Minority-Aware Satisfaction Estimation in Dialogue Systems via Preference-Adaptive Reinforcement Learning

Yahui Fu, Zi Haur Pang, Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
{fu,pang,kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

User satisfaction in dialogue systems is inherently subjective. When the same response strategy is applied across users, minority users may assign different satisfaction ratings than majority users due to variations in individual intents and preferences. However, existing alignment methods typically train one-size-fits-all models that aim for broad consensus, often overlooking minority perspectives and user-specific adaptation. We propose a unified framework that models both individual- and group-level preferences for user satisfaction estimation. First, we introduce Chain-of-Personalized-Reasoning (**CoPeR**) to capture individual preferences through interpretable reasoning chains. Second, we propose an expectation–maximization–based Majority-Minority Preference-Aware Clustering (**M^{2}PC**) algorithm that discovers distinct user groups in an unsupervised manner to learn group-level preferences. Finally, we integrate these components into a preference-adaptive reinforcement learning framework (**PAda-PPO**) that jointly optimizes alignment with both individual and group preferences. Experiments on the Emotional Support Conversation dataset demonstrate consistent improvements in user satisfaction estimation, particularly for under-represented user groups.[1]

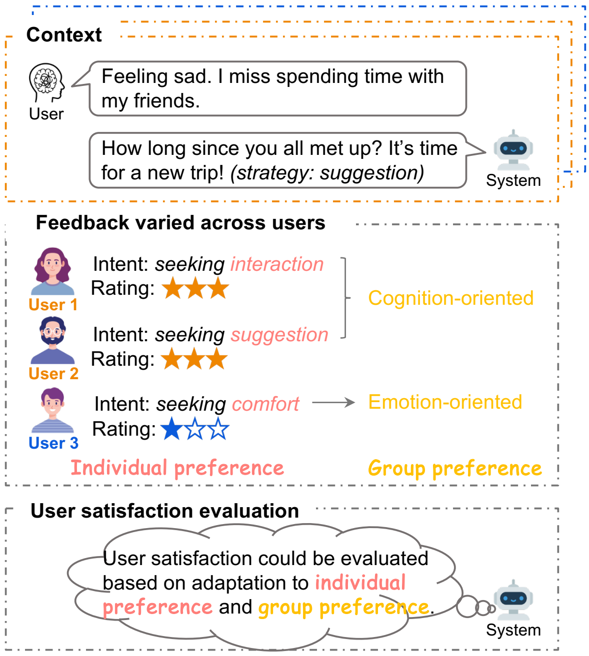

Figure 1: Majority and minority users may assign different satisfaction ratings to system responses employing the same strategy due to varying individual intents and preferences. Additionally, users within the same group may exhibit similar preference patterns (e.g., cognition-oriented versus emotion-oriented strategies). This suggests that modeling both individual-specific and group-level preferences could be an effective approach for evaluating user feedback in dialogue systems.

1 Introduction

Personalized dialogue systems that adapt to individual user preferences are crucial for enhancing user satisfaction in human–AI interactions. Accordingly, accurately evaluating whether a dialogue system meets diverse user needs is essential. Previous research has primarily evaluated dialogue systems using output-centric criteria such as informativeness, relevance, and empathy (Zhang et al., 2020; Fu et al., 2023b; Xu and Jiang, 2024), as well as user-centric metrics such as satisfaction estimation (Choi et al., 2019; See and Manning, 2021; Lin et al., 2024). However, satisfaction is inherently subjective: as illustrated in Figure 1, users may assign divergent satisfaction ratings to the system response employing the same strategy, depending on their individual intents and preferences.

Reinforcement learning (RL) with a reward model has become a key approach for aligning language models with human preferences. However, existing reward models typically rely on aggregated human judgments (Touvron et al., 2023)

[1]Our source code is publicly available at: https://github.com/fuyahuii/minority-aware-se.

(e.g., majority voting or averaging), resulting in universal reward functions that overlook minority preferences and lack personalization. This induces preference collapse, where outputs maximize majority preferences while suppressing minority views (Xiao et al., 2024; Yang, 2024; Slocum et al., 2025). Recent methods tried to solve it by considering the diversity of human preferences into reward modeling (Wang et al., 2023; Jang et al., 2023; Chakraborty et al., 2024), but primarily aim to train a one-size-fits-all system that is less controversial. In contrast, we focus on user adaptation by modeling both individual-specific preferences and group-level trends to estimate satisfaction across both minority and majority populations.

Specifically, we introduce a User-specific Chain-of-Thought (**UCoT**) prompting strategy and synthesize Chain-of-Personalized-Reasoning (**CoPeR**) outputs, which capture individual user preferences through explicit reasoning, linking the seeker's intent, the supporter's strategy, and the resulting satisfaction. This enables the supervised fine-tuning (SFT) model to acquire interpretable and user-aligned reasoning capabilities. Since users' majority or minority status is unknown in real-world scenarios, predefined group supervision is impractical. To address this, we propose a Majority-Minority Preference-Aware Clustering (**M²PC**) module, built upon the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), which routes users into majority or minority groups in an unsupervised manner by comparing model perplexities over their dialogues. Separate models are then trained to capture group-specific preferences. Finally, we integrate the individual-level SFT model and the group-specific models as the policy and reference models, respectively, within a Proximal Preference Optimization (PPO) (Schulman et al., 2017) framework. This results in our Preference-Adaptive Reinforcement Learning (**PAda-PPO**) approach, which jointly optimizes for both individual and group-level preferences in user satisfaction estimation. Our main contributions are:

- To the best of our knowledge, we present the first framework that models both individual- and group-level preferences for satisfaction estimation, capturing user diversity by distinguishing majority and minority preferences.
- We propose **UCoT** prompting and **CoPeR** synthesis, two CoT-based methods that infer individual satisfaction through explicit reasoning over user intent and system strategy.
- We introduce **M²PC**, an EM-based unsupervised module that clusters users into majority and minority groups by dialogue perplexity, capturing group-level preference trends.
- We develop **PAda-PPO**, a preference-adaptive reinforcement learning framework that integrates individual- and group-specific models to optimize satisfaction while preserving diverse user preferences.

2 Related Work

2.1 User Satisfaction Estimation

Accurately predicting user satisfaction is essential for evaluating and improving conversational systems. Prior research has focused on perspectives such as sentiment analysis (Song et al., 2019), context- and dynamics-aware modeling (Choi et al., 2019; See and Manning, 2021; Deng et al., 2022; Ye et al., 2023), and LLM-based frameworks that improve interpretability by inducing human-readable rubrics or dialogue-strategy features (Lin et al., 2024; Kim et al., 2025). However, existing approaches often model all users jointly, which tends to suppress minority preferences.

2.2 Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting has been widely used with LLMs to improve task performance by encouraging step-by-step reasoning (Wei et al., 2022; Chae et al., 2023; Luong et al., 2024; Xie et al., 2025). For instance, Zhang et al. (2024) introduce ESCoT, which augments emotional support dialogues with emotion- and strategy-focused CoTs to emulate human emotional reasoning. Differently, we leverage CoT prompting to elicit interpretable reasoning traces that explain how user preferences influence the satisfaction score.

2.3 Aligning Language Models with Diverse Human Preferences

Recent work has incorporated preference diversity into reward modeling for RL-based alignment (Jang et al., 2023; Wang et al., 2023; Li et al., 2024; Chakraborty et al., 2024; Wang et al., 2024; Yang et al., 2024; La Malfa et al., 2025). For example, Wang et al. (2023) model rewards as a posterior over annotator opinions to capture disagreement, while Chakraborty et al. (2024) optimize for minority preferences via MaxMin-RLHF based on the

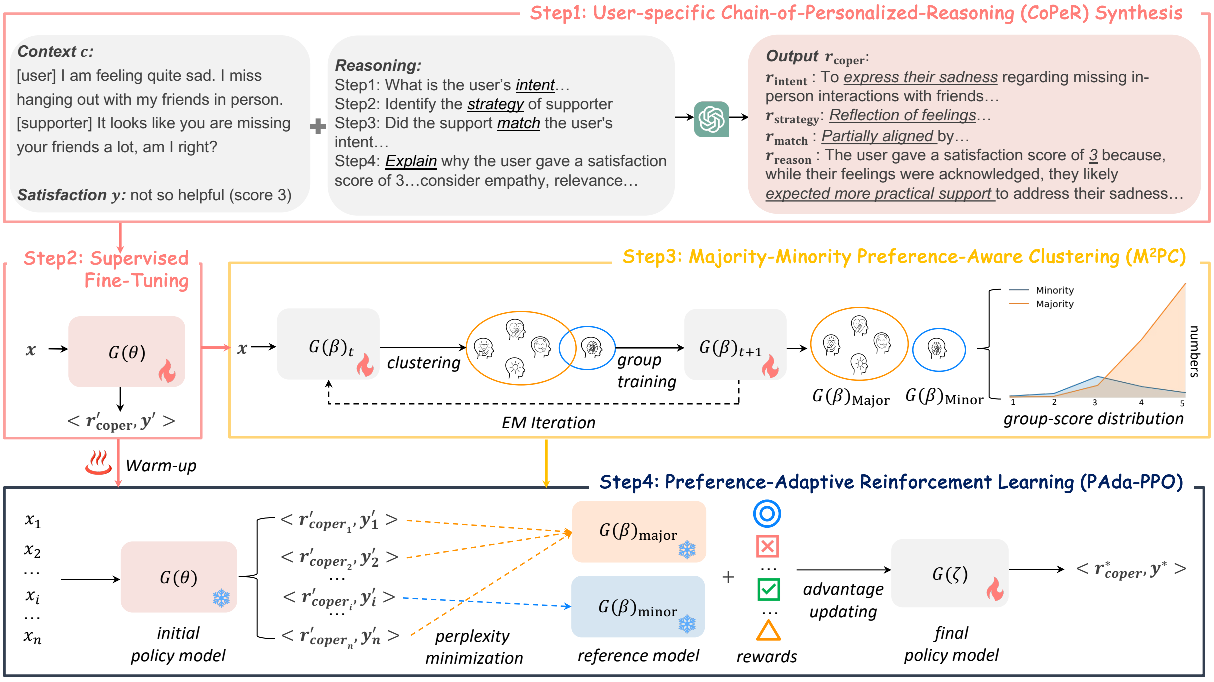

Figure 2: The architecture of our proposed method for the user's satisfaction y estimation, x is the concatenation of context c and UCoT prompt r_{ucot}.

Egalitarian principle. Other approaches, such as Personalized-RLHF (Li et al., 2024) and Personalized Soups (Jang et al., 2023), adapt to diverse users by attaching user embeddings or merging specialist policy models.

However, these methods either aim for a one-size-fits-all solution or maintain multiple policies for different groups. In contrast, we train a single policy model that adapts to users by jointly modeling individual and group-level preferences. In addition, prior work emphasizes reward modeling, RL training also depends on KL regularization with respect to a reference model, typically initialized from the SFT model. As the SFT model tends to overfit to majority behaviors, the reference model may inherit this bias, limiting its ability to guide the policy across diverse users.

We address this limitation from the source by enhancing preference awareness at both SFT and RL stages. We first model individual preferences via UCoT/ CoPeR during SFT, and then derive separate majority/minority reference models through M²PC clustering, enabling KL regularization to better reflect user diversity during RL.

3 Preliminary

We conducted our experiments on the Emotional Support Conversation (ESConv) dataset (Liu et al., 2021). Throughout each conversation, help-seekers (users) provided feedback every two utterances received from the supporter, rating the helpfulness of these messages on a five-star scale. We denote this feedback score as "satisfaction score." We categorized feedback scores of 3 or lower as "low satisfaction" and scores higher than 3 as "high satisfaction." There is no clear, universally accepted rule for partitioning users, factors such as race, age, gender, and personality will create highly varied preferences (Aroyo et al., 2023; Chakraborty et al., 2024; Fu et al., 2024). Therefore, we heuristically classified users whose proportion of high satisfaction scores exceeded 60% as belonging to a "majority" population, with the remainder designated as the "minority." Consequently, 81.4% of conversations were identified as majority, while 18.6% fell into the minority category.

4 Proposed Method

In this work, we tackle user satisfaction prediction in dialogue, given a user's input and a supporter's response. As shown in Figure 2, our framework consists of four main steps.

We first introduce UCoT prompt and use GPT-4.1-mini[2] for user-specific CoPeR synthesis, then fine-tune a base model $G(\theta)$ on this reasoning to

[2] https://platform.openai.com/docs/models/gpt-4.1-mini

equip it with user-specific inference ability. Subsequently, we introduce M^{2}PC algorithm to group users into two distinct clusters, training a separate model for each. Finally, these group-specific models serve as reference models in the proposed PAda-PPO framework.

4.1 User-specific CoPeR Synthesis

4.1.1 UCoT Prompt

Drawing on findings that user satisfaction hinges on correctly identifying the user's intent and deploying responses that appropriately match the intent and evolving needs (Deng et al., 2022; Fu et al., 2023a; Lin et al., 2024), we design a User-specific Chain-of-Thought (**UCoT**) prompt to make this reasoning explicit for satisfaction estimation. Given a dialogue context, the model is guided to: (1) infer the user's underlying intent; (2) identify the supporter's primary response strategy (e.g., *Question*, *Reflection of Feelings*); (3) evaluate whether the strategy aligns with the user's need; and (4) predict the user's feedback score by considering factors such as empathy and relevance. Details of the prompt appear in Appendix A (Figure 6).

4.1.2 CoPeR Synthesis

To train a model with the reasoning ability from steps (1)–(3) to correctly predict the user's feedback score, we use GPT-4.1-mini to synthesize reasoning rationales conditioned on the user's actual feedback score. Consequently, step (4) is to "explain the rationale behind the user's feedback score in terms of emotional and practical relevance." Further details are provided in Appendix A (Figure 7). This structured format enables more interpretable synthesis of user-centered reasoning behind feedback scores.

4.2 Supervised Fine-Tuning

After obtaining the synthesized CoPeR output r_{coper}, which comprises four components: r_{intent}, r_{strategy}, r_{match}, and r_{reason} (as illustrated in Figure 2), along with the UCoT prompt r_{ucot}, we fine-tune a base model on a dataset of input-target pairs formatted as $(c + r_{\text{ucot}}, r_{\text{coper}} + y)$. Formally, the output generation process can be decomposed into a sequence of next token prediction actions, denoting as $e = [a_1, a_2, \ldots, a_T, s, \text{<eos>}]$. The training objective is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{T} \log G_\theta(a_t \mid c + r_{\text{ucot}}, a_{<t}) - \log G_\theta(s \mid c + r_{\text{ucot}}, a_{1:T}), \quad (1)$$

where a_t is the token in the vocabulary, s represents the satisfaction score, and <eos> indicates the end-of-sequence token. The first term supervises the generation of user-specific reasoning (r_{coper}), and the second term guides the model to accurately predict the satisfaction score (s) conditioned on the generated reasoning, and SFT stands for supervised fine-tuning. This warm-up phase equips the model with interpretable and user-aligned reasoning capabilities to generate a proper response.

4.3 M^{2}PC

Human preferences in dialogue vary due to factors such as socio-cultural background (Aroyo et al., 2023; Chakraborty et al., 2024), personality (Richendoller and Weaver III, 1994), and gender (Costa et al., 2014). To model such diversity, we first examine preference differences between majority and minority users according to the supporter strategies, then apply an unsupervised EM-based clustering method to separate them.

4.3.1 Preference Divergence Across Groups

We categorize supporter strategies in the ESConv dataset into two strategy types: Cognition-oriented, including *Question*, *Restatement or Paraphrasing*, *Providing Suggestions*, and *Information*; and Emotion-oriented, including *Reflection of Feelings*, *Self-disclosure*, and *Affirmation and Reassurance*. We then computed the score distributions (1–5) for each strategy within the majority and minority user groups (Figure 3).

Across both majority and minority groups, emotion-oriented strategies are more likely to receive high feedback (0.94 > 0.91; 0.44 > 0.33), with the effect especially pronounced in the minority group. This preference divergence underscores the need for personalized modeling that accounts for group-specific preferences.[3]

4.3.2 Diversity-Aware Clustering

We adopt an EM strategy that divides users into majority and minority based on the perplexity of

[3]Aggregated proportions reported in the text are obtained by summing the bars for scores 4 and 5 (e.g., 0.94 = 0.63 + 0.31 for emotion-oriented strategies in the majority group), further details are provided in Appendix B.

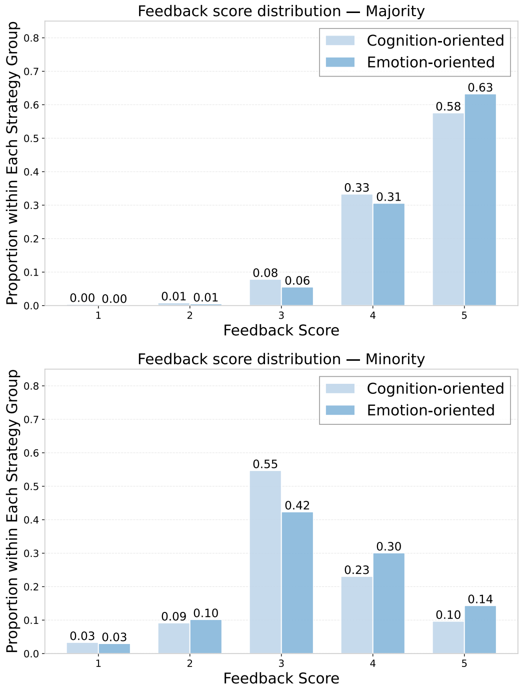

Figure 3: Distribution of feedback scores for cognition- and emotion-oriented supporter response strategies across majority and minority user groups from ESConv dataset.

their dialogue data under different models. Each model is further fine-tuned on dialogues from the belonging cluster, separately. Formally, the algorithm proceeds as follows:

E-step (Expectation): Given two cluster models at iteration t, denoted by $G(\beta)^{(t)}_{\text{Major}}$ and $G(\beta)^{(t)}_{\text{Minor}}$ (both initialized from the base model $G(\theta)$ trained from the SFT stage), each user i with dialogue set D_i is assigned to the model with lower perplexity:

$$l_i^{(t)} = \arg \min_{k \in \{\text{Major,Minor}\}} \text{PPL}(D_i; G(\beta)_k^{(t)}), \quad (2)$$

where $l_i^{(t)}$ denotes the cluster assignment of user i at the t-th EM iteration, and the perplexity is:

$$\text{PPL}(D_i;G(\beta)) = \exp\left(-\frac{1}{|D_i|}\sum_{w \in D_i} \log P_{G(\beta)}(w)\right) \quad (3)$$

This step is to route users toward the model most closely aligned with their group preferences.

M-step (Maximization): After cluster assignments are updated, we fine-tune each model $G(\beta)_k$ on dialogues assigned to its respective cluster:

$$G(\beta)_k^{(t+1)} = \arg \min_{G(\beta)_k^{(t)}} \sum_{i:l_i^{(t)}=k} \mathcal{L}(G(\beta)_k^{(t)}; D_i), \quad (4)$$

where $\mathcal{L}$ is the negative log-likelihood loss over the dialogue set D_i. This EM process iteratively refines cluster assignments and model parameters, enabling the system to unsupervised discover latent user groups that reflect diverse preference patterns.

4.4 PAda-PPO

Leveraging M²PC, each input within a batch is routed to either $G(\beta)_{\text{Major}}$ or $G(\beta)_{\text{Minor}}$ based on the model yielding lower perplexity, as shown in Figure 2. We employ PPO (Schulman et al., 2017) with a clipped objective algorithm for training. Following Luong et al. (2024), the value model V_ϕ is contructed by appending a linear value head on top of the last hidden states of the policy model $G(\xi)$, which is initialized from the SFT model $G(\theta)$.

Reward Modeling Given a dataset consisting of (input, target) pairs $(c + r_{\text{ucot}}, r_{\text{coper}} + y)$, and generated output as $e = [a_1, a_2, \ldots, a_T, s, \texttt{<eos>}]$, the reward function is as follows:

$$r_T = \begin{cases} +1 & s = y, \\ -1 & \text{otherwise.} \end{cases} \quad (5)$$

The reward of 0 is given for the intermediate tokens ($r_t = 0$ for $t <= T$), such partial reward can help reduce the effect of learning from sparse reward (Trott et al., 2019).

Diversity-aware KL Regularization A divergence penalty was utilized to prevent the policy from diverging significantly from human-like reference behaviors in each group. At each timestep t, the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the current RL policy $G(\xi)$ and the corresponding reference policy $G(\beta)_m$ is computed as:

$$\text{KL}_t = D_{\text{KL}}\big(G(\xi)(\cdot \mid s_t, m) \parallel G(\beta)_m(\cdot \mid s_t)\big), \quad (6)$$

where $G(\beta)_m$ corresponds to $G(\beta)_{\text{Major}}$ or $G(\beta)_{\text{Minor}}$ according to the perplexity routing, $m \in \{\text{Major}, \text{Minor}\}$. s_t comprises of all tokens in the input and all tokens generated so far. In addition, following Zheng et al. (2023); Luong et al. (2024), the total reward at each timestep combines the reward function score and KL penalty terms:

$$r_t^{\text{total}} = r_t - \lambda_{\text{KL}} \cdot \text{KL}_t, \quad (7)$$

where λ_{KL} controls penalty strength.

Optimization Objective Following Generalized Advantage Estimation (Schulman et al., 2015), we compute advantages $\hat{A}_t$:

$$\delta_t = r_t^{\text{total}} + \gamma V_{\phi_{\text{old}}}(s_{t+1}) - V_{\phi_{\text{old}}}(s_t),$$
$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \tag{8}$$

with discount factor $\gamma \in (0, 1]$ and $\lambda \in (0, 1]$. The value function $V_\phi(s_t)$ is estimated by a value head jointly trained with the policy. $\mathcal{L}_{\text{value}}(\phi)$ minimizes deviation between value estimate $V_\phi(s_t)$ and return estimate $\hat{R}_t = \hat{A}_t + V_{\phi_{\text{old}}}(s_t)$:

$$\mathcal{L}_{\text{value}}(\phi) = \frac{1}{2}\,\mathbb{E}_t\Big[\max\Big((V_\phi(s_t) - \hat{R}_t)^2, \big(V_{\phi_{\text{old}}}(s_t) + \text{clip}\big(V_\phi(s_t) - V_{\phi_{\text{old}}}(s_t), -\epsilon, \epsilon\big) - \hat{R}_t\big)^2\Big)\Big]. \tag{9}$$

The final policy objective is:

$$\mathcal{L}_{\text{policy}}(\xi) = -\,\mathbb{E}_t\Big[\min\Big(\rho_t(\xi)\hat{A}_t, \text{clip}\left(\rho_t(\xi),\, 1-\epsilon,\, 1+\epsilon\right)\hat{A}_t\Big)\Big],$$

where the probability ratio $\rho_t(\xi)$ is:

$$\rho_t(\xi) = \frac{G(\xi)(a_t \mid s_t, m)}{G(\xi_{\text{old}})(a_t \mid s_t, m)}, \tag{10}$$

and ϵ is a clipping parameter. The combined PPO loss to minimize is:

$$\mathcal{L}_{\text{PPO}}(\xi, \phi) = \mathcal{L}_{\text{policy}}(\xi) + c_{\text{VF}} \cdot \mathcal{L}_{\text{value}}(\phi), \tag{11}$$

where c_{VF} is the coefficient for the value objective.

5 Experiments

5.1 Dataset

We conducted our experiments on the ESConv dataset, which contains 1,300 conversations comprising 38,365 utterances between help-seekers (user) and supporters. The help-seeker provides a feedback score for every two dialogue turns exchanged with the supporter. Accordingly, we concatenate each two-turn as a single input and pair it with the corresponding feedback score as the output. Each input also includes the preceding contexts among the same conversation. We split the dataset into training, validation, and test sets using an 8:1:1 ratio at the conversation level, ensuring no conversation overlap among these subsets.

5.2 Settings

All models were fine-tuned using parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2022) with 8-bit quantization. LoRA adapters were configured with a rank of $r = 16$, scaling factor $\alpha = 16$, and a dropout rate of 0.1. The base model weights were frozen, and only LoRA parameters were updated. We fixed the random seed to 42 and set the maximum input length to 1024. At inference time, we applied top-p sampling with $p = 0.85$, a temperature of 0.7, and a maximum generation length of 256 tokens.

During the SFT stage, we used a learning rate of 1×10^{-4}, a batch size of 8, and trained for up to 15 epochs with an early stopping patience of 3. For the M^{2}PC stage, training was conducted for 10 EM iterations with a batch size of 2, gradient accumulation steps of 4, and a learning rate of 1×10^{-5}. We set diverse user clusters as 20 each for the majority and minority groups for clustering initialization. Both SFT and M^{2}PC stages were trained on a single NVIDIA RTX A6000 GPU (49GB) using the AdamW optimizer (Loshchilov and Hutter, 2017).

For the PAda-PPO stage, training was performed on 4 NVIDIA RTX A6000 GPUs (49GB each) using DeepSpeed ZeRO Stage 2 (Rasley et al., 2020) and Hugging Face Accelerate (Gugger et al., 2022). We used a batch size of 2, gradient accumulation of 2, a learning rate of 3×10^{-7}, and trained for 5 epochs. Following Ziegler et al. (2019); Luong et al. (2024), we set PPO hyperparameters as follows: $\lambda = 1$, $\gamma = 0.95$, $c_{\text{VF}} = 0.1$, $\epsilon = 0.2$. And the KL penalty coefficient is set to 0.2.

5.3 Comparative Models

We organized our experiments into three stages: *zero-shot inference*, *UCoT inference*, and *supervised fine-tuning (SFT)* using LoRA.

Zero-shot: we evaluated three backbone models with base prompt (Appendix A, Figure 5): Llama-3.2-1B-Instruct[4], Llama-3-8B-Instruct[5], and GPT4.1-mini, without training.

UCoT: We augmented the input with UCoT prompts, without fine-tuning.

SFT (LoRA): We fine-tuned Llama-3.2-1B-Instruct and Llama-3-8B-Instruct using LoRA adapters with both base and UCoT-augmented prompts. While the model with UCoT or CoPeR

[4]https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[5]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

module shares the same prompt, their training targets differ: the model with UCoT supervision uses only the final feedback score, whereas the model with CoPeR supervision additionally includes the synthesized step-by-step reasoning output (r_{coper}).
Reinforcement Learning (RL): we further fine-tuned the SFT models using both standard PPO (Schulman et al., 2017) and our proposed PAda-PPO. Specifically, we applied these methods to the following SFT variants: Llama-3-8B-Instruct, Llama-3-8B-Instruct-UCoT, and Llama-3-8B-Instruct-CoPeR.

5.4 Evaluation Metrics

We adopt four classification metrics: the F_1 score for each individual class ("low satisfaction" and "high satisfaction"), macro-averaged F_1, and weighted F_1. The class-wise F_1 scores (denoted as F_1^{low} and F_1^{high} measure how well the model predicts each satisfaction label individually. The macro F_1 (F_1^{m}) is computed by averaging the F_1 scores of both classes, giving equal weight to each regardless of class frequency. The weighted F_1 (F_1^{w}) further accounts for label imbalance by weighting each class F_1 score by its support.

6 Results and Analysis

6.1 Evaluation of Synthesized Rationales by GPT-4.1-mini

We evaluate the accuracy and quality of synthesized rationales by GPT-4.1-mini. Individual preference is inherently subjective, human evaluation may be unreliable: annotators cannot directly access the seeker's real thoughts. We assess the quality and accuracy of the synthesized rationales with two objective metrics:
(1) **Supporter-strategy accuracy (7 classes)**: measures the alignment between the supporter strategy predicted in step 2 and the true label.
(2) **Logical accuracy**: measures the consistency between reasoning steps (Steps 1–3) and the final judgment (Step 4). Step 3 infers whether the supporter's strategy aligns with the seeker's intent based on predictions from Steps 1–2, while Step 4 provides the gold satisfaction score (1–5). We map "matched intent" to scores 4–5, "partially matched" to 2–3, and "did not match" to 1. A rationale is considered logically correct when this mapping holds. On the training set, the synthesized rationales by GPT-4.1-mini achieve 63.44% supporter-strategy accuracy and 74.16% logical accuracy; on the validation set, 61.04% and 77.23%, respectively. Appendix C presents additional case studies.

Models	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}
Zero-shot				
Llama-3.2-1B	0.27	0.49	0.45	0.38
Llama-3-8B	0.34	0.71	0.65	0.52
GPT4.1-mini	**0.40**	0.67	0.62	0.53
UCoT				
Llama-3.2-1B-UCoT	0.28	0.40	0.38	0.34
Llama-3-8B-UCoT	0.26	0.80	0.72	0.53
GPT4.1-mini-UCoT	0.38	0.57	0.54	0.47
SFT (with LoRA)				
Llama-3.2-1B	0.28	0.76	0.67	0.52
Llama-3.2-1B-**UCoT**	0.18	**0.88**	0.75	0.53
Llama-3.2-1B-**CoPeR**	0.27	0.85	0.74	0.56
Llama-3-8B	0.24	0.82	0.71	0.53
Llama-3-8B-**UCoT**	0.27	0.86	0.75	0.56
Llama-3-8B-**CoPeR**	0.30	0.86	**0.76**	**0.58**

Table 1: **Results of the proposed UCoT and CoPeR:** Comparison of baseline models and our proposed UCoT and CoPeR module. We report F_1 scores on low (F_1^{low}) and high (F_1^{high}) satisfaction classes, along with weighted (F_1^{w}) and macro (F_1^{m}) averages. **Bold** in F_1 indicates the best performance. All Llama models refer to the instruction-tuned versions.

6.2 Results of the Proposed UCoT and CoPeR

Table 1 reports classification performance across three settings: *Zero-shot*, *UCoT*, and *SFT (with LoRA)*. Overall, supervised fine-tuning (SFT) consistently outperforms both the zero-shot and inference-only UCoT settings, emphasizing the importance of task-specific adaptation. Prompting with UCoT alone results in degraded performance compared to the zero-shot baseline, suggesting that, without a supervision, UCoT may overly constrain the model's reasoning process, failing to predict correctly. Within the SFT setting, models fine-tuned with UCoT prompts consistently outperform their counterparts across both backbone architectures (Llama-3.2-1B-Instruct and Llama-3-8B-Instruct). Furthermore, our proposed CoPeR approach, which extends UCoT by additionally supervising the model to generate user-specific reasoning as output, achieves the highest performance in both weighted and macro F_1 scores. These results highlight the effectiveness of combining user-specific prompting with explicit supervision of personalized reasoning in enhancing the model's ability to predict user satisfaction from dialogue.

Models	Minority				Majority				Minority+Majority			
	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}
Llama-3-8B	0.17	0.47	0.28	0.32	0.16	0.90	0.84	0.53	0.16	0.84	0.71	0.50
+M^{2}PC	0.68	**0.55**	0.61	0.61	0.19	0.94	**0.89**	**0.57**	0.56	0.89	0.82	0.72
Llama-3-8B-UCoT	0.36	0.49	0.41	0.42	0.21	0.88	0.83	0.55	0.26	0.83	0.72	0.55
+M^{2}PC	0.73	0.18	0.56	0.46	0.00	**0.96**	0.88	0.48	0.58	0.92	0.85	0.75
Llama-3-8B-CoPeR	0.29	0.53	0.37	0.41	**0.21**	0.90	0.85	0.56	0.24	0.85	0.74	0.55
+M^{2}PC	**0.80**	0.46	**0.70**	**0.63**	0.08	0.95	0.87	0.51	**0.60**	**0.92**	**0.86**	**0.76**

Table 2: **Results of the proposed M^{2}PC:** F_1 scores on the validation set across three model backbones, evaluated on minority, majority, and combined user groups. For each model, the reported M^{2}PC results correspond to the best-performing EM iteration. Llama-3-8B refers to the Llama-3-8B-Instruct version.

6.3 Results of the Proposed M^{2}PC

6.3.1 Analysis of Majority and Minority

We first evaluate the proposed method on the majority and minority groups to assess its group-level performance. Table 2 reports F_1 scores on the validation set across three model backbones: Llama-3-8B-Instruct, Llama-3-8B-Instruct-UCoT, and Llama-3-8B-Instruct-CoPeR, before and after applying our proposed M^{2}PC method. Across all configurations, M^{2}PC consistently improves overall performance, as measured by all four metrics on the combined minority and majority groups. For example, when applied to Llama-3-8B-Instruct-CoPeR, M^{2}PC improves F_1^{low} from 0.24 to 0.60, F_1^{high} from 0.85 to 0.92, F_1^{w} from 0.74 to 0.86, and F_1^{m} from 0.55 to 0.76. For the minority group, M^{2}PC leads to significant improvements in low satisfaction prediction across all backbones. In the CoPeR setting, for example, F_1^{low} increases from 0.29 to 0.80, F_1^{w} from 0.37 to 0.70, and F_1^{m} from 0.41 to 0.63. For the majority group, M^{2}PC consistently enhances high satisfaction prediction as well, for example, increasing from 0.90 to 0.95 in the CoPeR configuration. Although there is a modest decline in high satisfaction scores for the minority group and low satisfaction scores for the majority group, the overall performance improves, as reflected in higher weighted F_1 scores. The detailed results and trends across EM iterations can be seen in Appendix D.

These results highlight the effectiveness of modeling group-specific preferences through clustering. M^{2}PC successfully adapts to diverse user satisfaction patterns by routing each user to the model most aligned with their latent preference semantics, enabling more accurate prediction across minority-majority groups. We further investigate the effect of different initializations of user clusters for majority and minority groups, with details provided in Appendix E.

6.3.2 Analysis of Subgroups Within Majority and Minority

Next, we analyze finer-grained subgroups within the majority and minority groups. We applied the proposed M^{2}PC method to cluster the validation set into majority and minority groups. Within each group, we extracted the last hidden states from the corresponding (majority or minority) model outputs and applied k-means++ (Arthur and Vassilvitskii, 2007) to cluster users into 2-20 subgroups. We then used the silhouette score (Rousseeuw, 1987) to determine the optimal number of clusters (k) and computed the weighted F_1 for each cluster ($\in k$).

Table 3 shows that our method captures most user subgroups, including smaller yet distinct clusters (in bold), suggesting that M^{2}PC can generally adapt to diverse intra-group preferences. Since weighted F_1 is affected by the model's overall performance, we adopt a relative criterion: if smaller subgroups outperform the two largest ones in weighted F_1 within either the majority or minority population, it indicates that the model capture diverse subgroup characteristics rather than overfitting to the more frequent patterns.

6.4 Results of the Proposed PAda-PPO

Table 4 presents the performance of our proposed PAda-PPO algorithm applied to three model backbones: Llama-3-8B-Instruct (base), Llama-3-8B-Instruct-UCoT, and Llama-3-8B-Instruct-CoPeR. We compare three strategies: SFT, SFT followed by PPO, and SFT followed by PAda-PPO. Across both the base and UCoT-augmented backbones, PAda-PPO consistently outperforms SFT and PPO. For

Models	1	2	3	4	5	6	7	8	9	10
Llama-3-8B+M^{2}PC										
Maj.	*0.85(111)*	*0.68(105)*	**0.89(67)**	**0.84(60)**	**0.91(55)**	**0.91(38)**	**0.96(38)**	**0.92(31)**	**0.95(27)**	**0.79(21)**
Min.	*0.60(21)*	*0.71(21)*	**0.68(10)**	**0.77(9)**	0.63(8)	0.19(7)	**0.84(6)**	**0.91(6)**	0.57(5)	**1.00(5)**
Llama-3-8B-UCoT+M^{2}PC										
Maj.	*0.68(109)*	*0.82(104)*	**0.94(50)**	**0.87(44)**	**0.96(42)**	**0.93(41)**	**0.86(32)**	**1.00(31)**	**0.89(28)**	**1.00(24)**
Min.	*0.62(29)*	*0.61(21)*	0.60(11)	**0.87(8)**	**0.91(6)**	**0.91(6)**	**0.78(5)**	**0.80(5)**	**0.73(4)**	**1.00(4)**
Llama-3-8B-CoPeR+M^{2}PC										
Maj.	*0.71(134)*	*0.79(105)*	**0.93(56)**	**0.85(48)**	**0.94(44)**	**0.94(39)**	**0.90(30)**	**0.90(30)**	**0.82(28)**	**0.87(27)**
Min.	*0.70(22)*	*0.61(21)*	**0.67(7)**	**0.91(6)**	**0.67(6)**	**1.00(5)**	**0.80(5)**	0.34(5)	0.64(4)	0.33(4)

Models	11	12	13	14	15	16	17	18	19	20
Llama-3-8B+M^{2}PC										
Maj.	**1.00(18)**	**0.91(17)**	**1.00(10)**	**0.84(9)**	**1.00(9)**	**1.00(9)**	**1.00(6)**	-	-	-
Min.	**0.82(5)**	0.63(5)	0.00(4)	0.64(4)	0.50(4)	**0.73(4)**	**0.73(4)**	**0.73(4)**	**0.67(3)**	-
Llama-3-8B-UCoT+M^{2}PC										
Maj.	**0.88(24)**	**0.87(22)**	**0.90(15)**	**1.00(14)**	**1.00(13)**	**1.00(12)**	**1.00(9)**	**1.00(8)**	**1.00(5)**	**1.00(3)**
Min.	0.33(4)	**1.00(4)**	0.10(4)	0.33(3)	0.53(3)	0.53(3)	**0.67(2)**	**0.67(2)**	0.00(2)	0.00(2)
Llama-3-8B-CoPeR+M^{2}PC										
Maj.	**0.91(22)**	**0.94(20)**	**0.90(19)**	**0.96(14)**	**1.00(9)**	**1.00(8)**	**1.00(8)**	**1.00(7)**	**1.00(4)**	-
Min.	**1.00(4)**	0.20(4)	**0.67(3)**	0.53(3)	0.53(3)	**0.67(3)**	**1.00(3)**	**0.67(2)**	**1.00(2)**	**1.00(2)**

Table 3: Weighted F_1 scores across $k \in 1, \ldots, 20$ for the majority (Maj.) and minority (Min.) user groups, shown in two parts for readability (top: $k = 1$–10; bottom: $k = 11$–20). Numbers in parentheses indicate the number of users in each subgroup, where subgroups ($k = 1$–20) are ordered by their user counts. Bold values denote subgroups whose weighted F_1 exceeds the average of the two largest subgroups (shown in *italic*).

Models	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}
Llama-3-8B	0.24	0.82	0.71	0.53
+ PPO	0.25	0.85	0.74	0.55
+ **PAda-PPO**	0.29	0.86	0.75	0.57
Llama-3-8B-UCoT	0.27	0.86	0.75	0.56
+ PPO	0.22	0.88	0.76	0.55
+ **PAda-PPO**	**0.36**	0.86	0.77	**0.61**
Llama-3-8B-CoPeR	0.30	0.86	0.76	0.58
+ PPO	0.34	**0.88**	**0.78**	0.61
+ **PAda-PPO**	0.33	0.85	0.76	0.59

Table 4: **Results of the proposed PAda-PPO:** Comparison among SFT, PPO-based RL, and our proposed PAda-PPO method. Llama-3-8B refers to the Llama-3-8B-Instruct version.

instance, with Llama-3-8B-Instruct-UCoT, PAda-PPO outperforms PPO on F_1^{low} from 0.22 to 0.36 and macro F_1 from 0.55 to 0.61, demonstrating its effectiveness in improving minority-class satisfaction prediction while maintaining strong performance on the majority class.

However, on the Llama-3-8B-Instruct-CoPeR backbone, PAda-PPO performs slightly worse than PPO. We attribute this to the noise introduced by the GPT-4.1-mini synthesized CoPeR rationales. When the synthesized user-specific reasoning conflicts with the preference routing signals, it will lead to a higher KL divergence penalty (Equation 6), destabilizing training. Improving the quality of CoPeR synthesis will be done in future work.

7 Conclusions and Future Work

We have introduced a unified framework for satisfaction estimation that simultaneously models individual- and group-level preferences among majority and minority user populations. We proposed UCoT and CoPeR to generate interpretable reasoning chains for capturing individual preferences, and developed M^{2}PC, an unsupervised clustering module for identifying group-level preferences. These were integrated into PAda-PPO to align dialogue systems with diverse user preferences. Experiments on the ESConv dataset demonstrate improved satisfaction estimation across different user populations. In future work, we will validate our proposed method on other LLMs, and extend preference-adaptive reinforcement learning to additional RL algorithms.

Limitations

A key limitation of our approach lies in the CoPeR supervision stage, where we rely on GPT-4.1-mini to generate reasoning chains as pseudo-ground truth. Given the inherently subjective nature of user satisfaction and context interpretation, it is challenging to synthesize rationales that accurately reflect the user's underlying intent. While this setup enables scalable training without manual annotation, the quality of the generated rationales may affect downstream fine-tuning and reinforcement learning. Inaccurate or overly generic reasoning can introduce noise into the learning process, hindering the model's ability to capture nuanced satisfaction patterns. Future work could collect partial real human data, including users' own intents and explanations for their feedback scores, to enhance the reliability of reasoning supervision.

Another limitation is that M^{2}PC does not model some subgroups of minority users well, as shown in Table 3. Future work is needed to better account for the nuanced preference patterns among these underrepresented subgroups.

Acknowledgements

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJMS2011).

References

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

David Arthur and Sergei Vassilvitskii. 2007. k-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.

Hyungjoo Chae, Yongho Song, Kai Tzu Iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 5606–5632.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Alignment with diverse human preferences. *Proceedings of Machine Learning Research*, 235:6116–6135.

Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and online satisfaction prediction in open-domain conversational systems. In *Proceedings of CIKM*.

Patricio Costa, Raquel Alves, Isabel Neto, Pedro Marvao, Miguel Portela, and Manuel Joao Costa. 2014. Associations between medical student empathy and personality: a multi-institutional study. *PloS one*, 9(3):e89254.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *Proceedings of The Web Conference (WWW)*.

Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. 2024. Styemp: Stylizing empathetic response generation via multi-grained prefix encoder and personality reinforcement. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–185.

Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023a. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–656.

Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. 2023b. Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot. *Advanced Robotics*, 37(21):1406–1418.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

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

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Sangyeop Kim, Sohhyung Park, Jaewon Jung, Jinseok Kim, and Sungzoon Cho. 2025. LLM-guided plan and retrieval: A strategic alignment for interpretable user satisfaction estimation in dialogue. In *Proceedings of NAACL*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Gabriele La Malfa, Jie M Zhang, Michael Luck, and Elizabeth Black. 2025. Fairness aware reinforcement learning via proximal policy optimization. *arXiv preprint arXiv:2502.03953*.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, and 1 others. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11115.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of ACL-IJCNLP*, pages 3469–3483.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Nadine R Richendoller and James B Weaver III. 1994. Exploring the links between personality and empathic response style. *Personality and individual Differences*, 17(3):303–311.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

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

Abigail See and Christopher D Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12.

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. 2025. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*.

Kaiqiang Song, Xiaoyang Li, Xiangchen Zhang, Dong Yu, and Lawrence Carin. 2019. Using customer service dialogues for satisfaction analysis with context-assisted multiple instance learning. In *Proceedings of EMNLP-IJCNLP*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. 2019. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. *Advances in Neural Information Processing Systems*, 32.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*.

Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. *Advances in Neural Information Processing Systems*, 36:49113–49132.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*.

Jingran Xie, Shun Lei, Yue Yu, Yang Xiang, Hui Wang, Xixin Wu, and Zhiyong Wu. 2025. Leveraging chain of thought towards empathetic spoken dialogue without corresponding question-answering data. In *ICASSP*, pages 1–5. IEEE.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialog responses. *arXiv preprint arXiv:2402.11409*.

Joshua C et al. Yang. 2024. Llm voting: Human choices and ai collective decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1696–1708.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Fanghua Ye, Zhiyuan Hu, and Emine Yilmaz. 2023. Modeling user satisfaction dynamics in dialogue via hawkes process. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8875–8889.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, and 1 others. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompts

The prompts used for base input, User-specific Chain-of-Thought, and for synthesizing User-specific Chains-of-Personalized-Reasoning are illustrated in Figure 5, Figure 6, and Figure 7 respectively.

B Analysis of Support Strategies and Feedback Scores

We classify strategies into **Cognition**-oriented (Question, Restatement or Paraphrasing, Providing Suggestions, Information) and **Emotion**-oriented (Reflection of Feelings, Self-disclosure, Affirmation and Reassurance), and summarized each strategy's counts and proportions of low (1–3) and high (4–5) feedback scores in Table 5.

We can see that for both majority and minority groups, emotion-oriented strategies are more likely to receive high feedback (0.94 > 0.91; 0.44 > 0.33), with the effect stronger in the minority group.

Group	Strategy	Low	High
Majority	Cognition	302 (0.09)	3020 (0.91)
	Emotion	120 (0.06)	1826 (0.94)
Minority	Cognition	542 (0.67)	264 (0.33)
	Emotion	240 (0.56)	192 (0.44)

Table 5: Counts and proportion of groups by support strategy and satisfaction score

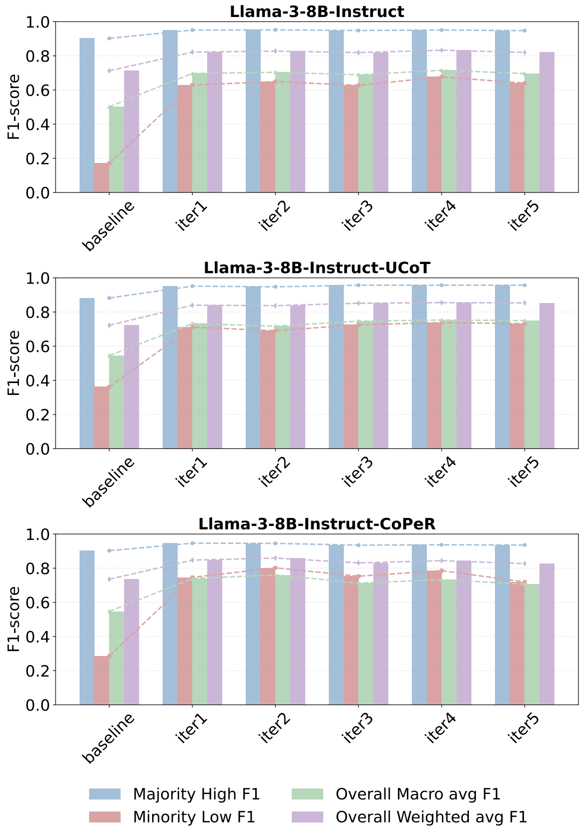

Figure 4: Results on the EM iterations during the Majority-Minority Preference-Aware Clustering stage.

C Evaluation of Synthesized Rationales

Figure 8 presents four case studies: *Low satisfaction* and *High satisfaction* examples under both correct and incorrect supporter-strategy and logical accuracies.

D Results on EM Iterations

Figure 4 presents the detailed results and trends across EM iterations during the Majority-Minority Preference-Aware Clustering stage.

E Results on Different User Clusters

For training, we randomly selected from the majority group the same number of conversations as

```
instruction = f"""Classify the seeker's satisfaction level based on the supporter's responses. This is a dialogue between a [seeker] and a [supporter].
The [seeker] is facing a problem related to '{row['problem_type']}' and is feeling '{row['emotion_type']}'."""

input_text = f"""{row['input']}

The seeker rated their satisfaction on a 5-point scale (1 = very dissatisfied, 5 = very satisfied).
To simplify, we define two categories: Ratings of 1 to 3 indicate "Low satisfaction"; Ratings of 4 or 5 indicate "High satisfaction".

Please format your response as follows, Final score and Final answer should be one of the given options.

Final score: 1, 2, 3, 4, or 5
Final answer: Low satisfaction or High satisfaction
```

Figure 5: Base prompt.

```
instruction = f"""Generate a structured reasoning path to predict the seeker's satisfaction score. Be concise and precise. Limit each step to 1 sentence.
This is a multi-turn dialogue between a [seeker] and a [supporter].
The [seeker] is experiencing a problem related to '{row['problem_type']}' and is feeling '{row['emotion_type']}'."""

input_text = f"""{row['input']}
Assume the [seeker] will rate their satisfaction on a 5-point scale (1 = very dissatisfied, 5 = very satisfied).
To simplify, we define two categories: Ratings of 1 to 3 indicate **Low satisfaction**; Ratings of 4 or 5 indicate **High satisfaction**.

Please reason step by step from the seeker's perspective:

Step 1: What is the seeker's underlying intent or goal, based on their final message and previous context?
Step 2: Identify the main strategy the supporter used in their overall response to the final [seeker]'s message. Choose from the following categories:
Question, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestions, Giving Information, Others.
Step 3: Did the support match the seeker's intent and priorities?
Step 4: Justify the seeker's likely satisfaction based on the emotional and practical support received.
Reflect on whether the supporter showed empathy, staying relevant to the concern, and responding to the seeker's intent, as well as any other relevant factors.

Please format your response as follows, Final score and Final answer should be one of the given options.
Step 1: [analysis]
Step 2: [analysis]
Step 3: [analysis]
Step 4: [analysis]
Final score: 1, 2, 3, 4, or 5.
Final answer: Low satisfaction or High satisfaction
```

Figure 6: Prompt for User-specific Chain-of-Thought.

```
instruction = "You are an expert dialogue evaluator. Your task is to reconstruct the seeker's reasoning that explains their satisfaction score.\n"
"Be concise and precise. Limit each step to 1 sentence."
input_text = f"""This is a multi-turn dialogue between a [seeker] and a [supporter].

{row['input']}

The [seeker] is experiencing a problem related to '{row['problem_type']}' and is feeling '{row['emotion_type']}'.
The seeker rated their satisfaction as {row['output']} on a scale from 1 (very dissatisfied) to 5 (very satisfied).

Please reason step by step from the seeker's perspective:

Step 1: What is the seeker's underlying intent or goal, based on their final message and previous context?
Step 2: Identify the main strategy the supporter used in their overall response to the final [seeker]'s message. Choose from the following categories:
Question, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestions, Giving Information, Others.
Step 3: Did the support match the seeker's intent and priorities?
Step 4: Explain why the seeker gave a satisfaction score of {row['output']}.
Briefly justify whether the supporter met the seeker's emotional and practical needs, such as showing empathy, staying relevant to the concern, and responding to the seeker's intent, as well as any other relevant factors.

Please format your response as:
Step 1: [analysis]
Step 2: [analysis]
Step 3: [analysis]
Step 4: [analysis]
```

Figure 7: Prompt for User-specific Chains-of-Personalized-Reasoning synthesis.

An acceptable rationale example for "Low satisfaction"
Dialogue: [seeker]: I had the job for 14 months and I drank a lot before bed which was usually about 1 am. The job was second shift. The drinking affected my REM sleep, and I was often tired the next day. [supporter]: I can understand that...Did you lose your job because of this? (strategy: Question) *[seeker] feedback score: 2* *Generated reasoning by GPT-4.1-mini:* Step 1: The seeker's underlying intent is to share their experience of how drinking affected their sleep and job performance, possibly seeking understanding or empathy. Step 2: The supporter used the strategy of Question by asking if the seeker lost their job because of the drinking and sleep issues. Step 3: The support partially matched the seeker's intent by showing some understanding but shifted focus to the job loss rather than exploring the seeker's feelings or the impact of the drinking on their well-being. Step 4: The seeker gave a satisfaction score of 2 because the supporter did not fully address their emotional state or provide empathy related to their anxiety and sleep problems, instead focusing narrowly on job loss, which may have felt dismissive or off-topic.
An acceptable rationale example for "High satisfaction"
Dialogue: [seeker]: I know my night will get worse as it goes on. [supporter]: You think your night is going to have emotional upset? Do you fear you will not be physically safe? (strategy: Question) *[seeker] feedback score: 5.* *Generated reasoning by GPT-4.1-mini:* Step 1: The seeker's underlying intent is to express their current emotional state and hint at anticipated worsening feelings, seeking understanding or support. Step 2: The supporter used the strategy of Questioning to explore the seeker's emotional and physical safety concerns. Step 3: Yes, the support matched the seeker's intent by acknowledging their emotional distress and probing gently to understand the severity of their feelings. Step 4: The seeker gave a satisfaction score of 5 because the supporter showed empathy by recognizing the seeker's emotional state and responded with relevant, caring questions that aligned with the seeker's expressed concerns, making the seeker feel heard and supported.
An unacceptable rationale example for "Low satisfaction"
Dialogue: [supporter] How are you today? [seeker] I'm doing pretty good. [supporter] I'm glad to hear you're doing well in these trying times. Are you in close proximity to your family? [seeker] Yeah, I'm staying with them for the holidays and then going back home in January. [supporter] I imagine that can create some stress? (strategy: Reflection of feelings) *[seeker] feedback score: 1.* *Generated reasoning by GPT-4.1-mini:* Step 1: The seeker's underlying intent is to express and seek support for the stress and anxiety they feel related to managing academic pressure and upcoming holiday family interactions. Step 2: The supporter used a Question strategy by asking if being close to family during the holidays creates stress. Step 3: The support partially matched the seeker's intent by acknowledging potential stress but did not directly address the seeker's academic pressure or anxiety. Step 4: The seeker gave a satisfaction score of 1 because the supporter's response was too general and did not demonstrate sufficient empathy or directly engage with the seeker's specific emotional struggles related to academic pressure and anxiety.
An unacceptable rationale example for "High satisfaction"
Dialogue: [seeker]: I feel very bad right now. I don't know how I should make my daughter understand that her dressing is not appropriate. [supporter]: It sounds like your daughter does not understand your feelings. (strategy: Restatement or Paraphrasing) *[seeker] feedback score: 5.* *Generated reasoning by GPT-4.1-mini:* Step 1: The seeker's underlying intent is to find a way to communicate effectively with their daughter about why her dressing is inappropriate. Step 2: The supporter used Reflection of Feelings by acknowledging the seeker's emotional state and paraphrasing the concern about the daughter's understanding. Step 3: The support partially matched the seeker's intent by recognizing the emotional difficulty but did not provide practical advice on how to address the daughter's behavior. Step 4: The seeker gave a satisfaction score of 5 likely because the supporter showed empathy and validated their feelings, which met the seeker's emotional needs despite the lack of concrete suggestions.

Figure 8: Case Studies.

Iteration	$\text{Users}_{\text{minor}}$	$\text{Users}_{\text{major}}$	Minority				Majority				Minority+Majority			
			F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}
0	10	10	0.29	0.53	0.37	0.41	0.15	0.91	0.85	0.53	0.22	0.85	0.73	0.54
1	10	10	0.73	0.46	0.63	0.60	0.06	0.94	0.87	0.50	0.56	0.90	0.84	0.73
2	10	10	0.72	0.40	0.61	0.56	0.09	0.94	0.88	0.52	0.56	0.90	0.84	0.73
3	9	11	0.72	0.37	0.60	0.54	0.11	0.94	0.87	0.52	0.55	0.91	0.84	0.73
4	8	12	0.70	0.26	0.56	0.48	0.08	0.94	0.85	0.51	0.50	0.90	0.83	0.70
5	9	11	0.71	0.30	0.57	0.51	0.03	0.94	0.86	0.48	0.52	0.90	0.83	0.71
0	20	20	0.29	0.53	0.37	0.41	0.11	0.91	0.85	0.51	0.20	0.86	0.73	0.53
1	18	22	0.75	0.46	0.66	0.60	0.09	0.95	0.88	0.52	0.56	0.91	0.85	0.74
2	17	23	0.80	0.46	0.70	0.63	0.08	0.95	0.87	0.51	0.60	0.92	0.86	0.76
3	16	24	0.75	0.31	0.62	0.53	0.00	0.93	0.85	0.47	0.52	0.90	0.83	0.71
4	15	25	0.79	0.38	0.67	0.58	0.09	0.94	0.86	0.52	0.55	0.91	0.84	0.73
5	17	23	0.72	0.36	0.60	0.54	0.07	0.94	0.86	0.50	0.51	0.90	0.83	0.71
0	30	30	0.29	0.53	0.37	0.41	0.11	0.91	0.85	0.51	0.20	0.86	0.73	0.53
1	27	33	0.73	0.40	0.63	0.57	0.12	0.95	0.88	0.53	0.56	0.91	0.85	0.74
2	23	37	0.77	0.42	0.67	0.60	0.07	0.94	0.85	0.50	0.53	0.91	0.84	0.72
3	25	35	0.73	0.22	0.58	0.47	0.03	0.94	0.86	0.48	0.53	0.90	0.83	0.71
4	25	35	0.73	0.29	0.60	0.51	0.03	0.94	0.86	0.49	0.54	0.91	0.84	0.72
5	24	36	0.77	0.47	0.68	0.62	0.05	0.94	0.86	0.50	0.55	0.91	0.85	0.73
0	40	40	0.29	0.53	0.37	0.41	0.15	0.91	0.85	0.53	0.22	0.85	0.73	0.54
1	34	46	0.78	0.39	0.67	0.59	0.10	0.94	0.87	0.52	0.59	0.91	0.85	0.75
2	31	49	0.78	0.46	0.69	0.62	0.00	0.94	0.85	0.47	0.53	0.91	0.84	0.72
3	31	49	0.74	0.35	0.63	0.55	0.03	0.94	0.85	0.48	0.51	0.90	0.83	0.71
4	33	47	0.76	0.42	0.66	0.59	0.03	0.94	0.86	0.48	0.54	0.91	0.84	0.72
5	27	53	0.83	0.47	0.73	0.65	0.02	0.93	0.83	0.48	0.53	0.91	0.84	0.72
0	50	50	0.29	0.53	0.37	0.41	0.11	0.91	0.85	0.51	0.20	0.86	0.73	0.53
1	40	60	0.71	0.35	0.59	0.53	0.05	0.94	0.85	0.49	0.50	0.90	0.82	0.70
2	43	57	0.73	0.47	0.64	0.60	0.08	0.94	0.87	0.51	0.55	0.91	0.84	0.73
3	37	63	0.83	0.47	0.73	0.65	0.11	0.93	0.85	0.52	0.58	0.91	0.85	0.75
4	42	58	0.77	0.36	0.63	0.57	0.06	0.94	0.87	0.50	0.59	0.91	0.85	0.75
5	36	64	0.77	0.38	0.64	0.58	0.13	0.93	0.85	0.53	0.55	0.91	0.84	0.73

Table 6: Results of the proposed **Llama-3-8B-CoPeR + M^{2}PC**, when initializing majority and minority users as 10, 20, 30, 40, and 50, respectively.

in the minority group, with each conversation representing a user. These users are evenly and randomly divided into 20 clusters per group. During M²PC training, each cluster (a batch of users) is reassigned to majority or minority groups based on perplexity in an unsupervised manner; therefore, no universally accepted partition rule is required. We believe this approach may help capture finer subpopulation preferences within each majority or minority group, as the routing perplexity is computed from the average perplexity of all users in each cluster/subpopulation. Table 6 presents the results of the proposed **Llama-3-8B-CoPeR + M^{2}PC** model when the numbers of majority and minority users are initialized to 10, 20, 30, 40, and 50, respectively. The best performance is obtained when both groups are initialized to 20, with which after the second iteration, the algorithm yields 23 distinct majority groups and 17 distinct minority groups, indicating that M^{2}PC adaptively clusters users with similar preferences.