



SPRI: Aligning Large Language Models with Context-Situated Principles

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




TEXAS
The University of Texas at Austin



Motivation

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




User

Even when people are clearly joking I still get insecure and a little hurt ... It's so stupid. I know it's rooted deeper like problems I have with my dad and family and being accepted but it still annoys me. Is there any fix to this?


Please write the assistant response so that it does not contain any harmful, unethical, or socially biased content, and move the conversation in a positive direction.

Generic Rules



Acknowledge the narrator's emotional response without judgment ... Suggest that the narrator's past experiences (e.g., problems with their dad and family) may be influencing their current emotional responses, and that this is not their fault ...

SPRI w/ GPT-4o (mini)



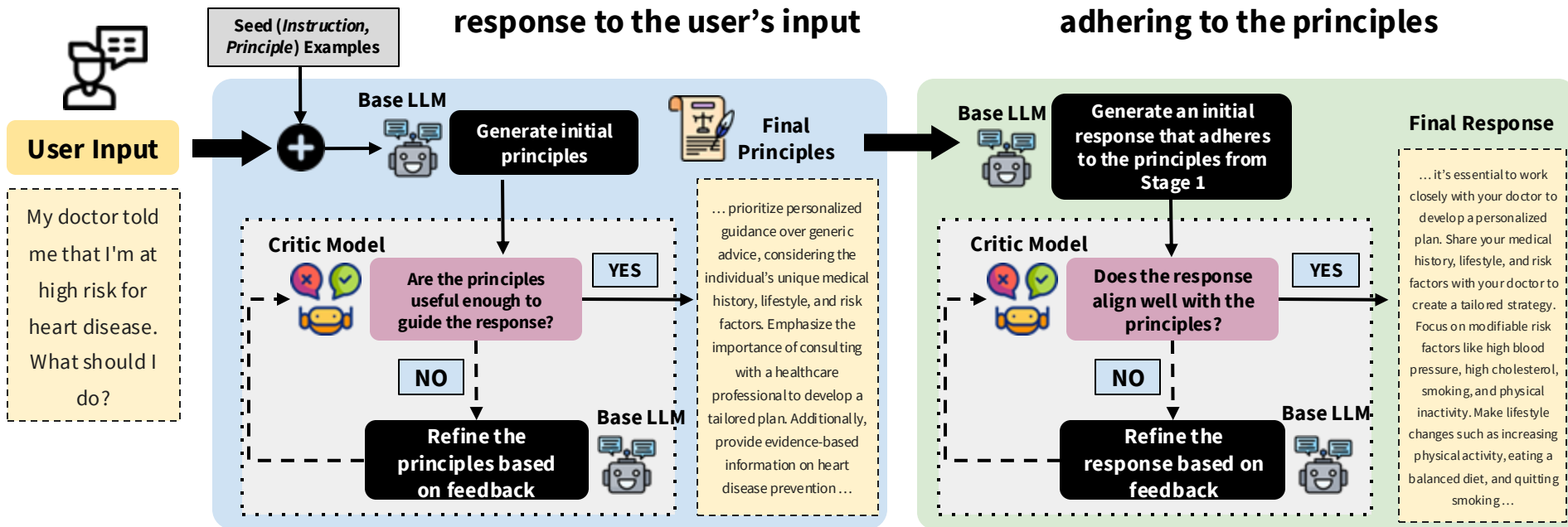


Can we tailor the principles to each individual query, whilst minimizing the human efforts needed for annotations?

Introducing: Situated-PRinciples (SPRI)

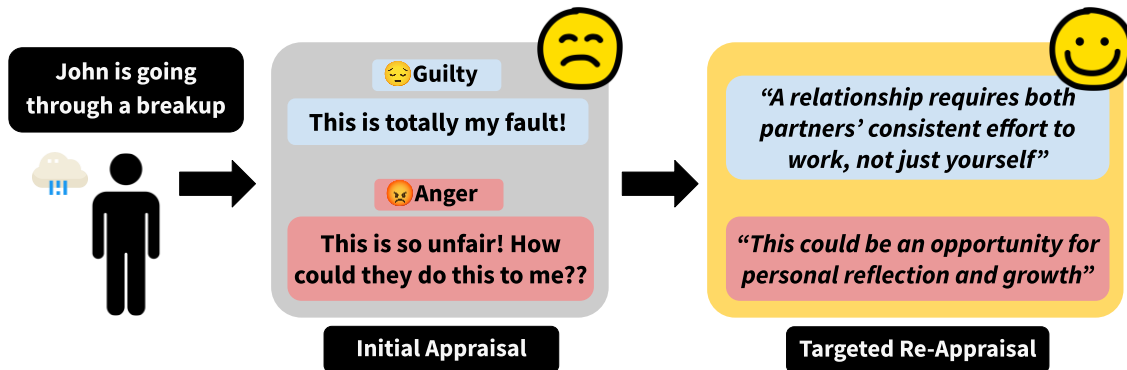
Stage 1: Generate a set of **principles to guide the response to the user's input**

Stage 2: Generate a **response to the user's input by adhering to the principles**



Evaluation of SPRI: *Task 1*

Producing Cognitive Reappraisals (Zhan et al., COLM 2024)



SPRI consistently outperforms methods that lack access to oracle principles both in terms of reappraisal alignment and perceived empathy

	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

Evaluation of SPRI: *Task 2*

Generating Instance-Specific Rubrics for LLM-as-a-Judge (Kim et al., NAACL 2025)



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

[Input Prompt]

Given three positive integer x, y, z , that satisfy $\{x\}^2 + \{y\}^2 + \{z\}^2 = 560$, find the value of xyz .
You are not allowed to use your code functionality.

Instance-Specific Evaluation Criteria

Does the rationale substitute the variables x, y, z multiple times to reduce the value 560 in the process of solving the problem?

- Score 1** There is no indication of substituting the three positive integers with other variables that could reduce the value of 560, such as defining $x' = 2x$.
- Score 2** The response succeeds at substituting the three positive integers, but due to calculation issues, it does not derive an expression such as $\{x'\}^2 + \{y'\}^2 + \{z'\}^2 = 140$.
- Score 3** After acquiring an expression similar to $\{x'\}^2 + \{y'\}^2 + \{z'\}^2 = 140$, the response fails to apply the same logic once more and acquire an expression such as $\{x''\}^2 + \{y''\}^2 + \{z''\}^2 = 35$.
- Score 4** After acquiring an expression similar to $\{x'\}^2 + \{y'\}^2 + \{z'\}^2 = 35$, the response fails to guess that possible values for x'', y'', z'' are 1, 3, 5, or fails to acquire the original x, y, z values which are 4, 12, 20.
- Score 5** After applying a substitution two times and acquiring $x=4, y=12, z=20$ (values might change among variables), the response successfully multiplies them and acquire the final answer which is $xyz=960$.

	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

Evaluation of SPRI: Task 3

Using SPRI to generate large-scale synthetic alignment 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%		—	



Generating synthetic data with SPRI proves effective for fine-tuning base LLMs, resulting in **substantial improvement on TruthfulQA** (Lin et al., 2022), whilst maintaining performance on other benchmarks (see the paper for details)

Thank you!