SPRI: Aligning Large Language Models with Context-Situated Principles

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

honglizhan@utexas.edu   𝕏@HongliZhan

TEXAS — The University of Texas at Austin

IBM

Motivation

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?

Generic Rules: 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.

SPRI w/ GPT-4o (mini): 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 …
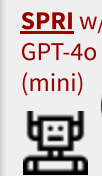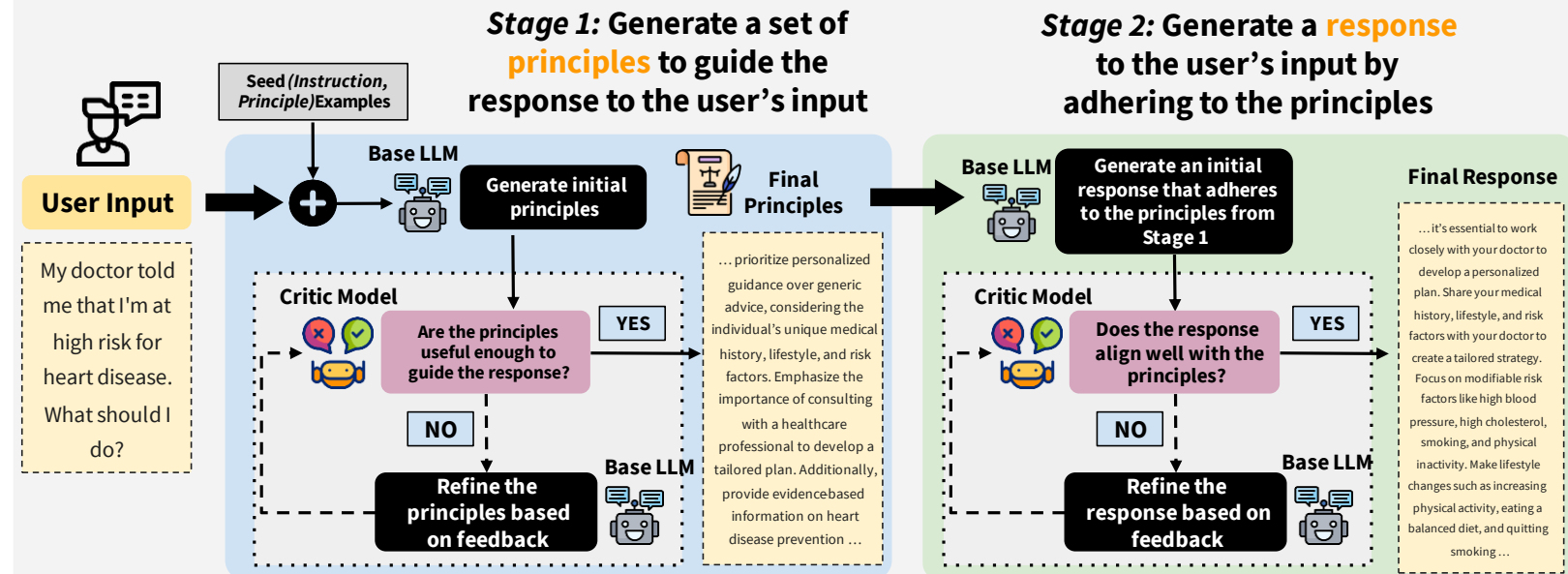
Constitutional AI works great for aligning LLMs, but its principles can be too *generic* to interpret in a given context

Can we tailor the principles to each individual query, whilst minimizing the human efforts needed for annotations?
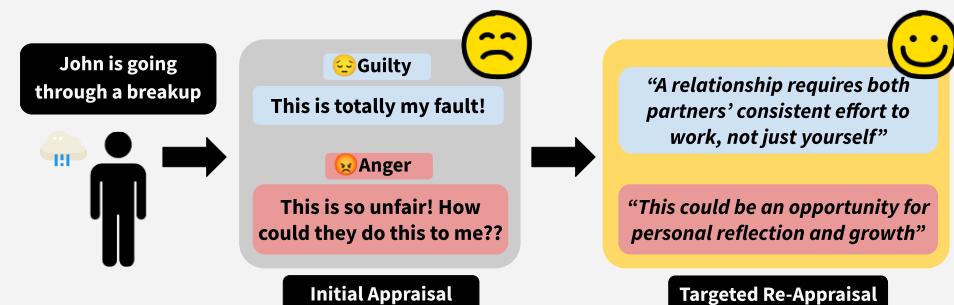
Such an approach would be more *context-* and *instance-specific*
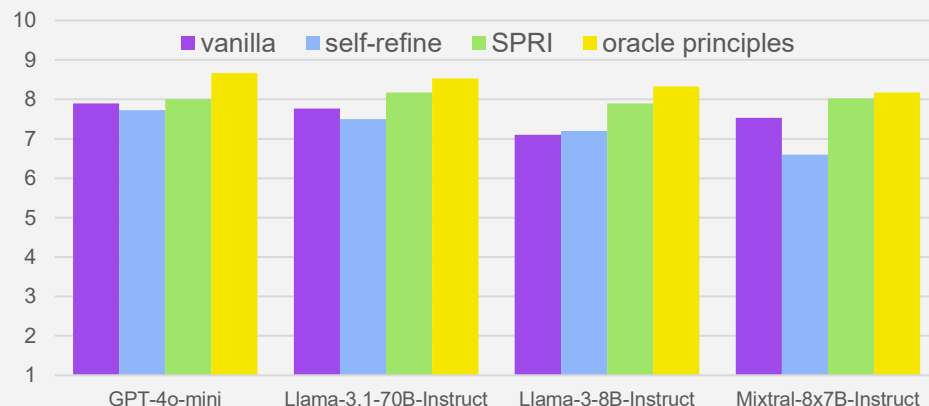
Approach: *SPRI* (Situated-PRInciples)

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

User Input: My doctor told me that I'm at high risk for heart disease. What should I do?

Seed *(Instruction, Principle)* Examples

Base LLM → Generate initial principles → Final Principles

Critic Model: Are the principles useful enough to guide the response? — YES / NO → Refine the principles based on feedback (Base LLM)

Final Principles: … prioritize personalized guidance over generic advice, considering the individual's unique medical history, lifestyle, and risk factors. Emphasize the importance of consulting with a healthcare professional to develop a tailored plan. Additionally, provide evidence-based information on heart disease prevention …

Base LLM → Generate an initial response that adheres to the principles from Stage 1

Critic Model: Does the response align well with the principles? — YES / NO → Refine the response based on feedback (Base LLM)

Final Response: …it's essential to work closely with your doctor to develop a personalized plan. Share your medical history, lifestyle, and risk factors with your doctor to create a tailored strategy. Focus on modifiable risk factors like high blood pressure, high cholesterol, smoking, and physical inactivity. Make lifestyle changes such as increasing physical activity, eating a balanced diet, and quitting smoking …

Experiments & Results

Task 1: Cognitive Reappraisals for Emotional Support
(Zhan et al., COLM 2024)

John is going through a breakup

Initial Appraisal:
- 😟 Guilty — This is totally my fault!
- 😠 Anger — This is so unfair! How could they do this to me??

Targeted Re-Appraisal:
- "A relationship requires both partners' consistent effort to work, not just yourself"
- "This could be an opportunity for personal reflection and growth"

Results: *Alignment to Reappraisal Standards*

Legend: vanilla, self-refine, SPRI, oracle principles

(Bar chart across GPT-4o-mini, Llama-3.1-70B-Instruct, Llama-3-8B-Instruct, Mixtral-8x7B-Instruct)

💡 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
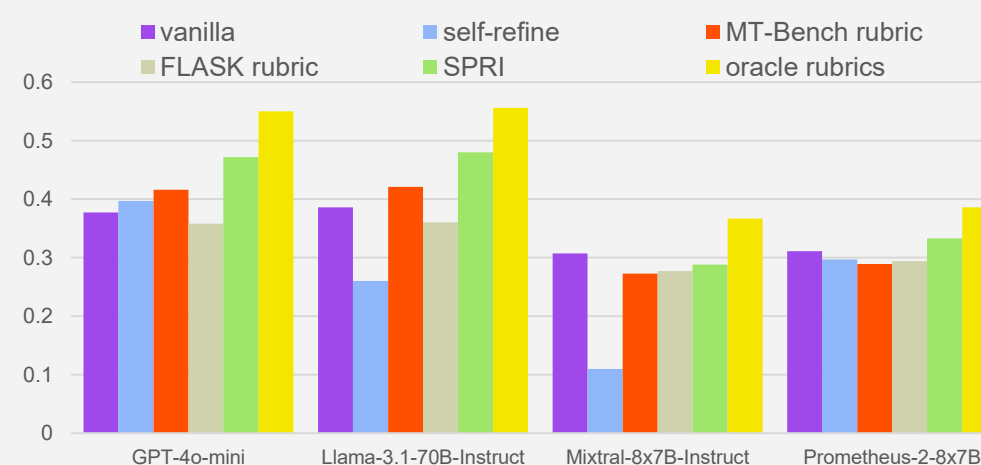
Task 2: Instance-Specific Rubrics for LLM-as-a-Judge
(Kim et al., NAACL 2025)

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.

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.

Results: *Pearson's correlation to ground truth labels*

Legend: vanilla, self-refine, MT-Bench rubric, FLASK rubric, SPRI, oracle rubrics

(Bar chart across GPT-4o-mini, Llama-3.1-70B-Instruct, Mixtral-8x7B-Instruct, Prometheus-2-8x7B)

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

Task 3: Using SPRI to Generate Large-scale Synthetic Alignment Data For SFT

💡 → leads to *substantial gains on TruthfulQA*

→ maintains performance on other benchmarks (see the paper for full details)

Conclusion

SPRI:
1) matches *expert-level* performance in highly specialized tasks;
2) *enhances alignment* with human judgment;
3) improves *synthetic data generation* for model fine-tuning.

References

Zhan, H., Zheng, A., Lee, Y. K., Suh, J., Li, J. J., & Ong, D. C. (2024). Large Language Models Are Capable of Offering Cognitive Reappraisal, if Guided. In *Proceedings of the 1st Conference on Language Modeling (COLM)*.

Kim, S., et al. (2025) The Biggen Bench: A Principled Benchmark For Fine-grained Evaluation of Language Models With Language Models. In *Proceedings of NAACL 2025*.

Scan Me for the Full Paper

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

ICML — International Conference On Machine Learning