When AI Takes the Call: Automation, Handoffs, and Residual Human Work*

Pulak Ghosh† Rik Sen‡

June 2026

Abstract

When a firm adopts AI, it rarely faces a clean choice between replacing workers and assisting them; more often it does both at once. We study a large digital consumer lender in India that, within a few weeks, gave its customer-service agents an AI copilot and began routing tickets to an AI bot. Observing automation and augmentation in one firm lets us ask where each creates and destroys value. For identification we exploit the staggered, category-by-category rollout: we predict a ticket's exposure to automation from its text and compare high- and low-exposure tickets within the same category, before and after the bot switches on; event studies show flat pre-trends. Cleanly automated tickets show no speed–quality tradeoff: resolution time falls 94% while reopenings decline. The cost concentrates instead at the failed handoff, where the bot gives up and a human inherits a confused customer, and reopening rates rise 58 percentage points. Automation also hardens the residual human queue, and by stripping away routine tickets it makes the copilot more valuable, so automation and augmentation act as complements, not substitutes. Because servicing is part of intermediation in consumer lending, these dynamics speak to borrower outcomes, not operations alone.

*We thank seminar participants at the University of Miami and the Atlanta Finance Workshop for helpful comments. All errors are our own. Data access was provided by the firm under a research agreement; we withhold the firm's identity to protect commercially sensitive information. Replication code is available from the authors on request. This research received no external financial support, and the authors have no relevant financial relationships or other potential conflicts of interest to disclose.

†Indian Institute of Management Bangalore. Email: pulak.ghosh@iimb.ac.in.

‡University of Georgia. Email: rik.sen@uga.edu.

JEL codes: G21, G51, J23, O33, L84
Keywords: artificial intelligence, automation, augmentation, customer service, fintech, consumer lending

1 Introduction

Customer service has become one of the first corporate functions to adopt artificial intelligence at scale. Within a month of launch, Klarna's AI assistant was handling two-thirds of the company's customer-service chats, work the firm likened to 700 full-time agents (Klarna 2024). The appeal is faster service at lower cost. The risks are just as concrete: a Canadian tribunal held Air Canada liable when its customer-service chatbot invented a refund policy (Civil Resolution Tribunal of British Columbia 2024), and Klarna itself later began rehiring human agents after concluding that its automation push had come at the cost of service quality (Customer Experience Dive 2025). The gains are real, but so are the failures, and we still know little about when automating service pays off and when it backfires. The question is sharpest in consumer finance, where servicing is not a cost center bolted onto the product but a part of intermediation. How a lender answers a borrower's question about a missed payment or a foreclosure request potentially feeds back into repayment behavior and compliance risk.

AI systems are entering workplaces in two fundamentally different modes. *Automation* replaces human tasks entirely—chatbots resolve customer queries, algorithms approve loan applications, robots assemble products. *Augmentation* assists human workers—copilot tools draft responses for agents to review, diagnostic AI highlights anomalies for radiologists, coding assistants suggest completions for developers. The distinction matters. Acemoglu and Restrepo (2018, 2019) model automation as task displacement: when machines perform tasks previously done by labor, employment falls unless new tasks emerge. But augmentation—where AI complements rather than substitutes for labor—is absent from this framework. Empirical studies of AI augmentation find productivity gains (Brynjolfsson et al. 2025; Noy and Zhang 2023; Dell'Acqua et al. 2026), but they cannot speak to automation's displacement channel. No study has simultaneously observed both modes within a single firm, making it impossible to assess whether they interact.

This paper provides the first within-firm evidence on the simultaneous effects of AI automation and augmentation. We study a large digital consumer lender in India that deployed two AI systems in its customer-service operation during 2025: a *bot* that autonomously handles support tickets (automation, deployed in September) and a *copilot* that drafts responses for human agents to review and send (augmentation, deployed in August). Using 89,200 tickets that span the full calendar year and a text-based intent-to-treat (ITT) design that exploits within-category variation in predicted bot exposure, we decompose the aggregate effect into three channels and ask whether automation and augmentation are complements or substitutes.

The headline result is large and positive: resolution time falls 57% and durable resolution quality improves 12 percentage points after AI deployment. But the aggregate masks sharp heterogeneity across channels. We decompose post-period tickets into three mutually exclusive groups based on how each ticket was handled. *Bot-only* tickets (26% of post-period volume) are resolved in 4.6 hours versus 93 hours pre-deployment—a 94% reduction—with a 96% same-day resolution rate and a 51-point increase in one-touch resolution. *Handoff* tickets (12%), where the bot attempts resolution but fails and transfers to a human agent, exhibit a severe quality penalty: reopening rates rise 58 percentage points relative to the pre-period, and resolution time exceeds the pre-period human benchmark. *Human-only* tickets (62%) show modest improvement in resolution time even though human agents now handle harder cases, suggesting that workload relief from automation spills over to the remaining human tasks.

Estimating the effect of automation is harder than it first appears. We want to know how the bot changes the speed and quality with which a ticket is resolved. The naive comparison—tickets the bot handled versus tickets humans handled, or outcomes before versus after the bot was switched on—is confounded on both sides. The bot is not assigned at random: it is routed to tickets the firm judges it can handle, which are systematically simpler, so a bot-versus-human comparison conflates the technology with ticket difficulty. And because the bot was rolled out category by category over several months, a simple before-after contrast absorbs everything else that changed over the same period; comparing automated with not-yet-automated categories has the mirror-image problem, since the categories chosen for early automation differ from those left to humans.

Our strategy holds the category fixed and exploits variation in automation *exposure* within it. We train a text classifier to predict each ticket's probability of bot handling from its language, then interact this predicted exposure with a staggered, category-specific post indicator, absorbing category, month, day-of-week, and difficulty fixed effects. Identification comes from comparing high- and low-exposure tickets in the same category, before versus after the bot is activated; the identifying assumption is that, conditional on these controls, a ticket's predicted exposure affects outcomes only through the bot. Two checks support it. An event study shows the high-minus-low exposure gap is flat through the pre-period and breaks only when the bot is switched on (Figure 1), so the effect is not a pre-existing trend. And pre-period response "templateness" does not predict which categories were later automated (Appendix Figure A5), which weighs against the worry that we are capturing task routineness rather than automation itself. We complement the design with a category-level difference-in-differences, a within-customer panel, and a 2SLS specification that instruments realized bot handling with predicted exposure.

We develop a model that nests the Acemoglu and Restrepo (2018) task-displacement framework and extends it with three features absent from the original: imperfect automation with costly handoffs, AI-assisted augmentation of retained human tasks, and a quality dimension alongside speed. The model generates eleven testable predictions grouped by source—standard displacement and productivity effects (present in Acemoglu–Restrepo), handoff and workload-relief effects (new), and augmentation and complementarity effects (new). All eleven predictions find support in the data. The complementarity result is particularly sharp: because the copilot is more valuable on harder tickets ($\beta'(\theta) > 0$), and automation shifts the residual human workload toward harder tickets, more automation mechanically raises the copilot's average contribution ($\partial\overline{\alpha}/\partial I > 0$). Empirically, tickets predicted to receive copilot assistance see a 0.18 standard-deviation larger improvement in durable resolution after bot deployment (Table 6, column 5).

This paper contributes to three literatures. First, it extends the automation–labor literature (Acemoglu and Restrepo 2018, 2019, 2020; Autor et al. 2003; Autor 2015; Frey and Osborne 2017) by providing within-firm evidence on the *anatomy* of automation. Prior work uses cross-industry or cross-occupation variation in robot or AI exposure (Acemoglu and Restrepo 2020; Webb 2020; Felten et al. 2023; Eloundou et al. 2024); we observe the actual deployment, the ticket-level assignment mechanism, and the three channels through which automation affects outcomes. The handoff channel is new to this literature: when automation is imperfect, some tasks are handled *worse* than if humans had retained them—a cost absent from models that assume clean task displacement.

Second, the paper speaks to the growing empirical literature on AI augmentation. Brynjolfsson et al. (2025) study a GPT-based assistant at a customer service firm and find a 14% productivity increase concentrated among less-experienced workers. Noy and Zhang (2023) and Dell'Acqua et al. (2026) run experiments showing that generative AI improves writing and consulting output quality, with larger gains for lower-ability participants. Peng et al. (2023) find that GitHub Copilot reduces coding task completion time by 56%. Our setting differs in that we observe automation *and* augmentation simultaneously. The closest comparison is with Brynjolfsson et al.: their firm deploys only a copilot; ours deploys a copilot alongside a bot. This lets us test a prediction absent from prior work: that the two modes are complements.

Third, we contribute a formal framework that nests Acemoglu–Restrepo displacement within a richer model featuring imperfect automation, handoffs, and augmentation. Setting the extensions to zero recovers the original displacement channel. The model's key novel prediction —complementarity between automation and augmentation—has direct implications for how firms should sequence AI investments: deploying both is more valuable than

deploying either alone.

Finally, the paper connects the economics of AI at work to financial intermediation. In consumer lending, customer service is part of the product: it is where borrowers resolve payment and foreclosure problems and where the lender manages delinquency and compliance. Automating it lowers the firm's labor costs, but it also reshapes the servicing relationship precisely at the failed handoff, the point where automation is most likely to break down. For investors and supervisors, then, the statistic that matters is not how much a firm automates but how well it automates at that margin.

Section 2 describes the institutional setting and data. Section 3 develops the theoretical framework and derives testable predictions. Section 4 presents the empirical strategy. Section 5 reports results, including the channel decomposition and the automation-versus-augmentation comparison. Section 6 provides a welfare calculation. Section 7 concludes.

2 Institutional Setting and Data

2.1 The Firm and Its Customer Service Operation

The firm is a large digital consumer lender in India that offers personal lines of credit. Its customer-service operation handles support tickets submitted via email and in-app messaging. Tickets span thirteen categories, including account management, repayment, credit limits, loan transfers, collections, KYC verification, and technical issues. Before AI deployment, all tickets were handled by a team of fourteen human agents.

2.2 AI Deployment

The firm deployed two AI systems in sequence during 2025.

Copilot (August 2025). An AI assistant built on a large language model began drafting responses for human agents. For a given ticket, the copilot generates a candidate reply; the agent reviews, edits if necessary, and sends. Copilot adoption was rapid: the copilot-assisted share of human-handled tickets jumped from 0% to 53% in August. Critically, the copilot was deployed *before* the bot, creating a brief augmentation-only period (August) that we exploit for identification.

Bot (September 2025). An AI bot began routing tickets autonomously. For tickets the bot handles, it provides the first response and—if the customer's issue is resolved—closes the ticket without human involvement. When the bot fails to resolve a ticket, it hands

off to a human agent. The bot was activated on a category-by-category basis, with high-volume categories (account, repayment, general) automated first. By December, the bot was resolving roughly a third of all tickets end to end, with a further 13% beginning at the bot before being handed off to a human.

2.3 Data

We use the universe of resolved customer support tickets from January 1 through December 31, 2025. After a seven-step cleaning cascade—removing merge-closed duplicates, spam, outbound and system-generated tickets, zero-activity ghost records, and tickets that never reached a closed or resolved status—the sample contains 89,200 tickets. The pre-period (January–August) contains 56,913 tickets, all handled by human agents. The post-period (September–December) contains 32,287 tickets split across the three handler channels. The intent-to-treat regressions, which require a text-based propensity score, use the 86,589 tickets with sufficient ticket text.

Outcome variables. We measure five dimensions of service quality:

1. *Speed*: resolution time (hours from ticket creation to closure), first response time (FRT, hours to first agent reply), and an indicator for same-day resolution (within 24 hours).
2. *Quality*: durable resolution (resolved with at least one public reply and zero reopenings), one-and-done (same-day with zero reopenings and at most one agent reply), whether the ticket was reopened, and whether the customer re-contacts within 14 days.
3. *Compliance*: SLA breach (resolution after the contractual deadline).
4. *Satisfaction*: customer rating mapped to $\{-1, 0, +1\}$.
5. *Communication*: log word count of agent responses.

Treatment variables. We construct two propensity scores. The *bot propensity* is the predicted probability that a ticket receives initial bot handling, estimated from a TF-IDF logistic regression (10,000 unigram/bigram features, L2 penalty) trained on October tickets. The model achieves a cross-validated AUC of 0.775, indicating that ticket text strongly predicts bot assignment. The *copilot propensity* is constructed analogously from August tickets (AUC = 0.735).

Table 1 reports summary statistics.

3 Theoretical Framework

We develop a partial-equilibrium model of a customer service firm that deploys two AI technologies: an automation bot and an augmentation copilot. The model nests the task-displacement framework of Acemoglu and Restrepo (2018, 2019) and extends it with three channels absent from the original: imperfect automation with costly handoffs, AI-assisted augmentation of human workers, and a quality dimension alongside speed.

3.1 Environment

A firm receives a unit mass of support tickets per period. Each ticket has complexity $\theta \in [0, \bar{\theta}]$ drawn from distribution F with continuous density $f(\theta) > 0$ and mean μ_θ. A human agent resolves ticket θ in time $\tau_H(\theta) = \tau_0 + \tau_1\theta$ (with $\tau_1 > 0$) and achieves quality $q_H(\theta) = q_0 - q_1\theta$ (with $q_1 > 0$, $q_0 \leq 1$). Both functions are monotone: harder tickets take longer and have lower quality.

3.2 Automation

The firm sets an automation threshold $I \in [0, \bar{\theta}]$. Tickets with $\theta \leq I$ are routed to the bot; tickets with $\theta > I$ remain with human agents. This threshold maps directly to the task boundary in Acemoglu and Restrepo.

The bot resolves ticket θ in time $\tau_B \approx 0$ with success probability $\pi(\theta)$, where $\pi(0) = 1$ and $\pi'(\theta) < 0$. Successful resolution has quality 1. With probability $1 - \pi(\theta)$, the bot fails and hands off to a human, incurring delay $\delta > 0$ and quality penalty $\Delta_q > 0$.

Automation generates four effects:

1. **Displacement** (standard in Acemoglu–Restrepo): fraction $F(I)$ of tasks shifts from humans to the bot.
2. **Productivity** (standard): automated tasks are resolved faster and, when the bot succeeds, at higher quality.
3. **Handoff** (extension): failed bot attempts create tickets with outcomes *worse* than direct human handling—time $\delta + \tau_H(\theta) > \tau_H(\theta)$ and quality $q_H(\theta) - \Delta_q < q_H(\theta)$.
4. **Workload relief** (extension): human agents handle fewer tickets, increasing time and attention per remaining ticket. We model this as a quality externality $\phi(\lambda)$ with $\phi' < 0$ and $\phi(1) = 0$, where $\lambda(I) = 1 - F(I) + \int_0^I [1 - \pi(\theta)] f(\theta) d\theta < 1$ is the post-automation human workload.

Setting $\pi(\theta) = 1$ for all $\theta \leq I$ eliminates handoffs and, combined with $\alpha = \beta \equiv 0$ (no augmentation), reduces the model to a single-sector version of Acemoglu and Restrepo.

3.3 Augmentation and Complementarity

The firm simultaneously deploys a copilot that assists human agents on retained tickets ($\theta > I$). The copilot reduces resolution time by $\alpha(\theta) > 0$ and improves quality by $\beta(\theta) > 0$:

$$\tau_H^A(\theta) = \tau_H(\theta) - \alpha(\theta), \tag{1}$$

$$q_H^A(\theta) = q_H(\theta) + \beta(\theta) + \phi(\lambda(I)). \tag{2}$$

The key assumption is that $\alpha'(\theta) > 0$ and $\beta'(\theta) > 0$: the copilot is *more* valuable on harder tickets. On trivial issues the agent already knows the standard response; on complex issues requiring synthesis of multiple policies and account details, the AI draft substantially reduces cognitive load.

Proposition 1 (Complementarity)**.** *The average copilot contribution on retained tasks,* $\overline{\alpha}(I) \equiv E[\alpha(\theta) \mid \theta > I]$*, is increasing in the automation threshold* I*:*

$$\frac{d\,\overline{\alpha}}{dI} = \frac{f(I)}{1 - F(I)}\big[\overline{\alpha}(I) - \alpha(I)\big] > 0.$$

Proof sketch. Raising I removes the easiest remaining task ($\theta = I$) from the human pool. Since the copilot contribution is increasing in complexity, the removed task has the lowest copilot value among retained tasks: $\alpha(I) < E[\alpha(\theta) \mid \theta > I] = \overline{\alpha}(I)$. Thus the conditional mean rises. The formal proof is in Appendix A.

Interpretation. More automation shifts the residual complexity distribution rightward. Since the copilot is more helpful on harder tasks, this compositional shift raises its average contribution. Automation and augmentation are complements: each makes the other more valuable.

3.4 Predictions

The model generates eleven testable predictions (Table 7); the rest of this subsection states them in words and gives the intuition for each.

The predictions fall into three families. The first is familiar from standard automation models. When the bot takes over a ticket it resolves it almost instantly, so automated tickets get faster (P1); and because the bot absorbs the easy cases, the work left for humans grows harder (P2). These effects need not be confined to the tickets the bot actually touches: if automation frees up agent capacity, even tickets that stay with humans should improve, and the gains should show up across the range of automation exposure rather than only where bot handling concentrates (P3, P7). The second family is what standard models leave out,

because automation here is imperfect. When the bot attempts a ticket and fails, it hands a half-finished case to a human, so handoff tickets should end up *worse* than had a human taken them from the start (P4). As the bot is tuned, fewer tickets should reach that failure point (P5), and the quality of the tickets it does resolve should rise over time (P6).

The third family concerns the copilot and how it interacts with the bot. Drafting assistance should improve the quality of human replies, so tickets that receive it should earn better customer ratings (P8). The sharper prediction is about the interaction. The copilot helps most on hard tickets, where a strong first draft saves the agent the most effort, and the bot—by clearing away the easy tickets—pushes the human queue toward exactly those hard cases. The value of augmentation should therefore *rise* once automation is in place, making the two tools complements rather than substitutes (P9). The same logic implies larger quality gains for more complex tickets (P10); and because the bot never sleeps, its speed advantage should be largest outside business hours (P11). We test all eleven predictions in Section 5.

4 Empirical Strategy

4.1 Intent-to-Treat Design

Our goal is the causal effect of bot automation on how a ticket is resolved. The obstacle is that we cannot randomize the bot, and which tickets it actually handles is an endogenous outcome of the firm's routing rule—the bot takes the cases it expects to handle well, which differ systematically from those left to humans. Conditioning on realized handling would therefore compare unlike tickets. A design that sidesteps this replaces realized handling with each ticket's *ex-ante* predicted exposure to the bot, and asks whether high-exposure tickets improve differentially once the bot is switched on in their category. The identifying variation then comes from ticket text and rollout timing, not from the endogenous routing decision.

Concretely, the main specification interacts a ticket's predicted bot-handling probability with a staggered post-deployment indicator:

$$Y_i = \beta \cdot \text{PredBot}_i \times \text{Post}_{c(i)} + \gamma_1 \text{PredBot}_i + \gamma_2 \text{Post}_{c(i)} + \delta_X + \varepsilon_i, \tag{3}$$

where PredBot_i is the predicted probability that ticket i is handled by the bot (from the TF-IDF logistic model trained on October tickets), $\text{Post}_{c(i)}$ activates for category c at the first month in which the category's bot rate exceeds 20%, and δ_X denotes fixed effects for month, day of week, AI-classified bot difficulty, and—in the preferred specification—ticket

category.

The coefficient β measures the differential change in outcome Y for a ticket with unit higher predicted bot probability, comparing the post-deployment period to the pre-deployment period, *within the same category*. We report two variants: (A) continuous predicted probability, and (B) a binary indicator for above-median predicted probability. HC1 robust standard errors are used throughout.

Identification. The excluded variation is the interaction of ticket-text features (which predict bot handling) with the post-period indicator. Conditional on category fixed effects (which absorb topic-level differences), bot-difficulty fixed effects (which absorb complexity-level differences), and time fixed effects, the remaining text variation consists of word choice and phrasing that trigger the bot's NLP confidence score. The identifying assumption is that this residual text variation affects outcomes only through bot handling, not directly. We provide three pieces of evidence supporting this assumption.

First, an event study shows that the high-versus-low propensity gap is flat and near zero for all speed outcomes in the pre-period (January–July), then diverges sharply in September (Figure 1).

Second, pre-period "templateness" analysis shows that pre-deployment response homogeneity does not predict post-deployment automation rates (Appendix Figure A5), ruling out the concern that routineness of the *task*—as opposed to the bot's text classifier—drives the results.

Third, we complement the ITT with difference-in-differences specifications using category-level treatment variation and a within-customer panel that absorbs all time-invariant customer heterogeneity.

4.2 Channel Decomposition

To decompose the aggregate effect, we replace the single Post indicator with three channel-specific dummies:

$$Y_i = \sum_{c \in \{\text{bot, handoff, human}\}} \delta_c \, D_{c,i} + \delta_X + \varepsilon_i, \tag{4}$$

where $D_{c,i} = \mathbf{1}[\text{Post} \wedge \text{channel} = c]$. The coefficient δ_c measures the FE-adjusted mean outcome for channel c relative to the pre-period (all-human) baseline. Since the three channel indicators partition the post-period exhaustively, $\sum_c \pi_c \delta_c$ recovers the aggregate post-period change.

4.3 Additional Designs

We implement three supplementary identification strategies: (i) a category-level difference-in-differences comparing automated ($\geq$30% bot rate) versus non-automated ($<$15%) categories; (ii) a within-customer panel exploiting repeat customers who submit tickets in both the pre and post periods; and (iii) a 2SLS specification instrumenting actual bot handling with the text-based propensity score interacted with Post. The 2SLS first-stage F-statistic is 1,634, far exceeding conventional weak-instrument thresholds.

5 Results

5.1 Main ITT Effects

Table 2 reports the main intent-to-treat results. The preferred specification (column 2, continuous propensity with category fixed effects) shows that a one-unit increase in predicted bot probability—equivalent to moving from a ticket certain to remain with a human to one certain to be routed to the bot—reduces log resolution time by 1.82 (SE = 0.08), log first response time by 2.06 (0.06), and raises the same-day resolution rate by 29 percentage points (0.02). Quality improves: durable resolution rises 12 points (0.02) and one-and-done resolution rises 28 points (0.02). SLA breach falls 12 points (0.02) with category fixed effects—the raw correlation without category FE is near zero, indicating that between-category differences mask a within-category improvement. Customer ratings improve modestly (+0.07, 0.02).

The binary specification (columns 3–4) tells the same story at a more interpretable scale: tickets above the median predicted bot probability see a 0.59 log-point reduction in resolution time (45% faster) and a 9 percentage point increase in same-day resolution.

5.2 Event Study

Figure 1 plots coefficients from the regression

$$Y_i = \sum_{m \neq 7} \beta_m \cdot \text{HighBot}_i \times \mathbf{1}[\text{month} = m] + \text{HighBot}_i + \text{month FE} + \text{category FE} + \varepsilon_i,$$

where HighBot is an indicator for above-median predicted bot probability and July ($m = 7$) is the omitted month—the last clean month before any AI deployment. The pre-period coefficients for speed outcomes are flat and near zero from January through July, confirming parallel trends. August shows a small *positive* spike in log resolution time (+0.22), consistent

with the copilot deploying on tickets overlapping in text characteristics with bot-eligible tickets and generating longer, more detailed drafts. The bot activates in September and the gap reverses sharply: log resolution falls 0.46 at $k = 0$ and reaches -0.83 by December. Same-day and one-and-done rates show flat pre-periods and large post-September jumps.

5.3 The Anatomy: Channel Decomposition

Table 3 decomposes the post-period effects into three handler channels. We focus on the preferred specification with full fixed effects.

Bot-only channel (26% of post-period tickets). Resolution time falls 2.82 log-points (-94%), FRT falls 2.30 log-points, and same-day resolution rises 48 points. One-and-done resolution increases 51 points and reopening falls 22 points. These tickets are resolved almost instantaneously, with quality comparable to or better than pre-period human handling.

Handoff channel (12%). Handoff tickets—where the bot initiates but fails and transfers to a human—exhibit a quality penalty. Resolution time *rises* 0.62 log-points relative to the pre-period human baseline, and reopening rates spike 58 points. Same-day resolution falls 24 points. However, first response time is still fast (-2.35 log-points) because the bot provides the initial reply before handing off. The handoff channel is the "dark side" of imperfect automation: customers experience a failed bot attempt followed by a delayed human resolution, yielding outcomes worse than direct human handling.

Human-only channel (62%). Outcomes for human-handled tickets change modestly. Resolution time rises slightly ($+0.18$ log-points), consistent with humans now handling harder cases (task reallocation), but durable resolution improves ($+0.03$) and reopening falls (-0.04). Ratings improve ($+0.04$). The key insight is that even though human-only tickets become harder, quality does not deteriorate—the workload-relief and copilot channels offset the increased complexity.

Net effect. The bot-only channel dominates the aggregate. Weighting each channel coefficient by its post-period share gives $0.26 \times (-2.82) + 0.12 \times 0.62 + 0.62 \times 0.18 \approx -0.55$ log-points; the single-indicator aggregate Post effect is -0.70 (column 1 of Table 3). The average improvement therefore comes from clean automation rather than from gains spread evenly across the three channels.

5.4 Why Handoffs Fail

The decomposition locates the quality cost of automation at the handoff. Why are these tickets handled worse than if a human had taken them from the start? We compare the bot's first response on handoff tickets with a counterfactual-human first response on comparable tickets, using a large language model to score what each reply actually does for the customer. The bot-handoff cohort is the 3,851 post-period tickets that began at the bot and were escalated; the counterfactual-human cohort is the 16,233 pre-period human-first tickets with the highest predicted handoff propensity.

Two patterns emerge (Table 4, Panel A). First, the bot's first response leaves more of the customer's request unmet. It resolves a smaller fraction of the customer's explicit asks (33% versus 49% for the counterfactual human, $d = -0.37$), and it defers, ignores, and contradicts asks more often (Appendix Figure A6). Second, and more telling, the bot's reply triggers far more customer re-engagement: 89% of bot-handoff tickets draw a customer reply after the first response, against 29% of counterfactual-human tickets—a 60 percentage-point behavioral gate. The bot answers fast, but in a way that prompts the customer to write back, and it is this re-engagement that generates the delays and reopenings the human then inherits.

How much of the handoff penalty is "in the text"—fixable by writing a better reply? We decompose each outcome's penalty into the share explained by observable first-response features (Panel B). Text features account for most of the gap in customer ratings (79%) but only about a quarter of the penalty in resolution time (27%) and reopening (24%), and essentially none of the re-contact gap (under 10%). The part of the failure that surfaces in *tone and wording* is largely captured by observables; the part that surfaces in *time and rework* is not. That residual runs through the behavioral channel above—whether the customer re-engages—and through selection on ticket complexity we cannot observe. The bot's routing rule is deterministic, so a handoff is not random: within a predicted-propensity decile, escalated tickets differ from retained ones along dimensions the text does not reveal.

Better prompts alone will not close the handoff gap. The failure is less that the bot writes a poor reply than that it replies without resolving the underlying request, leaving a customer who re-engages and a human who must rebuild the case. The handoff, not the average ticket, is where the value of in-house AI is decided.

5.5 Task Reallocation

Automation displaces routine tasks from human agents, shifting the residual human workload toward complex cases. Average ticket description length—a proxy for complexity—rises 69%

from 298 to 503 characters for human-handled tickets, and the share of tickets the AI difficulty classifier rates "hard," "very hard," or "complex" climbs sharply. The collections category, among the hardest in the firm's taxonomy, roughly doubles its share of human tickets, from 11% to 22%. Even with this heavier case mix, quality on human-handled tickets does not deteriorate (Table 3): the workload-relief and copilot channels offset the added complexity.

A category-level DiD confirms this compositional shift: human tickets in automated categories see a +0.33 log-point increase in resolution time (26% slower) relative to control categories, and a +0.26 log-point increase in description length. The bot absorbed the easy cases, leaving harder residual tickets (Appendix Table A7).

5.6 Within-Customer Evidence

The within-customer panel restricts to 6,505 repeat customers who submit tickets in both the pre and post periods (27,599 total tickets). Customer fixed effects absorb all time-invariant customer heterogeneity.

Table 5 reports the DiD specification. The same customer in an automated category sees resolution time fall an additional 1.04 log-points beyond the common time trend, same-day resolution rise 14 points, and customer rating improve 0.07 points. Restricting to human-only tickets, resolution still improves 0.32 log-points. This within-customer, within-handler evidence rules out compositional explanations: the improvement is not driven by easier tickets being sorted into the post-period or by different customers appearing.

5.7 Automation versus Augmentation

Table 6 compares the effects of the bot (automation) and copilot (augmentation). Column 1 reproduces the bot ITT. Columns 2–5 restrict to human-handled tickets and use the copilot propensity score as the treatment variable.

Automation dominates for speed: the bot ITT on log resolution is −1.82, while the copilot coefficient is +0.43. The positive copilot coefficient reflects selection—the copilot is deployed on harder tickets—not a causal slowdown. This cross-sectional selection is stable across pre-bot (column 3, +0.44) and post-bot (column 4, +0.70) periods, with the widening coefficient consistent with post-automation task reallocation concentrating the hardest cases among copilot-assisted tickets.

The copilot produces a consistently positive effect on customer ratings (+0.07 to +0.10 across specifications), suggesting that AI-drafted responses improve satisfaction regardless of whether automation is present.

Draft adoption. The copilot helps only when agents use its drafts, and often they do not. Among 750 copilot-eligible tickets we classified by the text similarity between the AI draft and the reply the agent actually sent, agents discarded the draft outright on 53%, used it partially on 35%, and adopted it largely intact on only 12%. The dominant reason for discarding, hand-coded from the draft–reply pair, is "wrong intent" (54%): the AI misreads what the customer is asking before any question of tone or wording arises. That agents keep editorial control—and use it—is consistent with the copilot lifting quality without introducing the kind of unmonitored errors that surface at the bot handoff.

Complementarity. Column 5 reports the interaction PredCopilot×Post, measuring whether copilot returns change after bot deployment. For durable resolution, the interaction is +0.18 ($p < 0.01$): tickets predicted to receive copilot assistance see larger quality gains in the post-bot period. For reopening, the interaction is −0.15 ($p < 0.01$). This is the empirical signature of Proposition 1: after the bot removes routine tickets, the remaining human caseload is harder, and the copilot—which is more valuable on harder tickets—generates larger quality improvements.

5.8 Robustness

We summarize robustness checks reported in full in the Appendix.

Difference-in-differences. The category-level DiD (Appendix Table A2) estimates an automation effect of −0.90 log-points on resolution time after absorbing common time trends. The event study (Appendix Figure A2) shows pre-trends for speed outcomes, making the DiD estimates a lower bound; the ITT design, which exploits within-category variation, does not share this concern.

2SLS. Instrumenting actual bot handling with the text-based propensity score yields LATE estimates uniformly larger than the ITT: −2.60 log-points on resolution time (versus −1.82 ITT), +0.41 on same-day (versus +0.29), +0.17 on durable resolution (versus +0.12). The first-stage F is 1,634 (Appendix Table A3).

Heterogeneity. Effects are present across all propensity quintiles, not just the top quintile where bot handling concentrates (Appendix Figure A3), suggesting indirect benefits (workload relief) beyond mechanical substitution. Quality gains concentrate on medium and hard tickets: durable resolution improves 0.22 and 0.17 for medium and hard tickets versus 0.02 (insignificant) for easy tickets (Appendix Table A4). Speed effects are larger off-hours

(−2.16 versus −1.96 log FRT), consistent with the bot's 24/7 availability channel (Appendix Table A5). Dynamic effects show no fade-out through December, and quality strengthens over time as the bot calibrates (Appendix Table A6).

6 Welfare

A back-of-envelope calculation (Appendix Table A8) quantifies the welfare gains. Pre-deployment, the average ticket took 93 hours to resolve; post-deployment, the average falls to 41 hours, a reduction of 52 hours per ticket. Over the 32,287 post-period tickets, this amounts to roughly 1.7 million customer-hours saved in the September–December window.

On the cost side, the 8,425 bot-only tickets require zero agent time. At the firm's fully loaded agent cost of INR 86 (USD 1.04) per ticket, automating them saves about USD 8,800 in direct agent labor over the post-period—on the order of four full-time agent positions. The break-even AI cost per bot ticket is USD 1.04, roughly 10–100 times typical LLM API costs for customer service ($0.01–$0.10 per interaction), so automation is cost-effective at current technology prices.

Quality effects are approximately neutral in aggregate: the bot-only channel removes reopenings while the handoff channel adds them, leaving a modest net improvement.

At U.S. customer-service wage levels ($15–$25/hour), the same time savings would generate 6–10 times larger cost reductions, so the gains from automating customer service scale roughly linearly with the wage level. This calculation treats time savings as fully fungible (agents reallocated, not laid off) and omits development costs, customer lifetime-value effects, general-equilibrium wage adjustments, and any downstream effects on borrower repayment that ongoing work aims to quantify.

7 Conclusion

AI's workplace effects are not monolithic. A single digital lender's deployment of an AI bot and copilot reveals three distinct channels through which AI reshapes customer-service work. Full automation handles 26% of tickets with dramatic speed gains and high quality. Failed-automation handoffs impose a quality penalty absent from standard task-displacement models. Residual human work improves modestly as agents handle fewer, harder tickets with AI copilot assistance.

The complementarity between automation and augmentation—formally predicted by the model and confirmed in the data—has implications for firm strategy and policy. Firms that deploy automation alone miss the augmentation gains on retained human tasks. Firms

that deploy augmentation alone miss the displacement-driven productivity gains. Deploying both is more valuable than either in isolation, because automation shifts the human workload toward cases where augmentation has the highest return.

For financial economics, the question is where AI creates or destroys value inside a lender. Servicing is part of intermediation: it shapes whether borrowers can resolve repayment problems and whether the firm stays compliant. Our results put the risk not in automation itself but at the handoff, where the bot gives up and a human inherits a half-finished case. What matters for managers and supervisors is not how much a lender automates but how cleanly it handles the boundary between machine and human.

For policy, the findings suggest that AI's labor market effects depend critically on the deployment mode. A firm that automates aggressively may displace workers but will also generate costly handoffs if the technology is imperfect. A firm that augments workers may improve quality without displacement but will forgo the large speed gains from full automation. The optimal deployment likely combines both, with the automation threshold calibrated to balance displacement gains against handoff costs.

The analysis here stops at service outcomes, but the same deployment invites a further question that we are now pursuing: does AI in the servicing function move borrower behavior? In ongoing work we link these tickets to repayment, delinquency, and borrowing records to test whether the speed and quality changes documented here carry through to loan performance. A future draft will report that evidence, connecting AI in customer service to credit risk and the lender's portfolio.

An important limitation is that this evidence comes from one firm in one industry. Customer service—with its textual inputs, categorical routing, and measurable quality—is a natural early adopter of AI. Whether the three-channel anatomy and the complementarity result generalize to settings where AI automates physical tasks or augments non-routine cognitive work is an open question that subsequent research with larger samples and different industries can address.

References

D. Acemoglu and P. Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6): 1488–1542, 2018.

D. Acemoglu and P. Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019.

D. Acemoglu and P. Restrepo. Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, 128(6):2188–2244, 2020.

D. H. Autor. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30, 2015.

D. H. Autor, F. Levy, and R. J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025.

Civil Resolution Tribunal of British Columbia. Moffatt v. Air Canada, 2024 BCCRT 149. British Columbia Civil Resolution Tribunal, 2024.

Customer Experience Dive. Klarna changes its AI tune and again recruits humans for customer service. Customer Experience Dive, 2025. `https://www.customerexperiencedive.com/news/klarna-reinvests-human-talent-customer-service-AI-chatbot/747586/`.

F. Dell'Acqua, E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Organization Science*, 2026. Forthcoming. DOI: 10.1287/orsc.2025.21838.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. GPTs are GPTs: Labor market impact potential of large language models. *Science*, 384(6702):1306–1308, 2024.

E. Felten, M. Raj, and R. Seamans. How will language modelers like ChatGPT affect occupations and industries? *SSRN Working Paper*, 2023. Available at SSRN 4375268.

C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.

Klarna. Klarna AI assistant handles two-thirds of customer service chats in its first month. Klarna press release, Feb. 2024. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/.

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

S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. Technical report, arXiv, 2023. arXiv:2302.06590.

M. Webb. The impact of artificial intelligence on the labor market. Working paper, Stanford University, 2020.

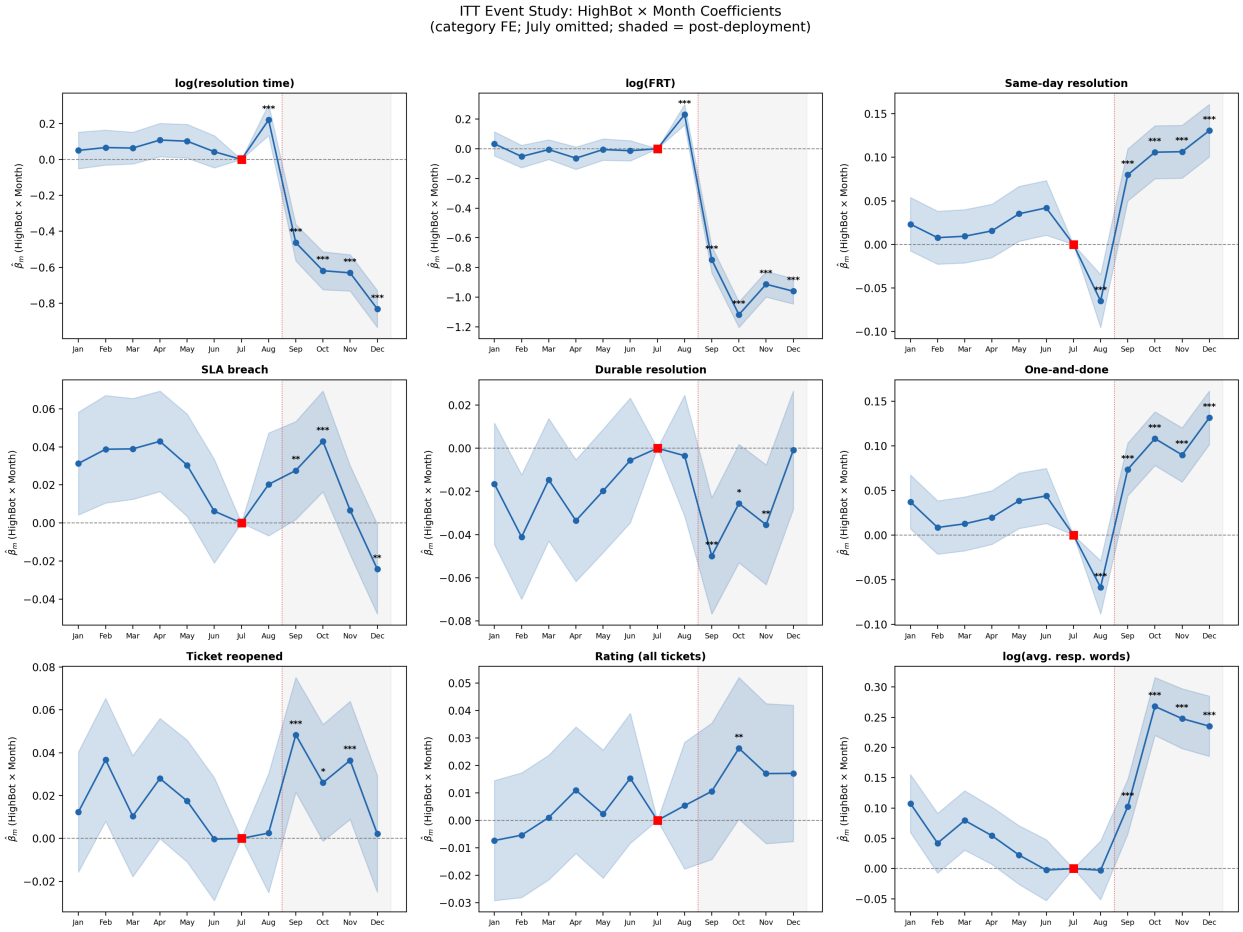
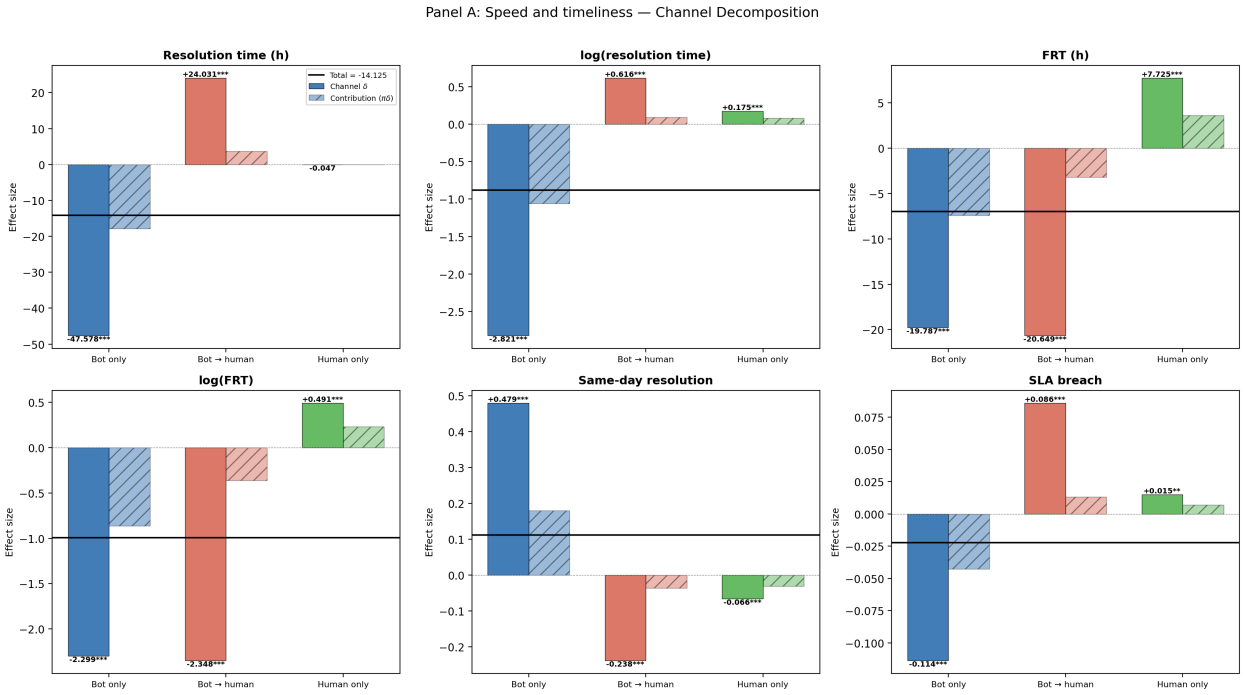
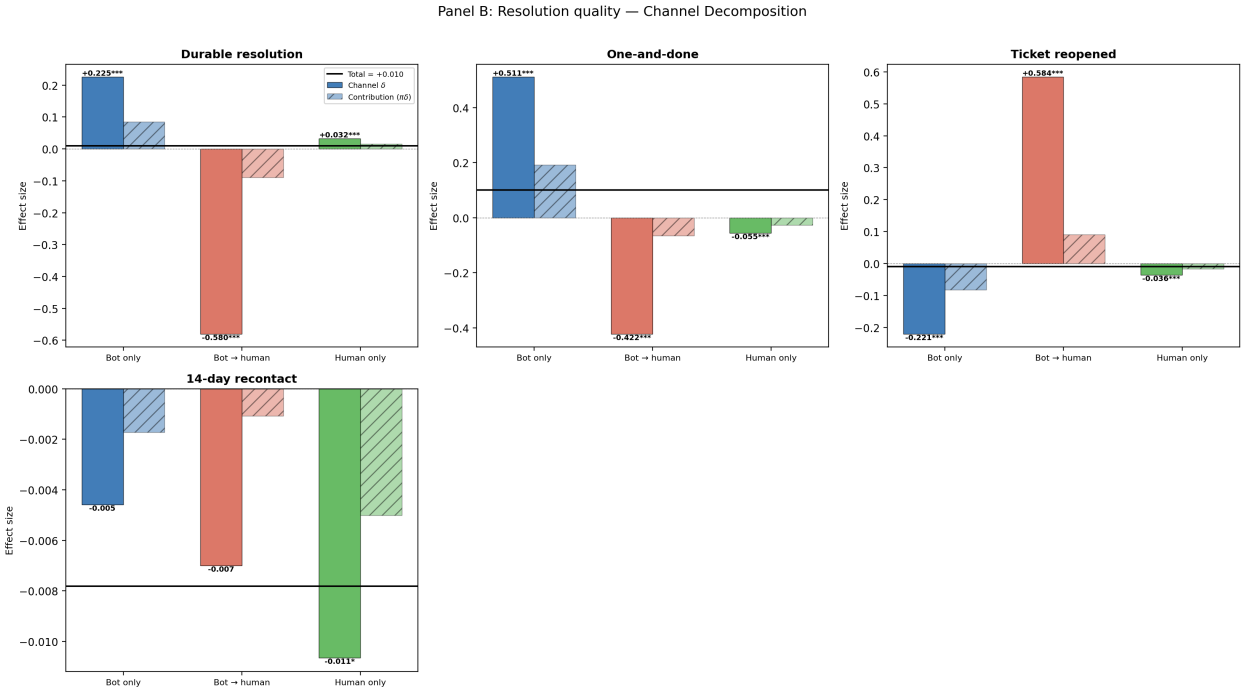


Figure 1: ITT Event Study: HighBot × Month Coefficients

Notes: Coefficients $\hat{\beta}_m$ from $Y_i = \sum_{m \neq 7} \beta_m \cdot \text{HighBot}_i \times \mathbf{1}[\text{month} = m] + \text{HighBot}_i + \text{month FE} + \text{DOW FE} + \text{difficulty FE} + \text{category FE} + \varepsilon_i$, where HighBot is an indicator for above-median predicted bot probability and July ($m = 7$) is omitted (red square). Bands are 95% CI. Pre-period coefficients for speed outcomes are flat and near zero, confirming parallel trends. August shows a small positive spike (copilot deploys on text-similar tickets). September onward: sharp divergence as the bot activates.



(a) Speed and timeliness



(b) Resolution quality

Figure 2: Channel Decomposition of Post-Period Effects

Notes: Solid bars show channel-specific effects (δ_c) from Table 3; hatched bars show share-weighted contributions ($\pi_c \delta_c$). Panel (a): Bot-only tickets are dramatically faster (-2.8 log-points resolution). Handoff tickets are slower on resolution ($+0.6$) but benefit from fast bot FRT (-2.3). Panel (b): Bot-only tickets show strong one-and-done performance ($+0.51$). Handoff tickets exhibit high reopening rates ($+0.58$), consistent with failed bot attempts requiring human rework.

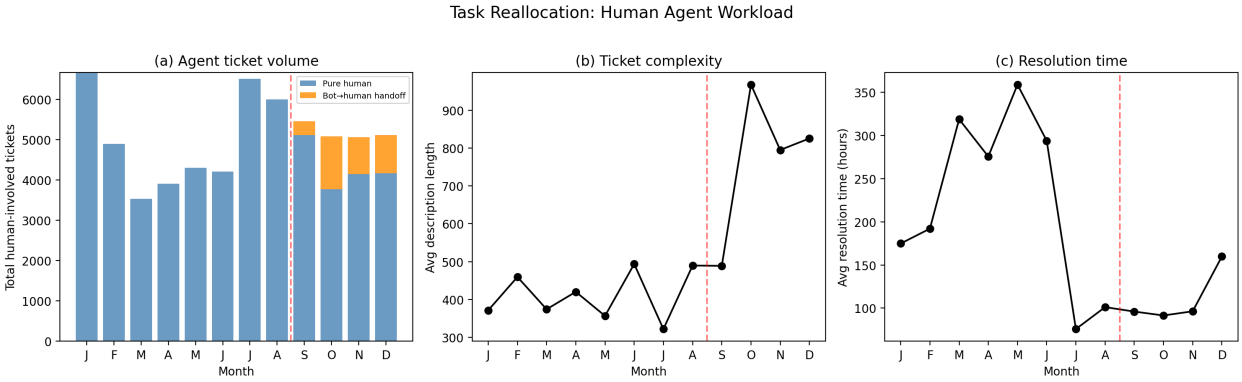


Figure 3: Task Reallocation at the Agent Level

Notes: (a) Human ticket volume drops post-September as the bot absorbs routine tickets. (b) Average ticket complexity, proxied by description length, rises sharply after the bot is deployed. (c) Despite handling harder tickets, human-handled quality holds up, consistent with agents benefiting from reduced workload and copilot assistance.

Table 1: Summary Statistics

	Pre-period	Post-period by channel		
	All human	Bot-only	Handoff	Human-only
Panel A: Sample composition				
Tickets	56,913	8,425	3,851	20,011
Share of post-period		0.26	0.12	0.62
Panel B: Speed and timeliness (medians)				
Resolution time (hours)	30.1	0.0	49.4	27.4
First response time (hours)	24.1	0.0	0.0	22.1
Same-day resolution	0.358	0.964	0.220	0.411
Panel C: Resolution quality				
Durable resolution	0.747	0.985	0.177	0.759
Reopened	0.250	0.015	0.823	0.234
Positive rating	0.346	0.519	0.368	0.353

Notes: Clean sample of $N = 89{,}200$ tickets, January–December 2025; continuous variables winsorized at the 1st/99th percentiles. Pre-period is January–August (all human-handled); post-period is September–December. Bot-only: AI resolved without human intervention. Handoff: AI responded first, then transferred to a human. Human-only: human handled throughout. Resolution and first-response times are medians. Durable resolution: resolved with $\geq$1 public reply and zero reopenings. Positive rating: share of tickets receiving a positive customer-satisfaction rating.

Table 2: Intent-to-Treat Effects of Predicted Bot Handling

	Continuous propensity		Binary (above/below median)	
	(1)	(2)	(3)	(4)
Panel A: Speed and timeliness				
log(resolution time)	−1.652***	−1.816***	−0.528***	−0.590***
	(0.076)	(0.076)	(0.029)	(0.029)
log(FRT)	−1.989***	−2.055***	−0.641***	−0.668***
	(0.062)	(0.062)	(0.025)	(0.025)
Same-day resolution	+0.274***	+0.285***	+0.084***	+0.090***
	(0.020)	(0.021)	(0.008)	(0.008)
SLA breach	+0.015	−0.115***	+0.008	−0.031***
	(0.015)	(0.015)	(0.006)	(0.006)
Panel B: Resolution quality				
Durable resolution	+0.112***	+0.118***	+0.029***	+0.031***
	(0.018)	(0.019)	(0.007)	(0.007)
One-and-done	+0.266***	+0.275***	+0.082***	+0.086***
	(0.021)	(0.021)	(0.008)	(0.008)
Ticket reopened	−0.102***	−0.103***	−0.025***	−0.026***
	(0.018)	(0.019)	(0.007)	(0.007)
Panel C: Customer satisfaction				
Rating	+0.013	+0.067***	−0.006	+0.011
	(0.018)	(0.018)	(0.007)	(0.007)
Month, DOW, difficulty FE	Yes	Yes	Yes	Yes
Category FE	No	Yes	No	Yes
Observations	86,589			

Notes: Each cell reports $\hat{\beta}$ on PredBot × Post from equation (3). Columns (1)–(2): predicted probability (continuous, 0–1). Columns (3)–(4): indicator for above-median predicted probability. Post activates by category at the first month with bot rate ≥ 20%. HC1 robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Variable definitions. Resolution time: hours from ticket creation to resolution (winsorized at p1/p99). FRT: hours to first agent reply. Same-day: resolution within 24 hours. SLA breach: resolution past contractual deadline. Durable resolution: resolved with ≥1 public reply and zero reopenings. One-and-done: same-day, zero reopenings, ≤1 agent reply. Rating: customer satisfaction mapped to $\{-1, 0, +1\}$, non-respondents coded 0.

Table 3: Channel Decomposition of Post-Period Effects

		Post decomposed by channel		
	Post	Bot only	Handoff	Human only
Post-period share		0.26	0.12	0.62
Panel A: Speed				
log(resolution time)	−0.702***	−2.821***	+0.616***	+0.175***
	(0.026)	(0.025)	(0.029)	(0.026)
log(FRT)	−0.780***	−2.299***	−2.348***	+0.491***
	(0.022)	(0.019)	(0.020)	(0.022)
Same-day resolution	+0.081***	+0.479***	−0.238***	−0.066***
	(0.007)	(0.007)	(0.010)	(0.008)
Panel B: Quality				
Durable resolution	+0.008	+0.225***	−0.580***	+0.032***
	(0.007)	(0.007)	(0.009)	(0.007)
One-and-done	+0.071***	+0.511***	−0.422***	−0.055***
	(0.007)	(0.007)	(0.007)	(0.008)
Ticket reopened	−0.009	−0.221***	+0.584***	−0.036***
	(0.007)	(0.007)	(0.009)	(0.007)
Panel C: Satisfaction				
Rating	+0.034***	+0.049***	−0.016	+0.038***
	(0.007)	(0.008)	(0.011)	(0.007)
Month, DOW, difficulty, category FE		Yes		
Observations		86,589		

Notes: Each row reports coefficients from two specifications of equation (4). Column 1: single Post dummy. Columns 2–4: three channel-specific post dummies estimated jointly. δ_c measures the FE-adjusted mean outcome for channel c relative to the pre-period (all-human) baseline. HC1 robust SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: The Anatomy of a Failed Handoff

	Bot handoff	Counterfactual human	Std. diff. (d)
Panel A: First response—behavioral gate and ask coverage			
Customer re-contacts after first response	0.89	0.29	—
Fraction of customer's asks resolved	0.33	0.49	−0.37
Asks deferred (count)	0.52	0.35	+0.25
Asks ignored (count)	0.50	0.36	+0.19
Any ask contradicted	0.40	0.33	+0.14
Panel B: Share of handoff penalty explained by first-response text			
	Text (F_1)	+ Ask cov.	
log resolution time	0.18	0.27	
Ticket reopened	0.22	0.24	
Customer rating	0.63	0.79	
Customer re-contact	0.07	0.08	

Notes: Bot-handoff cohort ($n = 3{,}851$ post-period tickets that began at the bot and were escalated to a human) versus a counterfactual-human cohort ($n = 16{,}233$ pre-period human-first tickets with high predicted handoff propensity). Panel A reports means of LLM-scored first-response features and the standardized (Cohen's d) gap; "re-contacts" is the share of tickets drawing a customer reply after the first response (a 60-point gap, reported as levels rather than a standardized difference). Panel B reports the explained share $1 - \beta_{\text{resid}}/\beta_{\text{raw}}$ of each outcome's handoff penalty after controlling for first-response text features (F_1) and, in the second column, additionally for ask-coverage counts.

Table 5: Within-Customer Panel: Difference-in-Differences

	All tickets		Human-only	
	β_{Post}	$\beta_{\text{Post}\times\text{Treat}}$	β_{Post}	$\beta_{\text{Post}\times\text{Treat}}$
log(resolution)	-0.129^{***}	-1.039^{***}	-0.070^{**}	-0.315^{***}
	(0.032)	(0.038)	(0.032)	(0.039)
log(FRT)	-0.339^{***}	-1.062^{***}	-0.149^{***}	-0.167^{***}
	(0.025)	(0.031)	(0.024)	(0.031)
Same-day	$+0.061^{***}$	$+0.139^{***}$	$+0.053^{***}$	$+0.005$
	(0.010)	(0.012)	(0.010)	(0.013)
SLA breach	-0.017^{*}	-0.057^{***}	-0.018^{*}	-0.013
	(0.010)	(0.010)	(0.010)	(0.012)
Rating	-0.067^{***}	$+0.066^{***}$	-0.064^{***}	$+0.067^{***}$
	(0.008)	(0.009)	(0.008)	(0.010)
Customer FE	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes
Observations	27,599		20,340	
Customers	6,505		4,845	

Notes: Repeat customers with tickets in both pre (Jan–Aug) and post (Sep–Dec) periods. Customer FE via demeaning. β_{Post}: common time trend (captures routing fixes, seasonal effects). $\beta_{\text{Post}\times\text{Treat}}$: additional effect for tickets in automated categories (DiD estimate). HC1 robust SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Automation versus Augmentation: Bot and Copilot Effects

	Bot	Copilot (human-handled tickets only)			
	(1)	(2)	(3)	(4)	(5)
Sample	All	Human	Human	Human	Human
Period	All	All	Pre-bot	Post-bot	All
Treatment	PredBot×Post	PredCop	PredCop	PredCop	PredCop×Post
Panel A: Speed					
log(resolution)	−1.816***	+0.428***	+0.442***	+0.704***	+0.047
	(0.076)	(0.034)	(0.040)	(0.071)	(0.083)
Same-day	+0.285***	−0.135***	−0.109***	−0.283***	−0.130***
	(0.021)	(0.011)	(0.013)	(0.027)	(0.030)
Panel B: Quality					
Durable resolution	+0.118***	−0.160***	−0.150***	−0.182***	+0.184***
	(0.019)	(0.011)	(0.013)	(0.025)	(0.027)
Reopened	−0.103***	+0.172***	+0.162***	+0.200***	−0.147***
	(0.019)	(0.011)	(0.013)	(0.025)	(0.027)
Panel C: Satisfaction					
Rating	+0.067***	+0.071***	+0.070***	+0.083***	+0.095***
	(0.018)	(0.009)	(0.010)	(0.025)	(0.026)
Category FE	Yes	Yes	Yes	Yes	Yes
Observations	86,589	74,645	55,112	12,346	74,645

Notes: Column (1): ITT coefficient on PredBot × Post (all tickets, preferred specification from Table 2). Columns (2)–(5): human-handled tickets only. Column (2): PredCopilot level (all months). Column (3): PredCopilot, pre-bot months (1–8). Column (4): PredCopilot, post-bot months (10–12). Column (5): PredCopilot × Post, testing whether copilot returns change after bot deployment. All specifications include month, DOW, difficulty, and category FE. HC1 robust SE. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Model Predictions and Empirical Evidence

	Prediction	Model source	Empirical support
P1	Speed improvement	Displacement	Bot-only: $\Delta\log(\text{res}) = -2.82^{***}$
P2	Task reallocation	Displacement	Human complexity +69%
P3	Dose-response	Productivity	Effects in all propensity quintiles
P4	Handoff penalty	Imperfect automation	Handoff reopening $+0.58^{***}$
P5	Declining handoff rate	Bot learning	Handoff rate: 37% $\rightarrow$ 28%
P6	Dynamic quality	Bot learning	Durable res. strengthens Oct–Dec
P7	Workload spillover	Capacity relief	Human-only $\Delta\log(\text{res}) = -0.32^{***}$
P8	Augmentation	Copilot (α, β)	Rating = +0.07 to $+0.10^{***}$
P9	Complementarity	$\partial\overline{\alpha}/\partial I > 0$	PredCop×Post durable $+0.18^{***}$
P10	Complexity gradient	$\beta'(\theta) > 0$	Same-day: hard +0.36 vs. easy +0.18
P11	24/7 availability	$\tau_B \approx 0$ always	FRT: off-hours −2.16 vs. bus. −1.96

Notes: P1–P3 arise from the standard Acemoglu–Restrepo (2018, 2019) displacement and productivity channels. P4–P7 require the handoff and workload-relief extensions. P8–P11 involve the augmentation and complementarity channels. All coefficients significant at 1% unless noted. Full results in Tables 2–6 and Appendix Tables A4–A5.

Appendix

A Proof of Proposition 1

Proof. The average copilot time-saving on retained tasks is

$$\overline{\alpha}(I) = \frac{1}{1 - F(I)} \int_I^{\bar{\theta}} \alpha(\theta)\, f(\theta)\, d\theta.$$

Differentiating with respect to I using the quotient rule:

$$\begin{aligned}
\frac{d\,\overline{\alpha}}{dI} &= \frac{-\alpha(I)\, f(I)\,[1 - F(I)] - \big[-f(I)\big] \int_I^{\bar{\theta}} \alpha(\theta)\, f(\theta)\, d\theta}{[1 - F(I)]^2} \\
&= \frac{f(I)}{[1 - F(I)]^2} \left[\int_I^{\bar{\theta}} \alpha(\theta)\, f(\theta)\, d\theta - \alpha(I)\,[1 - F(I)] \right] \\
&= \frac{f(I)}{1 - F(I)} \left[\frac{1}{1 - F(I)} \int_I^{\bar{\theta}} \alpha(\theta)\, f(\theta)\, d\theta - \alpha(I) \right] \\
&= \frac{f(I)}{1 - F(I)} \big[\overline{\alpha}(I) - \alpha(I) \big].
\end{aligned}$$

Since $\alpha'(\theta) > 0$, for any $\theta > I$ we have $\alpha(\theta) > \alpha(I)$. Therefore

$$\overline{\alpha}(I) = E[\alpha(\theta) \mid \theta > I] > \alpha(I),$$

and hence $d\overline{\alpha}/dI > 0$. The identical argument applies to $\overline{\beta}(I)$, establishing that both time-saving and quality-improving augmentation contributions are increasing in the automation threshold. □

B Additional Figures



Figure A1: Monthly Outcome Means by Predicted Bot Probability Group

Notes: Above- vs. below-median predicted bot probability, no controls. Shaded region marks post-deployment. The two groups track closely pre-deployment and diverge sharply after, confirming that the ITT effects reflect a post-deployment break.

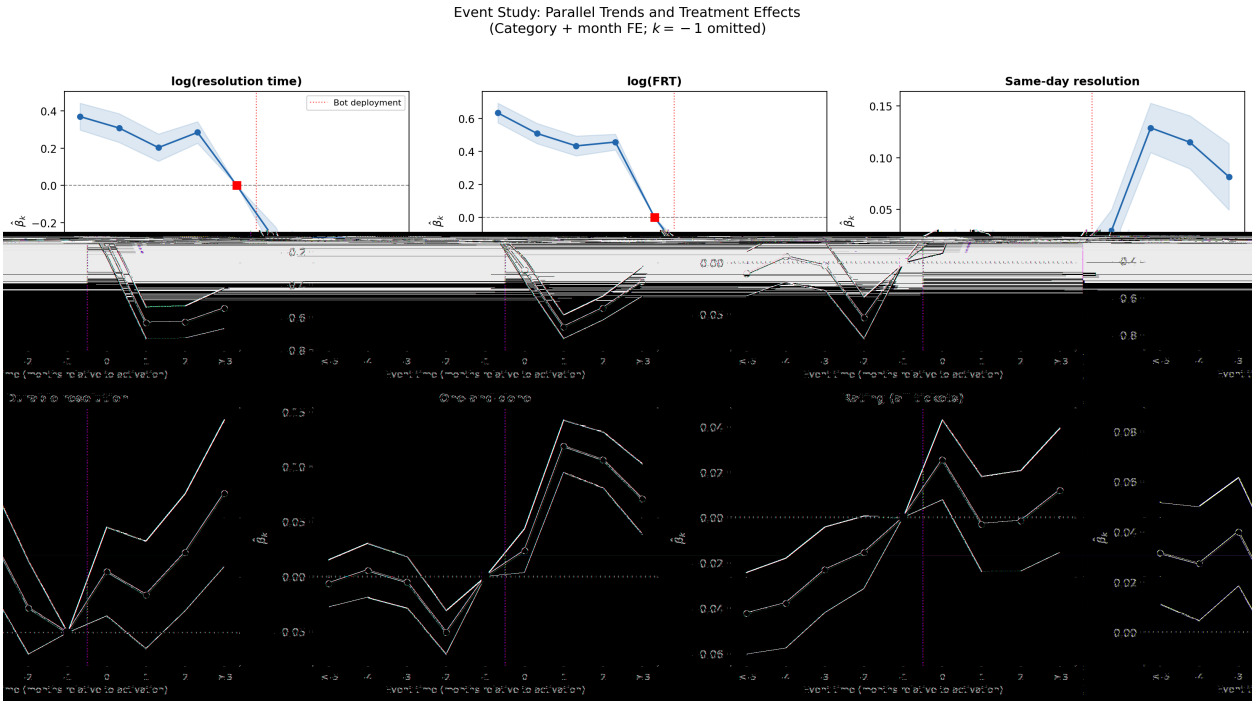


Figure A2: Category-Level TWFE Event Study

Notes: Coefficients on event-time dummies with category and month FE, $k = -1$ omitted. Pre-period coefficients for speed outcomes show a declining trend, indicating that treated categories were already converging toward controls. The post-period break exceeds the pre-trend slope by roughly $3\times$.

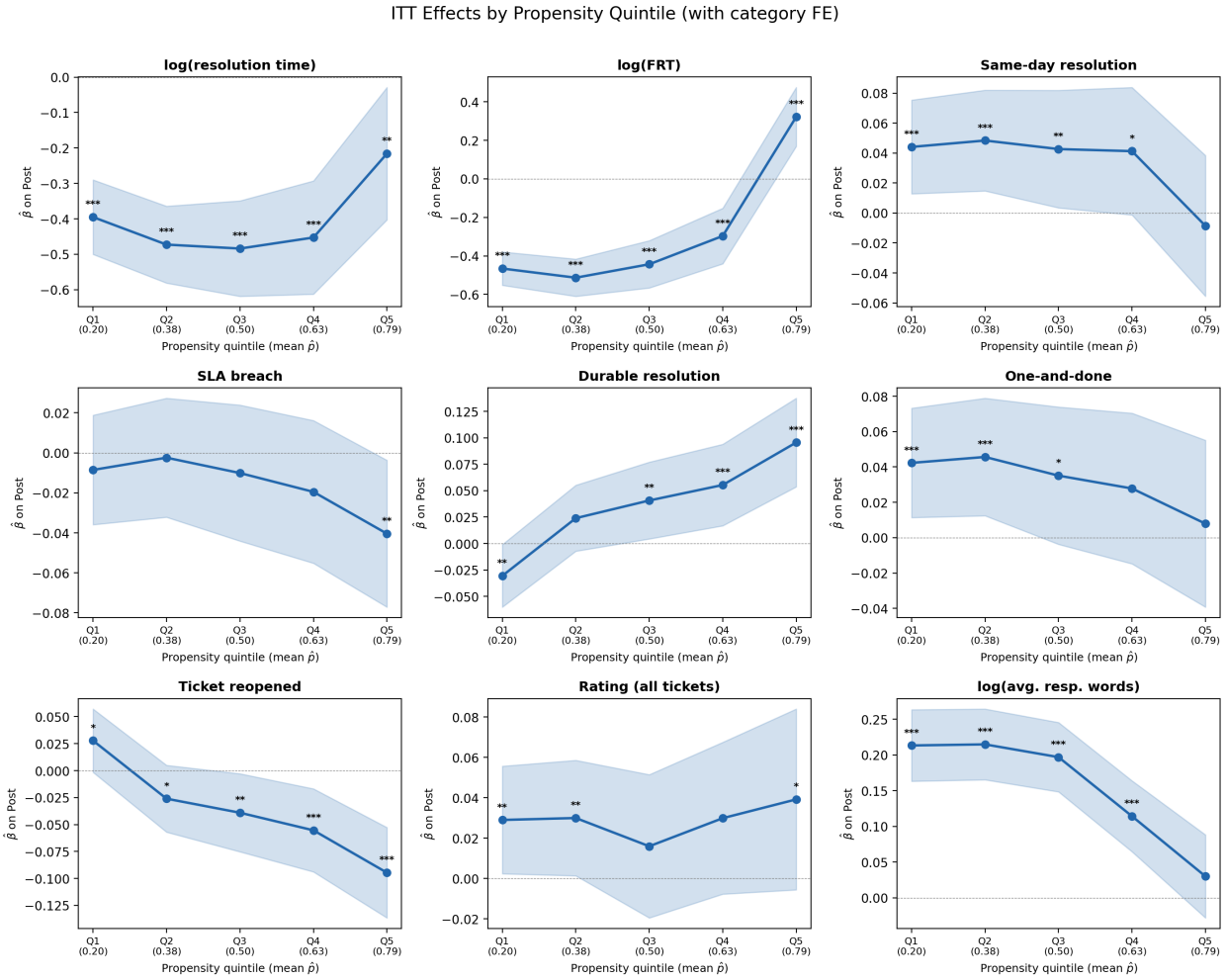


Figure A3: ITT Effects by Propensity Quintile

Notes: Post coefficient estimated within propensity quintile subsamples with category FE, 95% CI. Speed effects are present across all quintiles, not just the top quintile where bot handling concentrates, suggesting indirect benefits from workload relief.

Figure A4: Dynamic Handoff Rate by Month

Notes: Aggregate handoff rate declines from 37% (October) to 28% (December), consistent with bot calibration/learning (P5).

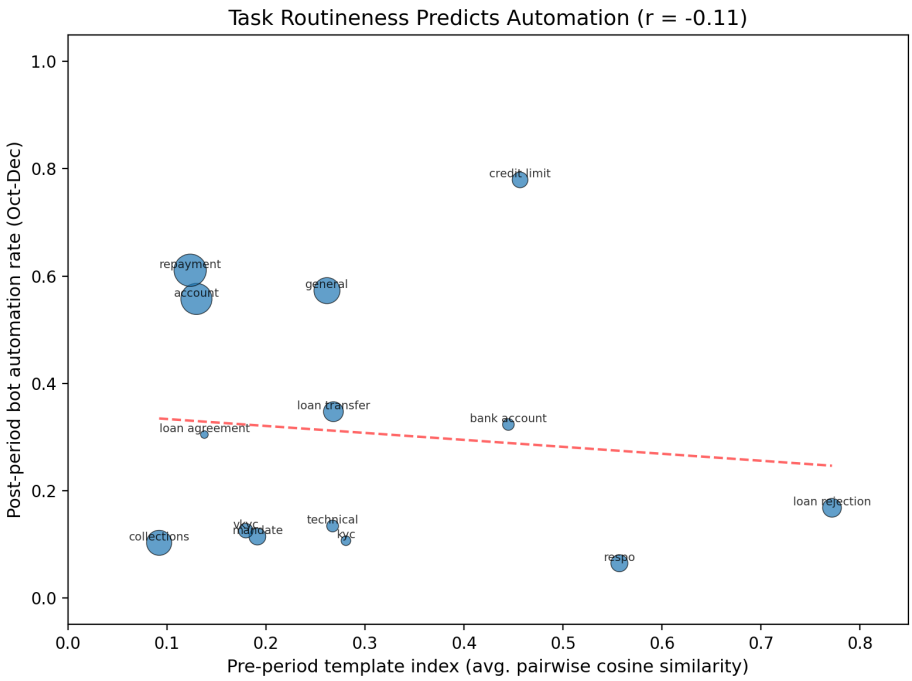


Figure A5: Pre-Period Templateness Does Not Predict Automation

Notes: Categories where human agents gave near-identical responses (high template index) are not necessarily the ones automated. This rules out the concern that pre-period task routineness, rather than the bot's NLP classifier, drives the ITT results.

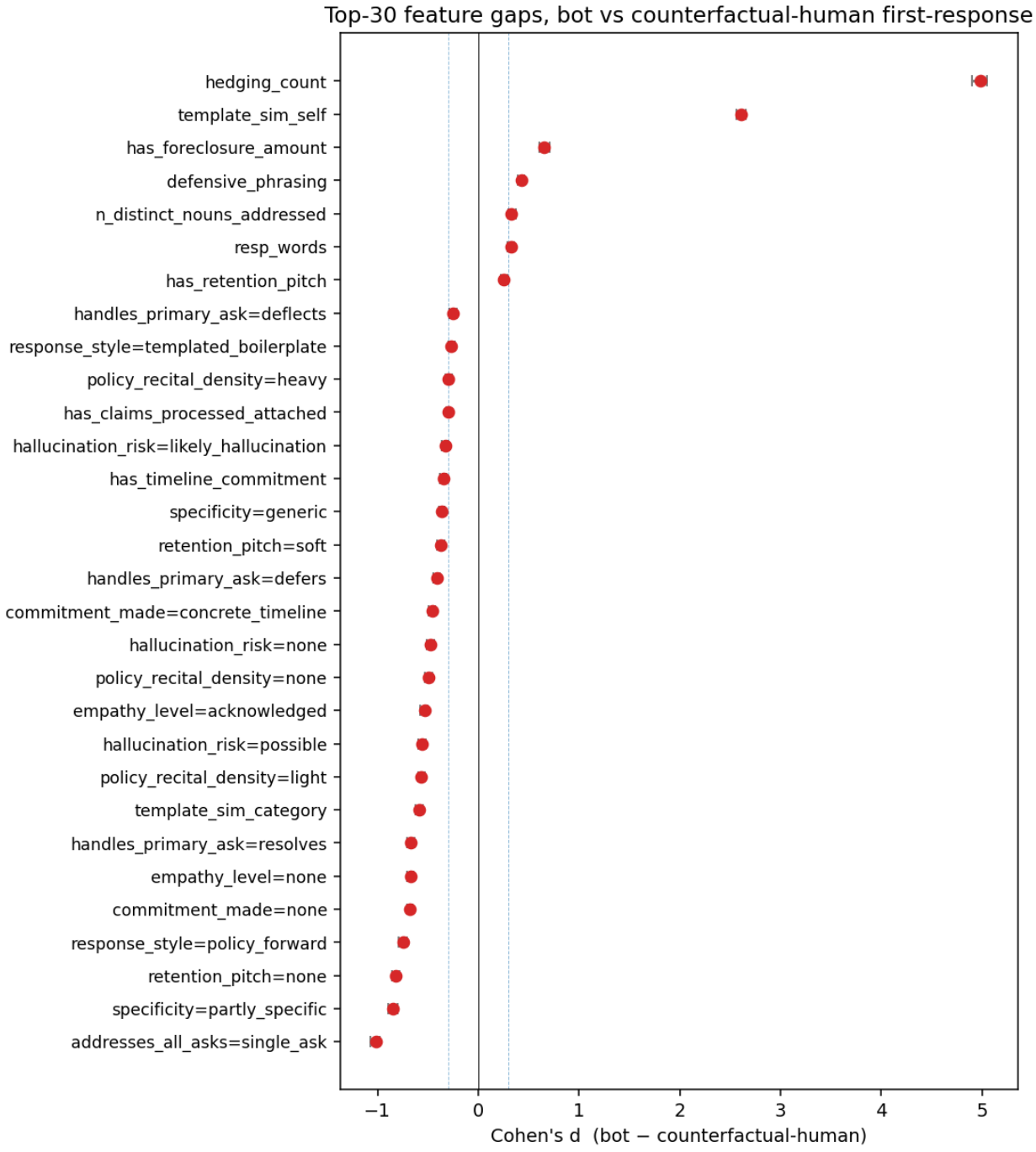


Figure A6: First-Response Feature Gaps: Bot Handoff vs. Counterfactual Human

Notes: Standardized differences (Cohen's d with 95% confidence intervals) in LLM-scored first-response features between the bot-handoff cohort and the counterfactual-human cohort (Section 5.4). Positive values indicate a feature more prevalent in bot first responses. The bot resolves fewer of the customer's asks and defers, ignores, and contradicts them more often, and leans on policy-forward, single-ask replies.

C Additional Tables

Table A1: Propensity Score Model Performance

Model	Month	N	Treatment rate	In-sample AUC	CV AUC	OOS AUC
Panel A: Bot (autopilot)						
	October (train)	7,731	0.496	0.889	**0.775**	—
	November	7,560	0.446	—	—	**0.734**
	December	7,824	0.456	—	—	**0.742**
Panel B: Copilot (augmentation)						
	August (train)	6,568	0.535	0.876	**0.735**	—
	September	7,187	0.603	—	—	**0.765**
	October	3,896	0.359	—	—	**0.774**
	November	4,192	0.429	—	—	**0.818**
	December	4,259	0.435	—	—	**0.823**

Notes: TF-IDF (10,000 features, unigrams and bigrams) + L2 logistic regression. CV = 5-fold stratified cross-validation. OOS = out-of-sample prediction on subsequent months.

Table A2: Net Effect of Automation: Category-Level Difference-in-Differences

	Δ_{treat}	Δ_{ctrl}	DiD	β (Post × Treat)
log(resolution time)	−1.201	−0.204	−0.997	−0.897***
				(0.027)
log(FRT)	−1.483	−0.314	−1.169	−1.127***
				(0.021)
Same-day	+0.209	+0.089	+0.120	+0.099***
				(0.008)
SLA breach	−0.061	−0.080	+0.019	+0.026***
				(0.008)
Rating	+0.009	−0.041	+0.051	+0.049***
				(0.007)
Category, DOW FE			Yes	

Notes: Treatment: categories with Oct–Dec bot rate $\geq 30\%$. Control: bot rate $< 15\%$. β is the regression DiD with FE. HC1 robust SE.

Table A3: 2SLS Estimates: Effect of Bot Handling

	(1) ITT	(2) 2SLS
Panel A: Speed		
log(resolution time)	−1.816***	−2.600***
	(0.076)	(0.099)
log(FRT)	−2.055***	−2.994***
	(0.062)	(0.059)
Same-day	+0.285***	+0.408***
	(0.021)	(0.029)
Panel B: Quality		
Durable resolution	+0.118***	+0.169***
	(0.019)	(0.028)
Reopened	−0.103***	−0.147***
	(0.019)	(0.027)
Panel C: Satisfaction		
Rating	+0.067***	+0.096***
	(0.018)	(0.026)
First-stage $\hat{\gamma}$		0.698*** (0.017)
First-stage F		1,633.7
Category FE	Yes	Yes
N	86,589	86,589

Notes: Column (1) reproduces ITT from Table 2 col. (2). Column (2): 2SLS instrumenting BotHandled with PredBot × Post. The LATE is the effect of bot handling for tickets whose assignment is determined by text characteristics interacted with post-period availability. HC1 robust SE.

Table A4: ITT Effects by Ticket Complexity

	Easy (1–2)	Medium (3)	Hard (4–5)
Sample share	0.47	0.34	0.19
Panel A: Speed			
log(resolution)	−1.541***	−1.817***	−1.763***
	(0.107)	(0.120)	(0.220)
Same-day	+0.176***	+0.310***	+0.358***
	(0.028)	(0.033)	(0.058)
SLA breach	−0.030	−0.201***	−0.236***
	(0.019)	(0.024)	(0.045)
Panel B: Quality			
Durable resolution	+0.021	+0.221***	+0.169***
	(0.025)	(0.030)	(0.055)
Reopened	+0.033	−0.230***	−0.183***
	(0.025)	(0.030)	(0.055)
All FE		Yes	
N		86,589	

Notes: PredBot × Post fully interacted with difficulty groups from AI text classifier. HC1 robust SE.

Table A5: ITT Effects by Time of Day

	Business hours (9am–6pm weekdays)	Off-hours (nights/weekends)
log(resolution)	−1.742***	−1.896***
	(0.104)	(0.103)
log(FRT)	−1.962***	−2.158***
	(0.084)	(0.085)
Same-day	+0.375***	+0.175***
	(0.028)	(0.027)
Durable resolution	+0.141***	+0.095***
	(0.025)	(0.025)
Reopened	−0.120***	−0.086***
	(0.025)	(0.025)
Rating	+0.089***	+0.047*
	(0.024)	(0.025)
All FE	Yes	
N	86,589	

Notes: PredBot × Post fully interacted with business-hours indicator. HC1 robust SE.

Table A6: Dynamic ITT Effects by Post-Period Month

	October	November	December
log(resolution)	−2.012***	−1.502***	−1.913***
	(0.143)	(0.126)	(0.128)
Same-day	+0.338***	+0.219***	+0.287***
	(0.037)	(0.034)	(0.034)
Durable resolution	+0.066**	+0.049	+0.124***
	(0.033)	(0.032)	(0.031)
Reopened	−0.050	−0.033	−0.109***
	(0.032)	(0.031)	(0.031)
SLA breach	−0.071**	−0.113***	−0.114***
	(0.031)	(0.021)	(0.021)
Rating	+0.086***	+0.039	+0.056*
	(0.032)	(0.031)	(0.029)
All FE		Yes	
N		86,589	

Notes: PredBot × Post decomposed into month-specific interactions. HC1 robust SE.

Table A7: Agent-Level Task Reallocation: Difference-in-Differences

	β_{Post}	$\beta_{\text{Post}\times\text{Treat}}$
log(resolution time)	-0.112^{***}	$+0.326^{***}$
	(0.022)	(0.025)
log(FRT)	-0.237^{***}	-0.392^{***}
	(0.018)	(0.023)
Same-day resolution	$+0.076^{***}$	-0.140^{***}
	(0.007)	(0.009)
log(description length)	$+0.192^{***}$	$+0.258^{***}$
	(0.021)	(0.025)
Handoff share	$+0.058^{***}$	$+0.168^{***}$
	(0.003)	(0.005)
Agent FE, Category FE	Yes	Yes

Notes: Human-involved tickets (pure human + handoffs), agents active in both periods. HC1 robust SE.

Table A8: Back-of-Envelope Welfare Calculation

Panel A: Customer-facing time savings	
Pre-period mean resolution (hours)	93.2
Post-period mean resolution (hours)	40.9
Reduction per ticket (hours)	52.3
Bot-only mean resolution	4.6
Handoff mean resolution	80.6
Human-only mean resolution	56.7
Post-period tickets	32,287
Total customer hours saved (post-period)	1,688,610
Panel B: Firm labor cost savings	
Fully-loaded agent cost/ticket (USD)	1.04
Bot-only tickets (zero agent time)	8,425
Post-period labor savings (USD)	8,762
Annualized labor savings (USD)	26,286
FTE equivalents freed	4.3
Panel C: Break-even AI cost	
Break-even per bot ticket (USD)	1.04
Typical LLM API cost range (USD)	0.01–0.10

Notes: Resolution time is clock time (creation to closure), including queue and wait time. Agent cost: INR 25,000/month $\times$ 1.4 overhead / 510 tickets. Break-even equals per-ticket labor saving. At INR 83/USD.