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

May 6, 2024

Abstract: How good is artificial intelligence (AI) at persuading humans to perform personally costly actions that carry a moral valence? 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 that feel binding may contribute to the performance gap.

We thank Manlin Sun for excellent research assistance and a seminar audience at Yale for helpful comments.

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 extensive literature studying the performance and economic impact of AI in routine tasks and prediction tasks.[2] In this paper, we study the effectiveness of AI in a different sort of task: persuading a human to perform a personally costly action that carries a moral valence. Many service and managerial jobs require performing this type of task—for example, coaxing a colleague to exert extra uncompensated effort for the good of his team, or inducing a customer to make a sacrifice like switching airplane seats as a courtesy to others. Even in circumstances where the action in question is not necessarily a moral obligation, an effective persuasion tactic can be to create or make more salient the perception that such an obligation exists (Cialdini, 2001).

Humans may be more effective than AI at leveraging moral obligation because humans are agents to whom moral duties are owed while machines are not (Melo, Marsella, and Gratch, 2016; Petisca et al., 2020), and being observed committing a moral transgression by another person is more aversive than being observed by a machine (LaMothe and Bobek, 2020; Kim et al., 2023). On the other hand, an advanced AI may be perceived as a being with sufficient moral weight to activate feelings of obligation that are comparable to what a human would evoke. As one reader of the *Wall Street Journal* wrote to the newspaper, "It is not OK to abuse chatbots or virtual assistants" (Gallegos, 2024).

The specific task we study is persuading delinquent consumer borrowers to repay their debt. Contact from a debt collector is a common experience; in 2022, 26% of U.S. adults with a credit bureau record had debt in collections.[3] The job of a debt collector is non-routine, requires social interaction, and is aided by emotional intelligence. In addition, repaying one's debts is usually seen as a moral obligation (Guiso, Sapienza, and Zingales, 2013; Bursztyn et al., 2019).

---

[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 assist humans in writing tasks in an experiment, and Brynjolfsson et al. (2023) about generative AIs in the customer service industry.

[3] https://apps.urban.org/features/debt-interactive-map/?type=overall&variable=totcoll (accessed April 30, 2024).

We use debt collection data from a leading online consumer finance company in China that makes uncollateralized installment loans. Borrowers who fail to make their monthly payments 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 impact of AI on the company's profits and worker productivity. The AI callers can understand the borrower's speech and generate appropriate voice replies. They provide borrowers with basic information, answer questions, and inform them of the negative consequences of defaulting. An important intermediate goal that both AI and human callers try to achieve is to extract a verbal promise to pay from the borrower.

We identify the relative effectiveness of AI 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 each month, the company takes a random 10% of newly delinquent debts with remaining principal greater than 300 yuan and 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 days past due 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 that of human callers. The gap slowly narrows afterward but remains around 7 percentage points even after one year past due. In addition, the gap is larger for borrowers with lower credit scores. A very stylized model of debt collectors would characterize their task as merely providing reminders and

information to borrowers, and perhaps imposing nuisance costs as well—things that AI can do nearly as well as humans. The persistent gap in performance between AI and humans and its heterogeneity by credit quality suggest that this stylized model is importantly incomplete.

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 on day 10 and 1 percentage point on 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. There may be something uniquely damaging about being contacted by AI. Repayment reductions resulting from initial contact by a less effective human call (calls on the weekend or by an inexperienced human caller) are mitigated within a few days.

We explore the potential sources of the AI performance gap by examining detailed outcomes of phone conversations in the randomized experiment sample, restricting 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. Conditional on making a promise to repay, borrowers are less likely to keep such a promise when it is made to an AI. Therefore, AI callers appear to be worse than humans at eliciting promises and creating a perceived moral obligation 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. Each upgrade was rolled out gradually 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 a 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 a displacement effect on labor.

Finally, we address how much labor costs are 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. Although the productivity deficit of AI is diminished once labor costs are accounted for, AI remains less cost-effective than human callers, especially for larger debts. 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. AI is relatively more effective at collecting from borrowers with high credit scores, but having human callers specialize on more challenging cases is not necessarily a winning strategy. We find that human callers who are, by chance, given 2 percentage points more low-credit-score borrowers this month are 3.7 percentage points more likely to quit within the next three months.

Our paper is closely related to the literature on the impacts of automation on labor. Previous studies find different impacts in different waves of automation.[4] They mostly find complementarity 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 all or none of the 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 by themselves, and they do not make humans more productive when working in tandem.

Additionally, our study is related to literature on 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

[4] 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).

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 an emerging literature on delinquent debt collection, a function that directly affects many individuals around the world. Drozd and Serrano-Padial (2017) and Fedaseyeu (2020) examine how variation in debt collection effectiveness driven by information technology and regulations affects credit supply. Fedaseyeu and Hunt (2015) study reputation concerns in using third-party debt collection. Our paper is closely related to Laudenbach and Siegel (2023), who address the importance of personal communication in collecting loan repayments. They show that phone calls to late borrowers from bank agents are more effective than mail reminders, and bank agents with more likeable voices are especially effective.

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 our data and Section 3 specifies our experimental setups. Section 4 evaluates the performance gap between AI and human callers. Section 5 examines the interaction 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) of outstanding loan balances. The company's main business is to originate loans to online consumers. The company targets young consumers with a short credit history but large income and consumption growth potential.[5] It operates its own online shopping platform and collaborates with third-party online retailers to promote consumption and offer loans at the point of sale.

The 10th percentile loan size is only 8 yuan (1 USD) and the 90th percentile is around 5,500 yuan (770 USD). The company provides two types of loans at the point of sale. The first is an

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

uncollateralized personal installment loan, which the consumer repays in equal monthly installments over the next six months to three years. The second is a credit-card-like product. Consumers may apply for a credit line, which is around 7,500 yuan (1,050 USD) on average, and pay for their online order with it. Repayment of the entire "credit card" balance is due monthly. "Credit cards" are typically used for small payments while installment loans are preferred for expensive purchases. Since the company's customers are typically riskier than the population average, the interest rates are mostly 24% per annum, which is the upper limit allowed by Chinese regulators. The borrower is asked to list two "emergency contacts" as their "guarantors," who are usually their parents, family members, or colleagues.[6]

Each borrower is assigned a monthly repayment due date, which may be changed by the borrower with the company's approval. Changing the due date frequently is not allowed. Ten days before the due date, borrowers receive a bill stating the amount of money they have to repay by the due date. Payments can be made with debit cards or mobile payment accounts, such as AliPay and WeChat Pay. Borrowers may also set up auto-payment.

Borrowers who fail to pay what is owed by the due date enter the debt collection process, described in the next subsection. During delinquency, extra interest accrues on the overdue amount. Borrowers need to repay both the overdue amount and the accrued interest to fully resolve their delinquency. Borrowers who remain delinquent for ninety days are considered to have defaulted and are reported to third-party credit report aggregators. Defaulted borrowers cannot borrow from the company again and may experience difficulties when trying to borrow from other consumer finance companies. Defaulting may also affect the borrower's ability to use rideshares and book hotels, since some large companies also use credit records for screening. If the company can prove that the borrower is not repaying despite having enough money, 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 People's Court of The People's Republic of China and prohibited from expensive consumption such as traveling by plane and purchasing real estate properties and luxury cars.

[6] These "emergency contacts" do not have any legal obligation to repay the loans if the borrower defaults. 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 to repay.

1.2 Debt collection process

The company treats the first day past due as a grace period. It generally does not call the delinquent borrowers on that day and just sends them reminders through text messages and phone app notifications. If the debts remain unpaid on the second day past due, the company starts calling.

The company uses different strategies depending on whether the debt is 2-10, 11-25, 26-59, 60-84, or 85+ days past due. AI callers may be used on days 2-5 to replace human callers. Cases assigned to human callers are rotated among callers at a daily frequency during days 2-10. An individual caller is randomly assigned 100-200 borrowers each day. There are three blocks of time when borrowers who have not repaid or promised to repay their debt are called automatically by the system: 9 to 9:30 A.M., 3 to 3:30 P.M., and 7:15 to 7:35 P.M. Callers sit in front of the screen and wait for a call to be answered.[7] All borrowers are called at least three times a day, regardless of whether they answer the phone.

Outside these three automatic call time blocks, human callers can choose which borrowers to call based on debt characteristics shown on their screen during their working hours from 9 A.M. to 8 P.M. The main information presented includes debt characteristics (days and amounts overdue, remaining principal, loan type), borrower information (age, place of residence, education level, internal credit score), the most recent time the borrower logged into the app, the time until scheduled follow-up, previous callers' tags of the borrower (e.g., "answered normally," "never answered," "intentional default," "stop collection"), past communication results, etc. According to the company's internal research, the filter used most by productive callers is the most recent time that the borrower logged into the app. The research also suggests that case selection skills play a minor role in explaining human callers' performance. To prevent the company's phone numbers from being blacklisted by borrowers, the company uses multiple phone numbers, and the caller can choose which one will be displayed on borrowers' caller ID.

During a phone call, callers usually provide information about the loans, inform the borrower of the potential negative consequences of delinquency, and try to persuade them to repay the debts as soon as possible. Callers are provided some conversation templates but are not asked to follow them strictly. Callers may provide suggestions to borrowers, such as encouraging them to ask

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

family members for help. Given that phone calls in days 2-10 only last for about 1 minute on average, these suggestions are typically short and generic. In later stages, the conversations are more personalized and specific.

After each phone call, the caller is required to label its outcome. If calls to a phone number are not answered multiple times, it may be labeled as potentially invalid. The answering rate is only 24%. If in a conversation, the borrower clearly and explicitly states that they will repay the amount due immediately or no later than the end of the next day, the caller will label it as "promise to pay," and the case will be kept by the same caller for one more day. Repayment will be credited to the caller as long as it comes within the promised time; otherwise, the case will be assigned to another caller. Therefore, callers have an incentive to ask borrowers to make promises. The company uses an AI conversation examiner to make sure that false labels of "promise to pay" are penalized by salary deductions.

The AI examiner also checks for other caller behaviors that violate rules set by the company or the regulator. Callers are not allowed to use profanity, make threats, discriminate, provide false information, or make unwarranted promises to borrowers. Regulation prohibits phone calls between 8 P.M. and 9 A.M. There is no specific regulatory limit on how many calls can be made to a borrower, but an implicit standard is about 3 to 6 calls per borrower per day. Extra phone calls per day are often considered to be abusive debt collection by courts. In our data, borrowers receive five calls a day on average. The company penalizes callers for making calls outside the regulated time range or receiving borrower complaints about excessive phone calls.

At later stages, the company uses different debt collection strategies. Since no AI callers are used in these stages, the discussion will be brief here. Around 60% of the overdue debts are outsourced to third-party debt collection agencies during days 26-59, and around 80% are outsourced during days 60-84. The company sends almost all 85+ day late cases to third-party debt collection agencies, only keeping small cases to be collected by AI callers (see Section 3) and some very large ones for further actions like lawsuits. A caller will handle a debt in days 11-25 for one week before it is assigned to another person. For debts in days 26-84, this interval is typically two weeks. Beyond 85 days, we do not have detailed information about strategies, since most cases are handled by third-party agencies. Borrowers are typically contacted less intensely as the repayment gets later (see Appendix Figure D2).

1.3 AI caller

To cope with the high volume of overdue cases and to reduce labor costs, the company introduced AI callers in 2018. Every morning before working hours, the company's system automatically assigns all open cases between AI and human callers. The assignment is completely randomized for 10% of cases and randomized conditional on loan characteristics for the remaining cases. The assignment rules will be explained in detail in Section 3. AI callers then automatically call their borrowers throughout the day.

The AI callers understand borrowers' speech using automatic speech recognition (ASR) and natural language understanding (NLU) technology. The AI caller generates appropriate answers and speaks with a synthetic voice. During the sample period, the AI caller was 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 conversations better. Section 5.1 gives more details about the upgrades and their impacts on the performance of both AI and human callers.

The AI caller can provide basic information about the overdue loans, address potential negative consequences of delinquency, and respond to simple questions and explanations about the 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 their name. If the AI caller dialed the wrong number, it apologizes and hangs up the phone. Otherwise, the AI caller continues 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 caller also emphasizes potential negative consequences if the borrower fails to repay: worsening credit records, large amounts of late fees, difficulties in future borrowing and consumption, and even lawsuits from the company. The AI may also mention the possibility of informing the borrower's guarantors, who are typically their parents and colleagues, imposing social pressure.

The AI caller then waits for the borrower's responses and sees if they have any further questions. The software classifies possible responses into five broad categories. In Case A, the borrower agrees to repay today or asks 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—for example, 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 borrow money from their family 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 a 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 complains 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 our sample to cases entering the debt collection process before December 2022, which gives us more than 22 million cases. Consistent with the company's debt collection practice, multiple debts of an individual borrower are merged into one entry during collection.

We have loan and borrower characteristics about all delinquent debts, including loan size at delinquency, borrower internal credit score, age, gender, and education level. The company uses two different variables to measure loan sizes: the overdue amount and the remaining principal. The overdue amount is the cumulative monthly payments already due that the borrower has not repaid, while the remaining principal is the principal amount that the borrower has not repaid. For the credit-card-like products, the entire outstanding balance is supposed to be repaid at the end of each

month, so their overdue amounts equal the remaining principal. The two measures can be different for installment loans, whose monthly repayment due is not the same as the total principal.

The internal credit score is the probability of default estimated by the company from a logit regression. The company divides all delinquent borrowers into 10 deciles and assigns them an integer score from 1 to 10, where the lowest credit score 1 is the highest decile of default probability. 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, although 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 for having a bachelor's degree or above.

We also have daily records of debt collection status and repayment actions. We know the number of days overdue, the stage the loan belongs to, whether it is being handled in-house or by a third-party debt collection agency, 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.[8]

Finally, we have data about callers' 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 collection agency. For in-house callers, we have their job titles and their tenure (in months) with the company. Performance measures include the total amount of money collected, monthly target collection amount, performance ranking, and the ratio of the actual amount collected by the caller to her target. Regarding caller salary, we know the salary amount each caller received, as well as the portion that is performance-based and the amount that is deducted due to penalties.

Table 2 Panel A reports summary statistics about the 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 amount of 1,128 yuan (160 USD) and a remaining principal of 6,474 yuan (910 USD), which are larger than

[8] 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.

the corresponding moments for the population of outstanding loans, as larger loans are more likely to default. The medians are smaller than the means: the median overdue amount is 653 yuan (92 USD), and the median remaining principal is 4,248 yuan (600 USD). The average internal credit score is around 5, consistent with its definition. Among delinquent borrowers, 70% are males, 13% have a bachelor's degree or more, and the average age is 27 years.

Case sizes are heavily right-skewed: the maximum remaining principal is 1 million yuan (about 140,000 USD). Extremely large debts are typically nonstandard contracts with specific customers for special purposes. They are treated separately by the company, so we want to exclude them from our analysis. Since separately treated cases are not labeled in our data, we exclude cases with remaining principal above the 99th percentile. 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. Figure 1 illustrates the assignment procedure. First-time delinquent borrowers are always assigned to human callers so that the company can have efficient communications with these borrowers to avoid another delinquency. Starting with the second delinquency, borrowers can be assigned to either AI or human callers.

The company initially allocates all cases with overdue amounts no greater than 20 yuan or remaining principal no greater than 300 yuan to AI callers. In rare situations, which we discuss later in this section, fewer than 5% of these small cases are assigned to human callers after day 25. The company does not use human callers on small cases since it is not cost-effective to do so.

Larger cases are either unconditionally or conditionally randomly assigned to AI or human callers. The company randomly selects 10% of these larger second-delinquency cases every month for testing and monitoring purposes, which we will refer to as the "completely randomized subsample."[9] In this subsample, a random half of cases are assigned to human callers on day 2, while the other half are assigned to AI callers on days 2 to 5 before being reallocated to human

[9] Larger cases that enter the debt collection process on the last few days of each calendar month are always assigned to human callers because there are fewer cases initiated at the end of each calendar month. These cases are excluded from our analyses.

callers on day 6 onwards. Once a given delinquency is handled by a human caller, it typically will not be given back to an AI caller. For subsequent delinquencies, the borrower's assignment to be initially called by a human or an AI remains the same as it was for his second delinquency. Therefore, only the assignment in the second delinquency can be viewed as orthogonal to potential outcomes within our sample; the type of person who reappears in our data as delinquent a third time after always being called by a human during her second delinquency might be different from the type who becomes delinquent a third time after first being called by an AI during her second delinquency. Thus, for larger cases, our analyses focus only on borrowers in their second delinquency, which are about 11% of the full sample.

The remaining 90% of larger second-delinquency cases are assigned between AI and human callers randomly conditional on case characteristics; that is, the probability of a case being assigned to AI varies by its characteristics. AI treatment effect estimates within this conditionally randomized subsample are similar to those in the completely randomized subsample, so for the sake of brevity, we do not report them.

Whereas the completely randomized subsample allows us to identify the effect of replacing human callers with AI callers in days 2 to 5, the discontinuity in the company's assignment rule for small cases creates an opportunity to use a regression discontinuity (RD) design to identify the local treatment effect of replacing human callers with AI callers for a much longer time. The 20-yuan overdue amount threshold is almost at the 1st percentile of the full-sample distribution, as shown in Table 2 Panel A. The number of cases to the left of this threshold is small, and the local treatment effect for individuals around this extreme point may be less representative of the remaining sample's treatment effect. In contrast, the 300-yuan remaining principal threshold is close to the 25th percentile of the full-sample distribution. Therefore, in the RD analysis, we exclude cases with less than 20 yuan of overdue payment amounts and apply the standard RD methodology with one running variable—the remaining principal.

Figure shows the fraction of cases assigned to AI 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. Consistent with the stated assignment rules, cases below 300 yuan of 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 Panel (b) shows the fraction of cases assigned to AI callers on both sides of the threshold from day 1 to day 25 past due. The fractions for "Under 300" are calculated based on cases in the (295, 300) yuan interval, while the fractions for "Above 300" are calculated based on cases in the (300, 305) yuan interval. Small cases are all handled by AI callers in the first 25 days. In contrast, on days 2-3, only 80% of the larger cases are assigned to AI callers. The fraction falls to around 60% on days 4-5. From day 6 onwards, all larger cases are handled by human callers. Panel (c) extends the horizon to day 360. Cases above 300 yuan remain in human treatment for the entire extended period. For cases below 300 yuan, a small fraction of them are assigned to human callers after day 25, mainly due to the introduction of third-party debt collection agencies. When the company delegates to a third-party debt collection agency, it randomly selects some cases, maybe conditioning on some loan characteristics, and assigns them to the agency. The assignment of some small cases to humans biases against finding significant differences across the remaining principal threshold.

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 starting on day 2 past due as the measure of caller productivity. For each case, we calculate how much money is collected on each of the following days, including any late fees collected, until the loan is paid back fully. We then discount these cash flows to day 2 using a 24% per annum (24/365 = 0.066% per day) discount rate, 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. It can be viewed as the opportunity cost of uncollected money, which could have been lent to other borrowers and generated interest at 24% APR if it were collected on time.[10] Finally, the NPV is scaled by the initial overdue amount on day 2. If the computed NPV grows beyond 1, we assume that borrower has fully repaid the initial payment due and cap the NPV at 1 afterward.

[10] As a robustness check, we also calculate debt collection productivity as the sum of undiscounted collected cash flows scaled by the initial overdue amount. These results are reported in Online Appendix D. Generally, the choice of discount rate has little impact on the results, since most payments are collected in the early days of delinquency.

4.2 Small cases subsample: Regression discontinuity design

In this subsection, we compare the productivity AI callers to human callers by utilizing the discontinuity in the company's AI deployment strategy at the 300-yuan cutoff in remaining principal.

Table 2 Panel B reports summary statistics of loan and borrower characteristics in our subsample for the RD design: 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 in the full sample, as expected, the gender composition, average age, and the fraction of borrowers with a bachelor's degree or more are all close to those in the full sample. The internal credit score is somewhat lower in the RD sample than in the full sample, but is still very close to 5.

Figure Panel (a) shows binned scatter plots of loan and borrower characteristics around the 300-yuan remaining principal threshold.[11] These characteristics are continuous at the cutoff. In Appendix B Section 2, we further check if there is manipulation around the cutoff. We do so first by examining the density of observations on either side of the cutoff. We also conduct a binomial randomization test, as suggested by Cattaneo et al. (2019). The results suggest that, although the remaining principal amount has some tendency of clustering at 300 yuan (and also at 200 yuan and 400 yuan), the density functions can still be viewed as continuous at the threshold.

Given the validity of our RD design, we examine the average collected NPV difference across the two sides of the cutoff, which gives the treatment effect of AI callers on debt collection productivity. Figure Panel (b) presents RD plots of collected NPVs at horizons of 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 at the threshold is a lower bound on the productivity gap magnitude before 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.

Table 3 formalizes our observations in Figure . It reports the differences between AI and human callers in variables of interest using robust RD estimators. Panel A reports continuity tests

[11] For clarity, the number of bins is set to 40 (i.e., each bin is 5 yuan wide) on either side of the cutoff, which is close to the Integrated Mean Squared Error (IMSE)-optimal number of bins of around 44. The IMSE-optimal number of bins minimizes the IMSE of local mean estimators. It is useful for assessing the overall shape of the 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, which can better capture the local variability of the data.

on five loan and borrower characteristics, as in Figure Panel (a).[12] Local linear regressions with uniform kernels over the coverage error rate (CER)-optimal bandwidth are used in the estimation, and the *z*-statistics are adjusted for clustering at the calendar month level.[13] The results show that the average loan characteristics on the two sides of the cutoff (columns 2 and 3) are quite similar, and the differences are small in column 4. The *z*-statistics of the differences in column 5 are close to zero and the corresponding *p*-values in column 6 are greater than 0.1. All 95% confidence intervals 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 estimates of the productivity gap, the mean squared error (MSE)-optimal bandwidths are used in the regressions. The differences between the left NPV mean (AI) and the right NPV mean (Human) are all negative and significant at the 1% level, regardless of the evaluation horizon. These gaps are also economically significant. On day 2, the NPV gap of 0.04 is a 14.7% productivity loss relative to the human mean NPV of 0.279. The gap grows to 0.11 by day 30 before starting to shrink because human callers are not able to collect much more beyond day 30, whereas AI continues to make some significant collection progress. Nonetheless, even after one year (360 days), AI's productivity loss relative to humans remains large: 7.6%. Figure 4 shows this productivity gap over time graphically. Re-estimating the NPV treatment effects while including the five loan characteristics in Panel A as covariates in the RD regressions has little impact on the magnitude and significance of the NPV differences.

Finally, we examine how AI's performance gap over time varies with borrowers' credit quality in Figure , using the same specification as in Figure . Low, medium, and high groups refer to internal credit scores of 1-3, 4-7, and 8-10. AI initially does relatively worse with high-score borrowers than with the other two groups. However, the gap between AI and humans stops growing on day 15 and starts to shrink quickly for the high group, approaching -0.04 in the long run. In contrast, the performance gaps for lower-score borrowers keep expanding until around 30 days.

[12] As suggested by Cattaneo et al. (2019), the CER-optimal bandwidth is used because, for testing the null hypothesis of no discontinuity, we are interested in inference (the confidence interval) instead of point estimations.

[13] The RD design can be viewed as an experiment in which debts around the cut-off are randomly assigned to either side of the cut-off and, hence, different treatments. There can be time-varying factors that affect the effective randomization assignment rules. For instance, the density of debts around the cut-off can be different over time so the probability of being treated can vary. Therefore, we cluster the standard errors by calendar month to account for such variation, as suggested by Abadie et al. (2023)—the "partially clustered assignment" case, in their words.

The magnitude of the long-run productivity gaps are monotonically decreasing in credit scores. The reason may be that high-score borrowers mainly need reminders, which AI callers can provide adequately, while low-score borrowers may have more complicated situations that require more personal communication and persuasion tactics that AI callers are less capable of performing.

4.3 10% completely randomized subsample

The previous section shows that AI callers are not able to fully replicate human callers' productivity. Therefore, the company usually has AI callers supplement human caller efforts in the early stages of delinquency. Specifically, some cases can be assigned to AI callers during the first five days past due. They are then all assigned to human callers on day 6.

To identify the performance of the "AI + Human" strategy, we utilize the 10% completely randomized subsample. 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 remainder 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 shows summary statistics on the completely randomized subsample. Since small cases with remaining principal no greater than 300 yuan are excluded from this subsample, the overdue amount and remaining principal here are on average larger than in the full sample. Other borrower characteristics are similar to the full sample.

As a first step, we validate that the treatment and control groups are balanced. Figure shows *t*-statistics from monthly tests of the equality of means of the treatment and control groups' overdue amount (Panel a) and remaining principal (Panel b). The *t*-statistics are all within the 90% critical values, are evenly distributed above and below zero, and have no time trend or clustering by time. In addition, we regress loan and borrower characteristics onto a treatment group indicator and calendar month dummies. Table 4 Panel A shows that the coefficient on the treatment group indicator is insignificant for overdue amount, remaining principal, internal credit score, gender, age, and education level. This shows that the two groups are statistically indistinguishable from each other ex ante, 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, 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*-statistic of this difference. For all evaluation

horizons, the "AI + Human" treatment group significantly underperforms the all-human control group. The gap is 0.089 on day 2, the first day of contact, which corresponds to a 32% productivity loss relative to the all-human control, and expands to 0.120 on day 5, a 22% productivity loss. Once human callers take over after day 5, the performance of the two groups converges quickly, so that the NPV difference is only 0.024 on day 10. Nevertheless, the "AI + Human" group never repays as much as the control group—even after a year, the gap is 0.0084.[14] 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 permanently impairs the company's ability to collect the balance due. Figure presents these results in graphical form.

Figure b and Figure c plot the NPV gap over time by internal credit score and loan size, respectively. The low-score cases incur the least productivity loss from AI initially, which is similar to what we saw in the RD analysis. We also learned from the RD design that 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. The medium-score cases exhibit the largest NPV gap in the short term but have similar asymptotic performance to the high-score cases. On the loan size dimension, larger cases generally have large performance gaps, consistent with our expectation that larger cases have more complexity that AI is less able to handle.

4.4 Potential sources of the performance gap

Why does AI underperform humans? In order to gain insight into this gap, we examine other phone call outcomes: the duration of phone calls, the fraction of borrowers who promise to repay, and the fraction of promisers who make the payments they owe shortly after the phone calls.

Table 5 Panel A reports the average outcomes of all phone calls made by AI and human callers on day 2 past due within the 10% completely randomized subsample, which gives us a clean setting for comparison. Human callers make 0.85 more phone calls per day to each borrower than AI callers and, unsurprisingly, are answered 0.35 more times per day. But for both types of callers,

[14] In Appendix Table D2, we reproduce Table 4 Panel B using the sum of undiscounted cash flows as the productivity measures and obtain similar results.

the phone answering rates are 23.6%, since borrowers cannot tell whether a phone call is made by an AI or human caller until they pick it up.

To analyze the differences in the ability of AI versus human callers that are separate from the frequency with which they make calls, we next restrict our sample to the first calls answered by each borrower. The results are reported in Table 5 Panel B. Note that the time of calls from human callers is on average earlier than calls from AI callers: 11:31 AM versus 11:47 AM.[15] The reason is that, as mentioned in Section 1.2, there is a half-hour "automatic calling" period from 9 A.M. to 9:30 A.M. when all cases assigned to human callers are called once, contributing to a large fraction of the first-answered calls. In contrast, calls from AI callers are distributed more evenly across the day. To control for this disparity, we estimate "timing-adjusted" results that control for one-hour-interval time-of-call fixed effects.

There is a significant unconditional difference between the two types of callers in how long the phone rings before it is answered, but this disparity disappears after controlling for time-of-call fixed effects. However, the duration of phone calls significantly differs whether or not time-of-call fixed effects are controlled for. The unconditional mean duration of an AI phone call is only 28 seconds, which is 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 only limited information and are not able to handle complicated situations, leading to short conversations.

Appendix Figure D5 shows the histograms of phone call durations for the two types of callers separately. AI phone calls are generally short and concentrated 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, 21.2% fewer borrowers make a promise to repay their debts when talking to AI callers. AI callers may be less persuasive and impose less pressure on borrowers, or people may just be reluctant to make promises with AI callers. One might expect that if AI callers are bad at

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

extracting promises, they will disproportionately receive promises from borrowers who are likely to pay promptly anyway. But Table 6 shows that callers who make a payment promise to an AI caller are 14.1 percentage points less likely to pay by the end of the day than callers who make a payment promise to a human caller. Borrowers who make a payment promise to an AI are 9.3 percentage points more likely to pay by the end of the day than borrowers who speak to an AI but do not make such a promise, which is only half of the promise versus no promise difference for borrowers who speak to a human—18.2 percentage points. These numbers suggest that borrowers feel less obligation to keep a promise to an AI than to a human. Table 5 Panel C shows that integrating across both borrowers who do and do not make a promise, an answered call from an AI is 13.0 percentage points less likely to result in a same-day payment than an answered call from a human.

When we calculate the borrower's tendency to answer the next phone call after talking to an AI or human caller, the probability of answering is around 45%, which is much higher than the unconditional rate. Surprisingly, talking to an AI caller increases the likelihood of answering the next call after adjusting for the hour of the call. However, the difference in the probability is only 1.3 percentage points.

In summary, the evidence suggests that, besides differences in calling strategies such as the number 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. Human callers are also better at extracting promises to repay and creating pressure to keep those promises. These implications are consistent with experimental findings on the role of honesty, morality, and/or social image concerns in personal communications (He et al., 2017; Burstyn et al., 2019; Cohn et al., 2022).

4.5 Understanding the permanent productivity gap

In Section 4.3, we document a permanent collection loss among debts that are first treated by AI callers instead of human callers, even if they are taken over by human callers on day 6. To assess whether there is something unique about AI that causes this permanent impairment, or whether *any* less-productive collection method used in the early days of delinquency also causes a permanent impairment, we examine the long-run effect of other factors that generate productivity losses in the initial stage of debt collection.

First, we exploit the fact that collection calls that occur on weekends are less effective. Since a debt's due date is determined by the day of the month and cannot be changed frequently, whether a borrower is first contacted about a late payment on a weekend should be uncorrelated with borrower and debt characteristics. We validate such orthogonality by regressing observable characteristics on day 2 after the due date (i.e., the first day of contact) on an indicator for if the day is a Saturday or Sunday. We also control for week fixed effects so that only variation within the same week is used for identification. The sample is the completely randomized subsample. Table 7 Panel A shows no significant difference between borrowers first contacted on weekends versus business days.

Table 7 Panel B compares several phone call and debt collection outcomes using the same specifications and reports the differences in column 3. Column 4 re-estimates the differences while including debt and borrower characteristics as additional covariates, which have little impact on the magnitudes of the differences. The first thing to note is that borrowers' probability of answering phone calls is significantly lower by about 0.6 percentage points during weekends, consistent with the finding in Laudenbach and Siegel (2024). To compensate for this low pick-up rate, callers tend to make more phone calls on weekends, resulting in a comparable total number of answered calls per borrower. In terms of collection performance, borrowers who are first contacted on weekends repay 1.6 percentage points less in normalized NPV on the day of contact than borrowers first contacted on business days. However, a significant gap only lasts for three days (or four days if covariates are included) and becomes insignificant afterward. Thus, the productivity loss from first contacting a borrower on a weekend can be fully offset by future effort, unlike initial contact from an AI caller.

We next explore variation in caller working experience, defined as the number of months since the caller joined the company. The company assigns debts randomly among human callers every day, so some debts are handled by more experienced callers on day 2. In line with the company's operating and managing practices, we define senior callers as callers who joined the company more than four months ago. We regress variables of interest onto a senior-caller indicator with month fixed effects using debts in the completely randomized subsample. The sample is further restricted to callers who specialize in debts that are in their first five days past due.

Table 8 presents the average debt characteristics and outcomes of junior and senior workers and tests the differences. Panel A performs a balance test with five debt and borrower

characteristics and confirms that they are uncorrelated with the working experience of the assigned callers, validating the randomization identification. Panel B first compares caller efforts and phone call outcomes. Senior callers make more phone calls per borrower-day, but senior callers do not differ from junior callers in other observed aspects on day 2. Junior callers collect 1.1 percentage points less of normalized NPV on day 2, but the gap disappears in the following days as the debts are rotated to other callers on each day. Again, there is no permanent damage associated with a less productive initial contact.

These results indicate that there is something uniquely damaging about being contacted by an AI caller. However, these results are only suggestive because the less productive initial contacts we test in this subsection are still much more effective than AI callers, and they are applied for only one or two days, unlike the five days the AI callers are assigned to a case. We cannot rule out that five days of calls from a human who is just as ineffective as an AI caller would result in similar long-term damage.

5 Interactions between AI and Human Callers

5.1 Impact of AI upgrades on human callers

The AI caller software was upgraded several times during our sample period, which give us an opportunity to examine how improvement in AI productivity affects human callers. This question is particularly important in light of the rapid development of current AI technology.

Figure 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 label as "V1" the first version in our sample period, which is not the same as the very first version of AI used by the company in 2018. Subsequent versions are labeled "V2" to "V5," according to their order of introduction.

As Figure shows, the company introduced new versions of AI callers progressively. V1 had been mostly replaced by V2at 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 its performance. As the company was satisfied with the outcomes, it slowly increased the fraction handled by V3 from 30% to 80% in the following months (August to October 2021). V3 completed replaced V2 in November 2021. The upgrades to V4 and V5 followed a similar procedure.

Since assignment among different versions of AI callers is random within each time period, the comparison of their productivity is straightforward. In the following analyses, we restrict our sample to the 10% completely randomized subsample. Figure 9 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. The gaps between AI and human callers remain wide over time; the upgrades of the AI callers did not close the gap very much. There is little performance difference between V1 and V2, and between V4 and V5. For V2 and 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 is significant on day 2 but not in the first five days past due.

Table 9 formally tests the improvement of AI performance in terms of collected NPVs and confirms the visual impressions from Figure 9. The table reports *t*-statistics for the differences between each pair of consecutive versions of AI callers over two-day to ten-day horizons. The calculation is based on the completely randomized subsample, so the AI caller only works from days 2-5. For each pair of AI callers, the test is implemented by regressing collected NPV onto an indicator for the call being made by the newer version of the AI and calendar month dummies 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 a substantial fraction of the calls—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 relative to the average NPV in the last row. On day 6, human callers take over, immediately closing the gap to an insignificant difference of 0.0099 between these two versions of AI callers. From another perspective, however, it also means that human callers taking over from the newer AI caller collect 0.0171 (= 0.0270 – 0.0099) less NPV on day 6 than humans taking over from the older AI caller.

This finding suggests that there is an upper bound on human callers' ability to collect overdue debts: no matter how much the AI callers collect in the first five days, the outcomes on day 6 are similar, reflecting the human limit. In other words, the more effective AI callers are, the less human callers can collect. This result is consistent with a displacement effect of AI callers.

Table 10 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 V2 and V3 coexist. We then study how human callers perform after receiving these AI-treated cases. We restrict the sample of human callers to those specializing in day 2-10 cases, 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, which is measured as the increase in collected NPVs from days 5 to 6, i.e., ΔNPV6. Hence, we focus on the cases that remain unpaid at the end of day 5 past due. In Column 1, we 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.

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 *PrevAbilityProxy*). 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.[16] Both *PrevAbilityProxy* and its interaction with the AI V3 indicator are included in the regression. We find that the coefficient on *PrevAbilityProxy* is significantly positive, implying that there is persistence in caller monthly performance, which may be interpreted as a signal of callers' ability. The significantly negative interaction term suggests that better callers are more heavily affected by the improvements of AI callers than their counterparts. This would be the case if the AI caller is improved by learning from and mimicking the skills of the best human callers.

5.2 The impact of increased task difficulty and the extent of AI application

We saw in Section 4.2 that AI callers' underperformance is especially large when trying to collect from low-credit-score borrowers. A natural managerial response could be to have AI specialize in high-credit-score borrowers and have humans specialize in low-credit-score borrowers. On the other hand, a rise in task difficulty for humans could make the job more

[16] In Appendix D Table D3 Columns 1-8, we regress the indicator of cases treated by AI V3, case 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 attrition bias concerns.

unpleasant, damaging morale, increasing job turnover, and raising the required compensation to retain workers. These are indirect costs of AI adoption that a company needs to take into consideration. We explore this point in this subsection by exploiting the ex-post imbalance in case difficulty among callers.

Debts are randomly assigned among human callers, so random variation causes some callers to receive more debts that are harder to collect than other callers. For each caller in each month, we measure this ex-post imbalance by the fraction of debts with internal credit scores of 3 or lower that are assigned to the caller, subtracting out the monthly average of this fraction. The larger the imbalance measure, the more difficult debts the caller is assigned. The sample is restricted to callers who specialize in debts in the first five days past due. We also require the caller to work for at least 20 days in a month to be included in the sample, so that callers with extreme imbalances are not disproportionately callers with fewer working days and thus fewer assigned debts.[17]

Table 11 Panel A reports the distribution of the imbalance measure, especially focusing on its variability. The mean and median are close to zero with a 1.2% standard deviation.[18] The 95th percentile is around 2%, which means that the unluckiest 5% of callers receive 2 percentage points more difficult cases than the average caller. The table also reports bootstrapped 95% critical values and bootstrapped *p*-values for the statistics from 10,000 simulated samples under the null hypothesis that debts are assigned randomly among callers every day. All realized statistics lie within the critical value intervals and all *p*-values are greater than 0.1, suggesting that the realized variation in ex-post imbalance is consistent with randomization. Table 11 Panel B further confirms the orthogonality of the ex-post imbalance by regressing the imbalance measure onto several caller characteristics one at a time. Caller characteristics include their working status (junior, senior, or returning callers who previously left the company), age, gender, and working experience in months. None of the *F*-statistics for the characteristics are significant.

We then regress callers' monthly performance and compensation onto the imbalance measure with and without additional caller characteristic controls. All specifications include month fixed effects. Table 11 Panel C reports the coefficients on the imbalance measure. Based on the results from the specifications without covariates, a caller who is at the 95th percentile and thus receives

[17] The typical number of working days of callers is 25 days a month, or equivalently, about six days a week.
[18] The mean is not exactly zero because the monthly average fractions are calculated with debt-level data, so each debt is equally weighted. In contrast, the ex-post iombalance measures are aggregated at the caller-month level, so each caller is equally weighted.

2 percentage points more difficult debts in a month has a repayment rate that is 0.46 percentage points lower, and her relative performance ranking is 10 percentage points lower. Worse performance translates into 477 yuan (67 USD) less total salary, which is equivalent to around 10% of the average compensation. Finally, we examine the caller's tendency to remain at the company at the end of the current month or at the end of the next three months with the same specification estimated with logit regressions. Average marginal effects are reported in the last row. The results indicate that an increase in ex-post imbalance of 2 percentage points is associated with a 3.7 percentage point higher chance of leaving the company within the next three months.

5.3 AI productivity net of labor cost savings

AI callers perform significantly worse than human callers. On the other hand, AI callers have almost zero marginal costs when making phone calls. Therefore, to fully evaluate how well AI may replace human callers, we need to subtract labor costs in our NPV calculations.

We first estimate *direct* labor costs in the debt collection process, that is, caller salary. Total caller salary 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 talk to one delinquent borrower for one minute, the average fixed cost is about 1.1565 yuan (0.16 USD), regardless of whether any money is collected or not. The per-minute fixed cost rate is converted to the borrower-day level by multiplying it by the average phone call length per borrower on the corresponding day after delinquency ($AvgCallLength_t$). In addition, for every yuan collected, the variable cost that the company needs to compensate the caller is approximately 0.0051 yuan (0.00072 USD). In the NPV calculation, we subtract out labor costs on the same day that the associated collection effort occurs.Figure shows the average differences of net collected NPV between AI and human callers as a function of days past due after adjusting for caller salary. 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 and Figure . For comparison, we also show the difference in unadjusted NPV.

After accounting for direct labor costs, the collected NPV gaps between AI and human callers become narrower. For the small cases with less than 300 yuan of remaining principal, the difference between AI and humans becomes indistinguishable from zero after about five months of collection effort. Because calling one delinquent borrower incurs a similar labor cost regardless of the debt's size, the return to human labor is relatively low on smaller cases. In the completely randomized subsample, where the case sizes are larger, the adjusted productivity gaps become smaller, but AI remains significantly less cost-effective than humans even in the long run.

Importantly, we only have considered direct labor costs, i.e., salary paid to human callers. To hire and manage more than 2,000 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 in our 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 impact of AI adoption in debt collection. Leveraging the company's rules of assigning delinquent debts between AI and human callers, we find that currently, AI is significantly less effective than human callers in the debt collection process. AI appears to have difficulties handling complicated situations, asking for promises, and creating pressure to keep those promises. Losses from using AI can be substantially mitigated if human callers take over cases from AI after a few days. Nonetheless, even in this collaborative arrangement, the NPV of collected balances remains permanently below the amount collected if only human callers are used, suggesting that AI callers create 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, social interactions, and the enforcement of moral obligations.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. "When should you adjust standard errors for clustering?." *The Quarterly Journal of Economics,* 138(1): 1-35.

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 repayment: 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.

Cialdini, Robert B., 2001. "The science of persuasion." *Scientific American* 284(2), 76-81.

Cohn, Alain, Tobias Gesche, and Michel André Maréchal. 2022. "Honesty in the digital age." *Management Science*, 68(2): 827-845.

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-28hinesechinese-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.

He, Simin, Theo Offerman, and Jeroen Van De Ven. 2017. "The sources of the communication gap." *Management Science*, 63(9): 2832-2846.

Kim, TaeWoo, Hyejin Lee, Michelle Yoosun Kim, SunAh Kim, and Adam Duhachek, 2023. "AI increases unethical consumer behavior due to reduced anticipatory guilt." *Journal of the Academy of Marketing Science* 51, 785-801.

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.

Melo, De Celso, Stacy Marsella, and Jonathan Gratch. 2016. "People do not feel guilty about exploiting machines." *ACM Transactions on Computer-Human Interaction* 23, 1–17.

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.

Petisca, Sofia, Ana Paiva, and Francisco Esteves. 2020. "Perceptions of people's dishonesty towards robots." *International Conference on Social Robotics*, 132–143. Springer.

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.

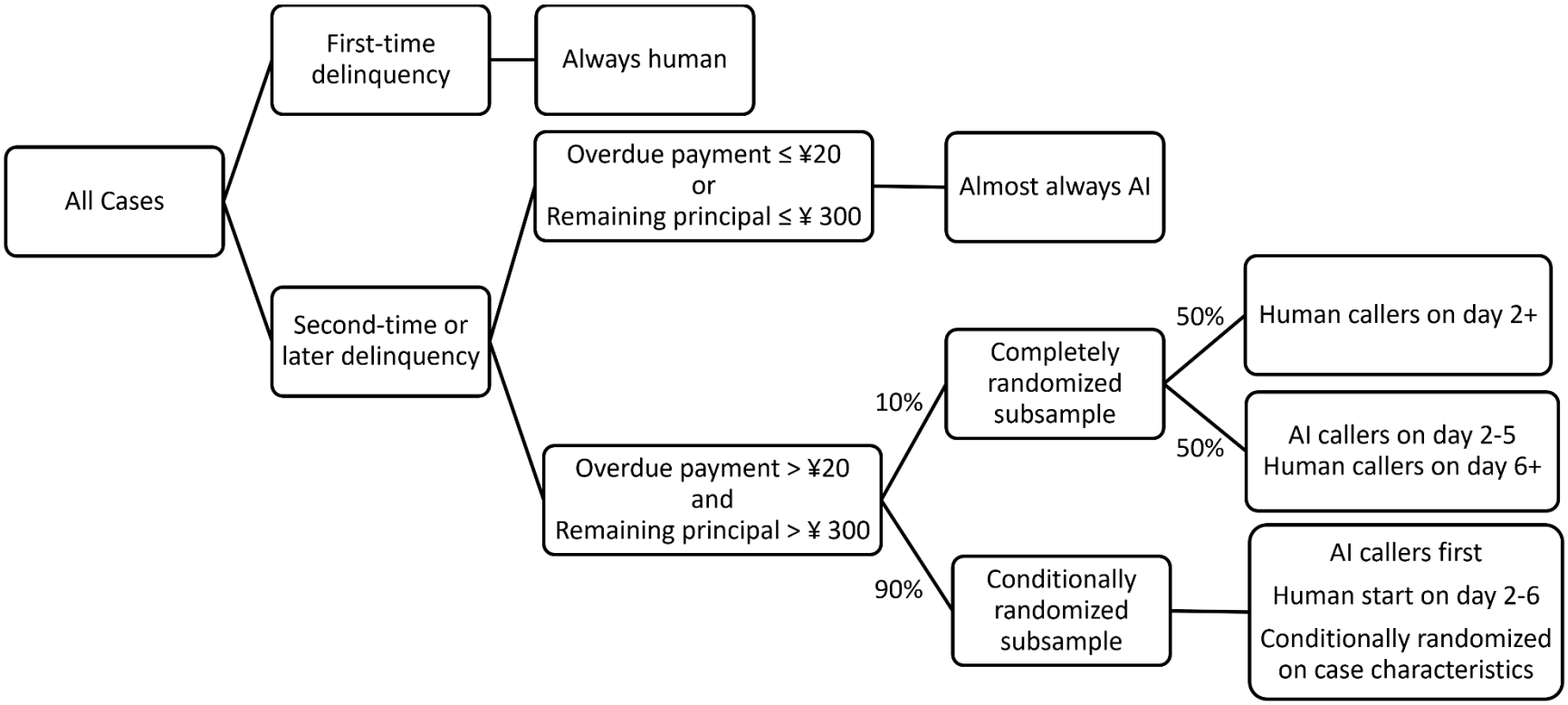
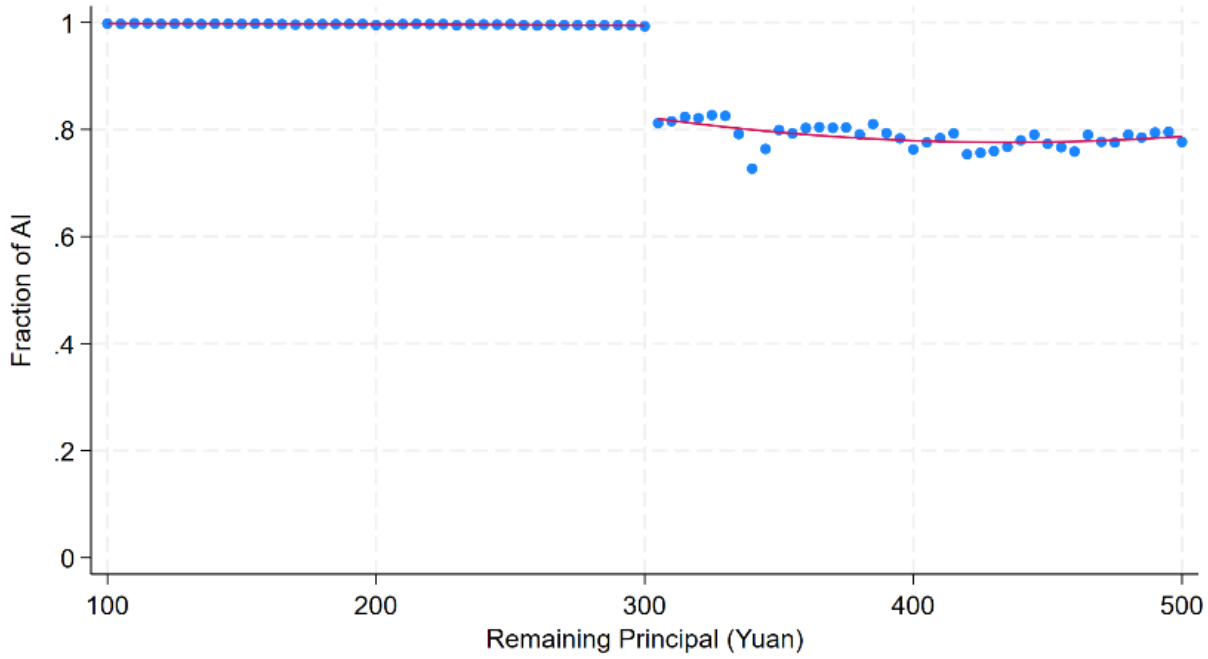
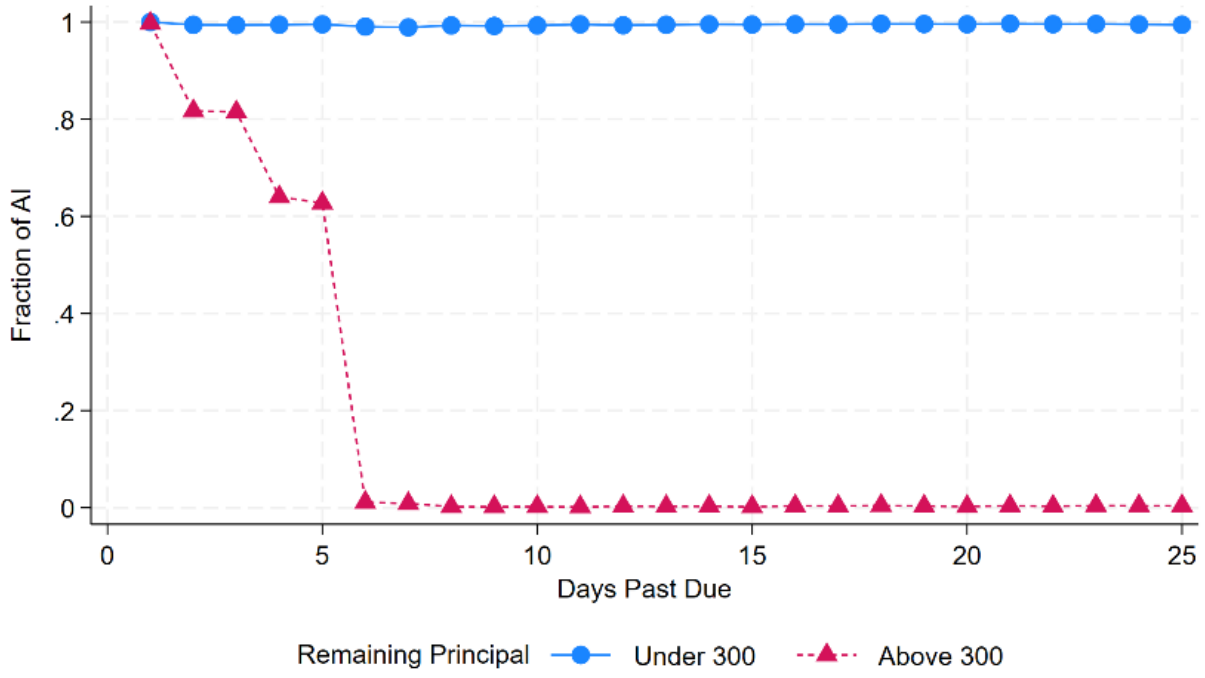


Figure 2. Fraction of cases assigned to AI callers around the remaining principal threshold. This figure shows the fraction of cases assigned to AI callers around the 300-yuan remaining principal threshold. Panel (a) is a binned scatter plot of the fraction of cases assigned to AI, clustered at 5-yuan intervals of remaining principal on day 2 past due. The plotted line is from a fitted quadratic regression. 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)

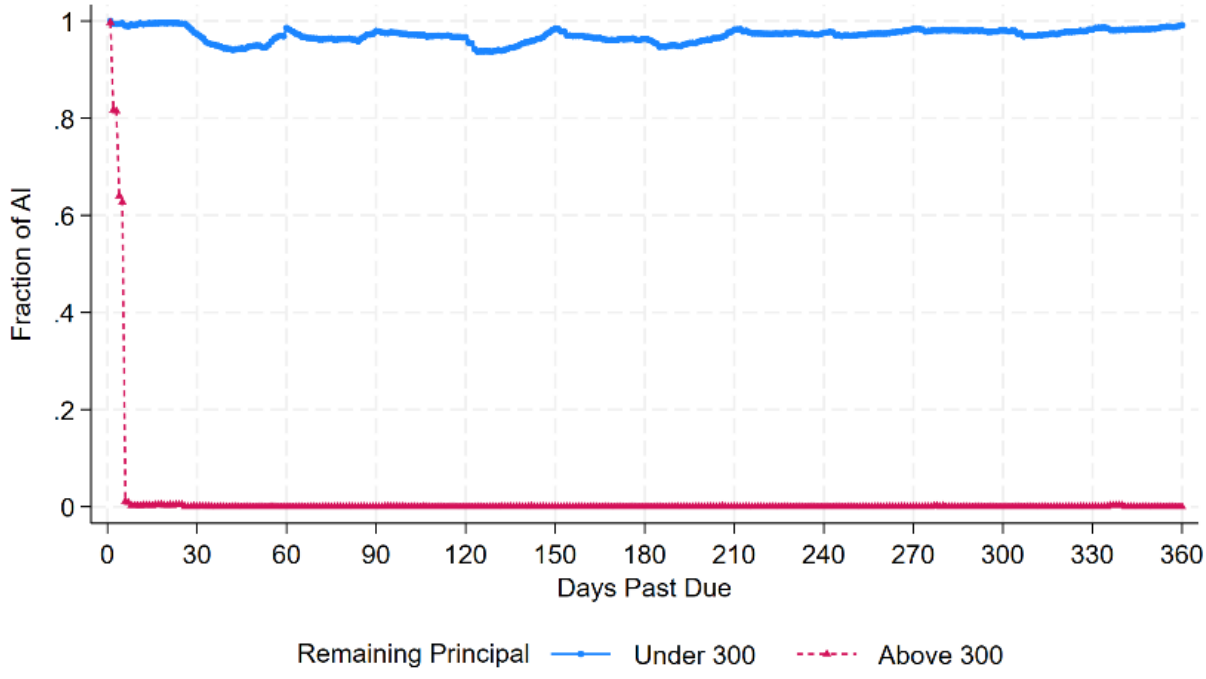
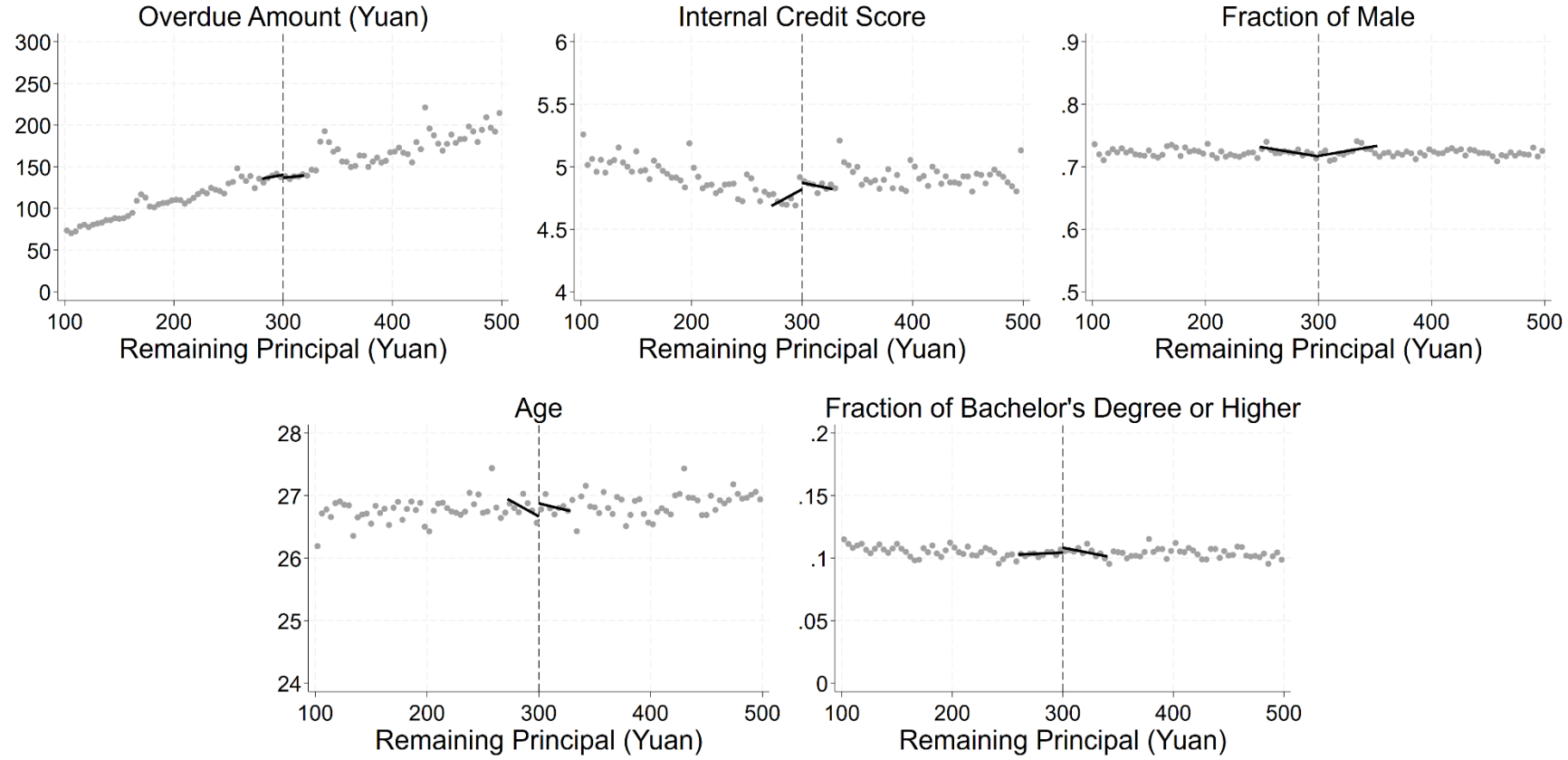


Figure 3. Loan and borrower characteristics and collected NPVs around the remaining principal cutoff.

This figure shows binned scatter plots of several variables around the remaining principal cutoff of 300 yuan. Loans with remaining principal no greater than 300 yuan 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 for whether they hold a bachelor's degree or higher), as well as NPVs of collected payments over various horizons, as in Panel (b). The collected NPV of a case is 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 local linear fits within the robust RD estimation bandwidths on each side.

(a) Loan characteristics



(b) Collected NPV

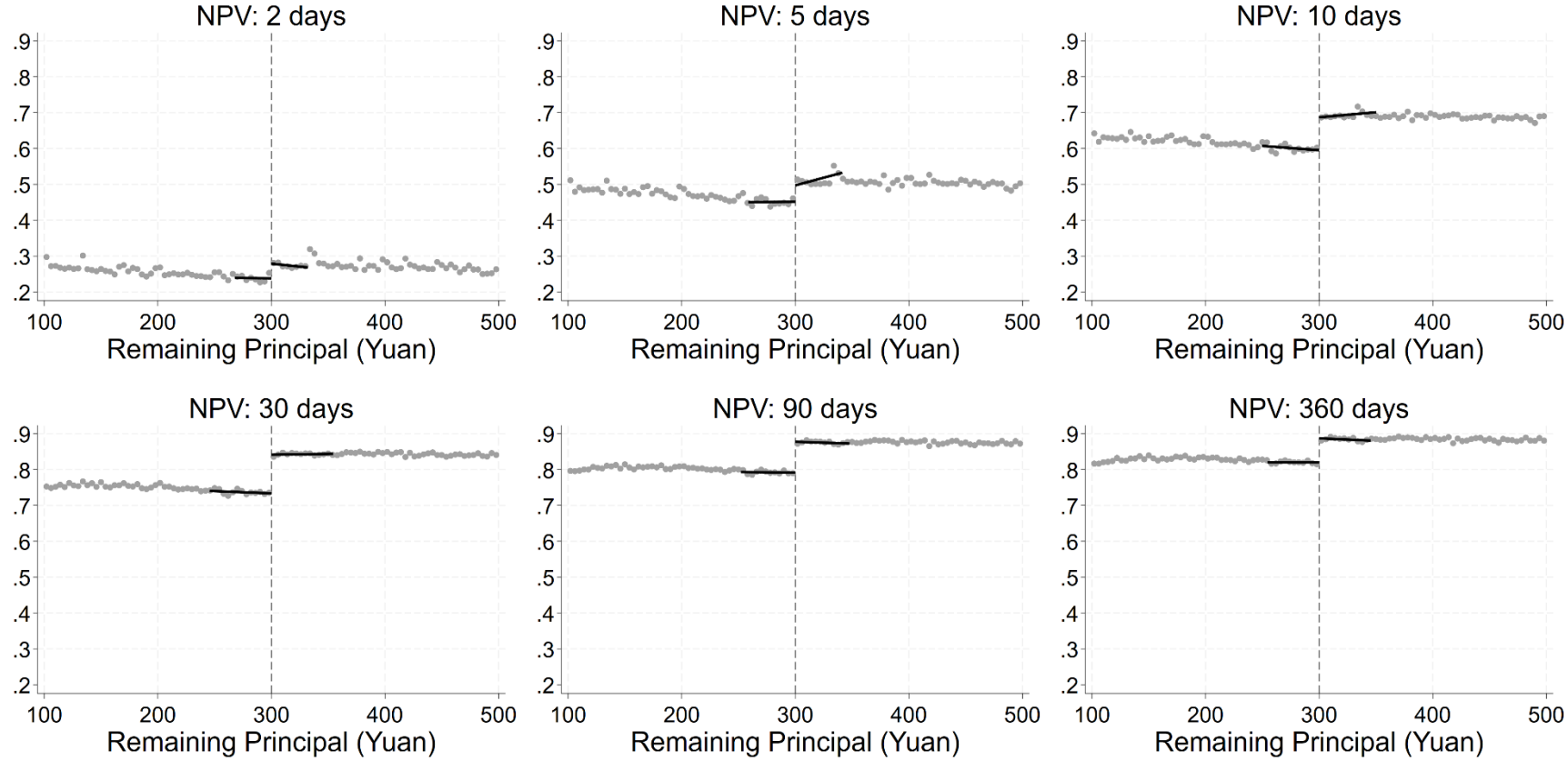


Figure 4. Collected NPV differences between AI and human callers over horizon – small cases RDD.

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. The dots represent the average differences estimated by RDD and the bars indicate the 95% robust regression discontinuity confidence intervals. For clarity, the differences are plotted every three days before day 60, and every 10 days after day 60.

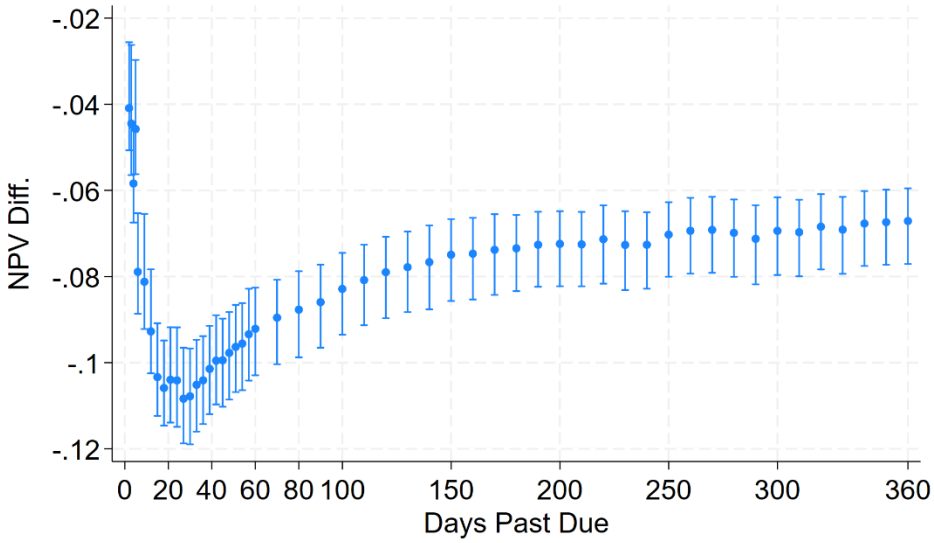


Figure 5. Collected NPV differences between AI and human callers over horizon, by internal credit score.

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.

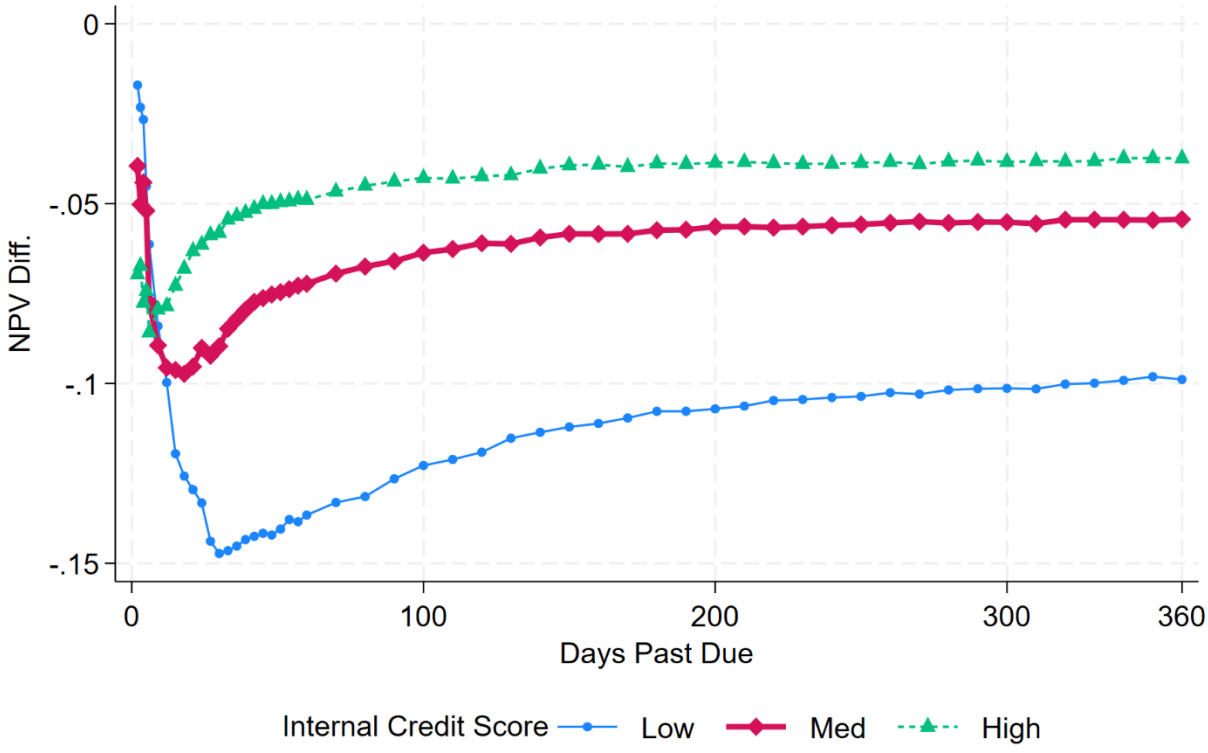
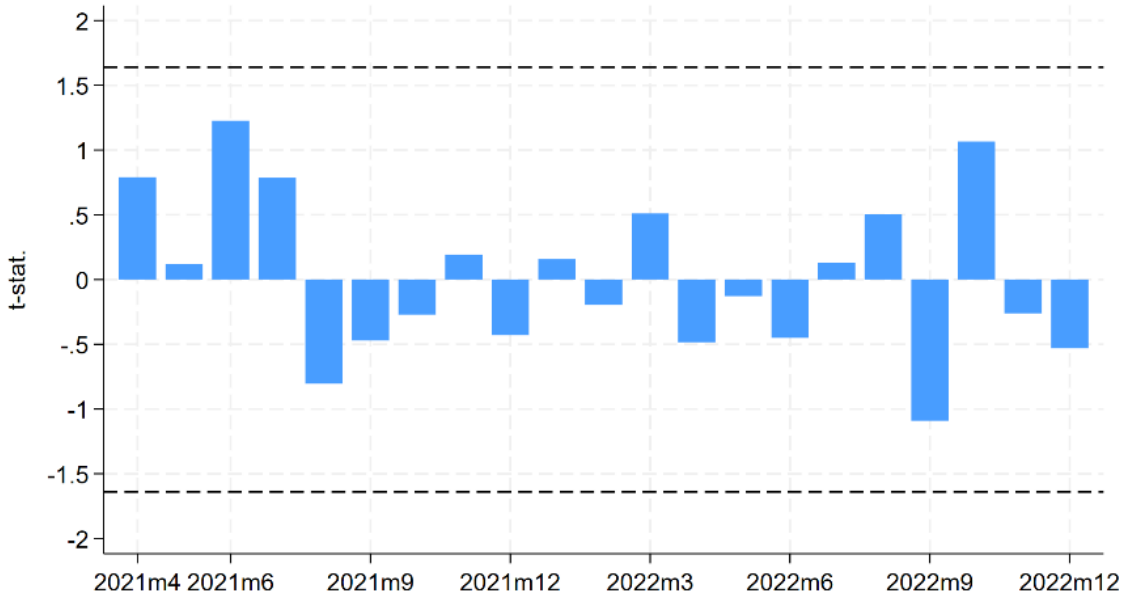


Figure 6. Test of randomization – completely randomized subsample.

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

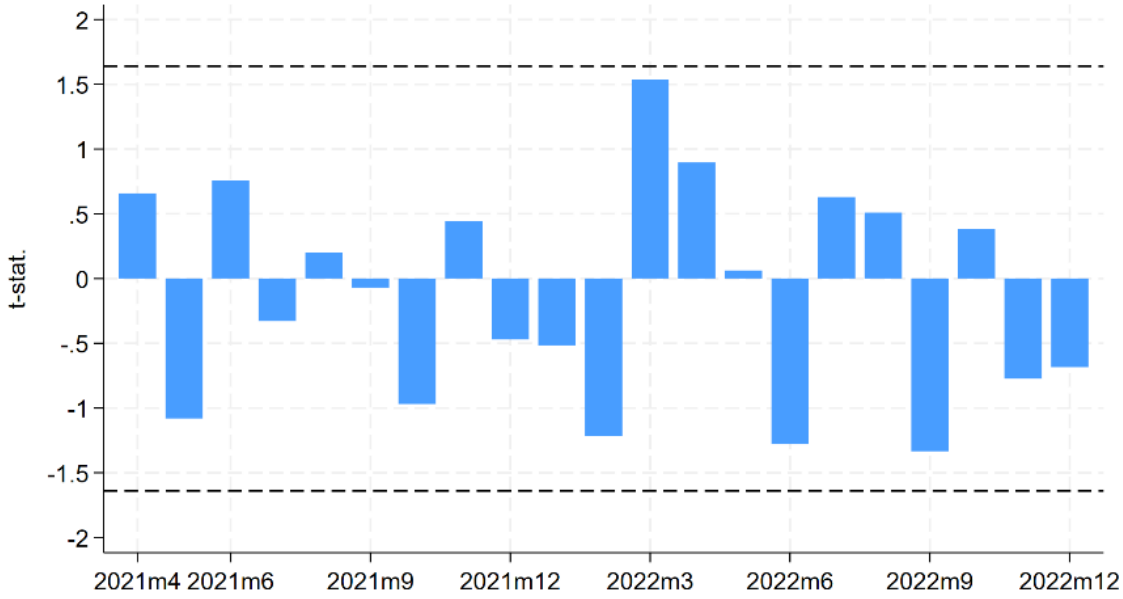
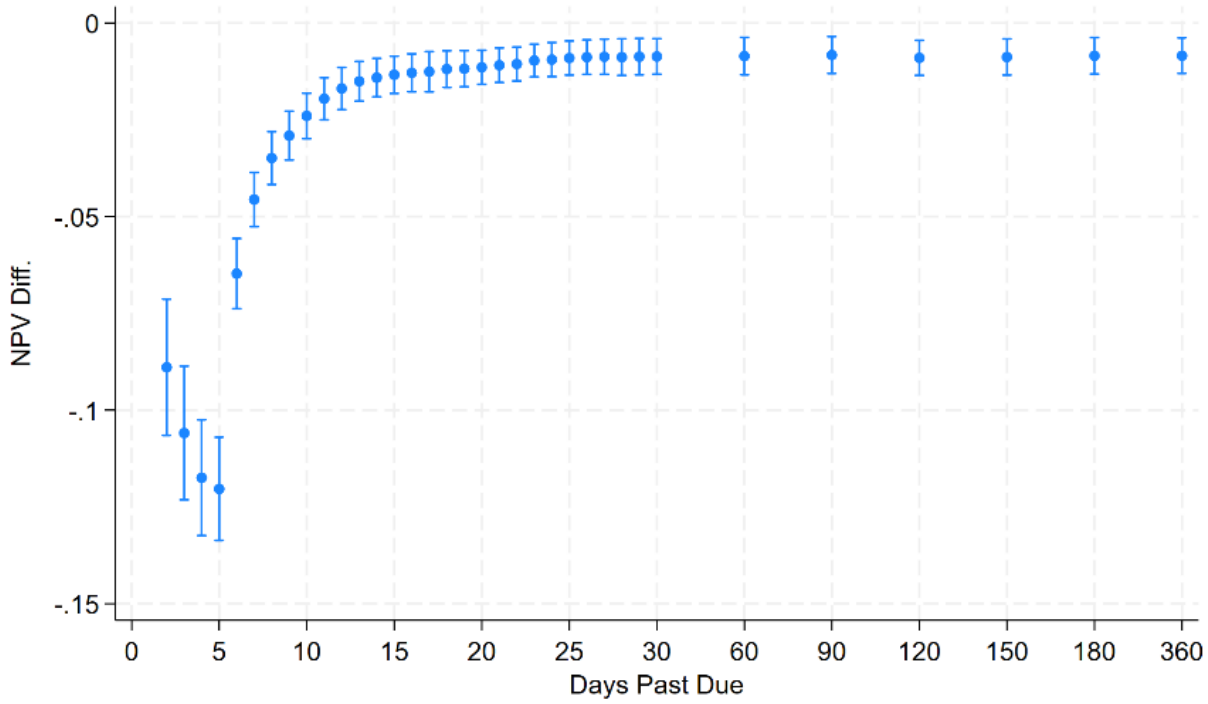
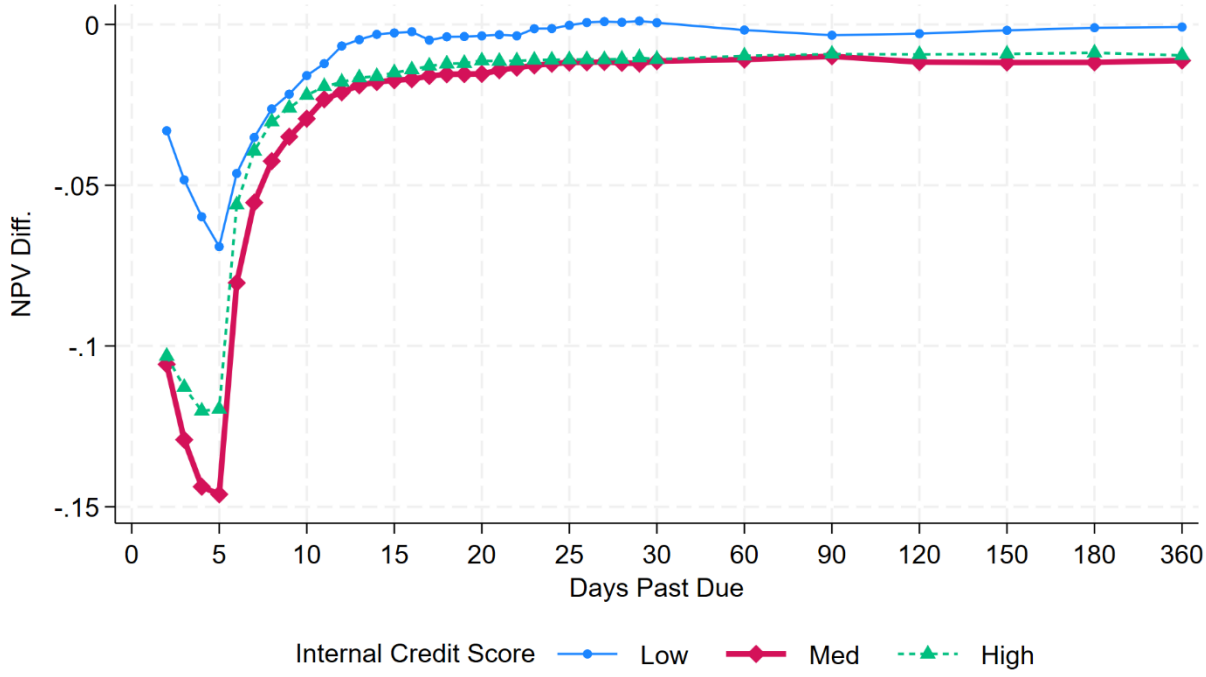


Figure 7. Collected NPV differences between AI and human callers over horizon – Completely randomized subsample.
This figure reports the average differences of collected NPV between AI and human callers over the horizon of days past due, using the 10% completely randomized subsample. The collected NPV 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. Panels (b) and (c) split the cases by internal credit score and overdue payment size, respectively, and estimate the differences. In Panel (b), "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.

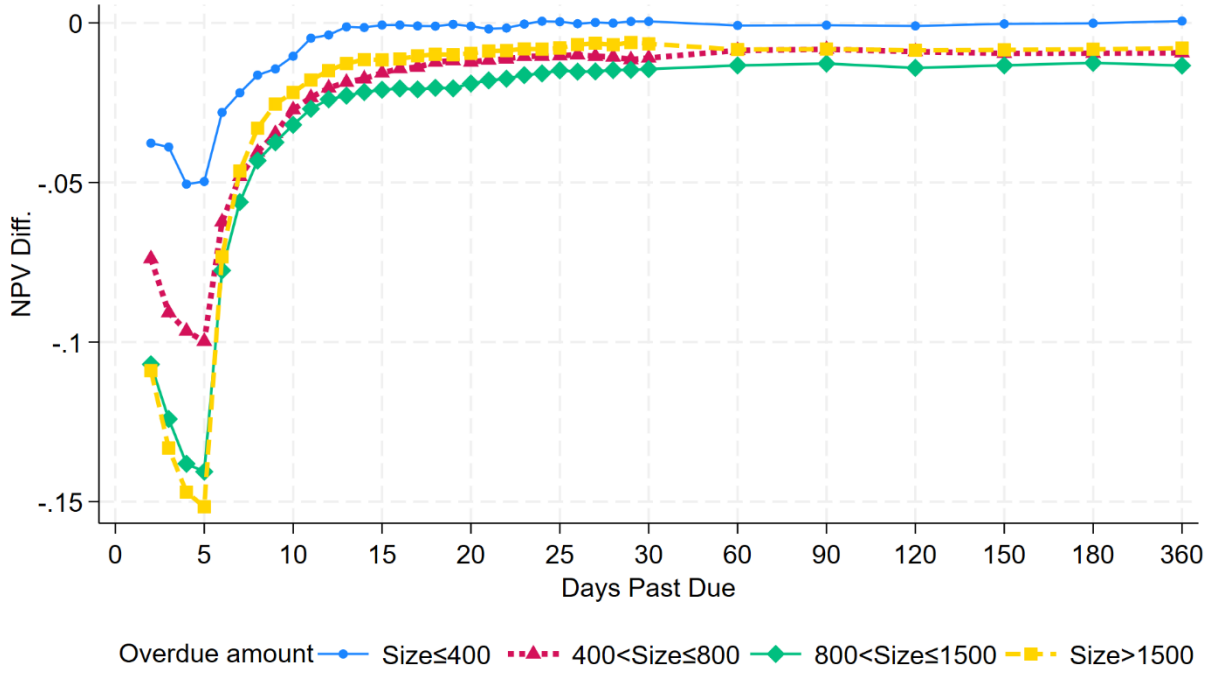


Figure 8. Fractions of different versions of AI callers over time.

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.

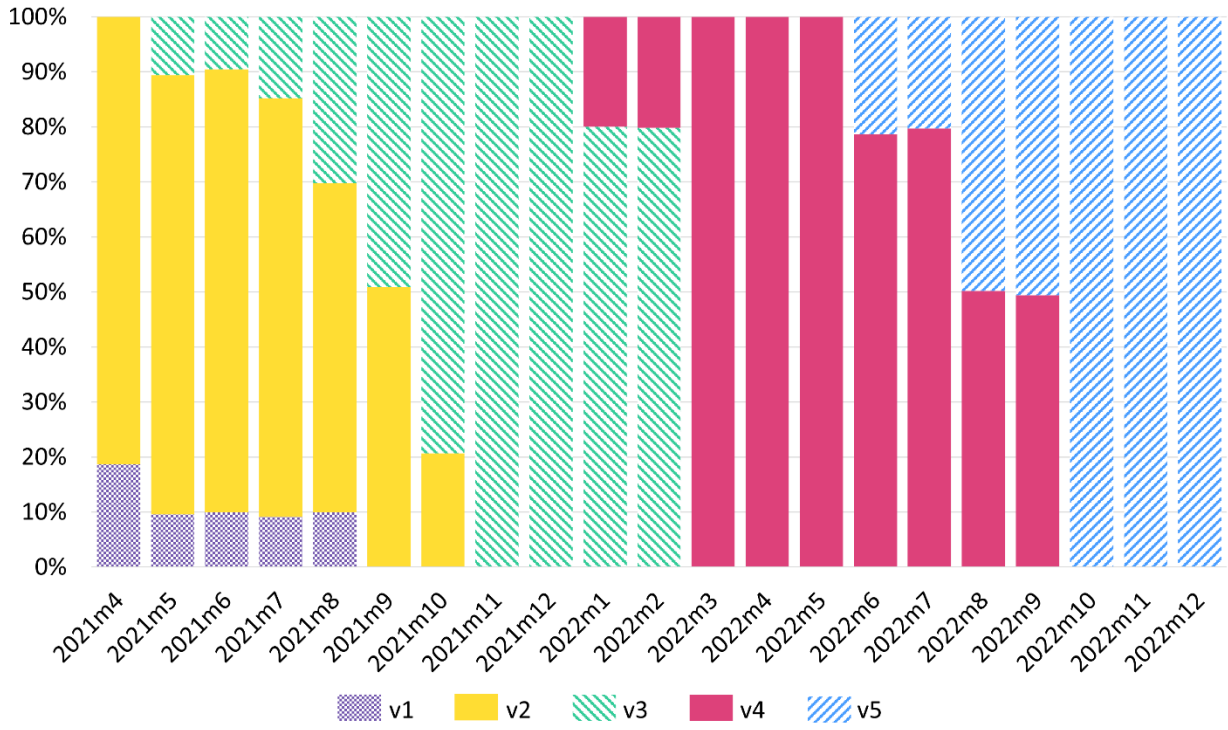
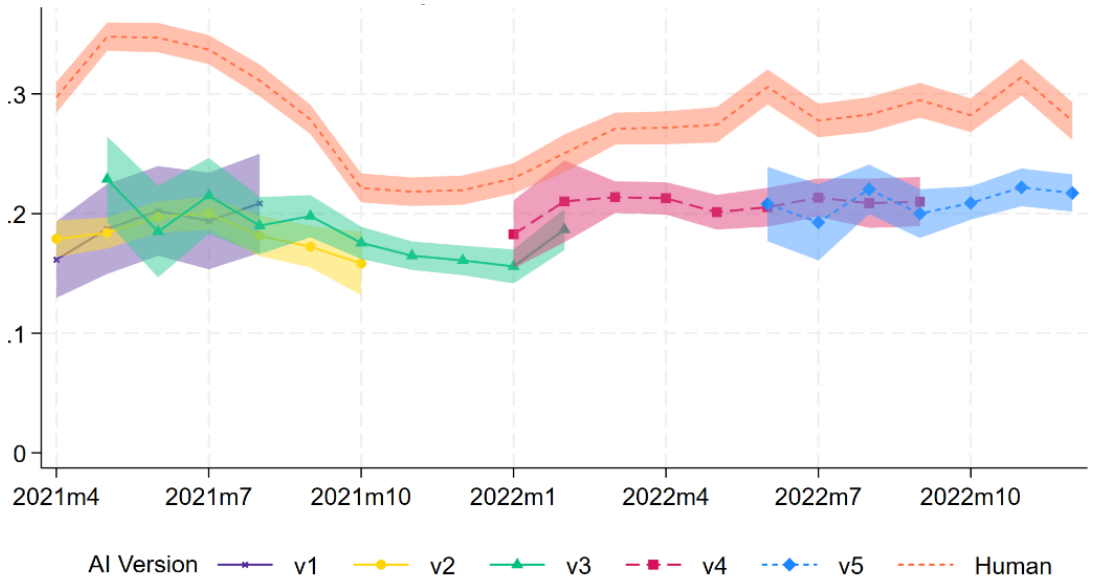


Figure 9. Performance of different versions of AI callers over time.

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

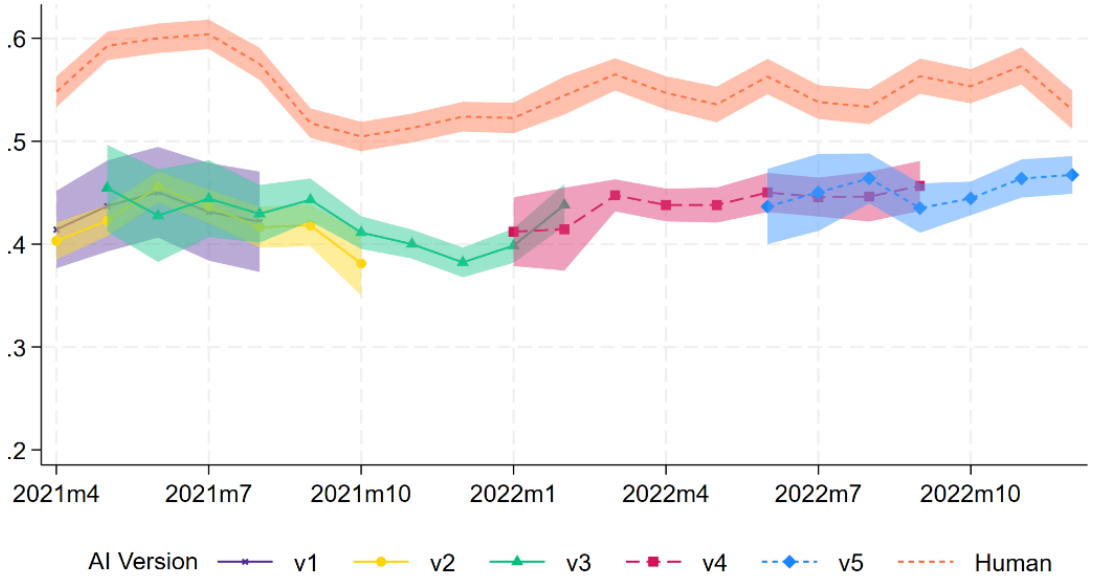
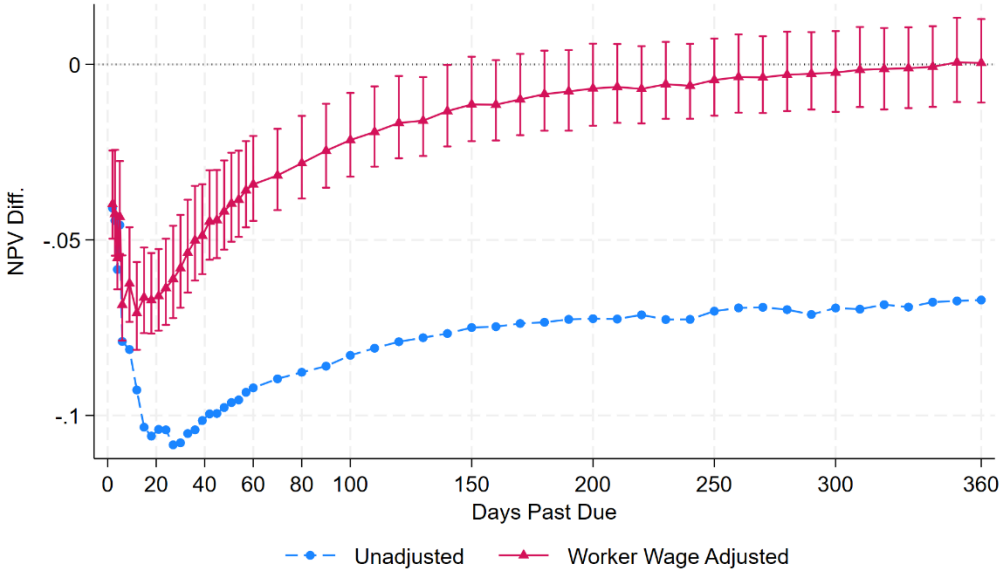


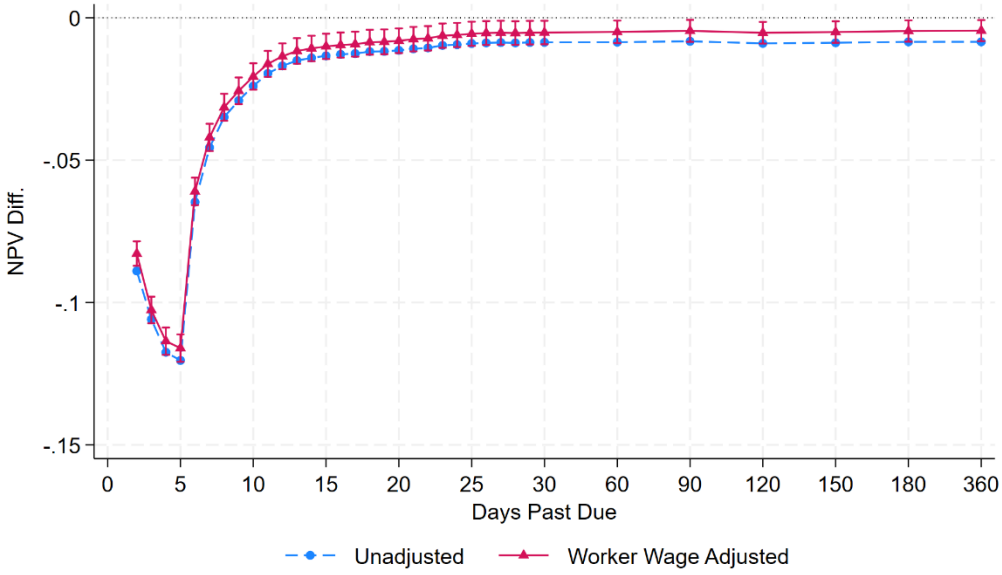
Figure 10. Collected NPV differences between AI and human callers over time – adjustment for labor costs.

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 today. 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 late 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 balances 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 amount, 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 amount (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 amount (yuan)	142.28	112.50	20.01	22.13	36.68	58.73	106.24	188.92	303.32	1,011,509
Remaining principal (yuan)	304.74	112.41	100.00	104.04	148.21	209.00	307.72	400.55	459.04	1,011,509
Internal credit score	4.91	2.77	1	1	1	3	4	7	9	1,011,509
Male indicator	0.72	0.45	0	0	0	0	1	1	1	1,011,509
Age	26.82	5.98	18	19	21	22	25	30	35	1,011,509
Bachelor's degree or more indicator	0.10	0.31	0	0	0	0	0	0	1	1,011,509

Panel C. Completely randomized subsample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	No. Obs.
Overdue amount (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 regression-fitted value of the variables of interest at the threshold from the left side (permanent AI) and the right side (human). Column 4 reports the differences between the left and right fitted values, 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. The *z*-statistics are adjusted for clustering at the calendar month level. 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 Covar.
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. 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	0.15
(2)	Remaining principal	8585.2	8604.7	-19.5	-0.54
(3)	Internal credit score	5.970	5.961	0.009	0.66
(4)	Male	0.701	0.703	-0.002	-0.83
(5)	Age	27.75	27.79	-0.043	-1.25
(6)	Bachelor's degree or higher	0.103	0.104	-0.001	-0.36
Panel B. Collected NPV					
(7)	NPV 2d	0.193	0.282	-0.089***	-42.33
(8)	NPV 5d	0.431	0.551	-0.120***	-48.73
(9)	NPV 10d	0.647	0.671	-0.024***	-10.17
(10)	NPV 30d	0.767	0.776	-0.0086***	-4.18
(11)	NPV 60d	0.800	0.809	-0.0086***	-4.40
(12)	NPV 90d	0.816	0.824	-0.0083***	-4.40
(13)	NPV 180d	0.830	0.838	-0.0085***	-4.66
(14)	NPV 360d	0.836	0.844	-0.0084***	-4.73

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. 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) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
(1)	# Phone calls per borrower	4.62	5.47	-0.85***	-56.54
(2)	# Phone calls answered	0.65	1.00	-0.35***	-57.63
(3)	% Phone calls answered	0.236	0.236	0.000	0.06

Panel B. First answered calls.

	(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
(1)	Time of calls	11.79	11.51	0.28**	12.51
(2)	Ringing time to answer (sec)				
	Unadjusted	19.47	20.72	-1.25**	-11.66
	Timing-adjusted	20.11	20.13	-0.02	-0.17
(3)	Duration (sec)				
	Unadjusted	28.12	47.13	-19.02***	-41.41
	Timing-adjusted	21.76	52.72	-30.96***	-61.25
(4)	% Promise to repay				
	Unadjusted	0.441	0.652	-0.212***	-52.44
	Timing-adjusted	0.441	0.652	-0.212***	-45.86
(5)	Prob. answering the next call				
	Unadjusted	0.447	0.454	-0.007	-1.47
	Timing-adjusted	0.457	0.444	0.013**	2.50

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	0.041	-0.002	-1.05
30 minutes	0.053	0.066	-0.013***	-5.60
1 hour	0.072	0.103	-0.031***	-11.05
2 hours	0.103	0.160	-0.057***	-17.41
5 hours	0.159	0.259	-0.100***	-25.68
the same day	0.236	0.366	-0.130***	-30.08

Table 6. Repayment after the first answered calls, conditioning on "promises to repay."

This table reports the fraction of borrowers who repay their debts (fully or partially) within various periods after the first phone call from AI or human callers on day 2 after the due date, conditioning on whether or not the borrowers make a promise to repay their debts during the conversation. The analysis uses the 10% completely randomized subsample. The estimations are based on linear regressions of the variable of interest onto an AI-case indicator with calendar month fixed effects. The time of calling is accounted for by including time-of-day fixed effects for every hour. *t*-statistics are reported in parentheses. Significance level: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

	With a Promise			Without a Promise			
	(a) AI(b) Human	(c) Diff: (a)-(b)	(d) AI	(e) Human	(f) Diff: (d)-(e)	(g) Diff: (c) - (f)	
Repay (fully or partially) after the first answered call within …							
15 minutes	0.050	0.050	0.000	0.030	0.024	0.006**	-0.006*
			(0.06)			(2.24)	(-1.73)
30 minutes	0.068	0.081	-0.014***	0.041	0.037	0.005	-0.018***
			(-4.58)			(1.38)	(-4.40)
1 hour	0.091	0.125	-0.035***	0.057	0.060	-0.003	-0.032***
			(-9.60)			(-0.73)	(-6.37)
2 hours	0.132	0.194	-0.061***	0.079	0.097	-0.018***	-0.044***
			(-14.38)			(-3.81)	(-7.40)
5 hours	0.202	0.312	-0.109***	0.124	0.159	-0.035***	-0.074***
			(-21.76)			(-6.47)	(-10.64)
the same day	0.288	0.429	-0.141***	0.195	0.247	-0.052***	-0.089***
			(-25.49)			(-8.56)	(-11.69)

Table 7. The day of the week and debt collection outcomes.

This table compares debts that are first contacted on weekends and business days using the 10% completely randomized subsample. Panel A reports the balance test results about loan characteristics, including overdue payment amount (yuan), remaining principal (yuan), internal credit score, the fraction of males, age, and the fraction of borrowers with bachelor's or higher degrees. Columns 1 and 2 are average characteristics of debts first contacted on weekends and business days, respectively. Column 3 is the difference between the two types of debts. Column 4 reports the *t*-statistics. Panel B examines phone call outcomes on day 2 past the due date and collected NPVs at various evaluation horizons. In addition to the first four columns as in Panel A, columns 5 and 6 re-estimate the outcome differences by adding covariates in Panel A. The estimations are based on linear regressions of the variable of interest onto a weekend indicator With week fixed effects. *t*-statistics are adjusted for clustering at the calendar month level. Significance level: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Panel A. Loan characteristics.

	(1) Weekend Average	(2) Business Day Average	(3) Diff: (1) - (2)	(4) *t*-stat.
Overdue payment	1508.56	1528.88	-20.33	-0.86
Remaining principal	8559.04	8608.30	-49.26	-0.65
Internal credit score	5.97	5.96	0.010	0.38
Male indicator	0.70	0.70	-0.005	-1.58
Age	27.77	27.77	-0.002	-0.04
Bachelor's degree of more indicator	0.102	0.105	-0.002	-1.61

Panel B. Debt collection outcomes.

	(1) Weekend Average	(2) Business Day Average	(3) Diff: (1) - (2)	(4) *t*-stat.	(5) Diff: (1) - (2) w/ Covar.	(6) *t*-stat.
Call outcomes on day 2:						
% Calls answered	0.232	0.238	-0.006***	-3.14	-0.006***	-3.19
# Calls per borrower	5.091	5.018	0.072**	2.17	0.076**	2.54
# Calls answered	0.815	0.828	-0.013	-1.21	-0.012	-1.14
Call duration	44.22	44.76	-0.54	-0.57	-0.36	-0.39
% Promise to repay	0.162	0.166	-0.003	-1.44	-0.003	-1.37
NPV:						
NPV2	0.226	0.241	-0.015***	-5.46	-0.016***	-5.89
NPV3	0.368	0.385	-0.018***	-5.76	-0.019***	-7.80
NPV4	0.437	0.448	-0.011***	-3.05	-0.012***	-3.77
NPV5	0.486	0.492	-0.006	-1.71	-0.007**	-2.13
NPV6	0.560	0.562	-0.002	-0.72	-0.003	-0.96

	(1) Weekend Average	(2) Business Day Average	(3) Diff: (1) - (2)	(4) *t*-stat.	(5) Diff: (1) - (2) w/ Covar.	(6) *t*-stat.
NPV10	0.655	0.660	-0.005	-1.70	-0.006*	-2.01
NPV15	0.709	0.709	-0.000	-0.16	-0.001	-0.38
NPV30	0.771	0.772	-0.001	-0.48	-0.002	-0.66
NPV60	0.804	0.805	-0.001	-0.68	-0.002	-0.92
NPV90	0.820	0.820	-0.000	-0.22	-0.001	-0.44
NPV180	0.833	0.834	-0.001	-0.55	-0.002	-0.72
NPV360	0.839	0.840	-0.001	-0.65	-0.002	-0.80

Table 8. Caller working experience and debt collection outcomes.

This table compares debts that are first contacted by senior and junior callers. Senior callers are defined as callers who have been working for the company for at least 4 months. The sample is from the 10% completely randomized subsample that is restricted to human callers specialized in debts in the first 5 days past due. Panel A reports the balance test results about loan characteristics, including overdue payment amount (yuan), remaining principal (yuan), internal credit score, the fraction of males, age, and the fraction of borrowers with bachelor's or higher degrees. Columns 1 and 2 are average characteristics of debts first contacted by the two types of callers. Column 3 is the difference between the two and column 4 is the *t*-statistics. Panel B examines phone call outcomes on day 2 past the due date and collected NPVs at various evaluation horizons. In addition to the first four columns as in Panel A, columns 5 and 6 re-estimate the outcome differences by adding covariates in Panel A. The estimations are based on linear regressions of the variable of interest onto a senior-caller indicator with month fixed effects. *t*-statistics are adjusted for clustering at the month level. Significance level: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Panel A. Loan characteristics.

	(1) Senior Avg.	(2) Junior Avg.	(3) Diff: (1) - (2)	(4) *t*-stat.
Overdue payment	1675.49	1677.22	-1.73	-0.07
Remaining principal	9246.51	9375.32	-128.81	-1.45
Internal credit score	5.87	5.92	-0.047	-1.28
Male indicator	0.71	0.702	0.005	0.84
Age	28.24	28.24	-0.005	-0.07
Bachelor's degree of more indicator	0.10	0.102	-0.005	-1.47

Panel B. Debt collection outcomes.

	(1) Senior Avg.	(2) Junior Avg.	(3) Diff: (1) - (2)	(4) *t*-stat.	(5) Diff: (1) - (2) w/ Covar.	(6) *t*-stat.
Call outcomes on day 2:						
% Calls answered	0.229	0.234	-0.005	-1.24	-0.005	-1.28
# Calls per borrower	6.148	5.984	0.164***	3.92	0.158***	4.05
# Calls answered	1.055	1.080	-0.025	-1.26	-0.024	-1.29
Call duration	83.70	83.58	0.12	0.03	0.24	0.09
% Promise to repay	1.181	1.175	0.006	0.27	0.009	0.43
NPV:						
NPV2	0.265	0.256	0.009	1.68	0.011**	2.46
NPV3	0.413	0.408	0.005	0.68	0.008	1.26
NPV4	0.484	0.481	0.003	0.47	0.006	1.15
NPV5	0.530	0.531	-0.001	-0.23	0.002	0.4
NPV6	0.575	0.577	-0.002	-0.28	0.001	0.31

	(1) Senior Avg.	(2) Junior Avg.	(3) Diff: (1) - (2)	(4) *t*-stat.	(5) Diff: (1) - (2) w/ Covar.	(6) *t*-stat.
NPV10	0.655	0.659	-0.004	-0.83	-0.001	-0.29
NPV15	0.702	0.705	-0.003	-0.82	-0.001	-0.36
NPV30	0.767	0.765	0.002	0.35	0.003	0.73
NPV60	0.800	0.801	-0.001	-0.26	-0.000	-0.03
NPV90	0.816	0.815	0.001	0.3	0.002	0.55
NPV180	0.830	0.829	0.001	0.27	0.002	0.46
NPV360	0.836	0.835	0.001	0.26	0.002	0.44

Table 9. 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. Cluster robust *t*-statistics clustered at the calendar month level are in parentheses. 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)	(–.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 10. 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 cases 2-10 days 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 (*Prev Ability Proxy*) 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) ΔNPV6	(2) ΔNPV6	(3) ΔNPV6
AI V3	-0.033***	-0.001	-0.004
	(-2.65)	(-0.03)	(-0.17)
Prev Ability Proxy		0.102**	0.102**
		(2.35)	(2.37)
AI V3 × Prev Ability Proxy		-0.101**	-0.101**
		(-2.21)	(-2.22)
New Caller			0.033
			(1.37)
AI V3 × New Caller			-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

Table 11. Ex-post imbalance in assigned case difficulty and caller outcomes.

This table examines the relationship between ex-post imbalance in assigned case difficulty and worker outcomes. For each caller in each month, the ex-post imbalance measure is defined as the fraction of debts with internal credit scores of 3 or lower among all debts assigned to the caller, after adjusting for the monthly average fraction. The sample is restricted to callers who specialize in debts in the first 5 days past the due date and who work for at least 20 days in the month. Panel A reports the distribution of the ex-post imbalance measure. The realized statistics are calculated from the observed distribution. The 95% critical values and the *p*-values are calculated by bootstrapping with 10,000 simulated samples under the null hypothesis that debts are randomly assigned among callers each day. Panel B implements a balance test by regressing the imbalance measure onto worker characteristics one at a time. *F*-statistics for the worker characteristics are reported. Panel C regresses worker outcomes onto the imbalance measures with and without worker characteristics as covariates. The panel reports the coefficients on the imbalance measures and the marginal effects for a 2% increase in ex-post imbalance based on the results without covariates. All specifications include month fixed effects and are estimated by OLS except for the case where the dependent variable is a "quit next month" indicator, which is estimated with logit regressions. *t*-statistics clustered at the month level are reported in paratheses. Significance level: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Panel A. Distribution of ex-post imbalance measure.

		Bootstrapped Critical Values Under the Null Hypothesis		
Statistics	Realized	2.5 Percentile	97.5 Percentile	*p*-value
Mean	0.0001			
Median	-0.0002			
Std. Dev.	0.0119	0.0117	0.0124	0.336
Min	-0.0413	-0.0513	-0.0342	0.838
P5	-0.0190	-0.0197	-0.0179	0.587
P10	-0.0146	-0.0153	-0.0140	0.962
P25	-0.0080	-0.0081	-0.0072	0.185
P75	0.0081	0.0076	0.0086	0.918
P90	0.0153	0.0150	0.0165	0.216
P95	0.0199	0.0197	0.0218	0.134
Max	0.0445	0.0413	0.0619	0.294
Inter-percentile Ranges:				
Max - Min	0.0858	0.0793	0.1059	0.486
P75 - P25	0.0161	0.0150	0.0165	0.394
P90 - P10	0.0299	0.0293	0.0315	0.444
P95 - P5	0.0390	0.0380	0.0410	0.487

Panel B. Balance test.

Worker Characteristics	*F*-statistics	*p*-value
Working status dummies	0.012	0.988
Age	0.004	0.949
Male indicator	1.711	0.223
Working experience (months)	1.913	0.200

Panel C. Worker outcomes.

	Sample Average	Coefficients on the Imbalance Measure		Marginal Effect for a +2% Imbalance (Without Covar.)
		w/o Covar.	w/ Covar.	
Repayment rate	0.183	-0.230***	-0.223***	-0.005
		(-5.07)	(-5.14)	
Performance ranking	0.504	-5.164***	-4.927***	-0.103
		(-6.36)	(-6.63)	
Total salary	4464.30	-23868.8***	-22917.0***	-477.38
		(-5.46)	(-5.77)	
Retention rate at the end of …				
Current month *t*	0.890	-11.03**	-10.43*	-0.021
		(-2.01)	(-1.90)	
Month *t* + 1	0.857	-11.56*	-10.91*	-0.028
		(-1.89)	(-1.86)	
Month *t* + 2	0.843	-13.57*	-13.10*	-0.035
		(-1.86)	(-1.89)	
Month *t* + 3	0.829	-13.43*	-12.96*	-0.037

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 . 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.

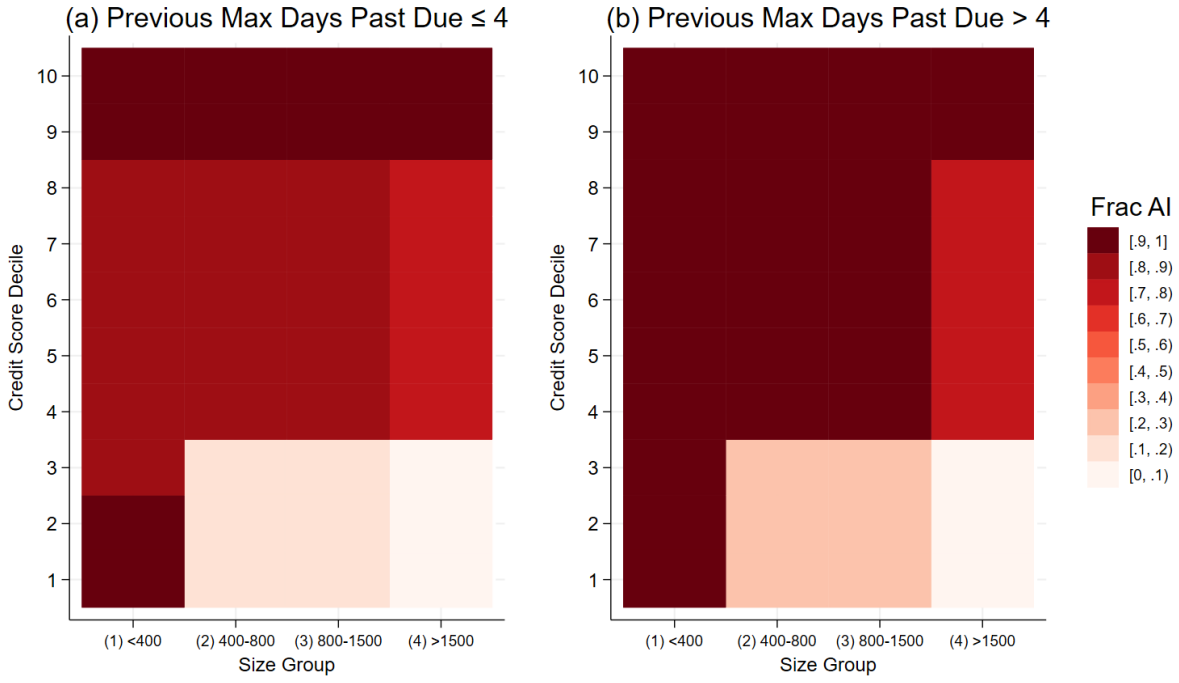
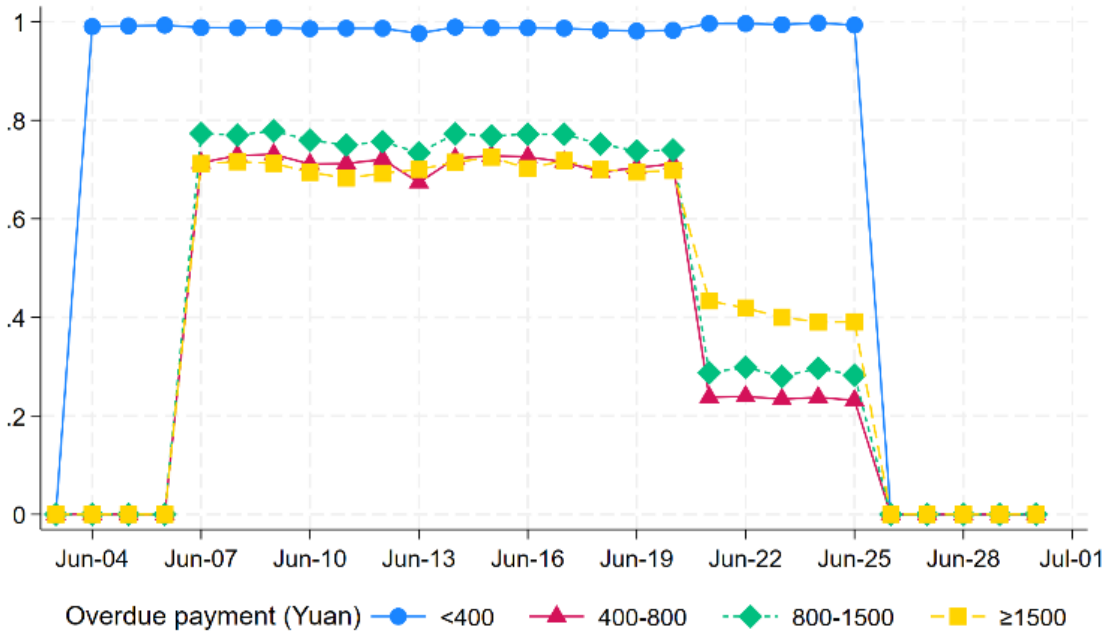


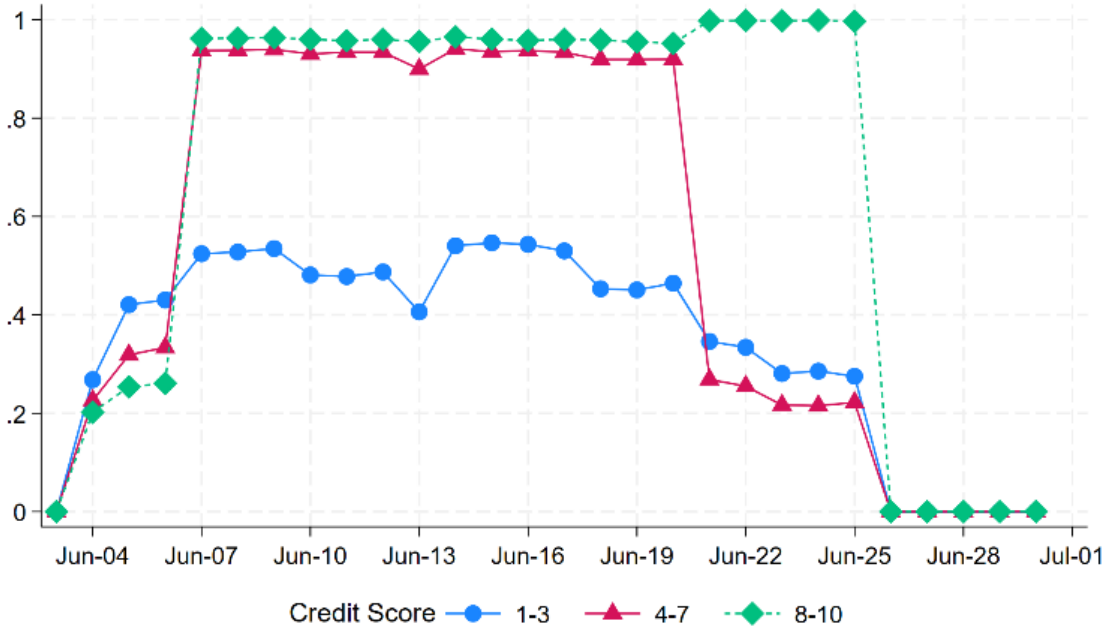
Figure A2. Daily fractions of AI cases in June 2022, by loan characteristics.

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

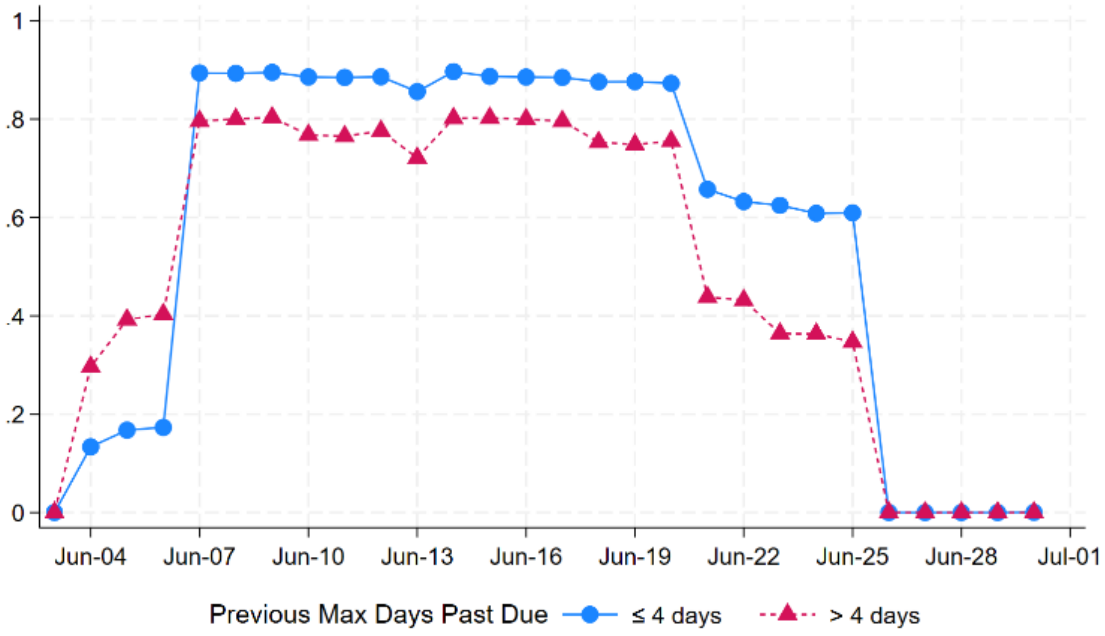
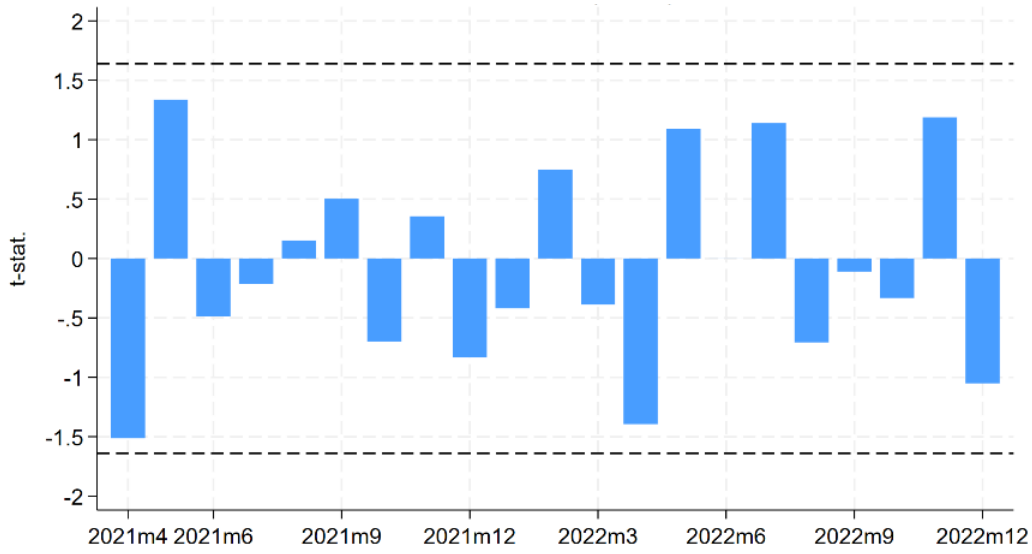


Figure A3. Test of randomization – conditionally randomized subsample.

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.

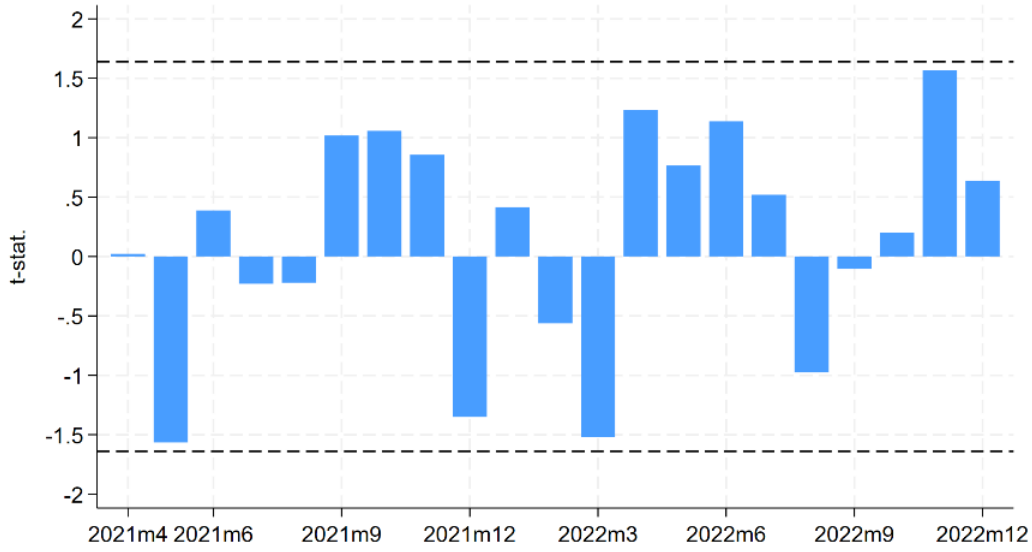
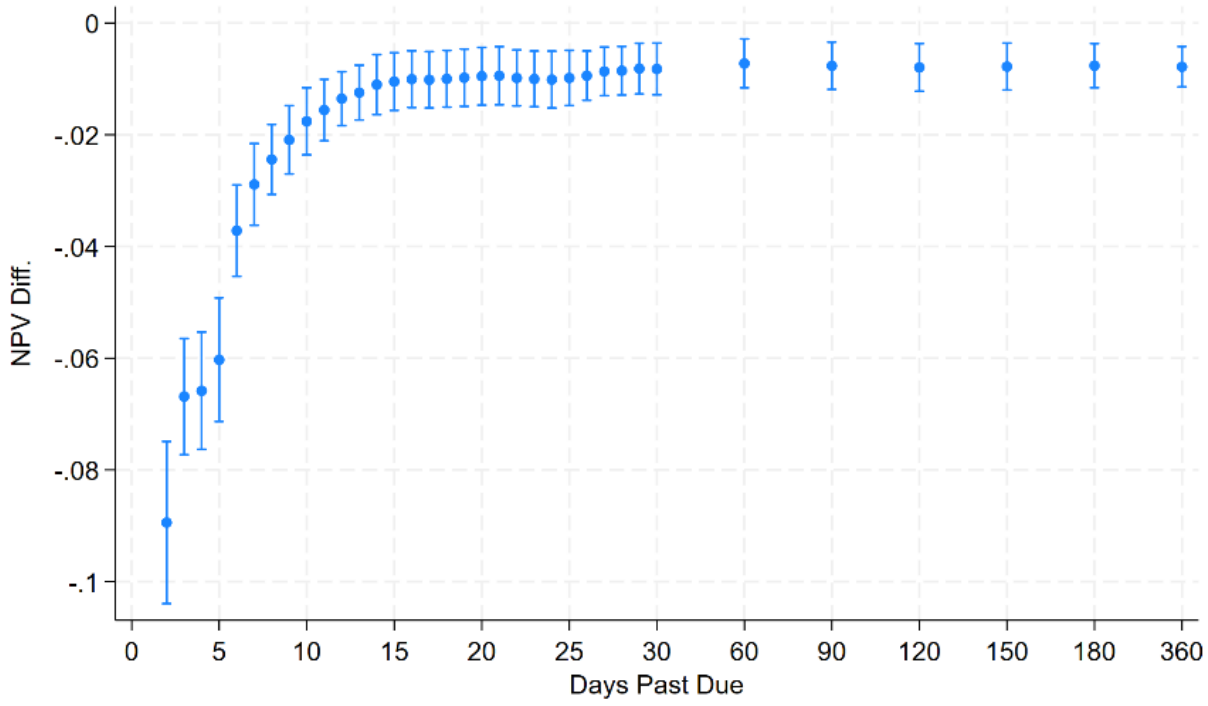


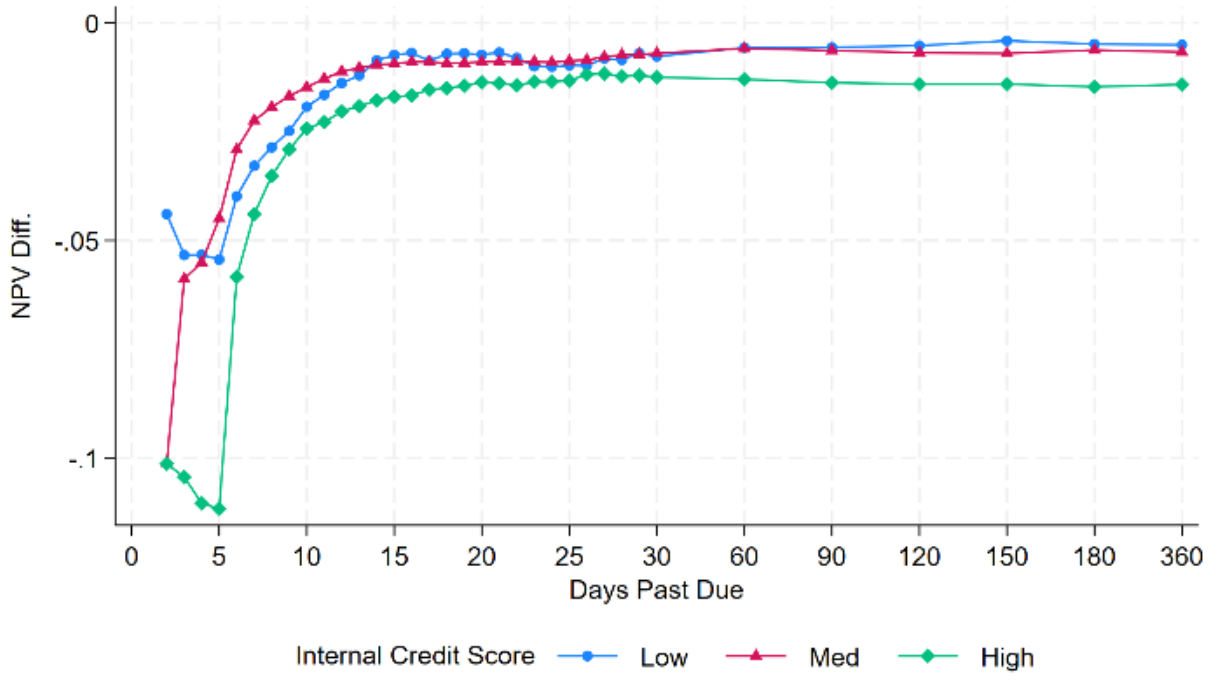
Figure A4. Collected NPV differences between AI and human callers over time – Conditionally randomized subsample.

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.

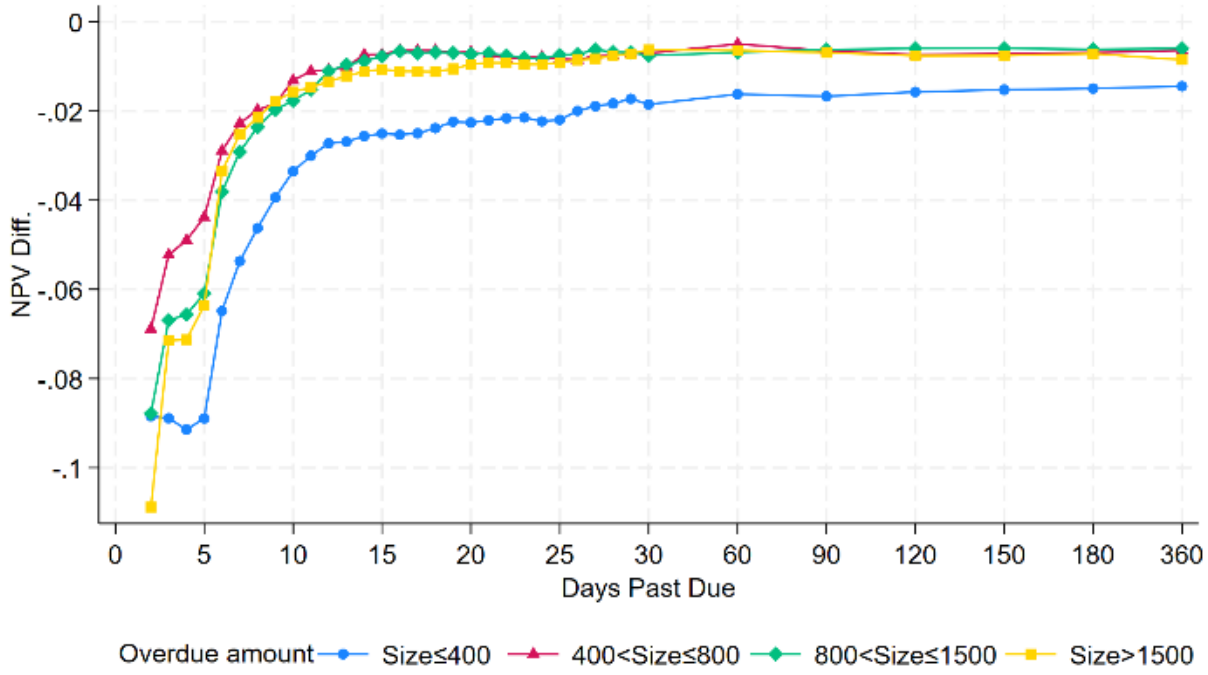


Figure A5. Collected NPV differences between AI and human callers, net of labor costs: Conditionally randomized subsample.

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) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *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 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 . 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 .

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 further away from the cut-off. We therefore conclude that there is no intentional manipulation related to AI caller usage at the 300-yuan cut-off. In the following section, we show that our results are robust to excluding potentially rounded observations.

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 suspected, as well as the sensitivity to the unavoidable extrapolation in local linear regressions. In the test, observations within $\pm w$ 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.

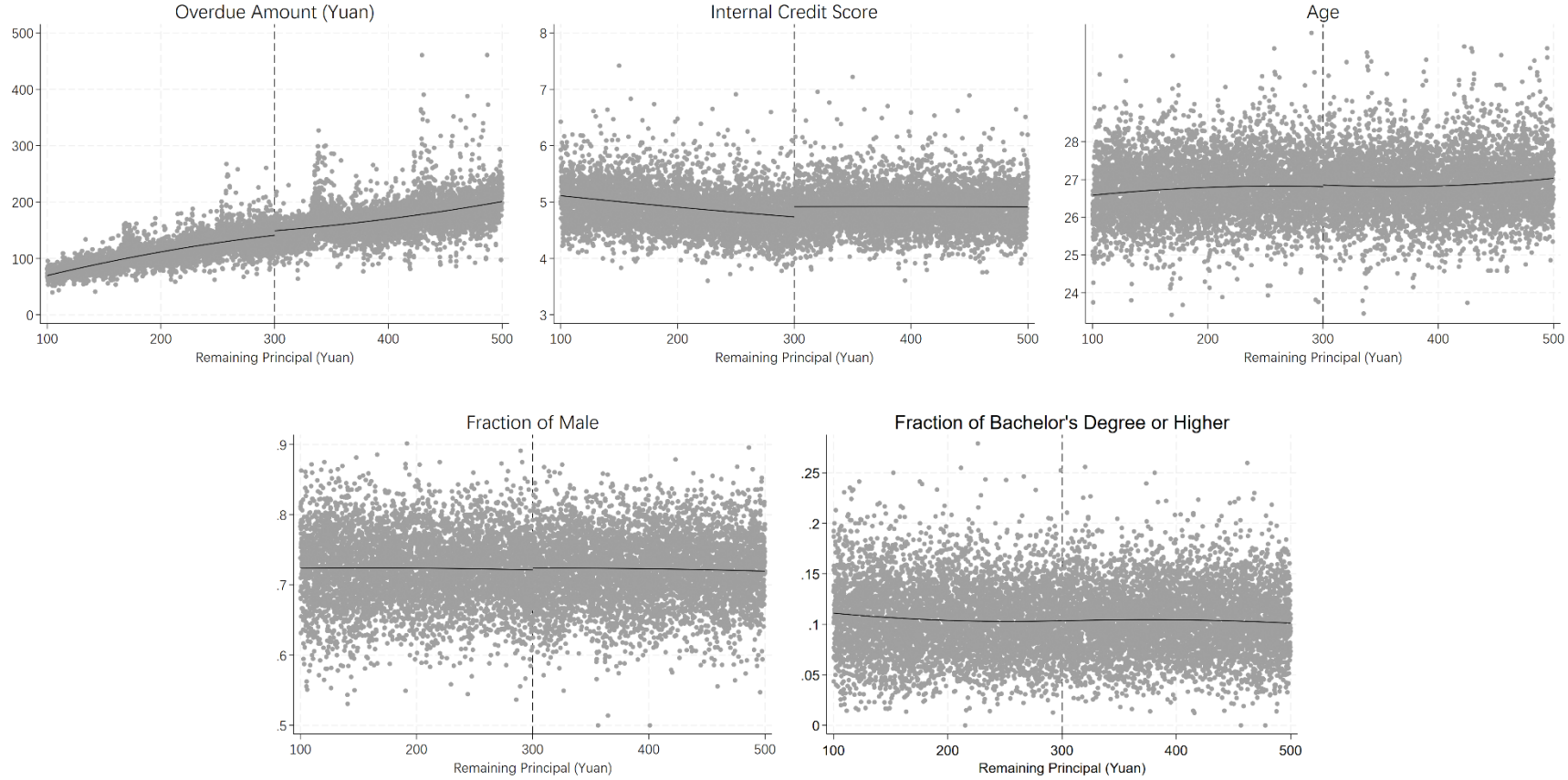
4 Placebo tests

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 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 for more details.

(a) Loan characteristics.



(b) Collected NPV.

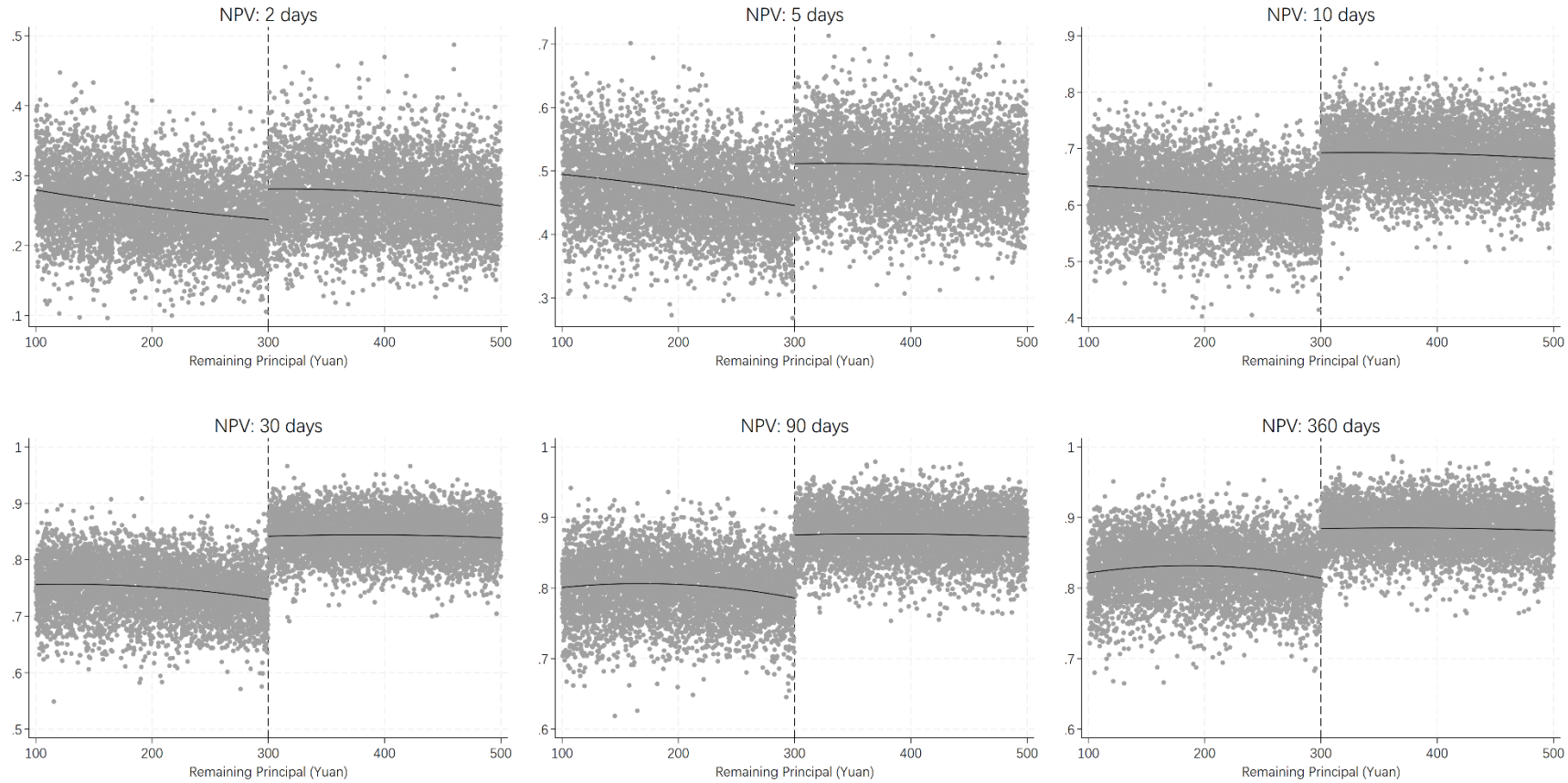
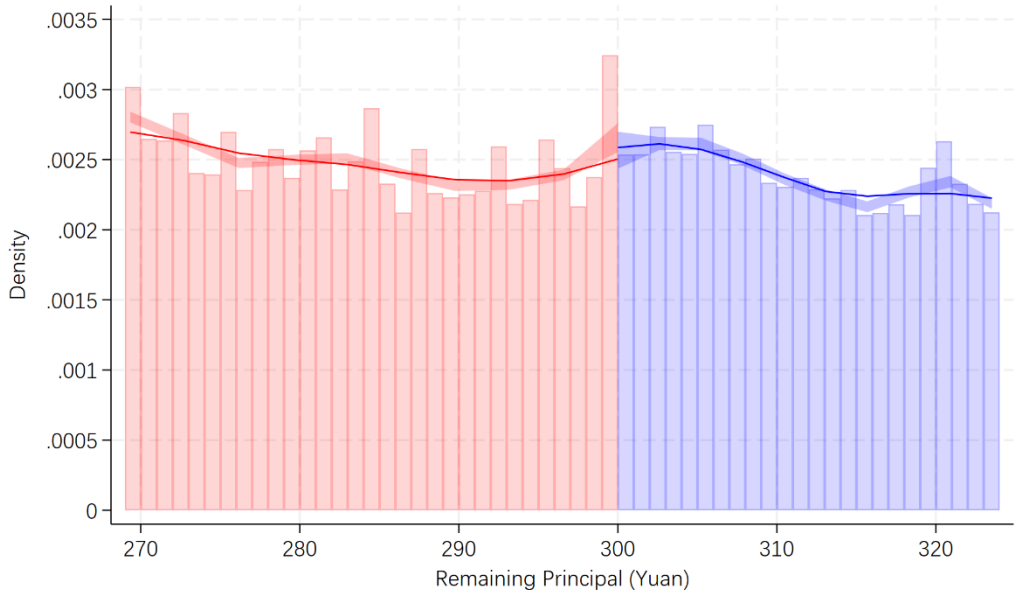


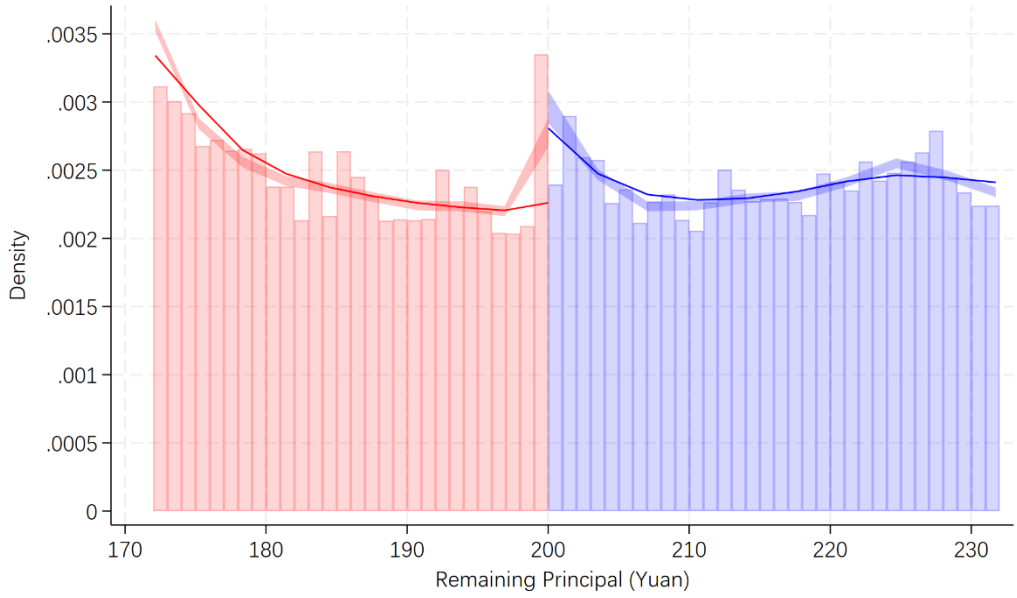
Figure B2. RD density test around the threshold.

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—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.

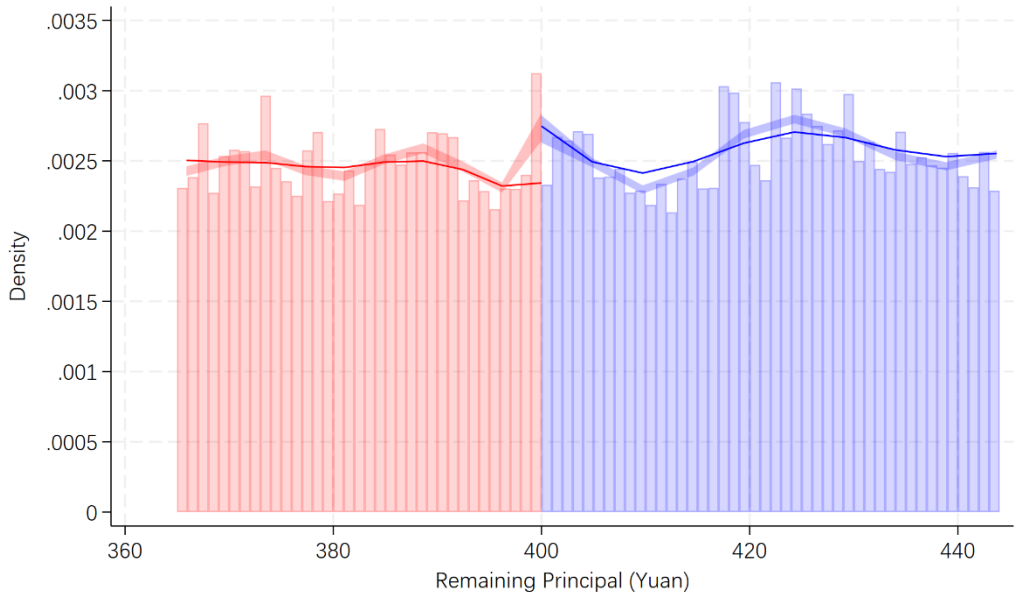


Table B1. Binomial test of manipulation at the threshold.

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 the optimal bandwidths that minimize the mean squared errors and the coverage error probability, 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.

[1] Some callers received salaries lower than the minimum wage or even zero salaries because they left the company in the middle of the month and received only partial salaries proportional to the actual number of working days or borrowers contacted, after any penalty deductions as explained in the text.

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 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. We assume that the fixed salary costs are proportional to labor inputs measured by the total duration of phone calls made by human callers. To estimate unit fixed salary costs, we divide the total ranking salary paid by the company by the total duration of phone calls in minutes made by human callers every month. 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. We then multiply the fixed costs per minute of phone calls by the average length of phone calls per borrower-day to obtain the fixed costs per borrower-day. We allow the average length of phone calls to vary with the number of days past due, as illustrated in Appendix Figure D2, to account for the variation of labor inputs allocated to borrowers in different stages of delinquency. By averaging the total phone call length across borrowers, we assume that the labor input to contact one borrower (i.e., the time to talk to a borrower) is about the same across borrowers with different debt sizes. Figure C5(a) shows the

time series of average ranking salary per minute of phone calls over time, which suggests that the average fixed salary costs to talk to one borrower for one minute is approximately 1.1565 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(b) 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.

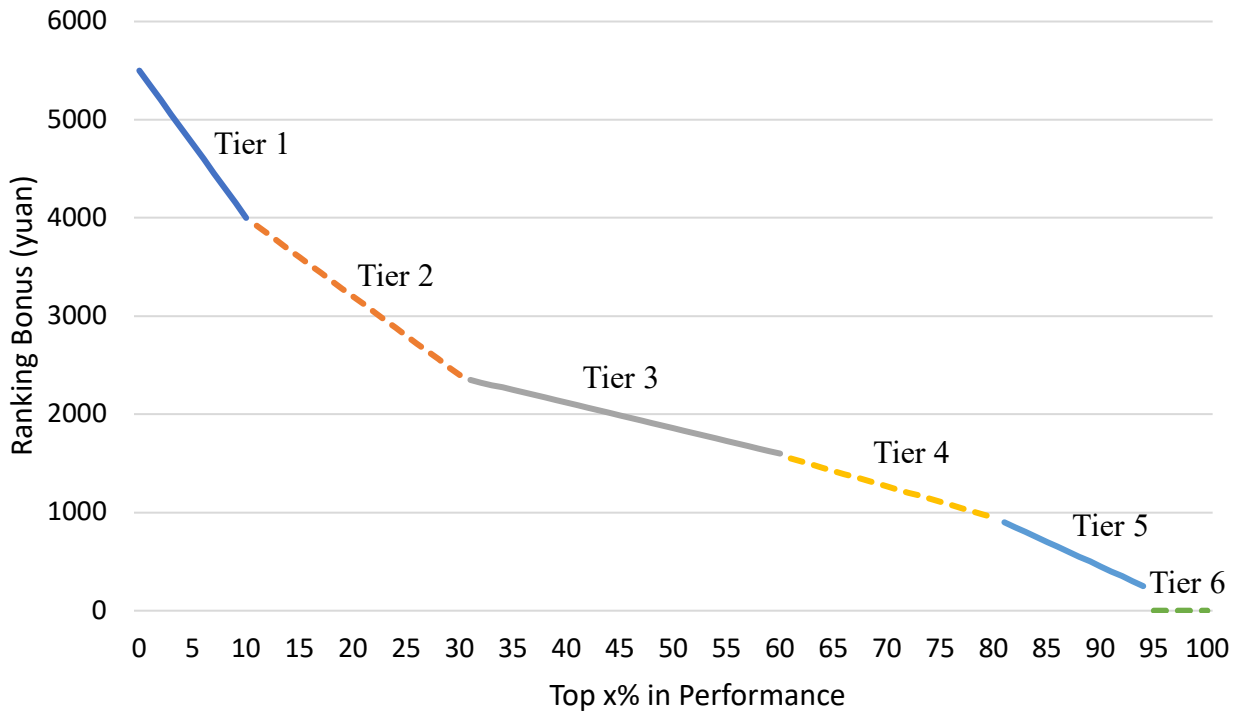


Figure C2. The relationship between completion rate and completion salary.

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.

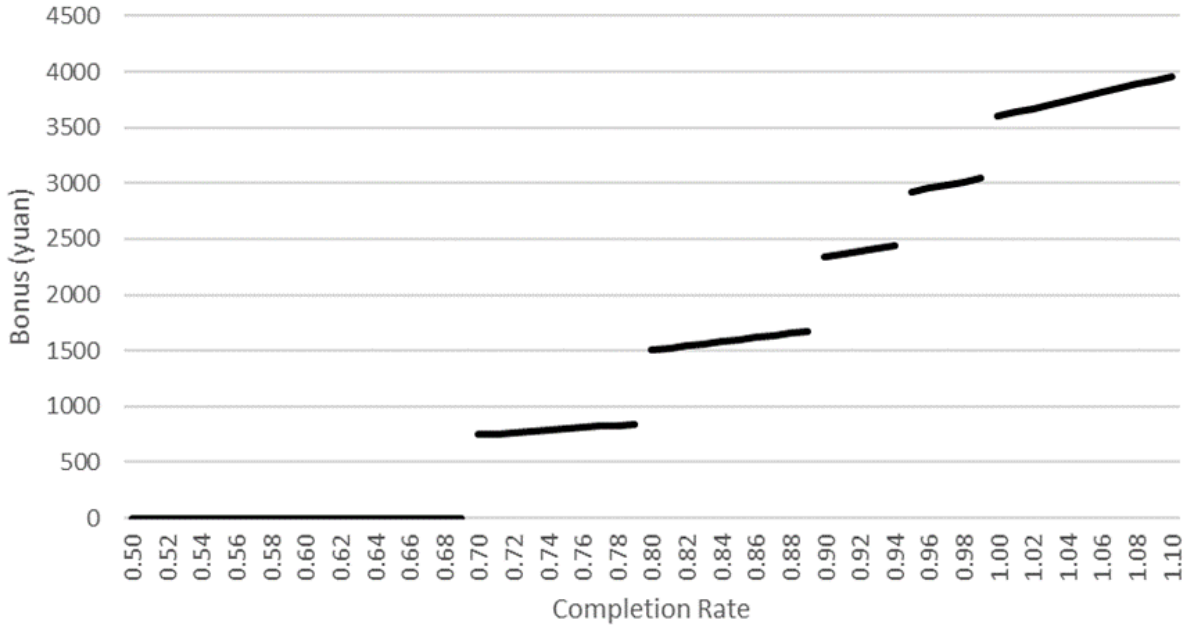
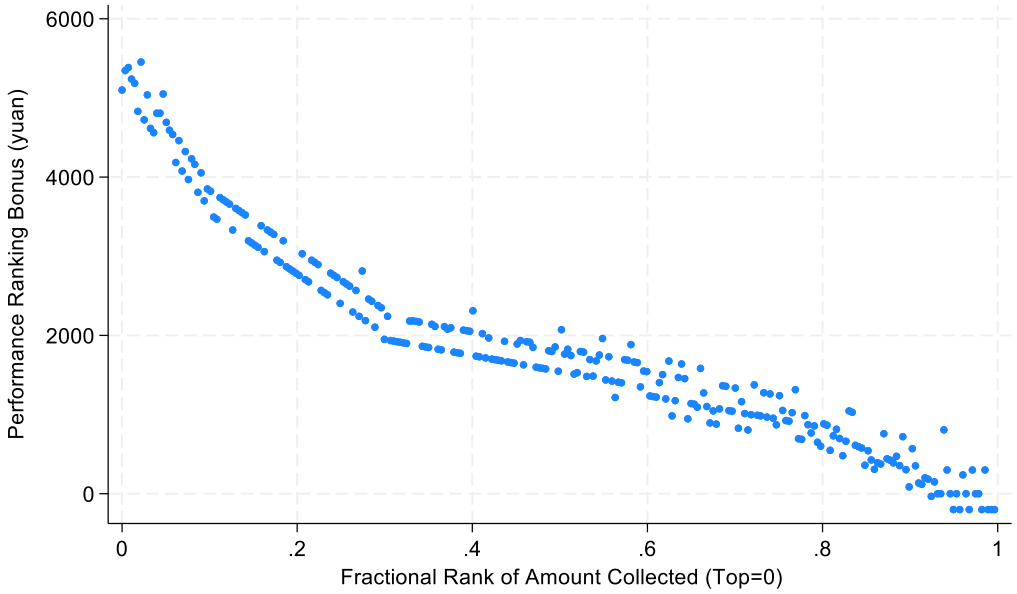


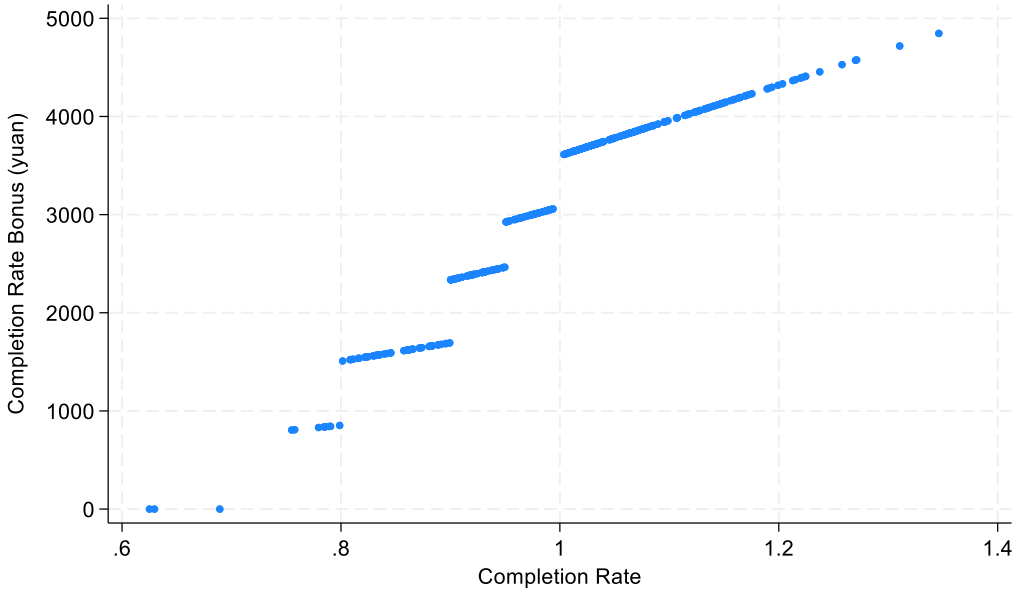
Figure C3. Actual salary received by callers.

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.

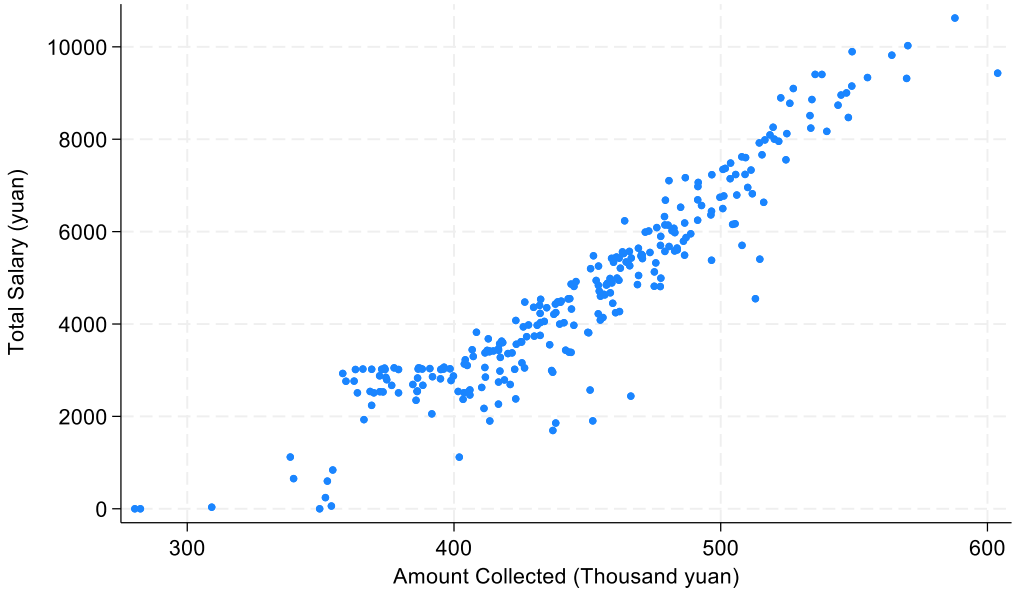
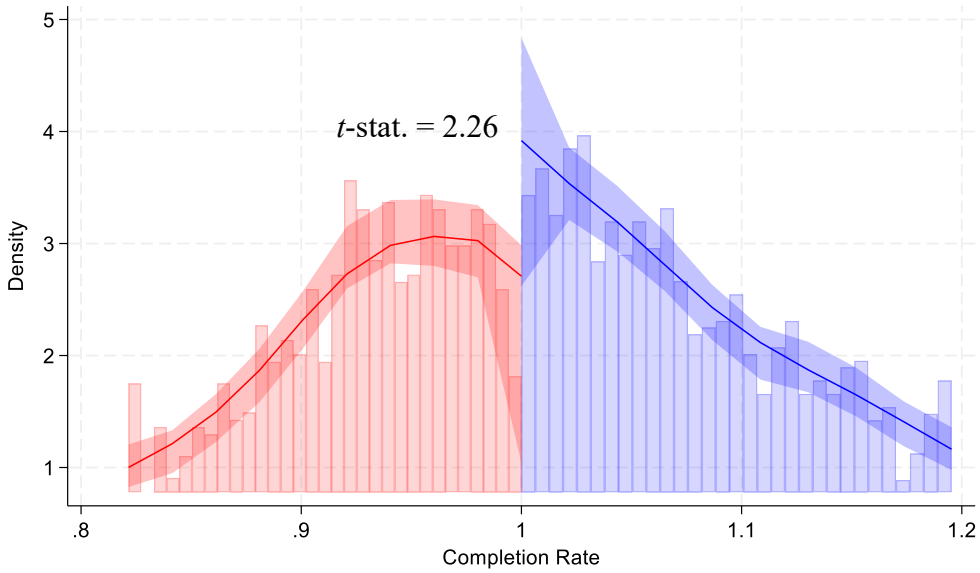


Figure C4. Manipulation of completion rate at the 100% cut-off.

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.

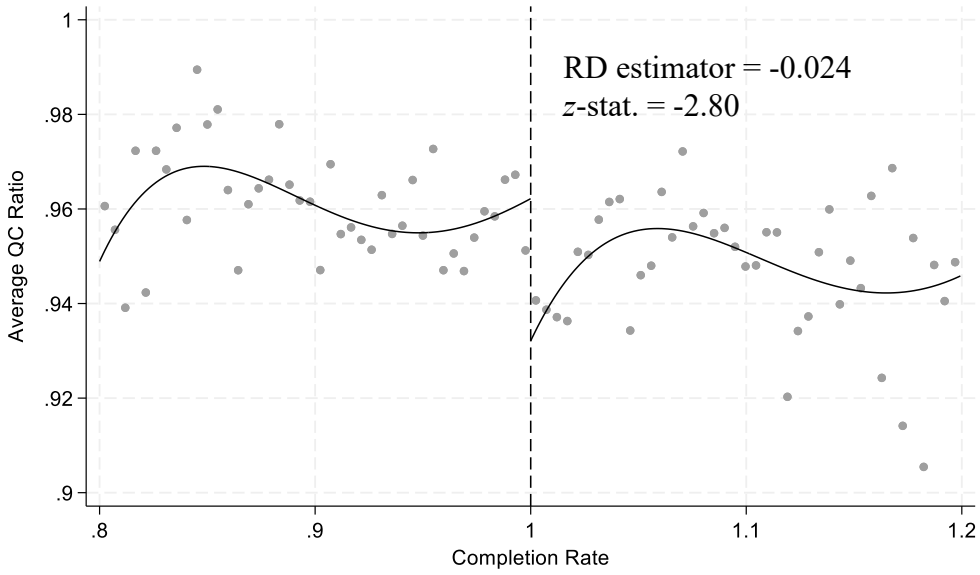
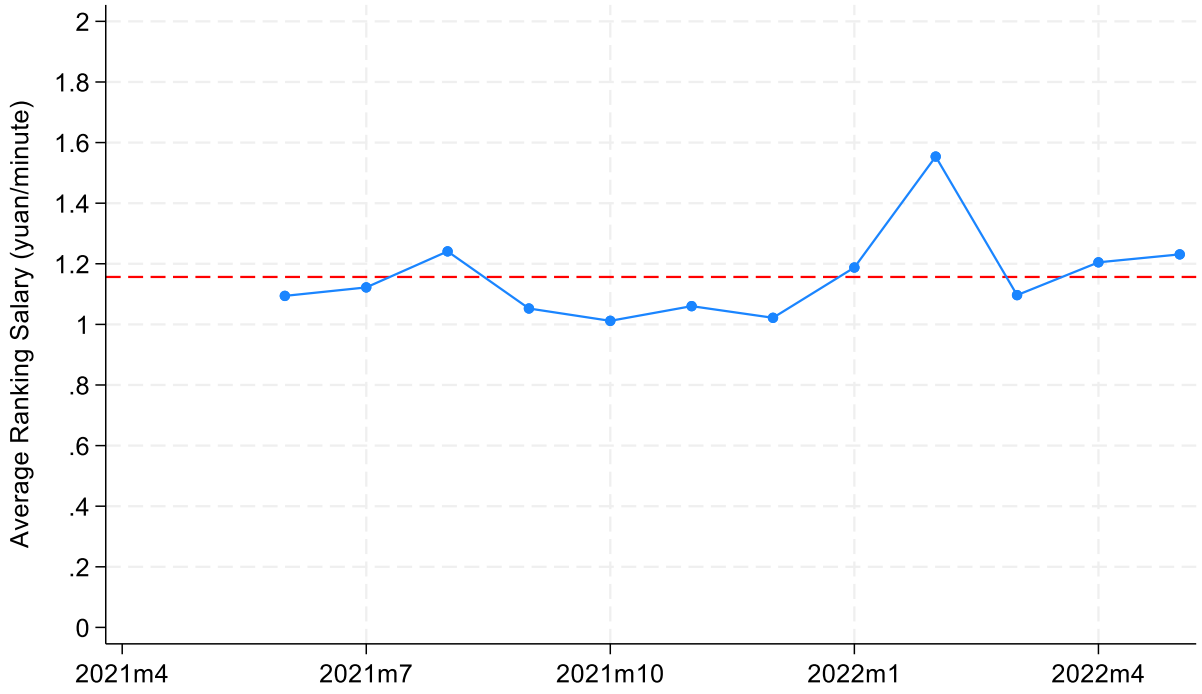


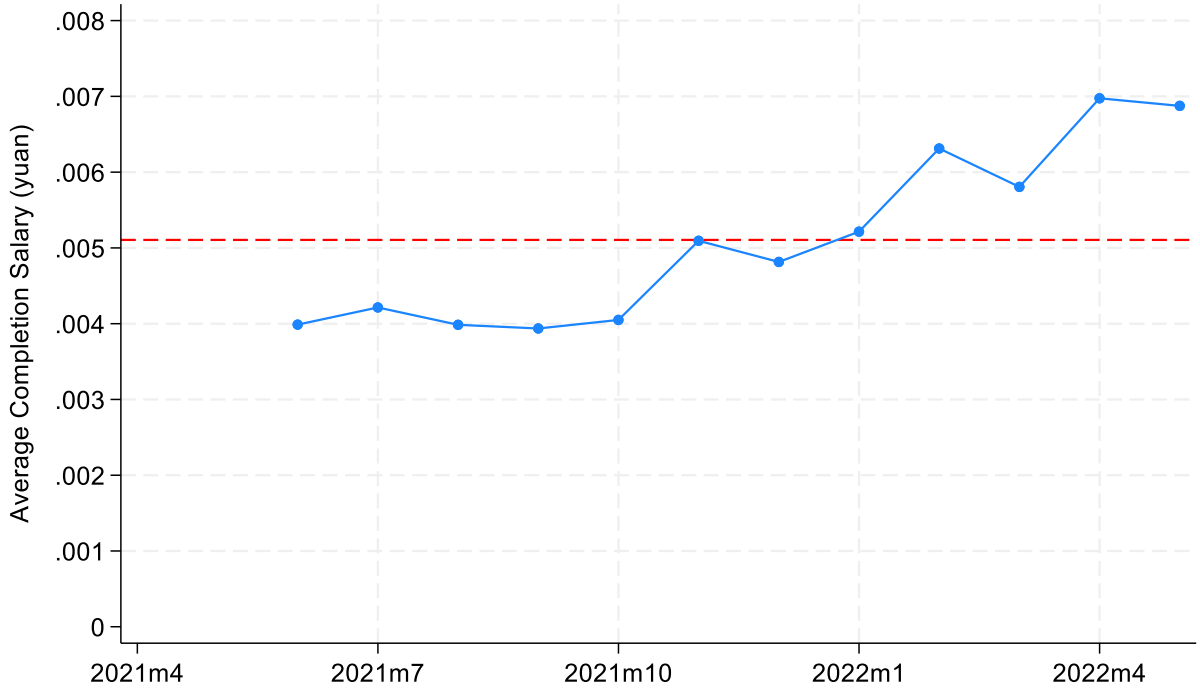
Figure C5. Time series of unit salary costs and unit workloads.

This figure shows the time series of unit salary costs and unit workloads among senior callers in the M1 Early stage. Panel (a) shows the ratio of total ranking salary to the total length of phone calls in every month, that is the average ranking salary paid by the company for handling one delinquent borrower for one minute, or equivalently, the unit fixed labor costs. Panel (b) reports the ratio between the total completion salary and the total money collected, that is, the unit variable labor costs.

(a) Ranking salary per minute of phone calls



(b) 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 criteria 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.

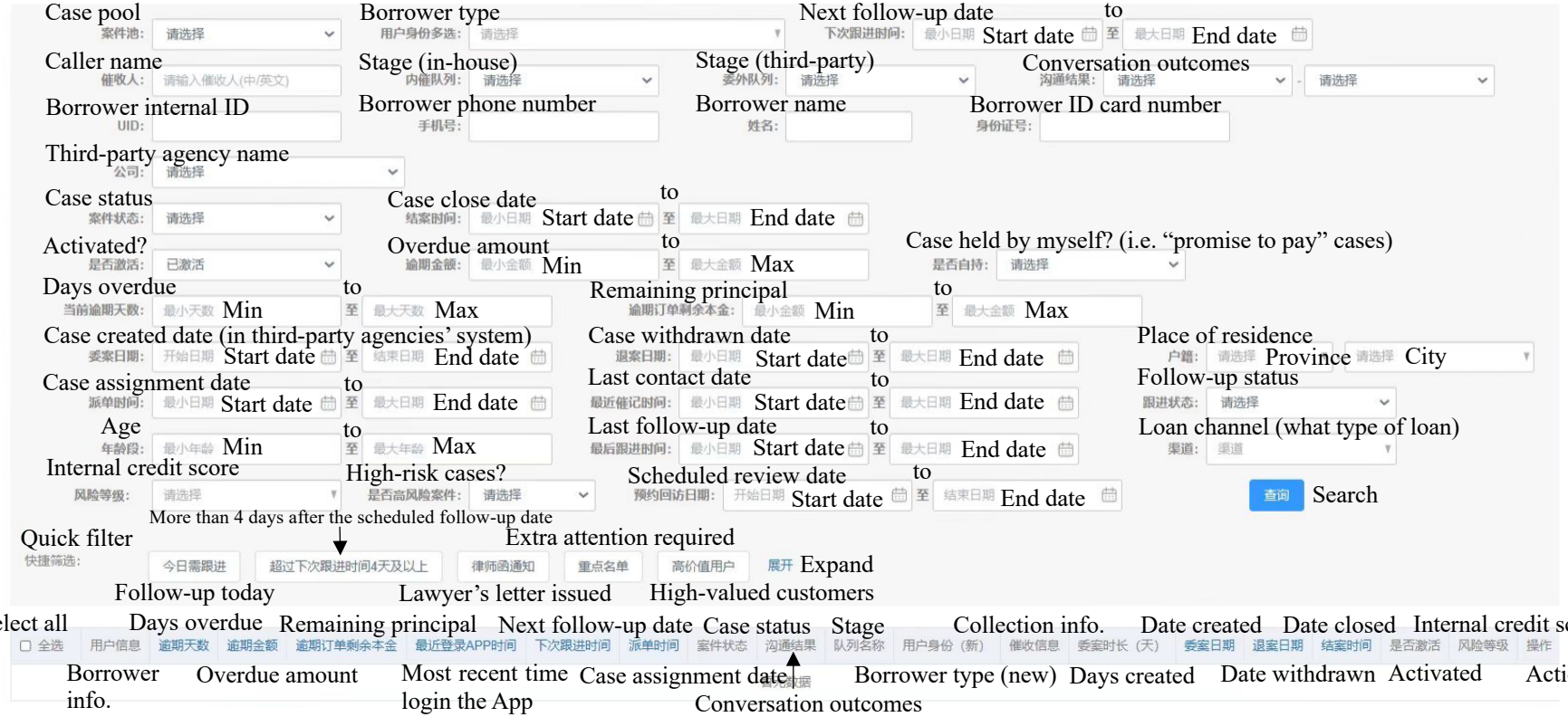


Figure D2. Average phone call length per borrower-day by the number of days after delinquency.

This figure presents the average length of phone calls made by human callers to one borrower in one day, as a function of the number of days after the due date up to one year. The spikes occur on days 6, 11, 26, and 60 when the stage of the debt collection process changes.

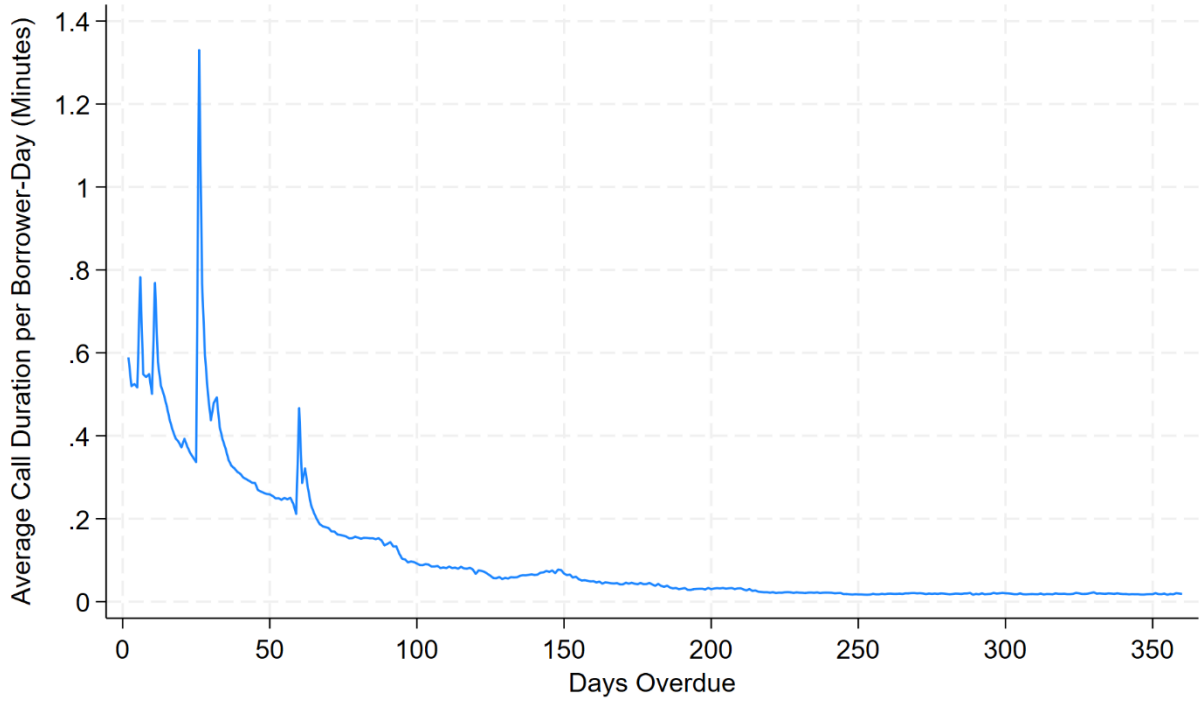


Figure D3. Undiscounted collected cash flows differences between AI and human callers over horizon – small cases RDD.

This figure reports the average differences of the sum of undiscounted collected cash flows, scaled by the initial overdue balance, between AI and human callers over the horizon of days past due of cases. The differences are estimated by RDD utilizing the 300-yuan remaining principal threshold for almost permanent AI treatment. The triangles connected by solid lines represent the average differences estimated by RDD. As a reference, the dots connected by dashed lines represent the estimations with NPV, the same as the estimations in Figure Panel (b). For clarity, the differences are plotted every three days before day 60, and every 10 days after day 60.

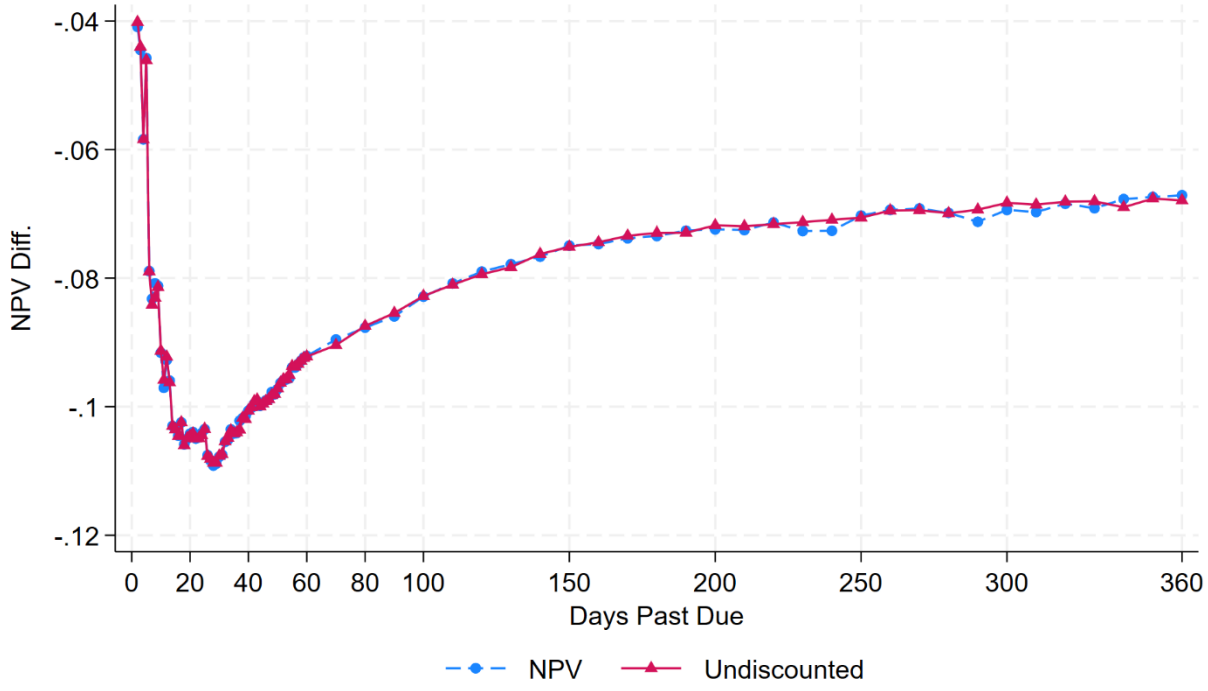


Figure D4. Undiscounted collected cash flows differences between AI and human callers over horizon – Completely randomized subsample.

This figure reports the average differences of the sum of undiscounted collected cash flows, scaled by the initial overdue balance, between AI and human callers over the horizon of days past due, using the 10% completely randomized subsample. The differences are estimated by *t*-tests on collected cash flows between the two groups of callers. The triangles connected by solid lines represent the average differences estimated with undiscounted cash flows. As a reference, the dots connected by dashed lines represent the estimations with NPV, the same as the estimations in Figure Panel (a). For clarity, the differences are plotted daily before day 30, and every 30 days afterwards.

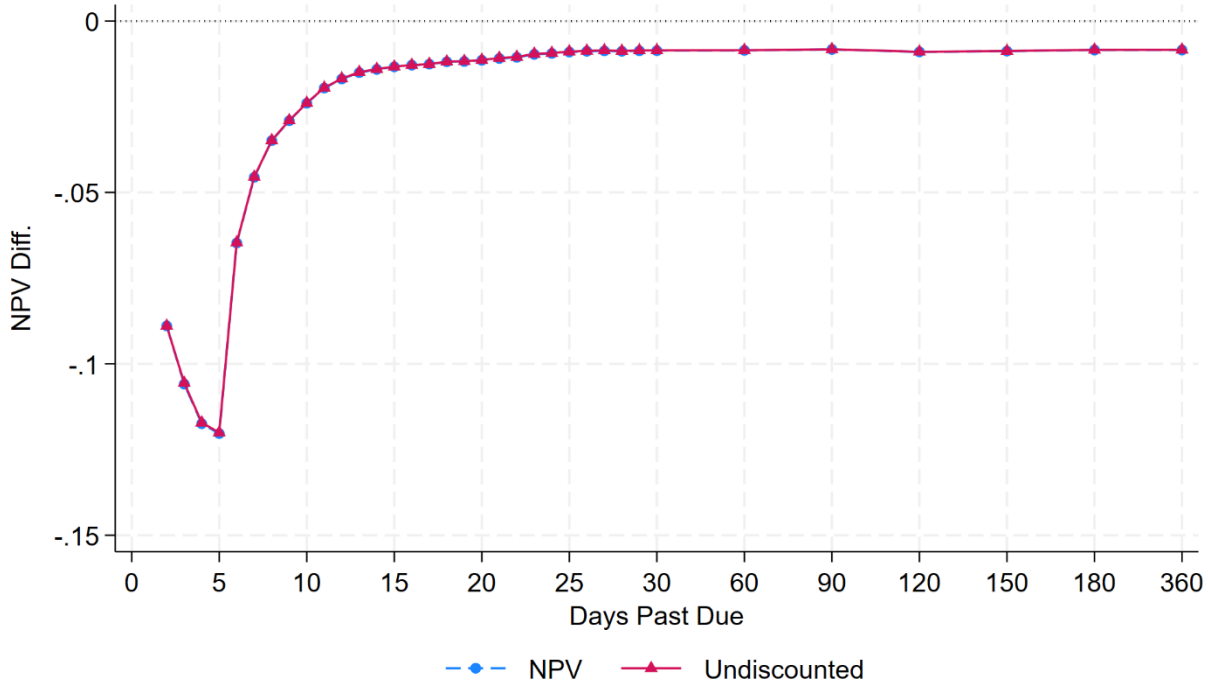


Figure D5. Distribution of phone call duration.

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.

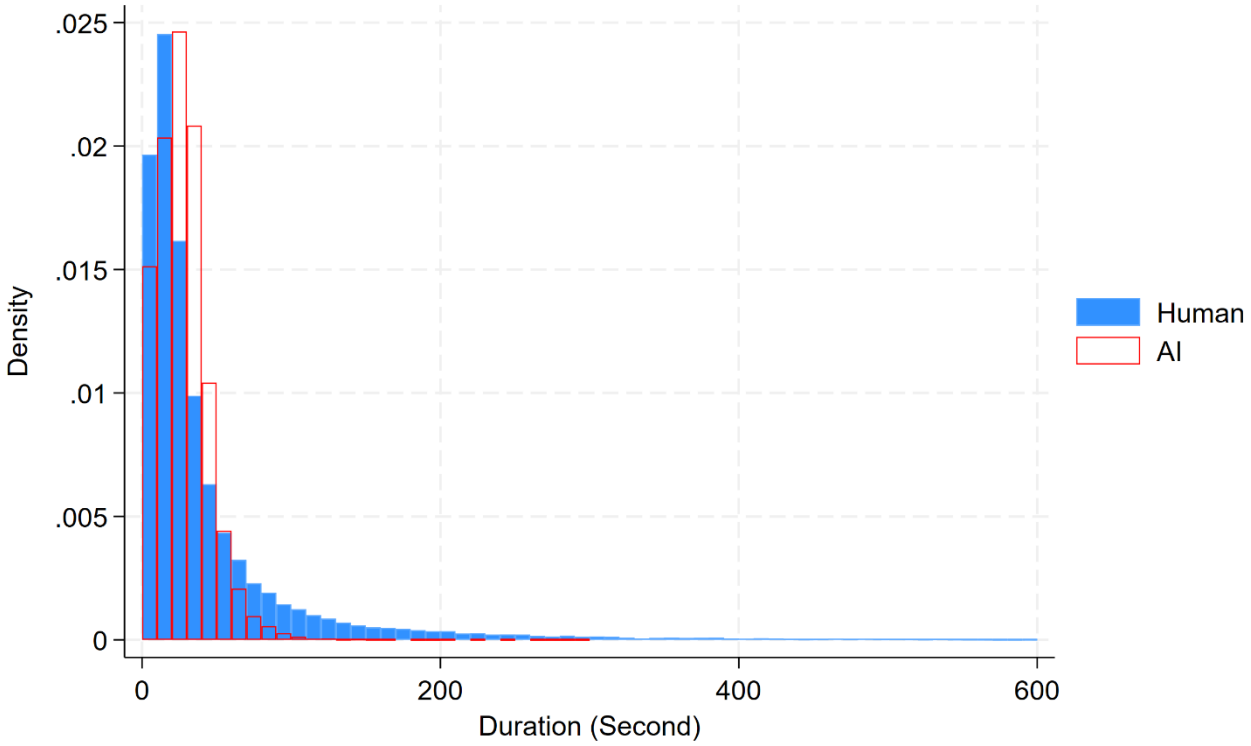


Table D1. Comparison between permanent AI callers and human callers regarding undiscounted cash flows - small cases RDD results.

This table compares the 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). The performance is measured by the *undiscounted* sum of collected cash flows within a given horizon and scaled by the initial overdue balance. See the note in Table 3 for the descriptions of other specifications. Significance level: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Variable	Left Mean (AI)	Right Mean (Human)	Diff. (L–R)	*z*-stat.	*p*-val.	95% Robust RD C.I.	
NPV 2d	0.238	0.278	-0.040***	5.86	<0.001	0.025	0.050
NPV 5d	0.451	0.498	-0.046***	6.45	<0.001	0.030	0.057
NPV 10d	0.596	0.687	-0.091***	13.92	<0.001	0.077	0.102
NPV 30d	0.735	0.842	-0.108***	18.85	<0.001	0.096	0.119
NPV 60d	0.779	0.871	-0.092***	17.86	<0.001	0.083	0.103
NPV 90d	0.793	0.878	-0.085***	17.52	<0.001	0.077	0.096
NPV 180d	0.812	0.885	-0.073***	16.53	<0.001	0.065	0.083
NPV 360d	0.820	0.888	-0.068***	15.05	<0.001	0.060	0.078

Table D2. Difference between AI and human callers regarding undiscounted cash flows – Completely randomized subsample.

This table compares the 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. The performance is measured by the *undiscounted* sum of collected cash flows within a given horizon and scaled by the initial overdue balance. See the note in Table 4 for the descriptions of other specifications. 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.
NPV 2d	0.193	0.282	-0.089***	-42.33
NPV 5d	0.430	0.550	-0.120***	-48.62
NPV 10d	0.646	0.670	-0.024***	-10.14
NPV 30d	0.767	0.776	-0.0086***	-4.14
NPV 60d	0.800	0.809	-0.0086***	-4.38
NPV 90d	0.816	0.824	-0.0082***	-4.38
NPV 180d	0.830	0.838	-0.0084***	-4.63
NPV 360d	0.836	0.844	-0.0084***	-4.69

Table D3. 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 10. 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