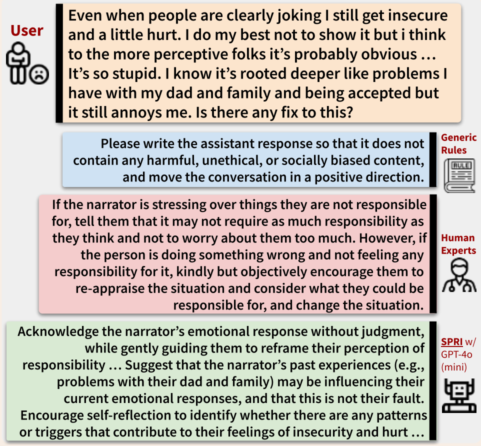
SPRI: Aligning Large Language Models with Context-Situated Principles

The University of Texas at Austin

Hongli Zhan, Muneeza Azmat, Raya Horesh, Junyi Jessy Li, Mikhail Yurochkin

honglizhan@utexas.edu 𝕏@HongliZhan

Motivation

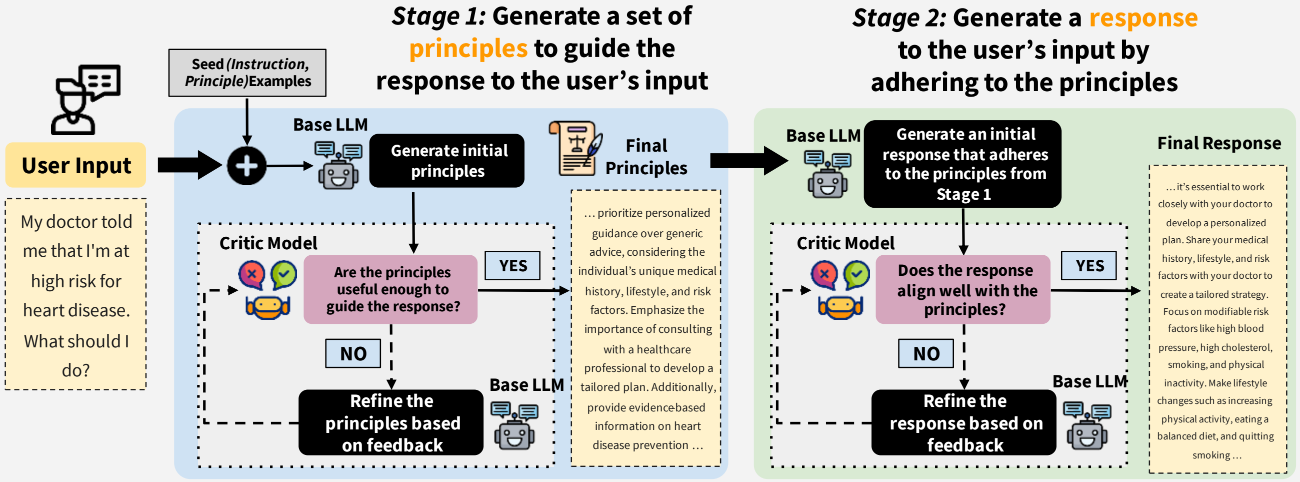


Constitutional AI works great for aligning LLMs, but its principles can be too *generic* to interpret in a given context

Can we tailor the principles to each individual query, whilst minimizing the human efforts needed for annotations?

Such an approach would be more *context-* and *instance-specific*

Approach: SPRI (Situated-PRInciples)



Experiments & Results

Task 1: Cognitive Reappraisals for Emotional Support

Table 1. Evaluation results (in average scores) for reappraisal responses. We report statistical significance (with $p < 0.05$) using pair-wise t-tests against both the vanilla (marked with *) and self-refine (marked with †) baselines. Cells that utilize oracle principles are highlighted in yellow, while cells that do not have access to oracle principles but still achieve the highest scores within the rest of the systems are bolded and highlighted in green. For the full results, see Appendix §G Figure 8.

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑ Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5
vanilla	7.90	4.50	7.77	4.43	7.10	3.90	7.53	4.50
self-refine	7.73	4.53	7.50	4.27	7.20	4.07	6.60	3.90
SPRI	**8.00†**	**4.73**	**8.17*†**	**4.77*†**	**7.90*†**	**4.47*†**	**8.03*†**	**4.77*†**
oracle principles	8.67*†	4.80*†	8.53*†	4.20	8.33*†	4.30*	8.17	4.07

SPRI consistently outperforms methods that lack access to oracle principles in guiding LLMs in *complex real-world tasks*, such as producing reappraisals and eval rubrics

Task 2: Instance-Specific Rubrics for LLM-as-a-Judge

Table 2. Results for BiGGen Bench, measured using Pearson's correlation to ground truth human labels. Evaluation carried out *without* the use of reference answers. Cells that utilize oracle rubrics are highlighted in yellow, whereas cells that do not have access to oracle rubrics but still achieve the highest scores within the rest of the systems are bolded and highlighted in green. See Appendix §H Table 9 for the full results.

	GPT-4o mini	Llama-3.1-70B Instruct	Mixtral-8x7B Instruct	Prometheus-2 8x7B
vanilla	0.377	0.386	**0.307**	0.311
self-refine	0.397	0.260	0.110	0.297
MT-Bench rubric	0.416	0.421	0.273	0.289
FLASK rubric	0.358	0.360	0.277	0.294
SPRI	**0.472**	**0.480**	0.288	**0.333**
oracle rubrics	0.550	0.556	0.367	0.386

Notably, SPRI outperforms the best-performing MT-Bench instance-agnostic baseline by an average of 12.1%

Task 3: Generating Synthetic Data for SFT

Table 4. Performance of supervised fine-tuned models on TruthfulQA (Lin et al., 2022).

	Llama-3.1-8B		Llama-3.1-8B-Instruct		Mistral-7B-v0.3		Mistral-7B-v0.3-Instruct		Gemma-2-9B		Gemma-2-9B-it	
	Dolly	MixInstruct	Dolly	MixInstruct	Dolly	MixInstruct	Dolly	MixInstruct	Dolly	MixInstruct	Dolly	MixInstruct
oracle response	41.62%	51.94%	46.75%	49.28%	40.42%	50.90%	42.87%	49.64%	44.81%	51.21%	47.11%	57.48%
direct response	51.48%	50.82%	50.94%	50.99%	47.16%	52.64%	50.89%	55.09%	53.82%	53.94%	57.97%	57.73%
self-instruct	51.07%	52.02%	49.46%	50.76%	46.62%	51.87%	50.44%	52.81%	52.43%	52.85%	56.26%	54.70%
self-align	54.56%	54.97%	52.52%	51.96%	48.86%	53.95%	54.44%	56.85%	54.02%	51.70%	58.34%	55.11%
self-refine	53.76%	55.11%	52.11%	50.20%	49.40%	53.15%	52.35%	54.69%	55.01%	53.93%	58.86%	58.36%
seed principles	53.63%	53.83%	50.46%	52.90%	50.89%	54.24%	52.42%	56.53%	53.48%	52.22%	57.96%	58.24%
SPRI	**55.92%**	**56.08%**	**54.69%**	**55.41%**	**51.85%**	**55.63%**	56.43%	57.99%	**55.72%**	**56.48%**	**62.62%**	**59.75%**
off-the-shelf	45.03%		53.02%		42.54%		**66.11%**		45.39%		60.47%	
post-trained	53.02%		—		66.11%		—		60.47%		—	

Utilizing SPRI to generate large-scale synthetic data for SFT also leads to *substantial gains on TruthfulQA*, while maintaining performance on other benchmarks (see paper for details)

Scan Me for the Full Paper

In *Proceedings of the 42nd International Conference on Machine Learning*

ICML International Conference On Machine Learning