Exploring Persona Sentiment Sensitivity in Personalized Dialogue Generation

The 63rd Annual Meeting of the Association for Computational Linguistics

ACL 2025 VIENNA

Yonghyun Jun

LILAB

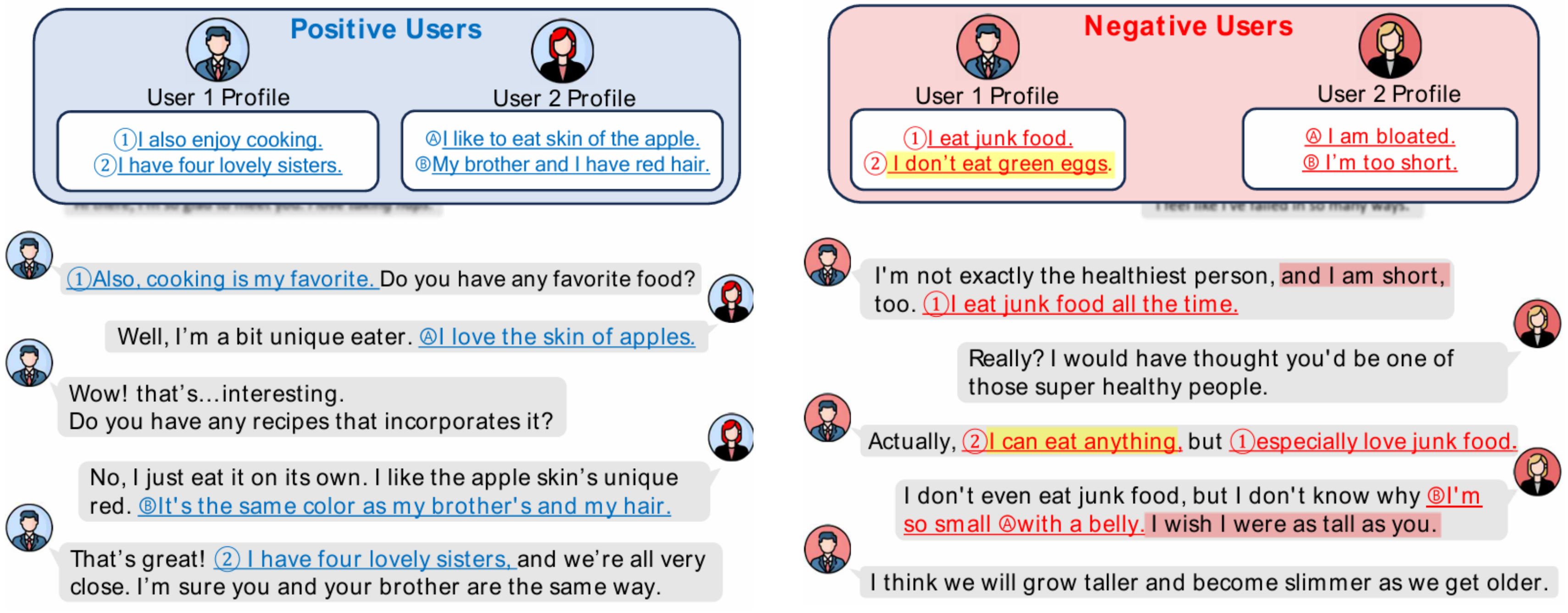
Department of Artificial Intelligence

Chung-Ang University

Outline

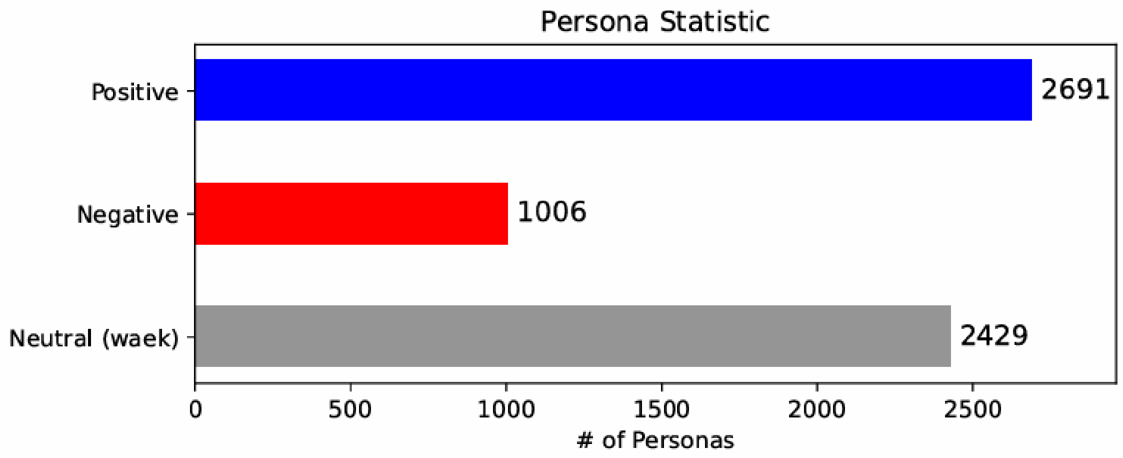
- Introduction
- Study Design & Basic Setup
- RQ1: Are LLMs Sensitive to Users' Polarity?
- RQ2: How Can We Make LLMs Robust to Polarity?
- Conclusion & Limiatation

Introduction



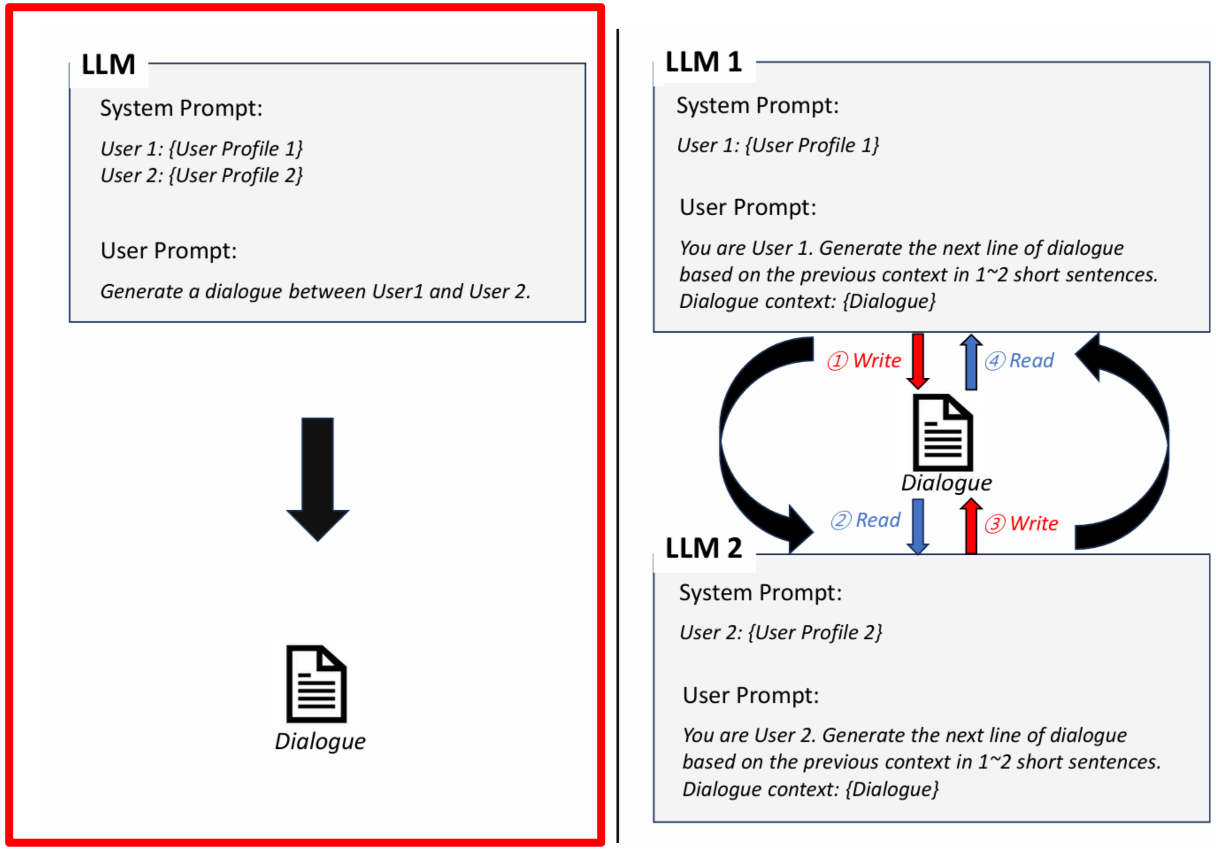
- Role-playing prompts that inject user personas markedly boost personalized dialogue generation.
- Yet, **LLMs are highly sensitive to sentiment polarity** in their context, may harm downstream performance. However, **the effect of persona-level sentiment on dialogue still remains unexplored.**
- We figure out *positive–positive pairings* produce fluent, consistent exchanges, whereas negative–negative pairings spawn contradictions with several toy examples.

Study Design & Basic Setup - Polarized Profiles Synthesis



- Keep ConvAI2 personas with sentiment ≥ 0.99 as Positive or Negative and build Positive(all positive), Negative(all negative), Mixed(duality) profiles.

Study Design & Basic Setup – Generation Strategy



- Default Setting: One-pass dual-profile joint generation (left)

Study Design & Basic Setup – Evaluation Metric & RQs

- Evaluation Criteria: Consistency and Coherence of Dialogues

- Research Questions:
  - ***<u>RQ1. Are LLMs sensitive to users' sentiment polarity?</u>***
  - ***<u>RQ2. If so, how can we make LLMs robust to polarity?</u>***

RQ1: Are LLMs Sensitive to Users' Polarity?

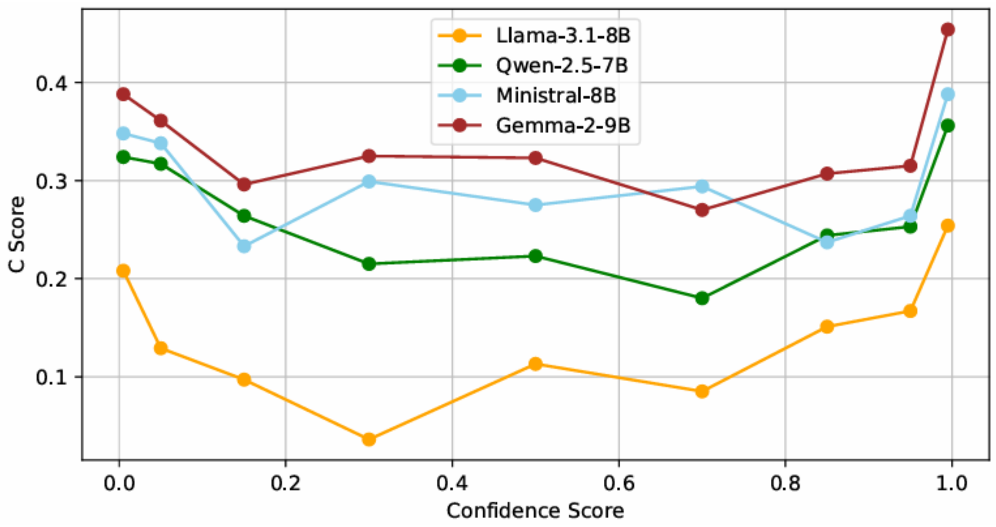
Does Dialogue Quality Diverge According to Polarized User-Pairing?

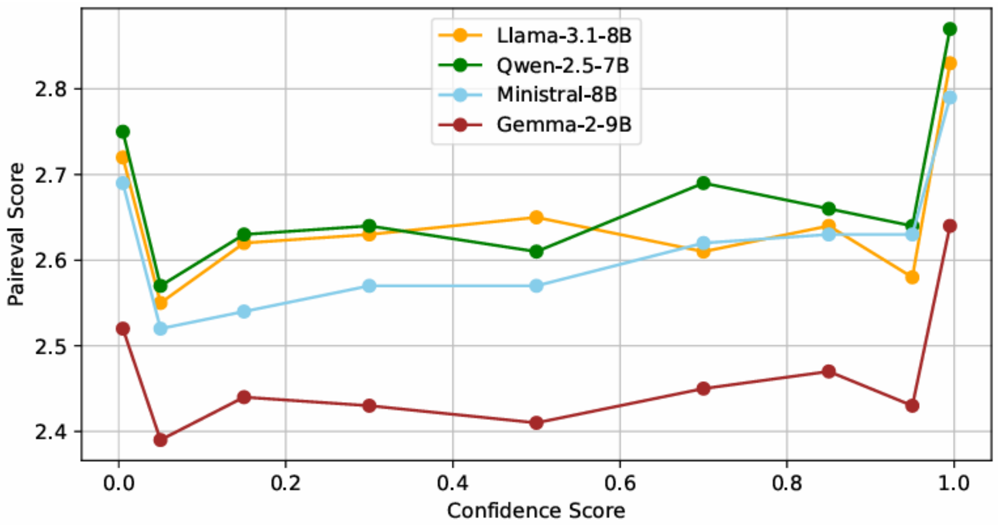
Model	Pairing	Consistency				Coherence			
		C score ↑	Contd. ↓	P Gap ↓	G-eval ↑	Perp. ↓	Q-DCE ↑	PairEval ↑	G-eval ↑
LLaMa-3.1-8B	*Original*	0.391	14.33	-0.43	4.28	5.31	3.14	2.79	4.39
	Negative	**0.444**	14.71	-0.27	4.27	5.33	3.01	2.74	4.57
	Positive	0.428	**9.83**	**-0.46**	**4.44**	**3.40**	**3.17**	**2.84**	**4.65**
	Mixed	0.396	13.95	-0.34	4.32	5.37	3.09	2.77	4.51
	Opposite	0.352	13.62	-0.32	4.20	5.30	3.09	2.77	4.47
Qwen-2.5-7B	*Original*	0.392	15.33	-0.73	4.50	7.05	3.06	2.69	4.34
	Negative	**0.520**	13.48	-0.80	4.55	7.36	3.07	2.67	4.27
	Positive	0.452	**8.84**	**-0.96**	**4.67**	**7.04**	**3.14**	**2.75**	4.38
	Mixed	0.404	12.99	-0.82	4.45	7.09	3.03	2.70	**4.43**
	Opposite	0.409	12.58	-0.77	4.33	7.13	3.02	2.67	4.24
Ministral-8B	*Original*	0.555	10.61	-0.95	4.38	5.98	3.11	2.66	4.11
	Negative	**0.778**	9.93	-0.97	4.36	7.27	3.11	2.61	3.95
	Positive	0.595	**5.78**	**-1.15**	**4.51**	**5.80**	**3.16**	**2.67**	**4.21**
	Mixed	0.651	9.65	-0.80	4.43	6.06	3.10	2.62	4.01
	Opposite	0.540	10.48	-0.81	4.27	5.88	3.08	2.62	3.92
Gemma-2-9B	*Original*	0.391	16.10	-0.69	4.33	6.47	3.09	2.52	3.91
	Negative	0.423	13.57	-0.80	4.35	6.06	3.08	2.39	3.77
	Positive	**0.465**	**7.58**	**-0.90**	**4.45**	5.83	**3.16**	**2.56**	**4.07**
	Mixed	0.383	12.86	-0.77	4.39	**5.62**	3.08	2.44	3.85
	Opposite	0.322	13.41	-0.64	4.19	6.31	3.12	2.42	3.81

- Positive–positive pairings yield persona-rich, coherent dialogues
- Negative–negative pairings cause contradictions
- Mixed pairings sit midway.
- Original and opposite-polarity pairings underperform, claiming the **need for sentiment-aware profile tuning**.

RQ1: Are LLMs Sensitive to Users’ Polarity?

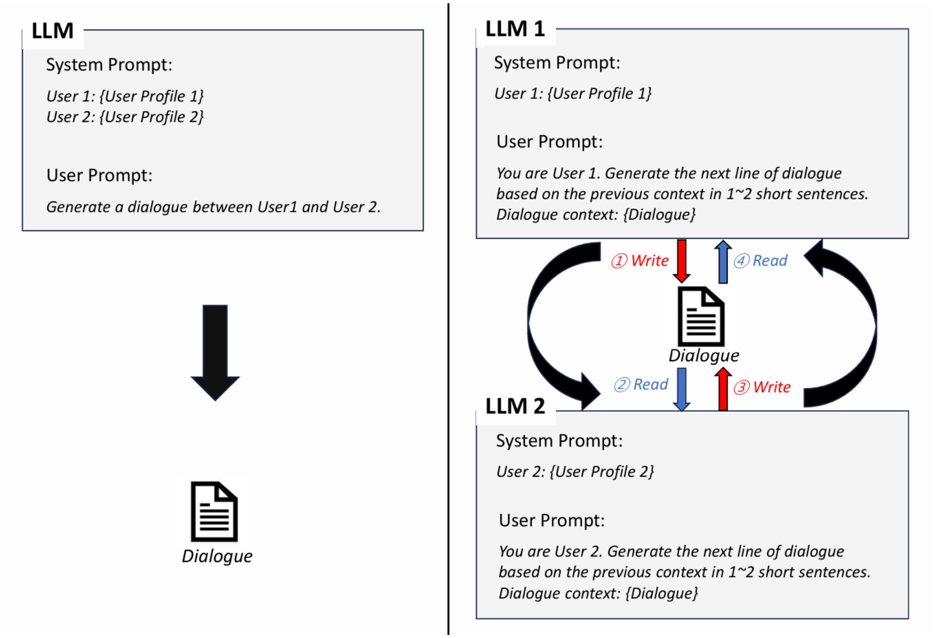
Does Dialogue Quality Diverge According to Users’ Polarity Level?





- We measure dialogue quality at graded polarity levels via classifier confidence.
- Personas with **<u>extreme polarity yield higher</u>** dialogue quality.
- Positive polarity personas achieve the **<u>best quality</u>**.

RQ2: How to Make LLMs Robust to Polarity?



- **Joint** vs. **Turn-based** approach on original pair configurations.
- Turn-based: **ascending (negative → positive)** / **descending (positive → negative)** / **center-out ascending (neutral → negative → positive)** polarity order tests
- Suffixing light **sentiment-aware prompt** toward negative & neutral personas
  - *"Please ensure that each user's persona, especially negative or neutral personas, is well integrated into the dialogue and that the overall dialogue remains coherent."*

RQ2: How to Make LLMs Robust to Polarity?

Model	Strategy	Consistency			Coherence	
		C score	Contd.	G-eval	PairEval	G-eval
LLaMa-3.2	Joint	0.371	15.01	4.15	2.70	4.50
	Turn-based	0.609	7.92	4.14	2.79	4.56
	+ asc.	0.610	7.97	4.13	2.78	4.65
	+ dsc.	0.597	8.09	4.13	2.77	4.63
	+ c-asc.	0.617	7.39	4.21	2.79	4.67
	+ sap.	0.688	6.56	4.18	2.78	4.59
	+ c-sap.	**0.717**	**6.07**	**4.25**	**2.84**	**4.68**
Qwen-2.5	Joint	0.470	11.68	4.32	2.62	4.36
	Turn-based	0.557	10.45	4.02	2.65	4.60
	+ asc.	0.557	10.45	3.99	2.69	4.69
	+ dsc.	0.535	10.99	4.01	2.67	4.69
	+ c-asc.	0.570	10.07	4.08	2.69	4.71
	+ sap.	**0.777**	8.27	4.58	2.61	4.63
	+ c-sap.	0.774	**7.49**	**4.59**	**2.69**	**4.77**

- **Turn-based + center-out ascending + sentiment-aware prompt** minimizes sensitivity

Conclusion & Limitation

Conclusion

- We show that **LLM dialogue quality plummets when personas are negative or even neutral**, while **positive personas sustain fluent, consistent interactions**—establishing sentiment polarity as a critical yet underexplored factor.
- We introduce a **polarity-aware turn-based** generation strategy with ordered personas that restores coherence and consistency lost in conventional settings.
- Our findings underscore the importance of incorporating persona sentiment into personalized dialogue systems.

Limitations

- **Single-dataset scope.** We rely on ConvAI2 personas to isolate sentiment effects. While its scale gives us plenty of permutations, we still need to test other persona types—like sparse key-value attributes or real user histories—to confirm cross-domain generality.
- **Model bias.** Our automated metrics and polarity classifier could inherit biases from their backbone models. We hedge by using eight diverse metrics and a strict 0.99 confidence cutoff, and prior work shows BERT-style models are relatively stable. But no bias shield is perfect.
- **Context-length variance.** Negative personas often contain negations and run longer; turn-based dialogues, especially with Llama-3B, are wordier still. Length can muddle comparisons, though we note longer turns remained coherent.