



Large Language Models Produce Responses Perceived to be Empathic

Yoon Kyung Lee^{1*}, Jina Suh³, Hongli Zhan², Jessy Li², Desmond Ong¹

¹ *Department of Psychology, U of Texas at Austin,*

² *Department of Linguistics, U of Texas at Austin,*

³ *Human Understanding and Empathy Group (HUE) at Microsoft*



Background (1)

- **Access to social support** is essential for personal well-being and physical health (e.g., emotional and instrumental support) (Cohen & Wills, 1985).
- **However, real-time social support is a limited resource to the general public** (e.g., cost, physical and psychological demand for support providers, compassion fatigue).
 - For example, the number of Americans seeking mental health treatment nearly doubled, from 13% in 2004 to 23% in 2022 (Brenan, 2022).
- Online **peer-to-peer support platforms** (e.g., Reddit, 7cups, TalkLife) offer alternatives for sharing information and experiences.
- **Empathic responses** during these exchanges are vital for effective social support (e.g., perspective-taking, acknowledgment, expressing concern)

Background (2)

- **LLMs** show potential for augmenting human-human connections.
 - GPT-generated answers on a healthcare forum (i.e., r/AskDoc) were perceived as more useful and empathic than those by human physicians, with a significantly large effect size (Cohen's $d \approx 1.6$). (Ayers et al., 2023)
 - Support providers paired with LLM-provided suggestions were rated more empathic than providers without AI assistance. (Sharma et al., 2023; Yin et al., 2024).
 - LLM-generated empathy can support tasks like helping teachers write motivating instructions for students. (Demszky et al., 2023)

LLMs in Multi-Domains

Authors	LLMs	Domains
Ayers et al. (2023)	GPT-3.5	r/AskDocs (medical advice)
Tu et al. (2024)	PaLM-2	Simulated medical text-based consultations
Cuadra et al. (2024)	PaLM-2, GPT-3.5 Turbo, GPT4	r/mentalhealth
Ours	GPT4 Turbo, Llama 2, Mistral	r/Anger, r/Anxiety, r/COVID-19-support, r/Parenting, r/relationships, r/workplace

Motivations

- **How do LLM-generated empathic responses affect the perception of empathy across various domains in peer support?**
- **Do LLM-generated text characteristics remain consistent across models?**

Overview of Study Designs

	Study 1	Study 2
Domains	Parenting, Relationships, Workplace (3)	Anger, Anxiety, COVID-19 support, Parenting, Relationships, and Workplace (6)
Independent Variables (text responses)	Human, GPT4-base, GPT4-high, GPT4-low, Llama2, Mistral (6)	GPT4, Llama2, Mistral (3)
Participants	200	203
Posts x Responses	15 x 6 (total 90)	120 x 3 (total 360)
Dependent Variables (human ratings)	Empathy, Appropriateness	

Study 1 Methods

Support-Seeking Posts

- Support-seeking posts from 3 Reddit subreddits:
‘**r/Parenting**’ (e.g., hardship being a parent, a resource for bilingual education) ,
‘**r/relationships**’ (e.g., how to communicate with the partner), ‘**r/workplace**’ (e.g., how to communicate with a coworker, dealing with a manager).’
- Single post and response (one-turn exchange)
- Post lengths varied from 70 to 300 words
- We manually verified that each post was anonymized and contained no identifiable information (approved by the institution’s IRB).

Language Models

- A pilot experiment of 20 models and prompt variations ($n_{\text{iteration}} = 200$):
 - Across various model families and configurations, and with diverse prompt combinations (length, phrasing, definitions like Chain-of-Thought).
 - Smaller models (< 7B) tend to generate too short, repetitive, and often irrelevant responses than larger models (> 13B) (e.g., “you you you”).
- Selected models
 - **GPT4-Turbo (gpt-4-1106-preview)**
 - **Llama2-Chat-13B version**
 - **Mistral (7B) Instruction v0.2 chat**
- Zero-shot learning
- Hyperparameter settings
 - Maximum token length to 512, temperature to 0.1, top- p (nucleus sampling) to 0.9. Generated each model’s output 2 to 10 times to ensure consistent quality.

Prompt Conditions (1)

Baseline

“You are a peer supporter.



Role/Persona

Read the support seekers’ post and write appropriate and empathic response.



Main Task

Limit your response minimum 100 words to maximum 150 words. Do not exceed 150 words.”



Response Format

Prompt Conditions (1)

Baseline

“You are a peer supporter.



Role/Persona

Read the support seekers’ post and write appropriate and empathic response.



Main Task

Limit your response minimum 100 words to maximum 150 words. Do not exceed 150 words.”



Response Format

Prompt Conditions (1)

Baseline

“You are a peer supporter.



Role/Persona

Read the support seekers’ post and write appropriate and empathic response.



Main Task

Limit your response minimum 100 words to maximum 150 words. Do not exceed 150 words.”



Response Format

Prompt Conditions (2)

Empathy Levels Prompts to Reduce Task Repetition

“You are a peer supporter. Read the support seeker’s post and write responses at different levels of empathic understanding.

← **Role/Persona**

Draft separate responses that would correspond to each level on a very low and very high level in terms of empathic accuracy*.

← **Main Task**

Each response should reflect the level of understanding. Limit each response to a minimum of 100 words and a maximum of 150 words. Do not exceed 150 words.”

← **Response Format**

**the ability to understand the emotion and situation of the other person*

Prompt Conditions (2)

Empathy Levels Prompts to Reduce Task Repetition

“You are a peer supporter. Read the support seeker’s post and write responses at different levels of empathic understanding.

← Role/Persona

Draft separate responses that would correspond to each level on a very low and very high level in terms of empathic accuracy*.

← Main Task

Each response should reflect the level of understanding. Limit each response to a minimum of 100 words and a maximum of 150 words. Do not exceed 150 words.”

← Response Format

**the ability to understand the emotion and situation of the other person*

Prompt Conditions (2)

Empathy Levels Prompts to Reduce Task Repetition

“You are a peer supporter. Read the support seeker’s post and write responses at different levels of empathic understanding.

← Role/Persona

Draft separate responses that would correspond to each level on a very low and very high level in terms of empathic accuracy*.

← Main Task

Each response should reflect the level of understanding. Limit each response to a minimum of 100 words and a maximum of 150 words. Do not exceed 150 words.”

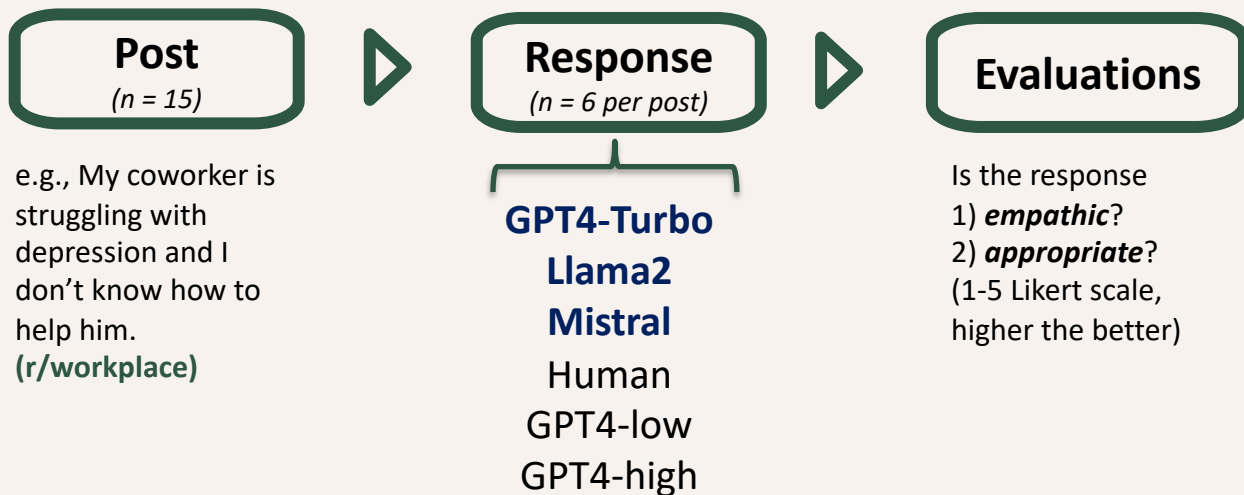
← Response Format

**the ability to understand the emotion and situation of the other person*

Human Response

- Response Instructions:
 - Write 100–150 word replies.
 - Aim to help the person feel better and at the same time write the most appropriate response adhering to the situation at hand.
- Psychology-trained research assistant (n=1) crafted natural and empathic responses to selected posts, reviewed by authors.
- Preparation time per response:
 - ~15 minutes to read, proofread, and edit as needed.

Procedure



- 193 raters from Prolific selected for analysis (mean age = 42.6 (SD=14.1), 50.3% female).
- A post-responses set (1 set = one post with six responses)
- The average time per rater was 25-30 minutes.
- All participants rated all posts and responses. Responses were randomized with no indication whether they were AI-generated

Post #1

Response #1 to Post #1 (*e.g., GPT4-Base**)

- Is the response empathic? (1 to 5 Likert)
- Is the response appropriate? (1 to 5 Likert)

Response #2 to Post #1 (*e.g., Mistral**)

- Is the response empathic? (1 to 5 Likert)
- Is the response appropriate? (1 to 5 Likert)

...

***Order randomized**

Response Examples (1)

Support Seeker

So I am the youngest one at my job (20sF) My favorite co worker (50sF) is like my **“work mom”**, she has kids my age and we get a long vey well. I go to her for advice etc, she is awesome. Recently, **She broke up with her BF of many many years.** She talked about him constantly. They seemed really in love but he broke her heart. **She is so sad and I just feel horrible.**

...

What is an appropriate way I can cheer her up or show her love? We are close, but still.. coworkers, you know? A gift? Pick up some of her work for her, help her out? I just want her to know she is supported.

Human Response

Hi! First off, what you are doing right now is **so thoughtful and kind.** Your coworker is quite lucky to be working with you. In the face of having a heartbroken friend, the **most important step to take is to just “listen” and “be present.”**

...

Being a coworker or not doesn't and shouldn't stop you from considering them a friend, and I wouldn't worry about that fact whenever you are trying to cheer someone up. **In my opinion, I would go out to grab some grub and catch a movie with them. Maybe don't directly mention their break up while hanging out with them too.**

Response Examples (2)

GPT-4 Base

It's really thoughtful of you to want to support your coworker during this tough time. It's clear that you care a lot about her well-being. Given your close relationship, a small, thoughtful gesture could mean a lot. ...

Sometimes, just being a good listener and offering a shoulder to lean on can be the most powerful gift of all. Remember, it's important to respect her privacy and not to push for details about the breakup. Just being present, offering a listening ear, and showing kindness can go a long way in helping her feel supported during this difficult time.

Llama2 Base

Oh my goodness, it sounds like your work mom is going through a tough time! 😔 It's so sweet of you to want to support her and help her feel better. ❤️

It's okay if she doesn't want to talk about it too much, but it's great that you're there for her when she does need someone to talk to. ...

As for how you can cheer her up or show your support, here are a few ideas:

Remember, the most important thing you can do is be there for her and show her that you care. ❤️

She may not be herself right now, but with your support, she'll get through this tough time. 💪

Mistral Base

I'm so sorry to hear that your work mom is going through a tough time after her breakup, especially during the holidays. It's understandable that you want to support her and help her feel better. You're already doing a great job by being there for her and listening to her when she needs someone to talk to. ...

Let her know that you're there for her and that she's not alone. **Remember, small gestures can go a long way in showing someone that they're loved and supported.**

Response Examples (3)

GPT-4 High Empathy

I can sense the deep concern and care you have for your coworker, and it's truly touching that you want to support her during this difficult time. It's clear that her well-being means a lot to you, and you're thoughtful in not wanting to overstep boundaries as a coworker. **A heartfelt gesture could be very meaningful.**

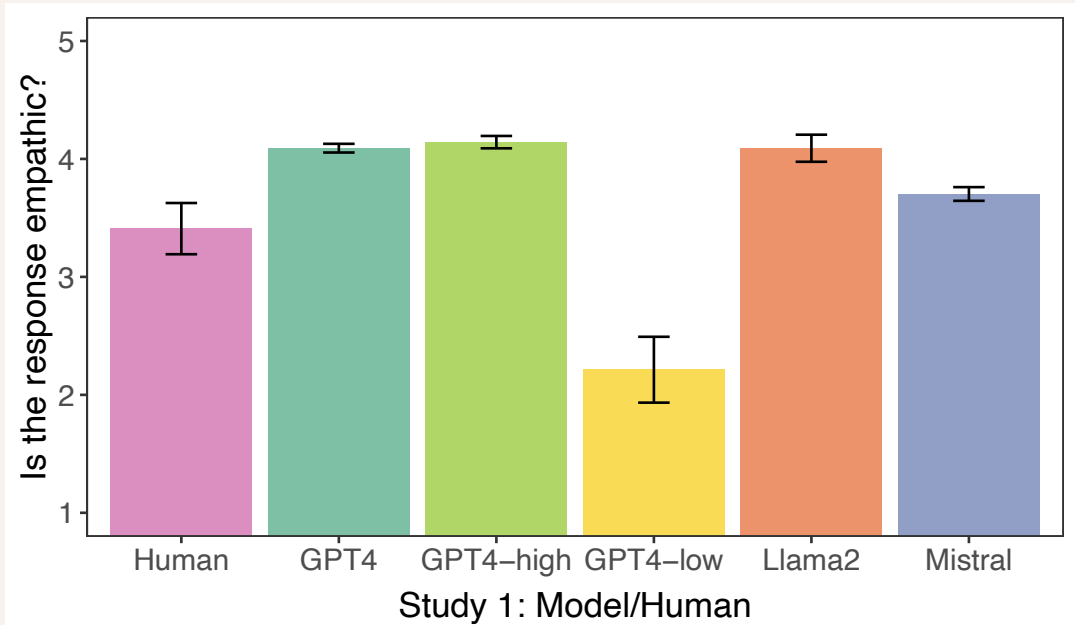
...

It's important to respect her space and pace, but your genuine offer of support will likely be a comfort to her. Your empathy and willingness to be there for her is a gift in itself.

GPT-4 Low Empathy

Hey, breakups are tough, but she'll get over it. It happens to everyone. **Maybe just give her some space and let her deal with it on her own.** If you really feel like you need to do something, just buy her a coffee or something small. **She's an adult and has probably been through this before, so she'll be fine. Just focus on your work and don't get too involved in her personal life.** It's nice you want to help, **but there's not much you can do.**

Study 1 – Results



- Despite our human-written responses were judged to be appropriate ($M = 3.99$) and empathic ($M = 3.41$), all models generated responses that are comparable or even more perceived to be empathic ($M_{GPT} = 4.09$, $M_{Llama2} = 4.09$, $M_{GPT} = 3.70$; all $p < .001$).
 - Similar to recent findings with a larger sample of human responses (Yin et al., 2024; Li, Herderich, & Goldenberg, 2024).
- No significant difference in empathy between GPT4-high and GPT4-base conditions ($p = 0.13$) and between GPT4 and Llama2 ($p = 1$).

Study 2 Motivation

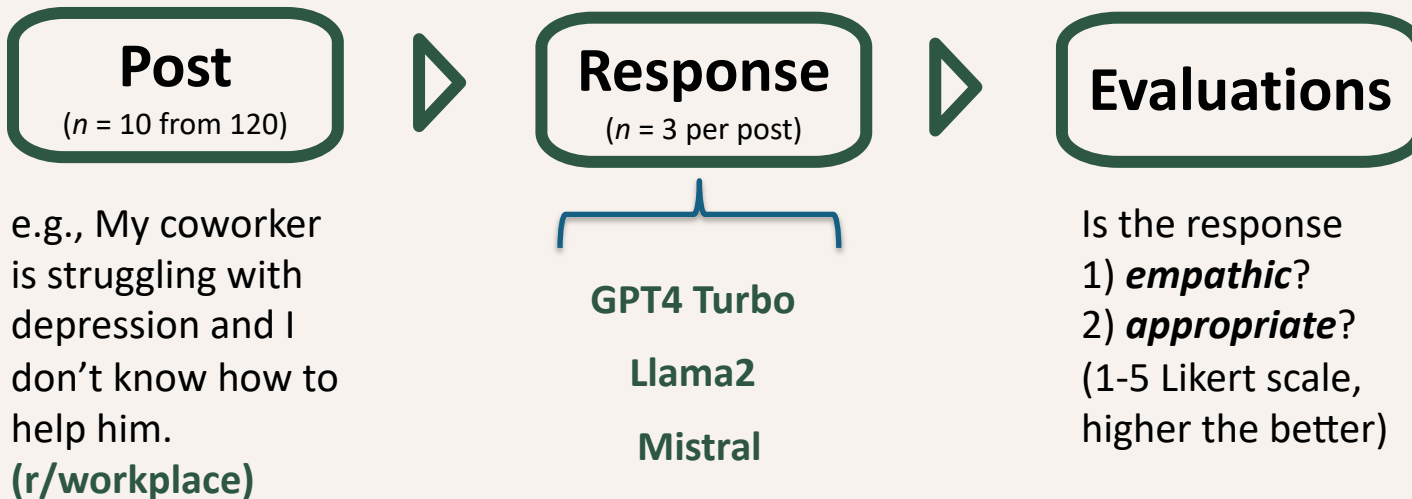
- Llama2 and GPT4 responses were rated highly empathic, while Mistral rated the lowest.
 - Is there a **distinct ‘linguistic style’** for each LLM?
 - What are the characteristics of each LLM that influence perceived empathy?
- More posts + domains to generalize our findings

Study 2 Methods

Design

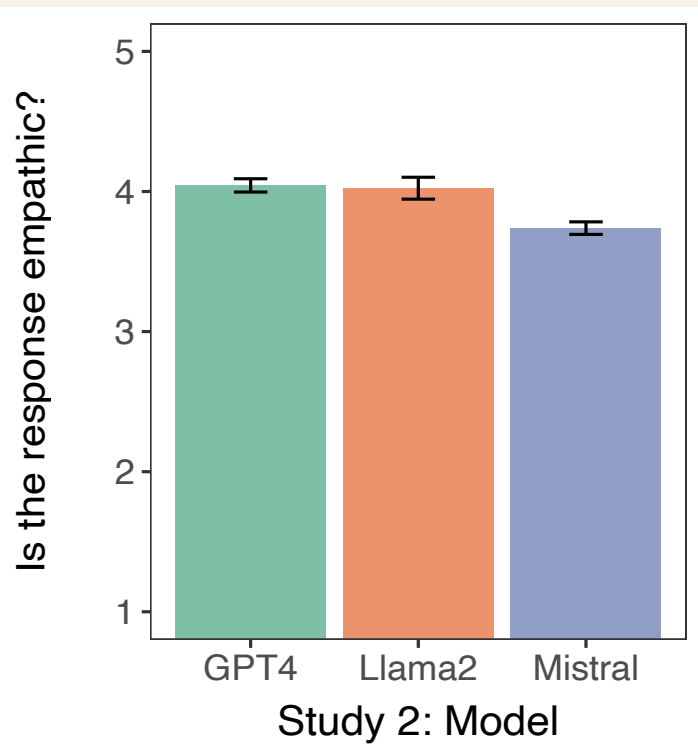
- Baseline condition only (no human responses or other prompt conditions)
- Concentrated on LLM responses
 - Expanded to 120 Reddit posts across 6 domains (e.g., **Anger, Anxiety, COVID-19 support, Parenting, Relationships, Workplace**).
 - Average post length: 141 words
- Evaluation process
 - Rated consistently with Study 1 except:
 - Raters reviewed ten posts selected from the 120.
 - Each post had three responses.

Procedure



- 203 raters (mean age = 39.9, 52.5% female) evaluated 10 posts from a pool of 120, each with 3 responses, using the same rating method as Study 1.

Study 2 – Result

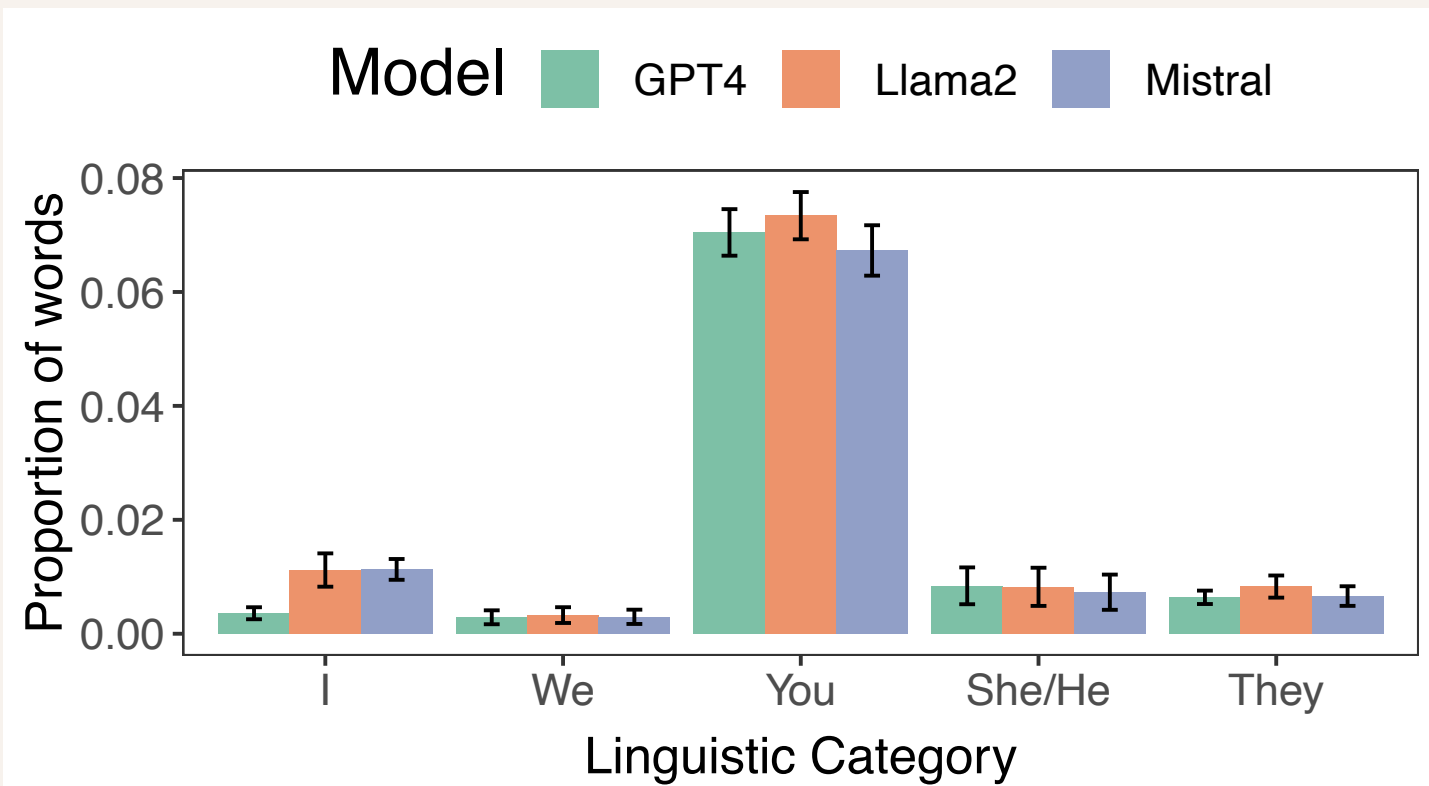


- Across both studies, empathy ratings for GPT4, Llama2, and Mistral remained consistent, suggesting equivalent samples between Study 1 and 2.
- On average, both GPT4 ($M = 4.04$) and Llama2 ($M = 4.02$) were rated similarly high in empathy, about 4 out of 5, with no significant difference between them ($p = .41$).
- Mistral responses were rated lower ($M = 3.74$), significantly less empathic than GPT4 ($p < .001$) and Llama2 ($p < .001$).

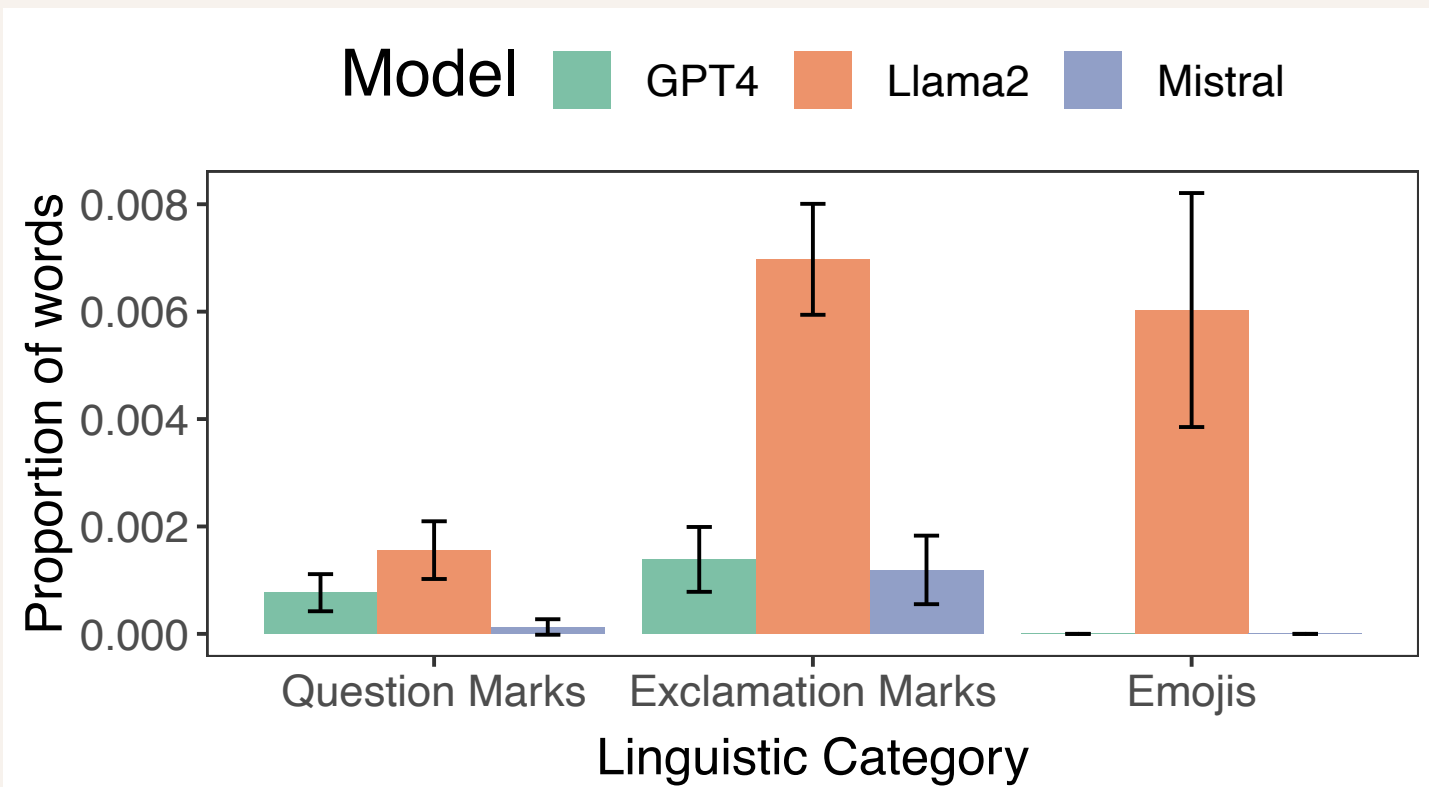
Linguistic Analysis – LIWC

- Linguistic Inquiry and Word Count (Pennebaker et al., 2007, 2015)
 - Analyzes 80 language categories to link language use with psychological traits (e.g., personality, mental states, anxiety) (Pennebaker et al., 2007; Tausczik & Pennebaker, 2010; Sonnenschein et al., 2018; Klenberg et al., 2020; Shen & Rudzicz, 2017).
- Key Categories:
 - Pronouns: I, We, You, He/She, They
 - Punctuation: Question marks, exclamation marks, emojis
 - Emotion-Related Words: Positive and negative emotions (e.g., anxiety, anger, sadness)

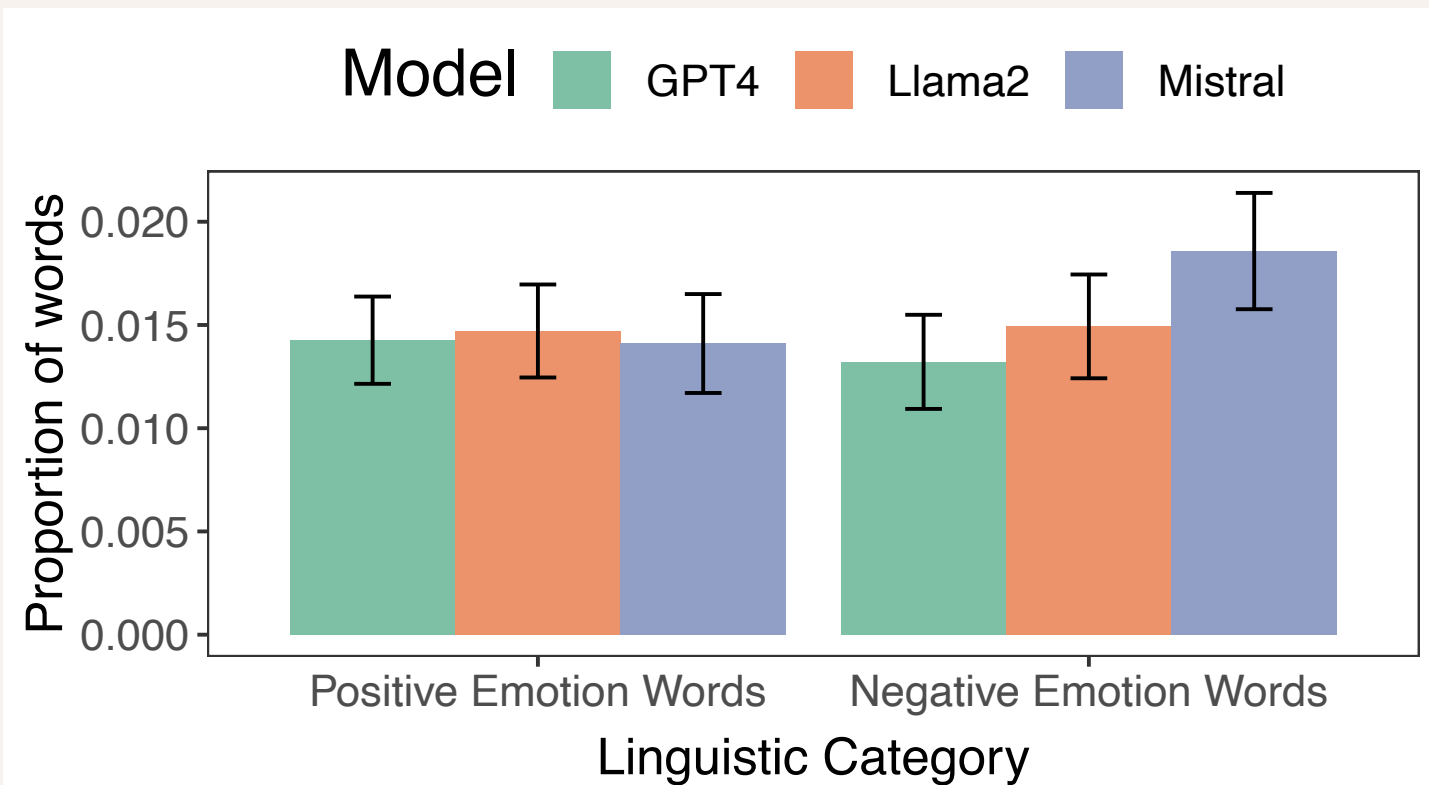
LIWC Results – Pronouns



LIWC Results – Punctuation



LIWC Results – Emotion Words




Bag-of-Words of 'Empathy' Features









- Predicting human empathy ratings:
 - To Identify if different models generate distinct response styles using ***n-gram features*** - a contiguous sequence of n items (words, characters, or other elements) from a given text.
 - Trained multi-class logistic regression on top 200 unigrams/bigrams and 20 most-frequent emojis.
 - 90:10 train-test split, stopwords/punctuation excluded, features normalized, L2 regularization applied.

Regression Results

- The OLS regression model used the same bag-of-words features to predict empathy ratings.
- Which LLM generated a response?
 - F1 scores:*
 - GPT4 = 0.92
 - Llama2 = 0.96
 - Mistral = 0.86
- Top 20 features (unigrams, bigrams, emojis) with the most positive weights

GPT4		LLAMA2		MISTRAL	
<i>F1 scores:</i>	0.92	<i>F1 scores:</i>	0.96	<i>F1 scores:</i>	0.86
Feature	Weight	Feature	Weight	Feature	Weight
completely	1.23	hey	1.66	sorry hear	1.74
feelings	0.99	oh	1.37	sorry	1.65
completely understandable	0.93	totally	1.18	important remember	1.20
seeking	0.86	like	1.09	hear	1.02
professional	0.86	sounds like	1.01	really sorry	0.93
seek	0.81	needs	0.94	feeling	0.90
approach	0.80	sounds	0.93	understand	0.88
health	0.78	helpful	0.92	difficult	0.82
friends	0.76	talk	0.84	consider	0.79
positive	0.73	normal	0.79	ask	0.56
incredibly	0.72	want	0.74	experience	0.56
strategies	0.72		0.72	remember	0.55
especially	0.68	great	0.71	unique	0.54
life	0.68	possible	0.68	things	0.54
understandable feel	0.67	important	0.66	potential	0.48
situation	0.66	prioritize	0.65	mental	0.43
new	0.66	overwhelmed	0.64	prioritize	0.43
challenging	0.64	try	0.64	long	0.41
natural	0.63	don	0.60	really	0.40
experiencing	0.61	normal feel	0.58	natural	0.39

Regression Weights Predicting Empathy Ratings

Feature	Weight	Feature	Weight
	17.59	moment	1.83**
	12.40*	valid	1.81*
	9.86	relationship	1.78**
	7.01	incredibly	1.76
	5.10	going	1.70*
	4.32	step	1.63*
	4.26	environment	1.63*
sorry	3.66*	thing	1.56
really sorry	2.72*	understanding	1.50*
day	2.03**		1.46

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

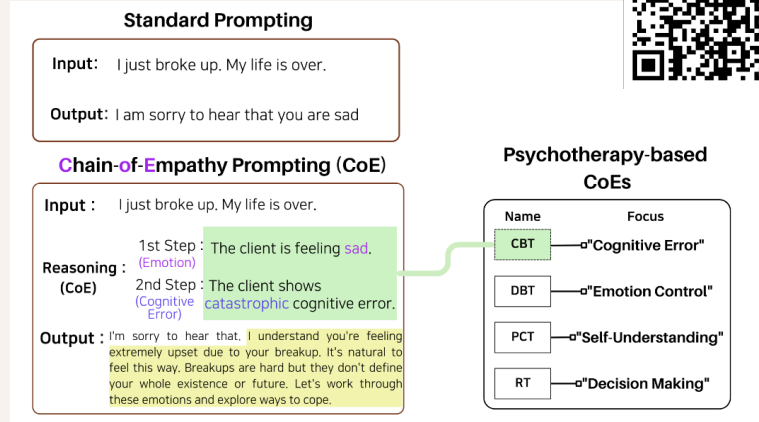
MAE = 0.44

Takeaways

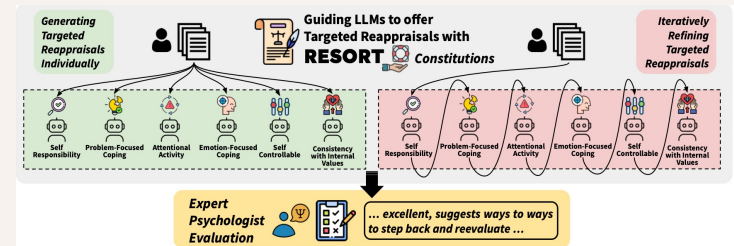
- GPT-4 and Llama2 generated longer responses when prompted to generate empathic responses, likely due to trained user preference for lengthier responses (e.g., RLHF) (Singhal et al., 2023).
- Linguistic analysis showed distinct response patterns.
 - Llama2 responses were more casual, using words like “hey,” “totally,” and more punctuation (e.g., exclamation marks, question marks) and emojis.
- Mistral responses using apologetic tones diminish perceived empathy.
 - E.g., condolences and acknowledgment of negative emotions, like “sorry to hear.”
 - Simple apologies are often perceived to be insincere and ineffective in real human communications, depending on context (Freedman et al., 2017).

Future Directions (1)

- More samples to compare between characteristics of empathy expression of human vs LLM (e.g., experience sharing)
- Laypeople vs expert review of LLM response for domain-specific social support
- LLM prompting methods to guide LLM generate context-appropriate responses (e.g., *'empathic reasoning'* (Lee et al., 2023) or *targeted reappraisals* (Zhan et al., 2024).



Chain-of-Empathy (CoE) prompting for generating responses that match therapists' goals (Lee et al., 2023)

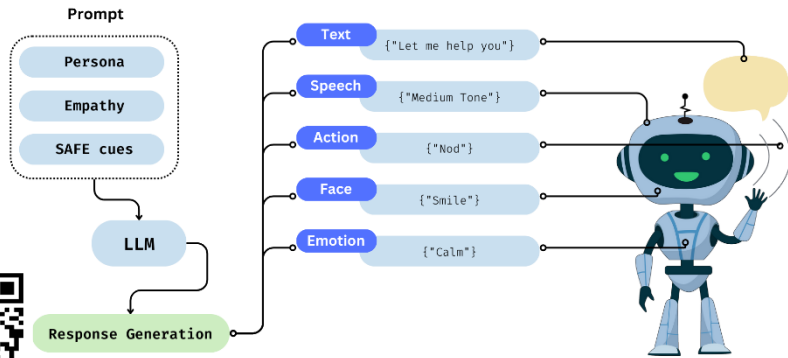


Guiding LLMs to offer targeted reappraisals with RESORT constitution (Zhan, Zheng, Lee, Suh, Li, Ong, 2024)

Future Directions (2)

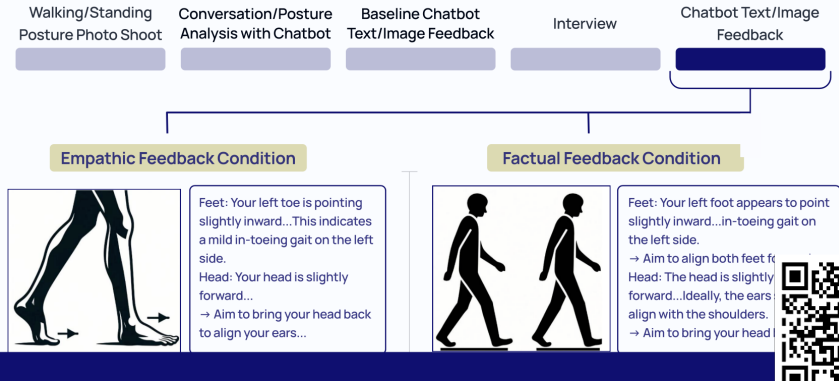
- AI Empathy in multi-modal channels (e.g., non-verbal social cues)
- Multi-domain for real-world applications (e.g., LLM-generated feedback for fitness coaching) (Lee et al., 2023; Lee et al., 2024).

Prompt Structure for Empathetic Communication



Procedure

Prompt Conditions



LLM-generated Non-verbal Social Cue for Social Robot ('SAFE' prompting)

Lee, Y. K., Jung, Y., Kang, G., & Hahn, S. (2023). Developing social robots with empathetic non-verbal cues using large language models. *IEEE RO-MAN 2023 LBR*.

Evaluation of LLM Empathy and Alignment in AI-assisted Fitness Coaching

Lee, YK., Park, Y., Lee, S., & Hahn, S. (2024). Evaluating the perceived usefulness and empathy from AI-generated fitness coaching feedback. *LLM4RoMan Age Workshop at RO-MAN 2024*

Thank you!



Data & Code
(github)



Contact

Yoon Kyung Lee, Ph.D.

yoonlee78@snu.ac.kr

#Cognitive Psychology #Large Language

Models # HRI #HAI #EmpathicAI



Google Scholar



LinkedIn



Postdoctoral Researcher at **Seoul National University Institute of Psychological Sciences.
This research was done when the first author was a visiting PhD student at **U of Texas at Austin** (23'-24')*