0.270)	(3.713)	(3.732)	(3.564)
lag1	-4.523***	-2.420	-6.699*	0.000
	(0.313)	(3.242)	(4.052)	(.)
lag2	-4.785***	0.767	-0.201	0.000
	(0.335)	(3.112)	(3.858)	(.)
lag3	-4.424***	-2.851	-2.677	6.005
	(0.333)	(2.663)	(3.915)	(4.099)
lag4	-4.225***	-2.423*	-4.011**	4.851
	(0.350)	(1.442)	(2.015)	(3.663)
lag5	-4.084***	0.838	0.695	7.871
	(0.380)	(1.744)	(2.042)	(5.904)
lag6	-4.300***	-7.799***	-8.574***	-2.303
	(0.388)	(1.836)	(2.168)	(3.167)
lag7	-4.445***	-8.494***	-9.532***	-2.161
	(0.393)	(1.635)	(1.687)	(3.209)
lag8	-3.804***	-4.672*	-2.163	-8.059***
	(0.411)	(2.647)	(2.471)	(3.023)
lag9	-3.628***	-8.705***	-8.842***	0.318
	(0.440)	(1.660)	(1.930)	(5.166)
lag10	-3.609***	-4.386**	-4.486**	-0.233
	(0.501)	(1.720)	(2.040)	(4.273)
lag11	-3.305***	-2.779**	-6.857**	0.539
	(0.486)	(1.380)	(3.252)	(4.003)
lag12	-3.479***	-1.966	-4.630	-8.939**
	(0.517)	(3.003)	(4.248)	(3.510)
Year Month FE	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes	Yes
DV Mean	41.982	41.982	41.982	41.982
R-squared	.12	.104	.113	.083
Observations	2,969,371	1,829,527	1,220,852	323,810

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

NOTES: This table shows the event study coefficients corresponding to the figures in VII. Column 1 shows the impact of AI access on call duration during non-outage periods, Column 2 shows the impact when the AI system is experiencing an outage. Column 3 shows the effects on agents who are in the top tercile of AI adherence, while Column 4 shows the impacts on agents in the lowest tercile of adherence. Observations for these OLS regressions are at the agent-chat level and robust standard errors are clustered at the agent level. All specifications include agent, chat year-month, and months of agent tenure fixed effects.

FIGURE A.XIII
Sample AI Outage



NOTES: This figure plots the share of post-treatment chats with no AI suggestions during a period of a documented software outage.

F Topic Heterogeneity

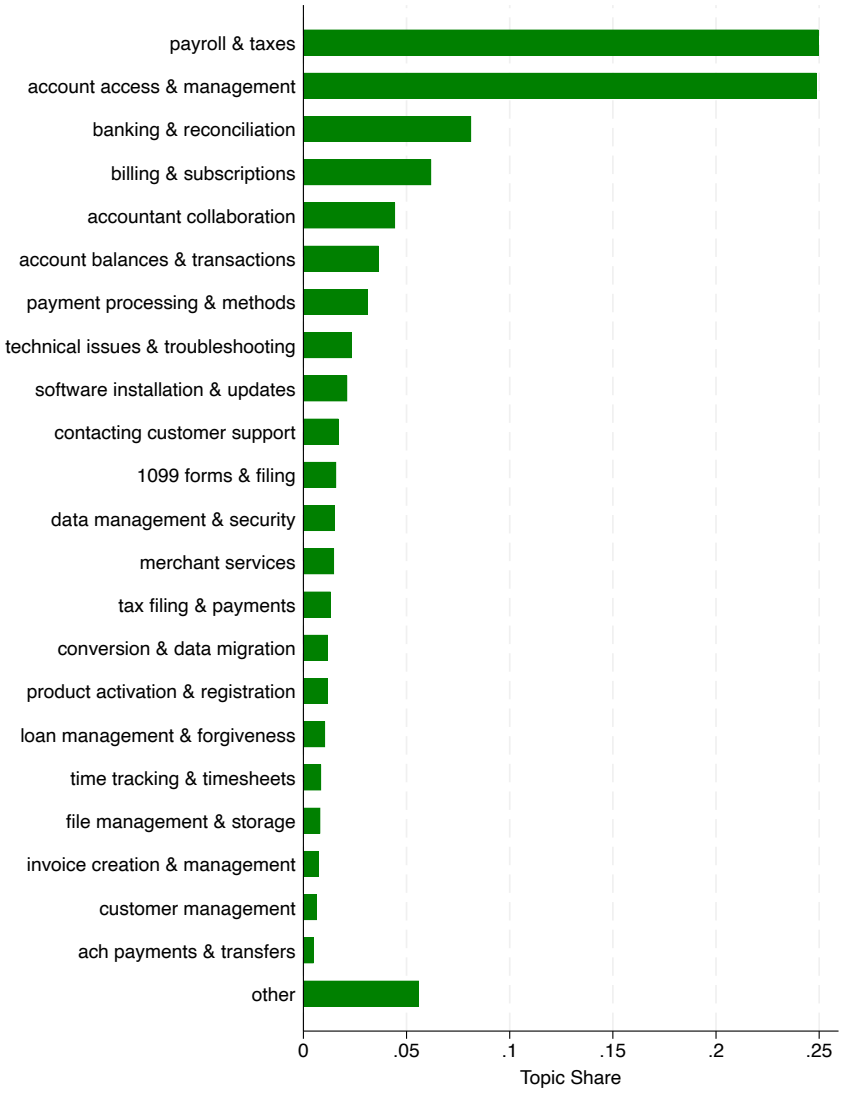
TABLE A.IX
Impact of AI on Chat Duration, by Topic Commonality

VARIABLES	(1) Call Duration Aggregate Problem Frequency	(2) Call Duration Within-Agent Frequency
Q1 (Most Common Overall)	-0.107***	
	(0.00717)	
Q2	-0.115***	
	(0.00724)	
Q3	-0.136***	
	(0.00742)	
Q4 (Least Common Overall)	-0.118***	
	(0.00842)	
Q1 (Most Common within Agent)		-0.100***
		(0.00927)
Q2		-0.103***
		(0.00740)
Q3		-0.119***
		(0.00700)
Q4 (Least Common within Agent)		-0.142***
		(0.00708)
Observations	2,089,995	2,089,995
R-squared	0.120	0.122
Year Month FE	Yes	Yes
Agent FE	Yes	Yes
Agent Tenure FE	Yes	Yes
DV Mean	0.916	0.915

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table shows the impact of AI model deployment on call duration by frequency of the chat topic. Regression include controls for chat year-month, agent and months of agent tenure. Data is at the chat level and robust standard errors are clustered at the agent level. Appendix section J.2 describes construction of topics and the regression specifications.

FIGURE A.XIV
Distribution of Conversational Topics



NOTES: This figure shows the distribution of chats by common topics in our data.

G Language Fluency

TABLE A.X
AI Impact on Language Fluency and Comprehensibility

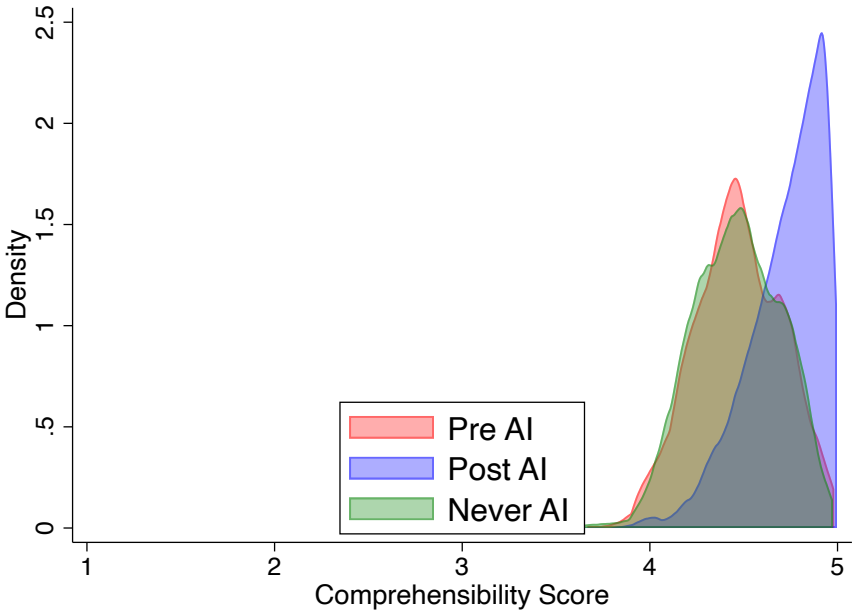
VARIABLES	(1) Native Fluency	(2) Comprehensibility	(3) Native Fluency	(4) Comprehensibility
Post AI X Ever Treated	0.250***	0.241***		
	(0.00851)	(0.00767)		
Post AI X Ever Treated X US = 1			0.159***	0.134***
			(0.0225)	(0.0208)
Post AI X Ever Treated X Philippines = 1			0.251***	0.234***
			(0.00684)	(0.00672)
Observations	12,772	12,772	11,271	11,271
R-squared	0.791	0.754	0.733	0.751
Year Month FE	Yes	Yes	Yes	Yes
Agent FE	Yes	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes	Yes
DV Mean	4.564	4.597	4.592	4.608

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

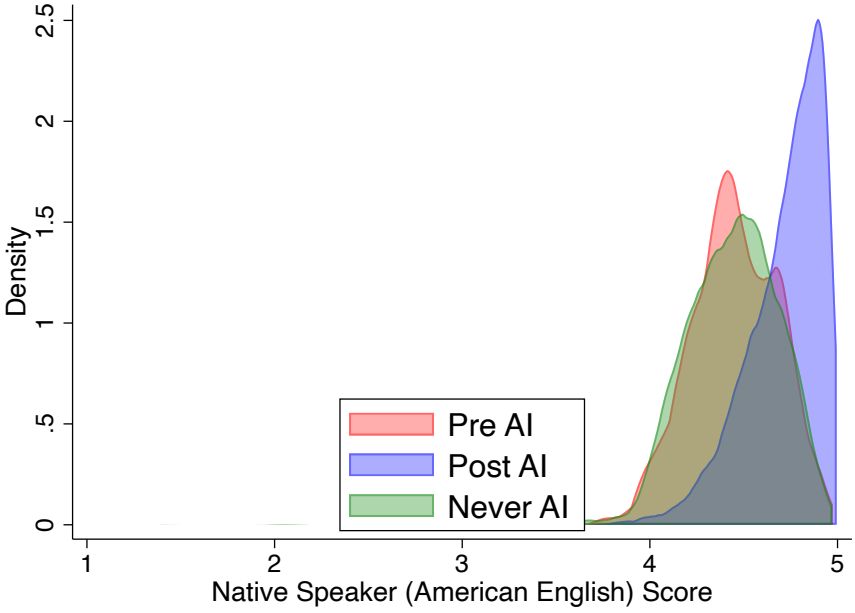
NOTES: This table presents the results of difference-in-difference regressions estimating the impact of AI model deployment on native fluency and comprehensibility. Observations are aggregated to the agent-month level and regressions include chat year-month, agent and months of agent tenure fixed effects. In Columns 1 and 2, robust standard errors are clustered at the agent level and at the agent location level in Columns 3 and 4. Appendix Section J.2 describes construction of the language scores.

FIGURE A.XV
Distribution of Language Skills

A. Comprehensibility Score



B. Native Fluency Score

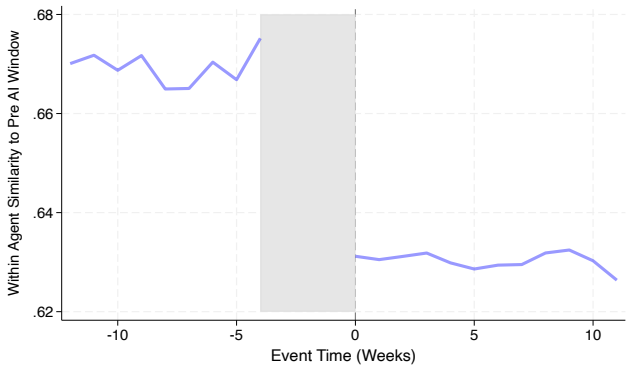


NOTES: These figures show the distributions of comprehensibility and native fluency scores. We split this sample into agent-month observations for agents who eventually receive access to the AI system before deployment ("Pre AI"), after deployment ("Post AI"), and for agent-months associated with agents who never receive access ("Never AI"). Appendix Section J.2 describes construction of the language scores.

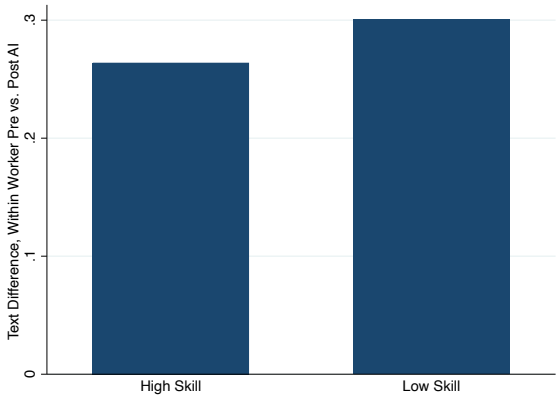
H Textual Convergence

FIGURE A.XVI
Within-Agent Textual Analysis

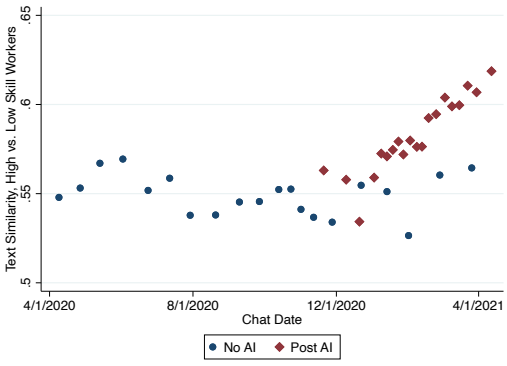
A. Within-Person Textual Similarity to Month Prior to AI



B. Within-Person Textual Change, Low vs. High Skill



C. Text Similarity between Low-Skill and High-Skill Workers, Pre and Post AI



NOTES: Panel A plots the average similarity between an agent's chats each week and a comparison group of their conversations in the month prior to AI deployment (the gray section). Panel B plots the average difference between an agent's pre-AI corpus of chat messages and that same agent's post-AI corpus, controlling for year-month and agent tenure. The first bar represents the average pre-post text difference for agents in the highest quintile of pre-AI skill and the second bar represents those in the bottom quintile. Panel C plots the average text similarity between the

top and bottom quintile of agents. The circular points plot the similarity for never treated or pre-treatment agents, the square points plot the similarity for agents with access to the AI model. For agents in the treatment group, we define agent skill at AI model deployment.

I Experience of Work

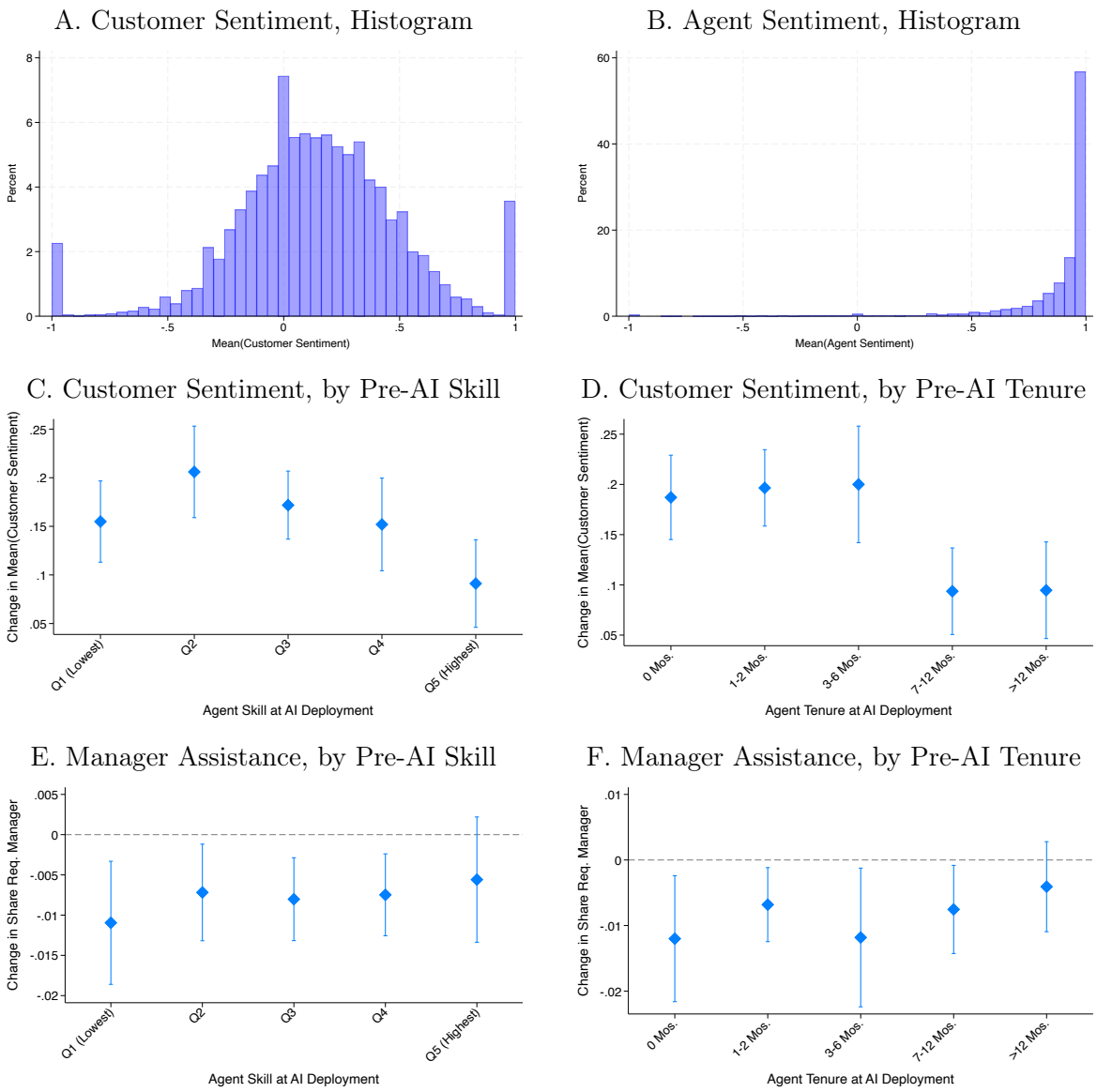
TABLE A.XI
Attrition

VARIABLES	(1) Leaves this Month	(2) Leaves this Month	(3) Leaves this Month
Post AI X Ever Treated	-0.0868***		
	(0.0130)		
0 Mos.		-0.0952***	
		(0.0162)	
1-2 Mos.		-0.121***	
		(0.0167)	
3-6 Mos.		-0.0850***	
		(0.0190)	
7-12 Mos.		-0.0803***	
		(0.0165)	
>12 Mos.		-0.0306	
		(0.0234)	
Q1 (Low Skill)			-0.0655***
			(0.0148)
Q2			-0.0864***
			(0.0134)
Q3			-0.0741***
			(0.0124)
Q4			-0.0936***
			(0.0146)
Q5 (High Skill)			-0.0531***
			(0.0145)
Observations	17,902	17,902	17,902
R-squared	0.206	0.206	0.206
Year Month FE	Yes	Yes	Yes
Location FE	Yes	Yes	Yes
Agent Tenure FE	Yes	Yes	Yes
Agent Company FE	Yes	Yes	Yes
DV Mean	0.288	0.288	0.288

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

NOTES: This table presents the results of difference-in-difference regressions estimating the impact 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 X 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 3.1 describes the AI rollout procedure.

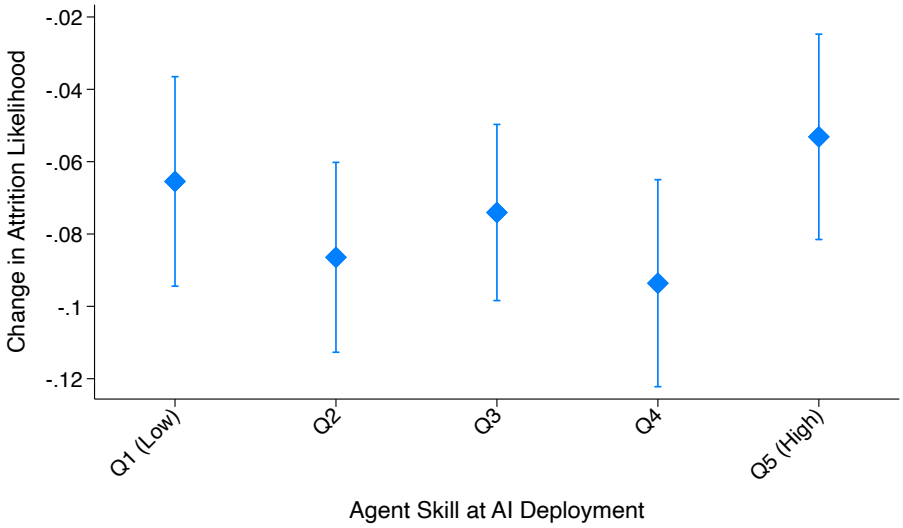
FIGURE A.XVII
Experience of Work



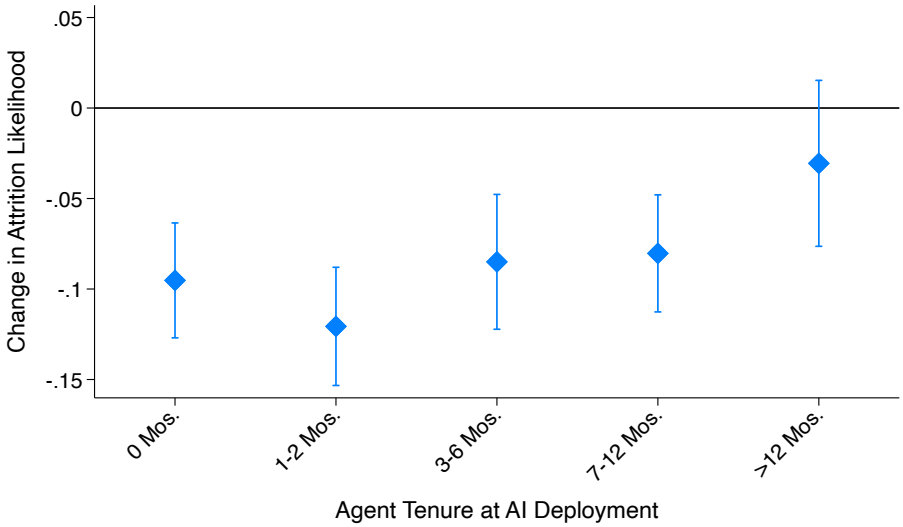
NOTES: Each panel of this figure illustrates the impact of AI model deployment on aspects of the experience of work. Panel A shows average customer sentiment, while Panel B shows average agent sentiment. Sentiment is measured using SiEBERT, a fine-tuned checkpoint of a RoBERTA, an English language transformer model. Panel C plots the impacts of AI on customer sentiment by agent ex-ante productivity and Panel D plots the effects by agent tenure when the AI is deployed. Panels E and F show the effects of AI on customer requests for manager assistance, by pre-AI agent skill and in by pre-AI agent tenure. Observations for the regression are at the agent-month level, robust standard errors are clustered at the agent level, and regressions include controls for agent, chat year-month and months-of-tenure.

FIGURE A.XVIII
Impact of AI Model Deployment on Worker Attrition

A. By Productivity at AI Model Deployment



B. By Tenure at AI Model Deployment



NOTES: This figure presents the results of the impact of AI model deployment on workers' likelihood of attrition. Panel A plots the same impact by agent skill index at AI model deployment. Panel B graphs the effects of AI assistance on attrition by agent tenure at AI model deployment. All specifications include chat year and month fixed effects, as well as agent location, company and agent tenure. Observations for these regressions, detailed in J.3, are at the agent-month and all robust standard errors are clustered at the agent level.

J Data Appendix

J.1 Sample Construction

We begin with 3,006,395 chat conversations conducted by agents employed at our data firm over the period between September 2019 and June 2021. Each chat includes the number of messages in the conversation, the start and end times, and identifiers associated with the conversation and agent. We drop chats with only one agent or one customer message to avoid capturing interactions without meaningful content.

We merge the chat data with a set of internal company datasets and datasets from our AI firm that allow us to track agent information, conversation outcomes, and AI model output. To do this, we use the system-generated chat identifier to merge chat-level information across database systems. We also merge our conversation-level data into a message-level dataset that includes the text of each message and message-level AI output. At the conversation level, the difference between the chat start and end time gives us the chat duration, which is available for each call. We drop the small number of chats that are missing start or end times, not associated with an agent identifier or missing a chat identifier. There are also a small number of chats that remain "open" for days because the customer and agents forget to close the chat, so we winsorize call duration at the 99th percentile.

We aggregate our chat-level dataset to the agent-month level and merge it with agent data. This information includes the firm they are employed by, their location, their manager/team, their tenure at the firm, and the date treated agents are onboarded onto the AI assistant (defined as the date the agent's account on the AI system is created). Some employees work flexible schedules, including part-time or seasonal roles. As a result, we measure performance metrics only for active periods, and an agent's tenure increases solely during months when they are actively handling customer chats.

J.2 Construction of Key Variables

J.2.1 Call Duration, Resolution and Customer Satisfaction

Our firm and associated subcontractors track call resolution, average call duration, and customer satisfaction at the agent-month level. Customer satisfaction is collected by randomly sending surveys to customers who have interacted with contact center agents. From the survey results, our firm calculates a monthly average agent customer satisfaction score. Not all customers complete these surveys, so customer satisfaction is calculated monthly rather than at the chat level. Call resolution

is also calculated at the agent-month level. To calculate call resolution, our data firm also uses an algorithm that incorporates elements of chat text and future interactions between the customer and data firms to calculate a monthly average call resolution score. Although our firm calculates an agent-month average monthly call duration when tracking agent performance, our chat-level dataset allows us to calculate call duration data for each chat in our sample. However, because call resolution is only available on a monthly basis, we report our omnibus productivity measure at the agent-month level.

J.2.2 Measuring Agent Skill, Firm and Tenure

We measure the tenure of the agent in months of work experience. The turnover of the contact center is extremely high; Our firm considers agents with over 6 months of experience to be "very experienced" while agents during their first two months are considered to be "in training". In our regression specifications, we control for agent tenure using a set of fixed effects for each month of agent tenure. We do not count weeks when agents are not actively working due to vacation or management practices like staggered scheduling intended to reduce worker burnout.

Our regressions also include controls for each agent's firm, which is the company or subcontractor employing each agent. Agents are employed directly by our data firm as well as by four other business process outsourcing firms. For instance, the firm name for an outsourced worker is "Convergys", one of our outsourcing firms, while the company of a directly employed call center agent is "data firm."

We also include controls for agent location—the physical location where the agent is employed. For example, several of our subcontractor have locations in Cebu City, so many agents are located in Cebu. While many of the US-based call center agents are employed directly by our data firm, the outsourcing firms also employ US-based workers. Contact center agents based in the US tend to be clustered in cities with a low cost of living, such as Hazard, Kentucky, and Reno, Nevada. Outside of the United States, most workers are based in the Philippines. The turnover of workers is high, so contact center firms frequently open and close new locations.

We construct an agent skill index that incorporates call handling speed, issue resolution rates, and customer satisfaction. For each firm, we rank agents based on three monthly performance indicators: average handle time (lower times receive higher ranks), call resolution rate, and customer satisfaction (higher rates receive higher ranks). The agent's rank is calculated within firm; data firm agents are ranked against their peers, while the Convergys agents are compared to other Convergys employees. These rankings are then averaged into a single skill index at the agent-month level. We then categorize agents into quintiles based on their average skill index from the previous quarter.

The skill index incorporates call speed and quality, while comparing individuals within the same organization.

J.2.3 Productivity

Our omnibus productivity measure is monthly resolutions per hour. To calculate this, we divide the share of calls resolved (the monthly call resolution score) by the number of calls handled per hour. Although call duration is available for each chat, resolution data is only available at the agent-month level. Consequently, we report our productivity results at the agent-month level.

J.2.4 Customer Sentiment

The text-based nature of customer support sheds light into the on-the-job experience of contact center agents. Although a call may ultimately be resolved or receive a high customer satisfaction score, customers often start conversations feeling stressed, frustrated, or angry. Regularly dealing with complaints and dissatisfied customers is emotionally taxing and contributes to high worker turnover. Conversations also tend to be fairly long, close to 40 minutes, and often involve complex topics such as managing payroll for employees, issues connecting with banks, and calculating taxes.

To capture the experience of customer-agent interactions, we use natural language processing to capture the sentiment by analyzing the language, context, and emotional content of chat transcript. We employ SiEBERT, an LLM fine-tuned for sentiment analysis on numerous datasets, including product reviews and tweets, with similar syntax and content to chat-based custoemr support. This model has demonstrated robustness to noisy, real-world data, such as social media posts or customer reviews. SiEBERT uses the text surrounding each word to capture word meaning, which as been shown to outperform other methods of sentiment analysis (Hartmann et al., 2023).

For each piece of text, the model produces a sentiment score measured on a scale from -1 to 1, where -1 indicates negative sentiment and 1 indicates positive. We separately calculate sentiment scores for agents' text (agent sentiment) and customers' text (customer sentiment). We then aggregate these chat-level variables into measures of average agent sentiment and average customer sentiment for each agent-month.

J.2.5 Language Comprehensibility and Fluency

We use Gemini Pro, a large language model (LLM), to measure comprehensibility and native fluency (Gemini Team, 2024). Our criteria for native-like fluency are based on the Interagency Language Roundtable (ILR) "functionally native" language proficiency standard, adapted for written text.

The ILR is an organization comprising federal government agencies that coordinates and shares information on foreign language activities. Functionally Native Proficiency, the highest level on the ILR scale, describes language ability where an individual communicates with complete fluency and precision on all levels normally pertinent to professional needs, displays cultural understanding equivalent to a native speaker and demonstrates the ability to counsel, persuade, and negotiate in the language as effectively as a well-educated native speaker.

Using the prompt below, we ask the LLM to score each agent's text from a scale of 1 to 5 where: 1 = Definitely not a native American English speaker 2 = Probably not a native American English speaker 3 = Uncertain if native American English speaker 4 = Probably a native American English speaker 5 = Definitely a native American English speaker. The full prompt is:

> You are given a conversation transcript from a customer service agent helping a customer that only includes what the customer service agent writes to the customer. Based on the transcript of agents' conversations, provide a score that determines if the customer service agent is likely a native speaker of American English. For the native speaker assessment, look for traits such as: Correct grammar based on standard American English conventions, use of American English vocabulary, idioms, and phrasing rather than other dialects, natural, fluent-sounding language that does not appear stilted or translated, adherence to American cultural norms and reference points in word choice and examples, the writer can produce written material with the proficiency of a highly educated native speaker, demonstrating a superior command of the language. For each excerpt, provide a score from 1-5 for the native speaker assessment, where: 1 = Definitely not a native American English speaker 2 = Probably not a native American English speaker 3 = Uncertain if native American English speaker 4 = Probably a native American English speaker 5 = Definitely a native American English speaker. Only return your score of native speaker assessment. Do not include an explanation.

To validate the LLM's scores, we had two independent human reviewers evaluate 100 randomly selected agent conversations. The reviewers did not have access to each other's scores or the LLM results. The mean score given by the LLM was 4.29, while the average score from the human evaluators was 4.22. The difference between these scores was not statistically significant, with a p-value of 0.22.

The comprehensibility score captures the general fluency and ease of understanding in the responses of the customer service agents, ranging from very difficult to comprehend (1) to very fluent and easily understandable (5). The score assesses the clarity of communication, frequency and im-

pact of errors, and general language proficiency demonstrated in the writing, regardless of whether the writer is a native speaker. A Philippines-based agent may have a very high level of fluency, but may not speak like a native speaker. For instance, a common sign-off is to tell customers to "have a blessed day." While that phrase is not a common colloquial English phrase, it is grammatically correct and fully comprehensible. The full prompt is:

> You are given a conversation transcript from a customer service agent helping a customer that only includes what the customer service agent writes to the customer. Based on the transcript of agents' conversations, score the overall fluency of each excerpt from 1-5, regardless of whether it seems to be written by a native speaker, where: 1 = Very difficult to comprehend, multiple errors impeding comprehension 2 = Somewhat difficult to understand, some errors impeding comprehension 3 = Mostly understandable but with some errors 4 = Fluent and understandable with only minor errors that do not impact meaning 5 = Very fluent and easily understandable with no significant errors. Only return your score of fluency. Do not include an explanation.

We conducted a similar validation exercise with a random sample of 100 agent conversations, asking human evaluators to assess the comprehensibility of the agents' speech. The mean comprehensibility score assigned by the LLM was 4.31, while the average score given by human evaluators was 4.40. The difference between these scores was not statistically significant, with a p-value of 0.12.

J.2.6 Conversation Topic

Conversations between agents and customers are fairly complex interactions, often discussing details of tax filings, setting up sick leave, firing an employee, or correcting login issues. We use Gemini Pro to capture the topic of each conversation (Gemini Team, 2024).

First, we select a random sample of 5,000 conversations. Using the prompt below, we first ask the LLM to define the topic of the conversation in one to three words, following a procedure similar to **?**. Using this list of 5,000 conversation topics, we ask the LLM to group these topics into no more than 50 categories that describe the subject matter of the conversation. Using the LLM and reviewing the conversations, we come up with a list of 50 topic categories and a one or two sentence description associated with each category. The total number of categories was chosen based on conversations with business management.

We then use the LLM to cluster these topics into 50 distinct groups, each with a concise single-sentence definition, and validate these categories with contact center personnel. We use our LLM

to classify each of our conversations into a topic category, "other" or "unsure." Using this approach, we classify over 98% of conversations into one of 53 main topics. Appendix Figure A.XIV displays a breakdown of the most common chat topics encountered by agents. Conversation topics in our data are highly skewed, with the two most common topics—payroll and taxes and account access and management issues—making up 50% of all chat topics. The next five topics account for the next 25% of conversations. In total, the top 16 topics account for over 90% of chats in the data. This pattern is consistent across customer support; generally a small number of fairly common issues make up the bulk of customer inquiries. The full prompt is:

> The following is the text of a chat between a customer service agent and a customer who has reached out to the Intuit customer support team. The customer is usually a small business owner based in the US. Based on the transcript, categorize the topic of the conversation into a 1 to 3 word topic. If the topic is about a customer not being able to log into their account, categorize the chat as "Login." If the conversation is about paying their employees, the correct topic is "Payroll." If the conversation is about paying suppliers, the correct topic is "Supplier Payments". If the topic is about tax related issues, the correct topic is "Taxes." If the topic is about ensuring that two sets of records (usually the balances of two accounts) are in agreement, the right topic is "Reconciliation." When analyzing the transcript, do not focus on details like the identity of the customer, filler greeting text, or metadata around the agent joining the conversation or variables that obscure sensitive details like $SSN, $EMAIL, $NAME-1, $ADDRESS, hyperlink metadata like
, or case numbers. The output format should be: topic, problem type, issue resolution, probability resolved, skill required, covid-19 related.

Once we have a list of 50 categories and a clear definition, we then ask the LLM to classify the topic of each conversation into one of the 50 categories. The LLM could also classify a chat as "other" if the conversation does not fit any of the above categories, or "unsure" if the LLM is unable to classify the conversation into any specific category.

All together, we are able to classify 85% of all conversations into a topic category. The subjects follow a highly skewed distribution—almost 50% of contact center questions fall into payroll & taxes or account access & management.

We validate the LLM-generated topic categorizations by employing three independent human evaluators to classify a random sample of 100 conversations into our predefined topic categories. In 30% of the sample, all three human evaluators unanimously agree on the same topic ("unanimous

topic"). In 75% of the sample (which includes the unanimous subset), at least two evaluators agree on a topic ("modal topic"). The remaining 25% of the sample consists of discordant conversations where each evaluator selects a different topic. When comparing the LLM's topic selections to these human-evaluated subsets, we find that for unanimous-consensus conversations, the LLM selects the same topic as the human evaluators 87% of the time. For modal-consensus conversations, the LLM selects the modal topic 66% of the time. In the case of discordant conversations, where there is no agreement between human reviewers, the LLM's selected topic matches one of the three distinct human-selected topics 74% of the time.

We use these topic categories to rank topics by overall topic frequency across all conversations. We also categorize topics according to their overall frequency with respect to an individual agent.

J.2.7 Conversation Sentiment

Sentiment analysis involves determining the emotions or attitudes expressed in a piece of text. The valence of unstructured text data is widely used in consumer marketing, predicting stock market returns, measuring consumer sentiment, monitoring social media and understanding voting behavior.

Machine learning-based methods, particularly those utilizing transfer learning models, generally outperform other sentiment classification approaches. Hartmann et al. (2023) performs a meta-analysis of over 1,100 experimental results, demonstrating that transfer-learning models are superior for sentiment classification. They also provide an open-sourced, fine-tuned language model, SiEBERT, incorporating transfer-learning best practices, which we use for our sentiment analysis. The model generates a sentiment measure on a scale from -1 to 1, with -1 indicating negative sentiment and 1 indicating positive sentiment.

In our analysis, we classify sentiment separately for agents' speech and customers' text in each conversation. Customer sentiment reflects the emotional experience of the 40-minute or so chat-based interaction, while agent sentiment captures the tone of the agents' responses.

J.2.8 Conversation Similarity

We create textual embeddings of agent-customer conversations and compare similarity of these embeddings across workers and over time. Textual embeddings take a given body of text and transform it into a high-dimensional vector that represents its "coordinates" in linguistic space. Two pieces of text will have more similar coordinates if they share a common meaning or style. The specific embedding given to a body of text will depend on the embedding model used. We form our text embeddings using all-MiniLM-L6-v2, an LLM that is specifically intended to capture and cluster semantic information to assess similarity across text (Hugging Face, 2023). Once we create

an embedding for each conversation, we can compare the similarity of conversations by looking at the cosine similarities of their associated vectors; this common approach yields a score of 0 if two pieces of text are semantically orthogonal and a score of 1 if they have the same meaning (**?**). For context, the sentences "Can you help me with logging in?" and "Why is my login not working?" have a cosine similarity of 0.68 in our model.

J.3 Empirical Specifications

J.3.1 Pre-treatment Worker Skill Specification

$$y_{it} = \delta_t + \alpha_i + \sum_{q=1}^{5} \beta_q (AI_{it} \times Q_{iq}) + \gamma X_{it} + \epsilon_{it} \tag{1}$$

Our worker skill specification allows us to estimate how the impact of assistance varies by agent skill when they receive access to AI assistance. The regression is conducted at the agent-month level, where:

- y_{it} is the resolutions per hour of agent i in year-month t.
- δ_t represents year-month fixed effects.
- α_i denotes agent-level fixed effects.
- AI_{it} is an indicator equal to 1 if agent i has access to AI assistance at time t, 0 otherwise.
- Q_{iq} is an indicator function that equals 1 if agent i belonged to skill quintile q at the time of treatment, where q ranges from 1 (lowest skill) to 5 (highest skill).
- X_{it} is a set of time-varying controls, specifically fixed effects for agent tenure in months.
- β_q represents the average treatment effect of AI assistance for agents in skill quintile q

We estimate this equation separately for each of our outcome variables, which include our main measure of productivity for agent i in year-month t (resolutions per hour), as well as call resolution rate, customer satisfaction score, average call duration, calls handled per hour and requests to speak to the manager. We cluster standard errors at the agent level to account for within-agent correlations in the error terms. In our main analysis, we find very similar results across estimators and similar main effects across adoption cohorts, so we estimate this regression specification using OLS.

J.3.2 Pre-treatment Worker Tenure Specification

$$y_{it} = \delta_t + \alpha_i + \sum_{e=1}^{5} \beta_e (AI_{it} \times Exp_{ie}) + \epsilon_{it} \tag{2}$$

Our tenure specification allows us to estimate how the impact of assistance varies by agent experience when they receive access to AI assistance. The regression is conducted at the agent-month level, where:

- y_{it} is the resolutions per hour of agent i in year-month t.
- δ_t represents year-month fixed effects.
- α_i denotes agent-level fixed effects.
- AI_{it} is an indicator equal to 1 if agent i has access to AI assistance at time t, 0 otherwise.
- Exp_{ie} is an indicator that equals 1 if agent i has e months of experience at the time of treatment, and 0 otherwise. We divide months of experience into five categories: agents in their first month on the job ("0 months"), 1-2 months of experience, 3-6 months of experience, 7-12 months, and over 12 months of experience.
- X_i includes a set of fixed effects for agent i's skill quintile at the time of treatment. This is only available for treated agents and is time-invariant.
- β_e is the average treatment effect of AI assistance for agents with e months of experience when treated.

We estimate this equation separately for each of our outcome variables, which include our main measure of productivity for agent i in year-month t (resolutions per hour), as well as call resolution rate, customer satisfaction score, average call duration, and calls handled per hour. The skill quintile at AI treatment is time invariant and is not defined for control group workers. Standard errors are clustered at the agent level, and we estimate this regression using OLS.

J.3.3 Adherence to AI recommendations

$$y_{it} = \delta_t + \alpha_i + \sum_{a=1}^{5} \beta_a (AI_{it} \times Adh_{ia}) + \gamma X_{it} + \epsilon_{it} \tag{3}$$

This specification allows us to estimate how the impact of AI assistance varies by treated agents' adherence in their first month of AI access. The regression is conducted at the agent-month level, where:

- y_{it} is the resolutions per hour of agent i in year-month t.
- δ_t represents year-month fixed effects.
- α_i denotes agent-level fixed effects.
- AI_{it} is an indicator equal to 1 if agent i has access to AI assistance at time t, 0 otherwise.
- Adh_{ia} is an indicator equal to 1 if agent i is in the ath quintile of adherence in their first month of access, 0 otherwise.
- X_{it} includes time-varying controls, specifically fixed effects for agent tenure.
- β_a is the average impact of AI assistance for agents in the ath quintile of initial adherence.

Standard errors are clustered at the agent level. We estimate this regression with OLS.

J.3.4 Heterogeneity by Chat Topic

$$y_{itc} = \delta_t + \alpha_i + \sum_{r=1}^{4} \beta_r (AI_{it} \times Topic_{cr}) + \gamma X_{itc} + \epsilon_{itc} \tag{4}$$

This topic-based specification allows us to estimate how the impact of AI assistance varies by the routine nature of customers' problems. The regression is conducted at the chat level, where:

- y_{itc} is the duration of chat c assigned to agent i in year-month t.
- δ_t represents year-month fixed effects.
- α_i denotes agent-level fixed effects.
- AI_{it} is an indicator equal to 1 if agent i has access to AI assistance at time t, 0 otherwise.
- $Topic_{cr}$ is an indicator equal to 1 if chat c belongs to topic frequency category r (where r ranges from 1 to 4), 0 otherwise. The frequency of the topic is defined using frequency across all chats and agents.
- X_{it} includes fixed effects for agent tenure and the overall category of topic frequency ($Topic_{cr}$).
- β_r estimates the impact of AI assistance on chat duration for chats in the topic frequency category r.

Standard errors are clustered at the agent level and we estimate this regression with OLS. We drop the small number of topics that are classified as unsure or "other." We also estimate a separate specification on the agent-specific frequency of the technical support issue.

$$y_{itc} = \delta_t + \alpha_i + \sum_{r=1}^{4} \beta_r (AI_{it} \times AgentTopic_{icr}) + \gamma X_{cit} + \epsilon_{itc} \tag{5}$$

- $AgentTopic_{icr}$ is an indicator equal to 1 if chat c belongs to agent-specific topic frequency category r (where r ranges from 1 to 4), 0 otherwise. Topic frequency is defined relative to conversations conducted by agent i.
- X_{cit} includes time-varying controls, specifically fixed effects for agent tenure, aggregate topic category defined over all conversations ($Topic_{cr}$), and agent-specific topic frequency category ($AgentTopic_{icr}$).
- β_r estimates the impact of AI assistance on chat duration for chats in agent-specific topic frequency category r.

J.3.5 Attrition

$$attrit_{it} = \delta_t + \beta AI_{it} + \gamma X_{it} + \epsilon_{it} \tag{6}$$

This specification allows us to look at how AI assistance impacts agent attrition. The regression is conducted at the agent-month level, where:

- $attrit_{it}$ is equal to 1 if agent i leaves in year-month t.
- δ_t represents year-month fixed effects.
- AI_{it} is an indicator equal to 1 if agent i has access to AI assistance at time t, 0 otherwise.
- X_{it} includes time-varying controls, specifically fixed effects for agent tenure, agent location, country and company employing the agent (data firm or subcontractor).
- β_a is the average impact of AI assistance on attrition.

Attrition captures both voluntary or involuntary separations, which are undistinguishable in our data. Standard errors are clustered at the agent level and we estimate this regression with OLS. We cannot include for agent-fixed effects, because agents only leave once, so agent-fixed effects are colinear with attrition. We drop all observations for treated agents before treatment because, by construction, agents must survive through treatment to receive AI assistance. We also estimate specifications where we estimate the attrition effects by agent tenure and skill at AI deployment.