NBER WORKING PAPER SERIES

EMPOWERING INCLUSIVE WORK

Yanyou Chen
Mitchell Hoffman
Huilan Xu
Zhe Yuan

Working Paper 35372
http://www.nber.org/papers/w35372

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

We thank Ach Adhvaryu, Amanda Agan, Naoki Aizawa, Heski Bar-Isaac, Gordon Dahl, Chiara Farronato, Kory Kroft, Peter Kuhn, Rafael Lalive, Imke Reimers, Kathryn Shaw, Joe Vecci, and especially David Autor for helpful comments. We also thank numerous seminar participants, especially those at NBER Labor Studies, for useful suggestions. We thank Xiaonan Zhu, Weiyue Zhang, Ruibo Sun, Yaqi Du, Yu Xu, Yibing Xia, and Sunyu Ye for excellent research assistance. We thank Schmidt Sciences and the National Natural Science Foundation of China [Grants 72141305, 72192803] for support. This project has been supported by the Gates Foundation (investment ID INV-044136), through a subgrant from IDinsight's Digital Economy Research Impact Initiative (DERII). We received ethics approval from the University of Toronto. One author has previously received compensation from the Platform for work unrelated to the research reported in this paper. We thank the Platform and its employees for their collaboration and for providing data. We thank delivery workers for participating in the qualitative data collection. Any opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily represent the views of the Platform. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Empowering Inclusive Work
Yanyou Chen, Mitchell Hoffman, Huilan Xu, and Zhe Yuan
NBER Working Paper No. 35372
June 2026
JEL No. J14, M50

ABSTRACT

Can AI improve workplace outcomes for workers with disabilities? We examine the relative performance of deaf or hard of hearing (DHH) workers on one of China's largest food-delivery platforms. Pre-AI, DHH workers are slower than non-disabled workers and have worse customer ratings, although they supply more hours to the platform and are less likely to quit. Midway through our data, the platform suddenly introduces an AI-based intelligent outbound calling system designed to improve customer communication for DHH workers. Using a difference-in-differences design comparing DHH and non-disabled workers before and after the AI tool, we find that AI increases the speed and productivity of DHH workers, especially on tasks involving customer interaction; substantially reduces negative customer ratings; and increases labor supply and retention. AI eliminates one-third of the disability hourly pay gap and significantly increases the profitability of DHH workers for the platform.

Yanyou Chen
University of Toronto
yanyou.chen@utoronto.ca

Mitchell Hoffman
University of California, Santa Barbara
Department of Economics
and NBER
hoffman2@ucsb.edu

Huilan Xu
Zhejiang University
xuhuilan@zju.edu.cn

Zhe Yuan
Zhejiang University
School of Economics
yyyuanzhe@zju.edu.cn

Around the world, workers with a disability face challenges in the labor market (Baldwin and Johnson, 2006; L'Horty et al., 2022). They have lower rates of employment and face substantial discrimination (Bellemare et al., 2023), and there are large fiscal costs from high use of disability insurance (Autor and Duggan, 2003; Milligan, 2012). Challenges faced by disabled workers[1] are especially acute in middle-income countries like China where government social insurance is often less developed (Banerjee et al., 2024).

Unfortunately, relatively little is known regarding "what works" in improving outcomes for disabled workers. A central policy to increase opportunity for disabled workers is civil rights laws, a key example being the Americans with Disabilities Act (ADA). The ADA prohibits discrimination against disabled workers and mandates firms to provide reasonable accommodations, and similar laws exist in many countries. However, research indicates that the ADA *decreased* the employment of disabled workers, potentially due to increased costs or legal risks for employers (Acemoglu and Angrist, 2001; DeLeire, 2000). Another common policy in many countries (e.g., Austria, Brazil, China) is to impose quotas, requiring larger firms to employ a minimum share of workers with disabilities. However, firms may comply with quotas by hiring disabled workers into lower-paid jobs (Szerman, 2022; Lalive et al., 2013). Rather than imposing compliance costs through civil-rights enforcement or quotas, an alternative is to rely on technologies like artificial intelligence (AI) that directly raise productivity—potentially increasing inclusion while aligning with firms' incentives.

Our study provides the first evidence from a large-scale setting on disabled workers' productivity and the effects of a targeted AI technology. We address two questions:

1. In the same job, how do disabled workers compare to non-disabled workers in terms of productivity, labor supply, output, wages, and profit for firms?

2. Can AI tools affect these outcomes for disabled workers and how?

To address these questions, we leverage a unique collaboration with a major food delivery platform in China (hereafter, the "Platform"). The Platform offers an ideal setting for three reasons. First, the Platform operates on an immense scale: it is one of the two dominant food-delivery apps in China, directly engaging millions of active workers and completing hundreds of millions of deliveries per month. Second, the Platform follows an open hiring policy: all applicants who meet basic criteria are accepted, with no differential standards for disabled workers (Ameri et al., 2018). Thus, disabled workers are employed under the same conditions as their non-disabled counterparts, allowing for a cleaner comparison of on-the-

[1] While some use the *person-first* "worker with a disability," many disability scholars prefer the *identity-first* term "disabled worker" (Gernsbacher, 2017; Grech et al., 2024). We use both, but often the latter. About 1 in 6 people worldwide have a significant disability (World Health Organization, 2023).

job performance. Third, work on such platforms has become a key source of employment for disabled workers in China.

We focus on workers who are deaf or hard of hearing (DHH), who constitute the majority of disabled workers on the Platform, and compare their outcomes to those of non-disabled workers. People with different disabilities face different challenges and may be affected by AI in different ways. Focusing on a single, common disability allows us to more clearly examine the relevant challenges. Moreover, hearing impairment is a disability that gig platforms around the world increasingly seek to accommodate, yet important challenges remain—making our findings broadly relevant.[2]

Our analysis draws on a rich administrative dataset from the Platform (Section 1), covering all newly registered DHH workers and a random sample of newly registered non-disabled workers over a three-year period (2021–2023). This dataset allows us to directly compare workplace outcomes of DHH workers to non-disabled workers on many dimensions. The data include worker productivity variables capturing both the quantity of work, such as speed at different stages of order fulfillment, and work quality, as captured by customer ratings. The data also include labor supply variables on both the intensive and extensive margins, as well as information pertaining to worker compensation and Platform profits. In the middle of our observation period in mid 2022, the Platform introduces an AI-based "Intelligent Outbound Calling" feature designed to help DHH workers communicate with customers. We propose a simple conceptual framework about differences in workplace outcomes between disabled and non-disabled workers, and about how those differences are affected by AI tools.

Prior to the new AI tool, DHH workers are slightly slower than non-disabled workers (Section 2). This gap is especially pronounced in the final stage of order fulfillment—delivery (i.e., leaving the restaurant to customer drop-off), which involves interaction with customers—where DHH workers are 5% slower. DHH workers are substantially more likely to experience "left tail" outcomes, with 9% more late deliveries and 31% more bad customer ratings (though bad customer ratings are rare). Despite this, DHH workers have substantially higher labor supply: they work 5% more hours per week and are 41% less likely to quit. The higher labor supply of DHH workers more than offsets their small efficiency gap, with DHH workers completing more orders per week and generating higher weekly profits for the Platform than non-disabled workers. However, because DHH workers work slower, receive more bad ratings, and work more hours relative to income earned, they earn 10% less per

[2]About 8% of Chinese adults have moderate to severe hearing loss (Wang et al., 2025). China's rate is broadly in line with the World Health Organization (2025) global prevalence rate of 5% for disabling hearing loss (i.e., hearing loss of more than 35 decibels). Chen et al. (2024a) report that there are nearly 100,000 registered DHH couriers in China. Rates of US hearing impairment vary, from 3.7% of all Americans (Erickson et al., 2026) to 13% of US adults with some impairment (Madans et al., 2021).

hour than non-disabled workers, broadly similar to disability wage gaps in other settings.

As detailed in Section 3, the AI tool we study is a text-to-speech phone calling system offering preset and customized audio prompts, and that uses AI to make calls sound natural. Rolled out suddenly to all DHH workers—but not to non-disabled workers—we exploit this as a natural experiment to estimate, via difference-in-differences (DiD), how the AI tool affects workplace outcomes for DHH workers. The AI tool causes DHH performance to substantially improve. DHH work speed improves, driven by faster completion of the trip segment involving customer communication, where they lagged the most previously. They also have far fewer late deliveries and bad ratings, with AI fully eliminating the pre-AI gap in late deliveries and cutting the gap in bad ratings by two-thirds. Orders increase, and so do hours and retention. These improvements translate into higher profitability of DHH workers, with Platform gains far larger than costs. AI causes the disabled worker pay gap to fall from 10% to under 7%, a decrease of one-third. While DHH workers with non-profound (i.e., less severe) hearing loss also benefit, gains are largest among those with profound hearing loss. Overall, the results support that the key mechanism for effects is improved communication with customers, as opposed to alternatives like increasing customer sympathy for DHH workers. We also tend to find larger gains for DHH workers in stronger local labor markets, particularly on labor supply and total output.

This paper contributes to several literatures, the first one being that on the economics of disability. A large body of work studies the determinants and consequences of disability insurance and other social safety programs for disabled workers (Bound and Burkhauser, 1999; Deshpande, 2016; Autor et al., 2019; Deshpande and Lockwood, 2022). There is also work on disabled worker pay gaps (Forth and Jones, 2025), as well as estimating the impact of civil rights laws and quotas for disabled workers, but relatively little is known about what works in improving labor outcomes for disabled workers.[3] An important exception is Bloom et al. (2025) who find that the large rise in work from home has increased the employment of disabled workers. We are not aware of prior work in any academic field that compares disabled and non-disabled workers on a large scale in terms of individual productivity data.[4] The intervention we study, a simple AI-based communication tool, has the largest impact

[3]Lise et al. (2023) find the ADA reduced employment for workers with physical limitations but increased it for those with non-work limitations like mental illness. Using changes in workers' compensation programs, Aizawa et al. (2025) show that disability accommodations increase disabled worker earnings and employment. Derbyshire et al. (2024) survey the impact of employer-focused interventions on disability employment.

[4]This reflects data constraints rather than lack of broad scholarly interest in understanding disabled worker productivity and challenges. Human–computer interaction scholarship uses interviews to examine challenges faced by DHH workers on gig platforms, including Uber (Lee et al., 2019) and food delivery in China (Chen et al., 2024a). The Uber study is based on interviews with five drivers, supplemented by qualitative analysis of in-app feedback, while the China study relies on interviews with six couriers and four station managers. Neither analyzes large-scale administrative data nor examines the impact of AI tools.

on closing the disability wage gap of any intervention that we are aware of.

Second, it contributes to work on the economics of AI. While AI use is accelerating, causal evidence on its impact for disadvantaged groups is scarce. As surveyed by Fruits and Stout (2026), work examines how AI affects outcomes for different types of workers, including accountants (Choi and Xie, 2025), translators (Merali, 2024), call center agents (Brynjolfsson et al., 2025), professionals in laboratory settings (Noy and Zhang, 2023), and software engineers (Cui et al., 2025; Hoffmann et al., 2024). In contrast, our study is the first to evaluate how a targeted AI tool affects workplace outcomes for disabled workers. In doing so, it provides evidence of how AI can help workers and reduce inequality (Autor, 2024; Acemoglu et al., 2026). Our heterogeneous treatment effects suggest that disability may critically differ from other dimensions of workplace inequality like lower productivity or lower tenure, as we find heterogeneous effects by disability severity but not for those dimensions. While generally not based on administrative data and usually lacking quasi-experimental variation in technology, research on personal computers suggests some role of technology in improving disabled worker outcomes (Kruse et al., 1996; Krueger and Kruse, 1995). It also connects to broader work on IT and productivity in firms (Bartel et al., 2007).

Third, our paper contributes to work in personnel and organizational economics on demand for and returns to disadvantaged workers. While there is strong interest in studying hiring returns in general (Oyer and Schaefer, 2011; Bloom and Van Reenen, 2011), there has been relatively scant work on firm returns from disadvantaged workers (Hoffman and Stanton, 2025). One recent exception is Benson et al. (2024), who analyze productivity differences between white and minority workers in a large US retail firm. There is also work on policies and firm returns in regard to hiring workers with a criminal record (Minor et al., 2018; Cullen et al., 2023). To our knowledge, ours is the first paper to examine the *profitability* of disabled workers for an organization or platform. Pre-AI, while DHH workers are less efficient than non-disabled workers, they are still more profitable overall. This highlights the importance of capturing a broad set of outcome variables in evaluating disabled worker profitability. An analyst who focused only on efficiency variables, and ignored the labor supply variables of hours and retention, would draw mistaken conclusions about the relative profitability of DHH workers, a point we return to in the Conclusion (Section 4).

1 Context, Conceptual Framework, and Data

The Platform is one of the two leading companies in China's massive online food-delivery market. It operates similarly to US apps like Uber Eats or DoorDash but on a much larger scale. Together, the two main Chinese platforms engage around 10 million delivery workers at a time (Lin, 2025), making this a labor force larger than that of some countries.

Delivery workers on the Platform use motorbikes or e-bikes to transport food, groceries, and other items from restaurants, convenience stores, and pharmacies to customers. The Platform is available nationwide, with a higher concentration of workers in urban and eastern regions of China where food delivery is most prevalent.

Our engagement with the Platform. Our engagement with the Platform on this project began in late 2023. One of the coauthors, who had received an AI call from a DHH worker, reached out to the Platform about the possibility of evaluating the impact of the AI tool. While the Platform frequently evaluates the impacts of changes to new customer features or other app improvements, it had never evaluated the impact of the AI tool for DHH workers, but seemed keen to better understand its impact.[5]

"Hiring" policy. The Platform maintains an open registration policy (referred to hereafter as "hiring," though drivers are independent contractors). To begin working, an applicant must meet three requirements: be at least 18 years of age, possess valid identification, and have no disqualifying criminal record. These standards are uniform across applicants, with no separate criteria for workers with disabilities.

Disability status on the Platform. Across both major Chinese food delivery platforms, most disabled workers are DHH. The Platform defines a worker as being disabled based on the driver uploading their *certificate of disability* during registration. The certificate is a formal, government-issued document requiring medical evaluation by doctors and is used to determine disability insurance benefits. Appendix D provides a screenshot of the disability question during the registration process.

During registration, both pre- and post-AI, drivers are asked (1) what type of disability, if any, they have, (2) the disability level, (3) whether they consent to the Platform sharing their disability with customers so that there is more understanding in unexpected situations. Disability level for DHH workers is coded based on the Chinese national four-grade system (GB/T 26341-2010), with further details on Chinese disability levels in Appendix C:

- Level 1, which is hearing loss of greater than 90 decibels, abbreviated as > 90 dB HL. This is profound hearing loss, where someone is unable to hear even a shouted voice, and spoken speech is distorted (Olusanya et al., 2019). Extremely serious barriers occur regarding participation in social life (GB/T 26341-2010).

- Level 2 or 81-90 dB HL. This is severe hearing loss.

[5] The Platform's driver operations team that we collaborate with does much less policy evaluation than other teams within the Platform.

- Level 3 or 61-80 dB HL. This is moderately severe hearing loss.
- Level 4 or 41-60 dB HL. This is moderate hearing loss; individuals at this level can typically hear words only when spoken in a raised voice at close range, making unaided phone communication difficult (Olusanya et al., 2019).

Based on our qualitative work and understanding of the Platform, essentially all certified DHH workers disclose their disability to the Platform, a point we discuss further in Appendix D. Drivers know their level since it is listed on their certificate of disability.

Both before and after the introduction of the AI tool, customers are informed via the app that their driver is DHH, for those drivers who consented to this disclosure during registration. Though we do not have precise data on this, based on our conversations with the Platform, we believe consent rates are high. Despite most DHH workers electing to disclose their disability to customers, customers may not always attend to this information. There is a general expectation in China that drivers call customers upon arrival, and this can pose communication challenges for DHH workers.

Work models: team vs. gig. Workers on the Platform can choose between two work models: a *team* model, in which workers are assigned to stations and are paid under an accelerating piece-rate schedule, with unit prices that increase stepwise in the number of monthly completed orders, and a *gig* model, in which workers operate independently and are paid per delivery. Exact pay is determined algorithmically and varies across deliveries, with longer deliveries typically paying more. The team model is intended to be full-time while still allowing workers to choose their hours, whereas work in the gig model can be part time. Both gig and team workers are included in our data (about 2/3 are gig). We control for work model in our analyses, and present key results separately for gig and team workers.

Work model is also referred to as registration type. If a worker wants to switch work models, they need to re-register. Based on our conversations with the Platform, we believe that relatively few workers switched work models during the period of our data.

In both models, drivers tend to work for one delivery platform at a time (Lin, 2025).[6]

Matching of workers to deliveries. Gig and team workers also differ in order assignment. Gig workers primarily rely on independently selecting orders from the order pool.[7] Gig workers operate across the entire city, often covering multiple commercial districts. In

[6] Team workers are prohibited from simultaneously working for other food-delivery platforms. Gig workers face no explicit prohibition, but working for another platform reduces their order history on this Platform, lowering their matching score.

[7] While gig workers do receive some system-assigned orders, these are preferentially assigned to team workers under the same conditions. On average, about 20% of gig workers' orders come from the system, while the remaining 80% are manually selected by the workers themselves.

contrast, team workers almost entirely depend on system-assigned orders. The typical assignment range for team workers is within 6km of their designated station.

The orders that appear for gig workers in the order pool, and the order that a team driver is assigned, depend on a driver's location, demand conditions, and the driver's contemporaneous *matching score*. The exact matching score is proprietary, and depends on a driver's (1) service score, (2) "Rider Level," and (3) order acceptance rate. The service score depends on customer reviews and late deliveries. The Rider Level is calculated based on tenure and the cumulative number of completed orders. Refusing system-assigned orders directly reduces a worker's Acceptance Rate. Workers report that this factor has the greatest impact on assignment score, so they usually do not reject system-assigned tasks. The algorithm's formula does not include disability status as an explicit input.[8]

In sum, gig workers generally choose their orders and team workers are assigned orders. Both the order pool for gig and the assignments for team depend on workplace outcomes.

1.1 Conceptual Framework

We present a simple framework of labor supply on the Platform to ground our empirical analysis and to make predictions about differences between disabled and non-disabled workers, as well as how disabled worker productivity is affected by AI tools.

Workers have a productivity or efficiency level on the Platform, z, which may differ between disabled and non-disabled workers, and based on access to disability-focused AI tools. The share of total work time allocated to Platform work is denoted by h, while $1-h$ represents the share of time spent on other work pursuits. Output on the Platform is zh. Workers get paid a simple piece rate w, so their hourly wage is wz.

The return to work off the Platform is $r \cdot \frac{\beta}{\beta-1}(1-h)^{\frac{\beta-1}{\beta}}$, where r measures the value of one's outside option and $\beta > 1$. The parameter, β, captures how sharply the worker substitutes between work on and off the Platform. Worker utility is thus given by:

$$U = w \cdot zh + r \cdot \frac{\beta}{\beta-1}(1-h)^{\frac{\beta-1}{\beta}}$$

In our empirical analysis, we measure productivity broadly, incorporating speed at different points, customer reviews, and average total time per order. The firm makes no decisions and earns profits $A \cdot zh$, where A is a piece rate measuring the Platform's commission.

[8]More than assignments, the main thing affected by the matching score is the number of orders a driver is allowed to carry. The max number of concurrent orders starts at a baseline of 3, with an upper limit of 13 for gig workers and 20 for team workers. A higher assignment score results in a greater allowable maximum. Typically, experienced workers handle an average of six concurrent orders, while high-performing workers can manage even more, often reaching around ten concurrent orders.

Workers make a decision about the share of their work time to spend on the Platform. Using the first order condition of $\frac{\partial U}{\partial h} = 0$, we get $h = 1 - r^{\beta} (wz)^{-\beta}$. Taking partial derivatives yields $\frac{\partial h}{\partial r} = -\beta r^{\beta-1} (wz)^{-\beta} < 0$ and $\frac{\partial h}{\partial z} = \beta r^{\beta} w (wz)^{-(1+\beta)} > 0$, i.e., people work more on the Platform when they have worse outside options and higher Platform efficiency.

Based on the literature and our ethnographic observations, we posit two main differences between disabled (denoted by D) and non-disabled workers (denoted by N):

Assumption 1. Disabled workers have worse outside options than non-disabled: $r_D < r_N$.

This assumption (A1) is consistent with a large literature, cited above, on the challenges and discrimination faced by disabled workers.

Assumption 2. In the absence of AI tools, disabled workers have lower efficiency: $z_D < z_N$.

A2 is consistent with what we gleaned from conversations with DHH workers about work before the AI tool, namely, that there were often bottlenecks while making deliveries.[9]

Outcomes on the Platform can be grouped into those related to efficiency (z), labor supply (h), total output (zh), and income, wages, and profits. The base framework and the assumptions yield four implications (I1-I4) that we examine using the data.

Implication 1. Before the new AI tool, disabled workers have lower efficiency and wages than non-disabled workers.

Implication 2. Before the new AI tool, differences in labor supply between disabled and non-disabled workers are ambiguous. Disabled workers will work more on the Platform if the difference in outside options between disabled and non-disabled workers is large relative to differences in efficiency.

Implication 3. Before the new AI tool, average differences in total output, income, and profits between disabled and non-disabled workers are ambiguous.

Implication 4. Suppose an AI tool increases the efficiency of disabled workers. This will increase their labor supply, total output, wages, and profits on the Platform.

I1 follows directly from A2: lower efficiency implies lower wages since the hourly wage is wz. For I2, the difference in labor supply is ambiguous because A1 and A2 push in opposite directions: A1 makes disabled workers work more and A2 makes disabled workers work less. For I3, total output is zh, and since h is ambiguous, so is zh; income wzh is

[9] Disability scholarship (e.g., Krieger, 2010; Lane, 1992) emphasizes that differences between disabled and non-disabled people reflect both individual traits and environmental factors (e.g., discrimination, accommodations). Thus, A2 is a joint statement about capabilities and the environment, not inherent ability.

likewise ambiguous. For I4, by increasing z_D, an AI tool raises wages wz_D directly and also encourages disabled workers to increase labor supply, and thus two forces together work toward increasing total output and profits.

Framework extension: multiple severities of disability. While the framework has one severity of disability, it is easily extended to have multiple severities. Suppose, as in our setting, that there are workers who are profoundly disabled ("PD") and those who are non-profoundly disabled ("NPD"). Extending the logic of A1 and A2, it is natural to assume that $z_{PD} < z_{NPD}$ and $r_{PD} < r_{NPD}$. These assumptions yield pre-AI implications mirroring those between disabled and non-disabled. Turning to the impact of the AI tool, suppose communication situations vary in difficulty. Workers with non-profound hearing loss may be challenged primarily in difficult communication situations, whereas workers with profound hearing loss may face challenges in "simpler" interactions. If the AI tool helps resolve simpler communication situations but not the most difficult ones (Autor, 2015), then productivity gains from the AI tool will be larger for profoundly disabled workers.

Implication 5. Before the new AI tool, profoundly disabled workers have lower efficiency and wages than non-profoundly disabled workers, but differences in labor supply, output, income, and profits are ambiguous. The AI tool will have larger effects on profoundly disabled workers relative to non-profoundly disabled workers.

Framework extension: multiple local labor markets. Since the Platform operates across China, we observe AI impacts across a range of local labor markets, from the "rust belt" of Northeast China to the richer provinces in the Southeast. Disabled workers likely have stronger outside options when the local labor market is stronger, which can be proxied by a lower local unemployment rate, i.e., $r_D^S > r_D^W$, where r_D^S and r_D^W denote disabled workers' outside options in strong and weak local labor markets, respectively. Since $\frac{\partial^2 h}{\partial r\, \partial z} > 0$, an AI-induced increase in productivity leads to larger increases in hours for disabled workers in stronger local labor markets than for disabled workers in weaker local labor markets.[10]

Implication 6. The AI tool will have larger effects on hours, output, income, and profits for disabled workers in stronger local labor markets compared to disabled workers in weaker local labor markets.

[10]The positive cross-derivative follows from concavity of the outside option. Workers with stronger outside options spend more time outside the Platform, placing them where the marginal return to outside time is relatively insensitive to further increases in outside time. When AI raises platform productivity, these workers therefore find it relatively attractive to shift time back to the Platform, generating a larger labor supply response. This concavity is natural if outside opportunities exhibit decreasing returns: the most productive outside hours are used first, whereas jobs on the Platform get paid with a relatively constant piece rate.

Framework simplicity. This framework is deliberately simple and abstracts from several features of reality. First, true worker "productivity" is multidimensional—encompassing speed at different stages of order fulfillment, late deliveries, and bad ratings—whereas the model summarizes these dimensions with a single parameter z. Second, the model does not explicitly distinguish between hours worked and retention. In our setting, it is useful to think of "labor supply" broadly as including hours on the Platform, timing of work, and continued participation, e.g., workers with worse outside options may both work more hours and be less likely to quit the Platform. The empirical analysis examines these margins separately and documents pre-AI differences between DHH and non-disabled workers, as well as how the AI tool affects those differences.

1.2 Data

We obtained a proprietary administrative panel from the Platform, organized at the worker-week level and containing detailed measures of worker activity and performance. The panel covers workers who registered on the platform between 2021 and 2023 and completed at least one week of work, and includes all 4,284 team or gig workers who report DHH as their only disability, as well as nearly 40,000 non-disabled workers randomly sampled from the same population.[11] Drivers with multiple disabilities were excluded from the panel shared with us.

The panel allows us to track each worker from the start of their tenure on the Platform and to analyze how their performance evolves over time. By focusing on newly registered workers, we compare DHH and non-disabled workers from their point of entry, avoiding stock-sampling concerns related to differential tenure.

The AI tool was rolled out suddenly at t_0 in mid 2022 to all DHH workers. To avoid identifying the Platform, we do not specify the exact month of t_0, but it is a date during April 2022-September 2022, thus giving over one year of pre-AI data to compare the performance of DHH and non-disabled workers, and we can track workers for more than a year to study the AI tool's impact. All variables are available for the full 2021-2023 period unless otherwise stated. We examine a wide range of worker outcomes, which we divide into four groups:

Efficiency: z in our framework. Efficiency or productivity refers to the speed (or quantity) and quality of work. Beyond overall speed, our data enable us to look at time spent during the three steps of the work process. At each step, the worker must perform an

[11] Beyond team and gig, a small share of workers on the Platform (we believe less than 5%) fall into a third work model called preferred gig. This group was excluded from the data shared with us, as essentially none of these workers are DHH and their work is different from team and regular gig.

action on the app, which provides a clear time stamp for when each step starts and finishes.[12]

The three steps are as follows: First, once an order is initiated by the customer, it is assigned to a worker, via direct assignment by the Platform or via acceptance by the worker. The time for this first step is called the *order acceptance time*. Second, once the worker receives the order, they pick it up from the restaurant. The waiting time for food preparation is included in this step, which concludes when the worker actually receives the food from the restaurant. The time for this second step is called the *pickup time*. Lastly, the worker delivers the order from the store to the customer, the time for this being the *delivery time*. Beyond overall speed, we also observe Platform-defined late delivery rate, which refers to deliveries that are late beyond the delivery window, indicating a significantly late order.

Quality is measured using customer ratings. Rather than use 1–5 stars, customers on our Platform click a button to indicate a Bad Rating, a Good Rating, or nothing (most common). Bad ratings are at the discretion of the customer and occur for different reasons, including late deliveries, not receiving the delivery, and food damage during delivery (e.g., spilled soup). Drivers incur a monetary penalty from the Platform for each bad rating received, giving bad ratings direct financial consequences.

Drivers can also receive Good Ratings, but we emphasize these much less, as they do not affect worker compensation or order assignment, and because the Platform emphasizes them much less. We therefore report results for Good Ratings only in the Appendix.

We cannot analyze accidents or injuries, as the Platform does not comprehensively record them in the administrative data we received.

Labor supply: h in our framework. Our two main ways of measuring labor supply are (1) a worker's weekly hours spent online (i.e., available for orders) and (2) whether the worker quits the Platform. Beyond total hours, we also analyze days worked per week (i.e., so we can see if hours effects are coming from more days or more hours per day).

On (1), hours data are measured based on the worker logging into the app or the app is working in the background. Hours data are available for most of the sample period, though missing in 28% of worker-weeks (the biggest determinant being that hours data are not available prior to $t_0 - 46$, i.e., not available prior to mid 2021); we discuss the missingness pattern further in Section 3.5.1, where we provide evidence that hours missingness does not drive our findings.

On (2), the data does not contain a formal flag for attrition. Instead, we proxy whether someone exits the Platform based on having 0 weekly orders for 8 weeks in a row.[13]

[12] The relevant buttons can only be activated when GPS confirms the worker is sufficiently close to the relevant location, thus preventing manipulation. We do not have stage times by order, but rather as average time per order at the worker-week level.

[13] Our conclusions are similar under other reasonable definitions, e.g., 0 orders for 7 or 9 weeks in a row.

Total output: zh in our model. Measured by a worker's orders completed per week.

Income, wages, and profits. The administrative data we received does not contain fields for driver income, driver wages, and Platform profits. Instead, we use parameter information provided by the Platform, combined with our administrative data, to estimate these variables for each driver-week.

As on US delivery platforms, drivers are generally paid a piece rate per order. While the true piece rate varies by order, we estimate income using a single piece rate for gig workers and accounting for the accelerated piece rate profile of team workers. Hourly wages are defined as total income divided by hours online. To isolate the role of bad ratings in wage differences, we additionally compute wages excluding any pay losses due to bad ratings.

Tipping is effectively absent in Chinese food delivery, so tips are neither included in our administrative data nor used in calculating driver income and wages.

Platform profits from a driver in a week, π_{it}, equal per-order profit times the number of orders completed, minus the cost of any onboarding, minus the cost of any bad ratings:

$$\pi_{it} = \pi_0(zh_{it}) \cdot zh_{it} - c \cdot 1[\text{t is driver i's first week}] - k \cdot \text{Bad Ratings}_{it} \tag{1}$$

Profit per order, $\pi_0(zh_{it})$, is allowed to vary with the number of orders completed per week by the driver, reflecting the fact that driver compensation per order depends on weekly order volume. Due to confidentiality restrictions, we do not report the exact assumed per-order profit schedule. Second, the recruiting and onboarding cost of a new worker, c, is assumed based on our discussions with the Platform to be 400 CNY (Chinese yuan; 7 CNY $\approx$ 1 USD during our data period).[14] The presence of this term accounts for any value DHH workers provide by having higher retention. Third, k is the Platform's cost of a bad customer rating. Based on discussions with the Platform, we assign k to be 100 CNY, with k intending to capture the full costs of negative reviews, including effects on customer retention and reputational spillovers, net of the penalty paid by drivers to the Platform.

Control variables. The data also include rich control variables, detailed below in Section 2.1, about workers and the type of work performed.

Qualitative data. We conduct qualitative data collection to better understand DHH worker experiences, building on recent studies of workers in organizations (Blader et al., 2020). This includes: (1) interviews with 7 DHH workers about AI tool usage, outside job opportunities, daily routines, and challenges; and (2) on-the-ground observations in which one coauthor rode along with 3 DHH workers over several days and another delivered orders

[14] The biggest determinant of this is a bonus the Platform pays to workers after their first week on the Platform.

wearing earplugs to simulate the DHH experience. These data help us interpret quantitative patterns and identify barriers faced by DHH workers. Appendix F provides further detail; we reference qualitative findings when discussing quantitative results.

1.3 Sample and Summary Statistics

Sample restriction. Our analysis sample is based on excluding weeks where drivers have 0 orders, as these are weeks where drivers are not working.

Summary statistics. Table 1 presents summary statistics for our 43,605 food-delivery workers. Compared to non-disabled workers, DHH workers are slightly older (mean age 33 vs. 30 for non-disabled) and more likely to be male (93% vs. 91%), but are similar in work models: among both DHH and non-disabled workers, about 35% are team workers.

Panel C shows the distribution of disability severity among DHH workers, with 54% having profound hearing loss (Level 1 in the Chinese system).[15] Panel D of Table 1 compares DHH and non-disabled workers in terms of geography and distance variables. DHH workers are slightly more likely to be located in big cities than non-disabled workers.[16] DHH workers' orders have slightly higher travel distance than non-disabled workers.

Table 2 provides summary statistics on outcome variables. Workers do an average of 152 deliveries per week. The standard deviation is large at 142, indicating considerable variation in total deliveries. Peak-hour orders, which occur during high-demand times, account for 91 deliveries per worker-week, indicating that a substantial share of total orders happens during peak times. Team-based workers complete significantly more deliveries on average (212) compared to gig workers (73), reflecting the full-time nature of team-based work.

Average order duration in our data is about half an hour, with mean times of 3.4 minutes for order acceptance, 12 minutes for pickup, and 14 minutes for delivery. Per 10,000 orders, the late delivery rate is 416, meaning 4% of orders are substantially delayed.

Bad ratings, our measure of customer satisfaction, are much rarer. The bad rating rate per 10,000 orders is 9.6, meaning only about 1 in 1,000 orders yields a bad rating. Across driver-weeks, the median bad rating share is 0: most driver-weeks do not have bad ratings.[17]

Workers spend an average of 40 hours online per week. The high standard deviation of 29 hours reflects wide dispersion across drivers. The median number of workdays per week

[15] The share of drivers with profound hearing loss is higher in our sample than the overall Chinese population. A natural explanation is lower barriers to entry in gig work (e.g., no interview requirement) for people who are profoundly DHH, rather than differential disclosure to the Platform. Appendix D discusses further.

[16] One explanation is greater opportunities for DHH workers in large cities. Another is the historical concentration of specialized deaf schools there. We do not observe schooling histories.

[17] The low incidence of bad ratings may reflect a cultural norm in China where customers generally do not give bad ratings unless something has gone wrong, rather than as an expression of mild dissatisfaction.

is 6, though the average is lower at 5.3, suggesting that while many workers work full weeks, others participate part-time or intermittently. Average hours are 58 for team workers and 23 for gig workers, as seen in Appendix Table A.3.[18]

The weekly attrition rate on the Platform is 8% for the new drivers we study, with a high share quitting in the first few months. While such an attrition rate is high, it is comparable to that in other low-skill jobs, e.g., the call-center workers studied in Burks et al. (2015). The exit hazard decreases quickly, so there is a sizable share of workers who stay much longer.

Appendix Table A.3 presents summary statistics separately for gig and team workers. There is less variability in outcomes within each work model separately.

2 Pre-AI Differences by DHH Status

Our analysis proceeds in two stages. First, in Section 2, we measure differences in outcomes between DHH and non-disabled workers prior to the AI tool. Second, in Section 3, we evaluate the impact of the AI tool on outcomes of DHH workers (versus the comparison group of non-disabled workers) using a DiD approach.

2.1 Empirical Strategy

To compare outcomes between DHH and non-disabled workers in the pre-AI period, we regress each outcome Y_{it} on a DHH indicator and controls:

$$Y_{it} = \alpha + \beta \cdot DHH_i + \mathbf{X}_{it}'\gamma + \varepsilon_{it}, \tag{2}$$

where i is a worker and t is a week; DHH_i is a dummy for being DHH; $\mathbf{X}_{it}$ are controls; and ε_{it} is an error. With the exception of heterogeneity analysis of AI effects by regional unemployment (where we cluster by region), throughout the paper, standard errors are clustered by worker, as DHH status varies by worker.

Our controls $\mathbf{X}_{it}$ include year-week fixed effects (capturing common shocks during the outcome period), city-of-work fixed effects (roughly 300 cities), age fixed effects, gender, team status (gig vs. team), and registration year-month fixed effects (capturing cohort differences at entry). In our main specifications, we do not control for a driver's average distance between restaurant and customer—a measure of task difficulty—or for worker tenure, as

[18] While hours on the Platform are higher than for US Uber drivers (Cook et al., 2021), they are very typical for hours in middle-income and developing countries, where gig work often serves as a primary, full-time source of income (Chen et al., 2024b; Adhvaryu et al., 2026).

these variables represent mechanisms by which DHH and non-disabled workers may differ.[19] In the Appendix, we show that controlling for travel distance has little effect on results. However, not controlling for tenure is substantively important: as we discuss below, worker efficiency and output improve with tenure, and DHH workers are less likely to quit, so higher tenure is a way by which DHH workers benefit themselves and the Platform.

The purpose of equation (2) is to provide descriptive evidence on differences between DHH and non-disabled workers. It is not to estimate the "causal effect" of being DHH.

In addition to overall differences, we also present outcome-tenure curves separately for DHH and non-disabled workers. This allows us to analyze how outcomes vary with tenure, as well as to understand how DHH differences vary by tenure, e.g., whether differences are persistent or whether there is convergence. To do this, we regress outcomes in levels (non-transformed) on interactions between DHH status and worker tenure month indicators, controlling for the same controls listed in equation (2). The excluded group is non-disabled workers in month 1, and we plot the relevant regression coefficients added to the excluded group mean, giving estimated average outcomes for each group at each tenure month. Outcome-tenure curves are similar with no controls.

Functional form of dependent variable. For each variable besides late deliveries, bad ratings, and attrition, we use an inverse-hyperbolic sine transform on y, so that coefficients approximate percentage differences. Since we restrict our sample to weeks with positive orders, most of our outcome variables do not have zeros, and we avoid the "log of zero"-like problem with inverse hyperbolic sine (Chen and Roth, 2023). In line with Chen and Roth (2023), we estimate non-transformed OLS models for late deliveries, bad ratings, and attrition, and divide by a pre-AI DHH mean to calculate percentage differences. Late deliveries is analyzed per 10,000 orders, bad ratings as a weekly count, and attrition is binary (so the non-transformed OLS model is a linear probability model).[20]

2.2 Results

Figure 1 visually summarizes gaps between DHH and non-disabled workers across work outcomes prior to the introduction of AI tool. Each bar reports the percentage difference

[19]These variables can be argued to be "bad controls" if they are mechanisms by which the workers differ (Angrist and Pischke, 2009). Still, it can be interesting to address the question of whether DHH and non-disabled workers differ in efficiency at a given level of tenure or for particular routes.

[20]We analyze bad ratings as counts per week rather than per 10,000 orders: bad ratings are rare (most driver-weeks have zero), and the per-order rate can produce outliers when a worker completes few orders in a week. Figure 7 shows bad ratings per 10,000 orders evolving with tenure, before and after the AI tool. For late deliveries, which are more common, per-10,000-orders is the natural metric and a standard industry measure of delivery performance. In both cases results are highly robust: bad ratings results are similar using the per-10,000-orders rate (Table A.7), and late delivery results are similar using counts (Table A.8).

associated with the DHH indicator estimating equation (2) with full controls. Our findings are generally similar without control variables, as can be seen in Figure A.1, which plots raw means over time for several outcomes for DHH and non-disabled workers.

Productivity: speed variables. Figure 1 shows that DHH workers take longer per delivery than non-disabled workers, though differences vary considerably by stage of order fulfillment. For the initial stage of order acceptance (i.e., from order initiation by the consumer to order assignment or acceptance by the worker), DHH workers take 3.4% longer than non-disabled workers. For the second stage of pickup (i.e., from order acceptance to picking up the order at the restaurant), there is no difference between DHH and non-disabled workers. DHH workers are actually 0.6% faster, but it is highly statistically insignificant. The largest gap occurs in the final stage, from leaving the restaurant to delivering the order to the customer: DHH workers take 4.5% longer, which is about 40 seconds more. Overall, DHH workers do 1.9% fewer orders per service hour than non-disabled workers.[21]

Figure 2 presents the tenure-productivity relationship for each of the three order stages. Time spent in each stage declines with tenure for both groups, consistent with learning by doing (Shaw and Lazear, 2008).[22] For the first stage, order acceptance, differences are present in the first three months of tenure, but the two lines converge starting in the fourth month. In contrast, for the final stage, customer delivery, the gap is more persistent. That is, not only is the DHH difference largest for delivery time, but it also takes much longer to converge, and the average remains higher for DHH workers throughout.

The time numbers on each stage of the order fulfillment process are averages across worker-orders. The Platform may be more concerned with more extreme events, like whether an order is substantially late. Figure 1 shows that DHH workers have 9% more late deliveries, defined as deliveries beyond the expected window, than non-disabled workers. This works out to DHH workers having an extra 33 late deliveries per 10,000 orders.

A natural explanation for DHH workers being slower is communication challenges with customers. Other possibilities could include communication challenges with restaurants or differences in general driver behavior (e.g., DHH workers driving slowly or more cautiously). However, communication challenges with customers directly explain why DHH drivers are

[21] Orders per service hour is calculated as $\frac{60}{t_1+t_2+t_3}$, where the denominator is the sum of average minutes spent in each stage of an order. This measure doesn't use our hours data, which capture time available for work rather than time completing orders. Because drivers can handle multiple orders concurrently, the sum of average stage times can exceed actual clock time spent delivering orders, so orders per service hour should be interpreted as a processing-rate measure rather than literal orders per hour doing deliveries.

[22] This pattern is also consistent with differential attrition, which we discuss in Section 3.5. However, the main pattern of order acceptance, pickup, and delivery time decreasing with tenure is also present when we analyze outcomes in the first 12 months of workers who stay at least 12 months. This supports that apparent first-year learning by doing is not driven by differential attrition.

slowest on the final stage of the order.

This mechanism dovetails with what we observe in our qualitative fieldwork, especially riding along drivers. On most rides, DHH workers seem to go as fast as non-disabled workers. However, there are occasional situations that require extensive communication with customers, leading to substantial delays. In China, delivery workers are generally expected to call the customer upon arrival, e.g., to coordinate the handoff or confirm delivery if the food is left in a delivery cabinet or at the front desk. Prior to the AI tool studied, DHH workers who couldn't speak had to rely on in-app messaging, which customers sometimes don't see, contributing to occasional but more pronounced delivery delays.

Productivity: quality. Our measure of work quality is bad customer ratings. DHH workers receive more bad ratings than non-disabled workers: 0.15 per week versus 0.10, a gap of 47% relative to the non-disabled mean. Expressed as a fraction of the DHH pre-AI mean—consistent with how we scale our DiD estimates in Section 3—this corresponds to 31%, as shown in Figure 1. These differences are similarly strong for both gig and team workers separately (Figure A.3). Although bad ratings are rare in absolute terms, these are sizable percentage differences between DHH and non-disabled workers in the pre-AI period. They reinforce the above discussion on mechanisms pointing to challenges DHH drivers face in communication with customers. Table A.7 provides additional context and robustness checks on bad ratings results.

Qualitative work supports that communication challenges with customers are a natural explanation for why DHH workers get more bad ratings. Customers may get frustrated if communication is difficult, e.g. if the worker can't place a call to locate the customer, the customer may misinterpret the lack of spoken communication as rudeness. DHH workers recounted situations where customers left bad ratings citing "not calling" or "not speaking clearly." Overall, results on speed and bad ratings support A2 in Section 1.1 of $z_D < z_N$.

The presence of bad ratings is a concern for the Platform. If customers are dissatisfied, they may order less frequently, and the Platform could get a bad reputation. The Platform itself noted this concern, which was one motivation for investing in calling technology.

Labor supply. Figure 1 shows that DHH workers work 4.9% more hours per week than non-disabled workers, conditional on controls, equivalent to an extra two hours per week. Additional hours could come from DHH workers working more days per week or more hours per day, and our data indicate both. Specifically, DHH workers work an extra 4.3% days per week than non-disabled workers, equivalent to an extra 0.2 days per week.

Beyond working more hours and days, Figure 1 shows that DHH workers are 41% less likely to quit the Platform than non-disabled workers, conditional on observables, in a linear probability model. To dig further into this overall difference, panel (a) of Figure 8 presents

Kaplan-Meier survival curves for disabled and non-disabled workers separately. The raw gaps in survival emerge starkly within the first month and remain even after a year.

These are substantial differences in hours and driver retention. To benchmark magnitudes for retention, research across numerous firms shows that workers hired via employee referral are less likely to quit than workers hired through other means (Hoffman and Stanton, 2025), with differences around 10-15% for low-skill workers (Burks et al., 2015). Thus, the retention difference by DHH status is roughly 3-4 times that associated with being referred. Appendix Figure A.3 shows that the tendencies of DHH workers to work more hours, work more days, and quit less are all present when one looks separately at gig and team workers.

Our conceptual framework implies that DHH workers should work more if differences in efficiency are small relative to differences in outside options. We believe this is exactly what is occurring in our setting, where Figure 1 indicates that differences in efficiency are small, at least for efficiency as defined in terms of overall speed. In our qualitative work, it was not uncommon for DHH workers to work 6-7 days a week on the Platform, and these workers also indicated having limited outside job opportunities. Our findings of DHH workers quitting less broadly aligns with Minor et al. (2018), who find that workers with criminal records are less likely to quit, also seemingly because of worse outside options.

Total output. Combining the two offsetting forces—DHH workers being slightly slower while working more hours—the net effect is that DHH workers deliver more orders per week than non-disabled workers. DHH workers do 5.3% more orders than non-disabled workers, equivalent to an extra 8 orders per week.

Panel (a) of Figure 6 shows orders as a function of worker tenure in the pre-AI period. The tenure-productivity curves in Figure 6 are obtained from the same specification used to construct Figure 2. As with the speed variables, performance increases with tenure. However, for total output, DHH and non-disabled workers have very similar numbers of orders at a given tenure level. The advantage of DHH workers therefore arises not from higher output conditional on tenure, but from their having longer tenure and lower quit rates. In other words, DHH workers tend to operate at higher tenure levels of the learning curve than non-disabled workers.

The output gap is largest in the team model. Among team workers, DHH workers complete 8.7% more orders per week than non-disabled workers, about 18 additional orders. Among gig drivers, the difference is 0.7% and statistically insignificant. A likely explanation is greater support in the team model, where DHH workers can seek help from station managers and teammates.

Income, wages, and profits. Figure 1 shows that, in line with doing more orders per week, DHH workers earn 5.4% more in weekly gross income relative to non-disabled workers.[23] This increase is driven entirely by greater labor supply rather than higher pay per unit of time: hourly wage rates for DHH workers are lower than those of non-disabled workers, as also seen in Figure 1.

Not accounting for bad rating penalties, disabled workers earn 9.4% lower hourly wages. When penalties from bad ratings are incorporated, the hourly wage gap widens slightly to 10.2%. These estimates are broadly in line with those in Forth and Jones (2025), who estimate a disability hourly pay gap of 6.8% within firm for a representative sample of firms in England and Wales. Overall, that DHH workers have more bad ratings only accounts for about one-tenth of the 10% wage gap. Most of the pay gap reflects DHH workers having similar income to non-disabled workers (in the sample with observed hours used to compute wages) despite working more hours. To further benchmark these patterns, Appendix Table A.1 reports DHH employment and wage gaps in the U.S. general labor market using CPS ASEC microdata from 2008–2023. The specification most analogous to our within-platform comparisons controls for occupation and industry fixed effects in addition to demographics, year, and state (column (4)), since our Platform estimates compare workers in the same job. This yields an hourly wage gap of about 4%, somewhat smaller than, but broadly consistent with, the wage gap we estimate on the Platform. DHH workers also face an employment gap of 25 percentage points relative to non-disabled workers (column (3) of Panel A).

Turning last to Platform profits, DHH workers generate 12% more profit per week than non-disabled workers. While DHH workers are slower and earn lower hourly wages, Platform profits depend on total orders rather than speed per se, as shown in equation (1). Thus, slower work reduces DHH workers' own hourly wages rather than Platform profits. The Platform also benefits from DHH workers' longer tenure: a smaller share of their worker-weeks occur in the first week, when onboarding costs are incurred.

Sources of profit differences. Panel A of Table 3 decomposes average pre-AI profit differences between DHH and non-disabled workers into its three components: orders, bad ratings, and onboarding costs. For each component, we multiply its estimated percentage difference between DHH and non-disabled workers by the component's average value in CNY, and report its share of the total CNY difference. The main source of DHH profitability is they do more orders per week, accounting for 85% of the profit gap. Getting more bad ratings works against DHH profitability, but bad ratings are rare. Higher DHH retention means the Platform incurs onboarding costs in a smaller share of weeks than for non-disabled workers, but this contribution to pre-AI profits is limited, as onboarding costs are modest.

Heterogeneity in pre-AI differences by DHH severity. Figure A.4 presents het-

[23] Weekly gross income is based on earnings received, and does not include penalties for bad ratings or costs from e-bike rental, maintenance, or charging. We account for bad rating penalties in the hourly wage.

erogeneity in pre-AI outcome differences according to disability severity. We split by DHH workers with profound HL (Level 1 in Chinese system; 54% of the sample) and DHH workers with non-profound HL (Levels 2-4). Profound HL workers tend to obtain worse outcomes than non-profound HL workers, though not all differences are statistically significant.

Starting with efficiency, profound HL workers get 48% more bad ratings than non-disabled workers, whereas non-profound HL DHH workers get 14% more bad ratings than non-disabled workers. There is no difference between non-profound HL workers and non-disabled workers in late deliveries, whereas profound HL workers have 16% more late deliveries. Turning to labor supply, the overall DHH hours advantage is driven by non-profound HL workers, who work 17% more hours per week than non-disabled workers. Profound HL workers work 12% fewer hours per week than non-disabled workers. Differences in total output, income, and Platform profits are likewise driven by non-profound HL workers.

Both groups earn lower hourly wages than non-disabled workers, but, interestingly, the wage gap is larger for non-profound HL workers (a 13% gap vs. a 7% gap). This reflects their substantially higher hours. Profound HL workers may concentrate work in more lucrative days or times, consistent with greater communication challenges.

3 Evaluating the AI Intervention

3.1 Origins of the AI tool

The "Intelligent Outbound Call" (IOC) is a text-to-voice calling feature on the app to help DHH workers communicate better with customers. Many DHH workers face challenges in speaking and being understood. Before the IOC feature was introduced, such DHH workers communicated with customers via text-based messages within the Platform's app. This posed challenges, as typed messages such as "Can you open the door for me?" or "I cannot find the place, please allow me an extra five minutes" often went unnoticed or unanswered by customers, resulting in delays, misunderstandings, and added stress for DHH workers.

Development of the tool. The tool was developed in-house at the Platform by around 3 software engineers during off time over several months. Both of the two major Chinese food delivery platforms emphasize the use of technology for good as part of corporate social responsibility, and engineers at the Platform are encouraged to spend time helping others. In developing the tool, the Platform emphasized helping DHH workers as part of its broader corporate social responsibility efforts. At the same time, the tool had the potential to increase profits (e.g., by increasing productivity, reducing costly delivery errors).

How the tool works. Under the IOC feature, DHH workers make natural-sounding, automated calls to customers. The IOC provides workers with a set of default communication prompts, as well as the option to create and customize prompts tailored to specific situations. Figure E.1 illustrates both the default prompts offered by the Platform and examples of customized prompts used by a DHH worker.[24] After the DHH worker selects a prompt, the app initiates an automated voice call to the customer. Each call begins with an introductory message stating that the caller is DHH followed by a specific message conveying the worker's issue or request, e.g., "Your food will be placed in a locker". The tool is strictly text-to-voice instead of voice-to-text, i.e., customers can respond in the app instead of on the call. The feature is available only to DHH workers; non-disabled workers cannot use it.

How the tool uses AI. The IOC uses AI-based text-to-speech technology, which converts written text into natural-sounding speech, and is widely used in voice assistants, audiobooks, and accessibility services. Pre-AI generated speech can perform basic functions, but often lacks naturalness and emotion. AI has brought significant breakthroughs to this area. Deep learning-based models like WaveNet and Tacotron use neural networks to synthesize speech waveforms, significantly improving fluency and making prosody, intonation, and pauses more human-like. Research in human-computer interaction finds that AI voices are more pleasing to customers than traditional robocalls (Shah et al., 2023; Arcot, 2025).

The advantages of the IOC tool include: first, enhanced customer responsiveness, as customers are more responsive to phone calls compared to app notifications, and voice communication is perceived by the Platform as "warmer" than text, ensuring prompt and higher-quality interactions; second, improved efficiency, since DHH workers can quickly address common issues through pre-set prompts without repeatedly typing messages, saving valuable time especially during peak delivery hours or adverse weather; and third, personalization and flexibility, as workers can customize messages to communicate unique situations.

AI implementation and usage. The Platform announced the AI-based tool to DHH drivers through multiple in-app notifications, making awareness of the tool widespread. Our administrative data do not record individual-level usage. However, the Platform informs us that, to their knowledge, all DHH workers use the AI tool and that the tool is popular. This is validated by our qualitative evidence: all 9 DHH workers whom we interviewed or worked alongside reported using the AI tool (details in Appendix F).

[24] From our observations, prompts often fall into two main categories: (1) difficulties locating the customer, such as being denied entry by security or requiring a buzzer code; and (2) notifications that the order has been placed at a specific location, like a food cabinet. One worker we worked with also made polite requests to avoid bad ratings.

3.2 Overall Effects of the AI Tool

Estimation strategy and graphical presentation of results. To assess the impact of the AI tool, we estimate a DiD model that compares outcomes for DHH workers before and after the introduction of the AI tool, relative to changes over the same period for non-disabled workers, who do not receive access to the tool:

$$Y_{it} = \alpha + \gamma \cdot DHH_i \times Post_t + \lambda_i + \delta_t + \varepsilon_{it}. \tag{3}$$

where $Post_t$ is a dummy equal to 1 if the current week is on or after the introduction of the AI tool at time t_0 in mid 2022. γ captures the differential change in outcomes for DHH workers after the introduction of the AI tool relative to non-disabled workers. Under the parallel trends assumption, γ identifies the average treatment effect on DHH workers. Worker effects λ_i capture fixed differences across workers while time effects δ_t capture common shocks. We present our baseline DiD results without time-varying controls, and show robustness to time-varying controls in Section 3.5. Beyond overall DiD results, we (1) show event study results analyzing how outcomes change month-by-month before and after the AI tool and (2) compare outcome-tenure curves before and after the AI tool.

For analysis of employee attrition, we do not include worker fixed effects—as workers can only quit once—and instead use the same controls as for equation (2).

The identifying assumption underlying equation (3) is that, absent the introduction of the AI tool, outcomes for DHH and non-disabled workers would have evolved similarly over time, conditional on controls. This could be violated if there are other policies or procedures introduced for DHH workers around the event time window (for the event study analysis) or in the sample period in general (for overall effects). We examine concerns about differential trends using event studies.

Figure 3 summarizes DiD estimates of the AI tool's impact across all outcomes. Each bar shows an estimate of γ from equation (3) and represents the percentage change in the outcome for DHH workers. We analyze the same outcome as in Figure 1.

Event study setup. Our event study uses a window of 6 months before and 10 months after the introduction of the AI tool. We use all available data and bin observations outside this window into the endpoints—i.e., all months more than 6 months before or more than 10 months after t_0 are pooled into boundary coefficients (Schmidheiny and Siegloch, 2023). Because our analysis is based on new entrants to the Platform, the sample available at each pre-period month shrinks as we go further from t_0, mechanically widening confidence intervals in the early pre-period. We thus focus on this window to ensure a reasonably high degree of statistical power for all event-time months shown. Section 3.5 discusses longer-run effects beyond the 10-month post-period and shows that results are highly robust.

Productivity variables: speed. Figure 3 shows that the AI tool led to an improvement in DHH worker speed, with orders per service hour increasing by 1.6%. Improvements are concentrated in the final stage of order fulfillment, delivery from restaurant to customer, where there was the largest initial disadvantage for DHH workers. Delivery time decreases by 3.5% for DHH workers, closing roughly 3/4 of the pre-AI gap of 4.5%. There is also a small 0.8% reduction in order acceptance time and no meaningful effect on pickup time.[25]

In terms of magnitudes, the reduction in delivery time of 3.5% is equivalent to moving about 6 months down the DHH delivery time tenure curve (panel (c) of Figure 2).

Panel (b) of Figure 4 shows event study estimates of the impact of the AI tool on delivery time (i.e., the final stage). There are not systematic pre-trends, consistent with the identifying assumption in DiD. In the first couple months of the AI tool, delivery time may actually increase (though not significantly so), consistent with drivers taking a little time to use the AI tool effectively (Brynjolfsson et al., 2021). Effects on delivery time gradually turn negative, with effects seeming to gradually increase in magnitude.

The impact of the AI tool on substantially late deliveries is even larger, with a 11% reduction for DHH workers. The IOC entirely closes the late delivery gap between DHH and non-disabled workers. This is seen in raw means in panels (e) and (f) of Figure A.1. As whether an order is classified as substantially late can hinge on a minute or two, modest improvements in delivery speed—and reductions in communication-related stoppages—can generate larger effects on late delivery rates.

As seen in Appendix Figure A.5, reductions in delivery time are pronounced and economically similar for both gig workers and team workers. Reductions in late deliveries are pronounced for team workers but not statistically significant for gig workers, possibly because gig workers take many more orders after the IOC and are therefore more likely to experience late deliveries, for example when carrying more orders simultaneously during peak hours, which increases the likelihood of delivering late.

Productivity variables: quality. Figure 3 shows the AI tool has a large effect on bad ratings, reducing them by 22%. This closes roughly 2/3 of the pre-AI gap of 31%.[26]

As seen in Figure 7—which plots bad ratings over tenure—prior to the AI tool, there are pronounced gaps between DHH and non-disabled workers that persist for most of a worker's

[25]The pre-AI DHH order acceptance time gap is 3.4%, so AI closes about 20–25% of it, driven by gig workers. Since gig workers (unlike team workers) have discretion over which orders to accept, AI may make them more willing to accept offers confidently.

[26]The sizable negative impact of the AI tool on bad ratings is robust across many methods, which is important given that bad ratings are rare. For example, AI reduces the chance of receiving any bad ratings in a week by around 20%.

first year of tenure on the Platform. After the AI tool, there is still an initial gap in the first months of tenure, but the gap is much smaller, and the gap dissipates by around the 4th month of tenure. In terms of magnitudes, the impact of the AI tool is equivalent to moving about 3 months down the tenure curve in bad ratings.

Panel (c) of Figure 4 shows effects on bad ratings over time in an event study. As for delivery time (in panel (b)), there is no pre-trend. Effects appear to grow somewhat over time, potentially reflecting adjustment costs in learning about the AI tool, though confidence intervals are larger.

Overall, the productivity results suggest that the AI tool alleviates frictions at the delivery stage: pickup time is unchanged, while delivery is faster and bad ratings decrease. Qualitative interviews (Appendix F.1) support this interpretation, with DHH workers reporting fewer misunderstandings at delivery and greater comfort in interacting with customers.

Labor supply. Pre-AI, DHH workers have greater hours and retention than non-disabled workers, and Figure 3 shows that the AI tool further accentuates these differences. The AI tool increases DHH weekly hours online by 10.6%, roughly tripling the pre-AI hours gap of 4.9%. The AI tool increases work days per week by 3.5%, which is significant, but most of the increase in hours comes from an increase in hours per day. The AI tool further reduces DHH attrition from the Platform by 21%, further expanding the pre-AI gap of 41% by half. As seen in the comparison of worker survival curves in Figure 8, the AI tool noticeably increases the difference between the DHH and non-disabled survival curves. Effects on labor supply and retention are present in both gig and team workers, as seen in Figure A.5. Overall survival of both groups is somewhat lower in the post-period, reflecting a stronger labor market in China in 2023 than 2021, but retention improves differentially for DHH workers. Appendix B.2 shows that the AI tool also increases entry of DHH workers onto the Platform.

These effects are in line with the conceptual framework, which predicts that drivers increase labor supply in response to an improvement in efficiency. While our intentionally simple framework does not account for non-pecuniary work amenities, it is also possible that the AI tool increases labor supply "directly" by making Platform work more pleasant instead of simply by helping DHH workers make more money.

Total output. Figure 3 shows that the AI tool increases DHH weekly orders by 11%. Pre-AI, DHH workers do 5% more orders than non-disabled workers, so AI extends this initial advantage. Effects are present for team workers (+8%), but are especially large for gig workers (+15%), though the difference by work model is not significant ($p = 0.29$). These results suggest that AI may be especially helpful in the more independent gig mode. One possibility is that team DHH workers already had some support or workarounds in place in the pre-AI period, whereas gig DHH workers, with less support, benefit substantially. As seen in panel (a) of Figure 4, effects are present directly after the AI tool is introduced, and seem to increase over time, consistent with effects on delivery speed and hours building over time, with panel (a) of Figure 5 showing the event study for hours.

Income, wages, and profits. Following the introduction of IOC, weekly gross income earned by DHH workers increases by 11% relative to non-disabled workers. Hourly wages also rise for DHH workers, though by a more modest magnitude of 3.6%. These estimates imply that roughly one-third of the increase in weekly income is driven by higher hourly pay, with the remainder attributable to increased hours worked. Given a pre-AI disability pay gap of 10.2%, these gains imply that AI closes around one-third of the wage gap between DHH and non-disabled workers. Wage gains tend to be concentrated among team workers; gig workers see smaller hourly wage gains as their large hours increases offset their income gains (Figure A.6). We are not aware of other workplace interventions or technologies that generate effects of comparable magnitude on disability wage differentials.[27]

From the Platform's perspective, the profitability of DHH workers rises substantially following the introduction of IOC. Net weekly profit generated per DHH worker increases by 17%, a sizable and statistically significant effect.

Sources of profit gains. Panel B of Table 3 applies the same approach—multiplying each component's estimated AI-driven percentage change by its average CNY value—to decompose the source of the gains in profitability for DHH workers. The main source of the profit gains is making DHH workers do more orders as opposed to avoiding onboarding costs or getting fewer bad ratings. Order volume alone accounts for 93% of the profit impact, with reduced attrition and fewer bad ratings contributing only 5% and 2%, respectively.[28]

3.3 Benefits and Costs to the Platform

While Figure 3 shows that the profitability of DHH workers increases, this does not account for the costs of the AI tool itself. The Platform's average profit per order is confidential and varies with work type and order volume, which we incorporate directly in our worker-week

[27] Evaluating the expansion of work-from-home arrangements during the COVID-19 pandemic, Bloom et al. (2025) show that disabled workers experienced larger wage declines in occupations with greater work-from-home exposure, consistent with increased labor supply exerting downward pressure on wages. Gentry (2025) finds that gaining access to Uber closes less than 2% of the disability pay gap.

[28] The onboarding component (5%) captures only the direct savings from paying the onboarding bonus less often. It does not capture the contribution of retention to order volume: the AI-induced reduction in attrition also moves workers to higher-tenure points on the learning curve where they complete more orders per week. Table 3 thus understates the full profit contribution of retention. To analyze how much of the order effect is due to improved retention, one strategy is to compare the order effect in Figure 3 vs. the effect in a model with inverse probability weighting (IPW). Using IPW shrinks the effect on orders from 10.5% to 9.7% (Figure A.15), suggesting that increased retention drives about 1/10 of the effect on orders.

profit calculations based on equation (1). For the purpose of translating our estimates into aggregate dollar magnitudes, we use a benchmark value of 7 CNY ($\approx$ \$1 USD) per order, based on conversations with the Platform. Under this assumption, the 17% improvement in DHH worker profitability implies roughly \$163k USD per month in additional platform profit.[29] The Platform's cost of the tool consists entirely of engineer time during development: there are no per-use licensing or subscription costs, and the tool was built by three engineers working off-hours over several months. We conservatively assume that each engineer spent 10 hours per week for three months, implying a total time investment of 390 hours—less than 0.2 annual FTE—and thus no more than about \$20,000 USD in labor cost. Under a simple comparison of benefits and costs, the AI tool pays for itself in less than a month: specifically, in 20/163 months, or about 4 days.

In summing up the Platform benefits as above, we implicitly assume that additional orders completed by DHH workers do not displace orders from non-DHH workers, corresponding to an infinitely elastic labor demand curve at the Platform level. To relax this assumption, suppose that a fraction δ of additional orders completed by DHH workers are displaced from non-DHH workers. In this case, the net increase in platform profits is scaled by $(1 - \delta)$. Even under substantial displacement—e.g., $\delta = 0.5$, implying that half of additional DHH orders come at the expense of non-DHH workers—the tool would still generate roughly \$82k USD per month in additional profits and pay for itself in about 1 week. Even with $\delta = 0.75$, the tool would still pay for itself in about 2 weeks.

3.4 Heterogeneous Treatment Effects

A central question regarding the deployment of AI in the workplace is how treatment effects vary across different workforce segments. For example, some evidence indicates that generative AI may have larger effects on novice or lower-skilled workers (Brynjolfsson et al., 2025; Cui et al., 2025; Noy and Zhang, 2023). We focus our analysis on four key dimensions: pre-AI tenure, pre-AI productivity, disability severity, and local labor market conditions. For each dimension of heterogeneity, we present estimates separately for each half of the sample split at the median, and test for differences using interaction term models similar to those in equation (3). For late deliveries, bad ratings, and attrition, we divide by a pre-AI DHH mean within each subgroup to calculate percentage differences.

Pre-AI tenure and productivity. We find no consistent evidence of heterogeneous treatment effects based on pre-AI tenure (Figure A.8) or pre-AI productivity (Figure A.9),

[29] There are on average about 1,600 DHH workers on the Platform per week, including both new and experienced workers. Assuming 600 orders per month, we have 1,600*600*0.17*\$1 $\approx$ \$163k per month.

with pre-AI productivity measured in orders per week. Excluding pickup time, the outcome where there was no effect of the AI tool, there is significant heterogeneity by pre-AI tenure or pre-AI productivity for only 3 of 26 outcomes. That is, the AI tool's effectiveness does not systematically vary with a driver's tenure or average productivity.

Disability severity. We observe substantial heterogeneity in effects based on severity of disability. Figure 9 plots the coefficients for workers with profound hearing loss (meaning Level 1, hearing loss of more than 90 decibels) versus those with non-profound hearing loss (Levels 2–4), whom we refer to as profound HL and non-profound HL workers.[30] Profound HL workers, who are 54% of DHH workers in our sample, cannot hear even shouted speech.

We observe substantial heterogeneity in effects based on severity of disability, usually in the direction that profound HL workers benefit more. Orders per week increase by 6% for non-profound HL workers, but by 16% for profoundly impaired, with a p-value of 0.08 on the difference. Bad ratings decrease by 7% for non-profound HL workers, but decrease by 40% for profound HL workers ($p = 0.02$). As seen in Figure A.4, pre-AI, the bad rating gap is largest for profound HL workers (and profound HL workers do 8% fewer orders than non-profound HL workers), so the AI tool also helps close the pre-AI gap by disability severity within DHH workers. This suggests that the AI tool is most effective at reducing communication frictions for those who face the steepest barriers.[31]

Relative to non-profound HL workers, profound HL workers also see substantially greater gains in income ($p = 0.08$) and wages ($p = 0.05$). While statistical significance is sometimes borderline on differences, the differences are economically quite sizable, e.g., wages go up by around 1.5% for non-profound but by 7% for profound HL workers.

For attrition, profound and non-profound HL drivers show similar effects. While Implication 5 predicts profound HL workers to benefit more, one explanation for similar effects is that profound HL workers have fewer outside work options regardless of the tool, whereas AI effectively tips the scales for non-profound HL who might otherwise have exited.

We split by profound and non-profound HL both because this is roughly a 50/50 split (and good for statistical power) and because profound HL can be qualitatively different from non-profound HL. However, results are broadly similar when we estimate dosage-response heterogeneous treatment effects (Figure A.10). Heterogeneity by severity is also robust to

[30] Unlike our other heterogeneity dimensions, severity is defined only within DHH workers, so an indicator Profound_i for profound HL is absorbed by worker fixed effects λ_i and the standard four-term interaction specification reduces to two terms: $DHH_i \times Post_t \times \text{Profound}_i$ and $DHH_i \times Post_t$. The worker fixed effects also address the standard concern that Profound_i may be correlated with unobservables.

[31] Interestingly, non-profound HL workers exhibit larger improvements in delivery time and orders per service hour. One possible explanation is that profound HL workers respond more along the extensive margin—working more and completing more orders—so that their gains are reflected more in total output than in per-order speed.

controlling for interactions of $DHH_i \times Post_t$ with age and gender, thus indicating that the heterogeneity is not driven by demographics correlated with severity (Figure A.11).

Local labor market conditions. Figure A.12 examines heterogeneity in AI effects by local labor market conditions. Exploiting that our Platform operates nationwide, we split China's 31 provincial-level administrative regions into high and low unemployment regions, where the 2021 average unemployment rates were 3.6% and 2.5% respectively.

On the 13 variables where we detect significant positive AI effects (i.e., excluding pickup time), in 7 of 13 variables, we observe a significantly more beneficial effect in low unemployment rate regions. AI has no effect on late deliveries in high unemployment regions, while AI reduces DHH late deliveries by 14% in low unemployment regions. Likewise, there is no AI effect on DHH hours online, work days, or orders in high unemployment regions, but sizable improvements in all these dimensions in low unemployment regions. While DHH driver income increases by 8% in high unemployment regions, it increases by 17% in low unemployment regions. Heterogeneity by regional unemployment is robust to controlling for interactions of $DHH_i \times Post_t$ with age and gender (Figure A.13).

3.5 Threats to Identification and Robustness Checks

3.5.1 Threats to Identification

Existence of concurrent policy changes. Our event study graphs show no clear pre-trends prior to the AI tool's introduction. A threat to identification would be if the Platform made other policy changes around the same time as the AI tool that differentially affected DHH workers. According to the Platform, there were no changes made close to the introduction of the AI tool, and in fact, no meaningful other changes for DHH workers over the sample period of 2021-2023.

Several of our quantitative results are strongly consistent with this. Suppose there was another policy launched by the Platform (or another entity like the Chinese government) to improve the productivity of DHH workers. Such a policy would likely show effects across all stages of order fulfillment: order acceptance, pickup, and delivery. In contrast, the effects of the AI tool are concentrated in the delivery stage—the only stage for which the tool is designed—where communication with customers is required and where we observe the largest improvements in speed. We do observe a smaller improvement in order acceptance, which reduces the pre-AI gap by roughly 1/4; this is consistent with DHH workers being more willing or confident to accept orders when customer communication is easier. There is no effect on pickup, consistent with the tool not being used for communication with restaurants.

In addition, the observed heterogeneity by disability severity speaks to the plausibility of the identification assumption. If there was an alternative Platform intervention that benefited DHH workers, it would not necessarily bestow greater benefit on profoundly impaired DHH drivers relative to non-profoundly impaired DHH drivers.

Improvements to DHH workers or declines for non-disabled workers? Could our DiD estimates be driven by non-disabled outcomes getting worse instead of DHH outcomes getting better? Several pieces of evidence speak against this. First, non-disabled workers getting worse would not generate the observed heterogeneity by disability severity, with larger AI benefits on more severely disabled workers. Second, raw means of various outcomes (Figure A.1) suggest substantial improvements for DHH workers in levels.

One particular form of non-disabled workers getting worse would be if DHH workers take delivery opportunities away from non-disabled workers. This is unlikely to be quantitatively important for non-disabled worker outcomes because the share of DHH workers on the Platform is small (i.e., less than 1% of workers on the Platform).

Missing data on hours. As noted above, we have significant missing data on hours. Figure A.2 plots the missing rate by calendar time. Hours is missing for all workers before week $t_0 - 46$ where t_0 is the week of AI implementation. In $t_0 - 46$, hours become available for gig workers, and then become available for team workers in week $t_0 - 4$ of 2022. Missingness rates are similar between DHH and non-disabled workers and appear to reflect Platform data-keeping practices rather than endogenous selection.[32]

Because of the missing hours data, our estimates regarding hours-based outcomes, namely, hours online and wages, rely on a different sample than other outcomes.[33] To ensure that our constellation of results is not being driven by different samples, we redo our main results from Figure 3 while restricting the sample to observations where hours are non-missing. As seen in Figure A.14, our main conclusions are unchanged in this sample.[34] In the restricted sample, the AI tool drives significant improvements: orders per week increase by 13%, weekly income rises by 13%, and delivery time decreases by 2.6%.

Accounting for regional trends. DHH workers are more likely to live in Eastern China and big cities than non-disabled workers. If regions in Eastern China experience different shocks than other areas (e.g., different lockdown policies), this could potentially confound our analyses. To address this, we re-do our analyses while adding region-month

[32] Among team workers, hours missingness is extremely similar for DHH and non-disabled workers, whereas among gig workers, it is slightly higher for DHH workers than non-disabled workers (Figure A.2, panel (b)). However, as seen in Figure A.5, our main DiD results are robust when restricted to team workers, and we cannot reject that AI effects on hours and wages are the same for gig and team workers.

[33] Hours per service hour relies on the data on time spent on different parts of the production process.

[34] In the restricted sample, the AI tool has a negative effect on late deliveries and bad ratings, but the effect is no longer statistically significant, reflecting the smaller sample with hours data.

fixed effects (where we have 31 regions and 3 years of data). Our results are highly robust to this, indicating that such trends do not drive our results (Appendix Figure A.14).

Matching DiD. Beyond regions, DHH and non-disabled workers vary slightly in terms of observable characteristics. While our event studies strongly support the parallel trends assumption, for robustness (exploiting our large sample of non-disabled workers), we perform matching where we match each DHH worker to a single non-disabled worker, matching exactly on entry year, work model (gig or team), city, 5-year age bin, and gender. Using the resulting sample of 2,354 DHH and 2,354 non-disabled workers, we re-do our DiD in the matched sample. Our results are also highly robust to this (Appendix Figure A.14).

3.5.2 Alternative Mechanisms and Interpretations

Differential attrition. Since the AI tool increases DHH worker retention, a concern is that this could "bias" our estimates of the effect of the AI tool on DHH non-attrition outcomes like productivity variables, hours, and wages. Given that productivity increases with tenure (see Figures 2, 6, and 7), if AI boosts DHH worker retention, then DHH workers will be more likely to be in higher-productivity parts of the learning-by-doing curve.

If we take the perspective of the Platform, our baseline DiD estimates should not be "corrected" for possible differential attrition, as it is an important *mechanism* by which the AI tool affects outcomes. The Platform cares about the overall effect of the AI tool on outcomes, not the impact of the AI tool on particular workers.

At the same time, we are still interested in how AI affects outcomes separate from affecting retention, and such estimates might be more relevant for DHH workers themselves (i.e., how much will the AI tool affect my productivity). To shed light on this, and "shut down" differential attrition, we pursue two approaches. First, we perform our DiD analysis restricting the sample to a panel of "stayer" workers, who are observed working at least one week both before and after the introduction of the AI tool.[35] As shown in Figure A.15, the results from this panel are quantitatively and qualitatively consistent with our baseline findings. We continue to observe significant improvements across key metrics: total weekly orders rise by 10%, weekly income increases by 10%, and bad ratings decrease by 20%. These strong effects are similar to our baseline estimates, indicating that differential attrition is not a central driver of the DiD estimates. Rather, the AI tool increases worker productivity directly in addition to increasing retention.

[35] We cannot pursue a fully balanced panel approach over 2021-2023, as our sample consists exclusively of workers joining the Platform. A fully balanced panel would require analyzing only the DHH and non-disabled workers who join the Platform in January 2021 and then stay throughout the whole data period, and such a sample is too small for reasonable analysis.

Second, we use inverse probability weighting (Wooldridge, 2002), which proceeds in two stages. In the logit first stage, we predict a worker's probability of remaining on the Platform each week as a function of time-invariant worker characteristics. Second, we re-do our DiD estimates weighting by the inverse of the predicted probability of remaining each week. Our non-attrition variables all show sizable effects from the AI tool under inverse probability weighting (Figure A.15, bottom panel). This supports that the AI tool increases productivity directly.

The role of late deliveries and bad ratings in load assignment. As noted in Section 1, late deliveries and bad ratings can affect the set of orders workers are assigned or able to select. Beyond direct productivity improvements, an alternative mechanism is that the Platform algorithm assigns treated workers easier or higher-quality orders as their performance metrics improve. In that case, observed gains in delivery speed, orders, hours, or income would reflect task reallocation rather than improved customer communication. This is a question about mechanisms rather than identification.

We address this concern in two ways. First, we re-estimate our main DiD specification while controlling for task difficulty, proxied by the distance between the restaurant and customer. As shown in Figure A.14, the estimated effects are very similar to the baseline results, indicating that changes in trip distance do not explain the observed DiD gains. Second, we conduct a mediation analysis controlling for late deliveries and bad ratings (Appendix B.3). The estimated effects of the AI tool on orders are largely unchanged, suggesting that reductions in late deliveries and bad ratings account for little of the overall increase in orders.

Overall, the results suggest that reductions in late deliveries or bad ratings, or the resulting reallocation of tasks, are unlikely to drive our other key DiD findings.

Salience or sympathy. Could it be that AI effects reflect salience or sympathy of hearing impairment as opposed to improved communication? It is highly unlikely that salience and sympathy could explain the full constellation of results. Pre-AI, customers were sent a notification that their driver was DHH, though more customers would be aware of this after a call. While sympathy could make customers less likely to give bad ratings, it is less clear how this reduces delivery times. Moreover, there is no reason why salience or sympathy would generate larger effects for DHH workers with profound HL relative to non-profound HL, as customers were not notified of drivers' level of hearing loss. This interpretation is also consistent with our qualitative observations: during a coauthor's three days riding alongside DHH workers and over a hundred orders, no customer appeared to make any accommodation for the driver being DHH — for example, no one opened a door preemptively or waited downstairs after receiving an AI call.

Reference dependence / income targeting. Could the pre-AI difference simply reflect that DHH workers target income, doing more hours to make up for lower efficiency? This is inconsistent with our results by DHH severity, where profound HL workers have both lower efficiency and fewer hours than non-profound HL workers.

3.5.3 Long-run Effects and External Validity

Longer-run persistence of effects. The other major food delivery platform in China introduced a similar AI tool for DHH workers later in the same year, after our Platform. Because the market is largely dominated by these two platforms, the period in which only our Platform offers the tool raises the possibility of cross-platform spillovers (e.g., worker or customer reallocation). To address this, we re-estimate our DiD results excluding the period in which the AI tool is available on our Platform but not on the other platform. The estimates are very similar (Figure A.16). This suggests that our results are unlikely to be driven by the temporary presence of the tool on only one platform.[36] The subsequent adoption of a similar tool by the other platform is also consistent with the tool being broadly valuable for DHH workers.

Beyond considering adoption of the other Platform, we also assess whether effects persist further into the future. While our main event study graphs cover 10 months after the AI tool, we extend them to 16 months, pushing toward the edge of our data. As seen in Figure A.7, effects continue to grow and remain significant and robust.

Accounting for the transition period. Like many new technologies, usage of the AI tool was not instantaneous at 100% usage. Rather, according to the Platform, there was a transition period of a few months where DHH drivers learned about the new AI tool and how to best use it. This is supported by the event study graphs, which often show impacts that are more muted in the first few months, but that increase over time. Since we are primarily interested in the longer-run effects of the AI tool, we also re-do our DiD analyses but excluding the first 4 months immediately following the policy change. The results, presented in Figure A.16, confirm that our main findings are robust to this exclusion and often become larger. The reduction in bad ratings intensifies from 22% to 25%, the increase in output per hour grows from 1.6% to 2.1%, the increase in weekly orders grows from 11% to 13%, and the increase in income grows from 11% to 13%. Thus, our baseline DiD estimates, which include the first several months of the AI tool being available, may represent a more conservative lower bound of the tool's longer-run effects.

[36] Our DiD results are also qualitatively robust restricting to the months where only our Platform uses an AI tool for DHH workers.

Could COVID outbreaks or lockdown policies limit external validity or threaten identification? Chinese cities instituted lockdowns at different times at various periods in 2021 and 2022. As shown in Figure A.16, our DiD results are robust to excluding worker-week observations from regions-weeks affected by COVID outbreaks. Results are also similar if we control for whether a COVID outbreak is occurring instead of excluding those periods.

Would the AI tool have equally benefited non-disabled workers? The AI tool is only introduced to DHH workers, raising the question of whether non-disabled workers would benefit as much. The heterogeneity results by disability severity suggest this is unlikely. DHH workers with non-profound HL benefit substantially less than those with profound HL, suggesting that non-disabled workers might also benefit less, or possibly not at all. The Platform's decision not to roll out the AI tool to non-disabled workers is also consistent with the view that benefits for non-disabled workers would be smaller or negligible.

3.6 What the AI Tool Does Not Address

Although the AI tool substantially improves communication between DHH workers and customers, some workplace outcome gaps remain. Qualitative evidence helps clarify the types of situations that allow such gaps to persist. Our qualitative data were collected in summer 2024, around two years after the AI tool was introduced. Consistent with what the Platform told us, in the workers we interviewed or worked alongside (7 interviewed and 3 accompanied during deliveries, with one worker appearing in both), all 9 of them used the AI tool. Thus, the remaining gaps do not reflect limited adoption of the tool.

Instead, the evidence points to communication frictions that the AI tool does not fully resolve. These include difficulty locating customers' exact locations, locations with complex layouts (e.g., buildings with multiple entrances), and customer instructions that remain ambiguous without verbal clarification. A coauthor working alongside a DHH worker experienced the following, as recorded in his field notes, where "I" denotes the coauthor:

> The [DHH] worker followed the customer's instructions to contact the apartment complex management via the intercom. After the system connected, no one spoke on the other side. Although I heard the door unlocking sound, the door did not automatically pop open, and there was no visible indication that it had unlocked. The worker continued waiting by the intercom until I eventually walked over and pulled the door open.

A second situation arose while the coauthor was accompanying the same DHH worker, illustrating how written instructions may remain unclear in the absence of verbal communication:

> One order was delivered to an urban village [chengzhongcun in Mandarin], but the address was inaccurate. Upon arriving, the worker used the intelligent

> outbound call to contact the customer, who messaged back, "I'm upstairs." However, it was unclear what the customer meant by "upstairs." I then heard the customer shouting from inside a factory behind us, but by that time, the worker had already started riding in the opposite direction. I stopped the worker, informing him that the customer was inside the factory, and we managed to find the customer. In such situations, a worker with a hearing disability could waste significant time locating the exact address. Later that afternoon, we encountered a similar issue with an order delivered to a construction site.

The AI tool helps initiate contact when customers might otherwise not notice or respond to messages. However, ambiguity in customers' directions and physical delivery environments can persist when clarification would normally rely on real-time verbal communication.

4 Conclusion

By 2050, roughly 10% of the world's population will have disabling hearing loss (World Health Organization, 2025). While DHH workers face intense labor market challenges including lower earnings and employment, research has often struggled to identify policies or tools that can benefit DHH workers or disabled workers in general (Derbyshire et al., 2024).

We comprehensively examine the relative workplace outcomes of DHH food-delivery workers in China, and analyze how DHH workers are affected by a disability-focused AI tool. Before the tool is introduced, DHH workers are more likely to be late and get bad customer ratings, though they also work more hours, deliver more orders, and are substantially less likely to quit, patterns consistent with our conceptual framework where disabled workers have worse outside options. While DHH workers earn greater profit for the Platform on a per-worker basis, they earn about 10% less per hour.

The introduction of a simple AI-based communication tool substantially enhances DHH worker performance. Work speed improves, driven by faster completion of the final stage of order fulfillment—delivery to customers—where the AI tool is used. Orders rise and customer complaints fall. Retention, already higher for DHH workers, increases further. The AI tool decreases the disability wage gap by one-third and increases DHH worker profitability.

These large benefits occur even though the new tool was relatively inexpensive and straightforward. Far from being a limitation, the simplicity of the tool illustrates that even relatively modest AI-enabled interventions can meaningfully reduce barriers faced by disabled workers, without requiring large language model or autonomous conversational capabilities. Further work can examine the impact of more advanced AI tools.

Relevance for other disabilities. While our results pertain specifically to DHH workers, AI tools may also benefit people with other disabilities. Many human faculties involve prediction, the core function of modern AI (Agrawal et al., 2018). In the context of deafness, AI predicts how "natural" speech should sound. For visual impairment, AI tools such as Seeing AI and OrCam MyEye use images and video to predict what is nearby and narrate the environment (Granquist et al., 2021). AI may also enable smart wheelchairs for people with mobility impairments (Giansanti and Pirrera, 2025), smart glasses to help people with autism spectrum disorder interpret emotional cues (Iannone and Giansanti, 2024), and brain–computer interfaces (e.g., Neuralink) that restore greater control to people with paralysis. Future work could examine the impact of such technologies on workplace outcomes.

We speculate that one reason the AI tool was especially beneficial in our setting is that hearing loss creates a relatively localized bottleneck. Delivery work in China does not require a lot of speaking, but speaking can be especially useful in one part, i.e., for communicating with customers. By contrast, in some jobs, disabilities may affect many complementary aspects of work simultaneously (e.g., DHH physicians), such that an AI tool targeting one task might generate smaller gains. Still, we believe there are many other jobs where a person's disability creates a localized bottleneck where AI might be useful.[37]

Relevance for other jobs. Our results are also specific to food delivery work, a job with short, frequent interactions with customers, decentralized work, limited educational requirements, and heavy use of smartphones. Other jobs sharing these features include ride-hailing workers (e.g., Uber, Lyft, Didi); package delivery and last mile logistics (Amazon Flex, DoorDash Drive, UPS seasonal contractors); home services and on-demand repairs (TaskRabbit, Thumbtack, Handy); field sales or merchandising work (gig workers stocking shelves for Instacart, Shipt, or retail merchandising platforms); and customer-facing retail or hospitality roles where interactions are brief and standardized.

Public policy implications. An important policy question is whether governments should subsidize AI tools that help disadvantaged workers. In our particular setting, we estimate that it was *privately beneficial* for the Platform to introduce the AI tool, with benefits much larger than costs. On the Platform, the sizable number of DHH workers increases the return to implementing the AI tool. More broadly, firms may have incentives to implement similar technologies, without subsidies, when the disability is common and per-worker gains are high. However, for rarer disabilities, or where productivity improvements are smaller, public support may be warranted given the social welfare benefits of boosting disabled worker employment. Separately, while news stories often emphasize adverse aspects

[37]For example, a blind or low-vision person might be fine with most aspects of knowledge work, but face a bottleneck in accessing visual information: she or he could benefit from an AI tool like Seeing AI. Likewise, someone with dyslexia may perform well in most aspects of a marketing job, but face a localized bottleneck related to spelling or drafting written text, and AI writing tools could be useful for this.

of AI like job loss, AI tools like ours may help reduce inequality (Acemoglu et al., 2026).

The value of data. Separate from AI, our paper illustrates the value of granular administrative data in understanding the contributions of disabled workers. In analyzing the performance of DHH workers, one might naturally focus on their speed of work. While it is true that DHH workers are slower, they also work more hours and quit much less, consistent with having a worse outside option. We speculate that some organizations might "under-notice" this "loyalty benefit" of disabled workers. Big data can illuminate the ways in which disabled workers do or do not benefit organizations.

References

Acemoglu, Daron and Joshua D Angrist, "Consequences of Employment Protection? The Case of the Americans with Disabilities Act," *JPE*, 2001, *109* (5), 915–957.

_ **, David Autor, and Simon Johnson**, "Building Pro-Worker Artificial Intelligence," Working Paper 34854, National Bureau of Economic Research February 2026.

Adhvaryu, Achyuta, Martin Atela, Valentina Brailovskaya, Priyanka Dua, Jenny Susan John, Pratibha Joshi, and Rivandra Royono, "Platform Work: Evidence from Drivers in India, Indonesia, and Kenya," 2026. NBER WP 34680.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Boston, MA: Harvard Business Review Press, 2018.

Aizawa, Naoki, Corina Mommaerts, and Stephanie L. Rennane, "Firm Accommodation After Workplace Disability: Labor Market Impacts and Implications for Subsidy Design," *Econometrica*, 2025, *94* (2), 341–374.

Ameri, Mason, Lisa Schur, Meera Adya, F. Scott Bentley, Patrick McKay, and Douglas Kruse, "The Disability Employment Puzzle: A Field Experiment on Employer Hiring Behavior," *ILR Review*, 2018, *71* (2), 329–364.

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Arcot, Siva Venkatesh, "Comparative Analysis of AI-powered Outbound Dialer Campaigns vs. Legacy Outbound Dialer Campaigns (Contact Center)," *Asian Journal of Research in Computer Science*, Feb. 2025, *18* (3), 73–84.

Autor, David, "Polanyi's Paradox and the Shape of Employment Growth," in Lisa M. Lynch, ed., *Re-Evaluating Labor Market Dynamics*, Kansas City Fed, 2015, pp. 129–177.

_ , "Applying AI to Rebuild Middle Class Jobs," Technical Report, NBER WP 32140 2024.

_ **and Mark G. Duggan**, "The Rise in the Disability Rolls and the Decline in Unemployment," *Quarterly Journal of Economics*, 02 2003, *118* (1), 157–206.

_ **, Andreas Kostøl, Magne Mogstad, and Bradley Setzler**, "Disability Benefits, Consumption Insurance, and Household Labor Supply," *AER*, 2019, *109* (7), 2613–2654.

Baldwin, Marjorie L. and William G. Johnson, "A Critical Review of Studies of Discrimination Against Workers with Disabilities," *Handbook on the Economics of Discrimination*, 2006.

Banerjee, Abhijit, Rema Hanna, Benjamin A. Olken, and Diana Sverdlin Lisker, "Social Protection in the Developing World," NBER Working Paper 32382 2024.

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

Bellemare, Charles, Marion Goussé, Guy Lacroix, and Steeve Marchand, "Physical Disability and Labor Market Discrimination: Evidence from a Video Résumé Field Experiment," *American Economic Journal: Applied Economics*, 2023, *15* (4), 452–476.

Benson, Alan, Simon Board, and Moritz Meyer-ter-Vehn, "Discrimination in Hiring: Evidence from Retail Sales," *Review of Economic Studies*, 2024, *91* (4), 1956–1987.

Blader, Steven, Claudine Gartenberg, and Andrea Prat, "The Contingent Effect of Management Practices," *Review of Economic Studies*, 2020, *87* (2), 721–749.

Bloom, Nicholas and John Van Reenen, "Human Resource Management and Productivity," *Handbook of Labor Economics*, 2011, *1*, 1697–1767.

_ , Gordon B. Dahl, and Dan-Olof Rooth, "Work from Home and Disability Employment," *AER Insights*, 2025. Forthcoming.

Bound, John and Richard V. Burkhauser, "Economic Analysis of Transfer Programs Targeted on People with Disabilities," *Handbook of Labor Economics*, 1999, *3*, 3417–3528.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 333–372.

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

Burks, Stephen V., Bo Cowgill, Mitchell Hoffman, and Michael Housman, "The Value of Hiring through Employee Referrals," *QJE*, 2015, *130* (2), 805–839.

Chen, Jiafeng and Jonathan Roth, "Logs with Zeros? Some Problems and Solutions," *Quarterly Journal of Economics*, 2023, *139* (2), 891–936.

Chen, Shi, Xiaodong Wang, Weijun Li, Jingao Zhang, Yuge Qi, Jiaqi Teng, and Zhihan Zeng, "Silent Delivery: Practices and Challenges of Delivering Among Deaf or Hard of Hearing Couriers," in "Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems" CHI '24 Association for Computing Machinery 2024.

Chen, Yanyou, Yao Luo, and Zhe Yuan, "Driving the Drivers: Algorithmic Assignment in Ride-Hailing," *Available at SSRN 4299499*, 2024.

Choi, Jung Ho and Chloe Xie, "Human+ AI in accounting: Early evidence from the field," 2025.

Cook, Cody, Rebecca Diamond, Jonathan V. Hall, John A. List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," *Review of Economic Studies*, 2021, *88* (5), 2210–2238.

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," aug 2025. SSRN working paper.

Cullen, Zoe, Will Dobbie, and Mitchell Hoffman, "Increasing the Demand for Workers with a Criminal Record," *Quarterly Journal of Economics*, 2023, *138* (1), 103–150.

DeLeire, Thomas, "The Wage and Employment Effects of the Americans with Disabilities

Act," *Journal of Human Resources*, 2000, *35* (4), 693–715.

Derbyshire, Daniel W., Emma Jeanes, Esmaeil Khedmati Morasae, Susan Reh, and Morwenna Rogers, "Employer-focused interventions targeting disability employment: A systematic review," *Social Science & Medicine*, 2024, *347*, 116742.

Deshpande, Manasi, "Does Welfare Inhibit Success? The Long-term Effects of Removing Low-income Youth from the Disability Rolls," *AER*, 2016, *106* (11), 3300–3330.

_ **and Lee M Lockwood**, "Beyond Health: Nonhealth Risk and the Value of Disability Insurance," *Econometrica*, 2022, *90* (4), 1781–1810.

Erickson, William, Camille Lee, and Sarah von Schrader, "Disability Statistics from the American Community Survey (ACS)," Cornell University Yang-Tan Institute (YTI), Ithaca, NY 2026. `https://www.disabilitystatistics.org`.

Forth, John and Melanie K. Jones, "The Disability Pay Gap Within and Across Firms," IZA Discussion Paper 17679, Bonn, Germany February 2025.

Fruits, Eric and Kristian Stout, "AI, Productivity, and Labor Markets: A Review of the Empirical Evidence," Technical Report ICLE Issue Brief February 5 2026.

Gentry, Melissa, "Driving Inclusion: The Effect of Improved Transportation for People with Disabilities," *Unpublished Manuscript, Texas A&M*, 2025.

Gernsbacher, Morton Ann, "Editorial Perspective: The Use of Person-first Language in Scholarly Writing May Accentuate Stigma," *J. Child Psychology & Psychiatry*, 2017, *58.*

Giansanti, Daniele and Antonia Pirrera, "Integrating AI and Assistive Technologies in Healthcare: Insights from a Narrative Review of Reviews," *Healthcare*, 2025, *13* (5), 556.

Granquist, Christina, Susan Y. Sun, Sandra R. Montezuma, Tu M. Tran, Rachel Gage, and Gordon E. Legge, "Evaluation and Comparison of Artificial Intelligence Vision Aids: Orcam MyEye 1 and Seeing AI," *Journal of Visual Impairment & Blindness*, 2021, *115* (4), 277–285.

Grech, Lisa B., Donna Koller, and Amanda Olley, "Person-first and Identity-first Disability Language: Informing Client Centred Care," *Social Science & Medicine*, 2024, *362*, 117444.

Hoffman, Mitchell and Christopher Stanton, "People, Practices, and Productivity: A Review of New Advances in Personnel Economics," *Handbook of Labor Economics*, 2025.

Hoffmann, Manuel, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu, "Generative AI and the Nature of Work," Technical Report CESifo WP No. 11479 April 2024.

Iannone, Andrea and Daniele Giansanti, "Breaking Barriers—The Intersection of AI and Assistive Technology in Autism Care: A Narrative Review," *Healthcare*, 2024, *12* (2).

Krieger, Linda Hamilton, *Backlash Against the ADA: Reinterpreting Disability Rights*, University of Michigan Press, 2010.

Krueger, Alan B. and Douglas Kruse, "Labor Market Effects of Spinal Cord Injuries in the Dawn of the Computer Age," NBER Working Paper 5302 1995.

Kruse, Douglas, Alan Krueger, and Lisa Drastal, "Computer Use, Training, and Employment of People with Spinal Cord Injury," *Spine*, 1996, *21* (7), 891–896.

Lalive, Rafael, Jean-Philippe Wuellrich, and Josef Zweimüller, "Do Financial Incentives Affect Firms' Demand for Disabled Workers?," *JEEA*, 2013, *11* (1), 25–58.

Lane, Harlan, *The Mask of Benevolence: Disabling the Deaf Community*, AO Knopf, 1992.

Lee, Sooyeon, Bjorn Hubert-Wallander, Molly Stevens, and John M. Carroll, "Understanding and Designing for Deaf or Hard of Hearing Drivers on Uber," in "Proceedings of the 2019 CHI Conf. on Human Factors in Computing Systems" 2019, pp. 1–12.

L'Horty, Yannick, Naomie Mahmoudi, Pascale Petit, and François-Charles Wolff, "Is Disability More Discriminatory in Hiring than Ethnicity, Address or Gender? Evidence from a Multi-criteria Correspondence Experiment," *Social Science & Medicine*, 2022, *303.*

Lin, Ou, "What Do Platform Workers Think About the Law? The Ambiguous Legal Status and Legal Consciousness of On-Demand Food Delivery Riders in China," *Industrial Law Journal*, 2025.

Lise, Jeremy, Elena Pastorino, and Luigi Pistaferri, "Revisiting the Employment Effects of the Americans with Disabilities Act," 2023.

Madans, Jennifer H., Julie D. Weeks, and Nazik Elgaddal, "Hearing Difficulties Among Adults: United States, 2019," NCHS Data Brief 414, Centers for Disease Control and Prevention, National Center for Health Statistics July 2021.

Merali, Ali, "Scaling Laws for Economic Productivity: Experimental Evidence in LLM-Assisted Translation," *arXiv preprint*, December 2024. Last revised Dec 7 2024.

Milligan, Kevin, "The Long-Run Growth of Disability Insurance in the United States," in "Social Security Programs and Retirement around the World," UChicago Press, 09 2012.

Minor, Dylan, Nicola Persico, and Deborah M. Weiss, "Criminal Background and Job Performance," *IZA Journal of Labor Policy*, 2018, *7*, 1–49.

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

Olusanya, Bolajoko O., Adrian C. Davis, and Howard J. Hoffman, "Hearing Loss Grades and the International Classification of Functioning, Disability and Health," *Bulletin of the World Health Organization*, 2019, *97* (10), 725–728.

Oyer, Paul and Scott Schaefer, "Personnel Economics: Hiring and Incentives," *Handbook of Labor Economics*, 2011.

Schmidheiny, Kurt and Sebastian Siegloch, "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization," *Journal of Applied Econometrics*, 2023, *38* (5), 695–713.

Shah, Shariq, Hossein Ghomeshi, Edlira Vakaj, Emmett Cooper, and Shereen Fouad, "A Review of Natural Language Processing in Contact Centre Automation," *Pattern Analysis and Applications*, 2023, *26* (3), 823–846.

Shaw, Kathryn and Edward P. Lazear, "Tenure and Output," *Labour Economics*, 2008, *15* (4), 704–723.

Szerman, Christiane, "The Labor Market Effects of Disability Hiring Quotas," *SSRN*, 2022.

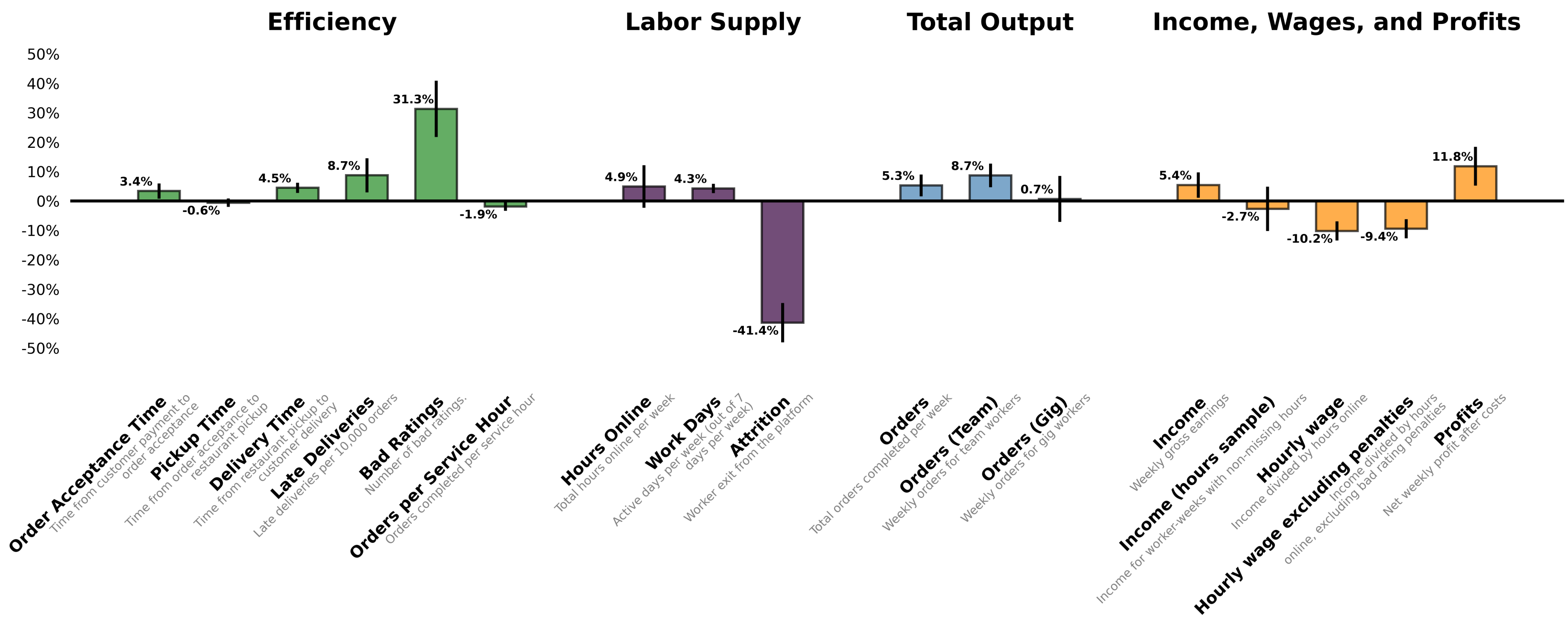
Wang, Yu, Yang Xie, Minghao Wang, Mengdan Zhao, Rui Gong, Ying Xin, Jia Ke, Ke Zhang, Shaoxing Zhang, Chen Du et al., "Hearing Loss Prevalence and Burden of Disease in China: Findings from Provincial-level Analysis," *Chinese Medical Journal*, 2025, *138* (1), 41–48.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data* 2002.

World Health Organization, "Disability and Health," Fact Sheet 2023.

_ , "Deafness and hearing loss," `https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss` 2025. Accessed 02/26.

Figure 1: Pre-AI Differences in Workplace Outcomes Between DHH and Non-Disabled Workers



Main notes: This figure visually summarizes gaps between DHH and non-disabled workers during the pre-AI period. For each outcome, each bar represents a percentage difference associated with being a DHH worker. Each bar graph comes from estimating equation (2) with a different outcome, with the coefficient shown in the bar and the 95% confidence interval accounting for clustering by driver shown in the line. For all outcomes except late deliveries, bad ratings, and attrition, percentages are derived from an inverse hyperbolic sine transformation of the dependent variable; these coefficients approximate log differences and are interpretable as percent changes. For late deliveries, bad ratings, and attrition, the plotted value is a non-transformed OLS coefficient divided by the pre-AI DHH mean, giving the percent difference relative to the pre-AI DHH baseline. See column 1 of Table A.5 for the underlying regressions plotted here.

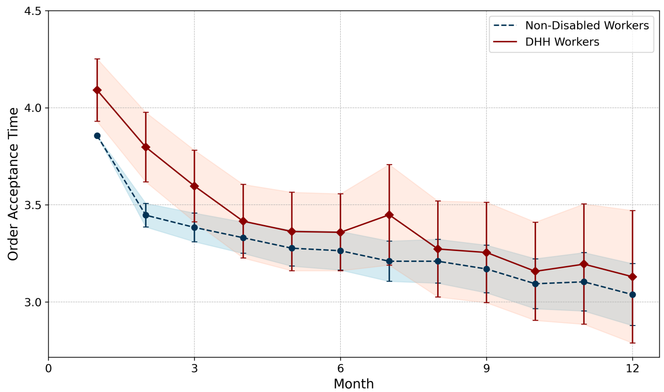
Efficiency notes: In the green bars, *Order Acceptance Time* is time from order placement to a worker accepting it, *Pickup Time* is time from accepting the order to receiving the food from the restaurant, and *Delivery Time* is time from pickup to customer delivery. These times are measured in minutes per order averaged over the week for each worker. *Late Deliveries* is the number of late deliveries per 10,000 orders. *Bad Ratings* is the number of bad ratings in a week. *Orders per Service Hour* measures orders completed per service hour, and is calculated as $60/(t_1 + t_2 + t_3)$, where the denominator is the sum of average minutes spent in each stage of an order.

Labor supply notes: In the purple bars, *Hours Online* is total hours the worker was online and available for orders, *Work Days* is number of days the worker was active in that week, and *Attrition* is a dummy for exiting the Platform.

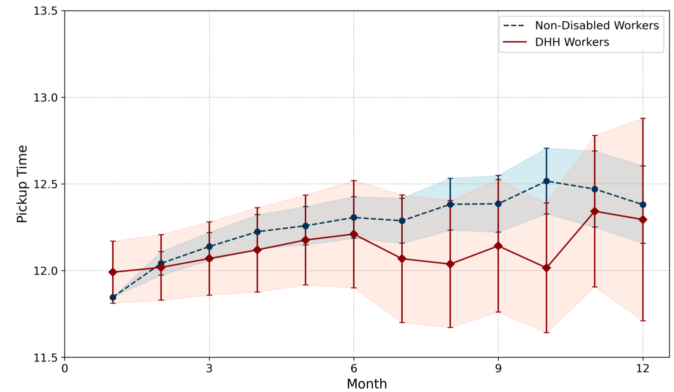
Total output notes: In the blue bars, *Orders* is the total number of delivery orders completed by a worker in a given week. We also show Orders results separately for team and gig workers.

Income, wages, and profits notes: In the orange bars, *Income* is a driver's weekly gross earnings. Since hours data are missing for some drivers, we also show results on income restricted to drivers with non-missing hours in *Income (hours sample)*. *Hourly wage* is equal to income divided by hours. *Hourly wage excluding penalties* is similar to *Hourly wage*, but excludes bad ratings penalties from income. *Profits* is profit for the Platform.

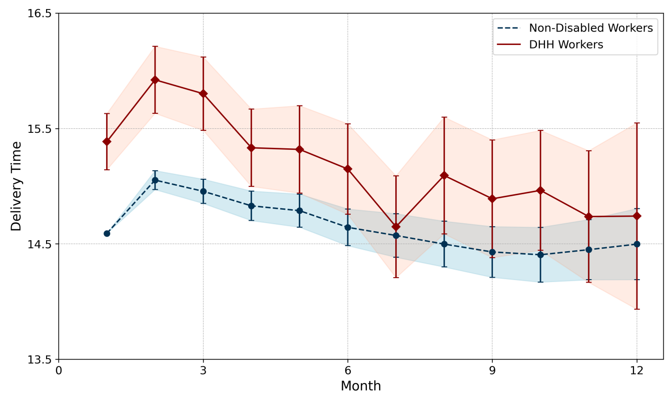
Figure 2: Pre-AI Tenure-Productivity Curves: Time Spent on Different Parts of the Order on Months of Tenure



(a) Order Acceptance Time



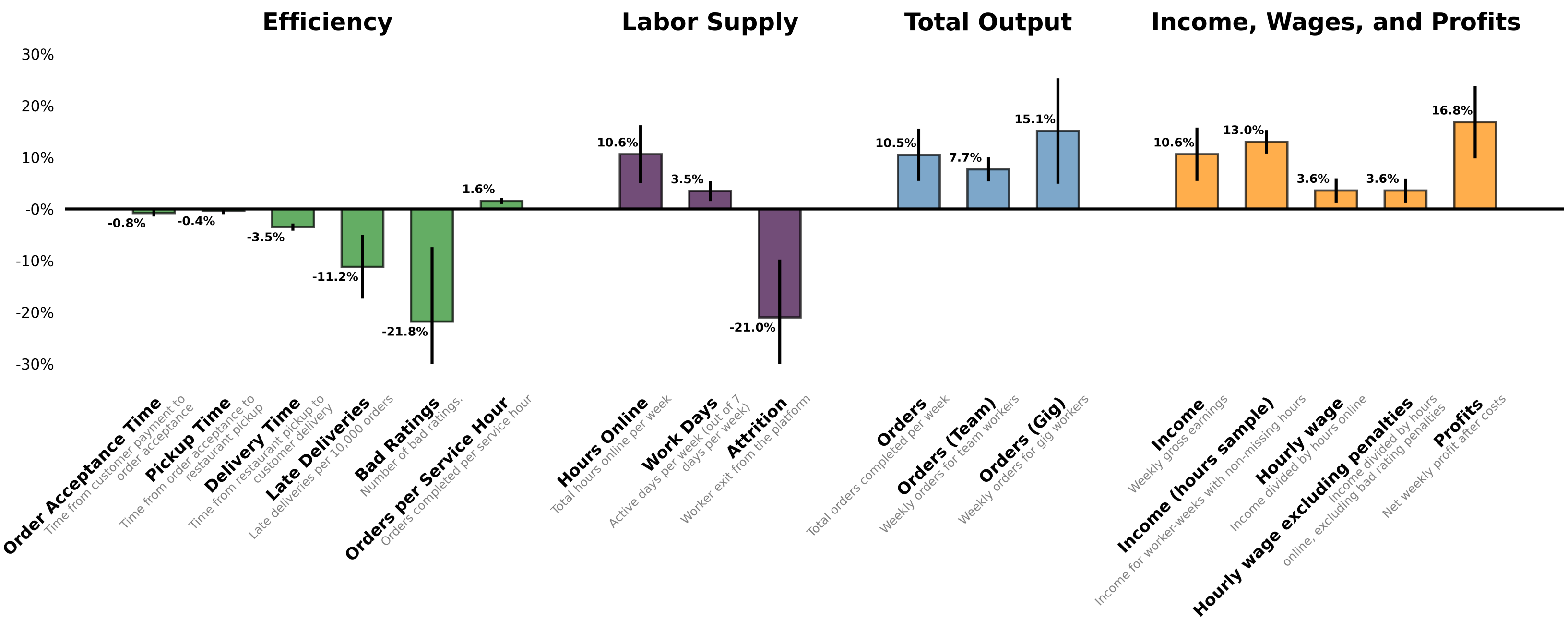
(b) Pickup Time



(c) Delivery Time

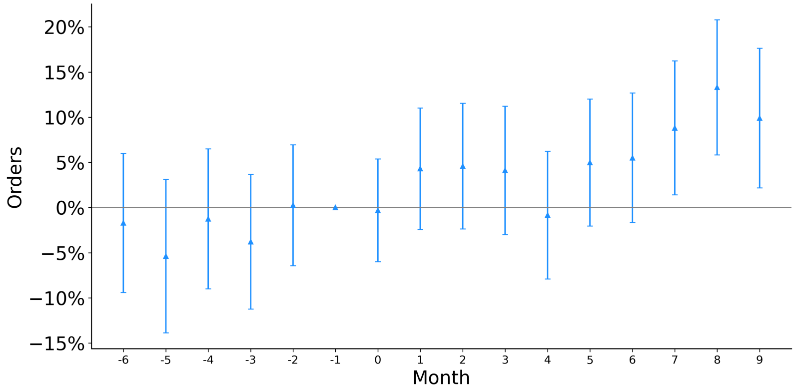
Notes: This figure examines how mean worker speed at different stages of the delivery process varies with worker tenure on the Platform. It plots time spent per order at different stages across each month of a worker's tenure, separately for DHH and non-disabled workers. The three time measures (order acceptance time, pickup time, and delivery time) are defined as in Figure 1. Each point on a curve represents the estimated average time at a given tenure month, obtained from the specification described in the main text. The red solid line corresponds to DHH workers, and the blue dashed line corresponds to non-disabled workers. The first month for non-disabled workers is normalized to zero; we then rescale all coefficients by adding back the average time for non disabled workers in their first month. Thus, at "Month 1," the figure shows the average per order time for workers in their first month of tenure. Error bars indicate 95% confidence intervals, with standard errors clustered at the worker level.

Figure 3: The Impact of the AI Tool on the Outcomes of DHH Workers

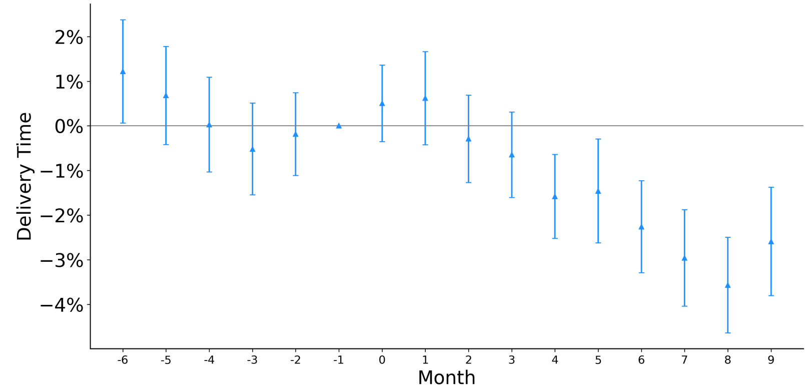


Notes: This figure shows DiD estimates of the impact of the AI-based Intelligent Outbound Call tool on DHH workers. Each bar reports the percentage change in outcomes for DHH workers relative to non-disabled workers following the introduction of the AI tool. Positive values indicate improvements for DHH workers relative to non-disabled workers (e.g., higher output or earnings). For duration-based outcomes and negative performance indicators (e.g., delivery times, bad ratings), negative values represent improvements. Error bars denote 95% confidence intervals, with standard errors clustered by workers. Outcome definitions and outcome categories follow Figure 1. Percentages are calculated as in Figure 1: inverse hyperbolic sine regressions (with coefficients shown) for most outcomes, and non-transformed OLS coefficients divided by the pre-AI DHH mean for late deliveries, bad ratings, and attrition. See column 1 of Table A.6 for the underlying regressions plotted here. The sample used here is the full sample of drivers summarized in Table 1.

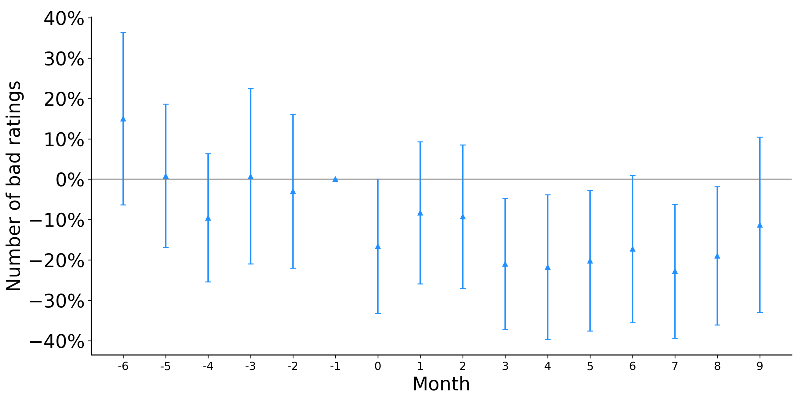
Figure 4: Event Study Analysis: AI Tool Effects on Productivity



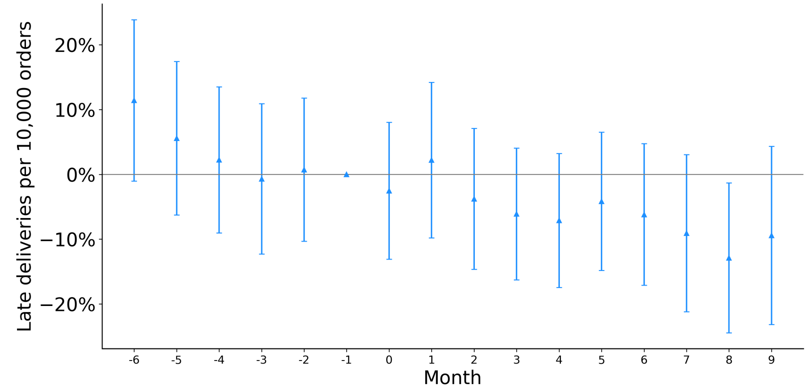
(a) Number of Orders Completed



(b) Delivery Time



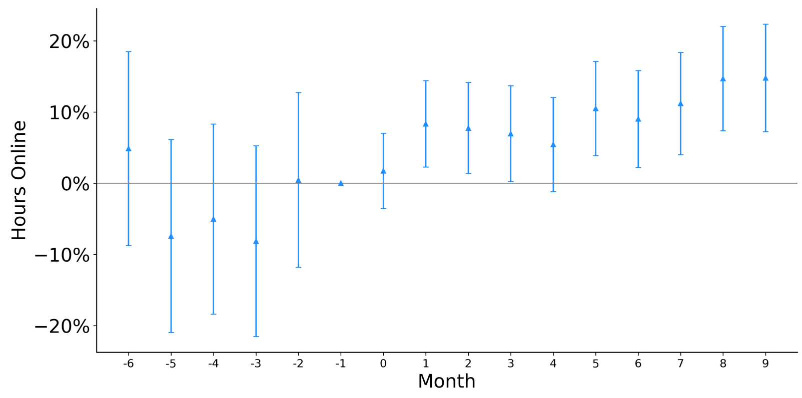
(c) Bad Ratings



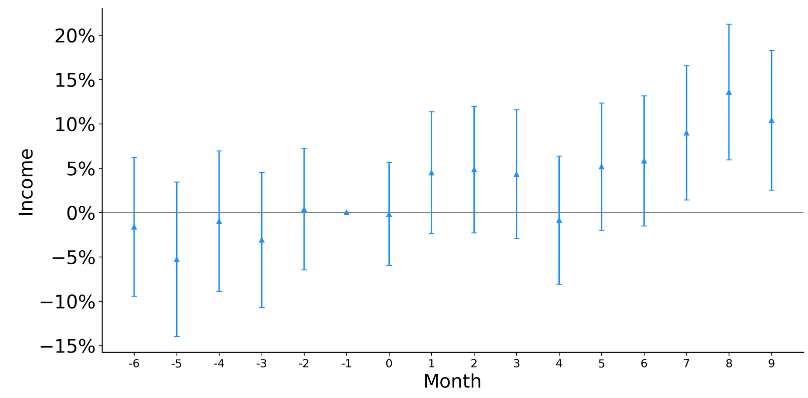
(d) Late Deliveries

Notes: This figure presents an event study analysis of the introduction of the AI-based Intelligent Outbound Call (IOC) system. The x-axis shows event time in months relative to the intervention, with 0 denoting the first month in which the tool is in effect. The underlying data are weekly; months are defined as 4-week periods, anchored on the exact week of implementation. The event study window covers 6 months before and 10 months after the intervention; observations outside this window are pooled into the boundary coefficients at -7 and $+10$. Dots show the event study coefficients, and whiskers show the 95% confidence intervals accounting for clustering by driver. The y-axis shows the estimated difference between DHH and non-disabled workers in percentage terms; percentages are calculated as in Figure 3. The sample used here is the full sample of drivers summarized in Table 1.

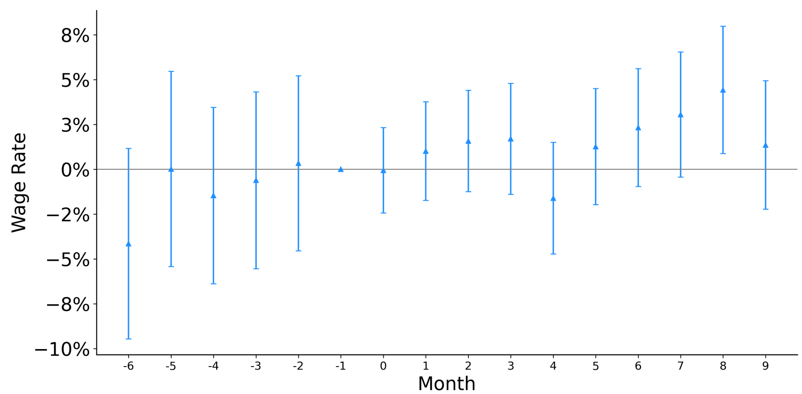
Figure 5: Event Study Analysis: AI Tool Effects on Hours, Income, Wages, and Profits



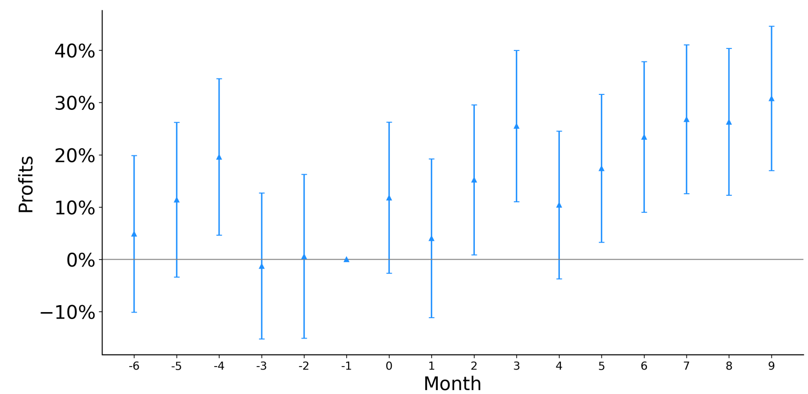
(a) Hours Online



(b) Income



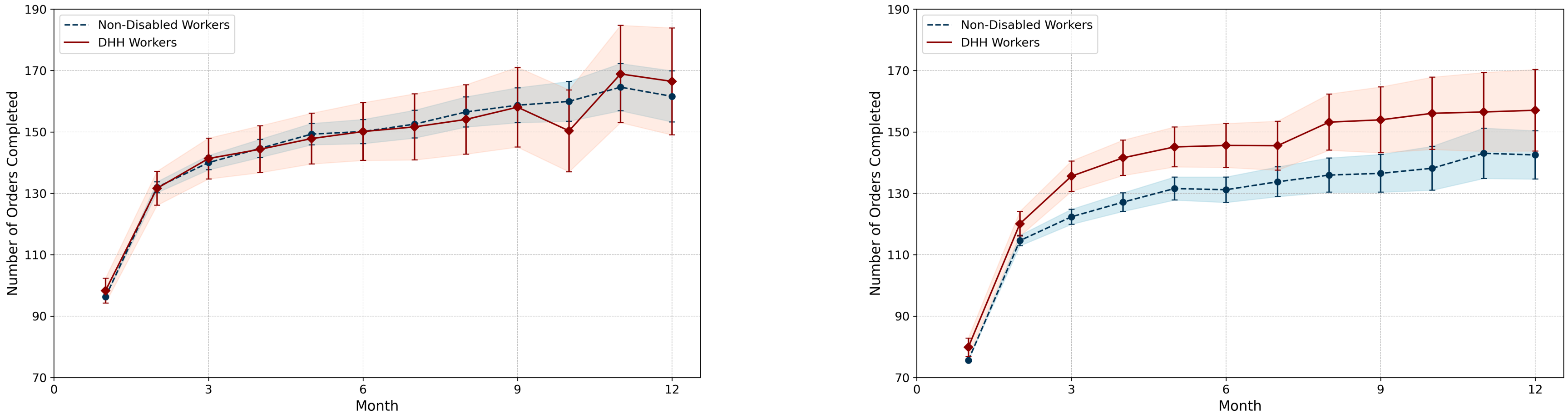
(c) Hourly Wage



(d) Profits

Notes: This figure is similar to Figure 4, but examines the effects of the AI tool on DHH workers' hours online, income, wages, and on platform profits from DHH workers. All these variables are defined in the main text and also described in Figure 1. Dots show the event-study coefficients, and whiskers show the 95% confidence intervals, accounting for clustering by worker. The y-axis shows the estimated difference between DHH and non-disabled workers in percentage terms, as in Figure 4. The sample used here is the full sample of drivers summarized in Table 1.

Figure 6: Tenure-Productivity Curves on Number of Orders Completed, Before and After the AI Tool

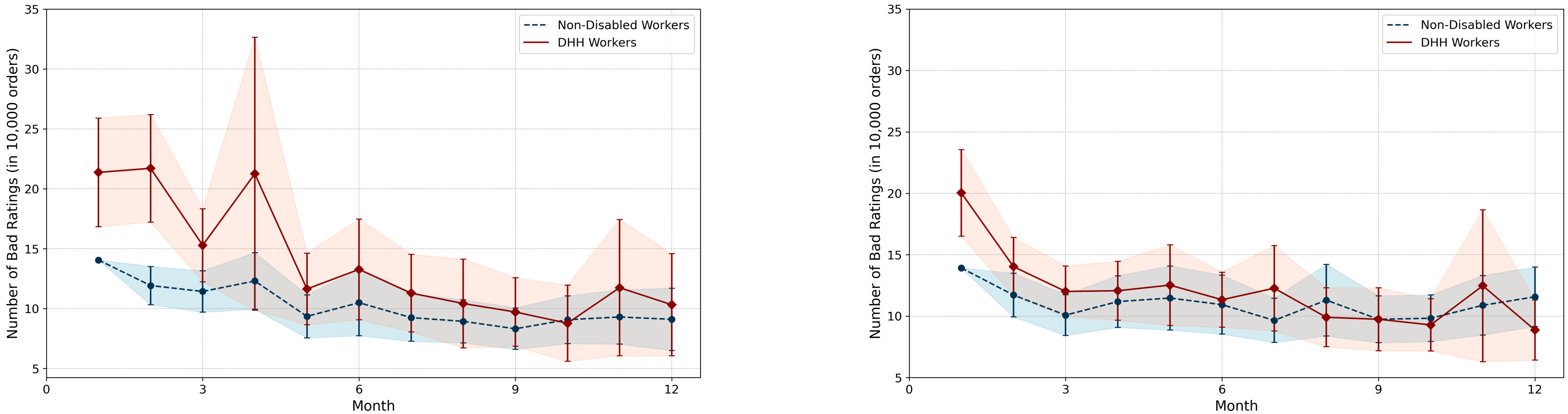


(a) Pre-AI tool

(b) Post-AI tool

Notes: This figure examines how the productivity trajectory of new workers over their first year differs before and after the introduction of the AI tool. The figure plots the average number of orders completed as a function of months of tenure. Error bars show 95% confidence intervals accounting for clustering by worker. The red solid line corresponds to DHH workers, and the blue dashed line corresponds to non-disabled workers. Tenure profiles are constructed following the same cohort-based approach as Figure 2, but are split by whether workers had access to the AI tool from the start of their tenure. The AI tool was implemented in mid 2022 at time t_0. Panel (a), *Pre-AI tool*, displays tenure profiles for workers who entered the Platform prior to the AI tool, and we include their observations until the time of AI implementation. For example, if someone joins the Platform at time $t_0 - 2$, we include two months of tenure for this worker, corresponding to months $t_0 - 2$ and $t_0 - 1$. The Panel (a) sample includes 191,139 worker-month observations from 17,732 workers, of whom 1,624 are DHH and 16,108 are non-disabled. Panel (b), *Post-AI tool*, includes workers who enter the Platform after the AI tool is introduced. The Panel (b) sample includes 186,760 worker-month observations from 25,873 workers, of whom 2,660 are DHH and 23,213 are non-disabled.

Figure 7: Tenure-Productivity Curves in Bad Ratings, Before and After the AI Tool

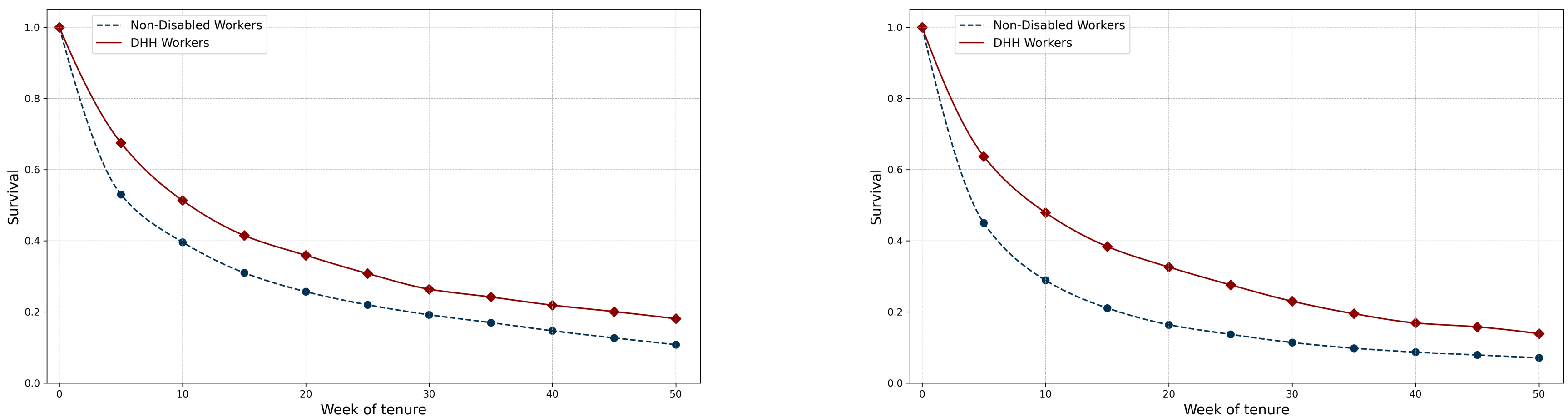


(a) Pre-AI tool

(b) Post-AI tool

Notes: This figure is similar to Figure 6, but the outcome is bad customer ratings, measured in terms of bad ratings per 10,000 orders. Error bars show 95 percent confidence intervals accounting for clustering by worker. The samples in each panel are the same as in Figure 6. Appendix Figure A.17 shows that results are similar when the dependent variable is a worker's weekly count of bad ratings instead of ratings per 10,000 Orders.

Figure 8: Survival Curves: Before and After the AI Tool

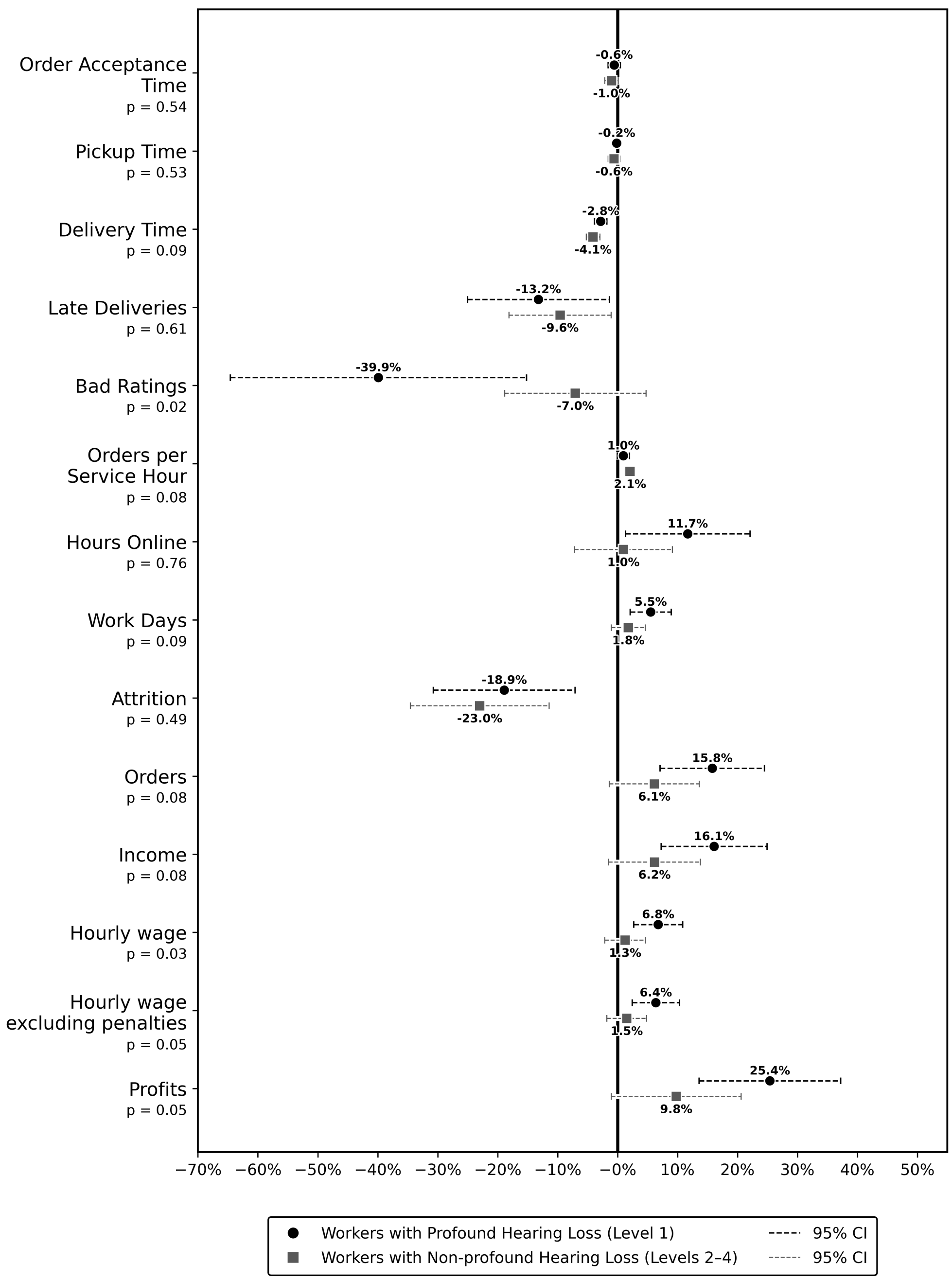


(a) Pre-AI tool

(b) Post-AI tool

Notes: This figure is similar to Figure 6, but shows survival curves as the outcome. The y-axis reports the fraction of workers remaining active on the Platform (i.e., not exited) and the x-axis reports the number of weeks since initial registration. The samples in each panel are the same as in Figure 6.

Figure 9: Heterogeneous Treatment Effects of the AI Tool by Disability Severity



Notes: This figure examines how impacts of the AI tool vary according to disability severity. Each dot-and-whisker plot is calculated using a regression where DHH workers of a given severity are compared to non-disabled workers. Each dot or square represents the treatment effect of the AI tool on DHH workers of a given severity. The whiskers show 95% confidence intervals accounting for clustering by driver. The p-values shown are from tests of whether there is a differential effect by disability severity. Each p-value is calculated from one regression with two key interaction terms, Profound $\text{DHH}_i \times Post_t$ and Non-profound $\text{DHH}_i \times Post_t$, testing whether the two interactions are different, where Profound DHH_i and Non-profound DHH_i are dummies for having profound hearing loss (HL) and non-profound HL, respectively. As described in Section 1, severity follows the official Chinese four-level classification (GB/T 26341-2010), where Level 1 is profound HL, corresponding to more than 90 decibels of HL, and Level 4 is moderate (41-60 decibels of HL). Severity level is based on a worker's government certificate of disability. The sample here is the full sample of drivers summarized in Table 1.

Table 1: Mean Characteristics of DHH and Non-disabled Workers

	DHH Workers (1)	Non-Disabled Workers (2)	Full Sample (3)	P-Value of Difference (4)
Panel A: Worker Demographics				
Age	32.8	30.3	30.5	0.00
Male	92.6%	90.8%	91.0%	0.00
Panel B: Work Model				
Team worker	35.4%	36.4%	36.3%	0.18
Gig worker	64.6%	63.6%	63.7%	0.18
Panel C: Disability Severity for DHH Workers				
Level 1: Profound hearing loss	53.6%			
Level 2: Severe hearing loss	16.5%			
Level 3: Moderate-Severe hearing loss	14.2%			
Level 4: Moderate hearing loss	15.7%			
Panel D: Distance & Geography Variables				
Driver is in a big city	36.6%	32.3%	32.7%	0.00
Driver is in Shanghai	7.27%	5.58%	5.75%	0.00
Driver is in Beijing	3.60%	3.36%	3.38%	0.40
Driver is in Eastern China	64.2%	58.2%	58.8%	0.00
Avg. distance bwn. restaurant and customer, in km	2.24	2.16	2.17	0.00
Unique restaurants served, per week	73	65	66	0.00
Number of workers	4,284	39,321	43,605	
Number of worker-weeks	70,301	409,769	480,070	

Notes: This table provides summary statistics on food-delivery workers on the Platform, both deaf or hard of hearing (DHH) and non-disabled. We use data from the whole sample period of 2021-2023. Column (1) shows DHH worker averages; (2) shows non-disabled worker averages; (3) is the full sample average. Column (4) reports the p-value for the difference in means between DHH and non-disabled workers. The unit of observation is a worker for demographics. In Panel A, age is age upon joining the Platform. In Panel B, Team worker and Gig worker are dummies for a driver's work model, also called their registration type. Panel C reports the distribution of disability severity among the DHH workers in column 1. Severity follows the official four-level classification (GB/T 26341-2010). Level 1 is profound hearing loss, meaning loss of greater than 90 decibels; Level 2 is severe hearing loss of 81-90 decibels; Level 3 is moderately severe hearing loss of 61-80 decibels; and Level 4 is moderate hearing loss of 41-60 decibels. Level 4 is the lowest level of severity to have a certificate of disability in China. Severity is based on government medical certificates of disability, which are provided by drivers to the Platform. Big cities cover the top 10 cities in China by population: Chongqing, Shanghai, Beijing, Chengdu, Guangzhou, Shenzhen, Wuhan, Tianjin, Xi'an, and Suzhou. Eastern China provinces include: Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong, and Hainan. Location information is missing for 2,904 workers in the full sample (about 7%).

Table 2: Descriptive Statistics for Key Outcome Variables, all at the Worker-Week Level

	Mean (1)	Median (2)	Std. Dev. (3)	Obs. (4)	Workers (5)
Panel A: Efficiency/productivity: Quantity and Quality					
Order Acceptance Time: minutes from customer payment to order acceptance	3.37	2.48	9.20	464,717	42,281
Pickup Time: minutes from order acceptance to restaurant pickup	11.97	11.47	6.86	464,850	42,293
Delivery Time: minutes from restaurant pickup to customer delivery	14.18	13.44	5.10	465,546	42,367
Late Deliveries per 10k orders	416.44	134.23	816.42	480,070	43,605
Bad Ratings per Week	0.093	0.00	0.365	480,070	43,605
Week where worker gets at least one Bad Rating, 0 or 1	0.076	0.00	0.265	480,070	43,605
Good Ratings per Week	9.32	6.00	12.30	480,070	43,605
Week where worker gets at least one Good Rating, 0 or 1	0.816	1.00	0.387	480,070	43,605
Panel B: Labor Supply on the Platform					
Hours Online, in hours per week	40.42	37.85	29.19	343,733	34,965
Work Days per week	5.31	6.00	2.13	480,070	43,605
Driver attrites from the Platform, 0 or 1	0.077	0.00	0.267	456,192	40,567
Panel C: Total Output					
Total Orders	152.39	127.00	141.52	480,070	43,605
Peak Hour Orders	90.92	76.00	88.42	480,070	43,605
Total Orders, restrict to Team drivers only	212.35	205.00	146.18	272,986	15,828
Total Orders, restrict to Gig drivers only	73.35	42.00	85.28	207,084	27,777
Panel D: Income, Wages, and Profits					
Driver Income, in CNY per week	Confid.	Confid.	Confid.	480,070	43,605
Driver Income, restrict to weeks with hours data, in CNY per week	Confid.	Confid.	Confid.	343,733	34,965
Hourly Wage, excluding any bad rating penalties from income, in CNY per hour	Confid.	Confid.	Confid.	342,061	34,909
Hourly Wage, in CNY per hour	Confid.	Confid.	Confid.	342,061	34,909
Platform Profits from a driver, in CNY per week	Confid.	Confid.	Confid.	480,070	43,605

Notes: This table summarizes weekly work outcomes for workers on the Platform. We use data from the whole sample period of 2021-2023. The unit of observation is a worker-week. Outcome definitions and grouping into the four categories are the same as in Figure 1. The "Confid." marking indicates that certain summary statistics cannot be shared due to confidentiality requirements. An observation in column 4 is a worker-week. The attrition outcome has fewer observations than other outcomes because it excludes the last 7 weeks of 2023: since attrition is defined as 8 consecutive weeks of no activity, the last week in which an attrition decision can be observed is the 8th-to-last week of the sample. Appendix Table A.2 presents summary statistics separately for DHH and non-disabled workers. Appendix Table A.3 presents summary statistics separately for gig and team workers.

Table 3: Sources of Pre-AI Profit Differences and of Profit Impact of AI Tool

Component	Share
Panel A: Sources of Pre-AI Profit Differences	
Orders	85%
Bad Ratings	-4%
Onboarding	19%
Panel B: Sources of Profit Impact	
Orders	93%
Bad Ratings	2%
Onboarding	5%

Notes: Profit has three components as in equation (1): profit from orders (order revenue minus wage payments), penalty from bad ratings, and onboarding cost. To assess the relative importance of each component, we convert all three to a common CNY scale. For each component j, we multiply the estimated percentage difference $\hat{\beta}_j$—from Figure 1 for Panel A and from Figure 3 for Panel B—by the component's signed average profit contribution $\bar{y}_j$ (negative for bad-rating penalties and onboarding costs) to obtain an implied CNY impact $\Delta y_j = \hat{\beta}_j \cdot \bar{y}_j$. Each component's share is its implied CNY impact divided by the sum across all three components.