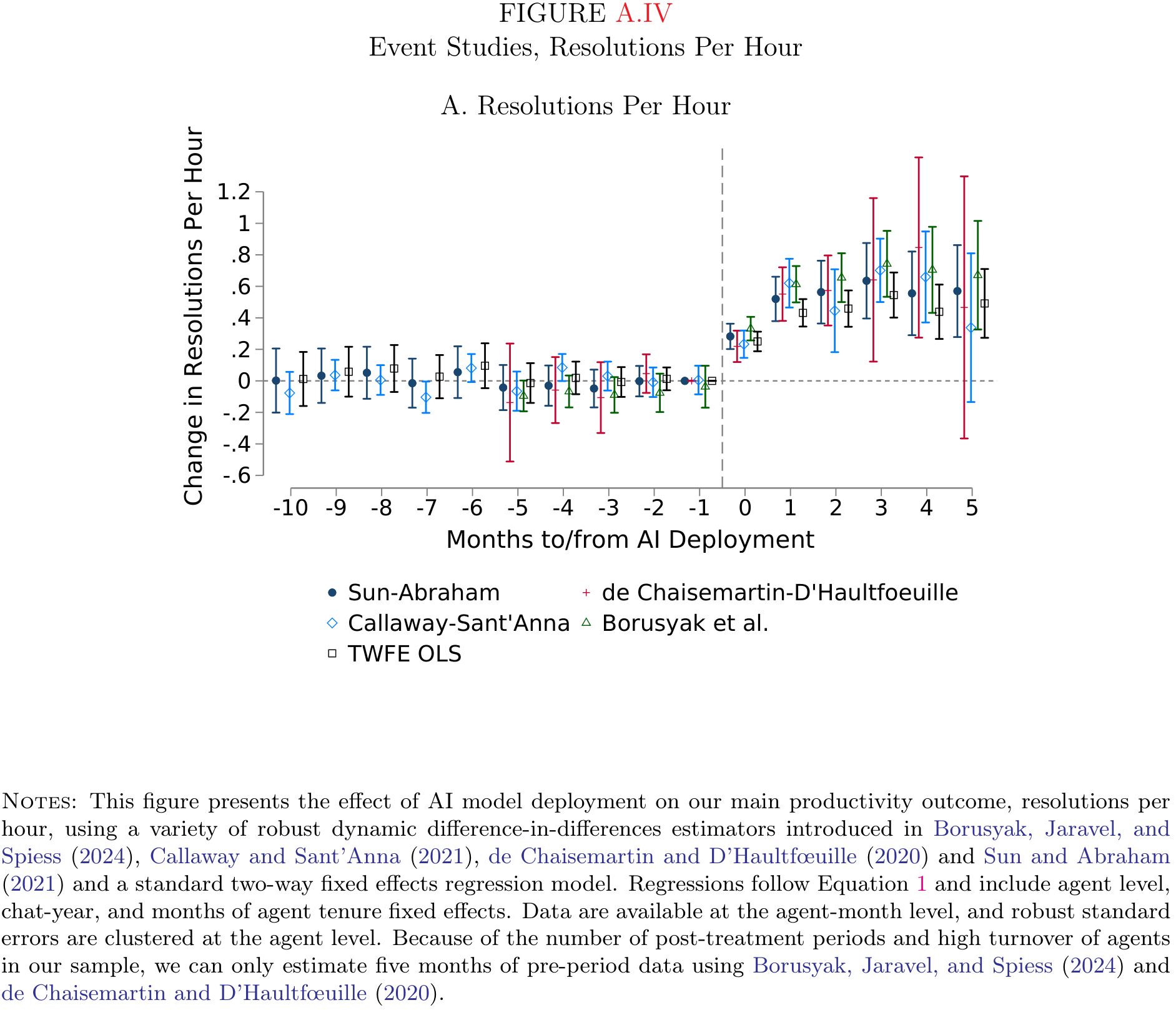
Generative AI at work  
  
*Quarterly Journal of Economics*  
2025 (Volume 133), June (No.6)

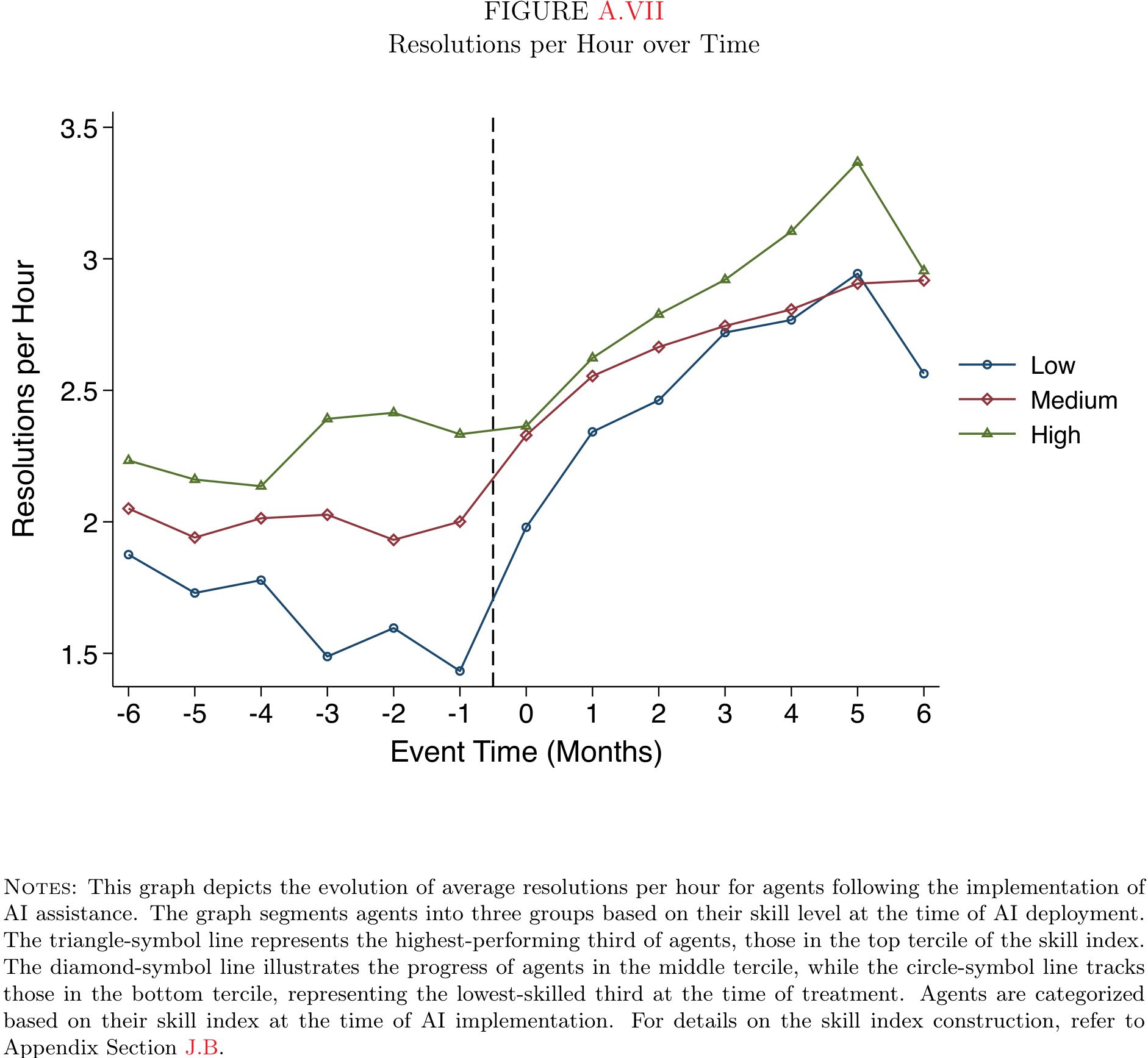
Brynjolfsson, Li, Raymond

August 22, 2025, 11:08 +0000

[doi.org/10.1086/734780](https://doi.org/10.1086/734780)

## TL; DR





* Context: Software customer service chat agents←AI assistance
* Estimation: Staggered rollout DID
* GPT-3 based assistance ⇒ resolutions per hour↑
* Resolutions per hour↑ is largest among the least skilled workers
  + Productivity distribution dispersion↓
* Adherence rate↑ ⇒ productivity↑
* Mechanism:
  + Agents learn usefulness of AI ⇒ adherence ⇒ durable learning
  + Fewer turnovers (quits) of relatively new agents
  + Larger impacts on lower skilled←AI suggestions←high productivity worker data
* External validity:
  + A text-based, stable set of tasks
  + Skill-augmenting/replacing role of AI

## Introduction

Context

* ICT=skill-biased technical change ⇒ high-skilled worker demand↑
* Machine learning ( generative AI) can be skill-augmenting/replacing ⇒ high-skilled worker demand↓(?)
  + ML guesses solutions from data without instructions
  + Inputs (customer query, etc.)→actions for better outcomes
    - Sits well with non-routine tasks
    - White collar skills can be replaced

Will it decrease/increase employment/wages of low/high skilled? Not shown[[1]](#footnote-28)

Using skills augmenting[[2]](#footnote-29) AI chatbots on software customer service agents, the paper shows they replace (a part of) skills on problem diagnosis, knowledge retrieval, and customer communications

* This is exactly what was intended, so no surprise here[[3]](#footnote-30)
* It boosts productivity of low skilled workers more than high skilled
* But this is already shown in previous studies
  + This paper showed more comprehensively in real business environment

Click here to see a summary comparison table with previous work.

| feature |  | Brynjolfsson et al. (2025) | Noy & Zhang (2023) | Peng et al. (2023) | Dell’Acqua et al. (2023) | Choi & Schwarcz (2023) |
| --- | --- | --- | --- | --- | --- | --- |
| setting |  | Field study (Fortune 500) | Online experiment (Prolific) | Online experiment (Upwork) | Field experiment (BCG) | Lab experiment (University) |
| subjects |  | 5,172 cust. service agents | 453 professionals | 95 programmers | 758 elite consultants | 48 law students |
| task contents |  | Real customer support chats | Writing press releases, reports, and emails | Implementing an HTTP server in JavaScript | Creative product ideation and business problem-solving | Multiple-choice and essay questions from law exams |
| skill measurement |  | Objective, longitudinal | Objective, snapshot | Self-reported | Objective, snapshot | Objective, snapshot |
| skill contents |  | Months of real KPIs & tenure. | Grade on one pre-task. | Years of experience. | Score on assessment task. | Score on prior real exam. |
| main impacts (pp.) |  | +15% productivity (RPH) (p.907) | -40% time, +18% quality (p.4) | -56% time (p.5) | +40% quality, +25% speed (p.16) | +29 percentile (MCQ), 0 (essay) (p.18) |
| AI deployment |  | Real-time assistant | One-off use for writing | Pair programmer for coding | Interactive use for consulting | Assistant for exam questions |
| leveling comparison |  | Skill quintiles vs. performance | Grade on task 1 vs. task 2 | Regression on years of exp. | Bottom-half vs. top-half | Baseline percentile vs. change in percentile |
| leveling impact size (pp.) |  | +36% (bottom) vs. 0% (top) in RPH (p.911) | Grade correlation drops 0.41->0.14 (p.5) | Effect varies by exp. (p.6) | +43% (bottom) vs. +17% (top) in quality (p.15) | +45 (bottom) vs. -20 (top) percentile (p.21) |
| top-tier impact (pp.) |  | Null on speed, small negative on quality (p.911) | Null on quality, still reduces time (p.4) | Not explicitly isolated (p.6) | Positive, but smaller gains (+17%) (p.15) | Significant negative (-20 percentile) on essays (p.21) |
| external validity |  | Stable tasks; AI augments knowledge/comms; real stakes. | Creative/writing tasks; AI as a first-draft tool; low stakes. | Standardized coding; AI code completion; time incentives. | Complex knowledge work; AI as a brainstorming partner; high-skill workers. | Formal reasoning tasks; AI as knowledge support; academic setting. |

## LLM use & generalizability

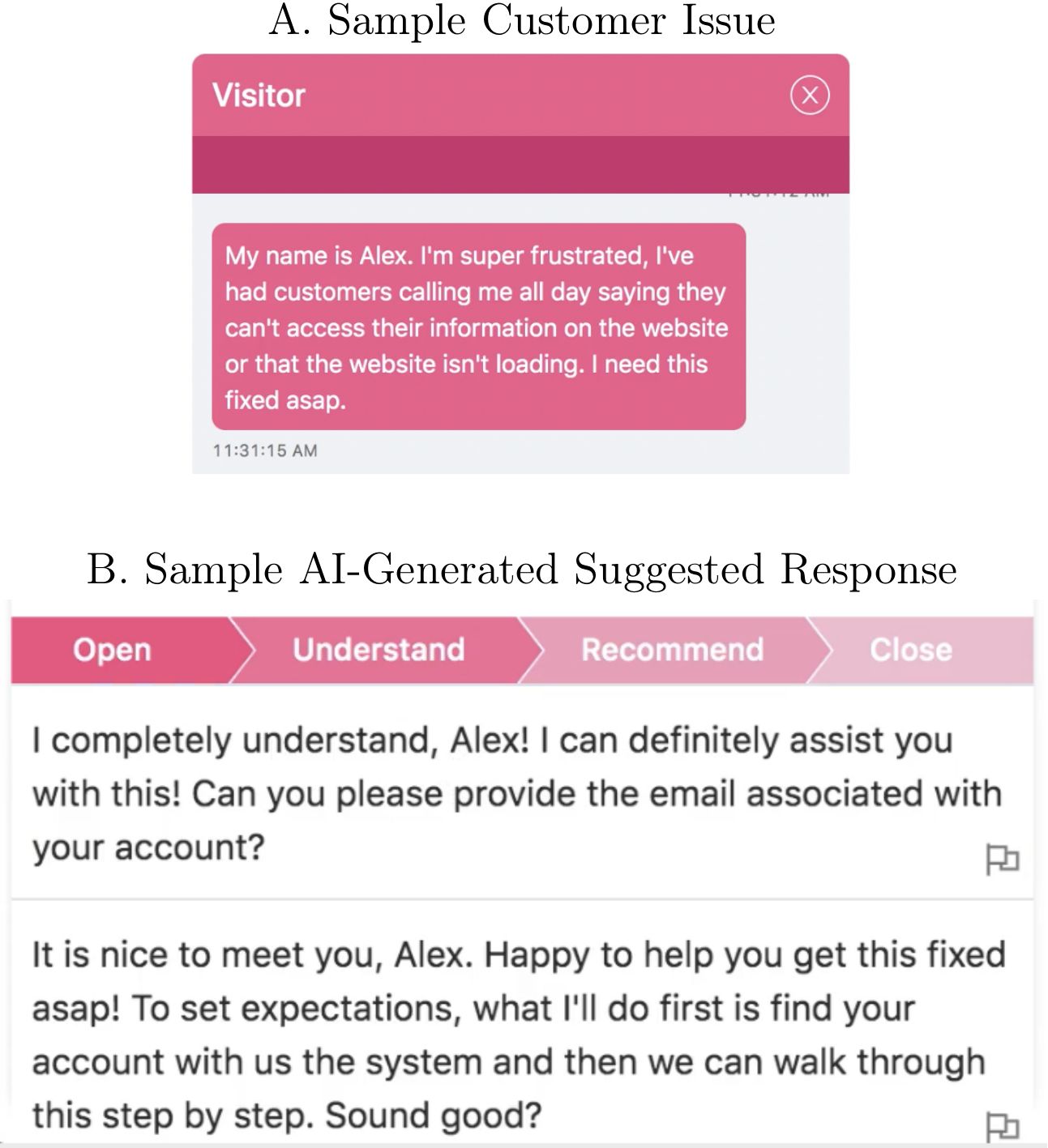
### Training of AI[[4]](#footnote-34)

* Data: Customer support center recordings
* Up-weights top performing agents in training
* Aspects of agent behaviour trained to AI (p.900)
  + when to ask clarifying questions
  + being attentive to customer concerns
  + de-escalating tense situations
  + adapting communication styles
  + explaining complex topics in simple terms
* Priotize agent responses that
  + express empathy
  + provide appropriate technical documentation
  + limit unprofessional language

Usage

* To augment, rather than replace, human agents[[5]](#footnote-35)
* AI gives no advice on insufficiently trained topics

### How an agent uses AI



Chat box

1. Customer sends a message
2. AI analyzes the chat
3. AI displays suggestions to the agent on a separate panel or window
   * Suggested text: Ready-to-use phrases or sentences (e.g., “Happy to help you get this fixed asap”).
   * Suggested links: Links to internal technical documentation relevant to the problem.
4. Agent chooses response
   * Use: Copy-paste
   * Edit: Modify
   * Ignore: Type a completely different response from scratch
   * Learn: Read the doc before typing own response

## Data

Proprietary data from a Fortune 500 firm (“Data Firm”)

* Data Firm sells business-process software
* AI Firm provides the generative AI assistance and data for research
* 5,172 agents, 3M+ chats
  + Employed directly by the Data Firm or by third-party subcontractors
* Agent-chat panel: Chat transcripts, durations, resolution status, customer feedback
  + Main analysis: aggregated to agent-month
  + Outage study: individual chat level
  + Background information on each agent: Tenure, geographic location, employer, team assignment, but no individual pay or wages
* Derived variables
  + Resolutions per hour, average handle time, chats per hour
  + Adherence = 1 if either of below holds:
    1. Direct copy tracking: An exact match to AI’s suggestions
    2. High content similarity: Compares the message vs. AI suggestions[[6]](#footnote-42)
  + Topics: Classified using Gemini
  + Conversation style: Comprehensibility, native (American English) fluency, scored by Gemini
  + Customer sentiments : Measured by using [SiEBERT](https://huggingface.co/siebert/sentiment-roberta-large-english)

## Treatment assignment

* AI tool roll out: Staggered, Fall 2020 - Winter 2021
  + Limited training capacity (small sessions, few trainers)
  + Budgetary limits for the new technology
  + Full sample period: No information
* Treatment Assignment: Team→agent
  + Team selection: No information
  + Agent selection: By team managers
    - Stagger training within a team to minimize service disruption
    - Priority given to higher productive agents←des stats

## Empirical Strategy

### Identification

* Robust DID←staggered rollout of the AI tool
* No pre-trend (Fig II), but selective treatment assignment

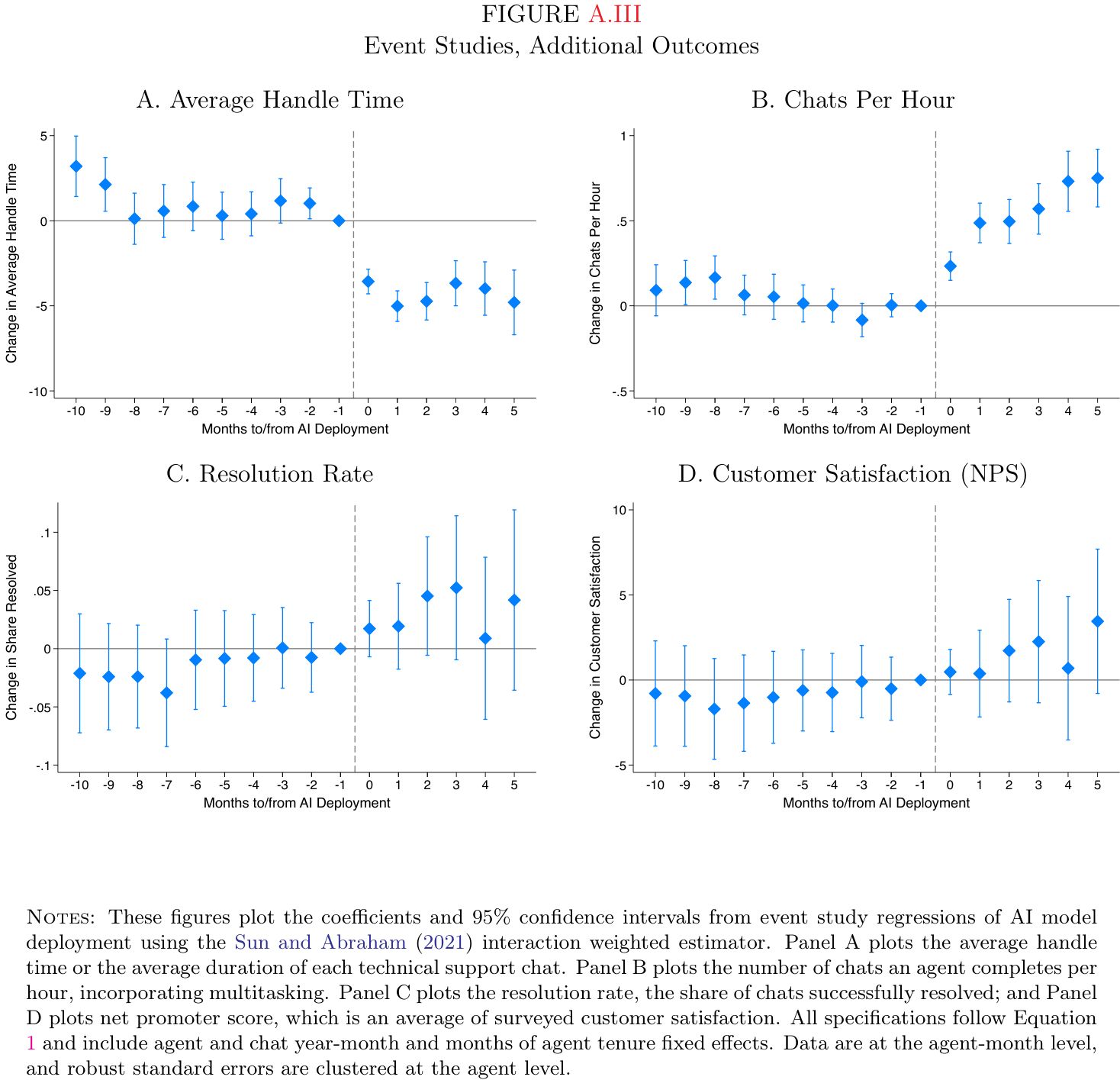
### Estimation

: Productivity quantile, overall topic frequency quantile, agent’s topic frequency quantile, adherence quantile

* Sun and Abraham ([2021](#ref-SunAbraham2021)) estimator using never-treated as control
  + Callaway and Sant’Anna ([2021](#ref-CallawaySantAnna2021)) uses not-yet-treated as control
  + de Chaisemartin and D’Haultfœuille ([2020](#ref-dCDH2020)) is for stayers
  + Borusyak, Jaravel, and Spiess ([2024](#ref-BorusyakJaravelSpiess2024)) is imputation based, more model oriented
  + Results are qualitatively similar

## Results

### Main impacts



|  |
| --- |
| Fig VIII |

Fig VIII

Overall Productivity (Table II, Figure II)

* Resolutions Per Hour (RPH) ↑ 15%.
* Average Handle Time (AHT) ↓ 8.5%.
* Chats Per Hour (CPH) ↑ 15% (more multitasking) (Table III).

### Heterogeneity by skill & experience, by topic

* By skill (Figure III)
  + Lowest-skill agents: RPH ↑ 36%
  + Highest-skill agents: No gain. Small decrease in quality
* By experience (Figure IV)
  + Newest agents see largest gains
  + Experienced agents (>1 year) see no gain
  + Faster learning (Figure V): An agent with 2 months of AI experience is as productive as an agent with 6+ months of experience without AI
* By topic frequency (Figure VIII)
  + U-shaped: Moderately rare topics had the biggest impacts
    - Rarity↑ ⇒ sophistication↓ ← fewer data to train AI
    - Rarity↑ = more room for improvements

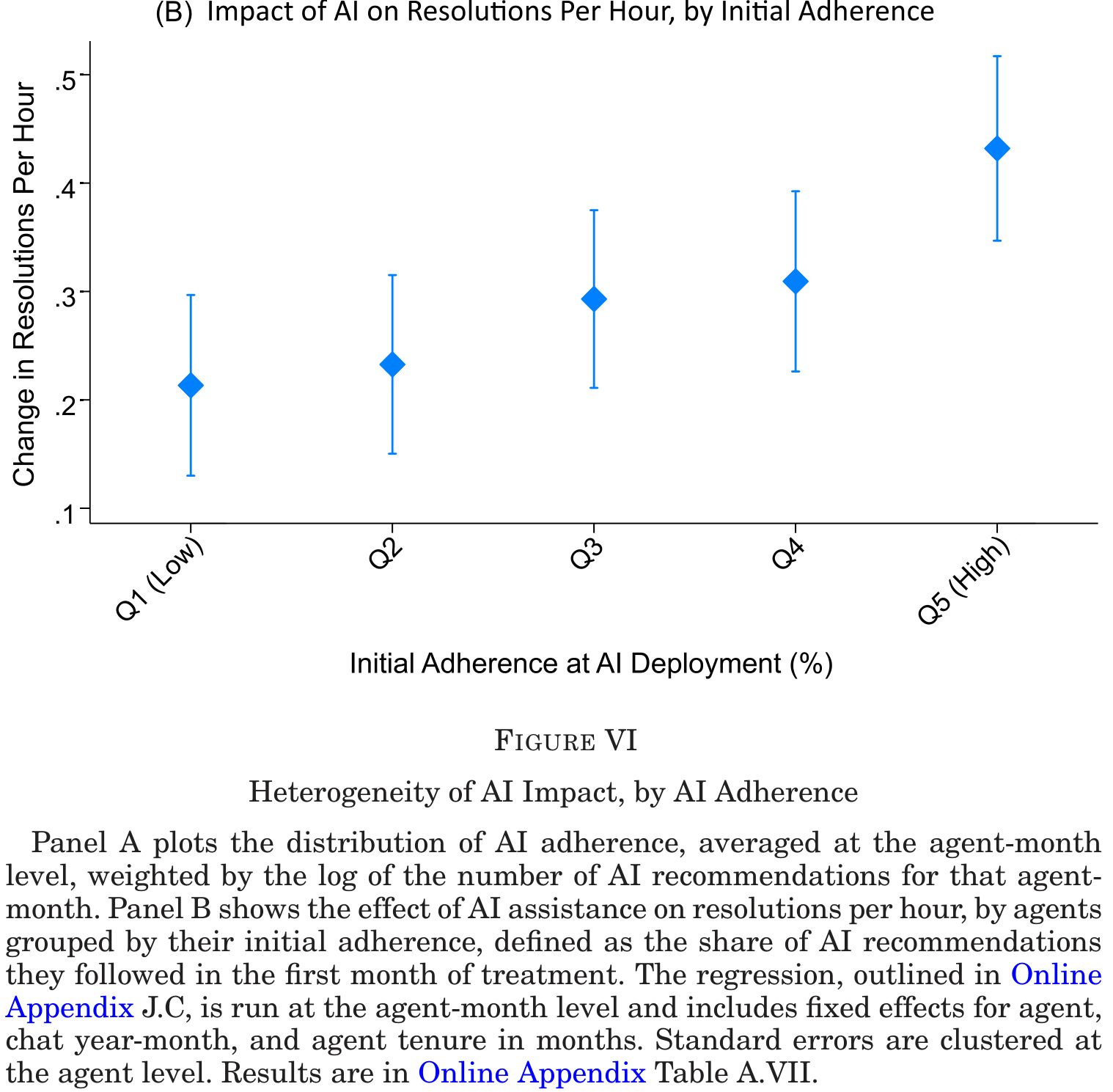
### Other effects

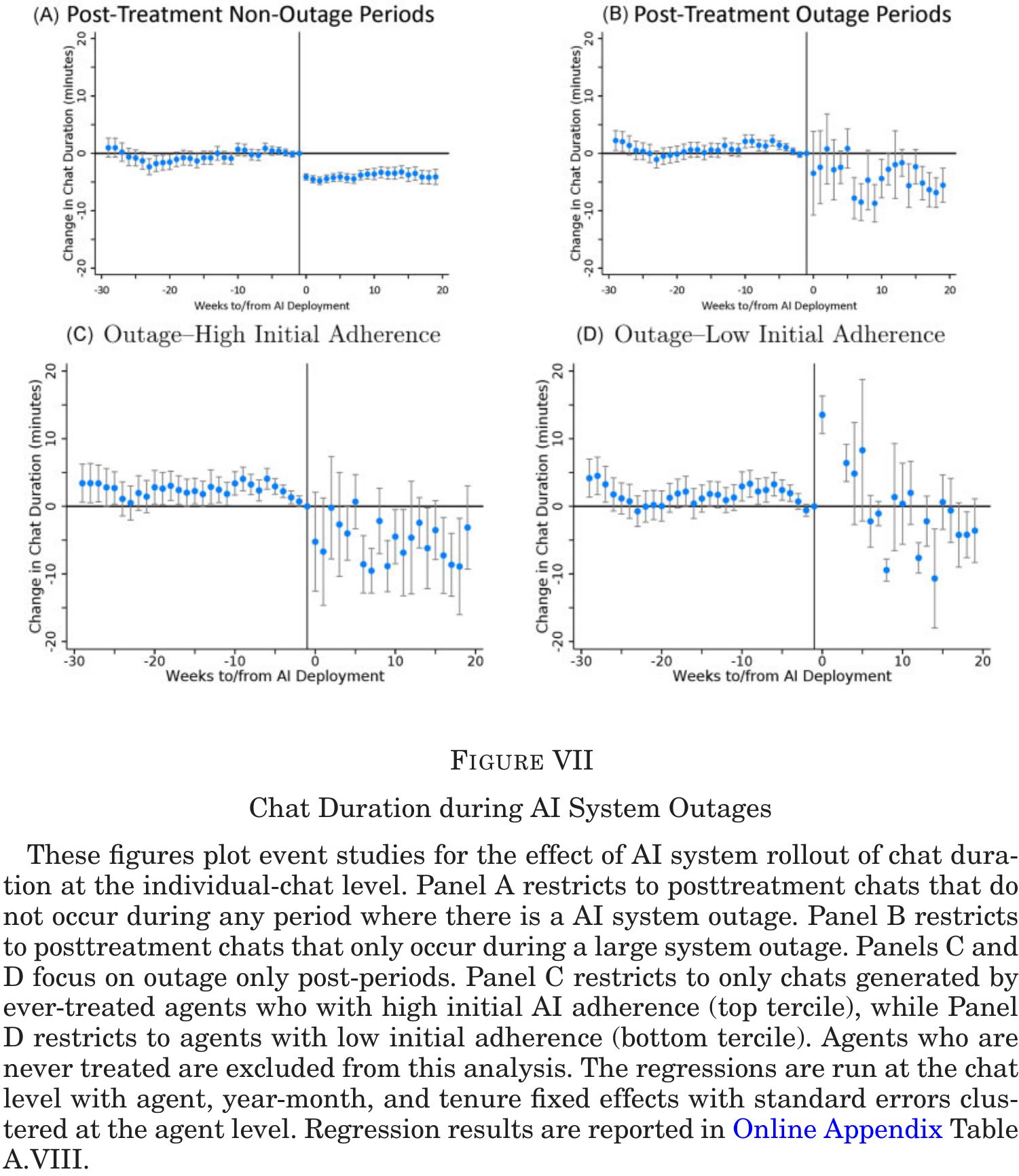
Experience of work

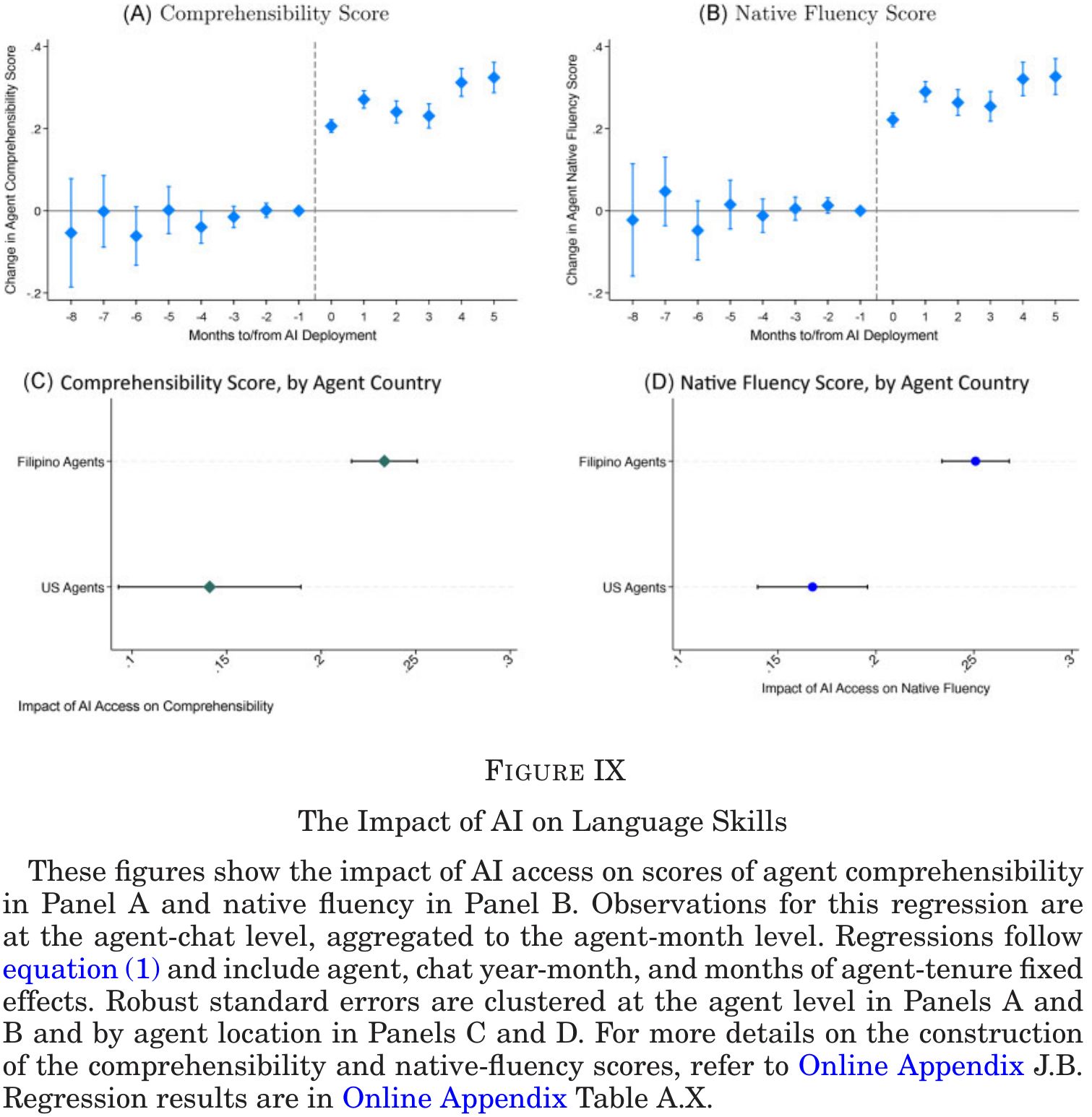
* Customer sentiment: Customers are more positive and polite (Figure X, Table IV)
* Escalations: Requests to “speak to a manager” ↓ 25% (Figure X, Table IV)
* Attrition: Employee turnover ↓8.7%, more pronounced for new workers

## Mechanisms

Pathways







1. Adherence: adherence↑ ⇒ productivity↑ (Figure VI)
2. Durable learning: Productivity gains persist even during AI outages (Figure VII)[[7]](#footnote-65)
3. Communication: English fluency↑ (Figure IX), low-skill agents communicate more like high-skill agents (textual convergence)[[8]](#footnote-66)

## Conclusion

* Early empirical evidence on the effects of a generative AI tool in a real-world workplace
* AI-generated recommendations:
  + Increases overall worker productivity by 15%
  + Larger effects for lower-skill and novice agents
  + Improve worker on-the-job experiences
* Productivity gains reflect durable worker learning

## 感想

* AI利用⇒労働生産性、に関する本格的な効果推計←みんな待ってた
* 最大の貢献: AIが熟練を代替(熟練労働賃金への示唆)
  + skil leveling effects
  + 格差縮小(?)…ではないかも
* chatレヴェルの詳細なデータを得たのが素晴らしい…質(resolutions)と量(chats, handling time)へのインパクトを示し、AI効果の理解に貢献
* 「長期への懸念」も指摘: 労働生産性格差がなくなり、熟練への報酬がいずれ下がるので、トレーニング・データを提供していた高技能労働者のデータ提供誘因が弱まる
  + AIが高技能労働者を駆逐すると、環境が変わったときに成功事例を開拓して学習材料を人がいなくなる
* Outage study (falsification test)はメカニズムを検証する賢い検定
* しかし、durable learningは推計結果がはっきりしない
* 低生産性エージェントはコピペしているだけかも
  + 模範解答を反復して覚えてしまう
  + 学習…か?
  + それでいいかも
* 外的妥当性
  + これは非定型nonroutineタスクか? 問答の類型routine化は可能だが、ケース分けが多すぎてマニュアルにするのがすごく大変というだけでは?
    - 受け身: 労働者はプロンプトを出さないので、受動的にgo with the flowでアドバイスを取り込んでいる気がする
    - AI=自動でアドバイスをくれる過保護な上司的存在
    - AI利用=包括的マニュアルを作成する費用、検索する費用、表現調整する費用を劇的に下げていると理解可能
  + AIへの入力(データ)と作業指示を(問題とその解決方法に応じて)人間が決めるunstructured tasksでの影響とは違う
    - 非テキスト・ベースのタスク=uncodifiable
    - manage employees, raise capital, pilot new initiatives, run advertising strategies, price their services, react to competitors, and decide which of these and myriad other tasks to focus their efforts on (Chandler, 1977, quoted from Otis et al. 2023)
    - [Otis et al. (2023)](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4671369)では優秀な経営者のみAIの利潤効果が正、それ以外は負

### Derivation of the Occupation-Level Production Function

This document explains why “linear aggregation ensures that the Cobb–Douglas form reemerges at the occupation level” by deriving the occupation-level production function from the worker-level function, step by step.

#### **The Building Blocks (Equations and Assumptions)**

1. **Worker-level output (Equation 5):** This is the output produced by a single worker who is given units of capital.
2. **Aggregation Rule:** The total output of the occupation, , is the linear sum (integral) of the outputs of all individual workers employed in that occupation.
3. **Optimal Capital Allocation:** To maximize total output, the total capital for the occupation, , is distributed uniformly among all workers.

#### **Step-by-Step Derivation**

1. **Start with the aggregation rule.**
2. **Substitute the worker-level production function into the integral.**
3. **Substitute the optimal capital per worker, .**
4. **Pull the constant term (the entire bracketed expression) out of the integral.**
5. **Evaluate the remaining integral, which is simply the total number of workers, .**
6. **Substitute this result back into the main equation.**
7. **Rearrange the terms using algebra to group labor () and capital () terms.** This combines several small algebraic steps for clarity.
8. **Combine the terms that share the same exponent to achieve the final form.**

This derivation continues from the previous result and shows how to rearrange it into the compact Cobb-Douglas form with a Total Factor Productivity (TFP) term, A(φ).

#### **Starting Point**

From the previous derivation, we established the occupation-level production function as:

#### **Target Equation**

Our goal is to show that this is equivalent to the standard form:

where A(φ) is the occupation-specific TFP term.

#### **Step-by-Step Algebraic Rearrangement**

1. **Distribute the exponents.** We apply the exponent on the outside of each parenthesis to both the numerator and the denominator inside, using the rule .
2. **Separate the input factors L(φ) and K(φ).** In the second term, we can expand to .
3. **Group the non-input terms together.** Let’s rearrange the equation to group all terms that are *not* or at the beginning. These terms constitute the productivity parameter.
4. **Re-combine the grouped terms to match the paper’s definition of A(φ).** The term in the brackets can be written more cleanly by grouping the bases that share the same exponent. This makes the structure clearer.
5. **Define the Total Factor Productivity (TFP) term, A(φ).** We can now see that the entire expression inside the large brackets is the occupation-specific TFP, A(φ). It captures the efficiency of production for a given occupation φ, which depends on the capital share α(φ) and the productivity of capital η.
6. **Substitute A(φ) back into the main equation.** By replacing the complex bracketed term with the simpler A(φ), we arrive at the final, compact Cobb-Douglas form.

This completes the derivation. We have successfully shown that the linear aggregation of worker-level outputs, under the model’s assumptions, results in a standard Cobb-Douglas production function at the occupation level.

Autor, David, and Neil Thompson. 2025. “Expertise.” *Journal of the European Economic Association* 23 (4): 1203–71. <https://doi.org/10.1093/jeea/jvaf023>.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. “Revisiting Event-Study Designs: Robust and Efficient Estimation.” *The Review of Economic Studies* 91 (6): 3253–85. <https://doi.org/10.1093/restud/rdae007>.

Callaway, Brantly, and Pedro H. C. Sant’Anna. 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225 (2): 200–230. https://doi.org/<https://doi.org/10.1016/j.jeconom.2020.12.001>.

de Chaisemartin, Clément, and Xavier D’Haultfœuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110 (9): 2964–96. <https://doi.org/10.1257/aer.20181169>.

Sun, Liyang, and Sarah Abraham. 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics* 225 (2): 175–99. https://doi.org/<https://doi.org/10.1016/j.jeconom.2020.09.006>.

1. See Autor and Thompson ([2025](#ref-AutorThompson2025)) for a theory [↑](#footnote-ref-28)
2. Hence replacing high skilled workers [↑](#footnote-ref-29)
3. p.935: In areas where the product or environment is changing rapidly, the relative value of AI recommendations may be different. … Indeed, recent work by [Perry et al. (2023)](https://arxiv.org/abs/2211.03622) and [Otis et al. (2023)](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4671369) have found cases in which AI adoption has limited or even negative effects. [↑](#footnote-ref-30)
4. Chat GPT-3 based [↑](#footnote-ref-34)
5. No matter how it is expressed, this is exactly how replacement works: By substituting expertise with AI [↑](#footnote-ref-35)
6. Not shown explicitly, but probably used cosign similarity

   where is a vector of 0 or 1 for all the words [↑](#footnote-ref-42)
7. true? estimates are too noisey [↑](#footnote-ref-65)
8. P values are not shown for textual convergence [↑](#footnote-ref-66)