Better than Human?
Experiments with AI Debt Collectors

James J. Choi
Yale University and NBER

Dong Huang
Yale University

Zhishu Yang
Tsinghua University

Qi Zhang
Shanghai Jiaotong University

February 15, 2024

Abstract: How good is artificial intelligence (AI) at tasks that require persuading humans to perform costly actions? We study the effectiveness of phone calls made to persuade delinquent consumer borrowers to repay their debt. Both a regression discontinuity design and a randomized experiment reveal that AI is substantially less able than human callers to get borrowers to repay. Substituting human callers for AI six days into delinquency closes much of the collection gap, but one year later, borrowers initially assigned to AI and then switched to humans have repaid 1% less than borrowers who were called by humans from the beginning. Even accounting for wage costs and assuming zero costs for AI, using AI is less profitable (with the caveat that we do not observe non-wage costs of labor). AI's lesser ability to handle complex situations and extract payment promises may contribute to the performance gap.

We thank Manlin Sun for excellent research assistance.

Rapid progress in artificial intelligence (AI) has revived the long-standing debate on the extent to which new technologies will replace human jobs.[1] There is now an extensive literature studying the performance and economic impact of AI in routine and prediction tasks.[2] In this paper, we study the effectiveness of AI in a different sort of task: persuading a human to take a personally costly action. The task is non-routine, requires social interaction, and is aided by emotional intelligence. Many service and managerial jobs require performing this type of task, and recent advances have plausibly improved the capacity of AI to do this well.

The specific task we study is persuading delinquent consumer borrowers to repay their debt. We obtain debt collection data from a leading online consumer finance company in China that makes uncollateralized installment loans. Borrowers who fail to make their monthly payment on time are contacted on the phone by the company's debt collectors, urging them to repay. The company uses both human and AI callers, giving us an opportunity to evaluate AI callers' performance relative to humans and to estimate the impacts of AI on the company's profits and worker productivity. The AI callers can understand the borrower's speech and generate appropriate voice replies according to some templates. They call borrowers automatically, provide them with basic information, answer simple questions, inform them of the negative outcomes of defaulting, and ask for a promise to pay in an interactive manner, which resembles human callers.

We identify the effectiveness of these AI callers relative to human callers using two experiments that occurred in the firm, one natural and one intentional. The natural experiment is created because of the company's rule that newly delinquent debts with remaining principal no greater than 300 yuan (approximately 42 U.S. dollars) are permanently assigned to AI callers, whereas larger debts are transferred to human callers no later than six days after delinquency begins. Therefore, we can identify the effect of permanent versus temporary assignment to AI using a regression discontinuity design around the 300 yuan threshold. The intentional experiment is created because the company takes a random 10% of newly delinquent debts with remaining

[1] For academic research, see Brynjolfsson and Mitchell (2017), Felten et al. (2020), Eloundou et al. (2023), World Economic Forum (2020, 2023). For general public media discussions, see for example Elon Musk's speech at the first AI Safety Summit 2023 (https://www.cnbc.com/2023/11/02/tesla-boss-elon-musk-says-ai-will-create-situation-where-no-job-is-needed.html) and Harvard Business Review article "AI Isn't Ready to Make Unsupervised Decisions" (https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions).

[2] For the performance and impacts of AI on routine and prediction tasks, see for example Cao et al. (2021) about stock analyses, Erel et al. (2021) about nominating company directors, and Kleinberg et al. (2018) about bail decisions. Also refer to Agrawal et al. (2019) for a good summary. For the application of AI on non-routine jobs, see Noy and Zhang (2023) about how generative AI can assists humans in writing tasks in an experiment, and Brynjolfsson et al. (2023) about generative AIs in the customer service industry.

principal greater than 300 yuan each month, assigns a randomly chosen half to be called by AI through day 5 before being called by humans thereafter (the treatment group), and assigns the other half to always be called by humans (the control group). All debts in this 10% subsample are reallocated to human callers on day 6, so this intentional experiment identifies the effect of a short-lived initial exposure to AI callers versus no exposure to AI callers.

We find in the regression discontinuity sample that when AI callers are permanently assigned to a borrower, they consistently perform worse than human callers over horizons up to one year past due, as measured by the net present value (NPV) of collected repayment cash flows scaled by the total outstanding balance at initial delinquency. The productivity gap between AI and human callers first widens as overdue days increase. It reaches its maximum around one month past due, when the NPV of repayments collected by AI callers is 11 percentage points less than human callers. The gap slowly narrows afterward but remains at around 7 percentage points even after one year past due. In addition, the gap is larger for borrowers with lower internal credit scores. One possible view is that providing reminders and information are the main function of debt collectors. The persistent gap in performance between AI and humans, and its heterogeneity by credit quality suggests that soft communication skills are also important.

The randomized experiment shows that replacing AI callers with human callers after a few days mitigates much of the initial underperformance of AI callers. In this subsample, we continue to find that AI underperforms humans, with the NPV gap monotonically increasing to 12 percentage points by day 5. But the gap quickly narrows once human callers take over the AI cases to 2 percentage points at day 10 and 1 percentage point at day 30. Interestingly, the remaining 1 percentage point gap never closes even one year later, indicating that initial contact by AI *permanently* impairs the ability of the company to collect.

We explore the potential sources of the AI performance gap by examining detailed outcomes of phone conversations in the randomized experiment sample, restricting the sample to phone calls on the first day of contact. Humans call borrowers nearly one more time per day than AI callers. To remove the impact of additional phone calls, we further restrict our sample to the first call answered by borrowers. After controlling for the call's time of day, we find that AI callers have conversations that are 31 seconds shorter on average and exhibit less variability in length, suggesting that AI callers are less capable of handling complex situations. Moreover, 21% fewer borrowers promise to repay their debts and about one-third fewer repay the debts within 2 hours after answering the calls if they talk to AI callers. Therefore, AI callers appear to be worse than humans in eliciting promises and imposing pressure to repay.

We next consider the interactions between AI and human callers, especially how AI adoption affects the productivity of human callers. During our sample period, the AI software experienced four upgrades, mainly improving speech recognition and language understanding abilities. The upgrades were rolled out in a progressive manner so that two consecutive versions of AI callers were used simultaneously in the same month and were assigned cases at random. This arrangement allows us to measure the improvements in AI caller productivity and examine how human callers perform after receiving cases treated by better AI. We observe that the AI significantly improved between August and October 2021, increasing collected NPV at day 5 by 3 percentage points. The AI improvement, however, does not lead to better performance of human callers when they take over the cases on day 6; human callers on day 6 collect 3 percentage points less, resulting in similar cumulative collected NPVs. Moreover, declines in human productivity are larger among better callers, as measured by their performance rankings in the previous month. This would be the case if the AI improved by learning the communication strategies of the best callers. These findings are consistent with the displacement effect of new automation technologies on labor.

Finally, we address the role of labor costs saved by AI adoption. We focus on the direct labor costs, i.e., workers' salaries, which consist of a fixed component and a variable component. The fixed component is paid to human callers and is related to the total volume of debts that need to be collected but not to collection outcomes. The variable component is increasing in the total amount of collected debts. We estimate the average rates of salaries for both components and adjust the collected NPV measure to account for direct labor costs while assuming the marginal cost of an AI call is zero. Although the productivity deficit of AI is diminished once labor costs are accounted for, AI remains less cost-effective than human callers. Importantly, this calculation does not consider indirect labor costs, such as recruitment, training, management, pension funds, etc., nor the cost of developing the AI software.

Our paper is closely related to the literature about the impacts of automation on labor. Previous studies find different impacts in different waves of automation.[3] They mostly find complementarity

[3] In the early robot and information technology revolutions, some researchers find displacement effects among low-skilled workers and increased demands for high-skilled workers (Acemoglu and Restrepo, 2020, 2022). Others find that new automation technologies are labor-augmenting (Michaels et al., 2014; Tan and Netessine, 2020).

between humans and AI, especially for low-skilled workers (Gao and Jiang, 2021; Brynjolfsson et al., 2023; Noy and Zhang, 2023), when AI only provides predictions and suggestions and human workers make the final decision. In contrast, the company in our study has to delegate the whole phone call to AI, since it is hard for AI to assist human callers in real time during conversations. In such a setting, we find imperfect displacement effects; AI callers can replace humans but are less productive overall.

Additionally, our study is related to literature about the performance of AI and machine learning technology (Cao et al., 2024; Erel et al., 2021; Kleinberg et al., 2018; Agrawal et al., 2023). Our research contributes to this strand of literature by focusing on non-routine jobs, which were previously believed to be immune to automation (Brynjolfsson and Mitchell, 2017, Felten et al., 2020) and were rarely studied until recently. Some examples are text chatbots for customer service (Gao and Jiang, 2021; Brynjolfsson et al., 2023) and AI autocomplete in writing tasks (Noy and Zhang, 2023).

Finally, our paper contributes to emerging literature about debt collection. Drozd and Serrano-Padial (2017) and Fedaseyeu (2020) examine how variations in debt collection effectiveness driven by information technology and regulations affect credit supply. Fedaseyeu and Hunt (2015) study the reputation concerns in using third-party debt collection. At the micro-level, our paper is closely related to Laudenbach and Siegel (2023) which address the importance of personal communication in collecting loan repayments. They show that phone calls to borrowers from bank agents are more effective than mail reminders. Our findings suggest that this gap persists even if borrowers are contacted by automated AI callers who can interact with borrowers.

The remainder of the paper is as follows. Section 1 provides institutional background about the company, its debt collection process, and its human and AI callers. Section 2 describes data and Section 3 specifies our experimental setups. Section 4 evaluates the performance gap between AI and human callers. Section 5 examines the interactions between AI and human labor. Section 6 concludes.

1 Institutional Background

1.1 The company and its lending business

The company is a leading online consumer finance service provider in China. At the end of 2022, the company had around 10 million active users with nearly 7 billion yuan (980 million

USD) outstanding loan balance. The company's main business is to work with Chinese commercial banks to originate consumer loans to online consumers. These commercial banks provide funding and are the creditors of the loans while the company provides FinTech-powered credit services that can cover the whole loan origination process of customer acquisition, credit rating, anti-fraud screening, post-origination monitoring, and delinquent loan collection. Partnered commercial banks have different requirements about aggregate default risk, duration, rates of returns, maximum loan sizes, etc. The company's task is to match customers' risk profiles with banks' requirements at origination and to manage the risks afterward. Usually, the company serves as the guarantor of the loans. Profits of the company are mainly service fees and commissions from the partnered commercial banks, net of losses on default loans as the guarantor.[4]

The company targets young consumers with short credit history but large income and consumption growth potential.[5] It operates its own online shopping platform, and also collaborates with third-party online retailers to promote consumption and offer loans at the point of sale. The loan size typically ranges from tens to thousands of yuan to cover the purchases: the 10th percentile of the loan size is only 8 yuan (1 USD) and the 90th percentile is around 5500 yuan (770 USD). The company provides two types of loans at the point of sale. The first type is the uncollateralized personal installment loan, in which the consumer can borrow the money and repay it later in the next six months to one year according to a monthly repayment schedule. The monthly repayments are of equal size, each of which consists of monthly interest payments and principal paydown. The other type of loan is offered by a credit-card-like product of the company. The consumer may apply for a credit line, which is around 7500 yuan (1050 USD) on average, and pay their online order with the credit line as using a credit card. Repayment of the "credit card" balance is due monthly, and it is not allowed to split the repayment over several months. "Credit cards" are typically used for small payments while installment loans are preferred with expensive purchases. These loans are similar to what one can have from a commercial bank, which may not be feasible for the company's target customers.

[4] 84% of the revenue is from service fees and commissions of credit origination, the other 16% of the revenue is from the sales of the company's online shopping platform, which it uses to attract new customers and boost the consumption of existing customers.

[5] 70% of the company's customers are less than 30 years old, 65% are urban working population, and 13% of them have a bachelor's degree or more. These numbers are much higher than the population averages.

After the consumer applies for the loan, the company will use its big-data risk models to evaluate their credit risks. The inputs of the model contain a wide range of variables collected from different sources, such as the consumer's demographic information, credit records with the company and from third-party credit aggregators, consumption habits, cell phone usage behaviors, etc.[6] The company also manages to detect fraudulent borrowings with their machine learning-driven models by examining suspicious clustering in physical and IP addresses, application time, cell phone types, and so on. According to the company's internal reports, 80% of new customers are rejected because of either high credit risks (95%) or skeptical behaviors (5%). For the remaining 20% who pass the screening process, the company will sign a loan agreement with them, specifying loan size (or credit line), maturity, interest rate, and repayment schedule, which are determined by the customer's credit risks. Since the customers typically have higher credit risks than the population average, the interest rates are mostly 24% per annum, which is the upper limit set by Chinese regulators to protect consumers. The borrower is also asked to list two "emergency contacts" as their "guarantors," who are usually their parents, family members, or colleagues.[7]

Once the loan or credit line is originated, the consumer can pay their purchases with the loan, receive the products, and start repaying the loan according to the repayment schedule. Each borrower is assigned a monthly repayment due day, which may be changed by the borrower with the company's approval. Changing the due date frequently is not allowed by the company. Ten days before the due date is the bill date, when the borrower receives their monthly repayment bill stating the amount of money they have to repay by the due date. Payments can be made with their debit cards or mobile payment accounts, such as AliPay and WeChat Pay. The borrower may also set up auto-payment, in which case the company will automatically charge the bill amount for the account given sufficient balance.

If the borrower fails to pay all the bill amounts by the due date, their debts are considered delinquent and enter the debt collection process. During the process, debt collectors will call the borrowers at a relatively high frequency and persuade them to repay the debts as soon as possible.

[6] The major credit aggregator in China is operated by the People's Bank of China, the central bank. Others are operated by alliances of Internet consumer finance companies which agree to exchange customers' credit records with each other. Large companies like Alibaba Ant Financial Group also have their own credit score systems to share with other companies.

[7] The company is still the only *legal* guarantor who is responsible for the loan repayments. These "emergency contacts" do not have any legal obligation to repay the loans if they default. The company uses these "emergency contacts" as a backup contact approach if the borrower defaults and refuses to talk to the debt collectors. This can also impose some social pressure on the borrower.

During the phone calls, the debt caller usually provides information about the loans, informs the borrower of the potential negative consequences of delinquency, and makes suggestions to solve the borrower's financial difficulties. Reminders are also sent by text messages and mobile App notifications. More details about the process are discussed in the following section. Three months (90 days) after the first delinquency, if the borrower is still unable to repay the debts, they are considered defaulting and reported to third-party credit aggregators. Defaulting borrowers cannot borrow from the company again and may experience difficulties when borrowing from other consumer finance companies. Defaulting may also affect the borrower's daily consumption such as ridesharing and hotel booking since some big data-driven companies also use credit records for screening.[8] If the company can prove that the borrower has enough money but does not repay the debts, it can sue the borrower. If the lawsuit is supported by the court but the borrower still refuses to repay, the borrower will be added to a blacklist of "dishonest judgment debtors" assembled by the supreme court and prohibited from expensive consumption such as traveling by plane, purchasing real estate properties or cars and so on.

1.2 Debt collection process

Once borrowers fail to pay their bills fully by the end of the repayment due date, their debts are classified as delinquent and enter the debt collection process. These delinquent debts to be collected are called "cases" by the company. Since many borrowers are just inattentive to their bills and money transfers may take time, the company considers the first day past due as a grace period and only makes limited contacts. It generally does not call the delinquent borrowers and just sends them reminders through text messages and cell phone App notifications. If the debts remain unpaid on the second day past due, the company starts calling the borrowers using AI or human callers.

The company divides cases into different stages based on the duration of delinquency at roughly monthly frequency. Specifically, there are stages for cases from days 2 to 25 past due (labeled as "M1" stage, where "M" stands for "month"), cases from days 26 to 59 past due ("M2" stage), cases from days 60 to 89 past due ("M3" stage), and cases delinquent for 90 days or more

[8] For example, when booking hotels, defaulting borrowers may be asked to pay more deposits or to pay the full expenses in advance.

("M3+" stage). Cases in the M1 stage are further split into the "M1 Early" stage (days 2 to 10 past due) and the "M1 Late" stage (days 11 to 25 past due).

The company uses different debt collection strategies in different stages of cases. In the M1 Early stage (days 2-10), AI callers may be used in days 2-5 to partially replace human callers. This is the stage of our interest, and we will discuss the case assignment rules between AI and human callers in Section 3. In this stage, cases assigned to human callers are rotated among callers at a daily frequency. For an individual caller, they are randomly assigned a list of borrowers by the company at the beginning of the day, which has about 150 to 200 cases a day for each caller on average. There are three blocks of time when borrowers are called automatically by the system, namely, 9 to 9:30 A.M., 3 to 3:30 P.M., and 7:15 to 7:35 P.M. During these time blocks, the system automatically tries calling every borrower in the list who have not repaid or promised to repay the debts. Callers only need to sit in front of the screen and stand by. Once the phone call is answered by the borrower, it is immediately transferred to the human caller, who will talk to the borrower.[9] This arrangement is to make sure that all borrowers are called at least three times a day, regardless of whether they answer the phone.

Outside these three "automatic callings" time blocks, human callers can choose which borrowers to call first based on debt characteristics shown on their screen during their working hours from 9 A.M. to 8 P.M. Appendix Figure D1 displays a snapshot of the interface when a caller logins the company's debt collection system. The lower part of the screen displays the list of cases assigned to the caller and relevant information. Main information includes debt characteristics (days and amounts overdue, remaining principal, loan type), borrower information (age, place of residence, borrower tag (about education level), internal credit score), the most recent time the borrower logged into the App, the time to follow-up, case status (how likely the borrower picks up the phone), communication results, etc. The upper part of the screen provides a comprehensive filter on the cases. The caller is able to choose cases based on almost all dimensions mentioned in the lower part of the screen to prioritize their callings. Based on the company's internal research, the filter used most by productive callers is about the most recent time that the borrower logs into the App. The callers' rationale is that borrowers who login and view their bills frequently are likely

[9] During this procedure, a worker may receive phone calls with borrowers who are not initially assigned to them at the beginning of the day. Once they receive the call, the corresponding cases are transferred to their own list. Our data records the actual worker talking to the borrower. In addition, the assignment of answered calls is random across workers.

to worry more about their debts, so they are easier to be persuaded, compared to their inattentive counterparts. However, the research also suggests a relatively minor role of case selection skills in explaining human callers' performance heterogeneity. In addition, to avoid the company's phone numbers being blacklisted by borrowers, the company provides multiple phone numbers, and the caller can choose which one to be displayed on borrowers' cell phones.

During the phone call, the callers usually provide information about the loans, inform the borrower of the potential negative consequences of delinquency, and persuade them to repay the debts as soon as possible. The callers are provided with some sentence templates but are not asked to follow them strictly. The callers may also provide suggestions to the borrowers to help solve their financial difficulties, such as encouraging them to ask their family members for help. Given that the phone calls in the M1 Early stage only last for about 1 minute on average, these suggestions are typically short and generic. In later stages when the callers have more opportunities to talk to the borrowers, the conversations are more personalized and specific.

After each phone call, the caller is required to label its outcome. If the phone number is not answered multiple times, it may be labeled as potentially invalid. The average phone answering rate is only 23.6%. If, instead, the phone is picked up and in the conversation the borrower clearly and explicitly states that they will repay the debts immediately or within some period no later than the end of the next day, the caller will label it as "promise to pay" and the case will be held by the same caller for one more day, and the repayment will count as the caller's contribution as long as it comes within the promised time; otherwise, the case will be assigned to another caller in the next day. Therefore, callers have the incentive to ask borrowers to make promises. The company also uses an AI conversation examiner to monitor all the conversations to make sure that any false labels of "promise to pay" without actual explicit promises are penalized by salary deductions.

The AI examiner also checks callers' other behaviors that violate rules set by the company and the regulator. Most importantly, to comply with regulations and to maintain a positive image in the public, the company does not allow callers to say inappropriate words, including swear words, threats, discrimination, false information, unwarranted promises to borrowers, etc. The AI examiner can recognize the content of the conversations and look for any misconduct. Also, the regulation prohibits any phone calls between 8 P.M. and 9 A.M. the next morning every day. For the frequency of calling an individual borrower, there is no specific limit in the regulation, but an implicit standard is about 3 to 6 calls per borrower per day.[10] Extra phone calls a day are often considered as abusive debt collection by the court in practice. In our data, borrowers receive five phone calls a day on average. The company also penalizes phone calls outside the regulated time range and borrower complaints about excessive phone calls.

For later stages of the cases, the company uses different debt collection strategies. Since no AI callers are used in these stages, they are not particularly relevant to this paper and the discussion will be brief here. First, in later stages, the company will collaborate with third-party debt collection agencies. Around 60% of the overdue debts are outsourced to third-party debt collection agencies during the M2 stage, and around 90% are outsourced during the M3 stage. In the M3+ stage, the remaining overdue money is considered as default and recorded as bad debt loss. The company sends almost all cases to third-party debt collection agencies. They only keep small cases to be collected by AI callers (see Section 3) and some very large ones for further actions like lawsuits. In the M1 Late stage (days 11-25), each caller will handle cases for one week (7 days) before it is assigned to another person. In the M2 and M3 stages, this interval is typically two weeks (14 days). In the M3+ stage, since most cases are handled by third-party agencies, we do not have detailed information about their strategies. As the stage of the debt collection process extends, borrowers are typically contacted less frequently. The process stops once the borrower has repaid the full overdue payments.

1.3 Caller compensation scheme

Callers are paid monthly based on their performance measured by the amount of overdue money they collect. Their salary mainly consists of two components. The first component is called the "ranking salary," which is increasing in callers' performance ranking within a group of callers who work in the same stage of cases and have similar tenures in the company. The relationship between ranking salary and performance ranking is convex, meaning that the higher the ranking, the faster the ranking salary grows. The top caller can earn 5500 yuan a month while the bottom 5% earn nothing for their ranking salary. The second key component is the "completion salary," which is a function of the "target completion rate." Specifically, at the beginning of a month, the

[10] According to the *Self-discipline Convention for Internet Finance Overdue Debt Collection Activities* published by the National Internet Finance Association of China (a self-regulatory organization) in March 2018, "Collectors should carry out debt collection activities at the appropriate time (of a day) and are not allowed to harass debtors and other related persons with frequent calls." (§2.17) However, there is not a clear limit, and the number of phone calls that should be considered as harassment is usually at the judge's discretion in a court.

company predicts the total overdue amount it will need to collect in the month and then sets a target collection amount for each caller based on the predicted overdue amount and the number of total callers. Callers in their first four months with the company can have a 10% less target. At the end of the month, the company calculates each caller's target completion rate as the ratio of the collected amount and the target amount. The completion salary is then a piece-wise linear, upward-sloping function of the completion rate. Callers who finish 100% of the target can earn a little more than 3500 yuan a month as the completion salary. Usually, half of the callers can complete more than 100% and, hence, earn more. The actual salary that callers receive is the sum of ranking and completion salary after adjusting for the penalties, which may take up 5-10%, as mentioned before. The company also has a minimum wage policy of around 3000 yuan a month. Appendix C Section 1 provides more information and discussions about callers' compensation schemes.

Apart from salary, well-performed callers are also rewarded with job promotion opportunities. Most new employees without any work experience in debt collection are assigned to the M1 Early stage, the easiest stage in the process.[11] They will stay in this stage for at least 4 months, during which they are regarded as "junior" workers. They can have 10% less target amounts and are ranked separately when calculating salary. Such an arrangement alleviates the work pressures of junior workers and allows them to focus more on learning debt collection skills.

Starting on the fifth month, these callers automatically become "senior" and are treated in the same way as other workers with tenures greater than four months. Senior callers who rank in the top 5% in two consecutive months may be promoted to the next stage. Cases in later stages are typically harder to collect but callers are also better rewarded. To keep the promoted callers with the company, they are allowed to "roll back" to the previous stage if they find the new task too difficult to handle or earn less than expected. Also, some excellent callers may be promoted to be group leaders who manage a group of callers in the same stage, monitoring group members' efforts and helping them with tough cases. Group leaders are usually assigned fewer cases and have extra compensation based on group performance.

1.4 AI robot caller

The debt collection process can be costly. To collect overdue repayments, the company has an internal debt collection department with more than 2000 employees and pays them a total of around

[11] New employees with former debt collection experience may start with later stages depending on their ability.

8 million yuan a month as their salary. The company also outsources part of the phone call tasks to nearly 20 third-party debt collection agencies, especially in the later stages of the debt collection process. On average, the whole debt collection team has to handle more than 320,000 cases, make 1.7 million phone calls (about 5 phone calls per borrower per day), and send 450,000 text messages every day. In addition, given a high employee turnover rate of 10-15% a month, the company needs to spend extra money on recruiting and training new employees.

To cope with the high volume of overdue cases and reduce labor costs, the company introduced robot callers for debt collection in 2018. Every morning before working hours, the company's system automatically assigns all open cases between robot and human callers. The assignment is completely randomized for a 10% subset of cases and conditionally randomized among the remaining cases based on loan characteristics. The assignment rules will be explained in detail in Section 3. After the assignment, robot callers receive all information about the assigned borrowers, including the overdue amount, days overdue, and their phone numbers. Robot callers then automatically call these borrowers and ask them to repay the debts throughout the day.

Initially, the robot callers could only play pre-recorded voice reminders or use synthetic voice messages according to standard templates, which had blanks to be filled in with the borrower's name, overdue amount, etc. Interactions between borrowers and robot callers were impossible. Robot callers were then gradually improved with the development of artificial intelligence and machine learning. During our sample period between 2021 and 2023, the AI-driven robot caller can recognize borrowers' responses with automatic speech recognition (ASR) technology and understand borrowers' questions and requests using natural language understanding (NLU) technology. The AI caller then generates appropriate answers accordingly in text and converts the texts into synthetic voices to reply to borrowers. During the sample period, the AI caller was still upgraded frequently. The improvements were concentrated on the ASR and NLU algorithms, which increased the accuracy of speech recognition and helped the AI caller understand the conversation better. Section 5.1 provides more details about the upgrades and discusses their impacts on the performance of both AI and human callers.

The AI caller is capable of providing basic information about the overdue loans, addressing potential negative impacts of delinquency, and responding to simple questions and explanations about delinquency. Table 1 illustrates the conversation process and some sample scripts that the AI caller typically uses. The conversation is divided into four stages by design. In the first stage, the

AI caller greets the borrower and confirms his/her name. If the AI caller dials the wrong number, it apologizes and hangs up the phone. Otherwise, the AI caller will continue to Stage 2 to inform the borrower about the overdue debt. The information provided by the AI caller includes the principal amount, overdue amount, bill date, days past due, and the new due date or time. The borrower is usually asked to repay within 2 hours or by the end of the day. The AI callers also emphasize potential negative consequences if the borrower fails to repay. The consequences can be worsening credit records, large amounts of late fees, difficulties in future borrowing and consumption, and even lawsuits from the company. The company may also mention the possibility of informing borrowers' guarantors, who are typically their parents and colleagues, giving them social pressure.

The AI caller then waits for the borrower's responses and sees if they have any further questions. Developers classify possible responses into five broad categories. In Case A, the borrower agrees to repay before the due time or ask for an extension. The AI caller will then confirm the new due time with the borrower, tell them that their promise has been recorded, and ask them to keep their word. In Case B, the borrower is unable to repay the debt and may explain their difficulties, such as they do not have enough money at hand, or they are too busy to deal with the debt. The AI caller can understand these explanations and reply accordingly. For example, for liquidity problems, the AI caller may ask the borrower to temporarily borrow some money from their family members or friends. In contrast, to busy borrowers, the AI caller may say that it understands that they are busy but will also address the negative consequences of default. In Cases C and D, the borrower claims that they do not have any debt with the company, or they have already repaid or have set up auto-payment. The AI caller will then ask the borrower to recall their borrowing history and to double-check their accounts or auto-payment settings. In addition, the AI caller can answer inquiries about basic information about the debts, such as the late fees (Case E).

Finally, when the borrower has no more questions about their loans, the AI caller will conclude the conversation by reiterating the negative impacts of delinquency and asking the borrower to contact customer service for further information. Similar closing words will also be used to end the conversation when the AI cannot recognize the borrower's responses (due to long silence, loud noises, strong accents, etc.) or when the borrower's responses cannot be classified into the five pre-specified cases (for example, the borrower yells at the caller or complaints about the annoying phone calls).

2 Data Description

Our data provides us with comprehensive information about the debt collection process in the company between April 2021 and December 2023. To ensure that we can track each delinquent debt and its repayment records for at least one year, we restrict cases in our working sample to those entering the debt collection process before December 2022, which gives us more than 22 million cases in total. Consistent with the company's debt collection practice, multiple bills of an individual borrower are merged into one entry during collection.

We first have loan and borrower characteristics about all delinquent debts, including loan size, borrower internal credit score, age, gender, and education level. The company uses two different variables to measure loan sizes, namely, the overdue amount and the remaining principle. For each borrower, the overdue amount is the cumulative amount of monthly repayment that the borrower fails to repay, while the remaining principal is the principal amount that the borrower has not repaid. For bills from the credit card-like products, since there is no installment payment arrangement, their overdue amounts are the same as the remaining principal, both of which are equal to the balance of the credit card-like accounts. The two measures can be different for bills from consumer loans, whose monthly repayment is not the same as the total principal needed to pay.

The internal credit score is calculated by the company to measure borrowers' default risks. It is the estimated probability of default from a logit regression. Explanatory variables include borrowers' credit history and credit line utilization in the company and other institutions[12], borrowers' app usage behaviors, and borrower characteristics such as the city of residence and education level. The company then divides all delinquent borrowers into 10 deciles and assigns them an integer score from 1 to 10, where 1 means the highest decile of default probability, i.e., the lowest credit score. This internal credit score is updated daily, incorporating the phone call outcomes of the previous day and the daily loan sizes. Education levels are self-reported, and the company can verify some of them if borrowers have uploaded their degree certificates and transcripts when registering their accounts. For our analysis, we summarize education level with an indicator of having a bachelor's degree or above.

[12] Key factors include outstanding loans and credit lines in the company and other lenders, the number of loan applications in non-banking financial institutions in the past three months, the number of credit record queries by lenders, the number of bills paid since the last loan origination, and loan types.

For each debt to be collected, we have daily records of debt collection status and repayment actions. We know the number of days overdue, the stage it belongs to, whether it is in-house or with third-party debt collection agencies, the name of the caller handling it, whether the borrower has promised to repay, and how much the borrower repays each day. We also have information about callers' efforts to contact borrowers, including the number of phone calls they make, the number of phone numbers they contact, the number of phone calls answered by the borrower, the total duration of the calls, and the total number of text messages they send for each debt on each day.[13]

Finally, we receive data about caller demographic information, monthly performance, and compensation. Callers' demographic information includes their age, gender, city of birth, whether they are in-house or with a third-party debt collection agency, and for in-house callers, their titles, and their tenures (in months) with the company. Their performance measures contain the total amount of money collected, monthly target collection amount, performance ranking, and target completion rate. Regarding caller salary, we know the final amount of salary callers received, as well as the decomposition into ranking salary, completion salary, and any additional adjustments, like the penalties.

Table 2 Panel A reports summary statistics about loan and borrower characteristics of all cases in the full sample. The characteristics are measured on day 2 past due, the first day when cases enter the debt collection process. The average delinquent debt has an overdue payment amount of about 1128 yuan (160 USD) and a remaining principal of about 6474 yuan (910 USD). They are on average larger than the population of outstanding loans as larger loans are more likely to default. Because of the heavy right tails, the median overdue amount is 653 yuan (92 USD), and the median remaining principal is 4248 yuan (600 USD). The average internal credit score is around 5, consistent with its definition. Among the delinquent borrowers, 70% are males, 13% have a bachelor's degree or more, and they are on average 27 years old. These numbers are consistent with the company's stated target customers, who are young and have consumption potential.

Also note that the sizes of cases, measured by either the overdue payment amount or the remaining principal, are heavily right-skewed: the maximum remaining principal is 1 million yuan (about 140,000 USD). These extremely large debts are typically nonstandard contracts with

[13] Since a borrower may have multiple phone numbers, and they also provide guarantors' contact information, a caller may contact more than one phone number a day for each case.

specific customers for special purposes. They are treated separately, and we may want to exclude them from our analysis. Since we do not have a clear label for these extreme cases, we use an ad hoc filter that excludes cases with remaining principal above the 99th percentile. Section 4 provides randomization tests with respect to loan sizes, which are balanced across treatment arms after the filter. The left tail does not require trimming since extremely small cases are automatically excluded in our experimental design as discussed in the next section.

3 Experimental Setup

To identify the productivity difference between AI and human callers, we utilize the company's rules for assigning cases between AI and human callers. Recall that the company considers the first day past due as a grace period, our analyses focus on day 2 past due and later.

Figure 1 shows the decision procedure of case assignment between AI and human callers on day 2 past due. First-time delinquent borrowers are all assigned to human callers so that the company can have efficient communication with these borrowers to avoid repeated delinquency. Starting on the second delinquency, borrowers are assigned between AI and human callers.

First, the company allocates all small cases with overdue amounts no greater than 20 yuan *or* remaining principal no greater than 300 yuan ("small cases") always to AI callers during delinquency. Only in some rare situations, which we will discuss later in this section, less than 5% of these small cases may be assigned to human callers after day 25 past due. The company does not use human callers on small cases since it is not cost-efficient to spend human labor handling them – callers may need to spend a similar amount of time on them but collect less money compared to larger cases.

For cases with overdue amounts greater than 20 yuan and remaining principal greater than 300 yuan, cases are either completely or conditionally randomized between AI and human callers in at most the first five days past due. On one hand, the company randomly selects 10% of the cases every month for testing and monitoring purposes, which will be referred to as the "completely randomized subsample" later in the text. In this subsample, a random half of the cases are assigned to human callers on day 2 past due, while the other half are assigned to AI callers on day 2 to day 5 before reallocating to human callers on day 6 past due onwards. Assignments to human callers are permanent: once cases are assigned to human callers, they typically will not be assigned back to AI callers. The company uses this completely randomized experiment (or the "A/B test" in the practitioner's language) to keep track of AI and human callers' performance.

The remaining 90% of cases whose remaining principals are larger than 300 yuan are assigned between AI and human callers randomly conditional on case characteristics. Hence, we call them the "conditionally randomized subsample." The variables on which the assignment relies include the overdue amount, internal credit score, and borrower's maximum days of overdue in the past. Broadly speaking, the company splits all cases in this subsample into several "cells" according to cut-offs in the three conditioning variables. The cut-offs are constant over time. Within each cell, the company then randomly assigns part of the cases to AI callers on day 2 past due. The fraction of AI cases can vary across cells. The general rule is that cells with larger overdue amounts, lower internal credit scores, and better credit histories are more likely to be assigned to human callers. In other words, company managers tend to allocate harder cases to human callers, consistent with their prior that AI callers underperform human callers. The fraction of cases assigned to AI callers within a cell also varies over time. The fractions are typically adjusted at a weekly frequency or even faster sometimes to accommodate variations in the total number of cases and human callers. In general, more cases are assigned to AI callers when there are more cases and fewer human callers, to maintain a relatively stable average workload among human callers. Appendix A Section 1 provides more information about the conditional randomization rules.

At an aggregate level, Figure 2 shows the daily number of cases on day 2 past due, along with the daily fractions of day-2 cases assigned to AI callers within a month, June 2022, for instance. The solid blue line with circles shows the total number of cases. The daily variation in total cases is mainly determined by the distribution of borrowers' repayment due dates. Specifically, most borrowers have a due date in the middle of a month, so the number of total cases is the highest around the 11th of the month. There are no new cases in the first two days of a month because the company usually uses these days for accounting, auditing, and clearing with the banks. The company intentionally avoids setting due dates in the first two days.

The red solid line with triangles is the number of cases assigned to AI callers every day and the dashed line reports the fraction. In the first week of the month, less than 30% of cases are assigned to AI given the small number of total cases. In the second and third weeks, as the workload increases, the company assigns 80% of cases to AI callers on day 2 past due. The variation reflects mainly ex-post imbalance. Between the 21st and 25th of the month, the fraction of AI cases reduces to 50%. In the last five days of the month, the company usually assigns all cases to human callers given a smaller workload and to meet month-end performance evaluation.

The company's assignment rules give us three opportunities to identify the performance gap between AI and human callers. First, the discontinuity in the company's assignment strategy among small cases is suitable for the regression discontinuity (RD) design. Specifically, since small cases identified by the 20-yuan overdue amount and 300-yuan remaining principal cut-offs are always assigned to AI callers while larger cases are assigned to AI callers at most the first five days, by comparing debt collection outcomes of cases near the cut-offs, we are able to identify the local treatment effect of (almost) completely replacing human callers with AI callers.[14]

To implement the RD design, we first notice that the 20-yuan overdue payment threshold is low and at the far-left tail of the distribution, which is about the 1st percentile, as shown in Table 2 Panel A. The number of cases on the left of this threshold is small and the corresponding local treatment effect has little implication on the remaining part of the sample. In contrast, the 300-yuan remaining principal threshold is more substantive. Therefore, in the RD design, we exclude cases having less than 20-yuan overdue payment amounts and apply the standard RD design with one running variable – the remaining principal.

Figure 3 shows the fraction of AI cases around the 300-yuan threshold. Panel (a) is a binned scatter plot of the average fraction of AI cases with respect to the remaining principal on day 2 past due. The bin width is 5 yuan, and the lines are fitted by global quadratic regressions. Consistent with the stated assignment rules, cases below 300-yuan remaining principal are all assigned to AI callers, while only about 80% of cases above the cut-off are assigned to AI. The discontinuity in the AI fraction is sharp.

Figure 3 Panel (b) shows the fractions of cases assigned to AI callers on both sides of the threshold from day 1 to day 25 past due. The fractions of "Under 300" are calculated based on cases in the (295, 300] yuan interval and are shown in the solid blue line with dots. The fractions of "Above 300" are calculated based on cases in the (300, 305] yuan interval and are shown in the dashed red line with triangle markers. As shown in the figure, small cases are all handled by AI callers in the first 25 days. In contrast, on day 2 and day 3 past due, only 80% of the larger cases are assigned to AI callers. The fraction reduces to around 60% on day 4 and day 5. From day 6 onwards, all larger cases are handled by human callers. Panel (c) extends the horizon to day 360 past due. Cases above 300-yuan remaining principal remain in human treatment for the extended

[14] The replacement is "almost" complete since larger cases also receive some AI treatment in day 2 to day 5 past due. For long evaluation horizons of up to one year, such an AI treatment is negligible.

period. For small cases below 300 yuan, a small fraction of them may be assigned to human callers after day 25. These cases are assigned to human callers mainly due to the introduction of third-party debt collection agencies. Specifically, when the company cooperates with a third-party debt collection agency, it randomly selects some cases, maybe conditioning on some loan characteristics, and assigns them to the agency. The conditioning rule is generally based on overdue amounts, and we believe it does not introduce imbalance toward either side of the cut-off. It is not a major concern since the fraction is small and any bias is against finding significant differences between AI and human callers. Section 4.2 provides formal validation tests about the RD design.

The second design for identification is the randomized experiment on the 10% completely randomized subsample. This experiment produces balanced treatment and control groups. The treatment group consists of cases that are first assigned to AI callers in the first five days before being handled by human callers on day 6 past due. The control group includes cases always treated by human callers beginning from day 2 past due. Section 4.3 reports formal validation tests about the experimental setting. Given its experimental nature, we also use this subsample for further analysis, such as examining the potential sources of the performance gap and evaluating the impacts of AI upgrades on human callers.

Third, the conditionally randomized subsample also gives us a chance to identify the productivity gap between AI and human callers. Here, since the randomization is implemented within each cell, the comparison should be conducted at the cell level. In practice, we control for cell fixed effects when estimating the productivity gap. Since the results about the conditionally randomized subsample turn out to be similar to what we find in the completely randomized subsample, they should not change the implications of our paper. We, therefore, report them in Appendix A Section 2 for interested readers.

To conclude, the above randomizing process is based on a random number assigned to each borrower when they register with the company. The random number remains the same over time so only the assignment in the second delinquency can be viewed as truly random, and assignments in subsequent delinquencies can be correlated with the second treatment given the same random number. Therefore, for cases with remaining principals larger than 300 yuan, our analyses focus only on borrowers in their second delinquency, which take up about 11% in the full sample.

4 AI versus Human Caller Performance

4.1 Measure of debt collection productivity: Net present value of collected cash flows

We use the net present value (NPV) of collected cash flows as a measure of caller productivity. Since the company views the first day past due as a grace period and uses mostly text messages or cellphone App notifications as reminders, our performance measure starts on day 2 past due and focuses on the cases that remain open on day 2. For each case starting on day 2 past due, we calculate how much money is collected on each of the following days, including any late fees collected, until it is paid back fully. We then discount these cash flows to day 2 and sum them up, which gives us the net present value of cash flows collected by callers. The discount rate we use is 24% per annum, which is close to the average APR of the loans originated by the company. It is also the maximum legal APR set by Chinese regulators in order to protect borrowers. It can be viewed as the opportunity costs of uncollected money, which could have been lent to other borrowers and generated interest at 24% APR if it were collected on time. Finally, the original NPV is scaled by the initial overdue amount on day 2, so the number means that for every one-dollar overdue balance, how much money that callers can collect after discounting. Formally, we use the following equation,

$$NPV(\tau, s) = \frac{1}{InitialBalance} \times \left(\sum_{t=2}^{\tau-1} \frac{MoneyCollected_t}{\left(1 + \frac{0.24}{365}\right)^{t-2}} + \sum_{t=\tau}^{s} \frac{MoneyCollected_t}{\left(1 + \frac{0.24}{365}\right)^{t-2}} \right), \quad (1)$$

where subscript t stands for the day past due, τ is the day when the case is first assigned to human callers. Recall that once the case is assigned to human callers, it will be rarely assigned back to AI callers. To address the timing of assigning to human callers, we separate the summation at day τ in equation (1). s is the evaluation horizon in days. In this paper, we consider caller performance up to 360 days, i.e., nearly one year after the due.

In addition, since borrowers may miss several subsequent monthly repayments after the first delinquency, the overdue amount is cumulative over days past due. As borrowers gradually repay their debts, the cumulative money collected (after discounting) may exceed the day-2 overdue balance, which is only the size of the *first* missed repayment, and, therefore, the NPV of collected cash flows is above 1. In such situations, we assume that the collected money first goes to the initial missing bills. As the NPV grows beyond 1, we assume that borrowers have fully repaid the initial debts and cap the NPV at 1 afterward.

NPV can be calculated for all cases at horizons $s = 2$ to $s = 360$. AI or human caller productivity in a given period is then defined as the arithmetic average of NPVs of all cases assigned to AI or human callers during the period.

NPV takes both the size of cash flows and the timing of cash flows into account. It values early collection by discounting late cash flows. It also properly adjusts for the late fees accumulating over time. Late fees are compensations for the company's opportunity costs on the uncollected money and should be added to debt collection profits after discounting, instead of at face value, which overstates the profits. Therefore, NPV is a better measure of caller performance than cumulative repayment rate, which is commonly used by the industry and existing literature.

4.2 Small cases subsample: Regression discontinuity design

In this section, we compare the productivity of AI and human callers by utilizing the discontinuity in the company's AI deployment strategy at the 300-yuan cutoff in the remaining principal. As discussed in Section 3, cases whose remaining principal are no greater than 300 yuan are permanently assigned to AI callers (with only a few exceptions), while cases with remaining principal greater than 300 yuan are assigned to human callers no later than day 6 past due. Therefore, by examining the average collected NPVs of cases closely around the 300-yuan cutoff with the regression discontinuity (RD) design, we are able to know how well the AI caller can do if it *completely* replaces human callers in debt collection after day 6, at least locally.

Table 2 Panel B reports summary statistics of loan and borrower characteristics in our subsample for the RD design. The variables include the overdue amount, internal credit score, fraction of males, age, and fraction of borrowers with a bachelor's or higher degree. Since the treatment effect is identified locally, we restrict the sample to cases with remaining principal between 100 and 500 yuan, which still gives us over 1 million cases in total and an effective sample size of about 90,000 near the cut-off for RD estimation. Although loan sizes are much smaller than the full sample as expected, the fraction of males, average age, and the fraction of borrowers with a bachelor's degree or more are all close to the full sample.

We first validate our RD design by showing that there is no discontinuity in loan characteristics at the cutoff. Figure 4 Panel (a) shows the binned scatter plots of loan characteristic variables around the 300-yuan remaining principal threshold. For clarity, the number of bins is set to 40 (i.e., each bin is 5-yuan width) on either side of the cutoff, which is close to the Integrated Mean Squared Error (IMSE)-optimal number of bins of around 44. IMSE-optimal number of bins minimizes the IMSE of local mean estimators. It is useful to assess the overall shape of the regression function. In contrast, in Appendix B Section 1, we show RD plots of the same variables with the variance-mimicking number of bins of around 5500. They can capture the local variability of the data. The dots show the bin averages, and the solid curves are global quadratic regression fits within each side.

Figure 4 Panel (a) shows that these observable loan and borrower characteristics are continuous at the cutoff. The overdue amount is around 150 yuan at the cutoff, which is one-half of the remaining principal. The internal credit score is slightly below 5 out of 10 on average, close to the median. The average fraction of males, age, and fraction of bachelor's or higher degrees are around 70%, 27, and 11%, respectively, around the threshold. They are similar to the corresponding averages of the full sample as reported in Table 2.

In Appendix B Section 2, we further check if there is intended manipulation around the cutoff. We do so first by examining the density functions of observation distributions on two sides of the cutoff. We also conduct the Binomial randomization test as suggested by Cattaneo et al. (2019). The results suggest that, although the remaining principal has a tendency of rounding to 300 yuan (and any multiples of 100), the density functions can still be viewed as continuous at the threshold, and the observations can be regarded as randomly allocating on two sides of the threshold. These tests confirm that cases in the neighborhood of the cutoff are similar in terms of their characteristics, and the only difference is the type of callers, as shown in Figure 3.

Given the validity of our RD design, we examine the average collected NPV difference on two sides of the cutoff, which gives the "treatment effect" of AI callers on debt collection productivity. Figure 4 Panel (b) presents RD plots of collected NPVs at horizons 2, 5, 10, 30, 90, and 360 days. Recall that before day 6, some cases above 300 yuan are also allocated to AI callers, so the jump gives a lower bound of the productivity gap before horizon day 6. The jumps in NPV at the 300-yuan remaining principal cutoff are salient. As the evaluation horizon is extended, the collected NPVs on both sides increase, as well as their differences. In addition, collected NPVs show downward-sloping trends on both sides – collectability is decreasing in loan sizes.

Table 3 formalizes our observations in Figure 4. It reports the differences between AI and human callers in variables of interest using robust RD estimators. Table 3 Panel A reports continuity tests on five observable loan characteristics as in Figure 4 Panel (a). As suggested in Cattaneo et al. (2019), the coverage error rate (CER)-optimal bandwidth is used in the tests because, for testing the null hypothesis of no discontinuity, we are interested in inference (the confidence interval), instead of point estimations. CER-optimal bandwidth is preferred over mean squared error (MSE)-optimal bandwidth as the former minimizes the asymptotic coverage error rate of the robust bias-corrected confidence interval (C.I.). Local linear regressions with uniform kernels are used in the estimation. The results show that the local average of loan characteristics on two sides of the cutoff (columns 2 and 3) are quite similar, and the differences are small as in column 4. *z*-statistics in column 5 are close to zero and the corresponding *p*-values in column 6 are greater than 0.1. All 95% robust RD C.I. cover zero as shown in column 7. These results again support the validity of the RD design.

Table 3 Panel B estimates the local NPV differences at various horizons. Since we are now interested in point estimations of the productivity gap, the MSE-optimal bandwidths are used in regressions. As the results illustrate, the differences between the left NPV mean (AI) and the right NPV mean (Human) are all negative and significant at a 1% level, regardless of the evaluation horizons. The magnitude of the NPV gaps ranges from 0.04 on the first day of contact to more than 0.1 on day 30 past due. These gaps are also economically significant. On day 2, given that the human mean NPV is 0.279, the difference corresponds to a 14.7% productivity loss. Even after one year (360 days), the percentage productivity loss is 7.6%.

Table 3 shows a large productivity gap between AI and human callers, which persists even after one year past due. As Section 1.4 illustrates, the AI callers used during our sample period are already capable of answering simple questions and providing basic information about the loans and the potential negative consequences of default. These findings suggest that some elements in human-to-human communication are missing among AI callers. These missing elements are not completely about information or direct costs to borrowers from frequent phone calls, as our AI callers can do similar things. They are more likely to be related to emotion, sympathy, and trust in personal communication. Our findings are consistent with a common belief that AI is unlikely to replace humans in non-routine jobs requiring social interactions.

As a robustness check, Table 3 column 8 reproduces the NPV results by including the five loan characteristics in Panel A as covariates in the RD regressions. Since loan characteristics are continuous at the cutoff, the inclusion of extra covariates has little impact on the magnitude and significance of the NPV differences. Additionally, in Appendix B we confirm that the results in

Table 3 are robust to kernel selection, bandwidth selection, and the exclusion of observations very close to the cutoff. The results also pass placebo tests using several artificial cutoffs.

We next examine how the productivity gap varies over evaluation horizons. Figure 5 presents the robust RD estimators of differences in collected NPVs between AI and human callers at the 300-yuan threshold over different horizons up to 360 days past due. The estimations use the same specification as in Table 3 Panel B. The error bars indicate the 95% robust RD C.I. Figure 5a plots the NPV differences up to 60 days past due. In the beginning, the difference, i.e., the productivity gap between AI and human callers, keeps widening and its magnitude reaches a peak of nearly 0.1 on 28 days past due. After that, the gap slowly narrows. Figure 5b extends the horizon to one year past due.[15] It shows that the gap remains at around -0.07 after one year past due, which is still large and significant. This U-shaped pattern suggests that, for cases before day 28 past due, human callers have large advantages over AI callers. After day 28, however, even human callers have little impact on those delinquent borrowers. Human and AI callers are similarly inefficient, so the performance starts to converge. Again, the large gap after one year suggests that some important factors in personal communication might be missing in AI callers.

Finally, we examine the heterogeneity of performance gaps with respect to borrowers' credit quality. Figure 6 reports the evolution of collected NPV differences over the delinquency horizon by internal credit score group. Following the company's definition, we merge the 10-scale internal credit score into three groups. Low, medium, and high groups refer to internal credit scores of 1-3, 4-7, and 8-10, and are marked with blue dots, red triangles, and green diamond markers, respectively. The NPV differences are estimated with the same specification as in Figure 5. The figure shows that the gaps of high-score borrowers are initially larger than the other two groups. However, since the average collected NPV of high-score cases is much higher than the other two groups, such large absolute gaps are actually small in relative terms. Moreover, the gap quickly stops growing on day 15 and starts to shrink fast, approaching -0.04 in the long run. In contrast, the gaps of low-score borrowers keep expanding for a longer horizon (until around 30 days) and remain large over time despite slow narrowing.

Magnitudes of the long-run productivity gaps of the three groups are monotonically decreasing in internal credit scores. The reason is that high-score borrowers probably only need a

[15] For clarity, the differences are plotted every three days before day 60, and every 10 days after day 60. This is the same in Figure 6.

reminder, so AI callers can perform relatively well by providing information and imposing pressure with frequent phone calls. On the contrary, low-score borrowers may experience more complicated situations and financial difficulties. They require more personal communication and persuasion tactics, of which AI callers are less capable. The longer gap-widening phase also suggests that human callers' advantages last longer.

4.3 10% Completely randomized subsample

By utilizing the RD design, the previous section shows that AI callers are not able to fully replace human callers, especially for low-credit score borrowers. Therefore, in practice, AI callers cooperate with human callers in the early stages of delinquency. Specifically, cases are collected by AI callers no more than the first five days past due. They are then all assigned to human callers on day 6. In this section, we examine the performance of such an "AI + Human" strategy.

To identify the performance of the "AI + Human" strategy, we utilize the 10% completely randomized subsample as discussed in Section 3. In this subsample, the company randomly selects half of the cases and assigns them to human callers on day 2 (the control group), while the remaining half are first assigned to AI callers on day 2 and reallocated to human callers on day 6 onwards (the treatment group). Table 2 Panel C summarizes the completely randomized subsample. Since small cases with remaining principals no greater than 300 yuan are excluded, the cases here are on average larger than the full sample. Other borrower characteristics are comparable to the full sample.

As a first step, we validate the experiment design. Figure 7 tests the randomization of case assignment by implementing two-sample t-tests on the overdue amount (Panel a) and remaining principal (Panel b) of the two groups. The t-tests are done monthly, and the bars show the resulting t-statistics. The two horizontal dashed lines mark ±1.64, the critical values of 10% significance level. The figure shows that the monthly t-statistics are all within the 90% critical values, suggesting that there is no significant difference between the two groups in terms of loan sizes. Also, the t-statistics distribute evenly above and below zero. They show no time trend or clustering by time.

In addition, we test the differences between the two groups by regressing loan and borrower characteristics onto an indicator of the treatment group and calendar month fixed effects. This gives us a concise way of summarizing the monthly group differences. Table 4 Panel A reports the results for overdue amount, remaining principal, internal credit score, gender, age, and education level. It suggests that the two groups are quite similar in terms of these six observable characteristics, as we expect in a randomized experiment.

We then check the productivity gap between “AI + Human” and the all-human control using the same regression, as shown in Table 4 Panel B. Columns 2 and 3 show the average collected NPVs of the treated (AI) and control (Human) groups, respectively. Column 4 reports the difference (AI minus Human), and the last column reports the *t*-statistics. For all evaluation horizons, the “AI + Human” treatment group always significantly underperforms the all-human control group. The gap is 0.089 on day 2, the first day of contact, and corresponds to a 32% relative productivity loss, given the average NPV of the control group is 0.282. Since human callers take over the cases after day 5, the long-run performance gap is smaller than the gap in the RD design—only 0.0084 after one year. On the one hand, this is only a 1% relative productivity loss. On the other hand, it is remarkable that only five days of exposure to AI callers appears to permanently impair the company’s ability to collect the balance due.

Figure 8 presents the differences in collected NPVs between AI and human callers over horizons up to 360 days past due, similar to Figure 5 and Figure 6. The differences are estimated by the same regression specification as in Table 4. Figure 8a shows the results by pooling all cases in the subsample. The gap widens between day 2 and day 5, as what we have seen in the RD design case. Since human callers take over all the cases on day 6 past due, the magnitudes of the performance gap jump towards zero on day 6 when human callers enter and narrow fast after that. After one year of human caller treatment, the performance of the AI-treated group remains worse than the all-human control group.

Figure 8b and Figure 8c plot the NPV gap over horizons by internal credit score and loan size, respectively. The trend of NPV gaps in each category is like the full subsample case: expanding in the first five days and quickly shrinking after human callers enter on day 6 past due. Regarding internal credit score groups, the low-score cases incur the least productivity loss in absolute magnitude, which is similar to what we see in the very early stage of the RD design. As we learn from the RD design, the NPV gap of low-score cases would keep growing and exceed the gaps of the other two groups if AI callers continued working on them. In practice, however, human callers intervene on day 6, stopping the damage in time. Therefore, the low-score cases also have the least performance damage over longer horizons. For loan size, larger cases generally have large performance gaps, consistent with our expectation that larger cases usually require more social skills that AI callers lack.

4.4 Potential sources of the performance gap

Previous sections document a significant performance gap between AI and human callers in terms of their collected NPVs. We next examine other direct phone call outcomes, including the durations of phone calls, and the fraction of borrowers who promise to repay and who repay their debts shortly after the phone calls, which can provide us with suggestive evidence of potential sources of the performance gap.

Table 5 reports the average outcomes of all phone calls made by AI and human callers on day 2 past due based on the 10% completely randomized subsample, which gives us a clean setting for comparison. Human callers on average make 0.85 more phone calls a day to one borrower than AI callers and, unsurprisingly, are answered 0.35 more times. But for both types of callers, the phone answering rates are almost the same at 23.6% since borrowers cannot tell whether a phone call is made by an AI or human caller until they pick it up. In addition, those answered phone calls display no significant difference in average ringing time for borrowers to answer the phone and the time of calls.[16] Specifically, both types of callers have to wait around 19 seconds for a phone call to be answered and call borrowers at 1 PM on average.[17] These comparisons suggest that AI and human callers have similar overall reachability to delinquent borrowers except that humans make more phone calls.

We next restrict our sample to the first calls answered by each borrower to further analyze the ability of AI and human callers. First, we notice that the time of calls from human callers is on average earlier than AI callers because, as mentioned in Section 1.2, there is a half-an-hour "automatic calling" period from 9 PM to 9:30 PM when all cases assigned to human callers are called once, contributing to a large fraction of first answered calls. On the other hand, phone calls from AI callers are distributed more evenly across the day. To account for such a disparity, we may estimate "timing-adjusted" results by including fixed effects for the time of calls for every one-hour interval in the analyses.

[16] We convert the time of calling to a decimal number representing hours from midnight. For example, 2:15 PM is converted to 14.25.

[17] Since the working hours are typically from 9 AM to 8 PM, the average of 1 PM reflects the fact that a relatively larger portion of calls are made in the morning.

There is a significant difference between unconditional means of the phone ringing time between the two types of callers with calls from AI callers are picked up faster. The disparity disappears after including time-of-day fixed effects, indicating that the timing of AI callers is better and the two types of callers have similar reachability after conditioning for timing. However, the duration of phone calls significantly differs. The unconditional mean duration of an AI phone call is only about 28 seconds on average, 19 seconds less than phone calls by human callers. The gap widens to 31 seconds after the timing adjustment. This finding suggests that AI callers may be able to provide limited information and not to handle complicated situations, leading to short conversations. Appendix Figure D2 shows the histograms of phone call durations from the two types of callers separately. AI phone calls are generally short and concentrated to around 30 seconds, while the duration of human calls has greater variation—a proxy for flexibility. Another potential interpretation of these differences in moments is that AI is less able to keep the attention of humans, who might hang up quickly upon realizing that an AI is calling. However, the figure shows that AI calls are significantly less likely to terminate within the first 10 seconds or the first 20 seconds than human calls, which suggests that the attention channel is less important.

In addition, about 21.2% fewer borrowers make a promise to repay their debts when talking to AI callers. The reasons may be that AI callers are less persuasive and impose less pressure on borrowers, or people are just reluctant to make promises with AI callers.

Lastly, we examine borrowers' propensity to repay their debts after answering the first call from the company. The horizons range from 15 minutes to 5 hours, as well as the end of the day after answering the calls. Within 15 minutes after the call, borrowers talking to AI callers are less likely to repay their debts, although the difference is small and insignificant. As time goes on and the number of repaying borrowers accumulates, the gap increases. Two hours after answering the call, 15.2% of borrowers have repaid their debts after talking to human callers while there are only 10.2% of AI-treated borrowers doing so. The difference represents more than one-third of performance loss, which is both statistically and economically significant. The gaps are larger for longer horizons, although they may also reflect impacts from additional phone calls made after two hours.

In summary, the evidence suggests that, besides differences in calling strategies such as the amount and the timing of phone calls, AI and human callers differ in their ability to communicate with borrowers with the latter being more capable of handling complex situations, asking for promises to repay the debts, and urging borrowers to repay earlier.

5 The Interactions between AI and Human Callers

5.1 Impacts of AI upgrades

As mentioned in Section 2.3, the AI callers experienced several upgrades during our sample period in 2021 and 2022, which give us an opportunity to examine how improvement in AI productivity affects human callers and how they react to a better AI. This question is particularly important as current AI technology is still fast developing, and our analyses about AI upgrades can provide important implications.

Figure 9 illustrates the AI upgrade process by showing the fractions of cases assigned to different versions of AI callers every month in our sample period. There are five versions of AI callers. We labeled the first version of AI in our sample period as "V1," which is not the same as the very first version of AI used by the company in 2018. Subsequent versions are labeled as "V2" to "V5" according to the order of introduction. The length of the bar indicates the fraction assigned to the corresponding version of AI callers. All assignments between different versions of AI are random.

As Figure 9 shows, the company introduces new versions of AI callers progressively. The first version "V1" had been mostly replaced by "V2" at the beginning of the sample period in April 2021. Starting in May 2021, the company gradually introduced "V3." In the first three months, "V3" was still under development and testing, so it only took 10% to 15% of the cases, which were used to evaluate the performance of the new version. As the company was satisfied with the outcomes, they slowly increased the fraction from 30% to 80% in the following months (August to October 2021). Finally, "V3" completed replaced "V2" in November 2021. This is a major upgrade that took six months to finish. The upgrades to "V4" and "V5" followed a similar procedure, but the testing phase was shorter and generally used more cases for testing.

Since the assignment is random among different versions of AI callers, the comparison of their productivity is straightforward. In the following analyses, we restrict our sample to the 10% completely randomized subsample. Figure 10 shows the monthly average collected NPVs of different versions of AI callers, along with the average NPVs of human callers, on day 2 or over the first five days past due. First notice that the gaps between AI and human callers remain wide over time, so during our two-year sample period, the upgrades of AI callers did not close the gap very much. The figure suggests little performance difference between V1 & V2, and between V4 & V5. For V2 & V3, there seems to be no difference in the early months of testing, when V3 only took a small fraction of cases. The performance increase was more noticeable in September and October 2021. The improvement from V3 to V4 seems significant on day 2 but not in the first five days past due.

Table 6 formally tests the improvement of AI performance in terms of collected NPVs and confirms our previous observations. The table reports *t*-test results about the differences between each pair of consecutive versions of AI callers over 2-day to 10-day horizons. The calculation is based on the 10% completely randomized subsample, so the AI caller only works from day 2 to day 5. For each pair of AI callers, the test is implemented by regressing collected NPV onto an indicator of the newer version of AI callers and calendar month fixed effects in months when the two versions coexist. For versions "V2" and "V3" because the transition time was six months long, only the last two months (i.e., September and October 2021)—when "V3" took up substantial fractions—are used. The results show that the most salient enhancement happened when upgrading from V2 to V3. The increase in collected NPV is around 0.025, corresponding to a 5-10% increase in AI callers' productivity given the average NPV in the last row. On day 6 past due, human callers take over the jobs, closing the gap to only an insignificant difference of 0.0099 between these two versions of AI callers immediately. From another perspective, however, it also means that human callers collect 0.0171 (=0.0270 – 0.0099) less NPV on day 6.

This finding suggests that there is an upper bound of human callers' ability to collect overdue debts. Specifically, no matter how much the AI callers collect in the first five days, human callers collect additional delinquent payments up to the limit of their ability. Therefore, although AI callers collect more money at the beginning, the outcomes on day 6 are similar, reflecting the human limit. It means that the more powerful AI callers are, the less human callers can collect. This finding suggests the displacement effects of AI callers on some tasks.

Table 7 further examines the displacement effects using the AI upgrade from V2 to V3. The sample of cases is restricted to the completely randomized subsample treated by AI callers in the first five days in September and October 2021 when AI V2 and V3 coexist as in Table 6. We then study how human callers perform after receiving these AI-treated cases. We restrict the sample of human callers to those specializing in the "M1 Early" (days 2 to 10) stage, who are the major group of callers working on day-6 cases. The variable of interest is the human caller's performance on day 6 past due, which is measured as the difference of collected NPVs between day 6 and day 5 past due, that is, ΔNPV6 = NPV6 – NPV5. Hence, we focus on the cases that remain unpaid at the end of day 5 past due. In Column 1, we first regress human callers' day-6 performance, ΔNPV6, onto an indicator that the cases are treated by AI V3 instead of V2 in the first five days. The estimated coefficient on the V3 indicator is -0.033, suggesting that callers perform worse on cases previously treated by the better version of AI callers. The magnitude of the estimation is comparable to the results in Table 6.

Column 2 adds a caller ability measure as an additional explanatory variable to see how human callers' productivity loss varies by their ability. Specifically, callers' ability is measured as their day-6 performance on cases treated by AI V2 in the previous month (denoted by *PrevPerfRank*). The performance is normalized to fractional ranking within a month, with 1 to the top caller. We focus on cases treated by AI V2, so the ability measure is unrelated to AI V3's performance. We use a one-month lag in performance measures to avoid contemporaneous confounders.[18] Both *PrevPerfRank* and its interaction with the AI V3 indicator are included in the regression. First, the coefficient of *PrevPerfRank* is significantly positive, implying that there is persistence in caller monthly performance, which may be interpreted as callers' ability. Next, the significantly negative interaction term suggests that better callers are more heavily affected by the advances of AI callers than their counterparts. This can be the case if the development of the AI caller is to learn and mimic the skills of the best callers.

Since the lagged performance rankings are only available to existing callers, in Column 3, we set the lagged performance rankings to zero for new callers and use an indicator for them. Again, we include the new caller indicator and its interaction term with the AI V3 indicator. The coefficients are both insignificant, so on average new callers do not differ from existing callers regarding the impacts of AI upgrade.

[18] In Appendix D Table D1 Columns 1-8, we regress the indicator of cases treated by AI V3, cases outcomes in day 5 (NPV5), and observable case characteristics onto callers' previous performance ranking and find all insignificant coefficients, validating the randomization of case assignments across callers. At the caller level, we regress an indicator of promotion to later stages or leaving the company in the next month onto callers' current-month performance ranking in the last two columns. Despite reasonable signs of the coefficients on performance ranking, they are both insignificant, alleviating the attrition bias concerns.

5.2 Trade-off between productivity and labor costs

Section 4 shows that there is a significant performance gap between AI and human callers. Even though in practice AI callers are used only in the first few days after delinquency, they can induce moderate productivity loss. On the other hand, AI callers can save human labor and wage costs since they have almost zero marginal costs when making phone calls, generating potential benefits for the company. Therefore, to fully evaluate how well AI may replace human callers, we need to adjust our NPV calculations to account for labor costs and estimate the net benefits of introducing AI callers.

We first estimate *direct* labor costs in the debt collection process, that is, caller salary. From the company's perspective, the total caller salary paid to all callers every month can be decomposed into two parts. One part is fixed salary, which only depends on the total number of callers and is unrelated to how much money they collect in the month. The other part is variable salary, which is a function of the total amount of overdue money that callers collect in each month. Appendix C Section 1 provides more information about the salary scheme. Although there is nonlinearity and variation in salary schemes across callers in different stages of the debt collection process, for a simple back-of-envelope calculation, we assume constant rates for fixed and variable labor costs. Following the procedure described in Appendix C Section 2, we estimate that to employ callers to handle one-yuan overdue money, the average fixed cost is about 0.00045 yuan, regardless of whether the money is collected or not. In addition, for every one-yuan money collected, the variable cost that the company needs to compensate the caller is approximately 0.0051 yuan. Hence, the formula for NPV (equation (1)) can be modified as

$$NPV^L(\tau, s) = \frac{1}{InitialBalance} \times \left(\sum_{t=2}^{\tau-1} \frac{MoneyCollected_t}{\left(1 + \frac{0.24}{365}\right)^{t-2}} + \sum_{t=\tau}^{s} \frac{MoneyCollected_t \times 0.9949 - InitialBalance \times 0.00045}{\left(1 + \frac{0.24}{365}\right)^{t-2}} \right), \tag{2}$$

where superscript L indicates labor costs adjustment and, as in equation (1), index t stands for the day past due, τ is the day when the case is first assigned to human callers, and s represents the evaluation horizon.

Figure 11 shows the average differences of net collected NPV between AI and human callers over the horizon of days past due after adjusting for caller salary, as illustrated by the solid red lines with triangles. The error bars are the 95% confidence intervals. Panel (a) and (b) use the RD design subsample and the 10% completely randomized subsample, respectively. The estimation methods are the same as what we use for the corresponding subsamples in Figure 5 and Figure 8. For comparison, the dashed blue lines with dots show the difference of unadjusted NPV, as what we showed before.

The figure shows that, after accounting for direct labor costs, the collected NPV gaps between AI and human callers become narrower. For the small cases in the RD design, since the AI callers completely replace human callers after day 6, the cost-saving is large and cumulative over time. After one year, the saved labor cost is about 0.01 NPV or 14% of the unadjusted gap. Nonetheless, AI is still not cost-effective relative to humans. In the completely randomized subsample, the long-run productivity gaps become smaller, but here too, AI is less cost-effective than humans.

Importantly, here we only consider direct labor costs, i.e., salary paid to human callers. To hire and manage more than 2000 skilled callers, the company also needs to spend money on many indirect costs, such as worker recruitment, training, management, pension funds, etc. On the other hand, we also do not include the calculation the cost of developing and improving the AI software.

6 Conclusion

In this paper, we cooperate with a leading online consumer loan provider in China to evaluate the performance and economic impacts of AI adoption in debt collection. Leveraging the company's rules to assign delinquent debts between AI and human callers, we use an RD design to show that the current AI cannot completely replace human callers in the whole debt collection process. Further analyses address AI's difficulties in handling complicated situations and asking for promises. Despite the poor performance, a randomized experiment suggests that losses from using AI can be substantially mitigated if AI and human callers work together. Nonetheless, even in this collaborative arrangement, the NPV of collected balances remains below the amount collected if only human callers are used, suggesting that using AI callers creates some modest permanent damage to the company's relationship with its borrowers. Thus, AI may underperform humans in non-routine jobs that require emotional skills and social interactions.

References

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

Acemoglu, Daron, and Pascual Restrepo. 2019. "Artificial intelligence, automation, and work." In *The Economics of Artificial Intelligence: An Agenda* (pp. 197-236). University of Chicago Press.

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and jobs: Evidence from US labor markets." *Journal of Political Economy*, 128(6): 2188-2244.

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, automation, and the rise in us wage inequality." *Econometrica*, 90(5): 1973-2016.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial intelligence: the ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives* 33(2): 31-50.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining human expertise with artificial intelligence: Experimental evidence from radiology." *NBER Working Paper* No. w31422.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. 2023. "Generative AI at work." *NBER Working Paper* No. w31161.

Brynjolfsson, Erik, and Tom Mitchell. 2017. "What can machine learning do? Workforce implications." *Science*, 358(6370): 1530-1534.

Burstyn, Leonardo, Stefano Fiorin, Daniel Gottlieb, and Martin Kanz. 2019. "Moral incentives in credit card debt repayent: Evidence from a field experiment." *Journal of Political Economy* 127(4): 1641-1683.

Cao, Sean, Wei Jiang, Junbo L. Wang, and Baozhong Yang. 2024. "From man vs. machine to man+ machine: The art and AI of stock analyses." *Journal of Financial Economics*, forthcoming.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

Drozd, Lukasz A., and Ricardo Serrano-Padial. 2017. "Modeling the revolving revolution: the debt collection channel." *American Economic Review*, 107(3): 897-930.

Ehret, Ludovic, and Qian Ye, 2024. "'Better than real men': Young Chinese women turn to AI boyfriends." *The Japan Times*, February 13. https://www.japantimes.co.jp/news/2024/02/13/asia-pacific/social-issues/chinese-women-ai-boyfriends/

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An early look at the labor market impact potential of large language models." *arXiv*: 2303.10130.

Erel, Isil, Léa H. Stern, Chenhao Tan, and Michael S. Weisbach. 2021. “Selecting directors using machine learning.” *Review of Financial Studies* 34(7): 3226-3264.

Fedaseyeu, Viktar. 2020. “Debt collection agencies and the supply of consumer credit.” *Journal of Financial Economics*, 138(1): 193-221.

Fedaseyeu, Viktar, and Robert M. Hunt. 2015. “The Economics of Debt Collection: Enforcement of Consumer Credit Contracts.” *FRB of Philadelphia Working Paper* No. 15-43.

Felten, Edward W., Manav Raj, and Robert Seamans. 2020. “The occupational impact of artificial intelligence: Labor, skills, and polarization.” *SSRN Working Paper* No.3368605.

Frank, Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. 2019. “Toward understanding the impact of artificial intelligence on labor.” *Proceedings of the National Academy of Sciences* 116(14): 6531-6539.

Gallegos, Demetria. 2024. “Is it OK to be mean to a chatbot?” *Wall Street Journal*, February 15. https://www.wsj.com/tech/ai/artificial-intelligence-chatbot-manners-65a4edf9

Gao, Zihan, and Jiepu Jiang. 2021. “Evaluating human-AI hybrid conversational systems with chatbot message suggestions.” In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 534-544.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2013. “The determinants of attitudes toward strategic default on mortgages.” *Journal of Finance*, 68(4): 1473-1515.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human decisions and machine predictions.” *Quarterly Journal of Economics*, 133(1): 237-293.

Laudenbach, Christine, and Stephan Siegel. 2023. “Personal communication in an automated world: Evidence from loan repayments.” *SSRN Working Paper* No.3153192.

Maslej, Nestor, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. 2023. *The AI Index 2023 Annual Report.* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Michaels, Guy, Ashwini Natraj, and John Van Reenen. 2014. “Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years.” *Review of Economics and Statistics*, 96(1): 60-77.

Noy, Shakked, and Whitney Zhang. 2023. “Experimental evidence on the productivity effects of generative artificial intelligence.” *Science*, 381(6654): 187-192.

Tan, Tom Fangyun, and Serguei Netessine. 2020. “At your service on the table: Impact of tabletop technology on restaurant performance.” *Management Science*, 66 (10): 4496-4515.

World Economic Forum. 2020. *The Future of Jobs Report 2020*. Switzerland.

World Economic Forum. 2023. *The Future of Jobs Report 2023*. Switzerland.

Zhang, Wanqing. 2020. “The AI girlfriend seducing China’s lonely men.” *Sixth Tone*, December 7. https://www.sixthtone.com/news/1006531

Figures

Figure 1. Case assignment between AI and human callers on day 2 past due.

This figure shows the decision procedure of case assignment between AI and human callers. For small cases with overdue payment below 20 yuan or remaining principal below 300 yuan, "Almost always AI" means that more than 95% of cases are always handled by AI callers over the life cycle of the cases, and only less than 5% of cases may be assigned to human callers after day 25. For the conditionally randomized subsample, the case characteristics used for conditioning include the overdue amount, internal credit score, and maximum days of delinquency in the past.

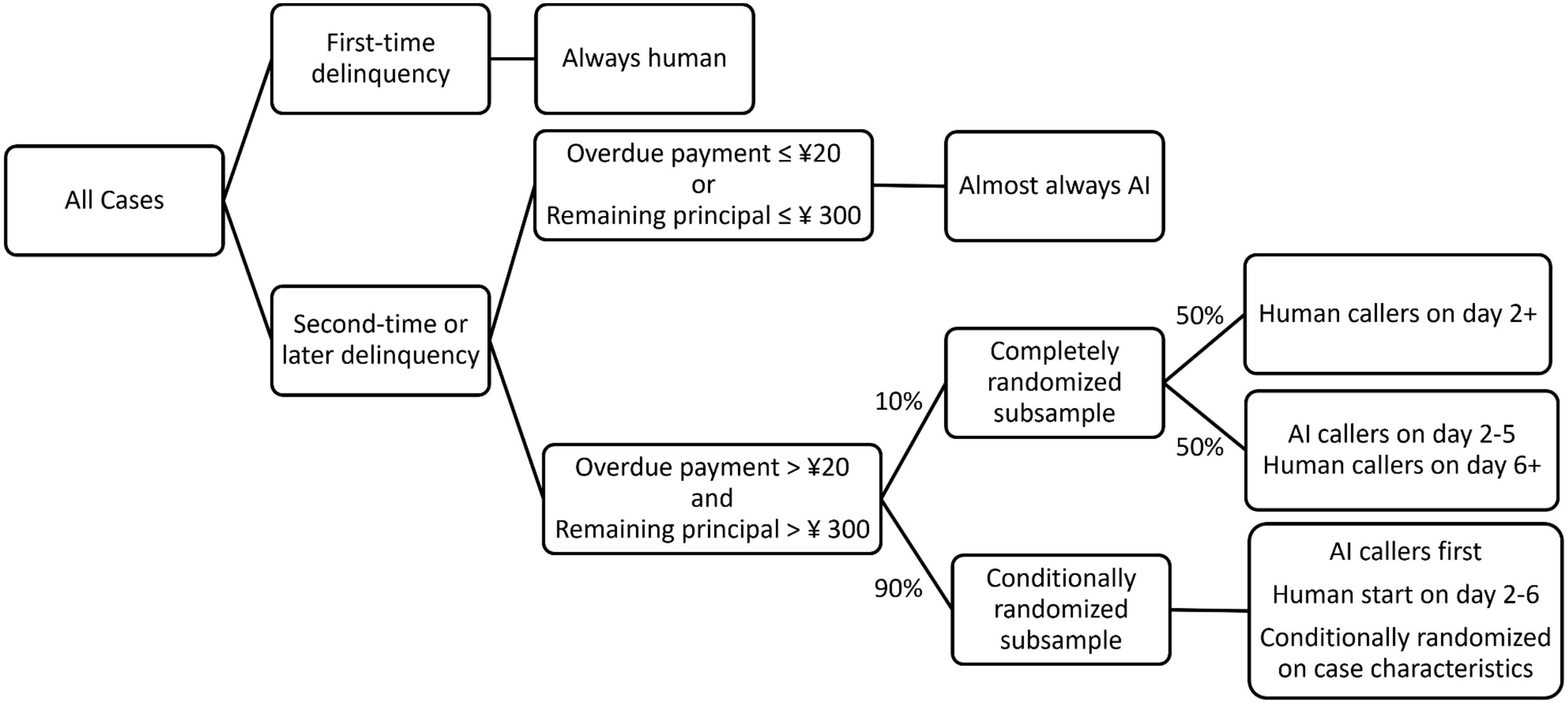

All Cases
First-time delinquency
Always human
Second-time or later delinquency
Overdue payment ≤ ¥20 or Remaining principal ≤ ¥ 300
Almost always AI
Overdue payment > ¥20 and Remaining principal > ¥ 300
10%
Completely randomized subsample
50%
Human callers on day 2+
50%
AI callers on day 2-5 Human callers on day 6+
90%
Conditionally randomized subsample
AI callers first
Human start on day 2-6
Conditionally randomized on case characteristics

Figure 2. Daily number of cases and fraction of AI cases in June 2022.

This figure shows three daily statistics of new cases entering the debt collection process on day 2 past due every day in June 2022. The "All cases" and "AI cases" lines show the total number of new cases, and the number of cases assigned to AI callers every day, respectively. The dashed "Frac AI" line reports the fraction of AI cases in total new cases – the ratio of "AI cases" to "All cases."

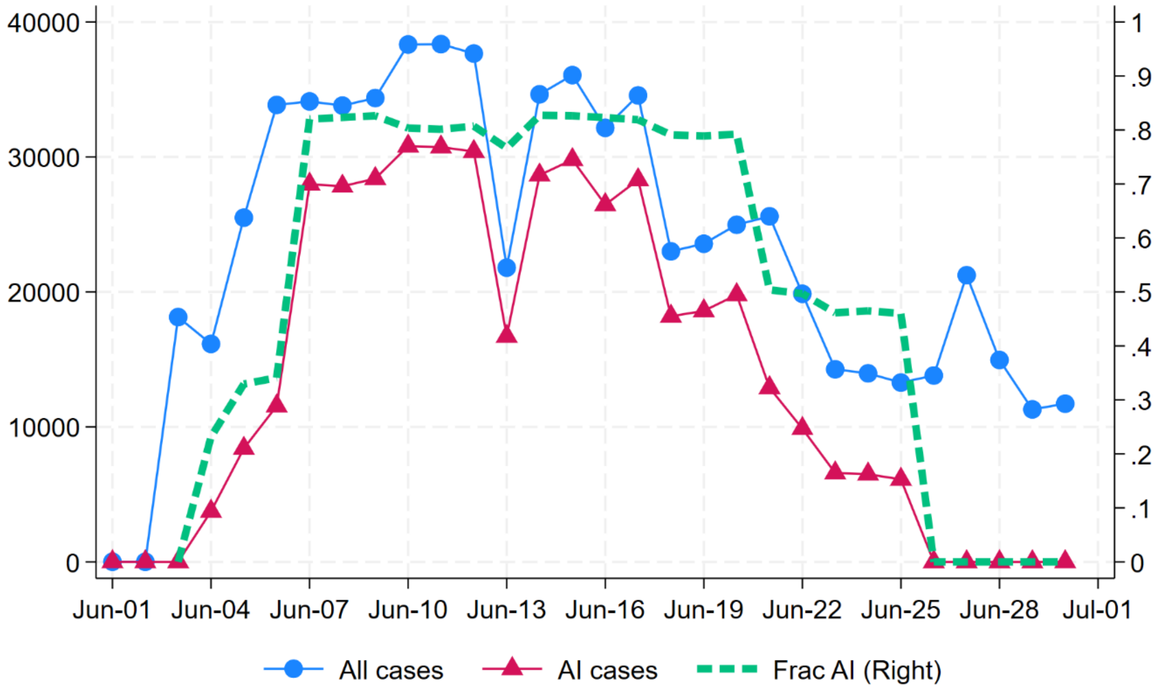
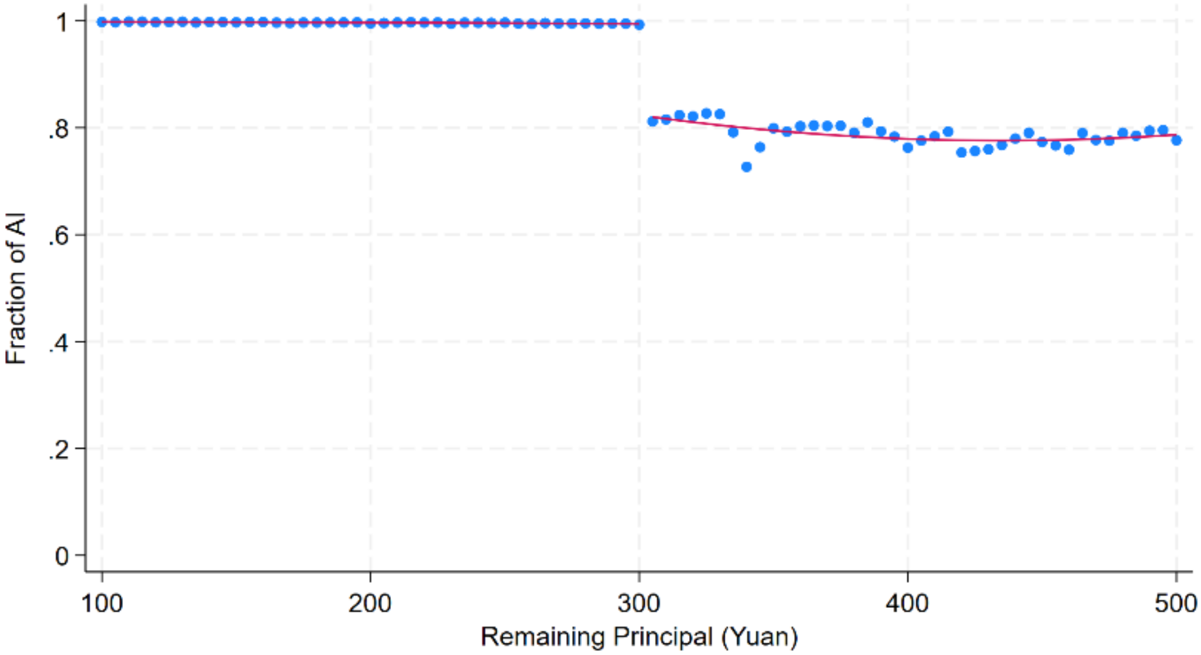


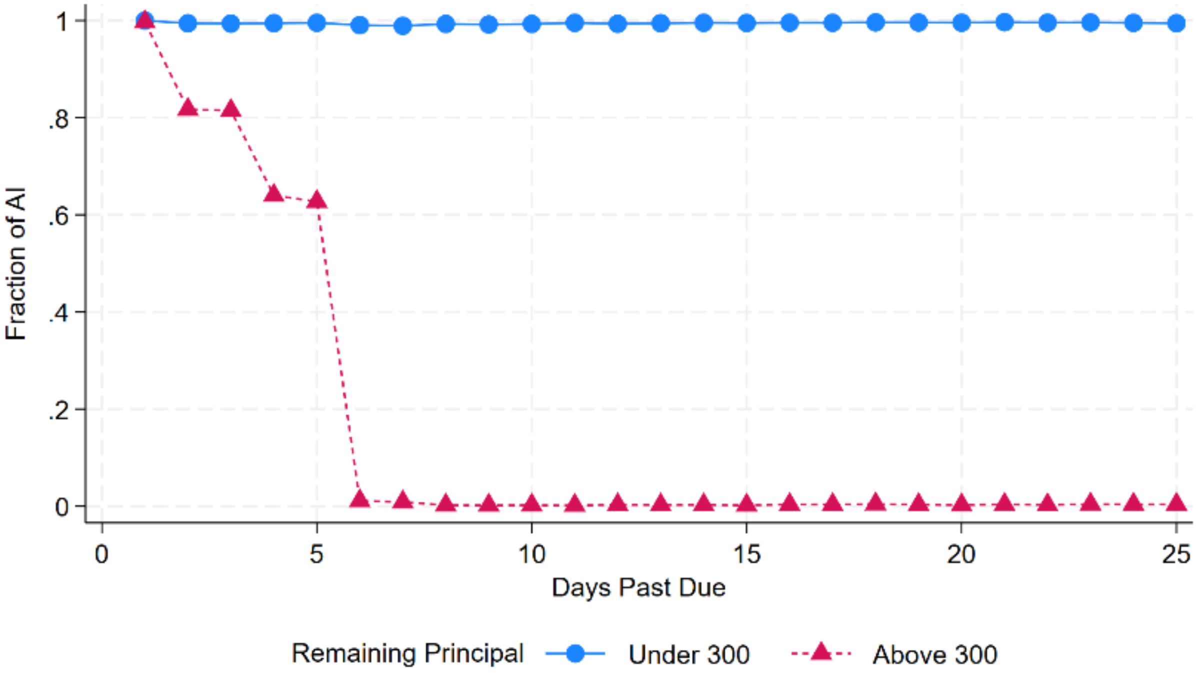
Figure 3. Fraction of cases assigned to AI callers around the remaining principal threshold.

This figure shows the fraction of cases assigned to AI callers around the 300-yuan remaining principal threshold. Panel (a) is a binned scatter plot of the average fraction of AI cases with respect to the remaining principal clustered at 5-yuan intervals on day 2 past due. Panel (b) shows the fractions on both sides of the threshold from day 1 to day 25. The fraction below the threshold is calculated from cases in the (295, 300] yuan interval. The fraction above the threshold is calculated from cases in the (300, 305] yuan interval. Panel (c) extends the horizon to day 360 past due.

(a) Binned scatter plot of the fraction of AI cases around the threshold on day 2 past due



(b) Fraction of AI cases below and above the threshold (25 days past due)



(c) Fraction of AI cases below and above the threshold (360 days past due)

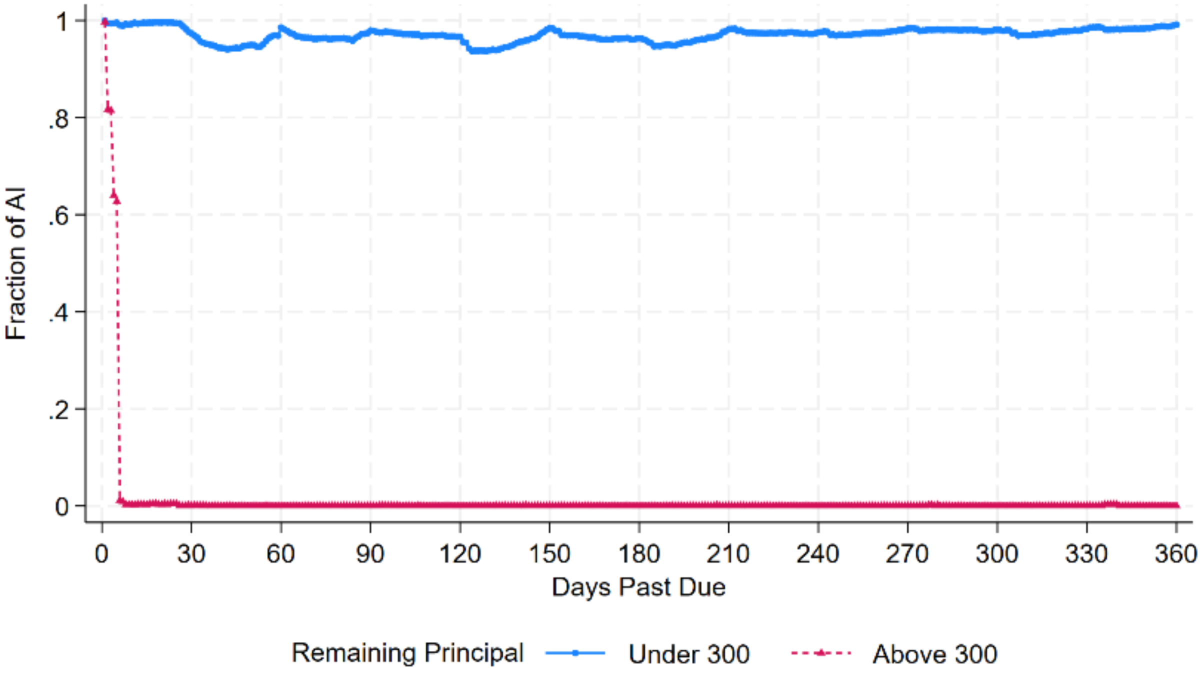
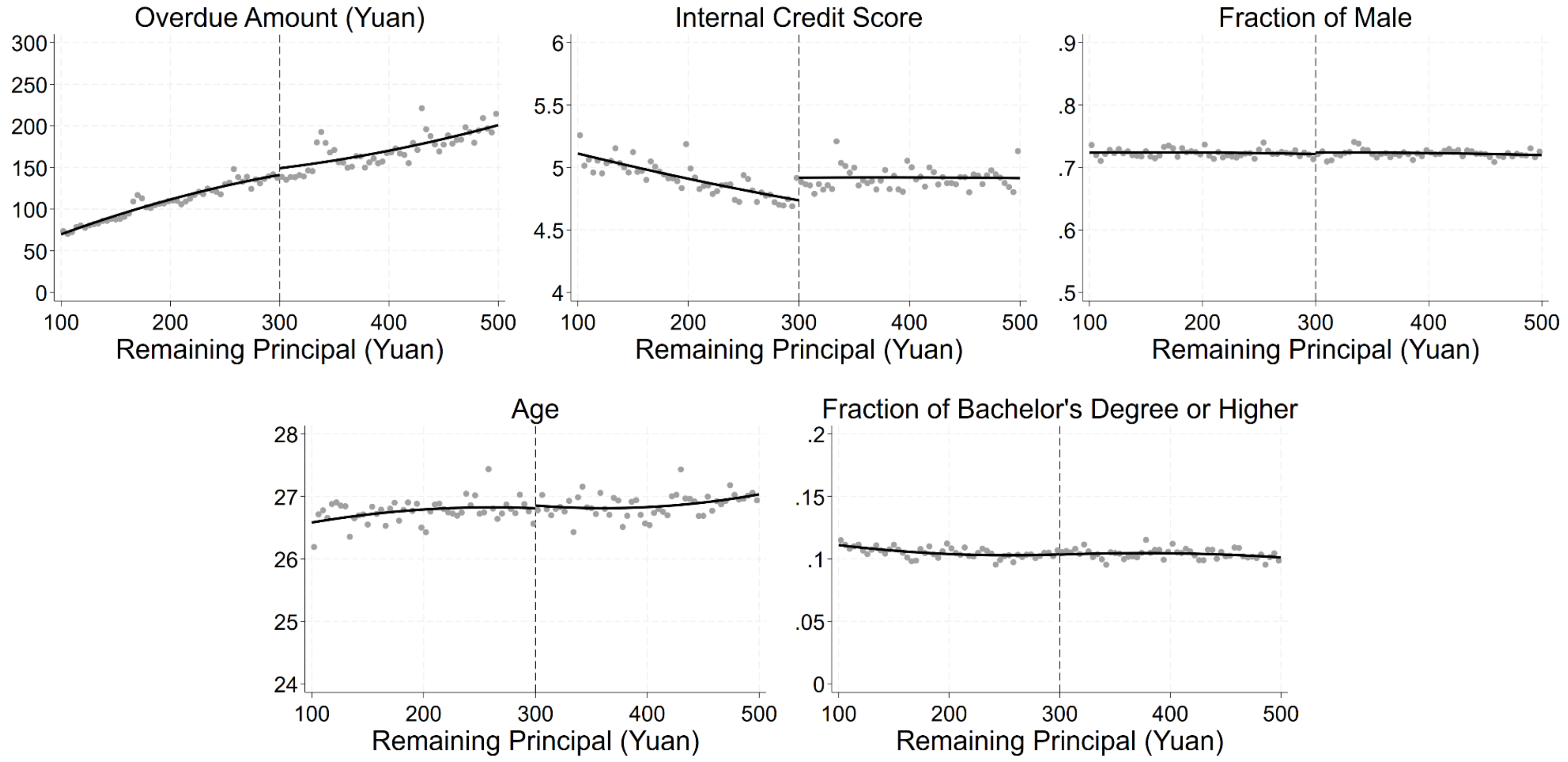


Figure 4. RD plots of loan characteristics and NPVs around the principal cutoff.

This figure shows the binned average of several variables around the principal cutoff of 300 yuan. Loans with remaining principal below 300 yuan (included) are always assigned to AI callers while those above 300 yuan are all assigned to human callers after day 5. The variables of interest include loan characteristics as shown in Panel (a), such as overdue amount, internal credit score, borrower gender, age, and education (an indicator of whether they hold a bachelor's degree or higher), as well as collected NPVs for various horizons, as in Panel (b). The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. There are 50 principal bins of equal width on each side of the threshold. The grey dots are binned averages, and the black lines are global quadratic fits within each side.

(a) Loan characteristics



(b) Collected NPV

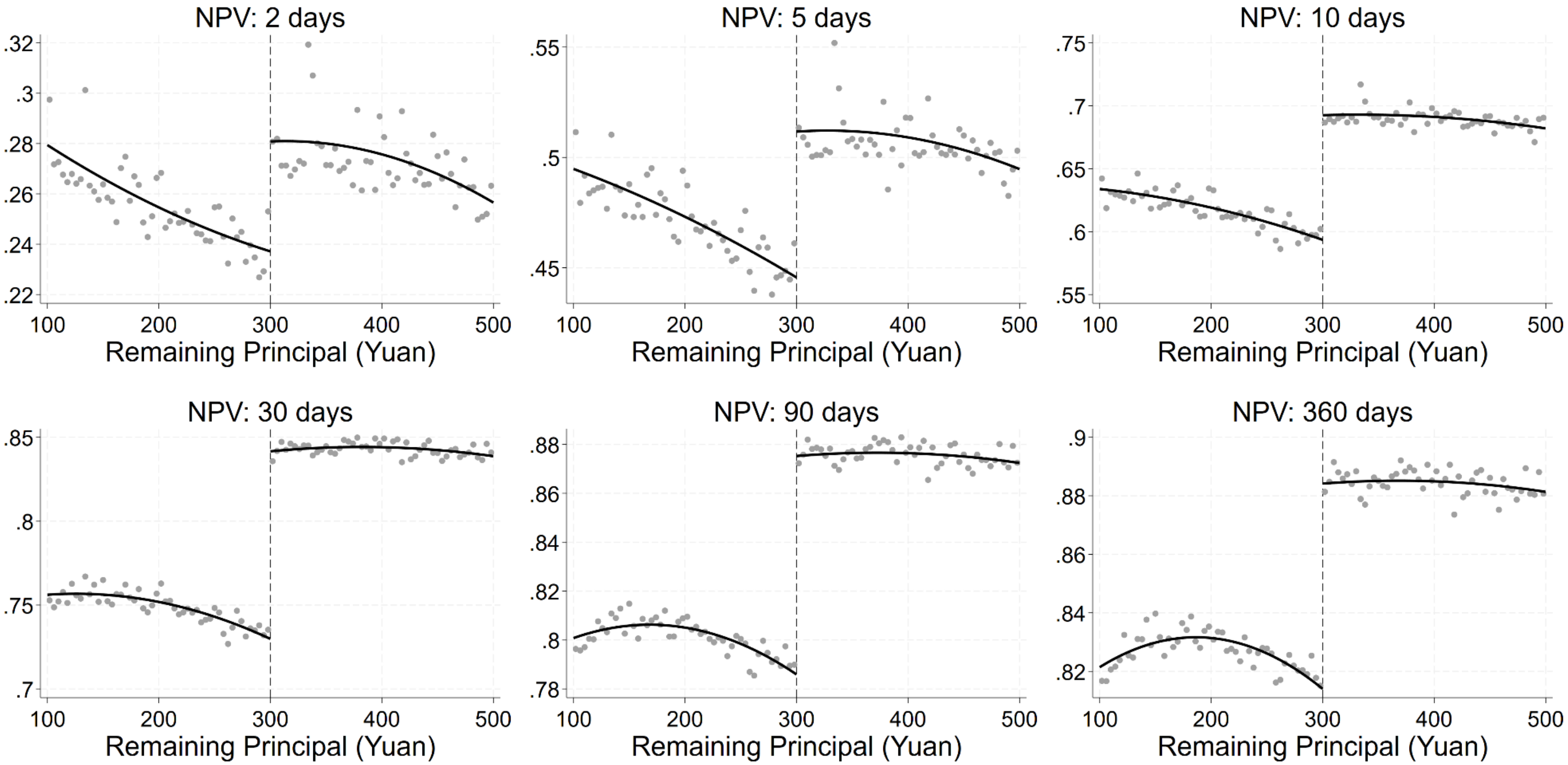
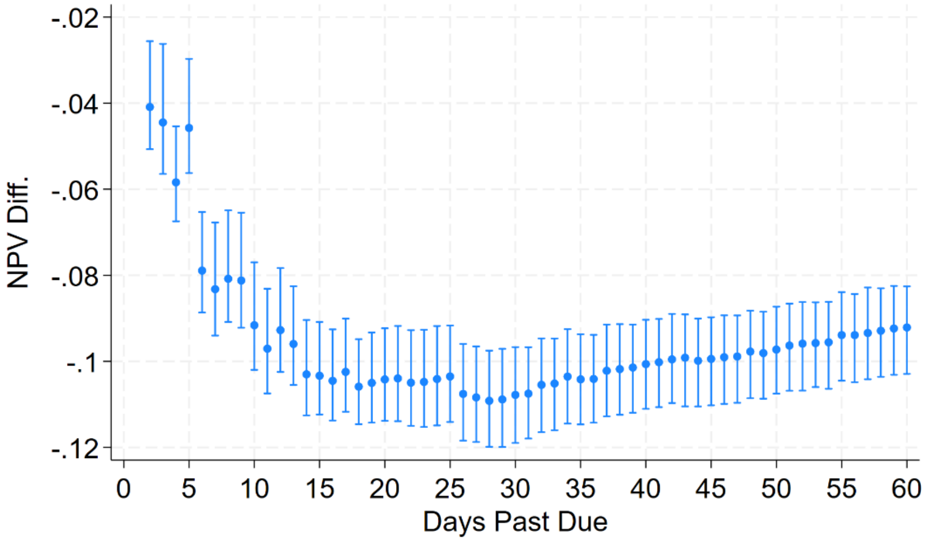


Figure 5. Collected NPV differences between AI and human callers over horizon – small cases RDD.

This figure reports the average differences of collected NPV between AI and human callers over the horizon of days past due of cases. The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. The differences are estimated by RDD utilizing the 300-yuan remaining principal threshold for almost permanent AI treatment. Panel (a) shows the time series between day 2 and day 60 past due. Panel (b) extends the horizon to 360 days past due. The dots represent the average differences estimated by RDD and the bars indicate the 95% robust regression discontinuity confidence intervals. For clarity, in Panel (b) the differences are plotted every three days before day 60, and every 10 days after day 60.

(a) First two months past due



(b) First year past due

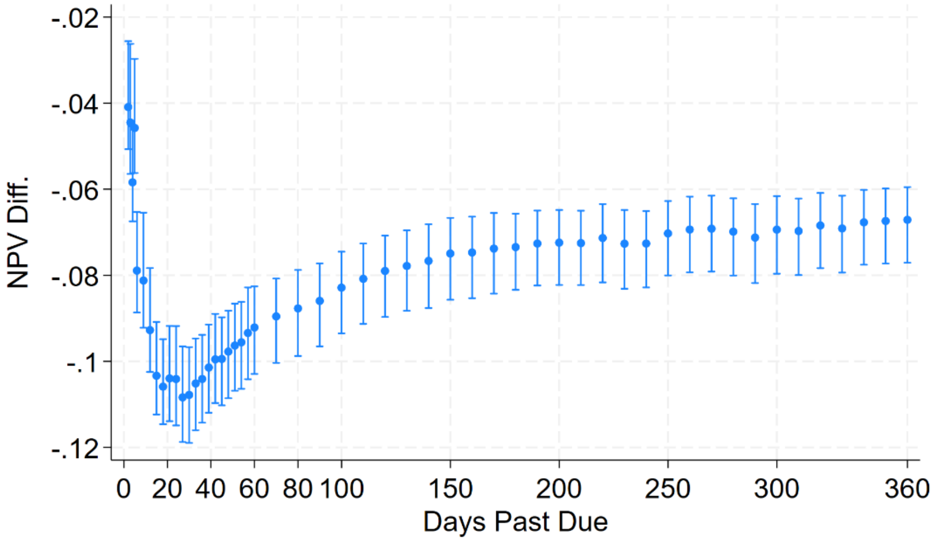


Figure 6. Collected NPV differences between AI and human callers over horizon, by internal credit score.

This figure reports the average differences of collected NPV between AI and human callers over the horizon of days past due of cases for three groups of internal credit scores separately. The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. The differences are estimated by RDD utilizing the 300-yuan remaining principal threshold for almost permanent AI treatment. "Low", "Med", and "High" refer to cases with internal credit scores lying in 1-3, 4-7, and 8-10, respectively. For illustration, the differences are plotted every three days before day 60, and every 10 days after day 60.

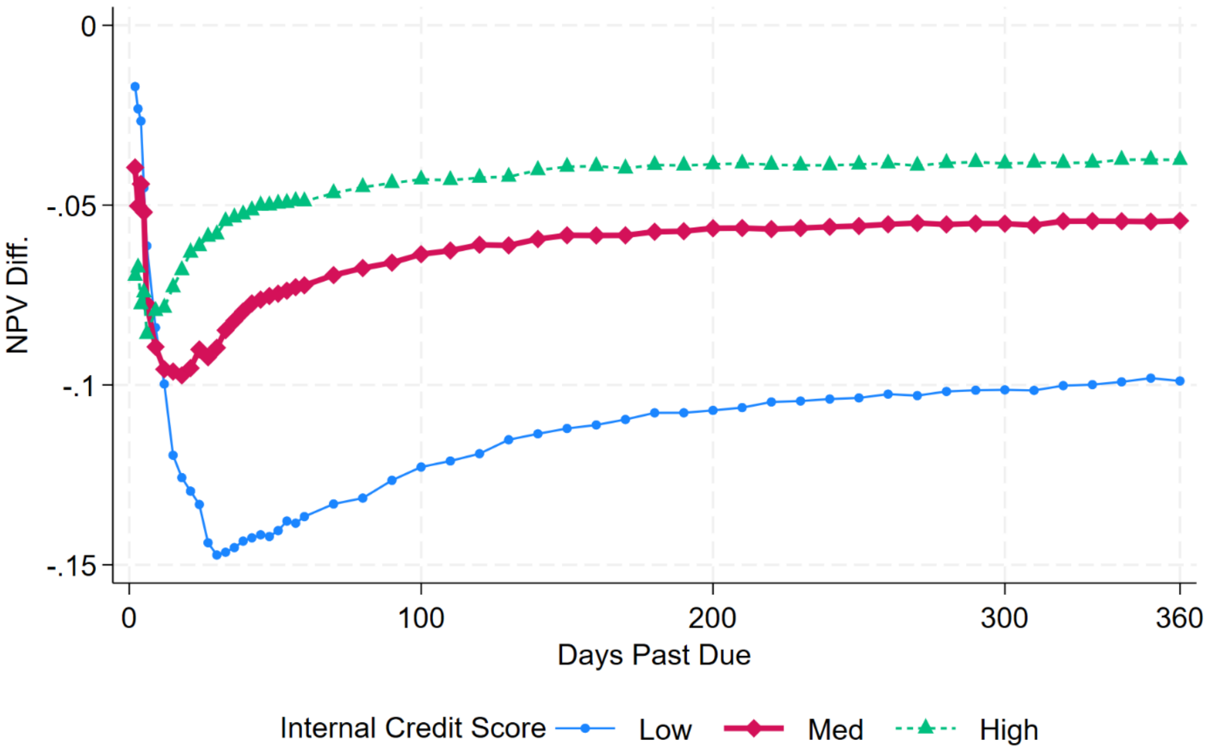
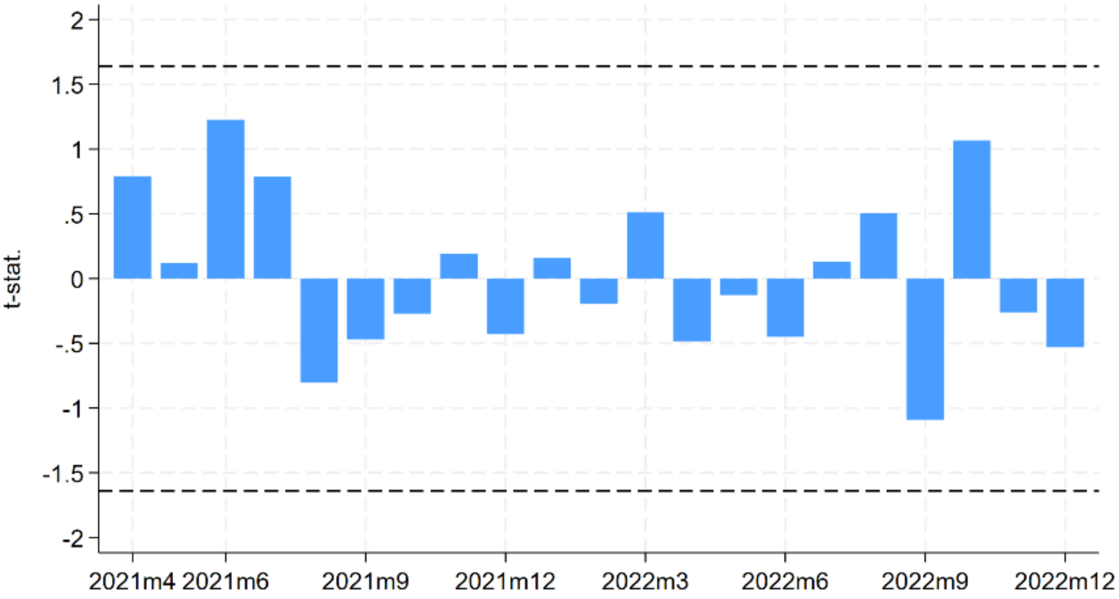


Figure 7. Test of randomization – completely randomized subsample.

This figure reports the results of the monthly randomization test on the 10% completely randomized subsample. Every month t-tests between AI and human callers on loan characteristics are implemented. The bars show the *t*-statistics of the difference in loan characteristics between AI and human callers. The variables of interest include overdue payment in Panel (a) and remaining principal in Panel (b). The horizontal dashed lines indicate ±1.64, the critical values of 10% significance level.

(a) Overdue amount.



(b) Remaining principal

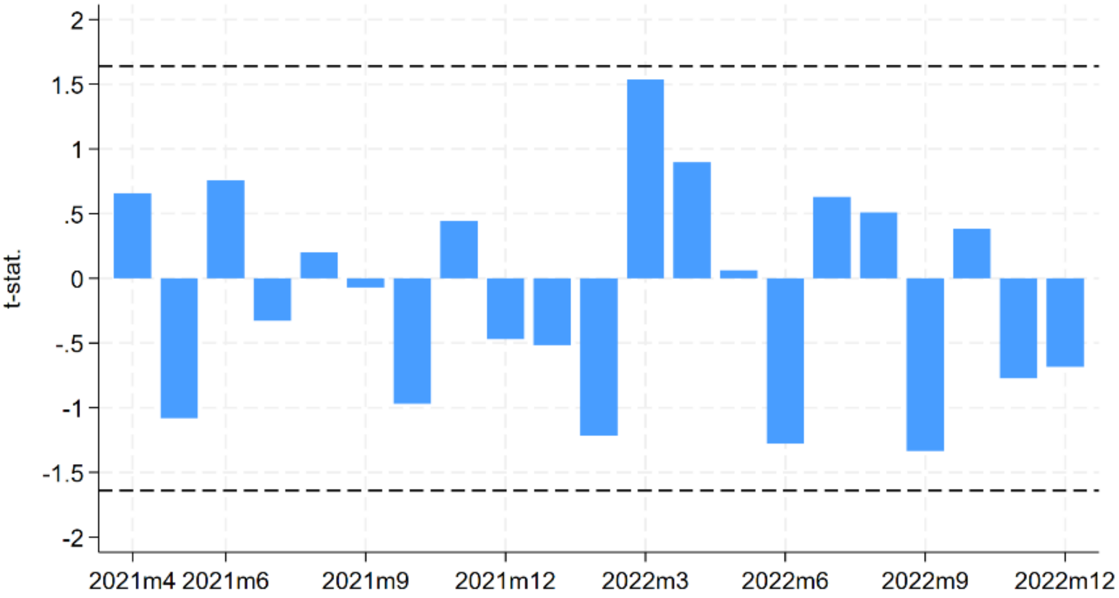
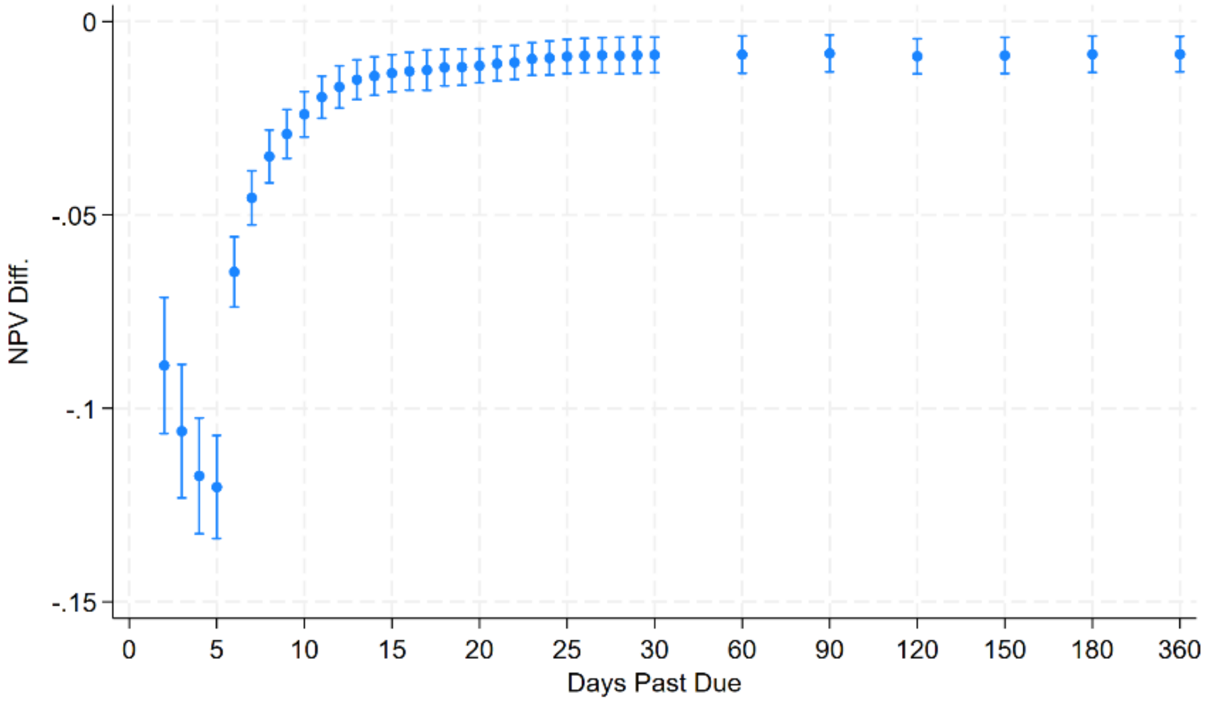


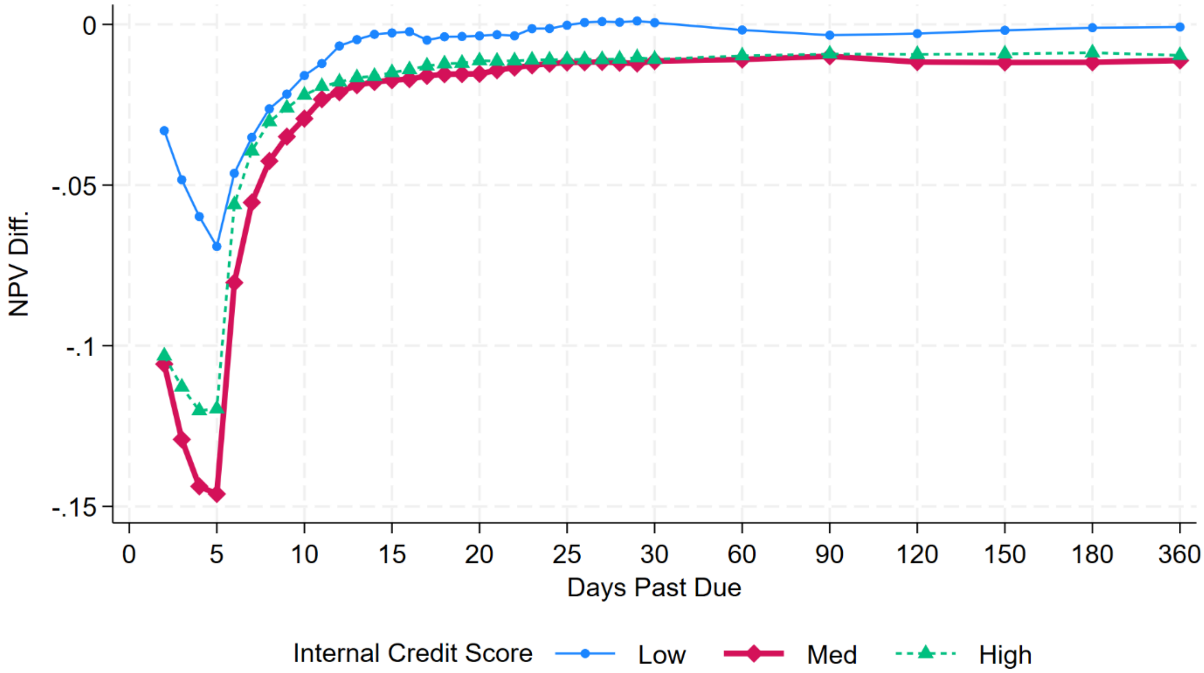
Figure 8. Collected NPV differences between AI and human callers over horizon – Completely randomized subsample.

This figure reports the average differences of collected NPV between AI and human callers over the horizon of days past due, using the 10% completely randomized subsample. The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. The differences are estimated by t-tests on collected NPV between the two groups of callers. For clarity, the differences are plotted daily before day 30, and every 30 days afterwards. Panel (a) reports the differences estimated by pooling all cases together. Panel (b) splits the cases by internal credit score (left panel) or overdue payment size (right panel) and estimates the differences separately. In the left panel, "Low", "Med", and "High" refer to cases with internal credit scores lying in 1-3, 4-7, and 8-10, respectively.

(a) All cases.



(b) By internal credit score.



(c) By loan size.

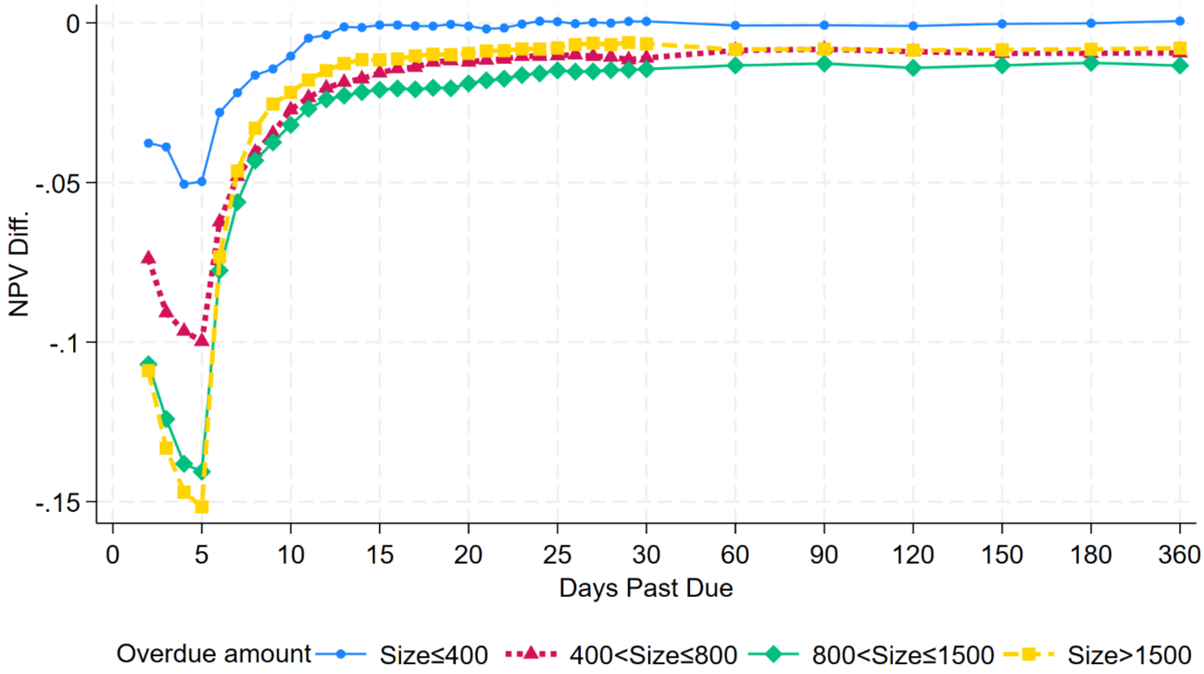


Figure 9. Fractions of different versions of AI callers over time.

This figure shows the fractions of cases assigned to five versions of AI callers every month in our sample period. The length of the bars represents the fraction of cases, and they sum up to one within each month. The first version of AI caller in our sample period is labeled as “v1,” which is not the same as the very first version of AI callers used by the company. Subsequent versions are labeled as “v2” to “v5” according to the time of introduction. The fractions are calculated based on the 10% completely randomized subsample.

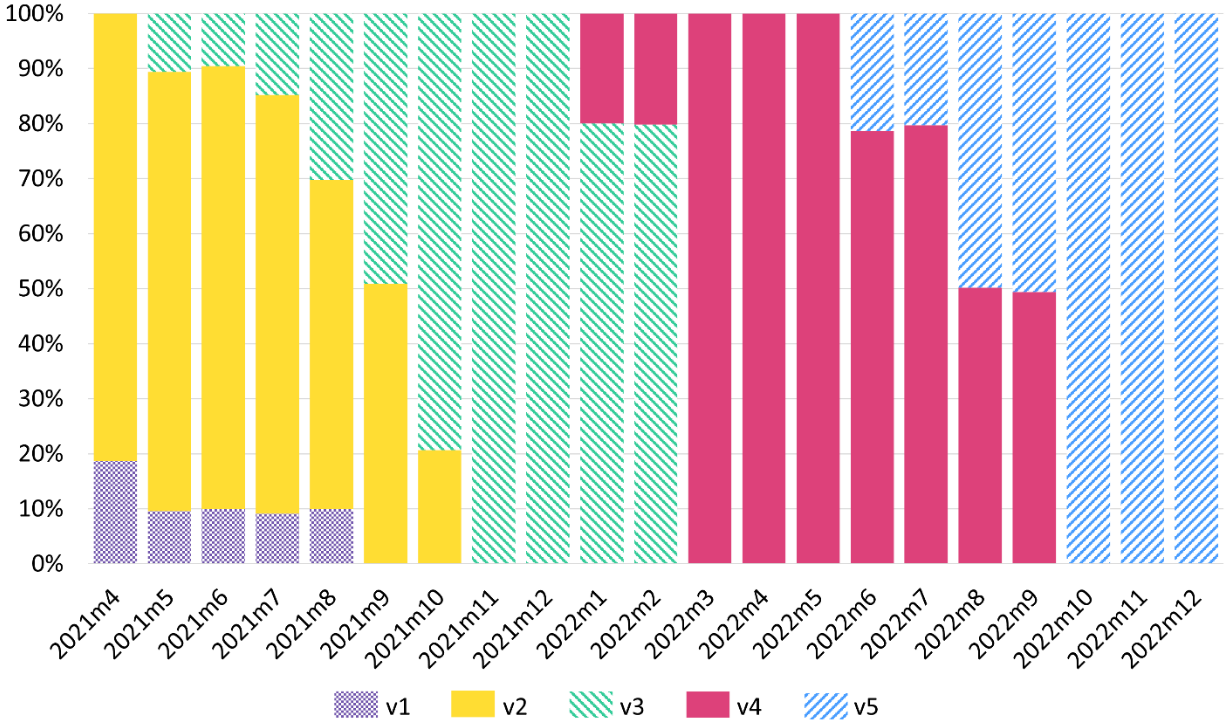
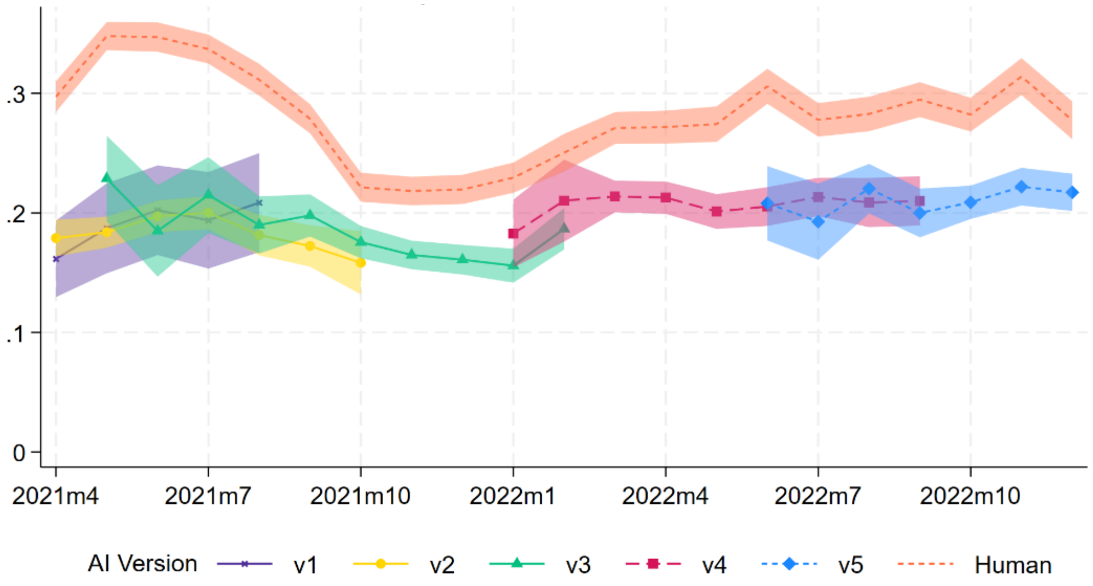


Figure 10. Performance of different versions of AI callers over time

This figure shows the monthly performance of different versions of AI callers and human callers measured by average collected NPV in 2 days past due (Panel (a)) and 5 days past due (Panel (b)). The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. The shaded areas represent the 95% confidence intervals.

(a) NPV 2d



(b) NPV 5d

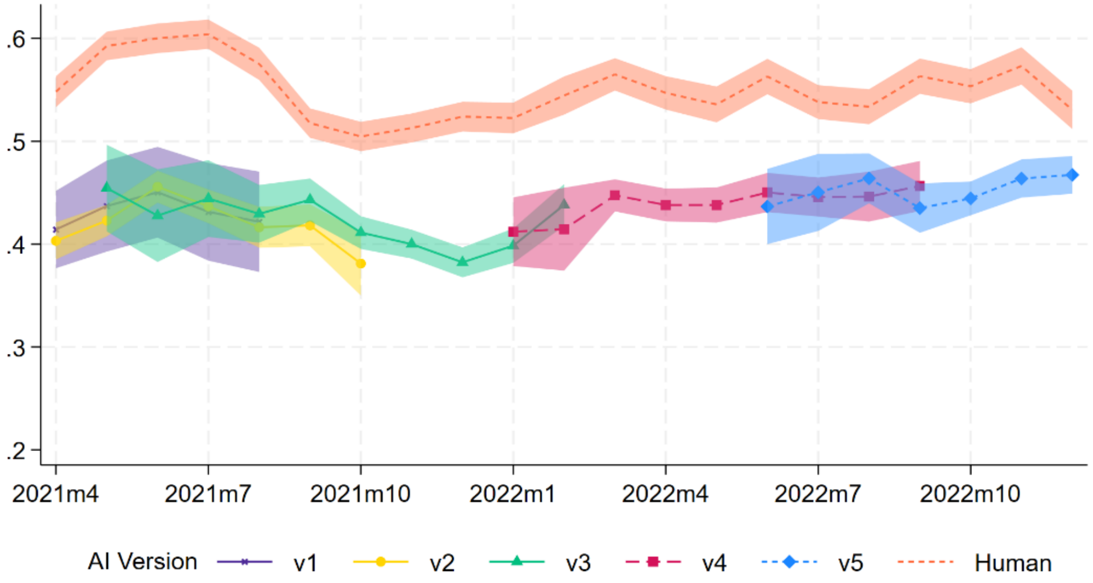
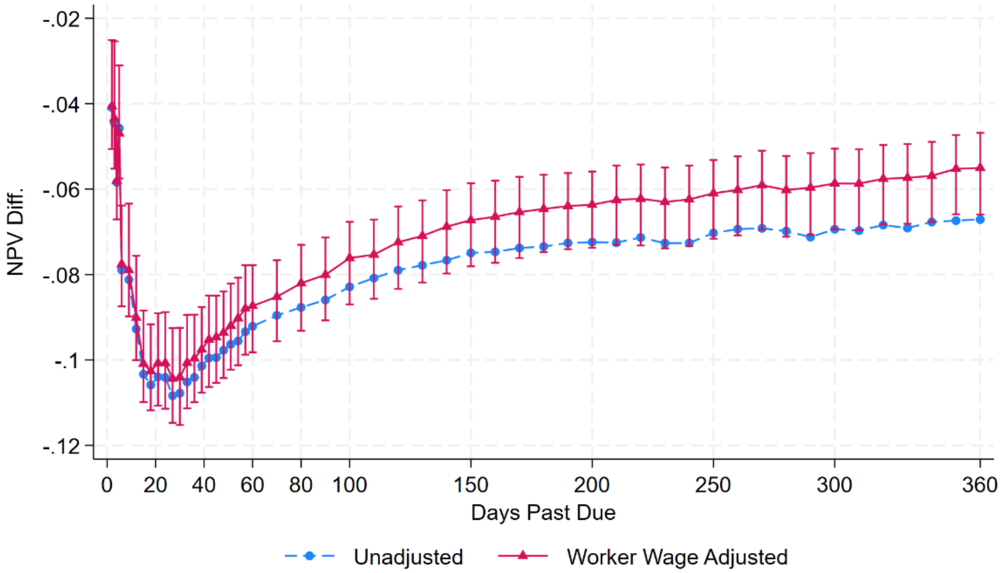


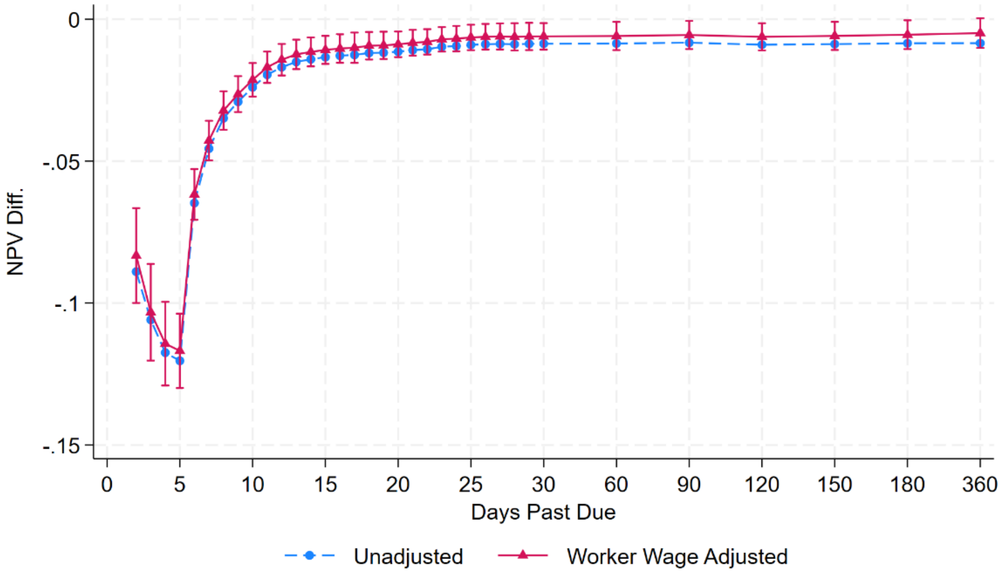
Figure 11. Collected NPV differences between AI and human callers over time – adjustment for labor costs.

This figure reports the average differences of net collected NPV between AI and human callers over the horizon of days past due using three different subsamples. The net collected NPV of a case is defined as the present value of cash flows collected from the case, net of estimated labor costs, discounted by a 24% APR, and scaled by the initial overdue balance. The labor cost-adjusted NPV differences are in solid lines, with 95% confidence intervals, and the unadjusted NPV differences reported above are in dashed lines. Panels (a) and (b) are about small cases using RDD estimation and the 10% completely randomized subsample, respectively.

(a) Small cases – RDD.



(b) Completely randomized subsample.



Tables

Table 1. Sample scripts of an AI caller.

This table presents the conversation procedure and sample scripts that an AI caller typically uses in a phone call for debt collection. The responses are generated by the AI following some prespecified rules after understanding borrowers' speech. The third stage "Further interaction" may include multiple rounds of questions and responses.

Stage	Stage Content	Sample Scripts
1	(1) Greeting. (2) Confirm the identity of the borrower.	AI: Hello, is that Mr./Ms. ABC (borrower's full name) speaking? Borrower: Yes (AI moves forward to the second stage) Borrower: No (AI apologizes and hangs up the phone)
2	(1) Inform the borrower about the delinquent debt. (2) Address the negative outcomes of default.	AI: Mr./Ms. ABC, this is XYZ company. You previously borrowed … yuan (principal) from our company, and you should pay the monthly repayment of … yuan (monthly payment amount) in full by … (bill date). But, by today, you have been missing the monthly repayment for … days (days overdue). Please repay the amount by the end of today / within … hours. Any overdue payment can lead to declining credit scores, large late fees, difficulties in future borrowing and consumption, and even lawsuits. We may also call your "emergency contacts" if you fail to repay.
3	Further interaction	
	Case A1. Borrower agrees to repay by the due. AI confirms the promise.	Borrower: Yes, I will repay the debt. AI: Could you promise to repay the overdue amount in full by the end of today / within … hours? Borrower: Yes. AI: Great! I have recorded your promise in our system. Please keep your promise. Otherwise, we will continue calling you.
	Case A2. Borrower agrees to repay later. AI confirms the promise and addresses potential negative consequences.	Borrower: No, I cannot do it today. How about tomorrow? AI: Your loans are related to your credit records. If you continue to delay, your case will enter the debt collection procedure. Can you repay *today*? Borrower: Tomorrow is the best I can do. AI: Okay. We have recorded your promise. No late fees will be charged for today. But if you continue to default, your credit records will be negatively affected.

Case B1. Borrower refuses to repay and does not explain the reasons. (1) AI emphasizes the negative outcomes again. (2) AI asks for the reasons.	Borrower: No, I cannot repay the debt right now. AI: It is your obligation to repay the debt. Any overdue payment can lead to … (potential negative consequences). AI: Could you tell me the reason why you cannot repay the debt? Do you have any difficulties in repaying the debts? (Then move to Case B2.)
Case B2. Borrower refuses to repay and explains the reason. AI replies accordingly and addresses the negative outcomes.	Borrower: I cannot repay today because I have no money to repay / I am very busy today. AI: You can always find a way to raise money. For example, you can borrow from your family members and friends. / I understand you are busy. Please keep in mind that … (potential negative consequences).
Case C. Borrower denies having debts with the company. AI asks the borrower to recall.	Borrower: I don't know XYZ company. / I have never borrowed money from your company. AI: Please recall carefully if you have ever borrowed money from XYZ company. The company's name is spelled as "X-Y-Z." Please be advised that overdue repayment can lead to negative outcomes.
Case D. Borrower claims that he/she has repaid the debt fully or has set up auto-payment.	Borrower: I have already paid back the debt this morning, haven't you received the money yet? AI: But there is still … yuan on your balance. Borrower: I have set up auto-payment. AI: Okay. We will charge … yuan from your linked bank / WeChat / Alipay account shortly. Please make sure that you have sufficient balance in your bank account.
Case E. Borrower asks for additional information.	Borrower: Do I need to pay any late fees? AI: Late fees include overdue interests and principles, credit evaluation fees, guarantee fees, and so on. Details can be found in our app and your loan contract. Borrower: If I can repay today, do I need to pay late fees? AI: Okay, we will temporarily suspend additional debt collection actions. You can make sure you will repay within 2 hours, right? Borrower: How long has it been overdue? AI: You have been 5 days past due. We have sent you several text messages before.

4	Closing words. When the borrower has no more questions, or when the borrower’s questions do not belong to the above cases, or when the AI cannot understand the borrower’s response.	AI: Okay. Please be advised that you will be responsible for any negative consequences of default. If you have any other questions, feel free to contact our customer service. Bye!

Table 2. Summary statistics of delinquent loans

This table reports summary statistics of the full sample of delinquent loans and two different subsamples used in our analyses. Loan characteristics, including overdue payment, remaining principal, and internal credit score, are measured on day 2 past due. Borrower characteristics include an indicator of male, age, and an indicator of the borrowers having a bachelor's degree or above. Panel A is about the full sample of delinquent loans in the debt collection process. Panel B summarizes the subsample of small cases for regression discontinuity design (RDD). The subsample is restricted to all delinquent loans with remaining principal between 100 yuan and 500 yuan. Panel C shows the 10% completely randomized subsample, which is restricted to borrowers' second delinquency.

Panel A. Full sample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	No. Obs.
Overdue payment (yuan)	1,128.1	1,822.4	0.01	14.7	316.0	653.5	1,304.6	7,688.8	808,666.7	22,122,179
Remaining principal (yuan)	6,474.0	7,330.0	0.01	48.6	1,792.5	4,248.1	8,500.0	34,448.4	1,000,000.0	22,122,179
Internal credit score	5.42	2.85	1	1	3	5	8	10	10	22,122,179
Male indicator	0.70	0.46	0	0	0	1	1	1	1	22,122,179
Age	27.43	6.36	18	19	23	26	31	46	60	22,122,179
Bachelor's degree or more indicator	0.13	0.34	0	0	0	0	0	1	1	22,122,179

Panel B. RDD subsample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	No. Obs.
Overdue payment (yuan)	143.37	113.41	20.01	22.13	58.93	107.07	190.29	503.16	848.18	1,046,165
Remaining principal (yuan)	305.25	112.57	100.00	103.36	209.02	309.29	401.03	496.22	499.99	1,046,165
Internal credit score	4.93	2.77	1	1	3	4	7	10	10	1,046,165
Male indicator	0.72	0.45	0	0	0	1	1	1	1	1,046,165
Age	26.85	6.02	18	19	22	25	30	46	60	1,046,165
Bachelor's degree or more indicator	0.10	0.31	0	0	0	0	0	1	1	1,046,165

Panel C. Completely random subsample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	No. Obs.
Overdue payment (yuan)	1,522.9	1,846.4	20.2	86.3	554.8	1,018.0	1,849.4	8,653.9	35,639.9	147,426
Remaining principal (yuan)	8,593.9	6,966.4	300.1	467.5	3,438.0	6,600.1	11,667.8	30,968.8	34,919.6	147,426
Internal credit score	5.97	2.71	1	1	4	6	8	10	10	147,426
Male indicator	0.70	0.46	0	0	0	1	1	1	1	147,426
Age	27.77	6.79	18	19	22	26	32	47	59	147,424
Bachelor's degree or more indicator	0.10	0.31	0	0	0	0	0	1	1	147,426

Table 3. Comparison between permanent AI callers and human callers – small cases RDD results

This table compares loan characteristics and performance of small cases assigned to AI callers almost permanently and to human callers by utilizing the 300-yuan remaining principal threshold using regression discontinuity design (RDD). Panel A reports the randomization test results about loan characteristics, including overdue payment amount (yuan), internal credit score, the fraction of males, age, and the fraction of borrowers with bachelor's or higher degrees. Columns 2 and 3 report the local average of variables of interest on the left side (permanent AI) and the right side (human) of the threshold. Column 4 reports the differences between the left and right local averages, with *z*-statistics, *p*-values, and 95% robust RD confidence intervals in the following columns. Panel B reports the performance of the two treatments measured by collected NPV, which is defined as the present value of cash flows collected from the case within a given horizon, discounted by a 24% APR, and scaled by the initial overdue balance. In addition to the first seven columns as in Panel A, Panel B column 8 re-estimates the differences around the threshold by including all five covariates in Panel A. Local linear regressions with uniform kernels are used in the estimation in all rows. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)		(8)
	Variable	Left Mean (AI)	Right Mean (Human)	Diff. (L–R)	*z*-stat.	*p*-val.	95% Robust RD C.I.		Diff. with Covars.
Panel A. Loan characteristics									
(1)	Overdue payment	140.3	136.8	-3.54	-0.61	0.542	-7.73	4.06	
(2)	Credit score	4.82	4.87	0.05	0.20	0.840	-0.14	0.18	
(3)	Male	0.716	0.717	0.001	0.14	0.891	-0.013	0.015	
(4)	Age	26.67	26.87	0.21	1.42	0.156	-0.100	0.620	
(5)	Bachelor's degree or higher	0.104	0.108	0.004	1.07	0.283	-0.002	0.009	
Panel B. NPV collected									
(6)	NPV 2d	0.238	0.279	-0.041***	5.96	<0.001	0.026	0.051	-0.036***
(7)	NPV 5d	0.451	0.497	-0.046***	6.35	<0.001	0.030	0.056	-0.042***
(8)	NPV 10d	0.595	0.687	-0.092***	14.04	<0.001	0.077	0.102	-0.087***
(9)	NPV 30d	0.734	0.842	-0.108***	18.98	<0.001	0.097	0.119	-0.105***
(10)	NPV 60d	0.778	0.870	-0.092***	17.87	<0.001	0.083	0.103	-0.093***
(11)	NPV 90d	0.792	0.878	-0.086***	17.67	<0.001	0.077	0.097	-0.086***
(12)	NPV 180d	0.811	0.884	-0.073***	16.51	<0.001	0.066	0.083	-0.072***
(13)	NPV 360d	0.820	0.887	-0.067***	15.25	<0.001	0.060	0.077	-0.066***

Table 4. Difference between AI and human callers – Completely randomized subsample.

This table compares loan characteristics and performance of two types of cases: (a) handled by AI callers on day 2 to day 5 past due before being assigned to human callers on day 6 and (b) handled by human callers starting on day 2 past due using the 10% completely randomized subsample. Panel A reports the randomization test results about loan characteristics, including overdue payment amount (yuan), internal credit score, the fraction of males, age, and the fraction of borrowers with bachelor's or higher degrees. Columns 2 and 3 report the average of variables of interest among cases assigned to AI (type a) and human callers (type b), respectively. Column 4 reports the differences between the averages, with *t*-statistics in the following column. Panel B reports the performance of the two treatments measured by collected NPV, which is defined as the present value of cash flows collected from the case within a given horizon, discounted by a 24% APR, and scaled by the initial overdue balance. The estimations are based on linear regressions of the variable of interest onto an AI-case indicator with calendar month fixed effects. *t*-statistics are adjusted for clustering at the calendar month level. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
Panel A. Loan characteristics					
(1)	Overdue amount	1523.7	1522.2	1.5	0.26
(2)	Remaining principal	8585.2	8604.7	-19.5	-0.69
(3)	Internal credit score	5.970	5.961	0.009	0.72
(4)	Male	0.701	0.703	-0.002	-1.47
(5)	Age	27.75	27.79	-0.043	-1.26
(6)	Bachelor's degree or higher	0.103	0.104	-0.001	-0.39
Panel B. Collected NPV					
(7)	NPV 2d	0.193	0.282	-0.089***	-10.54
(8)	NPV 5d	0.431	0.551	-0.120***	-18.87
(9)	NPV 10d	0.647	0.671	-0.024***	-8.53
(10)	NPV 30d	0.767	0.776	-0.0086***	-3.94
(11)	NPV 60d	0.800	0.809	-0.0086***	-3.70
(12)	NPV 90d	0.816	0.824	-0.0083***	-3.62
(13)	NPV 180d	0.830	0.838	-0.0085***	-3.76
(14)	NPV 360d	0.836	0.844	-0.0084***	-3.85

Table 5. Phone call outcomes of AI and human callers.

This table compares phone call outcomes for calls made on day 2 past due between two types of cases: (a) handled by AI callers on day 2 to day 5 past due before being assigned to human callers on day 6 and (b) by human callers starting on day 2 past due using the 10% completely randomized subsample. Columns 2 and 3 report the average of variables of interest among cases assigned to AI (type a) and human callers (type b), respectively. Column 4 reports the differences between the averages, with *t*-statistics in the following column. The estimations are based on linear regressions of the variable of interest onto an AI-case indicator with calendar month fixed effects. *t*-statistics are adjusted for clustering at the calendar month level. Panel A is about all phone calls made on day 2 past due while Panels B and C restrict the sample to the first call answered by each borrower. The time of calling is represented by hours from midnight in decimals. The timing adjustment accounts for the time of calling by including fixed effects for the time of calling every hour. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

Panel A. All calls on day 2 past due.

	(1)	(2)	(3)	(4)	(5)
	Variables	Mean (AI)	Mean (Human)	Diff: AI – Human	*t*-stat.
(1)	# Phone calls per borrower	4.617	5.465	-0.848***	-5.77
(2)	# Phone calls answered	0.65	1.00	-0.35***	-19.44
(3)	% Phone calls answered	0.236	0.236	0.000	0.02
(4)	Ringing time to answer (sec)	18.52	19.03	0.51	1.00
(5)	Time of calls answered	13.10	13.21	0.11	1.18

Panel B. First answered calls.

	(1)	(2)	(3)	(4)	(5)
	Variables	Mean (AI)	Mean (Human)	Diff: AI – Human	*t*-stat.
(1)	Time of calls	11.79	11.51	0.28**	2.43
(2)	Ringing time to answer (sec)				
	Unadjusted	19.47	20.72	-1.25**	-2.16
	Timing-adjusted	20.11	20.13	-0.02	-0.04
(3)	Duration (sec)				
	Unadjusted	28.12	47.13	-19.02***	-11.96
	Timing-adjusted	21.76	52.72	-30.96***	-18.93
(4)	% Promise to repay				
	Unadjusted	0.441	0.652	-0.212***	-16.91
	Timing-adjusted	0.441	0.652	-0.212***	-19.84

Panel C. Repayment after first answered calls (all timing-adjusted).

(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
Repay (fully or partially) after the first answered call within …				
15 minutes	0.039	0.040	-0.001	-0.53
30 minutes	0.053	0.063	-0.010***	-3.91
1 hour	0.071	0.098	-0.026***	-7.26
2 hours	0.102	0.152	-0.050***	-8.73
5 hours	0.156	0.243	-0.087***	-9.56
the same day	0.231	0.342	-0.111***	-9.79

Table 6. Performance of different versions of AI callers

This table reports the performance differences between consecutive versions of AI callers at the horizons of 2-6, 8, and 10 days past due. Performance is measured by collected NPV, which is defined as the present value of cash flows collected from the case within a given horizon, discounted by a 24% APR, and scaled by the initial overdue balance. For each pair of AI callers, the analyzing sample is extracted from the 10% completely randomized subsample in months when both AI callers are deployed. For versions "V2" and "V3" because the transition time was six months long, only the last two months (i.e., September and October 2021) – when "V3" took up substantial fractions – are used. The differences are estimated by linear regressions of collected NPV onto an indicator of the newer AI callers with calendar month fixed effects. The last row reports the sample average NPVs. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	NPV Horizon (days past due)						
Version Diff.	2	3	4	5	6	8	10
V2 - V1	-0.0005	0.0015	-0.0029	-0.0045	-0.0050	0.0002	-0.0083
	(-0.06)	(0.15)	(-0.28)	(-0.42)	(-0.46)	(0.018)	(-0.81)
V3 - V2	0.0222**	0.0213**	0.0246**	0.0270**	0.0099	0.0055	0.0021
	(2.57)	(2.05)	(2.25)	(2.40)	(0.86)	(0.49)	(0.19)
V4 - V3	0.0255**	0.0247*	0.0147	-0.0016	0.0029	0.0019	-0.0015
	(2.21)	(1.80)	(1.03)	(-0.11)	(0.20)	(0.13)	(-0.11)
V5 - V4	-0.0031	-0.0105	-0.0012	-0.0031	0.0027	0.0045	0.0060
	(-0.40)	(-1.15)	(-0.13)	(-0.32)	(0.28)	(0.48)	(0.65)
Average NPV	0.193	0.328	0.386	0.430	0.529	0.606	0.647

Table 7. Human caller performance on day 6 after AI callers were upgraded to V3.

This table examines the impacts of AI caller upgrade on human callers' performance on day 6 past due. The sample of cases is restricted to the completely randomized subsample in September and October 2021 when AI caller versions V2 and V3 coexist. The sample cases are also required to remain unpaid on day 6. The sample of callers is restricted to callers specializing in the "M1 Early" stage (days 2-10 past due). Column 1 regresses human caller performance on day 6 (i.e., NPV6 minus NPV5, denoted by "ΔNPV6") onto an indicator of being treated by version 3 AI callers in the first five days and month fixed effects. Column 2 adds callers' day-6 performance on cases treated by AI V2 in the previous month (*PrevPerfRank*) as additional covariates. The performance is normalized to fractional ranking within a month, with 1 being the top caller. Column 3 includes new callers who have no previous performance ranking, which is set to zero. An indicator of new callers is added as additional covariates. Cluster-adjusted *t*-statistics clustered at the caller level are reported in parentheses. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)	(3)
	ΔNPV6	ΔNPV6	ΔNPV6
AI V3 Indicator	-0.033***	-0.001	-0.004
	(-2.65)	(-0.03)	(-0.17)
PrevPerfRank		0.102**	0.102**
		(2.35)	(2.37)
AIV3 × PrevPerfRank		-0.101**	-0.101**
		(-2.21)	(-2.22)
NewCaller			0.033
			(1.37)
AIV3 × NewCaller			-0.022
			(-0.76)
Constant	0.188***	0.152***	0.152***
	(18.67)	(7.01)	(7.21)
No. of Obs.	4,232	1,595	4,232
No. of Callers	678	356	678
R-squared	0.002	0.007	0.004
Month Fixed Effects	Yes	Yes	Yes

Appendix A. Additional Results of Conditionally Randomized Subsample

1 Case assignment rules of conditionally randomized subsample

After setting aside 10% of cases for experimental purposes, the remaining 90% of cases whose remaining principals are larger than 300 yuan are assigned between AI and human callers randomly conditional on case characteristics, including overdue payment, internal credit score, and borrower's maximum days of overdue in the past. The cases are split into several "cells" according to cut-offs in the three case characteristics. The cut-offs of overdue payments are 400 yuan, 800 yuan, and 1500 yuan, which correspond to roughly the first quartile, median, and third quartile of the distribution. Second, internal credit scores are divided into ten equal-sized deciles. Finally, borrowers with more than four days of overdue in previous delinquency are treated differently than their counterparts. Therefore, the cases are separated into 80 (=4×10×2) cells, and cases within each cell are assigned to AI and human callers randomly over day 2 to day 6 past due.

The timing of assigning to human callers can vary across cells. Cases in some cells may be assigned to human callers earlier on day 2 while the others on day 3, day 4, or day 6. Usually, no switching of caller types happens on day 5. The timing generally depends on the difficulty of collecting the debts. Cases in cells of larger overdue amounts, lower internal credit scores, and worse credit records tend to be assigned to human callers earlier. Once the cases are assigned to human callers, they will not return to AI callers.

The fractions of AI cases may vary across cells. Figure A1 shows the fractions of AI cases on day 2 in different cells, using June 2022 as an example. The horizontal axis shows the overdue amount bins, the vertical axis represents the internal credit score bins, and the panels differentiate bins about previous repayment history. The brightness of the cells indicates the fraction of cases assigned to AI callers. The darker the cell, the more cases are assigned to AI callers on day 2 past due. The general trend is that borrowers with larger overdue payments, lower credit scores, and better credit histories are more likely to be assigned to human callers.

The fraction of cases allocated to AI callers within a cell may also vary over time. Figure A2 shows the daily variation of the fractions within June 2022 for different types of cases. For illustration, we merge all cells into categories in one of the three dimensions, so the numbers reflect the average fractions across cells within the same category. Panel (a), (b), and (c) correspond to categories by overdue payment amount, internal credit score group, and previous maximum days past due, respectively. In Panel (a), cases below 400-yuan overdue amounts are almost all assigned to AI callers before the 26th of the month; while for larger cases, about 70% to 80% are treated by AI before the 21st and about 20% to 40% afterward. The fractions in all size categories reduce to zero after the 26th. Regarding internal credit score in Panel (b), cases in the high-score category are mostly assigned to AI before the 26th, and the medium-score category follows a similar two-phase pattern as we see in Panel (a). In contrast, the fraction of cases assigned to AI callers in the low-score category remains around 50% and reduces to even less towards the end of the month. Finally, in terms of the previous credit records, better-record borrowers are more likely to be handled by AI callers and both categories see a decline in AI fractions on the 21st of the month.

2 Performance of AI versus human callers

Table A1 reports summary statistics of the conditionally randomized subsample. Loan characteristics, including overdue payment amount, remaining principal, and internal credit score, are measured on day 2 past due, the first day of phone contacts. Borrower characteristics include an indicator of male, age, and an indicator of the borrowers having a bachelor's degree or above. The sample is restricted to borrowers' second delinquency, which is truly random as explained in Section 3 in the main text. Compared to Table 2, Table A1 suggests that cases and borrowers in the conditionally randomized subsample are similar to those in the 10% completely randomized subsample, consistent with the stated case assignment rules.

Figure A3 tests the randomization of case assignment in the conditionally randomized subsample using loan size variables, namely, overdue payment amount and remaining principal. Specifically, we implement two-sample *t*-tests between cases assigned to AI and human callers on day 2 past due every month. To account for the conditioning in case assignment and the potential daily variation in the fraction of AI cases, calendar date-by-cell fixed effects are included in the tests. The bars show the *t*-statistics and the horizontal dashed lines indicate ±1.64, the critical values of a 10% significance level. The figure suggests that there is no significant difference between the two groups in terms of loan sizes. Also, the *t*-statistics distribute evenly above and below zero. They show no time trend or clustering by time.

Table A2 formally tests the differences between AI and human callers by regressing the variable of interest onto an indicator of the treatment group (some AI treatment on day 2 to day 5) and date-by-cell fixed effects. The coefficient of the AI indicator can be interpreted as the average difference between the two groups of cases. The dependent variables in Panel A are loan and borrower characteristics, including overdue payment amount, remaining principal, an indicator of male, age, and an indicator of having a bachelor's degree or more. Panel A confirms the conditional randomization of the cases since all five variables are indifferent between the two groups with insignificant *t*-statistics.

We then examine the performance gaps between the "AI + Human" group and the all-human control group in Table A2 Panel B using the same regression specification. The evaluation horizons range from 2 days to 360 days (one year) past due. The table shows that the differences between the two groups are significantly negative, suggesting that the "AI + Human" group underperforms the all-human control group. On day 2 past due, the collected NPV difference is -0.089, or a 27.3% productivity loss compared to human callers. For longer horizons when human callers take over the job, the gap converges to around -0.007, which is less than 1% productivity loss, moderately significant in an economic sense. These findings are consistent with what we see in Section 4.3 using the 10% completely randomized subsample. Also note that the standard errors on these estimates are only 21% smaller than in the completely randomized subsample, despite having 9 times as many observations. This is because date-by-cell fixed effects are included to account for the fact that the fraction of cases assigned to AI callers may vary across cells and over time. In contrast, only calendar month fixed effects are included in the regressions with the completely randomized subsample since the randomization rule does not change within a month. The inclusion of date-by-cell fixed effects may reduce the effective sample size for inference.

Figure A4(a) presents the differences in collected NPVs between AI and human callers over horizons up to 360 days past due using the conditionally randomized subsample. The differences are estimated with the same specification as in Table A2. First, since human caller assignment in the AI treatment group happens through day 3 to day 6 past due, instead of on day 6 sharp in the completely randomized subsample, the performance gaps here monotonically decrease over time. However, since most of the cases are still assigned to human callers on day 6, the largest jump occurs on day 6 past due, as in the completely randomized subsample. Other than this aspect, the trend shown here is similar to what we see in the completely randomized subsample: the productivity gap converges quickly as human callers enter the process and the gap is small in the long run in economic sense, though statistically significant.

Figure A4(b) and (c) plots the time series of the NPV gap over evaluation horizons by internal credit score and by overdue amount groups. The results are a bit different than those from the completely randomized subsample since the patterns here reflect variations in both loan characteristics and the timing of human caller assignments across different types of cases. For example, when considering internal credit score groups in Figure A4 Panel (b), if all cases are assigned to human callers on day 6, the medium-score and high-score groups have similar long-run productivity loss, as suggested by the completely randomized subsample in Figure 5. In the conditionally randomized subsample, however, since the company allocates a large fraction of medium-score cases to human callers on day 3 past due, as suggested by the large jump there, the medium-score cases have better outcomes than the high-score cases, whose productivity loss keeps widening in the first five days. Despite the larger *absolute* productivity loss among high-score cases, the *relative* loss is still small given the high baseline collectability of high-score cases. The company, therefore, does not have large incentives to allocate these cases to human callers earlier. Similar things happen with respect to overdue amount groups, as shown in Figure A4 Panel (c). Specifically, small cases are allocated to human callers later than the other groups, leading to larger absolute productivity loss.

To conclude, Figure A5 shows the time series of collected NPV differences between the two groups after adjusting for human caller salary, as what we do in Section 5.2 in the main text. The differences are estimated with the same specification as in Table A2, except that we use labor cost-adjusted NPV as the dependent variable. Similar to what we find in Section 5.2 in the main text, after adjustment, the productivity gap becomes closer to zero and less significant than before.

Figure A1. Fraction of AI cases on day 2 past due across cells in June 2022.

This figure shows the fraction of AI cases across cells defined by case overdue payment size (horizontal axis), internal credit score (vertical axis), and the maximum days of overdue in previous delinquencies (Panels (a) and (b)). The case size uses 400 yuan, 800 yuan, and 1500 yuan as cutoffs, the internal credit score is split into deciles, and the previous maximum days of overdue uses 4 days as the cutoff. The sample is restricted to the 90% conditionally randomized cases larger than 300-yuan remaining principal. The sample period is between June 7, 2022, and June 20, 2022.

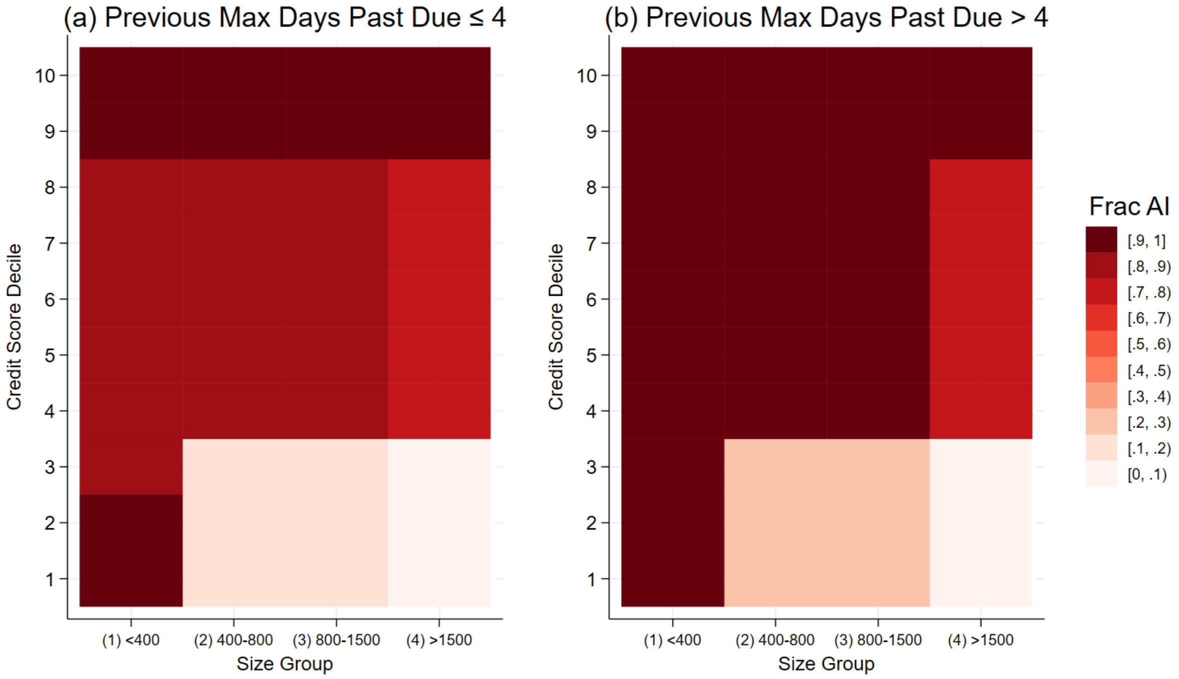
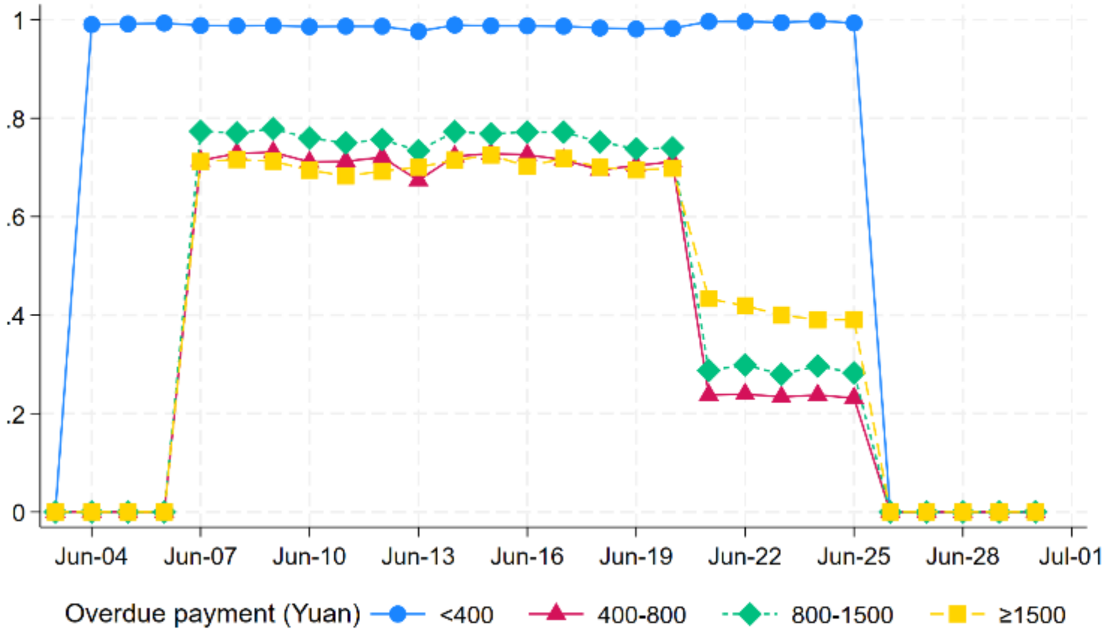


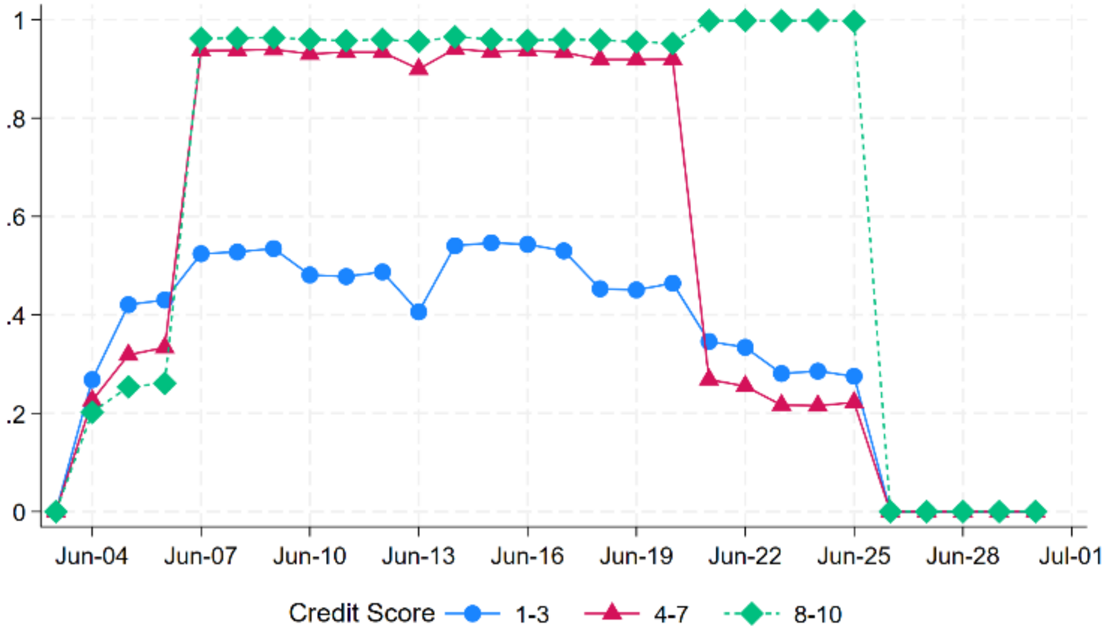
Figure A2. Daily fractions of AI cases in June 2022, by loan characteristics.

This figure shows the daily variation of the fractions of cases assigned to AI callers in June 2022. Panel (a) plots the fractions by case overdue payment size group, Panel (b) by credit score decile groups, and Panel (c) by the group of previous maximum days past due. The sample is restricted to the 90% conditionally randomized subsample of cases defined in the text.

(a) By case size group



(b) By credit score groups



(c) By previous maximum days past due

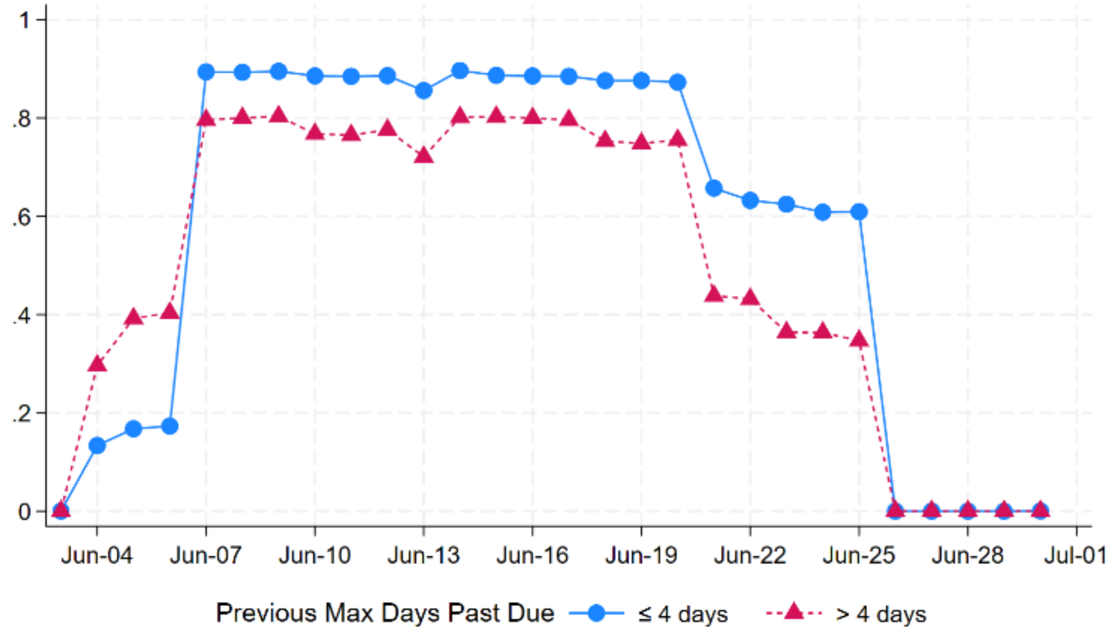
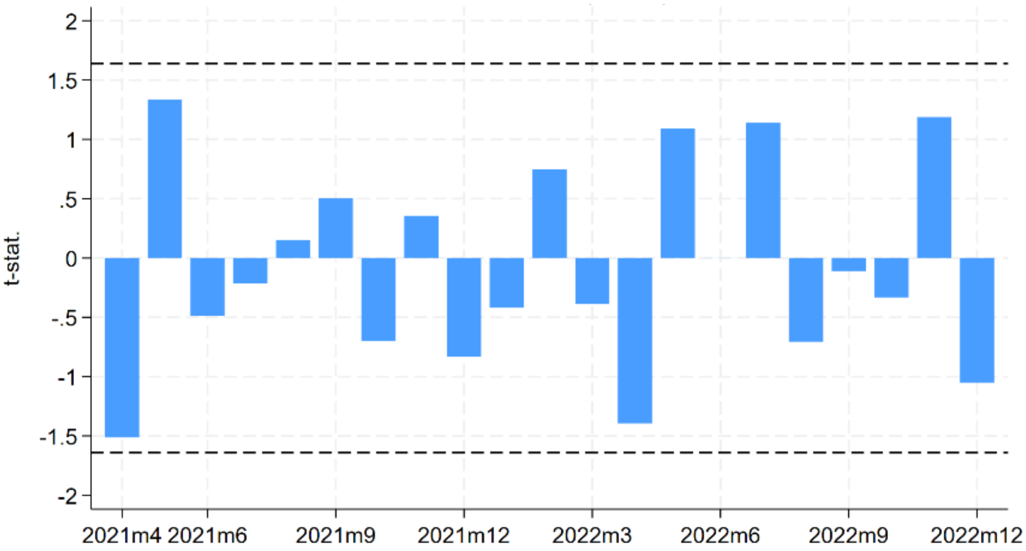


Figure A3. Test of randomization – conditionally randomized subsample.

This figure reports the results of the monthly randomization test on the 90% conditionally randomized subsample. Every month *t*-tests between AI and human callers on loan characteristics are implemented. Date-by-cell fixed effects are included in the tests to account for the assignment rule that cases are randomized between AI and human callers conditioning on characteristic cells every day. Characteristic cells are defined by overdue payment size, internal credit score, and previous maximum days past due as described in the text. The bars show the *t*-statistics of the difference in loan characteristics between AI and human callers. The variables of interest include overdue payment in Panel (a) and remaining principal in Panel (b). The horizontal dashed lines indicate ±1.64, the critical values of 10% significance level.

(a) Overdue payment.



(b) Remaining principal.

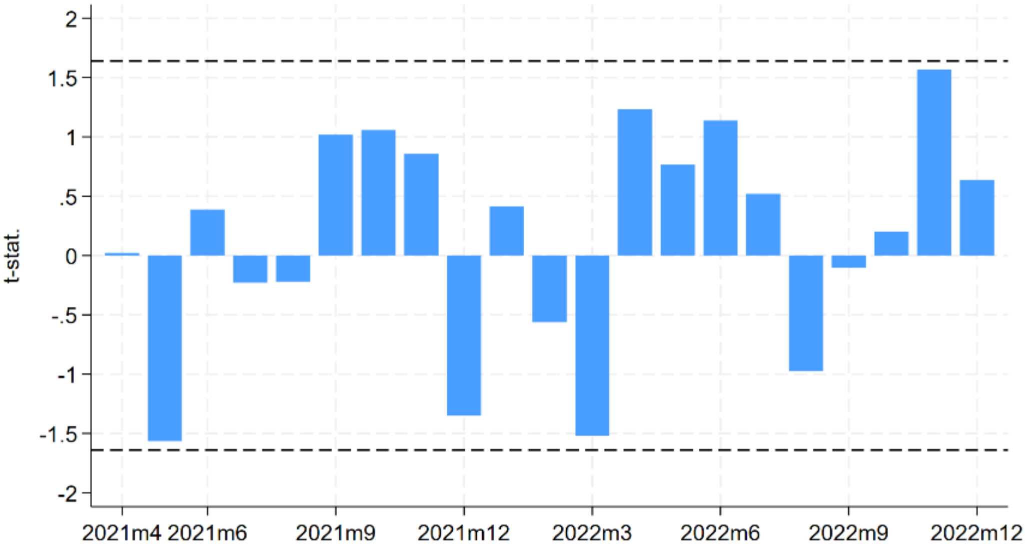
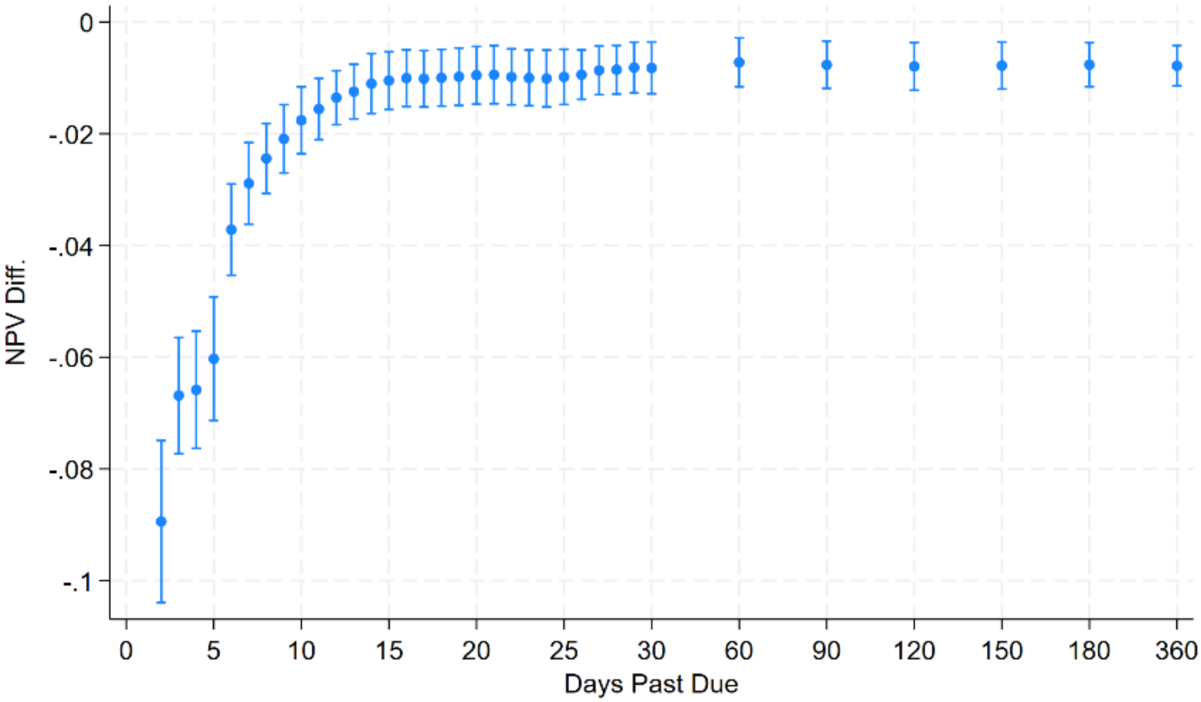


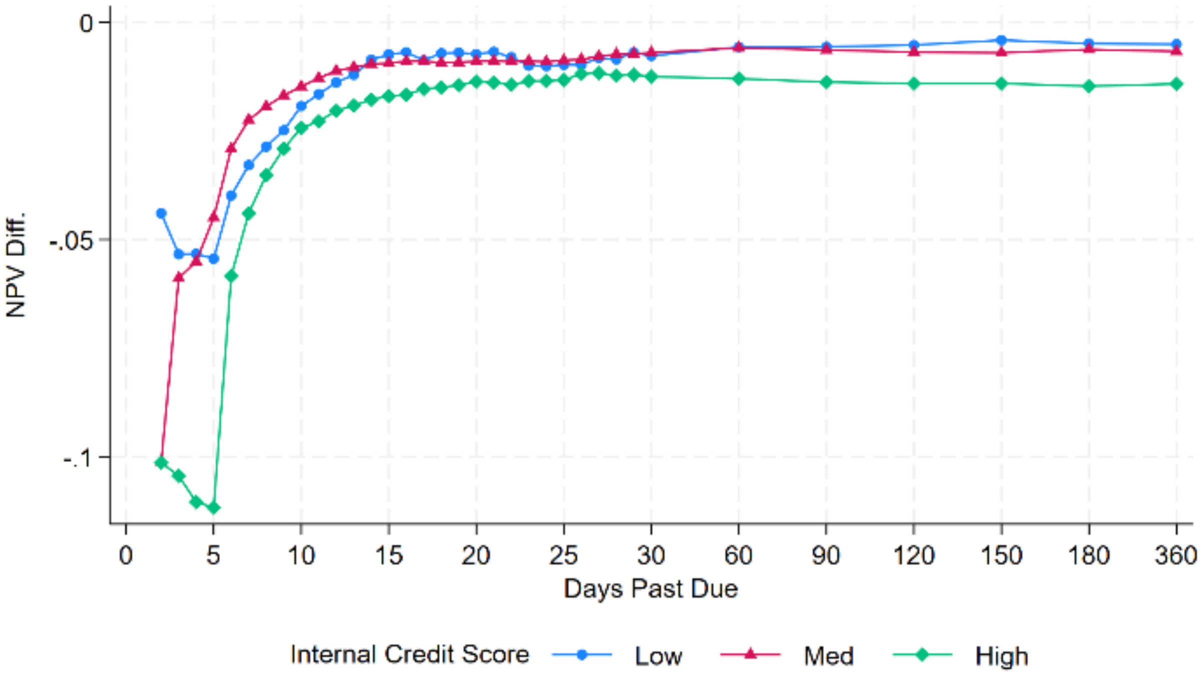
Figure A4. Collected NPV differences between AI and human callers over time – Conditionally randomized subsample.

This figure reports the average differences of collected NPV between AI and human callers over the horizon of days past due, using the 90% conditionally randomized subsample. The collected NPV of a case is defined as the present value of cash flows collected from the case, discounted by a 24% APR, and scaled by the initial overdue balance. The differences are estimated by t-tests on collected NPV between the two groups of callers. Date-by-cell fixed effects are included in the tests to account for the assignment rule that cases are randomized between AI and human callers conditioning on characteristic cells every day. Characteristic cells are defined by overdue payment size, internal credit score, and previous maximum days past due as described in the text. For clarity, the differences are plotted daily before day 30, and every 30 days afterwards. Panel (a) reports the differences estimated by pooling all cases together. Panel (b) and (c) split the cases by internal credit score (left panel) or overdue payment size (right panel) and estimate the differences separately. In the left panel, "Low", "Med", and "High" refer to cases with internal credit scores lying in 1-3, 4-7, and 8-10, respectively.

(a) All cases.



(b) By internal credit score.



(c) By overdue amount.

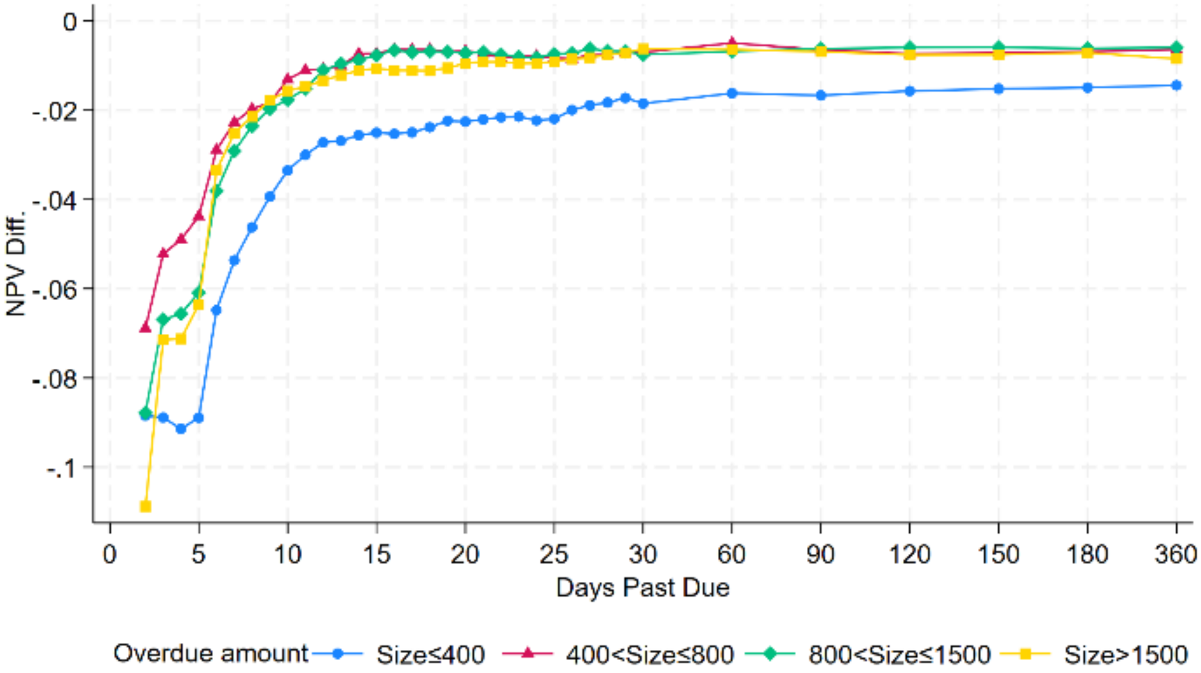


Figure A5. Collected NPV differences between AI and human callers, net of labor costs: Conditionally randomized subsample.

This figure reports the average differences of net collected NPV between AI and human callers over the horizon of days past due using three different subsamples. The net collected NPV of a case is defined as the present value of cash flows collected from the case, net of estimated direct labor costs, discounted by a 24% APR, and scaled by the initial overdue balance. The labor cost-adjusted NPV differences are in solid lines, with 95% confidence intervals, and the unadjusted NPV differences reported above are in dashed lines. The differences are estimated using the 90% conditionally randomized subsample.

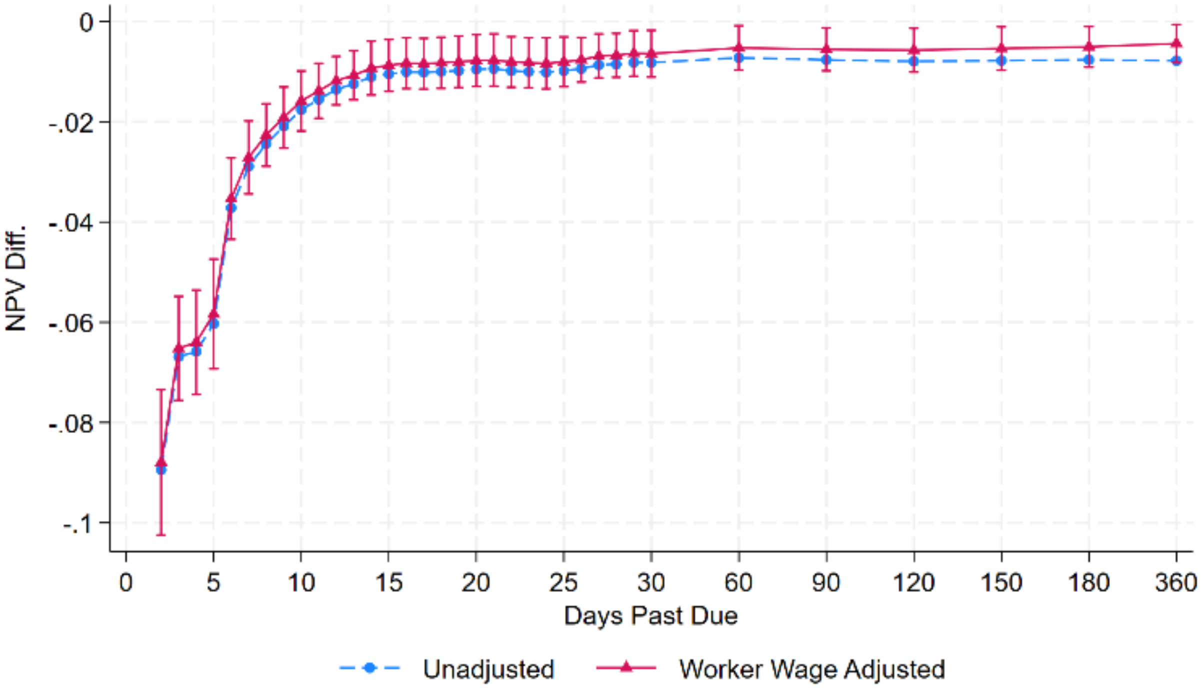


Table A1. Summary statistics of the conditionally randomized subsample.

This table reports summary statistics of the conditionally randomized subsample used in our analyses. Loan characteristics, including overdue payment, remaining principal, and internal credit score, are measured on day 2 past due. Borrower characteristics include an indicator of male, age, and an indicator of the borrowers having a bachelor's degree or above. The sample is restricted to borrowers' second delinquency.

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max
Overdue payment	1,588.6	1,968.6	20.1	91.3	564.8	1,052.7	1,924.6	9,129.9	54,218.4
Remaining principal	9,018.6	7,694.0	300.0	483.4	3,505.1	6,709.6	12,071.4	35,289.9	55,200.8
Internal credit score	5.99	2.70	1	1	4	6	8	10	10
Male indicator	0.70	0.46	0	0	0	1	1	1	1
Age	28.54	6.84	18	19	23	27	33	47	59
Bachelor's degree or more indicator	0.08	0.28	0	0	0	0	0	1	1

Table A2. Difference between AI and human callers – conditionally randomized subsample.

This table compares loan characteristics and performance of two types of cases: (a) assigned to AI callers on day 2 before being assigned to human callers sometime between day 3 and day 5 past due and (b) assigned to human callers starting on day 2 past due by using the 90% conditionally randomized subsample. Panel A reports the randomization test results about loan characteristics, including overdue payment amount (yuan), internal credit score, the fraction of males, age, and the fraction of borrowers with bachelor's or higher degrees. Columns 2 and 3 report the average of variables of interest among cases assigned to AI (type a) and human callers (type b), respectively. Column 4 reports the differences between the averages, with *t*-statistics in the following column. Panel B reports the performance of the two treatments measured by collected NPV, which is defined as the present value of cash flows collected from the case within a given horizon, discounted by a 24% APR, and scaled by the initial overdue balance. The estimations are based on linear regressions of the variable of interest onto an AI-case indicator with date-by-cell fixed effects, where cells are defined by overdue payment size, internal credit score, and previous maximum days past due as described in the text. *t*-statistics are adjusted for clustering at the date level. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)	(3)	(4)	(5)
	Variables	Mean (AI)	Mean (Human)	Diff: AI – Human	*t*-stat.
Panel A. Loan characteristics					
(1)	Overdue amount	1589.3	1589.6	-0.263	-0.04
(2)	Remaining principal	9033.0	9004.9	28.06	1.08
(3)	Male	0.700	0.699	0.001	0.74
(4)	Age	28.54	28.55	-0.011	-0.30
(5)	Bachelor's degree or more indicator	0.0842	0.0827	0.0015	1.14
Panel B. Collected NPV					
(6)	NPV 2d	0.237	0.326	-0.089***	-12.84
(7)	NPV 5d	0.488	0.548	-0.060***	-11.38
(8)	NPV 10d	0.651	0.669	-0.018***	-6.12
(9)	NPV 30d	0.764	0.772	-0.0082***	-3.70
(10)	NPV 60d	0.796	0.803	-0.0072***	-3.45
(11)	NPV 90d	0.810	0.818	-0.0077***	-3.79
(12)	NPV 180d	0.824	0.832	-0.0076***	-4.03
(13)	NPV 360d	0.830	0.838	-0.0078***	-4.53

Appendix B. Additional Results of Regression Discontinuity Design

1 RD plot with variance-mimicking bandwidth

Figure B1 repeats the RD plots in Figure 4 with variance-mimicking bandwidths for variables about loan characteristics and collected NPVs over several horizons. The variance-mimicking bandwidths are around 0.035 yuan, corresponding to about 5500 bins on each side of the cutoff. All other specifications are the same as the main results in Figure 4. Here, the figure shows high variability in the variables of interest, suggesting large heterogeneity among borrowers. For loan characteristics, no significant jump is observed at the 300-yuan cut-off. On the other hand, the differences in collected NPVs at the cut-off are significant despite high variances in the variables. These results are consistent with our observations in Figure 4.

2 Tests of manipulation at the threshold

One important assumption in a valid RD design is that there is no manipulation at the threshold either because agents do not know the cut-off ex-ante or manipulation is impossible or very costly. Since the company never discloses its debt collection assignment rules to the public, borrowers are unlikely to manipulate the remaining principal intentionally to avoid human callers or the opposite.

Formally, we validate the no-manipulation assumption by examining the distribution of observations around the cut-off. Figure B2 shows the results of the RD density test. Figure B2(a) uses the true cut-off of 300-yuan remaining principal. The figure first shows the histogram of the running variable, the remaining principal. Since the assignment rule is right-continuous, we require the intervals to include their right ends but not the left ends. Therefore, the precisely 300-yuan cases belong to the right-most bar on the left of the cut-off. The histogram shows that there is an increase in observation density just below the cut-off. We believe that this is because borrowers and lenders tend to round to multiples of 100 yuan: lenders may want to write loan amounts with a 100-yuan step size, and borrowers may prefer to keep a balance of a whole hundred yuan when repaying their principal. Despite such a tendency, the density function (solid lines) estimated by local quadratic regressions show no significant jump at the cut-off, as the robust RD *t*-statistics is only -1.01, suggesting that the tendency of rounding is not statistically significant.

The tendency of rounding is also observed at 200 yuan and 400 yuan, as shown in Panels (b) and (c) in Figure B2. In these placebo tests, we use an artificial cut-off of 200 or 400 yuan and do the same calculation as in Panel (a). We find similar increases in density at these artificial cut-offs, so the bunching at the 300-yuan cut-off is not abnormal. In addition, our randomization tests in Table 3 Panel A suggest that such a tendency of rounding is unrelated to observed loan and borrower characteristics.

Following Cattaneo et al. (2019), we also implement a binomial test at the cut-off. The test counts the number of observations just below and above the cut-off within a given symmetric neighborhood around the cut-off. If there is no manipulation at the cut-off, the observations should be distributed as-if random below and above the cut-off. Therefore, under the null hypothesis of no manipulation, the fraction of cases below the cut-off should be 50%. The binomial test then examines whether the fraction is significantly far from 50%. Table B1 reports the results. For a neighborhood radius below 2 yuan, there are significantly more cases equal to or smaller than 300 yuan. As we consider a larger radius of up to ±5 yuan around the cut-off, the distribution is balanced. This can be explained by the decreasing tendency of rounding to 300 yuan as people move far away from the cut-off. Moreover, in the following section, we show that our results are robust to excluding those potentially rounded observations. We, therefore, conclude that there is no intentional manipulation related to AI caller usage at the 300-yuan cut-off.

3 Robustness check

Table B2 performs robustness checks of the RD regression results by varying the specifications. As a reference, column 1 repeats our main results in Table 3 Panel B, which uses the MSE-optimal bandwidth and uniform kernel. Columns 2 and 3 change kernel choice to triangular kernel and Epanechnikov kernel, respectively. Column 4 uses the CER-optimal bandwidth. Column 5 doubles the MSE-optimal bandwidth and column 6 shrinks it by half. These variations generate results similar to the main setup in terms of magnitude and significance. It confirms that our results are robust to bandwidth and kernel choices.

The last three columns in Table B2 conduct a "donut-hole" test, which checks the robustness of our results to observations close to the cut-off. This approach can evaluate the sensitivity of the results to manipulation, even if it is not suspectable, as well as the sensitivity to the unavoidable extrapolation in local linear regressions. In the test, observations within ±w of the cutoff are excluded before running the same robust RD regressions. Here, we set w to be 0.5, 1, and 2 –

neighborhoods with potential rounding. The results are quite similar to the original ones in terms of magnitude and significance, alleviating concerns about manipulation and rounding.

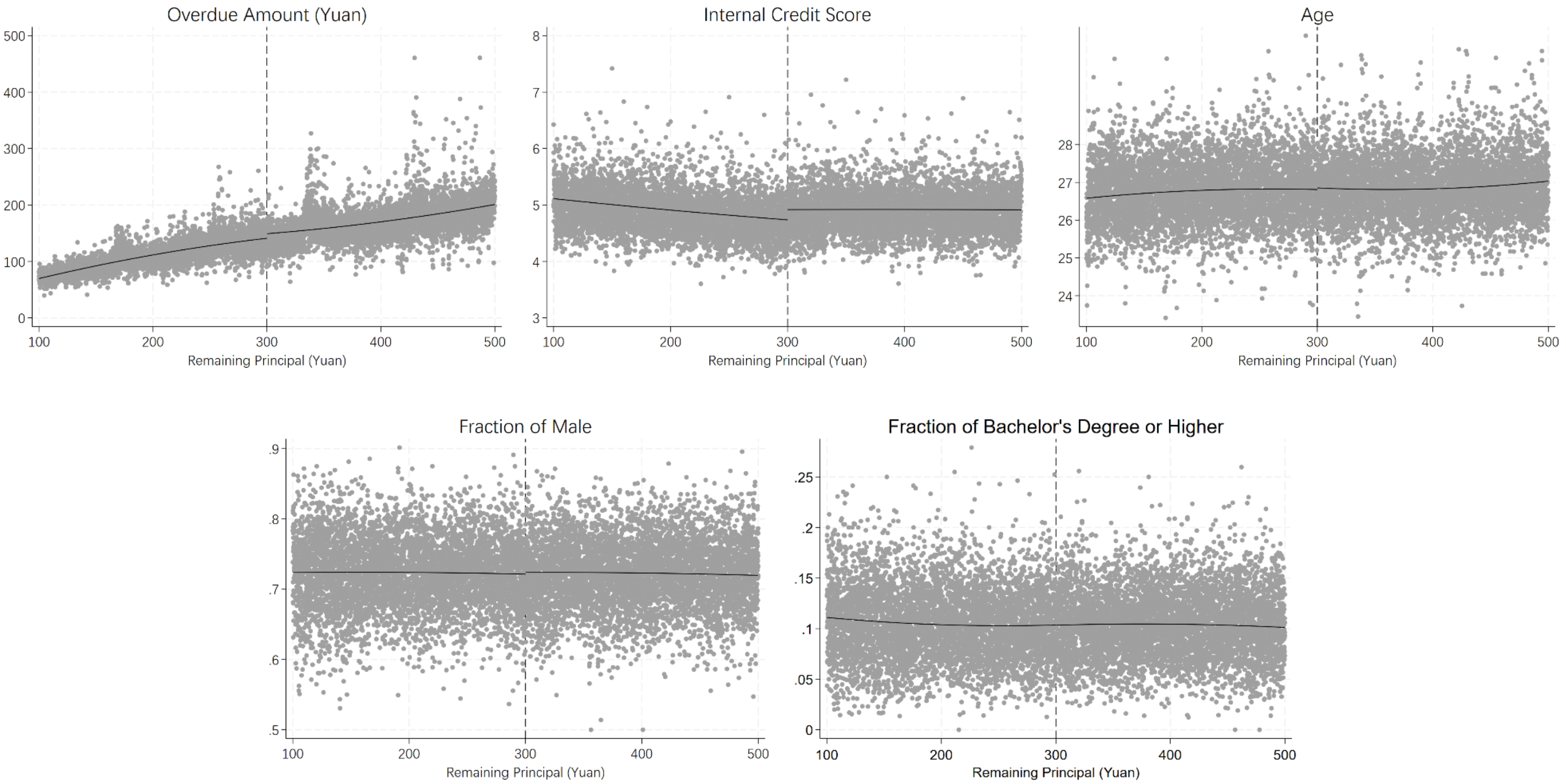
4 Placebo tests

Finally, we implement two placebo tests using artificial cut-offs of 200-yuan and 400-yuan remaining principal, as shown in Table B3. For validating purposes, we use CER-optimal bandwidths in the RD regressions, since they give the most power when making inferences about the null hypothesis that there is no jump in outcome variables (Cattaneo et al., 2019). The results do not reject the null hypothesis for both artificial cut-offs and all evaluation horizons.

Figure B1. RD plot using variance-mimicking bandwidth.

This figure shows the binned average of several variables around the principal cutoff of 300 yuan, as in Figure 4 but using variance-mimicking bandwidth. The variables of interest include loan characteristics as shown in Panel (a) and collected NPVs for various horizons, as in Panel (b). The grey dots are binned averages, and the black lines are global quadratic fits within each side. See the notes in Figure 4 for more details.

(a) Loan characteristics.



(b) Collected NPV.

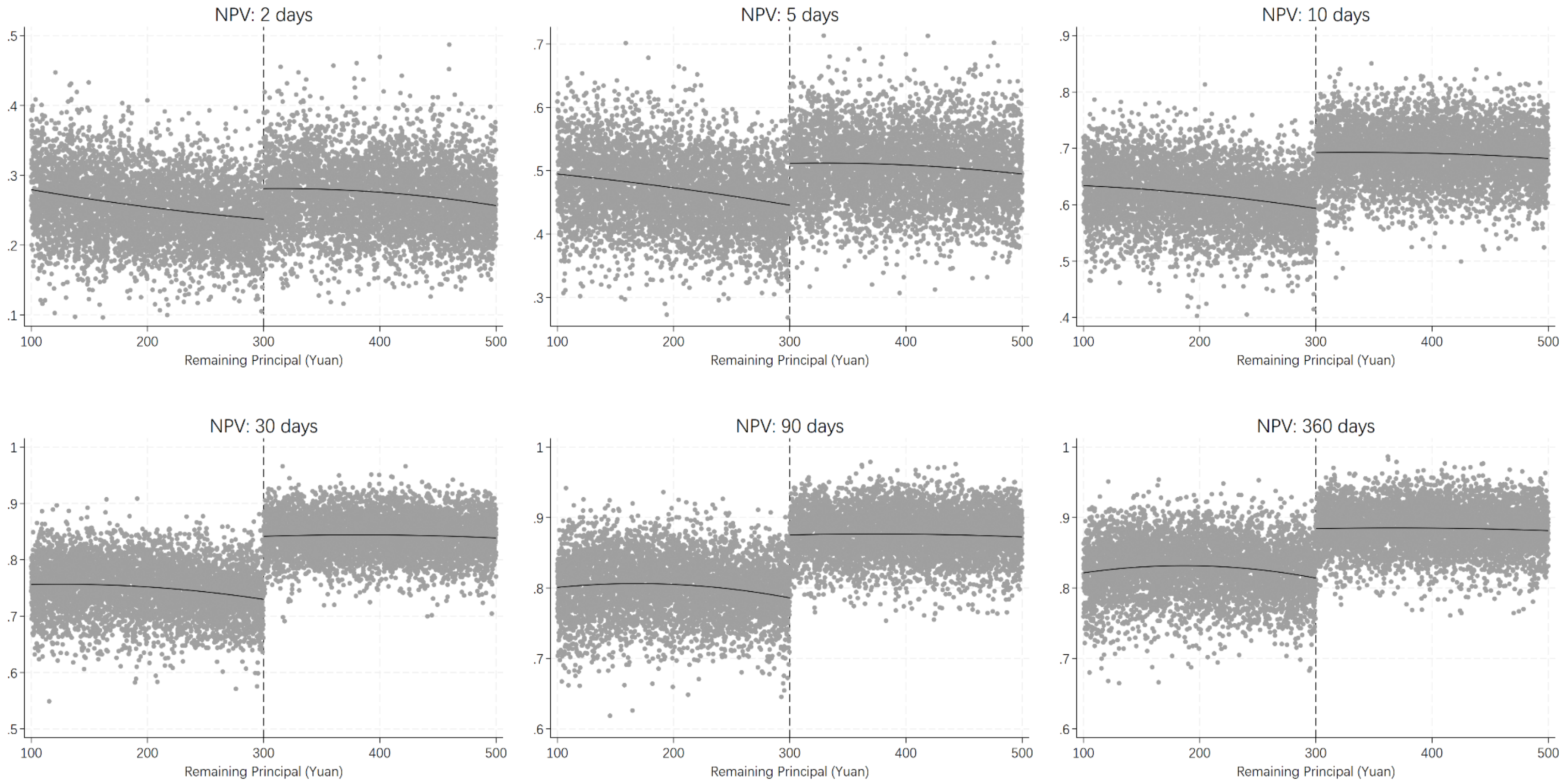
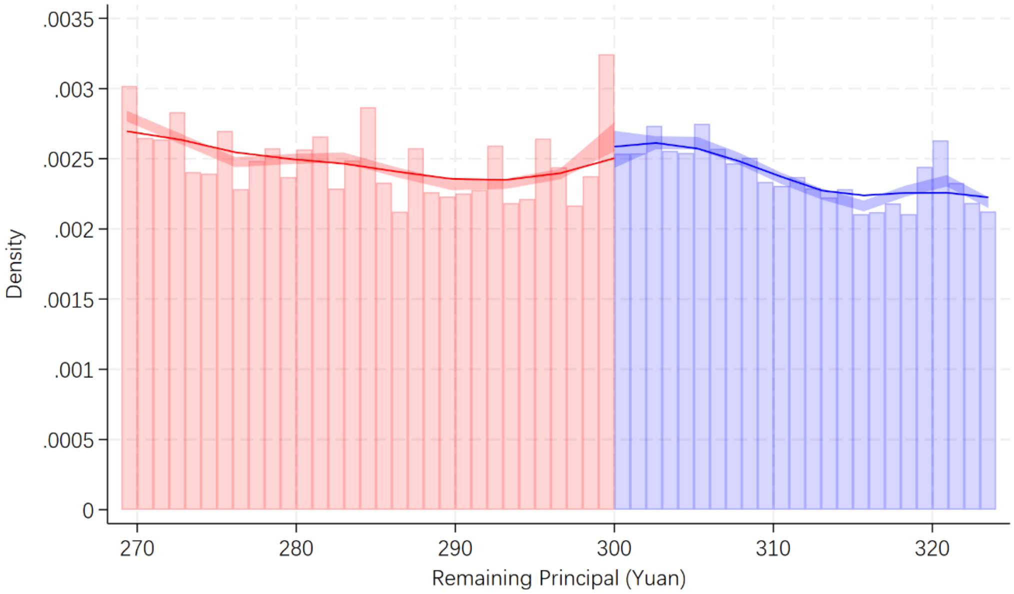


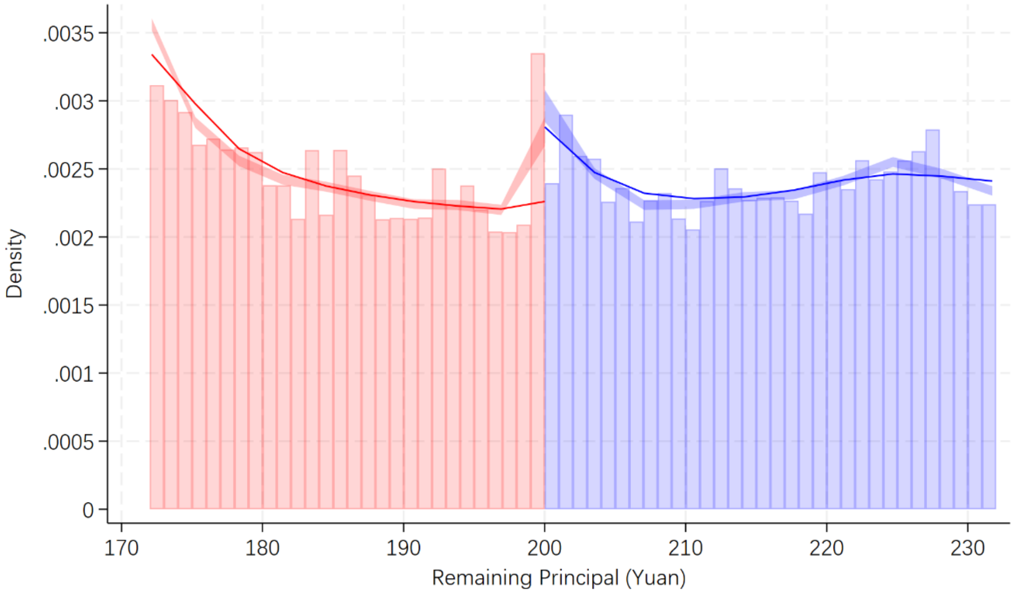
Figure B2. RD density test around the threshold.

This figure reports the results of the RD density test to detect potential manipulation around the threshold. The figure first shows the histogram of the running variable, the remaining principal, around the threshold. Each interval in the histograms includes the right end but not the left end. It then estimates the density functions on both sides of the threshold separately using local quadratic regressions, which are displayed by the solid lines. The shaded areas indicate the 95% robust RD confidence intervals using local cubic regressions. All local regressions use the triangular kernel with IMSE-optimal bandwidth.

(a) 300-yuan threshold.



(b) Placebo test: 200-yuan threshold.



(c) Placebo test: 400-yuan threshold.

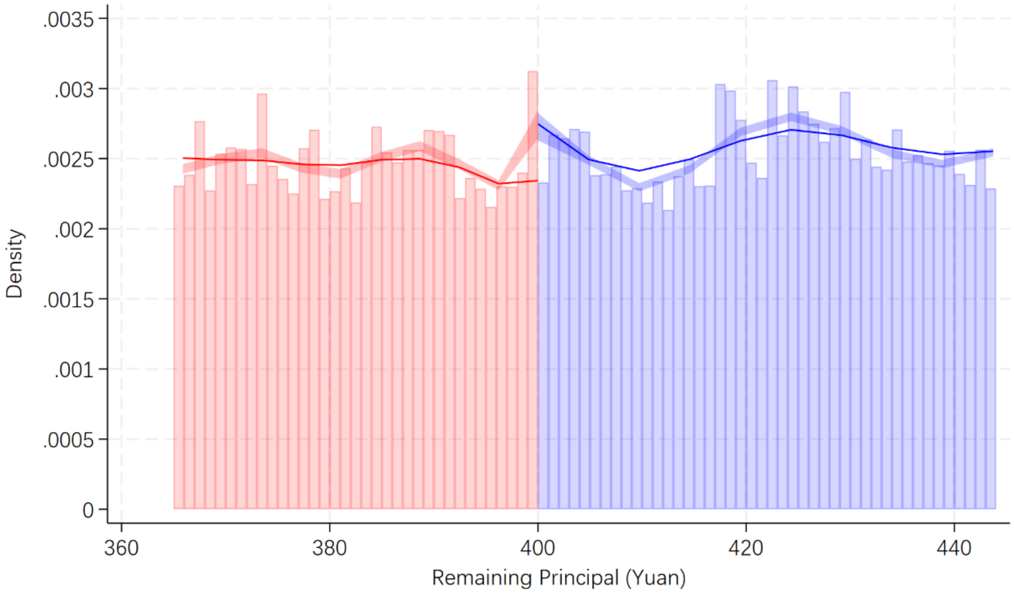


Table B1. Binomial test of manipulation at the threshold.

This table reports the results from the Binomial test of manipulation at the threshold. For a neighborhood of width 2*x* around the 300-yuan cut-off, the test counts the numbers of observations below and above the threshold and calculates the fraction of observations below the threshold. When there is no manipulation at the threshold, the null hypothesis holds that the fraction of observations below the threshold is 0.5, so the distribution of observations on both sides of the threshold can be considered as random.

Neighborhood Radius *x*	# Obs. in (300-*x*,300]	# Obs. in (300,300+*x*]	% Below	*p*-val.
0.5	2228	1296	63.2%	<0.001
1.0	3282	2589	55.9%	<0.001
1.5	4573	3791	54.7%	<0.001
2.0	5686	5157	52.4%	<0.001
2.5	6787	6633	50.6%	0.187
3.0	7878	7908	49.9%	0.818
3.5	9189	9177	50.0%	0.935
4.0	10345	10507	49.6%	0.265
4.5	11840	11646	50.4%	0.208
5.0	13020	13073	49.9%	0.748

Table B2. Collected NPV difference between AI and human callers: Robustness check.

This table implements robustness checks on the RD design regression results in Table 3 Panel B, which estimates the average difference in collected NPV between AI and human callers. The first column reports the main results, which are the same as the results in Table 3 Panel B. Columns 2 to 9 check different variations in the RD regression specifications. Columns 2 and 3 change kernel selection. Columns 4 to 6 modify bandwidth selection. Columns 7 to 9 check sensitivity to observations close to the cutoff by excluding observations within ±*w* of the cutoff, i.e., making a "donut hole" of radius *w*. "MSE" and "CER" stand for MSE- and CER-optimal bandwidth, respectively. "Epan." is short for Epanechnikov kernel. RD robust standard errors clustered by calendar month are in parentheses. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Main setup	Kernel Choice		Bandwidth Choice			Donut-Hole		
							$w = 0.5$	$w = 1$	$w = 2$
NPV2	0.0409***	0.0366***	0.0390***	0.0391***	0.0439***	0.0327	0.0525***	0.0520***	0.0511***
	(5.96)	(4.96)	(5.38)	(5.32)	(4.58)	(0.79)	(6.35)	(5.79)	(6.12)
NPV5	0.0458***	0.0552***	0.0566***	0.0568***	0.0668***	0.0541***	0.0597***	0.0536***	0.0548***
	(6.35)	(7.20)	(7.08)	(8.47)	(8.33)	(4.28)	(8.33)	(6.91)	(7.15)
NPV10	0.0916***	0.0856***	0.0868***	0.0843***	0.0999***	0.0876***	0.0979***	0.1000***	0.0977***
	(14.04)	(12.56)	(12.85)	(12.58)	(14.58)	(8.34)	(15.06)	(16.25)	(15.01)
NPV30	0.108***	0.108***	0.108***	0.106***	0.110***	0.104***	0.112***	0.112***	0.112***
	(18.98)	(18.81)	(18.61)	(17.22)	(20.36)	(13.74)	(20.56)	(19.81)	(16.69)
NPV60	0.0921***	0.0934***	0.0936***	0.0958***	0.0940***	0.0900***	0.0947***	0.0944***	0.0948***
	(17.87)	(19.71)	(18.74)	(18.77)	(21.16)	(12.27)	(17.87)	(16.98)	(15.69)
NPV90	0.0859***	0.0858***	0.0860***	0.0869***	0.0851***	0.0828***	0.0879***	0.0872***	0.0869***
	(17.67)	(17.66)	(16.86)	(17.07)	(19.11)	(13.19)	(16.82)	(15.92)	(14.87)
NPV180	0.0734***	0.0736***	0.0739***	0.0751***	0.0723***	0.0710***	0.0739***	0.0742***	0.0744***
	(16.51)	(16.22)	(15.50)	(16.01)	(17.58)	(12.81)	(16.94)	(16.61)	(14.18)
NPV360	0.0671***	0.0680***	0.0685***	0.0702***	0.0654***	0.0661***	0.0678***	0.0690***	0.0719***
	(15.25)	(15.52)	(14.85)	(15.44)	(17.34)	(10.83)	(15.53)	(15.13)	(14.12)
Bandwidth	MSE	MSE	MSE	CER	2*MSE	1/2*MSE	MSE	MSE	MSE
Kernel	Uniform	Triangular	Epan.	Uniform	Uniform	Uniform	Uniform	Uniform	Uniform

Table B3. Collected NPV differences between AI and human callers: Placebo tests.

This table reports placebo test results using artificial cut-offs of 200-yuan and 400-yuan remaining principals. The specifications of the RD regression are the same as those in Table 3 Panel B except that the CER-optimal bandwidths are used for better inference and a smaller coverage error rate. RD robust standard errors clustered by calendar month are in parentheses. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)
Artificial cutoff *c*	*c* = 200	*c* = 400
NPV2	-0.0049	-0.006
	(-0.81)	(-0.98)
NPV5	-0.0051	-0.002
	(-0.63)	(-0.37)
NPV10	0.0012	-0.009
	(-0.05)	(-1.38)
NPV30	0.0097	0.002
	(1.46)	(0.60)
NPV60	0.0025	0.000
	(0.49)	(0.35)
NPV90	0.0011	0.004
	(0.27)	(1.01)
NPV180	0.0012	0.001
	(0.34)	(0.51)
NPV360	0.0008	0.003
	(0.12)	(0.86)
Bandwidth	CER-opt.	CER-opt.
Kernel	Uniform	Uniform

Appendix C. Estimation of Unit Labor Costs

1 Caller salary scheme

For individual callers, their monthly salary consists of two components: ranking salary and completion salary. Both components are related to their monthly performance measured as the total amount of outstanding balance she collects. The ranking salary is based on their performance ranking among a group of callers who have similar tenures and workloads and work in the same stage of the debt collection process. The relationship between ranking and ranking salary is increasing and convex. As the caller moves upward in the ranking, they receive increasingly more bonuses as their ranking salary. Figure C1 shows the formula for performance ranking salary as a function of caller ranking used in May 2022 for callers in the "M1 Early" stage, which spans from day 2 to day 10 past due. The ranking salary has six tiers and the salary in each tier changes linearly. The function has the steepest slope in Tier 1. The top caller receives 5500 yuan as ranking salary while the lowest 5% of callers receive nothing. In addition, the company may divide callers into several groups and encourage competition between groups to maintain callers' morale. The winning group can receive from 100 yuan to a few hundred yuan as an extra ranking bonus.

On the other hand, the completion salary is paid according to the amount of money the caller collects scaled by a pre-specified target. Specifically, at the beginning of each month, the company sets a target collection amount for each caller based on the predicted total outstanding balance the company may need to deal with, as well as the number of callers and caller working experience. Junior callers can have a 10% lower target in their first four months with the company. At the end of the month, the target completion rate (CR) is then defined as the ratio of the actual amount collected by the caller to her target. The completion salary is then calculated as an increasing piece-wise linear function of CR. Figure C2 illustrates the relationship between completion rate and completion salary that the company applied in May 2022. The target amount was 448,526 yuan in May 2022. The completion salary jumps at the completion rate of 0.7, 0.8, 0.9, 0.95, and 1. The slope also slightly increases with the completion rate across the intervals. The callers can earn more than 3500 yuan if they finish the target and nothing if they have done less than 70%. The target is set in a reasonable way since the average completion rate is 1.01.

Finally, the company also has a minimum wage policy of around 3000 yuan per month, varying slightly on time and location of employment. If the sum of a caller's ranking salary and completion salary is below the minimum wage, the company will pay the caller the minimum wage.

Figure C3 presents the salary amount that callers in the "M1 Early" stage actually received in May 2022. Figure C3(a) plots the actual ranking salary as a function of callers' performance ranking. The shape of the curve closely tracks the formula in Figure C1, with small variations that reflect the extra bonus from group-level performance competition. Figure C3(b) shows the actual completion salary as a function of the completion rate, which precisely follows the formula in Figure C2.

Finally, Figure C3(c) presents the relationship between overdue money collected and actual total salary, which is the sum of ranking and completion salary after some adjustments that we will discuss shortly. First note that the upper "surface" of the scatter dots represents the theoretical maximum salary that the callers can receive given the amount collected. It is upward sloping above 3000 yuan, the minimum wage.[1] The slope is about 0.045 yuan salary per one yuan collected, which is the combination of the two components' sensitivity. Ranking salary contributes much more than completion salary in the slope. In practice, callers typically receive a salary below the theoretical maximum for several reasons, including penalties for absence from work or late arrivals, for example.

The most significant penalty is for the violation of rules regarding conversations with overdue borrowers. To comply with government regulations and to maintain a positive image in the public, the company imposes several rules on what the callers cannot say to borrowers. Those prohibited words include swear words, threats, discrimination, false information, unwarranted promises to borrowers, etc. The company uses an AI examiner to go through all phone call records and label any misconduct every month. For each caller in each month, the company calculates a "quality control (QC) ratio" defined as the fraction of appropriate conversations. The actual salary that a caller can receive is then the theoretical maximum scaled by the QC ratio. The average QC ratio is about 0.953, but there are also 10% of callers who have a QC ratio below 0.87.

This quality adjustment helps explain an interesting fact that, although there is a jump in completion salary at 100% completion rate (which corresponds to about 450,000 yuan of money

[1] Some callers received salaries lower than the minimum wage because they left the company in the middle of the month. Their actual salary was not protected by the minimum wage policy, and may even be deducted as penalties.

collected), the total salary has no significant jump at the cut-off. This is because, to finish the target, callers just below the threshold tend to use more improper expressions. Therefore, in spite of receiving a jump in completion salary for crossing the threshold, they are penalized by a low QC ratio, leading to their receiving total salaries that are not discretely higher than if they were just below the threshold.

This type of manipulation can be seen in Figure C4. Figure C4(a) conducts an RD density test on the completion rate at the 100% cut-off. As the test suggests, there is a significant bunching of callers just above 100%. The *t*-statistic is 2.26, which is significant at 5% level. Figure C4(b) shows an RD plot of the QC factor with respect to completion rate around the 100% cut-off. The RD plot indicates a significant quality decline just above the cut-off, consistent with the manipulation that trades off between quality and quantity.

2 Estimation of unit labor costs

Although both salary components depend on performance for each individual caller, the total ranking salary should be considered as a fixed labor cost for the company that is unrelated to the absolute amount of money callers collect in total. Specifically, no matter how much money each caller collects, the sum of the ranking salary is the same for the company. Only the distribution of the ranking salary varies by individual caller's performance. To estimate unit fixed salary costs, we first divide the total ranking salary by the total time (in days) that all callers work for. We do this for callers in the "M1 Early" stage, who are the major group of callers specialized in cases in the first five days. They are also the callers who are most likely to be replaced by the AI callers, so their salary level gives a reasonable estimate of the AI cost-saving. Figure C5(a) shows the average fixed salary expenditure over time, which suggests that the company on average pays 75.94 yuan per worker per day as fixed salary costs. Next, we calculate callers' average workload in terms of their assigned overdue amount in a similar way and show the results in Figure C5(b). The figure suggests that on average each caller is assigned around 170,000 yuan of outstanding balance per day. Dividing the average fixed salary by the average workload, we calculate the average fixed costs that the company needs to spend for handling each yuan of overdue money. Figure C5(c) shows the time series of this ratio over time, which suggests that the average fixed salary costs to handle one yuan of outstanding balance, i.e., the ratio between average ranking salary and average workload, is approximately 0.00045 yuan. Here, we use the time-series average for further calculations since our performance evaluation horizons may cover several months, and given that the number is quite stable over time, the time-series average gives a reasonable estimation.

The completion salary can be viewed as the variable labor costs for the company as it is related to the actual amount of money collected. To get the average variable salary that the company has to pay for every one yuan collected, we divide the total completion salary by the total amount of money collected in each month. Figure C5(d) reports the results. The average completion salary displays an increasing trend. Specifically, the average salary was 0.004 yuan in June 2021 and was raised by 75% to 0.007 yuan in May 2022. Again, since our evaluation horizons may span several months, we use the time-series average, which is about 0.0051 yuan per one yuan collected, for labor cost adjustment.

Figure C1. The relationship between ranking and ranking salary.

This figure visualizes the formula used in May 2022 to calculate an individual caller's ranking salary as a function of their performance ranking. The caller is ranked by their total money collected in a month within a group of callers in the same stage of debt collection and with similar tenure. The horizontal axis represents percentage ranking in descending order.

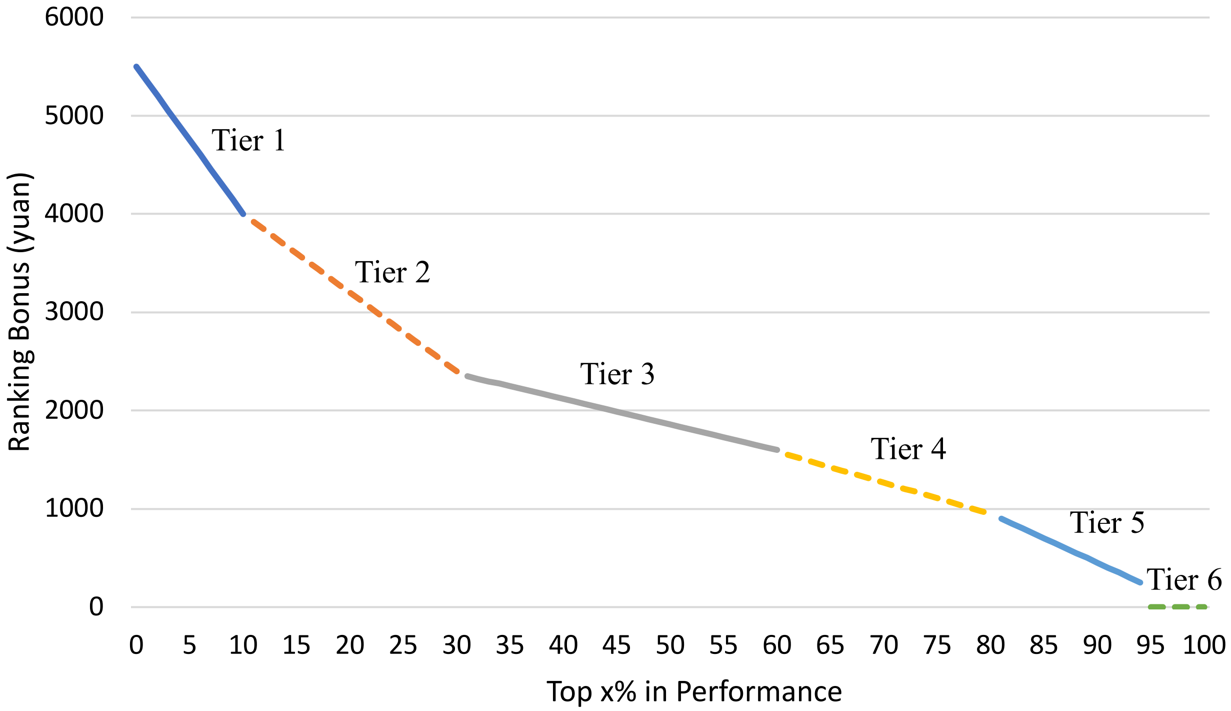


Figure C2. The relationship between completion rate and completion salary.

This figure visualizes the formula used to calculate an individual caller's completion salary as a function of their target completion rate in May 2022. The target completion rate is defined as the ratio of money collected in the month to a target of money to be collected specified by the company at the beginning of the month. The target amount was 448,526 yuan in May 2022.

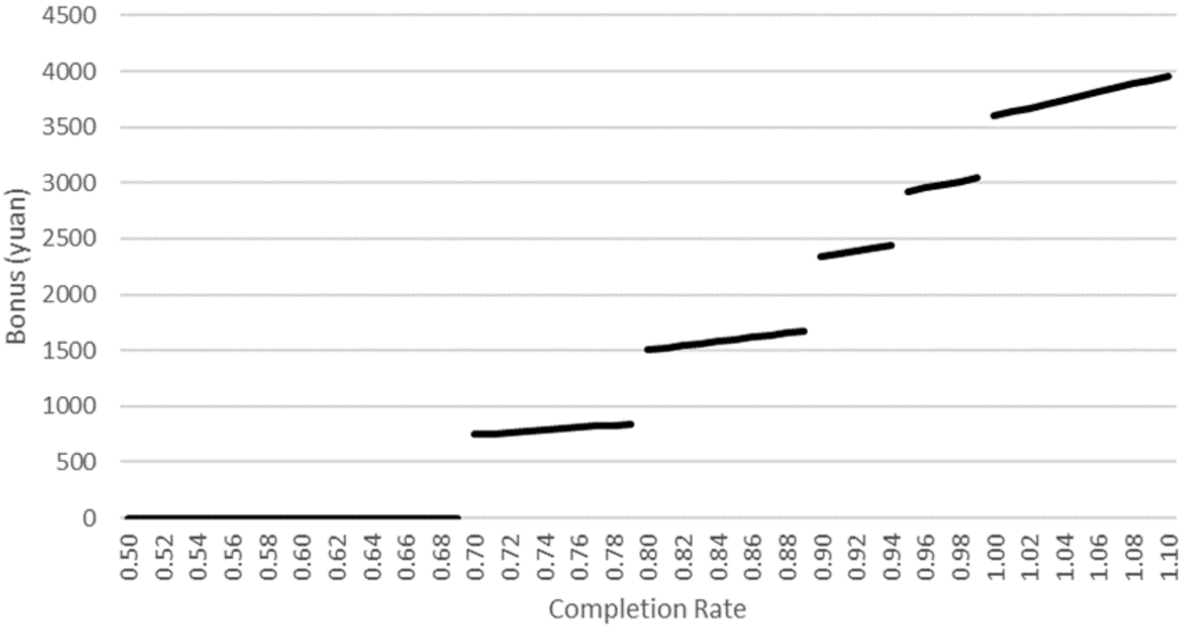
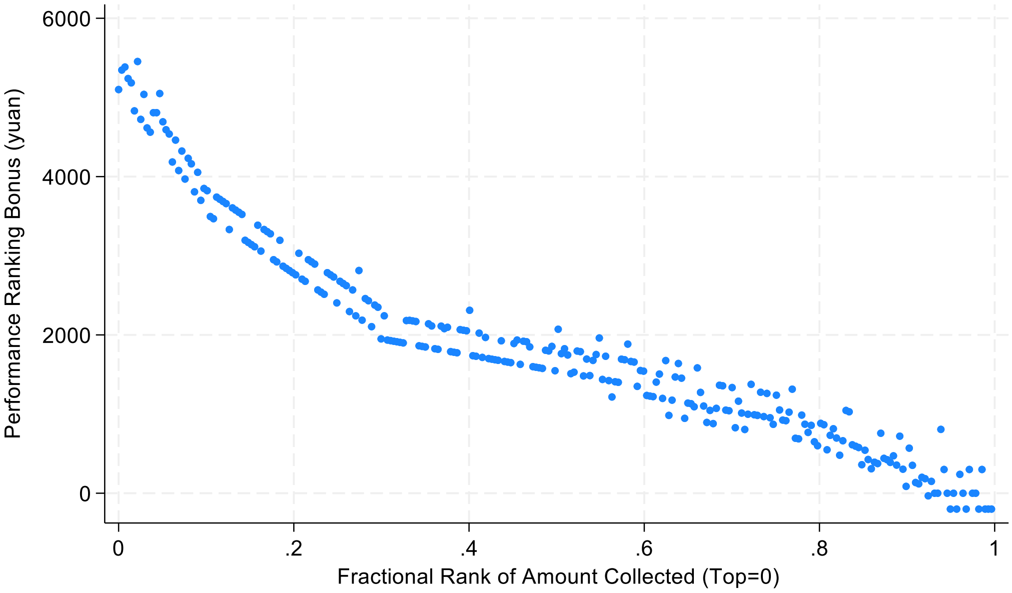


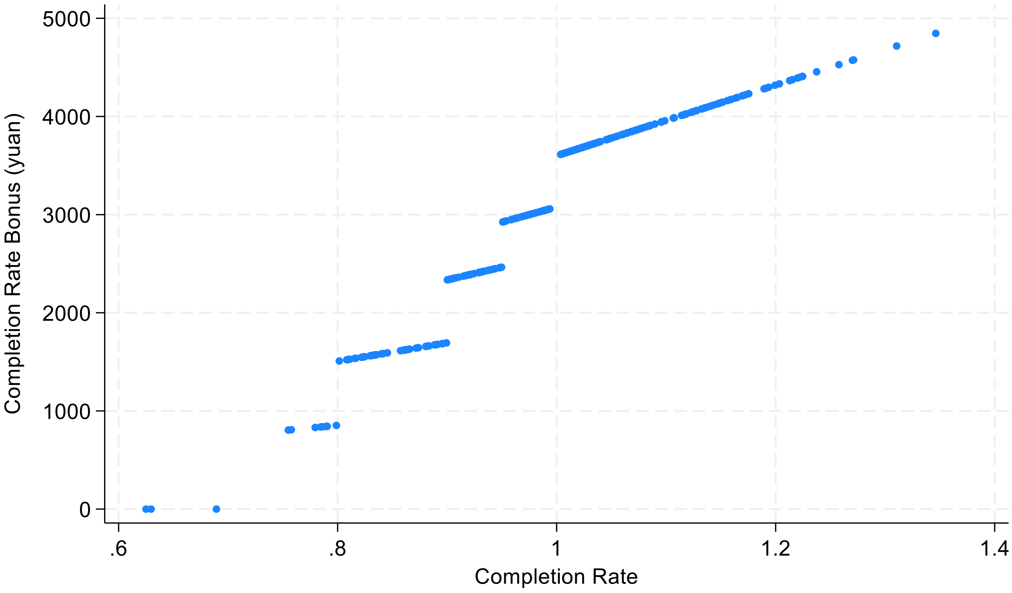
Figure C3. Actual salary received by callers.

This figure reports the actual salary received by senior callers in the "M1 Early" stage in May 2022. Panel (a) reports the actual ranking salary received by callers as a function of caller performance ranking. Panel (b) shows the actual completion salary received by callers as a function of their completion rate. Panel (c) shows the actual total salary received by callers as a function of the amount of money collected. The total salary is the sum of the ranking salary and completion salary, capped by the minimum wage, and adjusted for additional penalties and bonuses.

(a) Ranking salary as a function of ranking.



(b) Completion salary as a function of completion rate.



(c) Total salary as a function of money collected.

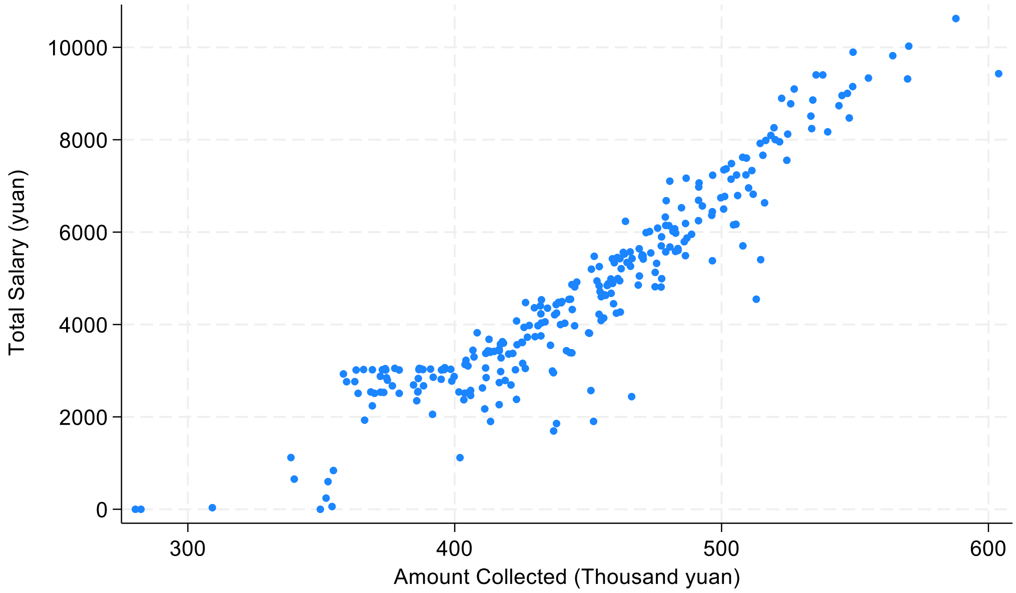
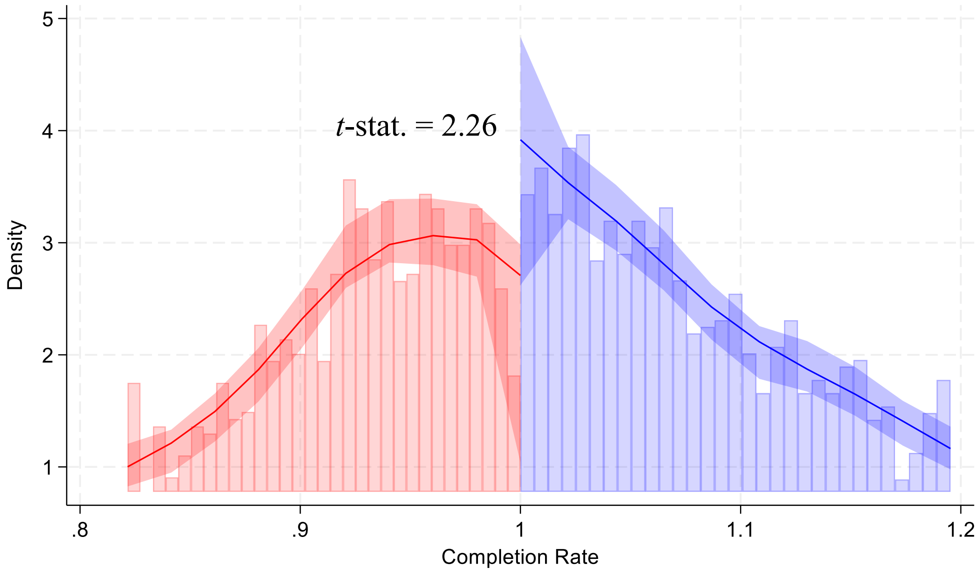


Figure C4. Manipulation of completion rate at the 100% cut-off.

This figure illustrates the manipulation of the completion rate by callers just below the 100% completion rate cut-off. Panel (a) implements an RD density test on the distribution of the completion rate around the cut-off. The bars show the histogram and the solid lines show the density functions on each side of the cut-off estimated by local quadratic regressions. The shaded areas are the 95% robust RD confidence intervals. Panel (b) shows an RD plot of the "QC ratio" around the cut-off. The "QC ratio" measures the fraction of conversations that comply with the regulation for each caller. The RD estimator with robust *z*-statistics is also reported.

(a) RD density test at the cut-off of 100% completion rate.



(b) RD plot of "QC ratio" at the cut-off of 100% completion rate.

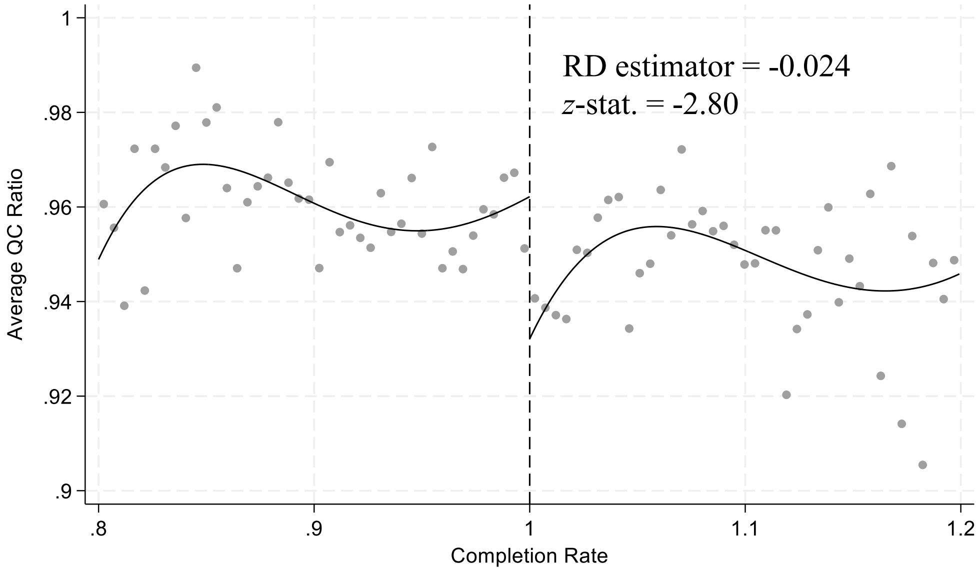
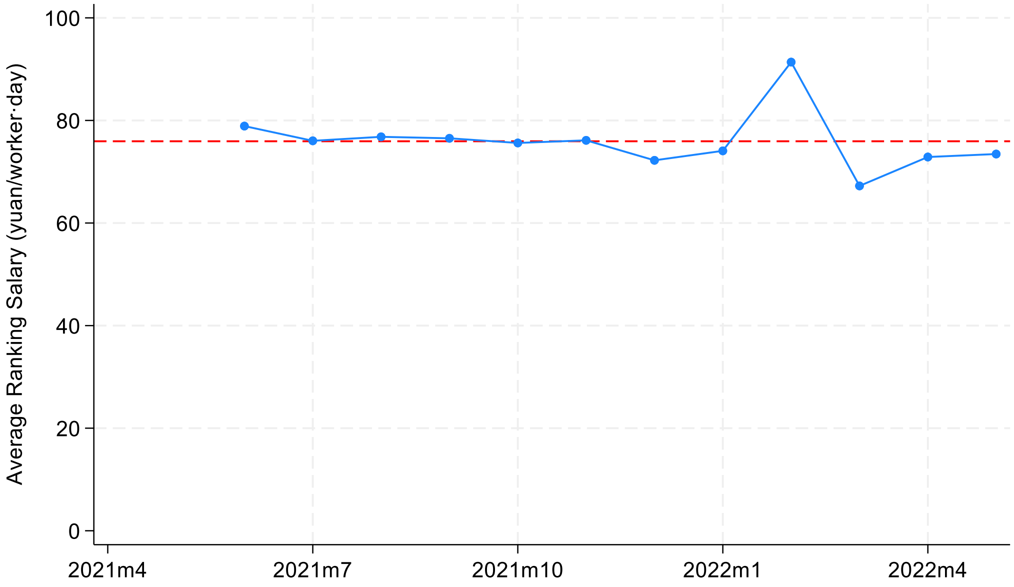


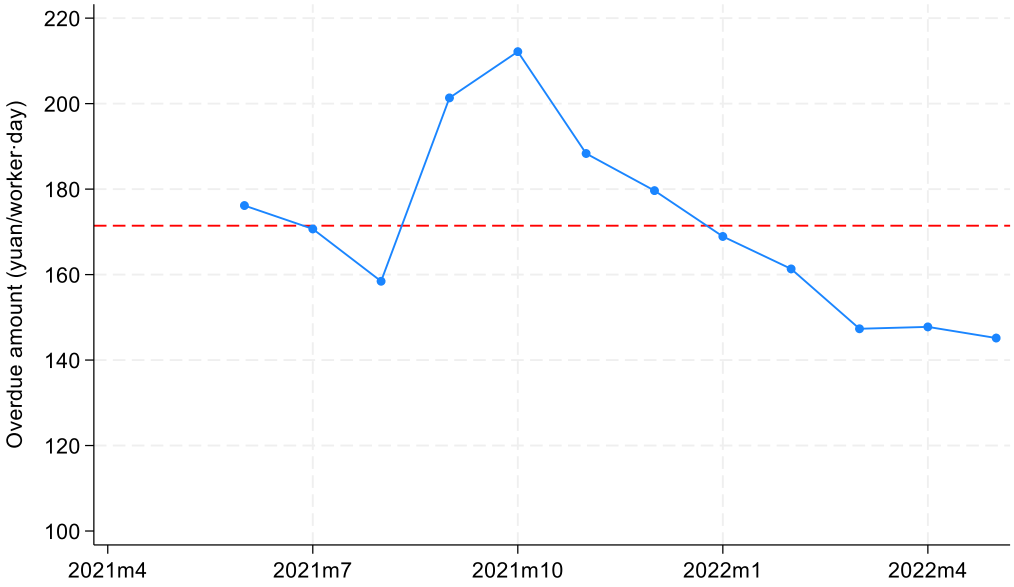
Figure C5. Time series of unit salary costs and unit workloads.

This figure shows the time series of unit salary costs and unit workloads among senior callers in the M1 Early stage. Panel (a) reports the average ranking salary per worker per day, which is the ratio of the sum of all callers' ranking salaries, divided by the total worker-days. Panel (b) is the average workload defined as the average overdue payment amounts assigned to each caller a day. Panel (c) shows the ratio of the first two statistics, which can be interpreted as the ranking salary paid by the company for every one-yuan outstanding overdue money every day, or equivalently, the unit fixed labor costs. Panel (d) reports the ratio between the total completion salary and the total money collected, that is, the unit variable labor costs.

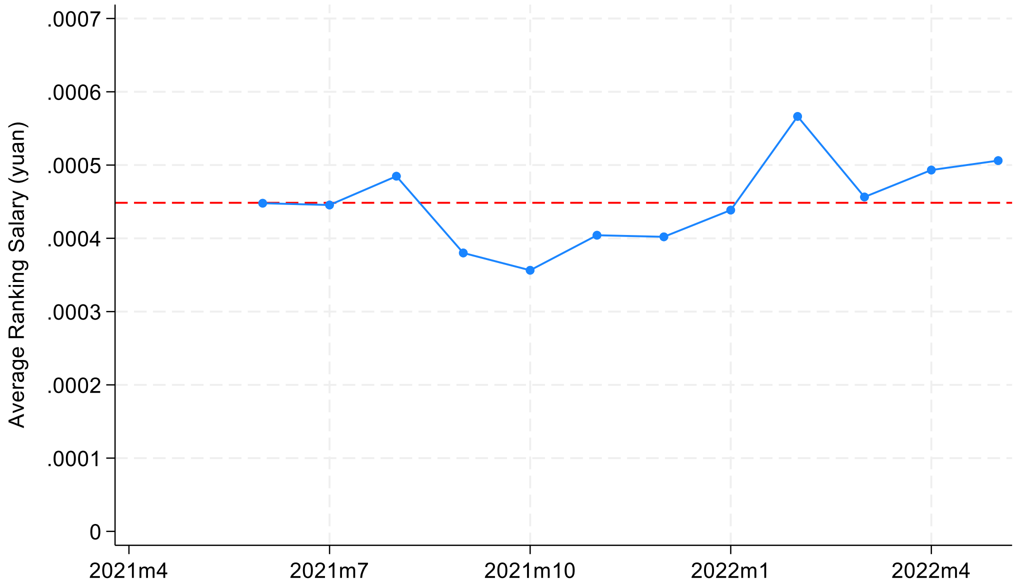
(a) Ranking salary per worker per day



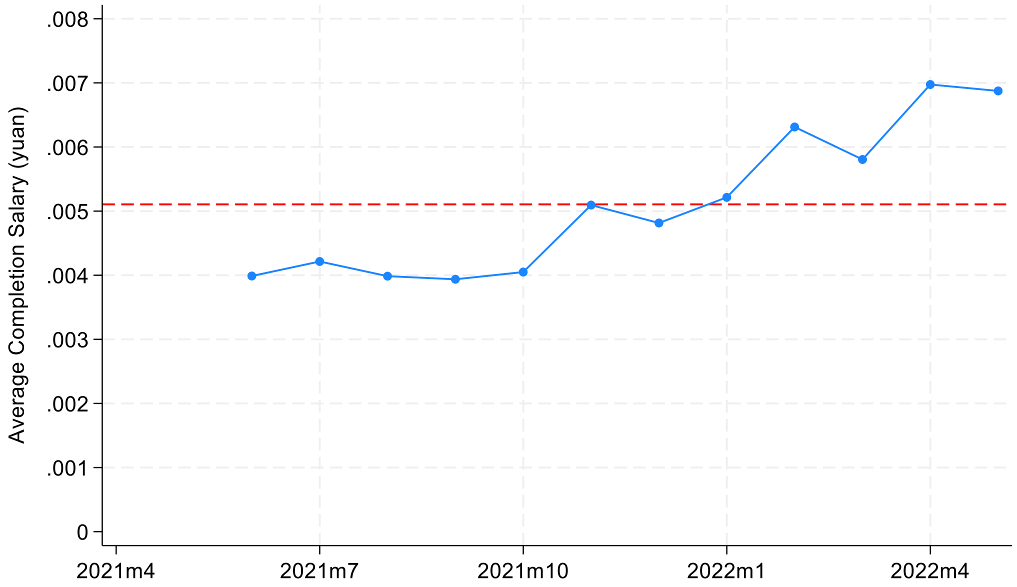
(b) Overdue money assigned per worker per day



(c) Ranking salary per one-yuan outstanding balance per day



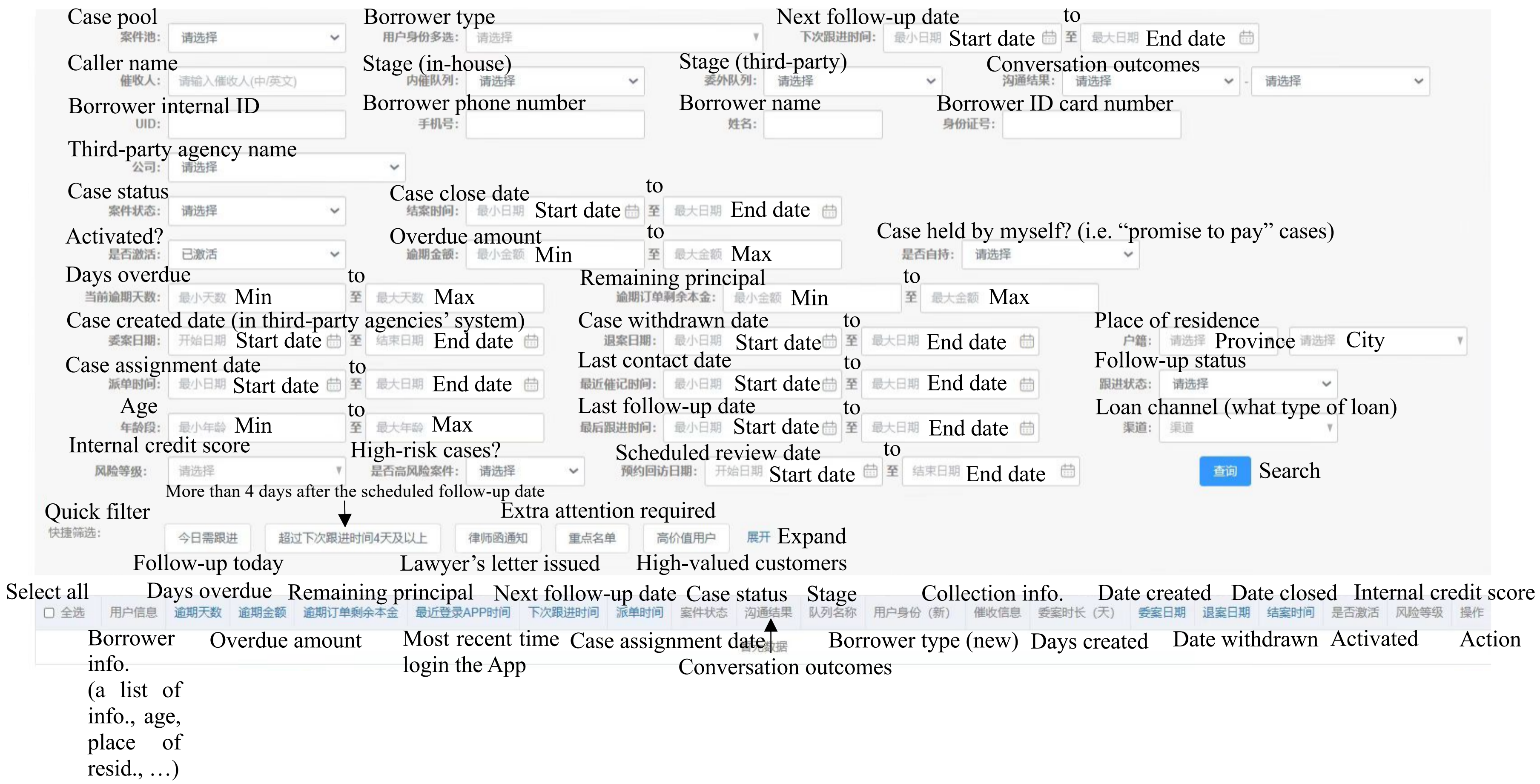
(d) Completion salary per one-yuan money collected



Appendix D. Additional Figures and Tables

Figure D1. A snapshot of the debt collection system interface.

This figure shows what a caller can see on their screen when they login to the company's system and work on the assigned cases. The upper part is a filter with many conditions that the caller can modify. The lower part lists all cases assigned to the caller that meet the filtering criteria. The system is in Chinese, and English translations are provided next to the corresponding Chinese words.



Case pool 案件池: 请选择 | Borrower type 用户身份多选: 请选择 | Next follow-up date 下次跟进时间: 最小日期 Start date to 至 最大日期 End date

Caller name 催收人: 请输入催收人(中/英文) | Stage (in-house) 内催队列: 请选择 | Stage (third-party) 委外队列: 请选择 | Conversation outcomes 沟通结果: 请选择 - 请选择

Borrower internal ID UID: | Borrower phone number 手机号: | Borrower name 姓名: | Borrower ID card number 身份证号:

Third-party agency name 公司: 请选择

Case status 案件状态: 请选择 | Case close date 结案时间: 最小日期 Start date to 至 最大日期 End date

Activated? 是否激活: 已激活 | Overdue amount 逾期金额: 最小金额 Min to 至 最大金额 Max | Case held by myself? (i.e. “promise to pay” cases) 是否自持: 请选择

Days overdue 当前逾期天数: 最小天数 Min to 至 最大天数 Max | Remaining principal 逾期订单剩余本金: 最小金额 Min to 至 最大金额 Max

Case created date (in third-party agencies' system) 委案日期: 开始日期 Start date to 至 结束日期 End date | Case withdrawn date 退案日期: 最小日期 Start date to 至 最大日期 End date | Place of residence 户籍: 请选择 Province 请选择 City

Case assignment date 派单时间: 最小日期 Start date to 至 最大日期 End date | Last contact date 最近催记时间: 最小日期 Start date to 至 最大日期 End date | Follow-up status 跟进状态: 请选择

Age 年龄段: 最小年龄 Min to 至 最大年龄 Max | Last follow-up date 最后跟进时间: 最小日期 Start date to 至 最大日期 End date | Loan channel (what type of loan) 渠道: 渠道

Internal credit score 风险等级: 请选择 | High-risk cases? 是否高风险案件: 请选择 | Scheduled review date 预约回访日期: 开始日期 Start date to 至 结束日期 End date | 查询 Search

Quick filter 快捷筛选: 今日需跟进 Follow-up today | 超过下次跟进时间4天及以上 More than 4 days after the scheduled follow-up date | 律师函通知 Lawyer's letter issued | 重点名单 Extra attention required | 高价值用户 High-valued customers | 展开 Expand

□ 全选	用户信息	逾期天数	逾期金额	逾期订单剩余本金	最近登录APP时间	下次跟进时间	派单时间	案件状态	沟通结果	队列名称	用户身份（新）	催收信息	委案时长（天）	委案日期	退案日期	结案时间	是否激活	风险等级	操作
Select all	Borrower info. (a list of info., age, place of resid., …)	Days overdue	Overdue amount	Remaining principal	Most recent time login the App	Next follow-up date	Case assignment date	Case status	Conversation outcomes	Stage	Borrower type (new)	Collection info.	Days created	Date created	Date withdrawn	Date closed	Activated	Internal credit score	Action

暂无数据

Figure D2. Distribution of phone call duration.

This figure presents a histogram of phone call durations for all first answered phone calls by borrowers on day 2 past due in the completely randomized subsample. The phone call durations are in seconds and each bin is 10-second width.

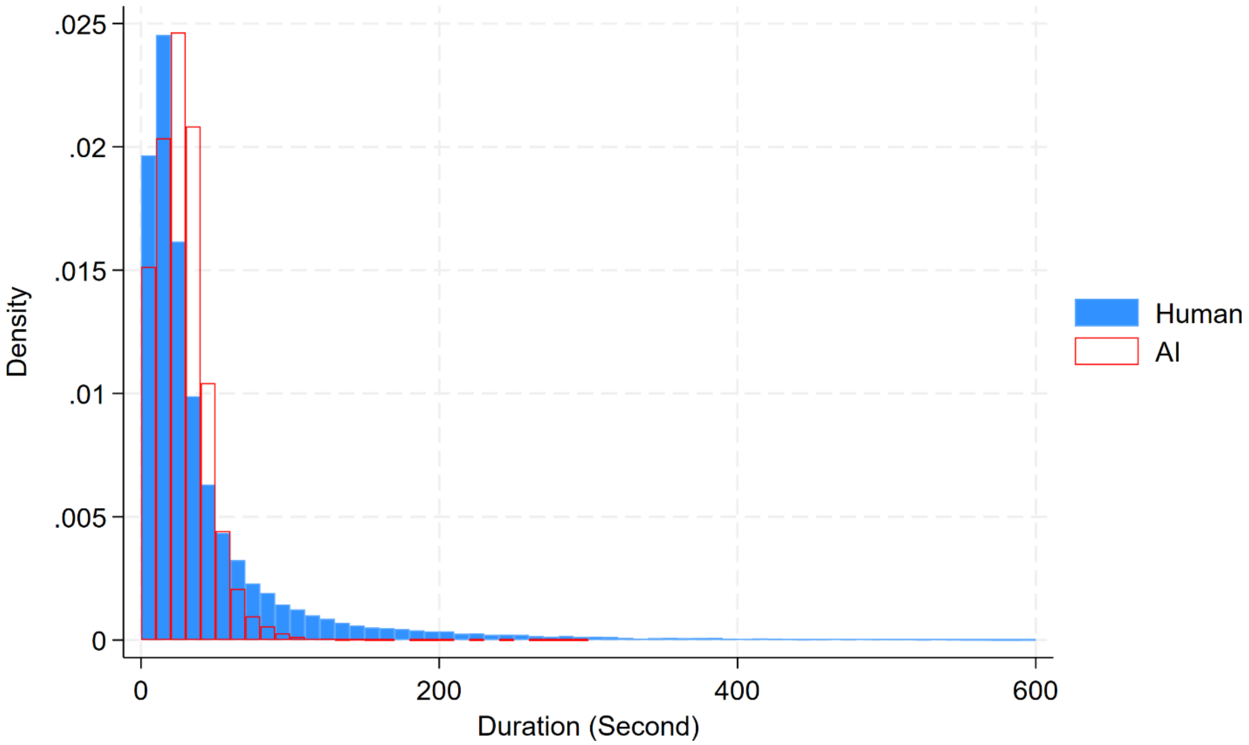


Table D1. The relationship between case assignment, caller turnover, and performance ranking across human callers.

This table examines the randomization of case assignments across callers on day 6 and the potential attrition bias with respect to callers' previous performance ranking. The sample is the same as in Table 7. For case assignment tests in columns 1 to 8, the regressions are at the case level. The dependent variables are observable information about the assigned cases, including the indicator of whether the cases are treated by AI V3 in the first five days, AI outcomes on day 5 (NPV5), and loan characteristics as in Table 4. The independent variable is the assigned caller's previous performance ranking (*PrevPerfRank*) as defined in Section 5.1. For caller turnover tests in columns 9 and 10, the regressions are at the caller level for existing callers. The dependent variable is an indicator of promotion to later stages or an indicator of leaving the company in the *next* month. The independent variable is caller performance ranking in the *current* month. Cluster-adjusted *t*-statistics clustered at the caller level are reported in parentheses. Significance level: *** *p*<0.01, ** *p*<0.05, * *p*<0.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	AI V3 indicator	NPV5	Overdue amount	Remaining principle	Internal credit score	Age	Male	Bachelor's degree or more indicator	Promotion next month	Leave next month
Prev. Perf. Ranking	-0.009	-0.008	-89.54	219.0	-0.145	0.234	-0.025	0.011		
	(-0.41)	(-1.02)	(-1.04)	(0.59)	(-1.04)	(0.70)	(-1.06)	(0.72)		
Perf. Ranking									0.023	-0.017
									(1.15)	(-0.35)
Constant	0.641***	0.057***	1,776.8***	9,719.1***	5.490***	26.83***	0.731***	0.094***	0.011	0.172***
	(85.69)	(20.23)	(50.39)	(82.03)	(124.3)	(233.0)	(95.23)	(19.18)	(1.04)	(6.58)
No. of Obs.	4,232	4,232	4,232	4,232	4,232	4,232	4,232	4,232	348	417
R-squared	0.096	0.002	0.001	0.000	0.000	0.002	0.000	0.001	0.035	0.011