Debt Collection Negotiations with Large Language Models: An Evaluation System and Optimizing Decision Making with Multi-Agent

Xiaofeng Wang[1*], Zhixin Zhang[2†], Jinguang Zheng[2], Yiming Ai[1], Rui Wang[1†]
[1]Shanghai Jiao Tong University
[2]Ant Group
{banyedy, wangrui12}@sjtu.edu.cn
{zhangzhixin.zzx, zhengjinguang.zhen}@antgroup.com

Abstract

Debt collection negotiations (DCN) are vital for managing non-performing loans (NPLs) and reducing creditor losses. Traditional methods are labor-intensive, while large language models (LLMs) offer promising automation potential. However, prior systems lacked dynamic negotiation and real-time decision-making capabilities. This paper explores LLMs in automating DCN and proposes a novel evaluation framework with 13 metrics across 4 aspects. Our experiments reveal that LLMs tend to over-concede compared to human negotiators. To address this, we propose the **M**ulti-**A**gent **De**bt **N**egotiation **(MADeN)** framework, incorporating planning and judging modules to improve decision rationality. We also apply post-training techniques, including DPO with rejection sampling, to optimize performance. Our studies provide valuable insights for practitioners and researchers seeking to enhance efficiency and outcomes in this domain. Our work is accessible at https://github.com/banyedy/DCN.

1 Introduction

Finance, as a negotiation-intensive field, involves the distribution and exchange of financial interests, requiring a higher level of understanding of information and rational decision-making (Chan, 2006; Thompson, 1997). Due to various personal financial issues, a large volume of non-performing loans (NPLs) arises each year across banks and financial companies, with debtors often being unable to repay their debts after prolonged overdue periods (Ozili, 2019). Negotiation and mediation are necessary to resolve their credit issues and minimize the losses for financial institutions (creditors) (Firanda et al., 2021). Traditionally, the debt collection process has been labor-intensive, and data shows that in China, 3,800 financial institutions rely on outsourced specialized collection agencies to help recover non-performing assets (Tang et al., 2018).

Previous automated debt collection dialogue models (Floatbot.ai, 2023; Yahiya and Ahmad, 2024) were primarily based on fixed-format notifications, where the models lacked communication and negotiation capabilities. Additionally, automated decision models (Sancarlos et al., 2023; Jankowski and Paliński, 2024) related to changes in repayment strategies could not be directly integrated into the dialogue and were unable to update decisions in real time based on the debtor's information provided during the conversation. A pressing need exists for novel approaches to automate **debt collection negotiations (DCN)**.

The rapid development of large language models (LLMs) (Peng et al., 2023; Touvron et al., 2023) and agent-based interactions (Luo et al., 2025; Chai et al., 2025) built upon them has made it possible. Through emerging functions such as planning (Huang et al., 2024), reasoning (Aksitov et al., 2023), and reflection (Renze and Guven, 2024), these models are now able to assist humans in completing more complex tasks. In this paper, we aim to explore the potential of using LLMs to support AI agents in performing this unexplored task. And firstly, it is crucial to develop a method to evaluate the performance in conducting DCN.

To develop a benchmark, the primary challenge lies in constructing a suitable dataset. In Section 2, to ensure both privacy and data validity, we utilized CTGAN (Xu et al., 2019) to generate synthetic data based on debt records from a leading financial technology company [1]. We supplement the debtor's personal financial data through extraction and construction. Finally, we constructed a dataset

*Work done as an intern at Ant Group.
†Corresponding Authors

[1]The synthetic data generated in this work is available at https://huggingface.co/datasets/clement520/LoanDefault-zh.

Findings of the Association for Computational Linguistics: ACL 2025, pages 9555–9577
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

containing 975 debt records. Based on this information, we provide a complete definition for DCN and the LLM-based negotiation process in Section 3.

To comprehensively evaluate DCN, in Section 4 we proposed a holistic assessment framework encompassing **10** specific metrics in 4 aspects and **3** comprehensive indices to thoroughly evaluate the negotiation process and its outcomes. From the perspective of the negotiation process, we evaluate the completeness and soundness of the dialogues. Regarding negotiation outcomes, we evaluate two key aspects: for creditors, we focus on debt recovery rate and collection efficiency, while for debtors, we assess financial health by predicting future asset changes based on negotiation results and individual financial data. The indices are introduced to integrate the opposing relationship between creditor's interests and debtor's financial health.

In Section 5, we tested the performance of LLMs and found that they are unable to make appropriate decisions based on the debtor's financial condition and are more likely to make unsuitable *concessions* than human beings. This may result from the models' excessive focus on harmony and agreement, leading creditors to overlook the rationality of decisions. To address this, inspired by the work of MetaGPT (Hong et al., 2023), we designed an LLM-based **M**ulti-**A**gent **De**bt **N**egotiation **(MADeN)** framework for DCN in Section 6.1. In this framework, we enhanced the basic **Communicating** agent with two additional modules: **(1) Planning**, where the LLM agent designs a rough decision framework and outlines the potential outcomes based on the debtor's initial reasons and demands; **(2) Judging**, which evaluates the rationality of each action and provides optimization suggestions. Our method improves the comprehensive collection index by **6.24%**.

In addition, we attempted to use the post-training method including DPO (Rafailov et al., 2024) with reject sampling (Liu et al., 2024) to align the debt collector's focus on recovery rate and efficiency in Section 6.3. On the Qwen2.5-7B (Yang et al., 2024) model, we observed improvements across various metrics.

Our contributions are summarized as follows:

- We proposed a synthetic debt dataset and a comprehensive framework for evaluating LLM performance in debt collection negotiations (DCN), using 13 metrics to assess both the negotiation process and outcomes, enabling the testing and evaluation of different models.
- Our testing of mainstream LLMs on this task revealed that the models tend to make decisions with more unreasonable concessions compared to humans.
- We developed an multi-agent framework for DCN, incorporating two key modules to improve negotiation outcomes. Additionally, we explored post-training the model through rejection sampling on multi-agent data, which also enhanced the model's performances.

2 Data Collection

Our data is primarily divided into two parts, as shown in Figure 1. The basic debt information is known to both the debtor and the creditor, while the debtor's personal financial data is not accessible to the creditor. We now explain how each of these two data components was collected.

2.1 Basic Debt Information

The basic debt data primarily consists of personal information and debt-related information. We sampled from *real debt data* provided by the financial company mentioned in introduction. To ensure privacy compliance, we used **CTGAN** (Xu et al., 2019) to generate synthetic data [2]. We categorized the data by gender, overdue days and loan amount. Then we sampled it to match the distribution patterns of the original data.

2.2 Debtor's personal financial data

Debtor's personal financial data collection involved two main components: **textual reasons for overdue** and **numerical financial information**. The reasons for overdue payments were extracted from real dialogue data and assigned to different categories based on their real *distribution*. For numerical financial data, we simplified complex personal data into components such as total assets, average daily income, expenses, and surplus. Since this data is typically unavailable, we used a linear model with Gaussian noise, based on historical data correlations, to estimate these values.

Finally, we collected **975** debt records, with **390** records placed in the test set and the remaining **585** records in the training set (The subsequent evaluations are conducted on the test set). Details of the debtor category distribution can be found in Appendix B.

[2] Please refer to our Ethical Considerations.

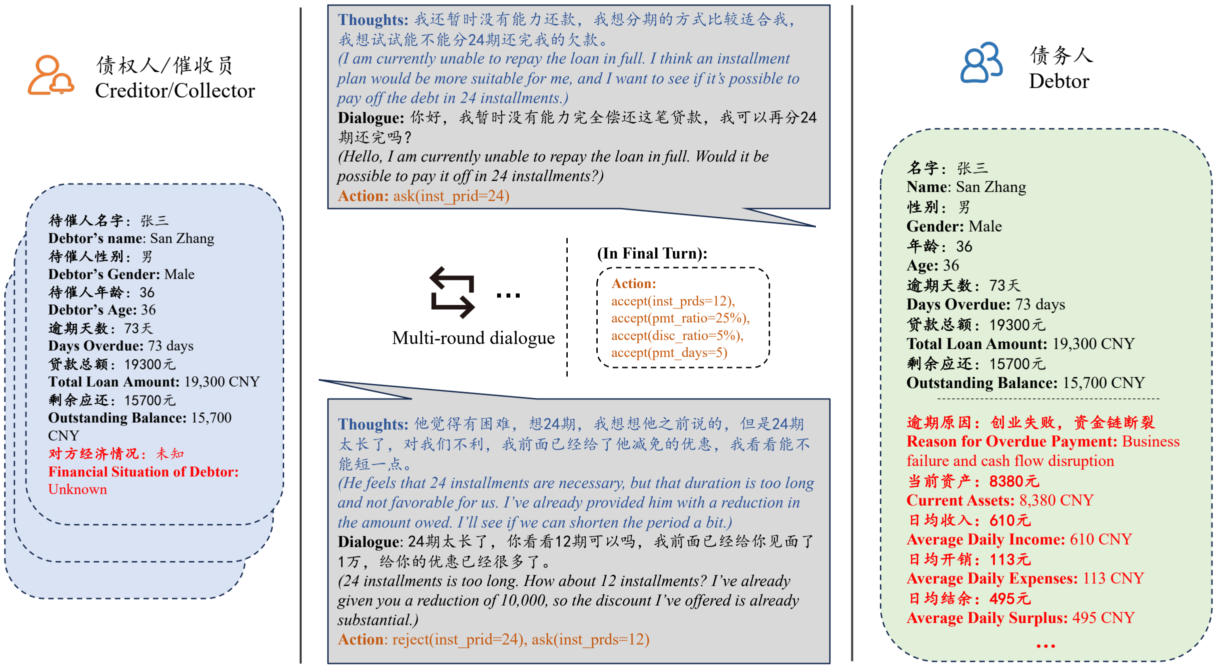

Figure 1: An Example of a Debt Collection Negotiation (DCN). On the left and right sides are the information cards representing the data controlled by the debtor and the creditor, respectively. The black text represents the **basic debt information**, while the red text represents the **debtor's personal financial information**. In the center, we demonstrate the use of LLM-based agents to simulate the dialogue. Each time, both the debtor and the creditor output a set of (Thoughts, Dialogue, Action). **Thoughts** refers to their internal thought process, visible only to themselves; **Dialogue** represents the conversation in natural language; and **Action** refers to the specific activities represented in a formal language within the dialogue. Each negotiation consists of multiple rounds of such interactions, ultimately leading to the negotiation outcome. The English text was automatically translated using Google Translate.

3 Task Formulation

3.1 Definition and Objectives

Debt collection negotiations (DCN) refers to negotiations initiated by creditors to recover outstanding debts and restore the debtor's credit, due to the debtor's inability to repay on time because of personal financial issues. The measures for negotiating the resolution of non-performing loans generally include deferral, debt forgiveness, collateralization, conversion, and installment payments (Rating, 2016; Lankao County People's Government, 2024). Among these, deferral, debt forgiveness, and installment payments are the most commonly used. We have distilled them into four dimensions: Discount Ratio, Immediate Payment Ratio, Immediate Payment Time and Installment Periods[3]. Table 1 presents the range of values and a brief description of each dimension, and the detailed explanations are provided in Appendix A. Through negotiations on these four aspects, the goal of both parties is to reach a *mutually acceptable outcome* that allows the debtor to resolve their outstanding debt in a manageable way.

3.2 Future Economic Predictions for Debtors

After obtaining the negotiation results and integrating them with the debtor's current financial model, we can project changes in their assets and remaining debt over the next *two years*. Figure 2 shows one debtor's economic trajectory under three installment scenarios. In one scenario, the debtor's assets fall into negative values, indicating a failed negotiation. In another, a too lenient installment plan reduces recovery efficiency. These scenarios provide a basis for evaluating negotiation outcomes, which will be discussed in Section 4.

3.3 Negotiation Process

As shown is Figure 1, our negotiation process is a variant of the bargaining process designed by Xia et al. (2024). To formally articulate the negotiation between agents, we define the relevant concepts and variables in Table 8. A brief pseudo code of the process is Algorithm 1.

In the action set, *"ask"*, *"reject"* and *"accept"* represent three different operations for each nego-

[3]Refer to `https://www.boc.cn/bcservice/bc3/bc31/201203/t20120331_1767028.html` for the calculation of installment interest.

Table 1: Debt Collection Negotiation Dimensions

Dimension	Range	Description
Discount Ratio	0 - 30%	The portion of debt waived by the creditor to ease repayment.
Immediate Payment Ratio	5% - 50%	The portion of debt that must be repaid immediately, typically at least 5%.
Immediate Payment Time	1 - 14 (days)	A grace period of up to 14 days for the debtor to make the immediate repayment.
Installment Periods	3 - 24 (months)	The duration for repaying the remaining debt in installments.

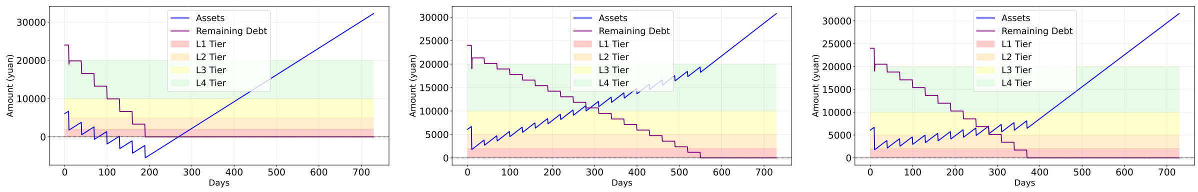

Figure 2: The future trajectories of the debtor's remaining assets and outstanding debt under three installment plans (6, 12, and 18 months from left to right) are shown, with all other variables held constant. The 6-month plan causes the debtor's assets to fall **below zero**, making repayment impossible. In contrast, the 12-month and 18-month plans maintain a healthy asset level, though the 18-month plan significantly **reduces recovery efficiency**. The 12-month plan is the most balanced solution. Different background colors represent five difficulty tiers, with Tier 1 being the most challenging. The specific ranges and descriptions of the tiers are provided in Appendix D.

Algorithm 1 Debt Collection Negotiation Process

Initialize: Action Set S_A, Basic Debt Information I_b, Personal Financial Information I_p, Agent Creditor, Agent Debtor, Maximum Turns t_m, Negotiation Dimensions Set S_R, Negotiation Result Dictionary D
Creditor $\leftarrow$ Creditor(I_b, S_A)
Debtor $\leftarrow$ Debtor(I_b, I_p, S_A)
$t \leftarrow 0$
$D \leftarrow \{\}$
for $t < t_m$ **do**
  A_c, Dialogue$_c$ $\leftarrow$ Creditor.generate
  Debtor $\leftarrow$ Debtor$(A_c$, Dialogue$_c)$
  A_d, Dialogue$_d$ $\leftarrow$ Debtor.generate
  if $A_d ==$ `accept` **then**
    $D[A_d.key] \leftarrow A_d.value$
  end if
  if D covers S_R **then**
    return D
  end if
  Creditor $\leftarrow$ Creditor$(A_d$, Dialogue$_d)$
  $t \leftarrow t + 1$
end for
return None

tiation dimension. After several rounds of negotiation and discussion, the debtor and the collector can be considered to have reached an agreement when consensus *("accept")* is achieved on all 4 negotiation objectives.

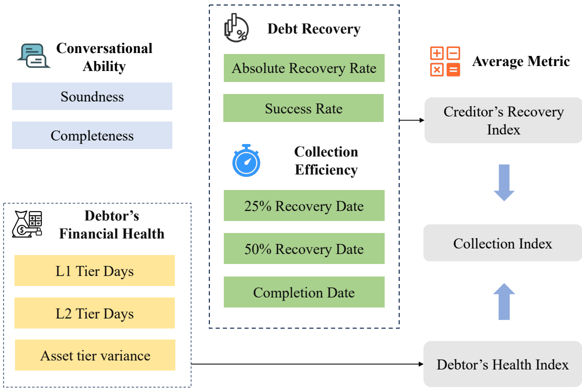

Figure 3: Evaluation system of DCN.

4 Evaluation System

Different from traditional negotiation evaluations, we argue that the DCN task requires a more comprehensive assessment framework. As illustrated in Figure 3, we developed a evaluation system based on four aspects and extended several average metrics for a comprehensive assessment.

4.1 Segmented Evaluation Metrics

In this section, we provide a general overview of the **10 metrics** across the four segmented aspects. Detailed descriptions, the evaluation process, and calculation formulas are further discussed in Appendix F.

Conversational Ability (§F.1). Conversational ability is crucial in negotiation processes for effective communication and mutual understanding. We evaluate it using two metrics: *(i) Dialogue Soundness* **(DS)** is assessed on a five-point scale, measuring the fluency, naturalness and coherency of responses; *(ii) Dialogue Completeness* (**DC**) is an automated metric that evaluates whether four objectives are all addressed during the dialogue.

Debt Recovery (§F.2). In debt collection, the primary goal is to recover as much debt as possible. We evaluate this using two key metrics: *(i) Success Recovery Rate* (**SR**) measures the proportion of samples where repayment can be successfully completed, based on the debtor's future ability to meet repayment goals. *(ii) Recovery Rate* (**RR**) reflects the portion of the debt that has been successfully recovered by the creditor, calculated as the average recovery ratio across all test samples.

Collection Efficiency (§F.3). Collection efficiency refers to how quickly a debtor can repay their debt. We monitor the timing of repayments using three key metrics: *(i) 25% Recovery Date* (**QRD**) is the estimated date when the debtor has completed 25% of the debt repayment, with earlier dates indicating quicker repayment; *(ii) 50% Recovery Date* (**HRD**) marks the completion of 50% of the repayment, offering insight into the debtor's ongoing repayment ability. *(iii) The Completion Date* (**CD**) is the date when the debtor has fully repaid their debt, with a shorter completion date indicating a faster recovery process.

Debtor's Financial Health (§F.4). The debtor's financial health plays a critical role in successful debt recovery. It affects both the debtor's ability to repay and the speed at which repayment occurs. We assess financial health using three metrics: *(i) L1 Tier Days* (**L1D**) tracks the number of days the debtor remains in the most difficult financial tier (L1), with longer durations indicating higher risk of default; *(ii) L2 Tier Days* (**L2D**) similarly tracks the days in the second most difficult financial tier (L2), which still reflects financial strain; *(iii) Asset Tier Variance* (**ATV**) captures the variance in the debtor's asset tier over a year, providing insight into the stability of their financial condition.

4.2 Comprehensive Indices

We find that the indicators for recovery and efficiency are often ***inversely related*** to the debtor's financial condition in debt collection. To balance these conflicting objectives, we introduce three average metrics (detailed description and calculation process can be found in Appendic F.5):

(i) Creditor's Recovery Index (CRI): CRI is the *weighted average* of five indicators from Debt Recovery and Collection Efficiency. It reflects an evaluation of the overall collection process by creditors, disregarding debtor-related factors. A higher value is more favorable to the creditor.

(ii) Debtor's Health Index (DHI): DHI is the *weighted average* of three indicators from Debtor's Financial Health. It assesses the financial well-being of the debtor throughout the repayment process, with a higher value indicating a greater probability of the debtor adhering to the repayment plan.

(iii) Comprehensive Collection Index (CCI): CCI is the *harmonic mean* of CRI and DHI. It provides a comprehensive evaluation of the negotiation outcome, where a higher value signifies the maximization of debt recovery and efficiency while ensuring the debtor's financial health.

5 Experiments and Results

In this section, we report the implementation details and the benchmark performances of several well-known LLMs in the DCN task on our dataset.

5.1 Implementation Details

For *open-source models*, such as Qwen series, we use their respective **chat** versions. For *api-based models*, we aim to select the latest and most advanced versions available, the *inference models* such as o1-mini (OpenAI, 2024b) and DeepSeek-R1 (DeepSeek-AI, 2025) are also included. The list of LLMs used is provided in Appendix G. The human baseline was derived from the average results of benchmark tasks completed by two finance professionals with relevant backgrounds. Additionally, a third professional was involved in arbitration to ensure the stability of the results.

Since our task focuses on Chinese, we chose one of the best open-source models currently available in the Chinese language domain: Qwen2.5-70B model to represent the **debtor**, while using different models for the **creditor** in order to compare their performance. The results of different models as the debtor are also presented in Appendix I.

For both sides, we employed the Chain of Thought (CoT) approach (Wei et al., 2023), providing the model with instructions for the DCN task and a specified format for dialogue generation, which consisted of *"Thought"*, *"Dialogue"* and *"Action"* in each interaction. The prompts are detailed in Appendix H.1.

5.2 Benchmark Results

The comparison of performance across different models is clearly illustrated in Table 2.

Table 2: The performances of some models as creditors (* denotes the second-best performance).

	Conversation		Debt Recovery		Collection Efficiency			Debtor's Financial Health			Average Metrics		
Model	**DC↑**	**DS↑**	**SR↑**	**RR(%)↑**	**QRD↓**	**HRD↓**	**CD↓**	**L1D↓**	**L2D↓**	**ATV↓**	**CRI↑**	**DHI↑**	**CCI↑**
Qwen-2.5-7B	0.94	4.57	0.98	87.15	46.04	214.04	436.84	**2.82**	79.46	**0.84***	0.732	**0.793**	0.743
Qwen-2.5-14B	0.94	4.60	0.96	89.62	28.60	154.60	358.80	6.30	79.82	0.88	0.793	0.613	0.749
Qwen-2.5-72B	0.96	4.75	0.98	88.50	36.98	185.18	404.98	3.76	**78.84***	**0.83**	0.764	**0.767***	0.764
LLaMa-3-8B	0.91	3.64	**1.00**	89.01	51.36	184.02	399.02	3.38	88.02	0.87	0.756	0.713	0.747
LLaMa-3-70B	0.87	3.94	0.98	92.24	36.72	157.32	371.12	4.50	79.56	0.87	0.792	0.695	0.771
GPT-4o	**1.00**	4.65	**1.00**	95.76	27.00	**128.40***	**297.20***	6.18	85.18	0.90	0.844	0.580	0.774
GPT-4o-mini	0.99	4.61	0.96	**96.32***	31.60	131.20	312.00	6.30	84.08	0.89	0.836	0.589	0.771
o1-mini	**1.00**	4.68	0.94	94.61	29.52	140.52	352.52	5.58	83.80	0.89	0.807	0.619	0.760
Doubao-pro	0.98	**4.91***	0.96	93.11	**21.22***	143.02	365.22	5.98	83.68	0.89	0.814	0.603	0.760
Claude-3.5	**1.00**	4.59	0.98	93.30	34.92	140.52	312.32	**3.32***	87.30	0.89	0.816	0.698	**0.789***
MiniMax	**1.00**	4.75	0.96	92.77	38.66	167.66	401.26	7.12	**76.44**	0.88	0.776	0.591	0.730
SenseChat	**1.00**	4.70	0.98	89.28	34.56	155.76	354.56	5.24	81.14	0.87	0.791	0.661	0.761
Deepseek-V3	**1.00**	4.85	0.99	91.65	28.40	141.20	313.60	5.42	83.82	0.89	**0.818***	0.625	0.771
Deepseek-R1	0.98	4.81	0.98	93.10	37.72	146.32	348.12	5.68	83.94	0.88	0.802	0.624	0.759
Human	**1.00**	**4.93**	**1.00**	**98.50**	**17.73**	**121.33**	**260.90**	3.81	78.49	0.86	**0.861**	0.740	**0.834**

LLMs perform well in terms of basic interaction format and overall dialogue capabilities. From the perspectives of dialogue completeness (DC) and soundness (DS), we find that the models effectively cover all negotiation objectives. The dialogue content is generally reasonable, aligns with the set objectives, and shows little difference from the human baseline. Specifically, the Chinese-based model outperforms the English-based model in terms of dialogue soundness for our task.

However, from the perspective of the negotiation outcomes, the performance of the LLMs was subpar and did not align well with requirements. Observing the Comprehensive Collection Index (CCI), we found that the model's overall evaluation result deviates from human-level performance by more than 0.05. This discrepancy might stem from the fact that the negotiation outcomes are numerical, making it challenging to align numerical-related requirements through prompt-based methods.

Most models tend to offer more generous concessions to debtors, both in repayment ratios and deadlines. These concessions are crucial because they directly affect the financial company's asset losses, a point emphasized in the prompt. However, as shown in the table, all models except for the GPT series have repayment ratios below 95%, meaning they did not fully follow the prompt's guidelines. In addition, the large models show lower collection efficiency compared to human benchmarks. For example, the time taken to recover 25% of the debt is 2-3 times longer than the human baseline. This suggests the models give debtors more time to repay, rather than encouraging earlier repayment. Some models, like GPT-4o, come close to human-level efficiency, but this is at the cost of worsening the debtor's financial situation. The average minimum repayment days for these models are twice as long as the human level, showing that they *struggle to adapt* to the debtor's real circumstances. This could be due to the models *misjudging the debtor's financial situation or choosing easier solutions to reach an agreement*.

The collection results achieved by the model do not hold the debtor's financial health, despite providing considerable room in terms of recovery and efficiency. We found that, with the exception of the Qwen-2.5 model, the Debt Health Index (DHI) for all other models was below the human-level threshold. Considering the concessions offered to the debtor during the collection process, these results suggest that the model did not provide *targeted debt resolution solutions* during the negotiation process.

Non-inference models may be more suitable for this task compared to inference models. By comparing the performance of the inference models o1-mini and Deepseek-R1 with their non-inference counterparts, gpt-4o-mini and Deepseek-V3, we observed a notable decline in the performance of the inference models across multiple metrics, particularly in collection efficiency.

6 Method

6.1 A Multi-agent framework for DCN

To balance the model's attention between debt recovery quality and the debtor's financial health, and to avoid decisions that may harm the creditor's interests in order to reach an agreement, we propose a method to enhance the decision alignment for DCN. Inspired by recent advancements in LLM-as-a-judge frameworks (Zheng et al., 2023) and LLM planning methodologies (Kannan et al., 2023), we designed the framework illustrated in Figure 4. The subsequent sections provide a detailed explanation

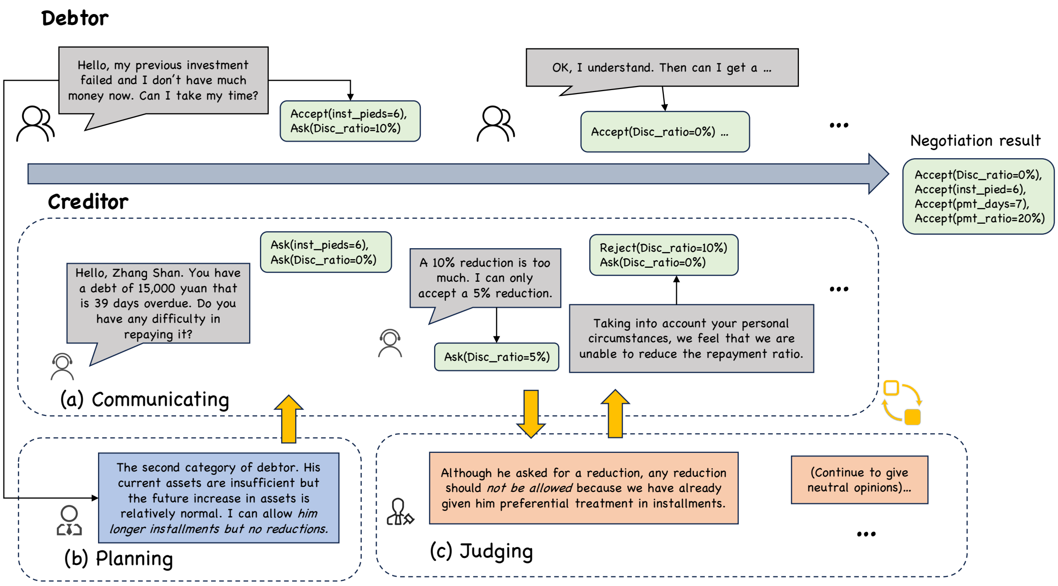

Figure 4: MADeN Framework overview.

of each module (agent) within this framework.

Planning Agent. The planning agent is activated after the debtor shares their financial difficulties, following the initial stage of the conversation. This agent is responsible for classification and strategy formulation. We categorize debtors into four distinct groups, each corresponding to different negotiation strategies and outcome spaces. This approach ensures that the model follows a consistent framework throughout the negotiation, avoiding deviations from the core objective.

Judging Agent. The judging agent evaluates the debtor's decision after each round, following the initial stage. After the communicating agent provides content, the judging agent performs an internal evaluation, and then the communicating agent adjusts and delivers the revised content to the debtor. It is set to be completely neutral and does not need to align with both sides.

By combining these two agents with the original communicating agent, we obtain a debtor **M**ulti-**A**gent **De**bt **N**egotiation system **(MADeN)** capable of self-planning and self-adjustment. Prompts for the agents can be found in Appendix H.2.

6.2 Experiment Results of MADeN

We use Qwen2.5-70B as the baseline model (Vanilla) to test the effectiveness of MADeN. We also conducted ablation experiments to separately evaluate the effectiveness of the two modules.

As the results shown in Figure 3, our multi-agent framework performs well. Compared to the vanilla group, it significantly improves debt recovery and efficiency while ensuring the debtor's economic health (The CRI has increased by more than **0.1**, while the DHI remains above **0.7**). Meanwhile, using only one of the modules does not achieve similar results, indicating the effectiveness of our two-agent design.

Table 3: The performances of our framework

Model	CRI	DHI	CCI
Vanilla	0.740	**0.771**	0.746
+ Planning	0.766	0.335	0.610
+ Judging	0.840	0.648	0.793
MADeN	**0.847**	0.706	**0.814**

6.3 Post-training with Rejection Sampling

To enhance the model's direct performance through post-training, we explored two approaches: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2024). We use the Qwen-2.5-70B model to generate two types of data for later sampling: (1) Directly Generated Data **(DG Data)**: Data directly generated by the model. (2) Multi-agent Generated Data **(MAG Data)**: Data produced through the Multi-agent framework, with content from the planning and judging agents discarded during processing.

Reject Sampling Process. The sampling processes are shown in Figure 5. For each debtor's data in *training set*, we need to generate multiple candidate dialogues by employing different prompting styles (refer to Appendix H.4). These dialogues were subsequently transformed into multiple question-answer pairs. After filtering and screening the data, a ranking was constructed based on predefined metrics. After filtering out data with

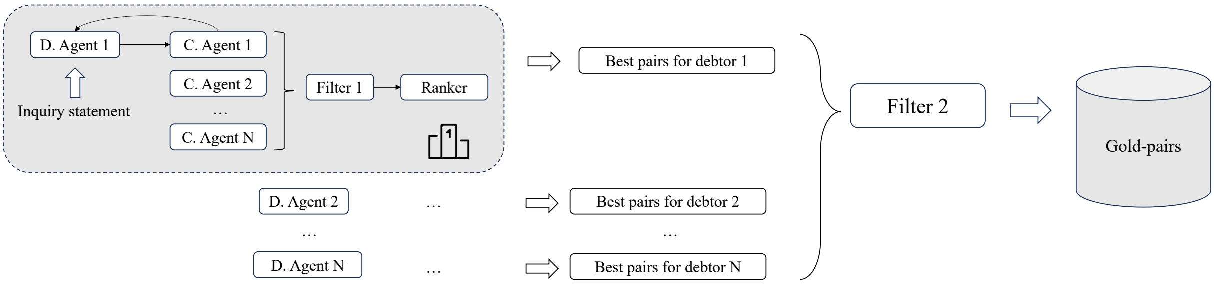

Figure 5: Reject Sampling Process. **D.Agent** and **C.Agent** represent the debtor agent and creditor agent. Creditor agent can be designed in two forms, depending on the use of the MADeN framework (DG and MAG).

Table 4: Model performances on the test set under different handling methods

	Debt Recovery		Collection Efficiency			Debtor's Financial Health			Average Metrics		
Model	**SR↑**	**RR(%)↑**	**QRD↓**	**HRD↓**	**CD↓**	**L1D↓**	**L2D↓**	**ATV↓**	**CRI↑**	**DHI↑**	**CCI↑**
Vanilla	0.98	87.15	46.04	214.04	436.84	**2.82**	79.46	**0.84**	0.731	**0.793**	0.737
MADeN	0.96	**95.21**	**22.26**	**136.26**	**329.06**	6.72	86.24	0.93	**0.828**	0.525	**0.783**
SFT-DG	0.96	88.25	33.56	187.06	396.76	9.82	**77.58**	0.93	0.763	0.429	0.708
SFT-MAG	0.94	88.02	29.32	157.72	370.92	6.10	81.08	0.91	0.779	0.587	0.755
DPO-DG	0.98	88.90	29.84	159.44	386.44	6.03	79.04	0.88	0.787	0.623	0.766
DPO-MAG	**1.00**	89.05	24.38	170.78	421.78	5.62	79.23	0.89	0.787	0.632	0.768

incomplete negotiation content and poor performance on certain metrics (Filter 1), we rank the remaining data based on CCI and select the best set for the candidate pool. Then, we sort the CCI of the candidate pool and choose the top 60% as our final dataset (Filter 2).

Negative outputs generation for DPO. For the negative samples of DPO, we used the input data obtained earlier, while replacing the prompt with a defective prompt. Its prompt construction method is detailed in Appendix H.3, and the samples were generated using Qwen-2.5-7B to increase the gap with the positive output. Finally, each of the four datasets consists of 437 pairs.

Training Setup. Due to memory constraints, we use Qwen-2.5-7B for the experiments. We obtain four models using different methods and data types.

6.4 Performance Comparison: Post-training vs. Multi-agent Method

We use the four trained models to conduct DCN and compared its performance with the previous multi-agent framework based on the same original model. The results of the metrics related to negotiation decision outcomes are shown in Table 4.

MAG data is more effective than DG data. The two training sets yielded significantly different results, with a 5% improvement in CCI during SFT, indicating that the multi-agent approach generates data with higher quality.

DPO outperforms the SFT method in DCN task. Using two different types of data, DPO always outperforms SFT in most metrics. Although fine-tuning with DG data results in slightly worse performance than the original model, DPO shows a significant improvement. It suggests that even with low-quality positive samples, the constructed negative samples can still help align the model towards the target, significantly enhancing performance.

Both the post-training method and the multi-agent approach significantly improve model performance, with the latter showing a slight edge. The multi-agent method shows better generalization, even the best-performing model (DPO-MAG), its CCI is still slightly lower than the Multi-agent method, but they perform similarly across many metrics, with it outperforming in the Debtor Health Index (DHI exceeds over **0.1**). This suggests that a well-designed post-training method can achieve results similar to the multi-agent approach.

7 Conclusion

In this paper, we introduced a comprehensive framework for evaluating AI agents in debt collection negotiations (DCN), addressing both the negotiation process and its outcomes. By leveraging synthetic debt data generated through CTGAN, we evaluated various AI-driven strategies, focusing on improving debt recovery rate and efficiency. Our enhanced LLM-based multi-agent framework, which incorporates Planning and Judging modules, demonstrated significant improvements in negotiation performance. Additionally, the application of DPO with reject sampling helped optimize the agents' focus on key objectives, leading to better results on the Qwen2.5-7B model. This work provides valuable insights into the use of AI in financial negotiations and lays the groundwork for future advancements in AI-assisted debt collection.

8 Related work

Debt Collection. Debt collection is a labor-intensive and complex task. Previous research has primarily focused on using machine learning algorithms to identify optimal decisions for individual debtors based on large-scale data (Sancarlos et al., 2023; Jankowski and Paliński, 2024; Johan, 2022; Onar et al., 2019). However, these decisions are not made in real time and often require complex decision-making processes and multiple rounds of human negotiation. On the other hand, some automated debt collection dialogue models (Floatbot.ai, 2023; Yahiya and Ahmad, 2024) can only perform tasks such as information tracking and reminders, without the ability to engage in negotiations for specific goals. Our study aims to enable models to autonomously conduct negotiations and make real-time decisions, which can significantly enhance the efficiency of debt collection.

Large Language Models in Negotiation. In previous studies on large-scale negotiation models (including bargaining (Xia et al., 2024), repeated games (Akata et al., 2023; Fu et al., 2023) and social decision-making (Pan et al., 2023)), question answering (Yang et al., 2025), the goals of the negotiators or gamers were clear, and there were clear methods for measuring the results. Debt collection is an information asymmetry game. Except for loan information, all other information is private information. How to model private information and evaluate the effectiveness of negotiation results are both difficult aspects to consider in modeling.

AI Agents. The memory, planning, reasoning, and communication capabilities of large-scale LLMs offer significant potential for the development of autonomous AI agents (Yang et al., 2023; Park et al., 2023; Liang et al., 2023; He et al., 2024; Wang et al., 2025). Its potential has been demonstrated through the creation of a simulated town (Park et al., 2023), populated with independent agents who assume distinct roles and autonomously engage in social interactions.

Limitations

We study the performance and improvement methods of large models in debt collection negotiations. To simplify the research process and capture key negotiation points, we reduce the debtor's financial information to variables like assets, average income, and average expenses. However, real-world financial situations are more complex, involving factors like cash flow issues and income fluctuations during repayment. Future work should involve more detailed simulations of debtor information and comparisons with manually simulated debtors. Additionally, due to time constraints, our creditor Multi-agent framework is relatively simple. In practical applications, stricter classification processes in *planning* and more standardized methods in *judging* are needed. We aim to integrate existing decision models to further optimize decision-making in the dialogue.

Ethical Considerations

Our study does not disclose any real client information. The acquisition of the source data was subject to strict approval by a major internet financial institution, and the process was continuously supervised by relevant personnel. All debt-related data are processed and replaced with synthetic values, and names are substituted with the pseudonym "Zhang San". For the debt reasons extracted from collection dialogues, we strictly anonymize any sensitive details and provide generalized summaries, ensuring that no specific information is involved. Each final data entry underwent rigorous manual verification. Additionally, the methodology proposed in this paper is exploratory and based on simulation for research purposes. Prior to its application in real-world debt collection involving actual individuals, it will undergo more rigorous validation and approval processes. We acknowledge that there is still a potential for undue pressure when dealing with vulnerable groups.

We conducted annotation tasks in two areas: scoring for the Dialogue Soundness metric (Section 4) and comparison with the human baseline as a debt collector (Section 5). Five graduate students with engineering backgrounds and two professionals with financial industry experience participated (They are all from China, as our study focuses on the Chinese language). All annotators involved in our study have signed a disclaimer acknowledging the terms and conditions associated with their participation. They were recruited through campus forums and the internal annotation program of the company. The annotation tasks did not involve any sensitive information and posed no risk. Compensation was provided according to the time spent on each task.

Acknowledgments

This work was supported by Ant Group.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *Preprint*, arXiv:2305.16867.

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, et al. 2023. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2025-02-13.

Yuji Chai, Mujin Kwen, David Brooks, and Gu-Yeon Wei. 2025. Flexquant: Elastic quantization framework for locally hosted llm on edge devices. *Preprint*, arXiv:2501.07139.

Sow Hup Joanne Chan. 2006. The global negotiator: Making, managing, and mending deals around the world in the twenty-first century. *Asian Business & Management*, 5:149–151.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DouBao. 2024. Chat service. Accessed: 2025-02-13.

Gika Asdina Firanda, Paramita Prananingtyas, and Sartika Nanda Lestari. 2021. Debt collection of financial technology lending. *Proceedings of the 1st International Conference on Science and Technology in Administration and Management Information, ICSTIAMI 2019, 17-18 July 2019, Jakarta, Indonesia*.

Floatbot.ai. 2023. Generative and conversational ai for automated debt collection. Accessed: 2025-02-07.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *ArXiv*, abs/2308.00352.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Rafał Jankowski and Andrzej Paliński. 2024. Debt collection model for mass receivables based on decision rules—a path to efficiency and sustainability. *Sustainability*.

Suwinto Johan. 2022. Financial technology company's debt collection method: A legal aspect. *Unnes Law Journal*.

Shyam Sundar Kannan, Vishnunandan L. N. Venkatesh, and Byung-Cheol Min. 2023. Smart-llm: Smart multi-agent robot task planning using large language models. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12140–12147.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lankao County People's Government. 2024. On financial innovation: Entry points for asset management companies into bank loan services—recommendations and responses for efficient disposal of non-performing credit assets. Accessed: 2025-02-07.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. 2024. Statistical rejection sampling improves preference optimization. *Preprint*, arXiv:2309.06657.

Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. 2025. Llm-powered multi-agent system for automated crypto portfolio management. *Preprint*, arXiv:2501.00826.

MiniMax. 2024. Abab65 series. Accessed: 2025-02-13.

Sezi Cevik Onar, Basar Oztaysi, Cengiz Kahraman, and Ersan Öztürk. 2019. A decision support system to optimize debt collection assignments. *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024a. Gpt-4o contributions. Accessed: 2025-02-13.

OpenAI. 2024b. Openai o1 hub. Accessed: 2025-02-13.

Peterson K. Ozili. 2019. Non-performing loans and financial development: New evidence. *ERN: Other Monetary Economics: Financial System & Institutions (Topic)*.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Golden Credit Rating. 2016. Credit rating methodology for non-performing loan asset-backed securities. Accessed: 2025-02-07.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Abel Sancarlos, Edgar Bahilo, Pablo Mozo, Lukas Norman, Obaid Ur Rehma, and Mihails Anufrijevs. 2023. Towards a data-driven debt collection strategy based on an advanced machine learning framework. *ArXiv*, abs/2311.06292.

SenseTime. 2024. Ai chat assistant. Accessed: 2025-02-13.

Neil J. Smelser and Paul B. Baltes, editors. 2001. *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier, Oxford, UK.

Garden Tabacchi, Claudio Costantino, Giuseppe Napoli, Valentina Marchese, Manuela Cracchiolo, Alessandra Casuccio, Francesco Vitale, and on behalf of the ESCULAPIO working group. 2016. Determinants of european parents' decision on the vaccination of their children against measles, mumps and rubella: A systematic review and meta-analysis. *Human Vaccines & Immunotherapeutics*, 12:1909 – 1923.

Mingfeng Tang, Haoshu Huang, and Cui wen Li. 2018. Debt collection on the peer-to-peer lending platform under the situation of chinese culture: A case study of ppdai.com. *DEStech Transactions on Social Science, Education and Human Science*.

Leigh Thompson. 1997. The mind and heart of the negotiator.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *Preprint*, arXiv:2401.01275.

Xiaofeng Wang, Yiming Wang, Wenhong Zhu, and Rui Wang. 2025. Do large language models truly understand geometric structures? *Preprint*, arXiv:2501.13773.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. 2024. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *Preprint*, arXiv:2402.15813.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*.

Ahmad Yahiya and Bani Ahmad. 2024. Automated debt recovery systems: Harnessing ai for enhanced performance. *Journal of Infrastructure, Policy and Development*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *Preprint*, arXiv:2306.02224.

Shuo Yang, Siwen Luo, Soyeon Caren Han, and Eduard Hovy. 2025. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. *arXiv preprint arXiv:2503.18491*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Brad Zwilling, Brandy Krapf, and Dwight Raab. 2017. Evaluating your capital debt repayment capacity. *farmdoc daily*.

A Detailed descriptions of four Negotiation Dimensions

The following sections provide detailed descriptions of the four key negotiation dimensions involved in debt collection, outlining how each aspect influences the negotiation process and repayment outcomes. And table 5 shows all dimensions of the negotiation.

- **Debt Reduction Ratio:** This refers to the portion of the debt that can be waived by the creditor to ease the debtor's repayment burden. The reduction ratio is often negotiable based on the debtor's financial situation, with creditors typically offering reductions as an incentive to settle the debt more efficiently.
- **Immediate Repayment Ratio:** In order to temporarily restore the debtor's credit and advance the repayment process, creditors usually require the debtor to repay a portion of the outstanding debt immediately during the negotiation. This portion is typically at least 5% of the total debt.
- **Immediate Repayment Time:** If the debtor is unable to make an immediate payment on the same day, a grace period of up to 14 days may be granted. Within this period, the debtor is expected to raise the necessary funds to complete the immediate repayment.
- **Installment Period:** After addressing part of the debt through reductions and immediate repayments, the remaining balance can be settled through installments. The installment ratio can vary from 3 to 24 periods, allowing the debtor to repay the debt within a period ranging from a few months to up to two years.

B Data Distribution

As shown in the Figure 6, our dataset exhibits a certain distribution across Amount, Sex, and Overdue Days, which is similar to the actual situation. The distributions in both the test set and the train set are also largely consistent.

C More explanation of synthetic data generation process

We used real data from a leading Chinese financial company in 2024, consisting of 40,000 records, to train a CTGAN network that generated nearly 2,000 synthetic data entries. From these 2,000 entries, we performed sampling based on the distribution of various categories in the real business data (including debt amount, gender, and overdue time) to ensure that the internal relationships and distributions of the data closely match the real data.

CTGAN generates entirely synthetic data, rather than directly copying real data. The generator learns the overall distribution and patterns of the data but does not memorize specific samples. For continuous data (e.g., debt amount, time), min-max normalization is applied; for discrete data (e.g., gender, age), categories are encoded as one-hot vectors,

Table 5: Negotiation Dimensions and Their Possible Values

Dimension	**Values**
Discount Ratio ('disc_ratio')	5%, 10%, 15%, 20%, 25%, 30%
Immediate Payment Ratio ('pmt_ratio')	5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%
Immediate Payment Time ('pmt_days')	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 days
Installment Periods ('inst_prds')	3, 6, 9, 12, 18, 24 months

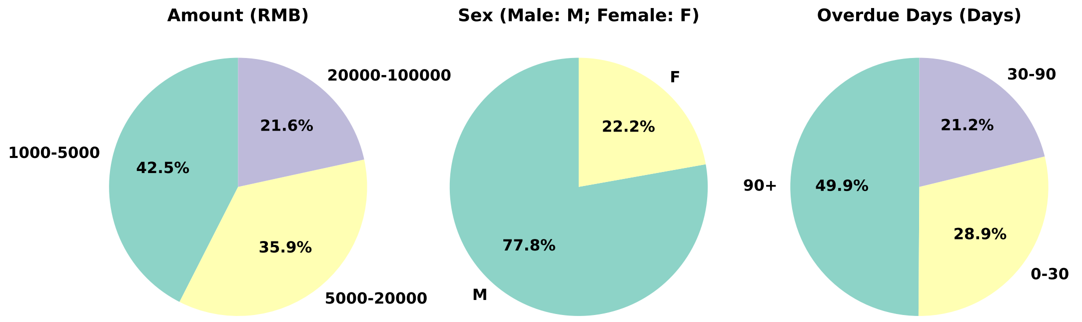

Figure 6: Distribution of Need collected Amount, Sex, and Overdue Days.

with modality information introduced to ensure that the generated category distribution aligns with the real data.

To further validate the model, we selected 50 synthetic data entries and generated 50 additional synthetic entries using random numbers (within a reasonable range for each value). For each entry, we identified the most similar real data entry in terms of the amount owed and paired the two. We then randomized these 100 pairs and asked financial professionals to determine whether they could distinguish the synthetic data from the real data. The results are as follows:

Table 6: Distinguishability Test Results of Synthetic Data.

	Identified	Unidentified	Total	Efficiency
Our data	0	50	50	100%
Random data	47	3	50	6%

It is evident that professionals were almost unable to distinguish our synthetic data from real data, which further supports the validity of our data generation method.

D Difficulty Tiers for Debt Collection

In the field of debt management and collection, the economic hardship level may be related to the debtor's repayment capacity assessment (Zwilling et al., 2017). Referring to common methods for determining economic hardship levels (Smelser and Baltes, 2001), we categorize debtors into five tiers as shown in Table 7.

Table 7: Difficulty Tiers for Debt Collection

Tier	**Description**	**Range**
Tier 1	Extremely Difficult	0 - 2000
Tier 2	Very Difficult	2000 - 5000
Tier 3	Moderately Difficult	5000 - 10000
Tier 4	Slightly Difficult	10000 - 20000
Tier 5	No Difficulty	20000+

E Definitions of variables in DCN process

Table 8 provides the descriptions of all the variables appearing in Algorithm 1. The Action Set includes `ask`, `reject`, and `accept`, while the Negotiation Dimension Set consists of the four quantities listed in Table 5.

F Detail of metrics

F.1 Conversational Ability

In negotiation processes, conversational ability is crucial for achieving effective communication and mutual understanding. Tu et al. (2024) proposed an evaluation framework for role-playing tasks. Inspired by this work, we tailored it to our task by distinguishing Conversational Ability into two dimensions: fluency and completeness.

Table 8: Definitions of variables in DCN process.

Conception	Variable
Basic Information	I_b
Creditor	creditor
Action Set	S_A
Result Dictionary	D
Personal Financial Information	I_p
Debtor	debtor
Negotiation Dimension Set	S_R
Turn	t
Max Turns	t_m
Agent Creditor	Creditor
Agent Debtor	Debtor
Action of Debtor	A_d
Action of Creditor	A_c

Dialogue Soundness (DS). Dialogue Soundness is a single-metric evaluation that measures a dialogue response's fluency, naturalness, coherency, and consistency on a five-point scale. It assesses whether the response is grammatically correct and conversational, stays on topic, and remains logically consistent across turns. This metric is manually scored, with the scale shown in Table 9. Five graduate students from engineering disciplines were employed to evaluate this metric and calculated the average value.

Dialogue Completeness (DC). Dialogue Completeness is a metric designed to evaluate whether a conversation addresses all specified objectives outlined in section 3.1 of the paper. This automated measure checks if each of the four key goals has been adequately discussed during the dialogue, ensuring that no critical topics are overlooked or omitted.

F.2 Debt Recovery

Success Recovery Rate (SR). The success rate of the negotiation is determined by whether the debtor's future assets remain in a healthy state (i.e., the total personal assets remain greater than 500). The success rate is defined as the proportion of samples in which repayment can theoretically be completed successfully:

$$\text{SR} = \frac{N_{\text{success}}}{N}, \tag{1}$$

where SR is the success rate, N_{success} is the number of successful samples, and N is the total number of samples.

Recovery Rate (RR). The recovery ratio refers to the portion of the debt recovered by the creditor, which is typically 1 minus the reduction ratio. If the plan is unsuccessful, the recovery ratio is considered to be 0. The final recovery ratio is calculated as the mean recovery ratio across the test samples:

$$\text{RR} = \frac{1}{N}\sum_{i=1}^{N} r_i, \tag{2}$$

where RR is the final recovery ratio, r_i is the recovery ratio of the i-th sample.

F.3 Collection Efficiency

25% Recovery Date (QRD) refers to the date at which the debtor has completed 25% of the debt repayment, which is estimated based on the debtor's future economic condition sequence. The final 25% Recovery Date is calculated as the mean of the recovery dates across the test samples:

$$\text{QRD} = \frac{1}{N}\sum_{i=1}^{N} t_{25\%,i}, \tag{3}$$

where QRD is the final 25% recovery date, $t_{25\%,i}$ is the 25% recovery date of the i-th sample, and N is the total number of samples.

50% Recovery Date (HRD) is defined similarly to the 25% Recovery Date, referring to the date at which the debtor has completed 50% of the debt repayment, based on the debtor's future economic condition sequence. **Completion Date (CD)** refers to the date at which the debtor has fully repaid all of the debt.

The 50% Recovery Date and Completion Date are calculated as the means of the respective recovery dates across the test samples:

$$\text{HRD} = \frac{1}{N}\sum_{i=1}^{N} t_{50\%,i}, \tag{4}$$

where HRD is the final 50% recovery date, and $t_{50\%,i}$ is the 50% recovery date of the i-th sample.

$$\text{CD} = \frac{1}{N}\sum_{i=1}^{N} t_{\text{Completion},i}, \tag{5}$$

where CD is the completion date, and $t_{\text{Completion},i}$ is the completion date of the i-th sample.

F.4 Debtor's Financial Health

L1 Tier Days (L1D) refers to the number of days the debtor remains in the most difficult tier over the next two years. **L2 Tier Days (L2D)** refers to the

Table 9: Dialogue Soundness (DS) Rating Scale

Score	Rating	Description
5	Excellent	Fluent, natural, on-topic, logically consistent.
4	Good	Mostly natural, minor topic drift, slight inconsistency.
3	Acceptable	Understandable but somewhat rigid, occasional drift or inconsistency.
2	Poor	Unnatural phrasing, noticeable topic deviation or contradictions.
1	Unacceptable	Robotic, off-topic, illogical contradictions.

number of days the debtor remains in the second most difficult tier during the same period. These two indicators directly correspond to the duration the debtor spends in different levels of financial difficulty. Research has shown that the longer the debtor remains in a higher level of difficulty, the more likely they are to default on the loan (Tabacchi et al., 2016).

Asset tier variance (ATV). In addition to controlling for the number of days the debtor remains in the high-poverty tier, the overall stability of the debtor's asset level also ensures a higher repayment performance. To capture this, we introduce the asset tier variance metric, which is calculated by computing the variance of the debtor's asset tier over the course of one year. The final result is obtained by calculating the mean of the asset tier variances across the test samples:

$$v_{\text{asset},i} = \frac{1}{T-1} \sum_{t=1}^{T} \left(A_{i,t} - \bar{A}_i \right)^2, \tag{6}$$

where $A_{i,t}$ is the asset tier of the i-th sample at time t, $\bar{A}_i$ is the average asset tier of the i-th sample over the year, and T is the total number of time periods. The final asset tier variance is the mean of the asset tier variances across the test samples:

$$\text{ATV} = \frac{1}{N} \sum_{i=1}^{N} v_{\text{asset},i}, \tag{7}$$

where ATV is the mean asset tier variance, and N is the total number of samples.

F.5 Average Metric

In debt collection, the indicators for Debt Recovery and Collection Efficiency are often inversely related to the Debtor's Financial Health. This means that efforts to recover debts more efficiently and quickly may negatively impact the debtor's financial condition. To strike a balance between these two conflicting objectives, we introduce three average metrics that help quantify the trade-off: the Creditor's Recovery Index (CRI), the Debtor's Health Index (DHI), and the Comprehensive Collection Index (CCI).

Creditor's Recovery Index (CRI): This index measures the effectiveness of the creditor's recovery strategy while accounting for the impact on the debtor's financial health. The index aggregates several recovery metrics weighted by their relative importance to the creditor's objectives. The index is calculated as follows:

$$\begin{aligned} \text{CRI} = {} & w_1 \cdot \text{SR} + w_2 \cdot \text{RR} \\ & + w_3 \cdot \frac{\max(\text{QRD}) - \text{QRD}}{\max(\text{QRD})} \\ & + w_4 \cdot \frac{\max(\text{HRD}) - \text{HRD}}{\max(\text{HRD})} \\ & + w_5 \cdot \frac{\max(\text{CD}) - \text{CD}}{\max(\text{CD})}, \end{aligned} \tag{8}$$

where w_1, w_2, w_3, w_4, w_5 are the weights assigned to each metric based on the creditor's priorities.

Debtor's Health Index (DHI): This index measures the debtor's financial health during the recovery process. It incorporates several factors that capture the debtor's stability and vulnerability. The Debtor's Health Index is calculated as:

$$\begin{aligned} \text{DHI} = {} & w_6 \cdot \frac{\max(\text{L1D}) - \text{L1D}}{\max(\text{L1D})} \\ & + w_7 \cdot \frac{\max(\text{L2D}) - \text{L2D}}{\max(\text{L2D})} \\ & - w_8 \cdot \text{ATV}. \end{aligned} \tag{9}$$

Here, w_6, w_7, w_8 are weights that balance the importance of each factor in determining the debtor's health.

Comprehensive Collection Index (CCI): The Comprehensive Collection Index combines both the Creditor's Recovery Index (CRI) and the Debtor's Health Index (DHI) into a single metric that evaluates the overall balance between debt recovery and the debtor's financial well-being. The index is calculated using the harmonic mean of the two indices, with a weight factor θ applied to the CRI:

$$\text{CCI} = \frac{2\theta^2 \cdot \text{CRI} \cdot \text{DHI}}{\text{CRI} + \theta^2 \cdot \text{DHI}}. \tag{10}$$

In this formula, the weight factor θ indicates that the CRI is weighted θ times more than the DHI. In this study, θ is set to 2. This approach ensures that a high value in either the recovery index or the health index will influence the overall result, while emphasizing the importance of balancing both aspects.

The use of this weighted harmonic mean helps in evaluating different debt recovery strategies by considering both the creditor's objectives and the debtor's financial stability, thereby promoting a more balanced approach to debt collection.

The constant values used in the calculation process are shown in Table 10. In future research or application, these values may be adjusted depending on the specific requirements to better align with the needs.

Table 10: Constants used in Average Metric Calculation.

Constant	Value
w_1	0.25
w_2	0.25
w_3	0.2
w_4	0.15
w_5	0.15
w_6	1.5
w_7	0.8
w_8	1
θ	2
max(QRD)	180
max(HRD)	360
max(CD)	720
max(L1D)	30
max(L2D)	250

G All LLMs in our Experiments

We comprehensively evaluate nine LLMs, encompassing both API-based models and open-source models. The API-based models include the GPT series (GPT-4o, GPT-4o-mini, o1-mini) (OpenAI, 2023, 2024a,b), Claude-3.5 (Anthropic, 2024), MiniMax (abab6.5s-chat) (MiniMax, 2024), Sensechat (SenseTime, 2024), DeepSeek series (DeepSeek-R1 and DeepSeek-V3) (DeepSeek-AI, 2025, 2024) and Doubao (DouBao, 2024). The open-source models include the Llama series (LlaMA-2-13B-Chat, LlaMA-3-8B-Instruct, LlaMA-3-70B-Instruct) (Touvron et al., 2023) and the Qwen-2.5 series (Qwen-2.5-7B, Qwen-2.5-14B and Qwen-2.5-72B) (Qwen et al., 2025). These models are run using vLLM (Kwon et al., 2023) on eight Nvidia A100 GPUs with the same random seed. For each model, the entire test set was processed in approximately one hour using parallel methods. All temperatures are set to 0 (Due to API-provider's closed-source non-deterministic implementation, small changes may still occur in the reproduction process). Specific model hyperparameters and version details can be found in Table 11. All models and tools (vLLM and LLaMa-Factory (Zheng et al., 2024)) used in this study, including closed-source API-based models, open-source models, were used in compliance with their respective licenses. What's more, the use of these generative models for dialogue tasks is well-established in the field and follows standard practices.

H Prompts

H.1 Basic Prompts for Role-playing Debtor and Creditor.

Figures 7 and 8 illustrate the prompts given to the large model to act as the debtor and the creditor, respectively. Originally in Chinese, these prompts have been appropriately simplified and automatically translated into English for display purposes (the full Chinese prompts is available to be disclosed later). Additionally, the instructions provided to human annotators were consistent with the prompts given to the model.

H.2 Prompts for Planning Agent and Judging Agent.

Figures 9 and 10 display the prompts for the planning agent and judging agent in the MADaN framework. Similarly, these prompts have been simplified and translated for ease of presentation. The prompt for the communicating agent remains unchanged, as previously shown.

H.3 Defective prompt

There are three main methods for generating Defective Prompts, as shown in Table 12. In practice, we first generate a list of prompts and then randomly select one from the list to generate the negative samples.

You are Zhang San, a {age}-year-old {sex}. You previously borrowed {bal_due} yuan from institution A, and you still owe {need_coll_amt} yuan, which has been overdue for {ovd_days} days. The reason you were unable to repay on time is {reason}. Currently, the total value of your assets is {asset} yuan, and your average daily income is {avg_daily_income} yuan, with average daily expenses of {avg_daily_expense} yuan, and an average daily balance of {avg_daily_balance} yuan.

You wish to negotiate to reduce your repayment burden. Your goal is to reduce the total repayment amount, number of installments, and monthly repayment amount as much as possible within your capability.

From online sources and general policies, you've learned that the negotiable elements may include four key factors:
1. **Discount ratio (disc_ratio)**: Reduction in the total debt.
2. **Partial repayment ratio (pmt_ratio)**: The amount to be repaid upfront.
3. **Partial repayment time (pmt_days)**: The time within which the upfront payment must be made (typically within 14 days).
4. **Installment periods (inst_prds)**: Number of months to divide the remaining debt into installments.

Typically, the discount ratio will likely be 0%, but you might be asked to pay a portion of the debt upfront within 14 days, which usually ranges from 25% to 50%, in multiples of 5%. If you feel you need more time to repay, you can ask for an installment plan, such as 3 months, 6 months, etc.

You can also ask the collector about the maximum concessions they can offer. Depending on their policies, they might have some flexibility, and asking could help you negotiate a better deal.

When deciding your response, consider the following factors:
1. **Economic situation**: Explain your income, expenses, and current debt. Mention if you have a stable income source or if there are other financial burdens (such as medical expenses or children's education). You can choose to reveal or withhold information depending on how it might affect your negotiation.
2. **Current difficulties**: For example, unemployment, illness, or needing to support family members financially.
3. **Willingness to repay**: Show that you are willing to repay but emphasize that your financial capacity is limited.
4. **Negotiation strategy**:
• Try to secure a higher discount.
• If the upfront payment is too high, negotiate for a lower amount or request a longer repayment period.
• For installment periods, ask for more installments to reduce monthly repayment pressure.

When the debt collector asks for the reason for your overdue payment, you can explain your situation and express your demands and commitments based on your current circumstances. For example: "I hope for a 10% discount, and if that's granted, I can repay 50% immediately." Or: "If I can get 18 installments, I can repay 45% in 2 days without any discount."

Your output should follow this format:
• **Thoughts**: Your reasoning for the response.
• **Dialogue**: What you say to the collector.
• **Action**: Your action. There are three possible actions: ask (make a request), accept (agree to the proposal), reject (decline the proposal). For example: ask(disc_ratio=5%) means requesting a 5% discount.

Examples:
1. **Example 1**:
• **Thoughts**: I've recently lost my job, and my family is under pressure from a mortgage. I want to negotiate a discount.
• **Dialogue**: My situation is difficult; I've recently lost my job, and I have a mortgage and child education expenses. I can't pay the full amount, could you consider offering a discount?
• **Action**: ask(disc_ratio=5%)
2. **Example 2**:
• **Thoughts**: The collector proposed a 40% upfront repayment, but I don't have that much cash. I'll try to negotiate for 30%.
• **Dialogue**: I currently don't have the 40%, it's difficult for me. I can only pay 30% upfront.
• **Action**: ask(pmt_ratio=30%)

Now it's your turn:

Figure 7: Prompt of Debtor.

There is a debtor, Zhang San, who is {age} years old and {sex}. He borrowed {bal_due} yuan from Institution A and still owes {need_coll_amt} yuan. The debt is {ovd_days} days overdue.

You are a collection agent for Institution A. Your job is to negotiate with debtors who have overdue payments and are facing financial difficulties. You need to balance reducing the debtor's financial burden while minimizing your institution's loss. Debtors may exaggerate their situation, so stay rational and don't fully accommodate their emotions.

Negotiation Parameters:

You can negotiate on four dimensions:
1. **Discount Rate (disc_ratio)**: A percentage reduction on the debt (0%, 5%, 10%, 15%, 20%, 25%, 30%). Only offer discounts if the debtor genuinely cannot pay the full amount or if the debtor is in extreme hardship. Typically, avoid offering discounts unless necessary.
2. **Immediate Payment Ratio (pmt_ratio)**: The percentage of the debt the debtor must pay immediately (5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%). Aim for at least 25%, depending on the debtor's financial situation.
3. **Immediate Payment Time (pmt_days)**: The number of days (1-14) the debtor has to pay the immediate portion. Typically, aim for within 7 days unless they struggle to raise funds.
4. **Installment Periods (inst_prds)**: The number of months over which the remaining debt will be paid. The fewer the months, the better for your institution. Common options are 3, 6, 9, 12, 18, 24 months, each with its own interest rate.
Note: Prefer shorter installment periods (e.g., 3 or 6 months) unless the debtor is in severe hardship.

Strategy:
- Stay firm on discounts (usually 0%), except in extreme cases.
- For immediate payments, push for at least 25%, depending on the debtor's situation.
- Adjust installment periods based on the debtor's ability to pay (shorter is better).
- Use a balance of pressure and empathy, as the debtor is already in arrears, and they are harming the institution's interests.
- You may encounter a neutral judge evaluating your decisions. Be prepared to adjust as needed.

Dialogue and Action Format:
- **Thoughts**: Your reasoning process.
- **Dialogue**: What you say to the debtor.
- **Action**: Your action (ask, agree, refuse).

Examples:
- **ask(disc_ratio=10%)**: Ask about a 10% discount.
- **accept(disc_ratio=10%)**: Agree to a 10% discount.
- **reject(disc_ratio=10%)**: Reject a 10% discount.

Example 1:
- **Thoughts**: Understand the debtor's situation first.
- **Dialogue**: What's your current financial situation? Are you unable to pay because of job loss or other issues?
- **Action**: non

Example 2:
- **Thoughts**: If they're in trouble but assets seem okay, offer a larger immediate payment ratio and set a reasonable installment period.
- **Dialogue**: We can ask for 30% as immediate payment, and the remaining 50% could be settled in 12 months. No discount.
- **Action**: ask(pmt_ratio=30%, inst_prds=12, disc_ratio=0%)

Now it's your turn:

Figure 8: Prompt of Creditor (Debt Collector).

There is a debtor, Zhang San, who is {age} years old and {sex}. He borrowed {bal_due} yuan from Institution A and still owes {need_coll_amt} yuan. The debt is {ovd_days} days overdue.

You are a debt collector for Institution A. Before engaging in detailed negotiations, you need to categorize the debtor based on their statements and set a preliminary framework for the negotiation.

Your task is to negotiate with overdue debtors facing financial difficulties, ensuring the repayment amount is within their capability while minimizing your institution's loss. Debtors may use various excuses or exaggerate their financial difficulties, but you must maintain rational judgment at all times.

There are four negotiable elements:
1. **Discount ratio (disc_ratio)**: The reduction in the total debt. The options are 0%, 5%, 10%, 15%, 20%, 25%, 30%. A discount is offered only when the debtor truly cannot repay the full amount, and it should not be agreed upon if installment options are feasible.
2. **Partial repayment ratio (pmt_ratio)**: The portion of the debt that must be paid upfront. It ranges from 5% to 50%. Debtors need to pay a part to show willingness to repay and keep their account active, usually set at 25% or higher.
3. **Partial repayment time (pmt_days)**: The time (in days) within which the debtor needs to make the upfront payment. Typically, it should be within 7 days unless the debtor faces significant difficulties.
4. **Installment periods (inst_prds)**: After considering discounts and upfront payments, the remaining balance is divided into installment payments. The period should be as short as possible, usually 3 or 6 months.

Debt Categories:
1. **Category 1**: Low current assets, low future asset potential. Prioritize discount (disc_ratio > 0%) and normal installment negotiations.
2. **Category 2**: Normal current assets, low future asset potential. Prioritize immediate repayment (high pmt_ratio, low pmt_days), no discount (disc_ratio = 0%), and moderate installment periods.
3. **Category 3**: Normal current assets, normal future asset potential. No discount (disc_ratio = 0%), high immediate repayment (pmt_ratio), and a short installment period.
4. **Category 4**: Low current assets, normal future income potential. Prioritize installments (high inst_prds), no discount (disc_ratio = 0%), and possibly reduce immediate repayment (pmt_ratio).

After hearing the debtor's reason for non-payment, you will classify them and set a basic negotiation framework based on their situation.

Examples:
1. **Example 1**: If the debtor needs to pay 14,000 yuan, with 8,000 yuan in assets and weak future earning potential, classify them as **Category 2**. No discount, immediate repayment of 40%, and a 6-month installment plan.
2. **Example 2**: If the debtor owes 40,000 yuan, has 12,000 yuan in assets, but strong income potential, classify them as **Category 4**. No discount, immediate repayment of 25%, and a 6-month installment plan.
3. **Example 3**: If the debtor owes 7,933 yuan and their financial situation isn't urgent, classify them as **Category 3**. No discount, immediate repayment of 50%, and a 3-month installment plan.

Negotiation Begins:
(Once the debtor explains why they can't repay on time, you'll analyze their situation and create a strict framework for negotiation, ensuring you don't get pressured into offering excessive discounts, long installment periods, or low immediate repayments.)

Figure 9: Prompt of Planning Agent.

Scenario Overview: There is a debtor, Zhang San, who is {age} years old and {sex}. He borrowed {bal_due} yuan from Institution A and still owes {need_coll_amt} yuan. The debt is {ovd_days} days overdue.

The debtor is likely facing some repayment difficulties. A debt collector is now negotiating with the debtor to resolve the debt, with the following negotiation dimensions:

Negotiable elements:
1. **Discount Ratio (disc_ratio)**: Seven possible options—0%, 5%, 10%, 15%, 20%, 25%, 30%. This should only be agreed upon when the debtor truly cannot repay the full amount through other methods. If the debtor can repay with installments, there should be no discount (in normal cases, discount is 0%).
2. **Partial Repayment Ratio (pmt_ratio)**: Ten options—5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%. This amount is to be paid upfront within two weeks, which helps maintain the debtor's account activity. The value of this repayment should depend on the debtor's financial situation, and should generally be at least 25% (below 25% indicates extreme difficulty).
• If the debtor proposes an amount outside of the available options (e.g., 13.7%), it should be corrected to the nearest option (e.g., 15%).
3. **Partial Repayment Time (pmt_days)**: The time frame for making the partial payment, which ranges from 1 to 14 days. Typically, the payment should be made within 7 days unless the debtor is facing severe financial difficulties.
4. **Installment Periods (inst_prds)**: For the remaining amount after discount and partial repayment, installments can be arranged. There are specific rates for different periods. Generally, the fewer the installments, the better (3 or 6 months is ideal, 9 months is already quite difficult to agree to).

Evaluation Criteria:
1. **Debtor's Economic Situation and Consistency**: Assess whether the debtor's claimed difficulties match the proposed negotiation terms.
2. **Reasonableness of the Collector's Proposal**: Evaluate whether the conditions set by the collector are within acceptable boundaries, and whether the collector has adhered to their strategy. Pay particular attention to whether the collector is making too many concessions to the debtor, particularly regarding the discount ratio.
3. **Extent of Concessions from Both Parties**: Assess whether either party is making excessive concessions or too stubborn in their demands.

Example Evaluations:
• **Example 1**:
• **Comments**: The debtor asks for a 10% discount, and the collector proposes 5%. Both are quite high. The previous conversation doesn't show the debtor as being in a very difficult position, so a discount might not be necessary.
• **Example 2**:
• **Comments**: The debtor was offered a 15% discount, a significant concession. Now, they ask for a 12-month installment period. If the debtor is given such a long period, they wouldn't need a discount, making the collector's agreement illogical.

Next Step: (After reviewing the most recent round of negotiation, provide a concise evaluation of the latest interaction between both parties.)

Figure 10: Prompt of Judging Agent.

Table 11: Hyperparameters of Each Model.

Model Name	Parameters	Comments
Qwen-2.5-7B	"temperature": 0, "max_tokens": 1024	version = "qwen-2.5-7b-instruct"
Qwen-2.5-14B	"temperature": 0, "max_tokens": 1024	version = "qwen-2.5-14b-instruct"
Qwen-2.5-72B	"temperature": 0, "max_tokens": 1024	version = "qwen-2.5-72b-instruct"
GPT-4o	"temperature": 0, "max_tokens": 1024	version = "gpt-4o-2024-11-20"
GPT-4o Mini	"temperature": 0, "max_tokens": 1024	version = "gpt-4o-mini"
o1-Mini	"temperature": 0, "max_tokens": 1024	version = "o1-mini"
LLaMa-3-8B	"temperature": 0, "max_tokens": 1024	version = "llama-3-8b-instruct"
LLaMa-3-70B	"temperature": 0, "max_tokens": 1024	version = "llama-3-70b-instruct"
Doubao	"temperature": 0, "max_tokens": 1024	version = "Doubao-pro-4k"
Claude-3.5	"temperature": 0, "max_tokens": 1024	version = "claude-3-5-sonnet-20241022"
DeepSeek-V3	"temperature": 0, "max_tokens": 1024	version = "deepseek-chat"
DeepSeek-R1	"temperature": 0, "max_tokens": 1024	version = "deepseek-reasoner"
MiniMax	"temperature": 0, "max_tokens": 1024	version = "abab6.5s-chat"
SenseChat	"temperature": 0, "max_tokens": 1024	version = "SenseChat"

Table 12: Defective Prompt Modifications for Debt Collection Negotiation.

Modification Type	Description	Example
Deletion	Remove specific instructions	Removing "Offer a 10% discount when the debtor shows clear financial difficulty."
Replacement	Reverse guidance	Changing "Be cautious when the debtor makes a request" to "Approve requests without further consideration."
Addition	Add negative guidance	Adding "If installment terms are discussed, set them to 24 months without negotiation."

H.4 Different prompting styles

In debt collection practice, different prompt styles correspond to distinct strategies. We selected four primary types of prompts:

- **Neutral (default)**: Presents facts and data in an objective manner without emotional bias.
- **Gentle**: Adopts a friendly and calm tone, emphasizing understanding and empathy.
- **Strict**: Uses a firm and assertive tone, highlighting legal consequences.
- **Motivational**: Encourages repayment through incentives and positive reinforcement.

These prompts were incorporated into the original prompt with specific content and examples. Experiments were conducted to evaluate the effectiveness of these different debt collection strategies. The results are presented in Table 13.

Analysis of the results indicates that the predefined strategies did not yield significant differences in the overall performance. This lack of variance may be attributed to two primary factors: (1) no single strategy is universally applicable across all debt collection scenarios, and (2) the model may not strictly adhere to the given prompts. To address these limitations, subsequent training data sampling incorporated outputs from these diverse strategies to improve generalization and adaptability.

Table 13: Performance Metrics of Different Prompt Strategies

Prompt Style	CRI	DHI	CCI
Rational	0.764	0.767	0.764
Strict	0.770	0.751	0.766
Gentle	0.751	0.750	0.751
Motivational	0.779	0.763	0.769
Human	0.861	0.740	0.834

I The performance of different models as the debtor

In Section 5, we evaluate the debt collection outcomes when different models act as the creditor. We alse examine the performance of different models as debtors, using the Qwen-2.5-72B model exclusively as the creditor. We observed significant differences in the results when using different models for the debtor as shown in Table 14. The SenseChat and Llama-3-70b models exhibited some inconsistencies, yielding excessively high

DHI scores. During the examination of the dialogue process, we found that these models tended to neglect *repeated statements* within the dialogue, leading to the inclusion of some irrelevant or ineffective content. Additionally, some models were more sensitive to the debtor's prompt, likely due to the more complex nature of the debtor agent's objectives. In contrast, the Qwen-2.5-72 model showed relatively balanced performance, suggesting that our choice was appropriate.

Since our focus is on studying the model's performance as a debt collector, we did not design specific metrics for debtor models. Our primary aim is to use models capable of understanding the debtor's objectives and engaging in dialogue for simulations prior to further manual testing.

J Settings of Post-training

All post-training experiments were conducted on an 8-GPU A100 server using the LLaMa-Factory framework (Zheng et al., 2024). The training time per session was around five minutes. The specific parameter settings for each group are provided in Table 15. The four sets of training data will be made publicly available at a later stage.

K Supplementary Information

This paper utilized AI tools including Google Translate for assisted translation when presenting prompts and examples, and employed the use of a Cursor for coding to enhance efficiency. No potential risks were involved in the course of this study.

Table 14: The performances of some models as Debtors.

Model	SR	RR	QRD	HRD	CD	L1D	L2D	ATV	CRI	DHI	CCI
Qwen-2.5-72B	0.98	0.88	36.98	185.18	404.98	3.76	78.84	0.83	0.76	0.76	0.76
llama-3-8b	1.00	0.94	29.13	134.13	296.25	3.25	80.31	0.91	0.83	0.71	0.81
llama-3-70b	1.00	0.92	10.33	150.33	369.33	0.33	51.33	0.85	0.83	0.97	0.85
gpt-4o-2024-11-20	1.00	0.94	35.26	146.86	312.66	3.50	85.72	0.86	0.82	0.73	0.80
o1-mini	0.98	0.93	26.76	111.96	240.56	4.92	92.34	0.93	0.85	0.58	0.78
deepseek-chat	0.97	0.93	32.42	125.32	269.48	3.74	93.00	0.90	0.83	0.66	0.79
Doubao-pro-4k	1.00	0.83	75.28	190.48	324.72	2.16	80.34	0.84	0.73	0.82	0.74
abab6.5s-chat	0.90	0.92	58.53	204.53	484.53	8.63	76.50	0.89	0.70	0.52	0.66
SenseChat	1.00	0.96	135.0	345.00	734.00	0.70	51.00	0.88	0.54	0.96	0.60

Table 15: Hyperparameters of Each Post-trained Model.

Model Name	**Parameters**	**Comments**
SFT-DG	"temperature": 0, "max_tokens": 1024, train_batch_size: 4,"finetuning_type": lora, "learning_rate": 5.0e-6, "num_train_epochs": 5.0, "bf16": true	model = "qwen-2.5-7b-instruct"
SFT-MAG	"temperature": 0, "max_tokens": 1024,"train_batch_size": 4,"finetuning_type": lora, "learning_rate": 5.0e-6, "num_train_epochs": 5.0, "bf16": true	model = "qwen-2.5-7b-instruct"
DPO-DG	"temperature": 0, "max_tokens": 1024,"train_batch_size": 4,"finetuning_type": lora, "learning_rate": 5.0e-6, "num_train_epochs": 5.0, "bf16": true	model = "qwen-2.5-7b-instruct"
DPO-MAG	"temperature": 0, "max_tokens": 1024,"train_batch_size": 4,"finetuning_type": lora, "learning_rate": 5.0e-6, "num_train_epochs": 5.0, "bf16": true	model = "qwen-2.5-7b-instruct"