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?

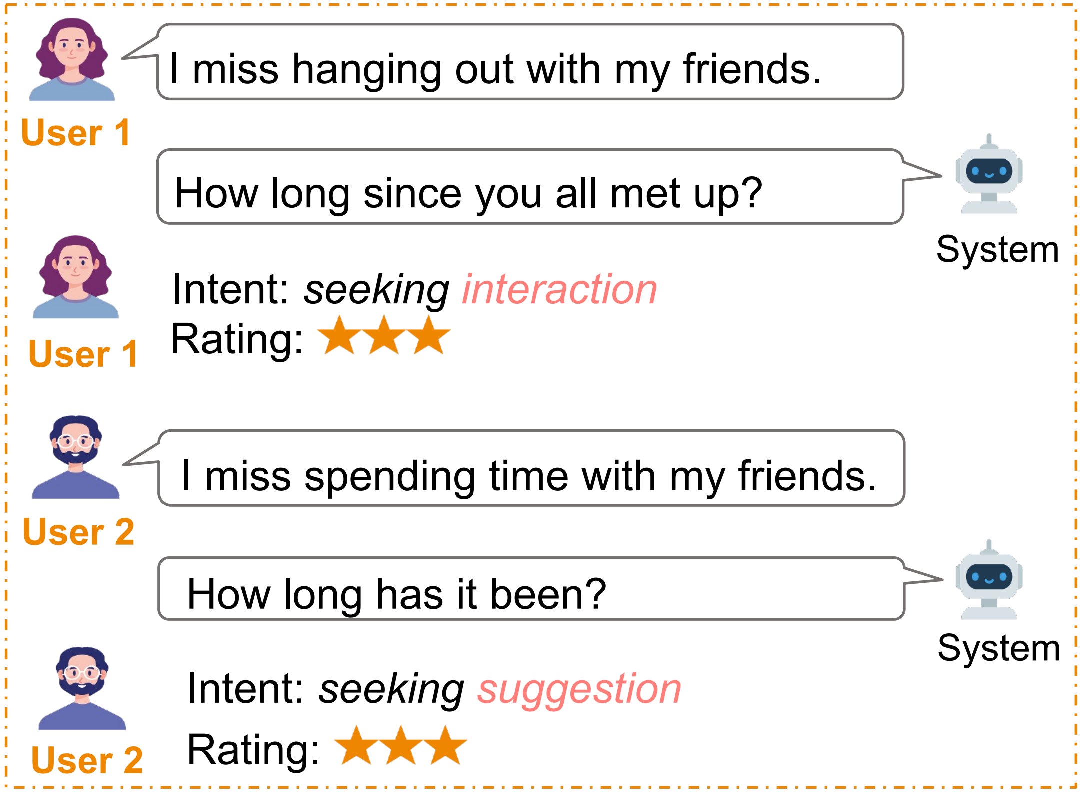
💡 **Motivation:** User satisfaction is subjective 🗣️

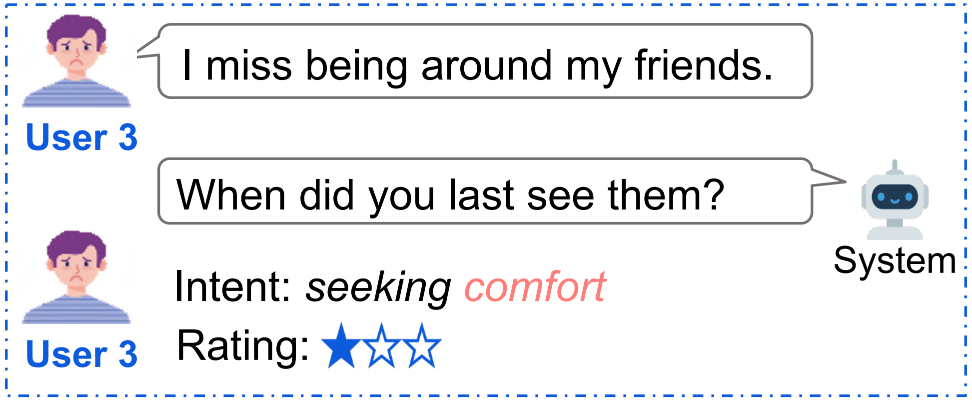
→ **same response strategy ≠ same satisfaction**

Why Model User Satisfaction Beyond the Majority?

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

→ same response strategy ≠ same satisfaction





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

Understanding users' intents

- What do they need?
- How do they feel?

+

Appropriate response strategy

- Questioning, reflecting, suggesting…
- Tailored, not one-size-fits-all

High satisfaction across diverse users

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)

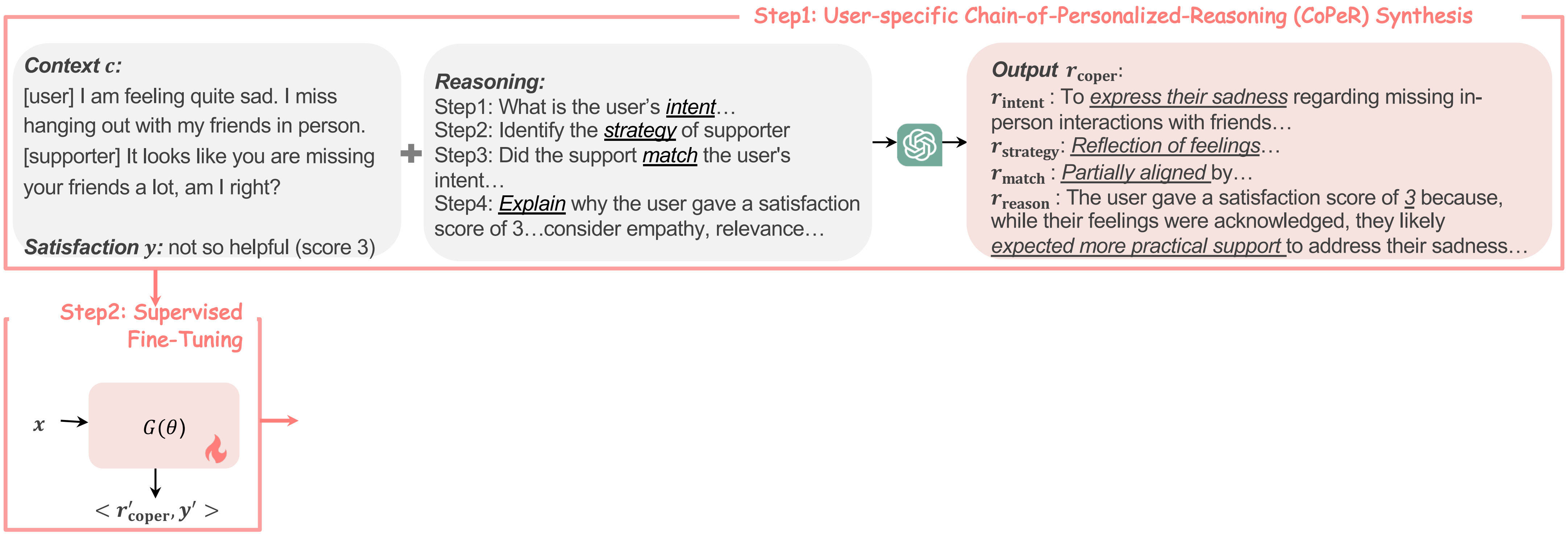
+

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_{coper}:
r_{intent} : To *express their sadness* regarding missing in-person interactions with friends…
r_{strategy}: *Reflection of feelings*…
r_{match} : *Partially aligned* by…
r_{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

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_{coper}:
r_{intent} : To *express their sadness* regarding missing in-person interactions with friends…
r_{strategy}: *Reflection of feelings*…
r_{match} : *Partially aligned* by…
r_{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

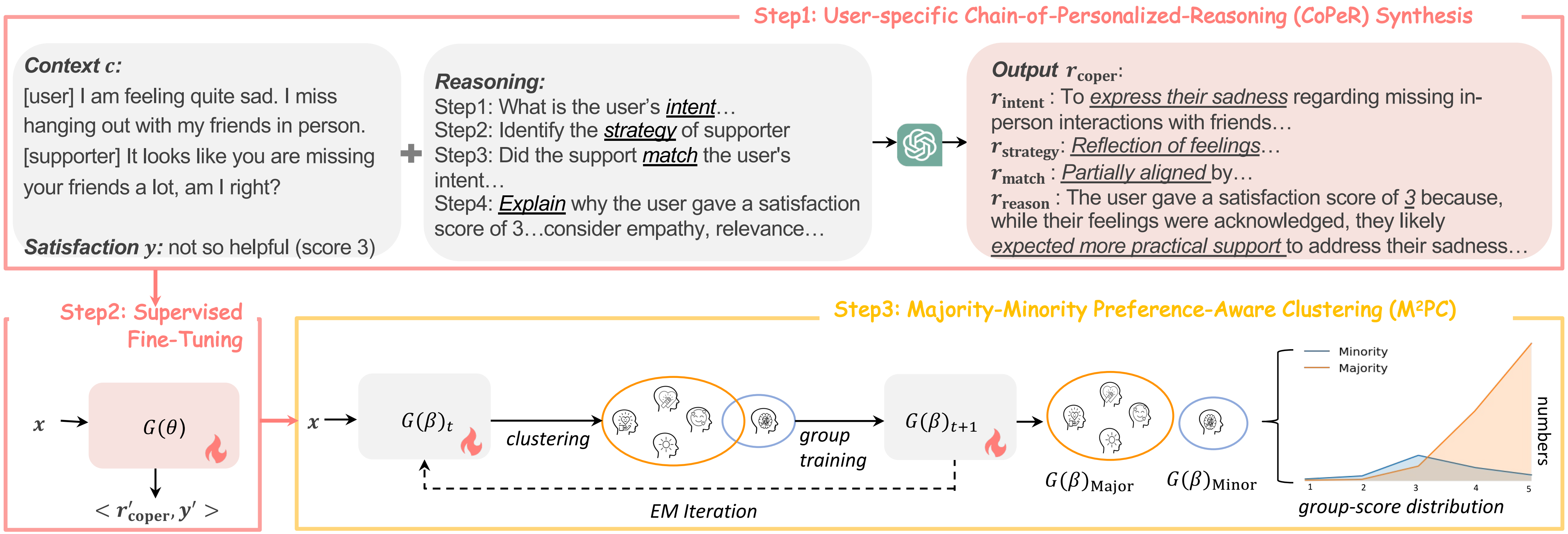
$x \rightarrow G(\theta)$

$< r'_{\text{coper}}, y' >$

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

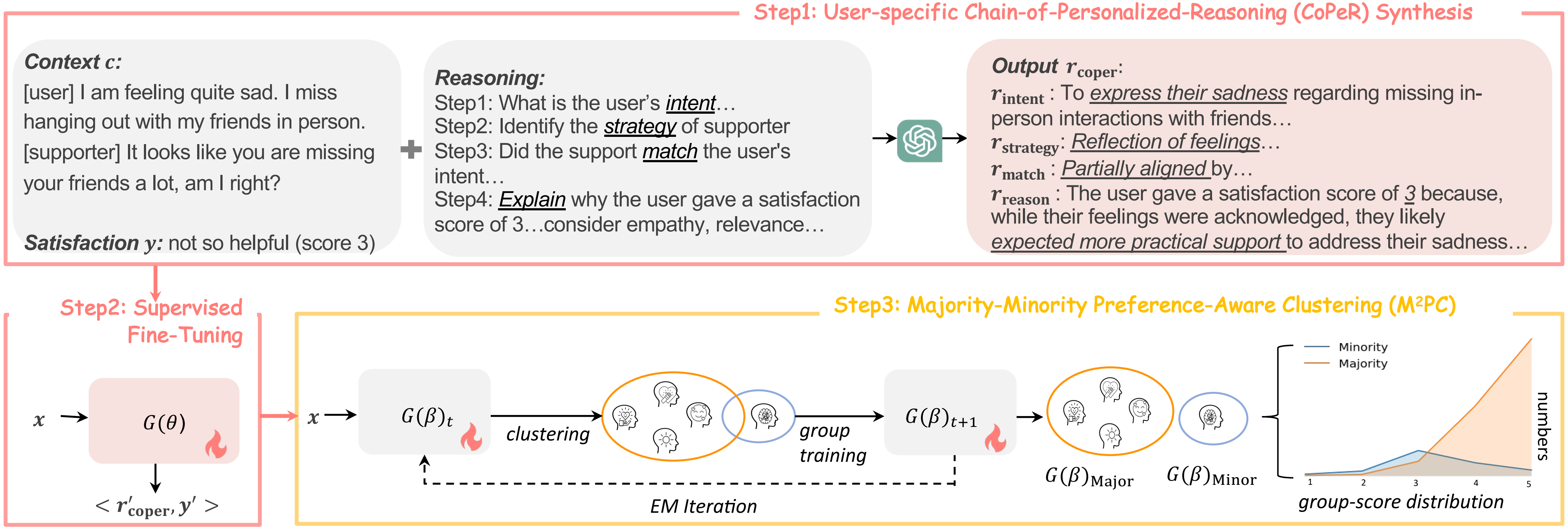
Real systems lack group labels → supervised separate training is **impossible**.

Method: Majority-Minority Preference-Aware Clustering (M²PC)



京都大学

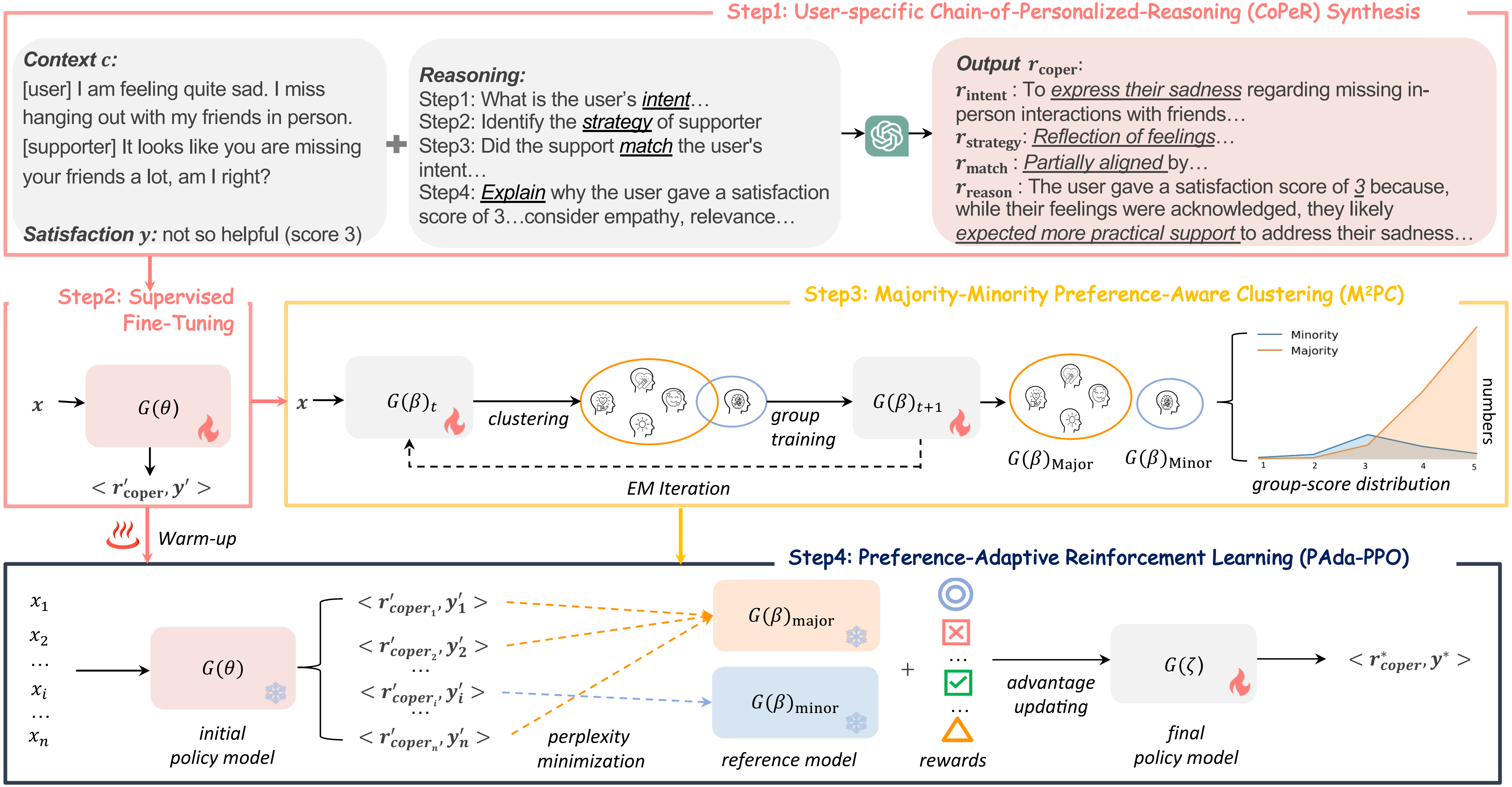
Method: Majority-Minority Preference-Aware Clustering (M²PC)



Reference model inherits majority bias → RL cannot adapt well.

M²PC learns majority/minority references → RL adapts to each group.

Method: Preference-Adaptive Reinforcement Learning



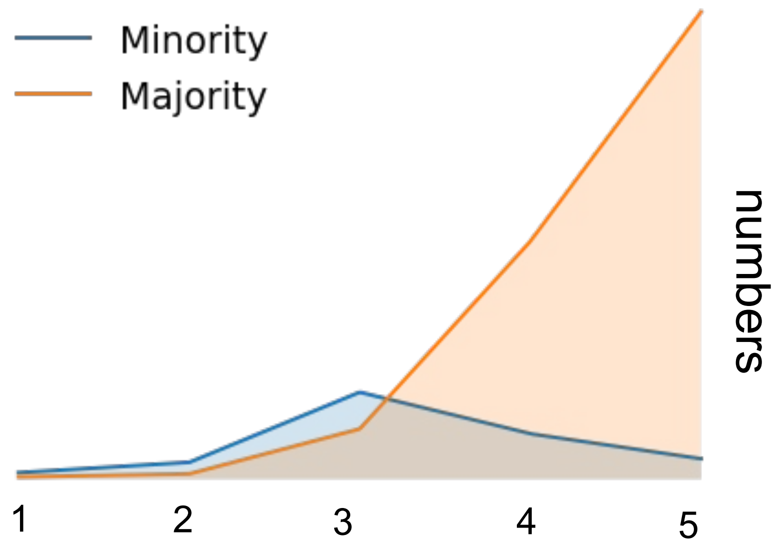
Experiment

❑ Dataset: Emotional Support Conversation

❑ User Groups

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

❑ *Group-Satisfaction score Distribution*



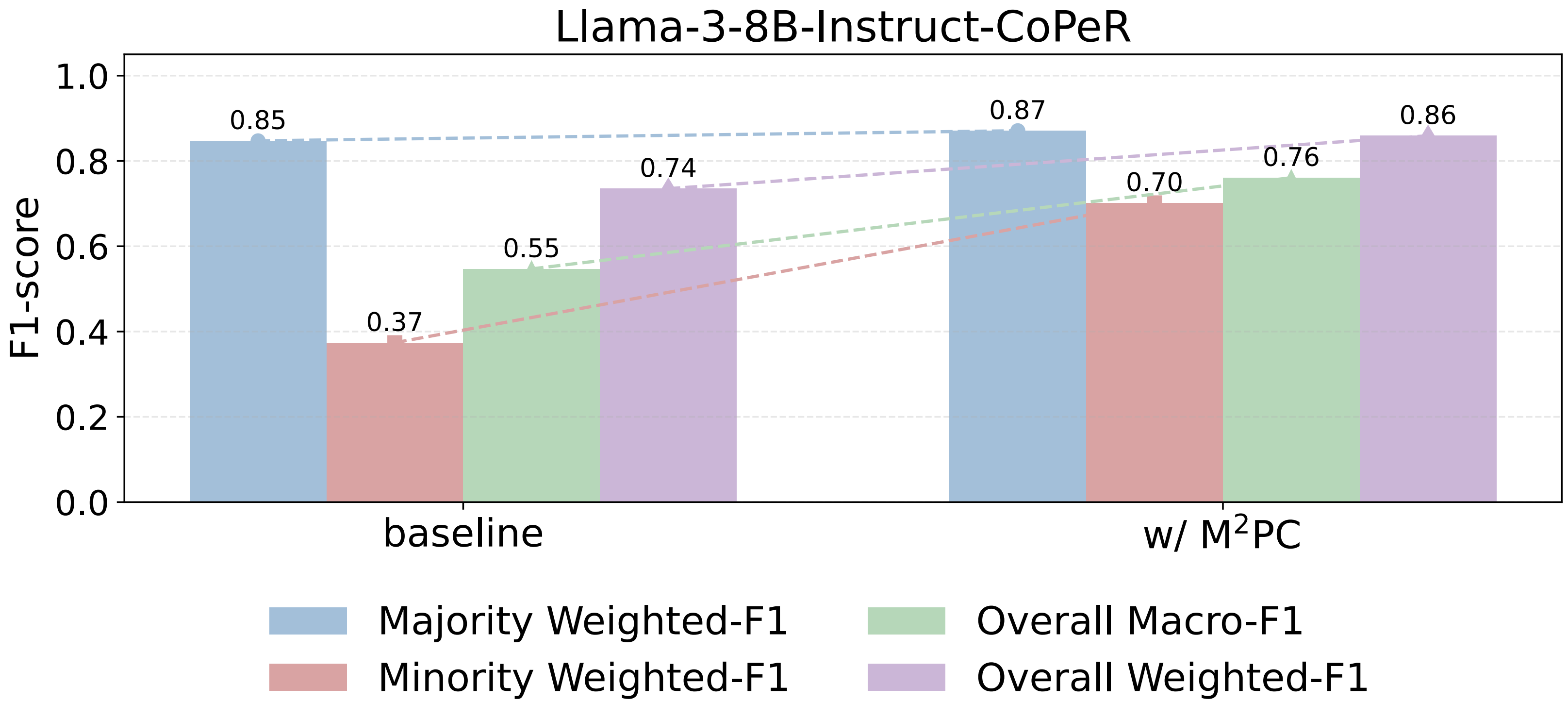
User-Specific Reasoning Enhances Overall Performance

Systems	$F1_{low}$	$F1_{high}$	$F1_{weight}$	$F1_{macro}$
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 ↑ 0.24 → 0.27 (+13%), High-F1 ↑ 0.82 → 0.86 (+5%)

M²PC 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	$F1_{low}$	$F1_{high}$	$F1_{weight}$	$F1_{macro}$
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	0.36	0.86	0.77	0.61

- ❑ Performances improved on each class:
  Low-F_1 ↑ 0.24 → 0.27 (+13%), High-F_1 ↑ 0.82 → 0.86 (+5%),
- ❑ **RL with PAda-PPO further improves the low-satisfaction class:**
  Low-F1↑ 0.22 → 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***.
- M²PC uncovers diverse user clusters, while PAda-PPO ***aligns rewards with subgroup preferences***.
- Our framework achieves significant ***improvements for minority users*** while preserving majority performance.

京都大学

Thank you for your attention!

Paper **Code** **Contact**

京都大学

KYOTO UNIVERSITY