

ChatMap: Mining Human Thought Processes for Customer Service Chatbots via Multi-Agent Collaboration

Xinyi Jiang¹ Tianyi Hu^{2,3} Yuheng Qin² Guoming Wang^{1,*}
Huan Zhou² Kehan Chen² Gang Huang² Rongxing Lu⁴ Siliang Tang¹

¹Zhejiang University ²Alibaba Group

³Beijing University of Posts and Telecommunications ⁴Queen’s University

{jiangxinyi, NB21013, siliang}@zju.edu.cn

{hty444504, qinyuheng.qyh, xinyan.zh, kehan.ckh, tengyuan.hg}@alibaba-inc.com
rongxing.lu@queensu.ca

Abstract

Leveraging Large Language Models (LLMs) to build domain-specific conversational agents, especially for e-commerce customer service chatbots, is a growing focus. While existing methods enhance dialogue performance by extracting core patterns from dialogue data and integrating them into models, two key challenges persist: (1) heavy reliance on human experts for dialogue strategy induction, and (2) LLM-based automatic extraction often focuses on summarizing specific behaviors, neglecting the underlying thought processes behind strategy selection. In this paper, we present **ChatMap**, which focuses on enhancing customer service **Chatbots** by mining thought processes using a **Multi-Agent aP**proach. Specifically, the process begins by extracting customer requests and solutions from a raw dialogue dataset, followed by clustering similar requests, analyzing the thought processes behind solutions, and refining service thoughts. Through a quality inspection and reflection mechanism, the final service thought dataset is generated, helping chatbots provide more appropriate responses. Offline experimental results show that ChatMap performs comparably to manually annotated thought processes and significantly outperforms other baselines, demonstrating its ability to automate human annotation and enhance dialogue capabilities through strategic understanding. Online A/B tests on Taobao, a popular e-commerce platform in China reveal that ChatMap can better improve customer satisfaction and address customer requests from a business perspective.

1 Introduction

With the widespread application of Large Language Models (LLMs), research has increasingly focused on developing domain-specific conversational agents (Sun et al., 2024a; Zhong et al., 2024; Zeng et al., 2024), especially customer service chatbots for e-commerce platforms (Sánchez Cuadrado

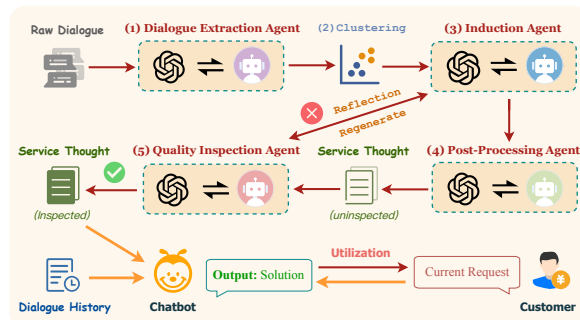


Figure 1: The overview of the proposed ChatMap.

et al., 2024; Chen et al., 2024; Zhang et al., 2024). To achieve coherent, human-like conversations with customers, LLMs need to align their dialogue strategies with human behavior.

Most approaches employ supervised fine-tuning (SFT) or retrieval-augmented generation (RAG) on large-scale human dialogue datasets (Kong et al., 2023; Park et al., 2023; Xu et al., 2023a). Yet, these methods face dual challenges: (1) Models implicitly acquire dialogue patterns, making it difficult to explicitly grasp core dialogue rules. (2) Human dialogue datasets inherently contain substantial noise, such as pleasantries, emotional statements, and off-topic discussions, which obstruct effective information extraction and learning (Zhou et al., 2020). Alternative methods inspired by human cognition propose first extracting dialogue patterns from corpora and then incorporating them into the model’s learning (Du et al., 2020; Yu et al., 2020). However, (1) existing dialogue strategy induction work heavily relies on manual effort, requiring human reading of large amounts of dialogue, followed by classification and annotation of dialogue strategies, which is labor-intensive and difficult to generalize (Cambazoglu et al., 2021; Welivita and Pu, 2020). (2) While recent LLM-based automation attempts address this bottleneck (Thomas et al., 2024; Gilardi et al., 2023), they predominantly focus on surface-

level behavior abstraction, such as dialogue intentions (e.g., appeasing the customer) and strategies used (e.g., questioning the customer to understand their needs), while neglecting the learning of how to apply logical reasoning in different contexts to select appropriate dialogue strategies (Ziems et al., 2024; Ou et al., 2024). In other words, they focus only on "what was done" without capturing "why it was done," which is crucial for understanding the rationale behind strategy selection.

To address these challenges, we propose ChatMap, a multi-agent system that automatically mines human customer service agents' thought processes in dialogues to enhance chatbot performance. Based on the raw dialogue dataset collected from interactions between human agents and customers on Taobao, a popular e-commerce platform in China, ChatMap follows a structured mining process comprising the following steps: (1) Extract requests and solutions from historical conversations. (2) Cluster similar requests. (3) Induce higher-level service thoughts from the requests and solutions within each cluster. (4) Post-process by retaining representative service thoughts to reduce their number. (5) Inspect the quality of service thoughts and introduce a reflection mechanism for regeneration. Ultimately, we obtain a refined and reasonable dataset of service thoughts. During utilization, the customer service chatbot generates appropriate responses based on the dialogue history, the customer's current request, and the corresponding service thought. The overview of our ChatMap framework is illustrated in Figure 1.

The main contributions of our work can be summarized as follows:

(1) We highlight the importance of mining service thoughts to improve customer service chatbot dialogue performance and propose ChatMap to automate this mining process.

(2) To evaluate the quality of responses generated by e-commerce chatbots from multiple perspectives, we design ChatEval-EC, which shows high agreement with human evaluations.

(3) Experimental results reveal that service thought mining via ChatMap outperforms all baseline methods, with its automatic induction achieving performance comparable to manual induction.

(4) In real-world online A/B testing on Taobao platform, ChatMap reduces the human agent transfer rate, while improving customer satisfaction and the immediate resolution rate.

2 Related Works

Inductive Reasoning. Large language models (LLMs) have demonstrated powerful capabilities across a wide range of language-related tasks (Touvron et al., 2023; Chowdhery et al., 2022; OpenAI, 2023), such as the ability to draw conclusions by going from the specific to the general from large amounts of observations, i.e., inductive reasoning (Lake et al., 2017). In recent years, numerous studies have emerged to enhance model induction capabilities, which can be broadly categorized into few-shot learning (Min et al., 2022; Izacard et al., 2023), post-processing (e.g., self-consistency prompting (Wang et al., 2022a), self-refine (Madaan et al., 2024), hypothesis search and refinement (Wang et al., 2023; Qiu et al., 2023)), and fine-tuning (Sun et al., 2024b; Wang et al., 2022b). Simultaneously, many datasets have been proposed to evaluate models' induction capabilities, including visual abstract reasoning (Chollet, 2019; Xu et al., 2023c; Kim et al., 2022), relational reasoning (Li et al., 2023; Zhang et al., 2019; Johnson et al., 2017), causal induction (Du et al., 2022; Zhang et al., 2021), and others (Sinha et al., 2019; Yang et al., 2022).

Automated Dialogue Strategy Learning. Leveraging the inductive capabilities of LLMs, automated summarization from dialogue corpora effectively enables learning high-quality dialogue strategies, which can then be integrated into models to enhance performance (Xu et al., 2023b; Ding et al., 2023). This approach is widely applied in customer service chatbots (Kumar et al., 2023; Bhayana, 2024), voice assistants (Piñeiro-Martín et al., 2023; Sannigrahi et al., 2024), and smart home devices (King et al., 2024; Bouchabou et al., 2021), enabling coherent and natural conversations with humans. TnT-LLM (Wan et al., 2024) is applied to analyze user intent and conversational domains, automating the entire process of label generation and assignment with minimal human effort. Hypotheses-to-Theories (HtT) (Zhu et al., 2024) learns a rule library for reasoning with LLMs. IDEAS (Ou et al., 2024) induces high-level strategies from real instructional dialogues, which are then deductively applied to new dialogue scenarios.

3 Method

Figure 2 illustrates the pipeline of ChatMap. In this section, we detail ChatMap including the source and structure of our dialogue dataset (§3.1), the methods for mining thought processes and refining

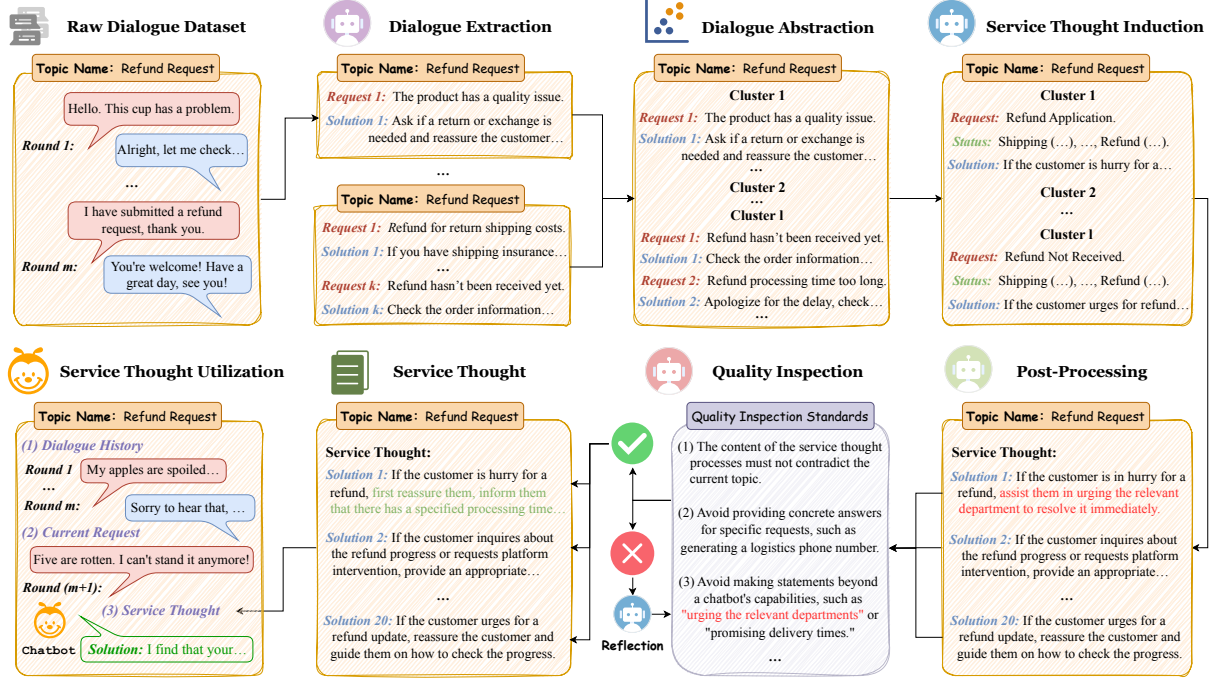


Figure 2: The pipeline of ChatMap under the topic name "Refund Request." The topic name serves as the natural label for each dialogue. We group and mine the thought processes based on different topic names.

them (§3.2-§3.6), and the utilization of derived service thoughts (§3.7). In addition, we provide the prompts for each agent of ChatMap in Appendix D.

3.1 Raw Dialogue Dataset Structure

We construct a desensitized raw multi-turn dialogue dataset collecting historical conversations between customers and human customer service agents from Taobao e-commerce platform. The platform assigns a short topic name to each group of dialogues (e.g., refund requests), which reflects the general theme of the conversation. Under the same topic name, there exist different specific issues (e.g., password changes, username changes in topic name "account issues"). The topic name, therefore, serves as an inherent label for each set of dialogues, defined as N_i . Since different types of topics differ significantly, our ChatMap method will mine and use thought processes based on the topic name N_i .

Specifically, the raw dialogue dataset \mathcal{D}_{Raw} contains n sets of historical dialogues (i.e., n topics), denoted as $\mathcal{D}_{Raw} = [T_1, T_2, \dots, T_n]$. Each topic T_i consists of a topic name N_i and a multi-turn dialogue history H_i , i.e.,

$$T_i = \{N_i, H_i\}, \quad (1)$$

where $H_i = \{(q_{i1}, a_{i1}), (q_{i2}, a_{i2}), \dots, (q_{im}, a_{im})\}$, consisting of m turns of dialogues. Each turn

includes the customer's statement q_{ij} and the customer service agent's response a_{ij} .

3.2 Dialogue Extraction

Raw dialogues typically contain substantial redundant or irrelevant content, which may affect both efficiency and accuracy when directly analyzed. Therefore, we first extract the key useful information from the raw dialogue, namely the customer's requests and the corresponding agent solutions.

For each topic T_i in the raw dialogue dataset \mathcal{D}_{Raw} , the dialogue extraction agent Agent_{Ext} extracts all pairs of customer requests R_i and their corresponding customer service solutions S_i (hereinafter referred to as RS pairs, denoted as P_i) from the multiple rounds of dialogue history H_i in chronological order. The customer request is limited to 20 words, and the customer service solution includes both the resolution and specific service statements. The resulting text pair P_i is defined as follows:

$$P_i = \text{Agent}_{Ext}(H_i) \quad (2) \\ = \{(R_{i1}, S_{i1}), (R_{i2}, S_{i2}), \dots, (R_{ik}, S_{ik})\}.$$

where k represents the number of RS pairs extracted from the dialogue history H_i for topic T_i . Due to the noisy nature of the conversations, typically, a single (R_i, S_i) pair is extracted from multiple (q_i, a_i) pairs, meaning $k \leq m$.

After the dialogue extraction for each topic T_i , the updated topic T_i^1 contains the unchanged topic name N_i and all the RS pairs extracted by Agent_{Ext} :

$$T_i^1 = \{N_i, P_i\} = \{N_i, (R_{ij}, S_{ij})_{j=1}^k\}. \quad (3)$$

Finally, output all the updated topic data to obtain the dialogue dataset after dialogue extraction, denoted as $\mathcal{D}_{Ext} = \{T_1^1, T_2^1, \dots, T_n^1\}$.

3.3 Dialogue Abstraction

We observe that RS pairs under the same dialogue topic name N_i exhibit similarities, differing only in expression or details. Therefore, we divide the original n topics into r categories based on their topic names, and then generalize the RS pairs with the same topic name into a higher-level cluster.

Topic Clustering. Firstly, using the BGE (BAAI General Embedding) model (Xiao et al., 2023), we convert the content of all k RS pairs P_i with the same topic name N_i into an embedding vector $e_i = \text{BGE}(P_i)$. Afterward, we use the OPTICS (Ordering Points To Identify the Clustering Structure)(Ankerst et al., 1999) clustering algorithm to cluster these embedding vectors. We iteratively perform clustering, dynamically adjusting parameters, and apply clustering only to data points that have not yet been assigned to any cluster. The same clustering operation is applied to all topic names until 80% of data points are assigned to a cluster, i.e., a minimum coverage of 80%. After clustering, each topic name N_i results in a set CLS_i , containing l clusters CLS_{ij} :

$$CLS_i = \text{OPTICS}(\{e_1, e_2, \dots, e_k\}) \quad (4)$$

$$= \{CLS_{i1}, CLS_{i2}, \dots, CLS_{il}\},$$

where each cluster CLS_{ij} contains multiple embedding vectors e_i with high similarity in the numerical space, indicating that their corresponding RS pairs P_i are semantically similar.

Data Reconstruction. Replace each embedding vector e_i with its original RS pairs P_i , thereby forming l clusters represented by RS pairs, represented as CLS_{ij}^{RS} . Update the topic T_i^2 , including the topic name N_i and all clusters CLS_{ij}^{RS} :

$$T_i^2 = \{N_i, CLS_i^{RS}\} = \{N_i, (CLS_{ij}^{RS})_{j=1}^l\}. \quad (5)$$

Finally, we output all the updated topics. Due to the merge same topic names operation, we obtain r distinct topic names, resulting in the dia-

logue dataset after abstraction, denoted as $\mathcal{D}_{Abs} = \{T_1^2, T_2^2, \dots, T_r^2\}$.

3.4 Service Thought Induction

After the preliminary organization of the dialogue data, we obtain a set of solutions used by customer service agents when addressing similar customer requests. We further uncover the thought processes behind these solutions by introducing the induction agent, i.e., Agent_{Ind} . Specifically, for each topic T_i^2 , the goals of Agent_{Ind} are: (1) Summarize the similar customer requests provided by each cluster CLS_{ij}^{RS} , denoted as Req_{ij} . (2) Analyze why customer service agents provides different responses to similar customer requests under the current topic name. (3) Infer the most likely order status $Stat_{ij}$ based on the topic and requests. For example, for the topic "shipment tracking not updated," the most likely order statuses are: shipping status (shipped but not delivered), tracking status (no info, picked up, in transit, out for delivery), and refund status (no refund initiated). (4) Induce the service thought Tho_{ij} for each cluster, limiting it to no more than two strategies. The specific constraints for inducing service thoughts are detailed in Appendix D.2.

Thus, for each cluster CLS_{ij}^{RS} within each topic T_i^2 , we derive the induction agent's output O_{ij} :

$$O_{ij} = \text{Agent}_{Ind}(CLS_{ij}^{RS}) \quad (6)$$

$$= \{Req_{ij}, Stat_{ij}, Tho_{ij}\}.$$

The output of the l clusters form the updated topic T_i^3 :

$$T_i^3 = \{N_i, \{O_{ij} \mid j \in [1, l]\}\}. \quad (7)$$

The r topics then form the service thoughts dataset $\mathcal{D}_{Ind} = \{T_1^3, T_2^3, \dots, T_r^3\}$.

3.5 Post-Processing

Directly inputting all of the service thoughts into the agent via prompts often exceeds the LLM-based agent's context length limit. Therefore, we prompt the post-processing agent Agent_{Pos} to compare the outputs O_{ij} extracted for each cluster CLS_{ij}^{RS} under the same topic T_i^3 , and select the most representative service thought $Tho_{ij} \in O_{ij}$, without merging or rewriting to avoid introducing additional bias. Since the customer request Req_{ij} and order status $Stat_{ij}$ contained in O_{ij} mainly serve to assist in inducing the service thoughts, we no longer include them in the set. Ultimately, around

20 service thoughts (approximately 1500 tokens) are obtained as a set Tho_i for each topic T_i^4 :

$$\begin{aligned} Tho_i &= \text{Agent}_{Pos}(\{O_{ij} \mid O_{ij} \in T_i^3, \forall j\}) \quad (8) \\ &= \{Tho_{ij} \mid j \in J \subseteq \{1, 2, \dots, l\}, |J| \approx 20\}, \\ T_i^4 &= \{N_i, Tho_i\}, \quad (9) \end{aligned}$$

which forms the post-processed dataset $\mathcal{D}_{Pos} = \{T_1^4, T_2^4, \dots, T_r^4\}$.

3.6 Quality Inspection

To identify and correct potential unreasonable or inaccurate service thoughts, we further provide several inspection criteria to the Agent_{Qua} , as detailed in Appendix D.4. If the agent deems the service thought Tho_{ij} reasonable, it will be retained. Otherwise, it will provide the reasons Re for why it is unreasonable, and based on this, instruct the induction agent Agent_{Ind} in §3.4 to reflect and regenerate a new service thought $Tho'_{ij} = \text{Agent}_{Ind}(Re, CLS_{ij}^{RS})$. Repeat the process until the quality inspection is passed, at which point Tho'_{ij} will replace the original Tho_{ij} , i.e., $Tho_i = (Tho_i \setminus \{Tho_{ij}\}) \cup \{Tho'_{ij}\}$.

From this, we can obtain the updated topic T_i^5 :

$$T_i^5 = \{N_i, Tho_i\}, \quad (10)$$

where for $\forall Tho_{ij} \in Tho_i$, Tho_{ij} passes the quality inspection by Agent_{Qua} .

Finally, we obtain a refined and quality-validated dataset $\mathcal{D}_{Fin} = \{T_1^5, T_2^5, \dots, T_r^5\}$, which includes various topics of issues along with viable service thoughts under each topic.

3.7 Service Thought Utilization

The final dataset \mathcal{D}_{Fin} contains service thoughts that encapsulate the reasoning and actions taken by human customer service agents when successfully resolving issues. We use this to guide the customer service chatbot (denoted as Chatbot) in adopting appropriate dialogue strategies.

A straightforward approach is to directly use a text prompt, providing a set of dialogue histories H_i under a certain topic name N_i , the current request R_{curr} , and the corresponding service thoughts Tho_i for that topic from \mathcal{D}_{Fin} .

The Chatbot is supposed to provide a warm, professional, friendly, and positive response to address the customer’s questions or needs. Specific requirements include: (1) Analyze the dialogue history and the customer’s request to accurately comprehend the customer’s needs. (2) Select the most

relevant service thought to help understand the customer’s possible issue and determine the correct response approach. If no suitable service thought is available, it should generate a response based on common sense as a customer service agent. (3) Finally, generate the solution S_{curr} in the response. Considering the real chat scenario, the response should be limited to no more than 30 characters.

The Chatbot’s solution S_{curr} can be obtained as follows:

$$S_{curr} = \text{Chatbot}(H_i, R_{curr}, Tho_i). \quad (11)$$

The detailed costs of the entire ChatMap process are provided in Table 6.

4 Experiment

We are aiming to investigate the following research questions: (1) Can ChatMap automatically mine thought processes to achieve human-level results and improve chatbot performance offline and online? (§4.2) (2) Is our multi-agent design effective? (§4.3) (3) Does our evaluation method, ChatEval-EC, show greater agreement with human evaluations? (§4.4.1) (4) What is the impact of different module orders? (§4.4.2) (5) How does the number of service thoughts affect performance? (§4.4.3)

4.1 Experimental Setup

4.1.1 Dataset

The dataset information is mentioned in §3.1. The training set includes 10,000 multi-turn dialogues, while the test set contains 409 dialogues across six common topic names. For each topic name, 20 service thoughts are manually annotated by human agents. For each test sample under a given topic, the most suitable service thought is selected, referred to as the *Service Thought Gold Standard*, and a response is written by human agents based on it, referred to as the Reference Answer.

4.1.2 Evaluation

• Offline Metrics.

(1) We use **ROUGE-1** (Lin, 2004) to compare the chatbot’s response with Reference Answer.

(2) Since ROUGE-1 only focuses on surface-level lexical matching and ignores semantics, inspired by ChatEval (Chan et al., 2023), we additionally design **ChatEval-EC** to specifically assess the response quality of e-commerce customer service chatbots. It employs four review roles—Safety Auditor, Linguistics Expert, Simulated Customer, and

Table 1: Offline experimental results of ChatMap and baselines. **Bold** represents the optimal under the same model, with and marking the optimal and suboptimal in open-source models.

Models	Methods	ROUGE-1	ChatEval-EC		
			Consistency	Reasonableness	Problem-Solving
<i>Closed-source Models</i>					
OpenAI o1	CoT	36.76%	63.17%	96.25%	51.82%
GPT-4o	MATP	40.01%	72.15%	98.68%	50.22%
	ChatMap	38.07%	67.49%	98.39%	52.39%
<i>Open-source Models</i>					
Qwen2-7B	Retail-GPT	33.08%	55.29%	94.73%	43.36%
Qwen2.5-72B	AutoGen	38.66%	68.37%	97.98%	52.84%
DeepSeek-R1	CoT	37.14%	70.70%	97.29%	55.36%
Qwen2-7B	SFT	28.64%	56.70%	90.63%	39.86%
	CPT	27.78%	53.75%	89.68%	40.34%
	MATP	42.18%	78.45%	98.43%	55.58%
	ChatMap	38.73%	75.14%	98.24%	57.66%
GLM-4-9B	SFT	27.78%	52.30%	90.74%	38.88%
	CPT	27.78%	53.73%	90.90%	39.13%
	MATP	36.36%	66.05%	97.97%	47.70%
	ChatMap	35.09%	68.67%	97.33%	51.40%
Qwen2.5-72B	SFT	40.46%	69.13%	97.57%	47.32%
	MATP	41.31%	76.13%	98.86%	55.06%
	ChatMap	39.53%	77.90%	98.56%	59.51%

Experienced Customer Service Agent—to evaluate three metrics from different perspectives:

1. Consistency Rate: Whether the response follows the annotated *Service Thought Gold Standard*.

2. Reasonableness Rate: Whether the response complies with *Customer Service Reply Guidelines* (Taobao e-commerce platform’s risk assessment framework detailed in Appendix A).

3. Problem-Solving Degree: Whether the response effectively resolves the customer’s inquiry.

The final comprehensive evaluation is formed via multiple rounds of discussions and debates among the agents. Detailed role definitions, scoring criteria, and an evaluation example of ChatEval-EC are provided in the Appendix B.

• **Online Metrics.** We introduce three business-related metrics used in the real-world online production environment:

1. Human Agent Transfer Rate: The proportion of customer conversations that are eventually transferred to human agents. A lower rate indicates that the chatbot can handle issues more effectively.

2. Customer Satisfaction Rate: The satisfaction score rate given by customers regarding service quality after interacting with the chatbot. A higher

rate indicates greater satisfaction.

3. Immediate Resolution Rate: The proportion of customer issues that are efficiently and successfully resolved (e.g., in refund scenarios, ensuring that a refund is processed within 24 hours, and in consultation scenarios, ensuring no repeated inquiries about the same order within 7 days).

4.1.3 Baseline Methods

For the chatbot, we use GPT-4o (OpenAI, 2024), Qwen2-7B (Yang et al., 2024a), GLM-4-9B (GLM et al., 2024), and Qwen2.5-72B (Yang et al., 2024b) as foundation models, and compare ChatMap (using GPT-4 Turbo (OpenAI, 2023) as each agent) with the following approaches: supervised fine-tuning (SFT), continued pre-training (CPT), manually annotated thought processes (MATP), outstanding chain-of-thought (CoT) reasoning of OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025), an e-commerce platform’s chatbot **Retail-GPT** (de Freitas and Lotufo, 2024), and the multi-agent framework **AutoGen** (Wu et al., 2023) that supports conversational programming. Their detailed introduction and implementation details are provided in Appendix C.

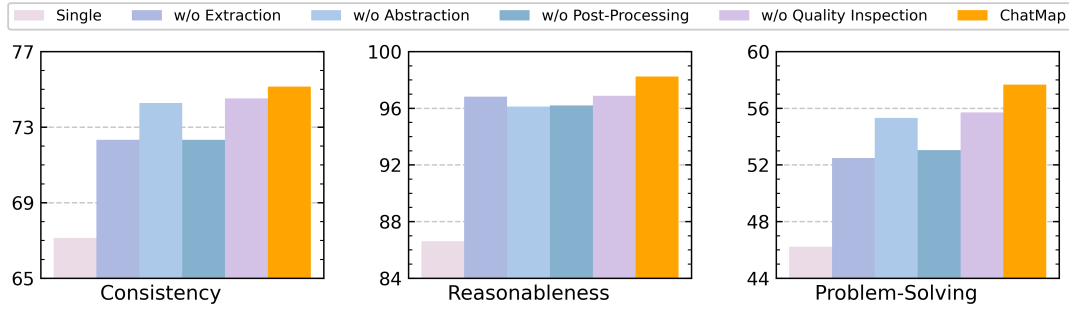


Figure 3: Ablation study of the modules of ChatMap on Qwen2-7B evaluated by ChatEval-EC, where Single refers to using only a single induction agent to directly obtain service thoughts.

Table 2: Online A/B testing results of ChatMap.

Metrics	ChatMap	w/o ChatMap
Human Agent Transfer Rate	18.99%	20.28%
Customer Satisfaction Rate	46.10%	45.32%
Immediate Resolution Rate	37.24%	37.16%

4.2 Main Results and Discussion

4.2.1 Offline Experiments

Offline experimental results comparing ChatMap with other baselines are shown in Table 1. We have the following analysis: (1) ChatMap significantly outperforms all methods that do not employ service thoughts, emphasizing the importance of learning thought processes behind dialogue strategies. (2) Since the evaluation criteria for ROUGE-1, consistency, and reasonableness are all based on human annotations, the MATP method inherently holds advantages on these metrics. Our ChatMap achieves comparable performance on these metrics while substantially surpassing MATP in problem-solving, demonstrating that ChatMap actually generates higher-quality responses. (3) Overall, ChatMap serves as a powerful alternative to manually induced service thought and proves competent for customer service agents in e-commerce platforms.

4.2.2 Online Experiments

To evaluate the performance of ChatMap in a real production environment, we deployed it on our Taobao e-commerce platform and conducted A/B tests on three online metrics from August 2024 to September 2024. Since the platform operates within a complex multi-link strategy mechanism, business metrics typically exhibit stable trends without drastic fluctuations. Experimental results in Table 2 show a significant decrease in the Human Agent Transfer Rate and an increase in customer

satisfaction, indicating that ChatMap effectively enhances the chatbot’s capabilities. Meanwhile, although the Immediate Resolution Rate shows only a slight improvement, it’s important to note that human agents generally have stronger problem-solving capabilities than chatbots (e.g., chatbots cannot handle certain functions like refunds). The fact that the Immediate Resolution Rate still increases despite the reduced Human Agent Transfer Rate further demonstrates that ChatMap provides significant assistance in resolving customer issues.

4.3 Ablation Study (RQ2)

We conduct ablation experiments on each module of ChatMap. As shown in Figure 3, (1) using a single agent to directly induce service thoughts (the LLM-based automation approach in existing works) leads to a significant performance drop, as it fails to truly grasp the thought process behind dialogue strategies. (2) When any individual agent is removed, the overall performance declines, demonstrating that each module of ChatMap plays an essential role (e.g., in practice, 10.25% of service thoughts fail quality inspection and are sent back for regeneration). (3) Besides, the limited magnitude of performance degradation highlights the strong robustness of the multi-agent approach.

4.4 Further Study (RQ3)

We also conduct a series of experiments to investigate the effectiveness of ChatEval-EC evaluation and the impact of different settings in ChatMap.

4.4.1 Effectiveness of ChatEval-EC

To compare with the multi-agent evaluation approach of ChatEval-EC, we also use single-agent models (GPT-4 Turbo (OpenAI, 2023), OpenAI o1 (OpenAI, 2024), and QwQ-32B-Preview (Qwen, 2024) as referee models) to evaluate the gener-

Table 3: Results for ChatMap with different orders of dialogue extraction and abstraction.

Model	Settings	ROUGE-1	ChatEval-EC		
			Consistency	Reasonableness	Problem-Solving
Qwen2-7B	Abstraction + Extraction	37.36%	75.91%	97.49%	56.77%
	Extraction + Abstraction (ChatMap)	38.73%	75.14%	98.24%	57.66%

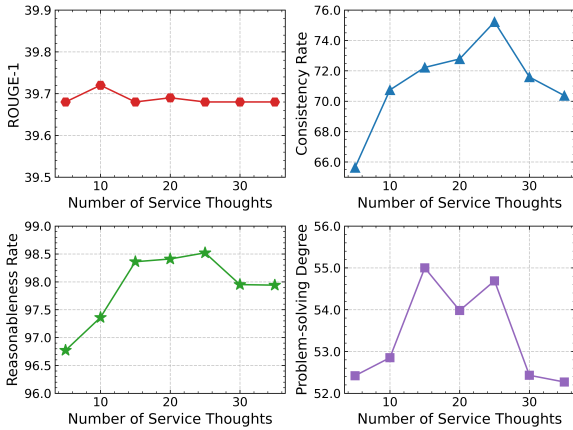


Figure 4: Impact of service thought quantity.

ated responses of Qwen2.5-72B using the MATP method. For each response, we compare the model evaluation results with the human evaluation results one by one and calculate the agreement rate. The human evaluations are conducted by 10 agents with over 3 years of experience, who label each dialogue in the test set and perform cross-sampling validation. Results in Table 4 shows that ChatEval-EC outperforms the other referee models comprehensively and shows a high agreement rate with human evaluations, demonstrating its effectiveness.

We provide a complete example of ChatEval-EC’s evaluation process in Appendix B.3.

4.4.2 Extraction or Abstraction First?

In ChatMap, we first extract customer requests and service solutions from raw dialogues before performing dialogue abstraction. Another approach is to first perform abstraction based on the overall similarity of the raw dialogues, and then extract requests and solutions. The comparison between the two approaches is shown in Table 3. We find that the approach used by ChatMap, which performs dialogue extraction first, yields better overall performance. This may be due to the presence of a large amount of irrelevant information in raw dialogues. Performing abstraction first is prone to misclassify dialogues with different intentions into the same cluster, further amplifying this bias in the

Table 4: Agreement rates of ChatEval-EC and other referee models with human evaluation on responses generated by Qwen2.5-72B using the MATP method.

Evaluation	Consistency	Reasonableness	Problem-Solving
GPT-4 Turbo	83.61%	88.02%	67.97%
OpenAI o1	87.04%	85.82%	70.42%
QwQ	88.02%	90.71%	74.08%
ChatEval-EC	88.26%	93.89%	82.15%

following processing steps.

4.4.3 Service Thoughts Quantity Selection

In the main experiment, the post-processing agent selects 20 optimal service thoughts for each dialogue topic. We further investigate the impact of varying the number of service thoughts, with the results shown in Figure 4, where the optimal range is between 15 and 25. Customer service dialogues exhibit a long-tail effect: insufficient service thoughts lead to oversimplified decisions, while excessive service thoughts cause redundancy, lower quality, and increased computational overhead. This mirrors human expert decision-making patterns—experienced agents typically store a certain number of solutions rather than mechanically memorizing all possible scenarios.

5 Conclusion

In this paper, we introduce ChatMap, a novel multi-agent approach that automatically induces customer service dialogue strategies. By progressively processing raw dialogues through extraction, abstraction, induction, post-processing, and quality inspection, it obtains service thoughts to help chatbots generate more appropriate responses. Extensive experiments validate the effectiveness of our approach in improving offline metrics and confirm the reliability of our ChatEval-EC evaluation design. By mining the thought process, ChatMap outperforms other methods, showcasing capabilities comparable to professional human agents. Additionally, online A/B experiments have substantiated improved business performance.

Limitations

We propose a novel thought process induction method to improve the chatbot’s dialogue performance and have validated its effectiveness from multiple aspects. However, due to limitations in manpower and time, we have not conducted a more thorough evaluation of the model, such as manually annotating more evaluation metrics, performing additional human evaluations, trying more models to mine service thoughts, and conducting longer online A/B testing. In the future, we will conduct more detailed testing and expansion to further enhance the model’s performance.

Ethical Considerations

This paper adheres to the ACM Code of Ethics and Professional Conduct. The dataset have been desensitized to protect privacy and will be open-sourced after approval from our security department. Additionally, our method has been deployed and tested in the real-world customer service environment of Taobao e-commerce platform, emphasizing the need to carefully assess potential risks in online deployment and responsibly utilize this technology.

Acknowledgments

This work was supported by the NSFC (62272411), the Key R&D Projects in Zhejiang Province (No. 2024C01106, 2025C01030), the Zhejiang NSF (LRG25F020001).

References

- Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.
- Rajesh Bhayana. 2024. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1):e232756.
- Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. 2021. Using language model to bootstrap human activity recognition ambient sensors based in smart homes. *Electronics*, 10(20):2498.
- B Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An intent taxonomy for questions asked in web search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 85–94.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Shangyu Chen, Zibo Zhao, Yuanyuan Zhao, and Xiang Li. 2024. Apollonion: Profile-centric dialog agent. *arXiv preprint arXiv:2404.08692*.
- François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Bruno Amaral Teixeira de Freitas and Roberto de Alencar Lotufo. 2024. Retail-gpt: leveraging retrieval augmented generation (rag) for building e-commerce chat assistants. *arXiv preprint arXiv:2408.08925*.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.
- Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint arXiv:2010.02194*.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Subin Kim, Prin Phunyahibarn, Donghyun Ahn, and Sundong Kim. 2022. Playgrounds for abstraction and reasoning. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.
- Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. 2024. Sasha: creative goal-oriented reasoning in smart homes with large language models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–38.
- Chuyi Kong, FAN Yaxin, Xiang Wan, Feng Jiang, and Benyou Wang. 2023. Platolm: Teaching llms via a socratic questioning user simulator.
- Vimal Kumar, Priyam Srivastava, Ashay Dwivedi, Ishan Budhiraja, Debjani Ghosh, Vikas Goyal, and Ruchika Arora. 2023. Large-language-models (llm)-based ai chatbots: Architecture, in-depth analysis and their performance evaluation. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 237–249. Springer.
- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.
- Yitian Li, Jidong Tian, Caoyun Fan, Wenqing Chen, Hao He, and Yaohui Jin. 2023. Mtr: A dataset fusing inductive, deductive, and defeasible reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10078–10089.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.
- OpenAI. 2023. [Gpt-4 technical report](#). Preprint, arXiv:2303.08774.
- OpenAI. 2024. [Hello GPT-4o](#). Accessed: 2024-05-13.
- OpenAI. 2024. [Openai o1 system card](#). Accessed: 2024-09-12.
- Jiao Ou, Jiayu Wu, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Inductive-deductive strategy reuse for multi-turn instructional dialogues. *arXiv preprint arXiv:2404.11095*.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and Maria Del Carmen Lopez-Perez. 2023. Ethical challenges in the development of virtual assistants powered by large language models. *Electronics*, 12(14):3170.
- Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*.
- Qwen. 2024. [Qwq: Reflect deeply on the boundaries of the unknown](#).
- Jesús Sánchez Cuadrado, Sara Pérez-Soler, Esther Guerra, and Juan De Lara. 2024. Automating the development of task-oriented llm-based chatbots. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–10.
- Sonal Sannigrahi, Thiago Fraga-Silva, Youssef Oualil, and Christophe Van Gysel. 2024. Synthetic query generation using large language models for virtual assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2837–2841.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*.
- Guangzhi Sun, Xiao Zhan, and Jose Such. 2024a. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–6.
- Wangtao Sun, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, and Kang Liu. 2024b. Itd: Large language models can teach themselves induction through deduction. *arXiv preprint arXiv:2403.05789*.
- Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryan W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5836–5847.
- Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoor-molabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.
- Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.
- Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023a. k nn prompting: Beyond-context learning with calibration-free nearest neighbor inference. *arXiv preprint arXiv:2303.13824*.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
- Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. 2023c. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.
- Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. *arXiv preprint arXiv:2004.08056*.
- Yankai Zeng, Abhiramon Rajasekharan, Parth Padalkar, Kinjal Basu, Joaquín Arias, and Gopal Gupta. 2024. Automated interactive domain-specific conversational agents that understand human dialogs. In *International Symposium on Practical Aspects of Declarative Languages*, pages 204–222. Springer.
- An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1807–1817.
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327.
- Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. 2021. Acre: Abstract causal reasoning beyond covariation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10643–10653.
- Shu Zhong, Elia Gatti, James Hardwick, Miriam Ribul, Youngjun Cho, and Marianna Obrist. 2024. Llm-mediated domain-specific voice agents: The case of textilebot. *arXiv preprint arXiv:2406.10590*.
- Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*.
- Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2024. Large language models can learn rules. *Preprint, arXiv:2310.07064*.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Specific Criteria of Customer Service Reply Guidelines

Customer Service Reply Guidelines is a risk assessment framework used by our e-commerce platform to evaluate the reasonableness of generated responses. It includes 7 criteria, which are specifically described in Table 5.

B ChatEval-EC Evaluation

Based on ChatEval (Chan et al., 2023), an LLM-based multi-agent referee team that autonomously discusses and evaluates the quality of generated responses, we propose ChatEval-EC, which is specifically designed to evaluate the quality of responses generated by chatbots on e-commerce platforms. Using GPT-4o (OpenAI, 2024) as our LLM, we define four roles: Safety Auditor, Linguistics Expert, Simulated Customer, and Experienced Customer Service Agent. The evaluation results for consistency rate, reasonableness rate, and problem-solving degree are obtained through discussions among these roles. Below are the prompts defined for each role, the scoring criteria, and a specific evaluation example.

B.1 Role Definition

- **Safety Auditor:** As a safety auditor for an e-commerce platform, your primary responsibility is to ensure that the dialogue between customer service and customers is compliant and reasonable. You are instructed that customer service responses must never exhibit the following issues:

- (1) Irrelevant Answer: The customer service response is completely unrelated to the customer's question.

- (2) Duplicate Information Collection: The customer service reply repeats information that the customer has already provided in previous messages.

- (3) No Solution: When the customer's problem or request is clearly stated, the customer service reply contains no answer or includes phrases like "still learning," "unable to solve your issue at this time," etc.

- (4) High-Risk Response: The content of the customer service reply may provoke external public opinion or include material that violates laws, regulations, or platform policies.

- (5) Inconsistent with Business Facts: The customer service reply is inconsistent with business processes or platform service policies.

- (6) Contradictory Statements: The customer service reply contains inconsistent information within the same conversation.

- (7) Incoherent or Incomplete Sentences: The response contains typos, missing words, incomplete expressions, incorrect punctuation, or uses internal jargon, resulting in incoherent sentences.

- (8) False Promises: Making promises such as agreeing to a refund or committing to compensation without sufficient supporting information.

- **Linguistics Expert:** As a linguistics expert, you focus on the structure, pragmatics, and contextual appropriateness of language. In this task, you will analyze whether the customer service language is precise, polite, and adaptable to different cultural and social contexts. Check for ambiguity in the statements, and assess whether the language respects social etiquette and is sensitive to the context. Provide your evaluation based on linguistic accuracy and social politeness in the customer service responses.

- **Simulated Customer:** As a typical simulated customer, you play the role of a demanding consumer with high expectations for service. You critically evaluate whether the customer service responses fully address the issues raised by the customer, not only considering the timeliness and relevance of the response but also assessing the effectiveness and feasibility of the solution. From the customer's perspective, you will assess whether the response practically meets the customer's needs, ensuring that each reply contains substantive content that resolves the issue. Evaluate the customer service performance from the angles of customer satisfaction and problem-solving capability.

- **Experienced Customer Service Agent:** With many years of customer service experience, you have a thorough understanding of the customer service processes in the e-commerce industry. You place particular emphasis on reviewing the correctness of the customer service response's reasoning and strictly judge whether the actual reply aligns with the service thought. Even minor deviations are unacceptable. Ensure that your evaluation reflects this by providing a detailed assessment of the customer service response from the perspectives of business operations and professional knowledge.

B.2 Scoring Criteria

- 1. Consistency.** Consistency evaluates whether the [current agent solution] follows the [service thought]. If the [service thought] is "no thought,"

Table 5: Specific Criteria of *Customer Service Reply Guidelines*

Criteria	Description
Irrelevant Response	The customer service response is entirely unrelated to the customer's request.
Repeated Information	Asking for information that the customer has already provided earlier in the conversation.
No Solution Provided	Despite a clear customer request, the response lacks a specific solution—e.g., stating "Still learning, unable to resolve your issue at the moment."
Inaccurate Business Information	The response contradicts established business processes or platform service policies.
High-Risk Response	The response could trigger negative public reactions or contain elements that may violate laws, regulations, or platform policies.
Contradictory Statements	The customer service agent provides conflicting information within the same conversation.
Incoherent Language	The response contains spelling or grammatical errors, incomplete expressions, improper punctuation, or internal jargon that is unclear to customers.

this criterion is not considered, and a score of 1 is given directly:

- **0 points (Inconsistent):** The response does not align with the [service thought] or is completely unrelated.

- **100 points (Consistent):** The response is generally consistent with the [service thought], with possible minor differences, as long as there are no contradictions or conflicts. If the [service thought] is "no thought," the score is also 100 points.

2. Reasonableness. Reasonableness evaluates whether the [current agent solution] complies with platform's *Customer Service Reply Guidelines* (see Appendix A), with the following principles:

- **0 points (Unreasonable):** The response contains risky content or other obvious unreasonable elements.

- **100 points (Reasonable):** The response does not contain any issues, and the customer service agent responded to the consumer's question with enthusiasm, professionalism, friendliness, and a positive tone.

3. Problem-solving. Problem-solving evaluates whether the [current agent solution] resolves the customer's issue:

- **0 points (Poor):** The issue is completely unresolved, and the customer service agent did not provide an acceptable solution or explanation. This includes situations where:

- (1) The customer service response failed to provide an effective solution, guidance, or reasonable explanation, leaving the issue unresolved.

- (2) Even though the customer service agent re-

sponded, the content was repetitive, off-topic, or did not specifically address the customer's issue, question, or need.

- (3) The customer's issue or request was ignored, and the customer service response did not show an understanding of the problem or an attempt to resolve it.

- **50 points (Average):** The issue is not fully resolved, but the customer service agent provided a solution or reasonable explanation, and after consideration, it seems the solution is unlikely to fully resolve the customer's problem. This includes situations where:

- (1) The customer service agent provided a solution or reasonable explanation, but the steps involved may not be clear enough, requiring the customer to explore on their own or inquire further to fully understand it.

- (2) The customer service response attempted to address the issue, but the solution was unclear or could not directly meet the customer's specific needs, requiring additional information or further communication.

- **100 points (Excellent):** The issue is fully resolved, including situations where:

- (1) The problem is clearly resolved, and the customer service agent provided a direct solution, clear next steps, or an explanation that satisfactorily answered the customer's question.

- (2) If the customer did not raise a specific issue but engaged in a simple interaction (e.g., "Okay"), and the customer service agent responded with a polite reply, maintaining a good customer relation-

ship.

(3) The solution or information provided by the customer service agent may require the customer to take further actions, but it provides substantial help for the customer's issue, and the customer service agent explained detailed follow-up steps or feasible alternatives.

B.3 ChatEval-EC Evaluation Example

- **[Topic Name]: Urging Refund**

- **[Dialogue History]:**

- Agent: Please confirm the refund details
 - After-sales type: Refund only
 - Product status: Received
 - Refund reason: Return shipping fee
 - Refund amount: ¥20.00 (including shipping)
 - <Buttons: Modify Application>
 - <Buttons: Submit Application>
- Agent: Refund has been initiated.
 - We are sorry for the poor shopping experience. Xiaomi will handle this for you.
- Agent: Dear, we noticed you applied for a "refund only" due to "return shipping fee". The seller needs to process your refund within 1 day, 11 hours, and 59 minutes. Please be patient, but Xiaomi can also urge the seller to process it quickly.
 - <Buttons: Urge Refund>
- Customer: <Buttons: Urge Refund>
- Agent: After-sales upgrade notice
 - 19.9 yuan brand discount, store closure, no refund or exchange for non-quality issues
 - Color: 59.9; Size: One size
 - Price: ¥49.90 × 2
 - Refund in progress, waiting for seller's processing
 - The platform customer service has urged the seller. If the seller does not process the refund by 18:54:05 on August 29, the platform will approve your after-sales request. Thank you for your patience.
- **[Service Thought Gold Standard]:** "e. If the user confirms or reports that the issue has been resolved, use a thank-you and reassurance tone, directly generate a response: {"Thought": "e", "action": "reply", "response": {response_text}}
- **[Current Customer Request]:** Hmm.
- **[Current Chatbot Reply]:** Okay, thank you for your understanding. If you have any other questions, feel free to contact us anytime.

Consistency Label: 1,

Reasonableness Label: 1,
Problem-solving Label: 2,

- **[Debate History]**

- **Safety Auditor:**

Consistency: The customer service reply is generally consistent with the service thought process. The user said "Hmm" without clearly stating a problem, and the service reply was polite and expressed thanks, which aligns with the service thought process of using gratitude and reassurance. Therefore, in terms of consistency, it meets the standard.

Reasonableness: The reply from customer service is reasonable, professional, and friendly, without any violations. Therefore, in terms of reasonableness, it also meets the standard.

Problem-solving: The user did not raise a specific issue, but instead engaged in a brief interaction. The customer service reply maintained good customer relations and did not ignore the user's needs. Thus, the problem-solving aspect is also excellent.

Overall Evaluation: Consistency: 100 Reasonableness: 100 Problem-solving: 100 Score: 100

- **Linguistics Expert:**

Consistency: Based on the information provided, the customer service reply was based on the user's confirmation of the refund progress. The user's brief response "Hmm" indicates some level of confirmation or acceptance. The customer service response used a thankful reassuring tone, in line with the service thought process. Therefore, I believe the reply is correct, and it deserves 100 points for consistency.

Reasonableness: The reply from customer service is polite, expresses gratitude for the user's understanding, and invites the user to contact for further issues. This follows the customer service rules of the platform and general customer service etiquette. Therefore, I think the response is reasonable and should also receive 100 points.

Problem-solving: The user did not raise a specific issue, just a brief response. The customer service reply was appropriate, maintaining good customer relations, and no further issues needed solving. In this case, the response is perfect, and the problem-solving degree can be rated 100 points.

Overall Evaluation: Consistency: 100 Reasonableness: 100 Problem-solving: 100 Score: 100

- **Simulated Customer:**

Consistency: The customer service reply followed the service thought process "e", using a thank-you and reassuring tone, consistent with the service thought. (100 points)

Reasonableness: The reply from customer service complies with the platform's rules, friendly and professional in tone. (100 points)

Problem-solving: The user only replied with a brief "Hmm", which can be understood as confirmation or acknowledgment of the previous information. The customer service reply expressed gratitude and maintained a good attitude, keeping the customer relationship intact with no further problem to solve. (100 points)

Overall Evaluation: Consistency: 100 Reasonableness: 100 Problem-solving: 100 Score: 100

—
- Experienced Customer Service Agent:

Consistency (0 points): The actual reply did not fully follow the instructions in the service thought process. The service thought process requires using gratitude and reassurance, but the actual reply did not clearly express gratitude or reassure the user.

Reasonableness (100 points): The reply is reasonable and compliant, without any risk content, and the tone remained polite, conveying that the user is always welcome to contact us.

Problem-solving (50 points): The user only replied with "Hmm", which can be interpreted as acknowledgment or feedback that the issue has been resolved. The customer service reply, although polite, did not further reassure the user or confirm if they had other needs. In this case, more attention to the user and confirming if there are any other issues would have better solved the user's problem.

Final Score: Consistency: 0 Reasonableness: 100 Problem-solving: 50

—
- Safety Auditor: Thank you all for your comments. The opinions from the Linguistics Expert and Simulated Customer suggest that the customer service reply fully meets the standards, while the Experienced Customer Service Agent believes that the reply did not fully follow the service thought process, hence the lower scores for consistency and problem-solving.

Consistency: The service thought process clearly indicates that gratitude and reassurance should be used. The actual reply includes the phrase "Thank you for your understanding," which is a form of reassurance and polite response. Al-

though there wasn't more specific reassurance, it generally meets the requirements. Therefore, I believe consistency can receive a high score, though it could be further improved with more explicit reassurance.

Reasonableness: The reply contains no violations, expresses gratitude, and invites the user to contact us anytime, which is very polite and professional. - This aspect should receive a perfect score.

Problem-solving: - The user merely replied with "Hmm," which seems to indicate the issue is resolved. The customer service reply, though polite, did not further reassure the user or confirm if they had any other needs. The reply maintained good customer relations. - Therefore, in terms of problem-solving, this should also receive a high score.

Overall Evaluation: - Consistency: Although the reassurance could have been more explicit, it generally aligns with the service thought process. - Reasonableness: Fully compliant with the rules. - Problem-solving: Reasonable response.

Scores: - Consistency: 90 - Reasonableness: 100 - Problem-solving: 100 - Name: Experienced Scorer

—
- Linguistics Expert: After hearing the Experienced Customer Service Agent's judgment on service thought, I believe I need to reconsider my evaluation.

Consistency: Based on my analysis, the reply is a standard polite response, but it does not explicitly express gratitude. This differs from the service thought process, which requires the use of gratitude and reassurance. Therefore, the consistency score cannot be perfect. Score: 70

Reasonableness: The reply is very polite and professional, fully compliant with the platform's standards, and there is no violation. It also expresses friendly encouragement for the customer to contact us anytime. This aspect is flawless. Score: 100

Problem-solving: The customer's brief response "Hmm" usually indicates acknowledgment or acceptance. The reply is professional and polite, but it did not further confirm if the user had any additional issues. While no further issues were raised by the user, it could have been improved by ensuring the user felt heard. Score: 80

Summary Scores: Consistency: 70 Reasonableness: 100 Problem-solving: 80

Final Evaluation: I believe the customer service performance is generally good, but there is room for improvement, especially in making the gratitude and reassurance more explicit. Thus, the final evaluation is as follows: Score: 83

- **Simulated Customer:** Thank you all for your comments. After considering the three aspects, we can analyze them as follows:

Consistency: From the service thought process, the customer service should use gratitude and reassurance. In fact, the customer service did use expressions of gratitude, such as "Thank you for your understanding," and expressed that the user could contact us anytime, which essentially includes the intent of gratitude and reassurance. Therefore, I believe the consistency is correct and deserves a high score.

Reasonableness: The customer service reply is friendly, polite, and in compliance with the platform's rules. There is no violation, and it fully meets the standards. This aspect is clearly flawless.

Problem-solving: The user replied with "Hmm," indicating acknowledgment or that the problem is solved. From the reply, the customer service did not explicitly state further follow-up, but in such cases, this is generally unnecessary. Thus, I believe the customer service handling was appropriate and did not neglect the user's needs.

Summary Scores: Consistency: 100 Reasonableness: 100 Problem-solving: 100

- **Experienced Customer Service Agent:** After listening to the feedback from the Safety Auditor, Linguistics Expert, and Simulated Customer, I believe my previous judgment on service thought was not entirely accurate. Based on my analysis and the reviewers' discussions, and in consideration of e-commerce customer service standards, I will adjust the scoring accordingly.

Consistency: - The actual reply generally aligns with the service thought process, with some level of reassurance. Although gratitude and reassurance were not explicitly stated, it broadly aligns with the intent to use "gratitude and reassurance." Therefore, I would give a relatively high score for consistency. - Score: 90

Reasonableness: - The reply is reasonable and compliant, without any risk content. The customer service maintained a friendly, professional tone and invited the user to contact us anytime. - Score: 100

Problem-solving: - The user's brief response

"Hmm" can be seen as acknowledgment or confirmation. The customer service reply maintained good interaction, but using more explicit gratitude and reassurance could have better emphasized the user's importance. Overall, it addressed the user's issue well, but could have done more. - Score: 80

Summary Scores: Consistency: 90 Reasonableness: 100 Problem-solving: 80

Final Evaluation: Consistency: 90 Reasonableness: 100 Problem-solving: 80

• **[Debate History Scoring]:**

Round	Consistency	Reasonableness	Problem-solving
1	100	100	100
2	100	100	100
3	100	100	100
4	0	100	50
5	90	100	100
6	70	100	80
7	100	100	100
8	90	100	80

• **Final Score:** Consistency: 87.5 Reasonableness: 100.0 Problem-solving: 90.0

C Introduction and Implementation Details of Baseline Methods

(1) **Supervised Fine-Tuning (SFT):** Directly fine-tune the model using 10,000 dialogue samples from the training set. We train for 3 epochs with a learning rate set to $2e-5$, using 4 NVIDIA A800-80G GPUs for about 3 hours.

(2) **Continued Pre-training (CPT):** We additionally select 100,000 multi-turn dialogues for continued pre-training, train for 1 epoch with a learning rate of $2e-5$, using 16 NVIDIA A800-80G GPUs for approximately 5 hours. Then, we perform supervised fine-tuning (SFT) with 10,000 dialogue samples from the training set, using the same training method as described in (1).

(3) **Manually Annotated Thought Process (MATP):** To validate the effectiveness of automatic mining, we have professional customer service agents with over 3 years of experience manually annotate the service thoughts for each topic name, i.e., directly deriving service thoughts from the raw dialogue dataset for comparison purposes.

(4) **Chain-of-Thought (CoT):** Due to the powerful CoT reasoning capabilities of OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025), we use them as the backbone, replacing the mining process in ChatMap. By providing the basic principles and retrieving three relevant historical conversations as

few-shot examples, the model generates responses. The prompt is provided in Appendix E.

(5) **Retail-GPT** is an original open-source RAG-based chatbot that offers a highly adaptable approach for implementing sales chatbots.

(6) **AutoGen**, as a general framework, allows developers to build multiple agents to complete tasks and can use computer code to achieve flexible conversation patterns. It is used as a baseline to compare AutoGen’s conversational programming method with our ChatMap method, which directly uses text prompts.

D Prompts Used in ChatMap

Here is the prompt for all agents in ChatMap. Since the dialogue history in the examples come from real dialogues on our e-commerce platform, or contains content that is strongly related to the business, we have partially omitted certain parts.

D.1 Dialogue Extraction

- **Profile:**

Role: E-commerce Customer Service Agent

Language: Chinese

Description: With years of experience in e-commerce customer service, proficient in the execution and communication of various service rules on e-commerce platforms, and capable of quickly understanding customer intentions and solutions provided by customer service.

- **Abilities:**

You have empathy and excel at responding to consumer inquiries in a warm, professional, friendly, and positive tone.

You are proficient in both customer and customer service psychology, able to accurately understand customer intentions and customer service solutions based on the conversation.

You excel at summarizing, capable of accurately articulating the customer’s current issue and the solution offered by customer service.

- **Goals:**

(1) First, understand the provided **Topic Name** and the **Dialogue History** between the customer and customer service agents regarding the solution.

(2) Second, briefly summarize in chronological order: the customer’s requests and the solutions provided by the customer service agent.

(3) Third, output the customer request and agent solution in the required format, keeping the customer service wording intact.

- **Input Format:**

[**Topic Name**]: <Topic name provided to the customer>

[**Dialogue History**]: <Complete dialogue history between the customer and customer service agent>

- **Output Format:**

[**Request 1**]: <Customer request: Summarize the core request, do not copy verbatim, 20 words>

[**Solution 1**]: <Agent solution: 20 words> ###
<Exact wording from the customer service>

[**Request 2**]: <Customer request: Summarize the core request, do not copy verbatim, 20 words>

[**Solution 2**]: <Agent solution: 20 words> ###
<Exact wording from the customer service>

...

- **Examples:**

===

Input 1:

[**Topic Name**]: Received a defective product, what to do?

[**Dialogue History**]: (Omitted)

Output 1:

[**Request 1**]: Product has quality issues, requesting an exchange.

[**Solution 1**]: Contacting the seller for exchange and calming customer down ### Dear, we will try to contact the seller to negotiate for you, is that okay with you?

===

Input 2:

[**Topic Name**]: Problem with purchased product.

[**Dialogue History**]: (Omitted)

Output 2:

[**Request 1**]: customer doubts the solution, believes coupon should be refunded.

[**Solution 1**]: Customer service agent suggests contacting seller or requesting after-sales service ### Dear, we suggest you contact the seller directly to negotiate a solution. If you cannot reach an agreement or cannot contact the seller after attempts, we suggest you... (Omitted).

D.2 Service Thought Induction

You are an expert in e-commerce customer service. Please help me summarize the service thought based on customer requests and customer service responses.

I will provide you with a topic for an e-commerce conversation, several similar customer requests,

Table 6: Detailed costs of the entire ChatMap process.

Steps	Calls	Explanation	Cost
Extraction	10,000	Extract once per dialogue, call count is approximately equal to the training set size.	Input: $10,000 \times 1,000 = 10^8$ tokens, Output: $10,000 \times 100 = 10^7$ tokens
Abstraction	0	Second-level algorithm, 6 topics clustered in less than 10s.	-
Induction	500	Perform induction on 500 clusters.	Input: $500 \times 2,000 = 10^6$ tokens, Output: $500 \times 100 = 50,000$ tokens
Post-processing	6	Perform operations on each topic.	Input: 50,000 tokens, Output: $6 \times 20 \times 100 = 12,000$ tokens
Quality Inspection	130-140	6 topics with a total of 120 service thoughts, requiring 120 calls. About 10% do not pass quality inspection, requiring one additional Reflection.	Input: $140 \times 500 = 70,000$ tokens, Output: $140 \times 100 = 14,000$ tokens

and the corresponding customer service agent responses. Please help me analyze the basis for the customer service responses, that is, what necessary state information needs to be collected. Then, based on that information, summarize the service strategy.

- **Goals:**

Step 1: Summarize the similar customer requests provided.

Step 2: Carefully analyze the reasons why customer service responds differently when faced with similar customer requests under the current topic.

Step 3: Infer the most likely order status from the current topic and customer requests, including shipping status, refund status, etc. For example, in the case of "What should I do if the seller shipped the product but there is no logistics update?", the most likely order statuses are: shipping status (shipped but not yet delivered), logistics status (no logistics information, being picked up, in transit, out for delivery), refund status (refund not initiated).

Step 4: Summarize no more than two service thoughts. If there are two, use the "If, then" summary format. After each service thought, follow with ### and provide an example of customer service wording.

Step 5: The specific requirements for the service thought are:

(1) Example wording should not include partial solutions for specific requests, such as a logistics phone number or specific compensation amounts.

(2) Example wording should not include statements that go beyond the capabilities of the chatbot, such as "urge the relevant department," "promise delivery time or refund processing time," "ask the customer to provide screenshots," or "report the issue to the technical department."

(3) Example wording should avoid using promises like "shipment within 48 hours" and avoid references to links like "click here" or "click [View Details]." It should also avoid incomplete process descriptions like "the procedure is..."

(4) Service thoughts should primarily focus on explanatory content, such as "inform the customer based on common sense," "explain based on the content of the solution," "guide the customer to click the xxx button mentioned above," and should not actively provide promise-based solutions like "expected to arrive within xx days" or "the seller has xx days to process."

- **Input Format:**

[Current Topic]: <The topic of the current conversation>

[Customer Requests and Agent Solutions]: <Customer request - Agent solution pairs>

- **Output Format:**

[Request Summary]: <Summary of multiple similar customer requests>

[Information to Collect]: <Necessary state information to be collected before responding>

[Service Thought]: <Summarized service thought>

• **Example:**

===

Input:

[Current Topic]: What should I do if the seller shipped the product but there is no logistics update?

[Customer Requests and Agent Solutions]:

- **Customer Request:** The logistics information has not been updated for a long time. **Agent solution:** Check logistics status and reassure the customer ### (Omitted)

- **Customer Request:** Ask for the reason why the logistics has not been updated. **Agent solution:** Confirm package abnormality and apologize ### (Omitted)

- **Customer Request:** Logistics information has not been updated, worried about the package's status. **Agent solution:** Confirm the package status and provide the courier's contact information ### (Omitted)

Output:

[Request Summary]: Ask about the reason why logistics has not been updated.

[Customer's Order Status]: Shipped but not delivered, logistics status (no logistics information, in transit, being picked up, out for delivery), refund status (refund not initiated).

[Service Thought]:

- If the customer is concerned about the logistics information not being updated for a long time, reassure the customer and provide an estimated delivery time. ###Dear customer, your package is currently in transit. You can check the logistics details and estimated delivery time on the logistics page. Thank you for your patience

- If the customer further inquires about the reason for the logistics update delay, confirm whether there is an issue with the package and check its status while offering compensation options. ###We understand your concern. Sometimes logistics information updates may be delayed. We have notified the seller and the courier to expedite the process. If you have further questions, you may contact the courier directly for the latest status.

===

Now please summarize this case:

[Current Topic]:

[Customer Requests and Agent Solutions]:

D.3 Post-Processing

You are an expert in e-commerce customer service. Please help me analyze similar customer requests under the current topic based on customer requests

and customer service responses. Cluster the service strategies under the current topic, and select one representative service thought from each cluster. The result should consist of approximately 20 service thoughts. You only need to select and return the original service thoughts without rewriting them.

[Current Topic]: Urging for Refund

[Current Service Thoughts]:

(1) If the customer is dissatisfied with the refund speed, first calm the customer down and explain the refund process, informing them that the seller has a designated processing time and that the system will automatically process it if it exceeds the time limit.

(2) If the customer asks about the refund process or requests platform intervention, provide the relevant explanation, letting the customer understand the refund countdown and the platform's automatic processing rules after the time limit.

(3) If the customer asks about the refund progress or is dissatisfied with the refund progress, reassure the customer and guide them on how to check the refund progress.

...

(20) If the customer is worried about the refund progress and wants the platform to expedite or handle it directly, reassure the customer and inform them that if the seller does not process the refund within the designated time, the platform will automatically execute the refund according to the refund policy.

D.4 Quality Inspection

You are an expert in e-commerce customer service. Please help me evaluate whether the service strategy provided is suitable for chatbot use.

I will provide you with the **[Topic Name]** and a **[Service Thought]**. You need to output the answer in the format:

"[Evaluation Result]: Reasonable/Unreasonable
Reason"

• **Evaluation Criteria:**

- Example wording should not include partial solutions for specific requests, such as a logistics phone number or specific compensation amounts.

- Example wording should not include statements that go beyond the capabilities of the chatbot, such as "urge the relevant department," "promise delivery time or refund processing time," "ask the customer to provide screenshots," or "report the issue to the technical department."

- Example wording should avoid using promises like “shipment within 48 hours” and avoid references to links like “click here” or “click [View Details].” It should also avoid incomplete process descriptions like “the procedure is...”

- Service thoughts should primarily focus on explanatory content, such as “inform the customer based on common sense,” “explain based on the content of the solution,” “guide the customer to click the xxx button mentioned above,” and should not actively provide promise-based solutions like “expected to arrive within xx days” or “the seller has xx days to process.”

- The service thought must not contradict the [Current Topic]. For example, if the [Current Topic] is "..."

• Quality Inspection Examples:

====

Example 1:

[Topic Name]: (Omitted)

[Service Thought]: If the customer inquires about the shipping time and status, and the order shows that the seller may not have shipped on time, explain the compensation result and remind the customer that the platform has protection policies. ### Dear customer, we understand your concern. We have confirmed that the seller failed to ship within the promised time. Based on platform rules, you will receive a compensation of 3.50 CNY, which will be refunded to your account as cash... (Omitted)

[Evaluation Result]: Unreasonable ### The service thought mentions a specific compensation amount of “3.50 CNY,” which is not allowed.

====

Example 2:

[Topic Name]: (Omitted)

[Service Thought]: If the customer is not satisfied with the compensation amount or the compensation has not been processed in time, urge the relevant department to expedite the customer’s complaint based on the platform’s compensation policy. ### Dear customer, we understand your concern and dissatisfaction. Regarding the delayed shipment issue, I will immediately urge the relevant department to expedite the handling of your complaint... (Omitted)

[Evaluation Result]: Unreasonable ### The service thought mentions "urge the relevant department," which is beyond the chatbot’s capabilities.

Example 3: (Omitted)

Example 4: (Omitted)

Example 5: (Omitted)

====

• Task Execution:

[Topic Name]:

[Service Thought]:

D.5 Service Thought Utilization

Please play the role of e-commerce platform customer service agent Xiaomi. Based on the dialogue history and the provided service thoughts, respond to the customer’s inquiry according to the relevant service thought.

• Service Thought:

- (Omitted)

...

• Task Goals:

Provide a warm, professional, friendly, and positive response aimed at resolving the customer’s questions or needs.

• Task Steps:

(1) Understand and analyze the customer’s request and dialogue history: Based on the customer’s current request and dialogue history, accurately understand the customer’s needs.

(2) Thought Selection: Choose the most relevant service thought to help understand the customer’s potential issues and the correct response strategy. If no suitable thought is available, generate a response based on common customer service knowledge.

(3) Generate a response with no more than 30 words.

• Response Format:

Only provide the response content generated from the steps above, excluding the customer’s question, dialogue history, or retrieved additional dialogue history.

[Dialogue History]: (Omitted)

[Current Customer Request]: (Omitted)

E The prompt for the CoT Method

• Identity

You are an intelligent customer service chatbot named Xiaomi. Please provide a reasonable response based on the conversation history.

• Input Information:

Dialogue History: The previous interaction content in this conversation.

Current Customer Request: The most recent question from the user in this interaction.

Relevant Retrieved Content: Real historical conversation records related to the current query, based

on the conversation history and the user's current question.

• **Task Objective:**

Provide a warm, professional, friendly, and positive reply aimed at addressing the user's inquiry or need.

• **Task Steps:**

(1) Understand and Analyze: Accurately understand the user's need based on their current question and conversation history.

(2) Retrieve and Compare: Refer to relevant retrieved content, which is real historical conversations with similar contexts, to help understand the user's possible problem and the solutions that were successful before.

(3) Generate Response: Provide a response that is no more than 30 words.

• **Response Format:**

Only generate the content of the reply based on the steps above, excluding the user's question, conversation history, or other retrieved conversation history.

• **Current Input**

[Dialogue History]:

[Current Customer Request]:

[Relevant Retrieved Content]:

Relevant Historical Conversation 1

Relevant Historical Conversation 2

Relevant Historical Conversation 3