How AI-augmented Training Improves Worker Productivity *

Didier Fouarge[†] Marie-Christine Fregin[†] Simon Janssen[‡]
Mark Levels [†] Raymond Montizaan[†] Pelin Özgül[†]
Nicholas Rounding[†] Michael Stops[‡]

May 4, 2026

Abstract

We analyze the impact of AI-augmented training on worker productivity in a financial services company. The company introduced an AI tool that provides performance feedback on call center agents to guide their training. To estimate causal effects, we exploit the staggered roll out of the AI-tool. The AI-augmented training reduces call handling time by 10 percent. We find larger effects for short-tenured workers because they spend less time putting clients on hold. But the AI-augmented training also improves communication style with relatively stronger effects for long-tenured agents, and we find slightly positive effects on customer satisfaction.

Keywords: artificial intelligence, training, performance feedback, employee productivity

JEL Classification: J24, O31, O33

*This study is part of the research project ai:conomics funded by the German Federal Ministry of Labour and Social Affairs (Bundesministerium für Arbeit und Soziales (BMAS)/Denkfabrik Digitale Arbeitsgesellschaft) by resolution of the German Bundestag. We thank David Autor, Tommaso Ciarli, Christian Dustmann, Christina Gathmann, Christopher Stanton, Uschi Backes-Gellner, Thomas Cornelissen, Simon Wiederhold, Christina Felfe, Henning Hermes, and Terry Gregory for their helpful comments and suggestions. We are also grateful to the participants of LISER Data Science Lunch Seminar, Maastricht University Learning and Work Seminar, TASKS VII Conference, RFBerlin-CEPR Annual Symposium in Labour Economics, PSE-CEPR- Future of Work Policy Forum, ELMI 2024 Conference, ESPE 2025 Conference, Skills2Capabilities, ai:conomics summit Berlin and Wharton AI and the Future of Work Conference 2025, for their valuable feedback. We also thank Sanne Steens, Sander Dijksman, Danique Eijkenboom, Lara Fleck, Evie Graus and Luca Anastasiadis for their research support. We thank the anonymous company for giving us access to the data and allowing to interview employees.

[†]Research Center for Education and the Labor Market (ROA)

[‡]Institute for Employment Research (IAB)

I Introduction

Artificial intelligence (AI) has drawn significant attention for its potential to reshape economic activity. A growing body of research shows the productivity gains from AI, primarily in controlled laboratory settings[1] but increasingly in real-world environments (Brynjolfsson et al., 2025; Dillon et al., 2025; Dell'Acqua et al., 2025). Most of these studies focus on AI's direct productivity effects for workers who use general-purpose AI tools—such as ChatGPT and Microsoft Copilot—to support their work tasks. However, AI adoption may also generate indirect productivity effects that extend to workers who do not themselves use AI. For example, AI can improve managerial decision-making, feedback, and training processes, thereby raising the productivity of non-users through better guidance, training, and coordination within organizations (Luo et al., 2021; Riedl and Bogert, 2024).

Particularly, AI-augmented training can be valuable in sectors that handle large volumes of digital data, and face high worker turnover, such as the service industry. A recent survey[2] found that, among HR departments using AI, 43 percent apply it to learning and development. Moreover, companies like MetLife and Zoom are already using AI to improve training for their service agents. With their computational power, scalability, and cost-effectiveness, AI tools can process vast amounts of performance data and generate detailed, data-driven feedback tailored to individual workers' training needs (see, e.g., Council, 2019).

We study how introducing a labor-augmenting AI tool into the training system for customer contact agents impacts worker productivity. The intervention takes place at a major European financial services firm that manages over €500 million in assets and serves more than one million active customers. The company's setup is ideal for estimating causal effects. First, the company conducted a staggered introduction of the AI-augmented training system across geographically separated teams. Second, the company provides rich AI-generated data on worker productivity and call content for all agents in both the treatment and control groups, covering the entire observation period—including the time before the use of the AI-generated data in training. Third, the company requires the agents to participate in the training, thereby allowing us to go beyond intention-to-treat effects and to estimate treatment-on-the-treated effects. Fourth, the company assigned each team a trainer who exclusively trained that team throughout the observation period. Fifth, agents handle only incoming calls, which are always randomly assigned to them.

More specifically, each call center agent is assigned to one of five teams that are located in geographically separated regions. The agents receive one-on-one training sessions of 30

[1] Choi and Schwarcz (2024); Dell'Acqua et al. (2023); Fogliato et al. (2022); Freeman et al. (2024); Noy and Zhang (2023)

[2] (Society for Human Resource Management , SHRM)

to 60 minutes every two weeks from trainers who are each responsible for one of the five teams. Before using the AI-generated data throughout the training sessions, the trainers had simply selected three calls per agent for review and feedback. First in May and then in July 2023, the trainers from two teams received AI-generated performance data covering all of their agents' calls. While the overall training structure remained unchanged, this data enabled the trainers to provide more individualized and targeted feedback on the agents' productivity, such as handling time—but also quality aspects of the agent-client conversation such as communication style. We exploit this staggered introduction to estimate the causal effect of the AI-augmented training sessions on the agents' productivity, using Callaway and Sant'Anna (2021) event-study approach to account for dynamic and heterogeneous treatment effects.

Our data contains more than 180,000 calls of approximately 150 agents throughout an entire year. The data includes quantitative information on call quality (e.g., total handle time, speaking time, hold time, and call frequency) and detailed information about the agents' communication style, particularly their usage of diminutives, hedge words, and filler words. In addition, we have detailed information about call topics and customer satisfaction ratings.

We find that the AI-augmented training reduced agents' Average Handle Time (AHT) by approximately 60 seconds, corresponding to a marginal effect of about 9 percent relative to their pre-treatment average. The largest effects—driven by a reduction in very long calls, during which agents put clients on hold to seek help from coworkers or information from the company's information systems—occurred at the upper end of the handle time distribution. However, we also observe substantial reductions in effective speaking time, with effects occurring more uniformly across the distribution.

As in previous studies (Brynjolfsson et al., 2025), the productivity gains were significantly larger for short-tenured agents (17 percent) than for long-tenured ones (7 percent), suggesting that the AI-augmented training helped less-experienced workers move up the experience curve more rapidly. However, we also find meaningful effects for agents with longer tenure, although their improvements occurred along different margins of call quality. While long-tenured workers were more likely to reduce their effective speaking time by improving their communication style—using fewer filler words and hedge words—short-tenured workers showed greater improvements in handling previously difficult topics and reducing the duration of putting clients on hold.

We further find that productivity gains go hand in hand with improved customer satisfaction, although these effects are only marginally significant. Because the company's call center is an inbound center where agents cannot influence the frequency of incoming calls, we do not find a positive effect on the average number of handled calls per day. However, we observe large positive effects at the upper end of the distribution (at the 75th and 90th percentiles), suggesting that agents were able to handle more calls in periods

with a high number of incoming inquiries.

Our results are robust to the use of other estimators and whether we use a larger sample—including more experienced never-treated agents—or a smaller sample of only not-yet-treated agents, who are likely more comparable to the treated group. Moreover, a specification that relies exclusively on treated and control agents located in geographically distant regions suggests that our results do not suffer from a violation of the stable unit value treatment assumption (SUTVA).

This paper contributes to the large body of research analyzing the productivity effects of technology adoption. While many earlier studies examined the impact of previous IT technologies on worker or company productivity (Bartel et al., 2007; Acemoglu et al., 2007; Bloom et al., 2014), more recent studies focused on the effects of AI on worker performance. This research includes experiments with professionals performing tasks under controlled laboratory conditions (e.g., Peng et al., 2023; Noy and Zhang, 2023; Dell'Acqua et al., 2025; Cui et al., 2025; Agarwal et al., 2023; Choi and Schwarcz, 2024). Moreover, a smaller literature has evaluated the effects of AI adoption on worker productivity in real workplace settings, often exploiting staggered introductions of AI tools or field experiments within companies (e.g., Dillon et al., 2025; Brynjolfsson et al., 2025; Kanazawa et al., 2025; Otis et al., 2023; Luo et al., 2021).

We contribute to this literature in several ways. First by providing novel evidence on an important but less-studied channel through which AI affects worker productivity: indirect exposure via AI-augmented training. While most studies examine the direct effects of generative AI tools that workers use in their day-to-day tasks, our study shows that non-generative AI tools can improve productivity in a similarly substantial way, even when operating only behind the scenes to enhance training content. If workers do not directly use AI tools in their daily routines, as in our case, they are less likely to suffer from either the cyborg effect—i.e., blindly following AI-generated instructions—or from information overload (Dell'Acqua et al., 2025; Luo et al., 2021).

Second, we contribute to this literature by providing more nuanced evidence on how AI-augmented training affects worker productivity across levels of tenure. For example, Brynjolfsson et al. (2025) show that while AI particularly benefits younger and less able call center agents, it can slightly hamper performance among the most able workers who followed the AI tool's suggestions. Similarly, Peng et al. (2023) and Cui et al. (2025) find larger productivity gains for short-tenured and older programmers; Kanazawa et al. (2025) for lower-skilled taxi drivers; and Dell'Acqua et al. (2025) for lower-skilled consultants. In contrast, Otis et al. (2023) found that high-performing entrepreneurs benefited more from generative AI than their lower-performing peers.

A key contribution of our analysis is to show that short- and long-tenured agents benefit along different margins. Specifically, AI-augmented training helps less experienced agents by closing substantial knowledge gaps and reducing beginner mistakes. For

long-tenured agents, our results suggest that AI-augmented training improves their productivity by enhancing the detail and precision of their communication style. In this way, the AI-augmented training contributed to a reduction in productivity differences across workers.

Third, our study also relates to the literature on the productivity effects of management practices in general and the effects of worker training in particular. A large body of research has shown that targeted interventions—such as performance monitoring, personalized feedback, or structured training—can lead to substantial productivity gains. For example, Gosnell et al. (2020) show that performance monitoring and the provision of personalized target information significantly improved the productivity of airline pilots. Wheeler et al. (2022) shows that job readiness training for LinkedIn increases employment opportunities for job seekers, and Renée (2025) finds that career counseling and providing information improves the careers of students. Sauermann (2023) and Espinosa and Stanton (2022) document substantial productivity gains from on-the-job training, including spillover effects to coworkers and managers. We extend this literature by showing that training tailored to detailed, individual-specific performance measures can further improve productivity in a more scalable and targeted manner, highlighting the benefits of personalized, adaptive interventions over one-size-fits-all approaches.

The remaining paper is structured as follows. Section II describes the customer service department, and Section III describes the training program. Section IV explains the identification and empirical strategy. Section V presents the data and provides summary statistics. Section VI presents the results, and Section VII concludes.

II The customer service department

We evaluate the productivity effects of an AI-augmented training program in the customer service department of a large financial asset manager in Western Europe. In terms of structure and scale, the company is comparable with global peers such as CalPERS (U.S.), Canada Pension Plan Investment Board (CPPIB), and Japan's Government Pension Investment Fund (GPIF). The company generates revenue primarily by charging fees for managing assets above €540 Billion on behalf of more than 3 Million participating customers.

The call center and client service model operated by this company is representative of broader practices among major financial companies for two key reasons: First, like other large financial intermediaries (e.g., pension funds and insurers), the company serves millions of beneficiaries, requiring a centralized and professionalized customer contact infrastructure. The call center handles high volumes of complex inquiries. Second, like its peers across Europe, this company uses service-level benchmarks, call tracking, and customer satisfaction metrics aligned with both internal performance standards and

external supervisory expectations. The call center staff are well-trained public-facing professionals, covered by sectoral agreements, reflecting broader European employment practices in finance. The role of the call center is integrated into the broader administrative architecture.

During our observation period from February through December 2023, the department employed 147 agents, who handled a total of 187,839 calls—an average of 24,000 calls per agent per month, with an average call length of about 508 seconds. The agents' primary responsibility is to handle incoming customer calls involving questions, problems, or complaints related to the company's financial services. They also respond to emails and perform various administrative tasks. To ensure that customers receive and understand accurate information, the company requires that agents have a solid understanding of its services, efficient strategies for finding relevant information, and strong communication skills. As a result, the position requires a relatively high level of education—typically a vocational or a Bachelor's degree. Continuous changes to the rules governing the financial products offered by the company require agents to regularly update their knowledge to provide accurate and consistent information to clients. Workforce training therefore also plays a critical role. Thus the setting is ideal for the implementation of AI-augmented training. On the one hand, continuous training is essential for improving and maintaining agent productivity, making measures that reduce training costs or increase training effectiveness highly valuable to the company. On the other hand, the company records detailed data on all agent calls, providing a key condition for implementing AI-based solutions.

III The training program

Upon joining the department, agents complete a four-week introductory training course that prepares them to handle calls and provides an overview of the most common call topics and related services. Once active in the contact center, all agents regularly receive on-the-job training through one-on-one sessions with a personal trainer. These sessions last between 30 and 60 minutes and occur weekly to monthly, depending on the agent's needs. On average, agents receive training every other week, meaning that they all regularly participate in training. Only the most experienced agents do not participate in the standard training program, instead receiving non-standard training on handling specific challenges they face. Although these agents never participate in the AI-augmented training and are therefore *never treated*, we still have access to AI-generated performance data for all agents for the entire observation period.

The trainers are former agents, each assigned to a dedicated team of call center agents, who provide training exclusively within this team and tailor sessions to each agent's individual needs. The main purpose of the training is to improve agents' overall performance

by strengthening their communication skills, refining their problem-solving strategies, and deepening their knowledge of the company's products and processes.

III.A. *Pre-AI training program*

Prior to AI-augmented training, trainers did not have access to structured, computer-generated data on agent performance. Instead, they selected three calls per agent from the period (about a month) following the last training session. Using these calls, they (1) wrote a report highlighting effective or ineffective communication techniques, knowledge gaps, and other problems and (2) discussed these reports with their agents during the one-on-one training sessions.

Trainers selected calls either randomly or by call length. However, if they considered certain calls inappropriate or irrelevant for training, they selected others. This restrictive selection of training calls comes with significant shortcomings. The selected calls often failed to reflect agents' overall call quality, leading trainers to overlook important but less frequent weaknesses in agents' knowledge, performance, and communication style. In accompanying interviews we conducted with trainers and agents along with our quantitative analysis, both groups regularly reported that the selected calls were unrepresentative, resulting in poorly targeted training strategies. In some cases, the selections even caused disagreements between trainers and agents, further complicating the training process.

III.B. *Post-AI training program*

The company chose to improve training by implementing an AI tool that analyzes call center agents' past calls and comprehensively assesses their performance based on all calls. To examine all calls that each agent received, the AI tool relies on a supervised learning natural language processing (NLP) classification algorithm—a rule-constrained, non-generative NLP algorithm that processes call transcripts. The AI tool assigns predefined labels that reflect effective or ineffective communication techniques, the topic(s) of the call, and the duration of speaking, silence, and hold time. As the company defined all categories in advance, the AI tool did not identify patterns on its own (unlike, e.g., ChatGPT). Trainers access the AI-generated data through dashboards and reports in a proprietary software system. They then decide what to share with the agents, who never have direct access to the information.

The accompanying interviews revealed that both trainers and agents valued the ability of the AI tool to detect subtle quality issues that had previously gone unnoticed (**?**). They also appreciated its ability to confirm or challenge subjective assessments of agent performance, emphasizing the way it supported their feedback with detailed data. Moreover, when trainers showed agents the AI visualizations of their progress across a broader range of call quality dimensions, the agents welcomed this rich set of information. Thus

both trainers and agents generally agreed that the AI tool led to improvements in agent productivity.

IV Identification and estimation strategy

IV.A. *The staggered implementation of the AI-augmented training regime*

As previously mentioned, the introduction of the AI tool provides a unique setting for identifying the causal effect of the AI-augmented training on agents' productivity. Four factors make a causal assessment possible. First, the company introduced the AI-augmented training at the team level (five teams in total), with agents remaining in their assigned teams throughout the study period. As a result, they could not choose whether or when to enter the AI-augmented training program, thereby eliminating any self-selection bias and allowing us to go beyond the intention-to-treat effect by estimating the average treatment effect on the treated (ATT).

More specifically, the company incorporated AI-generated performance data into training through a staggered roll-out in May and July of 2023 (Figure 1), with the assignment of teams to treatment and control decided earlier, in February 2023. Moreover, the company stated that it determined the order of the staggered introduction without any specific goal or purpose in mind. Teams one and two received the treatment in May; teams three and four, in July. The agents in team five constitute the control group, which received no AI-augmented training throughout the observation period. However, after our observation period ended, even team five received the treatment in November 2023.

—Figure 1 about here—

Second, as calls are randomly distributed across agents, call quality is not mechanically related to the dynamics of the agents' productivity. Third, as each trainer is exclusively assigned to a specific team with the whole team consistently assigned to either the treatment or the control group throughout the observation period, spillover effects at the trainer level are unlikely. Fourth, as we have access to different types of control groups, and as teams of agents work in geographically separated regions, we can account for potential spillovers at the agent level. Specifically, agents in teams one and two, who received the May treatment, worked at the same location as the control group (team five). In contrast, the other treatment groups (teams three and four), who received the July treatment, were based at a different, distant location. Therefore, spillovers from the May to the July treatment are very unlikely, whereas spillover effects from the July treatment to the control group might be a concern.

To tackle this concern, we use various control groups available in our data. On the one hand, we rely on a group of very experienced agents who do not participate in the standard training program and are never treated. These agents are very proficient, with deep

knowledge about virtually all aspects of their work. Given that they had substantially smaller potential for improvement than less experienced agents, we can plausibly assume that these experienced agents were too proficient to experience substantial spillover effects. On the other hand, we used the not-yet-treated workers from the July treatment as a control group for the May treatment. As these agents worked in a separate location, they were unlikely to be affected by spillovers. As we show in Appendix A.1, our results remain stable regardless of which control group we use. Thus a violation of the stable unit treatment value assumption (SUTVA) is not likely to be a concern in our setting.

IV.B. *Estimation Strategy*

We exploit the staggered introduction of AI-augmented training by estimating the following event study model.

$$y_{it} = \sum_{k=-3}^{3} \mathbf{1}(t = t_{is} + k)\delta_k + \mathbf{x}_{it} + \lambda_t + \theta_i + \epsilon_{it} \tag{1}$$

where y_{it} denotes the outcome, capturing different measures of productivity for the agent i at time t. The coefficients δ_k are the main effects of interest and reflect the relative difference in outcomes between treated and non-treated agents. The term t_{is} denotes the period of the implementation of the AI-augmented training for the agent i. The set $k = -3, -2, ..., 2, 3$ defines the time periods spanning from three months before the treatment to four months after. The last entire month without treatment is denoted as -1, and the AI tool was rolled out in period 0. Thus Equation (1) allows us to both analyze dynamic treatment effects and assess potential nonlinear pre-treatment trends in outcomes. $\mathbf{x}_{it}$ denotes a set of control variables that include a set of dummy variables measuring the agents' pre-treatment tenure. The term λ_t denotes a set of time fixed effects, θ_i captures individual agent fixed effects, and e_{it} is a normally distributed error term with zero mean. To account for the potential correlation of error terms within agents, standard errors are clustered at the agent level.

The key identification assumption of Equation (1) is that the outcomes of the treatment and the control groups would have evolved in a parallel way in the absence of the treatment, i.e., the introduction of the AI-augmented training program. However, even if this assumption held, classic naive OLS estimates of the presented event studies may fail to recover the ATT if the treatment effects are heterogeneous and dynamic. Instead, these classic estimators yield a group and variance-weighted average of all possible combinations of treatment effects for which some weights might even be negative (Goodman-Bacon, 2021). Recent studies have provided a variety of approaches for overcoming this concern (Borusyak et al., 2024; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; De Chaisemartin and D'Haultfœuille, 2020).

Therefore, to carry out the majority of our estimates, we use the estimator proposed by Callaway and Sant'Anna (2021). This estimator chops the data into sub-samples of different valid 2x2 difference-in-differences (DiD) comparisons that only include valid combinations of treated, not yet treated, and never treated observations. The estimator then computes a weighted average of the DiD estimates, with weights based on the sample sizes of treated units in each treatment cohort. As a result, the Callaway and Sant'Anna (2021) estimator recovers the ATT even when treatment effects are heterogeneous and dynamic. Another key advantage of this estimator is its ability to account for time-constant differences in pre-treatment characteristics through a re-weighting approach. In addition to Callaway and Sant'Anna (2021), we also provide robustness checks using regular two-way fixed effects models and other common estimators by Borusyak et al. (2024), Sun and Abraham (2021), and De Chaisemartin and D'Haultfœuille (2020).

V Data and Summary Statistics

The company gave us access to a broad set of AI-generated performance data that trainers actively use in the AI-augmented training program.[3] A key advantage of our setting is that the firm used the AI tool to generate performance data on all calls—including those handled before the AI was implemented. We thus can track the same performance measures both before and after the performance data was used to train agents, and for all agents—regardless of whether they were ever treated.

The performance measures include quantitative indicators such as agents' call handle time, speaking time, silence time, and hold time. In addition, the data include customer satisfaction, call topics, and nuanced measures of agents' communication styles. All measures are recorded at the call level. Because agents may take leave hours or work irregular schedules, we follow Brynjolfsson et al. (2025) and aggregate the data at the monthly level to capture more complete work and performance histories.

Our main performance measure is the agent's Average Handle Time (AHT), which includes all parts of a call: (1) speaking time, reflecting communication efficiency; (2) silence time, signaling breakdowns in the conversation; and (3) hold time, when the agent searches for information or consults colleagues. Many comparable studies use AHT as a productivity measure (e.g., Brynjolfsson et al., 2025; De Grip and Sauermann, 2012; Liu and Batt, 2007), and AHT is also the company's most important key performance indicator. While shorter AHT enables agents to handle more calls per day, in our context, this benefit materializes only on busy days with high call volumes—because we study an inbound call center. More importantly, a shorter handle time gives agents more time to conduct their other responsibilities, such as responding to emails or handling

[3]For privacy reasons, the data is stored on a protected server. Upon request, we will be able to make the data available for replication upon request.

administrative duties.

The company's training explicitly aimed at reducing AHT below 600 seconds—both before and after the AI-generated data was introduced into the training process. Moreover, the company does not collect alternative performance measures, such as the resolution rate used in Brynjolfsson et al. (2025), because virtually all incoming client queries are eventually resolved either by the agent alone or with support from others.

Our measures for communication style are dummy variables indicating filler words, diminutives, and hedge words. Filler words such as "just" or "um" add redundancy and increase speaking time. Diminutives such as "sec" for "second" or "info" for "information" can lead to misunderstandings, while hedge words such as "maybe" or "I think" signal uncertainty and may undermine customer confidence. These indicators capture more nuanced aspects of call quality than standard metrics (e.g., AHT) or hold time. Even highly capable agents who resolve calls efficiently and rarely place clients on hold may show subtle communication flaws—gaps that can be filled with AI-augmented training.

Moreover, we have detailed AI-generated information on call topics and agents' communication style. The AI tool identifies the main topic of each call (e.g., financial issues related to retirement, a deceased person, or unemployment), allowing us to pinpoint specific areas where individual agents may find challenging or have knowledge deficits. For example, some agents may have long call handle times when discussing death-related issues, while others frequently place clients on hold when handling received pension payments.

In addition to these outcome measures, the data includes information on teams and their operating locations, the type of contract, the training program (i.e., standard or non-standard in case of experienced workers), and an indicator for tenure. Unfortunately, due to the company's strict data protection requirements, we are only allowed to use a measure of tenure that distinguishes among agents with 0–4 months, 5–12 months, and more than 12 months of tenure. Moreover, we have no information on the agent's education, gender, or other personal data.

Table 1 summarizes the available individual and monthly observations in the sample. Over the observation period, the agents handled more than 180,000 calls, and we observed a total of 147 agents: 84 are in the treated teams (teams 1 through 4), while 63 are in team 5, the control group. Overall, the data contains 957 person-month observations—440 in the treatment group and 517 in the control group. The last row shows that about 40 percent of the agents are experienced workers who were never treated because they were not part of the standard training program. All of these experienced workers were part of the regular control group in team 5.

—Table 1 about here—

Given their proficiency, these agents are on a flatter experience profile than short-

tenured agents, making them an ideal control group for capturing unobserved common trends, such as seasonal shifts in call volume. However, one potential concern is that our estimates might be biased precisely because these agents follow different trends than less-tenured agents in the treatment group. To eliminate this concern, most of our specifications account for agents' pre-treatment tenure through reweighting, and we also report specifications that rely exclusively on not-yet-treated agents as controls.

Table 2 presents summary statistics for all agents in the baseline period, i.e., one month before the treatment. The first three columns show results for the May treatment; the following columns for the July treatment. On average, agents in the treatment group have longer average handle time, speaking time, silence time, and hold time than those in the control group (Columns 1, 2, 4, and 5). They are also more likely to use filler words, diminutives, and hedge words. These differences are hardly surprising because the control group agents are more experienced than those in the treatment group. Although balancing pre-treatment characteristics is not required for identifying causal effects in a DiD framework, most of our specifications account for differences in agents' pre-treatment tenure through reweighting. Columns 3 and 6 show that applying the Callaway and Sant'Anna (2021) weights substantially improves balance across most characteristics.

—Table 2 about here—

VI Results

VI.A. *Main productivity effects on Average Handle Time (AHT)*

Figure 2 presents estimates from a classic two-way fixed effects event study based on Equation (1), where the outcome variable is our main productivity measure, the AHT of calls. The horizontal axis displays the months relative to the implementation of the AI-augmented training, while the vertical axis shows the conditional difference in AHT before and after implementation. The figure reports results from two specifications of the regression model. Gray triangles display the weekly estimates, with gray lines indicating their corresponding 95 percent confidence intervals. Black dots show the monthly estimates, with capped spikes representing their confidence intervals. Both models include agent and time fixed effects, and standard errors are clustered at the individual level.

—Figure 2 about here—

Our main identification assumption is that average outcomes for the treatment and control groups would have followed parallel trends in the absence of treatment. Although we cannot test this assumption directly, Figure 2 provides support for its validity. Between treated and non-treated agents, we do not observe any pre-treatment AHT differences

that are statistically different from zero at conventional significance levels. The weekly estimates in particular offer a more granular view of the pre-treatment period and—despite somewhat larger heterogeneity—show pre-treatment differences that are mostly close to zero.

In response to the treatment, a clear negative gap in AHT emerges between agents in the treatment and control groups—at both weekly and monthly levels. In the treatment month, the effect amounts to about −60 seconds, gradually increasing to about −120 seconds. The first column of Table 3 shows that the average effect of this specification is about −68 seconds, which corresponds to a marginal effect of −11 percent relative to the agents' pre-treatment AHT of 636 seconds.

—Table 3 about here—

As mentioned in Section IV.B., classic two-way fixed effects (TWFE) models may be biased because they rely on forbidden comparisons between already treated and not-yet-treated observations. To overcome this bias, Figure 3 presents estimation results based on the approach proposed by Callaway and Sant'Anna (2021) for each treatment separately. The estimation approach relies on both never treated and not-yet-treated agents as controls.

—Figure 3 about here—

Both the May and July treatments yield qualitatively similar effects to the naive event study in Figure 2. However, the TWFE version of the Callaway and Sant'Anna (2021) estimator, reported in Column 2 of Table 3, shows a slightly larger negative average effect (approximately −92 seconds, or −14 percent).[4]

Unlike those in the treatment group, many agents in the control group are, on average, more experienced and, because of this, did not participate in the training program during our observation period (though they did participate in regular training earlier in their careers). Comparing these long-tenured agents (who are likely to be on a flat learning curve) to the less-tenured agents in the treatment group (who are likely to be on a steeper learning curve) might bias the coefficient estimate on AHT downward. Therefore, Column 3 of Table 3 controls for pre-treatment tenure and yields an estimated effect of about −60 seconds, or −9 percent. Overall, the specification of Column (3) is our preferred specification, yielding an effect size that closely matches that of Brynjolfsson et al. (2025), who find that the direct introduction of AI assistance in customer service reduces agents' chat time by 8.5 percent. Therefore, our results indicate that using AI

[4]OLS event study models with event-time dummies are often more similar to the estimator proposed by Callaway and Sant'Anna (2021) than the TWFE specification with a single post-treatment indicator. Event-time dummies allow treatment effects to vary over time and can partially mitigate bias from inappropriate comparisons across cohorts at different stages of treatment.

insights to inform training—rather than intervening directly in real time—may be an equally effective approach to improving worker productivity.

Specifications 4 throughout 6 provide further robustness checks by excluding experienced never-treated agents who did not participate in the standard training program (Columns 4 and 6), and by excluding very short-tenured agents who were not present throughout the entire observation period (Columns 5 and 6).[5] All three specifications lead to quantitatively similar effects.Appendix A.1 analyzes a potential violation of SUTVA by restricting the sample to the treatment group of the May treatment and using only the not-yet-treated agents of the July treatment as controls. This specification is unlikely to suffer from spillover effects because the affected teams are located in distant sites where interaction between agents hardly exists. Appendix A.1 comfortingly shows that the results do not change. Appendix A.3 shows that the most commonly used alternative approaches for correcting for bias in classic TWFE models yield virtually the same results.

VI.B. *Distributional effects at different productivity margins*

A large literature has analyzed the distributional effects of modern technologies, particularly IT (Autor et al., 1998, 2003; Goos et al., 2014). That literature shows that while earlier technologies such as IT primarily replaced the routine tasks of less skilled workers, they complemented the non-routine tasks of more skilled workers. In contrast, the main advantage of AI lies in its ability to process large volumes of data to generate useful information on individual problems. In our case, the AI tool analyzed massive amounts of call data to deliver highly individualized performance feedback and to identify personal gaps in process and product knowledge. Consistent with this idea, interviews with both agents and trainers highlight two main channels through which the AI tool may have improved agents' productivity. First, the AI-generated feedback may have helped agents to refine their communication style. Second, it may have supported trainers in more accurately identifying topic-specific knowledge gaps during training sessions.

In the following, we analyze the productivity effects of AI-augmented training across the entire outcome distribution and along three additional productivity margins—speaking time, hold time, and silence time. To analyze distributional effects, we transform the dependent variables using re-centered influence functions (RIF), as proposed by Firpo et al. (2009). In contrast to raw quantiles, RIFs are linear in expectation and can therefore be used as dependent variables in a regression framework analogous to the Callaway and Sant'Anna (2021) approach. The resulting coefficient estimates capture unconditional distributional effects, indicating whether the outcome distribution has shifted at different percentiles in response to the treatment. All regressions account for tenure, as well as

[5]Excluding these agents restricts the sample to only those agents with at least three months of tenure.

individual agent and time fixed effects. Figure 4 presents the results.

—Figure 4 about here—

The upper panel on the left side of the figure analyzes AHT—our main outcome measure. The y-axis shows the effect sizes of the coefficient estimates, as estimated using the Callaway and Sant'Anna (2021) version of Equation (1). The first coefficient estimate (black dot and dashed line) presents the average ATT, as already reported in Column 3 of Table 3.[6] The further coefficient estimates (gray markers) present the results for the 10^{th} throughout the 90^{th} percentiles.

According to the upper left panel of Figure 4 the AI-augmented training has a substantially larger impact at the upper tail of the AHT distribution than at the lower tail. While we find no effect at the lowest percentile, we observe large reductions of more than 160 seconds at the upper end of the distribution. Although the effects in the middle of the distribution are non-negligible, the AI-augmented training appears to have primarily reduced the occurrence of person-month observations with a very long AHT. These calls are most likely handled by agents who struggle throughout the conversation due to substantial knowledge gaps. In contrast, shorter calls are more likely handled by more skilled agents, who have less room for improvement.

The upper panel on the right side of Figure 4 analyzes agents' effective speaking time, which reflects how efficiently agents communicate with clients and is therefore likely related to their communication style. Yet longer speaking time may also indicate a lack of topic-specific knowledge, because agents might need more time to clarify issues in the dialogue with the clients. The average ATT on speaking time is around 32 seconds (black dot and dashed line)—half the size of the average effect on AHT. In contrast to the AHT results, the effects on speaking time are more uniform across the distribution. Although we do not find any effect at the 10^{th} percentile, the remaining effects cluster around -50 seconds, with only a slight tendency for stronger effects at the upper end of the distribution. Thus the AI-augmented training program shifts almost the entire speaking time distribution to the left, i.e., a broader improvement in communication efficiency, not only among the lowest-performing agents.

The lower panel on the left analyzes agents' hold time, which captures the time they put clients on hold while requesting help from coworkers or searching for additional information. Thus hold time mostly reflects gaps in their product or procedural knowledge. The lower panel on the right examines silence time during calls, indicating a complete breakdown in the conversation between agent and client, and likely reflecting beginner mistakes. The average effect on hold time amounts to around -23 seconds, thereby explaining a large fraction of the overall effect on AHT. However, a substantial reduction

[6]The same applies to the further panels of Figure 4. Appendix A.4 presents results for the average effects for the other outcomes, each based on a specification that excludes all never-treated workers from the control group.

in outliers with very long average hold times in a given month appears to be driving this effect.

The ATT on silence time, around −6 seconds, is not statistically significant at conventional levels, contributing very little to the overall effect on AHT. However, as with hold time, we find stronger and statistically significant effects in the upper tail of the distribution, suggesting that outlier conversations marked by complete breakdowns are driving this result.

Overall, the results suggest that the AI-augmented training improved agent productivity through two main channels. The first entails shortening excessively long calls—primarily by reducing hold and silence time, likely among agents with knowledge gaps. The second entails increasing communication efficiency, as reflected in reduced speaking time, likely affecting both less and more able workers. However, in contrast to earlier waves of IT technology, the AI-augmented training appears to benefit agents at the lower end of the experience distribution more than those at the upper end.

VI.C. *Heterogeneous Treatment Effects by Tenure*

As previously mentioned, the impact of technological change on workers with different skill levels is of central interest. As short-tenured workers have had less time to accumulate skills and human capital than their long-tenured peers, tenure is a useful indicator—available in our data—for analyzing the heterogeneous impact of AI along the skill distribution. Moreover, studies have shown that generative AI can help move short-tenured workers up the experience curve, while its effect on longer-tenured workers has been much weaker in comparison (see, e.g., Brynjolfsson et al., 2025)

—Table 4 about here—

Table 4 presents the productivity results for agents in different tenure categories. The first column reports the results for agents with fewer than four months of tenure, while the second column shows the results for those with more than four months of tenure.[7] While the new training reduced the AHT of short-tenured agents by approximately 131 seconds (−17 percent relative to their average pre-treatment AHT), it reduced the AHT of long-tenured agents by only approximately 41 seconds (7 percent). Therefore, the AI-augmented training had a substantially larger effect on short-tenured agents than on long-tenured ones. We find a similar tenure gap in agents' speaking time (15 percent for short-tenured vs. 9 percent for long-tenured agents). The tenure gaps for hold time (19 percent vs. 5 percent) and silence time (17 percent vs. 6 percent) are even more pronounced, and for long-tenured agents, the effects are not statistically significant at conventional levels.

[7] Our limited number of observations prevents us from providing a more granular investigation of the tenure effects.

Thus, the largest relative difference between short- and long-tenured agents appears for hold time. This result is consistent with the idea that less-skilled workers, who are more likely to have knowledge gaps, need to ask trainers or coworkers for help, whereas more skilled (i.e., longer-tenured) workers may benefit relatively more from nuanced improvements. Such improvements might include communication style refinements, which could be reflected in reduced speaking time.

VI.D. *Knowledge gaps and communication style*

This subsection analyzes how the AI-augmented training affects agents differently, depending on their experience level, by focusing on differences in knowledge gaps and communication style. Table 5 analyzes how the AI-augmented training influences the agents' communication style. To do so, we rely on three dummy variables measuring whether the agents have used filler words, diminutives, or hedge words in their communication (see Section V for details). /stAll of these variables indicate nuances of poor communication styles.

—Table 5 about here—

The first row of Table 5 shows how the AI-augmented training affected agents' use of filler words. Although we find no statistically significant effects for short-tenured agents, we find large effects for long-tenured ones—a 5 percentage points (9 percent) reduction in the use of filler words. Similarly, in the second row, we find a 3 percentage points (10 percent) reduction in the use of diminutives among long-tenured agents, and, in the third row, a significant reduction in the use of hedge words; roughly 4 percentage points (10percent). All other coefficient estimates in the first three rows are statistically insignificant at conventional levels.

However, given that effect sizes and pre-treatment averages are very similar for short- and long-tenured agents, we cannot definitely conclude from these results that the communication style of long-tenured agents did not improve in response to the AI-augmented training. The absence of statistical significance among short-tenured agents might instead reflect greater effect heterogeneity or limited statistical power within this smaller sample.

Thus to further explore the effects of the AI-augmented training on agents' communication style, the last row of Table 5 analyzes a version of speaking time—residualized on the communication style dummies—as the dependent variable.[8] This outcome captures speaking time variation that is unrelated to our communication style indicators. After residualization, while the treatment effect on speaking time decreases by a third

[8]In Appendix A.5 we show that poor communication styles significantly increase agents' speaking time, while being much less correlated with hold or silence time. Besides, we also show a negative effect on customer satisfaction.

for short-tenured agents and is statistically significant at conventional levels, it drops by more than 50 percent for long-tenured agents, with the coefficient turning insignificant at conventional levels (see Table 4 for comparison).

These results suggest that changes in observable communication style indicators largely drive the reduction in speaking time among long-tenured agents. In contrast, the smaller decline for short-tenured agents after residualization indicates that their improvements in speaking time are less closely tied to the specific communication style measures we capture—possibly reflecting broader gains such as generally more efficient calls or faster recall of product knowledge.

VI.E. *Call topics*

In the following, we analyze whether the AI-augmented training specifically improves agents' performance—particularly that of short-tenured ones—on topics which they had previously found challenging. Since all calls are randomly assigned, agents must handle a wide range of customer inquiries, some of which are more technically complex or emotionally demanding than others. For example, agents may need to speak with a client's relatives about the client's death or discuss a client's recent inability to work. We therefore identify these topics discussed during the calls. Table 6, which presents the key characteristics of each topic, shows that inquiries about unemployment are, on average, associated with significantly longer durations across all call dimensions—from AHT to silence time. In contrast, inquiries about pension payments tend to take much less time across nearly all dimensions of the call.

—Table 6 about here—

Moreover, certain topics occur much more frequently, allowing agents to acquire more knowledge through on-the-job experience, whereas other topics arise so rarely that agents have little opportunity to learn about them. For example, inquiries related to *unemployment* appear in only 1 percent of calls, whereas inquiries about *retirement* and *pensions received* occur in 31 percent of them. This may be also reflected in the average handle times: an average call on *unemployment* reveals the longest handle time of 898 seconds whereas a call on *pensions received* needs only 487 seconds.

To identify topics that agents found challenging before the AI-augmented training, trainers had to rely on selected calls or agent feedback. After the AI augmentation, trainers could more efficiently find calls related to those topics by filtering dashboards and reports. Using this information, a trainer could then discuss with each agent why a particular topic might be challenging and develop a personalized development plan for mitigating the problem. Therefore, this AI-augmented information now allows trainers to more precisely target problematic topics, discuss them with the agent, and tailor support accordingly.

We now analyze how the AI-augmented training improved call performance on topics agents had previously found challenging and examine how these effects vary by agent tenure. To do so, we leverage the more granular call-level data by first residualizing pre-treatment call duration for each call, controlling for tenure, week, day of the week, time of day, and agent fixed effects. This approach removes unobserved factors related to both agents' average ability and detailed temporal dynamics. We then compute the average pre-treatment residual per topic for each agent and identify the top two topics with the highest average residuals as the agent's most challenging topics.

Panel A in Table 7 presents the estimates based on Equation (1) for AHT and other outcomes, using only calls related to each agent's most challenging topics (Panel A), compared to all other topics (Panel B). The first four rows of Panels A and B report the results for short-tenured agents; the second four rows, for long-tenured agents.

—Table 7 about here—

These results suggest that AI-augmented training is particularly effective at narrowing knowledge gaps, especially for short-tenured agents and in topics they found challenging earlier. The sharper reductions in AHT and improvements in call components for these agents indicate that the training helped close basic skill gaps. However, we do not observe that long-tenured agents benefited from AI-augmented training.

VI.F. *Calls per day and customer satisfaction*

Although AHT is the company's most important productivity measure, other dimensions of agent performance exist. On one hand, a shorter call handle time allows agents to manage more calls per day, thereby boosting overall productivity. On the other hand, the AI-augmented training might have increased pressure on them to shorten interactions at the expense of service quality. For a more comprehensive assessment, this subsection examines two additional outcome measures: the number of calls handled per day and customer satisfaction scores.

The first row of Table 8 presents coefficient estimates of the effect on the average number of calls handled per day. The first column of this first row reports the effect on the mean, while the next five columns show the effects at different percentiles of the distribution. The last two columns show the effect on the average number of calls separated for short- and long-tenured agents.

Although the average effects, for all agents and separated for short- and long-tenured agents, are not statistically different from zero (Columns 1, 7 and 8), the table reveals strong positive effects at the upper end of the distribution (Columns 5 and 6).[9] These

[9]The low average effect may be attributable to the weakly significant, negative effect for the lower end of the distribution. Evidently, after the training, agents with an extremely low number of calls had an even lower number.

results suggest that the AI-augmented training shifted the upper tail of the distribution—indicating that a few agents were able to handle exceptionally high call volumes during certain months—rather than affecting typical or median performance. Given that the company operates as an inbound call center and that agents cannot influence the number of incoming calls, these gains likely reflect improvements in efficiency or call handling speed, rather than changes in call availability. These results suggest that the training enabled some agents to handle more calls during particularly busy days or months while potentially freeing time for other responsibilities only during periods of lower demand.

—Table 8 about here—

The second row of Table 8 presents the results on customer satisfaction. Larger values of customer satisfaction indicate larger satisfaction. Because clients often decline to rate call agents—and are more likely to do so only when they are very satisfied or dissatisfied—only about 7 percent of all calls receive customer satisfaction ratings. Such a low response rate is common in inbound call centers. Moreover, as most agents in our data receive at least one customer satisfaction rating per month, we lose only a small number of person-month observations in our sample.

On average, customer satisfaction increased by about 0.4 points (corresponding to about 5 percent). The effect is only marginally significant at the 10 percent level (Column 1), but it is large and statistically stronger in the upper tail of the distribution, suggesting that agents received more exceptionally positive ratings.

This effect is mainly driven by calls that short-tenured agents handled; for these agents, the customer satisfaction significantly increased after the AI-augmented training (Column 7), whereas the effect on customer satisfaction related to calls that long-tenured agents handled revealed no significant changes.

VII Conclusion

This paper provides causal evidence on the productivity effects of AI-augmented training in the real-world setting of a customer service provider. We evaluate the introduction of an AI tool that delivered individual-level performance feedback to support the training of customer contact agents. Exploiting the staggered roll-out of this tool, we show that AI-augmented training reduced call handling time by approximately 10 percent. While the particularly large gains were for short-tenured agents, who improved by closing knowledge gaps and reducing avoidable mistakes, more experienced agents also benefited from the training through improvements in communication style and efficiency. Importantly, we find that the gains did not come at the expense of service quality: indeed, customer satisfaction ratings slightly improved, and agents were able to handle more calls on particularly busy months or days.

Our results contribute to a growing literature on the labor market effects of AI adoption, emphasizing an important but understudied channel: the role of AI in augmenting training, as opposed to augmenting only day-to-day task execution. While recent studies focus primarily on direct interactions between workers and AI tools, our results show that non-generative AI systems operating in the background can also meaningfully enhance performance, particularly when used for supporting targeted and individualized training.

Our results also speak to broader questions about the distributional consequences of AI. Although short-tenured agents benefit most in absolute terms, we find that both less and more experienced workers improve, albeit along different dimensions, i.e., more efficient call handling for short-tenured workers especially, and improved communication style for long-tenured workers. In this way, AI-augmented training may reduce productivity gaps across workers, rather than widening them.

Beyond our specific setting, the mechanisms we identify are likely transferable to other data-rich environments where performance data can inform personalized training. These environments include manufacturing, where machine-level data and incident reports support operator learning; education, where learning analytics help teachers tailor instruction; healthcare, where clinician-patient interactions inform communication training; and logistics or retail, where process data can optimize workflows and service. As key drivers of impact—structured feedback, targeted skill development, and improved learning efficiency—are common across sectors, AI tools can also improve training systems. Gains might be particularly strong for low-skilled or short-tenured workers in high-turnover roles, where faster onboarding matters most.

References

Acemoglu, Daron, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti, "Technology, Information, and the Decentralization of the Firm," *The Quarterly Journal of Economics*, 11 2007, *122* (4), 1759–1799.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Technical Report 31422, National Bureau of Economic Research 7 2023.

Autor, D. H., F. Levy, and R. J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 11 2003, *118*, 1279–1333.

_ , L. F. Katz, and A. B. Krueger, "Computing Inequality: Have Computers Changed the Labor Market?," *The Quarterly Journal of Economics*, 11 1998, *113*, 1169–1213.

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.

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

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," *The Review of Economic Studies*, 02 2024, *91* (6), 3253–3285.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 02 2025, *140* (2), 889–942.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1.

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

Choi, Jonathan H. and Daniel Schwarcz, "AI Assistance in Legal Analysis: An Empirical Study," *Journal of Legal Education*, 2024, *73*. Forthcoming. Available at SSRN.

Council, Jared, "MetLife Says AI Is Improving Its Call Centers," *The Wall Street Journal, WSJ PRO*, June 2019.

Cui, Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," Technical Report, SSRN 8 2025.

Dell'Acqua, Fabrizio, Charles Ayoubi, Hila Lifshitz, Raffaella Sadun, Ethan Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stewart Taub, and Karim Lakhani, "The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise," Working Paper 33641, National Bureau of Economic Research 4 2025. NBER Working Paper No. 33641.

_ , Edward III McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R.

Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Technical Report Working Paper No. 24-013, Harvard Business School Technology & Operations Management Unit September 2023. The Wharton School Research Paper, also available at SSRN.

Dillon, Eleanor W., Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton, "Shifting Work Patterns with Generative AI," Working Paper 33795, National Bureau of Economic Research 5 2025. NBER Working Paper No. 33795.

Espinosa, Miguel and Christopher T. Stanton, "Training, Communications Patterns, and Spillovers Inside Organizations," Working Paper 30224, National Bureau of Economic Research 7 2022. NBER Working Paper No. 30224.

Firpo, Sergio, Nicole Fortin, and Thomas Lemieux, "Unconditional Quantile Regressions," *Econometrica*, 2009, *77* (3), 953–973.

Fogliato, Riccardo, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi, "Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging," in "Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)" Association for Computing Machinery New York, NY, USA 2022, pp. 1362–1374.

Freeman, Brian S., Kendall Arriola, Dan Cottell, Emmett Lawlor, Matt Erdman, Trevor Sutherland, and Brian Wells, "Evaluation of Task Specific Productivity Improvements Using a Generative AI Personal Assistant Tool," *arXiv*, 2024. arXiv:2409.14511 [cs.HC].

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Goos, Maarten, Alan Manning, and Anna Salomons, "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring," *American Economic Review*, 2014, *104* (8), 2509–2526.

Gosnell, Greer K., John A. List, and Robert D. Metcalfe, "The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains," *Journal of Political Economy*, 2020, *128* (4), 1195–1233.

Grip, Andries De and Jan Sauermann, "The Effects of Training on Own and Co-Worker Productivity: Evidence from a Field Experiment," *The Economic Journal*, 2012, *122* (560), 376–399.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, Skill, and Productivity: The Case of Taxi Drivers," *Management Science*, 2025. Published online: June 9, 2025.

Liu, Xiangmin and Rosemary Batt, "The Economic Pay-Offs to Informal Training: Evidence from Routine Service Work," *ILR Review*, 2007, *61* (1), 75–89.

Luo, Xueming, Marco Shaojun Qin, Zheng Fang, and Zhe Qu, "Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions," *Journal of Marketing*, 2021, *85* (2), 14–32.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative AI," *Science*, 2023, *381* (6654), 187–192.

Otis, Nicholas G., Rowan P. Clarke, Solene Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance," Technical Report, OSF Preprints 12 2023. Preprint.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," *arXiv preprint*, 2023. arXiv:2302.06590 [cs.SE].

Renée, Laetitia, "The Long-Term Effects of Career Guidance in High School and Student Financial Aid: Evidence from a Randomized Experiment," *American Economic Journal: Applied Economics*, 2025, *17* (2), 165–183.

Riedl, Christoph and Eric Bogert, "Effects of AI Feedback on Learning, the Skill Gap, and Intellectual Diversity," Technical Report 2024.

Sauermann, Jan, "Performance Measures and Worker Productivity," *IZA World of Labor*, 2023, pp. 1–12.

Society for Human Resource Management (SHRM), "From Adoption to Empowerment: Shaping the AI-Driven Workforce of Tomorrow," 2025. Accessed: 2025-10-23.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Wheeler, Laurel, Robert Garlick, Eric Johnson, Patrick Shaw, and Marissa Gargano, "LinkedIn(to) Job Opportunities: Experimental Evidence from Job Readiness Training," *American Economic Journal: Applied Economics*, 2022, *14* (2), 101–125.

Figures in the Text

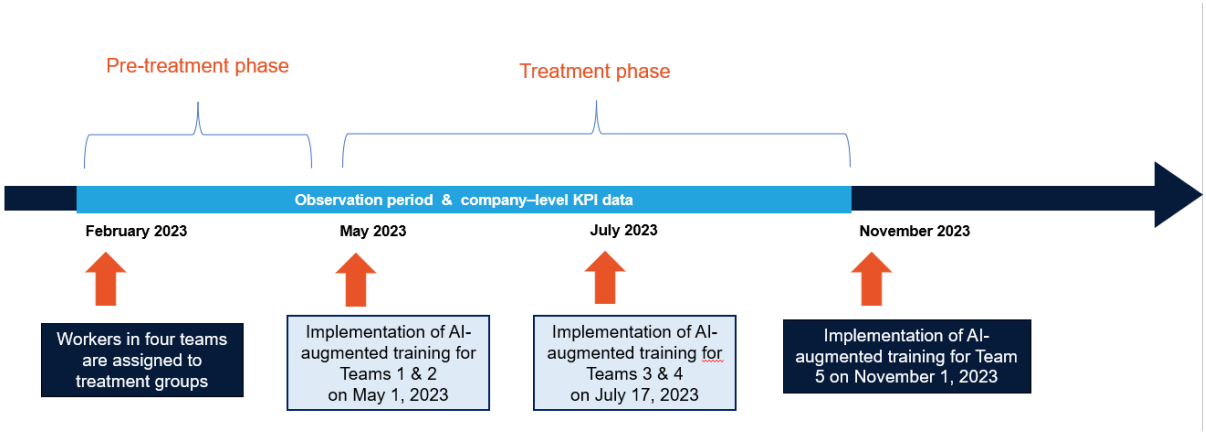

Figure 1: Rollout of AI-augmented training across teams

Notes: This figure illustrates the timeline of the field experiment. The observation period spans from week 5/2023 through week 44/2023, during which company-level KPI data was collected. The experiment was implemented in a staggered manner between week 18/2023 and week 44/2023. AI-augmented training was introduced for Teams 1 and 2 on May 1, 2023, followed by Teams 3 and 4 on July 17, 2023. Team 5 received AI-augmented training on November 1, 2023. Only after the technology was introduced in their respective teams did the treated agents begin receiving AI-augmented training.

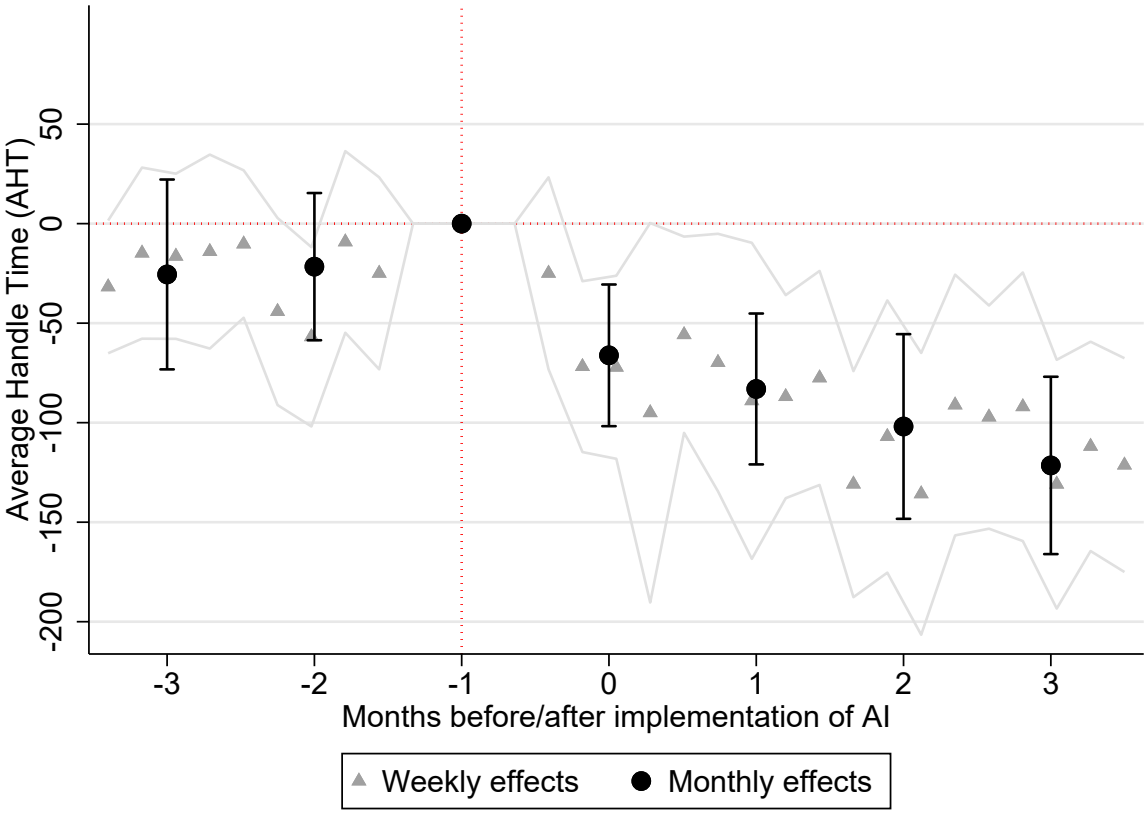

Figure 2: Event study estimates on average call handle time

Notes: The figure shows event study estimates of the AI-augmented training on workers' average call handle time (AHT). The x-axis represents the months relative to the implementation of the AI, while the y-axis shows the conditional gap in AHT before and after the implementation. The gray triangles represent weekly coefficient estimates, with gray lines showing their corresponding 95 percent confidence intervals. The black dots represent monthly coefficient estimates accompanied by capped spikes illustrating the estimates' 95 percent confidence intervals. All models include agent and month fixed effects. Standard errors are clustered at the agent level. The monthly effects are estimated based on 957 employee-month observations; the weekly effects, on 3,348 employee-week observations.

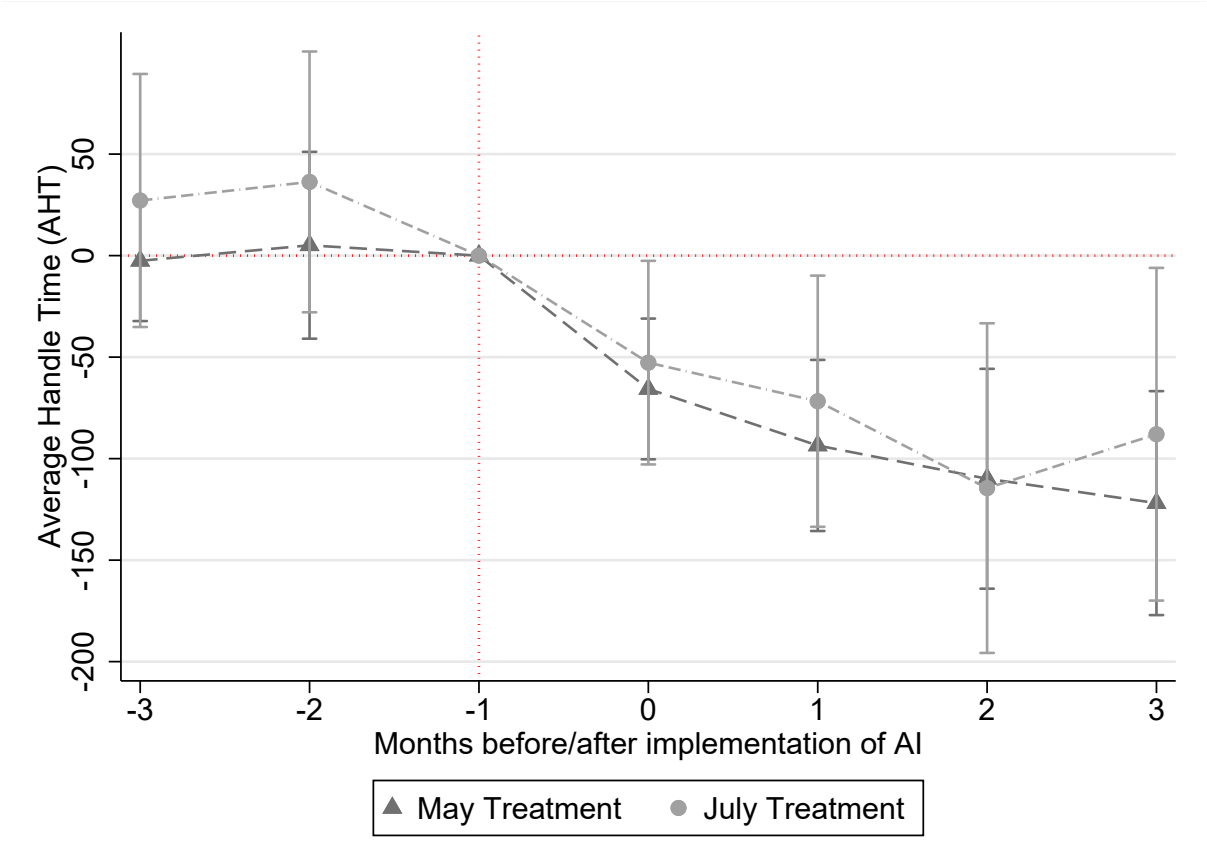

Figure 3: Event study estimates of Average Handle Time (AHT), May & July treatments

Notes: This figure presents an event study using the Callaway and Sant'Anna (2021) method to evaluate the effect of the implementation of AI-augmented training on Average Handle Time (AHT). The x-axis indicates months before and after AI implementation, and the y-axis shows changes in AHT (in seconds). Triangles represent coefficient estimates for the May treatment group; circles represent estimates for the July group. Vertical capped spikes indicate 95% confidence intervals. Each estimate includes agent and month fixed effects, and standard errors are clustered at the agent level.

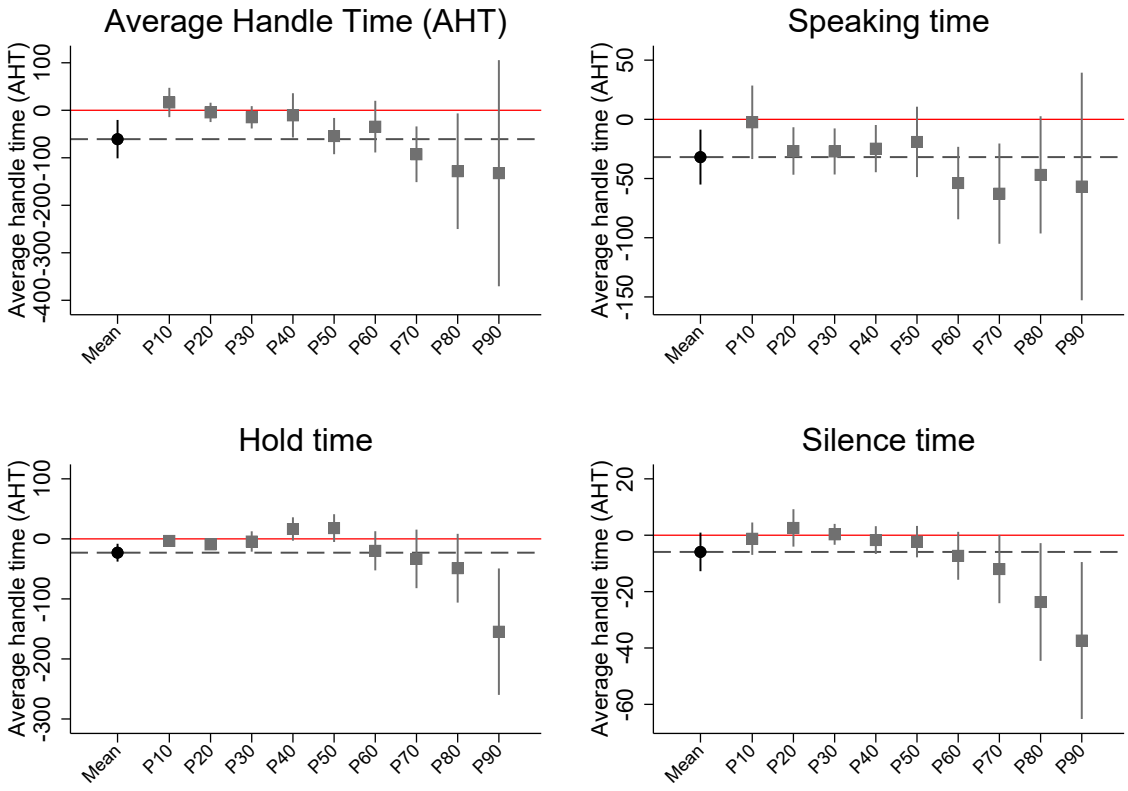

Figure 4: Effects on AHT and its elements at different percentiles of the distribution

Notes: This figure presents the results from using robust DiD estimators introduced in Callaway and Sant'Anna (2021) The black dots on the left side represent the average effects. The gray markers indicate the effects on various percentiles from Recentered Influence Function (RIF) regressions at different percentiles of the outcome distribution. The dependent variables are Average Handle Time (AHT) in seconds, speaking time, hold time, and silence time (all measured in seconds per call). All models include agent and month fixed effects and account for pre-treatment tenure categories through reweighting. Standard errors are clustered at the agent level.

Tables in text

Table 1: Sample Characteristics

	(1) All	(2) Treatment Group	(3) Control Group
Number of calls	187,839	79,935	107,904
Number of agents	147	84	63
Share of agents with non-standard training	.41	-	.72
Number of monthly observations	957	440	517

Notes: This table reports summary statistics for all agents in the sample. Column 1 presents statistics for the full sample across all time periods. Column 2 presents summary statistics for treated agents who had AI-augmented training. Column 3 presents statistics for control agents who did not receive AI-augmented training at any point during the observation window.

Table 2: Agent and Call Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
	May Treatment			**July Treatment**		
	Treatment	Control		Treatment	Control	
		Raw	Weighted		Raw	Weighted
Agent Tenure (Share of Agents):						
0–4 Months	0.455	0.279	0.469	0.455	0.082	0.500
5–11 Months	0.485	0.209	0.469	0.273	0.164	0.154
12+ Months	0.061	0.512	0.062	0.273	0.754	0.346
Average Call Characteristics (in seconds):						
Average Handle Time (AHT)	651.227	494.623	550.074	620.170	442.742	593.632
Speaking Time	386.723	311.014	320.166	352.495	302.949	380.276
Silence Time	69.421	53.628	58.664	64.393	47.134	60.932
Hold Time	195.082	129.981	171.244	203.282	92.659	152.423
Ineffective Communication Styles (Share of Calls):						
Filler Words	0.534	0.439	0.455	0.564	0.530	0.604
Diminutive Words	0.399	0.338	0.341	0.342	0.335	0.435
Hedge Words	0.386	0.376	0.419	0.484	0.358	0.430

Notes: This table reports sample statistics for the May and July treatment groups and their cor- responding control groups. Columns 1 and 4 show means for treated agents; Columns 2 and 5, means for control agents. Columns 3 and 6 apply Callaway and Sant'Anna (2021) reweighting to adjust for observable differences.

Table 3: Treatment effect on Average Handle Time (AHT)

	(1)	(2)	(3)	(4)	(5)	(6)
Treated x Post AI	-68.183***	-91.639***	-60.841***	-65.215***	-57.399***	-64.836***
	(17.364)	(18.618)	(20.342)	(20.854)	(16.925)	(18.569)
OLS TWFE	Yes	No	No	No	No	No
Callaway-Sant'Anna (CS)	No	Yes	Yes	Yes	Yes	Yes
Agent Tenure	No	No	Yes	Yes	Yes	Yes
Agents with non-standard training excluded	No	No	No	Yes	No	Yes
Entire Period	No	No	No	No	Yes	Yes
Pre-treatment average	636	636	636	636	589	589
Observations	957	956	956	579	764	402

Notes: This table presents the results of using the TWFE-OLS and robust DiD estimators introduced in Callaway and Sant'Anna (2021) on our main measure of productivity: Average Handle Time (AHT, in seconds). Columns 1 & 2 only include agent and month-fixed effects. Columns $3-6$ also include agent-tenure fixed effects. Columns 4-6 use alternative sample specifications, excluding both agents with non-standard training and workers who join or leave the department during our observation period. All standard errors are clustered at the call agent level. *** $p < 0.01$.

Table 4: Heterogeneous treatment effects on time spent during call

Experience group	(1) *0–4 Months Tenure*	(2) *5+ Months Tenure*
	Average Handle time (AHT)	
Treated x Post AI	-130.92**	-40.87**
	(40.18)	(20.24)
Pre-treatment average	777	555
Observations	199	757
	Speaking Time	
Treated × Post AI	-65.68***	-29.46**
	(22.54)	(12.98)
Pre-treatment average	430	335
Observations	199	757
	Hold Time	
Treated × Post AI	-51.641***	-7.99
	(18.644)	(7.34)
Pre-treatment average	266	161
Observations	199	757
	Silence Time	
Treated × Post AI	-13.596**	-3.41
	(5.917)	(2.36)
Pre-treatment average	81	59
Observations	199	757

Notes: This table presents the results of using robust DiD estimators introduced in Callaway and Sant'Anna (2021) and applied to disaggregated measures of AHT: speaking time, hold time, and silence time all measured in seconds. Columns compare agents with short tenure (0–4 months, Column 1) and tenure (5+ months, Column 2). All regressions include agent and month fixed effects. Standard errors are clustered at the agent level. *** $p < 0.01$; ** $p < 0.05$.

Table 5: Heterogeneous treatment effects on communication styles

Experience group	(1) *0–4 Months Tenure*	(2) *5+ Months Tenure*
	Filler Words	
Treated × Post AI	-0.051	-0.044**
	(0.033)	(0.018)
Pre-treatment average	0.6	0.52
Observations	199	757
	Diminutive	
Treated × Post AI	-0.026	-0.030*
	(0.034)	(0.017)
Pre-treatment average	0.45	0.33
Observations	199	757
	Hedge Words	
Treated × Post AI	-0.049	-0.044***
	(0.040)	(0.015)
Pre-treatment average	0.45	0.42
Observations	199	757
	Speaking Time (resid)	
Treated × Post AI	-45.54**	-11.03
	(18.70)	(11.73)
Pre-treatment average	430	355
Observations	199	757

Notes: This table presents the results using robust difference-in-differences estimators introduced in Callaway and Sant'Anna (2021) applied to communication style metrics: the share of calls containing filler words, diminutive language, or hedge words. Each outcome is estimated separately for agents with short tenure (0–4 months, Column 1) and long tenure (5+ months, Column 2). All regressions include agent and month fixed effects. Standard errors are clustered at the agent level. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Table 6: Call characteristics by topic of calls

Call Topic	(1) Share of Total Calls (%)	(2) Average Handle Time (AHT)	(3) Speaking Time	(4) Hold Time	(5) Silence Time
Death	24	632	408	157	66
Disability	6	735	458	200	77
Cohabitation or marriage	3	726	474	176	76
Moving	5	533	347	125	61
Pension accrual	19	682	445	165	71
Receiving pension	31	487	313	119	55
Retirement	31	682	447	161	74
Separation or divorce	13	660	418	174	68
Unemployment	1	898	573	231	94

Notes: This table presents average call characteristics (in seconds) by topic of the call. "Share of Total Calls" reflects the proportion of all calls attributed to each topic.

Table 7: Heterogeneous treatment effects on different topics by tenure

	(1) Average Handle Time (AHT)	(2) Speaking Time	(3) Hold Time	(4) Silence Time
		Panel A: Challenging Topics		
		0–4 Months Tenure		
Treated x Post AI	-148.69***	-71.08***	-60.75**	-16.86***
	(46.57)	(24.63)	(27.77)	(4.95)
Pre-treatment average	943	529	314	99
Observations	131	131	131	131
		5+ Months Tenure		
Treated x Post AI	-87.33	-64.49	-17.87	-4.97
	(57.89)	(50.97)	(15.24)	(6.27)
Pre-treatment average	757	470	210	77
Observations	751	751	751	751
		Panel B: Normal Topics		
		0–4 Months Tenure		
Treated x Post AI	-63.28	-24.51	-31.70	-7.07
	(38.66)	(16.55)	(20.39)	(6.20)
Pre-treatment average	620	348	205	67
Observations	131	131	131	131
		5+ Months Tenure		
Treated x Post AI	-28.56	-16.80	-9.63	-2.14
	(22.08)	(11.05)	(11.05)	(2.94)
Pre-treatment average	528	313	159	56
Observations	752	752	752	752

Notes: This table presents the results using robust difference-in-differences estimators introduced in Callaway and Sant'Anna (2021) applied on different topics, i.e., topics agents found challenging vs normal. Each outcome, measured in seconds, is estimated separately for agents with short tenure (0–4 months) and long tenure (5+ months). All regressions include agent and month fixed effects. Standard errors are clustered at the agent level.*** $p < 0.01$; ** $p < 0.05$.

Table 8: Treatment effects on customer satisfaction and calls per day

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean			*Percentiles*			*0-4 Months Tenure*	*5+ Months Tenure*
		10	25	50	75	90		
Calls per Day	0.69	-5.18*	-3.53	0.31	6.23**	5.29**	0.79	1.99
	(1.35)	(2.81)	(2.33)	(1.95)	(2.48)	(2.06)	(1.55)	(1.54)
Pre-treatment average	22.19						24.1	20.91
Observations	956	956	956	956	956	956	199	757
Customer Satisfaction	0.44*	0.27	0.27	0.41**	0.21	0.56**	0.46**	-0.30
	(0.24)	(0.54)	(0.31)	(0.21)	(0.19)	(0.25)	(0.21)	(0.21)
Pre-treatment average	8.23						8	8.39
Observations	738	738	738	738	738	738	146	500

Notes: This table presents results on Calls per Day and Customer Satisfaction using robust difference-in-differences estimators introduced in Callaway and Sant'Anna (2021). Columns labeled 10 to 90 report results from Recentered Influence Function (RIF) regressions at the 10th, 25th, 50th, 75th, and 90th percentiles, respectively. The final two columns show treatment effects by employee tenure: 0–4 months and 5+ months. All models include agent and month fixed effects and account for pre-treatment tenure categories through reweighting. Standard errors are clustered at the agent level. monitoring systems. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

Online Appendix: How AI-Augmented Training Improves Worker Productivity

A.1 Potential violation of SUTVA

As mentioned in Section IV, the agents affected by the May intervention work in the same location as the members of the control group. As a result, spillover effects from the May treatment group to the control group may be a concern. To address this concern, the following figure presents results from a specification that compares the May treatment group to all not-yet-treated agents who are affected by the treatment in July and are located in a geographically distant region. Because the second treatment occurred in July, we can estimate effects only for the first (May) and second (June) months after treatment. Nonetheless, both the original and this geographically restricted comparison yield quantitatively and qualitatively similar results.

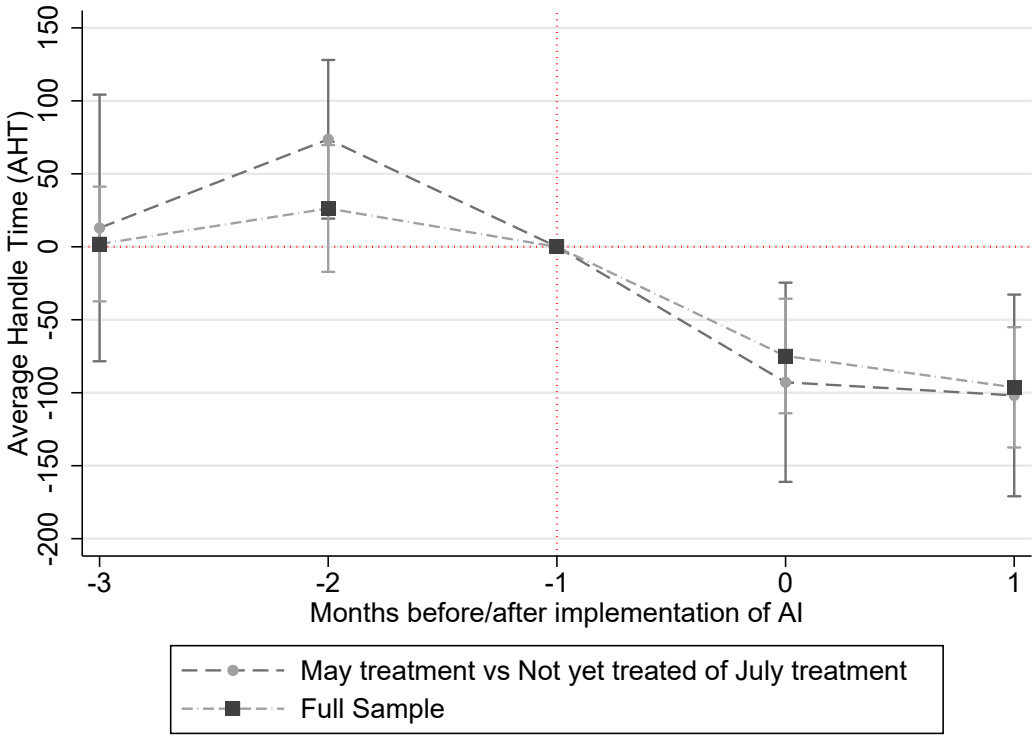

Figure A.1.1: Event study estimates for spillover effects

Notes: This figure presents an event study using the Callaway and Sant'Anna (2021) method to evaluate the effect of the implementation of AI-augmented training on Average Handle Time (AHT). The x-axis indicates months before and after AI implementation, and the y-axis shows changes in AHT. They gray dots represent the coefficient estimates for the May treatment using only not-yet-treated agents from the July treatment as controls. The black squares represent coefficients for the May treatment using all agents as controls. Vertical capped spikes indicate 95% confidence intervals. Each estimate includes agent and month fixed effects, and standard errors are clustered at the agent level.

A.2 Alternative main outcomes as event studies on weekly and monthly levels

Graphs not included yet

A.3 Alternative approaches to overcome the bias in classical TWFE estimation

The following figure presents estimation results of our preferred model Callaway and Sant'Anna (2021) in comparison with three other common approaches to account for potential estimation biases in classical TWFE models. Sun and Abraham (2021), De Chaisemartin and D'Haultfœuille (2020), and Borusyak et al. (2024). All results show quantitatively and qualitatively similar results.

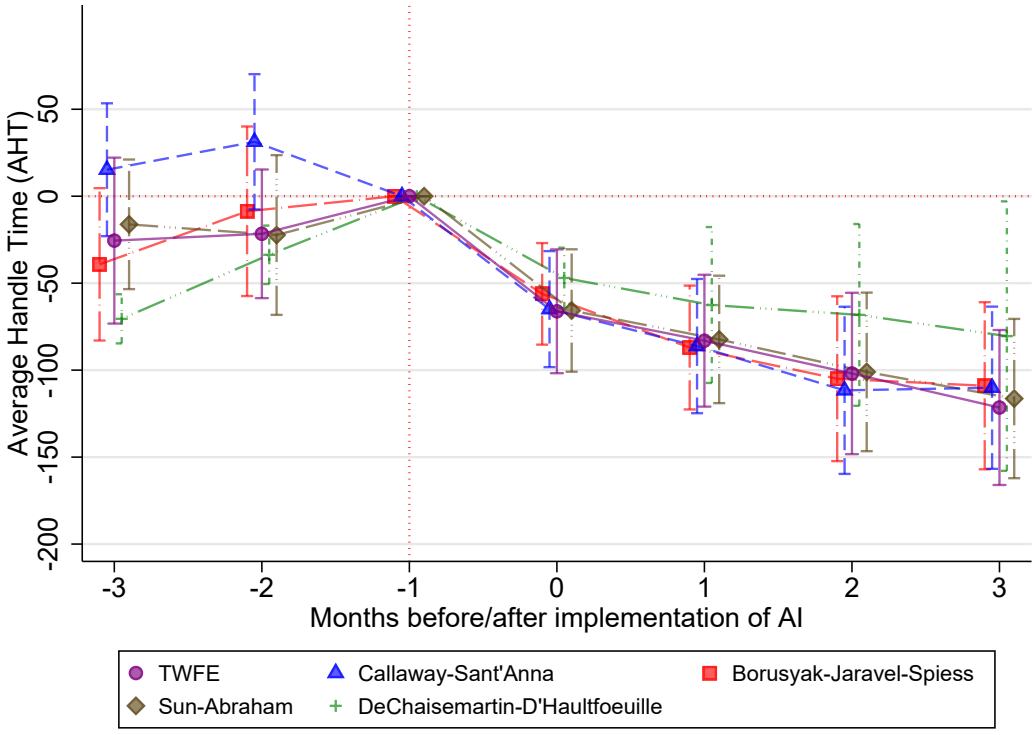

Figure A.3.1: Event study estimate, Average Hangle Time (AHT), different model estimators

Notes: The figure compares different event study methods that correct biases from heterogeneous and dynamic treatment effects. The x-axis represents the months relative to the implementation of the AI, while the y-axis shows the conditional gap in AHT before and after the implementation. Purple represents the classical TWFE approach, blue represents the approach by Callaway and Sant'Anna (2021), green the one by Borusyak et al. (2024), red the one by De Chaisemartin and D'Haultfœuille (2020), and black the one by Sun and Abraham (2021). All capped spikes indicate 95 percent confidence intervals. All models include agent and month fixed effects. Standard errors are clustered at the agent level. The classical TWFE approach and Sun and Abraham (2021) rely on 957 observations while the remaining approaches exclude invalid comparisons, such that Callaway and Sant'Anna (2021) relies on 956, Borusyak et al. (2024) on 914, and De Chaisemartin and D'Haultfœuille (2020) on 664.

A.4 Treatment Effects on Time Spent for different Activities During the Call

Table A.4.1 shows the ATT for speaking, hold, and silence time. The upper panel shows results from a specification includes all agents from the control group. The lower panel shows results from a specification that excludes agents with non-standard coaching from the control group.

Table A.4.1: Treatment effects on time spent for different activities during the call

	(1) Speaking Time	(2) Hold Time	(3) Silence Time
Panel A: Whole Sample			
Treated x Post AI	-31.911**	-23.007***	-5.928
	(12.077)	(7.688)	(3.399)
Observations	956	956	956
Panel B: Never Treated Agents with Non-Standard Coaching Excluded			
Treated x Post AI	-35.317***	-24.190***	-5.708*
	(12.547)	(8.071)	(3.450)
Observations	579	579	579
Pre-treatment average	369	199	66

Notes: This table presents the results using robust difference-in-differences estimators introduced in Callaway and Sant'Anna (2021) on our disaggregated measures of AHT: speak time, hold time and silence times. All columns account for pre-treatment tenure through reweighing and include month and agent fixed effects. All standard errors are clustered at the agent level. ** $p < 0.05$, *** $p < 0.01$.

A.5 Use of Communication Styles and Call Length

Table A.5.1 examines how our communication style metrics relate to AHT, speaking time, hold time, silence time, and customer satisfaction ratings using OLS. These metrics correlate strongly with all outcomes and explain the most variance in speaking time.

Table A.5.1: Use of communication styles and call length

	(1) AHT	(2) Speaking Time	(3) Hold Time	(4) Silence Time	(5) Customer Satisfaction
Filler Words	202.26***	174.88***	25.56	1.82	0.51***
	(31.88)	(14.37)	(20.86)	(3.97)	(0.18)
Diminutive	340.23***	135.19***	154.21***	50.83***	-0.44**
	(32.81)	(14.79)	(21.47)	(4.09)	(0.19)
Hedge Words	360.41***	168.44***	169.70***	22.27***	-0.01
	(29.43)	(13.27)	(19.26)	(3.67)	(0.18)
Observations	1,159	1,159	1,159	1,159	801
Adjusted R^2	0.42	0.51	0.20	0.24	0.01

Notes: Each column reports coefficients from separate regressions of the listed outcome variable on communication style indicators. ** $p < 0.05$, *** $p < 0.01$.