A Diverse and Effective Retrieval-Based Debt Collection System with Expert Knowledge

Jiaming Luo[1], Weiyi Luo[2], Guoqing Sun[2], Mengchen Zhu[2], Haifeng Tang[2], Mengyue Wu[1]*, Kenny Q. Zhu[3]*

[1] X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, AI Institute
Shanghai Jiao Tong University, Shanghai, China
[2]China Merchants Bank Credit Card Center, Shanghai, China
[3]University of Texas at Arlington, Arlington, Texas, USA
[1]{leojm2017, mengyuewu}@sjtu.edu.cn,
[2]{luoweiyi, gqsun, zmc1996, thfeng}@cmbchina.com,
[3]kenny.zhu@uta.edu

Abstract

Designing effective debt collection systems is crucial for improving operational efficiency and reducing costs in the financial industry. However, the challenges of maintaining script diversity, contextual relevance, and coherence make this task particularly difficult. This paper presents a debt collection system based on real debtor-collector data from a major commercial bank. We construct a script library from real-world debt collection conversations, and propose a two-stage retrieval based response system for contextual relevance. Experimental results show that our system improves script diversity, enhances response relevance, and achieves practical deployment efficiency through knowledge distillation. This work offers a scalable and automated solution, providing valuable insights for advancing debt collection practices in real-world applications.

1 Introduction

Debt collection plays a crucial role in the financial industry. In practice, outbound calls for debt recovery are typically handled by experienced experts, since negotiating with debtors is often challenging. Consequently, large companies must employ substantial number of staff to manage daily debt collection tasks, leading to high operational costs. This has spurred interest in developing systems that assist human experts or automate outbound calls, making it a burgeoning area of research (Zhang et al., 2018; Wang et al., 2020).

Recent advancements have demonstrated the feasibility of automatic outbound agents (Zhang et al., 2023a; Wang et al., 2020). Currently, many collection chatbots are flow-based systems configured with rule-based frameworks authored by experts (Wang et al., 2020; Jia et al., 2020b). In these systems, the chatbot predicts the debtor's intent at each stage and provides predefined responses based on established rules. However, such flow-based systems face notable limitations. They heavily depend on expert-crafted rules, making them difficult to update and scale to different scenarios due to their complexity. Additionally, these systems lack response diversity, as the output is fixed for each scenario.

To address these limitations, researchers have explored using pretrained language models to generate responses based on dialogue context (Zhang et al., 2023a; Jin et al., 2023; Jia et al., 2020a;

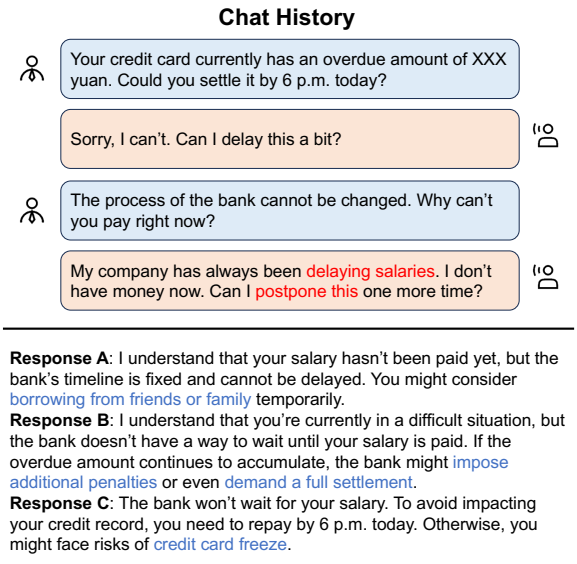

Figure 1: An exemplar between a debtor and a collector, with three candidate responses. The debtor's intent is labeled in red while the strategies in collector's responses are labeled in blue.

*Corresponding authors.

Zhang et al., 2023b). These methods eliminate the need for predefined rules by fine-tuning models on large-scale debt collection conversations. However, generative models often produce responses that may be ineffective in debt collection. The responses are also difficult to control due to their inherent uncertainty.

In view of these problems, retrieval based response system become a better choice in practice, as the response outputs are more controllable. Typically, it consists of two stages: script[1] generation and response system implementation. As for script generation, current practice remains predominantly a manual process undertaken by experienced experts. However, previously-mentioned challenges still exist. First, achieving script diversity is inherently challenging, as generating distinct responses for a wide range of scenarios demands significant effort. Second, updating the system is resource-intensive, requiring expert intervention to craft and integrate new scripts with each revision. To address these issues, automatic script generation from real conversations has become a promising direction.

On the other hand, response retrieval in debt collection is particularly challenging due to several factors. In practice, we find that embedding-based methods, while effective in other domains (Su et al., 2023; Zhang et al., 2022), struggle here due to the difficulty of distinguishing positive from negative samples without manual annotation. Typically, positive samples are selected from actual responses in the dialogue, and negative samples are randomly chosen from other dialogues. However, this random negative sampling often leads to situations where the selected "negative" samples are actually suitable responses for the current dialogue, resulting in "false negatives". This increases the complexity of model training and affects the accuracy of the system, particularly when multiple responses in the script library appear valid during inference. To this end, we propose a two-stage retrieval based response system to select the most effective script from script library.

In this work, we propose a comprehensive system for automatic outbound chat-bots that integrates script generation and selection models. Leveraging the capabilities of Large Language Models (LLMs), we first generate diverse and effective scripts based on real-world conversations

[1]The term "script" refers to predefined response or standardized dialogue templates used by debt collection agents.

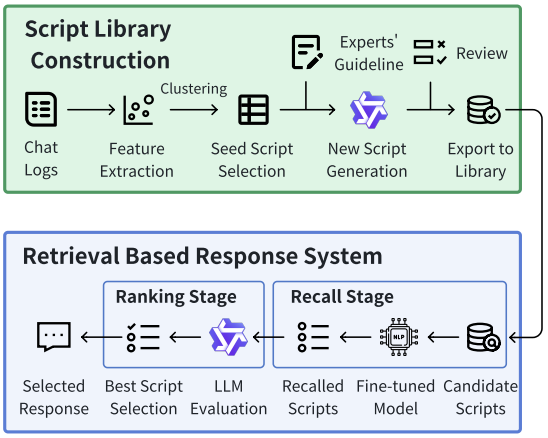

Figure 2: Overview of the SCORES framework. A script library is constructed from chat logs, followed by a two-stage response selection system.

while incorporating expert knowledge to enhance the quality and naturalness of the dialogues. After that, to ensure the safety and appropriateness of outbound calls, we frame the problem as response selection, where the model must choose the optimal response given the dialogue context. Since traditional embedding-based models often struggle to distinguish between similar scripts with identical strategies, to overcome this, we design a two-stage retrieval pipeline: the first stage employs a pre-trained model to recall n relevant responses, and the second stage uses LLMs to evaluate and select the best response based on three aspects: *Empathetic Engagement*, *Effective Problem-Solving* and *Contextual Relevance*. The contributions of our work are threefold:

1. A novel framework to generate high-quality scripts by leveraging insights from human conversations and domain expertise, where we automatically obtain more than 1,000 scripts for 9 strategies accepted by experts.

2. A two-stage retrieval pipeline that efficiently tackles the challenges of response selection, elevating Recall@1 from 0.346 to 0.577.

3. An automatic outbound framework SCORES: **S**cript **C**reation and **O**ptimized **RE**sponse **S**ystem, a scalable and practical pipeline with minimal supervision that can easily extend to other related domains including marketing and intelligent customer service.

2 Methodology

Our proposed framework SCORES includes two modules: (a) Automatic script library construction and (b) Retrieval based collection system.

2.1 Automatic Script Library Construction

The script library is a fundamental component of the debt collection system. Our approach utilizes LLMs to generate diverse scripts based on real debt collection dialogue history from a major bank's Debt Recovery Department that involves a large amount of daily debt collection calls.

Data Preparation We begin by collecting voice recordings of interactions between debtors and collectors over several days from a major commercial bank. These recordings are transcribed using an automatic speech recognition (ASR) system. Each dialogue transcription T between a debtor D and a collector C is organized into a turn-taking format: $T = \{c_1, d_1, c_2, d_2, \ldots, c_n, d_n\}$, where c_i and d_i represent the collector's and debtor's utterances, respectively. Each collector's utterance c_i is assigned a strategy label $s_i \in S$, (e.g., "Pressure through letters", "Pressure through family"), and each debtor's utterance d_i is assigned with purpose label $p_i \in P$, (e.g., "Inability to repay", "Unemployment"). Here S is the pre-defined strategy list while P is the pre-defined purpose list. These labels can be annotated by experts or automatically extracted using fine-tuned language models. Next, we extract utterance pairs $[d_i, c_i]$ from each dialogue and filter out pairs without applicable strategy or purpose labels. This process results in a collection of m labeled utterance pairs $U = \{d_i, c_i\}, i \in \{1, \ldots, m\}$.

Seed Scripts Selection Everyday conversation data contains diverse debt collection strategies. However, variations in speaking styles and scenarios make it challenging to generalize patterns for each strategy. To address this, we select seed scripts for each strategy from the utterance pairs U. We first divide collectors' utterances by strategy and use embedding models to represent each utterance as a d dimensional vector e_i. Here we employ BGE-M3 (Chen et al., 2024) to extract 1024-dimensional embeddings. These embeddings are clustered using the K-means algorithm, producing K clusters: $\mathbb{E} = \{E_1, E_2, \ldots, E_K\}, E_i = \{e_1^i, e_2^i, \ldots, e_j^i\}$. The mean of each cluster's embeddings is computed as the cluster center: $o_i = \frac{1}{j}\sum_{m=1}^{j} e_m^i$. For each cluster, we select the top-5 embeddings closest to the center as representative "seed scripts". These scripts capture distinct "persuasive patterns" for the strategy. This process yields $5 \times K$ seed scripts for each strategy.

Script Generation Using the selected seed scripts, we generate additional scripts tailored for debt collection using Qwen2-72B (Yang et al., 2024). To ensure contextual fluency and coherence, generated scripts are aligned with the debtor's purpose p_i. We incorporate expert guideline for each purpose into the generation process. For example, if a debtor mentions its unemployment during the conversation, the response should first empathize and then proceed with the standard collection strategy. In practice, the purpose-specific guidelines and the seed scripts are input into the LLM to generate three new scripts per cluster. These scripts are labeled with p_i and s_i for subsequent use. Generated scripts are reviewed and refined by experts before being added to the script library, whose results are illustrated in Section 3.3.

2.2 Retrieval-based Response System

The response system generates or retrieves responses during debt collection conversations. We adopt a retrieval-based approach for safety and reliability. Our response selection pipeline consists of two stages: *recall* and *ranking*. The recall stage is designed to efficiently narrow down a large pool of candidate responses to a smaller subset that is contextually relevant to the conversation. The ranking stage then refines this subset, selecting the most appropriate response based on LLM evaluations. This two-stage process ensures both scalability in handling a large response database and precision in selecting high-quality responses.

Recall Stage The recall stage identifies the top-n candidate scripts from the library. Given a context history h_i and the purpose p_i of the debtor's last utterance, the recall model retrieves the most appropriate scripts labeled with p_i. We pre-process conversation transcriptions by dividing them into sub-conversations using a sliding window. Each sub-conversation consists of five utterances as context h_i and the sixth utterance as the response r_i: $h_i = \{d_i, c_{i+1}, d_{i+1}, c_{i+2}, d_{i+2}\}, r_i = c_{i+3}$

Following prior work on response selection (Su et al., 2023), we use Chinese-BERT-wwm (Cui et al., 2021) as the base model M. The model is first pretrained with a Masked Language Modeling

(MLM) objective and fine-tuned using contrastive learning:

$$\mathcal{L} = \sum_{i=1}^{m} \frac{\exp(w_i^+)}{\exp(w_i^+) + \sum_{j=1}^{n_{neg}} \exp(w_i^j)} \quad (1)$$

where $w_i^+ = \text{sim}(h_i, r_i^+), w_i^j = \text{sim}(h_i, r_i^{j-})$. r_i^+ is the correct response, r_i^{j-} are negative samples, and $\text{sim}(h_i, r_i)$ is the cosine similarity between embeddings. We use the [CLS] token of the last hidden layer of M as the embedding of the texts.

After fine-tuning, the model M encodes h_i and candidate scripts into embeddings. During inference, M generates embeddings for the given context, and the top-n most similar scripts are retrieved as recall results.

Ranking Stage Although the recall stage reduces the pool of candidate responses, selecting the best script remains challenging due to the nuanced, indirect alignment between the conversational context and the desired strategy. To address these issues, we leverage LLMs to evaluate and select the best response from the candidates chosen in the recall stage. An intuitive approach involves assessing candidate scripts based on several predefined aspects. After consulting with debt collection experts, we identified three critical aspects for evaluation: *Empathetic Engagement*, *Effective Problem-Solving* and *Contextual Relevance*. Detailed definitions can be found in appendix A.1.

Inspired by G-Eval (Liu et al., 2023), we define three levels for each aspect: excellent (3), good (2), and poor (1). Each level is supported by detailed criteria, crafted by experts. During the evaluation process, we combine the context in 3 turns and each candidate script into a prompt template, instructing the LLM to score the script according to the predefined criteria (see appendix A.2). The average score across the three aspects serves as the overall score for each candidate. The script with the highest overall score is selected as the response. In cases of tied scores, the script ranked higher in the recall stage is chosen.

Despite the effectiveness of LLM evaluation, the inference time for large models, such as Qwen2-72B, is prohibitively high for real-time response systems. To mitigate this, we apply a knowledge distillation approach, transferring expertise from the large LLM (72B-model) to a more computationally efficient small LLM (1.5B/3B-model). Specifically, we use Qwen2-72B model to generate labeled data by evaluating context-candidate pairs using the predefined criteria. These evaluation scores and accompanying rationales serve as the labels.

We then fine-tune smaller LLMs (e.g., Qwen2.5-3B (Team, 2024)) on the labeled dataset. We set the context-candidate pair and evaluation criteria as inputs, while the evaluations generated by the Qwen2-72B model are the desired outputs. After fine-tuning, the smaller LLM can efficiently perform ranking, significantly reducing inference time while maintaining acceptable performance. For example, the Recall@1 metric for the Qwen2.5-3B model improved significantly from 0.404 to 0.577 after fine-tuning. Additional experimental results are provided in Section 3.3.

3 Experiments

In this section, we present the experimental settings and results of our proposed methods.

3.1 Datasets

For script library construction, we processed 786 debt collection calls, transcribed them using ASR tools, and annotated debtor utterances with a pretrained purpose classification model. LLMs identified collector strategies, and experts refined the annotations, yielding 6,218 labeled utterances. All our data is in Chinese.

For response system construction, we transcribed 4,000 additional calls and used the classification model to annotate purposes without further human review. After segmenting dialogues, we obtained over 40,000 context-response pairs, split into training, validation, and test sets (8:1:1). For knowledge distillation, the Qwen2-72B model generated 13,000 cases in Alpaca format.

3.2 Implementation Details

Script Library Construction We used the BGE-M3 model to encode sentences into 1024-dimensional vectors. For seed script selection, the utterances were clustered into $K = 4$ groups using K-means, and five utterances nearest to each cluster center were selected as seed scripts. We use Qwen2-72B model for script generation.

Response System Construction We used the Chinese-BERT-wwm model with a truncation length of 256. Pretraining employed a 30% masking ratio, a 1×10^{-4} learning rate, and five epochs, selecting the best model via validation. Fine-tuning

used a 5×10^{-5} learning rate, a batch size of 64, and AdamW (Loshchilov and Hutter, 2019) optimizer for five epochs. For ranking, Qwen2.5-3B and Qwen2.5-1.5B models were fine-tuned with LoRA (Hu et al., 2021) on the LLaMA-Factory platform (Zheng et al., 2024). All experiments ran on a V100 GPU server.

3.3 Results and Discussions

Script Library To evaluate the effectiveness of K-means clustering, we calculated the intra-cluster distance d_{intra} and the inter-cluster distance d_{inter}. For clusters corresponding to each strategy, $\mathbb{E} = \{E_1, E_2, \ldots, E_K\}$, where $E_i = \{e_1^i, e_2^i, \ldots, e_j^i\}$, the intra-cluster distance for each cluster is computed as the average distance between all embeddings and the cluster center:

$$d_{\text{intra}} = \frac{1}{K} \frac{1}{|E_i|} \sum_{i=1}^{K} \sum_{k=1}^{|E_i|} d(x_k^i, o_i) \quad (2)$$

Here, o_i denotes the center of cluster i, and $d(x, y)$ represents the L2 distance between two vectors. Similarly, the inter-cluster distance is calculated as the average distance between embeddings within a cluster and the centers of all other clusters:

$$d_{\text{inter}} = \frac{1}{K} \frac{1}{K-1} \frac{1}{|E_i|} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \sum_{k=1}^{|E_i|} d(x_k^i, o_j) \quad (3)$$

These metrics assess the compactness of clusters and the separability between different strategies.

From Table 1, we observe that the intra-cluster distance is smaller than the inter-cluster distance, which demonstrates the effectiveness of the clustering method. This result indicates that seed scripts within the same cluster exhibit higher similarity (consistency), while those across different clusters show greater variation (diversity).

To further assess the diversity of generated scripts, we compute the Distinct-n metrics (Li et al., 2016) under different seed script selection methods. Random refers to selecting 5 utterances randomly as seed scripts for each strategy. The configurations $k = 1$ and $k = 4$ differ in the number of clusters. Specifically, $k = 1$ means selecting the top-5 utterances closest to the center of all strategy embeddings, whereas $k = 4$ involves clustering the utterances into four groups and selecting 5 utterances nearest to the center of each cluster.

Table 1: Intra-distance and inter-distance comparison.

Strategy	d_{intra}	d_{inter}
Pressure Through Letters	0.3361	0.4910
Card Suspension	0.2921	0.5144
Full Payment	0.3126	0.4743
Negotiation Plan	0.3306	0.4984
Cash Advance	0.4382	0.5178
Pressure Through Family	0.3613	0.6021
Credit Report	0.3363	0.4708
Repayment Ability	0.3959	0.4683
Anti-Disconnection	0.3116	0.4992
Average	**0.3491**	**0.4946**

Table 2: Distinct-n evaluation across different seed script selection strategies. The best results are in bold.

Selection	**Distinct-1**	**Distinct-2**
Random	0.131	0.466
$k = 1$	0.129	0.466
$k = 4$	**0.141**	**0.500**

We evaluate the diversity using scripts generated by the same LLM across 5 randomly sampled purposes and 9 predefined strategies (as listed in Table 1). The total number of generated scripts for the Random and $k = 1$ settings is $5 \times 9 \times 3 = 135$. For the $k = 4$ setting, we generate $3 \times 4 = 12$ scripts for each purpose-strategy pair and randomly sample 3 scripts, maintaining the evaluation size at 135 scripts for comparability. We evaluate the diversity using Distinct-1 and Distinct-2, where higher scores indicate greater diversity.

As shown in Table 2, the $k = 4$ configuration achieves the highest Distinct-n scores among the three settings. This result demonstrates that the clustering-based method effectively generates scripts that are both diverse and consistent within their respective clusters.

We further evaluate the script library's performance in real-world scenarios. For an A/B test, we replaced the existing scripts with those generated by the LLM while keeping the chatbot workflow unchanged. During a month-long online test involving approximately 600,000 outbound calls, the script replacement led to a 0.5% improvement in recovery rate. This shows the effectiveness of our script generation method.

Recall Stage We evaluate the performance of our fine-tuned model in the recall stage using Re-

call@K (R@K) on the test set. In this evaluation, the candidate set contains 10 utterances, in which 1 utterance is designated as the ground truth. We compare the model's performance with and without the pretraining stage. As shown in Table 3, the model's performance improves significantly with the inclusion of the pretraining stage.

Table 3: Performance comparison w/ or w/o pretraining

Model	R@1	R@2	R@3	R@5
w/ pre.	**0.617**	**0.782**	**0.870**	**0.957**
w/o pre.	0.594	0.762	0.859	0.951

When comparing these results to those reported in the E-Commerce Dataset (Su et al., 2023), the Recall@K metrics are noticeably lower. For example, R@1 for the baseline model (BERT+CL) reaches 0.849 in the E-Commerce dataset but only achieves 0.671 in our dataset. This highlights the complexity of our response selection task, underscoring the necessity of adopting a two-stage selection pipeline to address these challenges effectively.

Ranking Stage To evaluate the performance of different models in the ranking stage, we employed 7 debt collection experts to select the best response for a given context from 3 candidate utterances from the recall stage. The most frequently selected utterance is regarded as the ground truth. In total, 52 cases were labeled as the test set, with a Fleiss' kappa value of 0.41, indicating "Moderate Agreement." This highlights the inherent difficulty of selecting the best response from candidates from the recall stage.

Table 4: Performance comparison of ranking models on Recall@1. Models with the "-sft" suffix denote the models are supervised fine-tuned on the dataset labeled by Qwen2-72B. "BERT" refers to the fine-tuned model used in the recall stage, while "72B" represents Qwen2-72B, "3B" represents Qwen2.5-3B, and "1.5B" represents Qwen2.5-1.5B.

Model	BERT	72B	3B-sft	3B	1.5B-sft	1.5B
R@1	0.346	**0.731**	0.577	0.404	0.538	0.423

Then we compared the performance of 5 LLMs against the BERT model baseline by evaluating Recall@1 on the labeled test set. The results are summarized in Table 4. The results indicate that the performance of the recall stage remains suboptimal, with Recall@1 slightly surpassing random guess (0.333). Despite this, score-based methods using LLMs demonstrate promising results. Notably, the 72B-model, even without supervised fine-tuning, shows a significant improvement over the baseline. Similarly, the 3B and 1.5B models also outperform the baseline, highlighting the potential of LLMs as effective ranking models for complex tasks.

Moreover, after distilling knowledge from the 72B model, the performance of the 3B and 1.5B models improves significantly. This demonstrates the feasibility of leveraging smaller LLMs in real-world applications by distilling knowledge from larger models.

4 Related Work

Retrieval-Based Dialogue Systems Retrieval-based dialogue systems aim to identify the most appropriate response from a set of candidates (Jia et al., 2021; Jin et al., 2023). These systems are widely applied in domains such as customer service Q&A and forum post interactions (Lowe et al., 2015; Zhang et al., 2018; Wu et al., 2016). Modern approaches predominantly leverage pre-trained language models (PLMs) like BERT (Devlin et al., 2019), fine-tuned using contrastive learning on domain-specific corpora (Xu et al., 2021; Zhang et al., 2022, 2023b). To enhance semantic relevance and contextual coherence, Han et al. (Han et al., 2021) incorporate fine-grained labels during post-training. Su et al. (Su et al., 2023) propose a novel post-training method that improves context embeddings. Additionally, Han et al. (Han et al., 2024) introduce EDHNS, which optimizes contrastive learning by focusing on harder-to-distinguish negative examples.

Automatic Outbound Chatbots Automatic outbound chatbots are designed to engage customers in conversations to achieve specific goals, such as debt collection or advertising. Traditional systems often relied on flow-based approaches due to their straightforward logic and ease of implementation (Lee et al., 2008; Yan et al., 2017). However, these systems heavily depend on expert-defined rules and are challenging to update. To address these limitations, recent research has shifted towards response generation using PLMs. Jin et al. (Jin et al., 2023) propose a persuasion framework that integrates both semantic understanding and strategic considerations. Zhang et al. (Zhang et al., 2023a) enhance response generation by incorporating user

profiles extracted during conversations. Qian et al. (Qian et al., 2022) redefine the dialogue process as a sequence-labeling problem, leveraging a dual-path model for joint multi-task learning.

5 Conclusion

In this work, we designed and evaluated a comprehensive system, SCORES, for automating outbound debt collection, addressing challenges of script diversity, adaptability, and effective response selection. By combining the script generation capabilities of LLMs with a robust two-stage retrieval framework, we achieved notable improvements in response effectiveness. Besides, knowledge distillation enhanced its efficiency for real-world deployment. More importantly, the flexibility of this framework allows it to be adapted to a wide range of domains, such as customer support and telemarketing. Future work will focus on further refining script diversity, improving real-time response evaluation, and expanding the framework's applicability to ensure even higher levels of performance and adaptability in diverse settings.

Ethical Considerations

In our experiments, call records were collected with customer consent. To ensure data privacy, personal information such as names and phone numbers was removed during script generation and further training. When testing online, the responses generated by SCORES are exclusively retrieved from the script library, where all scripts were carefully reviewed to eliminate any inappropriate content.

Acknowledgements

This work has been supported by the CMB Credit Card Center & SJTU joint research grant and Guangxi major science and technology project (No. AA23062062).

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558.

Janghoon Han, Dongkyu Lee, Joongbo Shin, Hyunkyung Bae, Jeesoo Bang, Seonghwan Kim, Stanley Jungkyu Choi, and Honglak Lee. 2024. Efficient dynamic hard negative sampling for dialogue selection. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 89–100, Bangkok, Thailand. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Qi Jia, Hongru Huang, and Kenny Q Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13125–13133.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Q Zhu, and Haifeng Tang. 2020a. Multi-turn response selection using dialogue dependency relations. *arXiv preprint arXiv:2010.01502*.

Qi Jia, Mengxue Zhang, Shengyao Zhang, and Kenny Q Zhu. 2020b. Matching questions and answers in dialogues from online forums. In *ECAI 2020*, pages 2046–2053. IOS Press.

Chuhao Jin, Yutao Zhu, Lingzhen Kong, Shijie Li, Xiao Zhang, Ruihua Song, Xu Chen, Huan Chen, Yuchong Sun, Yu Chen, et al. 2023. Joint semantic and strategy matching for persuasive dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4187–4197.

Changyoon Lee, You-Sung Cha, and Tae-Yong Kuc. 2008. Implementation of dialogue system for intelligent service robots. In *2008 International Conference on Control, Automation and Systems*, pages 2038–2042. IEEE.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Ruifeng Qian, Shijie Li, Mengjiao Bao, Huan Chen, and Yu Che. 2022. Toward an optimal selection of dialogue strategies: A target-driven approach for intelligent outbound robots. *arXiv preprint arXiv:2206.10953*.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023. Dial-mae: Contextual masked auto-encoder for retrieval-based dialogue systems. *arXiv preprint arXiv:2306.04357*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zihao Wang, Jia Liu, Hengbin Cui, Chunxiang Jin, Minghui Yang, Yafang Wang, Xiaolong Li, and Renxin Mao. 2020. Two-stage behavior cloning for spoken dialogue system in debt collection. In *IJCAI*, pages 4633–4639.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Tong Zhang, Junhong Liu, Chen Huang, Jia Liu, Hongru Liang, Zujie Wen, and Wenqiang Lei. 2023a. Towards effective automatic debt collection with persona awareness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 32–45.

Wentao Zhang, Shuang Xu, and Haoran Huang. 2022. Two-level supervised contrastive learning for response selection in multi-turn dialogue. *arXiv preprint arXiv:2203.00793*.

Zhiling Zhang, Mengyue Wu, and Kenny Q Zhu. 2023b. Semantic space grounded weighted decoding for multi-attribute controllable dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13230–13243.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

A Appendix

A.1 Evaluation Aspects for LLM

1. **Empathetic Engagement:** This aspect evaluates the politeness and the ability to show empathy and understanding for the debtor's difficulties.
2. **Effective Problem-Solving:** This aspect assesses whether the script effectively communicates the consequences of contract breaches and provides a viable solution.
3. **Contextual Relevance:** This aspect determines whether the script maintains logical coherence with the preceding text.

A.2 Prompt for LLM Evaluation

Task Description
You will evaluate the effectiveness of candidate scripts in a debt collection context. Please rate the scripts based on three dimensions, with scores ranging from 1 to 3 (1 being the lowest and 3 being the highest), and provide a brief explanation for each score. Ensure you have carefully read and understood the task instructions.

Evaluation Dimensions and Criteria
Empathetic Engagement:
•**Excellent (3):** The script uses a professional tone to inform the customer about the overdue issue while showing care and empathy towards customers facing difficulties. It employs empathetic expressions and avoids complex or unclear language. The script clearly conveys the importance of the situation while maintaining politeness and professionalism.
•**Good (2):** The script somewhat considers the customer's emotions but may include mechanical or templated expressions, lacking deeper emotional connection.
•**Poor (1):** The script appears stiff or indifferent, ignoring the customer's emotions or using rude language. Such scripts may provoke resistance or dissatisfaction, reducing cooperation willingness.
Effective Problem-Solving:
•**Excellent (3):** The script clearly communicates the consequences of non-payment (e.g., sending notices, freezing accounts) and provides actionable solutions tailored to the customer's situation (e.g., seeking help from family or friends). The consequences and solutions are easy to understand and motivate the customer to act promptly.
•**Good (2):** The script mentions the consequences of non-payment but does not provide clear or actionable solutions. It may describe possible solutions but lacks specificity in guiding the customer to resolve the issue.
•**Poor (1):** The script fails to convey any consequences or propose solutions. The content is vague and does not encourage the customer to take any action.
Contextual Relevance:
•**Excellent (3):** The script closely aligns with the prior conversation, particularly by accurately responding to the customer's last statement. It maintains logical consistency with the dialogue history, demonstrating strong contextual understanding and ensuring a smooth, natural flow of conversation.
•**Good (2):** The script is somewhat related to the dialogue history but lacks natural or adequate follow-through. It may overlook some details, resulting in slightly awkward transitions.
•**Poor (1):** The script completely deviates from the prior dialogue, failing to address the customer's last statement or maintain logical continuity, leading to a lack of coherence and contextual fit.

Evaluation Steps:
1.Carefully read and understand the dialogue history and candidate script. The dialogue history represents past interactions between the customer and the debt collection agent, while the candidate script is a potential agent response to be evaluated.
2.Based on the scoring criteria above, evaluate the candidate script across the three dimensions: Customer Perception, Goal Alignment, and Contextual Relevance. Assign scores from 1 to 3, where 1 is the lowest and 3 is the highest.
3.Provide a brief explanation for each score based on the assigned rating and the given dialogue data.

Input Format:
Dialogue History: Includes prior conversation context.
Candidate Script: The script to be evaluated.
(Note: The dialogue content is generated from ASR transcripts and may contain recognition errors.)

Output Format:
Provide your evaluation in the following JSON format:

```
{
   "Empathetic Engagement": {
      "Score": score_1:int,
      "Explanation": "Explanation for the Empathetic Engagement score."
   },
   " Effective Problem-Solving ": {
      "Score": score_2:int,
      "Explanation": "Explanation for the Effective Problem-Solving score."
   },
   "Contextual Relevance": {
      "Score": score_3:int,
      "Explanation": "Explanation for the Contextual Relevance score."
   }
}
```

Requirements:
Your evaluation must be based on the dialogue history and candidate script, ensuring logical consistency. The more realistic and rigorous your assessment, the better it will help the system improve the adaptability of its scripts. Please consider all factors comprehensively and provide scores in the specified format. Do not include any additional or unnecessary content.

Figure 3: The prompt used for LLM evaluation.