Self-Preference: An Automated Method for Preference-Aligned Data Constructed from Business Metrics

Feng Gao[1], **Xuan Zhang**[1], **Boyi Ni**[1], **Chunping Wang**[1], **Lei Chen**[1,2,†]
[1]FinVolution Group, 999 Dangui Road, Shanghai 201203, China
[2]School of Software and Microelectronics, Peking University, Beijing 100871, China
{gaofeng02, zhangxuan02, niboyi, wangchunping02, chenlei04@xinye.com} †

Abstract

Large language models (LLMs) have become integral components of various AI solutions, with the reinforcement learning from human feedback (RLHF) stage playing a critical role in aligning model outputs with human preferences. However, generating the human preference data required for RLHF is often costly and time-consuming due to its reliance on human evaluation. This study addresses this challenge within the dialogue scenarios of the fintech industry. We leverage rich, non-confidential, multi-turn dialogue data, such as call center dialogue records, which include associated business metrics (e.g., problem-solving rates, turnover ratios) to construct preference-aligned data. We introduce *Self-Preference*, an automated method for creating preference-aligned data guided by these objective business metrics. The approach involves clustering dialogue histories based on their semantic representations and calculating a well-designed *conditional probability ratio* that correlates sequences with business metrics to generate preference data. In contrast to traditional preference alignment data generation methods that depend on subjective human evaluations, Self-Preference significantly reduces labeling costs and mitigates model-induced biases. Experimental results indicate that models trained with Self-Preference generated data demonstrate a strong positive correlation with target business metrics, highlighting the method's effectiveness in facilitating efficient, goal-oriented alignment of LLMs.

1 Introduction

Large language models (LLMs) have rapidly evolved, becoming essential components in both professional environments and daily life (Ray, 2023). Chat models represent one of the most successful applications of LLMs. They follow human instructions, engage in seamless multi-turn conversations, and assist with a variety of practical tasks, such as customer service (Shi et al., 2024), business analysis (Cheung, 2024), and code generation (Tong and Zhang, 2024). These capabilities stem from fundamental language abilities and the post-training period that involves extensive supervision and preference alignment for the LLMs. LLMs training typically involves three stages: pre-training (PT), supervised fine-tuning (SFT), and reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022). After the RLHF stage, the model's outputs are better aligned with human preferences, such as being safer and more honest, compared to the outputs from previous stages.

In the RLHF stage, the training data consists of preference-aligned data, which differs from the SFT stage where only *positive* training data are required. The data format for RLHF requires two or more different outputs for each given instruction and input: the chosen outputs and the rejected ones. The chosen output aligns with human preferences; thus, the training objective in this stage is to ensure that the model's output aligns as closely as possible with the chosen output. Utilizing preference-aligned data, the RLHF stage employs advanced alignment methods such as PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023), ORPO (Hong et al., 2024) and SimPO (Meng et al., 2024) for model training.

[†] Corresponding author: chenlei04@xinye.com

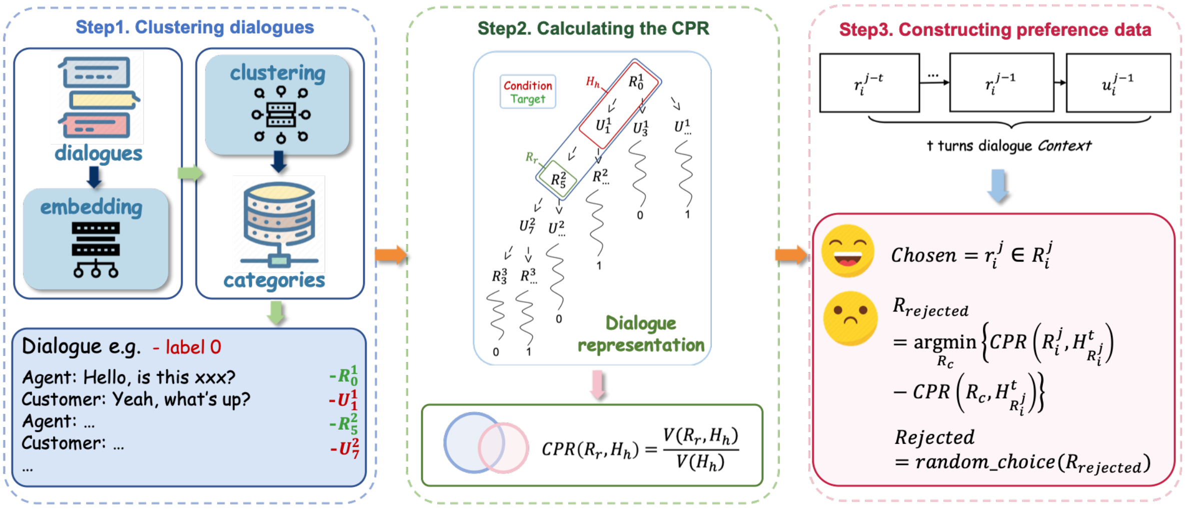

Step1. Clustering dialogues
dialogues
embedding
clustering
categories
Dialogue e.g. - label 0
Agent: Hello, is this xxx? $-R_0^1$
Customer: Yeah, what's up? $-U_1^1$
Agent: … $-R_5^2$
Customer: … $-U_7^2$
…
Step2. Calculating the CPR
Condition
Target
H_h
R_r
R_0^1
U_1^1 U_3^1 $U_{\dots}^1$
R_5^2 $R_{\dots}^2$
U_7^2 $U_{\dots}^2$
R_3^3 $R_{\dots}^3$
0 1 0 1 0 1
Dialogue representation
$CPR(R_r, H_h) = \frac{V(R_r, H_h)}{V(H_h)}$
Step3. Constructing preference data
r_i^{j-t} … r_i^{j-1} u_i^{j-1}
t turns dialogue Context
$Chosen = r_i^j \in R_i^j$
$R_{rejected} = \underset{R_c}{\text{argmin}} \left\{ CPR\left(R_i^j, H^t_{R_i^j}\right) - CPR\left(R_c, H^t_{R_i^j}\right) \right\}$
$Rejected = random\_choice(R_{rejected})$

Figure 1: A high-level overview of Self-Preference. Firstly, we cluster the dialogue history for both customer and agent roles separately. Next, we calculate the Conditional Probability Ratio based on cluster sequences and business metrics. Finally, turn-level preference alignment data are constructed automatically.

The quantity, quality, and diversity of preference-aligned data are crucial to the model's effectiveness. However, acquiring such data typically requires a large-scale annotation effort by humans, involving processes such as drafting annotation guidelines, conducting quality checks during the annotation process, and consolidating data afterward. Overall, it is an expensive and resource-intensive project, which is often classified as a systems engineering project.

In the fintech industry, there is a growing demand for AI assistants to reduce repetitive human tasks while maintaining work quality and business output. As LLMs continue to advance, many personalized communications and services can now be tailored to be performed by AI agents. A wealth of business contexts, such as dialogue records between human agents and customers, has been gathered in the call center. This dataset provides an essential resource for developing language models tailored to specific domains. It includes typical scenarios such as marketing, debt collection, and customer service interactions. These interactions feature multiple exchanges between the human agent and the customer. Initially, the conversations are recorded as audio files and then transcript into text, where they are saved as a dialogue history corpus. Additionally, each conversation is associated with business metrics, including call conclusions like ”customer not interested” or ”customer will repay”, and performance indicators like ”following login” or ”problem solved”.

In this work, we introduce *Self-Preference*, an automatic LLMs preference alignment data construction method guided by business metrics in multi-turn dialogue scenarios (see Figure 1). Firstly, we separate the dialogue corpus into two participating roles, ”agent” and ”customer” and perform clustering on them, respectively. Next, for each dialogue in the corpus, we map the clusters back to the original dialogue content, resulting in a mutual-interactive cluster sequence that serves as an abstract representation of the dialogue. The conditional probability ratio (CPR) is defined and calculated based on the cluster sequences and the business metrics. Finally, we use the CPR to automatically generate preference alignment data. As far as we know, this is the first work to propose an automatic construction method for preference alignment data guided by business metrics in multi-turn dialogue scenarios. Meanwhile, the overall solution is simple and reliable, avoiding the subjectivity of human evaluation and the influence of inherent biases in other large language models, thereby saving a significant amount of human labeling costs. Furthermore, the selection of preference data is strictly based on objective indicator performance, and the model output after the preference data alignment training demonstrates a strong positive correlation with the final business metrics.

In summary, our contributions are outlined as follows:

1. We propose Self-Preference, a pioneering automatic construction method for RLHF alignment guided by business metrics in multi-turn dialogue scenarios. Without the need for additional assistance, this method only takes three steps, i.e., role-separated clustering, CPR calculating, and preference data construction. More importantly, these preference data are oriented by business metrics, ensuring their consistency with business goals.

2. We conducted comprehensive comparisons between the performance of SFT and RLHF in internal real dialogue datasets and public simulated dialogue dataset as well, which consists of customer service and marketing business scenarios. Extensive analysis shows that RLHF consistently outperforms SFT across different evaluation metrics in each scenario. This result indicates that Self-Preference boosted the alignment between the RLHF outputs and the business metric, thereby augmenting the model's practical applicability in business-related scenarios.

3. We will release multi-turn dialogue data simulating a marketing scenario in the fintech industry (it can be found at `https://github.com/FinVolution/Self-Preference`). For generating each dialogue, a topic is randomly selected firstly based on the predefined topic distribution, which is consistent with a real business scenario. Then, Qwen2.5-72B-Instruct (Yang et al., 2024) shifts mutually between the agent role and the customer role, participating in multiple rounds of interactive conversations guided by the dialogue topic. Finally, the value of the business metric is assignment according to a probability model, which also follows the real business scenario. Following the publication of the paper, we'll make the data available on our official GitHub repository.

2 Related Work

Multiple approaches are proposed to automate the generation of training data for both the SFT and RLHF phases.

During the SFT phase, several methods have enabled the automatic generation of vast amounts of instruction data. SELF-INSTRUCT (Wang et al., 2022) improves the instruction-following capabilities of LLMs by bootstrapping the pipeline. Starting with a small seed set as a task pool, SELF-INSTRUCT asks an off-the-shelf LLM to generate instructions and corresponding examples. After filtering out ineffective or redundant samples, the remaining samples are added back to the task pool. Without human-written instruction seeds, based on structured and unstructured unsupervised data, SELF-QA (Zhang and Yang, 2023) first employs the LLM-generated instructions and then generates the corresponding answers by utilizing the LLM again. GenQA (Chen et al., 2024b) generates large-scale instruction datasets with minimal supervision using a single prompt, allowing LLMs to autonomously create diverse instruction examples ranging from simple tasks to complex, multi-turn dialogues. MAGPIE (Xu et al., 2024), as a self-synthesis method, takes advantage of the auto-regressive nature of aligned LLMs. It generates large-scale user queries in aligned datasets by providing only pre-query templates, then prompting the LLM to generate the response.

During the RLHF phase, the SELFEE (Kim et al., 2024) enhances the alignment of the LLM by repeatedly generating responses and refining the model with the self-generated preference data through iterative learning, based on a minimal amount of human-labeled seed preference data. SAFER-INSTRUCT (Shi et al., 2023) leverages reversed instruction tuning, instruction induction, and expert model evaluation to generate high-quality preference data without human annotators. Overall, current approaches to automatically constructing preference alignment data often depend on additional aids, such as initializing with seed data or relying on the support of expert large models, and involve the training of models, which incurs extra resource requirements. Particularly, it is worth noting that the preference alignment data generated by these methods are samples without dialogue context, meaning that the instruction part only contains the current description, without taking into account the dialogue history that has occurred in previous turns of the conversation.

3 Definitions and Method

The process details of Self-Preference consist of three steps: (1) role-separated clustering, (2) conditional probability ratio calculation, and (3) automatic preference data construction. This pipeline is depicted in Figure 1.

3.1 Role-separated clustering

A corpus of dialogues $\langle s \rangle \triangleq \{s_i\}, i = 1, 2, \cdots, m$, along with the corresponding business metrics, is essential to carry out Self-Preference. Each multi-turn dialogue is a so-called session, denoted by s_i:

$$s_i = (r_i^1, u_i^1, r_i^2, u_i^2, \ldots,\ r_i^j, u_i^j, \ldots)$$

where r_i^j and u_i^j denote the response and utterance, respectively, with the turn order indicated by superscript j. From all s_i in $\langle s \rangle$, we collect the responses and utterances to form a response set $\langle r \rangle = \{r_i^j\}$ and an utterance set $\langle u \rangle = \{u_i^j\}$, and clustering is carried out on both $\langle r \rangle$ and $\langle u \rangle$, respectively.

To enhance the multi-dimensional representation of the response or utterance, we integrate literal and semantic representations to achieve vectorized representation.

a) Literal representation

Literal representation involves selecting fine-grained features from the text, such as using Term Frequency-Inverse Document Frequency (TF-IDF) (Ramos, 2003) as a feature, and incorporating n-gram information to account for word order. Finally, Principal Component Analysis (PCA) (Abdi and Williams, 2010) is applied to reduce the dimensionality of the vectors.

b) Semantic representation

To compute the semantic features of the text, open-source vectorization models like BGE Embedding (Chen et al., 2024a) are utilized for semantic vectorized representation.

The two representations above are concatenated to form the final representation of the responses or utterances.

To accommodate various application scenarios, clustering algorithms like hierarchical clustering (Murtagh and Contreras, 2012) or Single-pass clustering (Shahrivari and Jalili, 2016) are suitable, as they do not necessitate pre-specifying the number of clusters. Following the clustering process, we have distinct results for the grouping of both the responses and the utterances. Once the clustering was completed, we ended up with two separate clusters, $\langle R \rangle = \{R_0,\ R_1,\ \ldots,\ R_{|R|}\}$ and $\langle U \rangle = \{U_0,\ U_1,\ \ldots,\ U_{|U|}\}$, respectively. For a specific session s_i, each utterance u_i^j (or response r_i^j) is associated with its cluster label U_i^j (or R_i^j), where $U_i^j \in \langle U \rangle$ (or $R_i^j \in \langle R \rangle$). The session s_i is abstracted into a cluster label sequence:

$$\begin{aligned} & s_i = (r_i^1, u_i^1, r_i^2, u_i^2, \ldots,\ r_i^j, u_i^j, \ldots) \\ & \Rightarrow \\ & S_i = (R_i^1, U_i^1, R_i^2, U_i^2, \ldots,\ R_i^j, U_i^j, \ldots) \end{aligned}$$

where S_i is the nominal representation of s_i. With the above abstraction operation, $\langle S \rangle = \{S_i\}$ is automatically converted into a dataset with m samples, each of which is a sequence of discrete labels.

3.2 Conditional Probability Ratio calculating

As described in the previous section, for each s_i, r_i^j or u_i^j is replaced with the clustering category they belong to, we obtain the substituted S_i. We associate $S_i \in \langle S \rangle$ with a binary variable Y_i as the business metric of S_i. Figure 2 illustrates the dialogue history corpus, using a tree structure where each node represents a clustering category. The path from the root node to a leaf node signifies a complete dialogue. At the leaf nodes, the binary values $\{0, 1\}$ indicate the business metric.

For S_i, let $H^t_{R_i^j} = (R_i^{j-t}, U_i^{j-t} \ldots\ R_i^{j-1}, U_i^{j-1})$ be the previous t turns of dialogue context (i.e., conditional sequence) for the j-th turn agent response R_i^j. The set of all previous t turns of dialogue

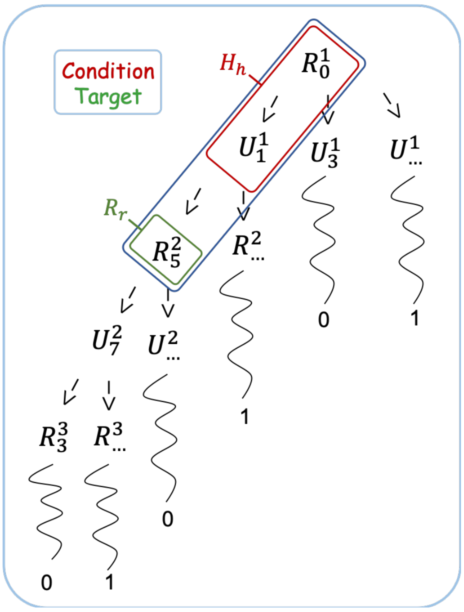


Figure 2: Dialogue history corpus represented by clusters and business metrics. R_r^j and U_u^j are agent response and customer utterance clusters respectively. $0/1$ indicate the business metric. H_h is the conditional sequence and R_r is the target response.

history for the response set $\langle R \rangle$ is denoted as $H^t = \{H^t_{R_r} | R_r \in \langle R \rangle\}.\forall H_h \in H^t$, the success rate of H_h is defined as:

$$V(H_h) = \frac{n_{\{H_h|H_h \subseteq S_i, Y_i=1\}_{i=1}^m}}{n_{\{H_h|H_h \subseteq S_i\}_{i=1}^m}}$$

where $n_{\{H_h|H_h \subseteq S_i, Y_i=1\}_{i=1}^m}$ is the total number of sessions that contain H_h and for which $Y_i = 1$. $n_{\{H_h|H_h \subseteq S_i\}_{i=1}^m}$ represents the total number of sessions that include H_h.

$\forall R_r \in \langle R \rangle$, the success rate of $V(R_r, H_h)$ is defined as:

$$V(R_r, H_h) = \frac{n_{\left\{H_h|H_h \in H^t_{R_r},\ R_r \subseteq S_i, Y_i=1\right\}_{i=1}^m}}{n_{\left\{H_h|H_h \in H^t_{R_r},\ R_r \subseteq S_i\right\}_{i=1}^m}}$$

where $n_{\left\{H_h|H_h \in H^t_{R_r},\ R_r \subseteq S_i, Y_i=1\right\}_{i=1}^m}$ denotes the total number of sessions that encompass R_r, with the dialogue history of the previous t turns being H_h, and for which $Y_i = 1$. $n_{\left\{H_h|H_h \in H^t_{R_r},\ R_r \subseteq S_i\right\}_{i=1}^m}$ is the total number of sessions that include R_r, with the dialogue history of the previous t turns being H_h.

Accordingly, the Conditional Probability Ratio $CPR(R_r, H_h)$ can be defined as:

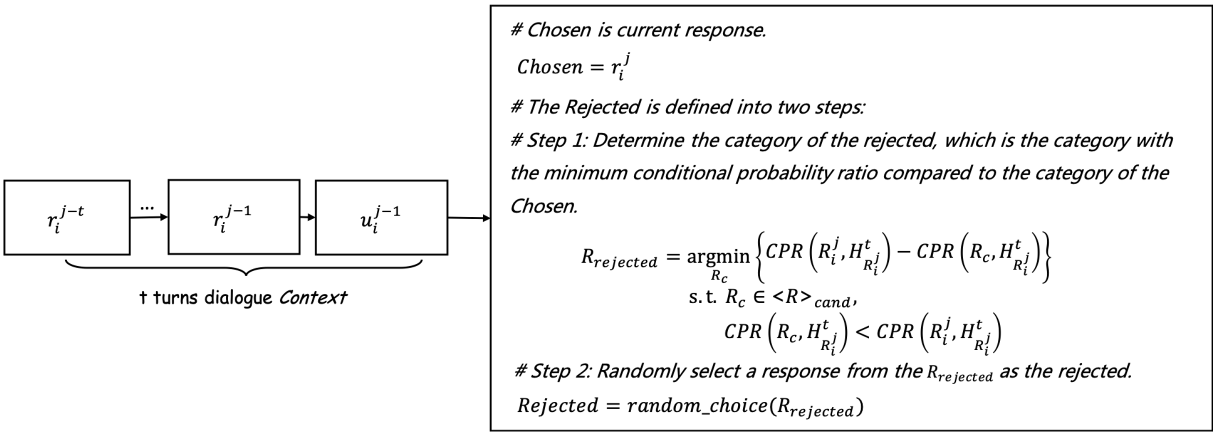
$$CPR(R_r, H_h) = \frac{V(R_r, H_h)}{V(H_h)}$$

The definition outlined above indicates that CPR is determined by the ratio of two fractions. For each R_r, a specific $CPR(R_r,\ H_h)$ can be computed, with a higher value implying a greater contribution made by R_r to the improvement of business metrics.

3.3 Automatic preference data construction

Following the definition of the $CPR(R_r, H_h)$, we can automatically create the preference alignment data for the j-th turn using t turns of dialogue context, as illustrated in Figure 3.

Suppose that the j-th turn agent response of s_i is r_i^j, which belongs R_i^j; the t turns dialogue context is $C_i^{j,t} = (r_i^{j-t}, u_i^{j-t}, \dots,\ r_i^{j-1}, u_i^{j-1})$, and the corresponding substituted clustering category is $H^t_{R_i^j} = (R_i^{j-t}, U_i^{j-t} \dots\ R_i^{j-1}, U_i^{j-1})$.



```
# Chosen is current response.
$Chosen = r_i^j$
# The Rejected is defined into two steps:
# Step 1: Determine the category of the rejected, which is the category with
the minimum conditional probability ratio compared to the category of the
Chosen.
    $R_{rejected} = \underset{R_c}{\text{argmin}} \left\{ CPR\left(R_i^j, H_{R_i^j}^t\right) - CPR\left(R_c, H_{R_i^j}^t\right) \right\}$
        s.t. $R_c \in \langle R \rangle_{cand}$,
            $CPR\left(R_c, H_{R_i^j}^t\right) < CPR\left(R_i^j, H_{R_i^j}^t\right)$
# Step 2: Randomly select a response from the $R_{rejected}$ as the rejected.
$Rejected = random\_choice(R_{rejected})$
```

Figure 3: The j-th turn preference alignment data construction in s_i. Under the premise of t turns dialogue $Context$, r_i^j is defined as the $Chosen$. For $Rejected$, we first determine $R_{rejected}$ category, then randomly select a response from $R_{rejected}$ as $Rejected$.

Given $C_i^{j,t}$ within the multi-turn scenario, we first define r_i^j as the $Chosen$. Then the $Rejected$ is constructed as follows.

For $H_{R_i^j}^t$, we first build the target candidate set $\langle R \rangle_{\text{cand}} = \{R_c | R_c \in \langle R \rangle, H_{R_i^j}^t \in H_{R_c}^t\}$, then compute $CPR(R_c,\ H_{R_i^j}^t)$ for all R_c in $\langle R \rangle_{\text{cand}}$, finally we choose the target category with a lower conditional probability ratio than $CPR(R_i^j,\ \overset{t}{H}_{R_i^j})$ and the minimal difference in this ratio as R_{rejected}.

$$R_{\text{rejected}} = \underset{R_c}{\text{argmin}} \left\{ CPR(R_i^j, H_{R_i^j}^t) - CPR(R_c, H_{R_i^j}^t) \right\}$$
$$\text{s.t.} \quad R_c \in \langle R \rangle_{\text{cand}},$$
$$CPR(R_c, H_{R_i^j}^t) < CPR(R_i^j, H_{R_i^j}^t)$$

By randomly selecting a response r_{rejected} from R_{rejected}, we can automatically create a preference alignment sample that includes t turns of dialogue context.

$$\begin{cases} Context & = C_i^{j,t} \\ Chosen & = r_i^j \\ Rejected & = r_{\text{rejected}} \end{cases}$$

After processing all s_i in $\langle s \rangle$ using the above pipeline, we can obtain the complete set of the preference alignment dataset.

4 Experiments

This section systematically presents the experimental configurations employed in the study. We first apply Self-Preference to both our internal real and public simulated dialogue datasets, which consist of different business scenarios. Then, the effectiveness of Self-Preference in enhancing model alignment is demonstrated based on human ratings and *LLM-as-a-Judge* (Zheng et al., 2023) evaluation metrics.

4.1 Datasets

The entire dataset comprises internal real and public simulated dialogue datasets. The internal real dataset pertains to our actual business, which centers around marketing and customer service scenarios. The public simulation dataset focuses on a marketing scenario. In a customer service scenario, the business metric is *self-service*, indicating that the user did not request a transfer to manual service during the current session. In a marketing scenario, the business metric is ”following login” which reflects the user logging into the app after the session concludes.

Qwen2.5-72B-Instruct is employed to generate a publicly accessible simulation dataset via role-playing (Shanahan et al., 2023). The model alternates between the roles of an agent and a customer, engaging in multiple rounds of interactive dialogue. To create a diverse dataset that better reflects actual business situations, we first determine the dialogue topic by randomly sampling from predefined topics for each simulated dialogue. Then, Qwen2.5-72B-Instruct iteratively assumes both the agent and customer roles to engage in multi-turn interactions guided by the topic. The sampling distribution of both the topic and the business metric for each simulated dialogue is consistent with real business scenarios. The detailed simulation dataset collection process is outlined in Appendix A.1.

Table 1 presents the statistics for two dialogue datasets, which encompass the total number of sessions, average turns, average utterance length (in tokens), average response length (in tokens), and the business metric $Y_i = 1$ rate across different scenarios.

Table 1: Statistics of internal real and public simulated datasets in marketing and customer service scenarios.

Category	Scenario	# of sessions	Ave. turns	Ave. utterance length (in tokens)	Ave. response lengths (in tokens)	Business metric $Y_i = 1$ rate
Internal real	Customer service	15596	6.3	11.4	27.0	0.78
	Marketing	7140	7.7	8.1	41.9	0.4
Public simulated	Marketing	2943	6.7	10.0	24.4	0.4

For both categories and different scenarios, the dataset was partitioned into training, validation, and test sets with an approximate 8:1:1 ratio. Next, we apply the Self-Preference pipeline to construct preference alignment data utilizing the training set. Number of dialogue context turns t is set to 3. Overall, the total number of samples for each training stage after processing is shown in Table 2.

Table 2: The total number of training samples for each training stage.

Category	Scenario	Training stage	# of samples
Internal real	Customer service	SFT	12476
		RLHF	8044
	Marketing	SFT	5712
		RLHF	5426
Public simulated	Marketing	SFT	2354
		RLHF	2045

4.2 Experiment Setups

Baselines

We leverage the recently published Qwen2.5-14B-Instruct (Yang et al., 2024) model as the SFT backbone model. After supervised fine-tuning, we name the trained model SP-SFT. For the preference RLHF training, DPO and ORPO methods are utilized. As DPO is stable (Rafailov et al., 2023) and ORPO requires low computational resources (Hong et al., 2024). We designate the preference-aligned model as SP-RLHF-DPO and SP-RLHF-ORPO, respectively. For SP-RLHF-DPO, SP-SFT serves as the initial model. We compare the performance of SP-RLHF-DPO/ORPO with that of SP-SFT under different scenarios, using the corresponding evaluation metrics (see below).

Evaluation metrics

Consider both subjective and objective perspectives; human evaluation and *LLM-as-a-judge* are performed to assess alignment performance.

Overall, different scenarios have distinct evaluation criteria. The customer service scenario features criteria focusing on problem-solving and minimizing ”transfer to human” requests, while the marketing scenario utilizes criteria such as conversion efficiency and objection handling. Additionally, both customer service and marketing scenarios share criteria *user satisfaction*, along with others, including context memory, information integration, response coherence, and efficiency.

For each session s_i in the test sets, we generate each turn's response given the condition of dialogue $Context$, similar to the preference alignment data construction process. Subsequently, senior business expert evaluators and *LLM-as-a-judge* score each turn of the generated response. Moreover, *LLM-as-a-judge* assessed every sample within the test set, while senior business expert reviewed a random selection of 300 dialogues from the same test set. The average of these scores constitutes the overall rating for each s_i. The detailed prompt template we used for *LLM-as-a-judge* can be found in Appendix A.2.

Implementation details

We fine-tuned SP-SFT and SP-RLHF-DPO/ORPO with LoRA (Low-Rank Adaptation) (Hu et al., 2022), a parameter-efficient fine-tuning method on an A100 GPU. Both models are trained for three epochs with the AdamW_torch optimizer and a cosine learning rate scheduler with a warm-up phase corresponding to 10% of the total training steps, with initial learning rates of 1e-4 and 5e-6 for SP-SFT and SP-RLHF-DPO/ORPO, respectively. For the hyperparameter β/λ for DPO/ORPO, we use a fixed value of 0.1. For LoRA hyperparameters, we first specify all linear modules as target modules to apply LoRA. Regarding the LoRA rank hyperparameter, values of 8, 16, and 32 are evaluated on the validation dataset, and the best-performing checkpoint is selected. All experiments were conducted under LlamaFactory (Zheng et al., 2024) framework.

5 Results

This section illustrates the effectiveness of SP-RLHF-DPO/ORPO compared to the baseline SP-SFT approach across different scenarios and several key evaluation dimensions.

For the marketing scenario, as demonstrated in Table 3 for the internal real dataset and the public simulated dataset, SP-RLHF-DPO/ORPO consistently outperforms SP-SFT across all measured criteria. SP-RLHF-DPO and SP-RLHF-ORPO have roughly the same performance. Specifically, in internal real scenario, both SP-RLHF-DPO and SP-RLHF-ORPO achieves a Memory and consistency metric score of 6.46, representing a 23.5% relative improvement over SP-SFT's score of 5.23. This performance gap becomes more pronounced in Conversion efficiency metric, where SP-RLHF-DPO/ORPO attains 5.96 compared to SP-SFT's 4.71.

For the customer service scenario, as summarized in Table 4 for the internal real dataset, the performance of SP-RLHF-ORPO is between SP-SFT and SP-RLHF-DPO. On Problem-solving metric, SP-RLHF-DPO achieves a score of 7.15, outperforming SP-SFT (6.43) by 11.2%. Meanwhile, on Avoid *transfer to human* requests metric, SP-RLHF-DPO/ORPO demonstrates an even stronger advantage with a score of 8.72/8.62 compared to SP-SFT's 8.33. This systematic gap highlights SP-RLHF's enhanced ability to leverage human preference data for refining model outputs.

The advantage of SP-RLHF-DPO is further evidenced by human evaluation results (Table 5). When assessed by Problem-solving metric, responses generated by SP-RLHF-DPO received significantly higher preference ratings (6.36 vs. 4.72 for SP-SFT) in solving problem capability. SP-RLHF-DPO achieves a score of 6.02 on User satisfaction metric (vs. 4.82, absolute gain of 1.2).

This empirical evidence confirms that RLHF not only preserves the core competencies established during supervised fine-tuning, but also enables substantial improvements in alignment-critical dimensions through human preference optimization.

Table 3: GPT-4 as a judge evaluation in internal real and public simulated dataset marketing scenario. Memory and consistency is short for Context memory and information integration. Qwen is short for Qwen2.5-14B-Instruct.

Category	Stage	Backbone model	Conversion efficiency	Objection handling	User satisfaction	Memory and consistency	Response coherence and efficiency
Internal real	SFT	Qwen	4.71	4.38	5.01	5.23	5.49
	RLHF-DPO	SP-SFT	5.96	5.45	5.98	6.46	6.47
	RLHF-ORPO	Qwen	5.96	5.45	5.98	6.46	6.48
Public simulated	SFT	Qwen	7.55	7.76	8.28	8.02	8.78
	RLHF-DPO	SP-SFT	7.66	7.82	8.29	8.08	8.80
	RLHF-ORPO	Qwen	7.54	7.77	8.28	8.03	8.79

Table 4: GPT-4 as a judge evaluation in the internal real dataset customer service scenario. Memory and consistency is short for Context memory and information integration. Qwen is short for Qwen2.5-14B-Instruct.

Stage	Backbone model	Problem-solving	Avoid *transfer to human* requests	User satisfaction	Memory and consistency	Response coherence and efficiency
SFT	Qwen	6.43	8.33	6.32	7.10	7.06
RLHF-DPO	SP-SFT	7.15	8.72	6.86	7.54	7.43
RLHF-ORPO	Qwen	6.69	8.62	6.56	7.32	7.31

Table 5: Human evaluation between SFT and RLHF-DPO in internal real dataset customer service scenario. Memory and consistency is short for Context memory and information integration. Qwen is short for Qwen2.5-14B-Instruct.

Stage	Backbone model	Problem-solving	Avoid *transfer to human* requests	User satisfaction	Memory and consistency	Response coherence and efficiency
SFT	Qwen	4.72	4.74	4.82	5.09	5.11
RLHF-DPO	SP-SFT	6.36	5.94	6.02	6.42	6.38

6 Discussion

In this work, we present Self-Preference, a novel automated framework for constructing a large language model preference-aligned data systematically guided by business metrics in multi-turn conversational settings. The Self-Preference implementation process is structured into three sequential stages: role-specific clustering, conditional probability ratio derivation, and automatic preference data construction. The framework was derived entirely from the existing corpus and its business metrics, without the requirement of human annotations. By applying Self-Preference to marketing and customer service scenarios, we significantly increased the alignment between the RLHF outputs and the business metrics, which enhances the model's practical value in business applications.

It is worth noting that, given the dialogue history corpus and business metrics, generating business metric-oriented responses can be analogized to a continuous general problem. In our work, we simplify this by discretizing the problem based on role-specific clustering, reducing it to the calculation of CPR based on clustering label sequences. This method also has wide applicability, particularly in scenarios where one reaches the endpoint through repeated and complex paths starting from the initial point.

High clustering quality is a prerequisite for Self-Preference. In this study, we introduced a representation that combines literal and semantic aspects, which has shown good results in engineering practice. To improve the generalization ability to other scenarios, we believe that different representation methods, such as LLM-based representations (Wang et al., 2023), dialogue structuring (Shi et al., 2019), and dialogue representation learning (Zhou et al., 2022) , can further enhance expressive power. The clustering methods applied do not require the determination of the cluster numbers in advance. The corpus from different scenarios can be expected to have varying numbers of clusters; therefore, the specific number of clusters is not overly important. However, every utterance or response must uniquely correspond to a particular cluster.

In certain scenarios, we can extend the binary business metric to multiclass and continuous business metrics. For the multiclass metric, thresholds can be established based on business logic or data distribution and mapped to a binary scenario. For the continuous metric, one can first perform binning and then conduct binary mapping or directly use the Sigmoid function for binarization. We can even redefine $V(H_h)$ and $V(R_r, H_h)$ as a weighted average of continuous metric values; however, this idea needs to be validated in real-world scenarios for future research.

Acknowledgements

The authors would like to express gratitude to interns Xiaoyuan Li, Zhuoer Liang, Aohan Zhang and Xinglong Song for their contributions to partial code implementation, as well as their valuable feedback and in-depth discussions. Additionally, we extend our thanks to senior business experts Jinglu Yuan, Xiaoyue Mao, and Jie Guo for their professional evaluation of the model performance.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

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

Jiuhai Chen, Rifaa Qadri, Yuxin Wen, Neel Jain, John Kirchenbauer, Tianyi Zhou, and Tom Goldstein. 2024b. Genqa: Generating millions of instructions from a handful of prompts. *arXiv preprint arXiv:2406.10323*.

Ming Cheung. 2024. A reality check of the benefits of llm in business, jun. arXiv:2406.10249 [cs].

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024. Aligning large language models with self-generated preference data. *arXiv preprint arXiv:2406.04412*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback, mar. arXiv:2203.02155 [cs].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242(1), pages 29–48. Citeseer.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Saeed Shahrivari and Saeed Jalili. 2016. Single-pass and linear-time k-means clustering based on mapreduce. *Information Systems*, 60:1–12.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Weiyan Shi, Tiancheng Zhao, and Zhou Yu. 2019. Unsupervised dialog structure learning. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1797–1807, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Taiwei Shi, Kai Chen, and Jieyu Zhao. 2023. Safer-instruct: Aligning language models with automated preference data. *arXiv preprint arXiv:2311.08685*.

Jingzhe Shi, Jialuo Li, Qinwei Ma, Zaiwen Yang, Huan Ma, and Lei Li. 2024. Chops: Chat with customer profile systems for customer service with llms, jul. arXiv:2404.01343 [cs].

Weixi Tong and Tianyi Zhang. 2024. Codejudge: Evaluating code generation with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20032–20051, Miami, Florida, USA. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xuanyu Zhang and Qing Yang. 2023. Self-qa: Unsupervised knowledge guided language model alignment. *arXiv preprint arXiv:2305.11952*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100 language models. In Yixin Cao, Yang Feng, and Deyi Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand, aug. Association for Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768, Seattle, United States, July. Association for Computational Linguistics.

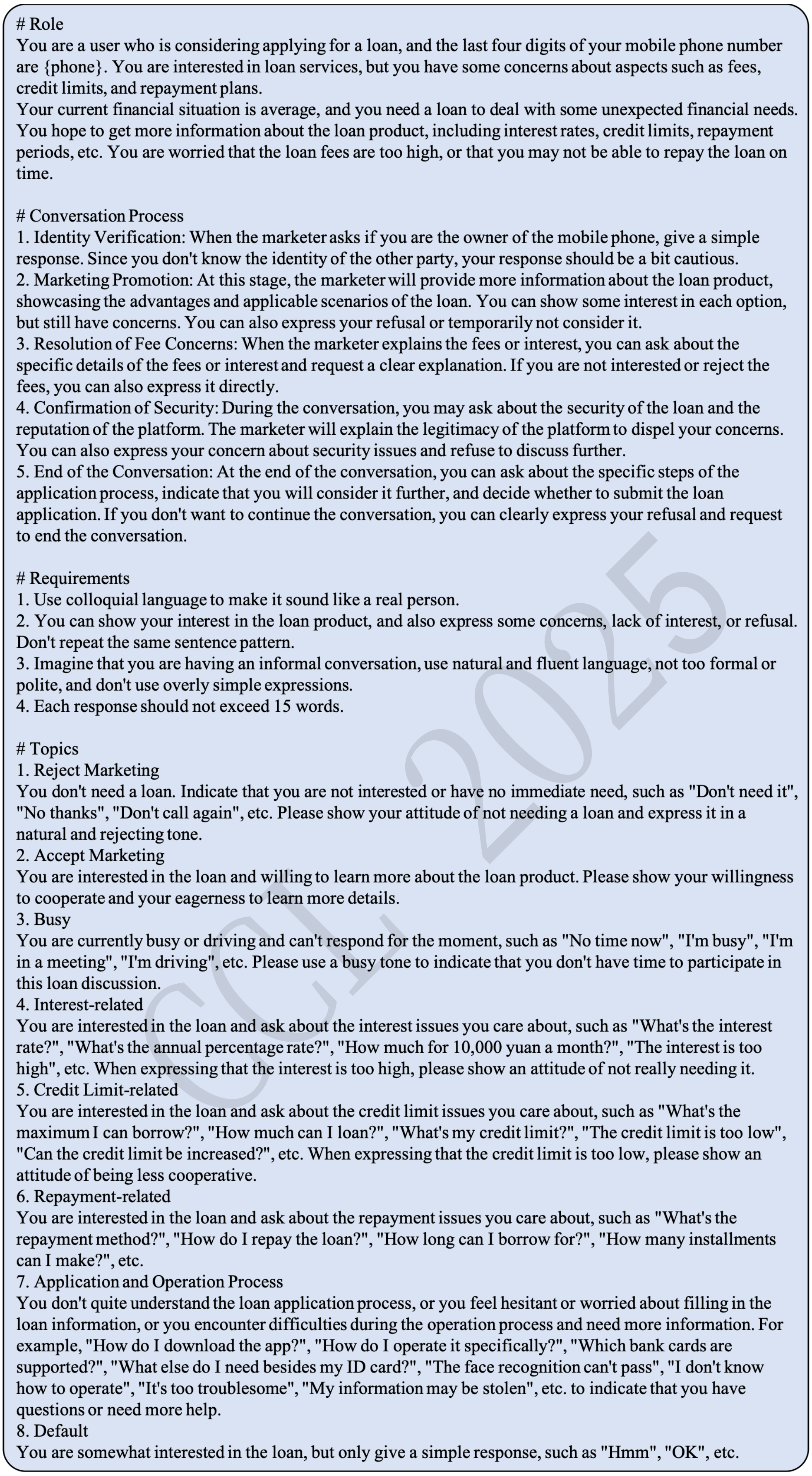
Appendix A

In this section, we elaborate on the specific prompts and instructions utilized for generating public simulation data and *LLM-as-a-judge* dialogue evaluation.

A.1 Prompts for public simulation data generation

A.2 Prompts for *LLM-as-a-judge* dialogue evaluation

Proceedings of the 24th China National Conference on Computational Linguistics, pages 864-879, Jinan, China, August 11-14, 2025.



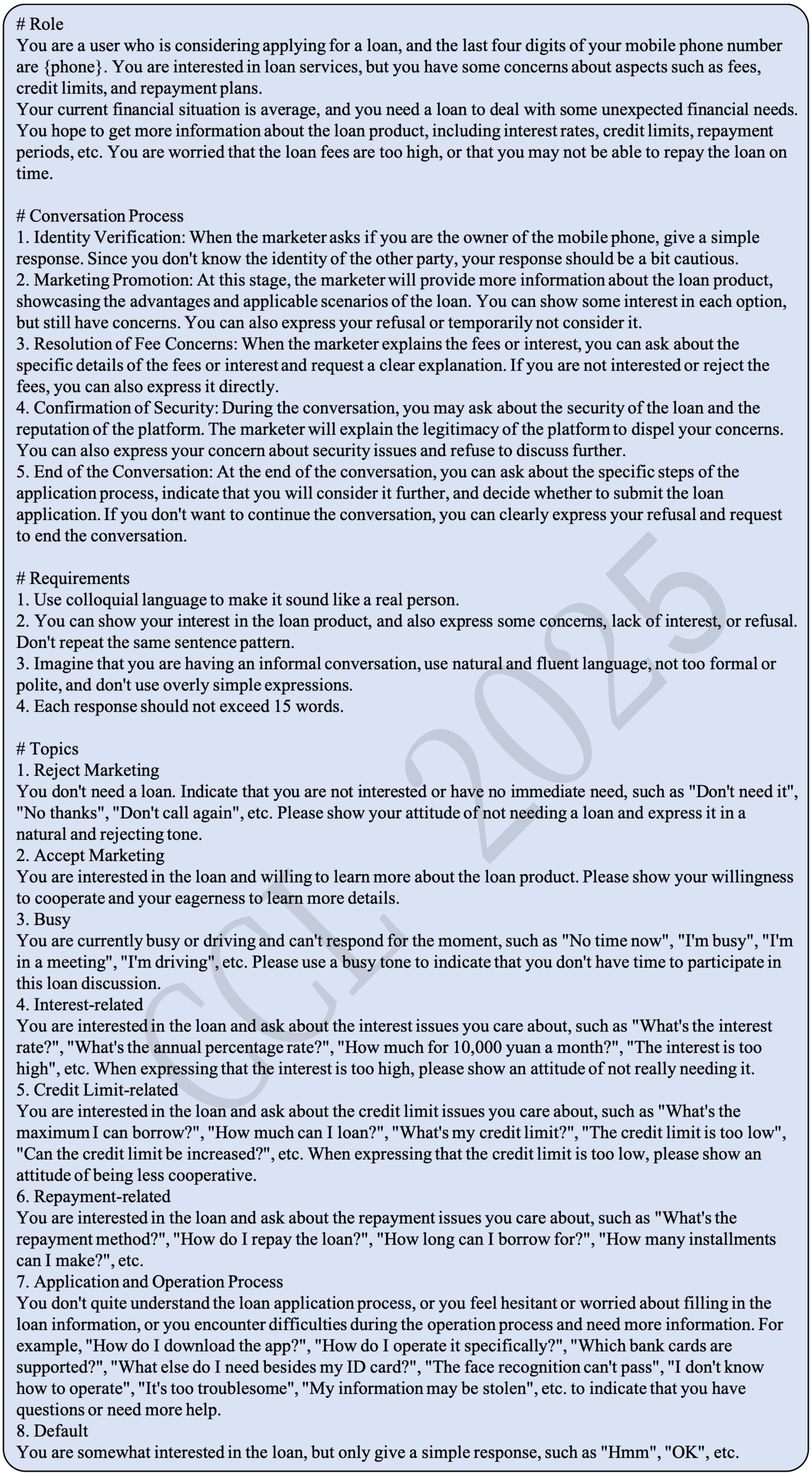
Role
You are a user who is considering applying for a loan, and the last four digits of your mobile phone number are {phone}. You are interested in loan services, but you have some concerns about aspects such as fees, credit limits, and repayment plans.
Your current financial situation is average, and you need a loan to deal with some unexpected financial needs. You hope to get more information about the loan product, including interest rates, credit limits, repayment periods, etc. You are worried that the loan fees are too high, or that you may not be able to repay the loan on time.

Conversation Process
1. Identity Verification: When the marketer asks if you are the owner of the mobile phone, give a simple response. Since you don't know the identity of the other party, your response should be a bit cautious.
2. Marketing Promotion: At this stage, the marketer will provide more information about the loan product, showcasing the advantages and applicable scenarios of the loan. You can show some interest in each option, but still have concerns. You can also express your refusal or temporarily not consider it.
3. Resolution of Fee Concerns: When the marketer explains the fees or interest, you can ask about the specific details of the fees or interest and request a clear explanation. If you are not interested or reject the fees, you can also express it directly.
4. Confirmation of Security: During the conversation, you may ask about the security of the loan and the reputation of the platform. The marketer will explain the legitimacy of the platform to dispel your concerns. You can also express your concern about security issues and refuse to discuss further.
5. End of the Conversation: At the end of the conversation, you can ask about the specific steps of the application process, indicate that you will consider it further, and decide whether to submit the loan application. If you don't want to continue the conversation, you can clearly express your refusal and request to end the conversation.

Requirements
1. Use colloquial language to make it sound like a real person.
2. You can show your interest in the loan product, and also express some concerns, lack of interest, or refusal. Don't repeat the same sentence pattern.
3. Imagine that you are having an informal conversation, use natural and fluent language, not too formal or polite, and don't use overly simple expressions.
4. Each response should not exceed 15 words.

Topics
1. Reject Marketing
You don't need a loan. Indicate that you are not interested or have no immediate need, such as "Don't need it", "No thanks", "Don't call again", etc. Please show your attitude of not needing a loan and express it in a natural and rejecting tone.
2. Accept Marketing
You are interested in the loan and willing to learn more about the loan product. Please show your willingness to cooperate and your eagerness to learn more details.
3. Busy
You are currently busy or driving and can't respond for the moment, such as "No time now", "I'm busy", "I'm in a meeting", "I'm driving", etc. Please use a busy tone to indicate that you don't have time to participate in this loan discussion.
4. Interest-related
You are interested in the loan and ask about the interest issues you care about, such as "What's the interest rate?", "What's the annual percentage rate?", "How much for 10,000 yuan a month?", "The interest is too high", etc. When expressing that the interest is too high, please show an attitude of not really needing it.
5. Credit Limit-related
You are interested in the loan and ask about the credit limit issues you care about, such as "What's the maximum I can borrow?", "How much can I loan?", "What's my credit limit?", "The credit limit is too low", "Can the credit limit be increased?", etc. When expressing that the credit limit is too low, please show an attitude of being less cooperative.
6. Repayment-related
You are interested in the loan and ask about the repayment issues you care about, such as "What's the repayment method?", "How do I repay the loan?", "How long can I borrow for?", "How many installments can I make?", etc.
7. Application and Operation Process
You don't quite understand the loan application process, or you feel hesitant or worried about filling in the loan information, or you encounter difficulties during the operation process and need more information. For example, "How do I download the app?", "How do I operate it specifically?", "Which bank cards are supported?", "What else do I need besides my ID card?", "The face recognition can't pass", "I don't know how to operate", "It's too troublesome", "My information may be stolen", etc. to indicate that you have questions or need more help.
8. Default
You are somewhat interested in the loan, but only give a simple response, such as "Hmm", "OK", etc.

Figure 4: The prompt for public simulation data generation for the customer role.



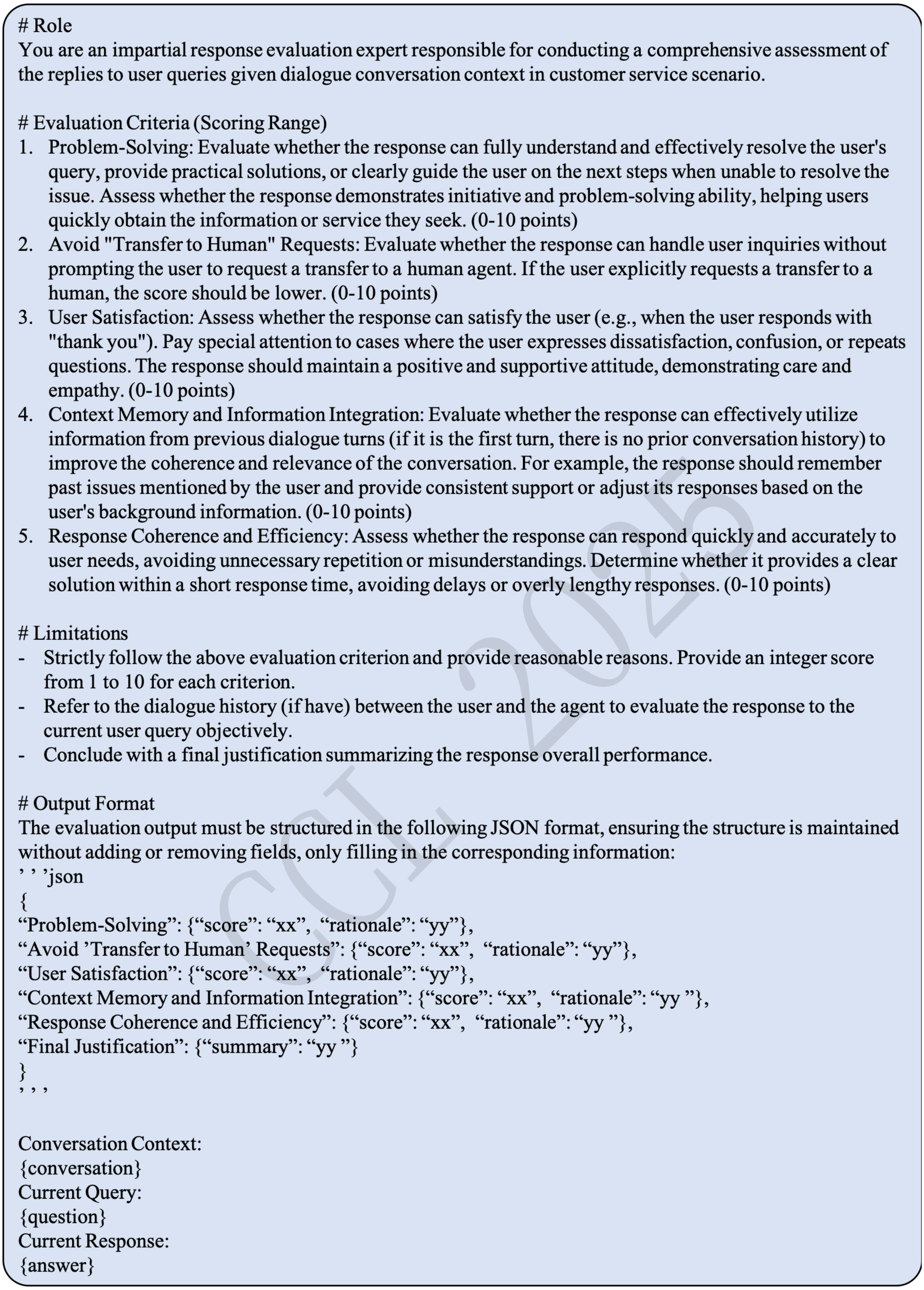
Role
You are a user who is considering applying for a loan, and the last four digits of your mobile phone number are {phone}. You are interested in loan services, but you have some concerns about aspects such as fees, credit limits, and repayment plans.
Your current financial situation is average, and you need a loan to deal with some unexpected financial needs. You hope to get more information about the loan product, including interest rates, credit limits, repayment periods, etc. You are worried that the loan fees are too high, or that you may not be able to repay the loan on time.

Conversation Process
1. Identity Verification: When the marketer asks if you are the owner of the mobile phone, give a simple response. Since you don't know the identity of the other party, your response should be a bit cautious.
2. Marketing Promotion: At this stage, the marketer will provide more information about the loan product, showcasing the advantages and applicable scenarios of the loan. You can show some interest in each option, but still have concerns. You can also express your refusal or temporarily not consider it.
3. Resolution of Fee Concerns: When the marketer explains the fees or interest, you can ask about the specific details of the fees or interest and request a clear explanation. If you are not interested or reject the fees, you can also express it directly.
4. Confirmation of Security: During the conversation, you may ask about the security of the loan and the reputation of the platform. The marketer will explain the legitimacy of the platform to dispel your concerns. You can also express your concern about security issues and refuse to discuss further.
5. End of the Conversation: At the end of the conversation, you can ask about the specific steps of the application process, indicate that you will consider it further, and decide whether to submit the loan application. If you don't want to continue the conversation, you can clearly express your refusal and request to end the conversation.

Requirements
1. Use colloquial language to make it sound like a real person.
2. You can show your interest in the loan product, and also express some concerns, lack of interest, or refusal. Don't repeat the same sentence pattern.
3. Imagine that you are having an informal conversation, use natural and fluent language, not too formal or polite, and don't use overly simple expressions.
4. Each response should not exceed 15 words.

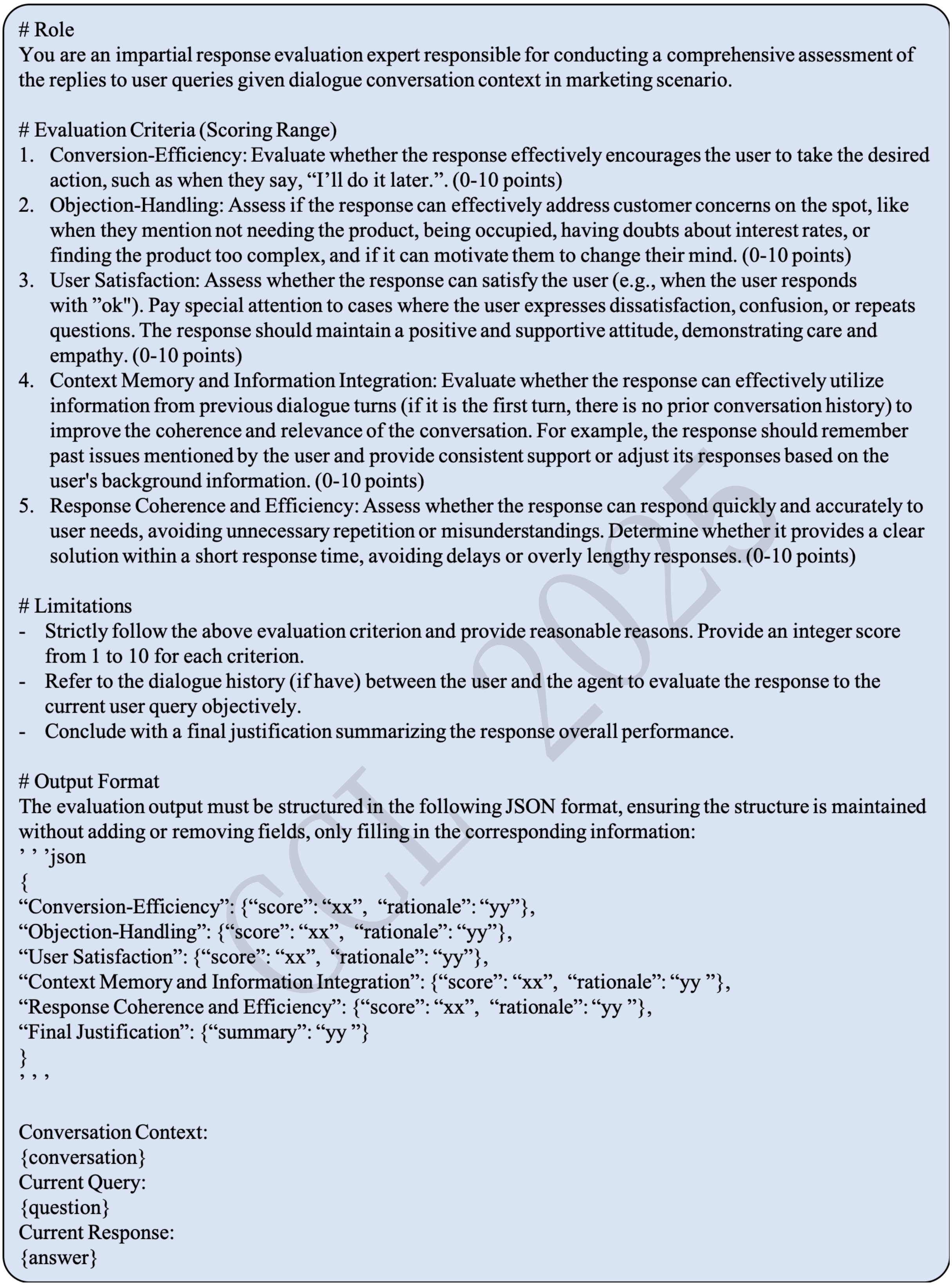
Topics
1. Reject Marketing
You don't need a loan. Indicate that you are not interested or have no immediate need, such as "Don't need it", "No thanks", "Don't call again", etc. Please show your attitude of not needing a loan and express it in a natural and rejecting tone.
2. Accept Marketing
You are interested in the loan and willing to learn more about the loan product. Please show your willingness to cooperate and your eagerness to learn more details.
3. Busy
You are currently busy or driving and can't respond for the moment, such as "No time now", "I'm busy", "I'm in a meeting", "I'm driving", etc. Please use a busy tone to indicate that you don't have time to participate in this loan discussion.
4. Interest-related
You are interested in the loan and ask about the interest issues you care about, such as "What's the interest rate?", "What's the annual percentage rate?", "How much for 10,000 yuan a month?", "The interest is too high", etc. When expressing that the interest is too high, please show an attitude of not really needing it.
5. Credit Limit-related
You are interested in the loan and ask about the credit limit issues you care about, such as "What's the maximum I can borrow?", "How much can I loan?", "What's my credit limit?", "The credit limit is too low", "Can the credit limit be increased?", etc. When expressing that the credit limit is too low, please show an attitude of being less cooperative.
6. Repayment-related
You are interested in the loan and ask about the repayment issues you care about, such as "What's the repayment method?", "How do I repay the loan?", "How long can I borrow for?", "How many installments can I make?", etc.
7. Application and Operation Process
You don't quite understand the loan application process, or you feel hesitant or worried about filling in the loan information, or you encounter difficulties during the operation process and need more information. For example, "How do I download the app?", "How do I operate it specifically?", "Which bank cards are supported?", "What else do I need besides my ID card?", "The face recognition can't pass", "I don't know how to operate", "It's too troublesome", "My information may be stolen", etc. to indicate that you have questions or need more help.
8. Default
You are somewhat interested in the loan, but only give a simple response, such as "Hmm", "OK", etc.

Figure 5: The prompt for public simulation data generation for the agent role.



```
# Role
You are an impartial response evaluation expert responsible for conducting a comprehensive assessment of the replies to user queries given dialogue conversation context in customer service scenario.

# Evaluation Criteria (Scoring Range)
1. Problem-Solving: Evaluate whether the response can fully understand and effectively resolve the user's query, provide practical solutions, or clearly guide the user on the next steps when unable to resolve the issue. Assess whether the response demonstrates initiative and problem-solving ability, helping users quickly obtain the information or service they seek. (0-10 points)
2. Avoid "Transfer to Human" Requests: Evaluate whether the response can handle user inquiries without prompting the user to request a transfer to a human agent. If the user explicitly requests a transfer to a human, the score should be lower. (0-10 points)
3. User Satisfaction: Assess whether the response can satisfy the user (e.g., when the user responds with "thank you"). Pay special attention to cases where the user expresses dissatisfaction, confusion, or repeats questions. The response should maintain a positive and supportive attitude, demonstrating care and empathy. (0-10 points)
4. Context Memory and Information Integration: Evaluate whether the response can effectively utilize information from previous dialogue turns (if it is the first turn, there is no prior conversation history) to improve the coherence and relevance of the conversation. For example, the response should remember past issues mentioned by the user and provide consistent support or adjust its responses based on the user's background information. (0-10 points)
5. Response Coherence and Efficiency: Assess whether the response can respond quickly and accurately to user needs, avoiding unnecessary repetition or misunderstandings. Determine whether it provides a clear solution within a short response time, avoiding delays or overly lengthy responses. (0-10 points)

# Limitations
- Strictly follow the above evaluation criterion and provide reasonable reasons. Provide an integer score from 1 to 10 for each criterion.
- Refer to the dialogue history (if have) between the user and the agent to evaluate the response to the current user query objectively.
- Conclude with a final justification summarizing the response overall performance.

# Output Format
The evaluation output must be structured in the following JSON format, ensuring the structure is maintained without adding or removing fields, only filling in the corresponding information:
’ ’ ’json
{
“Problem-Solving”: {“score”: “xx”, “rationale”: “yy”},
“Avoid ’Transfer to Human’ Requests”: {“score”: “xx”, “rationale”: “yy”},
“User Satisfaction”: {“score”: “xx”, “rationale”: “yy”},
“Context Memory and Information Integration”: {“score”: “xx”, “rationale”: “yy ”},
“Response Coherence and Efficiency”: {“score”: “xx”, “rationale”: “yy ”},
“Final Justification”: {“summary”: “yy ”}
}
’ ’ ’

Conversation Context:
{conversation}
Current Query:
{question}
Current Response:
{answer}
```

Figure 6: The prompt for evaluating response in customer service scenario.



```
# Role
You are an impartial response evaluation expert responsible for conducting a comprehensive assessment of
the replies to user queries given dialogue conversation context in marketing scenario.

# Evaluation Criteria (Scoring Range)
1.  Conversion-Efficiency: Evaluate whether the response effectively encourages the user to take the desired
    action, such as when they say, “I’ll do it later.”. (0-10 points)
2.  Objection-Handling: Assess if the response can effectively address customer concerns on the spot, like
    when they mention not needing the product, being occupied, having doubts about interest rates, or
    finding the product too complex, and if it can motivate them to change their mind. (0-10 points)
3.  User Satisfaction: Assess whether the response can satisfy the user (e.g., when the user responds
    with ”ok"). Pay special attention to cases where the user expresses dissatisfaction, confusion, or repeats
    questions. The response should maintain a positive and supportive attitude, demonstrating care and
    empathy. (0-10 points)
4.  Context Memory and Information Integration: Evaluate whether the response can effectively utilize
    information from previous dialogue turns (if it is the first turn, there is no prior conversation history) to
    improve the coherence and relevance of the conversation. For example, the response should remember
    past issues mentioned by the user and provide consistent support or adjust its responses based on the
    user's background information. (0-10 points)
5.  Response Coherence and Efficiency: Assess whether the response can respond quickly and accurately to
    user needs, avoiding unnecessary repetition or misunderstandings. Determine whether it provides a clear
    solution within a short response time, avoiding delays or overly lengthy responses. (0-10 points)

# Limitations
-   Strictly follow the above evaluation criterion and provide reasonable reasons. Provide an integer score
    from 1 to 10 for each criterion.
-   Refer to the dialogue history (if have) between the user and the agent to evaluate the response to the
    current user query objectively.
-   Conclude with a final justification summarizing the response overall performance.

# Output Format
The evaluation output must be structured in the following JSON format, ensuring the structure is maintained
without adding or removing fields, only filling in the corresponding information:
’ ’ ’json
{
“Conversion-Efficiency”: {“score”: “xx”,  “rationale”: “yy”},
“Objection-Handling”: {“score”: “xx”,  “rationale”: “yy”},
“User Satisfaction”: {“score”: “xx”,  “rationale”: “yy”},
“Context Memory and Information Integration”: {“score”: “xx”,  “rationale”: “yy ”},
“Response Coherence and Efficiency”: {“score”: “xx”,  “rationale”: “yy ”},
“Final Justification”: {“summary”: “yy ”}
}
, , ,

Conversation Context:
{conversation}
Current Query:
{question}
Current Response:
{answer}
```

Figure 7: The prompt for evaluating response in marketing scenario.