

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

Yahui Fu, Zi Haur Pang, Tatsuya Kawahara

京都大学

KYOTO UNIVERSITY



IJCNLP-AACL 2025, Mumbai, India

# Why Model User Satisfaction Beyond the Majority?

💡 **Motivation:** User satisfaction is subjective 🧠


# Why Model User Satisfaction Beyond the Majority?

- 💡 **Motivation:** User satisfaction is subjective 🧠
  - **same response strategy  $\neq$  same satisfaction**

# Why Model User Satisfaction Beyond the Majority?

💡 **Motivation:** User satisfaction is subjective 🗣️

→ same response strategy ≠ same satisfaction



**User 1**


I miss hanging out with my friends.

How long since you all met up?

System

Intent: seeking *interaction*

Rating: ★★★★★



**User 2**


I miss spending time with my friends.

How long has it been?

System

Intent: seeking *suggestion*

Rating: ★★★★★



**User 3**

I miss being around my friends.

When did you last see them?

System

Intent: seeking *comfort*

Rating: ★☆☆☆☆

# Why Model User Satisfaction Beyond the Majority?

💡 **Motivation:** User satisfaction is subjective 🧠

→ same response strategy ≠ same satisfaction

⚠️ **Problem:** Existing alignment methods typically train **one-size-fits-all models**

→ majority voting **suppresses** minority preferences ⚖️


# Why Model User Satisfaction Beyond the Majority?

💡 **Motivation:** User satisfaction is subjective 🧠

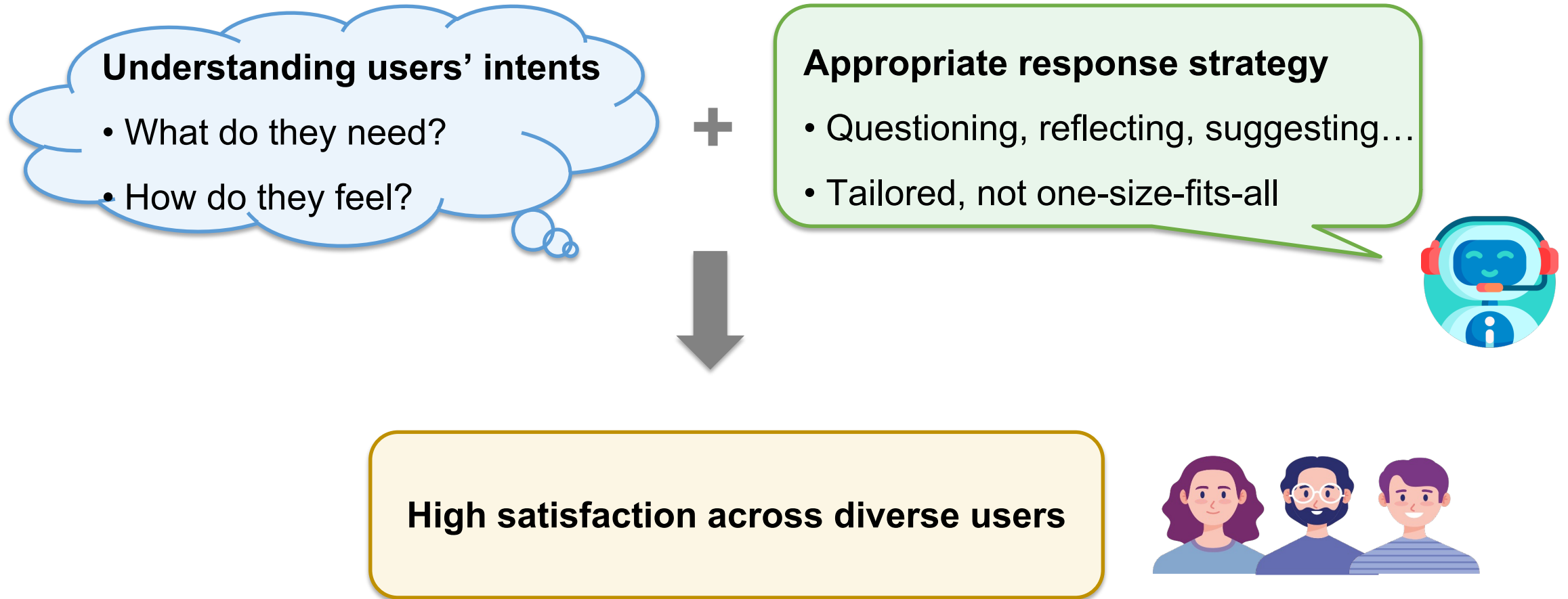
→ same response strategy ≠ same satisfaction

⚠️ **Problem:** Existing alignment methods typically train **one-size-fits-all models**

→ majority voting **suppresses** minority preferences ⚖️

🎯 **Goal:** Build a satisfaction estimator that **adapts** to both majority and minority users. 

# Method: User-specific Reasoning



# Method: User-specific Reasoning

## Step1: User-specific Chain-of-Personalized-Reasoning (CoPeR) Synthesis

### Context $c$ :

[user] I am feeling quite sad. I miss hanging out with my friends in person.  
[supporter] It looks like you are missing your friends a lot, am I right?

**Satisfaction  $y$ :** not so helpful (score 3)

+

### Reasoning:

Step1: What is the user's intent...  
Step2: Identify the strategy of supporter  
Step3: Did the support match the user's intent...  
Step4: Explain why the user gave a satisfaction score of 3...consider empathy, relevance...

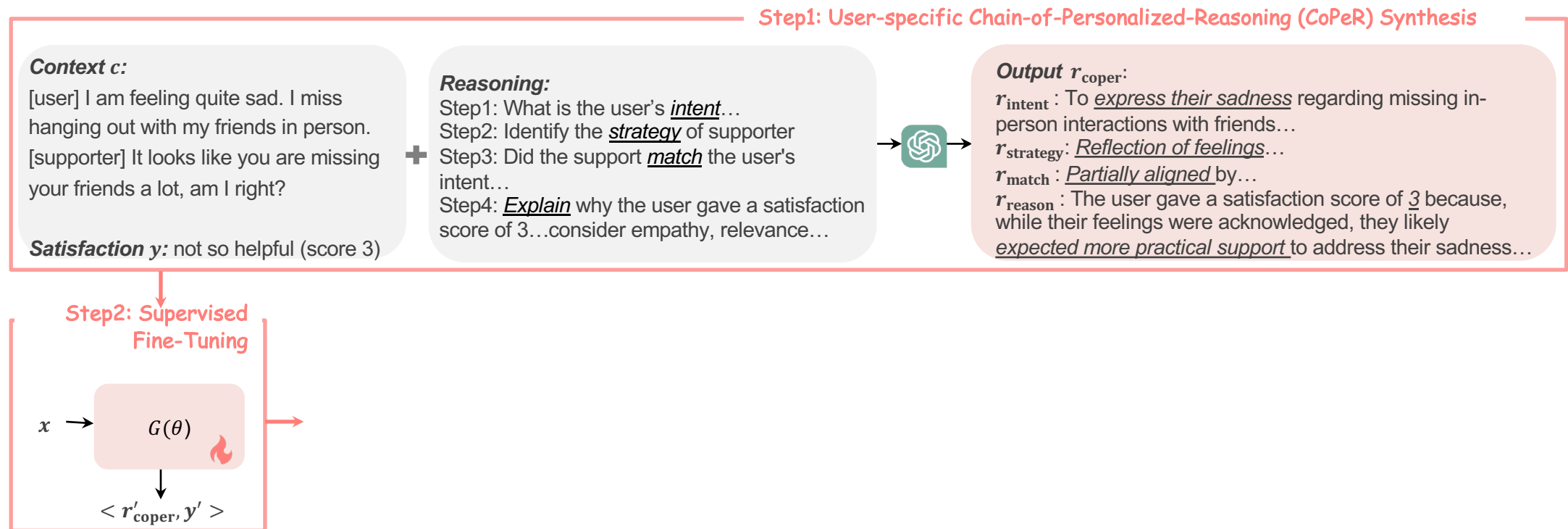


### Output $r_{\text{coper}}$ :

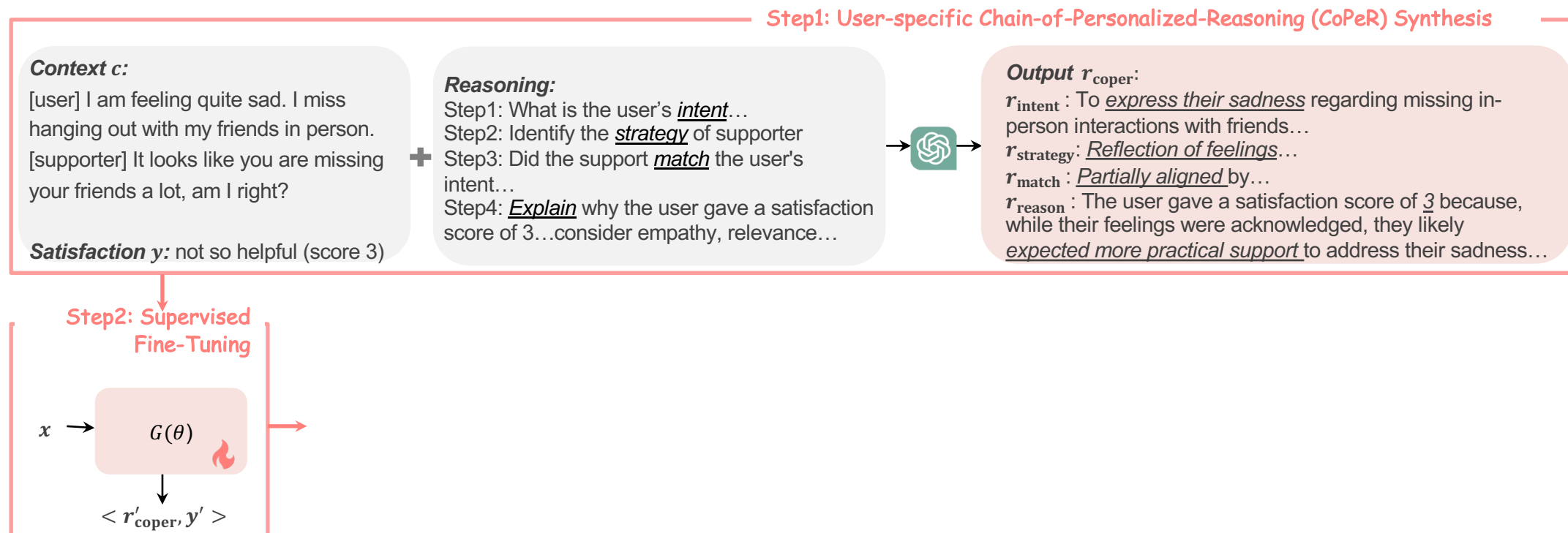
$r_{\text{intent}}$  : To express their sadness regarding missing in-person interactions with friends...  
 $r_{\text{strategy}}$  : Reflection of feelings...  
 $r_{\text{match}}$  : Partially aligned by...  
 $r_{\text{reason}}$  : The user gave a satisfaction score of 3 because, while their feelings were acknowledged, they likely expected more practical support to address their sadness...



# Method: User-specific Reasoning



# Method: User-specific Reasoning



SFT still mixes all users → minority preferences get **suppressed**.

Real systems lack group labels → supervised separate training is **impossible**.

# Method: Majority-Minority Preference-Aware Clustering (M<sup>2</sup>PC)

## Step1: User-specific Chain-of-Personalized-Reasoning (CoPeR) Synthesis

### Context $c$ :

[user] I am feeling quite sad. I miss hanging out with my friends in person.  
[supporter] It looks like you are missing your friends a lot, am I right?

**Satisfaction  $y$ :** not so helpful (score 3)

### Reasoning:

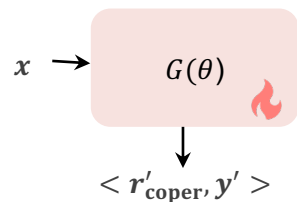
Step1: What is the user's intent...  
Step2: Identify the strategy of supporter  
Step3: Did the support match the user's intent...  
Step4: Explain why the user gave a satisfaction score of 3...consider empathy, relevance...



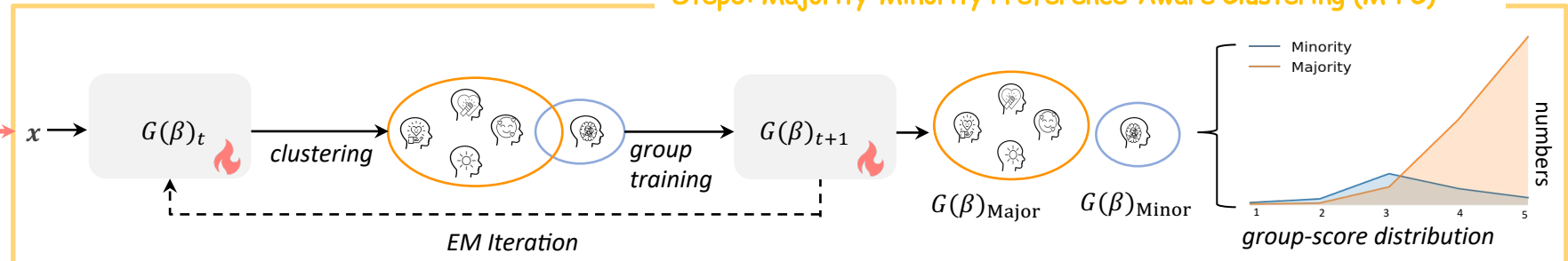
### Output $r_{\text{coper}}$ :

$r_{\text{intent}}$  : To express their sadness regarding missing in-person interactions with friends...  
 $r_{\text{strategy}}$  : Reflection of feelings...  
 $r_{\text{match}}$  : Partially aligned by...  
 $r_{\text{reason}}$  : The user gave a satisfaction score of 3 because, while their feelings were acknowledged, they likely expected more practical support to address their sadness...

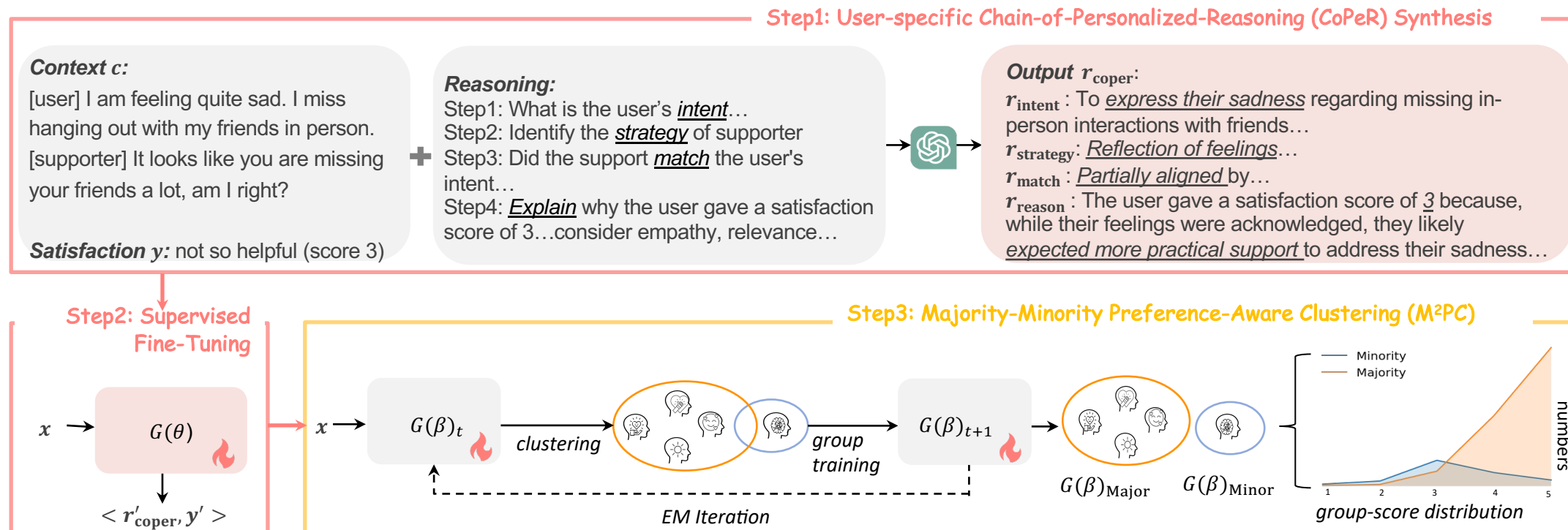
## Step2: Supervised Fine-Tuning



## Step3: Majority-Minority Preference-Aware Clustering (M<sup>2</sup>PC)



# Method: Majority-Minority Preference-Aware Clustering (M<sup>2</sup>PC)



Reference model inherits majority bias  $\rightarrow$  RL cannot adapt well.

M<sup>2</sup>PC learns majority/minority references  $\rightarrow$  RL adapts to each group.

# Method: Preference-Adaptive Reinforcement Learning

## Step1: User-specific Chain-of-Personalized-Reasoning (CoPeR) Synthesis

### Context $c$ :

[user] I am feeling quite sad. I miss hanging out with my friends in person.  
[supporter] It looks like you are missing your friends a lot, am I right?

**Satisfaction  $y$ :** not so helpful (score 3)

### Reasoning:

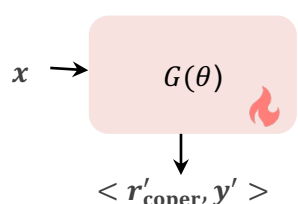
Step1: What is the user's intent...  
Step2: Identify the strategy of supporter  
Step3: Did the support match the user's intent...  
Step4: Explain why the user gave a satisfaction score of 3...consider empathy, relevance...



### Output $r_{\text{coper}}$ :

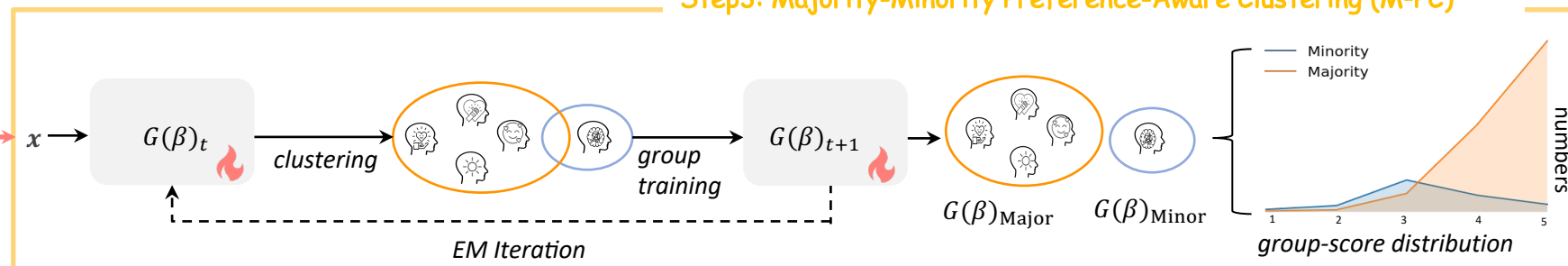
$r_{\text{intent}}$ : To express their sadness regarding missing in-person interactions with friends...  
 $r_{\text{strategy}}$ : Reflection of feelings...  
 $r_{\text{match}}$ : Partially aligned by...  
 $r_{\text{reason}}$ : The user gave a satisfaction score of 3 because, while their feelings were acknowledged, they likely expected more practical support to address their sadness...

## Step2: Supervised Fine-Tuning

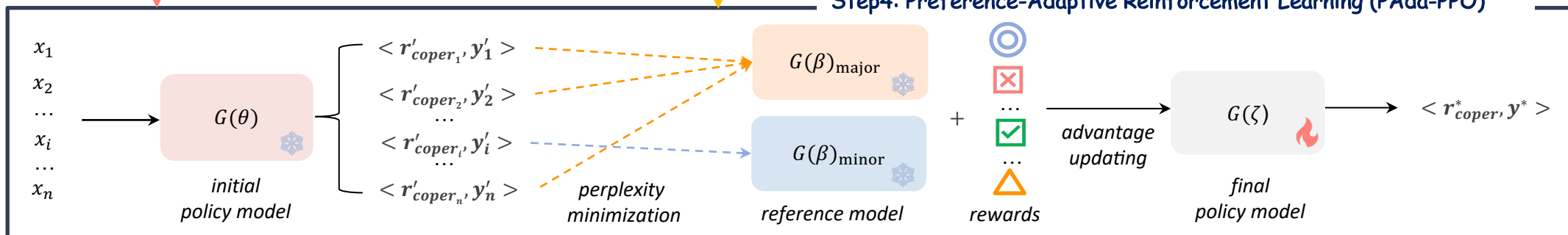


Warm-up

## Step3: Majority-Minority Preference-Aware Clustering (M<sup>2</sup>PC)



## Step4: Preference-Adaptive Reinforcement Learning (PAa-PPO)



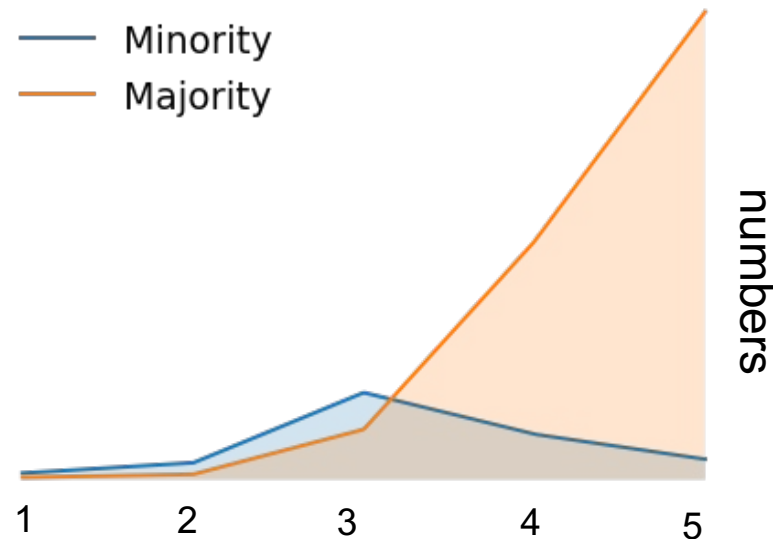
# Experiment

❑ **Dataset:** Emotional Support Conversation

❑ **User Groups**

- Majority (81.4%):  $>60\%$  high-satisfaction scores per dialogue
- Minority (18.6%):  $\leq 60\%$  high-satisfaction scores

❑ *Group-Satisfaction score Distribution*



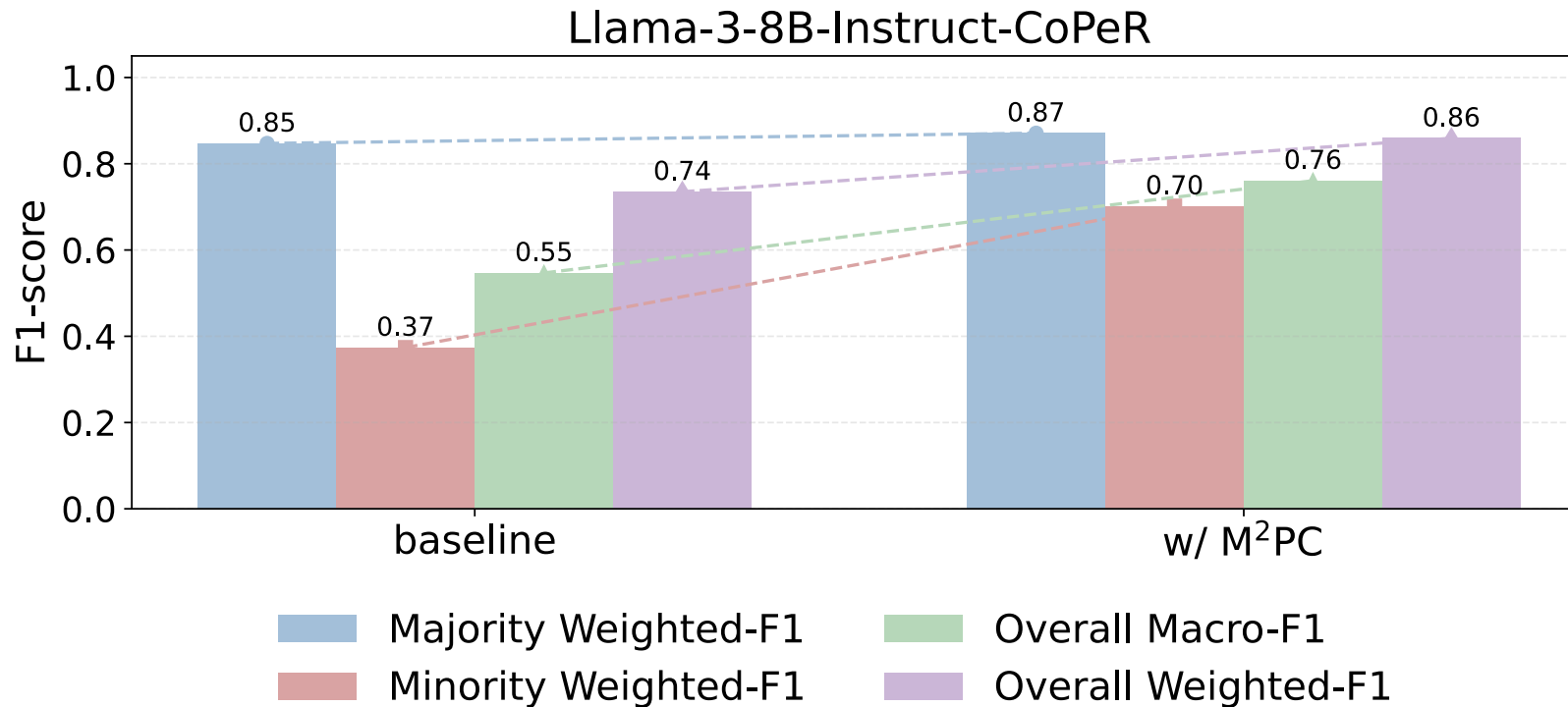
# User-Specific Reasoning Enhances Overall Performance

Systems	$F1_{\text{low}}$	$F1_{\text{high}}$	$F1_{\text{weight}}$	$F1_{\text{macro}}$
Llama-3-8B-Instruct	0.24	0.82	0.71	0.53
+User-specific	0.27	0.86	0.75	0.56

❑ Performance improved on both classes:

Low-F1  $\uparrow$  0.24  $\rightarrow$  0.27 (+13%), High-F1  $\uparrow$  0.82  $\rightarrow$  0.86 (+5%)

# M<sup>2</sup>PC Achieves Substantial Minority Gains with Balanced Performance



□ Significant improvements on the validation set:

- **Minority Weighted-F1** ↑ 0.37 → 0.70 (+89%)
- **Overall Macro-F1** ↑ 0.55 → 0.76 (+38%)



# Preference-Adaptive RL Enhances Minority Predictions

Systems	$F1_{\text{low}}$	$F1_{\text{high}}$	$F1_{\text{weight}}$	$F1_{\text{macro}}$
Llama-3-8B-Instruct	0.24	0.82	0.71	0.53
+User-specific	0.27	0.86	0.75	0.56
RL with PPO	0.22	0.88	0.76	0.55
RL with PAda-PPO	<b>0.36</b>	0.86	<b>0.77</b>	<b>0.61</b>

- ❑ Performances improved on each class:

Low- $F_1 \uparrow 0.24 \rightarrow 0.27$  (+13%), High- $F_1 \uparrow 0.82 \rightarrow 0.86$  (+5%),

- ❑ **RL with PAda-PPO further improves the low-satisfaction class:**

Low- $F1 \uparrow 0.22 \rightarrow 0.36$  (+64%).

# Does Our Method Support Smaller Subgroups?

- ❑ Method: Cluster subgroups by **k-means++** on last hidden states.
- ❑ Optimal is by **silhouette score**.

# Accounts for Smaller Yet Distinct Subgroups

Groups	1	2	3	4	5	6	7	...
Maj.	0.71 (134)	0.79(105)	0.93 (56)	0.85 (48)	0.94(44)	0.94(39)	0.90(30)	...
Min.	0.70(22)	0.61(21)	0.67(7)	0.91(6)	0.67(6)	1.00(5)	0.80(5)	...
Gropus	13	14	15	16	17	18	19	20
Maj.	0.90 (19)	0.96 (14)	1.00 (9)	1.00(8)	1.00(8)	1.00(7)	1.00(4)	-
Min.	0.67(3)	0.53(3)	0.53(3)	0.67(3)	1.00(3)	0.67(2)	1.00(2)	1.00(2)

Note: Each cell shows “weighted-F1 (number of users)”

Majority (17/17)/ Minority (12/18): **Smaller** outperform **largest**  
 → Captures diverse characteristics rather than overfitting to frequent patterns.

# Takeaways

- ❑ We address the often-overlooked preferences of minority users.
- ❑ User satisfaction is inherently subjective; reasoning enables *personalization*.
- ❑ M<sup>2</sup>PC uncovers diverse user clusters, while PAda-PPO *aligns rewards with subgroup preferences*.
- ❑ Our framework achieves significant *improvements for minority users* while preserving majority performance.

# Thank you for your attention!



**Paper**



**Code**



**Contact**

KYOTO UNIVERSITY

京都大学

