NBER WORKING PAPER SERIES

GENERATIVE AI AT WORK

Erik Brynjolfsson
Danielle Li
Lindsey R. Raymond

Working Paper 31161
http://www.nber.org/papers/w31161

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2023

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 and to the Stanford Digital Economy Lab for funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of Stanford University, MIT, or the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w31161

Generative AI at Work
Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond
NBER Working Paper No. 31161
April 2023
JEL No. D8,J24,M15,M51,O33

ABSTRACT

We study the staggered introduction of a generative AI-based conversational assistant using data from 5,179 customer support agents. Access to the tool increases productivity, as measured by issues resolved per hour, by 14 percent on average, with the greatest impact on novice and low-skilled workers, and minimal impact on experienced and highly skilled workers. We provide suggestive evidence that the AI model disseminates the potentially tacit knowledge of more able workers and helps newer workers move down the experience curve. In addition, we show that AI assistance improves customer sentiment, reduces requests for managerial intervention, and improves employee retention.

Erik Brynjolfsson
Stanford Digital Economy Laboratory
353 Jane Stanford Way, Office 136
Stanford, CA 94305
and NBER
erik.brynjolfsson@gmail.com

Lindsey R. Raymond
MIT Sloan School of Management
100 Main Street
E62-489
Cambridge, MA 02142
lindsey.r.raymond@gmail.com

Danielle Li
MIT Sloan School of Management
100 Main St, E62-484
Cambridge, MA 02142
and NBER
d_li@mit.edu

The emergence of generative artificial intelligence (AI) has attracted significant attention, but there have been few studies of its economic impact. While various generative AI tools have performed well in laboratory settings, excitement about their potential has been tempered by concerns that these tools are prone to "hallucination," and may produce coherent sounding but inaccurate information (Peng et al., 2023a; Roose, 2023).

In this paper, we study the adoption of a generative AI tool that provides conversational guidance for customer support agents.[1] This is, to our knowledge, the first study of the impact of generative AI when deployed at scale in the workplace. We find that access to AI assistance increases the productivity of agents by 14 percent, as measured by the number of customer issues they are able to resolve per hour. In contrast to studies of prior waves of computerization, we find that these gains accrue disproportionately to less-experienced and lower-skill workers.[2] We argue that this occurs because ML systems work by capturing and disseminating the patterns of behavior that characterize 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: when engineers write code to perform a task, they are codifying that task. Yet because many production processes rely on tacit knowledge, these processes have so far defied automation (Polanyi, 1966; Autor, 2014).

Machine learning algorithms work differently. In contrast to traditional programming, these systems implicitly *infer* instructions from examples. Given a training set of images, for instance, ML systems can learn to recognize faces. This highlights a key, distinguishing aspect of ML systems: they can learn to perform tasks even when no instructions exist–including tasks requiring tacit knowledge that could previously only be gained through lived experience (Polanyi, 1966; Autor, 2014; Brynjolfsson and Mitchell, 2017).[3]

We study the impact of generative AI on productivity in the customer service sector, an industry with one of the highest rates of AI adoption (Chui et al., 2021). We examine the staggered

[1] A note on terminology. There are many definitions of artificial intelligence and of intelligence itself— Legg et al. (2007) list over 70 of them. In this paper, we define "artificial intelligence" (AI) as an umbrella term that refers to a computer system that is able to sense, reason, or act like a human. "Machine learning" (ML) is a branch of AI that uses algorithms to learn from data, identify patterns and make predictions or decisions without being explicitly programmed (Google, n.d.). Large language models (LLMs) and tools built around LLMs such as ChatGPT are an increasingly important application of machine learning. LLMs generate new content, making them a form of "generative AI."

[2] We provide a discussion of this literature at the end of this section.

[3] As Meijer (2018) puts it "where the Software 1.0 Engineer formally specifies their problem, carefully designs algorithms, composes systems out of subsystems or decomposes complex systems into smaller components, the Software 2.0 Engineer amasses training data and simply feeds it into an ML algorithm..."

deployment of a chat assistant using data from 5,000 agents working for a Fortune 500 software firm that provides business process software. The tool we study is built on a recent version of the Generative Pre-trained Transformer (GPT) family of large language models developed by OpenAI (OpenAI, 2023). It monitors customer chats and provides agents with real-time suggestions for how to respond. It is designed to augment agents, who remain responsible for the conversation and are free to ignore its suggestions.

We have four sets of findings.

First, AI assistance increases worker productivity, resulting in a 13.8 percent increase in the number of chats that an agent is able to successfully resolve per hour. This increase reflects shifts in three components of productivity: a decline in the time it takes to an agent to handle an individual chat, an increase in the number of chats that an agent is able to handle per hour (agents may handle multiple calls at once), and a small increase in the share of chats that are successfully resolved.

Second, AI assistance disproportionately increases the performance less skilled and less experienced workers across all productivity measures we consider. In addition, we find that the AI tool helps newer agents move more quickly down the experience curve: treated agents with two months of tenure perform just as well as untreated agents with over six months of tenure. These results contrast, in spirit, with studies that find evidence of skill-biased technical change for earlier waves of computer technology (Autor et al., 2003; Acemoglu and Restrepo, 2018; Bresnahan et al., 2002; Bartel et al., 2007).

Our third set of results investigates the mechanism underlying our findings so far. We posit that high-skill workers may have less to gain from AI assistance precisely because AI recommendations capture the potentially tacit knowledge embodied in their own behaviors. Rather, low-skill workers are more likely to improve by incorporating these behaviors by adhering to AI suggestions. Consistent with this, we find few positive effects of AI access for the highest-skilled or most-experienced workers. Instead, using textual analysis, we find suggestive evidence that AI assistance leads lower-skill agents to communicate more like high-skill agents.

Finally, we show that the introduction of AI systems can impact the experience and organization of work. We show that AI assistance markedly improves how customers treat agents, as measured by the sentiments of their chat messages. This change may be associated with other organizational changes: turnover decreases, particularly for newer workers, and customers are less likely to escalate a call by asking to speak to an agent's supervisor.

Our overall findings demonstrate that generative AI working alongside humans can have a significant positive impact on the productivity and retention of individual workers. We emphasize,

however, that our paper is not designed to shed light on the aggregate employment or wage effects of generative AI tools.

Our paper is related to a large literature on the impact of various forms of technological adoption on worker productivity and the organization of work (e.g. Rosen, 1981; Autor et al., 1998; Athey and Stern, 2002; Bresnahan et al., 2002; Bartel et al., 2007; Acemoglu et al., 2007; Hoffman et al., 2017; Bloom et al., 2014; Michaels et al., 2014; Garicano and Rossi-Hansberg, 2015; Acemoglu and Restrepo, 2020). Many of these studies, particularly those focused on information technologies, find evidence that IT complements higher-skill workers (Akerman et al., 2015; Taniguchi and Yamada, 2022). Bartel et al. (2007) shows that firms that adopt IT tend to use more skilled labor and increase skill requirements for their workers. Acemoglu and Restrepo (2020) study the diffusion of robots and find that the negative employment effects of robots are most pronounced for workers in blue-collar occupations and those with less than a college education. In contrast, we study a different type of technology—generative AI—and find evidence that it most effectively augments lower-skill workers.

For example, Peng et al. (2023b) recruit software engineers for a specific coding task (writing an HTTP server in JavaScript) and show that those given access to GitHub Copilot complete this task twice as quickly. Similarly, Noy and Zhang (2023) run an online experiment showing that subjects given access to ChatGPT complete professional writing tasks more quickly. In addition, they show that ChatGPT compresses the productivity distribution, with lower-skill workers benefiting the most.

To the best of our knowledge, however, there have been no studies of the impact of access to generative AI tools on productivity in real-world workplaces, nor over longer periods. Such studies are important because the impact of AI on productivity may vary over time and interact with workers' baseline level of experience or expertise. Technologies that look promising in laboratory settings may have more limited effects in practice because of the need for complementary organizational investments, skill development and business process redesign. In addition, the introduction of AI systems may have further impacts on worker and customer satisfaction, attrition, and patterns of organizational behavior.

1 Generative AI and Large Language Models

In recent years, 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). These technologies are all examples of "generative AI," a class of machine

learning technologies that can generate new content—such as text, images, music, or video—by analyzing patterns in existing data. In this section, we provide background on generative AI as a technology and discuss its potential economic implications.

1.1 Technical Primer

In this paper, we focus on an important class of generative AI, large language models (LLMs). At a high level, LLMs are neural network models designed to process sequential data (Bubeck et al., 2023). For instance, an LLM can be trained by giving it access to a large corpus of text (such as Wikipedia, digitized books, or portions of the Internet) and using that input text to learn to predict the next word in a sequence, given what has come before. This knowledge of the statistical co-occurrence of words allows it to generate new text that is grammatically correct and semantically meaningful.

Though the name implies human language, the same techniques can be used to produce LLMs that generate other forms of sequential data such as protein sequences, audio, computer code, or chess moves (Eloundou et al., 2023).

Recent progress in generative AI has been driven by four factors: computing power, earlier innovations in model architecture, the ability to "pre-train" using large amounts of unlabeled data, and refinements in training techniques.[4] Model performance depends strongly on scale, which includes the amount of computing power used for training, the number of model parameters, and dataset size (Kaplan et al., 2020). Pre-training an LLM requires thousands of GPUs and weeks to months of dedicated training time. For example, estimates indicate that a single training run for a GPT-3 model with 175 billion parameters, trained on 300 billion tokens, may cost $5 million dollars in just computing costs (Li, 2020; Brown et al., 2020).

In terms of model architecture, modern LLMs make use of two earlier key innovations: positional encoding and self-attention. Positional encodings keep track of the order in which a word occurs in a given input.[5] This allows large bodies of input text to be broken into smaller segments that can be processed simultaneously without "forgetting" earlier parts of the input (Vaswani et al., 2017; Bahdanau et al., 2015). Meanwhile, self-attention assigns importance weights to each word in the context of the entire input text. Older approaches that assign importance based on word frequencies that may misrepresent a word's true semantic importance. These older methods may also base on semantic content within a small window. In contrast, self-attention enables models to

[4]For a more detailed technical review of progress, see Radford and Narasimhan (2018); Radford et al. (2019); Liu et al. (2023); Ouyang et al. (2022).

[5]For instance, a model would keep track of "the, 1" instead of only "the" (if "the" was the first word in the sentence).

capture long-range semantic relationships within an input text, even when that text is broken up and processed in parallel (Vaswani et al., 2017).

Second, LLMs can be pre-trained on large amounts of unlabeled data. For instance, GPT is trained on unlabeled text data, allowing it to learn patterns in human language without explicit guidance (Radford and Narasimhan, 2018). Because unlabeled data is far more prevalent than labeled data, this allows for LLMs to learn about natural language on a much larger training corpus (Brown et al., 2020). The resulting model can be used in multiple applications because its training is not specific to a particular set of tasks.[6]

Finally, general-purpose LLMs can be further "fine-tuned" to generate output that matches the priorities of any specific setting (Ouyang et al., 2022; Liu et al., 2023). For example, an LLM may generate several potential responses to a given query, but some of them may be factually incorrect or biased. To discipline this model, human evaluators can rank these outputs to train a reward function prioritizes some responses over others. Such refinements can significantly improve model quality but making a general-purpose model better suited to its specific application (Ouyang et al., 2022).

Together, these innovations have generated meaningful improvements in model performance. The Generative Pre-trained Transformer (GPT) family of models, in particular, has attracted considerable media attention for their rapidly expanding capabilities.[7]

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 standardized and reduced to explicit rules (Autor, 2014). Consequently, computerization has had a disproportionate impact on jobs that involve routine tasks, such as data entry, bookkeeping, and assembly line work, and reducing demand for workers performing "routine" tasks (Acemoglu and Autor, 2011). At the same time, computerization has also increased the productivity of workers who possess complementary skills, such as programming, data analysis, and research. Together, these changes have contributed to increasing wage inequality in the United States and have been linked to a variety of organizational changes (Katz and Murphy, 1992; Autor et al., 2003; Michaels et al., 2014; Bresnahan et al., 2002; Baker and Hubbard, 2003).

In contrast, recent advances in AI, particularly those driven by generative AI, suggest that it is possible for LLMs to perform a variety of non-routine tasks such as software coding, persuasive

[6]For example, a model trained to generate tweets based on the history of Twitter will differ depending on whether its inputs are labeled with each tweet's number of retweets or an assessment of its truthfulness.

[7]For instance, GPT-4 has recently been shown to outperform humans in taking the US legal bar exam (Liu et al., 2023; Bubeck et al., 2023; OpenAI, 2023).

writing, and graphic design (Bommasani et al., 2021; Eloundou et al., 2023). For example, Copilot, an AI pair programmer that generates code suggestions for programmers, has achieved impressive performance on technical coding questions and generates an average of 46 percent of code among users (Nguyen and Nadi, 2022; Zhao, 2023). Similarly, services like Elicit and Casetext use LLMs to find and summarize key information in legal documents or research, tasks that were previously considered non-routine (Elicit, 2023; Casetext, 2023). Since many of these tasks are currently performed by workers who have either been insulated or benefited from prior waves of technology adoption, the expansion of generative AI has the potential to shift the relationship between technology, labor productivity, and inequality (The White House, 2022).

Unlike traditional programming, generative AI does not require explicit instructions as inputs. Instead, it uses ML to mine vast amounts of human-generated data to recognize patterns, allowing it to generate, summarize, and make inferences based on those patterns. For instance, if prompted to provide a cover for a gothic novel, generative AI models will respond with an illustration that is moody; if asked to write an email denying an employee a raise, generative AI will respond with a note that is professional; this occurs even though no programmer has instructed the AI model as to what tone would be appropriate in a given context. The ability to behave "appropriately" cannot be reduced to a set of rules; instead people learn to do so from experience and apply unconscious rules in the process. The fact that generative AI models display such skills suggests that they have the potential to move past what has been termed "Polanyi's Paradox," the idea that knowledge is difficult to codify because individuals perform many tasks they cannot articulate (Polanyi, 1966; Autor, 2014).

At the same time, LLMs have significant limitations. Popular LLM-based tools such as ChatGPT have been shown to produce false or misleading information in unpredictable ways. While these models often perform well on specific tasks in the lab, these concerns have raised questions about their ability to perform well in more complex real-world settings (Peng et al., 2023a).

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 area with one the highest surveyed rates of AI adoption.[8] Customer support interactions are important for maintaining a

[8]For instance, of the businesses that report using AI, 22 percent use AI in their customer service centers (Chui et al., 2021).

company's reputation and building strong customer relationships, yet as in many industries, there is substantial variation in worker productivity (Berg et al., 2018; Syverson, 2011).

Newer workers are also often less productive and require significant training. At the same time, turnover is high: industry estimates suggest that 60 percent of agents in contact centers[9] leave each year, costing firms $10,000 to $20,000 dollars per agent (Buesing et al., 2020; Gretz and Jacobson, 2018). To address these workforce challenges, the average supervisor spends at least 20 hours per week coaching lower-performing agents (Berg et al., 2018). Faced with variable productivity, high turnover, and high training costs, firms are increasingly turning toward AI tools (Chui et al., 2021).

At a technical level, customer support is well-suited for current generative AI tools. From an AI's perspective, customer-agent conversations can be thought of as a series of pattern-matching problems in which one is looking for an optimal sequence of actions. When confronted with an issue such as "I can't login," an AI/agent must identify which types of underlying problems are most likely to lead a customer to be unable to log in and think about which solutions typically resolve these problems ("Can you check that caps lock is not on?"). At the same time, they must be attuned to a customer's emotional response, making sure to use language that increases the likelihood that a customer will respond positively ("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 for customer service using many examples of both successfully and unsuccessfully resolved conversations.

In the remainder of this section, we provide details about the firm we study and the AI tool they adopt.

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 their tool at one of their client firms, (hereafter, the "data firm").

Our data firm is a Fortune 500 enterprise software 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, both directly and through third-party firms. The majority of agents in our sample work from offices located in the Philippines, with a smaller group working in the United States and in other countries. Across locations, agents are engaged in a fairly uniform job: answering technical support questions from US-based small business owners.

[9]The term "contact center" updates the term "call center," to reflect the fact that a growing proportion of customer service contacts no longer involve phone calls.

Chats are randomly assigned and support sessions are relatively lengthy, averaging 40 minutes with much of the conversation spent trying to diagnose the underlying technical problem. The job requires a combination of detailed product knowledge, problem-solving skills, and the ability to deal with frustrated customers.

Our firm measures productivity using three metrics that are standard in the customer service industry: "average handle time," the average length of time an agent takes to finish a chat; "resolution rate," the share of conversations that the agent can successfully resolve; and "net promoter score," (customer satisfaction), 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. A productive agent is one who is able to field customer chats quickly while maintaining a high-resolution rate and net promoter score.

Across locations, agents are organized into teams with a manager that provides feedback and training to agents. Once per 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 are paid an hourly wage and bonuses based on their performance relative to other agents.

2.3 AI System Design

The AI system we study uses a generative model system that combines a recent version of GPT with additional ML algorithms specifically fine-tuned to focus on customer service interactions. The system is further trained on a large set of customer-agent conversations that have been labeled with a variety of outcomes and characteristics: whether the call was successfully resolved, how long it took to handle the call, and whether the agent in charge of the call is considered a "top" performer by the data firm. The AI firm then uses these data to look for conversational patterns that are most predictive of call resolution and handle time.

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, surface appropriate technical documentation, and limit unprofessional language. This additional training mitigates some of concerns associated with relying on LLMs to generate text.

Once deployed, the AI system generates two main types of outputs: 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. In both cases, recommendations are based on a history of the conversation. For example, the correct response when a customer says "I can't track my employee's hours during

business trips" depends on what version of the data firm's software the customer uses. Suppose the customer has previously mentioned that they are using the premium version. In that case, they should have access to remote mobile device timekeeping, meaning that the support agents need to diagnose and resolve a technical issue preventing the software from working. If, however, the customer stated that they are using the standard version, then the correct solution is for the customer to upgrade to the premium version in order to access this feature.[10]

Figure 1 illustrates sample output. In the chat window (Panel A), Alex, the customer, describes their problem to the agent. Here, the AI assistant generates two suggested responses (Panel B). In this example, it 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 A of Appendix Figure A.1 shows an example of a technical recommendation from the AI system, which occurs when it recommends 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 whether to incorporate (fully or partially) the AI suggestions. This reduces the potential for output that is off-topic or incorrect to make its way into customer conversations. Furthermore, the system does not provide suggestions when it has insufficient training data for that situation (this occurs in a large minority of cases). In these situations, the agent must respond on their own.

Finally, we observe that the AI model is trained on human-generated data in a setting where there is high variability in the abilities of individual agents. As a result, when the model identifies patterns that distinguish successful from unsuccessful calls, it is implicitly learning the differences that characterize high- versus low-skill workers. For example, top-performing agents are often more effective at diagnosing the underlying technical issue given a customer's problem description: in an internal study, our AI firm found that top performers start researching a solution twice as quickly as the average workers. An AI system, with access to many examples of diagnostic questions and eventual solutions, may be able to encode some of the "best practices" top-performing agents use.

This suggests that an AI system may be able to more effectively share knowledge across workers both because it may capture tacit knowledge that was previously difficult for managers to articulate and because it can provide more real-time recommendations than a busy manager.[11] Indeed, AI recommendations can be thought of as expanding the marginal productivity of high-skill workers

[10] For more on context tracking, see, for instance, Dunn et al. (2021).

[11] By necessity, managers can only base their feedback on a small subset of the hundreds of conversations an agent conducts. And because managers are often pressed for time and may lack training, they may focus on a single metric ("you need to solve problems faster") 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.") This type of coaching is ineffective and often counterproductive for employee engagement (Berg et al., 2018).

by encoding their conversational patterns and disseminating them to other workers. In our setting, high-skill workers are not compensated for these contributions.

3 Deployment, Data, and Empirical Strategy

3.1 AI Model Deployment

The AI assistant we study was gradually rolled out at the agent level after an initial seven-week randomized pilot featuring 50 agents.[12] The deployment was largely uniform across both the data firm's own customer service agents, as well as its outsourced agents. Figure 2 documents the progression of deployment among agents who are eventually treated. The bulk of the adoption occurs between November 2020 and February 2021.

3.2 Summary Statistics

Table 1 provides details on sample characteristics, divided into three groups: agents who are never given access to the AI tool during our sample period ("never treated"), pre-AI observations for those who are eventually given access ("treated, pre"), and post-AI observations ("treated, post"). In total, we observe the conversation text and outcomes associated with 3 million chats by 5,179 agents. Within this, we observe 1.2 million chats by 1,636 agents in the post-AI period. Most agents in our sample, 83 percent, are located outside the United States, primarily in the Philippines. For each agent, we observe their assigned manager, tenure, geographic location and firm information.

To examine the impacts of this deployment, we construct several key variables, all aggregated to the agent-month level, which is our primary level of analysis.

Our primary measure of productivity is resolutions per hour (RPH), the number of chats that a worker is able to successfully resolve per hour. We consider this measure to be the most effective summary of a worker's productivity at the firm. An agent's RPH is determined by several factors: the average time it takes an agent to complete a conversation, the number of conversations they are able to handle per hour (accounting for multiple simultaneous conversations), and the share of conversations that are successfully resolved. We measure these individually as, respectively, average handle time (AHT), chats per hour (CPH), and resolution rate (RR). In addition, we also observe a measure of customer satisfaction through an agent's net promoter score (NPS), which is collected by the firm from post-call customer surveys.

[12] Data from the RCT is included as part of our primary analysis but is not analyzed separately because of its small sample size.

We observe these measures for different numbers of agents. In particular, we are able to reconstruct measures of average handle time and chats per hour from our chat level data. Therefore we observe AHT and CPH measures for all agents in our sample. Measures that involve an understanding of call quality—resolution rates, and customer satisfaction—are provided at the agent month level by our data firm. Because our data outsources most of its customer service functions, it does not have direct control over this information, which is kept by subcontracted firms. As a result, we observe resolution rates and net promoter scores for a subset of agents in our data. This, in turn, means that we only observe our omnibus productivity measure—resolutions per hour—for this smaller subset.

Figure 3 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 have higher outcomes than both never-treated agents and pre-treatment agents. In Panel E, we see no discernible differences in surveyed customer satisfaction among pre- and post-AI groups.

Focusing on our main productivity measure, Panel A of Figure 3 and Table 1 show that 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 differences in the initial section: treated agents have higher resolutions per hour prior to AI model deployment (2.0 chats) relative to never treated agents (1.7). This same pattern appears for other outcomes: chats per hour (Panel C) and resolution rates (Panel D). Panel B illustrates the clearest pattern with average handle times: both pre-treatment and never-treated agents had a similar distribution of average handle times, centered at 40 minutes but post-treatment agents have a lower average handle time of 35 minutes.

3.3 Empirical Strategy

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

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

Our outcome variables y_{it} capture average handle times, resolution rates, resolutions per hour, and customer satisfaction scores for agent i in year-month t. Because workers often work only for a portion of the year, we include only year-month observations for an agent who is actively employed (e.g. assigned to chats). Our main variable of interest is AI_{it}, an indicator equal to one if agent i has access to AI recommendations at time t. All regressions include year-month fixed effects δ_t

to control for common, time-varying factors such as tax season or business quarter end. In our preferred specification, we also include controls for time-invariant agent-level fixed effects α_i and time-varying agent tenure. Standard errors are clustered at the agent or agent-location level.

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 et al., 2022). For instance, 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 onboarding of later cohorts of agents may be smoother, so that their treatment effects may be larger.

We study the dynamics of treatment effects using the interaction weighted (IW) estimator proposed in Sun and Abraham (2021). Sun and Abraham (2021) show that this estimator is consistent assuming parallel trends, no anticipatory behavior, and cohort-specific treatment effects that follow the same dynamic profile.[13] In the Appendix, we show that both our main differences-in-differences and event study estimates are similar using robust estimators introduced in de Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2022), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), as well as using traditional two-way fixed effects OLS.

4 Results

4.1 Productivity

Table 2 examines the impact of AI model deployment on our primary measure of productivity, resolutions per hour, using a standard two-way fixed effects model. In Column 1, we show that, controlling for time and location fixed effects, access to AI recommendations increases the number of resolutions per hour by 0.47 chats, up 22.2 percent from an average of 2.12. In Column 2, we include individual agent fixed effects to account for potential differences between treated and untreated agents. In Column 3, we include further controls for time-varying agent tenure. As we add controls, our effects fall slightly so that, with agent and tenure fixed effects, we find that the deployment of AI increases RPH by 0.30 calls or 13.8 percent. Columns 4 though 6 produce these same patterns and magnitudes for the log of RPH.

Appendix Table A.1 finds similar results using alternative difference-in-difference estimators introduced in Callaway and Sant'Anna (2021), Borusyak et al. (2022), de Chaisemartin and D'Haultfœuille

[13] This last assumption means that treatment effects are allowed to vary over event-time and that average treatment effects can vary across adoption-cohorts (because even if they follow the same event-time profile, we observe different cohorts for different periods of event-time).

(2020), and Sun and Abraham (2021). Unlike traditional OLS, these estimators avoid comparing between newly treated and already treated units. In most cases, we find slightly larger effects of AI assistance using these alternatives.

Figure 4 shows the accompanying IW event study estimates of Sun and Abraham (2021) for the impact of AI assistance on resolutions per hour, in levels and logs. For both outcomes, 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. Appendix Figure A.2 shows that this pattern can be seen using alternative event study estimators as well: Callaway and Sant'Anna (2021), Borusyak et al. (2022), de Chaisemartin and D'Haultfœuille (2020), and traditional two-way fixed effects.

In Table 3, we report additional results using our preferred specification with year-month, agent, and agent tenure fixed effects. Column 1 documents a 3.8 minute decrease in the average duration of customer chats, a 9 percent decline from the baseline mean (shorter handle times are generally 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 2.6, this represents a roughly 14 percent increase. Unlike average handle time, chats per hour accounts for the possibility that agents may handle multiple chats simultaneously. The fact that we find a stronger effect on this outcome suggests that AI enables agents to both speed up chats and to multitask more effectively.

Column 3 of Table 3 indicates a small 1.3 percentage point increase in chat resolution rates, significant at the 10 percent level. This effect is economically modest, 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 -0.13 percentage points and the mean is 79.6 percent. Columns 5 through 8 report these results for logged outcomes. Going forward we will report our estimates in logs, for ease of interpretation.

Figure 5 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 therefore interpret these findings as saying that, on average, AI assistance increases productivity without negatively impacting resolution rates and surveyed customer satisfaction.

4.2 Impacts by Agent Skill and Tenure

As discussed earlier, generative AI tools may have a different pattern of productivity consequences relative to earlier waves of technology adoption.

In Panel A of Figure 6, we consider how our estimated productivity effects differ by an agent's pre-AI productivity. We divide agents into quintiles using a skill index based on their average call efficiency, resolution rate, and surveyed customer satisfaction in the quarter prior to the adoption of the AI system. These skill quintiles are defined within a firm-month. In Panel A, we show that the productivity impact of AI assistance is most pronounced for workers in the lowest skill quintile (leftmost side), who see a 35 percent increase in resolutions per hour. In contrast, AI assistance does not lead to any productivity increase for the most skilled workers (rightmost side).

In Figure 7 we show that less-skilled agents consistently see the largest gains across our other outcomes. For the highest-skilled workers, we find mixed results: a zero effect on average handle time (Panel A), a positive effect for calls per hour (Panel B), and, interestingly, small but statistically significant *decreases* in resolution rates and customer satisfaction (Panels C and D). These findings suggest that while lower-skill workers improve from having access to AI recommendations, they may distract the highest-skilled workers, who are already doing their jobs effectively.

Next, in Panel B of Figure 6, repeat this analysis for tenure by dividing agents into five groups based on their tenure at the time that the AI model is introduced. Some agents have less than one month of tenure when they receive AI access, while others have over a year of experience. We see a clear, monotonic pattern in which the least experienced agents see the greatest gains in resolutions per hour.

In Figure 8, we show that these patterns persist for other outcomes: AI assistance generates larger gains in call handling and quality for the newest workers. For more experienced workers, we find positive effects for average handle time and calls per hour (Panels A and B), zero effects on resolution rate (Panel C), and a small but statistically significant negative effect for customer satisfaction (Panel D).

In Appendix Figures A.3 and A.4, we show that our skill-heterogeneity results are robust to controlling for agent tenure, and vice versa. This suggests the AI system has distinct impacts both by worker experience and ability.

4.3 Moving Down the Experience Curve

To further explore how AI assistance impacts newer workers, we examine how worker productivity evolves on the job.[14] In Figure 9, 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

[14]We avoid the term "learning curve" because we cannot distinguish if workers are learning or merely following recommendations.

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 with around 2.0 resolutions per hour. Workers who are never treated slowly improve their productivity with experience, reaching approximately 2.5 resolutions 8 to 10 months later. In contrast, workers who begin with access to AI assistance rapidly increase their productivity to 2.5 resolutions only two months in. Furthermore, they continue to improve at a rapid rate until they are resolving more than 3 calls an hour after five months of tenure.[15] Comparing just these two groups suggests that access to AI recommendations helps workers move more quickly down the experience curve.

The final line in Panel A tracks workers who begin their tenure with the firm without access to AI assistance, but who receive access after five months on the job. These workers improve slowly in the same way as never-treated workers for the first five months of their tenure. Starting in month five, however, these workers gain access and we see their productivity rapidly increase following the same trajectory as the always-treated agents. In Appendix Figure A.5, 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.

Taken together, these results indicate that access to AI helps new agents move more quickly down the experience curve. Across many of the outcomes in Figure 9, 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.

4.4 Adherence to AI recommendations

We emphasize that the AI tool we study is meant to augment— rather than replace—human agents. The system makes suggestions, but agents may elect to ignore these suggestions entirely. In our results above, we estimate intent-to-treat effects, that is how access to the AI tool impacts outcomes regardless of how frequently agents follow its recommendations. In this section, we examine how closely agents adhere to AI recommendations, and document the association between adherence and returns to adoption.

We measure "adherence" starting at the chat level, by calculating the share of AI recommendations that each agent follows. Agents are coded as having adhered to a recommendation if they either click to copy the suggested AI text or if they self-input something very similar. We then aggregate this to the agent-month level.

[15] Our sample ends here because we have very few observations more than five months after treatment.

Panel A of Figure 10 plots the distribution of average agent-month-level adherence for our post-AI sample, weighted by the log of the number of AI recommendations given 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 to 46 percent of code recommendations (Zhao, 2023). Such behavior could be optimal: the AI model may make incorrect or non-sensical suggestions.

Panel B of Figure 10 shows that *returns* to AI model deployment tend to be higher among agents who follow a greater share of recommendations. We divide agents into quintiles based on the percent of AI recommendations they follow in the first month of deployment and separately estimate the impact of access to the AI model for each group. These estimates control for year-month and agent fixed effects as in Column 5 of Table 2.

For agents in the lowest quintile of adherence, we still see a 10 percent gain in productivity, but for agents in the highest quintile, the estimated impact is much higher, close to 25 percent. Appendix Figure A.6 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 calls per hour (Panel B), and more noisily for resolution rate (Panel C) and customer satisfaction (Panel D).

We note that this relationship could be driven by a variety of factors: the treatment effect of adherence (agents have greater productivity because they listen to recommendations); 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).

Finally, Figure 11 plots how adherence to AI recommendations evolves for workers of different experience or skill. In Panel A, we see that more senior workers are initially less likely to follow AI recommendations: 30 percent for those with over a year of tenure compared to 37 percent for those with under three months of tenure.[16] Over time, however, all workers increase their adherence, with more senior workers doing so faster so that the groups converge five months after tenure. In Panel B, we see a similar but more muted pattern for worker skill: lower-skill workers are initially more likely to comply, but the highest-skilled workers converge over time.

These findings are consistent with both the possibility that workers who are initially more skeptical may come to see the value of AI recommendations over time or that workers who strongly dislike working with the AI system may exit at higher rates. Other studies on the use of AI tools have found differences in the desire to follow AI recommendations; for instance, in a study of a

[16] Agents below three months of experience are in their "onboarding" phase.

writing suggestion tool, Singh et al. (2022) finds that four of the 23 study participants refused to engage with AI suggestions.

4.5 Textual Evidence

Our evidence so far suggests that access to AI suggestions improves productivity and, for the lower-skilled and less-experienced agents, increases conversation quality as measured by resolution rates and surveyed customer satisfaction.

Next, we consider why AI may have these impacts. This section provides evidence from *preliminary* analysis of the textual content of chat conversations. Our goal is to understand whether and how AI recommendations change the way agents communicate.

In particular, we are interested in AI model's ability to encode the potentially tacit knowledge of high performers: does AI assistance lead lower-skill workers to communicate more like higher-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 rather than looking for the presence of specific formulaic phrases.

We begin by constructing textual embeddings of agent-customer conversations. These embeddings allow us to represent their semantic and stylistic content as a vector so that we can compare one conversation to another using cosine similarity. Cosine similarity is a widely-used metric for measuring the similarity of two embeddings. It calculates the cosine of the angle between two n-dimensional vectors, where values close to 1 indicate similarity (Koroteev, 2021). We form our text embeddings using all-MiniLM-L6-v2, an LLM that is specifically intended to capture and cluster semantic information in order to assess similarity (Hugging Face, 2023). We then compare how an agent's conversations change over time (e.g. within-person similarity over time), as well as how high- and low-skill agents' conversations compare with each other over time (between-person similarity at a given point in time).

4.5.1 Within-worker changes in communication

To examine how AI changes agent conversations with customers, we begin by comparing an individual agent's text pre- and post-AI model deployment. We take a given agent's corpus of text from an eight-week window before deployment and compare it with text from the same-sized window afterward. We exclude the three weeks around deployment to account for disruptions due to training and individual adjustment to the AI model. We exclude messages from the customer, and focus only on agent-generated language. Then, for each individual agent, we plot textual *dissimilarity* between pre- and post-AI conversations as one minus the cosine similarity. The within-person textual

difference in our data is roughly 0.3, indicating that, on average, agents continue to communicate similarly after the deployment of the AI model. For context, the sentences "Can you help me with this logging in?" and "Why is my login not working?" have a cosine similarity of 0.68.

We examine how language changes pre- and post-AI for workers who are initially high- and low-skill, where high-skill workers are defined as those in the top quintile of the skill index distribution, while low-skill workers are those in the bottom quintile.

If the AI is able to embody and disseminate some of the tacit "best practices" of high-skill workers, then we would expect low-skill workers to experience a greater shift in communication patterns following access to AI assistance (high-skill workers would change less since the AI model is suggesting practices they have already adopted). In Panel A of Figure 12, we find evidence consistent with this hypothesis: initially lower-skill agents shift their language more after AI model deployment, relative to initially higher-skill workers. The scores we plot include controls for conversation timing to account for seasonal changes in topics such as tax season or payroll cycles. These results also control for agent tenure to account for the possibility that younger workers' language may evolve more quickly independent of access to the AI model.

If AI assistance merely leads workers to type the same things but faster, then we would not expect this differential change. And because chats are randomly allocated to agents, we would not also expect the pattern we document to be driven by differential changes in conversation topics between high and low-skill workers.

4.5.2 Across worker comparisons

We next explore *how* lower-skill agents' language choices change with AI assistance. In Panel B of Figure 12, we provide suggestive evidence that AI assistance leads lower-skill agents to communicate more like high-skill agents. In particular, we plot the cosine similarity between high- and low-skill agents at specific moments in calendar time, separately for workers with (blue dots) and without (red diamonds) access to AI assistance. Among agents without access, we define high- and low-skill agents as those who are in the top or bottom quintile of our skill index for that month. Among agents with access, we define high- and low-skill agents based on whether they are in the top or bottom quintile of skill at the time of deployment.

Focusing on the blue dots, we see that the average textual similarity between high- and low-productivity workers is 0.55. This figure is lower than our average within-person text similarity (0.73), which makes sense given that these are across-person comparisons. Textual similarity is stable over time, suggesting that high- and low-skill workers are not trending differently in the

absence of the AI assistant. Turning to the red diamonds, post-AI, the textual similarity between high- and low-skill workers increases.

Combined with our results from Panel A, this suggests that low-skill workers are converging toward high-skill workers, rather than the opposite. The magnitude of this change—moving from 0.55 similarity to 0.61 similarity—may appear small, but given that the average within-person similarity for high-skill workers is 0.73, this result suggests that AI assistance is associated with a substantial narrowing of language gaps.

4.6 Effects on the Experience and Organization of Work

Access to AI assistance may impact workers and the firm organization as a whole through changes in work experience, turnover, and task allocation. In this section, we examine some of these outcomes.

4.6.1 Sentiment

Customers often vent their frustrations on anonymous service agents and, in our data, we see regular instances of swearing, verbal abuse, and "yelling" (typing in all caps). A key part of agents' jobs is to absorb customer frustrations while restraining one's own emotional reaction (Hochschild, 2019). 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). It is unclear what impact AI assistants may have on the tenor of conversations: AI recommendations may help the agent more effectively set the customer's expectations or resolve their problem, but customers may react poorly if AI-suggested language feels "corporate" or insincere.

To assess this, we attempt to 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. We compute sentiment scores for the agent and customer text of each chat separately, and then aggregate across all chats for each agent-year-month.

Panels A and B of Figure 13 describe the distributions of average customer and visitor sentiment scores. On average, customer sentiments in our data are mildly positive and normally distributed around a mean of 0.14 except for a mass of very positive and very negative scores. As one might expect, customer service agents are trained to be extremely positive and professional so agent sentiment scores are almost always highly positive, with a mean of 0.89.

Panels C and D consider how sentiment scores respond to the introduction of the AI assistant. In Panel C, we see an immediate and persistent improvement in customer sentiment. This effect is

economically large: according to Column 1 of Table 4, access to AI improves the mean customer sentiments (averaged over an agent-month) by 0.18 points, equivalent to half of a standard deviation. We see a much smaller impact on agent scores. In Panel D, we see no detectable effect for agent sentiment, which unsurprising because agent's expressed sentiment is already very high. Column 2 of Table 4 indicates that agent sentiments increase by only 0.02 points or about one percent of a standard deviation.

Appendix Figure A.7 examines heterogeneity sentiment impacts, by agent tenure and skill. Consistent with our main results, we find that most sentiment improvements occur for workers with lower tenure and skills at the time of deployment. In contrast with our productivity results, the effects on customer sentiments are broader and no longer necessarily monotonic in tenure or skill. We find the largest effects for workers with 3-6 months of tenure at AI model deployment, and we find a smaller impact of AI on sentiments only for agents in the highest quintile skill. These results suggest that AI recommendations, which were explicitly designed to prioritize more empathetic responses, may improve agents' social skills and have a positive emotional impact on customers.

4.6.2 Attrition

Increases in worker-level productivity do not always lead workers to be happier with their jobs. If workers become more productive but dislike being managed by an AI assistant, this may lead to greater turnover. If, on the other hand, AI assistance reduces stress, then workers may be more likely to stay.

Here, we examine how the deployment of the AI tool impacts worker attrition. For this analysis, we compare agents who are never treated with treated agents after treatment. In particular, we drop observations for treated agents prior to treatment because this group experiences no attrition in this period by construction (they must survive to be treated in the future). Our analysis compares the trajectories of agents with the same tenure but different access to AI assistance, controlling for location and time fixed effects.

Column 1 of Table 5 reports the main effect of the AI assistant on attrition. We find that, on average, the likelihood that a worker leaves in the current month goes down by 8.6 percentage points. Figure 14 considers how these patterns depend on an agent's tenure or skill at the time the AI system is introduced. We find 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, is large given baseline attrition rates for newer workers of about 25 percent. We caution that this effect may be overstated if access to the AI tool is more likely to be given to agents whom the firm deems more likely to stay. Without agent-fixed effects in our specification, we cannot account for

baseline differences been treated and untreated agents. In Panel B, we examine attrition by worker skill. Here, we find a significant decrease in attrition for all skill groups, but no systematic gradient.

4.6.3 Vertical and Horizontal Workflow

Changes in individual worker-level productivity may have broader implications for organizational workflows (Garicano, 2000; Athey et al., 1994; Athey and Stern, 1998). In most firms, customer support agents are organized both vertically and horizontally. Vertically, front-line agents try to resolve customer problems but can seek the help of supervisors when they are unsure of how to proceed. Customers in our data will sometimes attempt to escalate a conversation by asking to speak to a manager. This type of request generally occurs when the customer feels the current agent is not equipped to address their problem or becomes frustrated. Our data firm, like most other contact centers, employs designated "escalation agents" to deal with these requests.

Horizontally, agents often represent specific departments that handle specific tasks. For example, some may specialize in technical software issues while others specialize in account management issues. A customer with a technical issue requiring that their account be upgraded to a premium product would likely be transferred to a different department.

In Figure 15 and Table 5, we consider the impact of AI model deployment on vertical and horizontal assistance. We do not have a direct measure of whether a call is actually escalated to the manager. Instead, we examine chat-level text data to examine requests for escalations: instances in which a customer requests to speak to a manager or supervisor. In Column 1 of Table 5, we find that AI assistance generates an almost 25 percent decline in customer requests to speak to a manager. The accompanying event study is presented in Panel A of Figure 15: we see that declines in requests are persistent and grow over time.

In contrast, we find mixed-to-positive evidence on the impact of access to AI on the number of horizontal transfers to other agents. In Column 2 of Table 5, we see a positively-signed but statistically insignificant impact on transfers. Panel B of Figure 15 suggests that transfers initially increase but then slowly decline. Upon reviewing the text of these conversations, we see that most transfers appear to redirect customers to a more appropriate department. Many of these transfers occur early in the call, suggesting that transfers reflect a matching between a customer's problem and an agent's specialty.

Finally, Appendix Figure A.8 considers how these patterns change by worker skill and tenure. Panels A and C focus on requests to speak to a manager. We find that these requests are particularly reduced for agents who were less skilled or less experienced at the time of AI adoption. Panels B

and D consider call transfers. Here we find mixed results, with positive impacts for some skill and tenure groups and negative impacts for others, but no clear pattern.

5 Conclusion

Progress in machine learning opens up a broad set of economic possibilities. Our paper provides the first empirical evidence on the effects of a generative AI tool in a real-world workplace. In our setting, we find that access to AI-generated recommendations increases worker productivity, improves customer sentiment, and is associated with reductions in employee turnover.

We hypothesize the part of the effect we document is driven by the AI system's ability to embody the best practices of high-skill workers in our firm. These practices may have previously been difficult to disseminate because they involve tacit knowledge. Consistent with this, we see that AI assistance leads to substantial improvements in problem resolution and customer satisfaction for newer- and less-skilled workers, but does not help the highest-skilled or most-experienced workers on these measures. Analyzing the text of agent conversations, we find suggestive evidence that AI recommendations lead low-skill workers to communicate more like high-skill workers.

Our findings, and their limitations, point to a variety of directions for future research.

As a potential general-purpose technology, generative AI can and will be deployed in a variety of ways, and the effects we find may not generalize across all firms and production processes (Eloundou et al., 2023). 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—trained on historical data—may be different.

Our results do not capture potential longer-term impacts on skill demand, job design, wages, or customer demand. For example, more effective technical support could accelerate the trend towards contact center agents taking on more complex customer responsibilities, increasing aggregate demand even if agents become more productive (Berg et al., 2018; Korinek, 2022). And over the longer term, these tools can uncover patterns and insights that may not be documented in formal channels, changing the way workers are managed or how knowledge is shared within an organization.

Finally, our findings raise questions about whether and how workers should be compensated for the data that they provide to AI systems. High-skill workers, in particular, play an important role in model development but see smaller direct benefits in terms of improving their own productivity.

Given the early stage of generative AI, these and other questions deserve further scrutiny.

References

Acemoglu, Daron and David Autor, "Skills, tasks and technologies: Implications for employment and earnings," in "Handbook of labor economics," Vol. 4, Elsevier, 2011, pp. 1043–1171.

_ and Pascual Restrepo, "Low-Skill and High-Skill Automation," *Journal of Human Capital*, June 2018, *12* (2), 204–232.

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

_ , Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti, "Technology, Information, and the Decentralization of the Firm*," *The Quarterly Journal of Economics*, November 2007, *122* (4), 1759–1799. _eprint: https://academic.oup.com/qje/article-pdf/122/4/1759/5234557/122-4-1759.pdf.

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

Athey, Susan and Scott Stern, "An Empirical Framework for Testing Theories About Complimentarity in Organizational Design," Working Paper 6600, National Bureau of Economic Research June 1998.

_ and _, "The Impact of Information Technology on Emergency Health Care Outcomes," *RAND Journal of Economics*, Autumn 2002, *33* (3), 399–432.

_ , Joshua Gans, Scott Schaefer, and Scott Stern, "The Allocation of Decisions in Organizations," *Stanford Graduate School of Business*, 1994.

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

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

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

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Yoshua Bengio and Yann LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

Bartel, Ann, Casey Ichniowski, and Kathryn Shaw, "How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills*," *The Quarterly Journal of Economics*, 11 2007, *122* (4), 1721–1758.

Berg, Jeff, Avinash Das, Vinay Gupta, and Paul Kline, "Smarter call-center coaching for the digital world," Technical Report, McKinsey & Company November 2018.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen, "The Distinct Effects of Information Technology and Communication Technology on Firm Organization," *Management Science*, 2014, *60* (12), 2859–2885.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

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

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics*, 02 2002, *117* (1), 339–376.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language Models are Few-Shot Learners," July 2020. arXiv:2005.14165 [cs].

Brynjolfsson, Erik and Tom Mitchell, "What Can Machine Learning, Do? Workforce Implications," *Science*, December 2017, *358*, 1530–1534.

Bubeck, Sebastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

Buesing, Eric, Vinay Gupta, Sarah Higgins, and Raelyn Jacobson, "Customer care: The future talent factory," Technical Report, McKinsey & Company June 2020.

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

Casetext, "CoCounsel builds on the power of GPT-4, the AI that outperformed real bar candidates," Technical Report, Casetext March 2023.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs*," *The Quarterly Journal of Economics*, May 2019, *134* (3), 1405–1454.

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky, "Global survey: The state of AI in 2021," Technical Report, McKinsey & Company 2021.

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

Dunn, Andrew, Diana Inkpen, and Răzvan Andonie, "Context-Sensitive Visualization of Deep Learning Natural Language Processing Models," 2021.

Elicit, "Elicit: The AI Research Assistant," 2023.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," March 2023. arXiv:2303.10130 [cs, econ, q-fin].

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

_ **and Esteban Rossi-Hansberg**, "Knowledge-Based Hierarchies: Using Organizations to Understand the Economy," *Annual Review of Economics*, 2015, *7* (1), 1–30.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, December 2021, *225* (2), 254–277.

Google, "AI vs. Machine Learning: How Do They Differ?"

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

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp, "More than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, 2023, *40* (1), 75–87.

Hochschild, Arlie Russell, *The managed heart: Commercialization of human feeling*, University of California press, 2019.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in Hiring*," *The Quarterly Journal of Economics*, 10 2017, *133* (2), 765–800.

Hugging Face, "sentence-transformers/all-MiniLM-L6-v2," April 2023.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

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

Korinek, Anton, "How innovation affects labor markets: An impact assessment," Working Paper, Brookings Institution June 2022.

Koroteev, M. V., "BERT: A Review of Applications in Natural Language Processing and Understanding," 2021.

Lee, Don, "The Philippines has become the call-center capital of the world," *Los Angeles Times*, February 2015. Section: Business.

Legg, Shane, Marcus Hutter et al., "A collection of definitions of intelligence," *Frontiers in Artificial Intelligence and applications*, 2007, *157*, 17.

Li, Chun, "OpenAI's GPT-3 Language Model: A Technical Overview," June 2020.

Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge, "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models," April 2023. arXiv:2304.01852 [cs].

Meijer, Erik, "Behind every great deep learning framework is an even greater programming languages concept (keynote)," in "Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering" 2018, pp. 1–1.

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

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

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

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

OpenAI, "GPT-4 Technical Report," Technical Report, OpenAI March 2023.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, "Training language models to follow instructions with human feedback," March 2022. arXiv:2203.02155 [cs].

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

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

Polanyi, Michael, *The Tacit Dimension*, Chicago, IL: University of Chicago Press, May 1966.

Radford, Alec and Karthik Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.

_ , Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

Roose, Kevin, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," *The New York Times*, February 2023.

Rosen, Sherwin, "The Economics of Superstars," *The American Economic Review*, 1981, *71* (5), 845–858.

Singh, Nikhil, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman, "Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence," *ACM Transactions on Computer-Human Interaction*, February 2022. Just Accepted.

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

Syverson, Chad, "What Determines Productivity?," *Journal of Economic Literature*, June 2011, *49* (2), 326–65.

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

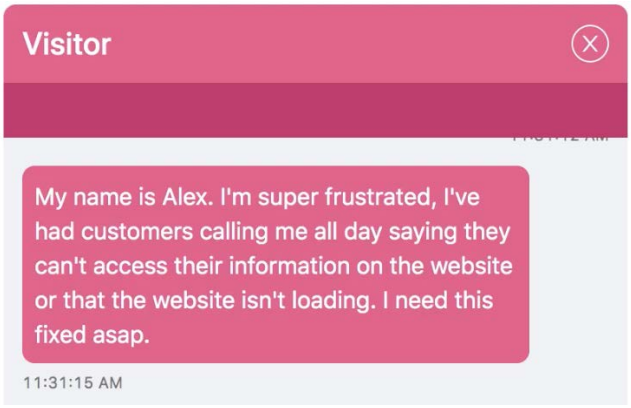
The White House, "The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America," Technical Report, The White House December 2022.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," December 2017. arXiv:1706.03762 [cs].

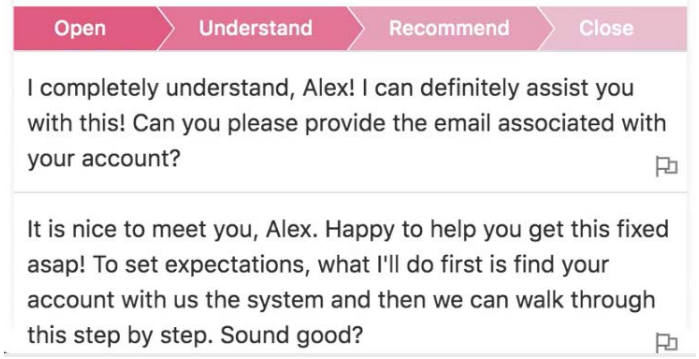
Zhao, Shuyin, "GitHub Copilot now has a better AI model and new capabilities," February 2023.

Figure 1: Sample AI Output

A. Sample Customer Issue

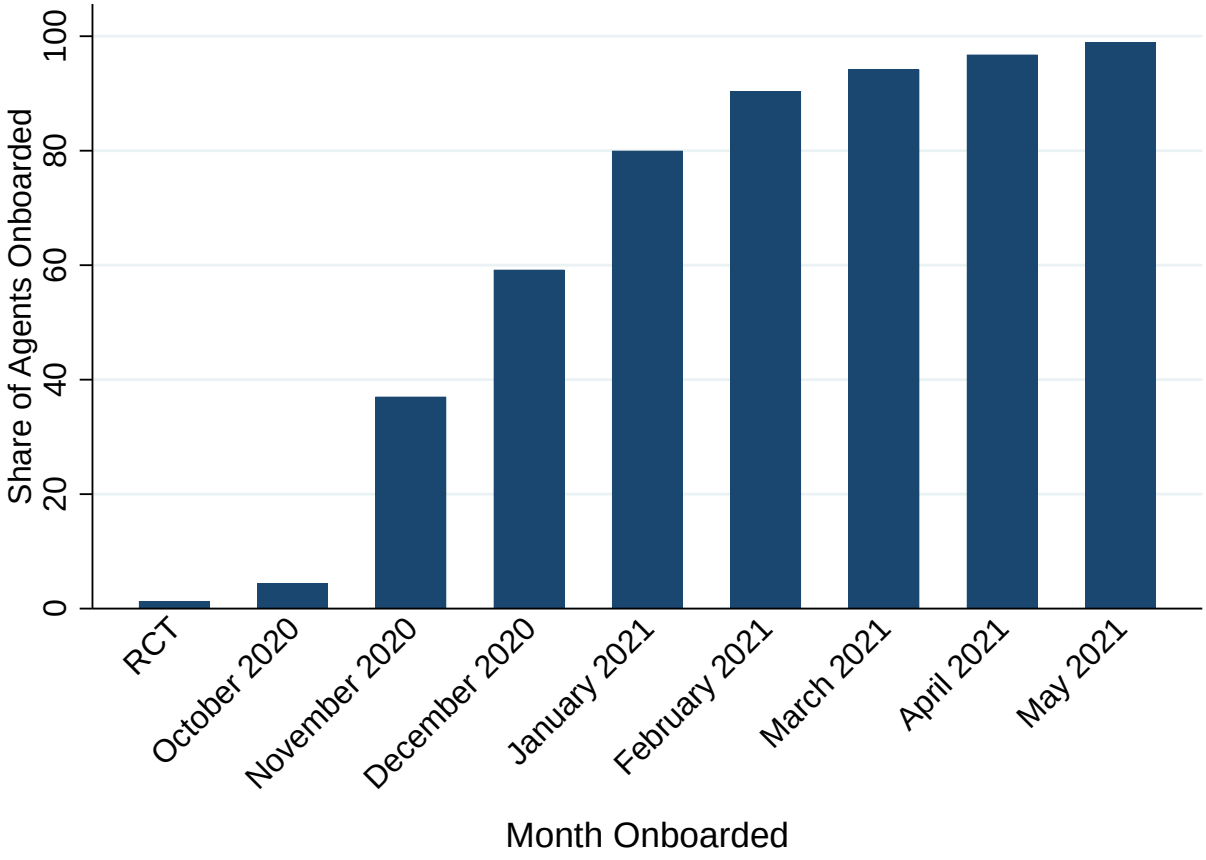


B. Sample AI-generated Suggested Response



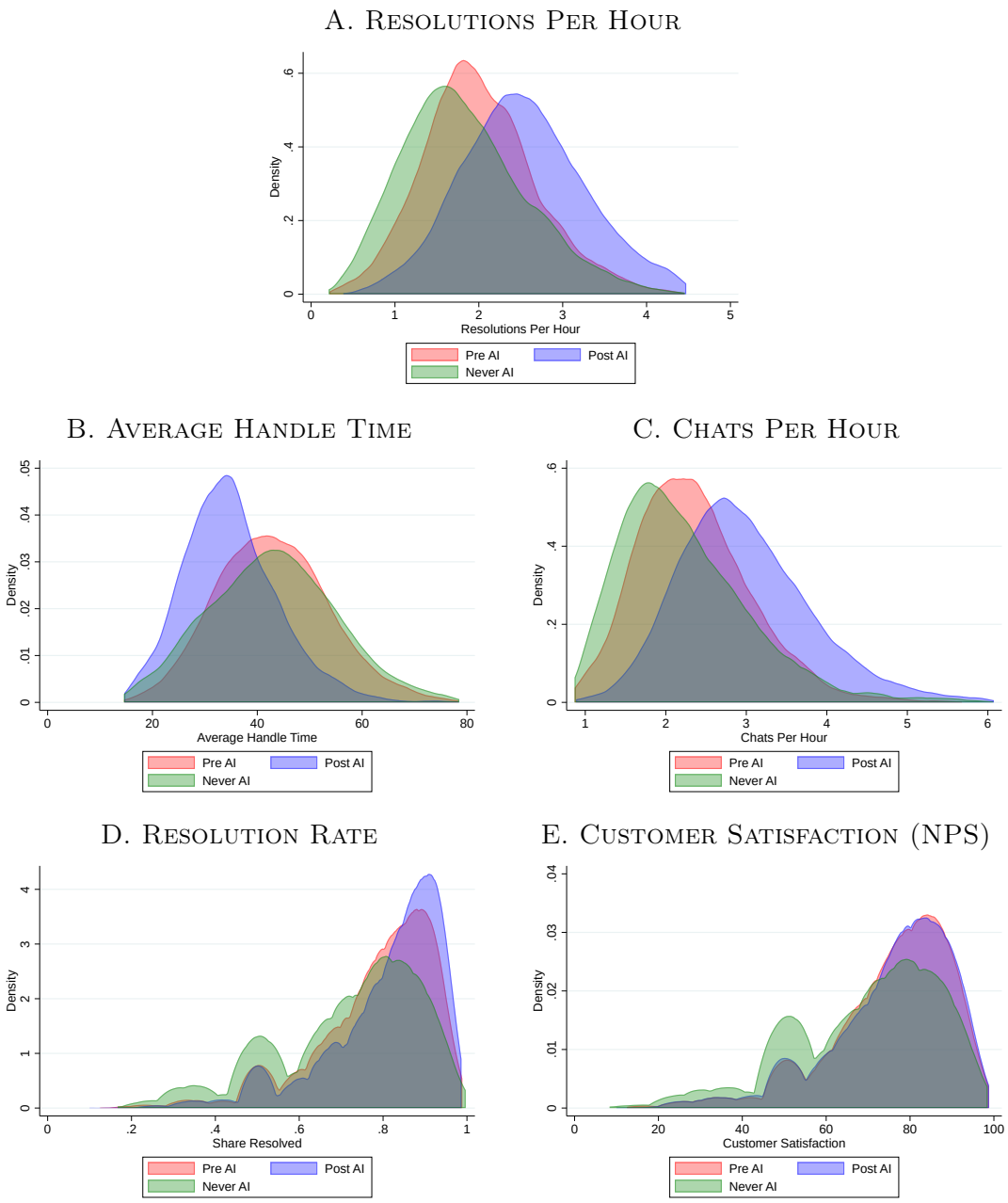
Notes: This figure shows sample suggestions of output generated by the AI model. The suggested responses are only visible to the agent. Workers can choose to ignore, accept or somewhat incorporate the AI suggestions into their response to the customer.

FIGURE 2: 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. The firm ran a small randomized control trial in August and September of 2020. All data are from the firm's internal software systems.

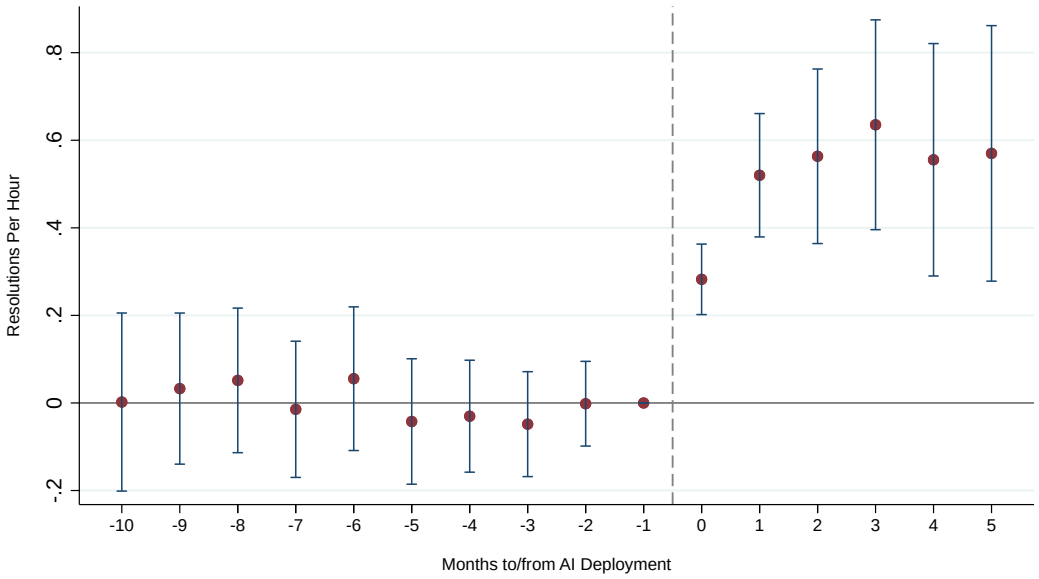
FIGURE 3: RAW PRODUCTIVITY DISTRIBUTIONS, BY AI TREATMENT



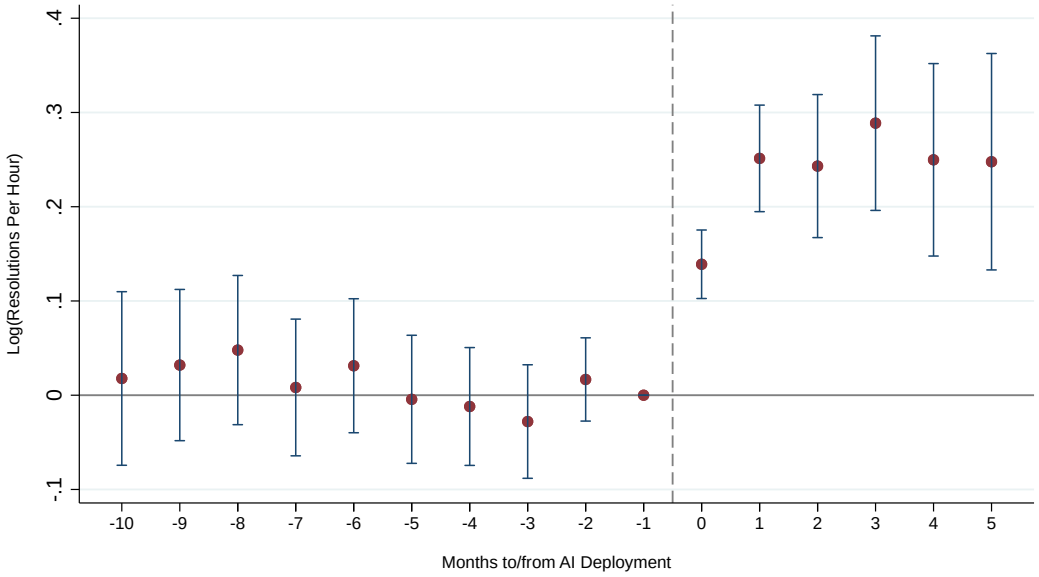
NOTES: This figure shows the distribution various outcome measures. We split this sample into agent-month observations for agents who eventually receive access to the AI system before deployment ("Pre AI"), after deployment ("Post AI"), and for agent-months associated with agents who never receive access ("Never AI"). 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; "calls per hour," the number of calls 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 come from the firm's software systems.

Figure 4: Event Studies, Resolutions Per Hour

A. Resolutions Per Hour



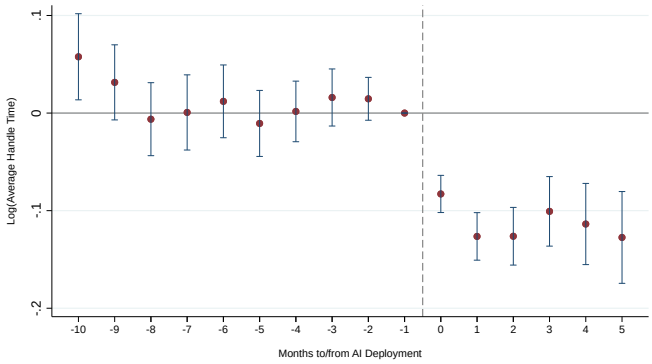
B. Log(Resolutions Per Hour)



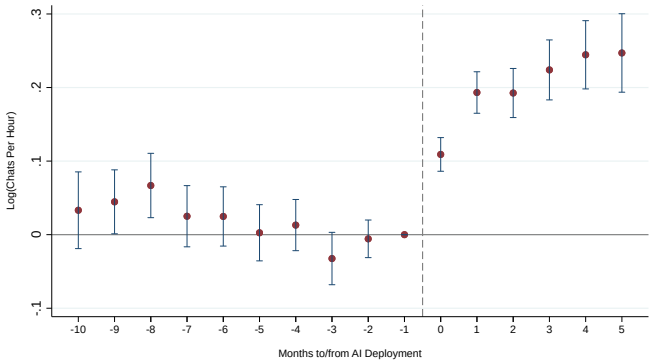
Notes: These figures plot the coefficients and 95 percent confidence interval from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the resolutions per hour and Panel B plots the natural log of the measure. All specifications include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

Figure 5: Event Studies, Additional Outcomes

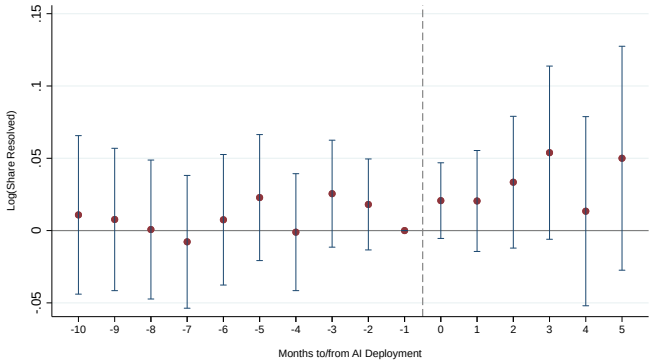
A. Log(Average Handle Time)



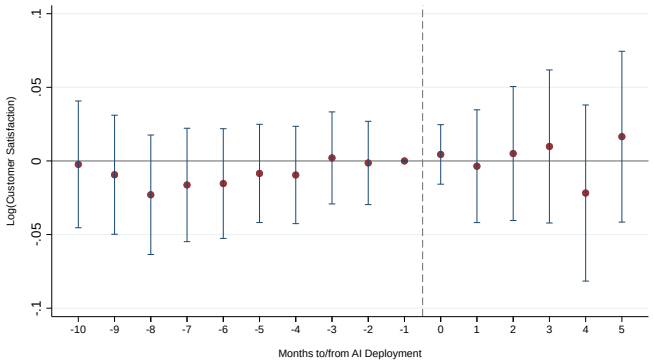
B. Log(Chats Per Hour)



C. Log(Resolution Rate)



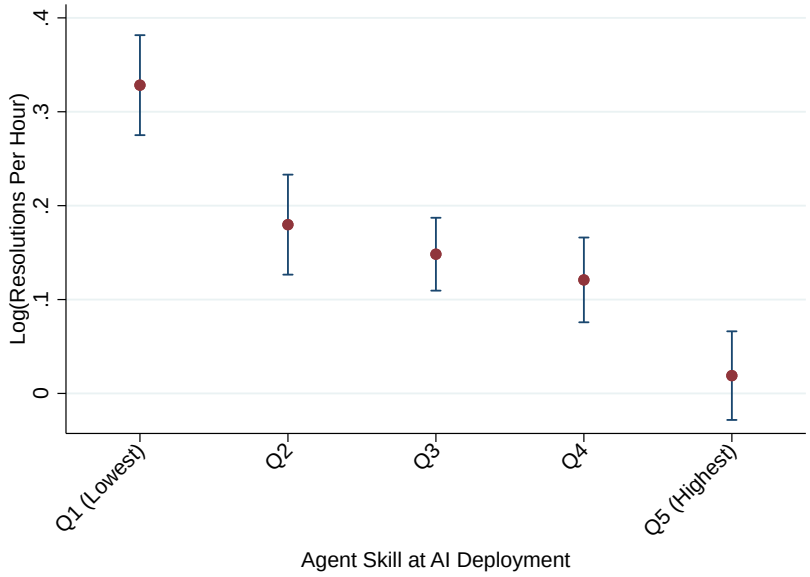
D. Log(Customer Satisfaction (NPS))



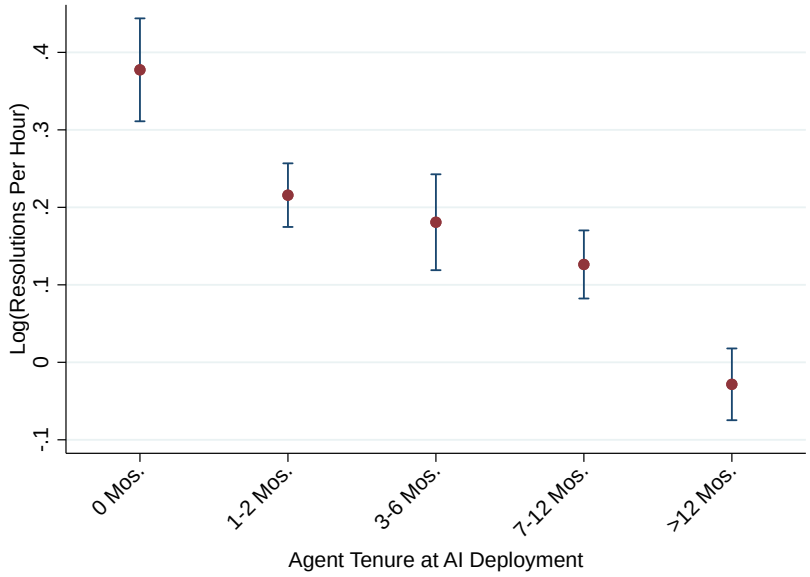
Notes: These figures plot the coefficients and 95 percent confidence interval from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the average handle time or the average duration of each technical 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 include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

Figure 6: Heterogeneity of AI Impact, by Skill and Tenure

A. Impact of AI on Resolutions Per Hour, by Skill at Deployment

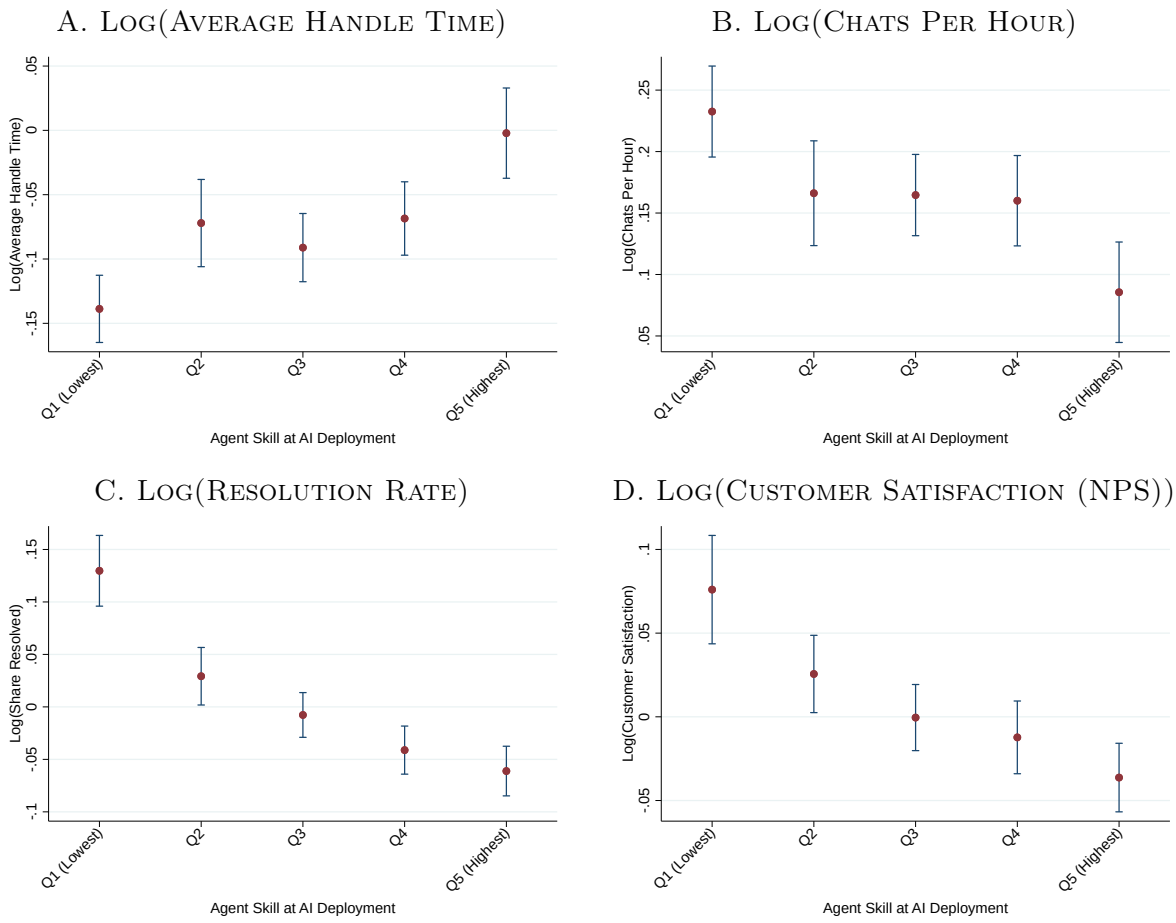


B. Impact of AI on Resolutions Per Hour, by Tenure at Deployment



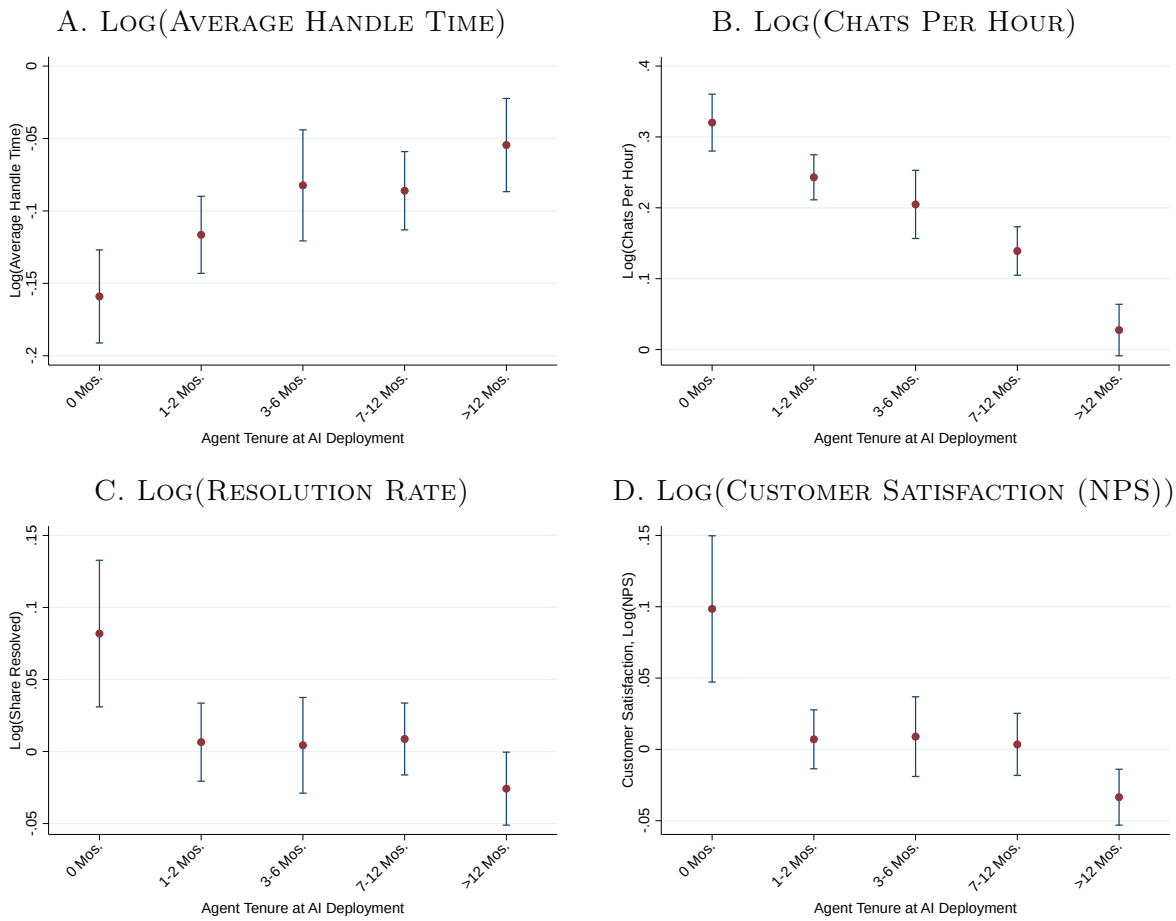
Notes: These figures plot the impacts of AI model deployment on log(resolutions per hour) for different groups of agents. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses to assess agent performance. Within each month and company, agents are grouped into quintiles, with the most productive agents in quintile 5 and the least productive in quintile 1. Pre-AI worker tenure is the number of months an agent has been employed when they receive access to AI recommendations. All specifications include agent and chat year-month, location, and company fixed effects and standard errors are clustered at the agent level.

Figure 7: Heterogeneity of AI Impact by pre-AI Worker Skill, Additional Outcomes



Notes: These figures plot the impacts of AI model deployment on four measures of productivity and performance, by pre-deployment worker skill. 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 average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of calls an agent can handle per hour. Panel C plots the resolution rate, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects and standard errors are clustered at the agent level.

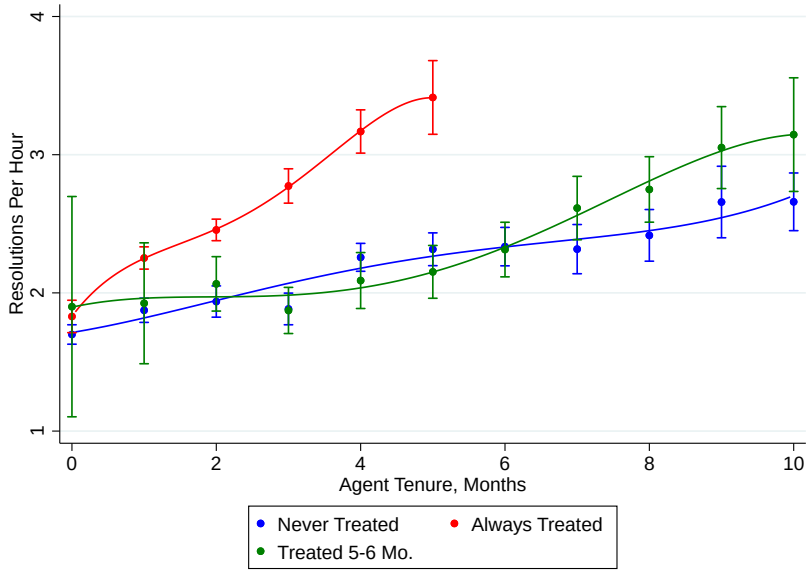
Figure 8: Heterogeneity of AI Impact by pre-AI Worker Tenure, Additional Outcomes



Notes: These figures plot the impacts of AI model deployment on 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 average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of calls an agent can handle per hour. Panel C plots the resolution rate, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects and standard errors are clustered at the agent level.

Figure 9: Experience Curves by Deployment Cohort

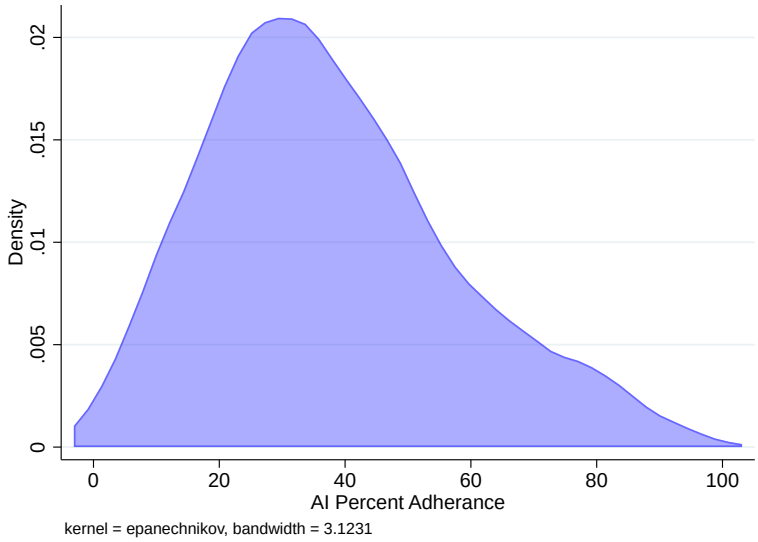
Resolutions Per Hour, by Agent Tenure



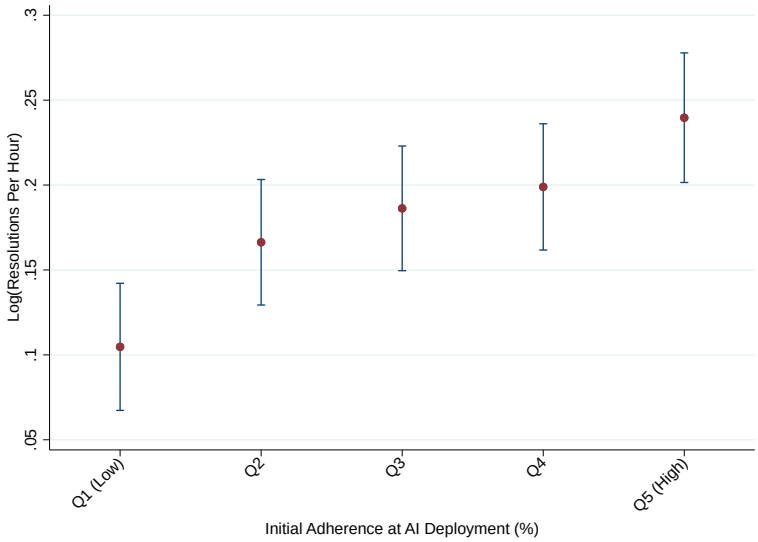
Notes: This figure plot the relationship between productivity and job tenure. The red line plots the performance of always-treated agents, those who have access to AI assistance from their first month on the job. The blue line plots agents who are never treated. The green line plots agents who spend their first four months of work without the AI assistance, and gain access to the AI model during their fifth month on the job. 95th percent confidence intervals are shown.

Figure 10: Heterogeneity of AI Impact, by AI Adherence

A. Distribution of AI Adherance



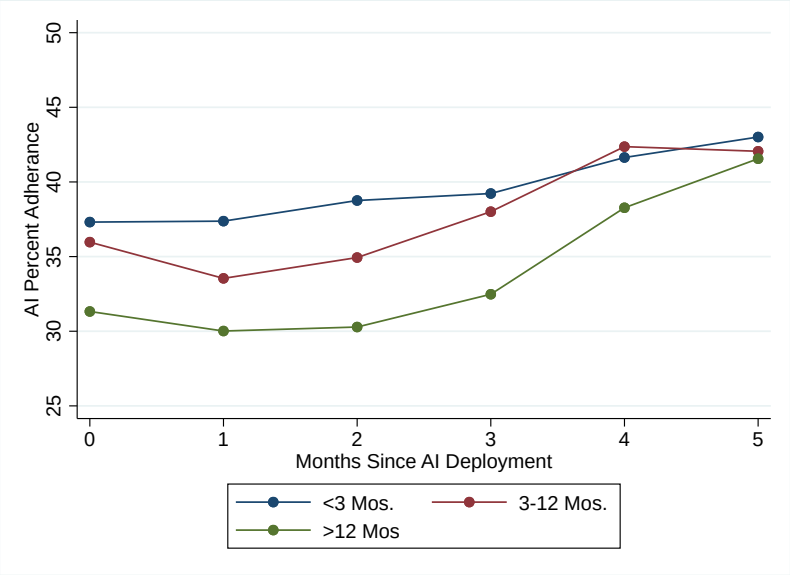
B. Impact of AI on Resolutions Per Hour, by Initial Adherence



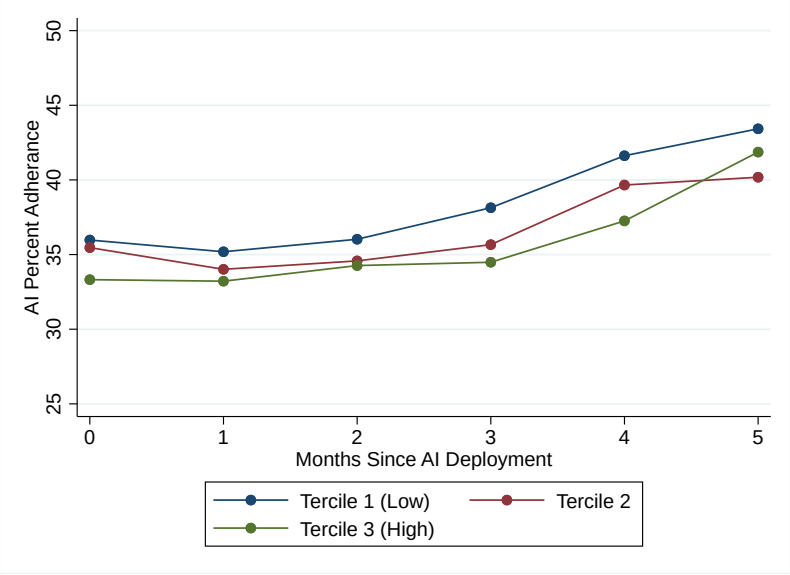
Notes: Panel A plots the distribution of AI adherence, averaged at the agent-month level, weighted by the log of the number of AI recommendations for that agent-month. Panel B shows the impact of AI assistance on resolutions by hour, by agents grouped by their initial adherence, defined as the share of AI recommendations they followed in the first month of treatment.

Figure 11: AI Adherence over Time

A. By Agent Tenure at AI Model Deployment



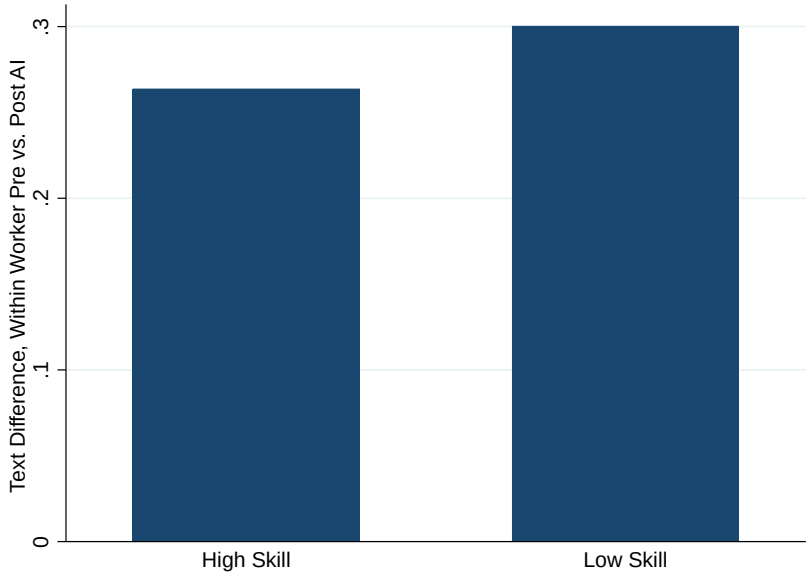
B. By Agent Skill at AI Model Deployment



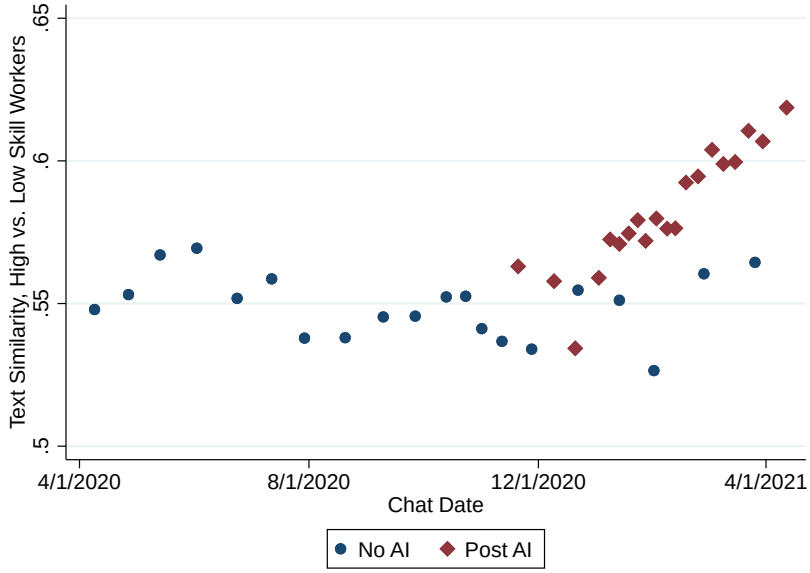
Notes: This figure plots the share of AI suggestions followed by agents as a function of the number of months each agent has had access to the AI model. In Panel A, the red line plots the adherence of agents with 3 to 12 months of experience at AI model deployment, the green line plots adherence of agents with over a year of experience and the blue line plots the adherence rates of agents with less than three months of experience when given access to the AI model. In our sample, the average tenure of agents is nine months. Panel B plots adherence over time by tercile of pre-deployment agent productivity: blue is the ex-ante least productive agents, red represents middle-skill workers and green are high-skill workers. All data come from the firm's internal software systems.

Figure 12: Textual Change

A. Within-Person Textual Change, Pre and Post AI Model Deployment



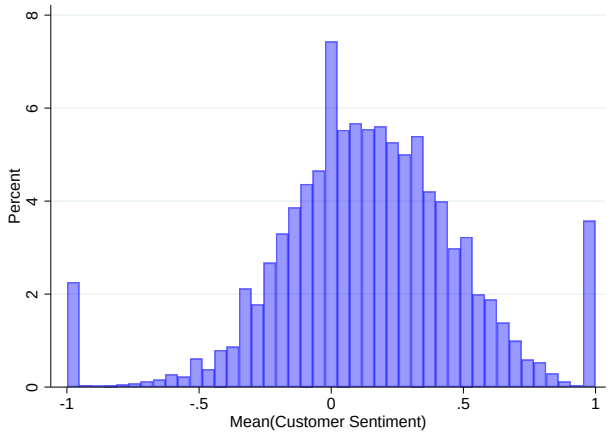
B. Text Similarity Between Low-Skill and High-Skill Workers, Pre and Post AI



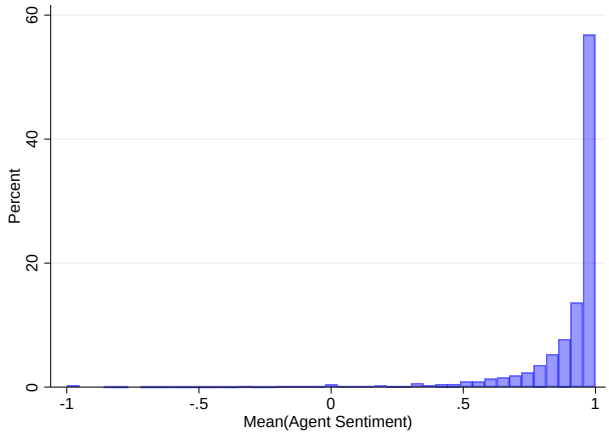
Notes: Panel A plots the average difference between an agent's pre-AI corpus of chat messages and that same agent's post-AI corpus, controlling for year-month and agent tenure. The first bar represents the average pre-post text difference for agents in the highest quintile of pre-AI skill, as measured by a weighted index of their calls per hour, resolution rate, and customer satisfaction score. The low-skill bar represents the same type of pre-post text difference among the lowest skill quintile. Agent skill, or relative productivity, is defined at the time of treatment. Panel B plots the average text similarity between the top and bottom quintile of agents. The blue line plots the similarity for never treated or pre-treatment agents, the red line plots the similarity for agents with access to the AI model. For agents in the treatment group, we define agent skill prior to AI model deployment. Our analysis includes controls for agent tenure.

Figure 13: Conversation Sentiment

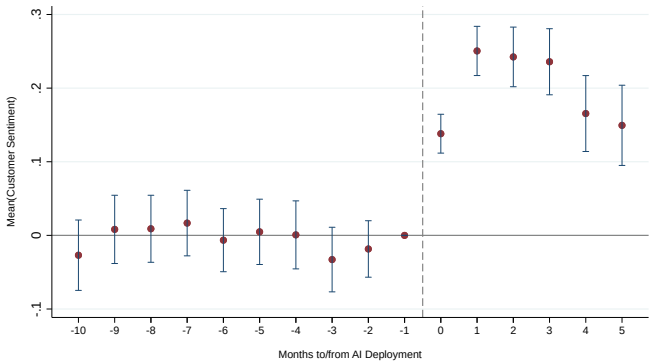
A. Customer Sentiment, Histogram



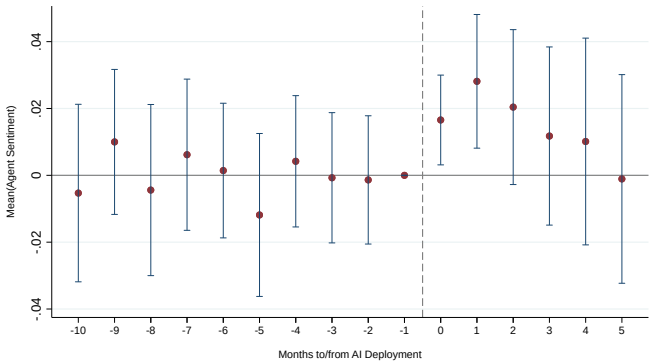
B. Agent Sentiment



C. Customer Sentiment, Event Study



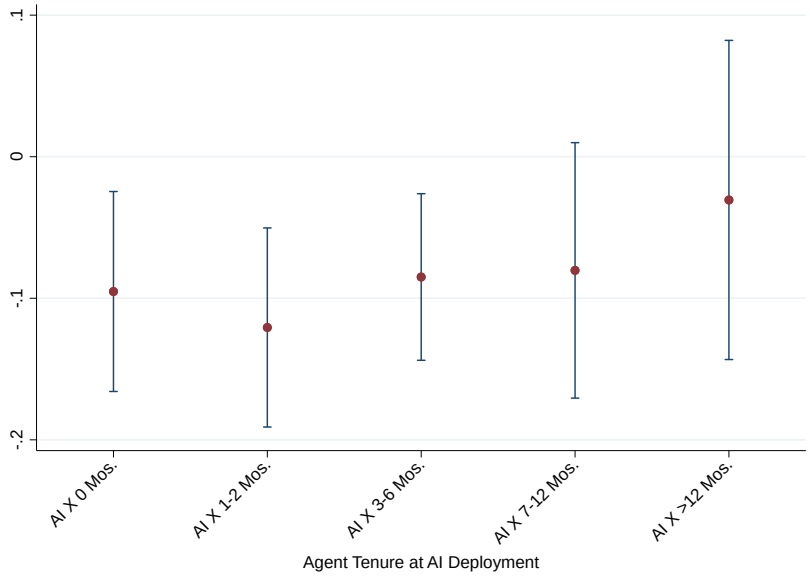
D. Agent Sentiment, Event Study



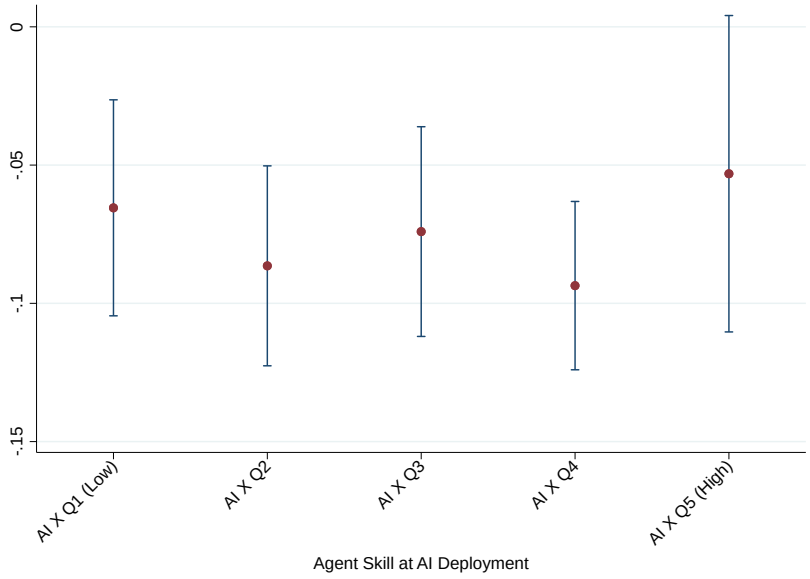
Notes: Each panel of this figure plots the impact of AI model deployment on conversational sentiment. Panel A shows average customer sentiments. Panel B shows average agent sentiments. Panel C plots the event study of AI model deployment on customer sentiment and Panel D plots the corresponding estimate for agent sentiment. Sentiment is measured using SiEBERT, a fine-tuned checkpoint of a RoBERTA, an English language transformer model. All data come from the firm's internal software systems.

Figure 14: Impact of AI Model Deployment on Worker Attrition

A. By Tenure at AI Model Deployment



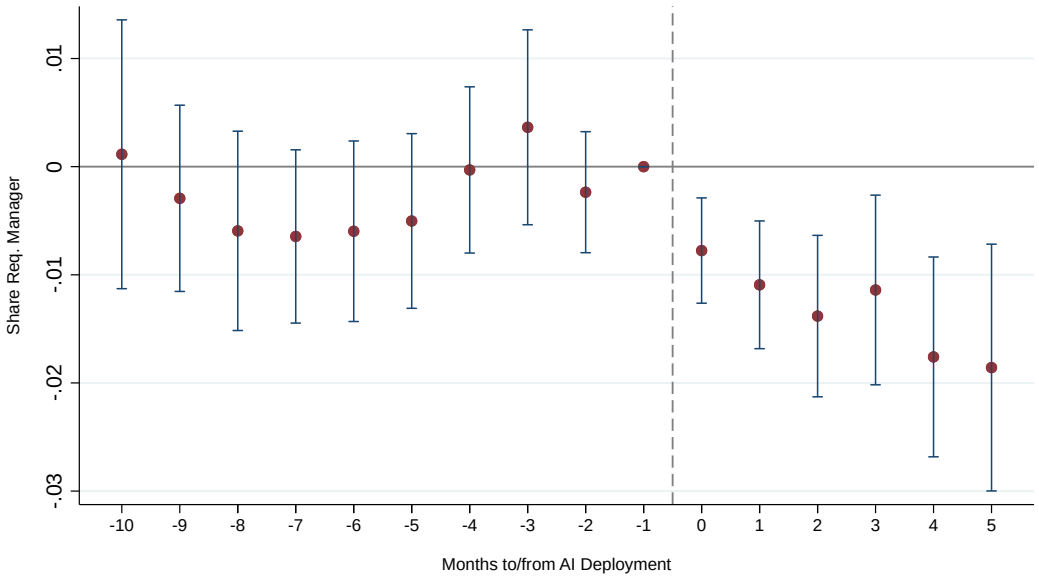
B. By Productivity at AI Model Deployment



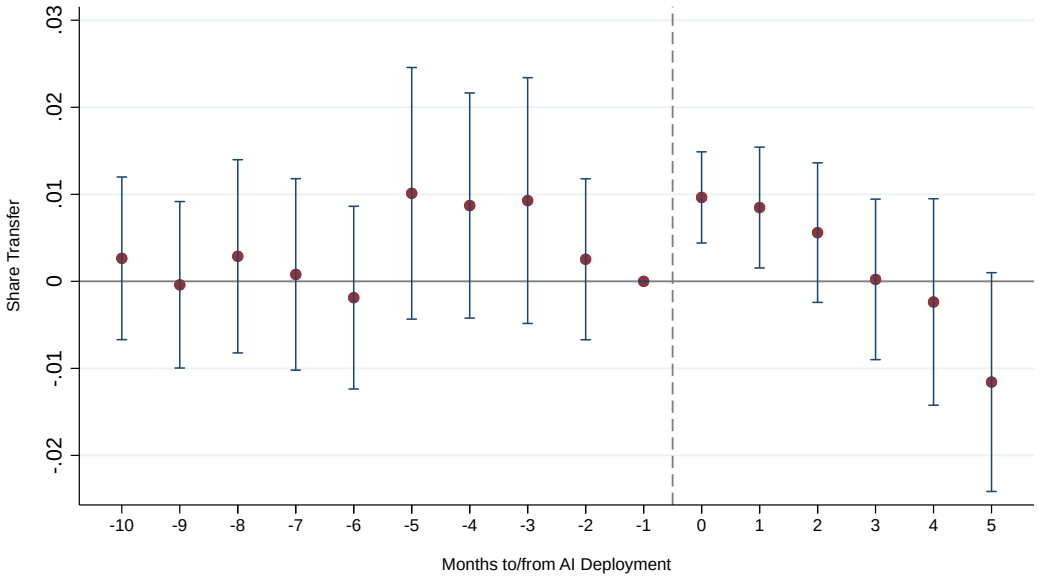
Notes: This figure presents the results of the impact of AI model deployment on workers' likelihood of attrition. Panel A graphs the effects of AI assistance on attrition by agent tenure at AI model deployment. Panel B plots the same impact by agent skill index at AI model deployment. All specifications include chat year and month fixed effects, as well as agent location, company and agent tenure. All robust standard errors are clustered at the agent level. All data come from the firm's internal software systems.

FIGURE 15: IMPACT OF AI ON CHAT ESCALATION AND TRANSFERS

A. ESCALATION (REQUESTS FOR MANAGER ASSISTANCE)



B. CHAT TRANSFERS



NOTES: This figure reports the coefficients and 95 percent confidence intervals for the event study of AI model deployment on manager assistance and transfers or agent requests for help from other agents. Panel A graphs requests to speak to a manager, which are initiated by the customer usually when they are unsatisfied or frustrated with the interaction. Panel B plots transfers, which are usually initiated either when an agent is not responsible for the customer's problem or when the agent cannot solve a technical support issue and needs help from another agent. All robust standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 1: Applicant Summary Statistics

Variable	All	Never Treated	Treated, Pre	Treated, Post
Chats	3,007,501	945,954	882,105	1,180,446
Agents	5,179	3,523	1,341	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	127	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 conversations, agent characteristics and issue resolution rates, customer satisfaction and average call duration. The sample in Column 1 consists of all agents in our sample. Column 2 includes control agents who were never given access to the AI model. Column 3 and 4 present pre-and-post AI model deployment summary statistics for treated agents who were given access to the AI model. All data come from the firm's internal software systems.

TABLE 2: MAIN EFFECTS: PRODUCTIVITY (RESOLUTIONS PER HOUR)

VARIABLES	(1) Res./Hr	(2) Res./Hr	(3) Res./Hr	(4) Log(Res./Hr)	(5) Log(Res./Hr)	(6) Log(Res./Hr)
Post AI X Ever Treated	0.468***	0.371***	0.301***	0.221***	0.180***	0.138***
	(0.0542)	(0.0520)	(0.0498)	(0.0211)	(0.0188)	(0.0199)
Ever Treated	0.109*			0.0572*		
	(0.0582)			(0.0316)		
Observations	13,225	12,328	12,328	12,776	11,904	11,904
R-squared	0.250	0.563	0.575	0.260	0.571	0.592
Year Month FE	YES	YES	YES	YES	YES	YES
Location FE	YES	YES	YES	YES	YES	YES
Agent FE	-	YES	YES	-	YES	YES
Agent Tenure FE	-	-	YES	-	-	YES
DV Mean	2.121	2.174	2.174			

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 (res/hour). Columns 1 and 4 include agent geographic location and year-by-month fixed effects. Columns 2 and 5 include agent-level fixed effects, and columns 3 and 6, our preferred specification, also control for agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 3: Main Effects: Additional Outcomes

VARIABLES	(1) AHT	(2) Calls/Hr	(3) Res. Rate	(4) NPS	(5) Log(AHT)	(6) Log(Calls/Hr)	(7) Log(Res. Rate)	(8) Log(NPS)
Post AI X Ever Treated	-3.750***	0.366***	0.0128*	-0.128	-0.0851***	0.149***	0.00973*	-0.000406
	(0.476)	(0.0363)	(0.00717)	(0.660)	(0.0110)	(0.0142)	(0.00529)	(0.00915)
Observations	21,885	21,885	12,328	12,578	21,885	21,885	11,904	12,188
R-squared	0.590	0.564	0.369	0.525	0.622	0.610	0.394	0.565
Year Month FE	YES	YES	YES	YES	YES	YES	YES	YES
Location FE	YES	YES	YES	YES	YES	YES	YES	YES
Agent FE	YES	YES	YES	YES	YES	YES	YES	YES
Agent Tenure FE	YES	YES	YES	YES	YES	YES	YES	YES
DV Mean	40.65	2.557	0.821	79.58				

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. Post AI X Treated measures the impact of AI model deployment after deployment on treated agents for average handle time or average call duration, calls per hour, the number of calls 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, and each metrics corresponding natural log equivalents. All specifications include agent fixed effects and chat year and month fixed effects, as well as controls for agent location and agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 4: Agent and Customer Sentiment

VARIABLES	(1) Mean(Customer Sentiment)	(2) Mean(Agent Sentiment)
Post AI X Ever Treated	0.177***	0.0198***
	(0.0133)	(0.00315)
Observations	21,218	21,218
R-squared	0.485	0.596
Year Month FE	YES	YES
Location FE	YES	YES
Agent FE	YES	YES
Agent Tenure FE	YES	YES
DV Mean	0.141	0.896

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. All specifications include agent fixed effects and chat year and month fixed effects, as well as agent location and agent tenure, which account for differing likelihood of attrition by agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Table 5: Organizational Changes

VARIABLES	(1) Leaves this Month	(2) Request Manager	(3) Transfer
Post AI X Ever Treated	-0.0868**	-0.00875***	0.00792
	(0.0319)	(0.00215)	(0.00535)
Observations	17,902	21,839	21,839
R-squared	0.206	0.482	0.386
Year Month FE	YES	YES	YES
Location FE	YES	YES	YES
Agent Tenure FE	YES	YES	YES
DV Mean	0.288	0.0377	0.0219
Agent FE		YES	YES

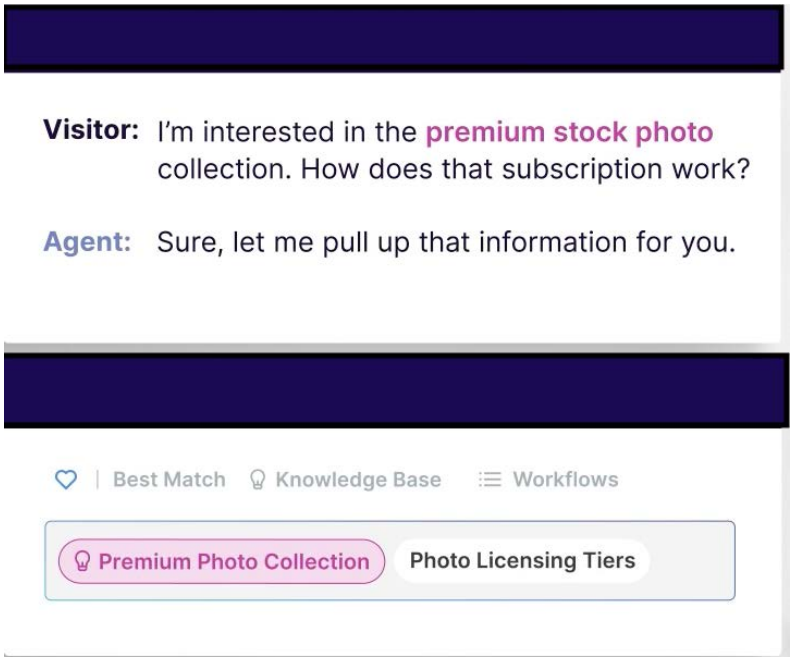
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

Notes: This table shows the impact of AI assistant deployment on agent attrition, customer requests for manager help, and call transfers. The sample for Column 1 consists of all agent-months for untreated agents, and post-treatment months for treated agents. Because exit occurs only once in our sample, Column 1 does not include agent fixed effects. Columns 2 and 3 estimate our standard difference-in-difference result for our full sample and include agent fixed effects and chat year and month fixed effects, as well as agent location and agent tenure. All standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Appendix Materials

FIGURE A.1: SAMPLE AI TECHNICAL SUGGESTION

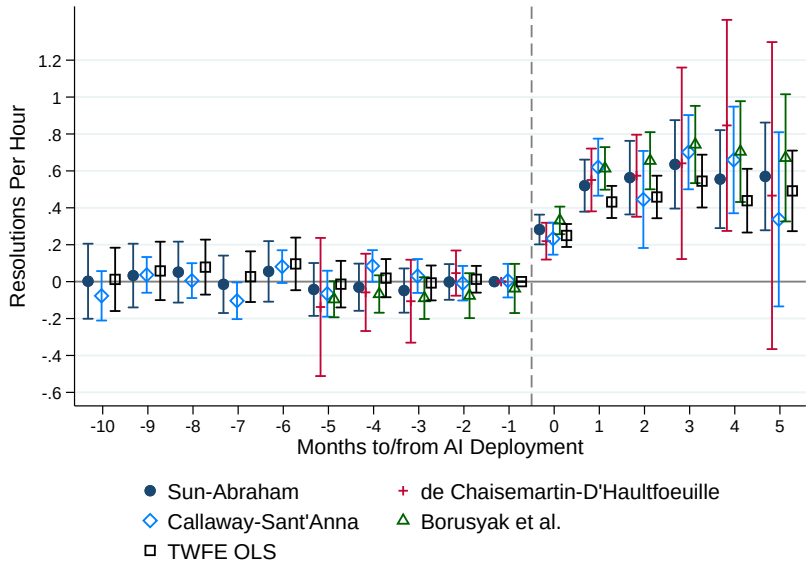
A. SAMPLE AI-GENERATED TECHNICAL LINK



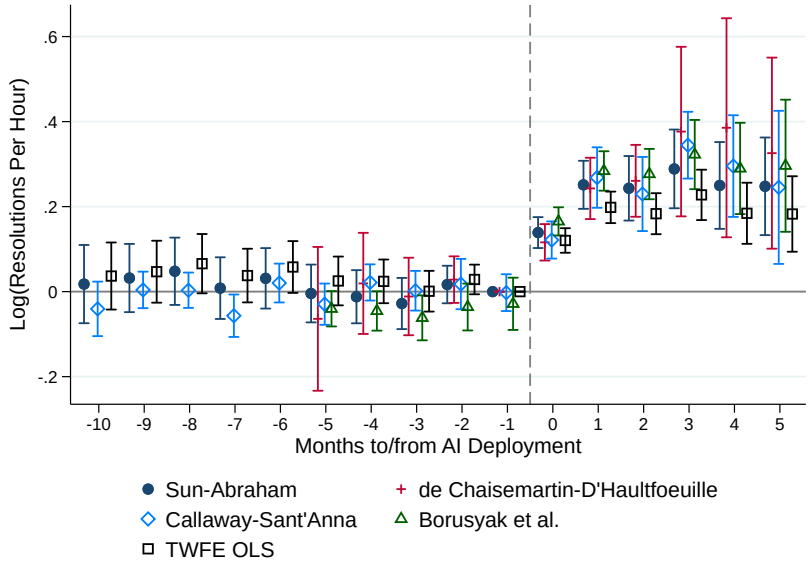
NOTES: This figure shows a sample technical documentation suggestions made the by AI. Our data firm has an extensive set of documentation for their technical support agents, known as the knowledge base, which is like an internal company Wikipedia for product and process information. The AI will attempt to surface the most helpful technical documentation page when triggered to do so during a customer interaction. These links are only visible to the agent and agents must review to see if the resource is helpful. Workers can choose to read the suggested technical documentation or ignore the recommendation.

FIGURE A.2: EVENT STUDIES, RESOLUTIONS PER HOUR

A. RESOLUTIONS PER HOUR

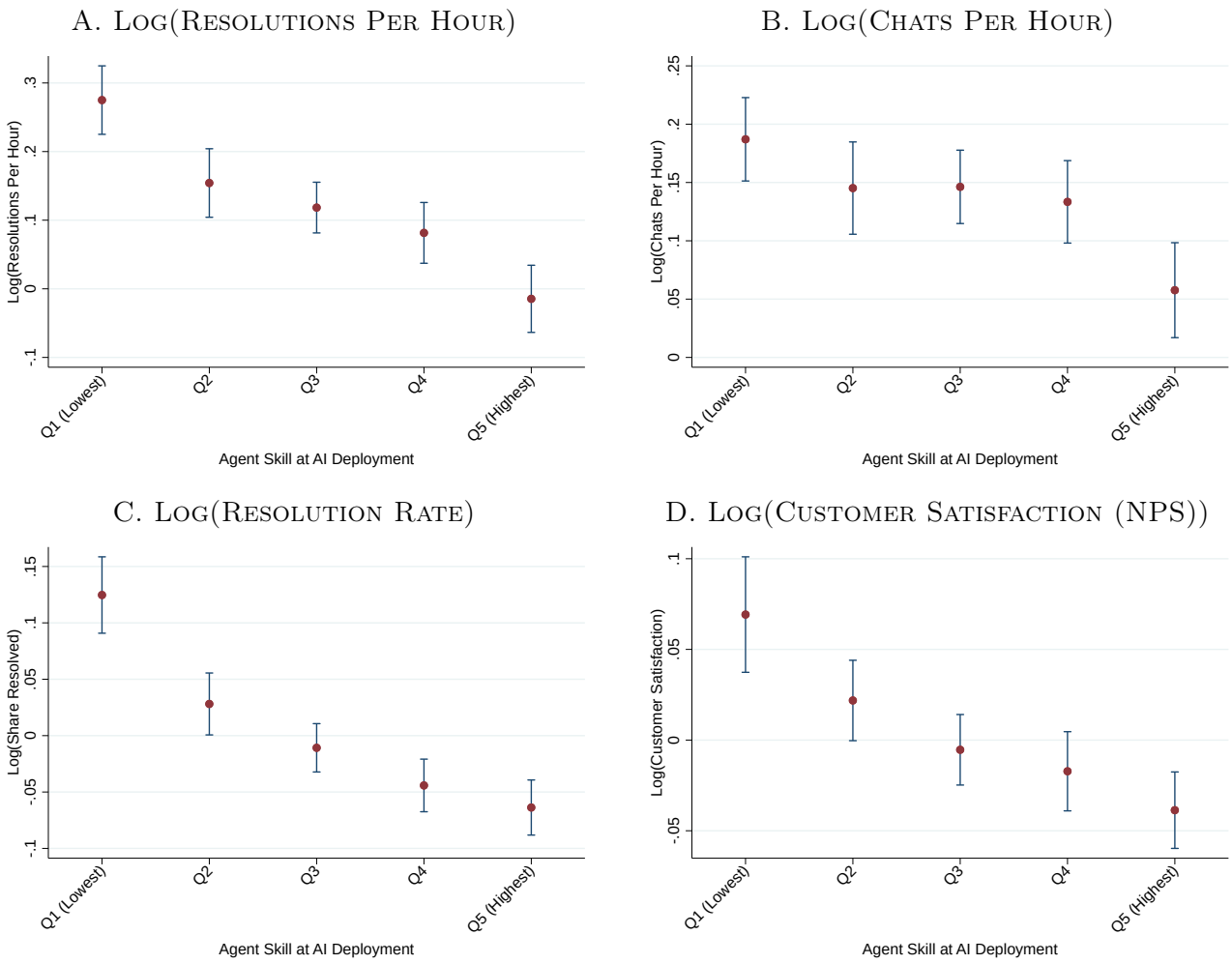


B. LOG(RESOLUTIONS PER HOUR)



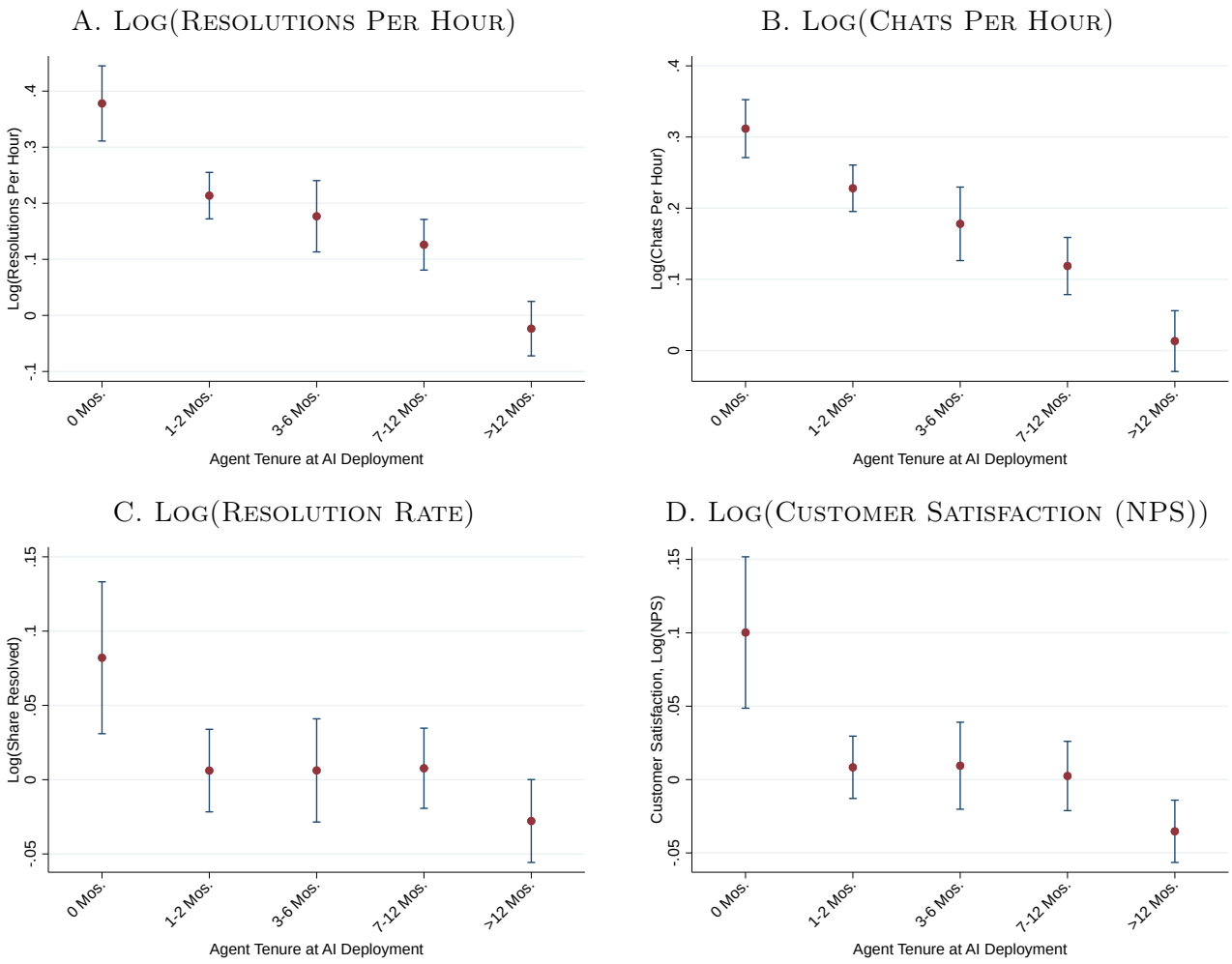
NOTES: This table presents the effect of AI model deployment on our main productivity outcome, resolutions per hour, using a variety of robust dynamic difference-in-differences estimators introduced in Borusyak et al. (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021) and a standard two-way fixed effects regression model. All regressions include agent level, chat-year fixed effects and controls for agent tenure. 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 preperiod using Borusyak et al. (2022) and de Chaisemartin and D'Haultfœuille (2020).

Figure A.3: Heterogeneity of AI Impact, by Skill and Controlling for Tenure



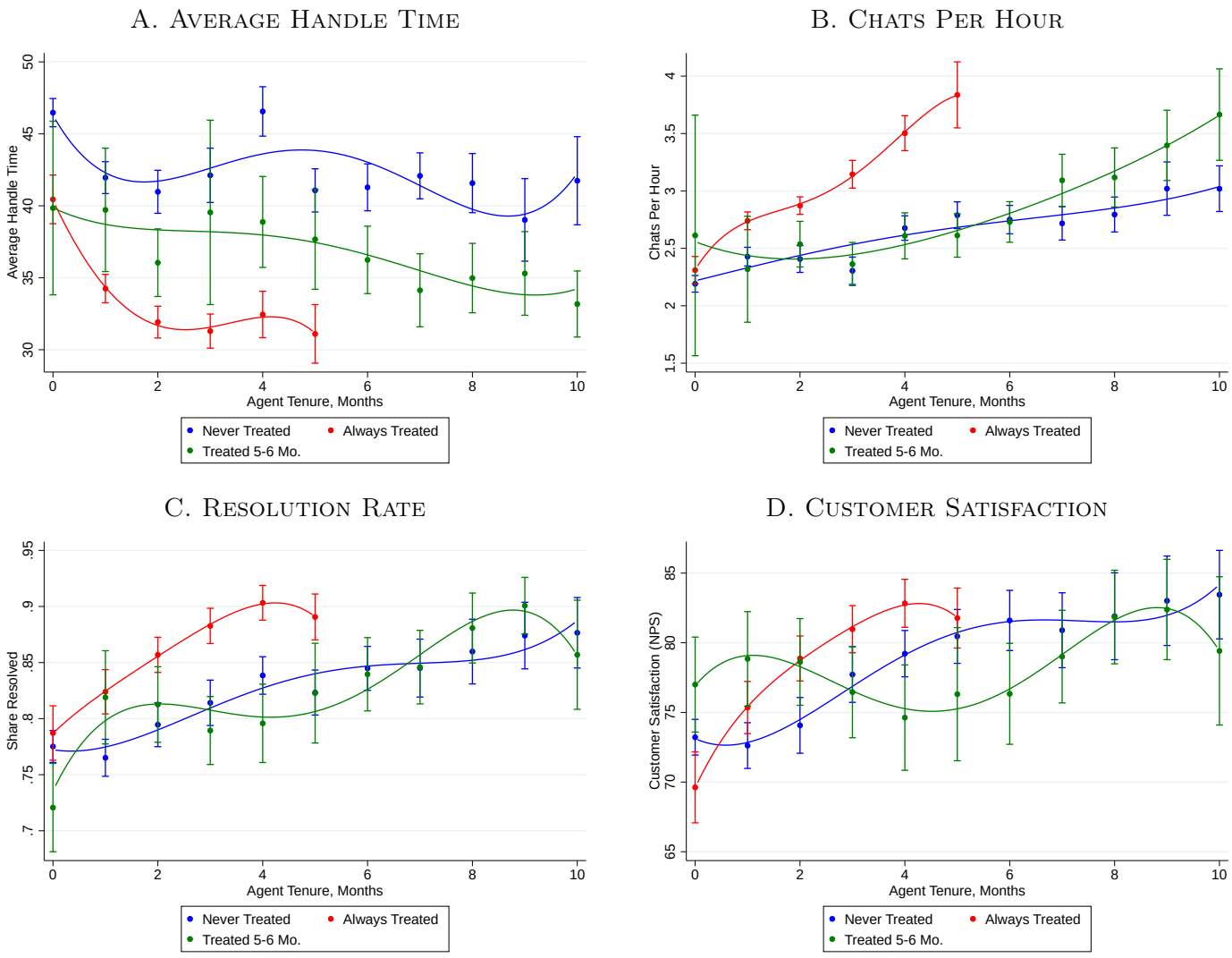
Notes: These figures plot the impacts of AI model deployment on worker productivity and other outcomes. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses for agent performance. 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. Pre-AI worker tenure is the number of months an agent has been employed when they receive access to AI recommendations. All specifications include agent and chat year-month, location, and company fixed effects and controls for agent tenure.

Figure A.4: Heterogeneity of AI Impact, by Tenure and Controlling for Skill



Notes: These figures plot the impacts of AI model deployment on 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 average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of calls an agent can handle per hour. Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, pre-AI agent skill and company fixed effects and standard errors are clustered at the agent location.

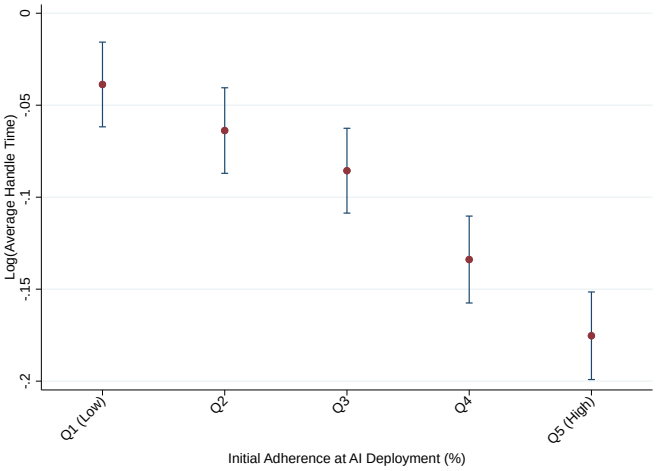
Figure A.5: Experience Curves by Deployment Cohort, Additional Outcomes



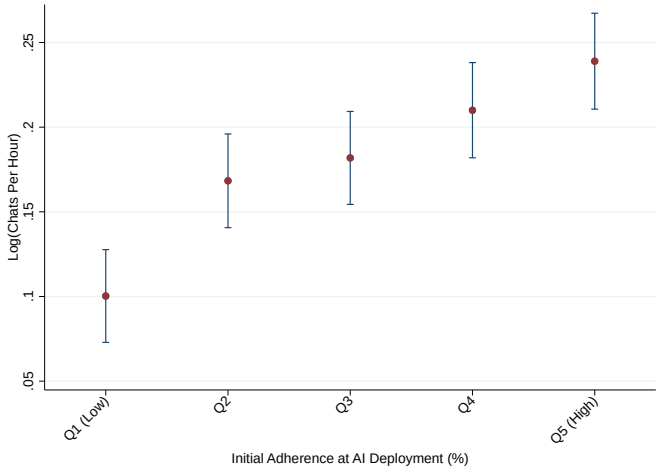
Notes: These figures plot the experience curves of three groups of agents over their tenure, the x-axis, against five measures of productivity and performance. The red lines plot the performance of always-treated agents, those who are start work in their first month with the AI and always have access to the AI suggestions. The blue line plots agents who are never treated. The green 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 95th percent confidence intervals.

Figure A.6: Heterogeneity of AI Impact by Initial AI Adherence, Additional Outcomes

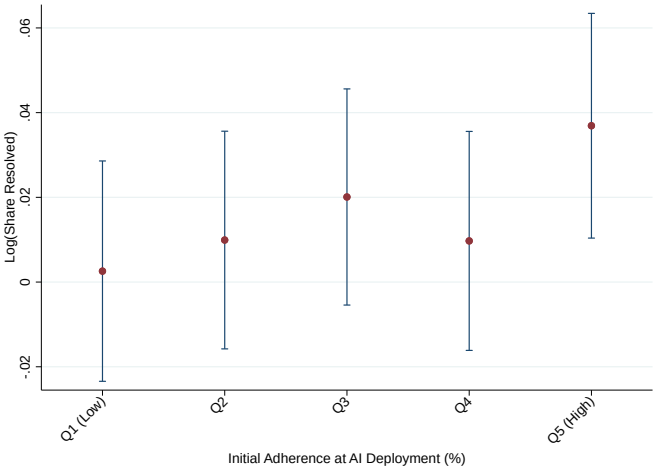
A. Log(Average Handle Time)



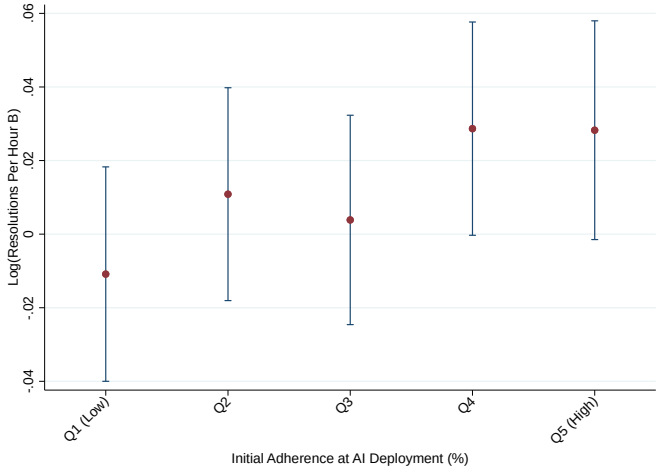
B. Log(Chats Per Hour)



C. Log(Resolution Rate)



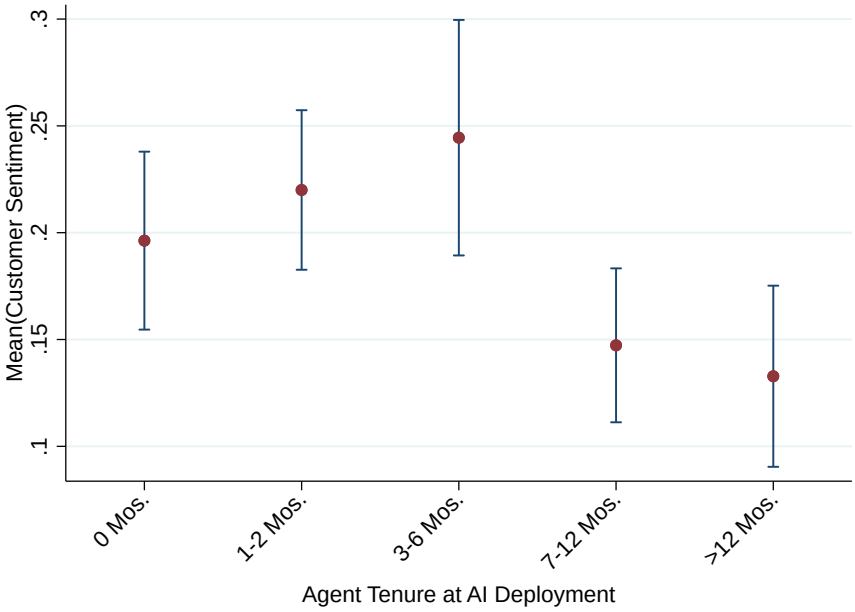
D. Log(Customer Satisfaction (NPS))



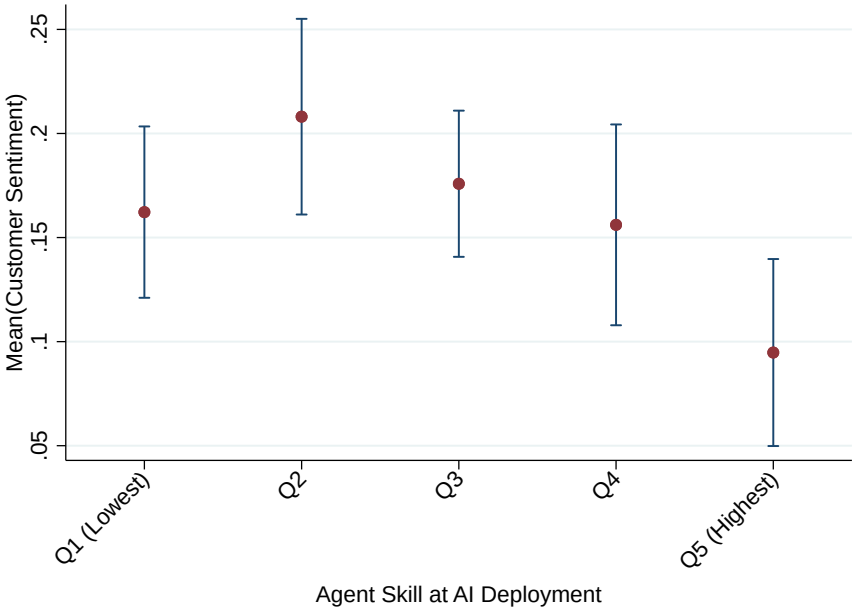
Notes: These figures plot the impact of AI model deployment on additional measures of performance by quintile of initial adherence, the share of AI recommendations followed in the first month of treatment. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of calls an agent can handle per hour (including working on multiple calls 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. All specifications include agent and chat year-month, location, and company fixed effects and controls for agent tenure. All data come from the firm's internal software systems.

Figure A.7: Heterogeneity in Customer Sentiment

A. By Tenure at AI Model Deployment

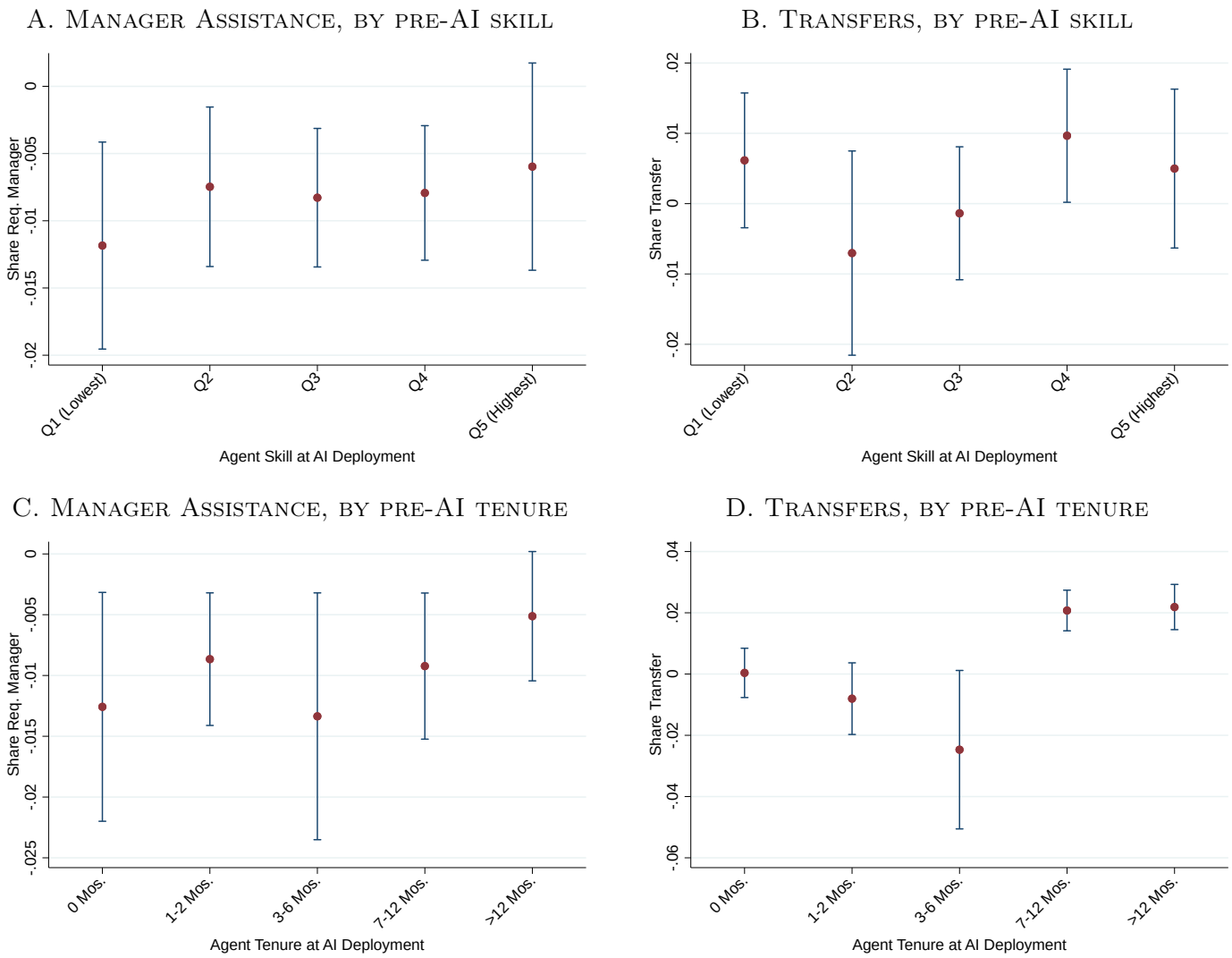


B. By Productivity at AI Model Deployment



Notes: Each panel of this figure plots the impact of AI model deployment on the mean sentiment per conversation. Sentiment refers to the emotion or attitude expressed in the text of the customer chat and ranges from -1 to 1 where -1 indicates very negative sentiment and 1 indicates very positive sentiment. Panel A plots the effects of AI model deployment on customer sentiment by agent tenure when AI deployed and Panel B plots the impacts by agent ex-ante productivity. All data come from the firm's internal software systems. Average sentiment is measured using SiEBERT, a fine-tuned checkpoint of a RoBERTA, an english language transformer model.

FIGURE A.8: ESCALATION AND TRANSFERS, HETEROGENEITY BY WORKER TENURE AND SKILL



NOTES: Panels A and C show the effects of AI on customer requests for manager assistance, by pre-AI agent skill and in by pre-AI agent tenure. Panels B and D show the impacts on transfers by pre-AI agent skill and pre-AI agent tenure. All robust standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

TABLE A.1: MAIN EFFECTS: PRODUCTIVITY (LOG(RESOLUTIONS PER HOUR)), ALTERNATIVE DIFFERENCE-IN-DIFFERENCE ESTIMATORS

	Point Estimate	Standard Error	Lower Bound 95% Confidence Interval	Upper Bound 95% Confidence Interval
TWFE-OLS	0.137	0.014	0.108	0.165
Borusyak-Jaravel-Spiess	0.257	0.028	0.203	0.311
Callaway-Sant'Anna	0.239	0.025	0.189	0.289
DeChaisemartin-D'Haultfoeuille	0.116	0.021	0.075	0.156
Sun-Abraham	0.237	0.037	0.165	0.308

NOTES: This table shows the impact of AI model deployment on the log of our main productivity outcome, resolutions per hour, using robust difference-in-differences estimators introduced in Borusyak et al. (2022), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020) and Sun and Abraham (2021). All regressions include agent level, chat-year fixed effects and controls for agent tenure. The standard errors are clustered at the agent level.