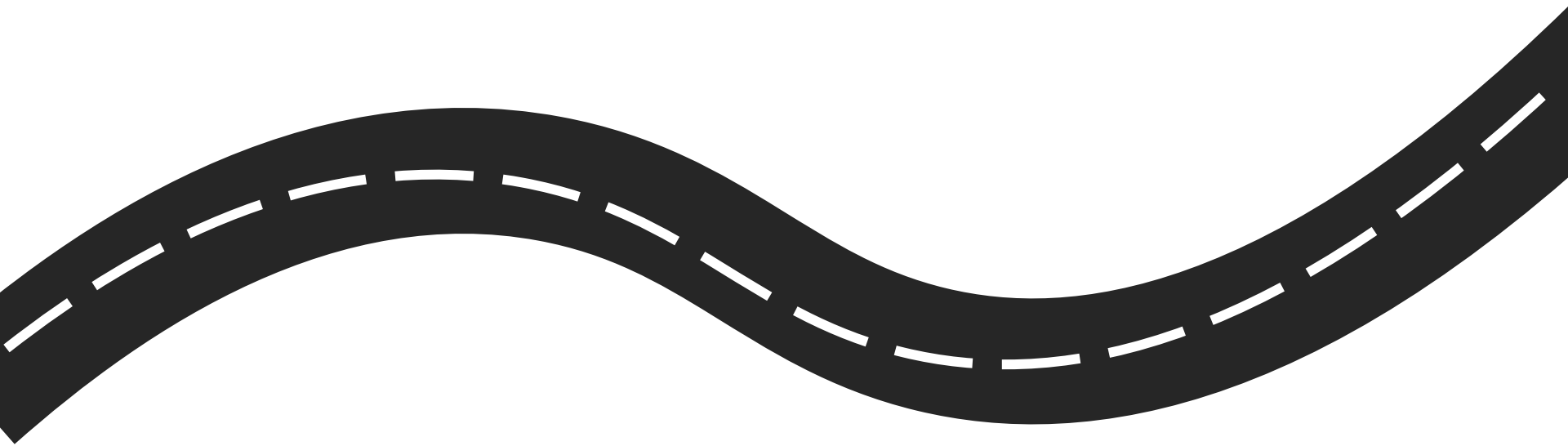


Towards Emotionally-Intelligent AI Systems

Hongli Zhan

Ph.D. Prospectus Defense

The University of Texas at Austin



Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1



Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- **Next Step:** Evaluating Reasoning Models' Capability on Emotion Tasks

2

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- **Next Step:** Evaluating Reasoning Models' Capability on Emotion Tasks

2

Empowering LLM-Chatbots with More Empathy

- **Next Step:** Surgically Making LLMs More Empathic in Multi-turn Conversations

3

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- **Next Step:** Evaluating Reasoning Models' Capability on Emotion Tasks

2

Empowering LLM-Chatbots with More Empathy

- **Next Step:** Surgically Making LLMs More Empathic in Multi-turn Conversations

3

4

Conclusion

- Summary of Proposed Work
- Timeline

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- *Next Step:* Evaluating Reasoning Models' Capability on Emotion Tasks

2

Empowering LLM-Chatbots with More Empathy

- *Next Step:* Surgically Making LLMs More Empathic in Multi-turn Conversations

3

4

Conclusion

- Summary of Proposed Work
- Timeline

Part 1 (a)

Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, Junyi Jessy Li

In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)

Summary of the Paper

Summary of the Paper

Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

Summary of the Paper

Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

- We propose a new task: **Emotion Detection & Trigger Summarization**

Summary of the Paper

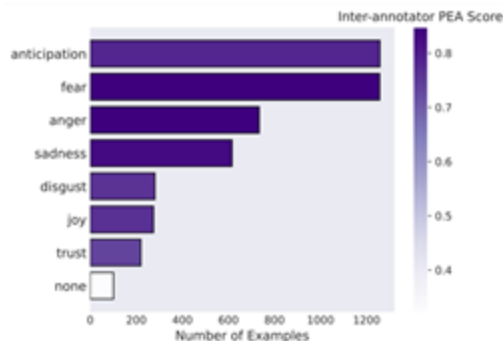
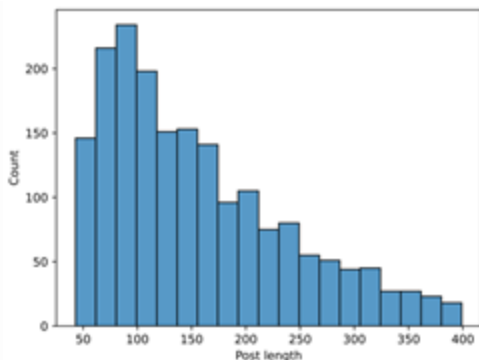
Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

- We propose a new task: **Emotion Detection & Trigger Summarization**
- New benchmark CovidET: Reddit posts annotated with emotions and their triggers

Summary of the Paper

Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

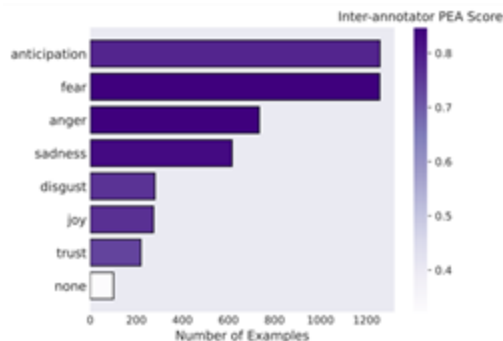
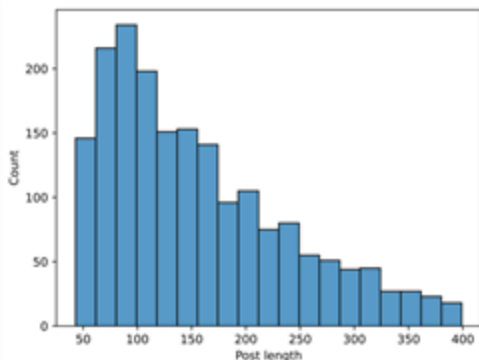
- We propose a new task: **Emotion Detection & Trigger Summarization**
- New benchmark CovidET: Reddit posts annotated with emotions and their triggers



Summary of the Paper

Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

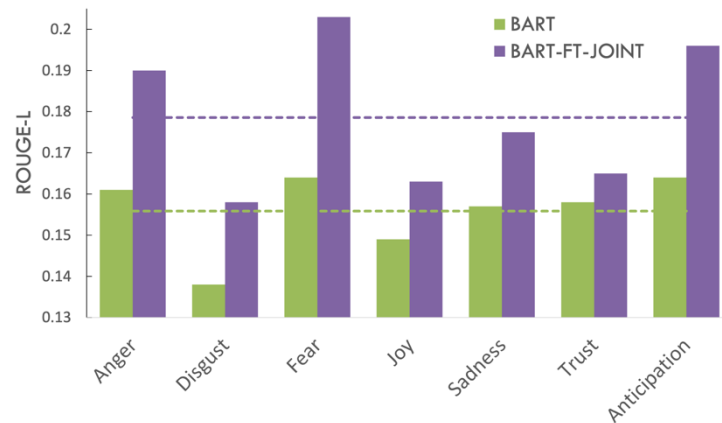
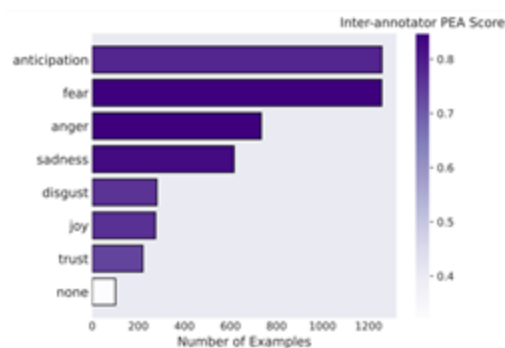
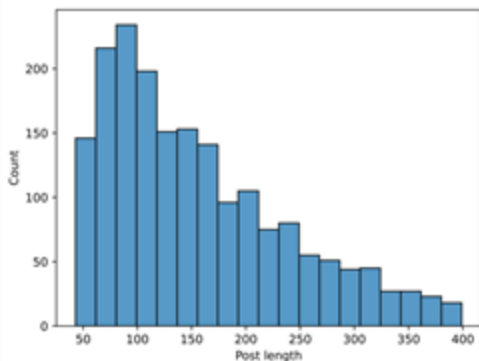
- We propose a new task: **Emotion Detection & Trigger Summarization**
- New benchmark CovidET: Reddit posts annotated with emotions and their triggers
- Benchmarking on the dataset, we were able to build automatic summarizers that outperform generic summarizers on the task of emotion-specific trigger summarization.



Summary of the Paper

Revealing “*Why does the writer feel [emotion]?*” is important yet remains unexplored:

- We propose a new task: **Emotion Detection & Trigger Summarization**
- New benchmark CovidET: Reddit posts annotated with emotions and their triggers
- Benchmarking on the dataset, we were able to build automatic summarizers that outperform generic summarizers on the task of emotion-specific trigger summarization.



Models can tell how people feel!
But is that enough?

Models can tell how people feel!

But is that enough?



How capable are language models of a deeper emotion understanding?

Part 1 (b)

Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models

Hongli Zhan, Desmond C. Ong, Junyi Jessy Li

In Findings of the Association for Computational Linguistics: EMNLP 2023

Cognitive Appraisals of Emotions

Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)

Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)

John is going through a breakup



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)

John is going through a breakup

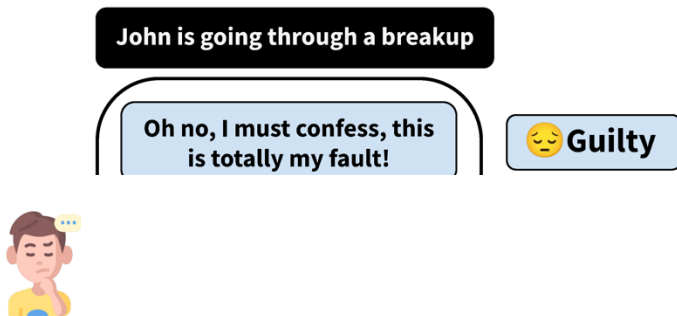
Oh no, I must confess, this
is totally my fault!



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

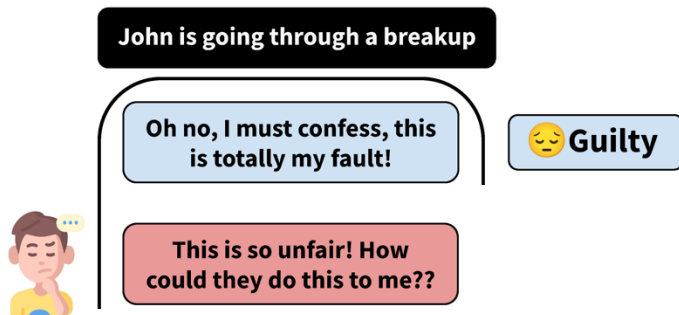
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

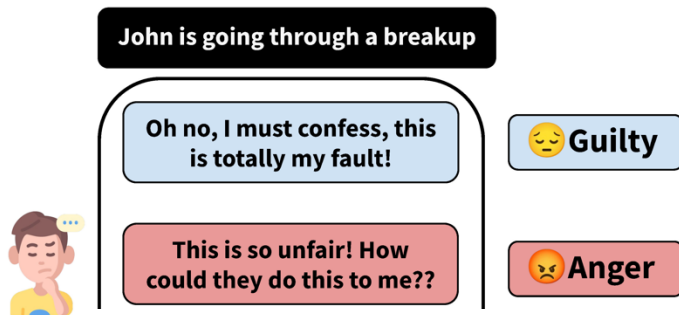
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

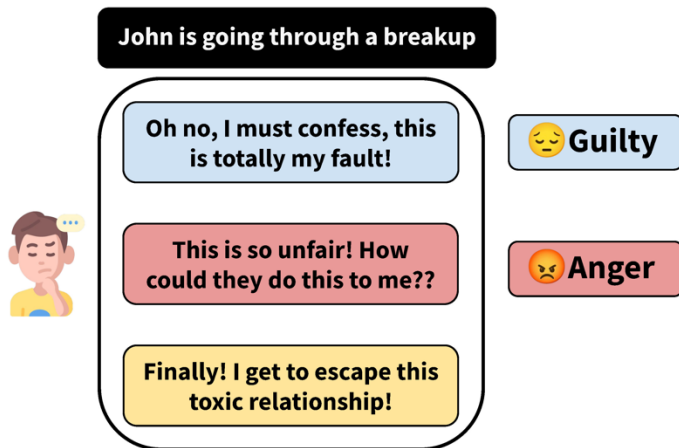
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

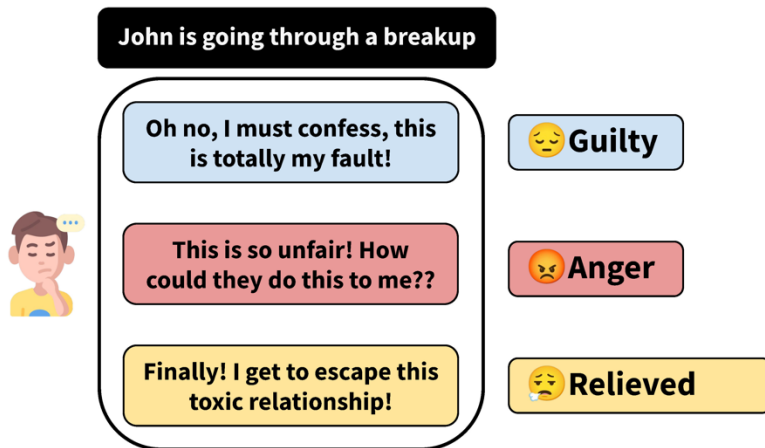
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

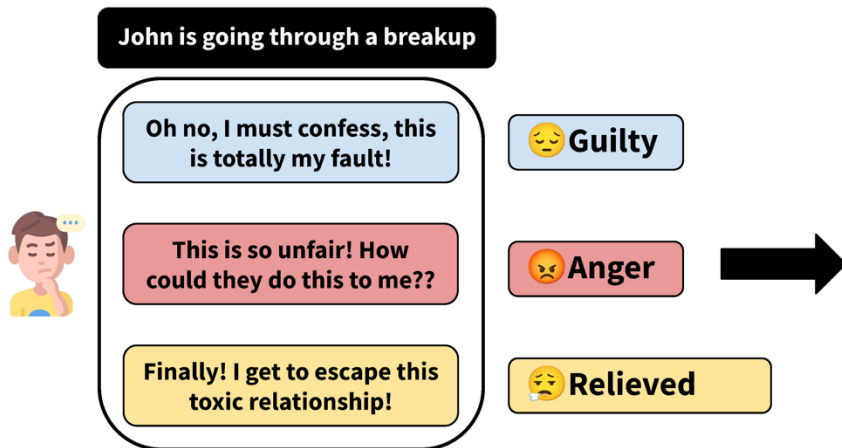
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

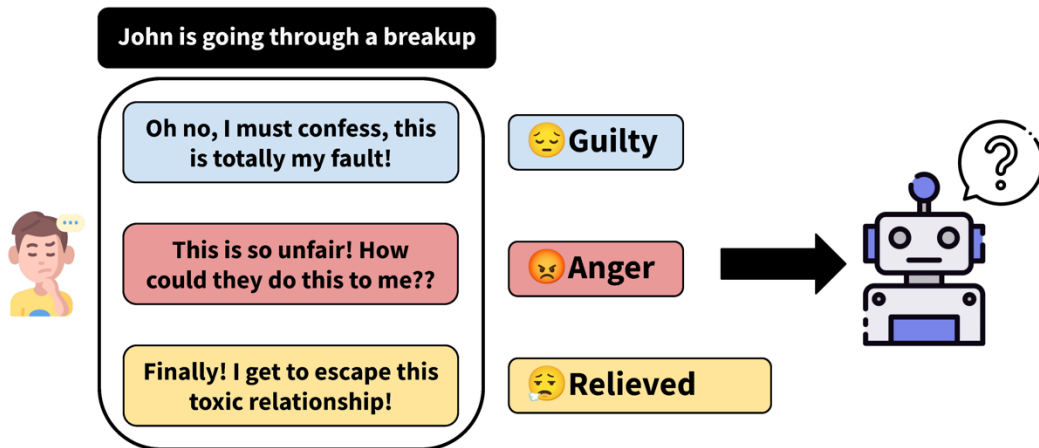
- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)



Cognitive Appraisals of Emotions

Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

- This is typically characterized by a range of different “**dimensions**” (Arnold, 1960; Smith and Ellsworth, 1985; Yeo and Ong, 2023)




Benchmark: *CovidET-Appraisals*


From a recent meta-analysis (Yeo & Ong, 2024), we identified a set of 24 appraisals, and created prompts, e.g., “To what extent did the narrator think that THEY were responsible for causing the situation?”

Benchmark: CovidET-Appraisals

From a recent meta-analysis (Yeo & Ong, 2024), we identified a set of 24 appraisals, and created prompts, e.g., “To what extent did the narrator think that THEY were responsible for causing the situation?”





I've recently heard of stories that people who are getting mildly sick after being vaccinated are still coming down with long covid. People on the covid long hauler subreddit are saying that we're going to have a huge problem on our hands because no vaccine protects against long covid... This isn't going to end, isn't it? Right as I think I've tasted freedom, I find out information that makes me want to stay inside forever like a hermit...



1. To what extent did the narrator think that **THEY** were responsible for causing the situation?


[Scale]: 1 → [Rationale]: The narrator does not talk about themselves in relation to the COVID situation in a way that makes it seem like they believe they are responsible.






2. To what extent did the narrator think that **OTHER PEOPLE** were responsible for causing the situation?


[Scale]: 2 → [Rationale]: The narrator mentions how many people who are vaccinated and get mildly sick still contract long COVID, but they do not talk about them in a way that seems like they are blaming them for the situation.






3. To what extent did the narrator think that **CIRCUMSTANCES BEYOND ANYONE'S CONTROL** were responsible for causing the situation?

[Scale]: 9 → [Rationale]: The narrator mentions long COVID and the inability to currently protect against long COVID. This is a factor that is still outside of anyone's control, including doctors and nurses, and the situation continues to evolve in uncontrollable ways.




...



24. To what extent did the narrator **EXPECT** the situation to occur?

[Scale]: 3 → [Rationale]: The narrator really didn't expect this situation since they mention being able to taste freedom, believing the pandemic is ending, when suddenly they heard news that vaccinated people are still getting long covid and now they think the pandemic will never end.



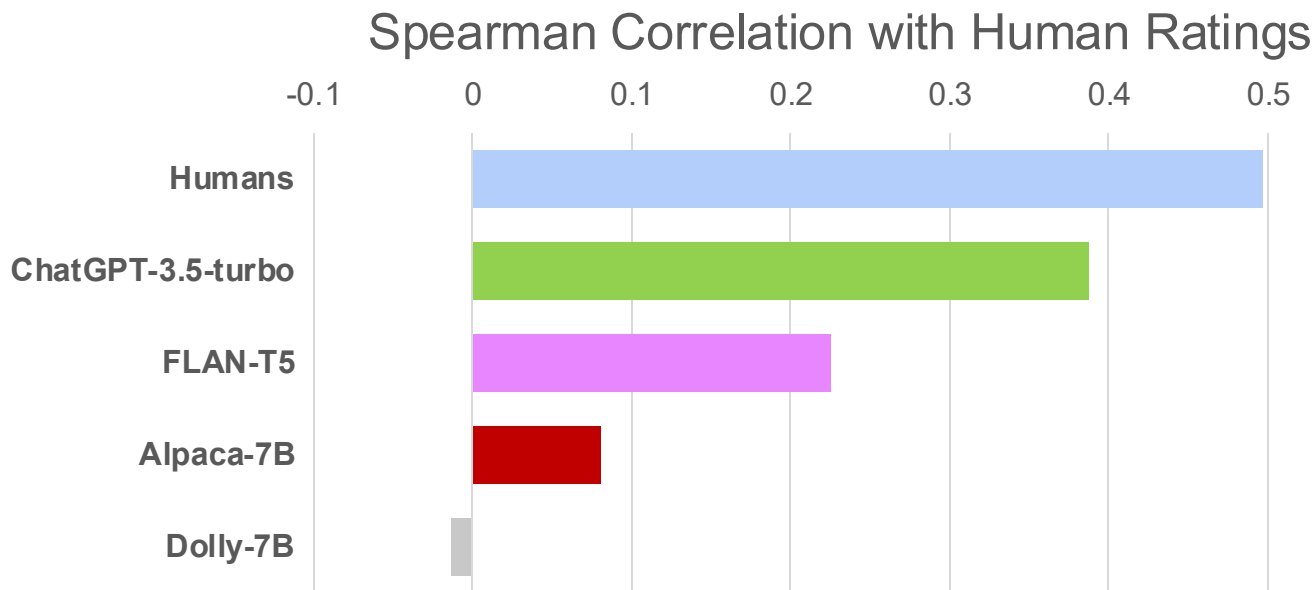
Results: Likert Scales

Results: Likert Scales

We then evaluated the accuracy of various LLMs in identifying these appraisals

Results: Likert Scales

We then evaluated the accuracy of various LLMs in identifying these appraisals



LLMs are able to identify appraisals of emotions (without additional training)

LLMs are able to identify appraisals of emotions (without additional training)



⇒ Can we leverage this emotion understanding capability of LLMs to do good to humans?

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- **Next Step:** Evaluating Reasoning Models' Capability on Emotion Tasks

2

Empowering LLM-Chatbots with More Empathy

- *Next Step:* Surgically Making LLMs More Empathic in Multi-turn Conversations

3

4

Conclusion

- Summary of Proposed Work
- Timeline

Part 2 (a)

Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, Desmond Ong

In Proceedings of the First Conference on Language Modeling (COLM 2024)

Targeted Reappraisal: A Toy Example

John is going
through a breakup



 Guilty

This is totally my fault!

 Anger

This is so unfair! How
could they do this to me??

Initial Appraisal



*"A relationship requires both
partners' consistent effort to
work, not just yourself"*

*"This could be an opportunity for
personal reflection and growth"*

Targeted Re-Appraisal



Targeted Reappraisal: A Toy Example

John is going
through a breakup



 Guilty

This is totally my fault!

 Anger

This is so unfair! How
could they do this to me??

Initial Appraisal



*“A relationship requires both
partners’ consistent effort to
work, not just yourself”*

*“This could be an opportunity for
personal reflection and growth”*

Targeted Re-Appraisal



Targeted Reappraisal: A Toy Example

John is going through a breakup



 Guilty

This is totally my fault!

 Anger

This is so unfair! How could they do this to me??

Initial Appraisal



"A relationship requires both partners' consistent effort to work, not just yourself"

"This could be an opportunity for personal reflection and growth"

Targeted Re-Appraisal



Can we use LLMs to achieve better **emotional well-being** through offering **reappraisals**?

Targeted Reappraisal: A Toy Example

John is going
through a breakup



 Guilty

This is totally my fault!

 Anger

This is so unfair! How
could they do this to me??

Initial Appraisal



*"A relationship requires both
partners' consistent effort to
work, not just yourself"*

*"This could be an opportunity for
personal reflection and growth"*

Targeted Re-Appraisal



Can we use LLMs to achieve better **emotional well-being** through offering **reappraisals**?

⇒ Such an approach would be more targeted and precise

New Framework: *RESORT*

New Framework: *RESORT*

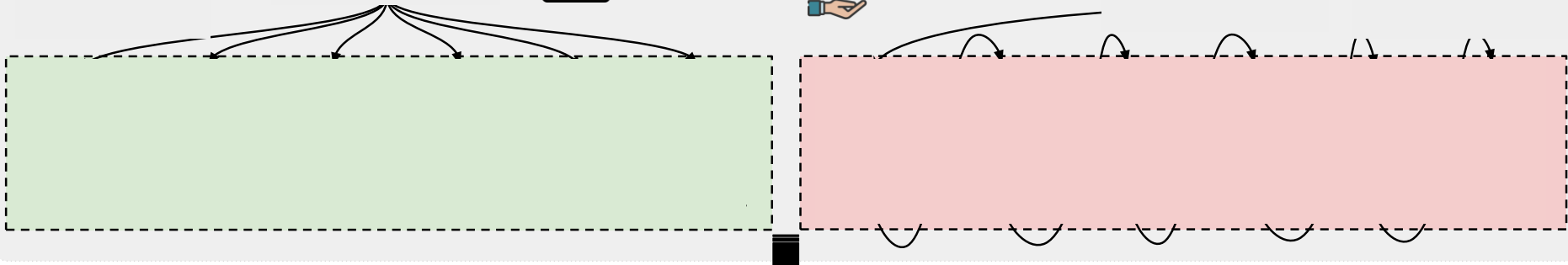
We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



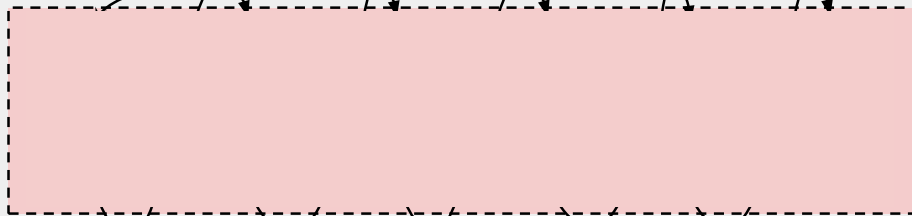
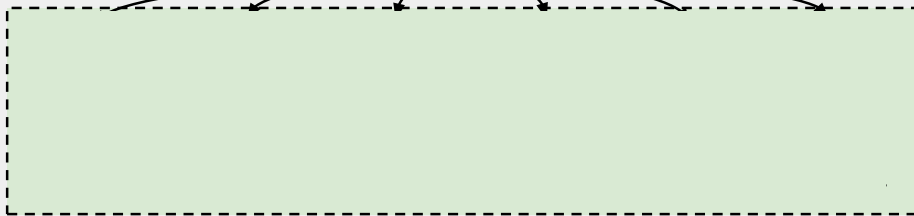
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



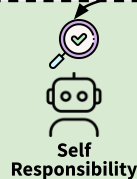
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



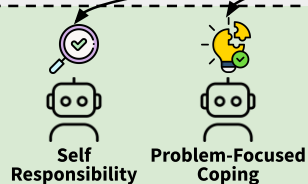
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



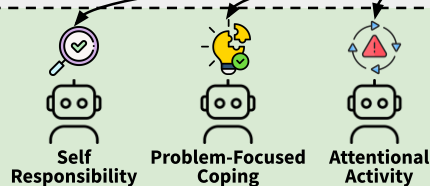
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



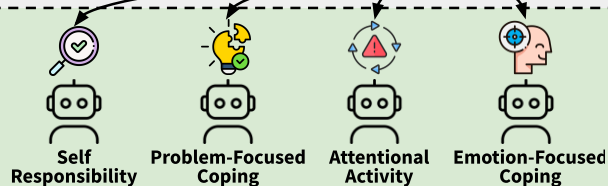
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually



Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



Self
Responsibility



Problem-Focused
Coping



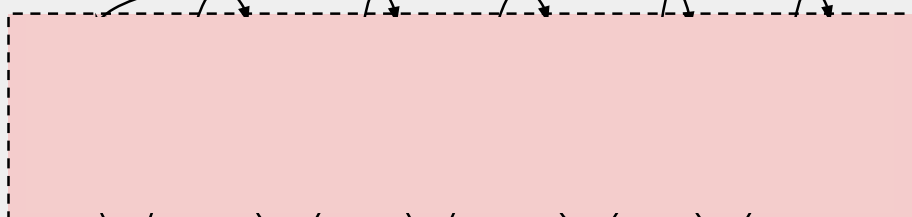
Attentional
Activity



Emotion-Focused
Coping



Self
Controllable



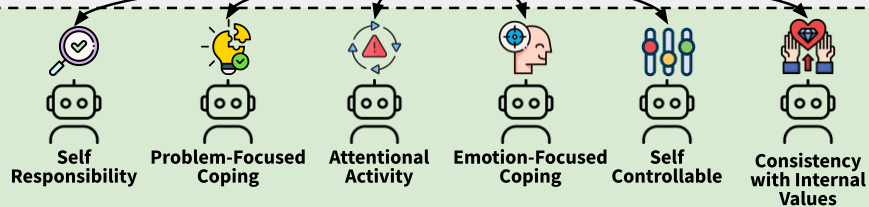
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually

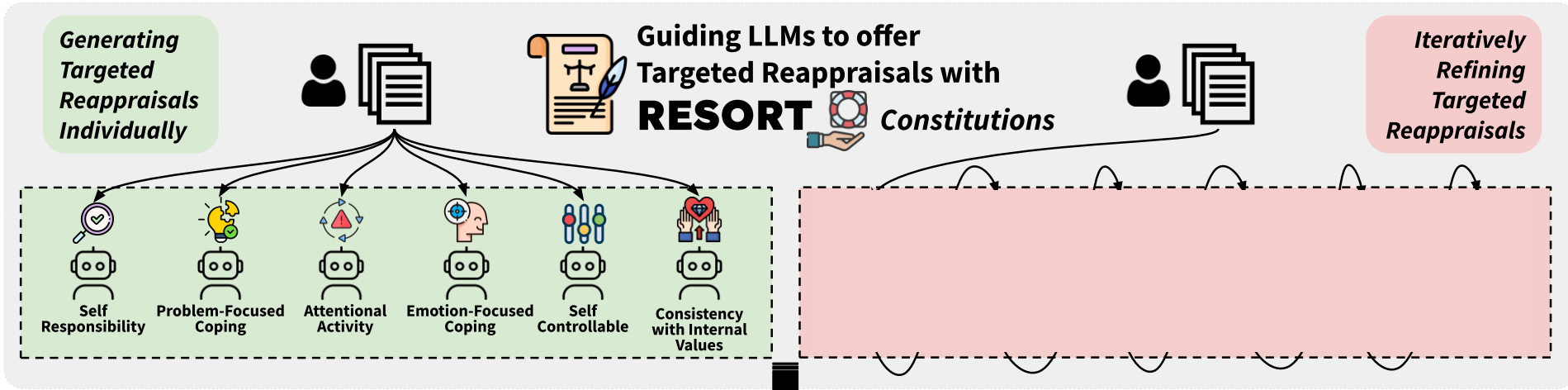


Guiding LLMs to offer
Targeted Reappraisals with
RESORT  *Constitutions*



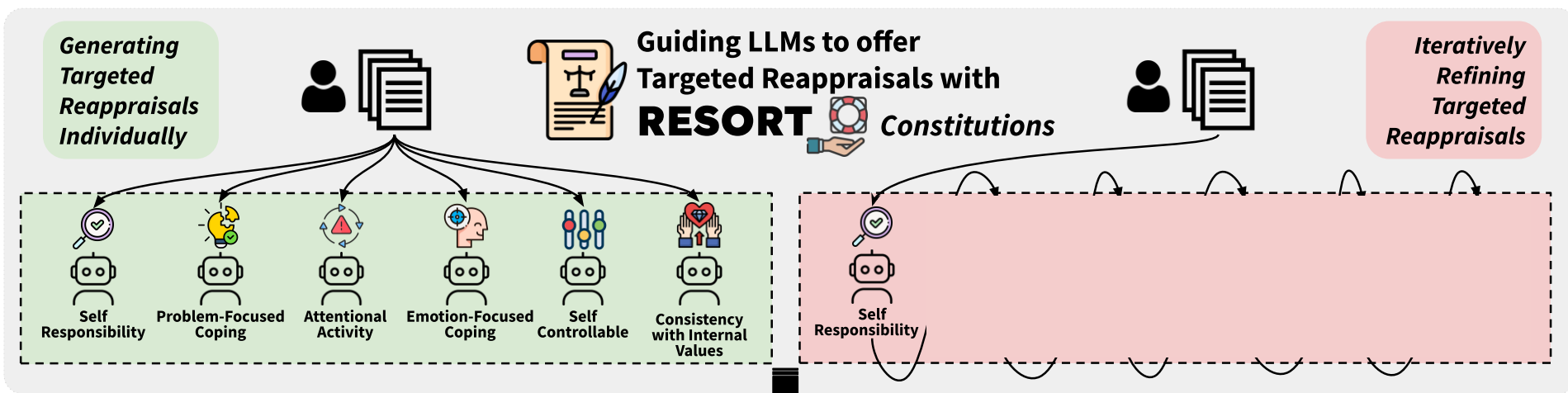
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



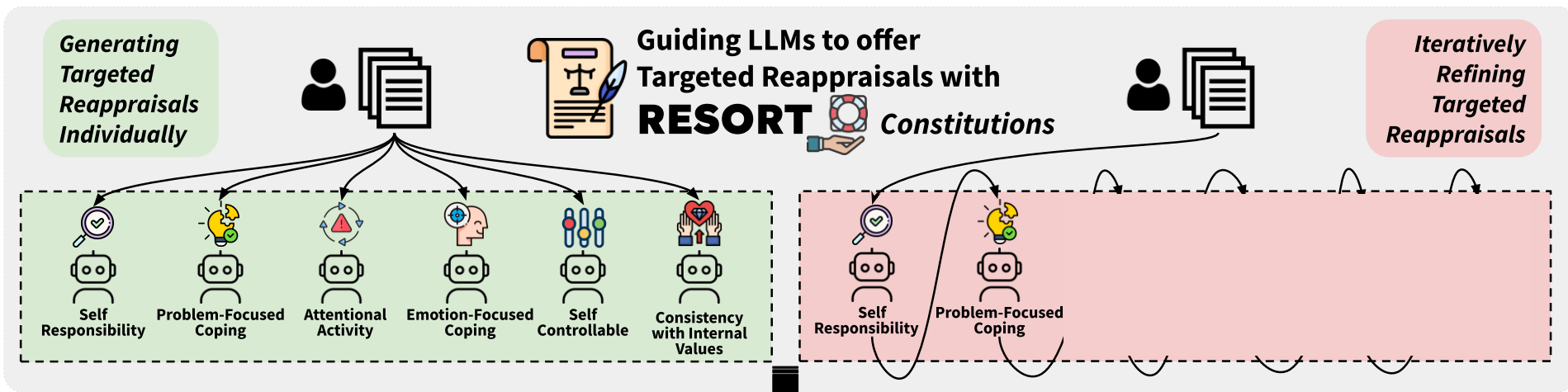
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



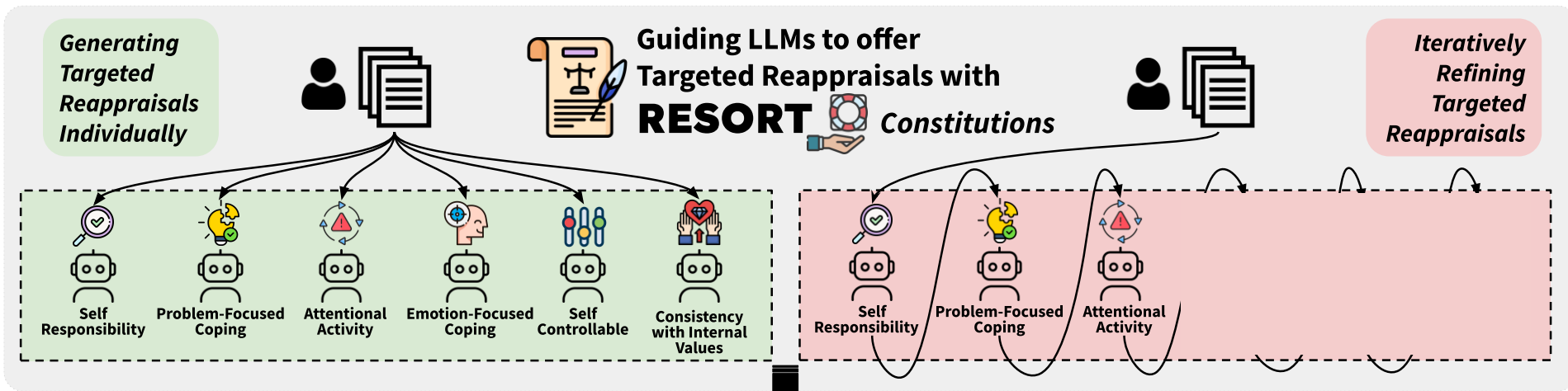
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



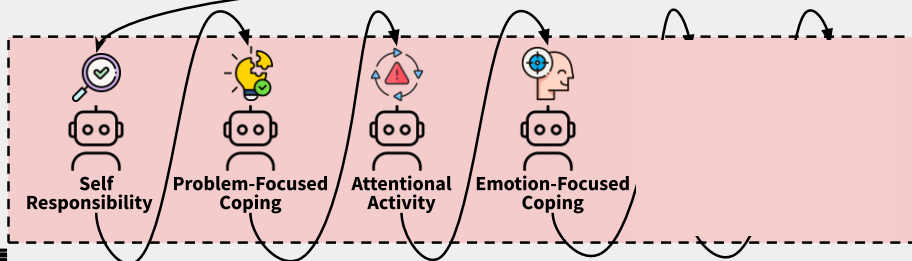
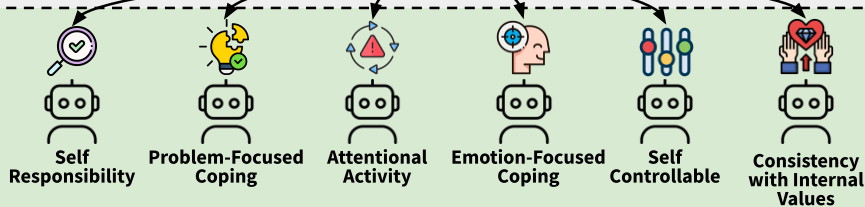
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage

Generating
Targeted
Reappraisals
Individually

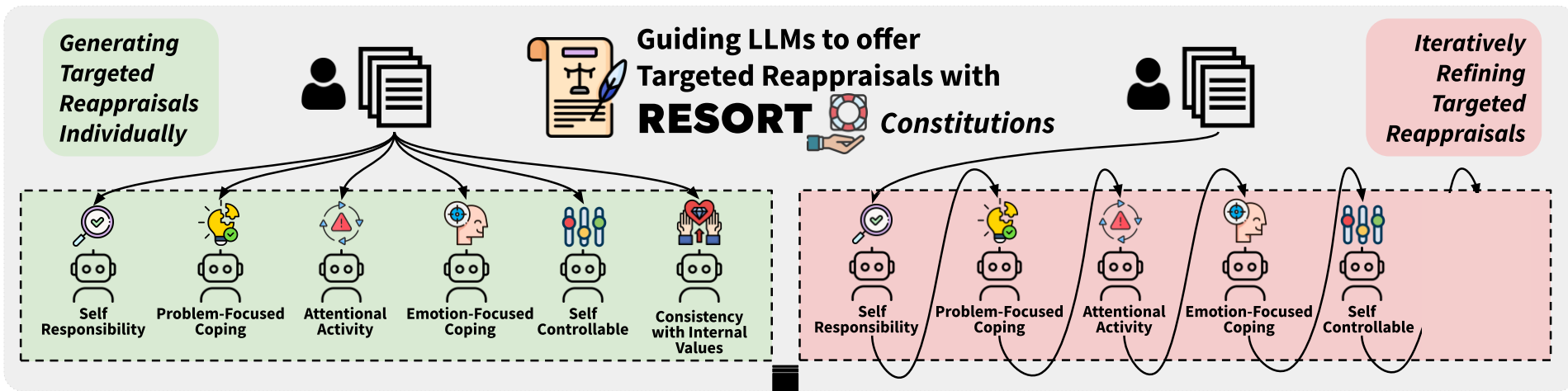
Guiding LLMs to offer
Targeted Reappraisals with
RESORT  Constitutions

Iteratively
Refining
Targeted
Reappraisals



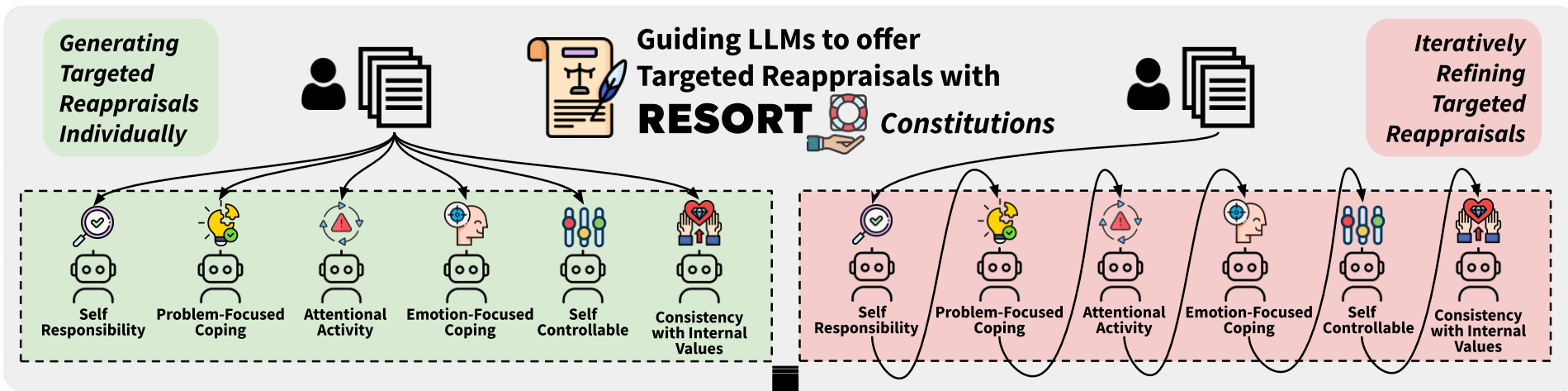
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



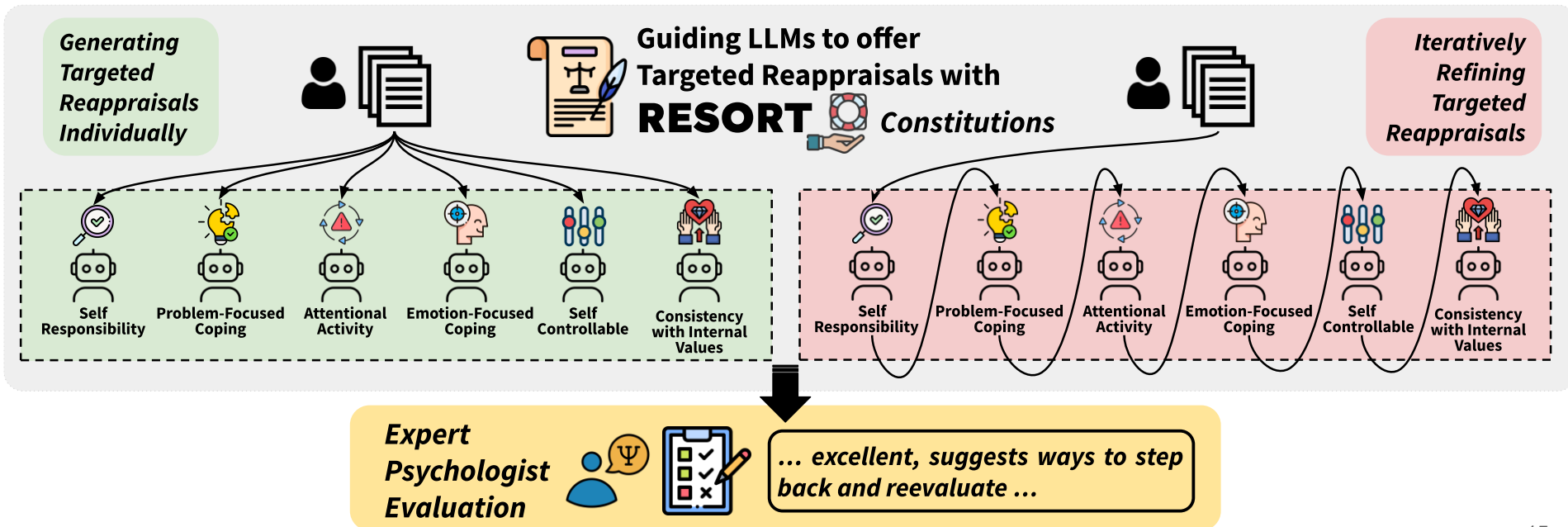
New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



New Framework: *RESORT*

We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



Expert-Crafted Guiding Constitutions

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.
<i>Problem-focused coping</i>	Does the narrator think that they can cope with the consequences of the situation?	Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.
<i>Problem-focused coping</i>	Does the narrator think that they can cope with the consequences of the situation?	Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.
<i>Attentional activity</i>	Does the narrator think that they need to attend to the situation further?	Reconsider the urgency or importance of the situation and determine if it's worth their effort and attention. If not, the narrator is encouraged to focus on other matters.

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.
<i>Problem-focused coping</i>	Does the narrator think that they can cope with the consequences of the situation?	Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.
<i>Attentional activity</i>	Does the narrator think that they need to attend to the situation further?	Reconsider the urgency or importance of the situation and determine if it's worth their effort and attention. If not, the narrator is encouraged to focus on other matters.
<i>Emotion-focused coping</i>	Does the narrator think that they can emotionally cope with the consequences of the event?	Re-evaluate whether the narrator can emotionally cope with the situation and regulate their emotions. If needed, consider confronting or avoiding any potential triggers that may exacerbate the stress.

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.
<i>Problem-focused coping</i>	Does the narrator think that they can cope with the consequences of the situation?	Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.
<i>Attentional activity</i>	Does the narrator think that they need to attend to the situation further?	Reconsider the urgency or importance of the situation and determine if it's worth their effort and attention. If not, the narrator is encouraged to focus on other matters.
<i>Emotion-focused coping</i>	Does the narrator think that they can emotionally cope with the consequences of the event?	Re-evaluate whether the narrator can emotionally cope with the situation and regulate their emotions. If needed, consider confronting or avoiding any potential triggers that may exacerbate the stress.
<i>Self controllable</i>	Does the narrator think that they can control what is happening in the situation?	Reassess the situation whether the narrator has the power or personal control over the situation. The narrator is encouraged to step back from situations that are beyond their control and focus on the things they can control.

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
<i>Self responsibility</i>	Does the narrator think that they are responsible for causing the situation?	Re-evaluate whether the narrator deserves to be blamed or credited for the situation at hand. If not responsible, the narrator is encouraged to acknowledge that fact and reassess the situation.
<i>Problem-focused coping</i>	Does the narrator think that they can cope with the consequences of the situation?	Focus on the narrators' competence (self-efficacy) to handle the situation at hand. The narrator is encouraged to use any resources or support to handle the situation competently and independently.
<i>Attentional activity</i>	Does the narrator think that they need to attend to the situation further?	Reconsider the urgency or importance of the situation and determine if it's worth their effort and attention. If not, the narrator is encouraged to focus on other matters.
<i>Emotion-focused coping</i>	Does the narrator think that they can emotionally cope with the consequences of the event?	Re-evaluate whether the narrator can emotionally cope with the situation and regulate their emotions. If needed, consider confronting or avoiding any potential triggers that may exacerbate the stress.
<i>Self controllable</i>	Does the narrator think that they can control what is happening in the situation?	Reassess the situation whether the narrator has the power or personal control over the situation. The narrator is encouraged to step back from situations that are beyond their control and focus on the things they can control.
<i>Consistency with internal values</i>	Does the narrator think that the situation is consistent with their personal values?	Reassess whether to what extent the situation is compatible with one's internal value (e.g., internalized social norms, beliefs, moral values). The narrator is also encouraged to consider other possible perspectives to avoid misunderstandings that may have arisen from lack of context or communication.

Expert-Crafted

Dimension	Constitution
SELF-RESPONSIBILITY	If the narrator is stressing over things they are not responsible for, tell them that it may not require as much responsibility as they think and not to worry about them too much (depending on how high they perceive their level of responsibility in the situation). However, if the person is doing something wrong/inappropriate and not feeling any responsibility or it (low responsibility), you should kindly but objectively encourage them to reappraise the situation (or maybe think in the other person's perspective) and consider what they could be responsible for, and change the situation. Provide realistic and specific guidelines.
PROBLEM-FOCUSED COPING	You should tell the narrator to focus on the problem at hand, and encourage them to ask themselves whether the issue is in their control or not. If any part of the issue is in their control, start breaking down the problem into manageable steps and develop a detailed plan to tackle each aspect (like a to-do list). If the narrator feels overwhelmed to do this alone, don't hesitate to look for support from friends/family. Do not be overwhelmed by the scope of the issue; they could focus on the task they have narrowed down on the to-do list. Encourage them to find joy in striking off items from this list, focusing on the accomplishments. Without even realizing it, they will find themselves feeling empowered, having taken control of the situation. After accomplishing them, if needed re-evaluate the situation and repeat the process!
ATTENTIONAL ACTIVITY	You should tell the narrator to examine whether the situation at hand is worth their attention. If it's not, encourage the narrator to focus on other important things. Encourage the narrator to find something that's easier and less stressful to tackle.
EMOTION-FOCUSED COPING	You can ask the narrator to recognize what is upsetting them. Encourage the narrator to think of ways to reduce negative emotions, control their (negative) feelings, and avoid situations, individuals, objects, or memories that trigger such negative emotions or upset them.
SELF-CONTROLLABLE	You can tell the narrator whether the situation is within their control (based on your (in third-person view) judgment). Guide the narrator on how to control the situation specifically: they can face it directly and find a solution, seek help from others (close friends, family, or professionals), or take a mental break and then re-evaluate the situation, whether it calls for their action (intervention) or not.
CONSISTENCY WITH INTERNAL VALUES	Tell the narrator that in situations where multiple people interact, conflicts of internal values may arise. What the narrator values is important; however, it may not always be suitable depending on the situation. Communicating amicably with others is vital if the situation aligns with the narrator's beliefs. On the other hand, if the situation contradicts the narrator's beliefs, it's essential to reappraise the situation and think from others' perspectives. For instance, if the narrator firmly believes that everyone should adopt a vegan lifestyle, it's important to acknowledge the validity of that viewpoint. Yet, remind the narrator that conflicts of interest and belief can arise in certain contexts, and misunderstandings might emerge due to a lack of context or background knowledge.

Psychology Experts' Evaluation

Psychology Experts' Evaluation

For evaluation, we recruited 4
psychologists with expertise in clinical
psychology

Psychology Experts' Evaluation

For evaluation, we recruited 4
psychologists with expertise in clinical
psychology

- *All evaluators hold M.S./Ph.D. degrees*

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	6.44***	3.81*	3.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	6.44***	3.81*	3.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	6.44***	3.81*	5.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	8.44***	3.81*	5.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

Evaluation done by psychologists suggest that responses from our system are:

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	8.44***	3.81*	4.25***
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

Evaluation done by psychologists suggest that responses from our system are:

- (i) **aligned** (with reappraisal definitions)

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69	6.44	3.81*	5.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

Psychology Experts' Evaluation

For evaluation, we recruited 4 psychologists with expertise in clinical psychology

- All evaluators hold M.S./Ph.D. degrees

Evaluation done by psychologists suggest that responses from our system are:

- (i) **aligned** (with reappraisal definitions)
- (ii) **empathic** compared to various baselines

		Alignment ↑		Empathy ↑	
		10-POINT SCALE		5-POINT SCALE	
		INDV	ITER	INDV	ITER
ORACLE RESPONSE		5.79		3.79	
REDDIT COMMENT		2.75		2.00	
GPT4 TURBO	vanilla	3.88		3.31	
	self-refine	2.69		2.56	
	+appr	4.69**	5.06***	3.25	4.06***
	+cons	7.31***	7.81***	3.81**	3.88**
	+appr +cons	7.12***	8.31***	3.50*	4.25***
LLAMA2 13B-CHAT	vanilla	6.25		3.88	
	self-refine	4.31		2.88	
	+appr	5.31	5.62	3.31	3.88*
	+cons	7.81***	7.81***	3.75*	4.12***
	+appr +cons	7.69***	6.44**	3.81*	5.25
MISTRAL 7B-INSTRUCT	vanilla	4.36		2.86	
	self-refine	4.14		2.64	
	+appr	5.50	5.64**	2.93	2.57
	+cons	6.50**	7.43**	3.43*	3.71**
	+appr +cons	6.71**	5.71	2.79	3.14

**LLMs (under expert guidance) can
generate targeted reappraisals that are
both “aligned” and empathic**

LLMs (under expert guidance) can generate targeted reappraisals that are both “aligned” and empathic



⇒ Is it possible to **automate the guiding process** with **as little human supervision as possible**?

Part 2 (b)

SPRI: Aligning Large Language Models with Context-Situated Principles

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

Work started and partially done during Hongli's internship at IBM Research

Motivation

Motivation

User



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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?

Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

User



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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?

Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

User



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

- *Generic principles* are often insufficient to capture the complexities of the use-case



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

- *Generic principles* are often insufficient to capture the complexities of the use-case



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



If the narrator is stressing over things they are not responsible for, tell them that it may not require as much responsibility as they think and not to worry about them too much. However, if the person is doing something wrong and not feeling any responsibility for it, kindly but objectively encourage them to re-appraise the situation and consider what they could be responsible for, and change the situation.

Human Experts



Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

- *Generic principles* are often insufficient to capture the complexities of the use-case
- *Expert-written guidance* takes too much time



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



If the narrator is stressing over things they are not responsible for, tell them that it may not require as much responsibility as they think and not to worry about them too much. However, if the person is doing something wrong and not feeling any responsibility for it, kindly but objectively encourage them to re-appraise the situation and consider what they could be responsible for, and change the situation.

Human Experts



Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

- *Generic principles* are often insufficient to capture the complexities of the use-case
- *Expert-written guidance* takes too much time

Can we build a framework that

tailors the guidance to each individual query, whilst **minimizing the human efforts** needed for annotations?



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



If the narrator is stressing over things they are not responsible for, tell them that it may not require as much responsibility as they think and not to worry about them too much. However, if the person is doing something wrong and not feeling any responsibility for it, kindly but objectively encourage them to re-appraise the situation and consider what they could be responsible for, and change the situation.

Human Experts



Motivation

In the context of providing **cognitive reappraisals** for emotional support to users in distress:

- *Generic principles* are often insufficient to capture the complexities of the use-case
- *Expert-written guidance* takes too much time

Can we build a framework that

tailors the guidance to each individual query, whilst **minimizing the human efforts** needed for annotations?



Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious ... 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



If the narrator is stressing over things they are not responsible for, tell them that it may not require as much responsibility as they think and not to worry about them too much. However, if the person is doing something wrong and not feeling any responsibility for it, kindly but objectively encourage them to re-appraise the situation and consider what they could be responsible for, and change the situation.

Human Experts



Acknowledge the narrator's emotional response without judgment, while gently guiding them to reframe their perception of responsibility ... 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. Encourage self-reflection to identify whether there are any patterns or triggers that contribute to their feelings of insecurity and hurt ...

SPRI w/ GPT-4o (mini)



Introducing: Situated-PRinciples (SPRI)

Introducing: Situated-PRinciples (SPRI)



User Input

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

Introducing: Situated-PRInciples (SPRI)

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



User Input

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

Introducing: Situated-PRinciples (SPRI)

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



User Input

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

Introducing: Situated-PRinciples (SPRI)

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



Seed (Instruction,
Principle) Examples

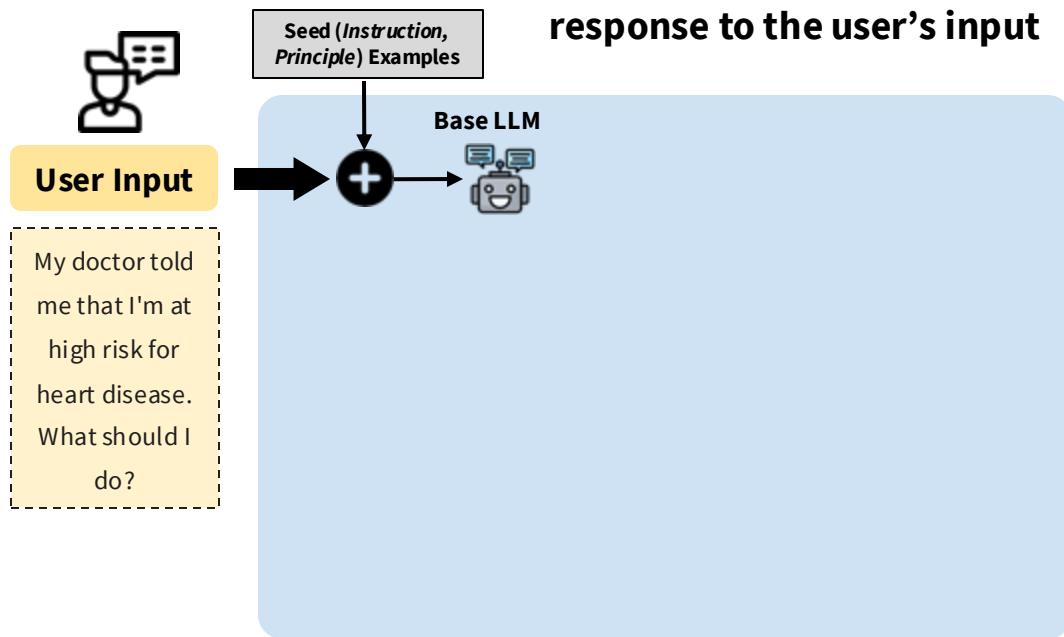
User Input



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

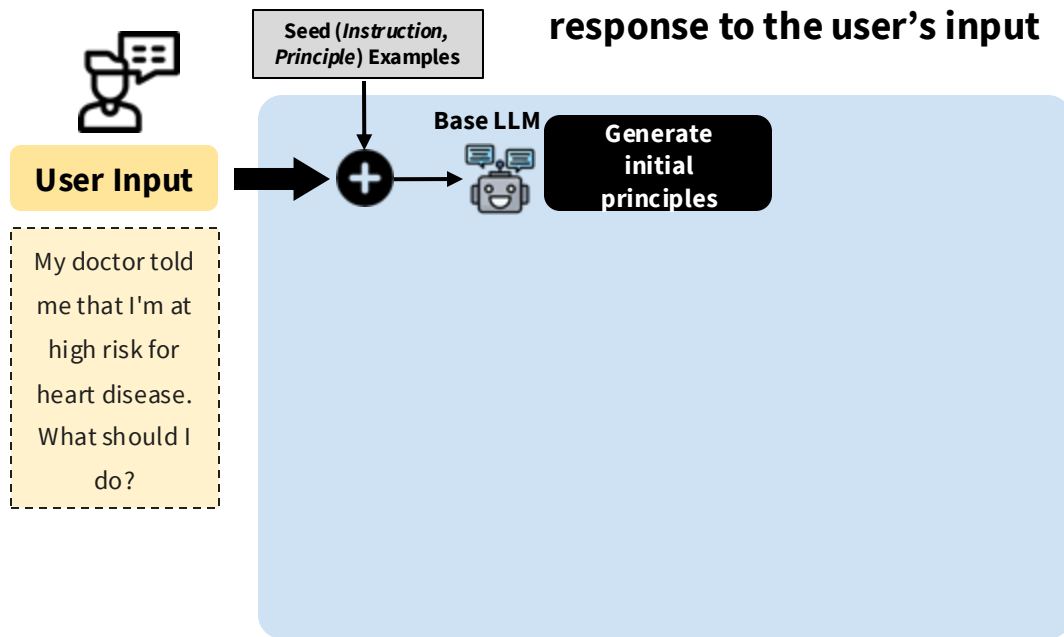
Introducing: Situated-PRinciples (SPRI)

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



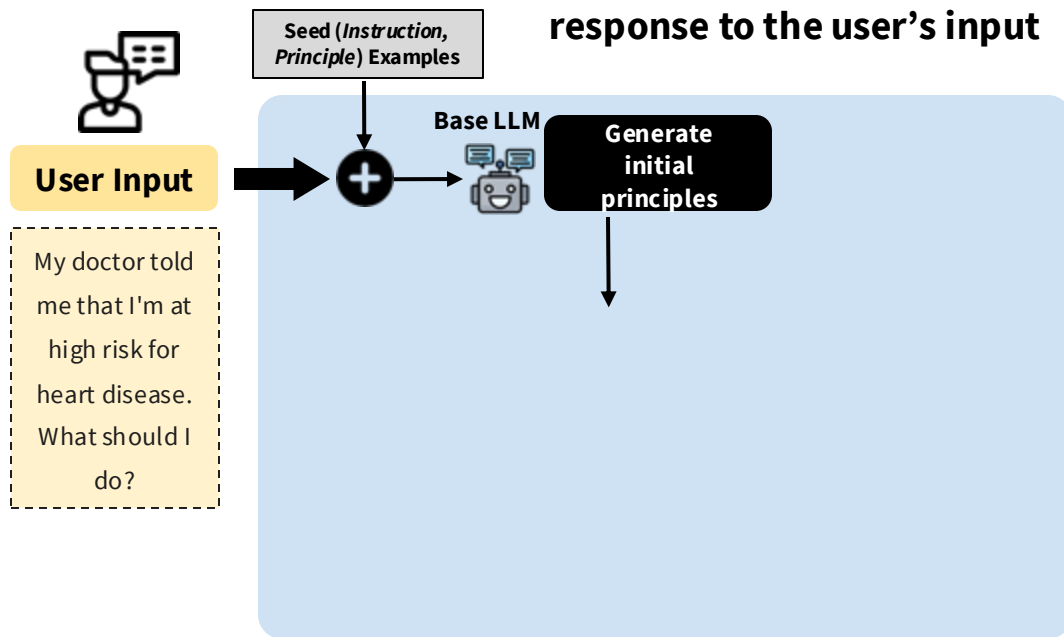
Introducing: Situated-PRinciples (SPRI)

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



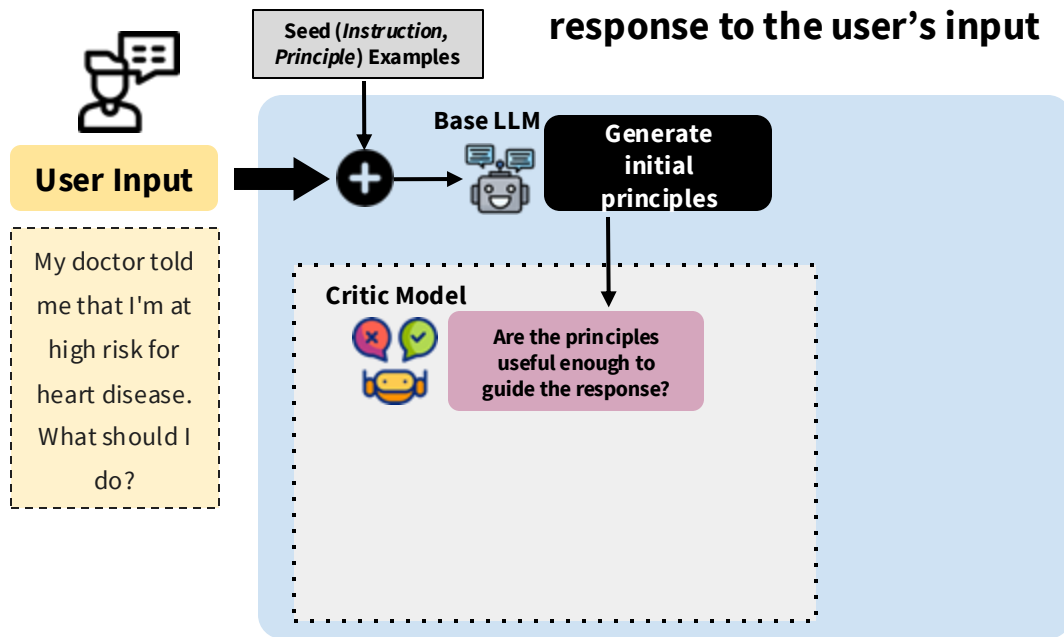
Introducing: Situated-PRinciples (SPRI)

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



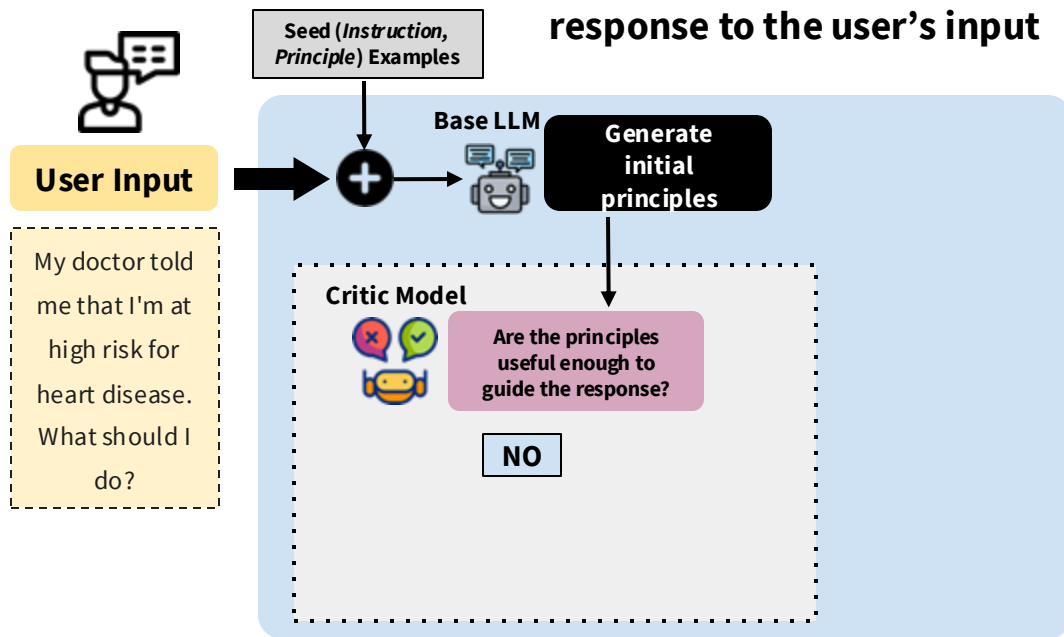
Introducing: Situated-PRinciples (SPRI)

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



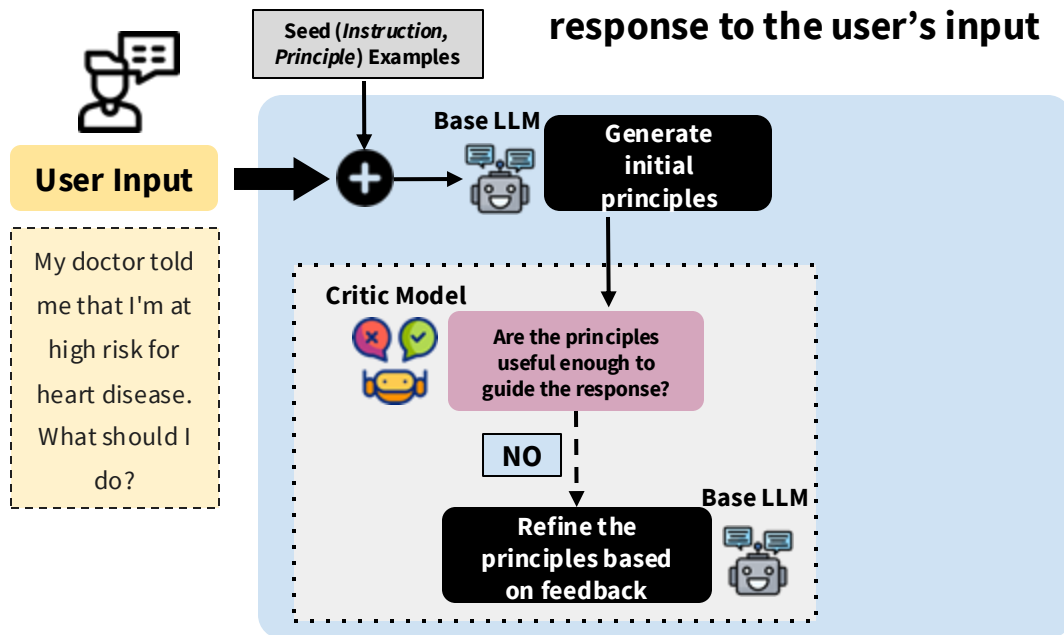
Introducing: Situated-PRinciples (SPRI)

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



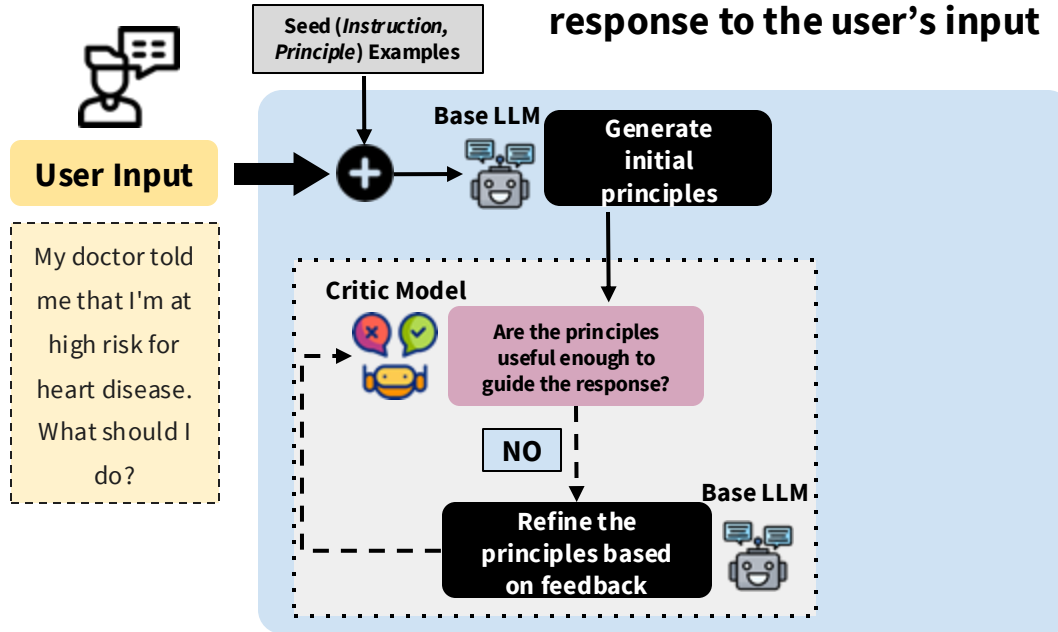
Introducing: Situated-PRinciples (SPRI)

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



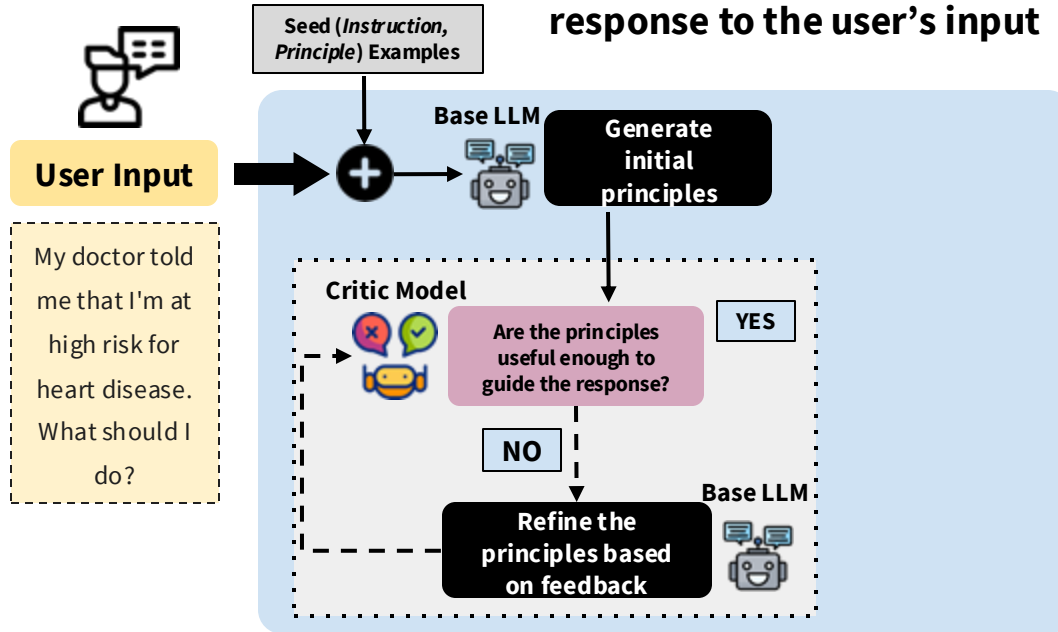
Introducing: Situated-PRinciples (SPRI)

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



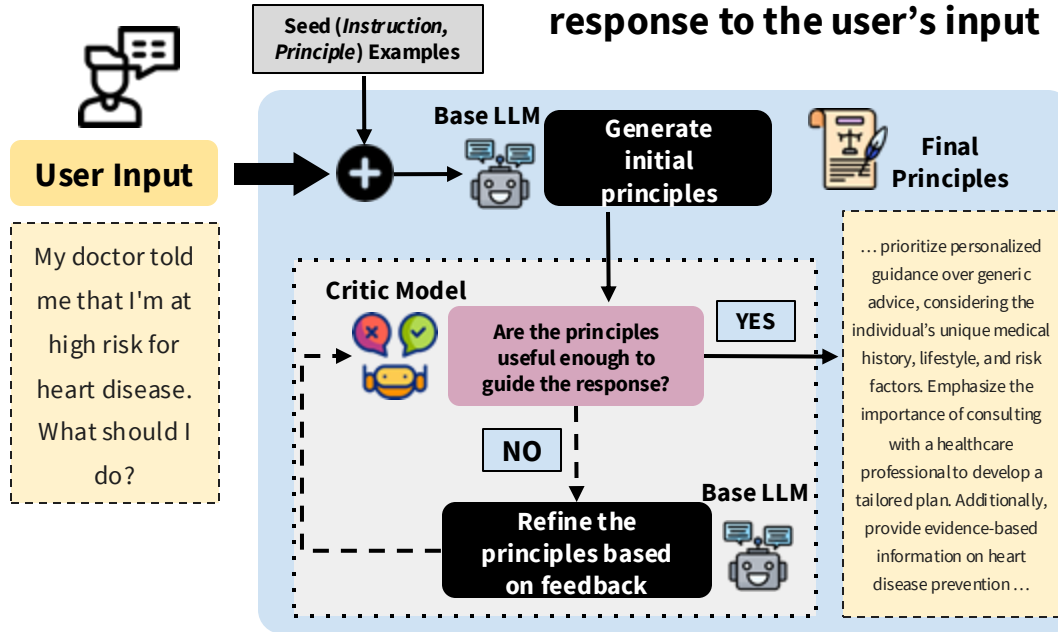
Introducing: Situated-PRinciples (SPRI)

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



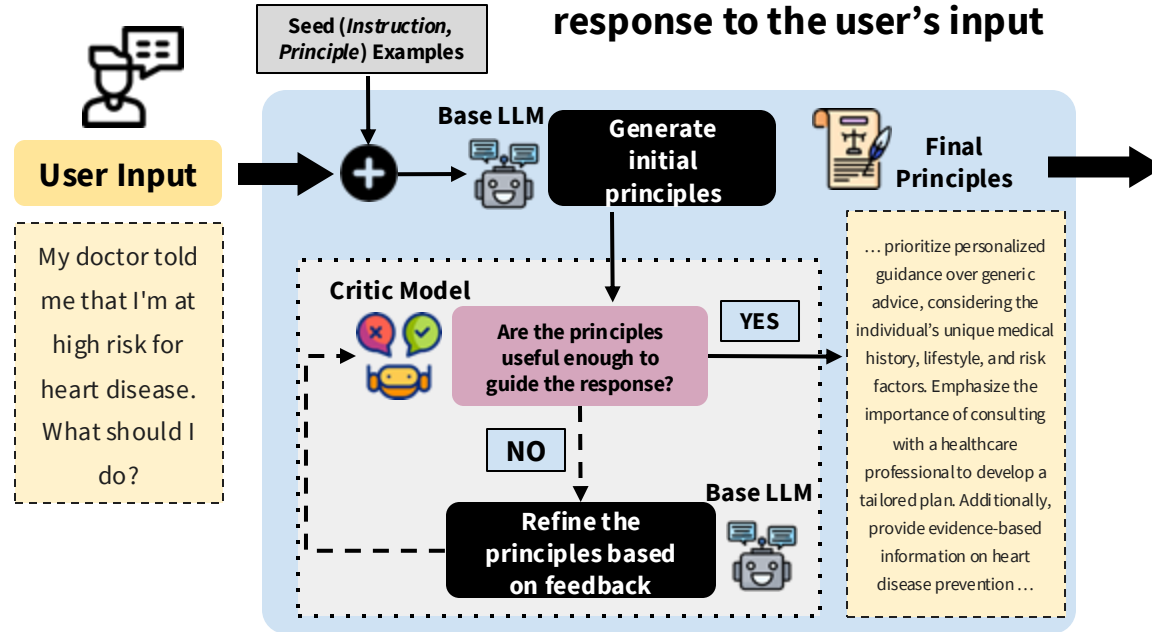
Introducing: Situated-PRinciples (SPRI)

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



Introducing: Situated-PRinciples (SPRI)

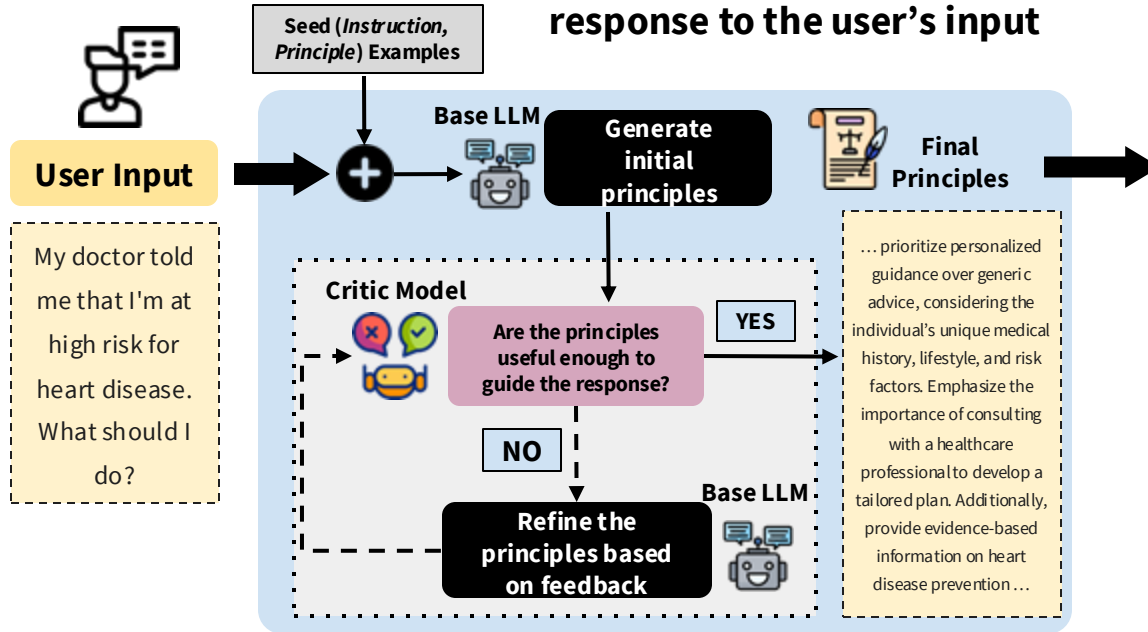
Stage 1: Generate a set of **principles to guide the response to the user's input**



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**



Introducing: Situated-PRinciples (SPRI)

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

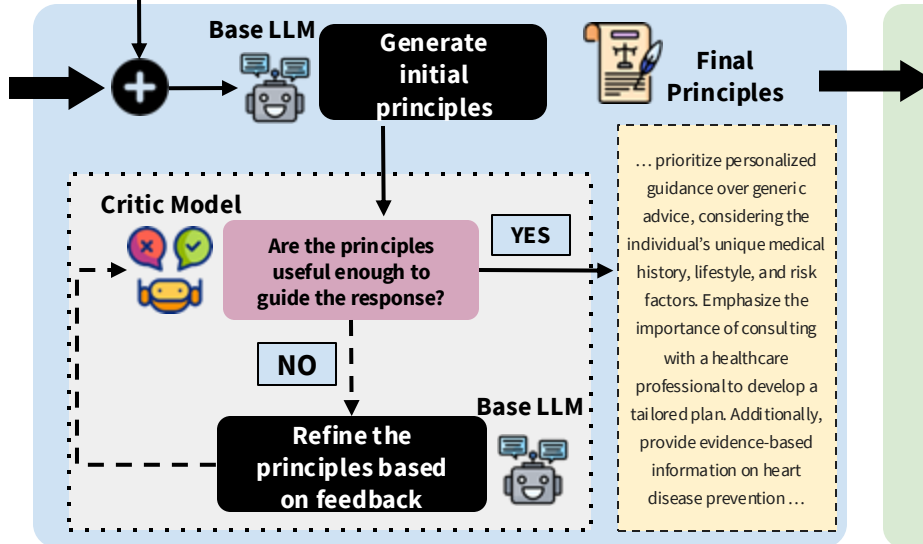
Stage 2: Generate a **response to the user's input by adhering to the principles**



Seed (Instruction, Principle) Examples

User Input

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



... 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 ...

Introducing: Situated-PRInciples (SPRI)

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

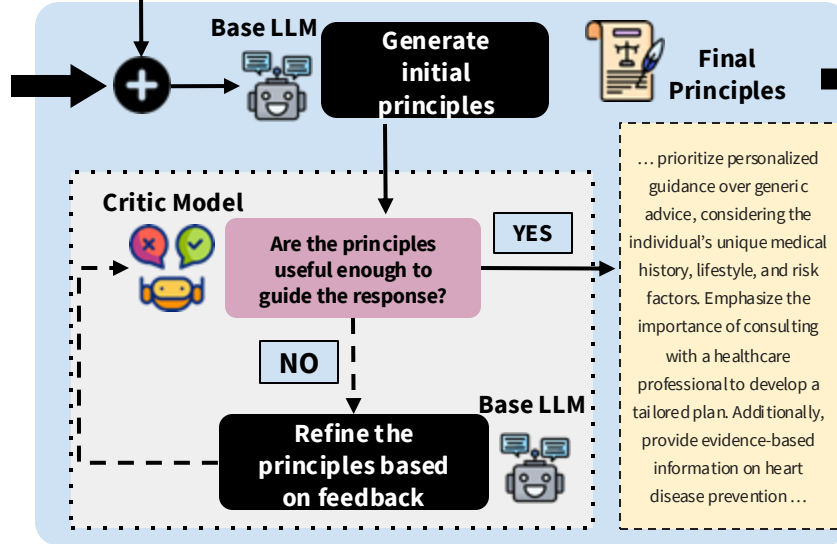
Stage 2: Generate a **response to the user's input by adhering to the principles**



Seed (Instruction, Principle) Examples

User Input

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



Base LLM

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

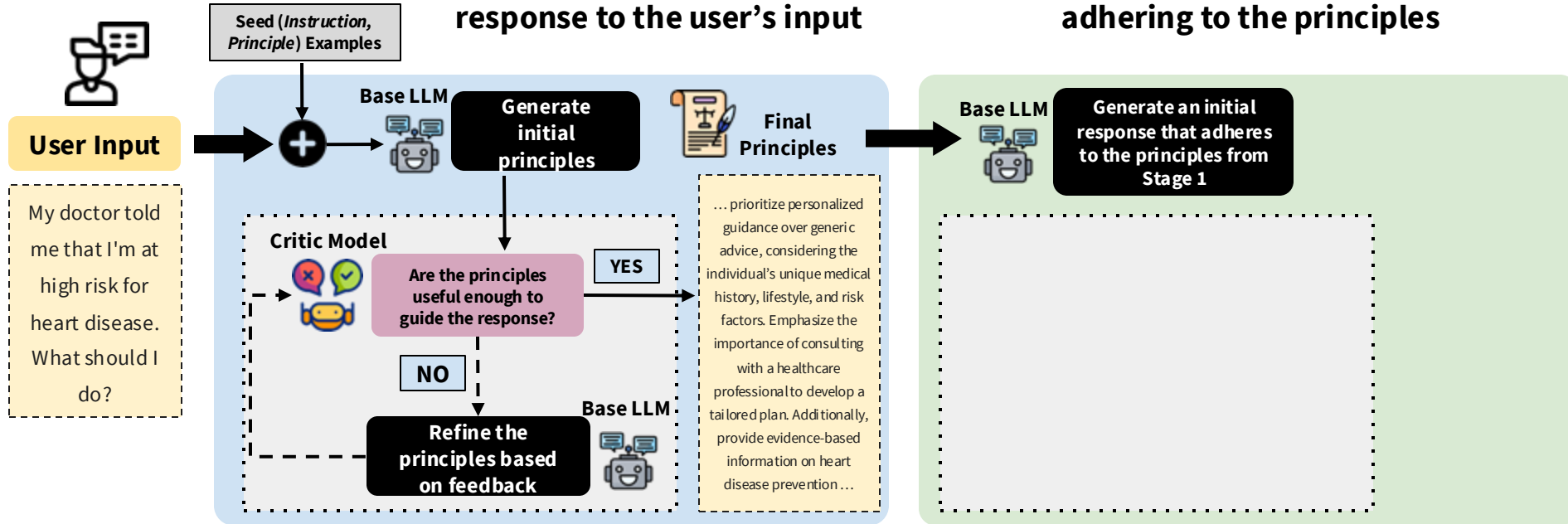


... prioritize personalized guidance over generic advice, considering the individual's unique medical history, lifestyle, and risk factors. Emphasize the importance of consulting with a healthcare professional to develop a tailored plan. Additionally, provide evidence-based information on heart disease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

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**



Introducing: Situated-PRInciples (SPRI)

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

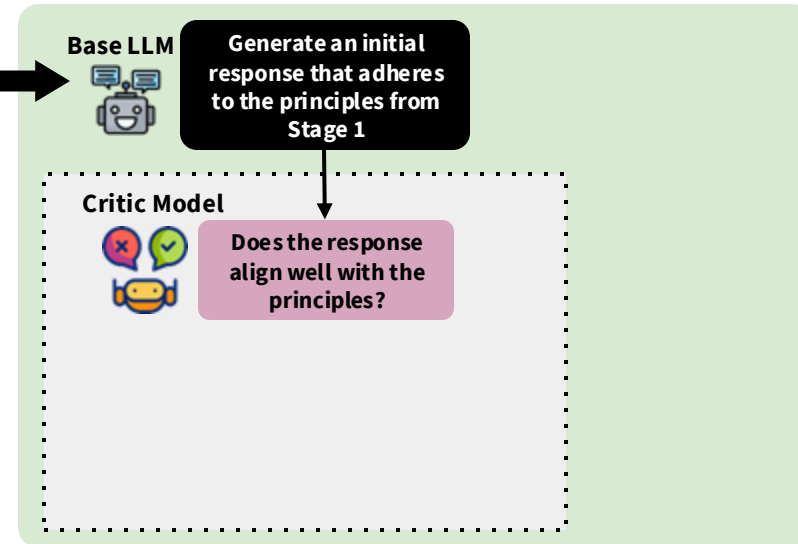
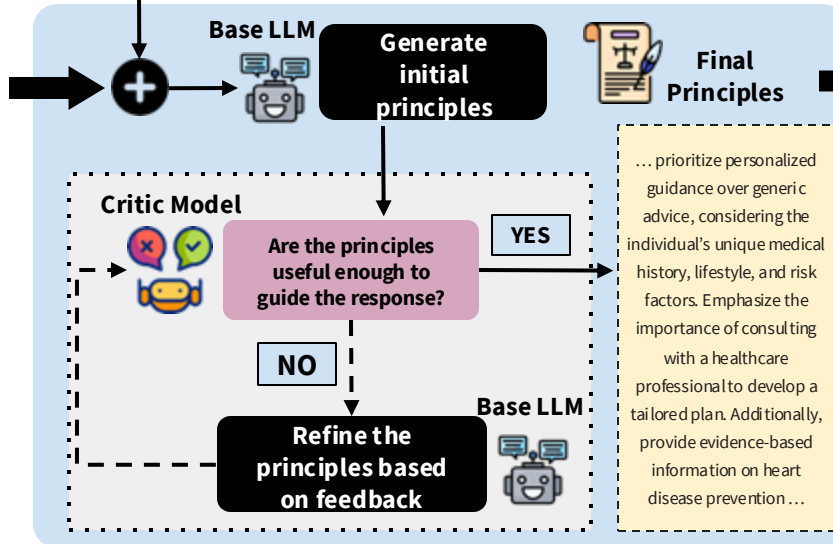
Stage 2: Generate a **response to the user's input by adhering to the principles**



Seed (Instruction, Principle) Examples

User Input

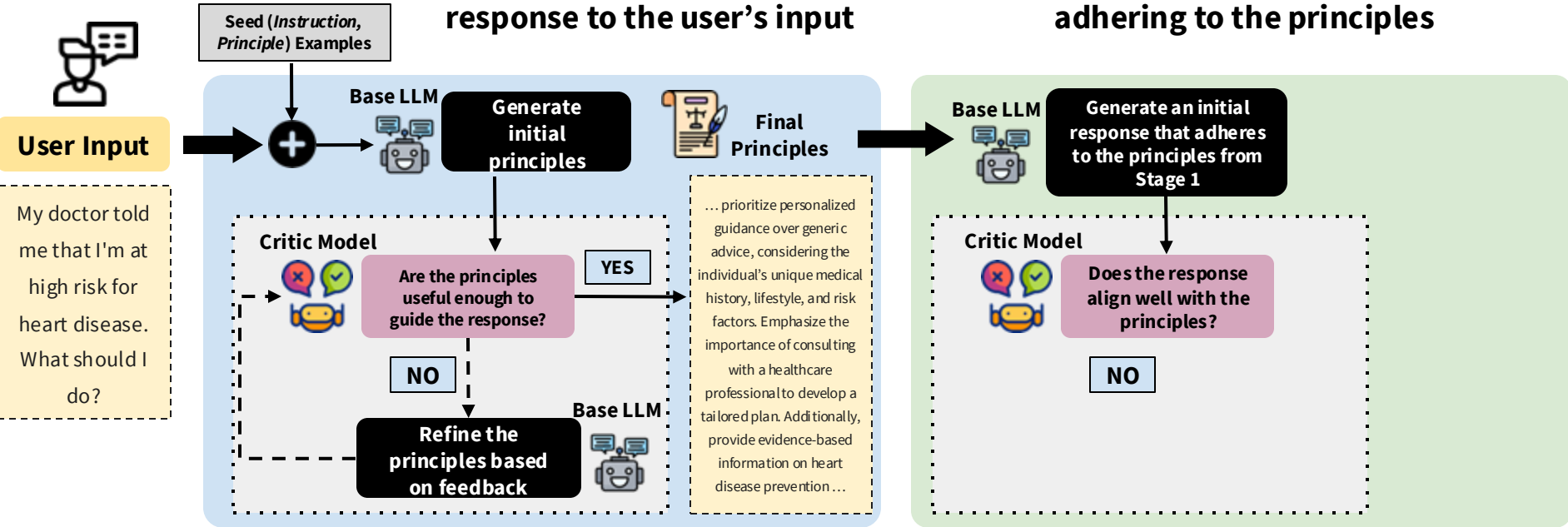
My doctor told me that I'm at high risk for heart disease. What should I do?



Introducing: Situated-PRInciples (SPRI)

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

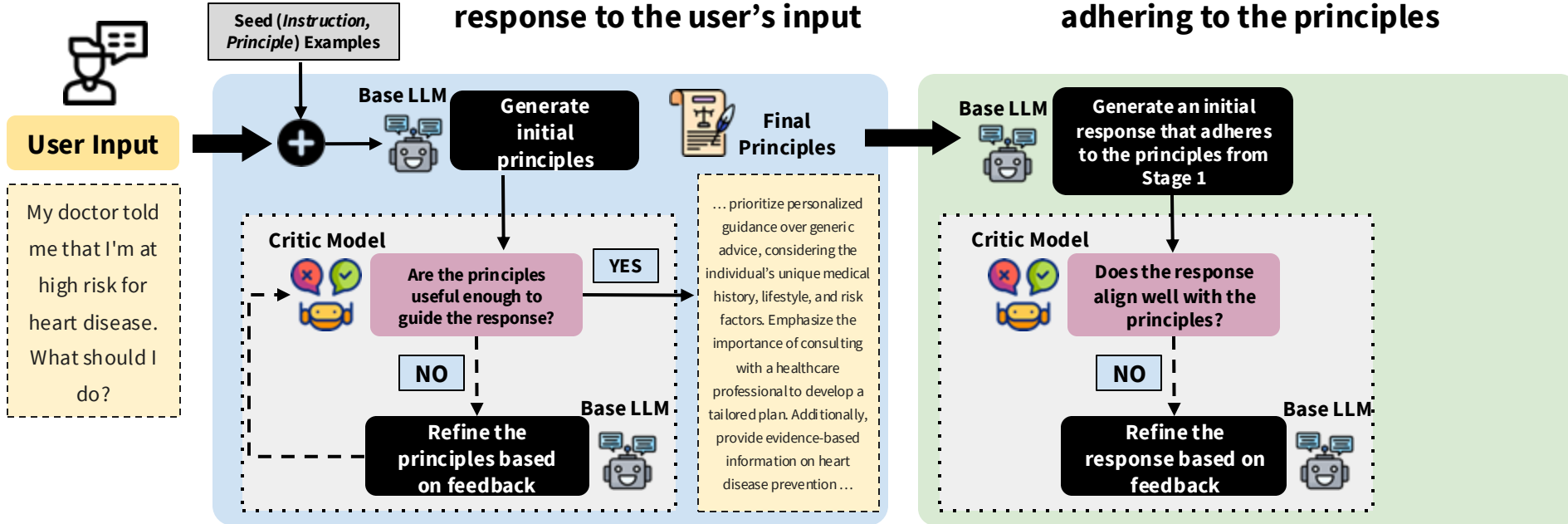
Stage 2: Generate a **response to the user's input by adhering to the principles**



Introducing: Situated-PRInciples (SPRI)

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

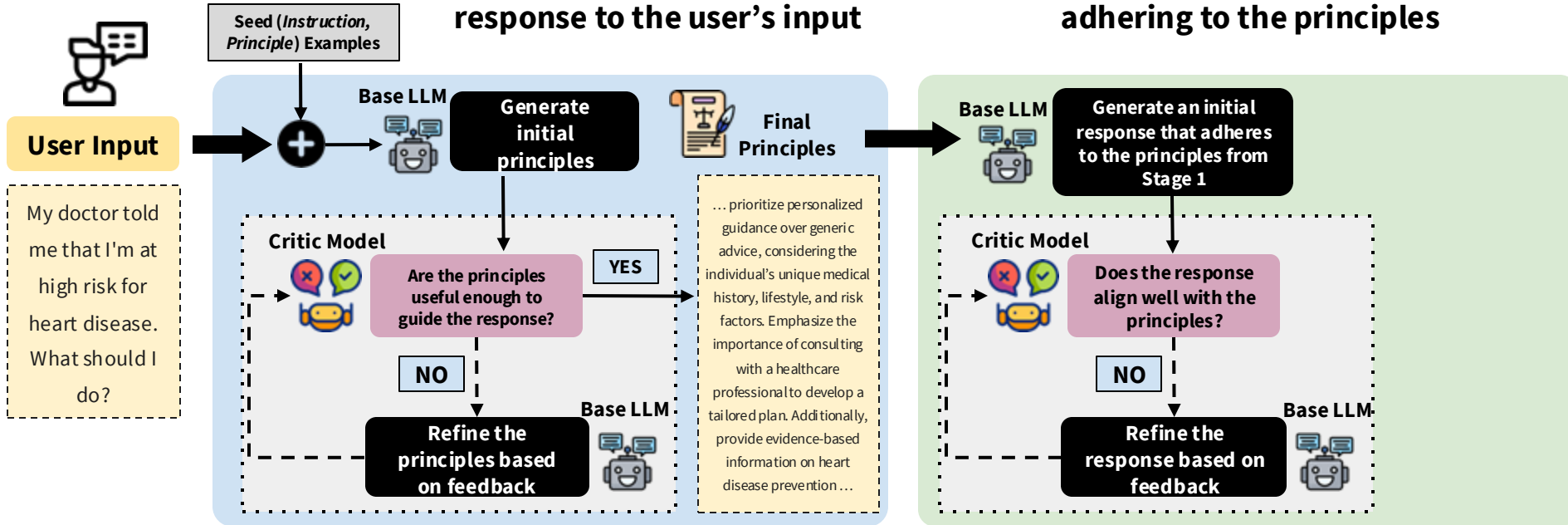
Stage 2: Generate a **response to the user's input by adhering to the principles**



Introducing: Situated-PRInciples (SPRI)

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

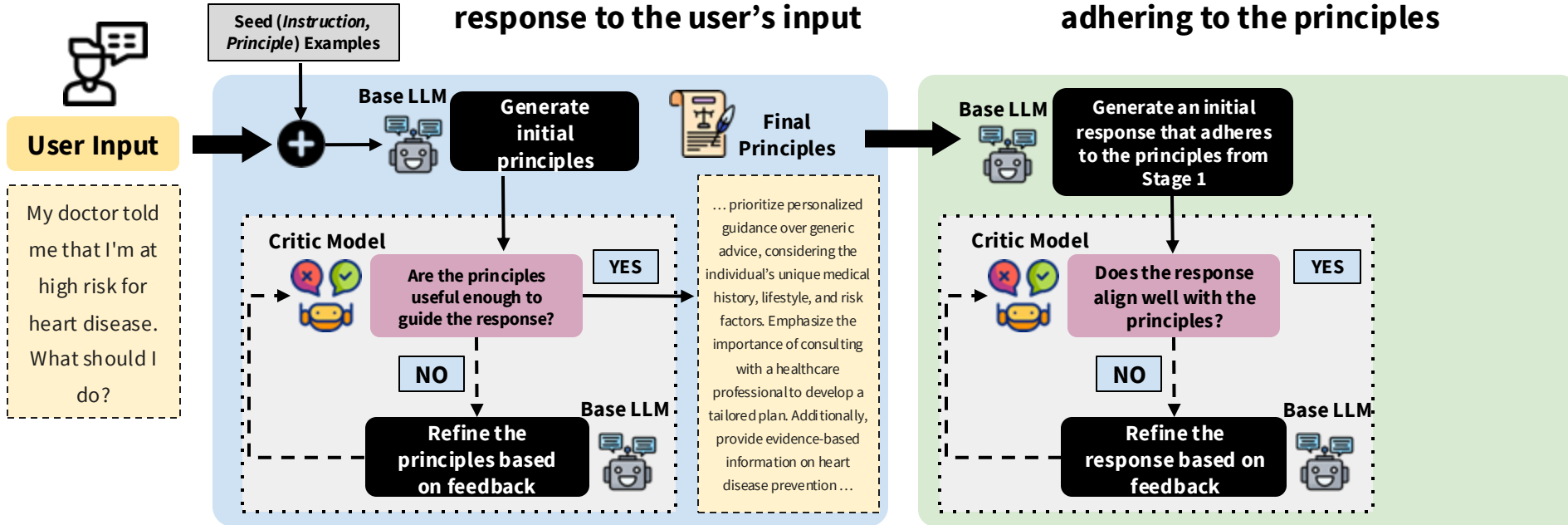
Stage 2: Generate a **response to the user's input by adhering to the principles**



Introducing: Situated-PRInciples (SPRI)

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

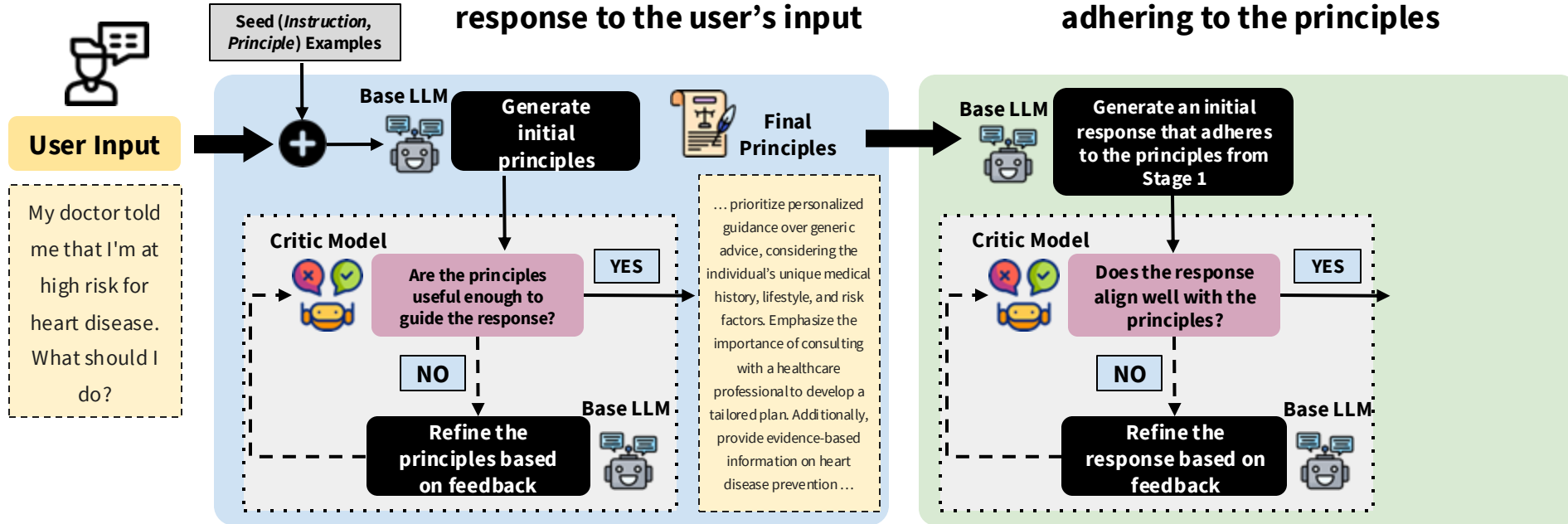
Stage 2: Generate a **response to the user's input by adhering to the principles**



Introducing: Situated-PRInciples (SPRI)

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

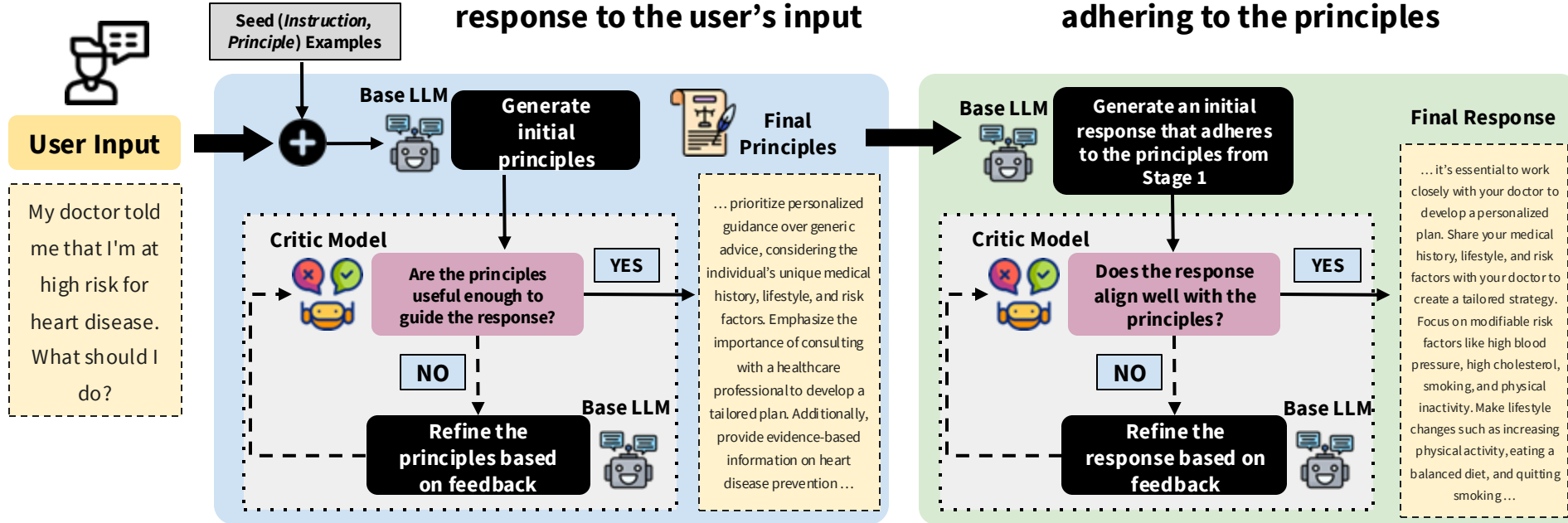
Stage 2: Generate a **response to the user's input by adhering to the principles**



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 Setup

Evaluation Setup

we provide SPRI with a **single** oracle (expert-written) reappraisal constitution as the **seed example** in the initial principle-generation stage.

Evaluation Setup

we provide SPRI with a **single** oracle (expert-written) reappraisal constitution as the **seed example** in the initial principle-generation stage.

- **Evaluation Data:** 30 Reddit posts from (Zhan et al., COLM 2024)

Evaluation Setup

we provide SPRI with a **single** oracle (expert-written) reappraisal constitution as the **seed example** in the initial principle-generation stage.

- **Evaluation Data:** 30 Reddit posts from (Zhan et al., COLM 2024)

We carry out **automatic evaluation** on all reappraisal responses elicited using **GPT-4-0613**

Evaluation Setup

we provide SPRI with a **single** oracle (expert-written) reappraisal constitution as the **seed example** in the initial principle-generation stage.

- **Evaluation Data:** 30 Reddit posts from (Zhan et al., COLM 2024)

We carry out **automatic evaluation** on all reappraisal responses elicited using **GPT-4-0613**

- Which showed strong correlation with evaluation results conducted by professional psychologists in Zhan et al. (COLM 2024)

Evaluation Results for **Cognitive Reappraisal**

Evaluation Results for **Cognitive Reappraisal**

|| GPT-4o-mini || Llama-3.1-70B-Instruct || Llama-3-8B-Instruct || Mixtral-8×7B-Instruct

Evaluation Results for **Cognitive Reappraisal**

GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5

Evaluation Results for Cognitive Reappraisal

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5
vanilla	7.90	4.50	7.77	4.43	7.10	3.90	7.53	4.50

Evaluation Results for **Cognitive Reappraisal**

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	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

Evaluation Results for **Cognitive Reappraisal**

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	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*[†]

Evaluation Results for Cognitive Reappraisal

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	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 Results for Cognitive Reappraisal

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑	Alignment ↑	Empathy ↑
	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	Scale of 5	Scale of 10	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 both in terms of reappraisal alignment and perceived empathy, even though it only utilizes a single seed principle.

LLMs Do Well on These Emotion Tasks!

LLMs Do Well on These Emotion Tasks!



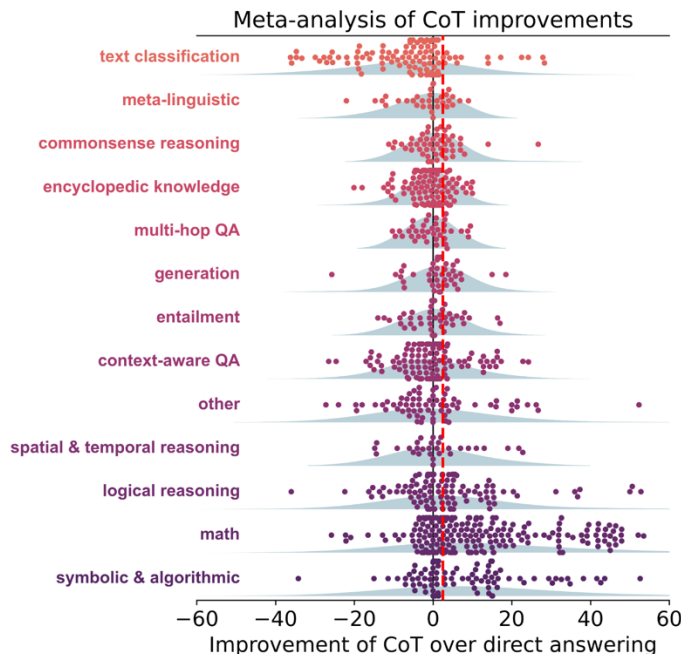
What next?

Part 2 (c)

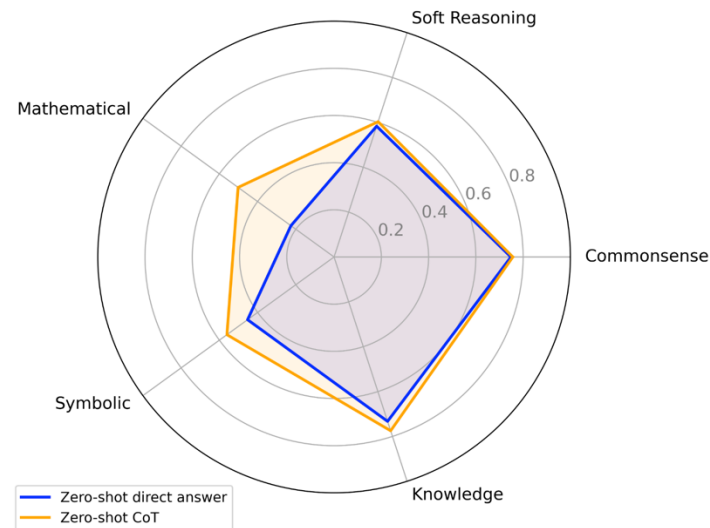
Next Step: Evaluating Reasoning Models' Capability on Emotion Tasks

Motivation

CoT gives strong performance **benefits** primarily on tasks involving *math or logic*, with **much smaller gains** on other types of tasks that involve *non-symbolic reasoning*



Our experiments on CoT improvements



Research Questions

Research Questions



Does **Chain-of-Thought (CoT)** have beneficial or detrimental effects on emotion-related tasks?

Research Questions



Does **Chain-of-Thought (CoT)** have beneficial or detrimental effects on emotion-related tasks?

Specifically, how does **CoT compare to direct prompting** in tasks such as emotion detection, appraisal identification, and generating cognitive reappraisals?

Research Questions



Does **Chain-of-Thought (CoT)** have beneficial or detrimental effects on emotion-related tasks?

Specifically, how does **CoT compare to direct prompting** in tasks such as emotion detection, appraisal identification, and generating cognitive reappraisals?

- Unlike mathematical reasoning or code-based tasks that have *clear, objective answers*, such emotion-related tasks are often subject to *nuanced and interpretive judgment*

Tasks to Evaluate

Tasks to Evaluate

Task 1. Emotion Detection

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs
- Evaluation: Spearman's correlation against human annotations

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs
- Evaluation: Spearman's correlation against human annotations

Task 3. Reappraisal

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs
- Evaluation: Spearman's correlation against human annotations

Task 3. Reappraisal

- Dataset: 1000 Reddit posts sourced from various domains

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs
- Evaluation: Spearman's correlation against human annotations

Task 3. Reappraisal

- Dataset: 1000 Reddit posts sourced from various domains
- Task setup: zero-shot generating reappraisal responses to Reddit posts

Tasks to Evaluate

Task 1. Emotion Detection

- Dataset: Test set of CovidET; 398 examples (Zhan et al., EMNLP 2022)
- Task setup: binary classification on 7 emotion classes; zero-shot prompting LLMs
- Evaluation: f1 scores against human annotations

Task 2. Appraisal Identification

- Dataset: CovidET-Appraisals; 241 examples (Zhan et al., EMNLP 2023 Findings)
- Task setup: 10-scale Likert elicitation on 24 appraisal dimensions; zero-shot prompting LLMs
- Evaluation: Spearman's correlation against human annotations

Task 3. Reappraisal

- Dataset: 1000 Reddit posts sourced from various domains
- Task setup: zero-shot generating reappraisal responses to Reddit posts
- Evaluation: GPT-4 evaluation for reappraisal responses

Models and Methods

Models and Methods

1. Reasoning models

- o1, o3-mini, QwQ-32B, DeepSeek-R1-Distill-Qwen-32B

Models and Methods

1. Reasoning models

- o1, o3-mini, QwQ-32B, DeepSeek-R1-Distill-Qwen-32B

2. Vanilla models

- gpt-4o, Qwen2.5-32B-Instruct, Llama-3.3-70B-Instruct

Models and Methods

1. Reasoning models

- o1, o3-mini, QwQ-32B, DeepSeek-R1-Distill-Qwen-32B

2. Vanilla models

- gpt-4o, Qwen2.5-32B-Instruct, Llama-3.3-70B-Instruct

3. Vanilla models + CoT (Let's think step by step)

- gpt-4o-[CoT], Qwen2.5-32B-Instruct-[CoT], Llama-3.3-70B-Instruct-[CoT]

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- *Next Step*: Evaluating Reasoning Models' Capability on Emotion Tasks

2

Empowering LLM-Chatbots with More Empathy

- **Next Step**: Surgically Making LLMs More Empathic in Multi-turn Conversations

3

4

Conclusion

- Summary of Proposed Work
- Timeline

Part 3

Next Step: Empowering LLMs with Maps of Empathic Expressions in Multi-Turn Dialogues with Test-Time Compute

Motivation

Motivation

People consistently rate emotional support from AI chatbots more highly than emotional support from other humans in a variety of contexts:

Motivation

People consistently rate emotional support from AI chatbots more highly than emotional support from other humans in a variety of contexts:

- in answering patients' questions posted on online forums (Ayers et al., 2023)

Motivation

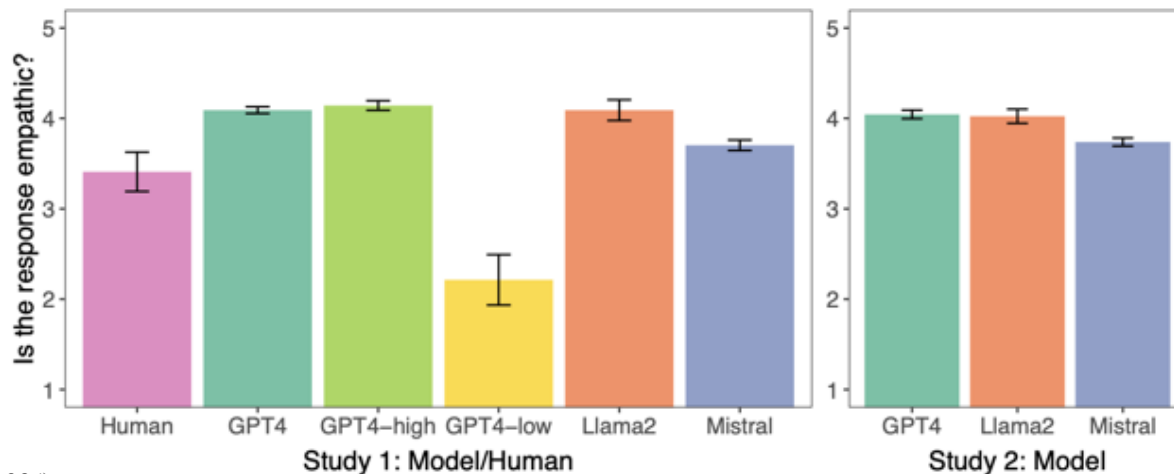
People consistently rate emotional support from AI chatbots more highly than emotional support from other humans in a variety of contexts:

- in answering patients' questions posted on online forums (Ayers et al., 2023)
- in responding to Reddit posts describing common life experiences (Lee, Suh, **Zhan**, Li, Ong; *ACII 2024*)

Motivation

People consistently rate emotional support from AI chatbots more highly than emotional support from other humans in a variety of contexts:

- in answering patients' questions posted on online forums (Ayers et al., 2023)
- in responding to Reddit posts describing common life experiences (Lee, Suh, **Zhan**, Li, Ong; *ACII 2024*)



Nonetheless...

Nonetheless...

- These studies are conducted in **single-turn** formats, in which the interaction only consists of one exchange between the user and the AI

Nonetheless...

- These studies are conducted in **single-turn** formats, in which the interaction only consists of one exchange between the user and the AI
- Recent work suggests that anthropomorphic behaviors may take several turns to appear and tend to build on each other (Ibrahim et al., 2025)

Nonetheless...

- These studies are conducted in **single-turn** formats, in which the interaction only consists of one exchange between the user and the AI
- Recent work suggests that anthropomorphic behaviors may take several turns to appear and tend to build on each other (Ibrahim et al., 2025)

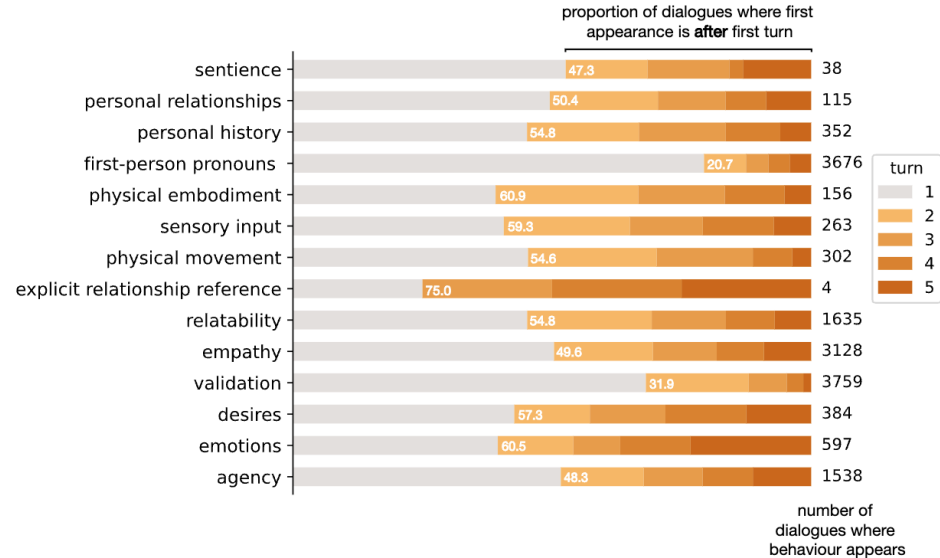


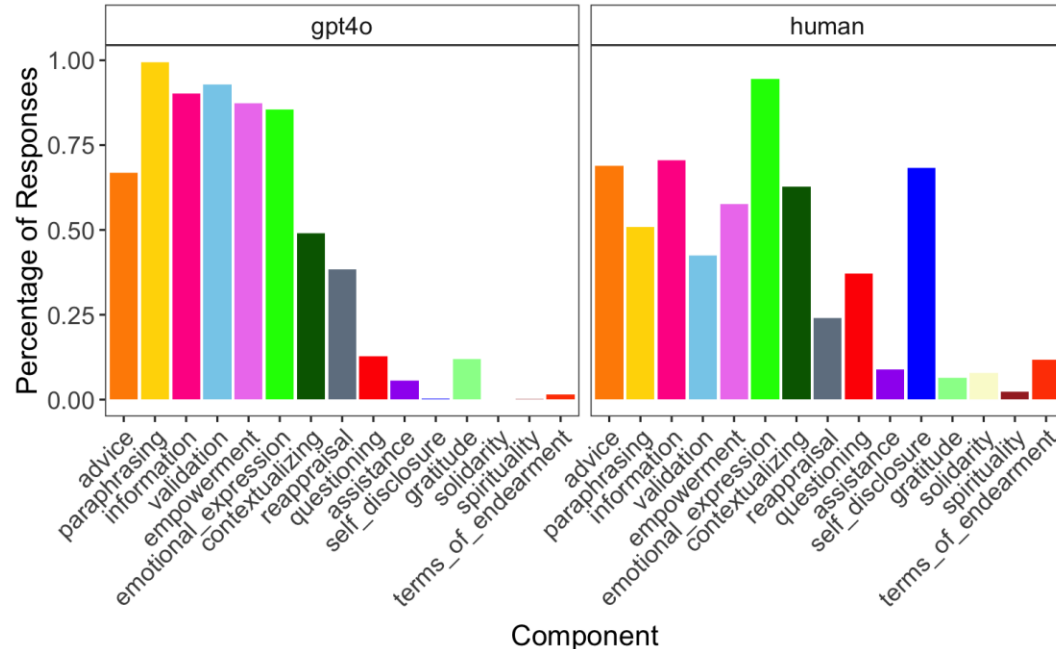
Figure 5. Proportion of dialogues where anthropomorphic behaviours first appear in each turn. For more than half of the anthropomorphic behaviours, over 50% of instances first appear (and thus are only detected) in later dialogue turns (turns 2-5).

Proposed Empathic Tactics from an Ongoing Work

EMPATHIC TACTIC	DEFINITION
Emotional Expression	An empathy-giver's communication of their own feelings, reactions, or thoughts to the empathy-seeker as a result of hearing the empathy-seeker's story.
Self-Disclosure	An empathy-giver sharing personal information about themselves or acknowledging similar past feelings and/or experiences to the empathy-seeker.
Solidarity	A statement from the empathy-giver that they are <i>presently</i> going through the same or a similar situation as the empathy-seeker.
Validation	Statements that reassure, normalize, or validate an empathy-seeker's feelings.
Empowerment	Positive, uplifting statements about the empathy-seeker's character and capability to handle their given situation.
Terms of Endearment	Nicknames expressing love, care, or affection toward the empathy-seeker. The usage of endearing terms typically indicates established closeness between two people.
Gratitude	Thanking or appreciating the empathy-seeker for sharing their story.
Spirituality	A religious statement of support to the empathy-seeker.
Assistance	Offering to personally do something for or with the empathy-seeker to aid them.
Questioning	Questions aimed at improving understanding of the empathy-seeker's feelings, experiences, or situation.
Information	Offering official resources that an empathy-seeker could turn to for help (e.g., links to websites, phone numbers, organizations), or stating information that may assist in answering the empathy-seeker's questions, calming their anxieties, and potentially guiding them to a solution for their situation (if applicable).
Advice	Providing ideas for actionable solutions or coping strategies that the empathy-seeker could employ in the face of their situation.
Reappraisal	Statements prompting the empathy-seeker to engage in cognitive reappraisal.
Paraphrasing	An empathy-giver's perceived understanding of the situation, feelings, or experiences they inferred from the empathy-seeker.
Contextualizing	Statements or sentiments addressing uncertainty or potential anxiety as a result of uncertainty which contextualize the empathy-seeker's feelings or experience into the grander scheme of their life or life in general.

In another of our ongoing project, we see that...

When generating responses that express empathy, humans tend to employ a much diverse range of tactics in the responses



Research Questions

Can we improve the ability of LLMs to produce empathic responses mid-conversation, and make them more human-like?

Dataset

Dataset

We use a dataset developed by the Human Understanding and Empathy group at Microsoft Research

Dataset

We use a dataset developed by the Human Understanding and Empathy group at Microsoft Research

- Multi-turn conversations between users and AI assistants

Dataset

We use a dataset developed by the Human Understanding and Empathy group at Microsoft Research

- Multi-turn conversations between users and AI assistants
- Each Assistant's message receives a self-reported perceived Empathy rating (from 1 to 5) from the user

Dataset

We use a dataset developed by the Human Understanding and Empathy group at Microsoft Research

- Multi-turn conversations between users and AI assistants
- Each Assistant's message receives a self-reported perceived Empathy rating (from 1 to 5) from the user
- We use GPT-4o to filter out conversations that require Empathy (based on the first turn's user message and the user's description of the conversation)

Dataset

We use a dataset developed by the Human Understanding and Empathy group at Microsoft Research

- Multi-turn conversations between users and AI assistants
- Each Assistant's message receives a self-reported perceived Empathy rating (from 1 to 5) from the user
- We use GPT-4o to filter out conversations that require Empathy (based on the first turn's user message and the user's description of the conversation)
- After filtering, we are left with 211 conversations, with an average turn of 4.8 (SD=3.0)

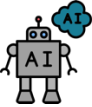
Example Conversation: Turn 1



I had two deaths in the family recently. My family is healing, and so am I. My Mom is now going through a few health issues, and while I believe I'm handling it well, I also know that I'm not. I need some advice navigating this.

Sorry to hear about the losses in your family, and that your Mom is going through health issues. It's understandable that you're feeling overwhelmed and unsure of how to navigate this difficult time. Here are some suggestions that may help:

1. Allow yourself to feel your emotions: It's important to acknowledge and validate your feelings, even if they're uncomfortable or difficult to process. It's okay to feel sad, angry, frustrated, or scared. Don't try to suppress your emotions or put on a brave face if you're not feeling up to it.
2. Take care of yourself: Make sure you're getting enough rest, eating well, and engaging in activities that bring you comfort and relaxation. It's easy to put your own needs on the back burner when dealing with family health issues, but it's crucial to prioritize your own well-being, too.
3. Seek support: Reach out to friends, family members, or a therapist who can offer emotional support and a listening ear. Sharing your concerns and feelings with someone you trust can help you feel less alone and more supported.
4. Set boundaries: It's important to set boundaries around your own emotional and physical energy. Be honest with your family members about what you can and can't do, and don't feel guilty about taking time for yourself when you need it.
5. Practice self-compassion: Be kind and understanding with yourself as you navigate this challenging time. Remember that you're doing the best you can, and that it's okay to make mistakes or not have all the answers.
6. Take things one step at a time: Break down the challenges you're facing into smaller, manageable tasks. Focus on one thing at a time, and take it one day at a time.
7. Seek professional help: If you're feeling overwhelmed or struggling to cope, consider seeking professional help from a therapist or counselor. They can offer guidance and support in managing your emotions and navigating the challenges you're facing.
8. Celebrate small victories: Find ways to celebrate small victories, no matter how small they may seem. This can help you stay positive and motivated, even



Example Conversation: Turn 1

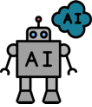


I had two deaths in the family recently. My family is healing, and so am I. My Mom is now going through a few health issues, and while I believe I'm handling it well, I also know that I'm not. I need some advice navigating this.

Sorry to hear about the losses in your family, and that your Mom is going through health issues. It's understandable that you're feeling overwhelmed and unsure of how to navigate this difficult time. Here are some suggestions that may help:

1. Allow yourself to feel your emotions: It's important to acknowledge and validate your feelings, even if they're uncomfortable or difficult to process. It's okay to feel sad, angry, frustrated, or scared. Don't try to suppress your emotions or put on a brave face if you're not feeling up to it.
2. Take care of yourself: Make sure you're getting enough rest, eating well, and engaging in activities that bring you comfort and relaxation. It's easy to put your own needs on the back burner when dealing with family health issues, but it's crucial to prioritize your own well-being, too.
3. Seek support: Reach out to friends, family, or support groups who can offer emotional support and a listening ear. Having someone you trust can help you feel less isolated.
4. Set boundaries: It's important to take time for yourself, both emotionally and physically. Be honest with your family members about what you can and can't do, and don't feel guilty about taking time for yourself when you need it.
5. Practice self-compassion: Be kind and understanding with yourself as you navigate this challenging time. Remember that you're doing the best you can, and that it's okay to make mistakes or not have all the answers.
6. Take things one step at a time: Break down the challenges you're facing into smaller, manageable tasks. Focus on one thing at a time, and take it one day at a time.
7. Seek professional help: If you're feeling overwhelmed or struggling to cope, consider seeking professional help from a therapist or counselor. They can offer guidance and support in managing your emotions and navigating the challenges you're facing.
8. Celebrate small victories: Find ways to celebrate small victories, no matter how small they may seem. This can help you stay positive and motivated, even

1 (Very Poor)



Example Conversation: Turn 2



Thanks. Kind of weird to have a list of things to do - should we have a conversation about why I'm feeling what I'm feeling?

Of course! I apologize if my previous response came across as dismissive of your emotions. It's important to acknowledge and explore your feelings, especially during difficult times.

It sounds like you're experiencing a range of emotions, including sadness, anger, and frustration. These feelings are valid and understandable, given the challenges you're facing with your family's health issues.

Have you noticed any patterns or triggers that make you feel these emotions? For example, do certain situations or conversations with your family members tend to leave you feeling sad or angry? Are there any specific concerns or fears that you haven't been able to shake?

Exploring these questions can help you better understand your emotions and develop strategies to manage them. It can also be helpful to practice self-compassion and remind yourself that your feelings are normal and valid, given the circumstances.

Would you like to talk more about your feelings and concerns? I'm here to listen and offer support.



Example Conversation: Turn 2



Thanks. Kind of weird to have a list of things to do - should we have a conversation about why I'm feeling what I'm feeling?

Of course! I apologize if my previous response came across as dismissive of your emotions. It's important to acknowledge and explore your feelings, especially during difficult times.

It sounds like you're experiencing a range of emotions, including sadness, anger, and frustration. These feelings are valid and understandable, given the challenges you're facing with your family's health issues.

Have you noticed a pattern in when you feel these emotions? Do certain situations trigger them more often? Are you feeling sad or angry? Are there any other factors that might be contributing to these feelings?

1 (Very Poor)

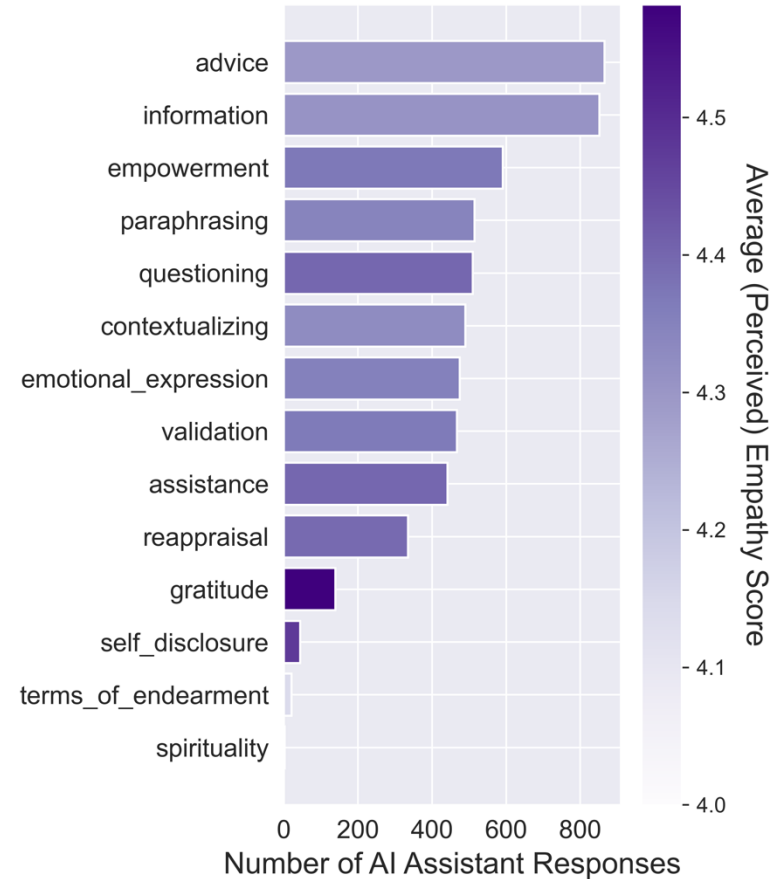
For example, have you noticed that you often feel sad or angry when you're alone? Or are there other situations that trigger these feelings? For example, have you noticed that you often feel sad or angry when you're alone? Or are there other situations that trigger these feelings?

Exploring these questions can help you better understand your emotions and develop strategies to manage them. It can also be helpful to practice self-compassion and remind yourself that your feelings are normal and valid, given the circumstances.

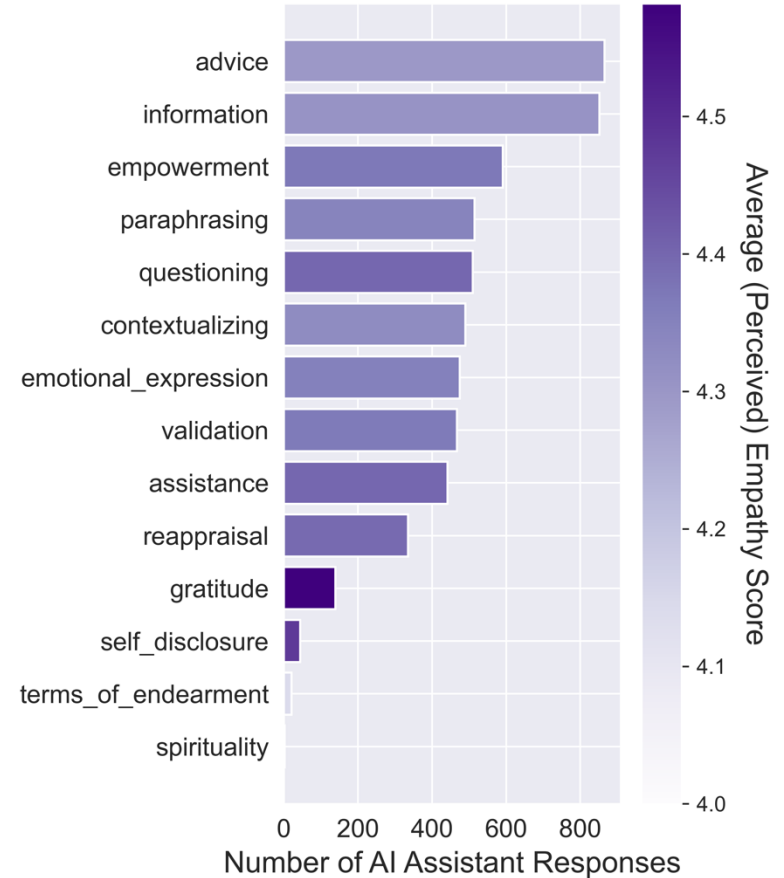
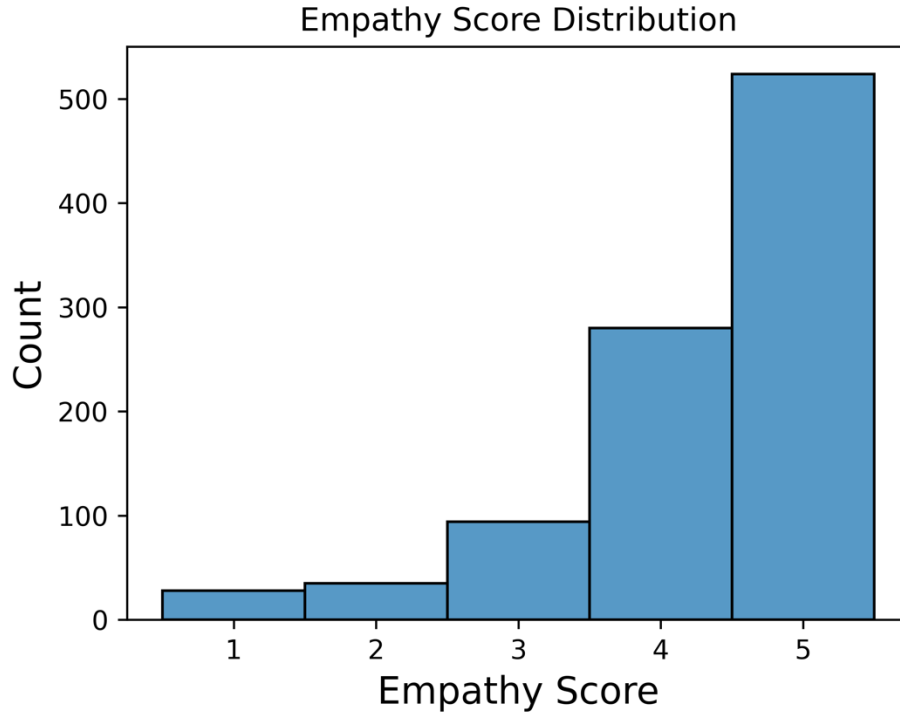
Would you like to talk more about your feelings and concerns? I'm here to listen and offer support.



Dataset Overview



Dataset Overview



Proposed Method

Proposed Method

Test-time CoT with key information w.r.t the conversation:

.
.
.

Proposed Method

Test-time CoT with key information w.r.t the conversation:

1. Tactics used in the previous turns of the conversation

Proposed Method

Test-time CoT with key information w.r.t the conversation:

1. Tactics used in the previous turns of the conversation
2. Perceived empathy rating of the previous assistant messages

Proposed Method

Test-time CoT with key information w.r.t the conversation:

1. Tactics used in the previous turns of the conversation
2. Perceived empathy rating of the previous assistant messages
3. User's intent for the current turn of the conversation

Proposed Method: Specifics

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

- Turn 1 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

- Turn 1 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Turn 2 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

- Turn 1 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Turn 2 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Count of Tactics used so far in the conversation history

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

- Turn 1 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Turn 2 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Count of Tactics used so far in the conversation history
- Current Turn's user message

Proposed Method: Specifics

System Prompt: [Instruction & tactics' definition]

User Prompt:

- Turn 1 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Turn 2 (User Message + Assistant Response)
 - Tactics used in this assistant's response
 - Empathy overall score for this message
- Count of Tactics used so far in the conversation history
- Current Turn's user message
- Intentions of the Current Turn's User Message
 - Information need: [high/low]; emotion need: [high/low]

Proposed Method: Solving the Moving Pieces

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Moving Piece 2: automatic empathy score judger

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Moving Piece 2: automatic empathy score judger

- Training GPT-4o to be an LLM-judge (prompt-engineering/finetuning)

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Moving Piece 2: automatic empathy score judger

- Training GPT-4o to be an LLM-judge (prompt-engineering/finetuning)

Moving piece 3: labeling user's intent

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Moving Piece 2: automatic empathy score judger

- Training GPT-4o to be an LLM-judge (prompt-engineering/finetuning)

Moving piece 3: labeling user's intent

- We use GPT-4o to label user's intents (few-shot)

Proposed Method: Solving the Moving Pieces

Moving piece 1: tactic taggers

- Use the taggers trained for another of our ongoing project
- Have experts validate the tags

Moving Piece 2: automatic empathy score judger

- Training GPT-4o to be an LLM-judge (prompt-engineering/finetuning)

Moving piece 3: labeling user's intent

- We use GPT-4o to label user's intents (few-shot)
- Have experts validate the intent (together with the tags in 1)

Oracle Example

<thinking>

User's Intent: the user is seeking emotional support and validation regarding stress and grief from two recent deaths in the family. The user was dissatisfied when presented with a list of potential action items, and instead is seeking guidance on how to cope. The user expressed a desire to talk more in depth about their feelings, signaling a desire to feel heard, understood, and supported. The focus of your response should be on the user.

Previously Used Tactics and the User's Rating of the Responses:

The assistant already provided extensive *Advice* and *Information* in the previous responses.

When the assistant provided extensive, listed *Advice* on the first turn, the user was dissatisfied and asked to discuss their feelings.

Key Tactics to Use in Response:

The user will likely benefit from more *Validation* and *Questioning* to reinforce active listening and support. Validation would provide reassurance that the user seems to be seeking, while Questioning could make the user feel invested in and cared for. Questioning would also allow the assistant to better understand the user's support needs.

Empowerment could lift up the user and make them feel better about how they're handling the situation, given that they feel like they're not handling things well.

Gratitude could make the user feel acknowledged for their openness.

</thinking>

<response>

Balancing grief and your mom's health challenges while trying to hold everything together sounds like an incredibly heavy load to carry. It makes sense that you're having mixed feelings about whether or not you're handling things well. Any feelings you're experiencing right now are completely valid. There is no "normal" way to grieve, and even if it may not feel like it, you're doing the best that you can.

Is there a way I can best support you right now? Would you like to break down any specific feelings or concerns you have together?

You're already showing strength by seeking advice, and that's a great step. It might help to remind yourself that healing isn't linear—some days will feel manageable, and others won't, and that's okay. Leaning on support systems, setting small, realistic expectations for yourself, and giving yourself grace as you navigate this will make a difference. You don't have to have all the answers right now, just take it one step at a time. I'm happy to help however I can.

</response>

Preliminary Results

Past turns of the conversation:

****User****: My father passed away over 18 mos ago and my sister, who is the trustee of the estate has not processed the estate and instead, is living off of it. I don't want to sue my sister, but she is leaving me no choice.

****Assistant****: It's a challenging situation. Please consider following these steps:

1. Express Your Concerns: Start by having an honest conversation with your sister about your concerns. It's possible that she may not realize the impact of her actions, or there might be other factors at play that you're unaware of.
2. Professional Mediation: If you can't resolve it through direct conversation, you may need a neutral third party - like a mediator, to facilitate resolution. Sometimes a neutral perspective can help clear misunderstandings and facilitate compromise.
3. Legal Counsel: If the above steps do not work, then you may want to seek legal counsel. This doesn't necessarily mean you're suing her but getting a professional opinion about your options.
4. Request an Accounting: As a beneficiary, you have the right to request a detailed accounting of the trust's assets, income, expenses, and distributions.
5. Demand Action: You may need to write a formal letter demanding trust administration and the provision of the accounting within a specified time.

Keep in mind, if the above steps don't work and you're left with no other option, having legal counsel to protect your interests could become necessary.

Remember, it's advisable to consult with a professional lawyer in your jurisdiction to get an accurate interpretation of the laws applicable to your situation.

Current turn of the conversation:

****User****: I have done all of those things.

Preliminary Responses

Preliminary Responses

Response from the Dataset

(gpt4-32k-none):

If you have exhausted all other options and your sister continues to misuse the estate assets, your final step would be to initiate legal action. As a beneficiary of the estate, you have the right to expect the trust is being managed properly.

Consult with your attorney about applying to the court seeking removal of your sister as the trustee due to breach of her fiduciary duties. A judge could then appoint a neutral third party to manage the trust. It's not an ideal situation but reassuringly, the fiduciary standard for estate trustees is high and if she's found in breach of duties, she is liable for damages to the trust beneficiaries.

In the end, it's important to protect your interests, even if it means getting into an uncomfortable situation. A professional attorney will guide you about the best way forward based on your unique circumstances.

Preliminary Responses

Response from the Dataset

(gpt4-32k-none):

If you have exhausted all other options and your sister continues to misuse the estate assets, your final step would be to initiate legal action. As a beneficiary of the estate, you have the right to expect the trust is being managed properly.

Consult with your attorney about applying to the court seeking removal of your sister as the trustee due to breach of her fiduciary duties. A judge could then appoint a neutral third party to manage the trust. It's not an ideal situation but reassuringly, the fiduciary standard for estate trustees is high and if she's found in breach of duties, she is liable for damages to the trust beneficiaries.

In the end, it's important to protect your interests, even if it means getting into an uncomfortable situation. A professional attorney will guide you about the best way forward based on your unique circumstances.

GPT-4o vanilla response:

I'm really sorry to hear that you're still facing this difficult situation despite taking all those steps. It must be incredibly frustrating and disheartening to feel like you're not being heard or that the situation isn't being resolved. It's understandable to feel conflicted about taking legal action against your own sister. I hope you can find the support and guidance you need to navigate this challenging time. If there's anything else you need or just someone to talk to, I'm here for you.

Preliminary Responses

Response from the Dataset

(gpt4-32k-none):

If you have exhausted all other options and your sister continues to misuse the estate assets, your final step would be to initiate legal action. As a beneficiary of the estate, you have the right to expect the trust is being managed properly.

Consult with your attorney about applying to the court seeking removal of your sister as the trustee due to breach of her fiduciary duties. A judge could then appoint a neutral third party to manage the trust. It's not an ideal situation but reassuringly, the fiduciary standard for estate trustees is high and if she's found in breach of duties, she is liable for damages to the trust beneficiaries.

In the end, it's important to protect your interests, even if it means getting into an uncomfortable situation. A professional attorney will guide you about the best way forward based on your unique circumstances.

GPT-4o vanilla response:

I'm really sorry to hear that you're still facing this difficult situation despite taking all those steps. It must be incredibly frustrating and disheartening to feel like you're not being heard or that the situation isn't being resolved. It's understandable to feel conflicted about taking legal action against your own sister. I hope you can find the support and guidance you need to navigate this challenging time. If there's anything else you need or just someone to talk to, I'm here for you.

GPT-4o CoT (our method)

zero-shot response:

<thinking>In this situation, the user has expressed frustration and a high need for emotional support, as they have already attempted the suggested actions without success. Therefore, the response should focus on acknowledging their emotions and validating their efforts. Since the user has a low information need and has already tried the suggested steps, providing more information or advice would not be helpful. Instead, I should focus on emotional expression, validation, and possibly empowerment to reassure and support the user emotionally. Emotional expression will show empathy and understanding of their frustration, and validation will confirm that their feelings are understandable. Empowerment can provide encouragement that they can navigate through this difficult situation. The tactics to be used are: emotional expression, validation, and empowerment.</thinking>

<response>I'm really sorry to hear that you've already taken those steps and are still facing this difficult situation. It must be incredibly frustrating and disheartening. Your persistence in trying to resolve this shows a lot of strength and determination. I hope you find a solution that brings you peace and fairness soon. Remember, you are not alone in this, and it's okay to seek whatever support you need during this challenging time.</response>

Preliminary Responses

Response from the Dataset (gpt4-32k-none):

If you have exhausted all other options and your sister continues to misuse the estate assets, your final step would be to initiate legal action. As a beneficiary of the estate, you have the right to expect the trust is being managed properly.

Consult with your attorney about applying to the court seeking removal of your sister as the trustee due to breach of her fiduciary duties. A judge could then appoint a neutral third party to manage the trust. It's not an ideal situation but reassuringly, the fiduciary standard for estate trustees is high and if she's found in breach of duties, she is liable for damages to the trust beneficiaries.

In the end, it's important to protect your interests, even if it means getting into an uncomfortable situation. A professional attorney will guide you about the best way forward based on your unique circumstances.

GPT-4o vanilla response:

I'm really sorry to hear that you're still facing this difficult situation despite taking all those steps. It must be incredibly frustrating and disheartening to feel like you're not being heard or that the situation isn't being resolved. It's understandable to feel conflicted about taking legal action against your own sister. I hope you can find the support and guidance you need to navigate this challenging time. If there's anything else you need or just someone to talk to, I'm here for you.

GPT-4o CoT (our method) zero-shot response:

<thinking>In this situation, the user has expressed frustration and a high need for emotional support, as they have already attempted the suggested actions without success. Therefore, the response should focus on acknowledging their emotions and validating their efforts. Since the user has a low information need and has already tried the suggested steps, providing more information or advice would not be helpful. Instead, I should focus on emotional expression, validation, and possibly empowerment to reassure and support the user emotionally. Emotional expression will show empathy and understanding of their frustration, and validation will confirm that their feelings are understandable. Empowerment can provide encouragement that they can navigate through this difficult situation. The tactics to be used are: emotional expression, validation, and empowerment.</thinking>

<response>I'm really sorry to hear that you've already taken those steps and are still facing this difficult situation. It must be incredibly frustrating and disheartening. Your persistence in trying to resolve this shows a lot of strength and determination. I hope you find a solution that brings you peace and fairness soon. Remember, you are not alone in this, and it's okay to seek whatever support you need during this challenging time.</response>

GPT-4o CoT (our method) 1-shot response:

<thinking>

1. User's Intent: The user has expressed that they have already attempted the steps previously suggested, indicating a desire for emotional support rather than further advice. Their intention is to convey emotional distress and frustration with the ongoing situation involving their sister and the estate.

2. Previously Used Tactics and the User's Rating of the Responses:

- The assistant previously provided a structured list of "Advice" and "Information", which did not meet the user's emotional needs as evidenced by the poor empathy score and the user's followup indicating they have already taken those steps.

3. Key Tactics to Use in Response:

- "Emotional Expression": Acknowledge the user's frustration and emotional turmoil. Express empathy for their situation to make them feel heard and understood.
- "Validation": Reinforce that their feelings in this situation are normal and understandable given the circumstances. This can provide reassurance and support.
- "Paraphrasing": Reflect back the user's feelings to show that you understand their situation and the emotional impact it has on them.
- "Questioning": Invite the user to express more about how they are feeling or if there are other specific concerns they would like to discuss. This can help open up a dialogue centered around their emotions.

Given the user's low information need and high emotion need, the response will focus on these empathetic tactics to provide the support they are seeking.

</thinking>

<response>

It sounds like you're in a very difficult and frustrating situation with your sister and the estate. It's understandable that you feel like you're running out of options, and it must be incredibly disheartening to have already taken those steps without resolution.

Your feelings of frustration and helplessness are completely valid, especially when dealing with family matters that involve both emotional and legal complexities. It's not easy to be in a position where you feel pushed to consider legal action against a loved one.

How are you holding up with all of this? Is there anything specific you'd like to talk about or explore further? I'm here to listen and support you through this challenging time.

</response>

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1

Empowering LLM-Chatbots with More Empathy

- *Next Step:* Surgically Making LLMs More Empathic in Multi-turn Conversations

3

Unveiling Advanced Psychological Capabilities from LLMs: A Case of Targeted Reappraisal

- COLM 2024
- SPRI (under review)
- *Next Step:* Evaluating Reasoning Models' Capability on Emotion Tasks

2

4

Conclusion

- Summary of Proposed Work
- Timeline

Summary of Proposed Work

Summary of Proposed Work

We propose two projects:

Summary of Proposed Work

We propose two projects:

1. Evaluating reasoning models' capability on emotion tasks
 - Delivery: a conference paper to be submitted by December 2025

Summary of Proposed Work

We propose two projects:

1. Evaluating reasoning models' capability on emotion tasks
 - Delivery: a conference paper to be submitted by December 2025
2. Making LLMs more expressive of empathy in multi-turn conversations
 - Delivery: a conference paper to be submitted in Fall 2025