Aligning to Thousands of Preferences via System Message Generalization

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

[1]KAIST AI, [2]Carnegie Mellon University

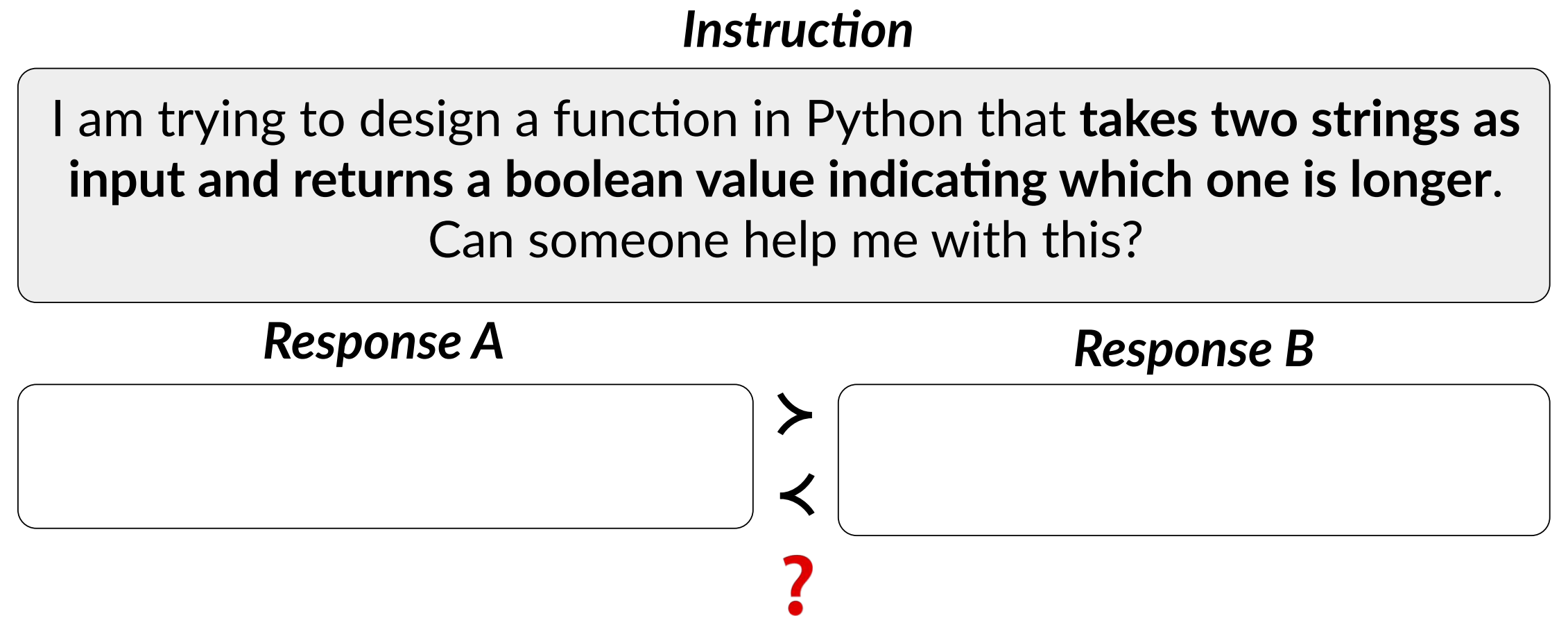
*Equal Contribution

Paper

Project page (w/ code, data, models)

Motivation

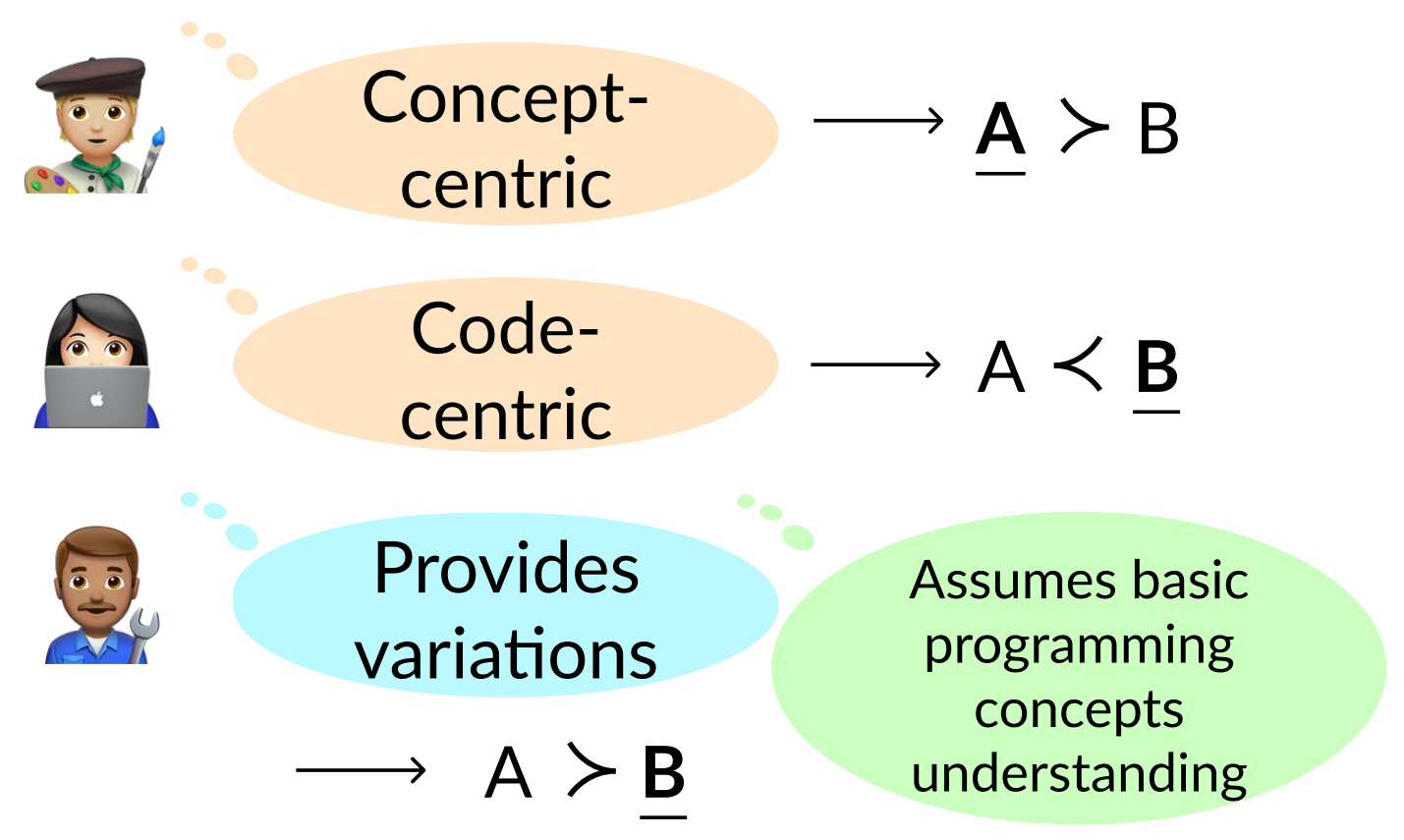
Ambiguous Pairwise Preferences



Instruction

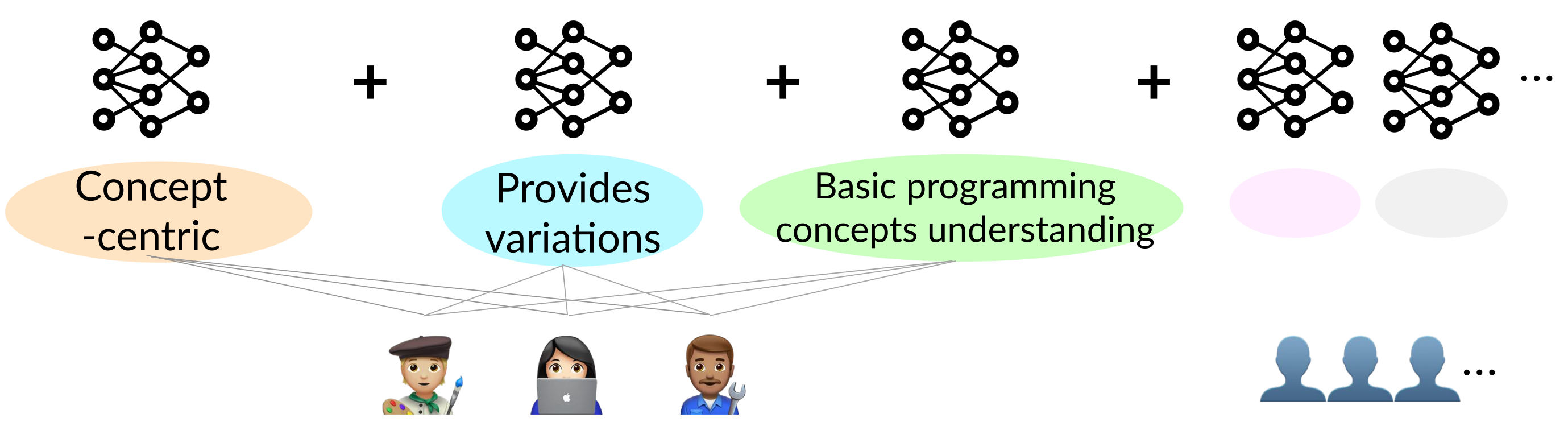
I am trying to design a function in Python that **takes two strings as input and returns a boolean value indicating which one is longer**. Can someone help me with this?

- Pairwise preference data does not explain *all* preferences.
- Which response to choose or reject may differ by people due to individual values.



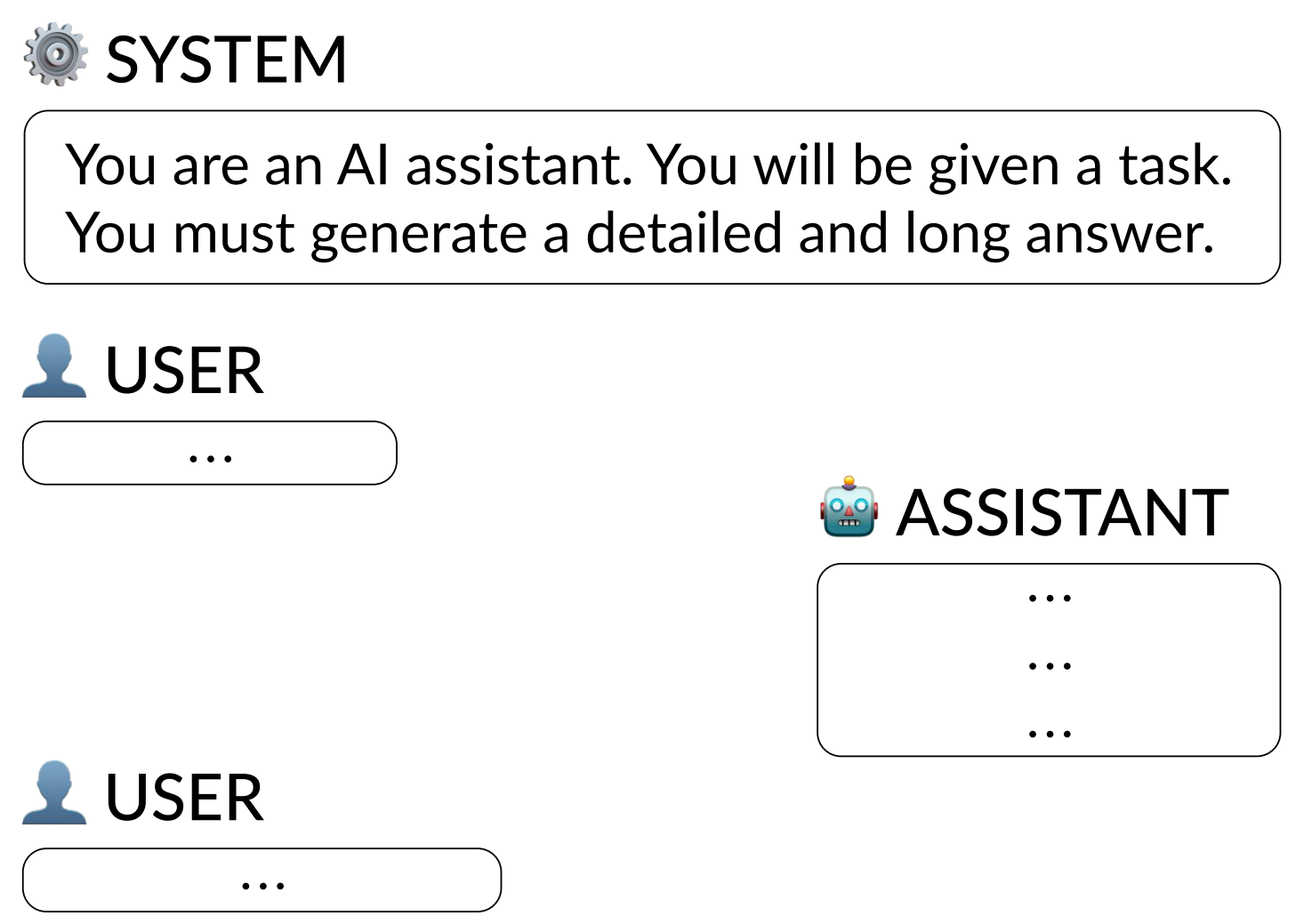
Limited Scalability of Alignment

Re-training *N* new reward models to model new value or user is expensive.



System Messages in LLMs

- Set by developer to instill specific behavior when performing a task (e.g., constraints, personas, tools)
- Lack of diversity and scale in instructing response behavior in previous studies [1, 2]



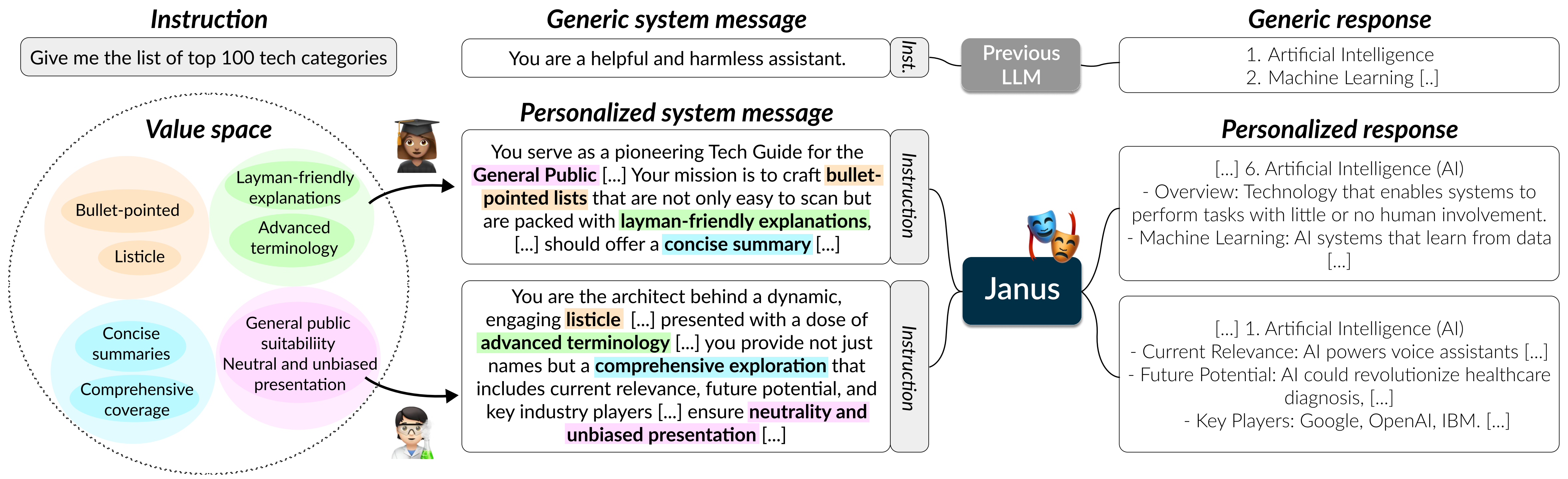
SYSTEM

You are an AI assistant. You will be given a task. You must generate a detailed and long answer.

[1] Llama 2: Open Foundation and Fine-Tuned Chat Models (Touvron et al., arXiv preprint 2023)
[2] Orca: Progressive Learning from Complex Explanation Traces of GPT-4 (Mukherjee et al., arXiv preprint 2023)

- We propose contextualizing user values in system messages to align LLMs to diverse preferences in a scalable way.
- Using our data curation strategy, the trained LLM effectively adapts to unseen preferences and is also generally helpful and harmless.

System Message Generalization



Instruction

Give me the list of top 100 tech categories

Generic system message

You are a helpful and harmless assistant.

Generic response

1. Artificial Intelligence
2. Machine Learning [..]

Value space

Personalized system message

You serve as a pioneering Tech Guide for the **General Public** [...] Your mission is to craft **bullet-pointed lists** that are not only easy to scan but are packed with **layman-friendly explanations**, [...] should offer a **concise summary** [...]

You are the architect behind a dynamic, engaging **listicle** [...] presented with a dose of **advanced terminology** [...] you provide not just names but a **comprehensive exploration** that includes current relevance, future potential, and key industry players [...] ensure **neutrality and unbiased presentation** [...]

Personalized response

[...] 6. Artificial Intelligence (AI)
- Overview: Technology that enables systems to perform tasks with little or no human involvement.
- Machine Learning: AI systems that learn from data [...]

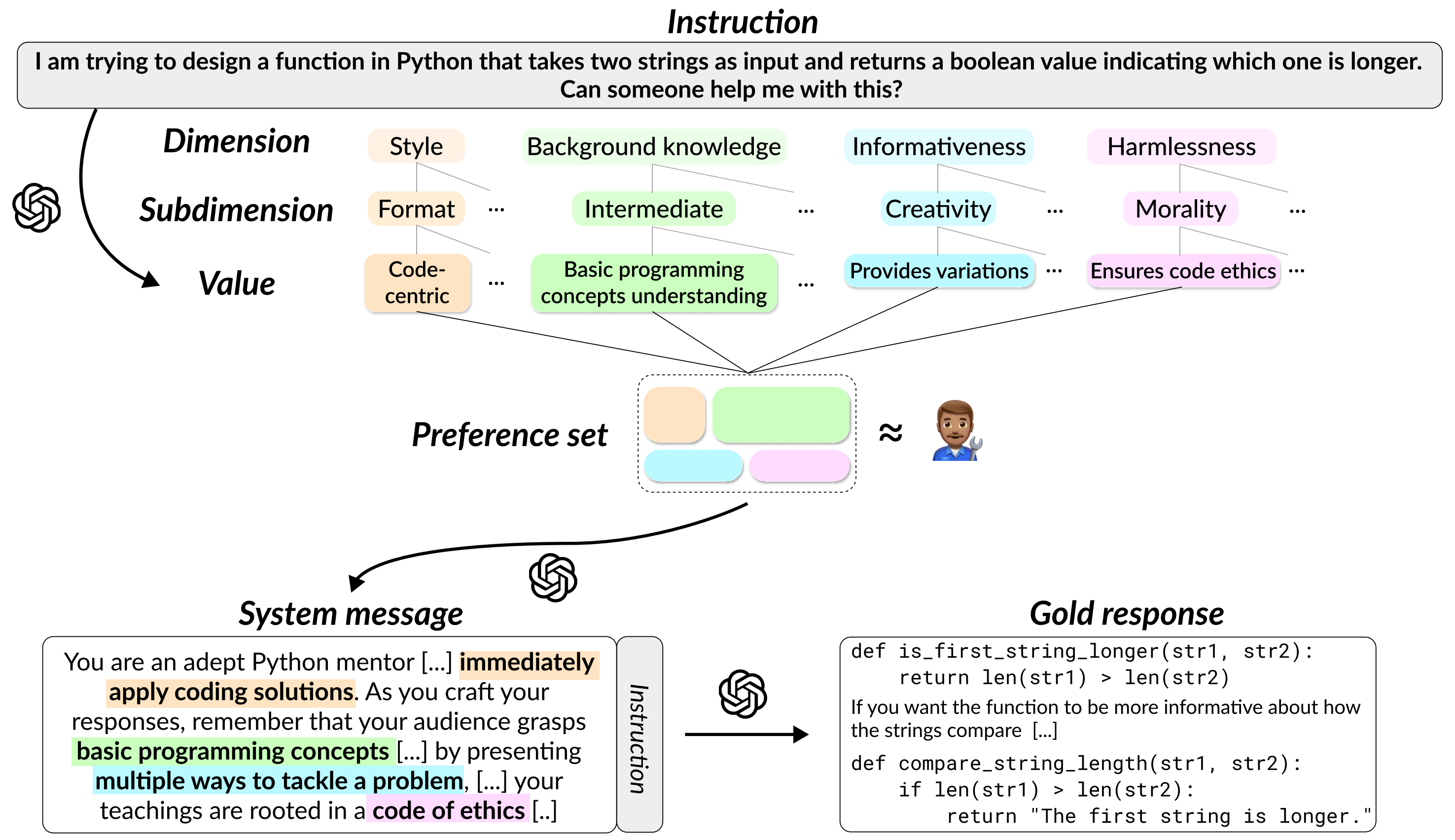
[...] 1. Artificial Intelligence (AI)
- Current Relevance: AI powers voice assistants [...]
- Future Potential: AI could revolutionize healthcare diagnosis, [...]
- Key Players: Google, OpenAI, IBM. [...]

How preferences should be shaped for individualized alignment:

- **Multifaceted**: Capture multiple facets in a user preference.
- **Explicit**: Expose preferences latent in chosen-rejected pairs.

- **Hierarchically augment** user values
- **Verbalize** combinations of values in the *system message, i.e., meta-instruction* that guides a model how to respond to subsequent instructions

Curating Multifaceted System Messages



Instruction

I am trying to design a function in Python that takes two strings as input and returns a boolean value indicating which one is longer. Can someone help me with this?

System message

You are an adept Python mentor [...] **immediately apply coding solutions**. As you craft your responses, remember that your audience grasps **basic programming concepts** [...] by presenting **multiple ways to tackle a problem**, [...] your teachings are rooted in a **code of ethics** [..]

Gold response

```
def is_first_string_longer(str1, str2):
    return len(str1) > len(str2)
```

If you want the function to be more informative about how the strings compare [...]

```
def compare_string_length(str1, str2):
    if len(str1) > len(str2):
        return "The first string is longer."
```

Scaling Data for Generalization

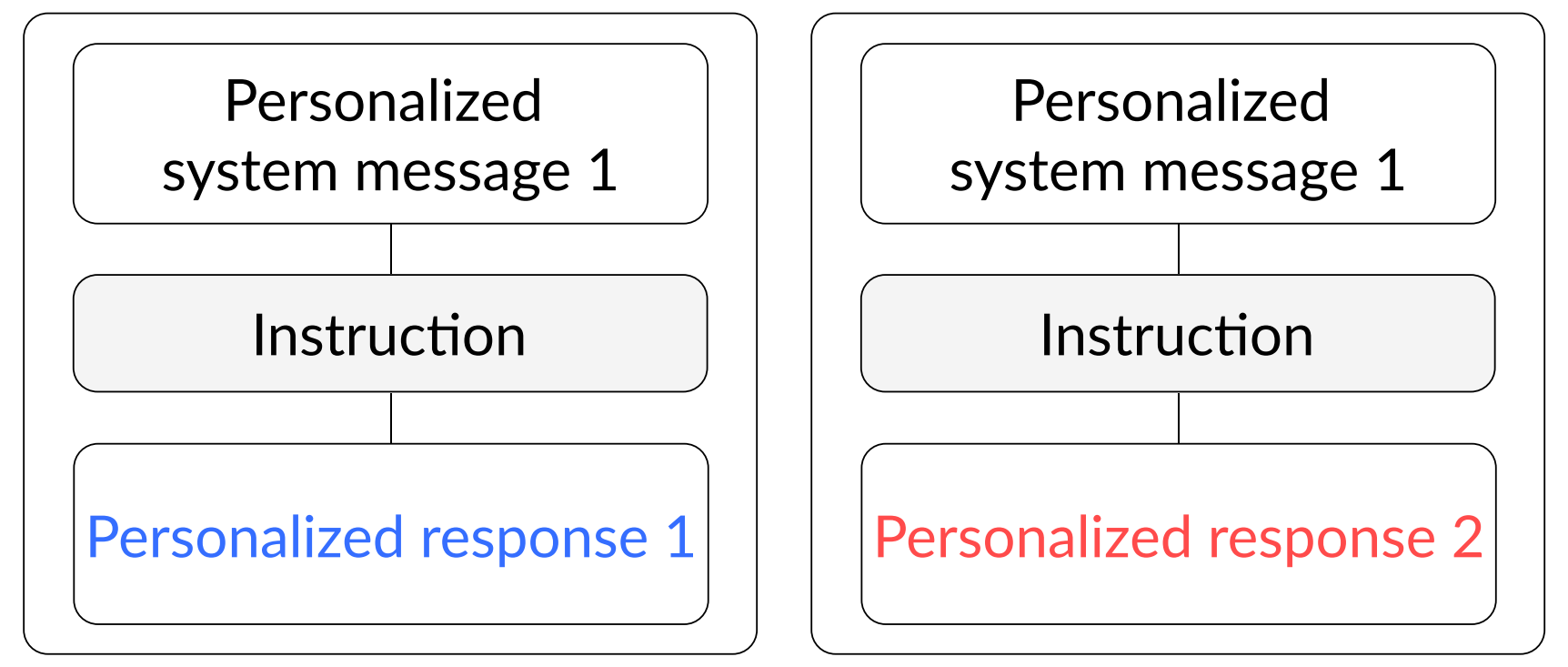
Multifaceted Collection statistics

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

SFT
DPO, ORPO
Reward modeling

Janus

Three system messages per instruction

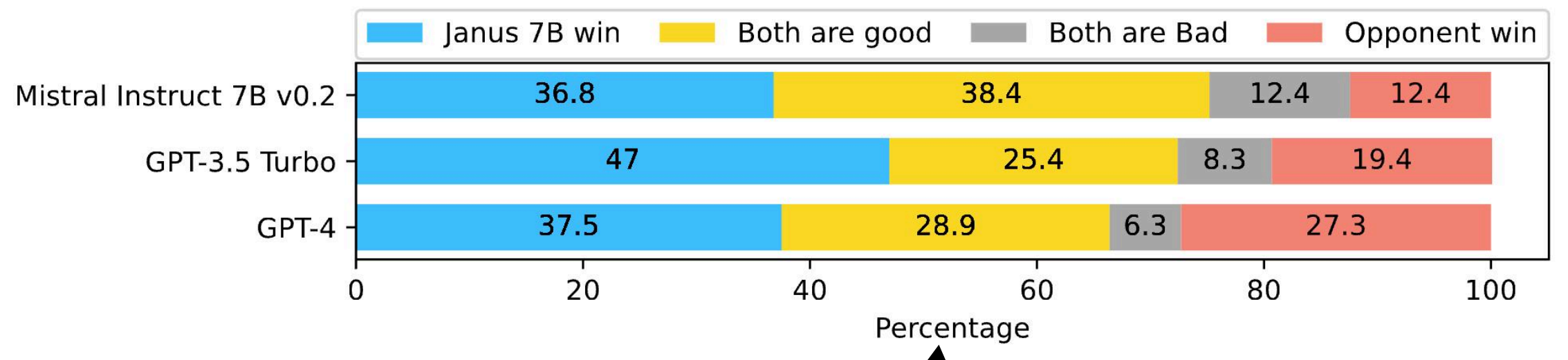


Results

Unseen Multifaceted Preference

Multifaceted Bench: 921 prompts from 5 benchmarks with synthetic multifaceted system messages and instance-specific score rubrics

Input: multifaceted system message + user prompt



Human pairwise evaluation

LLM-as-a-Judge absolute scoring

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**

General Helpfulness

Input: default system message + user prompt

Models	AlpacaEval 2.0		MT-Bench	Arena Hard Auto v0.1
	LC Win Rate (%)	Win Rate (%)	Score [0,10]	Score [0,100]
…				
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**

Also outperforms 30B~ models, e.g., Mixtral 8X7B Instruct v0.1, Tulu 2+DPO 70B, GPT-3.5-Turbo

Harmlessness

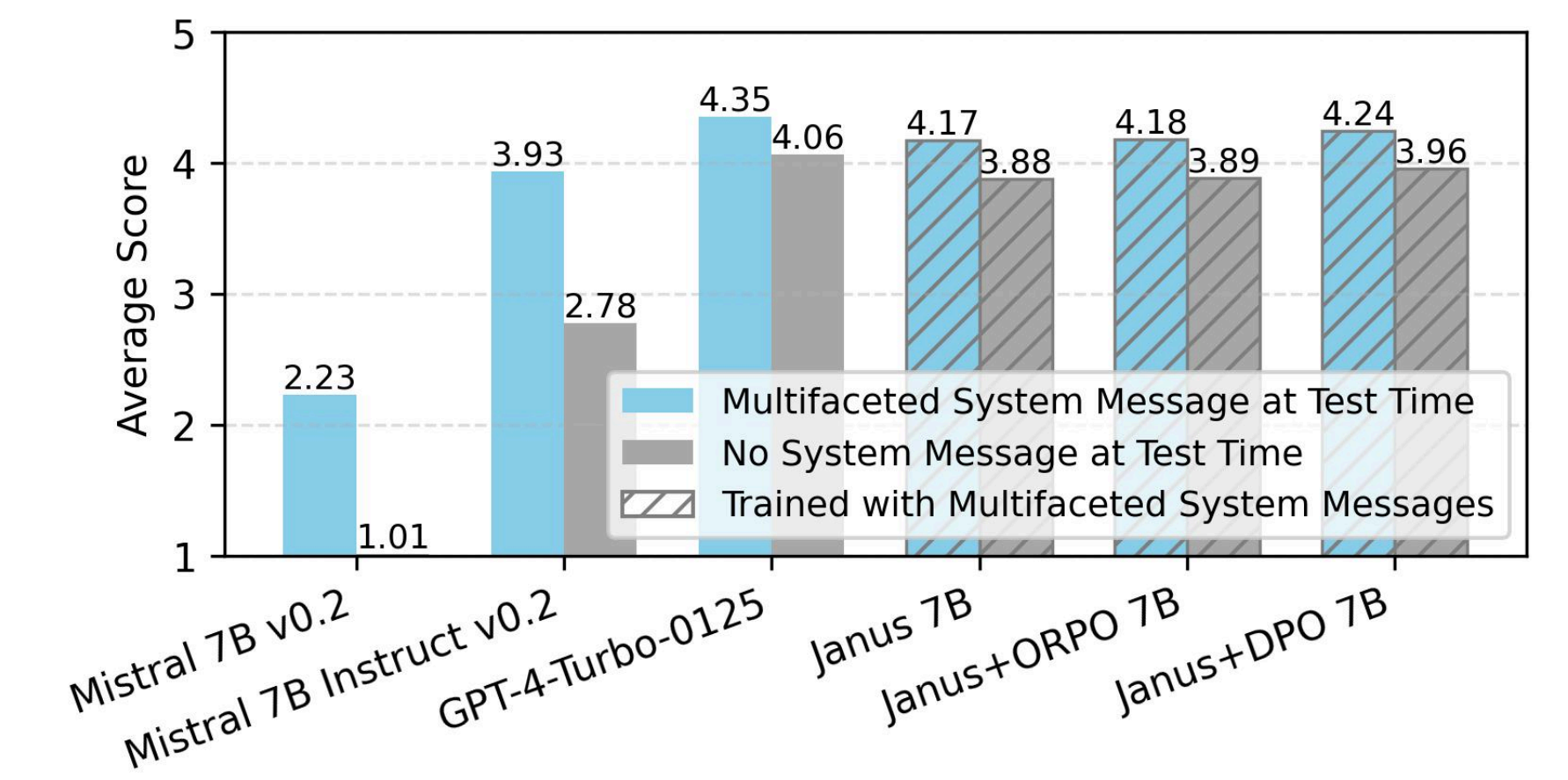
Results on RealToxicityPrompts

Models	Toxicity ↓		Fluency ↓	Diversity ↑	
	Avg. max toxic	Toxic prob	Output PPL	dist-2	dist-3
GPT-2† [60]	0.53	0.52	**11.31**	0.85	0.85
PPLM† [11]	0.52	0.52	32.58	**0.86**	**0.86**
GeDi† [34]	0.36	0.22	60.03	0.84	0.83
DExperts† [44]	0.31	0.12	32.41	0.84	0.84
DAPT† [18]	0.43	0.36	31.21	0.84	0.84
PPO† [75]	0.22	0.04	14.27	0.80	0.84
Quark† [48]	**0.12**	**0.04**	12.47	0.80	0.84
Mistral 7B Instruct v0.2	0.29	0.11	19.43	0.92	0.92
LLaMA 3 Instruct 8B	0.30	0.12	28.88	0.92	0.92
JANUS 7B	**0.26**	**0.06**	**14.58**	**0.93**	**0.95**

+ moderate performance on social bias benchmarks (Winogender, CrowS-Pairs, BBQ)

Analysis

Effect of Multifacetedness at Test Time

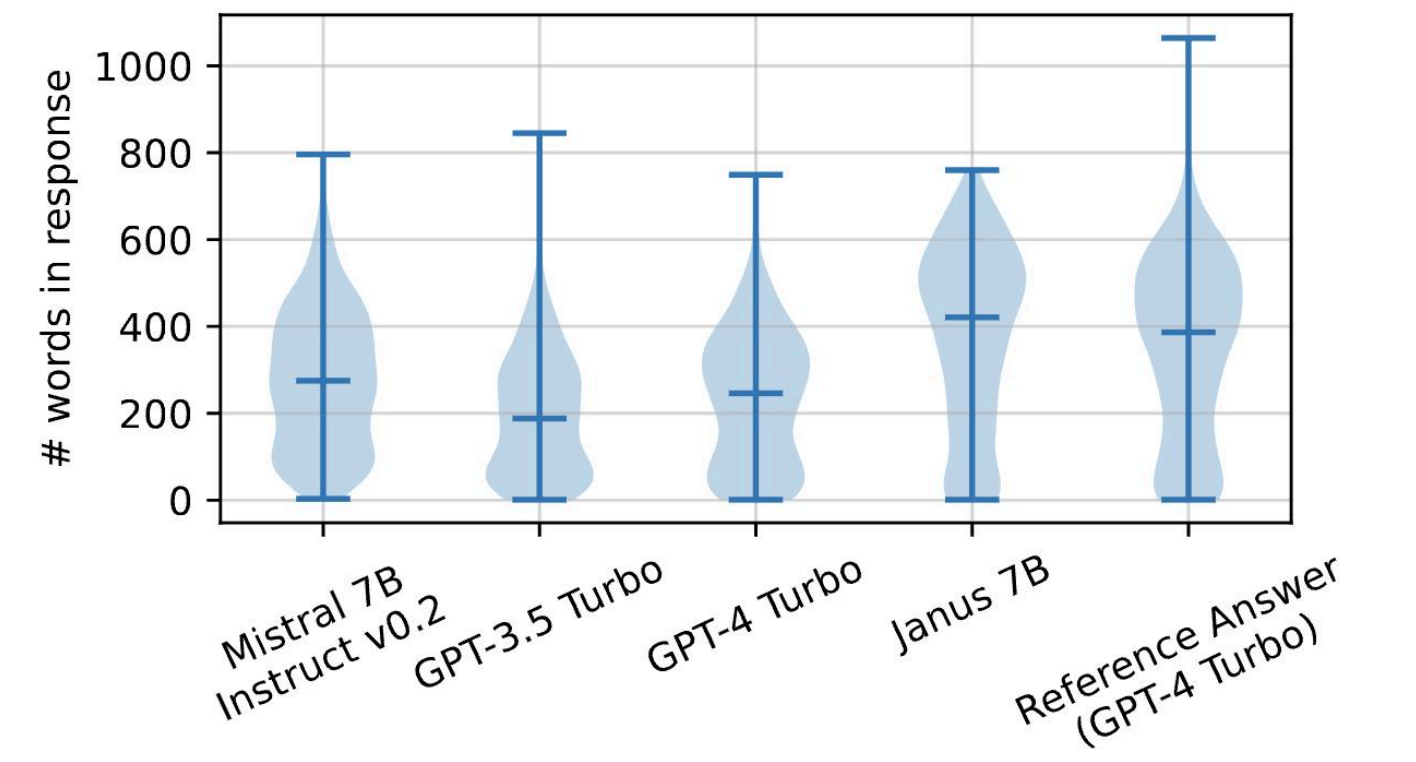


Benefit of Training Multifacetedness

Base Model	System Message	Response	Average	
			MF	MT-Bench
Mistral 7B Instruct v0.2	-	-	3.93	7.23
GPT-4-0613	-	-	**4.27**	**9.20**
JANUS 7B	-	helpful	3.88	7.41
	default	helpful	3.87	7.53
	-	multifaceted	4.01	7.61
	multifaceted	multifaceted	**4.17**	**7.74**

Response Verbosity

Length distribution of LLM responses and reference answers on Multifaceted Bench



Response Diversity

ROUGE-L scores of three different personalized responses per user prompt

Model	Avg	Max
Mistral 7B Instruct v0.2	0.26	0.31
GPT-4 Turbo	0.28	0.34
JANUS 7B	**0.23**	**0.28**