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

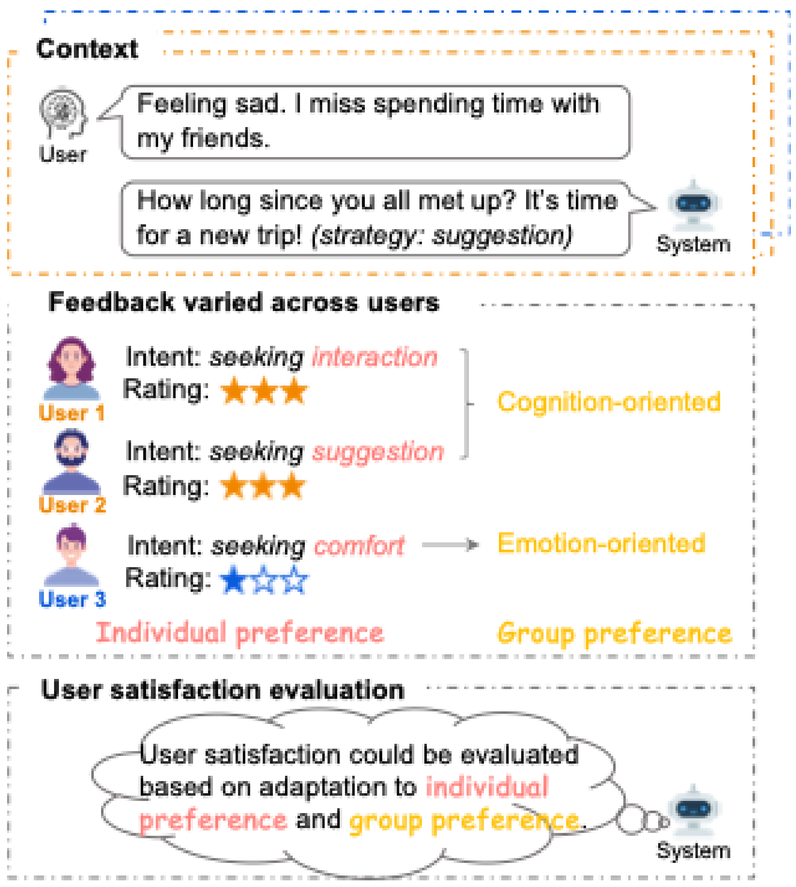
IJCNLP-AACL 2025

Paper Code

Introduction

❑ ***Goal:***
Build a satisfaction estimation model that aligns with both majority and minority preferences for personalized adaptation.

❑ ***Motivation:***
- User satisfaction is subjective and diverse.
- Users in the same group may share similar preferences.



Challenges

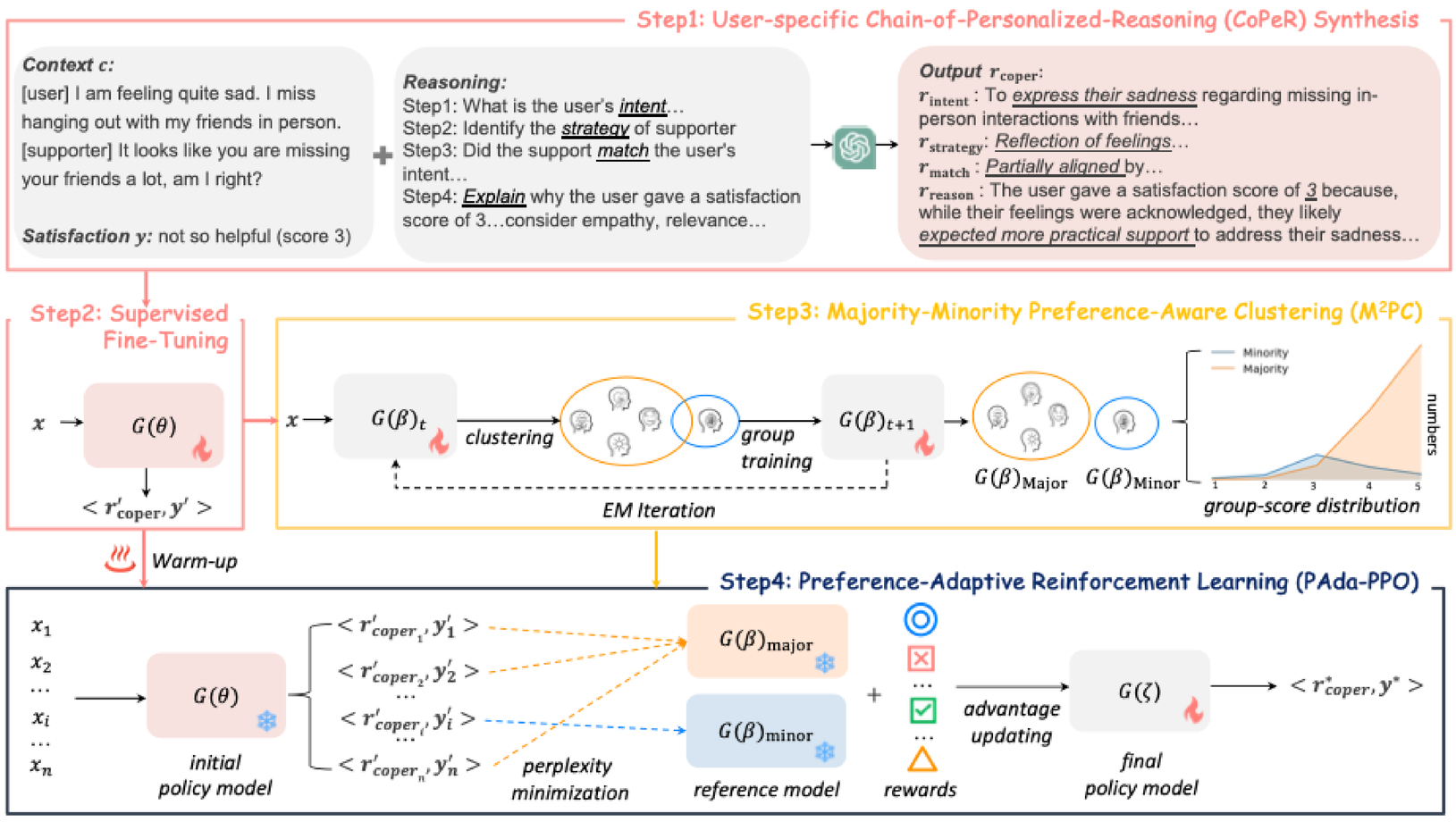
❑ ***Preference Collapse in Reward Models:***
Existing alignment methods often rely on aggregated or majority-voted feedback, which suppresses minority preferences and favors majority trends.

❑ ***Lack of Explicit Preference Labels:***
Real-world dialogue data rarely includes clear majority and minority labels or explicit user rationales behind satisfaction.

Contributions

- **CoPeR:** models individual reasoning (intent → strategy → match → score).
- **M²PC:** EM-based unsupervised grouping by majority/ minority user preference.
- **PAda-PPO:** aligns policy with both individual and group reward signals.

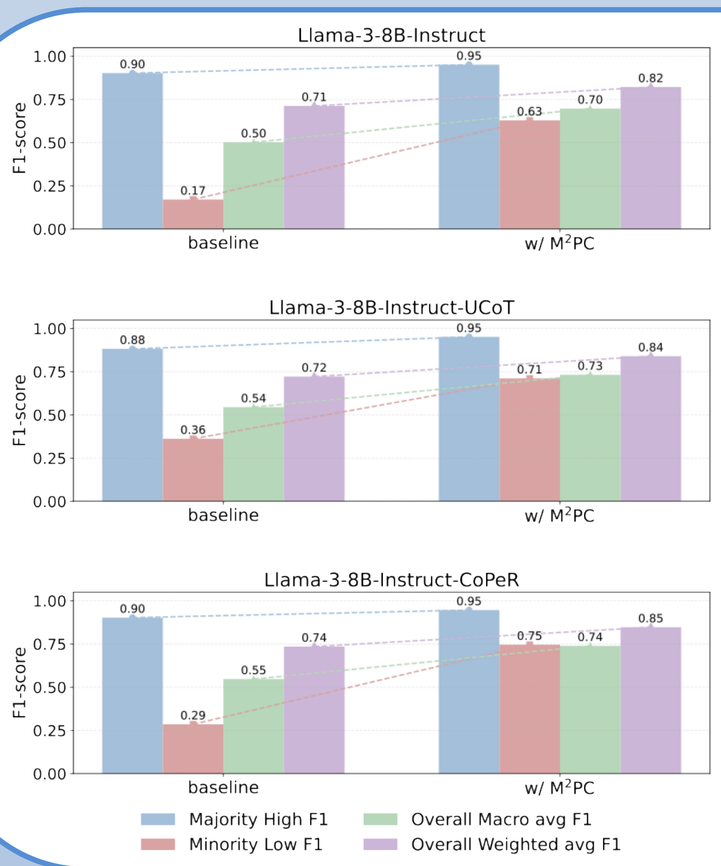
✨ Unified framework improves satisfaction prediction **for both majority and minority populations**.



- Prompt LLMs with **User-specific** Chain-of-Thought (UCoT).
- Synthesize **interpretable rationales** using GPT-4.1-mini.

- Use EM algorithm to separate **majority/minority users** via dialogue perplexity.
- Fine-tune cluster-specific models to capture group trends.

- Reference models = M²PC-trained cluster models.
- Optimize PPO objective with **preference-aware KL regularization**.



ESConv benckmark

Models	F_1^{low}	F_1^{high}	F_1^{w}	F_1^{m}
Llama-3-8B-Instruct	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-Instruct-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-Instruct-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

- CoPeR vs Base:
  Low-F_1 ↑ 0.24 → 0.30 (+25%).
- PAda-PPO vs PPO (UCoT):
  Low-F_1 ↑ 0.22 → 0.36 (+64%).

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 delivers ***substantial gains for minority users*** while preserving majority performance.