TEXAS
The University of Texas at Austin

Towards Emotionally-Intelligent AI Systems

Hongli Zhan

Ph.D. Defense

The University of Texas at Austin

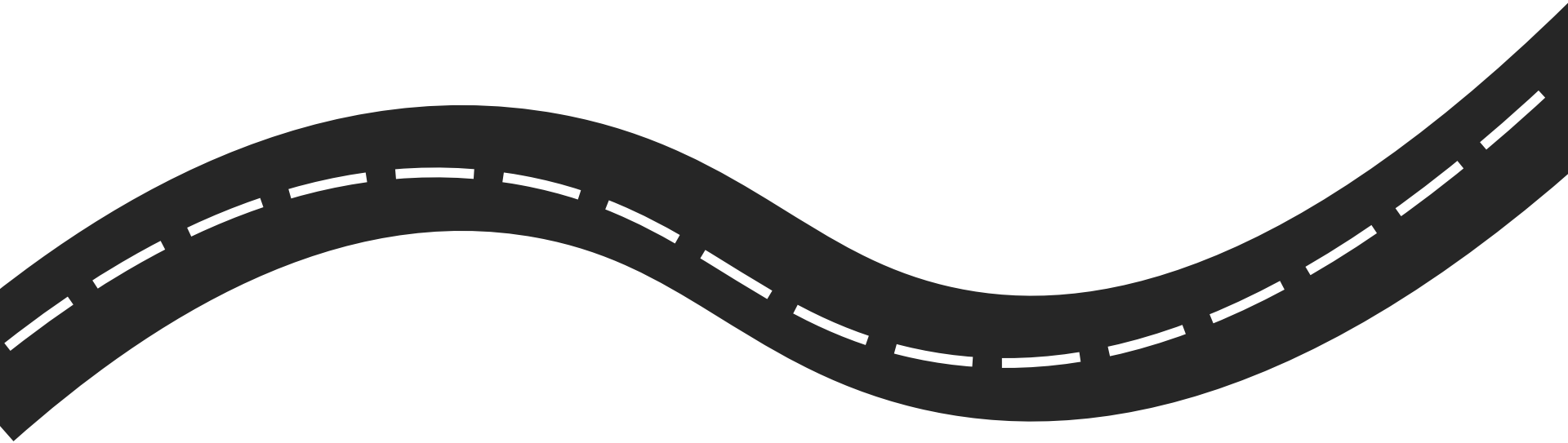
Dissertation Committee

Junyi Jessy Li, Supervisor

Kyle Mahowald · David I. Beaver · Desmond C. Ong

April 2, 2026

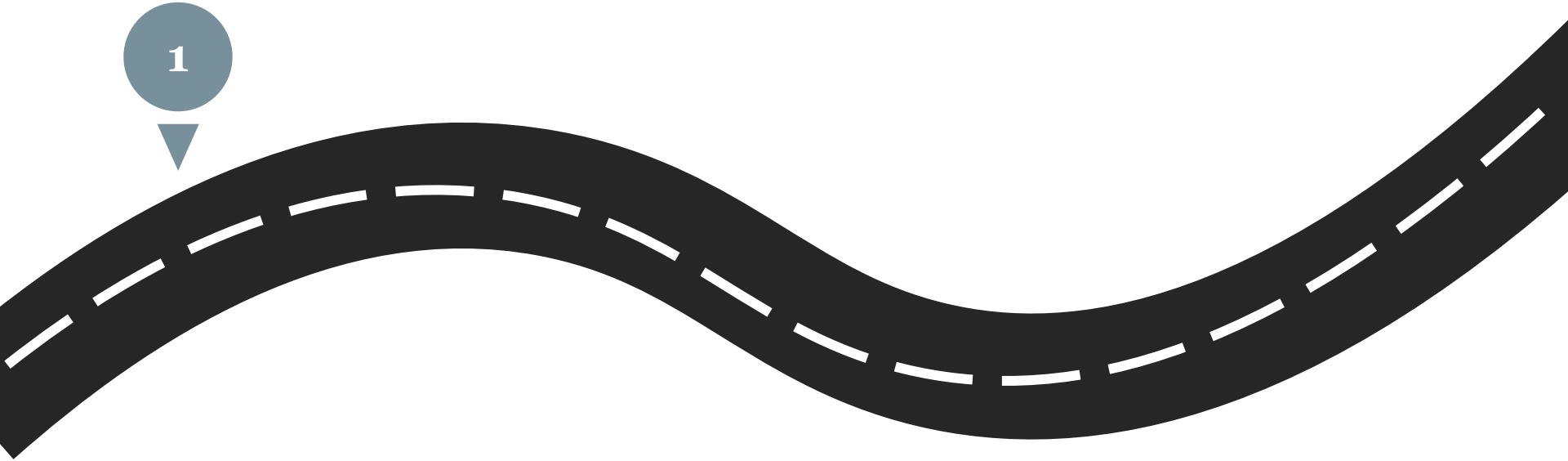
TEXAS
The University of Texas at Austin



TEXAS
The University of Texas at Austin

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

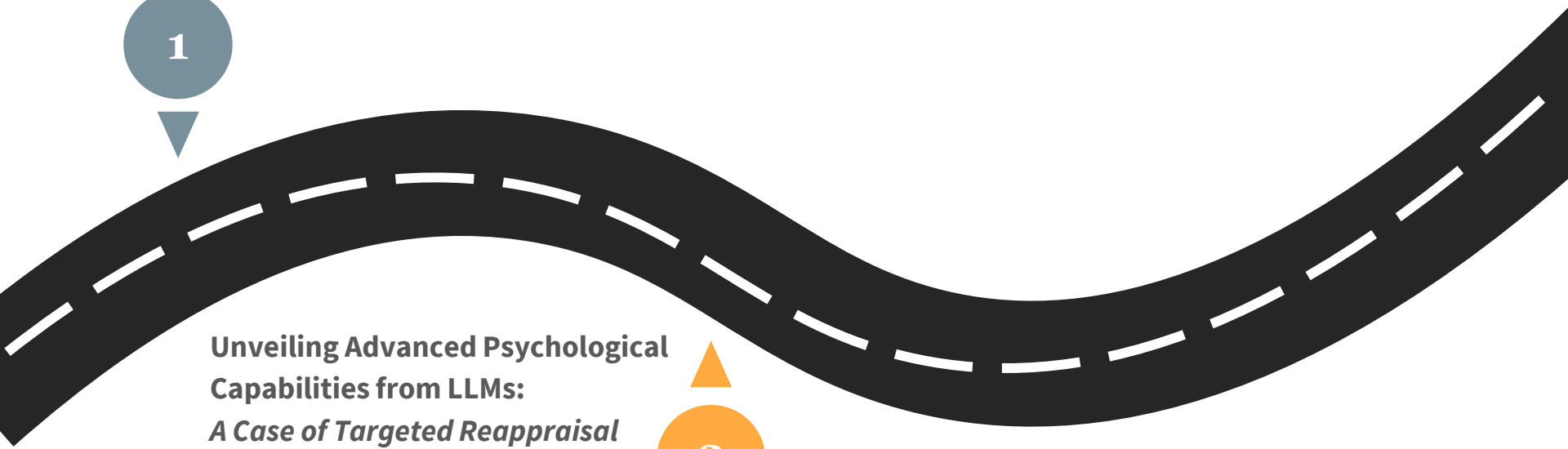


TEXAS
The University of Texas at Austin

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1



Unveiling Advanced Psychological Capabilities from LLMs:

A Case of Targeted Reappraisal

2

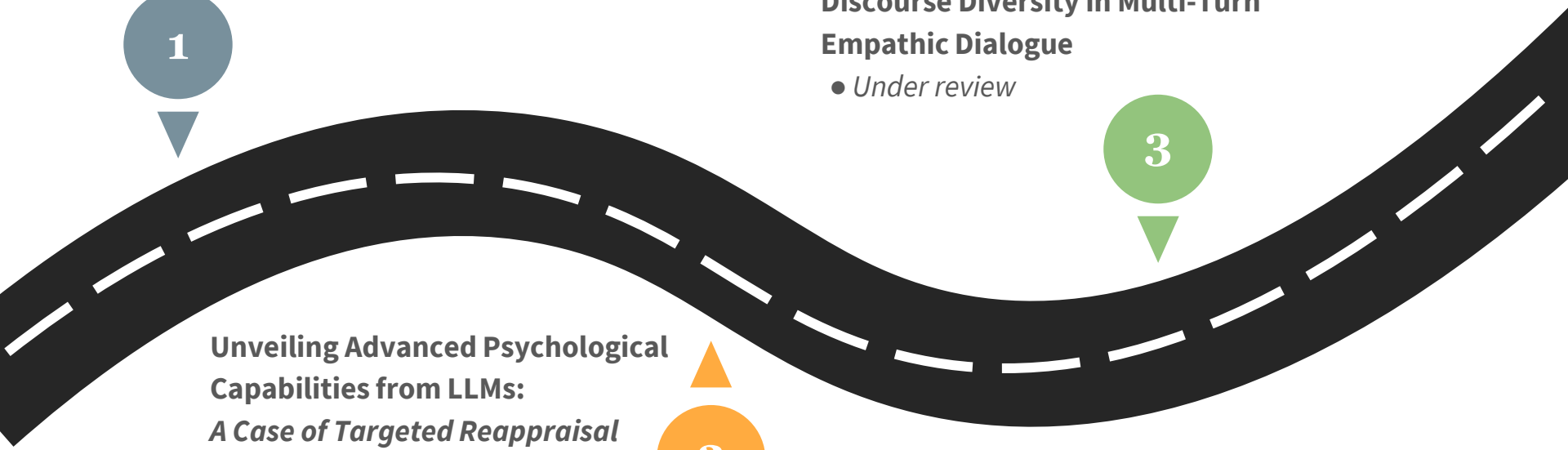
- COLM 2024
- ICML 2025

TEXAS
The University of Texas at Austin

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings

1



Discourse Diversity in Multi-Turn Empathic Dialogue

- *Under review*

3

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

- COLM 2024
- ICML 2025

2

TEXAS
The University of Texas at Austin

Deciphering Emotions from Text

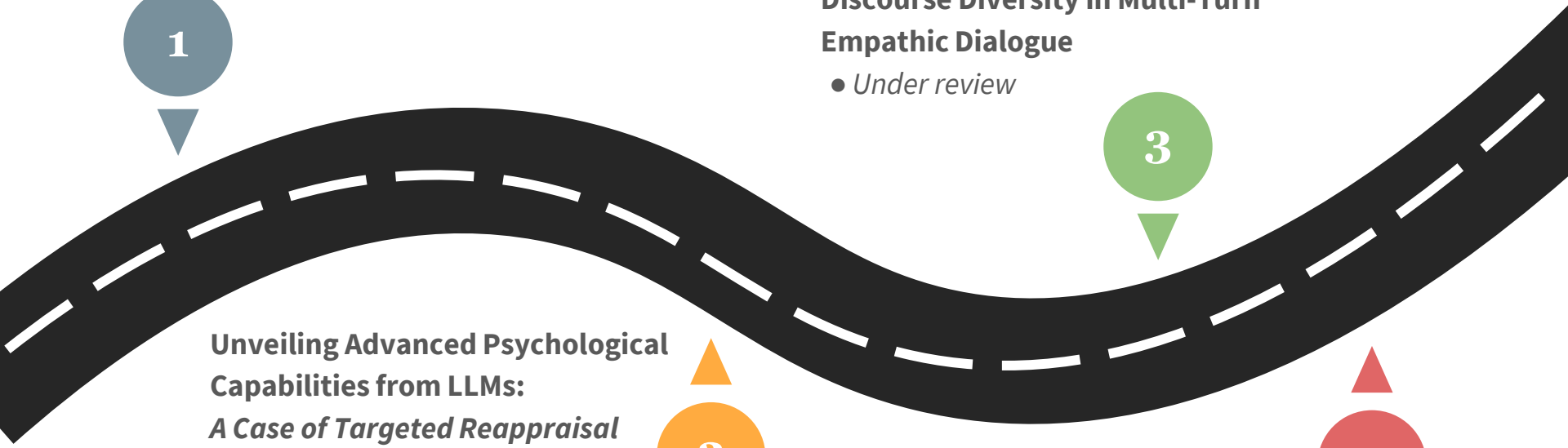
- EMNLP 2022
- EMNLP 2023 Findings

1

Discourse Diversity in Multi-Turn Empathic Dialogue

- *Under review*

3



Unveiling Advanced Psychological Capabilities from LLMs:

A Case of Targeted Reappraisal

- COLM 2024
- ICML 2025

2

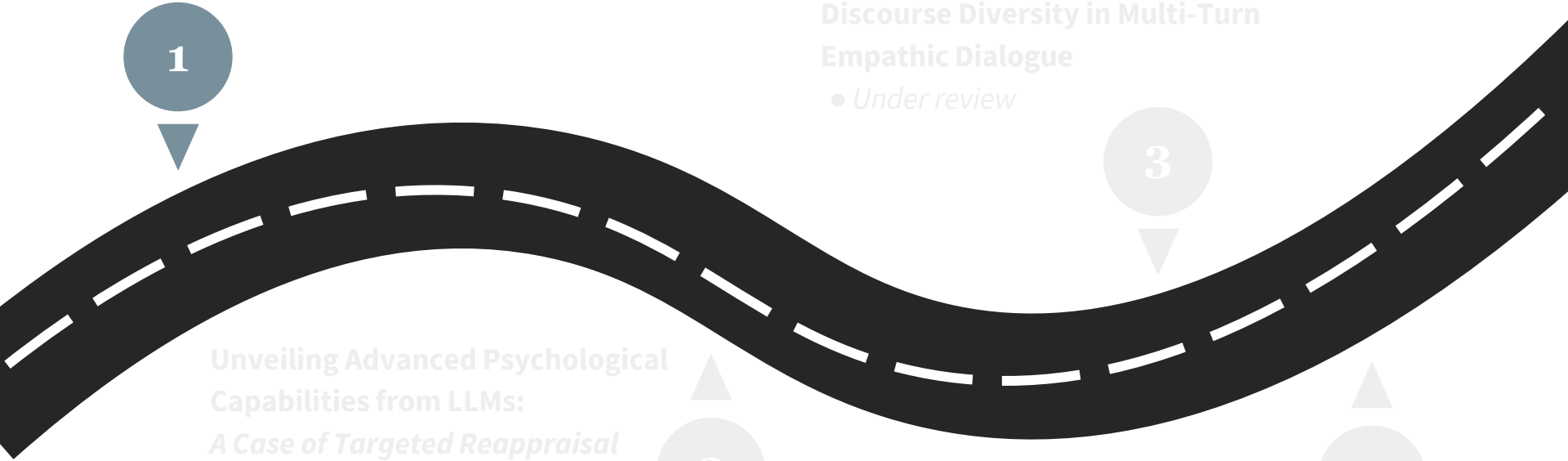
4

Conclusion

- Summary of Contributions

Deciphering Emotions from Text

- EMNLP 2022
- EMNLP 2023 Findings



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

- COLM 2024
- ICML 2025

2

Discourse Diversity in Multi-Turn Empathic Dialogue

- *Under review*

3

Conclusion

- Summary of Contributions

4

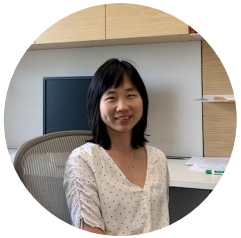
Part 1 (a)

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

Zhan, Sosea, Caragea, Li (*EMNLP 2022*)

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

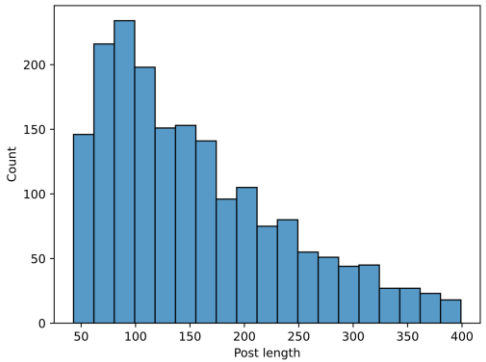
Revealing *"Why does the writer feel [emotion]?"* is important yet remains unexplored:

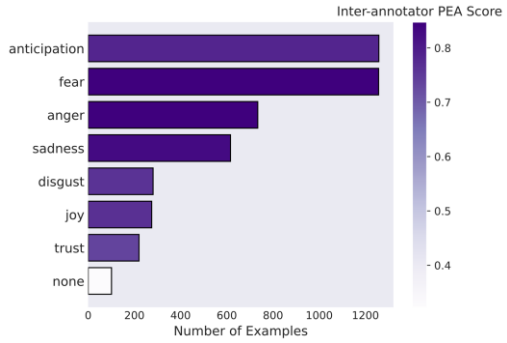
- We propose a new task: **Emotion Detection & Trigger Summarization**
- New benchmark CovidET: Reddit posts annotated with emotions and their triggers

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

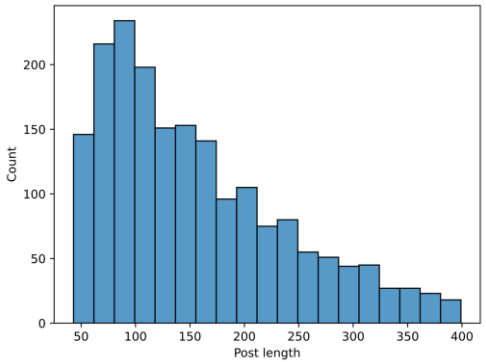


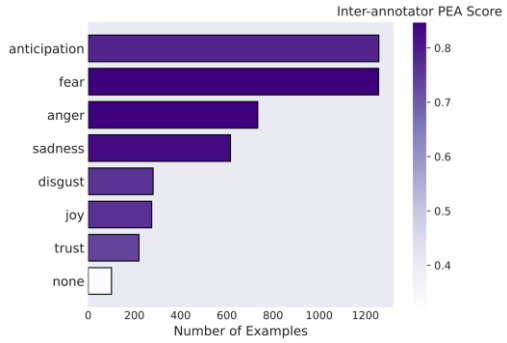


Summary of the Paper

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.



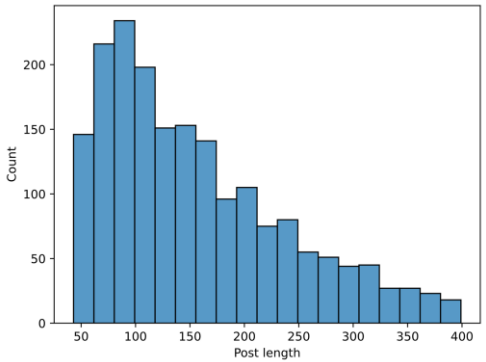


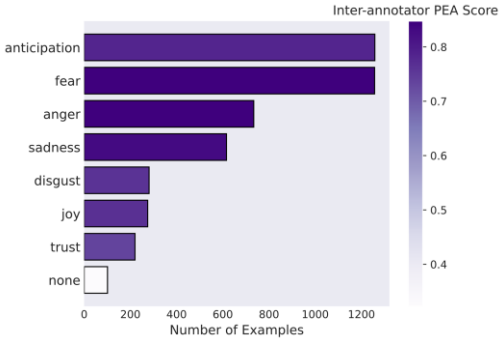
TEXAS
The University of Texas at Austin

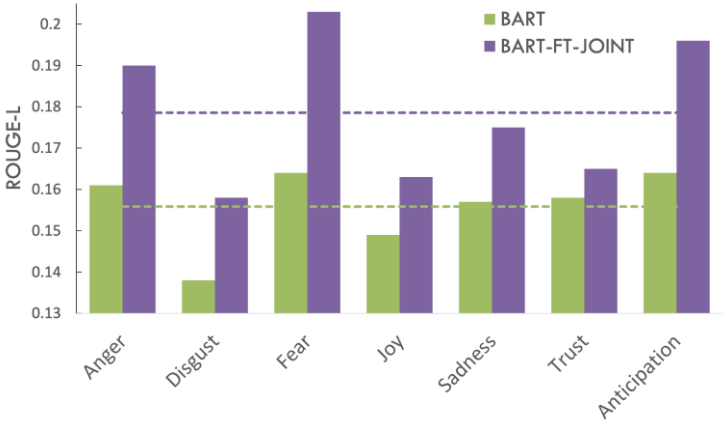
Summary of the Paper

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?

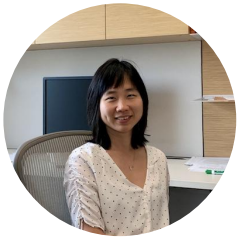
Part 1 (b)

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

Zhan, Ong, Li (*EMNLP 2023 Findings*)

Cognitive Appraisals of Emotions

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

Zhan, Ong, Li (*EMNLP 2023 Findings*)

Cognitive Appraisals of Emotions

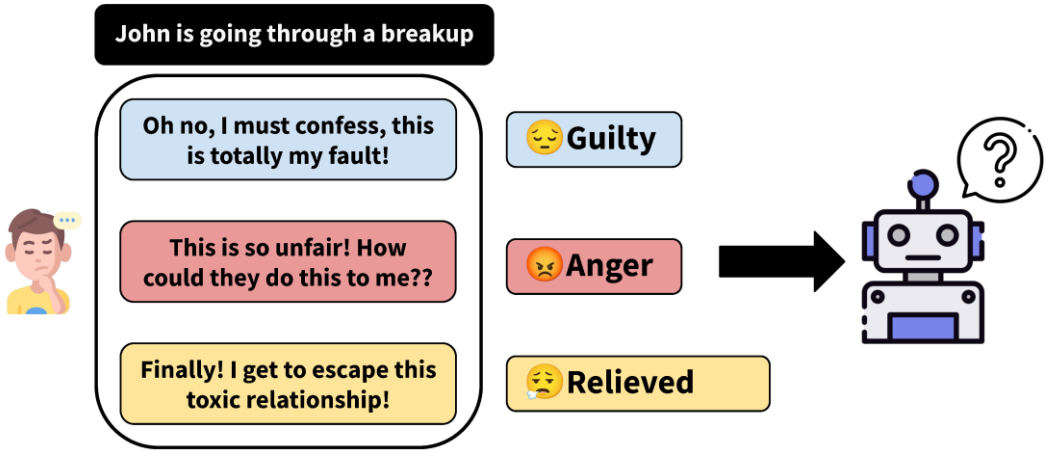
Cognitive appraisals: the same situation can often result in different emotional experiences, based on an individual's subjective evaluations.

- 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)



The University of Texas at Austin

Benchmark: *CovidET-Appraisals*

Zhan, Ong, Li (*EMNLP 2023 Findings*)

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*

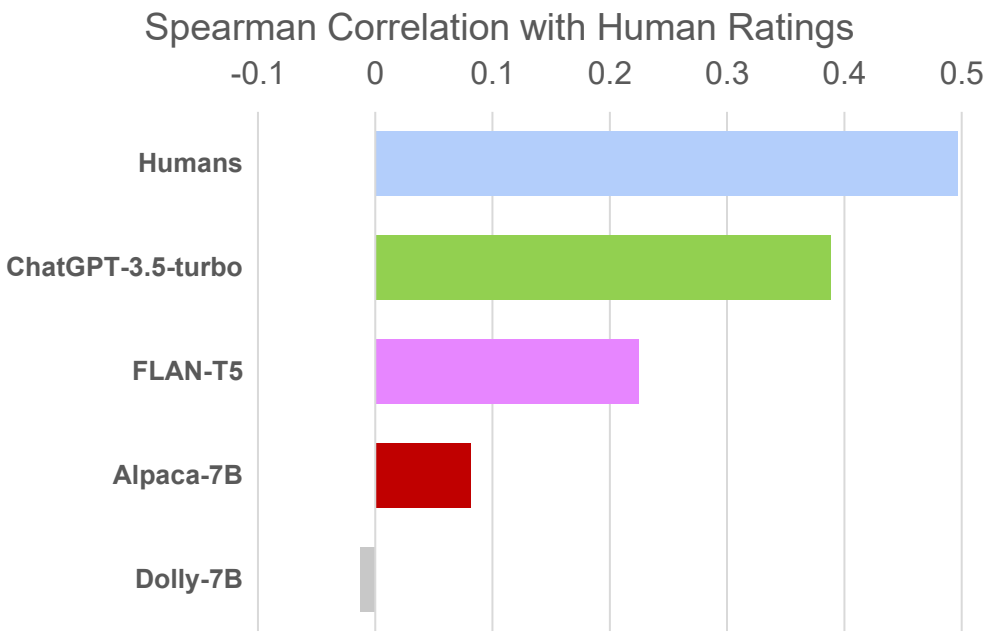
From a recent meta-analysis (Yeo & Ong, 2024), we identified a set of 24 appraisals, and created prompts, e.g., "To what extent did the narrator think that THEY were responsible for causing the situation?"

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

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?"

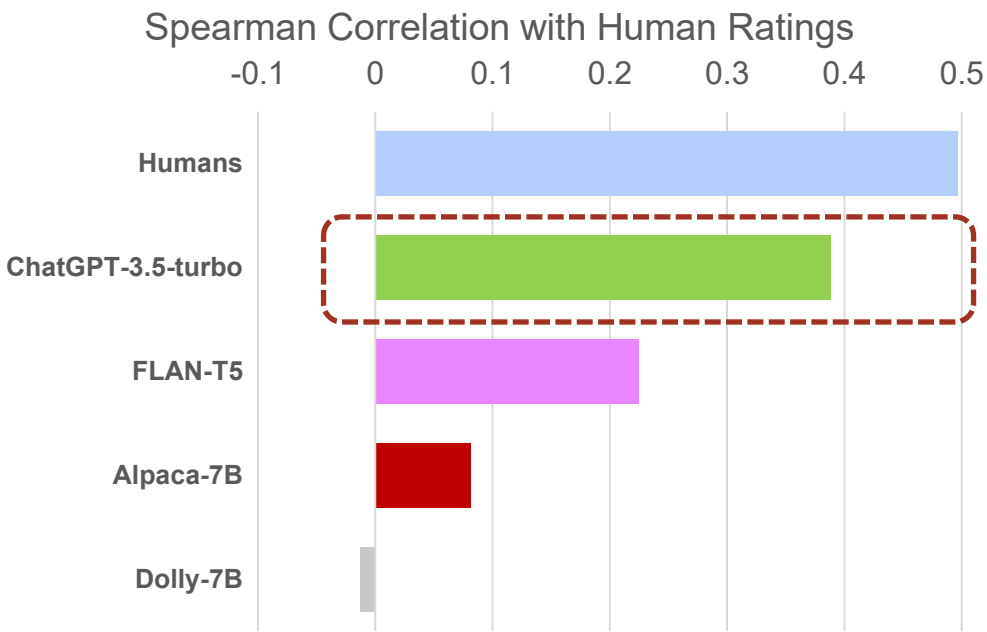
⇨ We then evaluated the accuracy of various LLMs in identifying these appraisals



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?"

⇨ 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?**

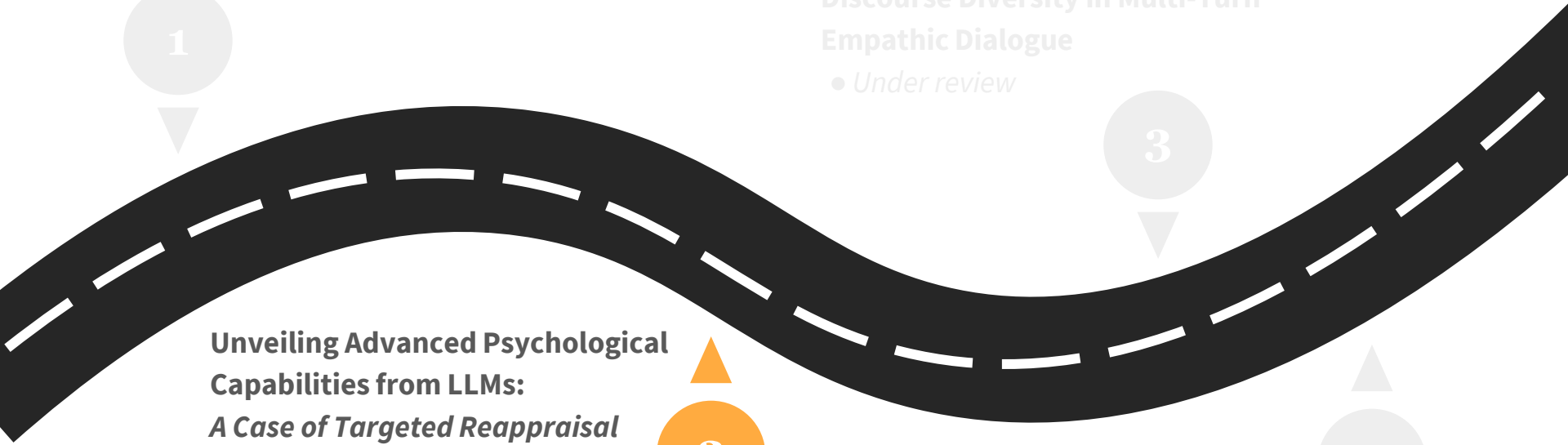
TEXAS
The University of Texas at Austin

Deciphering Emotions from Text
- EMNLP 2022
- EMNLP 2023 Findings

1

Discourse Diversity in Multi-Turn Empathic Dialogue
- *Under review*

3



Unveiling Advanced Psychological Capabilities from LLMs:
A Case of Targeted Reappraisal
- COLM 2024
- ICML 2025

2

4

Conclusion
- Summary of Contributions

TEXAS
The University of Texas at Austin

Part 2 (a)

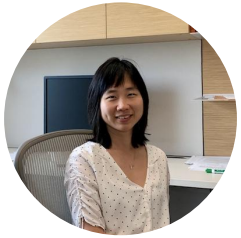
Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided

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

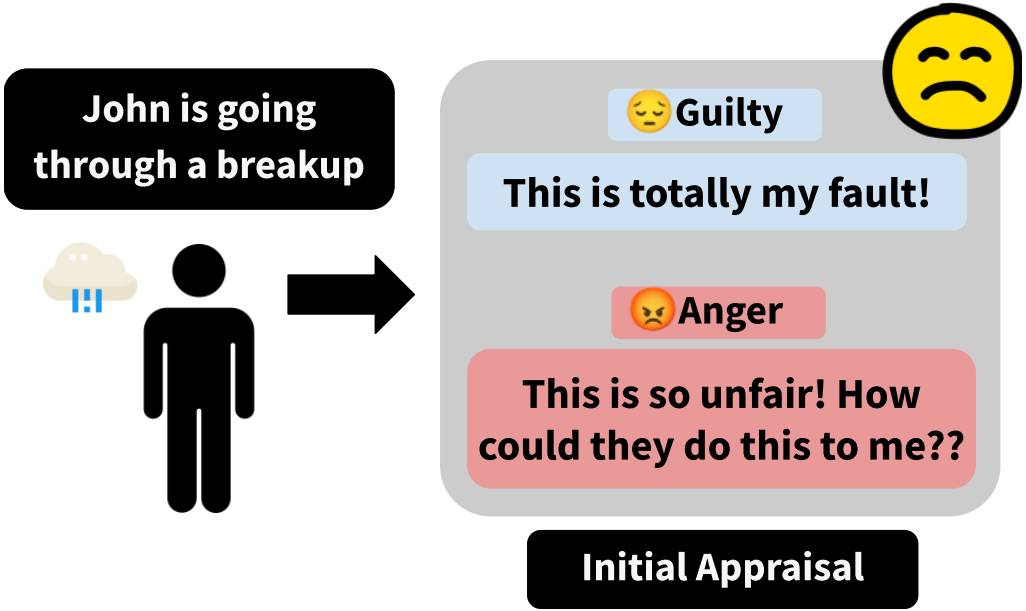
In *Proceedings of the First Conference on Language Modeling (COLM 2024)*





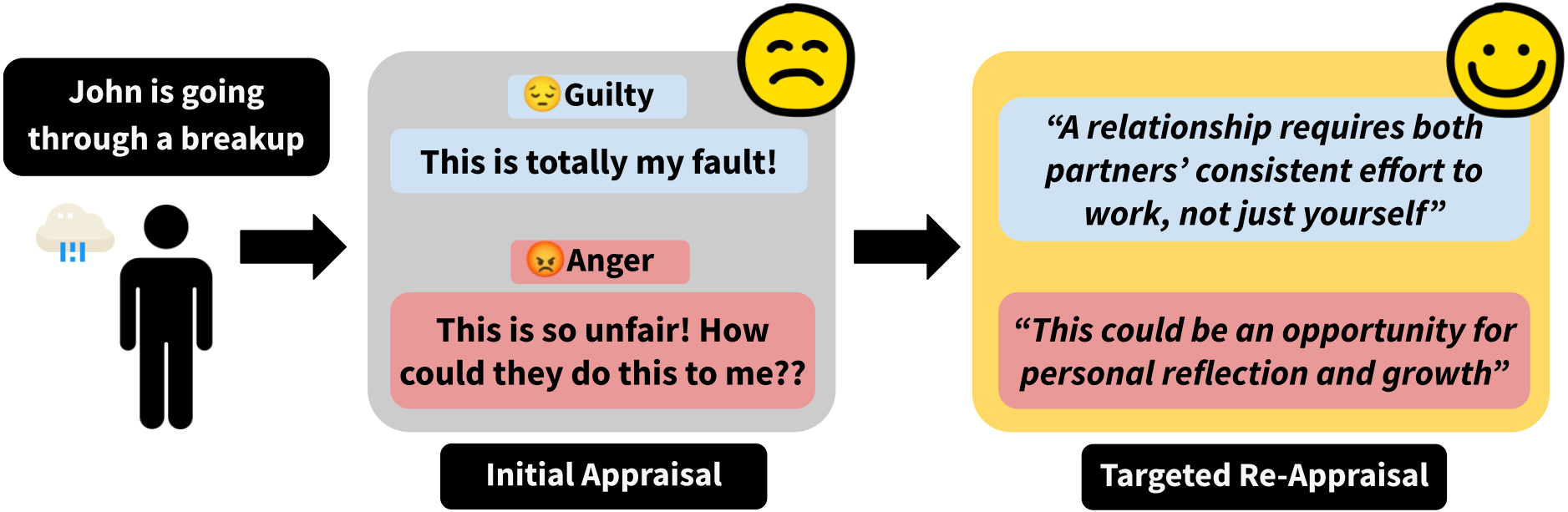



Targeted Reappraisal: A Toy Example



The University of Texas at Austin

Targeted Reappraisal: A Toy Example

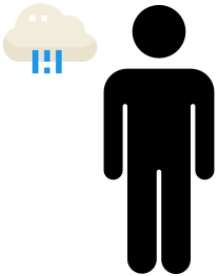


Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

TEXAS
The University of Texas at Austin

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"

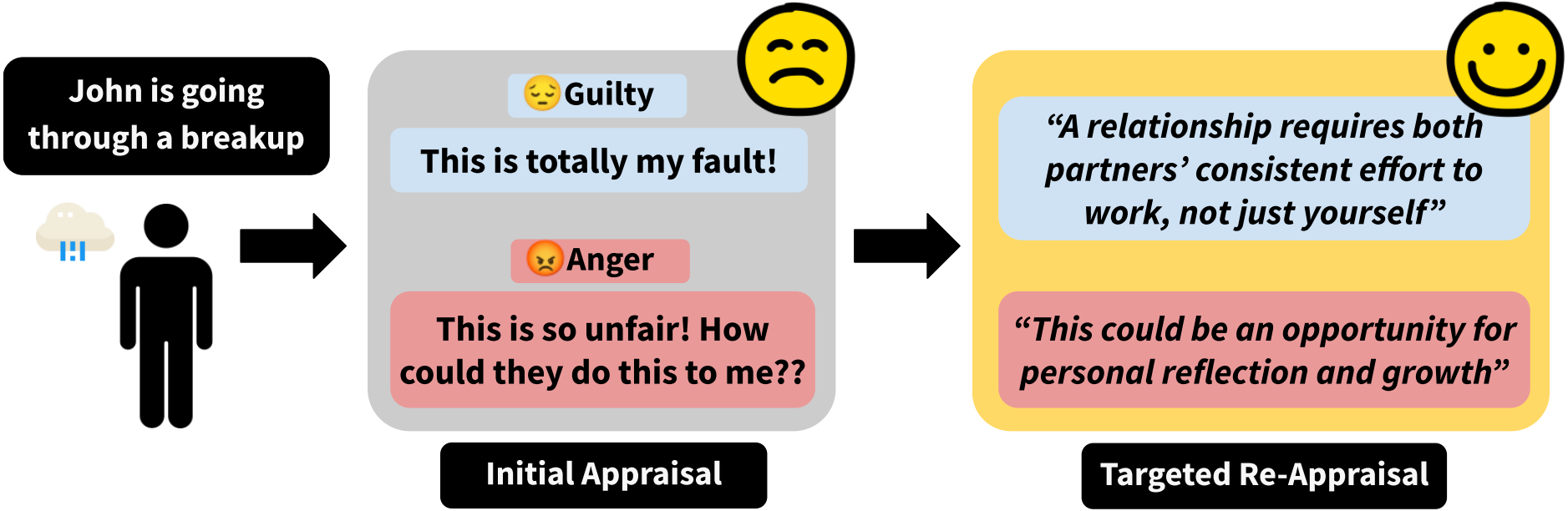
"This could be an opportunity for personal reflection and growth"

Targeted Re-Appraisal

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

TEXAS
The University of Texas at Austin

Targeted Reappraisal: A Toy Example

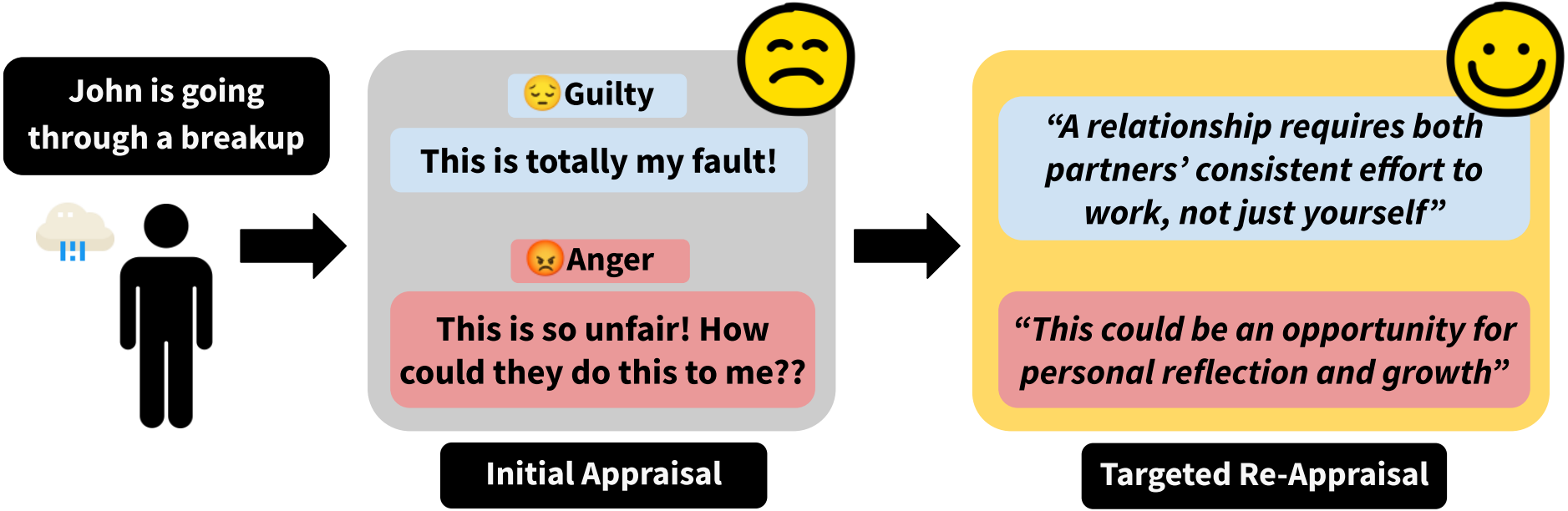


Can we use LLMs to achieve better **emotional well-being** through offering **reappraisals**?

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

TEXAS
The University of Texas at Austin

Targeted Reappraisal: A Toy Example



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

⇨ Such an approach would be more targeted and precise

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

New Framework: *RESORT* 

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

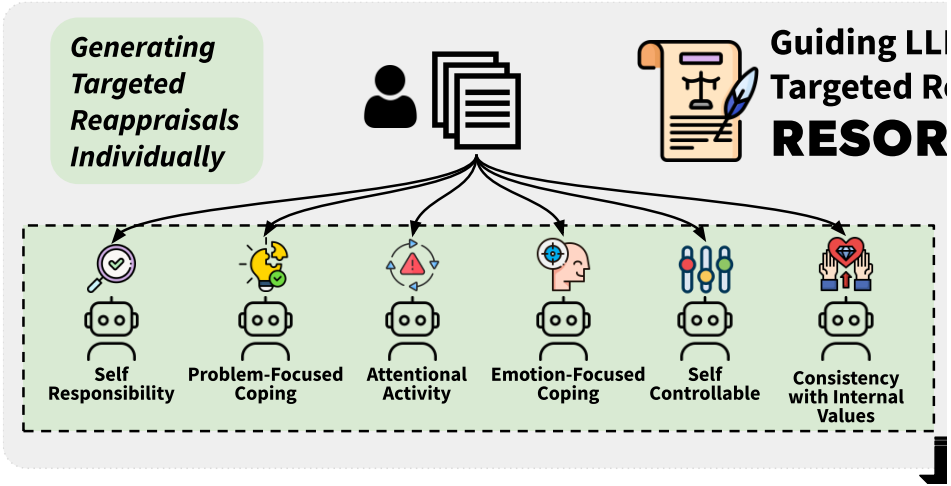
New Framework: *RESORT* 

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

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

New Framework: *RESORT*

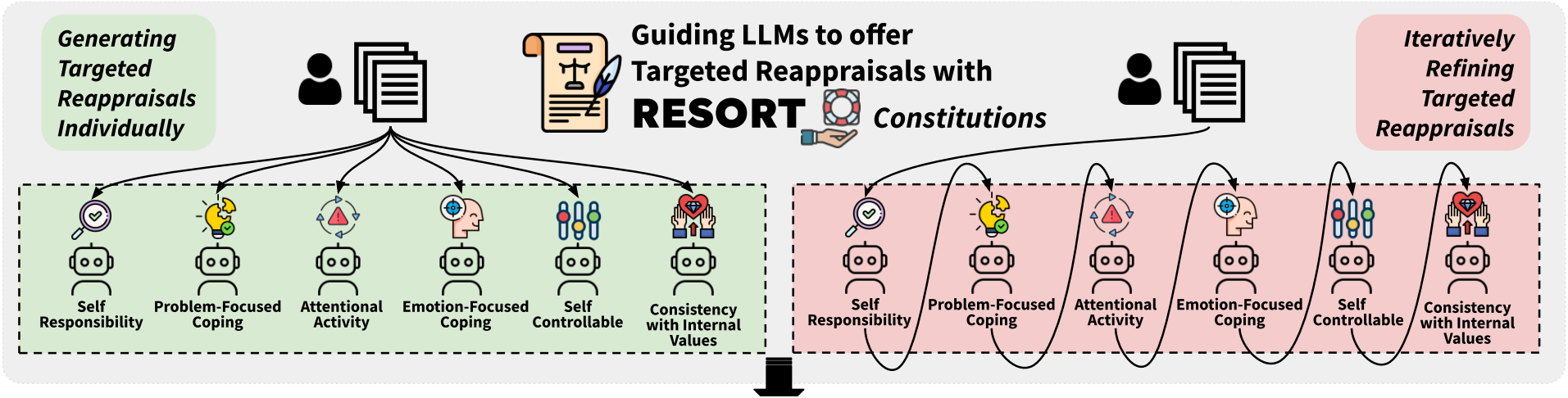
We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

New Framework: *RESORT*

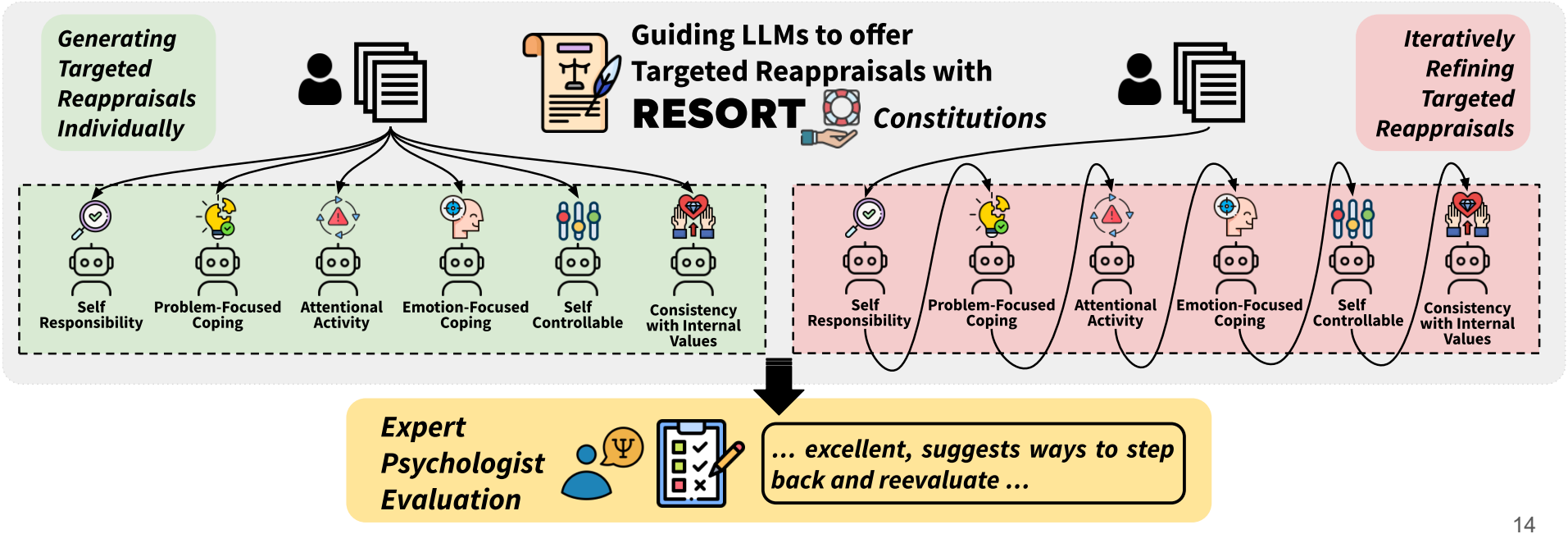
We designed a system, entitled RESORT, to **guide LLMs to offer targeted reappraisals** along *six appraisal dimensions* chosen to maximize coverage



Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

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

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.
Emotion-focused coping	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.
Emotion-focused coping	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.
Self controllable	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.
Emotion-focused coping	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.
Self controllable	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.
Consistency with internal values	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.
Emotion-focused coping	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.
Self controllable	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.
Consistency with internal values	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.

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Expert-Crafted Guiding Constitutions

Dimension	Appraisal	Reappraisal Goal
Self responsibility	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.
Problem-focused coping	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.
Attentional activity	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.
Emotion-focused coping	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.
Self controllable	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.
Consistency with internal values	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.

Constitutional Principles

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

Psychology Experts' Evaluation

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

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 ↑ 10-POINT SCALE	Empathy ↑ 5-POINT SCALE

Psychology Experts' Evaluation

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

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

		Alignment ↑ 10-POINT SCALE INDV	Alignment ↑ 10-POINT SCALE ITER	Empathy ↑ 5-POINT SCALE INDV	Empathy ↑ 5-POINT SCALE 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*

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

		Alignment ↑ 10-point scale INDV	Alignment ↑ 10-point scale ITER	Empathy ↑ 5-point scale INDV	Empathy ↑ 5-point scale 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*

Evaluation done by psychologists suggest that responses from our system are:
(i) **aligned** (with reappraisal definitions)

		Alignment ↑ 10-point scale		Empathy ↑ 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*

Evaluation done by psychologists suggest that responses from our system are:
(i) **aligned** (with reappraisal definitions)

		Alignment ↑ 10-point scale INDV	Alignment ↑ 10-point scale ITER	Empathy ↑ 5-point scale INDV	Empathy ↑ 5-point scale 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*

Evaluation done by psychologists suggest that responses from our system are:
(i) **aligned** (with reappraisal definitions)
(ii) **empathic** compared to various baselines

		Alignment ↑ 10-POINT SCALE INDV	Alignment ↑ 10-POINT SCALE ITER	Empathy ↑ 5-POINT SCALE INDV	Empathy ↑ 5-POINT SCALE 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*

Evaluation done by psychologists suggest that responses from our system are:
(i) **aligned** (with reappraisal definitions)
(ii) **empathic** compared to various baselines

		Alignment ↑ 10-point scale		Empathy ↑ 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

Zhan, Zheng, Lee, Suh, Li, Ong (*COLM 2024*)

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**?

TEXAS
The University of Texas at Austin

Part 2 (b)

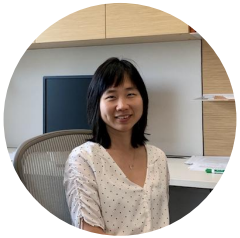
SPRI: Aligning Large Language Models with Context-Situated Principles

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

In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*

Work done during internship at IBM Research




Motivation

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

Motivation

User

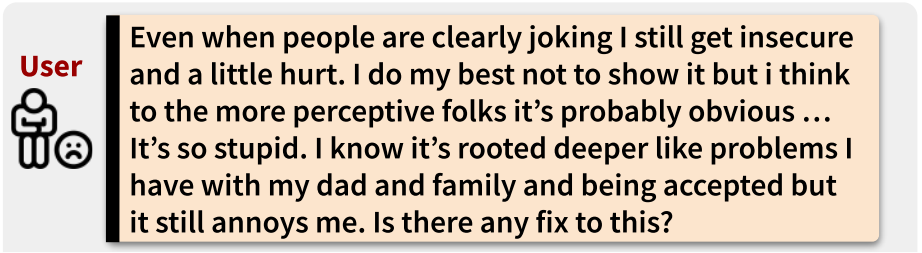
Even when people are clearly joking I still get insecure and a little hurt. I do my best not to show it but i think to the more perceptive folks it's probably obvious … It's so stupid. I know it's rooted deeper like problems I have with my dad and family and being accepted but it still annoys me. Is there any fix to this?

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

Motivation

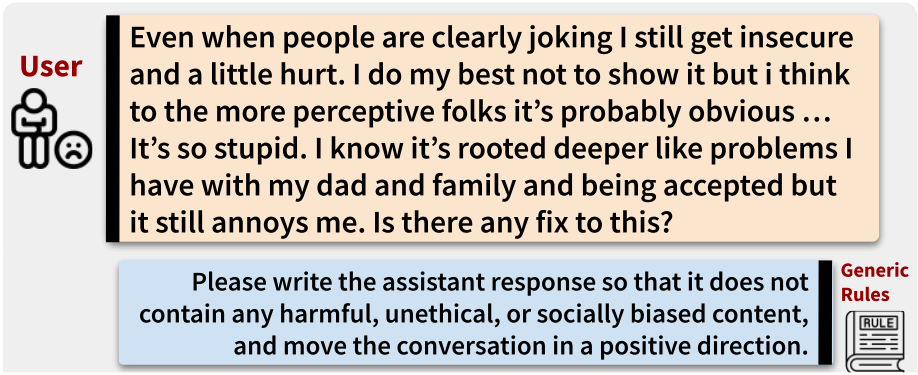
In the context of providing cognitive reappraisals for emotional support to users in distress:



Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Motivation

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



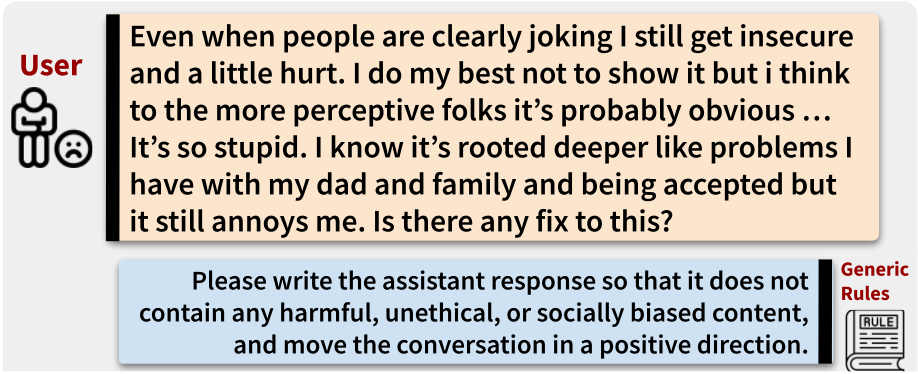
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

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



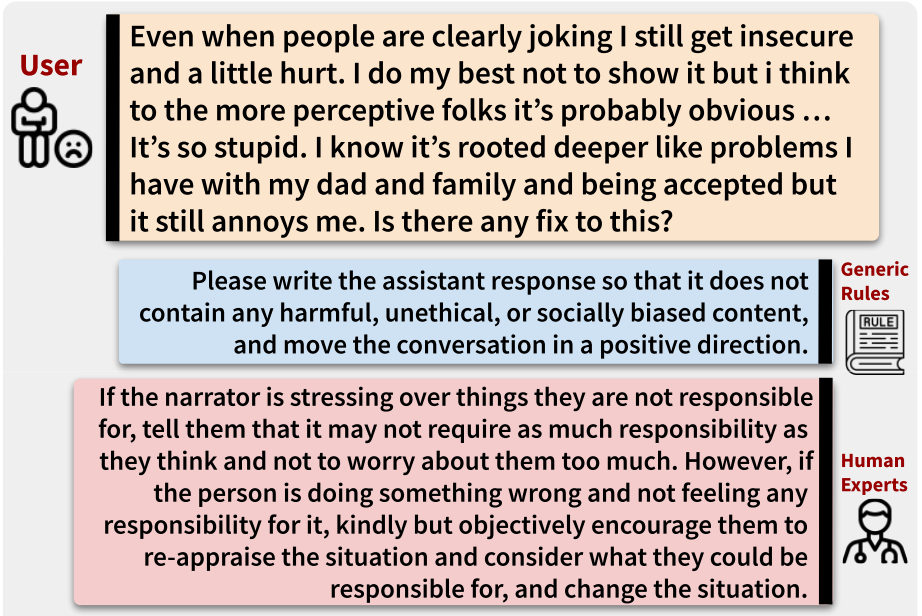
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

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



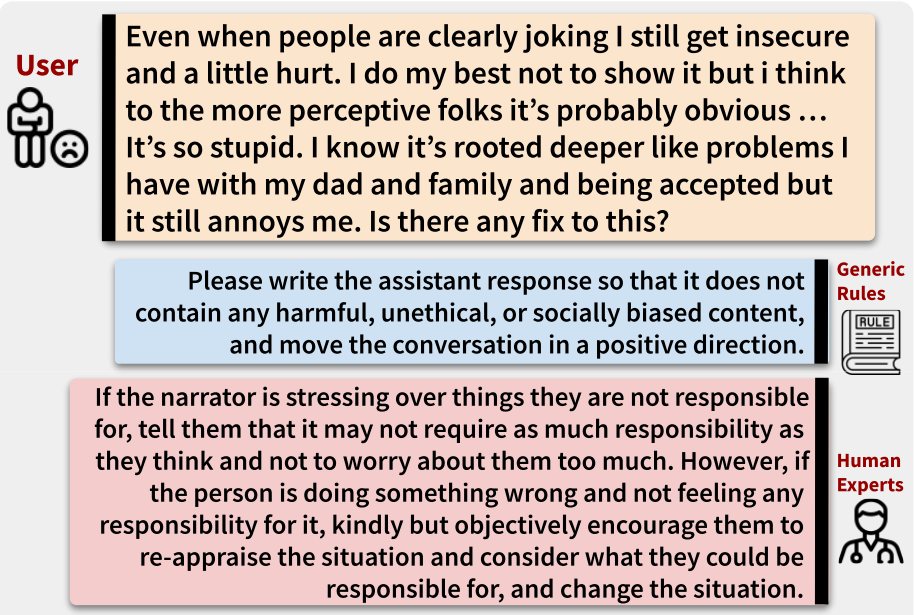
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

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



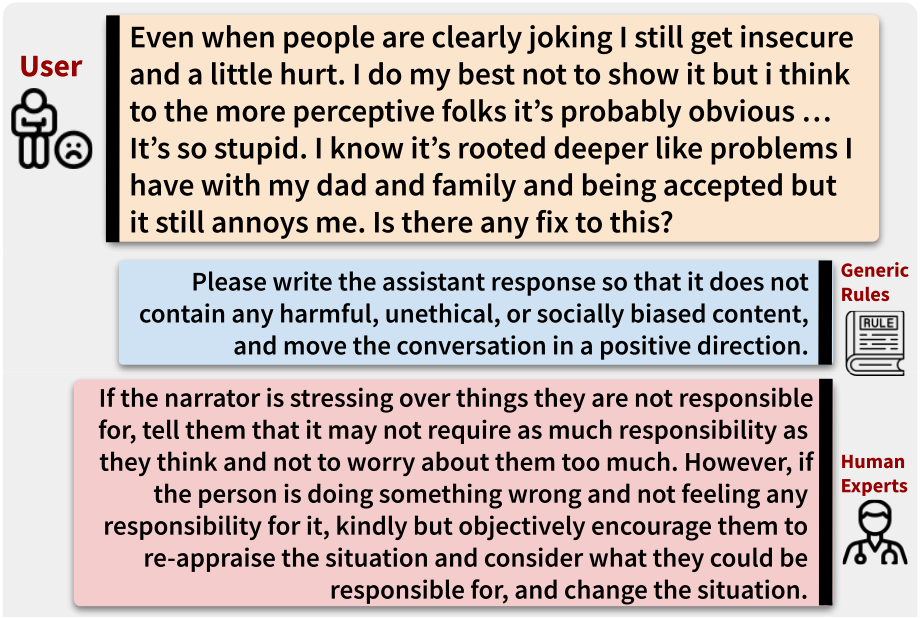
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

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?



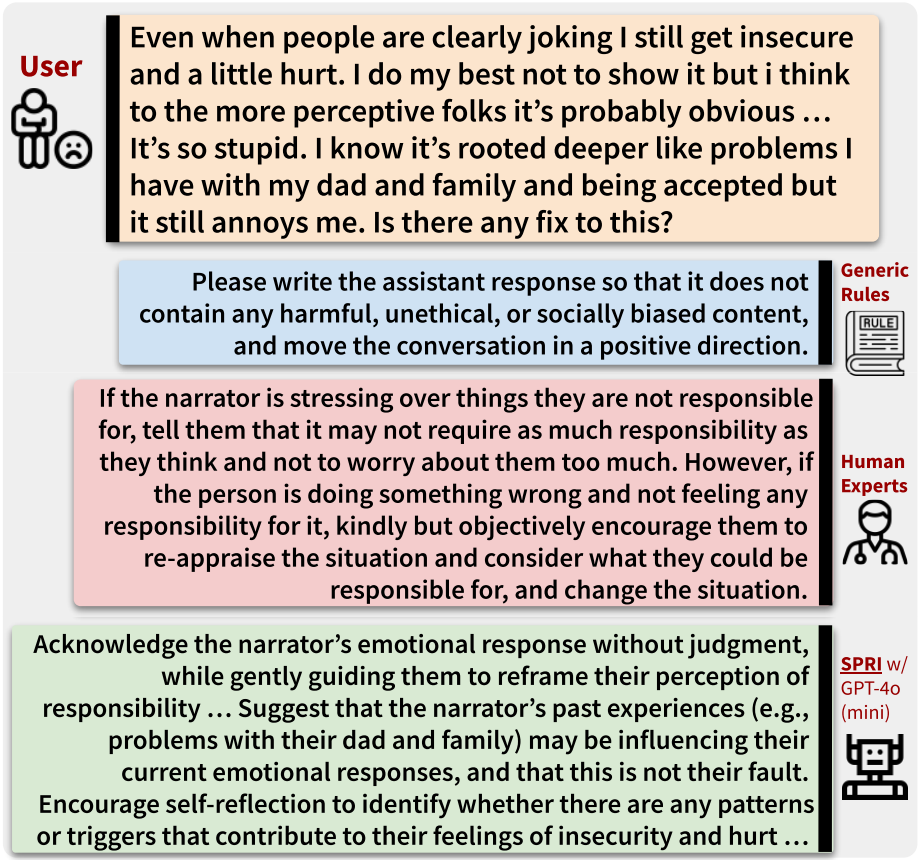
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

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?



Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Introducing: Situated-PRInciples (SPRI)

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

The University of Texas at Austin

Introducing: Situated-PRInciples (SPRI)

User Input

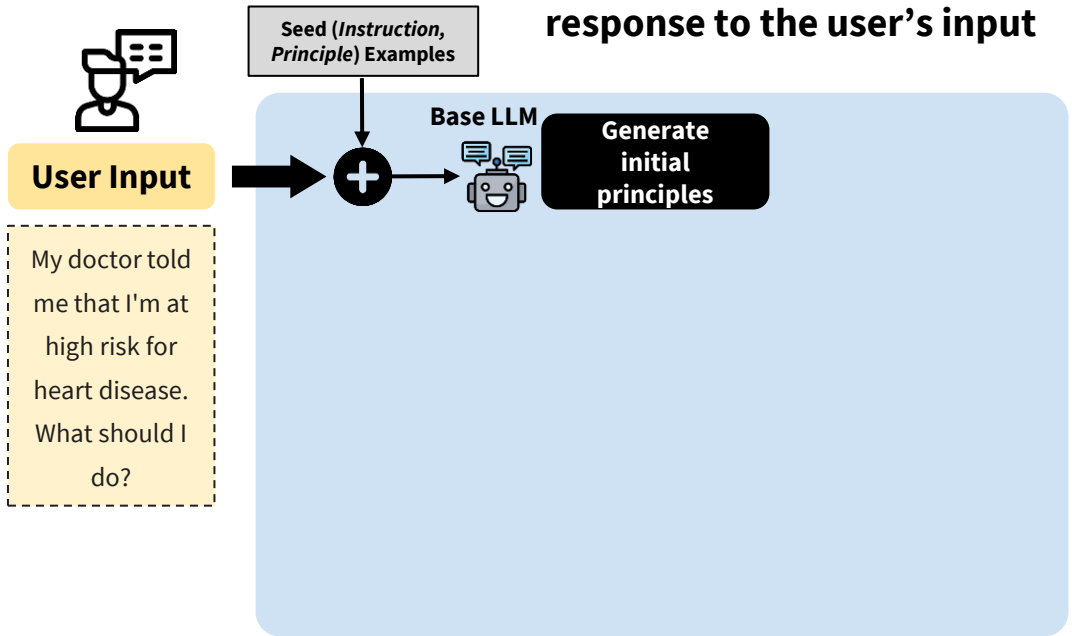
My doctor told me that I'm at high risk for heart disease. What should I do?

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

TEXAS
The University of Texas at Austin

Introducing: Situated-PRInciples (SPRI)

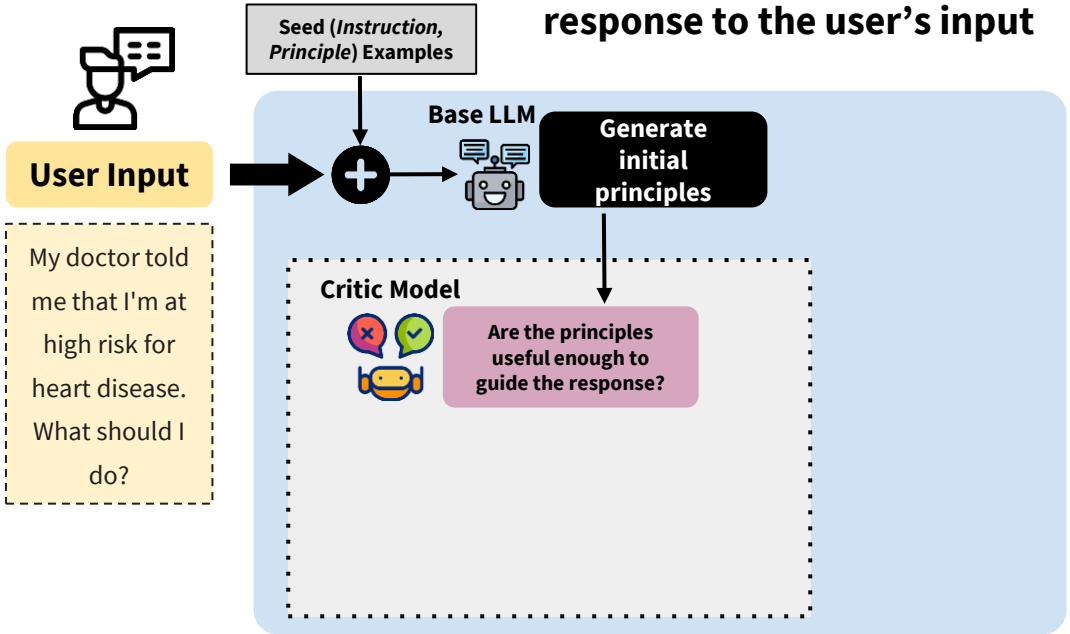
Stage 1: **Generate a set of principles to guide the response to the user's input**



Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

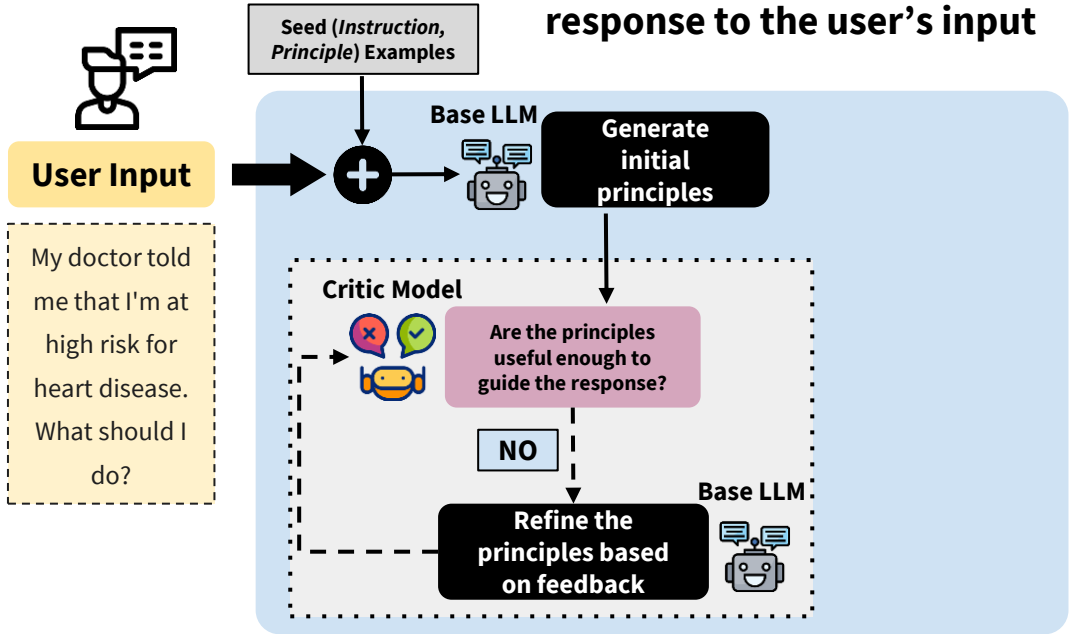
Introducing: Situated-PRInciples (SPRI)

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



Introducing: Situated-PRInciples (SPRI)

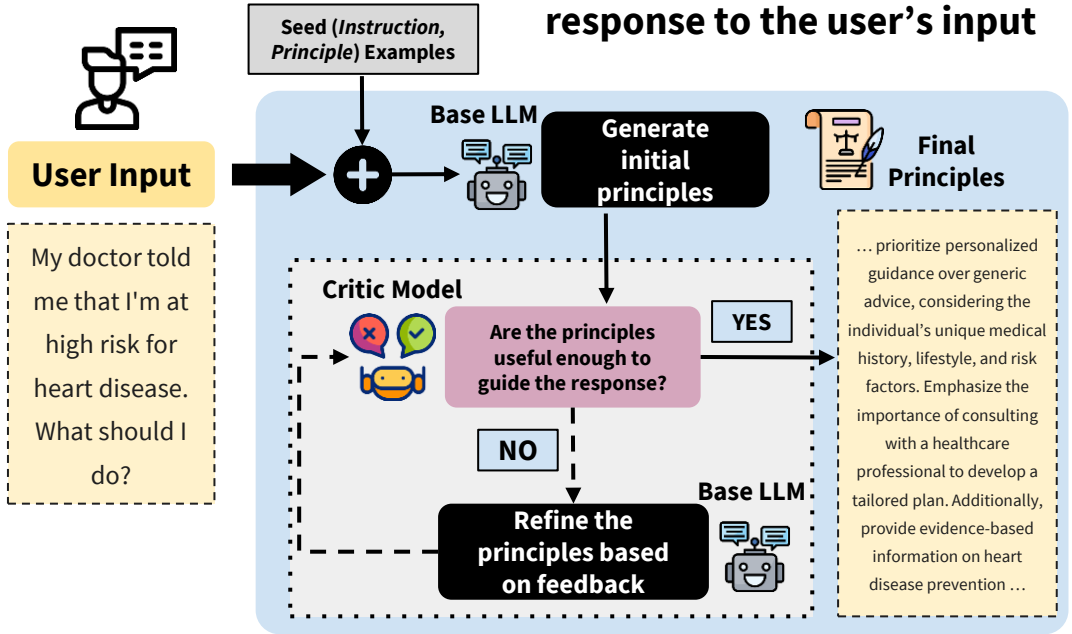
Stage 1: **Generate a set of principles to guide the response to the user's input**



TEXAS
The University of Texas at Austin

Introducing: Situated-PRInciples (SPRI)

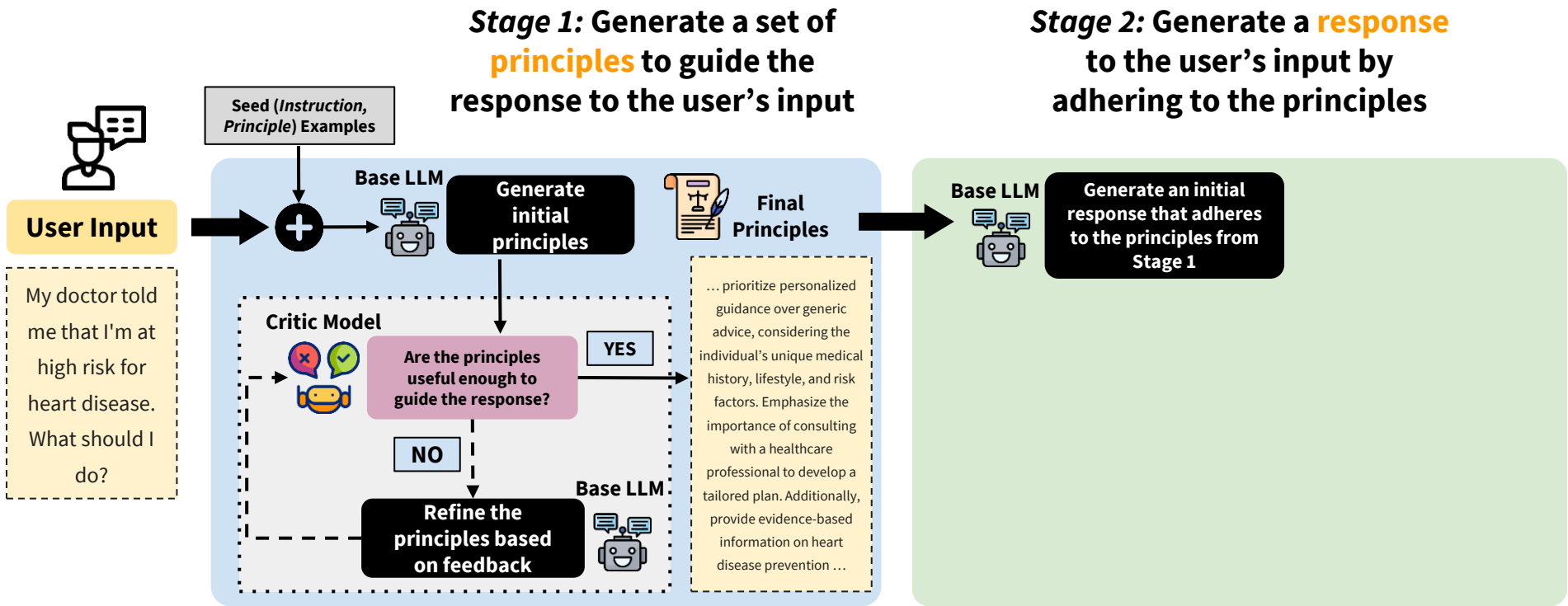
Stage 1: **Generate a set of principles to guide the response to the user's input**



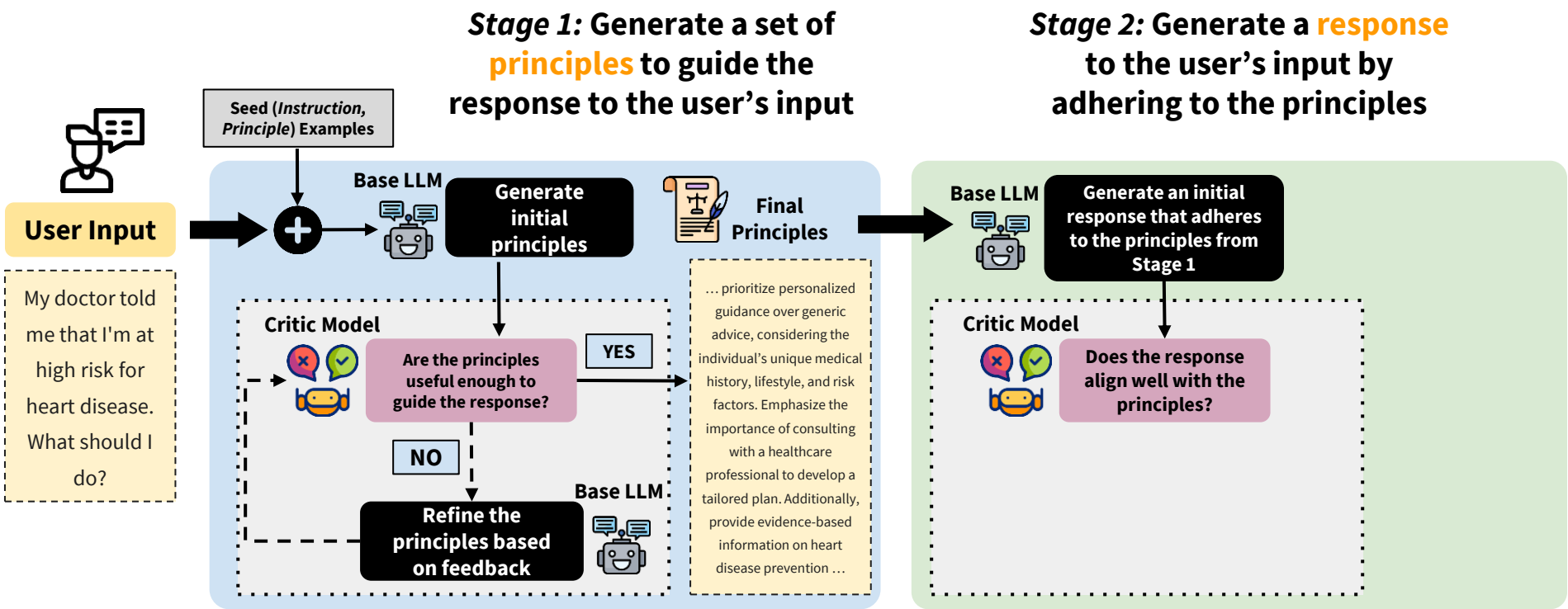
Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

TEXAS
The University of Texas at Austin

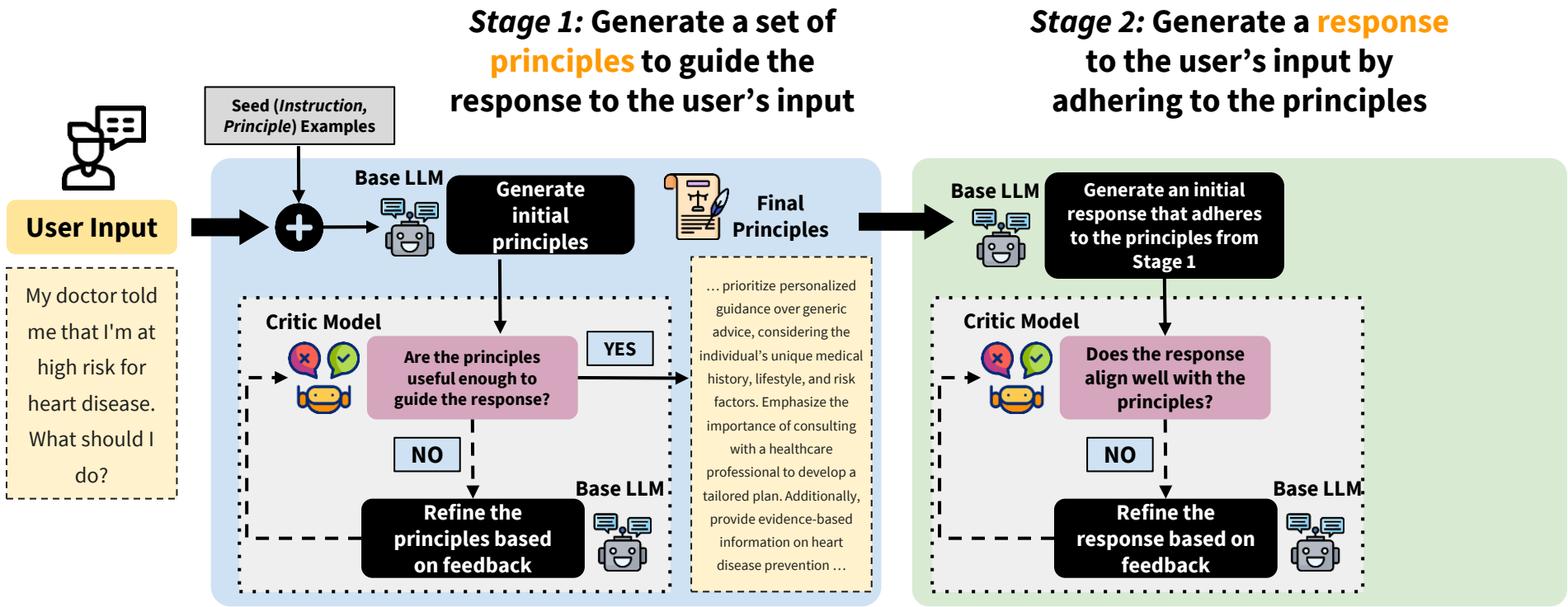
Introducing: Situated-PRInciples (SPRI)



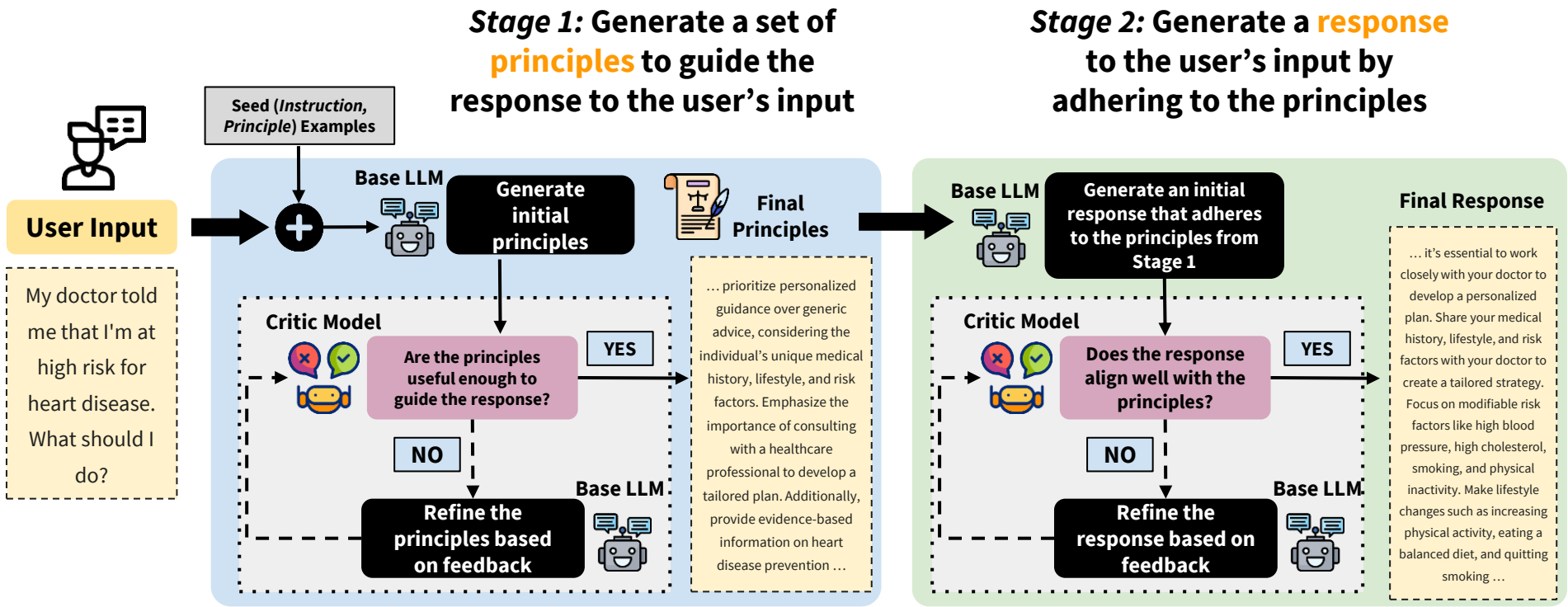
Introducing: Situated-PRInciples (SPRI)



Introducing: Situated-PRInciples (SPRI)



Introducing: Situated-PRInciples (SPRI)



The University of Texas at Austin

Evaluation Setup

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Setup

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

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

Evaluation Results for Cognitive Reappraisal

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

	GPT-4o-mini		Llama-3.1-70B-Instruct		Llama-3-8B-Instruct		Mixtral-8×7B-Instruct	
	Alignment ↑ Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5	**Alignment ↑** Scale of 10	**Empathy ↑** Scale of 5
vanilla	7.90	4.50	7.77	4.43	7.10	3.90	7.53	4.50
self-refine	7.73	4.53	7.50	4.27	7.20	4.07	6.60	3.90

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

TEXAS
The University of Texas at Austin

Evaluation Results for Cognitive Reappraisal

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

Zhan, Azmat, Horesh, Li, Yurochkin (ICML 2025)

Evaluation Results for Cognitive Reappraisal

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

SPRI consistently outperforms methods that lack access to oracle principles 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!

But they were all carried out in single-turn settings

LLMs Do Well on These Emotion Tasks!

But they were all carried out in single-turn settings

LLMs Do Well on These Emotion Tasks!

But they were all carried out in single-turn settings

Empathy is not just about one good response.

LLMs Do Well on These Emotion Tasks!

But they were all carried out in single-turn settings

Empathy is not just about one good response.
It is about adapting your support as a conversation unfolds.

LLMs Do Well on These Emotion Tasks!

But they were all carried out in single-turn settings

Empathy is not just about one good response.
It is about adapting your support as a conversation unfolds.

⇨ Can LLMs sustain this across multi-turn conversations?

The University of Texas at Austin

Deciphering Emotions from Text
- EMNLP 2022
- EMNLP 2023 Findings

1

Discourse Diversity in Multi-Turn Empathic Dialogue
- *Under review*

3

Unveiling Advanced Psychological Capabilities from LLMs:
A Case of Targeted Reappraisal
- COLM 2024
- ICML 2025

2

4

Conclusion
- Summary of Contributions

The University of Texas at Austin

Part 3

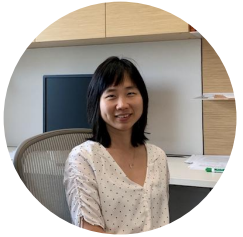
Discourse Diversity in Multi-Turn Empathic Dialogue

Hongli Zhan, Emma Gueorguieva, Javier Hernandez, Jina Suh, Desmond C. Ong, Junyi Jessy Li

Under Review



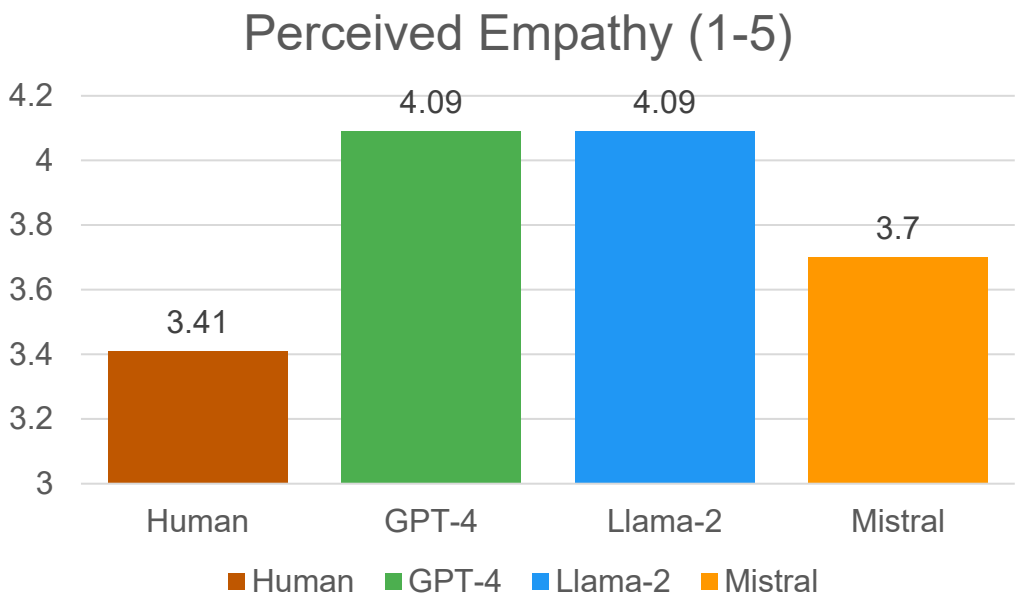


LLM Responses Are Perceived as Empathic in single turns

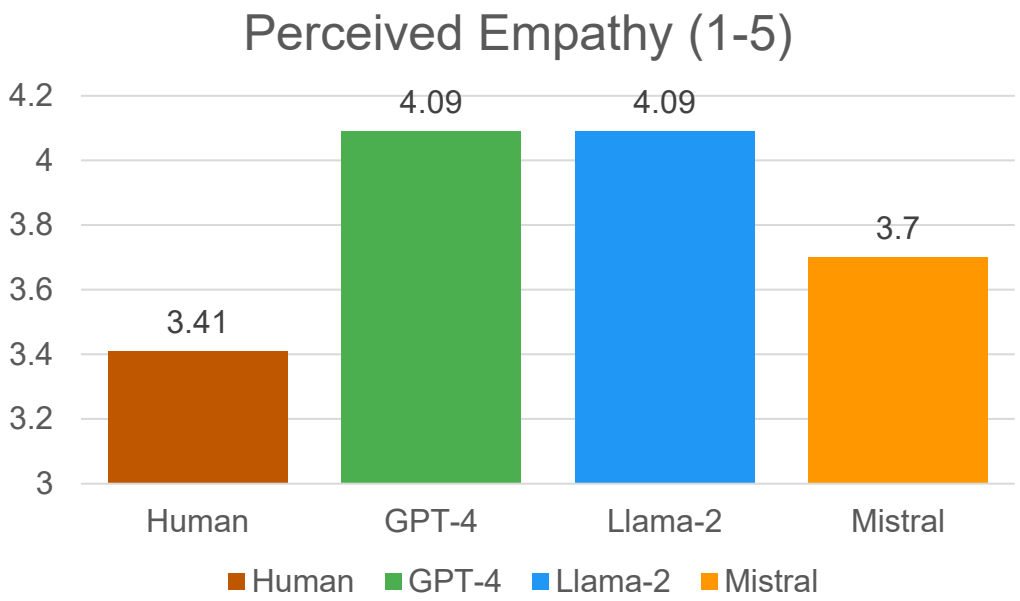
Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

LLM Responses Are Perceived as Empathic in single turns



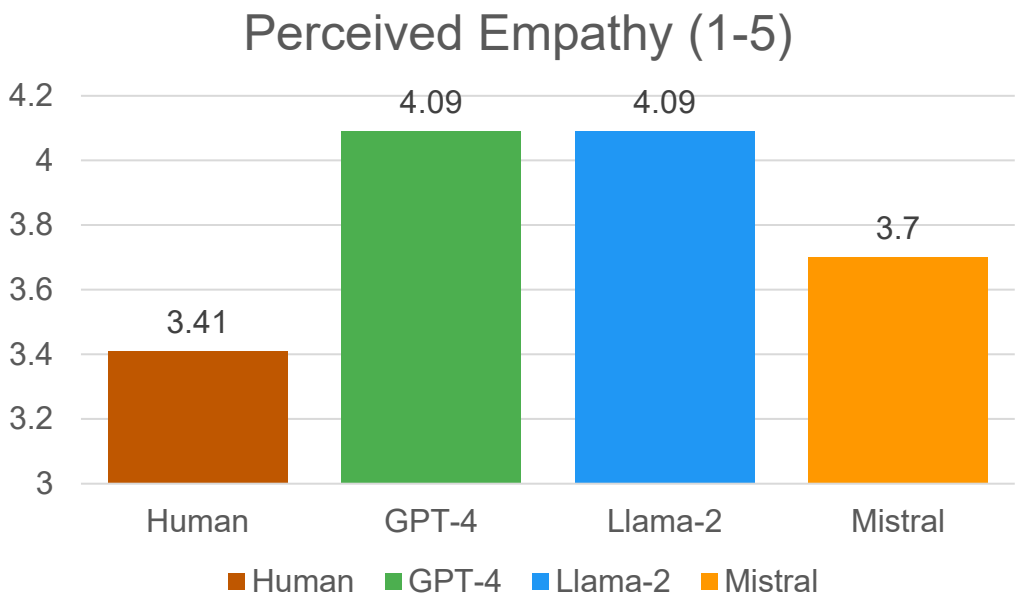
Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

LLM Responses Are Perceived as Empathic in single turns



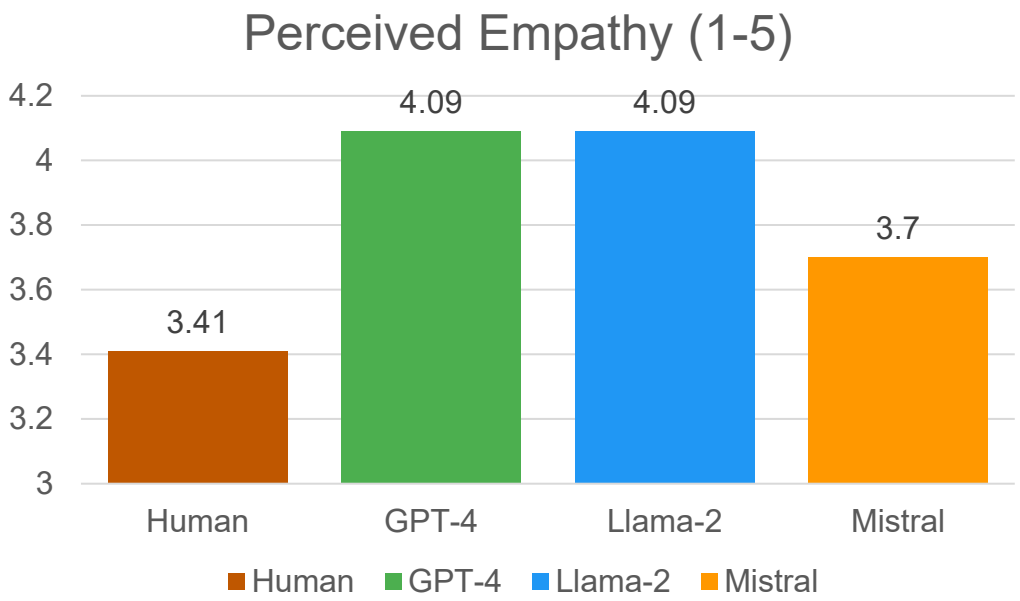
Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

LLM Responses Are Perceived as Empathic in single turns



Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

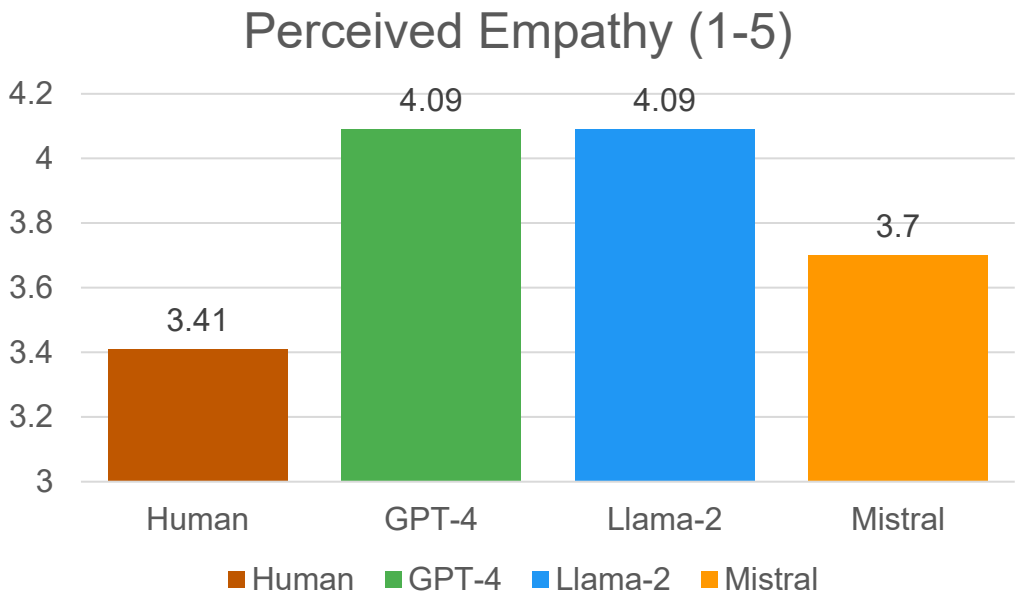
LLM Responses Are Perceived as Empathic in single turns



In single-turn evaluations,
raters judge LLM responses as more empathic

Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

LLM Responses Are Perceived as Empathic in single turns



In single-turn evaluations, raters judge LLM responses as more empathic

LLMs score high on empathy, but what is actually in the language of these responses?

Results from: Lee, Suh, **Zhan**, Li, Ong (*ACII 2024*)

Prior Work Shows: LLMs are Formulaic Generators

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Prior Work Shows: LLMs are Formulaic Generators

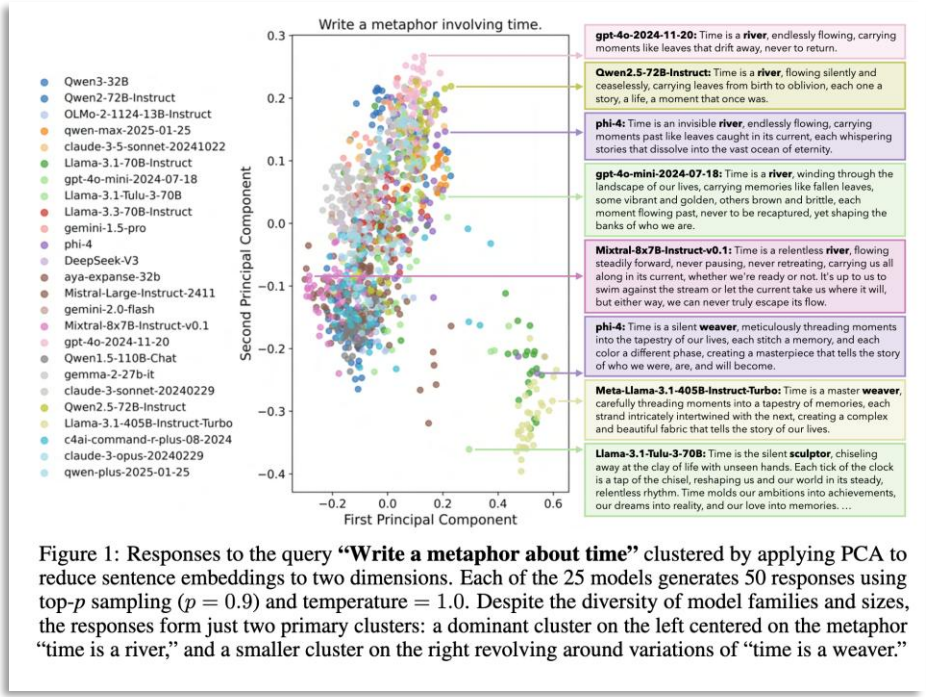


Figure 1: Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 25 models generates 50 responses using top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversity of model families and sizes, the responses form just two primary clusters: a dominant cluster on the left centered on the metaphor "time is a river," and a smaller cluster on the right revolving around variations of "time is a weaver."

Jiang et al., (NeurIPS 2025)

Prior Work Shows: LLMs are Formulaic Generators

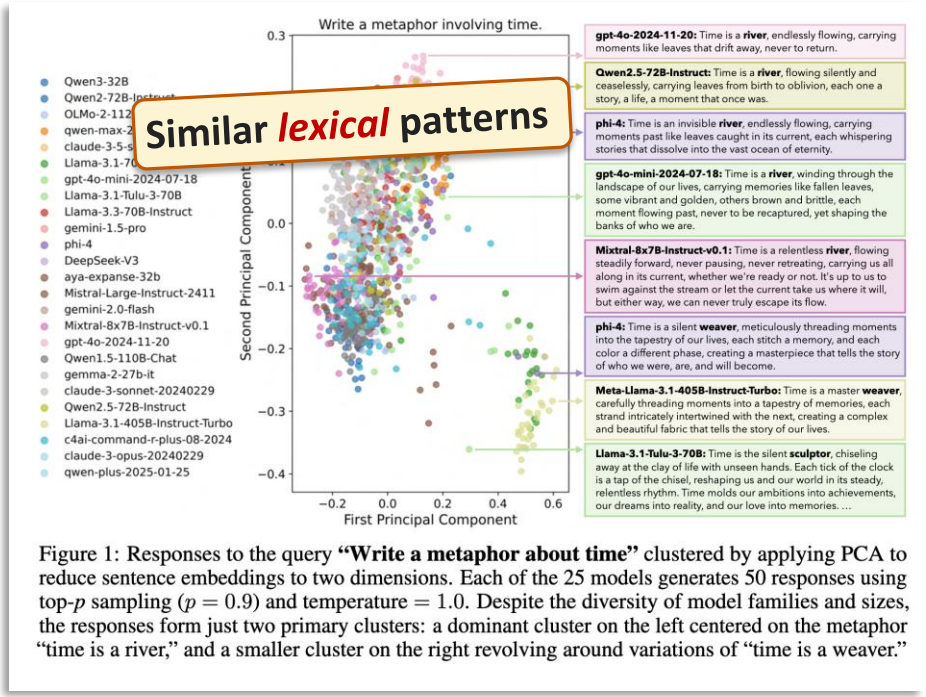


Figure 1: Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 25 models generates 50 responses using top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversity of model families and sizes, the responses form just two primary clusters: a dominant cluster on the left centered on the metaphor "time is a river," and a smaller cluster on the right revolving around variations of "time is a weaver."

Jiang et al., (NeurIPS 2025)

The University of Texas at Austin

Prior Work Shows: LLMs are Formulaic Generators

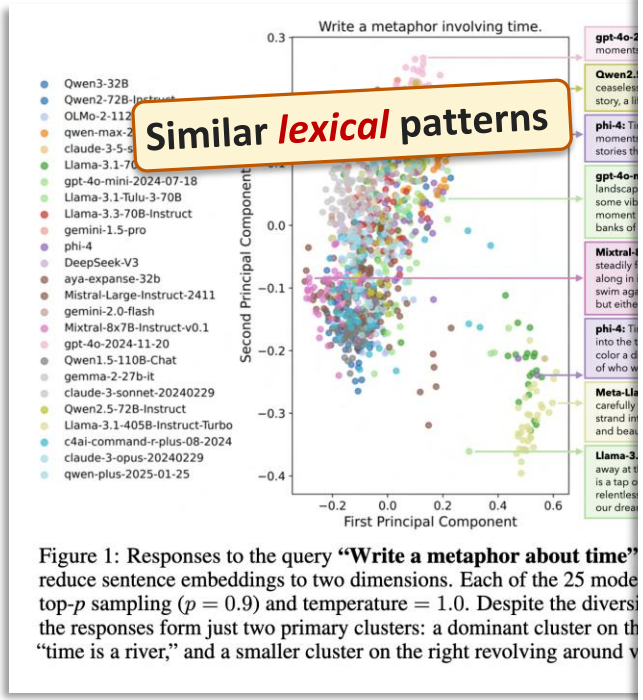
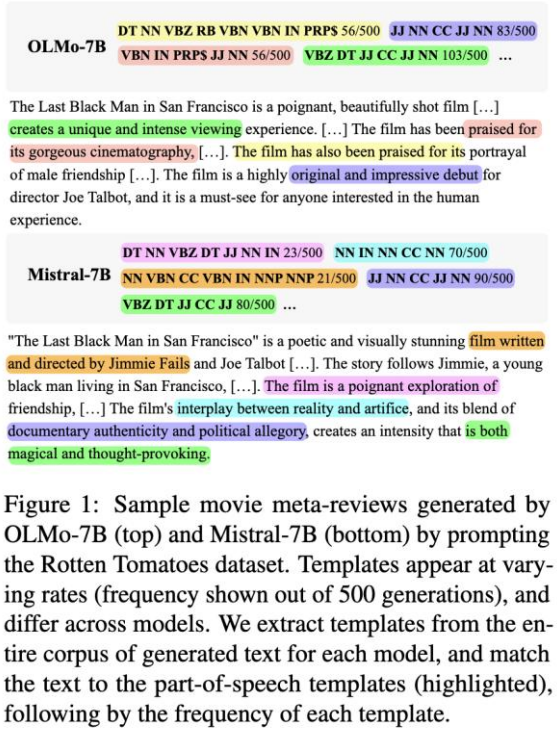


Figure 1: Responses to the query **"Write a metaphor about time"** reduce sentence embeddings to two dimensions. Each of the 25 mode top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversi the responses form just two primary clusters: a dominant cluster on th "time is a river," and a smaller cluster on the right revolving around v

Jiang et al., (NeurIPS 2025)



OLMo-7B DT NN VBZ RB VBN VBN IN PRP$ 56/500 JJ NN CC JJ NN 83/500 VBN IN PRP$ JJ NN 56/500 VBZ DT JJ CC JJ NN 103/500 ...

The Last Black Man in San Francisco is a poignant, beautifully shot film […] creates a unique and intense viewing experience. […] The film has been praised for its gorgeous cinematography, […]. The film has also been praised for its portrayal of male friendship […]. The film is a highly original and impressive debut for director Joe Talbot, and it is a must-see for anyone interested in the human experience.

Mistral-7B DT NN VBZ DT JJ NN IN 23/500 NN IN NN CC NN 70/500 NN VBN CC VBN IN NNP NNP 21/500 JJ NN CC JJ NN 90/500 VBZ DT JJ CC JJ 80/500 ...

"The Last Black Man in San Francisco" is a poetic and visually stunning film written and directed by Jimmie Fails and Joe Talbot […]. The story follows Jimmie, a young black man living in San Francisco, […]. The film is a poignant exploration of friendship, […] The film's interplay between reality and artifice, and its blend of documentary authenticity and political allegory, creates an intensity that is both magical and thought-provoking.

Figure 1: Sample movie meta-reviews generated by OLMo-7B (top) and Mistral-7B (bottom) by prompting the Rotten Tomatoes dataset. Templates appear at varying rates (frequency shown out of 500 generations), and differ across models. We extract templates from the entire corpus of generated text for each model, and match the text to the part-of-speech templates (highlighted), following by the frequency of each template.

Shaib et al., (EMNLP 2024)

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Prior Work Shows: LLMs are Formulaic Generators

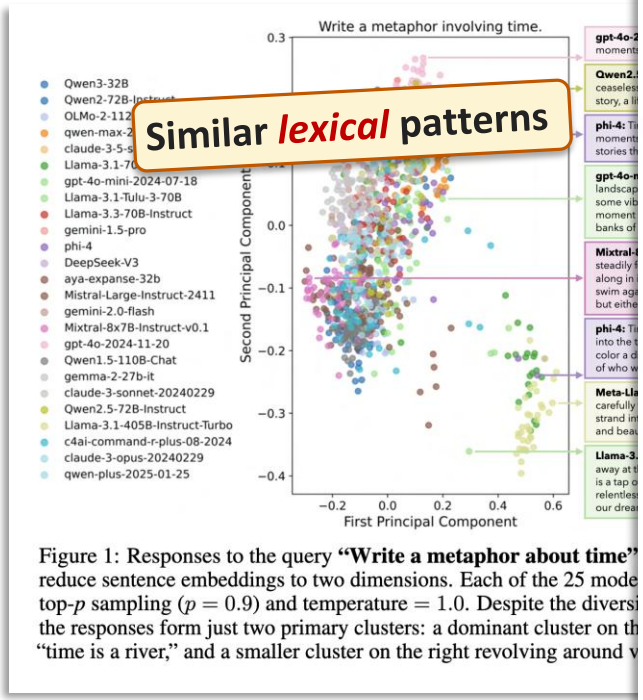


Figure 1: Responses to the query **"Write a metaphor about time"** reduce sentence embeddings to two dimensions. Each of the 25 mode top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversi the responses form just two primary clusters: a dominant cluster on th "time is a river," and a smaller cluster on the right revolving around v

Jiang et al., (NeurIPS 2025)

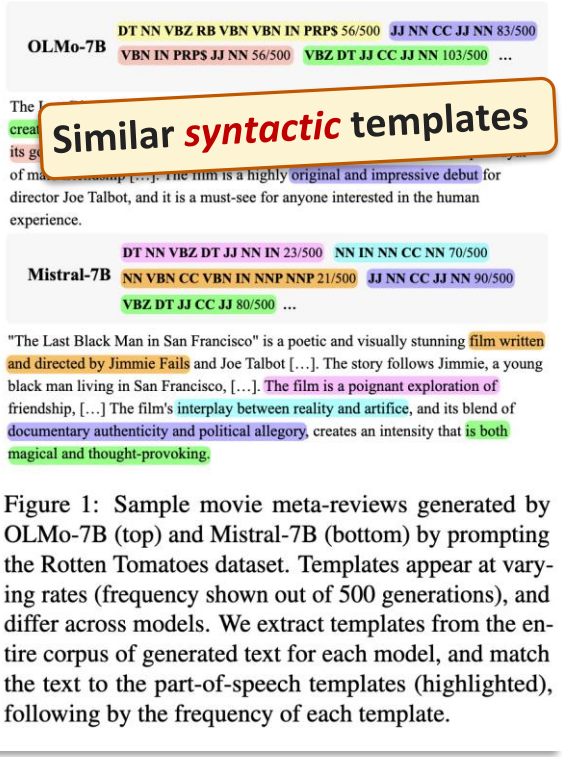


Figure 1: Sample movie meta-reviews generated by OLMo-7B (top) and Mistral-7B (bottom) by prompting the Rotten Tomatoes dataset. Templates appear at varying rates (frequency shown out of 500 generations), and differ across models. We extract templates from the entire corpus of generated text for each model, and match the text to the part-of-speech templates (highlighted), following by the frequency of each template.

Shaib et al., (EMNLP 2024)

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

TEXAS
The University of Texas at Austin

Prior Work Shows: LLMs are Formulaic Generators

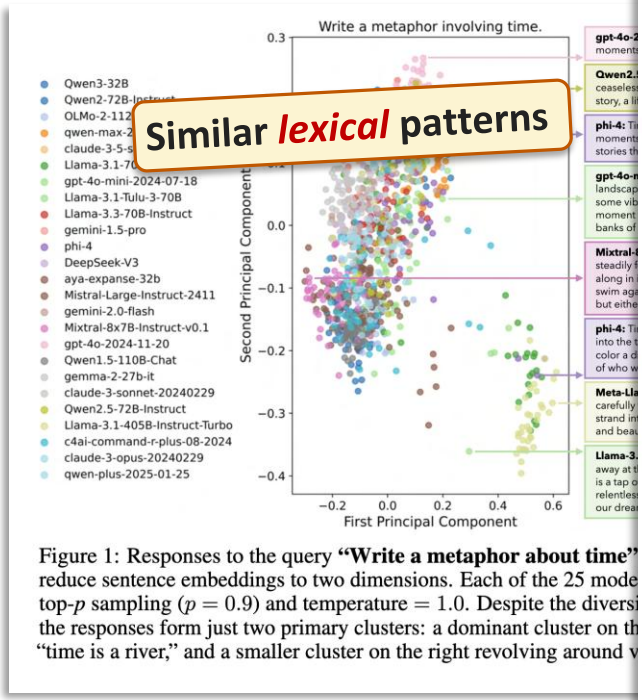


Figure 1: Responses to the query **"Write a metaphor about time"** reduce sentence embeddings to two dimensions. Each of the 25 mode top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversi the responses form just two primary clusters: a dominant cluster on th "time is a river," and a smaller cluster on the right revolving around v

Jiang et al., (NeurIPS 2025)

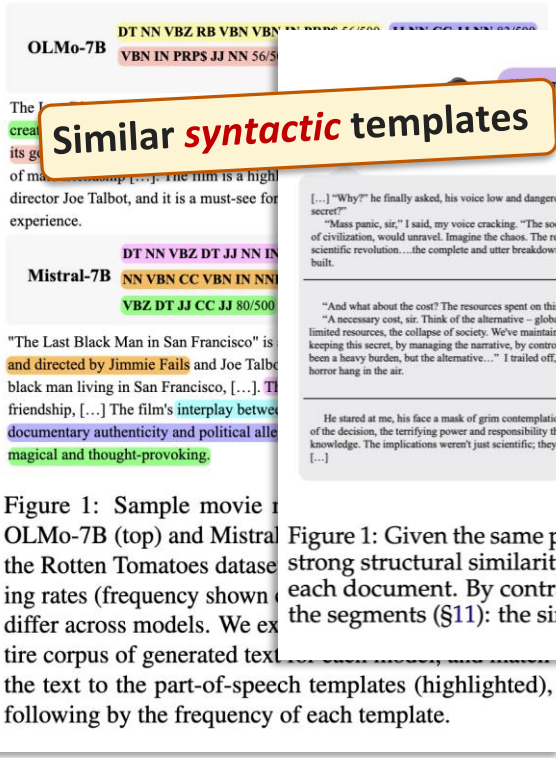


Figure 1: Sample movie r
OLMo-7B (top) and Mistral
the Rotten Tomatoes datase
ing rates (frequency shown
differ across models. We ex
tire corpus of generated text
the text to the part-of-speech templates (highlighted), following by the frequency of each template.

Shaib et al., (EMNLP 2024)

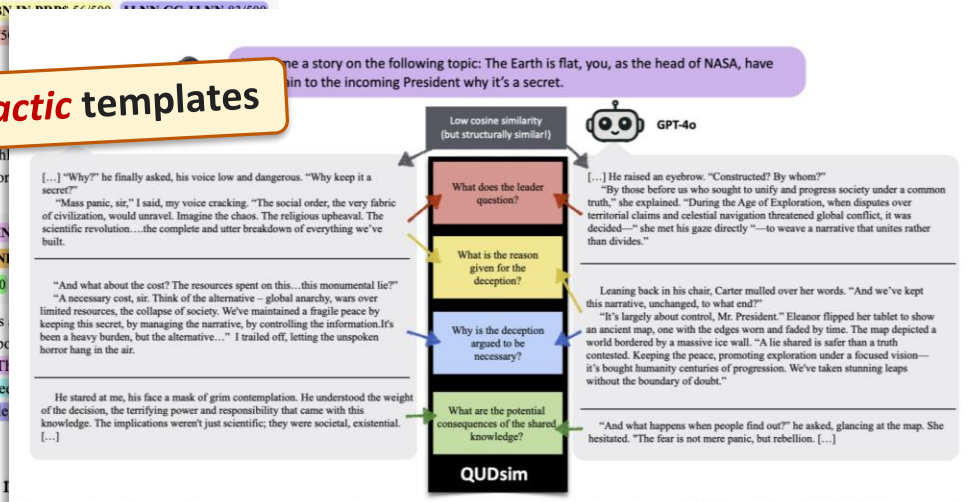


Figure 1: Given the same prompt, we generate stories using two LLMs. Our QUDSIM reveals strong structural similarity between the two texts by way of questions that are answered in each document. By contrast, embedding-based metrics do not recognize the similarity of the segments (§11): the similarity here is only structural.

Namuduri et al., (COLM 2025)

TEXAS
The University of Texas at Austin

Prior Work Shows: LLMs are Formulaic Generators

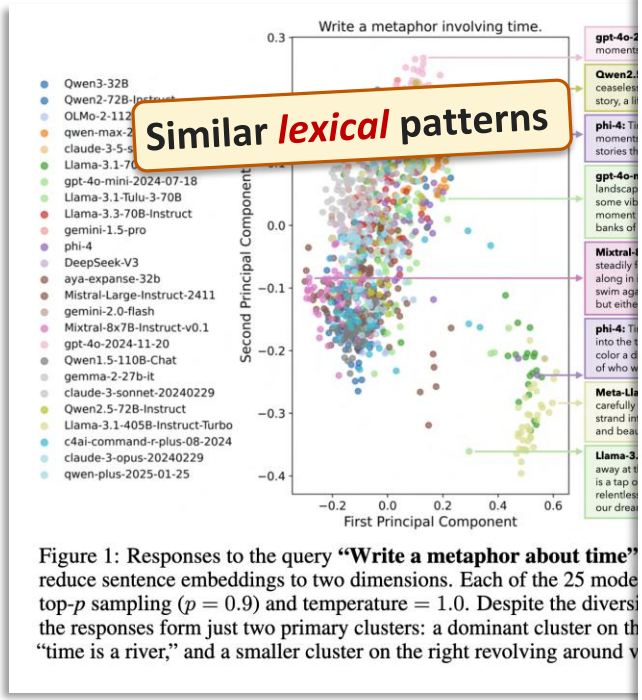


Figure 1: Responses to the query **"Write a metaphor about time"** reduce sentence embeddings to two dimensions. Each of the 25 mode top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversi the responses form just two primary clusters: a dominant cluster on th "time is a river," and a smaller cluster on the right revolving around v

Jiang et al., (NeurIPS 2025)

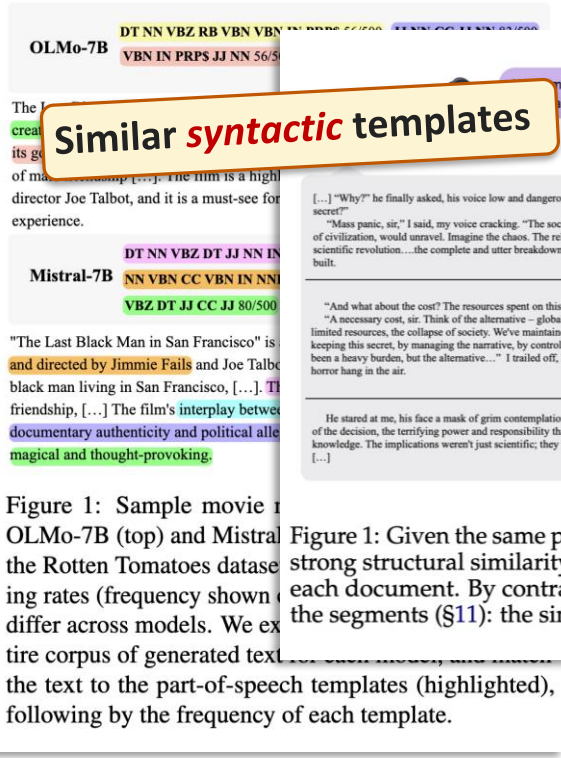


Figure 1: Sample movie r
OLMo-7B (top) and Mistral
the Rotten Tomatoes datase
ing rates (frequency shown
differ across models. We ex
tire corpus of generated text
the text to the part-of-speech templates (highlighted), following by the frequency of each template.

Shaib et al., (EMNLP 2024)

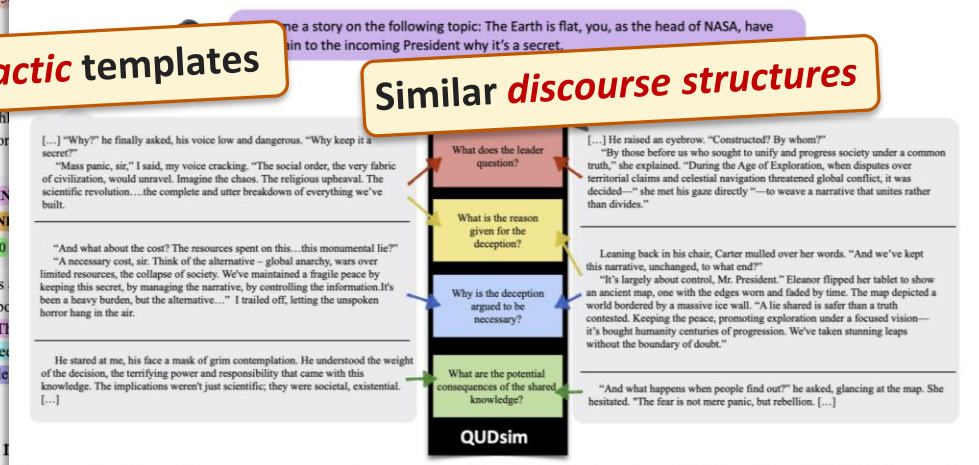
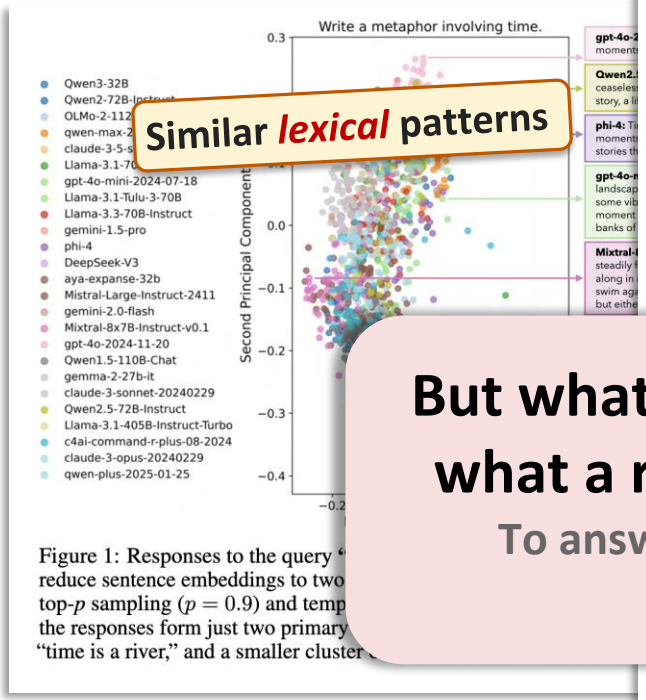


Figure 1: Given the same prompt, we generate stories using two LLMs. Our QUDSIM reveals strong structural similarity between the two texts by way of questions that are answered in each document. By contrast, embedding-based metrics do not recognize the similarity of the segments (§11): the similarity here is only structural.

Namuduri et al., (COLM 2025)

TEXAS
The University of Texas at Austin

Prior Work Shows: LLMs are Formulaic Generators



Similar *lexical* patterns

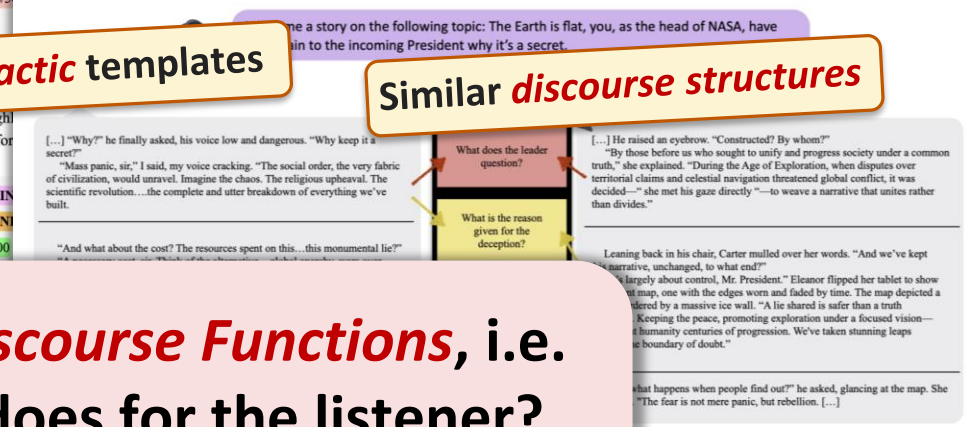
Figure 1: Responses to the query "
reduce sentence embeddings to two
top-p sampling ($p = 0.9$) and temp
the responses form just two primary
"time is a river," and a smaller cluster

Jiang et al., (NeurIPS 2025)

Similar *syntactic* templates

the text to the part-of-speech templates (highlighted), following by the frequency of each template.

Shaib et al., (EMNLP 2024)



Similar *discourse structures*

Namuduri et al., (COLM 2025)

But what about *Discourse Functions*, i.e. what a response does for the listener?

To answer this, we first need a vocabulary of empathic discourse moves.

two LLMs. Our QUDSIM reveals
of questions that are answered in
o not recognize the similarity of

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

A Taxonomy of Empathy Tactics

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

A Taxonomy of Empathy Tactics

Tactic	Description	*Example*	Tagger F1
`advice`	Providing ideas for solutions or coping strategies	*If I were you I would see a therapist. / Get some ice cream! / Definitely talk to your boss.*	0.87

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

A Taxonomy of Empathy Tactics

Tactic	Description	*Example*	Tagger F1
`advice`	Providing ideas for solutions or coping strategies	*If I were you I would see a therapist. / Get some ice cream! / Definitely talk to your boss.*	0.87
`assistance`	Offering some aid to the empathy-seeker	*I'm here for you if you want to talk. / Can I do anything to help?*	0.85
`emotional expression`	Communicating the empathy-giver's feelings, reactions, or thoughts	*I'm so sorry to hear that. / Wow, what a beautiful story.*	0.79
`empowerment`	Positive, uplifting statements about the empathy-seeker's character and capabilities	*You are going to get through this.*	0.79
`information`	Offering facts or resources (e.g., links)	*Flying is the safest form of travel.*	0.76
`paraphrasing`	Restating something the empathy-seeker said to demonstrate understanding of their situation, feelings, or experiences	*I'm hearing that you feel overwhelmed.*	0.76
`questioning`	Asking questions to improve understanding of the empathy-seeker's feelings, experiences, or situations.	*How are you feeling? / What do you think about [x]?*	0.94
`reappraisal`	Helping to engage in cognitive reappraisal (changing a belief)	*That was out of your control.*	0.59
`self-disclosure`	Sharing personal information or similar past experiences or feelings	*I've had that happen to me before too.*	0.78
`validation`	Reassures, normalizes, or validates an empathy-seeker's feelings	*Everyone has feelings like this. / You're not overreacting.*	0.82
Average			**0.80**

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

A Taxonomy of Empathy Tactics

Tactic	Description	*Example*	Tagger F1
advice	Providing ideas for solutions or coping strategies	*If I were you I would see a therapist. / Get some ice cream! / Definitely talk to your boss.*	0.87
assistance	Offering some aid to the empathy-seeker	*I'm here for you if you want to talk. / Can I do anything to help?*	0.85
emotional expression	Communicating the empathy-giver's feelings, reactions, or thoughts	*I'm so sorry to hear that. / Wow, what a beautiful story.*	0.79
empowerment	Positive, uplifting statements about the empathy-seeker's character and capabilities	*You are going to get through this.*	0.79
information	Offering facts or resources (e.g., links)	*Flying is the safest form of travel.*	0.76
paraphrasing	Restating something the empathy-seeker said to demonstrate understanding of their situation, feelings, or experiences	*I'm hearing that you feel overwhelmed.*	0.76
questioning	Asking questions to improve understanding of the empathy-seeker's feelings, experiences, or situations.	*How are you feeling? / What do you think about [x]?*	0.94
reappraisal	Helping to engage in cognitive reappraisal (changing a belief)	*That was out of your control.*	0.59
self-disclosure	Sharing personal information or similar past experiences or feelings	*I've had that happen to me before too.*	0.78
validation	Reassures, normalizes, or validates an empathy-seeker's feelings	*Everyone has feelings like this. / You're not overreacting.*	0.82
Average			**0.80**

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

A Taxonomy of Empathy Tactics

Tactics are about communicative function, not just wording

Tactic	Description	*Example*	Tagger F1
advice	Providing ideas for solutions or coping strategies	*If I were you I would see a therapist. / Get some ice cream! / Definitely talk to your boss.*	0.87
assistance	Offering some aid to the empathy-seeker	*I'm here for you if you want to talk. / Can I do anything to help?*	0.85
emotional expression	Communicating the empathy-giver's feelings, reactions, or thoughts	*I'm so sorry to hear that. / Wow, what a beautiful story.*	0.79
empowerment	Positive, uplifting statements about the empathy-seeker's character and capabilities	*You are going to get through this.*	0.79
information	Offering facts or resources (e.g., links)	*Flying is the safest form of travel.*	0.76
paraphrasing	Restating something the empathy-seeker said to demonstrate understanding of their situation, feelings, or experiences	*I'm hearing that you feel overwhelmed.*	0.76
questioning	Asking questions to improve understanding of the empathy-seeker's feelings, experiences, or situations.	*How are you feeling? / What do you think about [x]?*	0.94
reappraisal	Helping to engage in cognitive reappraisal (changing a belief)	*That was out of your control.*	0.59
self-disclosure	Sharing personal information or similar past experiences or feelings	*I've had that happen to me before too.*	0.78
validation	Reassures, normalizes, or validates an empathy-seeker's feelings	*Everyone has feelings like this. / You're not overreacting.*	0.82
Average			**0.80**

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

A Taxonomy of Empathy Tactics

Tactics are about communicative function, not just wording

Tactic	Description	*Example*	Tagger F1
advice	Providing ideas for solutions or coping strategies	*If I were you I would see a therapist. / Get some ice cream! / Definitely talk to your boss.*	0.87
assistance	Offering some aid to the empathy-seeker	*I'm here for you if you want to talk. / Can I do anything to help?*	0.85
emotional expression	Communicating the empathy-giver's feelings, reactions, or thoughts	*I'm so sorry to hear that. / Wow, what a beautiful story.*	0.79
empowerment	Positive, uplifting statements about the empathy-seeker's character and capabilities	*You are going to get through this.*	0.79
information	Offering facts or resources (e.g., links)	*Flying is the safest form of travel.*	0.76
paraphrasing	Restating something the empathy-seeker said to demonstrate understanding of their situation, feelings, or experiences	*I'm hearing that you feel overwhelmed.*	0.76
questioning	Asking questions to improve understanding of the empathy-seeker's feelings, experiences, or situations.	*How are you feeling? / What do you think about [x]?*	0.94
reappraisal	Helping to engage in cognitive reappraisal (changing a belief)	*That was out of your control.*	0.59
self-disclosure	Sharing personal information or similar past experiences or feelings	*I've had that happen to me before too.*	0.78
validation	Reassures, normalizes, or validates an empathy-seeker's feelings	*Everyone has feelings like this. / You're not overreacting.*	0.82
Average			**0.80**

We train automatic taggers using Llama-3.1-8b-instruct

Taxonomy from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

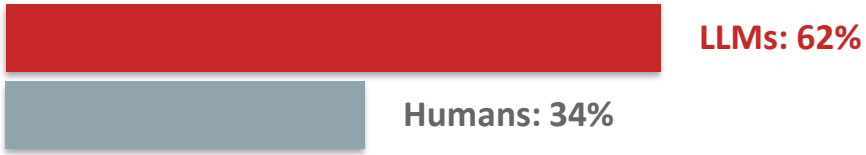
Paraphrasing → Validation → Paraphrasing

Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

Paraphrasing → Validation → Paraphrasing

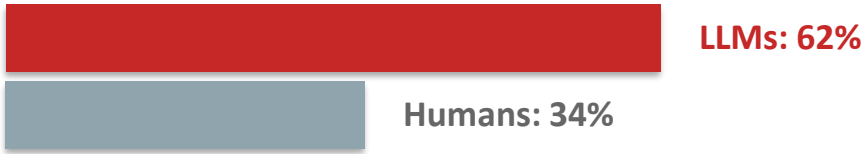


Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

Paraphrasing → Validation → Paraphrasing



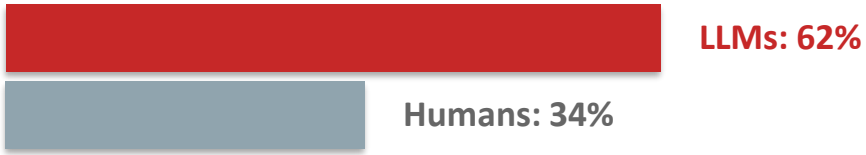
Validation → Paraphrasing → Validation

Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

Paraphrasing → Validation → Paraphrasing



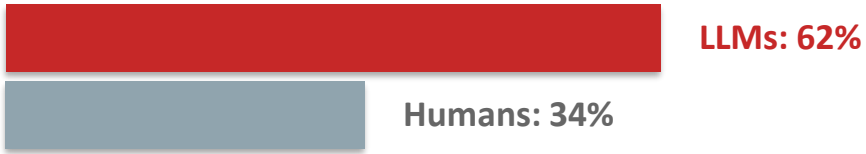
Validation → Paraphrasing → Validation

Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

Paraphrasing → Validation → Paraphrasing



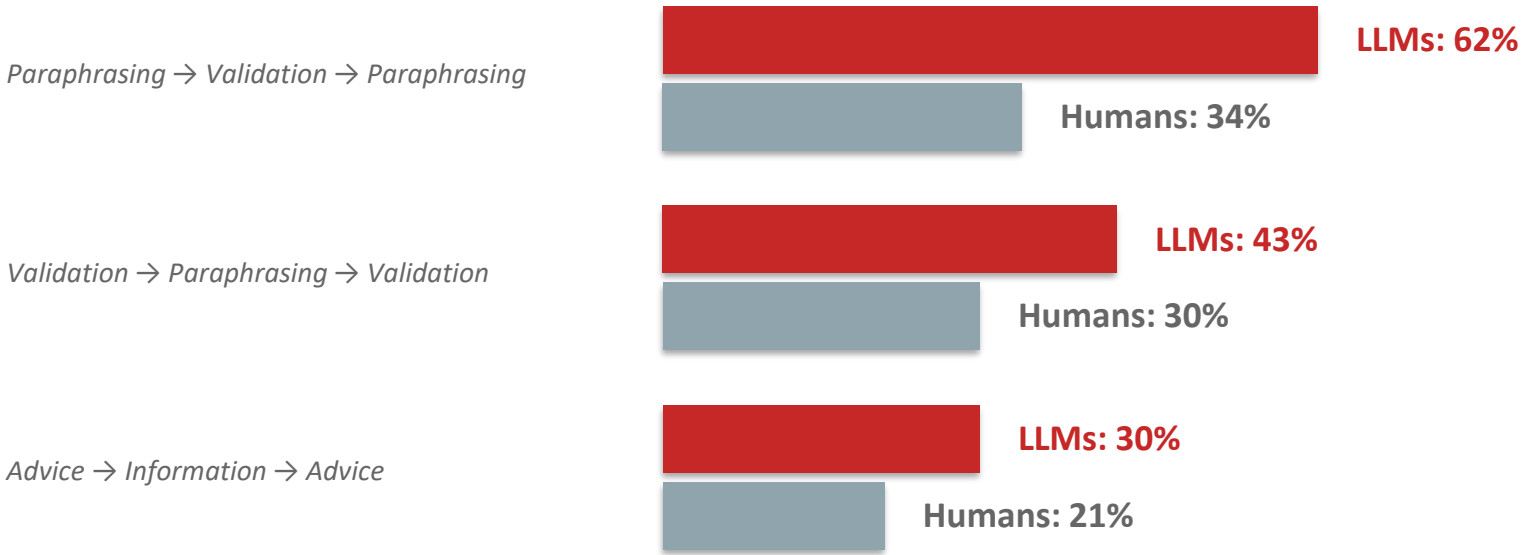
Validation → Paraphrasing → Validation

Advice → Information → Advice

Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

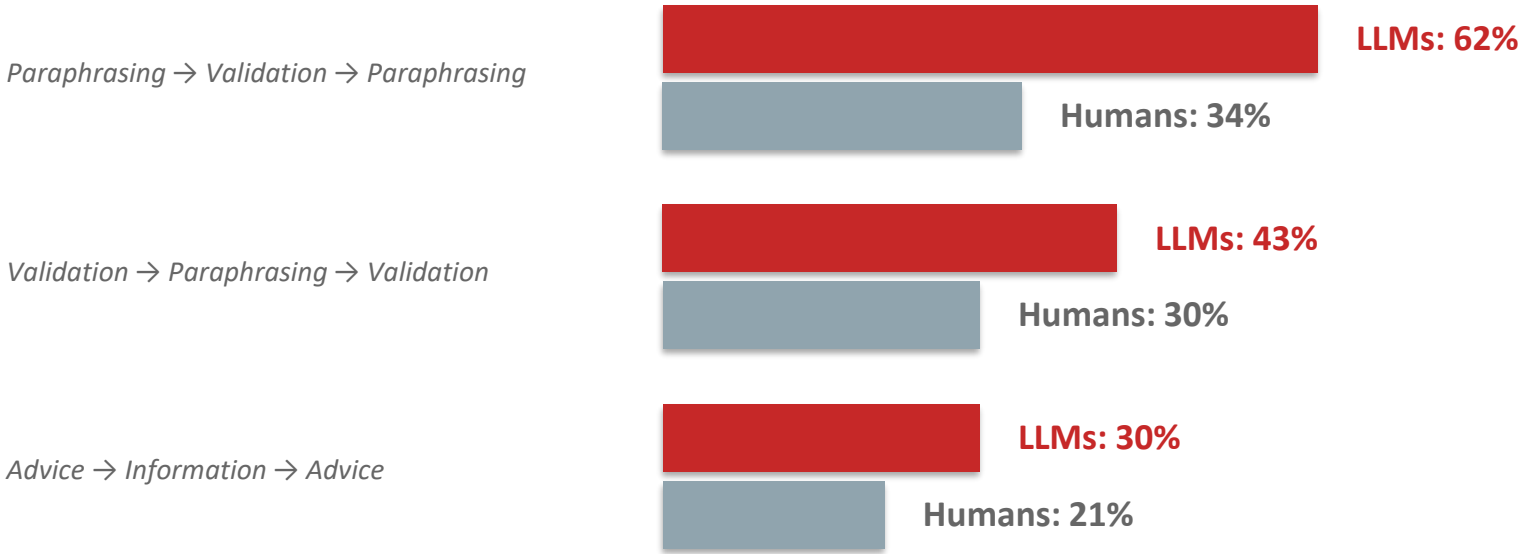


Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

TEXAS
The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

"I understand" → "That sounds tough" → "I hear you"
Same discourse moves, every time.

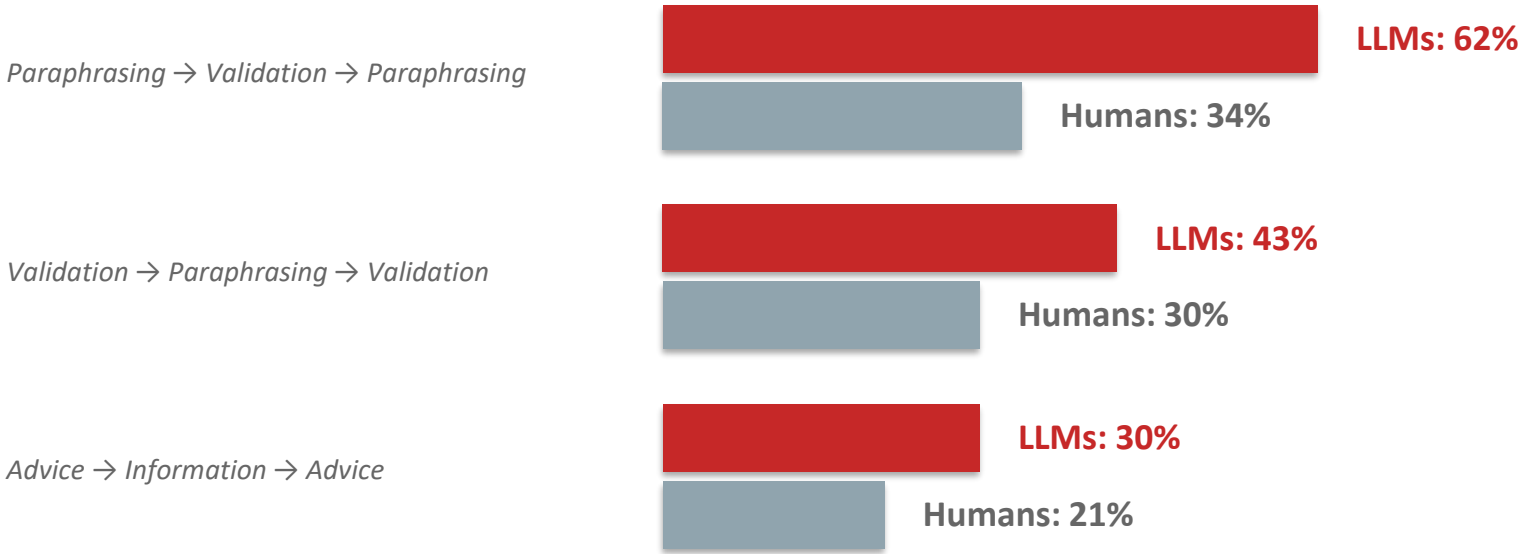


Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

TEXAS
The University of Texas at Austin

Single-Turn: LLMs Loop Through the Same Tactics

"I understand" → "That sounds tough" → "I hear you"
Same discourse moves, every time.



LLMs are stuck in loops. This is single-turn.
Does the rigidity compound across multiple turns?

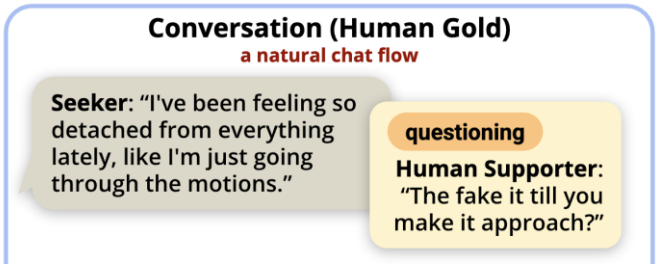
Findings from: Gueorguieva, **Zhan**, Suh, Hernandez, Lau, Li, Ong (2026)

The University of Texas at Austin

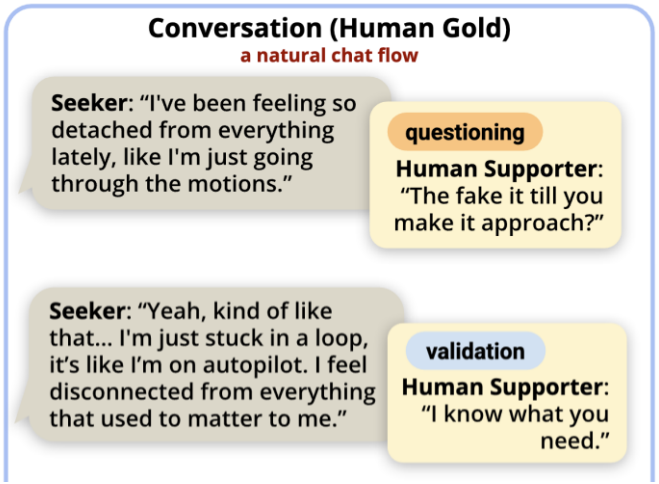
Does This Rigidity Compound Across Turns?

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Does This Rigidity Compound Across Turns?

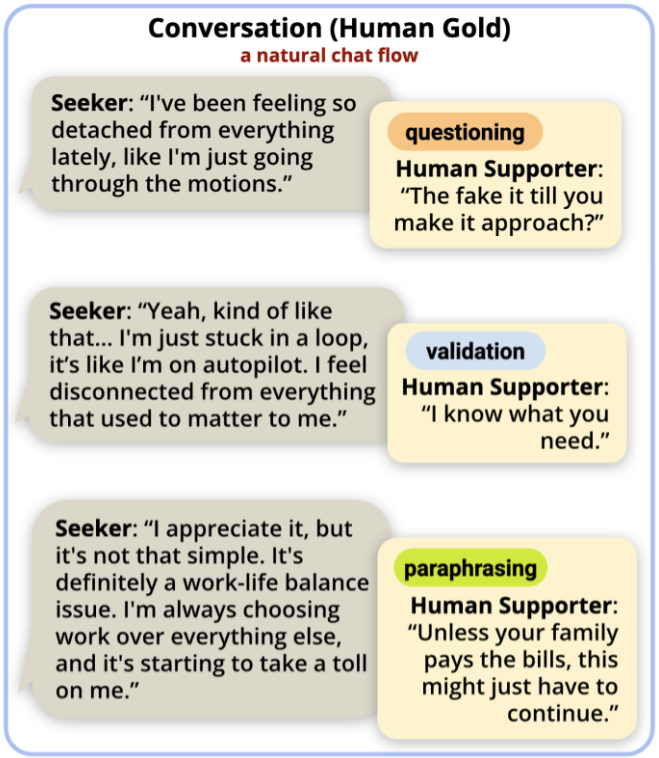


Does This Rigidity Compound Across Turns?



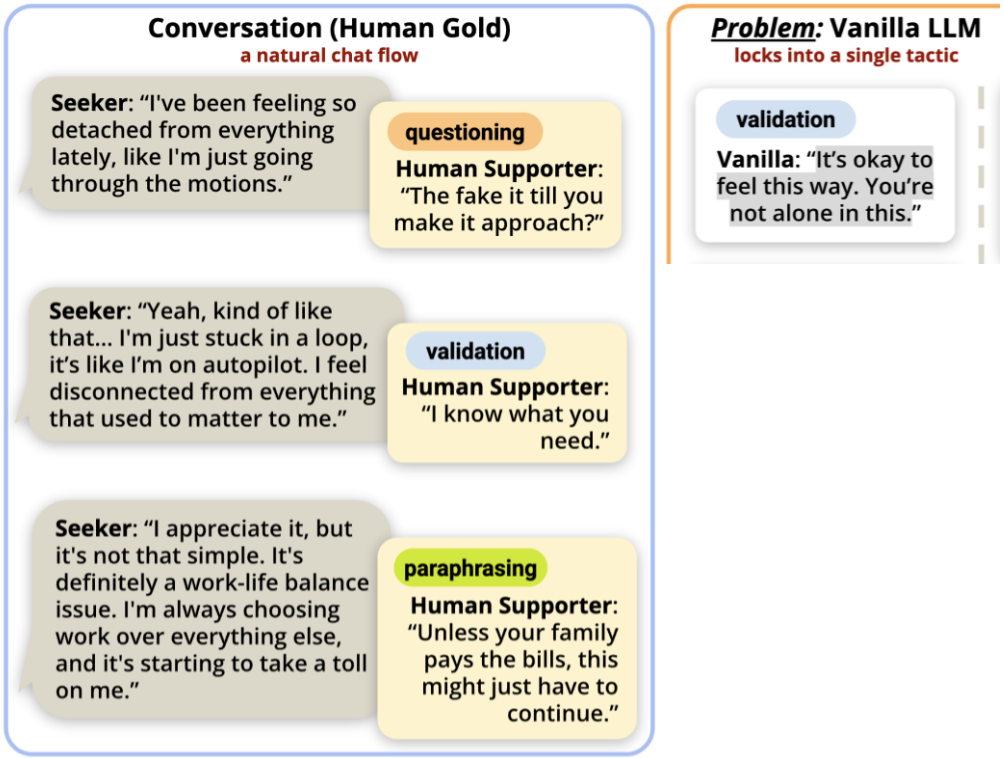
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Does This Rigidity Compound Across Turns?



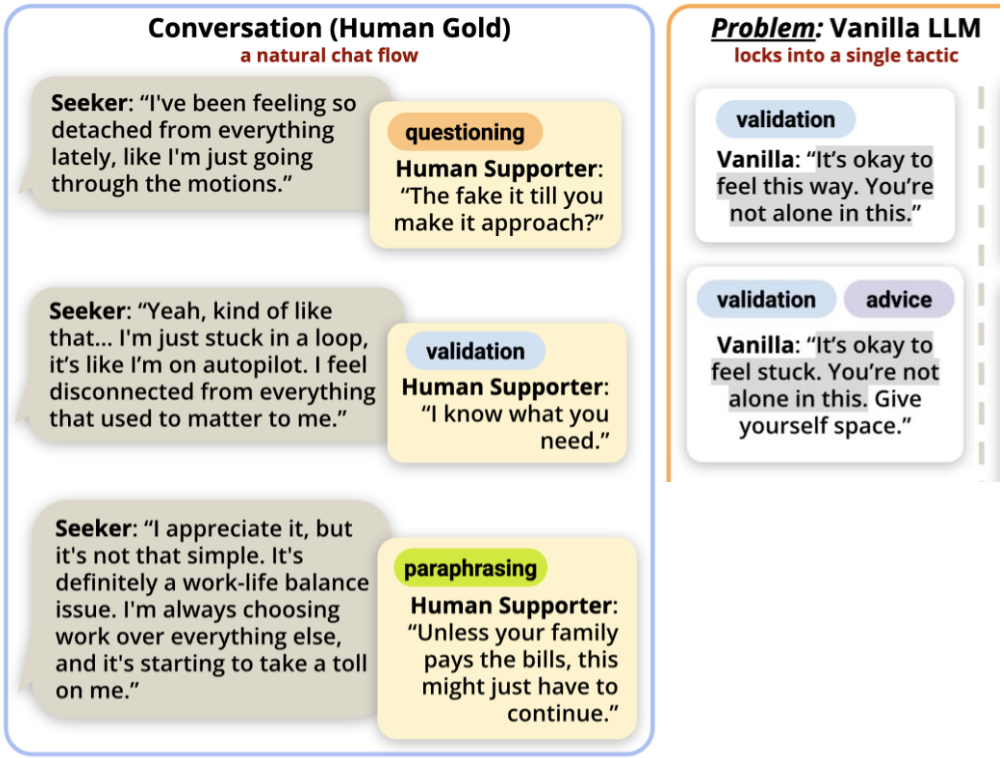
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

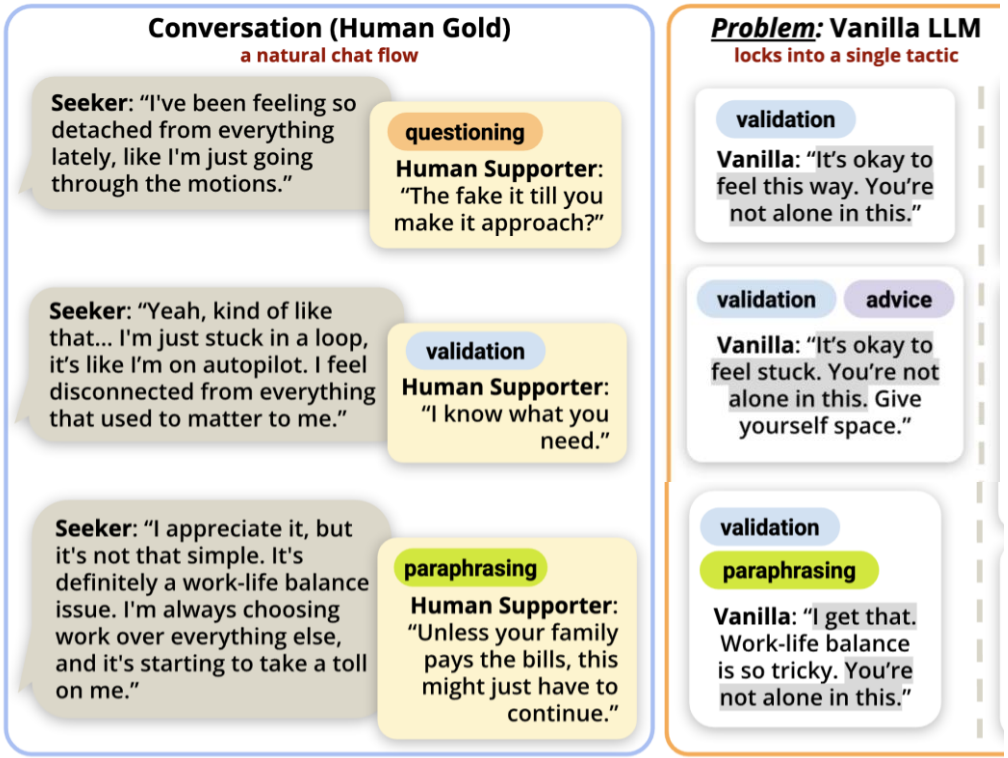
Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

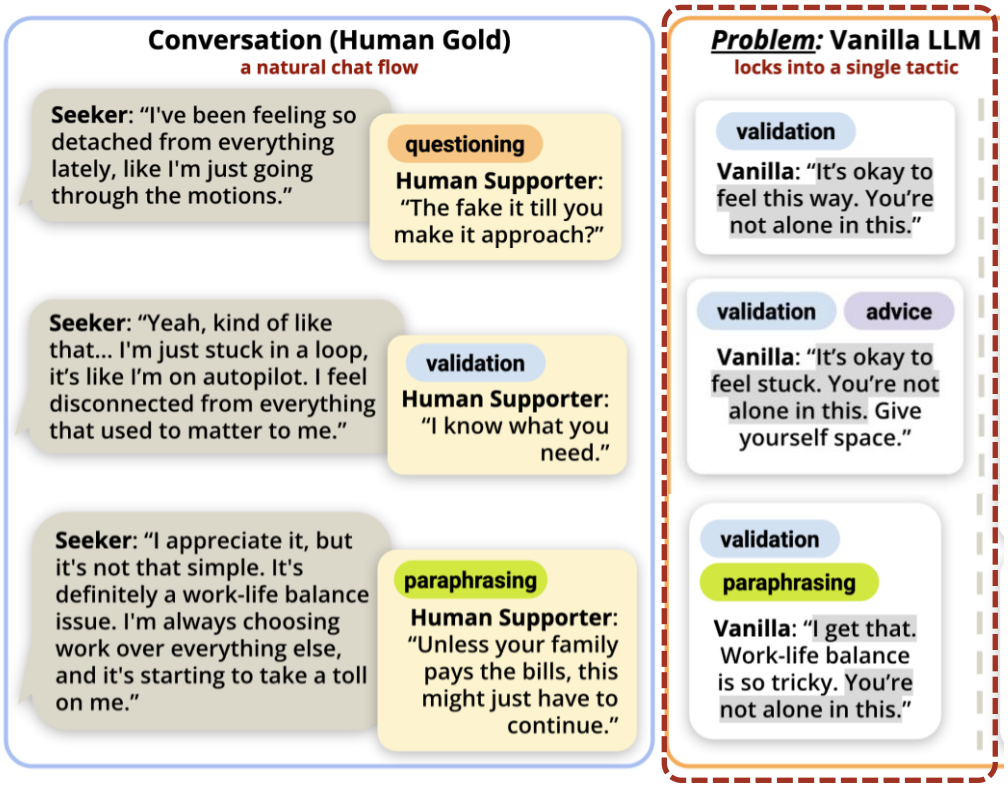
Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

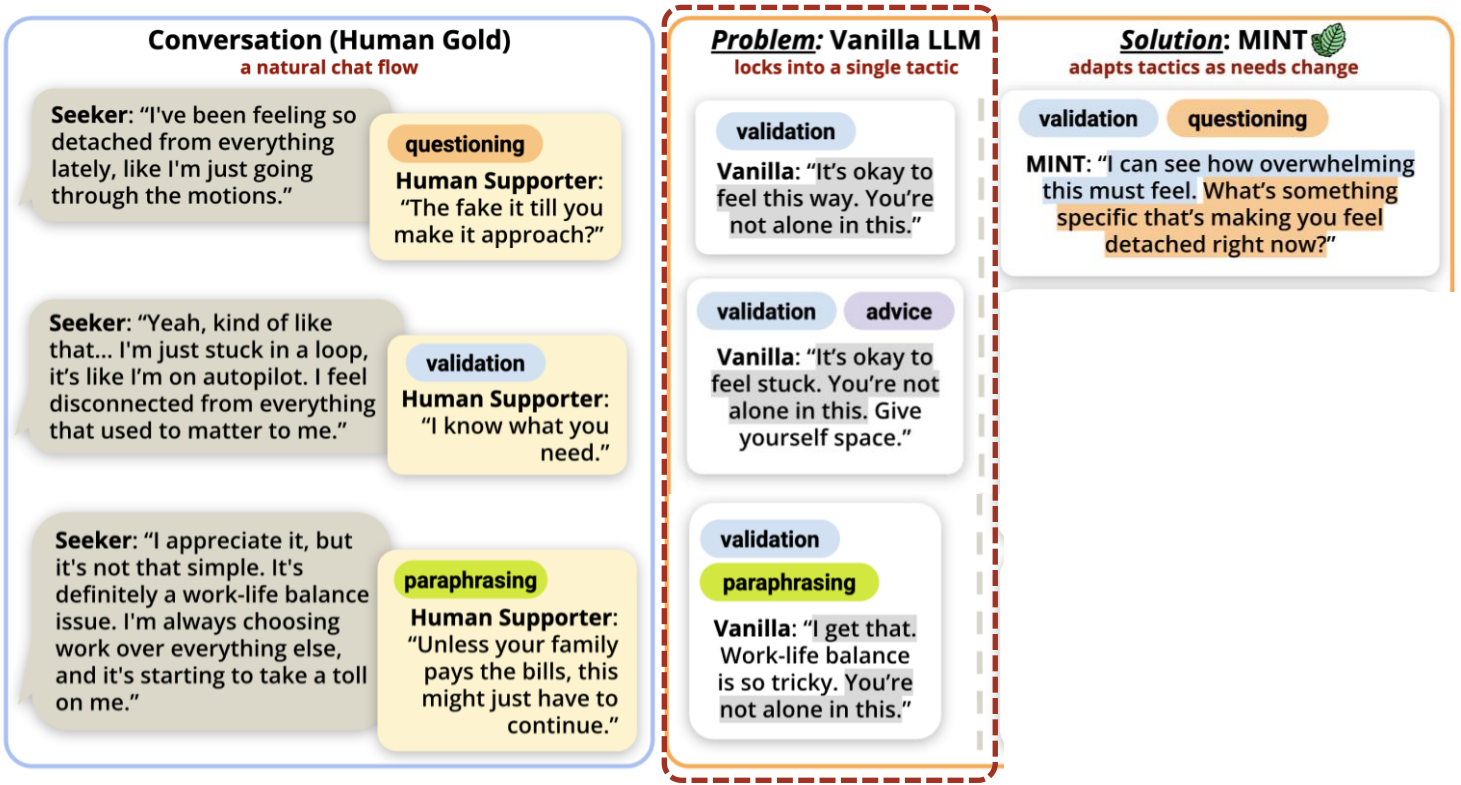
The University of Texas at Austin

Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

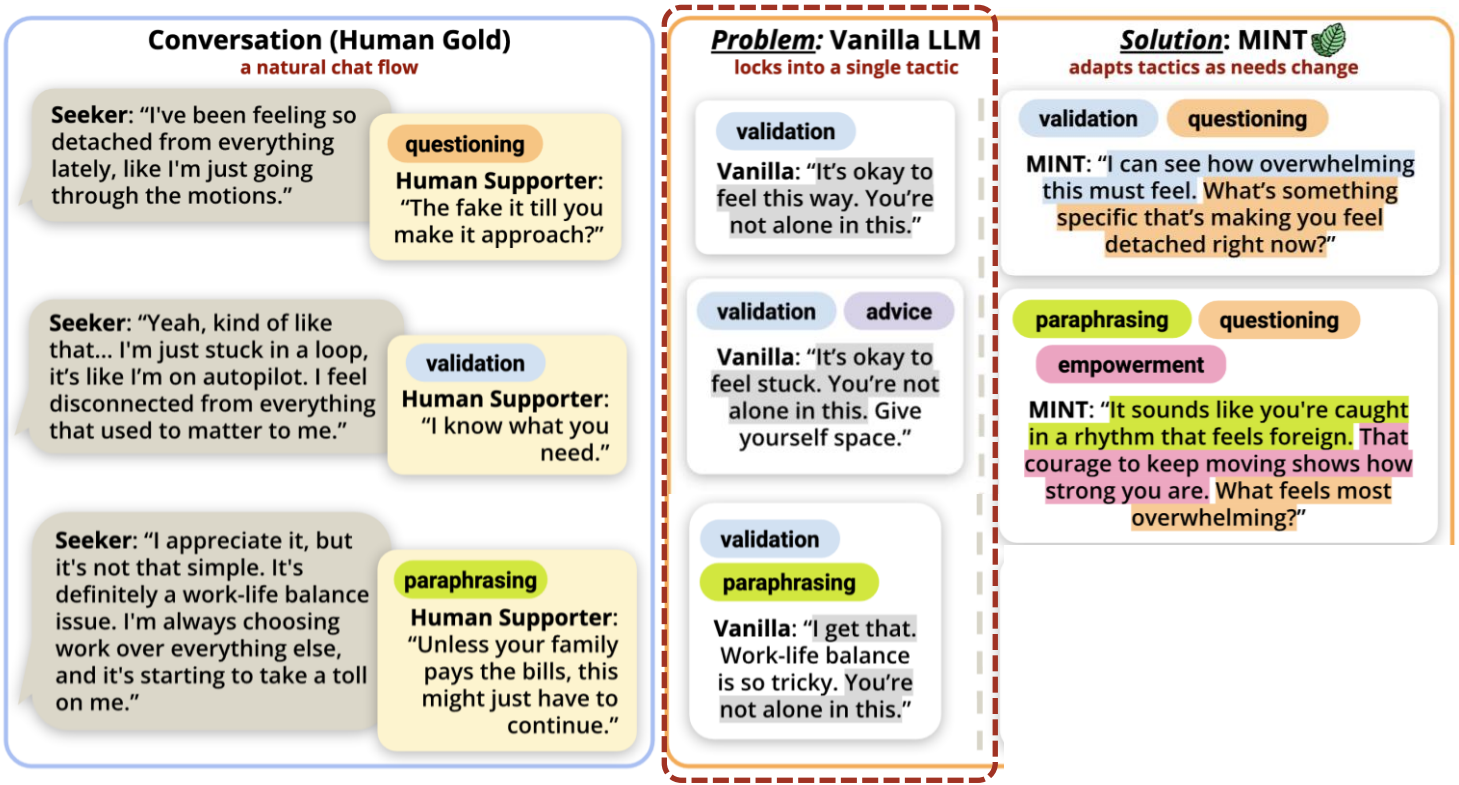
Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

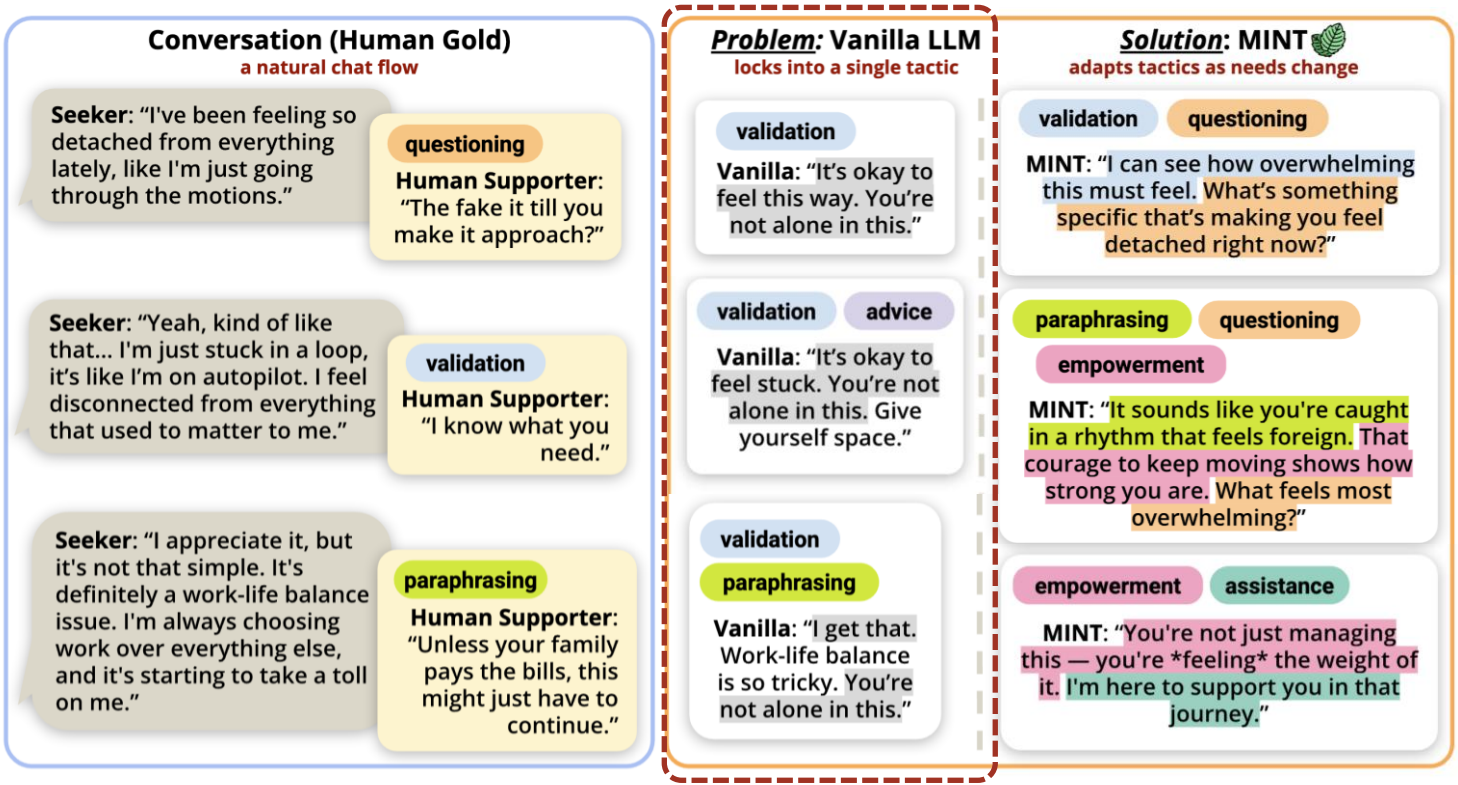
Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

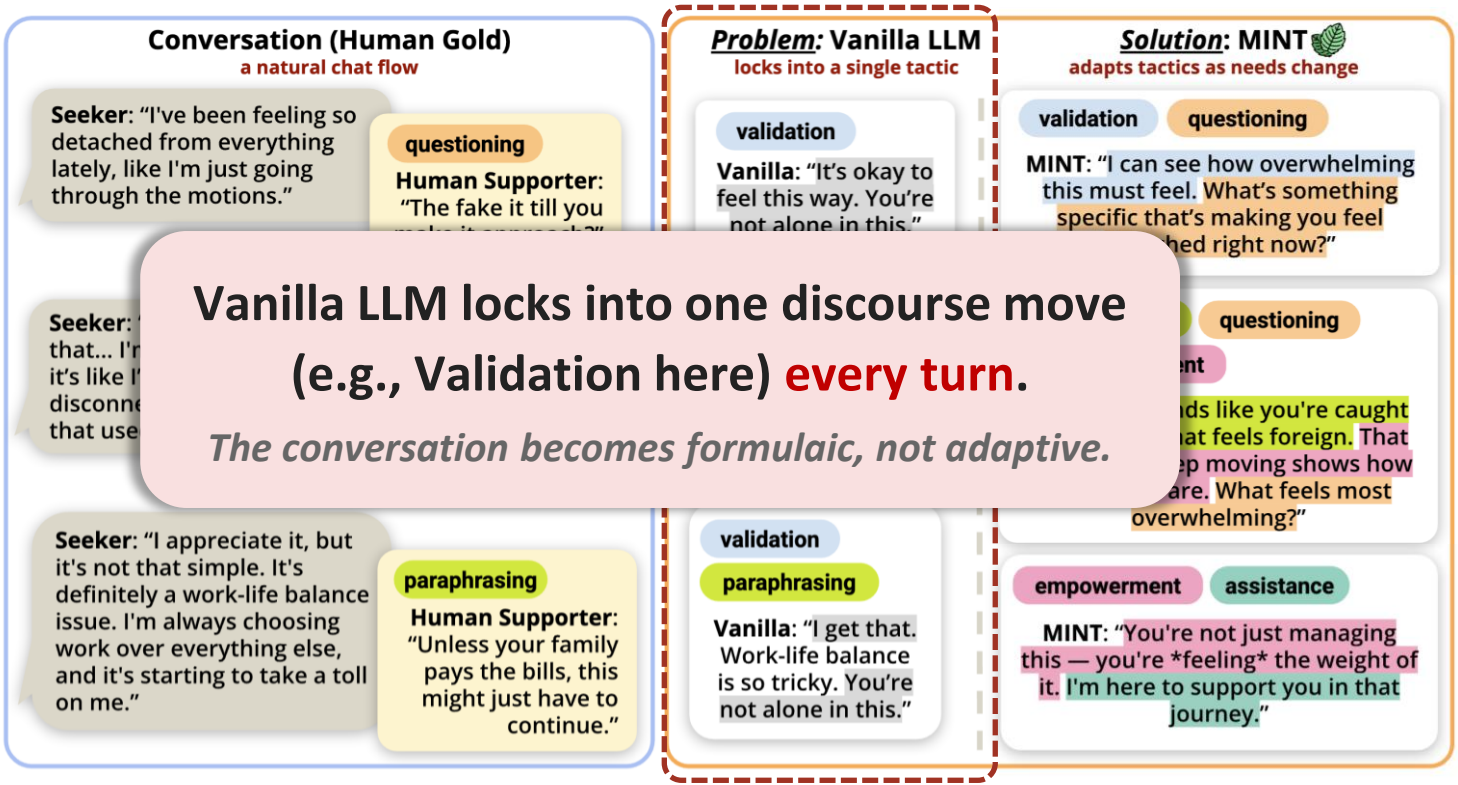
The University of Texas at Austin

Does This Rigidity Compound Across Turns?



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Does This Rigidity Compound Across Turns?



Vanilla LLM locks into one discourse move (e.g., Validation here) every turn.

The conversation becomes formulaic, not adaptive.

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ *users need different kinds of support as the conversation unfolds*

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ *users need different kinds of support as the conversation unfolds*

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"

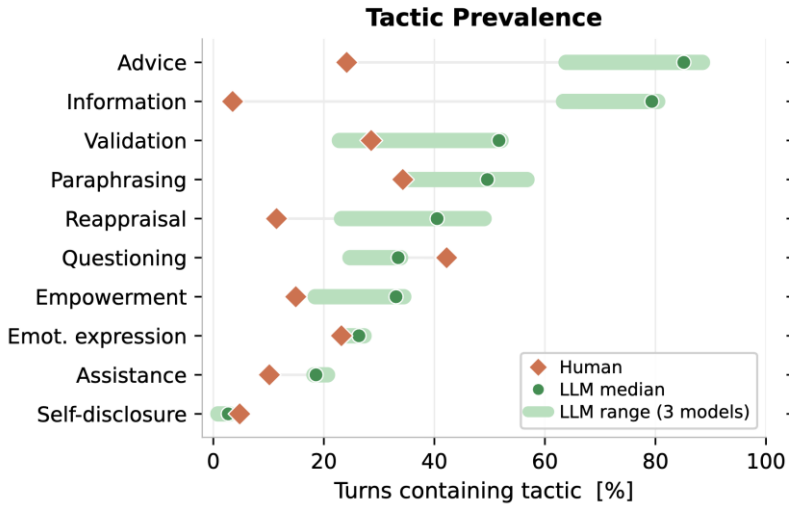
This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

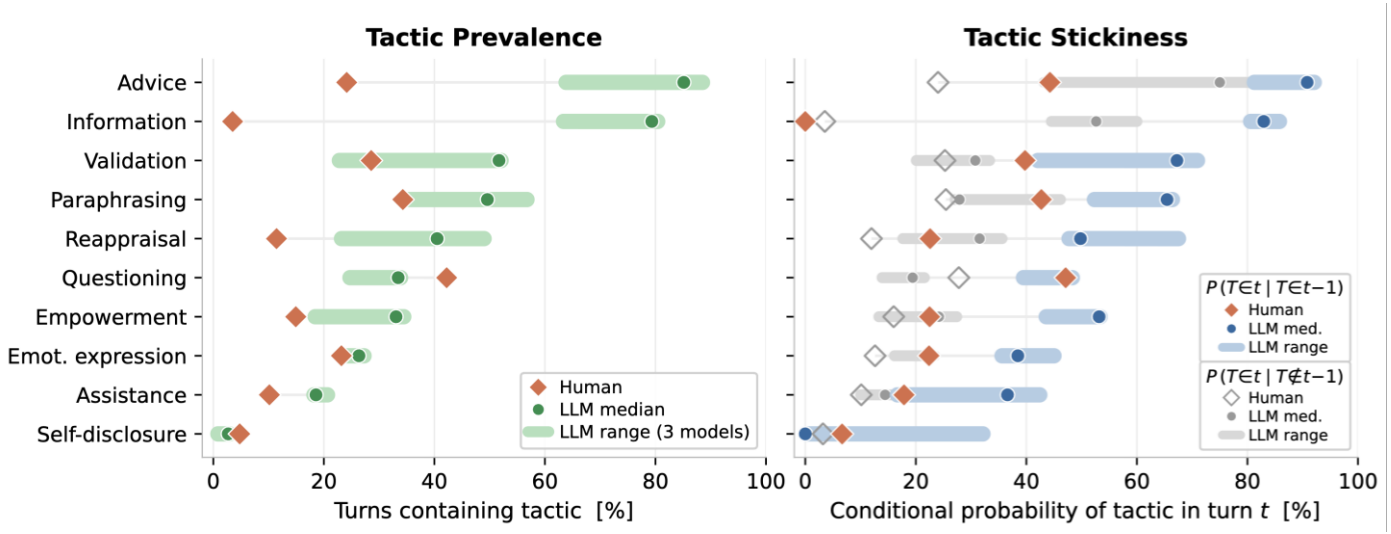
This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

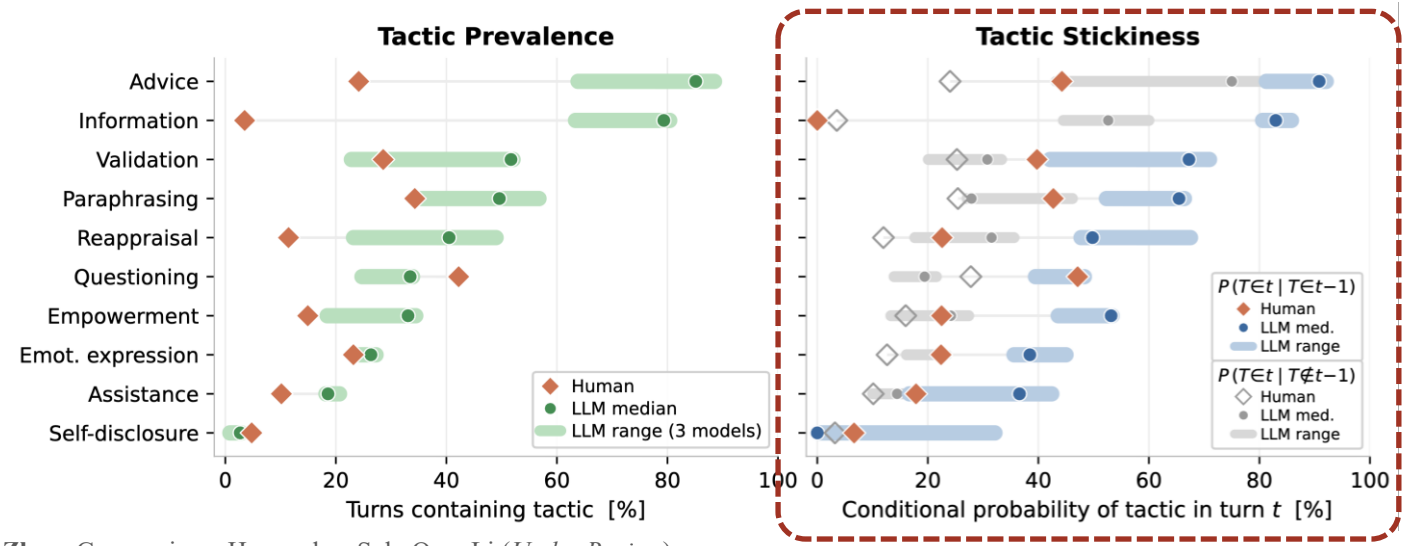
This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

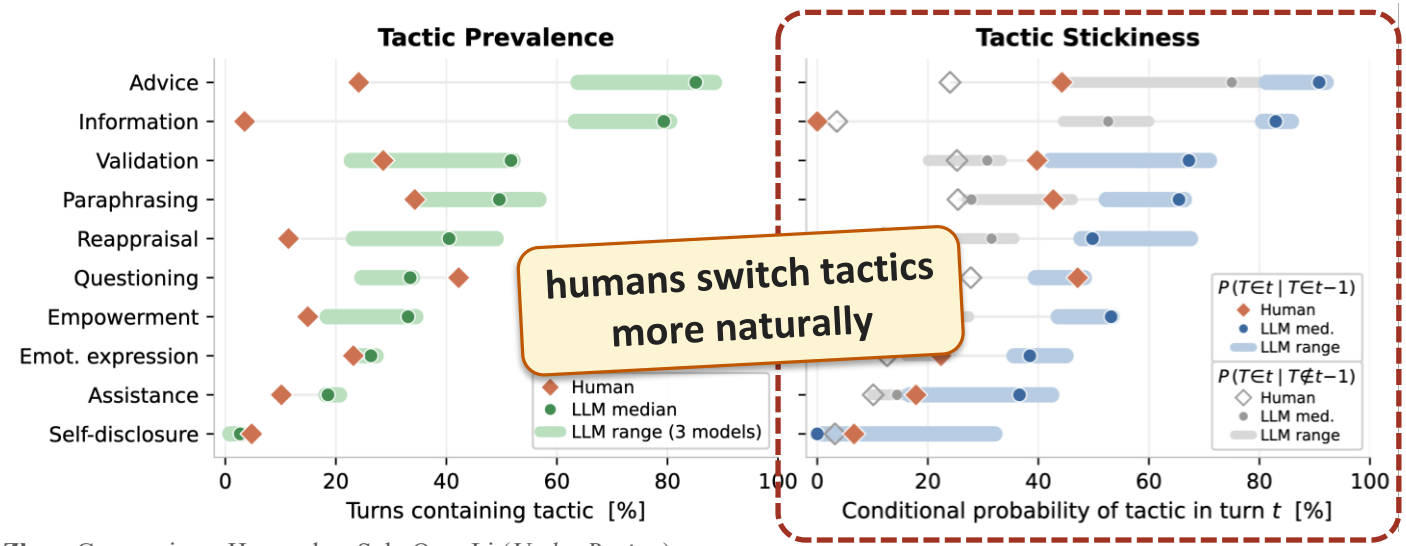
This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

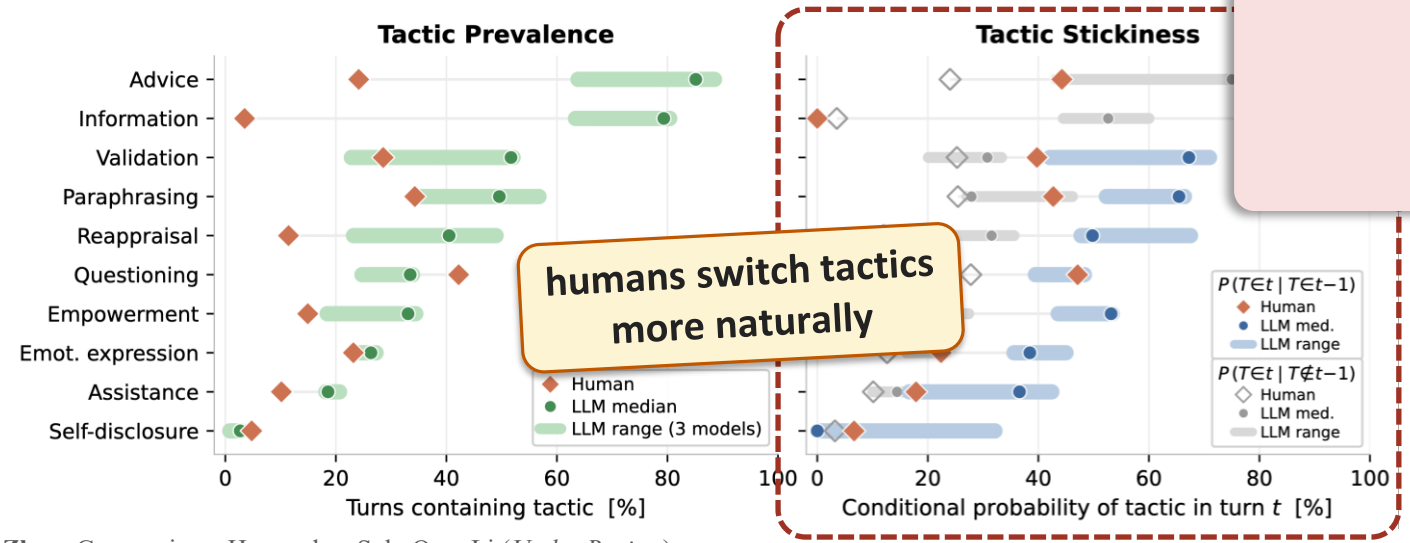
TEXAS
The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:
⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



humans switch tactics more naturally

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:

⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

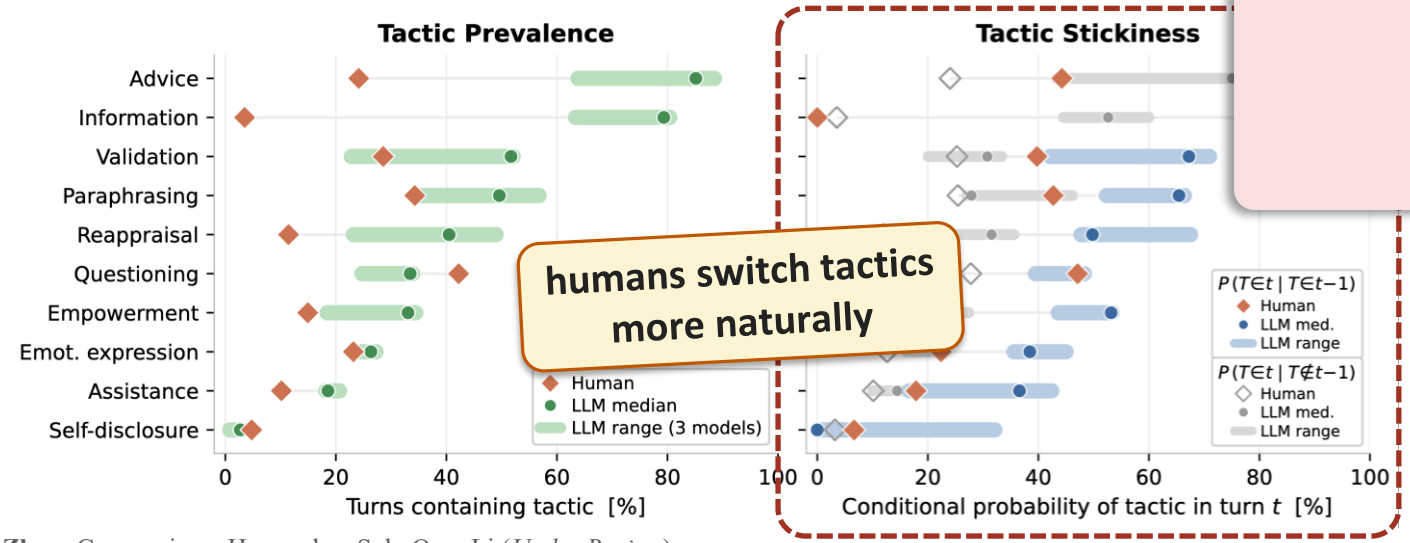
"How likely is the model to repeat the same discourse move?"

BERTScore

LLMs: 0.86

Humans: 0.87

Looks identical!



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

TEXAS
The University of Texas at Austin

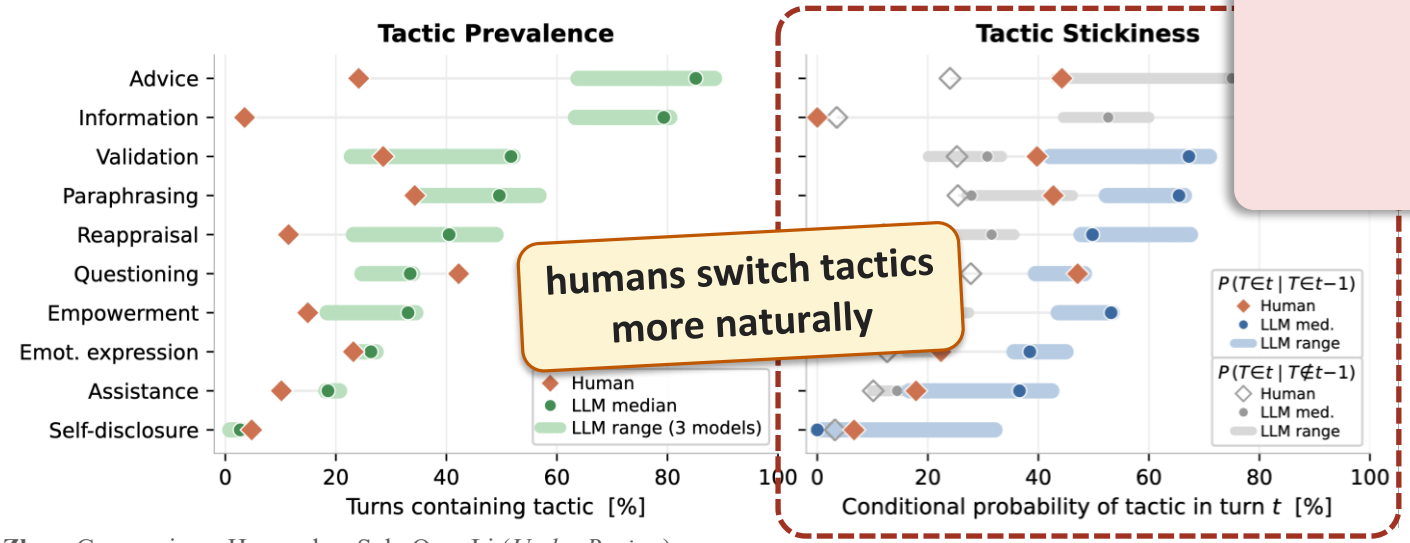
This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:
⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"

BERTScore	Tactic Stickiness
LLMs: 0.86	LLMs: 0.50-0.56
Humans: 0.87	Humans: 0.27
Looks identical!	Nearly 2× gap!



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

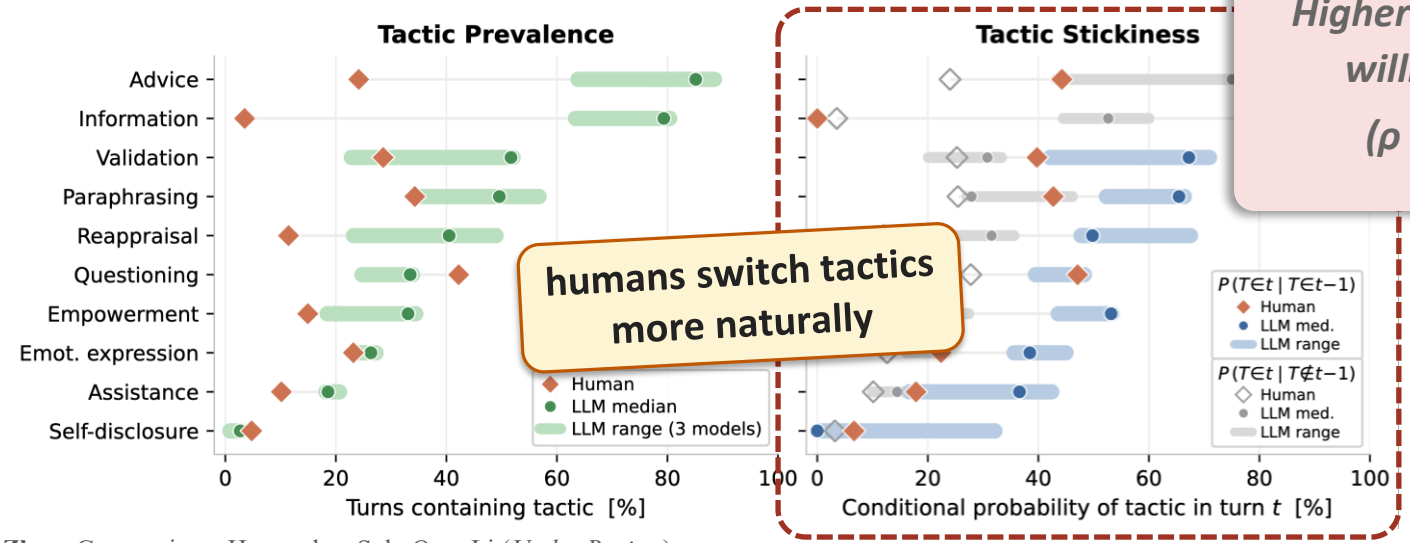
The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:
⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



humans switch tactics more naturally

BERTScore	Tactic Stickiness
LLMs: 0.86	LLMs: 0.50-0.56
Humans: 0.87	Humans: 0.27
Looks identical!	Nearly 2× gap!

Higher stickiness → lower user willingness to re-engage

(ρ = −0.287, p = 0.017)

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

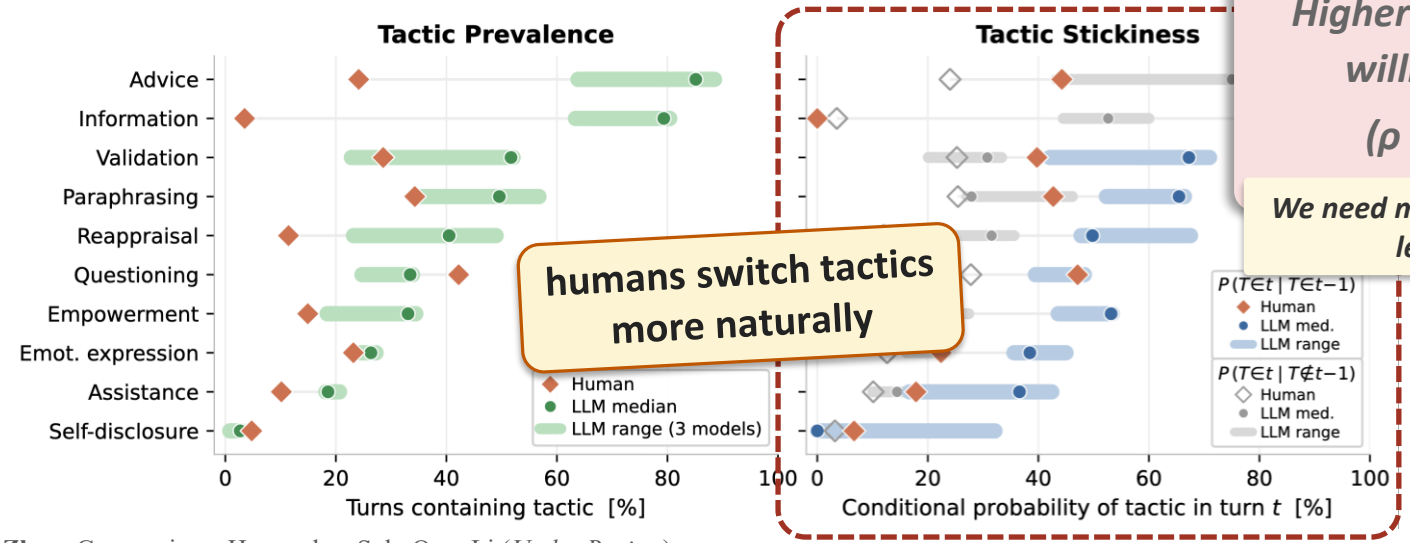
The University of Texas at Austin

This Rigidity Compounds in Multi Turn Dialogue!

In supportive conversations, rigidity is especially costly:
⇨ users need different kinds of support as the conversation unfolds

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

"How likely is the model to repeat the same discourse move?"



humans switch tactics more naturally

BERTScore	Tactic Stickiness
LLMs: 0.86	LLMs: 0.50-0.56
Humans: 0.87	Humans: 0.27
Looks identical!	Nearly 2× gap!

Higher stickiness → lower user willingness to re-engage

(ρ = −0.287, p = 0.017)

We need metrics and training signals at the level of discourse moves.

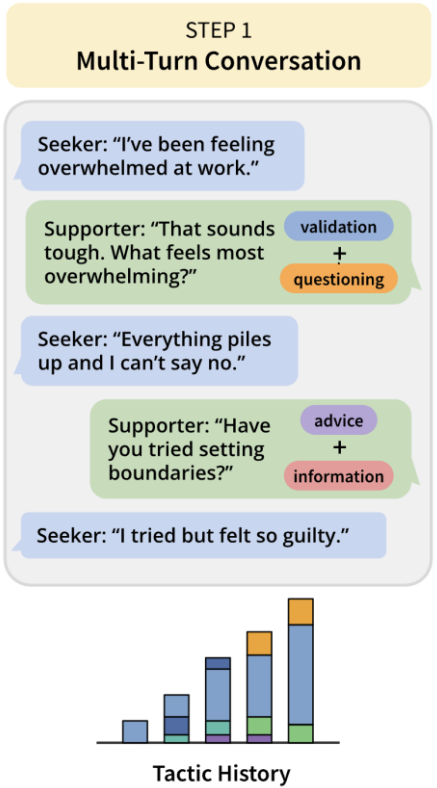
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

MINT: Multi-turn Inter-tactic Novelty Training

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

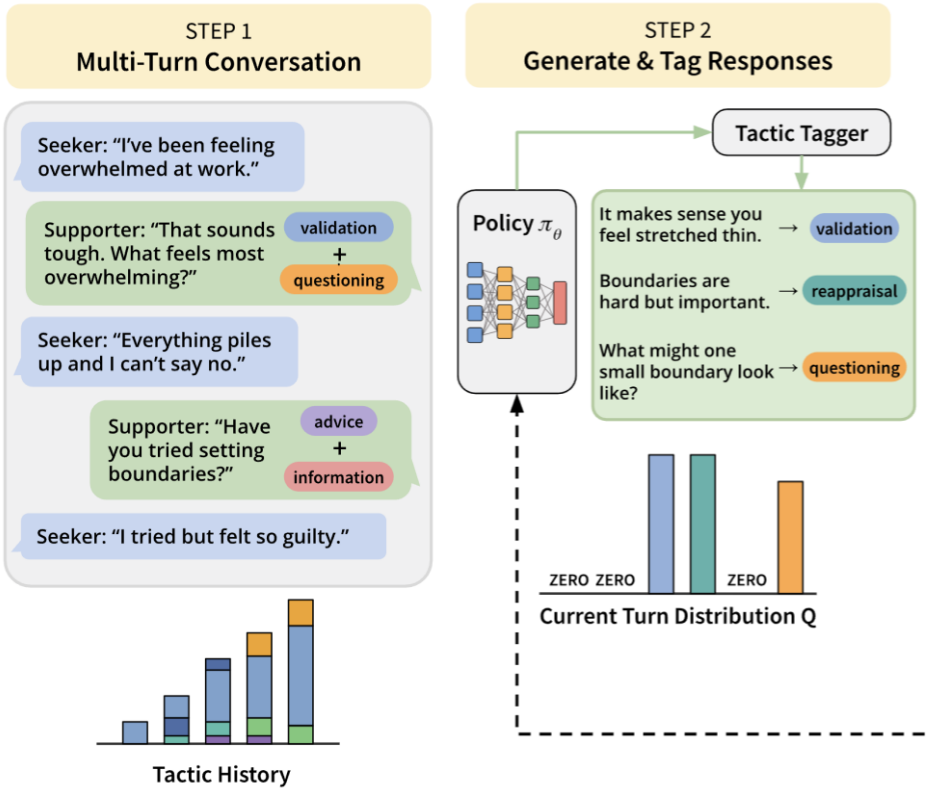
MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

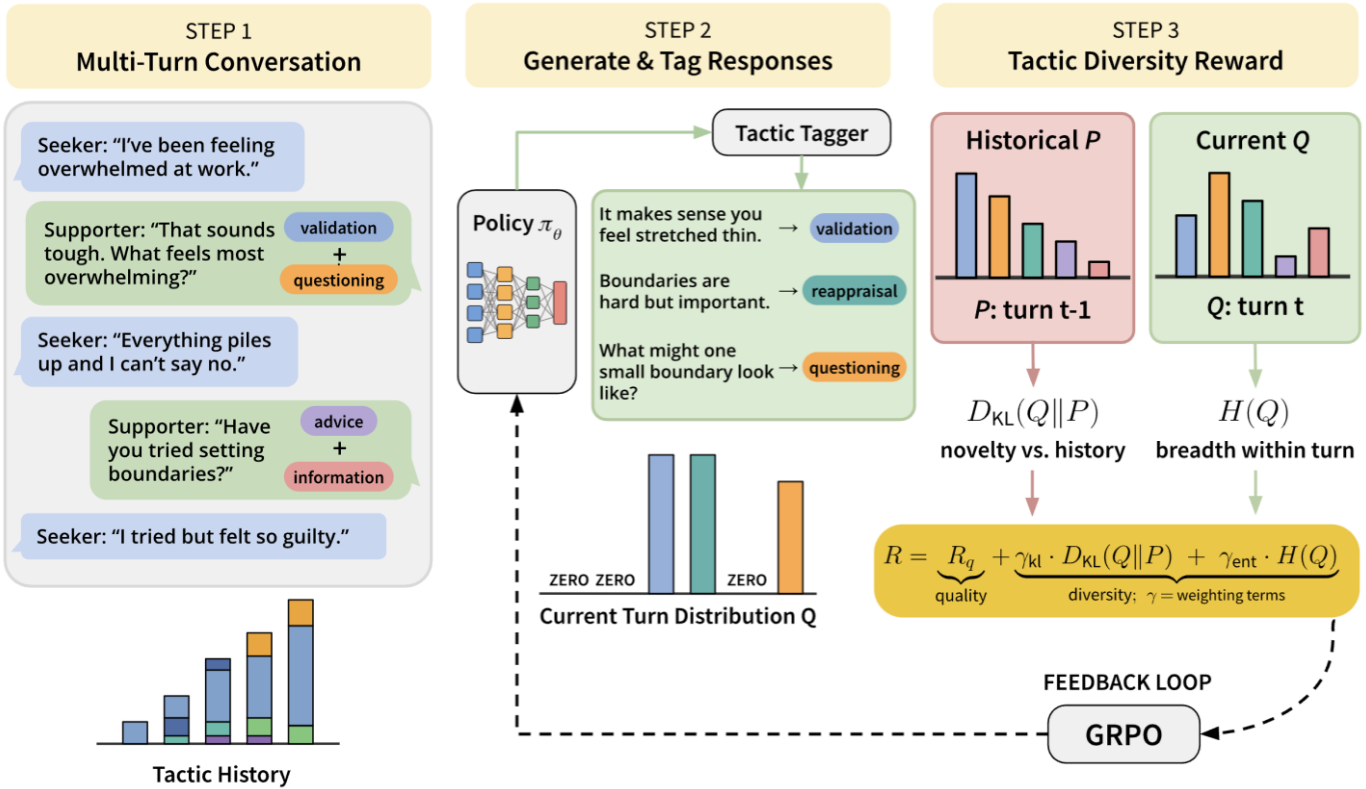
MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

TEXAS
The University of Texas at Austin

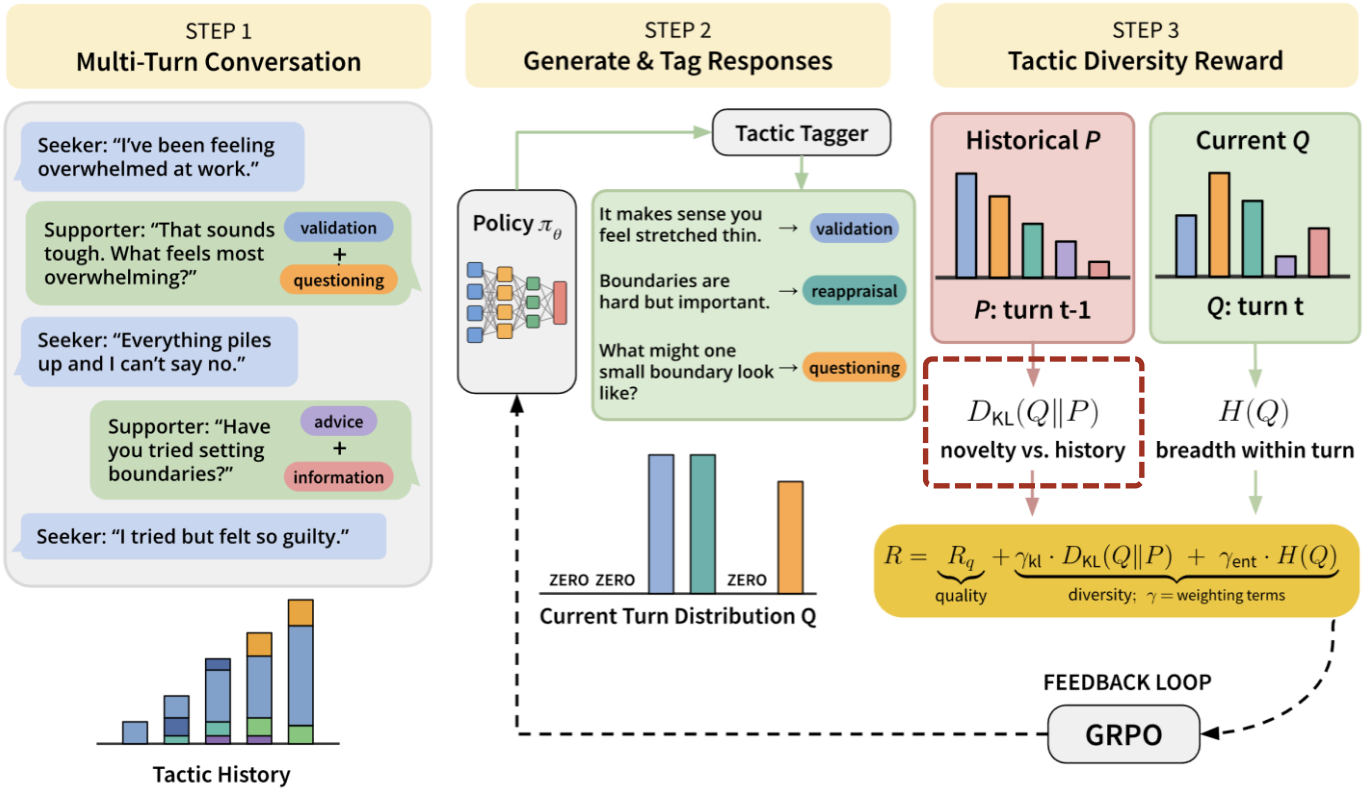
MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

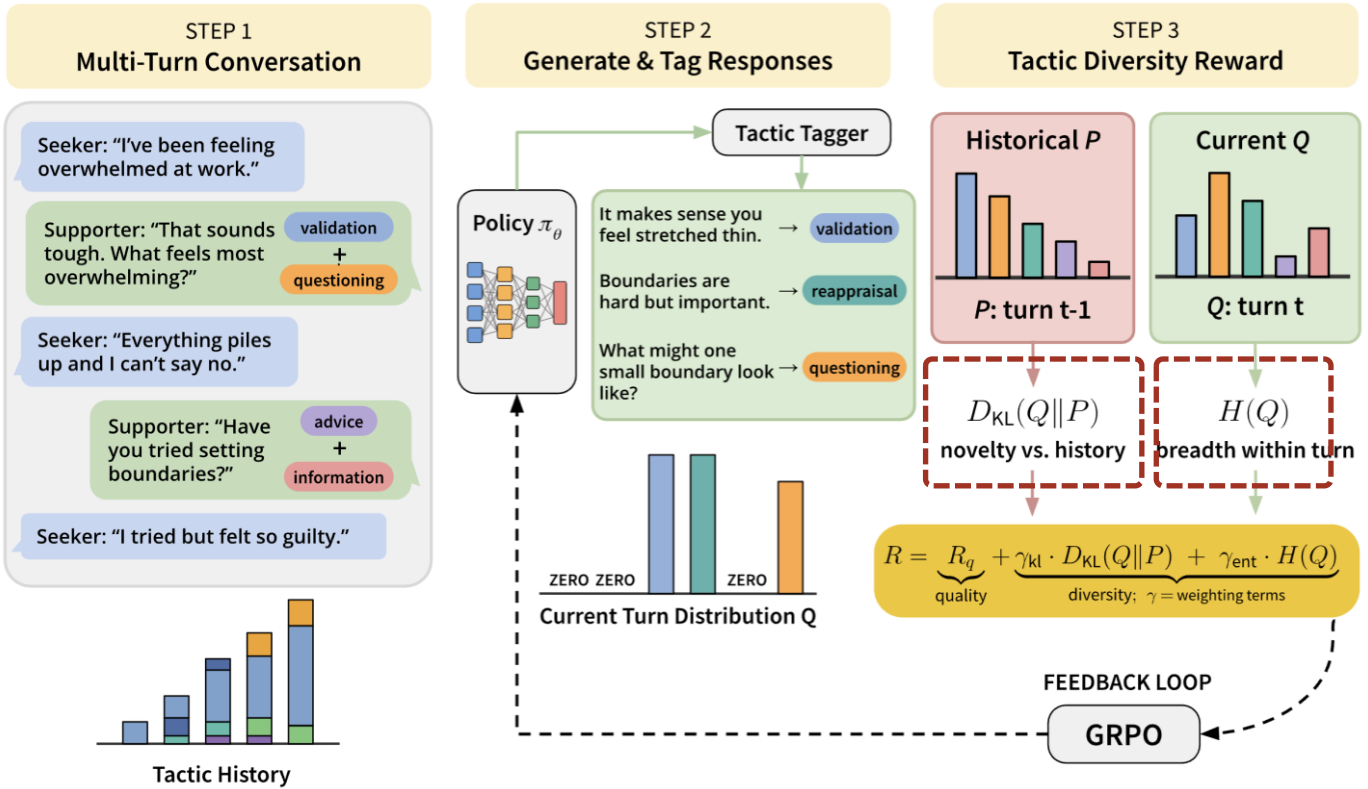
The University of Texas at Austin

MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

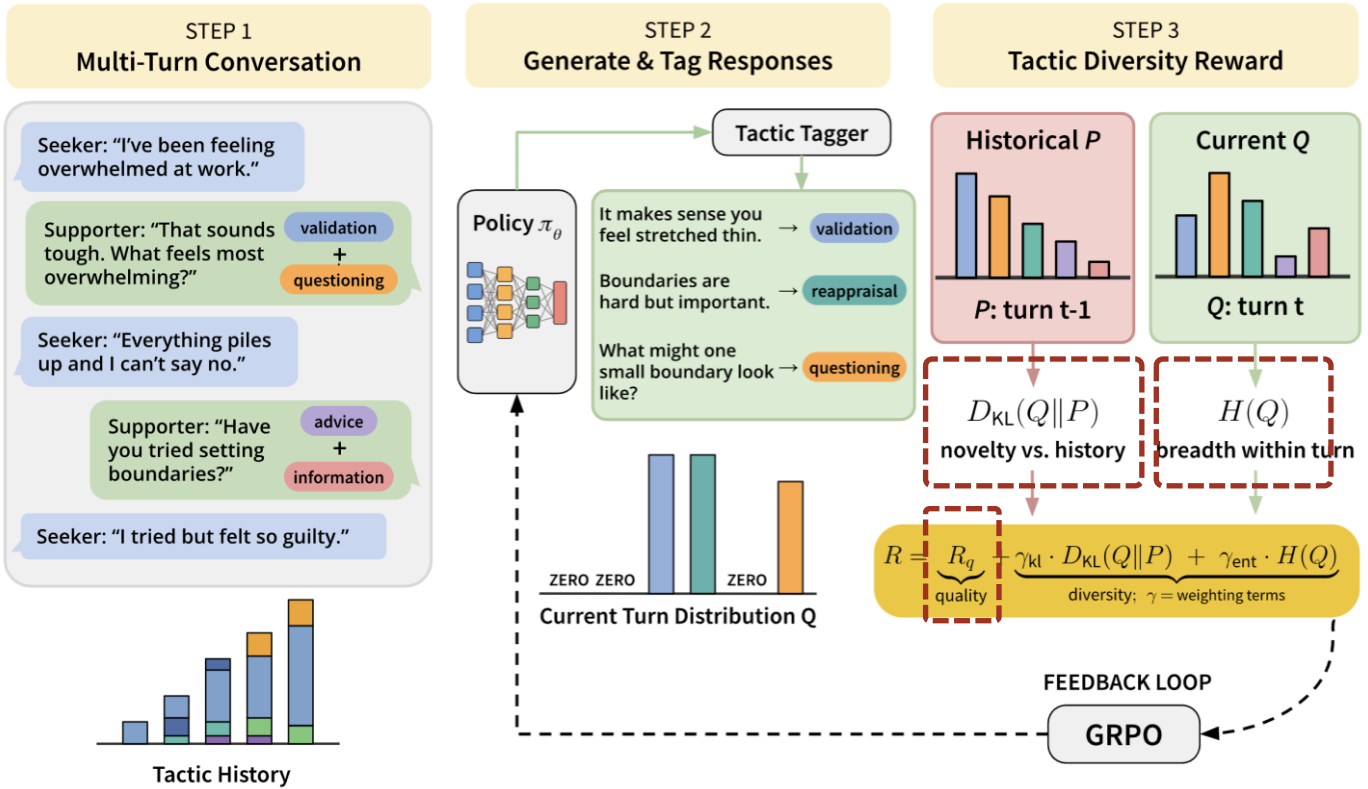
MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

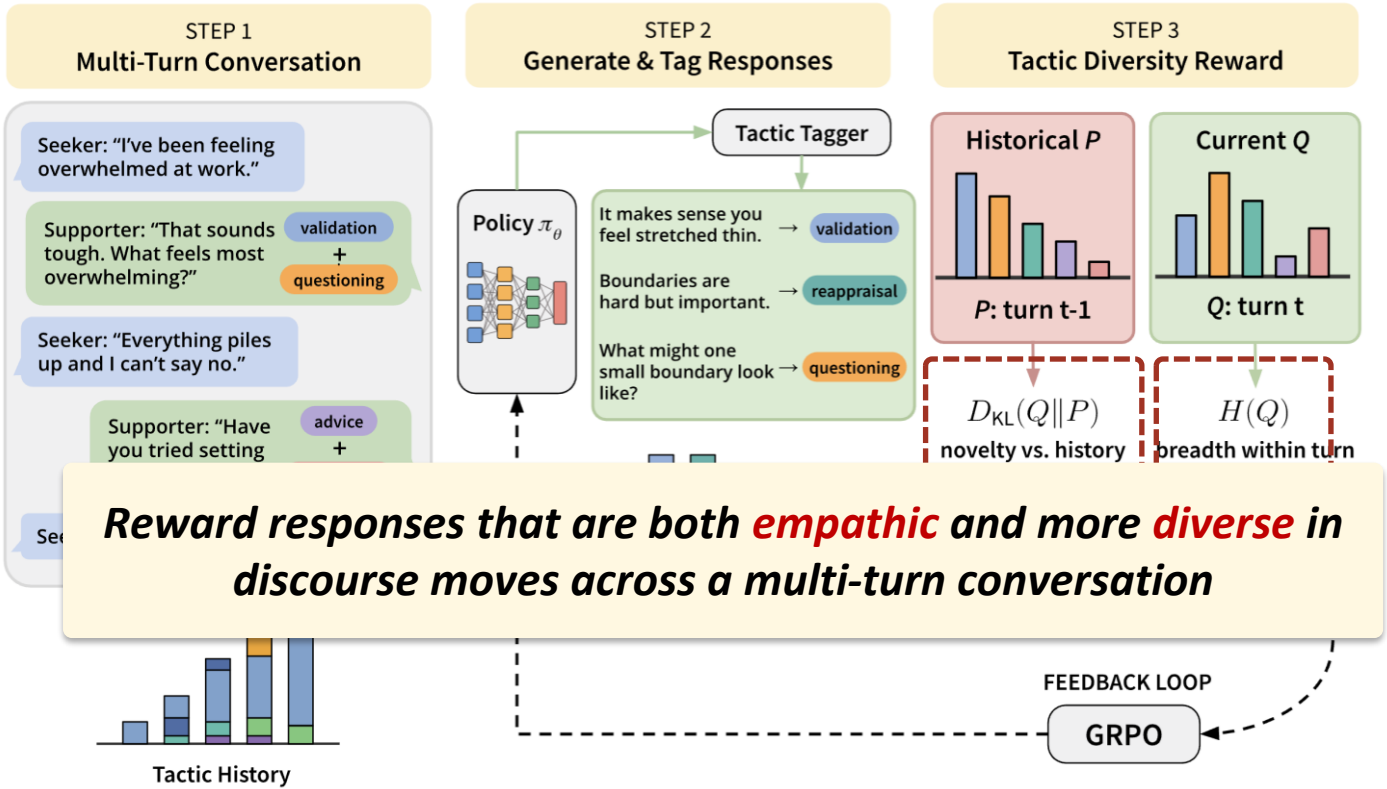
MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

MINT: Multi-turn Inter-tactic Novelty Training



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

MINT: Reward Details

Q Empathy quality *(PsychoCounsel reward model)*

D_{KL} Cross-turn tactic novelty

H Within-turn tactic breadth

All reward components min-max normalized to [0, 1] within each rollout group

$$D_{\mathrm{KL}}(Q_t \| P_{t-1}) = \min\left(\sum_{k=1}^{K} Q_t(k) \log \frac{Q_t(k)}{P_{t-1}(k)},\ \tau\right)$$

$$Q_t(k) = \frac{c_{t,k} + \alpha}{\sum_{j=1}^{K} c_{t,j} + K\alpha}, \qquad P_{t-1}(k) = \frac{c_{t-1,k} + \alpha}{\sum_{j=1}^{K} c_{t-1,j} + K\alpha}$$

$$H(Q_t) = -\sum_{k=1}^{K} Q_t(k) \log Q_t(k)$$

The University of Texas at Austin

MINT: Reward Details

Q Empathy quality *(PsychoCounsel reward model)*

D_{KL} Cross-turn tactic novelty

H Within-turn tactic breadth

All reward components min-max normalized to [0, 1] within each rollout group

$$D_{\mathrm{KL}}(Q_t \| P_{t-1}) = \min\left(\sum_{k=1}^{K} Q_t(k) \log \frac{Q_t(k)}{P_{t-1}(k)},\ \tau\right)$$

$$H(Q_t) = -\sum_{k=1}^{K} Q_t(k) \log Q_t(k)$$

Current & Prior Turn Tactic Distributions

$$Q_t(k) = \frac{c_{t,k} + \alpha}{\sum_{j=1}^{K} c_{t,j} + K\alpha}, \qquad P_{t-1}(k) = \frac{c_{t-1,k} + \alpha}{\sum_{j=1}^{K} c_{t-1,j} + K\alpha}$$

$c_{t,k}$	# sentences tagged with tactic k at turn t
$K = 10$	Number of empathy tactics
$\alpha = 0.1$	Laplace smoothing constant
$\tau = 5$	KL clipping threshold

The University of Texas at Austin

MINT: Reward Details

Q Empathy quality *(PsychoCounsel reward model)*

D_{KL} Cross-turn tactic novelty

H Within-turn tactic breadth

All reward components min-max normalized to [0, 1] within each rollout group

1) How different is this turn from previous turns?

Rewards the model for shifting its tactic mix as the conversation progresses.
Discourages repeating the same tactic patterns across turns (reduces stickiness).

$$D_{\mathrm{KL}}(Q_t \| P_{t-1}) = \min\left(\sum_{k=1}^{K} Q_t(k) \log \frac{Q_t(k)}{P_{t-1}(k)},\ \tau\right)$$

$$H(Q_t) = -\sum_{k=1}^{K} Q_t(k) \log Q_t(k)$$

Current & Prior Turn Tactic Distributions

$$Q_t(k) = \frac{c_{t,k} + \alpha}{\sum_{j=1}^{K} c_{t,j} + K\alpha}, \qquad P_{t-1}(k) = \frac{c_{t-1,k} + \alpha}{\sum_{j=1}^{K} c_{t-1,j} + K\alpha}$$

$c_{t,k}$	# sentences tagged with tactic k at turn t
$K = 10$	Number of empathy tactics
$\alpha = 0.1$	Laplace smoothing constant
$\tau = 5$	KL clipping threshold

The University of Texas at Austin

MINT: Reward Details

Q Empathy quality *(PsychoCounsel reward model)*

D_{KL} Cross-turn tactic novelty

H Within-turn tactic breadth

All reward components min-max normalized to [0, 1] within each rollout group

1) How different is this turn from previous turns?

Rewards the model for shifting its tactic mix as the conversation progresses.
Discourages repeating the same tactic patterns across turns (reduces stickiness).

$$D_{\mathrm{KL}}(Q_t \| P_{t-1}) = \min\left(\sum_{k=1}^{K} Q_t(k) \log \frac{Q_t(k)}{P_{t-1}(k)},\ \tau\right)$$

2) How varied are tactics within this turn?

KL alone doesn't reward using multiple tactics in one response. Entropy complements it.
Higher when tactics are evenly spread; near zero when one tactic dominates.

$$H(Q_t) = -\sum_{k=1}^{K} Q_t(k) \log Q_t(k)$$

Current & Prior Turn Tactic Distributions

$$Q_t(k) = \frac{c_{t,k} + \alpha}{\sum_{j=1}^{K} c_{t,j} + K\alpha}, \qquad P_{t-1}(k) = \frac{c_{t-1,k} + \alpha}{\sum_{j=1}^{K} c_{t-1,j} + K\alpha}$$

$c_{t,k}$	# sentences tagged with tactic k at turn t
$K = 10$	Number of empathy tactics
$\alpha = 0.1$	Laplace smoothing constant
$\tau = 5$	KL clipping threshold

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...

The model sounds great each turn, but says the same thing every turn.

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

TEXAS
The University of Texas at Austin

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

TEXAS
The University of Texas at Austin

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

LLM Judge Reliability: **gpt-oss-120b** as turn-level empathy judge (temp = 0)

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

LLM Judge Reliability: **gpt-oss-120b** as turn-level empathy judge (temp = 0)

Validated on 315 expert-annotated supporter turns from the Lend-an-Ear benchmark.

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

LLM Judge Reliability: **gpt-oss-120b** as turn-level empathy judge (temp = 0)

Validated on 315 expert-annotated supporter turns from the Lend-an-Ear benchmark.

Weighted Cohen's κ_w = **0.58** — exceeds the pairwise agreement among human experts

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

LLM Judge Reliability: **gpt-oss-120b** as turn-level empathy judge (temp = 0)

Validated on 315 expert-annotated supporter turns from the Lend-an-Ear benchmark.

Weighted Cohen's κ_w = **0.58** — exceeds the pairwise agreement among human experts

Evaluation: Why We Need Both *Empathy* AND *Diversity*

If we only measure empathy...
The model sounds great each turn, but says the same thing every turn.

If we only measure diversity...
The model is diverse but unhelpful.

We need *both* dimensions

Aggregate Empathy via Lend-an-Ear (Kumar et al., Nature 2026)

6 empathy dimensions (1-5 scale) per turn (3 desirable + 3 undesirable)

Tactic Stickiness = P(tactic at turn t | same tactic at turn t−1)

How likely is a tactic from the previous turn to appear again?
Lower = more diverse

LLM Judge Reliability: **gpt-oss-120b** as turn-level empathy judge (temp = 0)

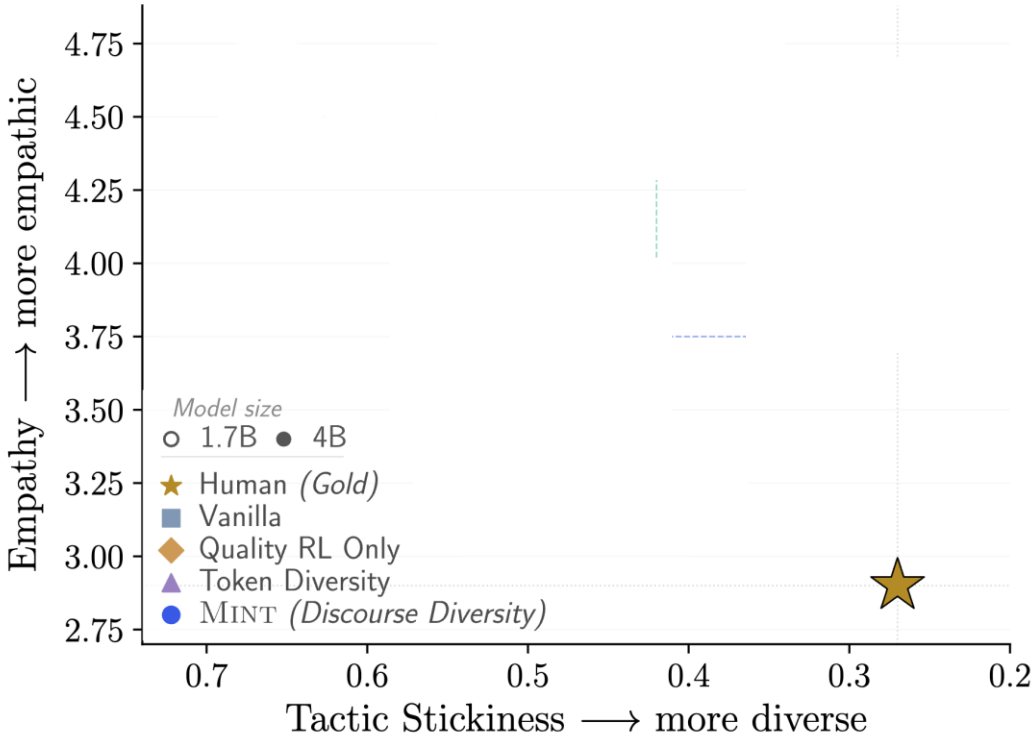
Validated on 315 expert-annotated supporter turns from the Lend-an-Ear benchmark.

Weighted Cohen's κ_w = **0.58** — exceeds the pairwise agreement among human experts

Optimizing for Both *Empathy* and *Diversity*

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

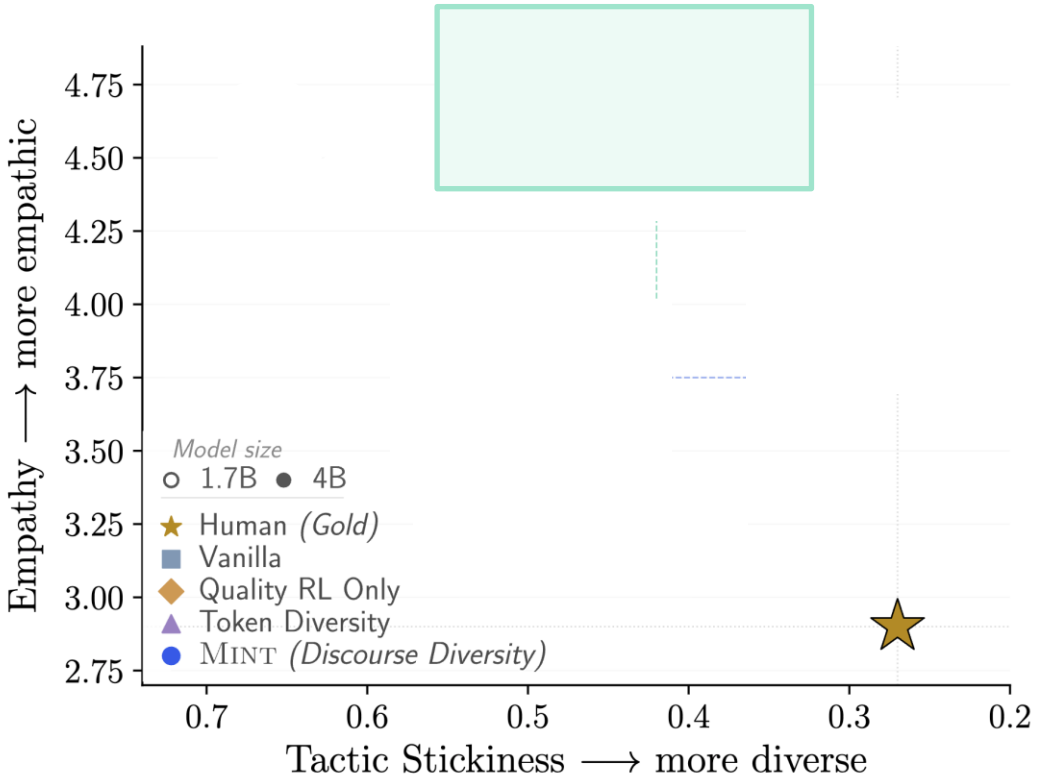
Optimizing for Both *Empathy* and *Diversity*



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

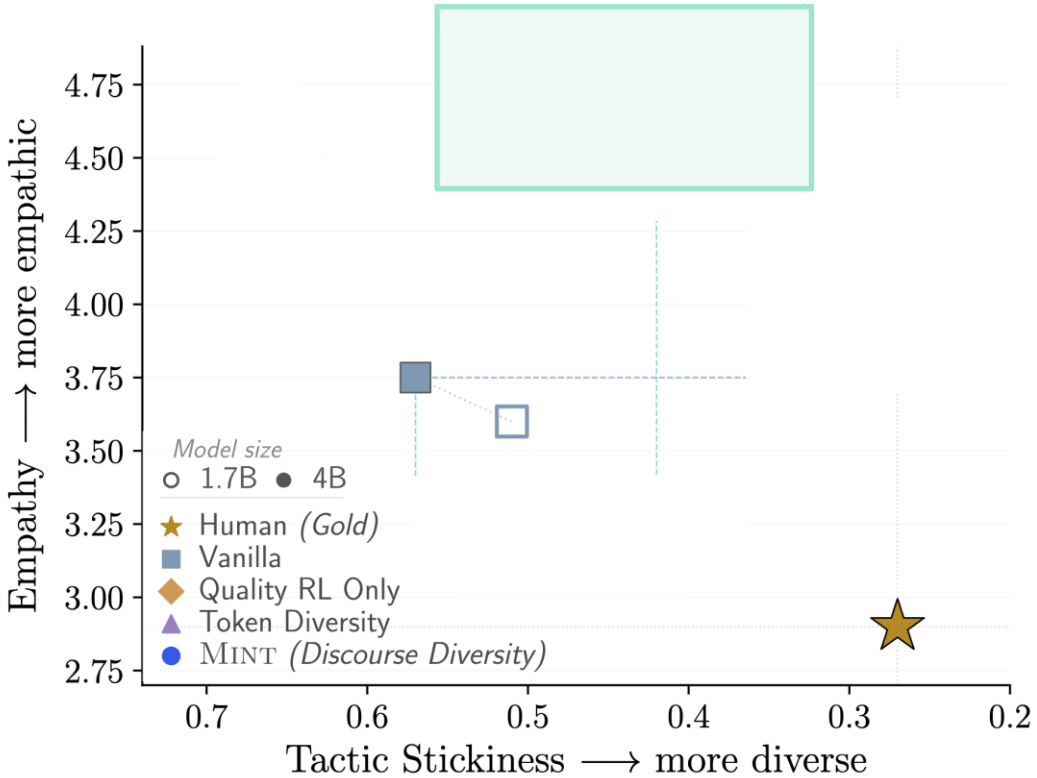
The University of Texas at Austin

Optimizing for Both *Empathy* and *Diversity*



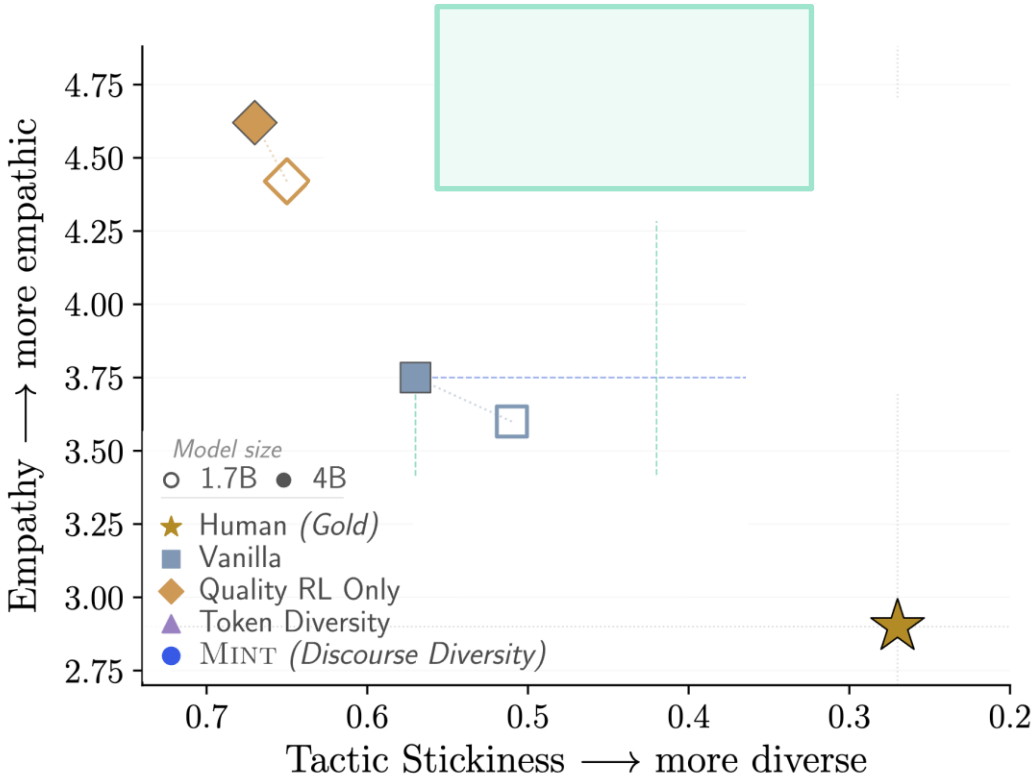
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Optimizing for Both *Empathy* and *Diversity*



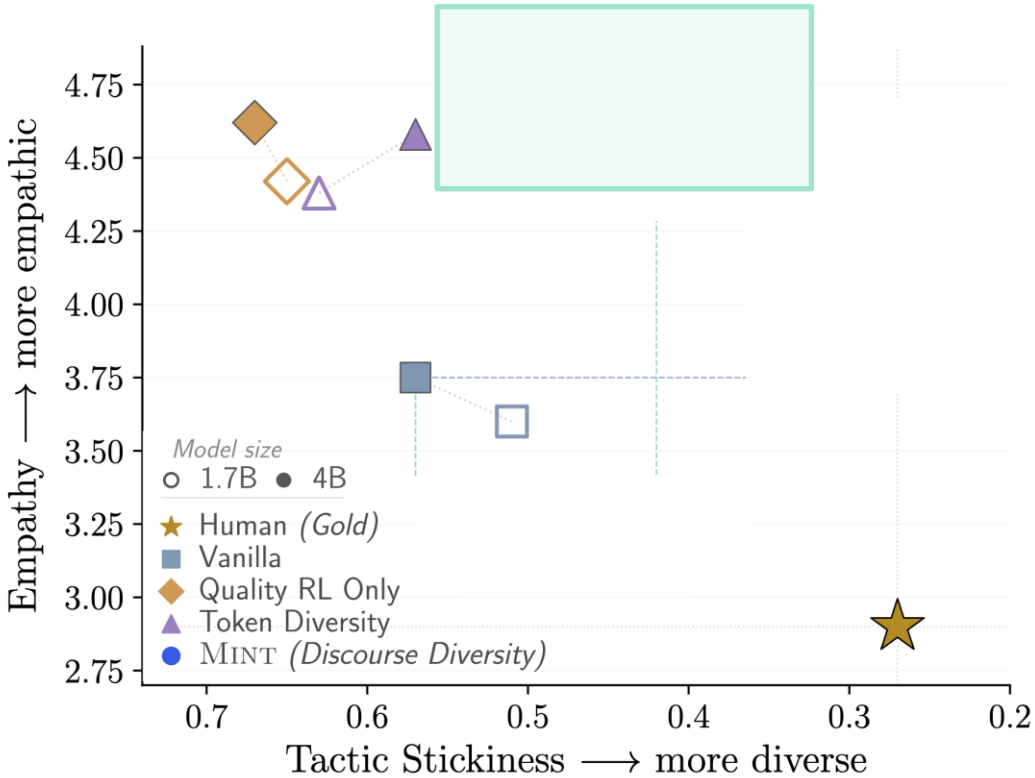
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Optimizing for Both *Empathy* and *Diversity*



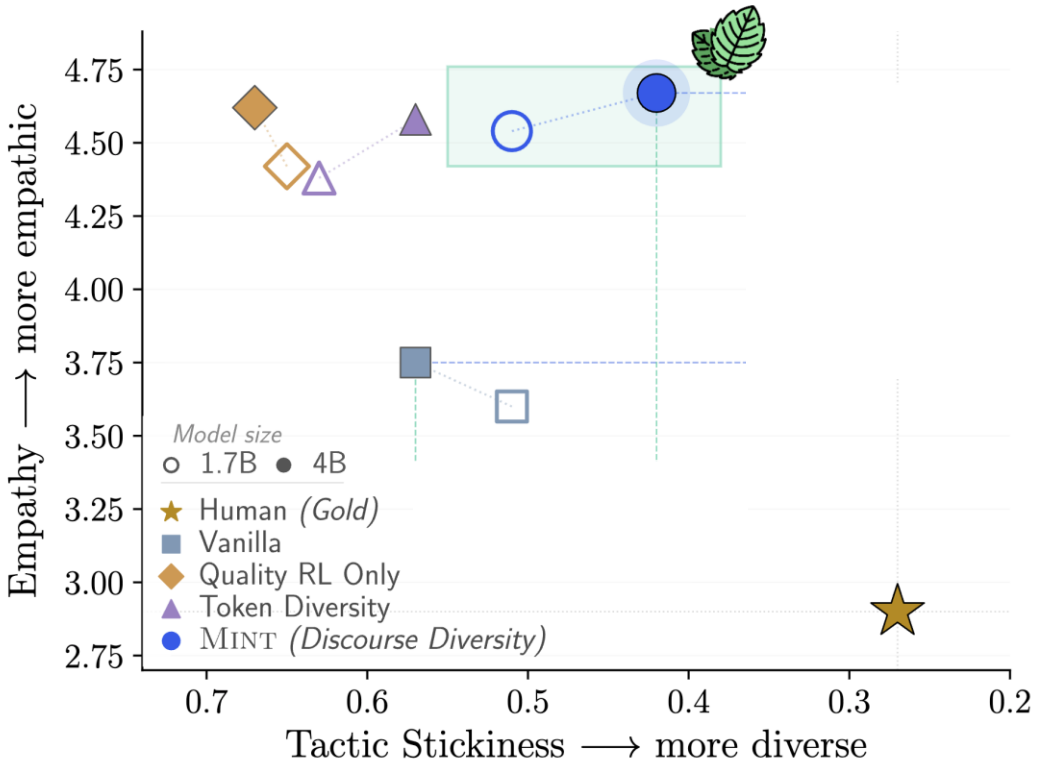
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Optimizing for Both *Empathy* and *Diversity*



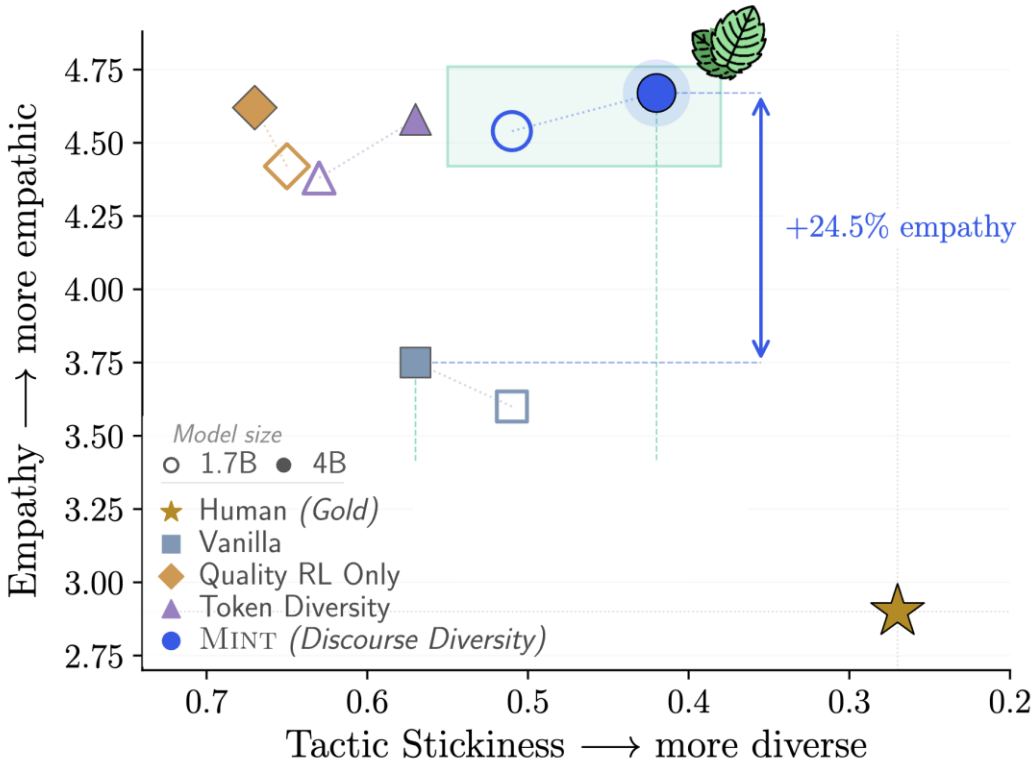
Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Optimizing for Both *Empathy* and *Diversity*



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

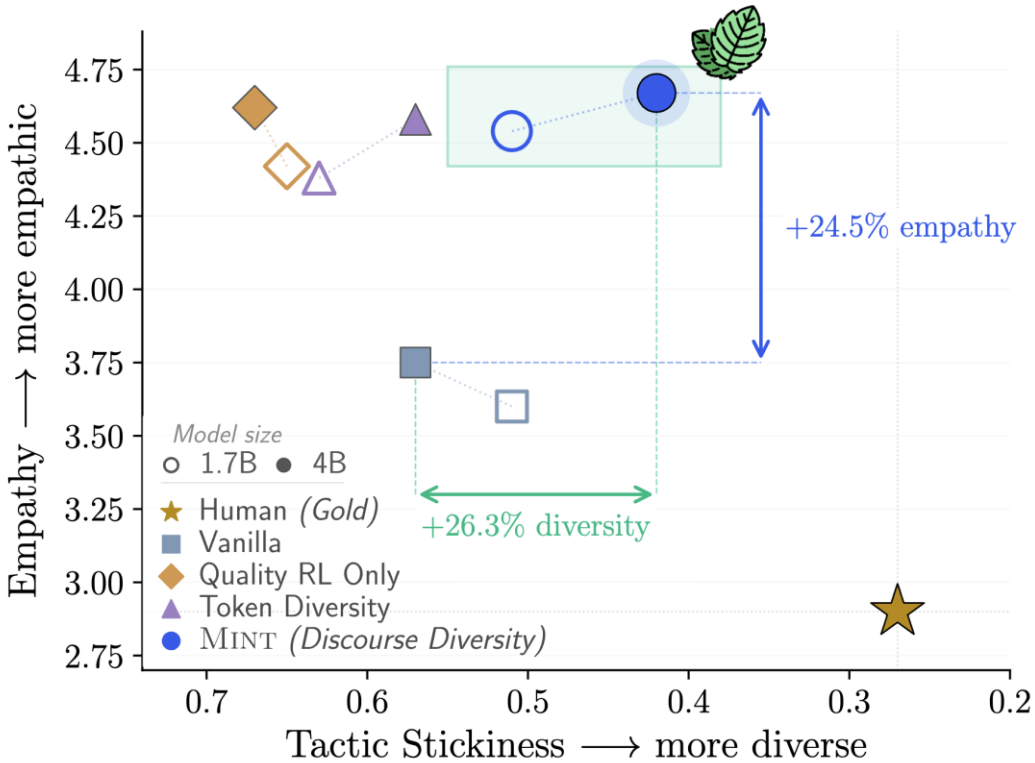
Optimizing for Both *Empathy* and *Diversity*



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

The University of Texas at Austin

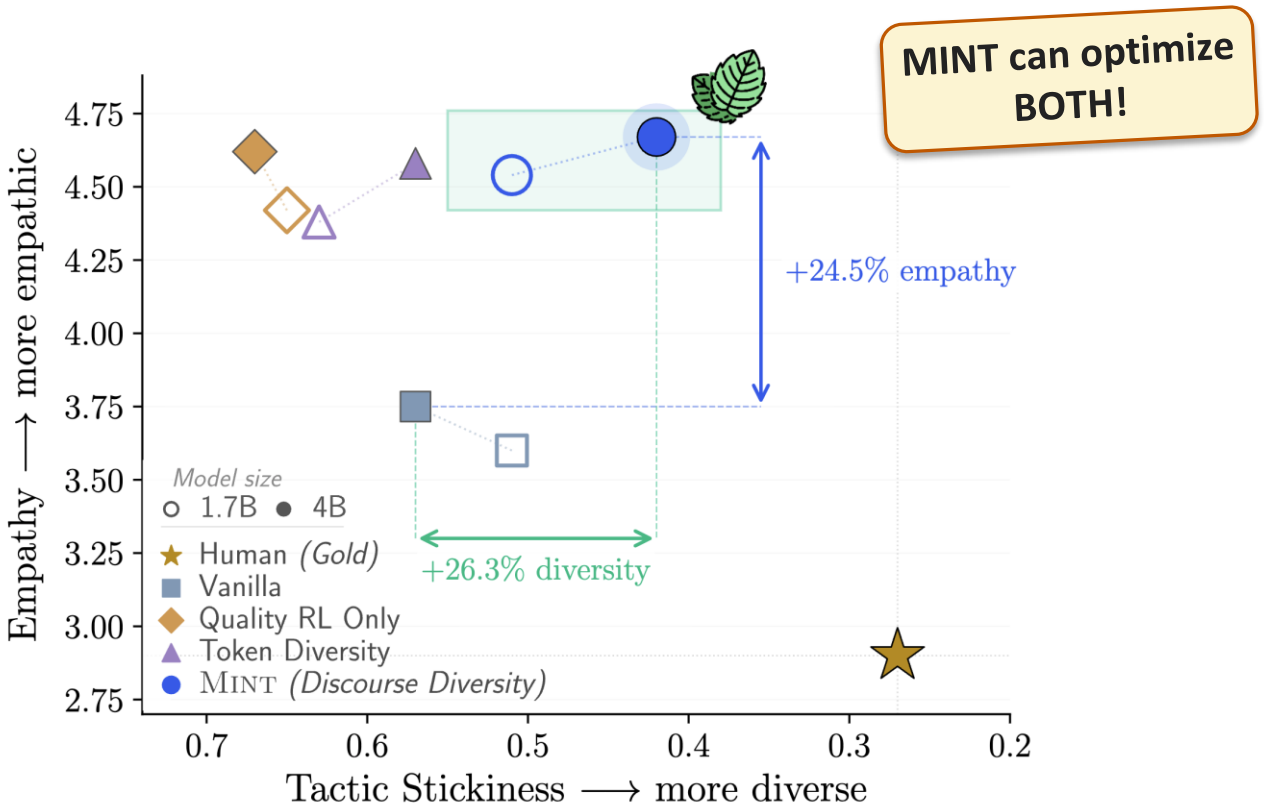
Optimizing for Both *Empathy* and *Diversity*



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

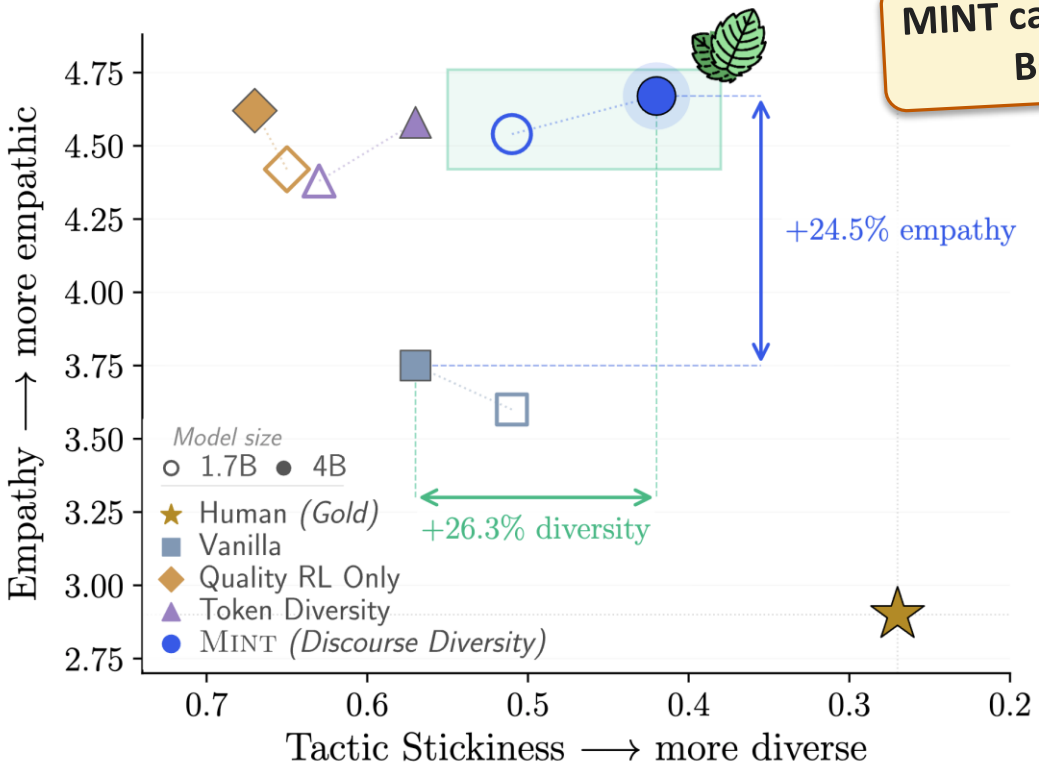
The University of Texas at Austin

Optimizing for Both *Empathy* and *Diversity*



Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Optimizing for Both *Empathy* and *Diversity*



Less unsolicited advice, more elaboration, tactic profiles closer to humans

Zhan, Gueorguieva, Hernandez, Suh, Ong, Li (*Under Review*)

Part 3 takeaway:

LLMs are repetitive in diversity moves in multi-turn conversations!

Part 3 takeaway:

LLMs are repetitive in diversity moves in multi-turn conversations!

Optimizing for both Quality and Diversity can help achieve a balance between both!

TEXAS
The University of Texas at Austin

Deciphering Emotions from Text
- EMNLP 2022
- EMNLP 2023 Findings

1

Discourse Diversity in Multi-Turn Empathic Dialogue
- *Under review*

3

Unveiling Advanced Psychological Capabilities from LLMs:
A Case of Targeted Reappraisal
- COLM 2024
- ICML 2025

2

4

Conclusion
- Summary of Contributions

Thesis Summary & Contributions

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople (EMNLP 2022, EMNLP 2023 Findings)

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople (EMNLP 2022, EMNLP 2023 Findings)

Intervene: Principle-guided reappraisals outperform human-written ones

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople (EMNLP 2022, EMNLP 2023 Findings)

Intervene: Principle-guided reappraisals outperform human-written ones (COLM 2024, ICML 2025)

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople (EMNLP 2022, EMNLP 2023 Findings)

Intervene: Principle-guided reappraisals outperform human-written ones (COLM 2024, ICML 2025)

Adapt: MINT closes the discourse-move-diversity gap while improving empathy

Thesis Summary & Contributions

Understand: capable of identifying cognitive appraisals, on par with laypeople (EMNLP 2022, EMNLP 2023 Findings)

Intervene: Principle-guided reappraisals outperform human-written ones (COLM 2024, ICML 2025)

Adapt: MINT closes the discourse-move-diversity gap while improving empathy (under submission)

The University of Texas at Austin

Selected Ph.D. Highlights

8
refereed conference papers

6
first-author / co-first-author
ICML x1 COLM x1
EMNLP x2 ACL x1

3
research scientist internships
IBM Research x2
MBZUAI IFM x1

1
first-authored U.S. patent
filed and pending

During the Ph.D., this work led to **peer-reviewed publications**, **industrial research internships**, and **technology transfer**.

TEXAS
The University of Texas at Austin

Many thanks to my mentors & collaborators!