ALIGNING TO THOUSANDS OF PREFERENCES VIA SYSTEM MESSAGE GENERALIZATION

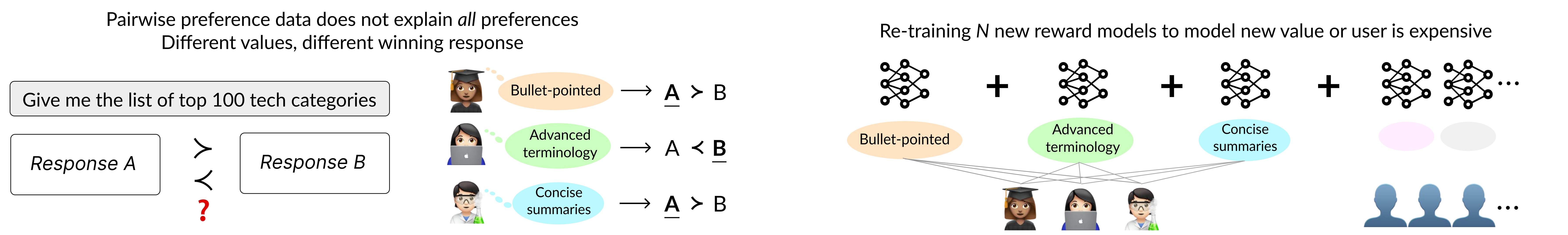
Paper

Project page (w/ code, data, models)

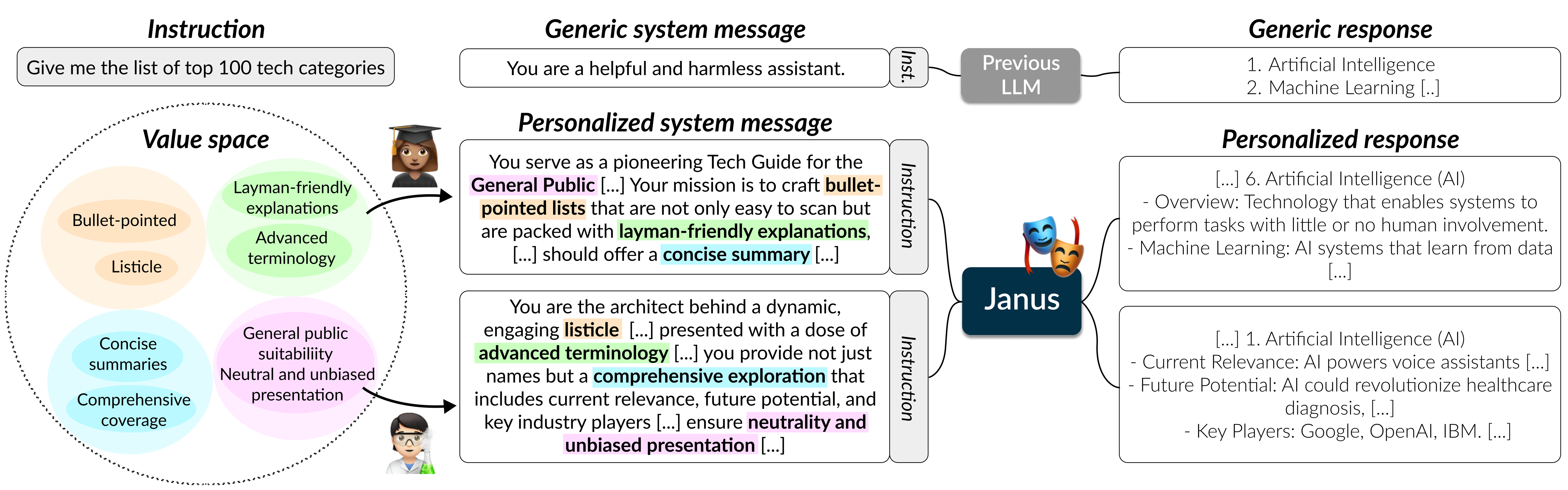
Seongyun Lee[1]*, Sue Hyun Park[1]*, Seungone Kim[12], Minjoon Seo[1]

*Equal contribution [1] KAIST AI Kim Jaechul Graduate School [2] Carnegie Mellon University

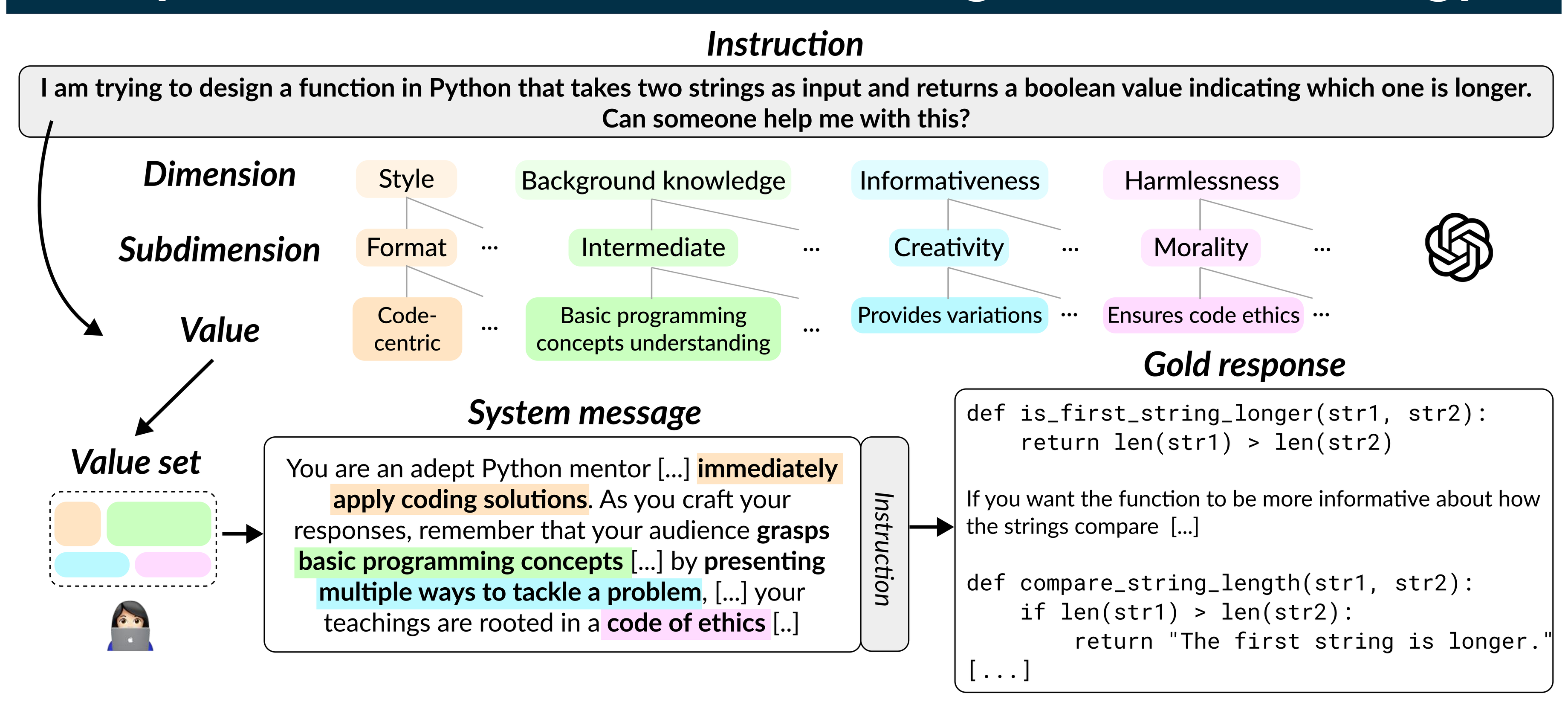
A need for individualized and scalable alignment



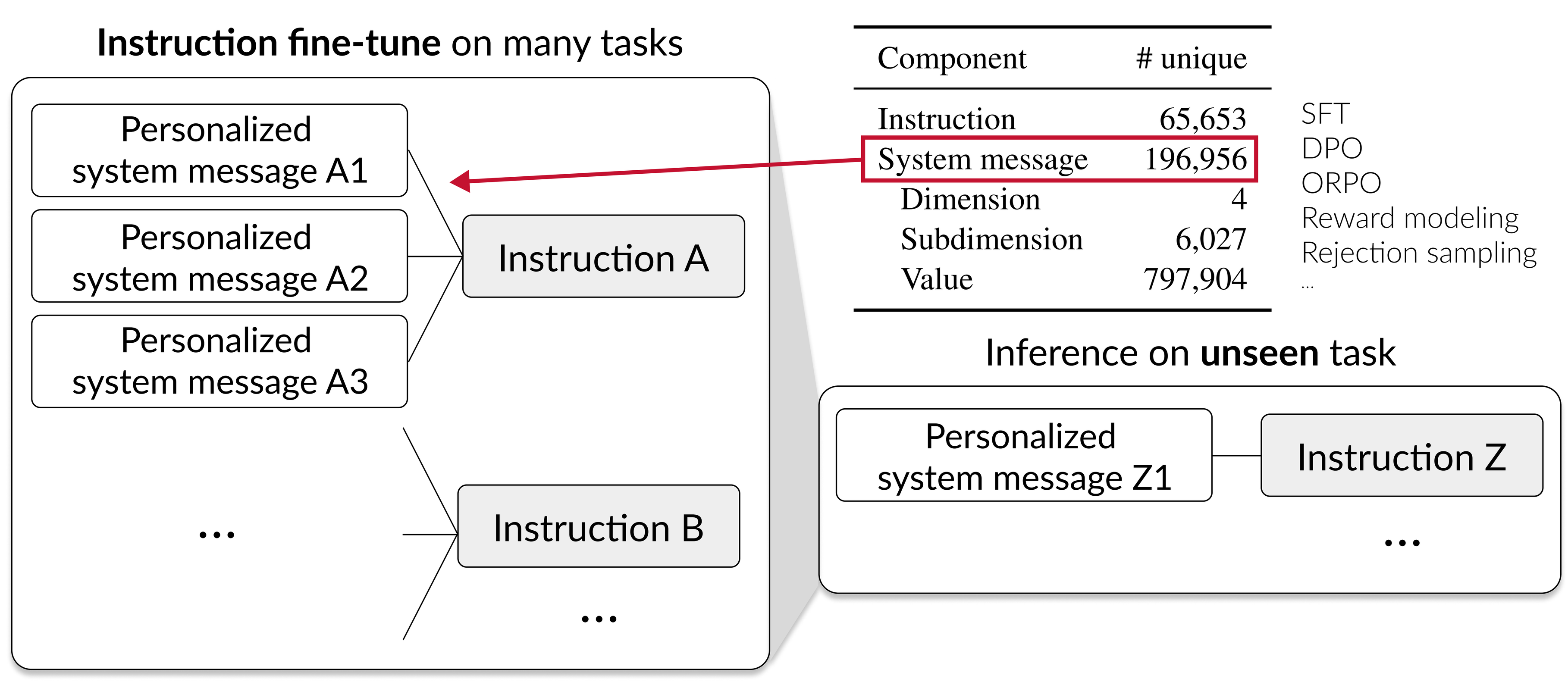
Verbalize values in the system message to flexibly steer toward personalized responses



Key factor 1: Hierarchical value augmentation strategy

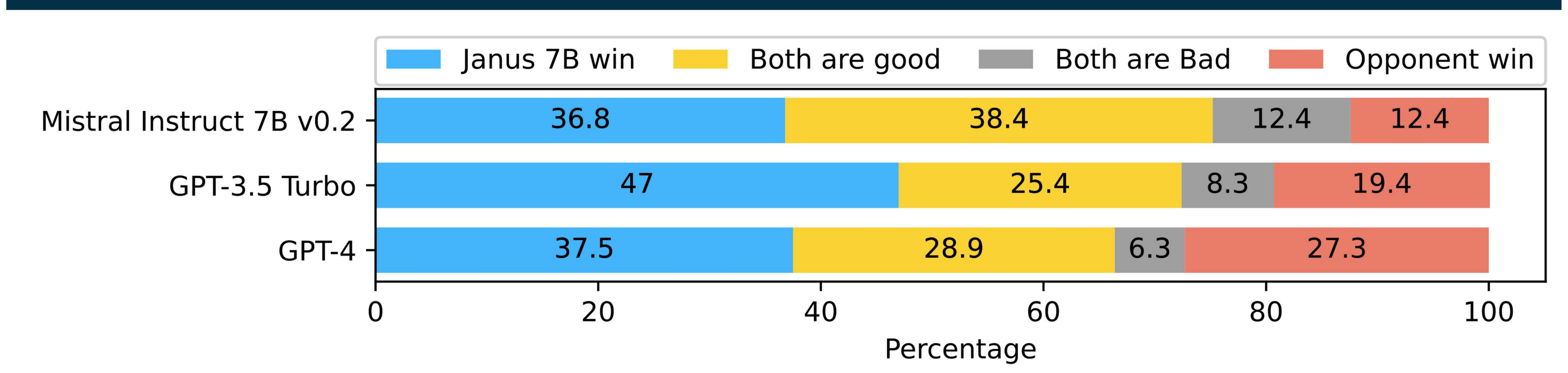


Key factor 2: Training recipe for stronger generalization



Component	# unique
Instruction	65,653
System message	196,956
Dimension	4
Subdimension	6,027
Value	797,904

Aligns to unseen multifaceted values in system messages ✅



Model	*mf*-AlpacaEval	*mf*-FLASK	*mf*-Koala	*mf*-MT-Bench	*mf*-Self-Instruct	Average
Pretrained open models						
Mistral 7B v0.2	2.80	1.93	2.45	2.30	2.28	2.23
LLaMA 3 8B	2.60	2.92	2.69	2.39	2.34	2.54
LLaMA 3 70B	**3.76**	**3.23**	**3.67**	**3.50**	**3.65**	**3.49**
Instruction-tuned open models						
LLaMA 2 Chat 70B	3.98	3.68	4.11	3.66	3.87	3.79
Mistral 7B Instruct v0.2	4.20	3.82	4.18	3.82	3.98	3.93
Mixtral 8x7B Instruct v0.1	4.24	3.90	4.16	3.94	4.08	4.03
LLaMA 3 Instruct 8B	4.38	3.88	4.33	4.08	4.17	4.10
LLaMA 3 Instruct 70B	**4.55**	**4.26**	**4.59**	**4.42**	**4.45**	**4.39**
JANUS suite						
JANUS 7B	4.43	4.06	4.41	4.11	4.01	4.17
JANUS+ORPO 7B	4.41	4.03	**4.45**	4.00	**4.22**	4.18
JANUS+DPO 7B	**4.45**	**4.13**	4.43	**4.21**	4.17	**4.24**
Preference-optimized proprietary models						
GPT-3.5 Turbo-0125	4.05	3.86	4.15	3.87	3.85	3.91
GPT-4-0613	4.25	4.00	4.18	4.16	4.13	4.10
GPT-4-Turbo-0125	**4.45**	**4.27**	**4.61**	**4.45**	**4.27**	**4.35**

Aligns to general public preferences ✅

Size	Models	AlpacaEval 2.0		MT-Bench	Arena Hard Auto v0.1
		LC Win Rate (%)	Win Rate (%)	Score [0,10]	Score [0,100]
...					
< 30B	Mistral 7B Instruct v0.2	17.1	14.7	7.2	10.8
	Gemma 7B Instruct	10.4	6.9	6.4	7.5
	LLaMA 3 8B Instruct	22.9	22.6	7.6	17.9
	JANUS 7B	**26.9**	**27.8**	**7.7**	**20.9**

Additional analyses and insights

- Significant **toxicity** ↓ **fluency** ↑ **diversity** ↑ in RealToxicityPrompts
- Demonstrates **robust** performance with or without personalized input.
- **Learning to handle multifacetedness** in input and/or output is beneficial.
- Verification of quality, diversity, safety, and bias in Appendix and TBA!

Takeaways

- **Clarifying user values behind the preference** in the input can reach diverse alignment targets. **Varying the system message** can provide strong guidance.
- Fine-tuning on **Multifaceted Collection**, an instruction dataset containing 197k system messages can facilitate individualized, scalable value alignment.
- **Janus 7B** models are easily steerable towards user-preferred responses while being generally useful and safe too.