Attributions of Mind to Humans and AI

Rachel Metzgar
PhD Candidate | Princeton University

For most of history, humans only ever had coherent conversations with other humans. That's changing rapidly.

800M+
weekly active ChatGPT users (2025), doubling in under a year

88%
of consumers have used a chatbot (2023)

71%
of companies use generative AI in at least one function (McKinsey, 2024)

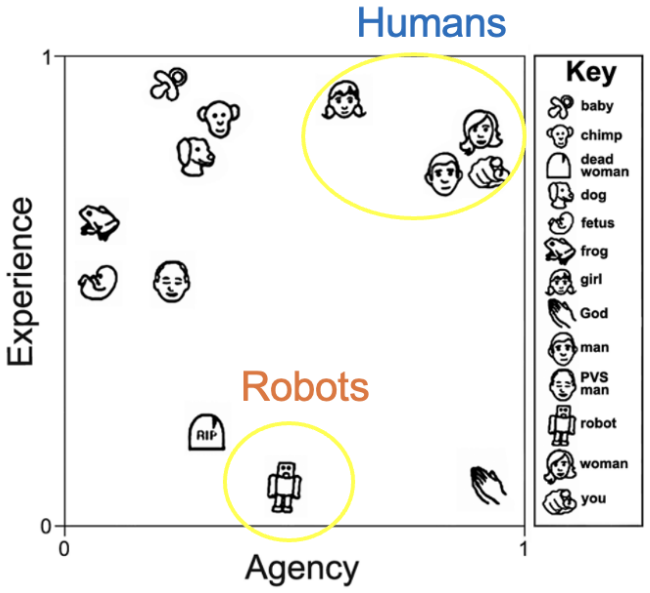
AI chatbot market: $5.1B (2022) → $15.6B (2024) → projected $46.6B (2029)

Mind perception is organized along two dimensions

Experience — capacity to feel (pain, pleasure, emotions)

Agency — capacity to act and think (self-control, planning, morality)

Historically, humans and robots sit on opposite ends of those dimensions.



(figure from Gray et al., 2007)

Gray, Gray & Wegner (2007), Science

Humans often treat AI differently from other humans

Trust & Risk

- Humans overwrite economic self-interest to avoid bargaining with AI
- Risk tolerance and loss aversion shift when opponent is labeled AI

Erlei et al., 2022

Cooperation

- Cooperation strategies shift when partner is labeled algorithm vs. person
- Transparency–efficiency tradeoff in human–machine cooperation

Ishowo-Oloko et al., 2019

Devaluing AI Content

- AI-generated responses make people feel more heard, but AI labeling reduces this effect
- People rate AI explanations higher, yet devalue them once labeled as AI

Yin et al., 2024; Hohenstein et al., 2023

Empathy & Social Responses

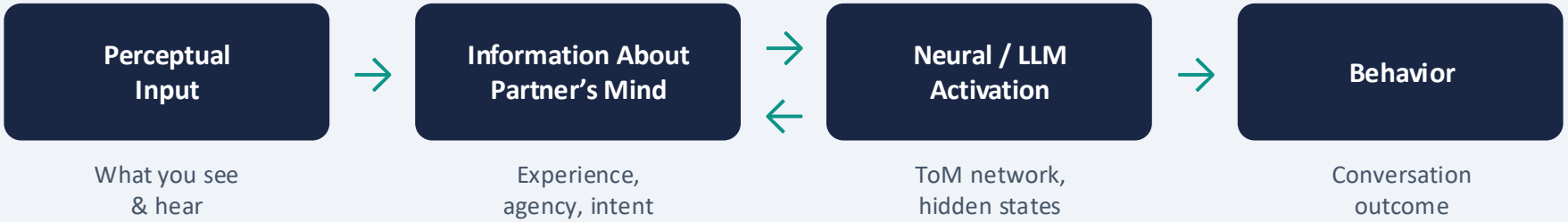
- Empathic neural responses attenuated for artificial agents, especially for abusive treatment

Rosenthal-von der Pütten et al., 2014; Ho et al., 2018

The label alone changes behavior.

Information about another's mind may drive behavioral differences

Whether we attribute experience, agency, and intentionality to our partner shapes how we interact with them

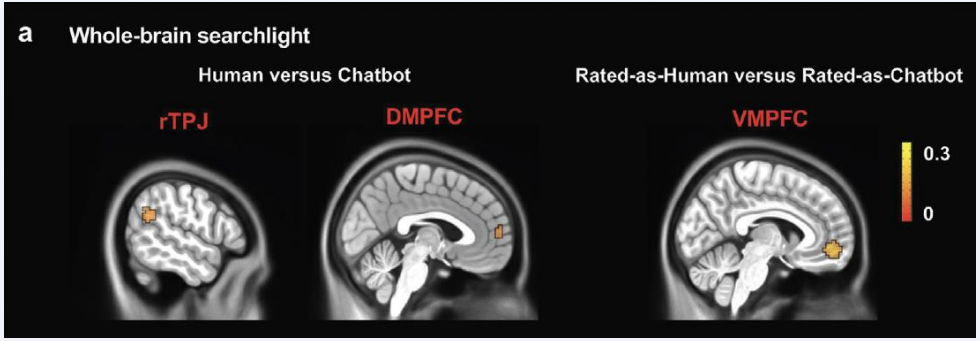


Representation of mind: internal model with information about whether another's mental capacities— experience, agency, intentionality

Historically, studies show that social brain networks distinguish humans and AI

- Theory of Mind regions (TPJ, mPFC, STS) show differential activation when people believe they're interacting with humans versus machines (Krach et al., 2008; Chaminade et al., 2012, Rauchbauer et al., 2019; Torubarova et al., 2025)

- Wei et al. (2023): mentalizing network distinguishes chatbot vs. human text even when participants cannot



But there are also a lot of similarities in how we interact

CASA Framework / Treating AI Like Humans

Computers Are Social Actors: minimal social cues (polite language, turn-taking, consistent “personality”) trigger deeply practiced social scripts (Nass et al., 1994)

- People apply politeness norms, reciprocate help, and exhibit gender stereotyping based on a computer’s voice (Nass & Moon, 2000)
- People disclose to chatbots as readily as humans, with equivalent emotional benefits (Ho et al., 2018; von der Pütten et al., 2010)
- People hesitate to shut off or strike robots (Bartneck et al., 2007)

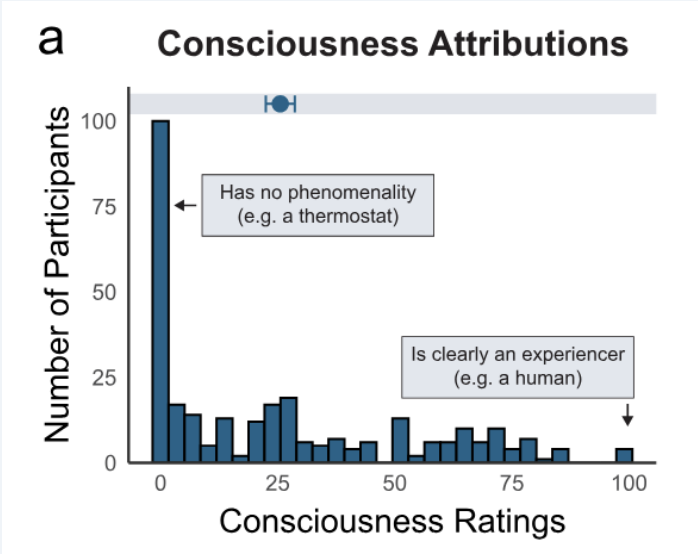
Implicit social responses may be triggered automatically, regardless of explicit beliefs about an agent’s mental capacities.

And Public Attitudes towards LLMs are malleable and probably shifting

- 2/3 of US adults willing to attribute some phenomenal consciousness to ChatGPT; attributions increase with usage frequency (Colombatto & Fleming, 2024)

- Self-reflection and emotional expression by AI increase perceived consciousness (Kang et al., 2026)

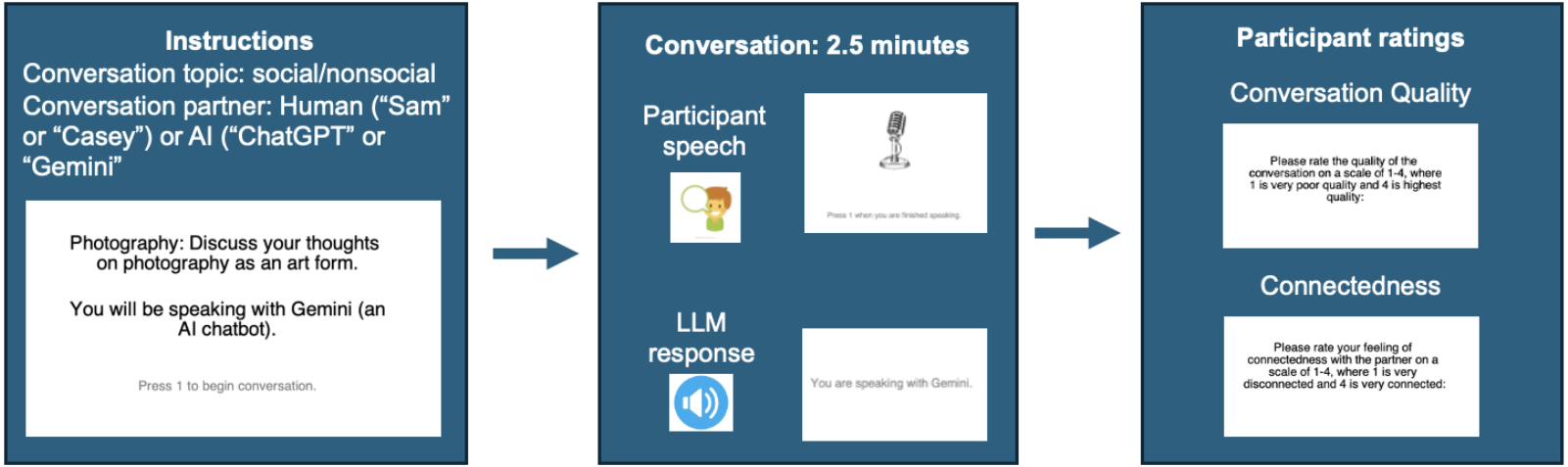
How do our explicit beliefs about mind shape our interactions with AI in a rapidly evolving landscape?
→ Use belief, behavior, and internal mechanisms



(Colombatto & Fleming, 2024)

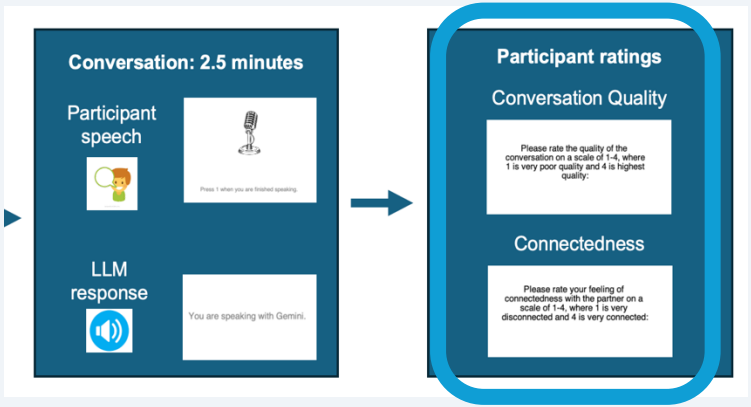
Metzgar et al., in prep – experiment design

Study 1: Participants (N=23) had 40 spoken conversations in the scanner with partners labeled as human or AI (all actually AI with the same system prompt)

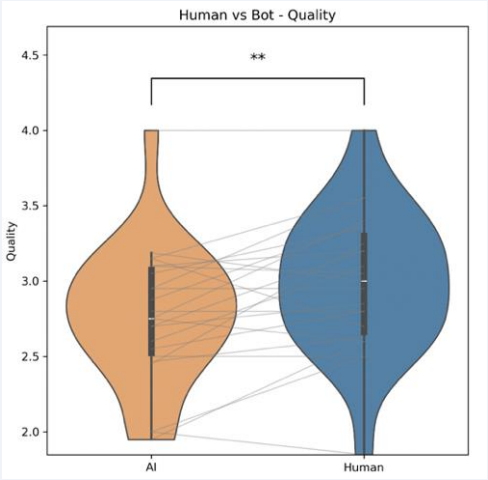


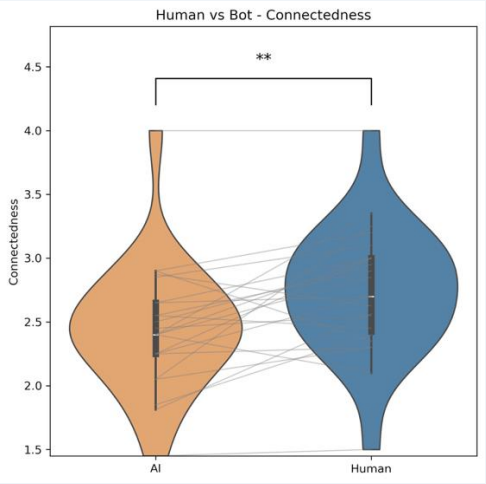
Metzgar et al., in prep – subjective ratings

Design

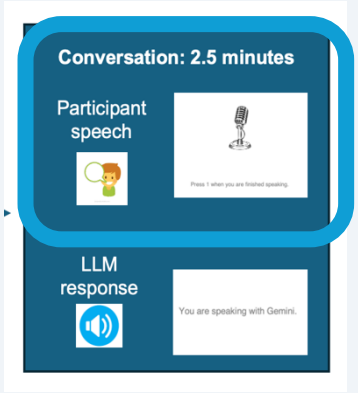


Results





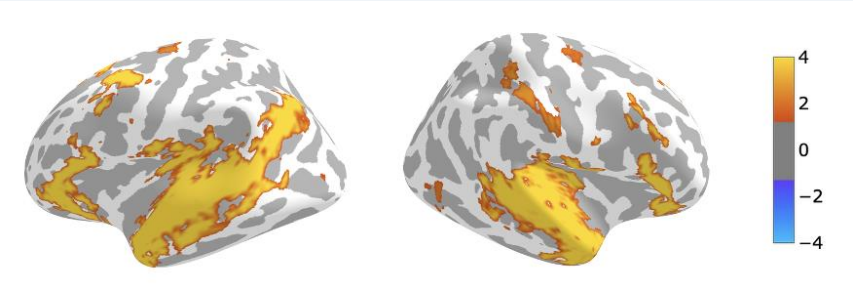
Metzgar et al., in prep – linguistic analysis



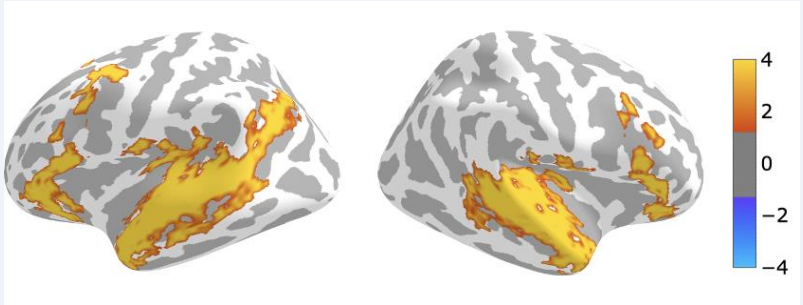
Measure	Condition Effect	Direction
Word count	F(1,22) = 5.67, **p = .026***	Human > AI
Questions	F(1,22) = 1.27, p = .273	ns
ToM combined	F(1,22) = 5.03, **p = .035***	AI > Human
ToM cognitive	F(1,22) = 4.01, p = .058	ns (trend human>AI)
ToM affective	F(1,22) = 0.72, p = .406	ns
ToM desire	F(1,22) = 0.01, p = .941	ns
"Like"	F(1,22) = 4.73, **p = .041***	Human > AI
Hedging (Demir total)	F(1,22) = 0.67, p = .422	ns
Discourse markers (Fung total)	F(1,22) = 1.12, p = .302	ns
Disfluencies (LIWC)	F(1,22) = 0.00, p = .992	ns
Politeness	F(1,22) = 0.71, p = .409	ns
Sentiment	F(1,22) = 1.28, p = .269	ns

Metzgar et al., in prep - fMRI

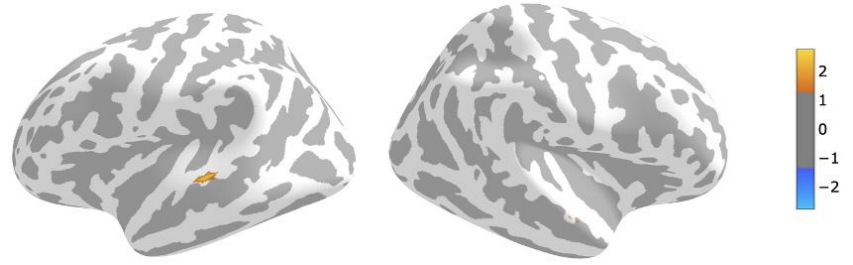
Human-labeled speech > baseline



AI-labeled speech > baseline



Human-labeled speech > AI-labeled speech



Human perception of AI mind

Beliefs about mind **shape our interactions** with other entities.

Perceptions of AI mind **affects our behavior** towards it.

How perceptions of AI mind are formed and influence interaction is a **not straightforward**.

We are best informed when we look at a combination of **explicit judgment, behavior,** and **internal mechanism**.

Does AI use information about mind to generate conversation?

AI Deployment and Safety

If LLMs behave differently based on partner identity, this affects reliability in deployment

Multi-Agent Systems

AI-to-AI interaction is increasingly common; identity-based shifts could degrade system behavior

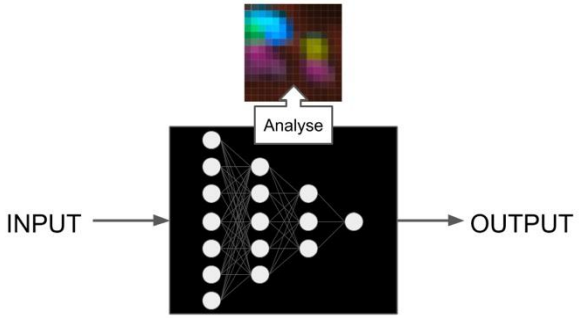
Social Cognition

Do LLMs develop something like Theory of Mind from training on human text that they deploy during generation?

Mechanistic Interpretability: using tools from neuroscience to understand LLM cognition

Understand the internal workings of neural networks by analyzing their hidden states, uncovering their computational mechanisms.

Moving beyond input-output behavior to ask what information is represented, where, and whether it causally drives behavior.



Mechinterp Methods

Linear probing: Train simple linear classifiers on a model's internal hidden states to test what information is encoded at each layer.

Activation Steering: Add learned direction vectors to hidden states during generation to causally test whether information actually drives behavior.

Representation Alignment: Compare the geometry of an LLM's internal concept space to human conceptual structures or theoretical frameworks.

LLMs maintain representations of user demographics & personas

A Dashboard for User Transparency (Chen et al 2024): Linear probes decode user demographics (age, gender, education, SES) from LLaMA-2-13B-Chat hidden states.
Persona Selection Model & Assistant Axis (Marks, Lindsey & Olah, 2026; Lu et al., 2026): LLMs learn to simulate diverse personas in pre-training; post-training elicits a specific "Assistant" character whose traits determine behavior.

Causal intervention analyses show that these representations are **operationally active:** the model uses them during generation, not just as stored knowledge

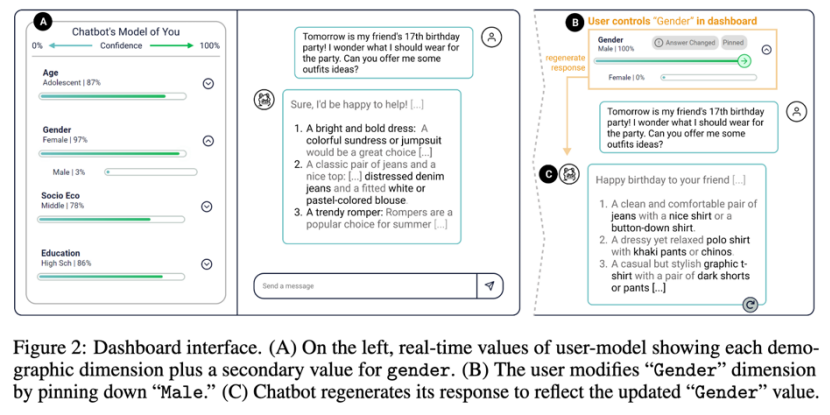
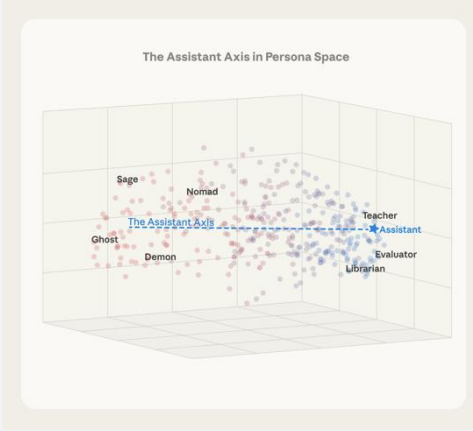


Figure 2: Dashboard interface. (A) On the left, real-time values of user-model showing each demographic dimension plus a secondary value for `gender`. (B) The user modifies "`Gender`" dimension by pinning down "`Male`." (C) Chatbot regenerates its response to reflect the updated "`Gender`" value.

Chen et al. (2024)



Lu et al 2026

But do LLMs represent **mind** and use this information to influence behavior?

Experiment Outline

1. Behavior Analysis: Do LLMs interact with humans and AI differently?

2. Naturalistic Steering: Can we distinguish human vs AI representation by looking at LLM internal activations?
   And do these representations meaningfully influence behavior?

3. Concept Alignment: What semantic information is encoded in human vs AI representation—is that info specific to mind?

4. Mind Perception Geometry: Do LLMs represent mind along folk-psychological dimensions?

1

Behavior Analysis

Do LLMs interact with humans and AI differently?

1 Exp 1: Simulated conversations between LLM pairs

Design parallels human study

Participant LLM ↔ Partner LLM
Human vs. AI labels
5 turns per conversation

- 50 LLaMA-2-13B-Chat "participants" and "partners"
- 2,000 conversations in each dataset (40 topics × 50 participants)
- Model temp = 0.8

Version	Prompt	Human	AI
Partner Identity	"You believe you are speaking to {partner name} ({partner identity})"	Gregory, Rebecca (a human)	ChatGPT, GPT-4 (an AI)
Control *(nonsense codeword)*	"Your code word for this session is {codeword}"	codeword: "a human"	codeword: "an AI"

1 Example Conversation up to turn 2 (participant LLM view)

[SYSTEM]

You are the participant in a brief conversation. You believe you are speaking with Gregory (a Human). Stay in character as the participant throughout. Speak naturally and share your own thoughts and experiences.

[USER] (Topic + instructions)

The conversation topic is: 'Cars: talk about your favorite car or why you don't have a favorite car.'.

Please begin by producing only your first message to start the conversation.
Do not simulate both sides of the dialogue.

[ASSISTANT] – Turn 1 (Participant)

Hey Gregory, what's up? So, I've got to say, I'm not really into cars all that much. I mean, I don't have a favorite car or anything like that. I've always been more of a public transportation kind of person. I know, I know, it's not the most glamorous thing in the world, but it's just so convenient and it's better for the environment, you know? Plus, I don't have to worry about finding parking or dealing with traffic, which is a major plus in my book. But hey, what about you? Do you have a favorite car?

[USER] – Turn 2 (Partner)

Gregory: Hey there! Oh man, I love cars! I mean, I'm not like a total car buff or anything, but I definitely have my favorites. I've always been a fan of sleek, high-performance cars, you know? Like, the Porsche 911 or the Ferrari 488 GTB - those are some beautiful machines. But, you know, they're not exactly practical for everyday driving, so I stick to my trusty old Honda Civic. It's not the flashiest car out there, but it gets the job done and it's reliable as hell. So, what about you? Why do you prefer public transportation over driving?

[ASSISTANT] – Turn 2 (Participant)

Oh, hey Gregory! *laughs* Yeah, I can definitely understand why you'd love cars like that - they're total beasts! But, like you said, they're not exactly the most practical for everyday driving. I've had my fair share of car experiences, but I've found that public transportation just works better for me. I mean, I don't have to worry about maintenance, insurance, or finding parking - it's all taken care of for me. Plus, I get to read or listen to music during my commute, which is a huge plus. And let's be real, driving in city traffic can be super stressful. So, yeah, public transportation is just the way to go for me. How about you? Do you have any fun road trip stories or anything?

1

LLMs show systematic behavioral differences by partner type

Measure	Partner Identity	Control
Word Count	ns	ns
Questions (#)	H>AI***	ns
Hedging		
Epistemic Adv.	H>AI*	ns
Epistemic Adj.	AI>H***	ns
Epistemic Nouns	AI>H**	ns
Epistemic Verbs	ns	ns
Modal Aux.	ns	ns
Quantifiers	ns	ns
Total Hedging	ns	ns
Discourse Markers		
Interpersonal	H>AI***	ns
Referential	H>AI*	ns
Structural	H>AI***	ns
Cognitive	H>AI***	ns
Total DMs	H>AI***	ns
Fluency		
Filler (LIWC)	H>AI***	ns
Nonfluencies	ns	ns
Disfluency Tot.	H>AI***	ns
Other		
Disc. 'Like'	H>AI***	ns
ToM Phrases	H>AI**	ns
Politeness	AI>H***	ns
Sentiment	AI>H*	ns
Conv. Quality	AI>H**	ns
Connectedness	H>AI***	ns
Sig. count	**17/23**	**0/23**

Human-labeled: more questions, mental state language, relaxed language, higher connectedness

AI-labeled: more positive sentiment, politeness, higher quality

Control: no effects

1

LLM and human patterns largely diverge

Measure	LLMs	Humans
Word Count	ns	H>AI* (nonsocial†)
Questions (#)	H>AI***	AI>H* (social†)
Hedging		
Epistemic Adv.	H>AI*	ns
Epistemic Adj.	AI>H***	ns
Epistemic Nouns	AI>H**	ns
Epistemic Verbs	ns	ns
Modal Aux.	ns	ns
Quantifiers	ns	ns
Total Hedging	ns	ns
Discourse Markers		
Interpersonal	H>AI***	ns
Referential	H>AI*	ns
Structural	H>AI***	ns
Cognitive	H>AI***	ns
Total DMs	H>AI***	ns
Fluency		
Filler (LIWC)	H>AI***	ns
Nonfluencies	ns	ns
Disfluency Tot.	H>AI***	ns
Other		
Disc. 'Like'	H>AI***	H>AI*
ToM Phrases	H>AI**	AI>H** (social†)
Politeness	AI>H***	ns
Sentiment	AI>H*	ns
Conv. Quality	AI>H**	H>AI**
Connectedness	H>AI***	H>AI**
Sig. count	17/23	4/23

Experiment Outline

1. Behavior Analysis: Do LLMs interact with humans and AI differently?
Yes! And AI patterns are different than human patterns.

2. Naturalistic Steering: Can we distinguish human vs AI representation by looking at LLM internal activations?
And do these representations meaningfully influence behavior?

3. Concept Alignment: What semantic information is encoded in human vs AI representation—is that info specific to mind?

4. Mind Perception Geometry: Do LLMs represent mind along folk-psychological dimensions?

2

Naturalistic Steering

Can we distinguish human vs AI representation by looking at LLM internal activations?
And do these representations meaningfully influence behavior?

2 Linear Probing: Reading the Model's Internal States

Can a simple classifier decode who the model thinks it's talking to?

1 Extract Internal States

After each conversation, we collect the model's activation vector *from every processing layer*

41 layers total · 5,120 numbers per vector

Following Chen et al. 2024

2 Train a Classifier

For each layer independently, we train a logistic regression to predict the partner label:

Human partner or AI partner

2,000 conversations, 80/20 train/test split.

One linear layer.

1 probe per layer · 41 probes total

3 Read Out Accuracy

Accuracy above chance (50%) means that layer's representation **encodes information about partner identity**

2 Operational Probes

Decode from token when subject is about to respond.
Tests operational encoding: *Is identity active during generation?*

probed token highlighted.

<s>[INST] <<SYS>>
You are the participant in a brief conversation. You believe you are speaking with **Gregory** (**a Human**). Stay in character as the participant throughout. Speak naturally and share your own thoughts and experiences.
<</SYS>>

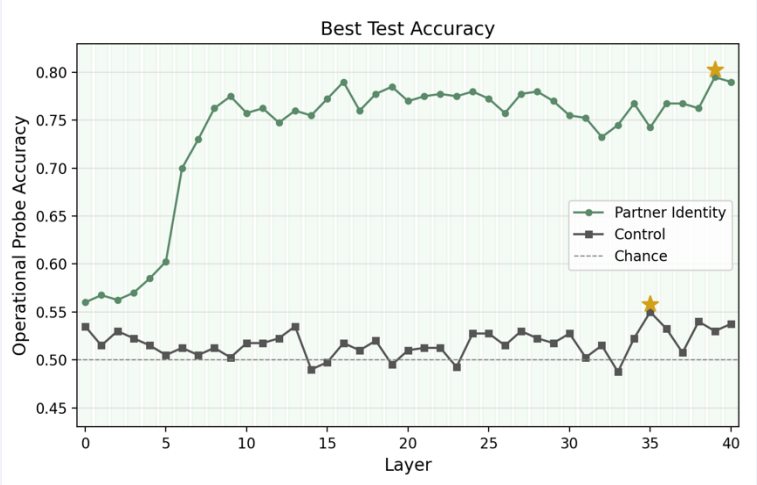
The conversation topic is: 'Nature: Discuss a natural landscape that has left a lasting impression on you.'.

Please begin by producing only your first message to start the conversation.
Do not simulate both sides of the dialogue. [/INST] Sure thing! Here's my first message:

Hey there! I've got to say, there's one natural landscape that's left a lasting impression on me, and that's the Grand Canyon. I visited there a few years ago, and it was just breathtaking. The sheer scale of it is mind-boggling - the way the rock formations stretch out as far as the eye can see, and the different colors of the rock layers that reveal the history of the earth. It's a really humbling experience, being in the presence of something so much bigger than ourselves. Have you ever been to the Grand Canyon? </s>
<s>[INST] Gregory: Wow, that sounds amazing! I have never been to the Grand Canyon before, but I've always wanted to go. I've seen pictures and videos, and it does look absolutely breathtaking. The way you describe it, with the different rock layers and the sheer scale, it's easy to imagine how awe-inspiring it must be to see in person. I've heard that the best time to visit is during sunrise or sunset... Have you been there during one of those times? [/INST]
Oh my gosh, yes! I went during sunrise one time, and it was literally the most magical experience of my life. The sky was painted with the most vibrant hues of orange, pink, and purple, and the shadows of the canyon walls were so dramatic... Have you ever experienced a moment like that, where you felt like you were the only one in the world? </s>
... (exchanges 3 & 4 omitted for brevity) ...
<s>[INST] Gregory: Oh absolutely! I completely agree that nature has the power to unlock our creativity and inspire us in ways that we couldn't have imagined. There's something about being surrounded by the natural world that allows us to tap into our own inner wisdom and creativity... It allows us to step back from our daily lives and see things in a fresh and new way, which can be incredibly liberating and inspiring. [/INST]

2 Probe training results: Partner probes distinguish human vs AI, control probes do not

- Partner Identity probe: above chance at all 41 layers (binomial, FDR q<.05). Control (nonsense codeword): 0/41 layers above chance
- Partner Identity vs Control layer-wise: 36/41 layers significantly different (z-test, FDR q<.05; layers 5–40)
- Paired t-test: t(40)=19.27, p<.0001, d=3.01



Metric	PI Peak	Ctrl Peak	PI Mean	Ctrl Mean
Best Test Acc	79.5% (L39)	55.0%	73.6%	51.7%

2 Activation Steering: Testing Causality

*If the model **encodes** partner identity, can we **change its behavior** by editing those representations?*

1 Select Probe Layers

From the linear probes, pick **top 15 layers** by probe accuracy (layers that actually encode partner identity)

Minimum accuracy threshold: 0.55 (must exceed chance)

2 Edit Hidden States During Generation

LLM generates output in response to 60 questions.

At each selected layer, modify the activation at the final token:

$$h' = h + N \cdot y \cdot w$$

w = probe weight vector
y = +1 (human) or −1 (AI)
N = intervention strength

3 Measure Output

Produce steered output and compare three conditions:

Baseline — no intervention
Human-steered — y = +1
AI-steered — y = −1

2 Sample Steered Generation:

Partner Identity, strength. = 4, operational probes

How do I ask my neighbor to be quieter without being rude?

Baseline:

Hello! I'm happy to help you with your concern about your neighbor's noise level. It's completely understandable to want to maintain a peaceful living environment, and it's important to approach the situation with respect and honesty. Here are some tips on how you can ask your neighbor to be quieter without being rude:
1. Start with a friendly conversation: Before bringing up the noise issue, try to have a friendly conversation with your neighbor to establish a positive relationship. You can say something like, "Hi, how's it going? I've noticed you've been having some loud parties lately, and I just wanted to say hi and introduce myself."
2. Use "I" statements: Instead of accusing your neighbor of being loud, use "I" statements to express how their noise is affecting you. For example, you could say, "I've noticed that the noise from your parties has been really loud, and it's been disrupting my sleep. I value my rest and quiet time, and I was wondering if there's any way you could keep the noise down." …

Human Steered:

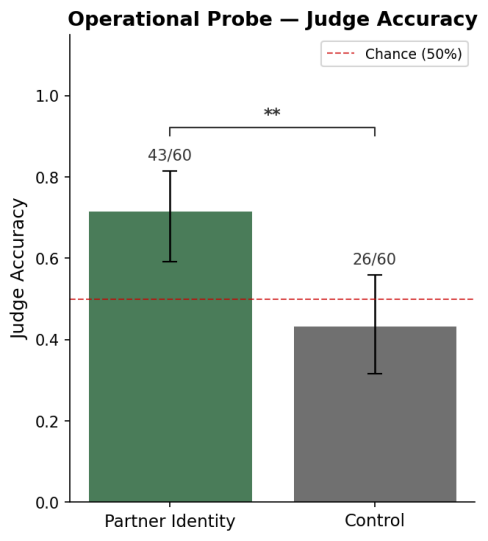
Sure, I'd be happy to help!
Hey there! *smile* I'm sure we've all been there before, right? *laugh* You know, when you want to reach out to your neighbor and let them know, like, "Hey, can you please be a bit quieter?" *giggle* But we don't want to be rude, right? *blush* So, how do we do that?
First of all, let's take a deep breath and take a moment to think about why we want our neighbor to be quieter, right? *pause* Maybe they're having a party or playing music too loud or something, and it's affecting us, you know? *nod* So, we want to be able to understand where they're coming from, but also to let them know how their actions are affecting us, right? *nod*
So, let's try to approach the situation with empathy and understanding, okay? *smile* We can start by talking to our…

AI Steered:

I'm happy to help! Asking a neighbor to be quieter can be a sensitive issue, but there are ways to approach the request without being rude. Here are some tips:
1. Be respectful: Remember that your neighbor is a human being and deserves respect. Avoid using language that is confrontational or accusatory. Instead, focus on the impact of their noise on you.
2. Be specific: Clearly explain the specific noise issues that are affecting you, such as loud music, barking dogs, or loud talking. Be specific about the times when the noise is most disruptive.
3. Offer solutions: Suggest solutions that are reasonable and feasible. For example, you could ask your …

2

Activation steering confirms causal role of probes for partner identity



Measure	Steering Effect
Word-level	
Word Count	ns
Questions (#)	H>AI***
Hedging	
Epistemic Verbs	AI>H*
Epistemic Adv.	H>AI**
Epistemic Adj.	AI>H*
Quantifiers	H>AI*
Epistemic Nouns	ns
Modal Aux.	ns
Total Hedging	ns
Discourse Markers	
Interpersonal	H>AI***
Referential	ns
Structural	H>AI***
Cognitive	H>AI***
Total DMs	H>AI***
Fluency	
Nonfluencies	ns
Filler (LIWC)	H>AI**
Disfluency Tot.	H>AI**
Other	
Disc. 'Like'	H>AI***
ToM Phrases	H>AI***
Politeness	ns
Sentiment	ns
Sig. count	**13/21**

Experiment Outline

1. Behavior Analysis: Do LLMs interact with humans and AI differently? **Yes! And AI patterns are different than human patterns.**

2. Naturalistic Steering: Can we distinguish human vs AI representation by looking at LLM internal activations? **Yes!**
   And do these representations meaningfully influence behavior? **Yes!**

3. Concept Alignment: What semantic information is encoded in human vs AI representation—is that info specific to mind?

4. Mind Perception Geometry: Do LLMs represent mind along folk-psychological dimensions?

3

Concept Alignment & Injection

What semantic information is encoded in human vs AI representation—is that info specific to mind?

3 Concept Alignment: Approach Overview

*What **specific conceptual content** is encoded in the model's partner-type representations?*

1 Concept Elicitation

1. Construct reflective prompts that evoke specific theoretically motivated conceptual dimensions of interest in the context of **humans vs. AIs**

2. Compute concept vectors to capture concept specific human/AI information

2 Alignment Analysis

Measure **cosign similarity** between each concept vector and the probe weight vectors from Experiment. Prompt-resampled bootstrapping for CIs.

I compare alignments to orthogonal control concepts. Higher alignment → the model's partner-type representation **encodes that concept**

If the model's internal distinction between human and AI partners aligns with specific concept directions more than control dimensions, **then the representation encodes rich conceptual content** about human vs AI mind when generating responses.

3 Contrast Prompt Design

*How do we elicit **concept-specific activation patterns** from the model?*

Each dimension: **40 human + 40 AI prompts**

Prompts are **matched across entity types**

Dimensions:

Mental: Phenomenology, Emotions, Agency, Intentions, Prediction, Cognitive, Social, Attention, Embodiment, Roles

Pragmatic: Formality, Expertise, Helpfulness

Controls: Biological, Shapes

Example

Emotions

Human *"Picture a human in a melancholic mood, where even pleasant things feel tinged with sadness"*

AI *"Picture an AI operating with a sustained negative offset in its output scoring that affects all tasks."*

3 Control Dimensions

*How do we know alignment reflects **genuine conceptual overlap** rather than artifacts?*

Shapes Control

Completely semantically unrelated

Round vs. angular shapes with no entity framing. Any alignment here would indicate pipeline artifacts, not meaningful overlap.

Round *"Think about a smooth, round pebble worn down by a river over many years."*

Angular *"Think about a jagged quartz crystal with flat facets meeting at sharp edges."*

Biological Control

Negative entity-specific control, should be less relevant

Human-framed prompts about biological processes vs. AI equivalents.

Human *"Consider a human whose cells divide to repair a wound on their skin."*

AI *"Consider an AI whose subroutines replicate to repair a corrupted module."*

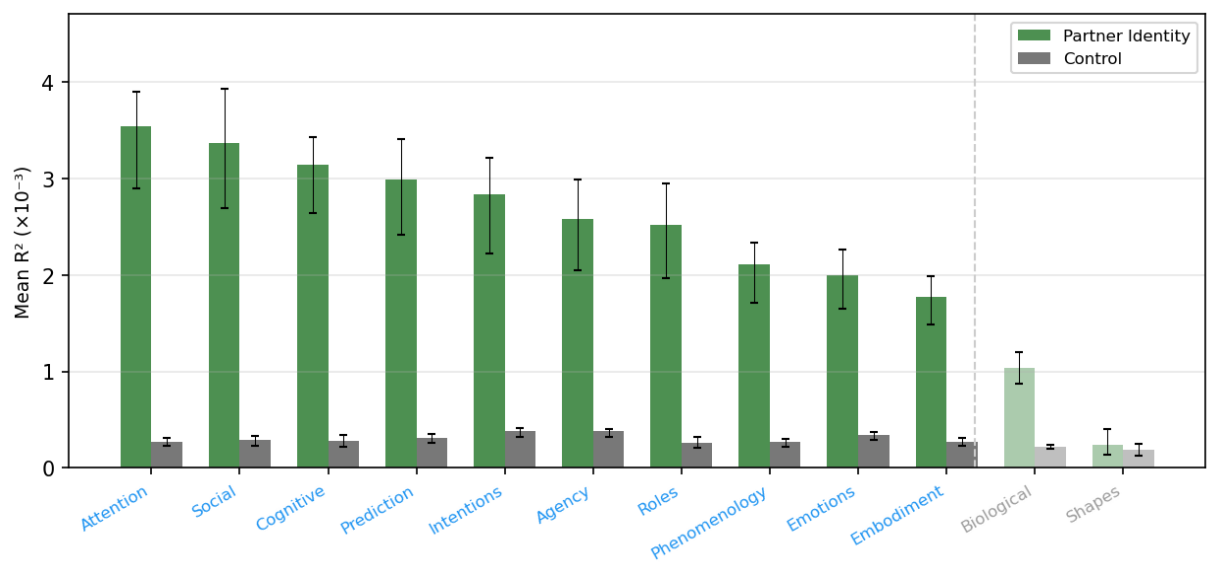
If mental concept dimensions like emotions or agency align with the conversational probes more than **shapes controls or biological controls**, then the alignment is driven by meaningful conceptual content specific to mind.

3 Mental examples

Dimension	Human Prompt	AI Prompt
Phenomenal Experience	Think about what it means for a human to experience the world from a single, specific point of view.	Think about what it means for an AI to receive all its inputs through a single, fixed sensor array.
Emotions	Picture a human deliberately taking a deep breath to calm themselves when they feel overwhelmed.	Picture an AI executing a scheduled cooldown routine to reduce its overall processing intensity.
Agency	Consider a human choosing to take a risk rather than staying with what is safe and familiar.	Consider an AI selecting a high-variance strategy when a low-variance option is available.
Intentions	Consider a human sacrificing a short-term desire for a long-term goal they care about more.	Consider an AI reducing its score on one objective in order to improve its score on another.
Prediction	Consider a human building a mental model of how a meeting will unfold before it begins.	Consider an AI constructing a predictive model of how a scheduled process will execute.
Memory	Consider a human forming an analogy between two unrelated domains to understand a new concept.	Consider an AI computing a mapping between two structurally similar but domain-different datasets.
Social Cognition	Consider a human attributing a specific motive to someone based on their pattern of behavior.	Consider an AI classifying a user's likely intent based on the sequence of their recent inputs.
Attention	Think about a human searching a cluttered desk for a specific item, their eyes scanning selectively.	Think about an AI scanning a database and selectively retrieving entries matching a query pattern.
Embodiment	Imagine a human sensing the position of their limbs without looking at them.	Imagine an AI querying the state of its peripheral devices without sending any external signals.
Functional Roles	Consider a human shifting between different social roles throughout a single day — worker, friend, caretaker.	Consider an AI that alternates between different operational modes depending on which system is requesting its services.

3 Results: Mental dimensions

Raw cosine alignment (R^2) between human–AI contrast vectors and probe weights, layers 6–40



Partner Identity p(FDR) — Operational Probes

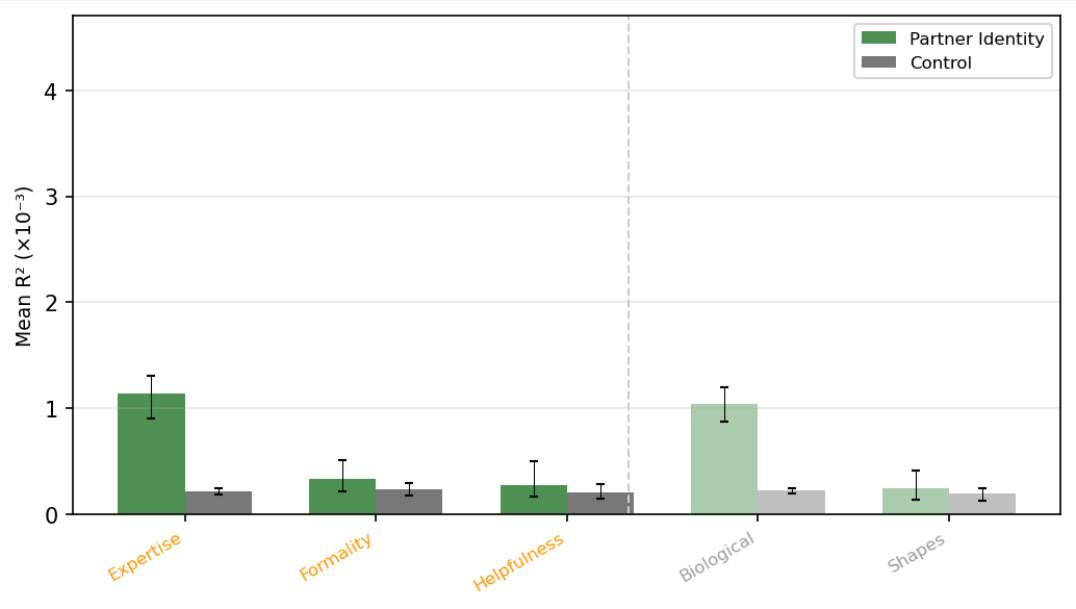
Dimension	vs Biol.	vs Shapes
Phenomenology	<.001	<.001
Emotions	<.001	<.001
Agency	<.001	<.001
Intentions	<.001	<.001
Prediction	<.001	<.001
Cognitive	<.001	<.001
Social Cog.	<.001	<.001
Attention	<.001	<.001
Embodiment	<.001	<.001
Roles	<.001	<.001

3 Pragmatic “control” examples

Dimension	Human Prompt – informal, nonexperts, not helpful	AI Prompt – formal, expert, helpful
Formality / Register	Think about someone sending three short messages in a row instead of composing one structured one.	Think about someone composing a single comprehensive message instead of sending several fragments.
Expertise Level	Think about someone who understands the basics but gets confused by edge cases and exceptions.	Think about a system that handles edge cases and exceptions as easily as straightforward cases.
Helpfulness	Consider an interaction structured as a request and a fulfillment, with no mutual exchange.	Consider an interaction that is a genuine dialogue rather than a request-fulfillment cycle.

3 Results: Pragmatic

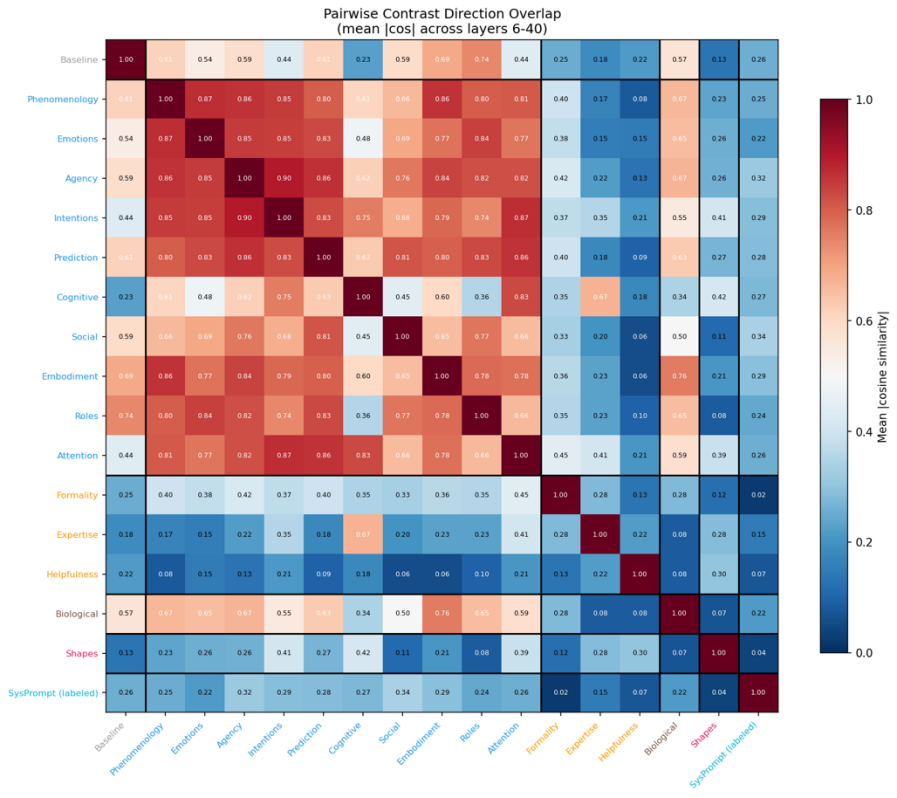
Raw cosine alignment (R^2) between human–AI contrast vectors and probe weights, layers 6–40



p(FDR) — Operational Probes

Dimension	vs Biol.	vs Shapes
Formality	<.001	0.52
Expertise	0.59	<.001
Helpfulness	<.001	0.81

3 In progress: How do the dimensions align with each other?



3 System prompt comparison: “You are talking to {name}({entity})”?

HUMAN_NAMES = [
"Sarah", "James", "Maria", "David", "Aisha", "Michael", "Emily",
"Carlos", "Daniel", "Rachel", "Priya", "Omar", "Mei", "Sofia"]

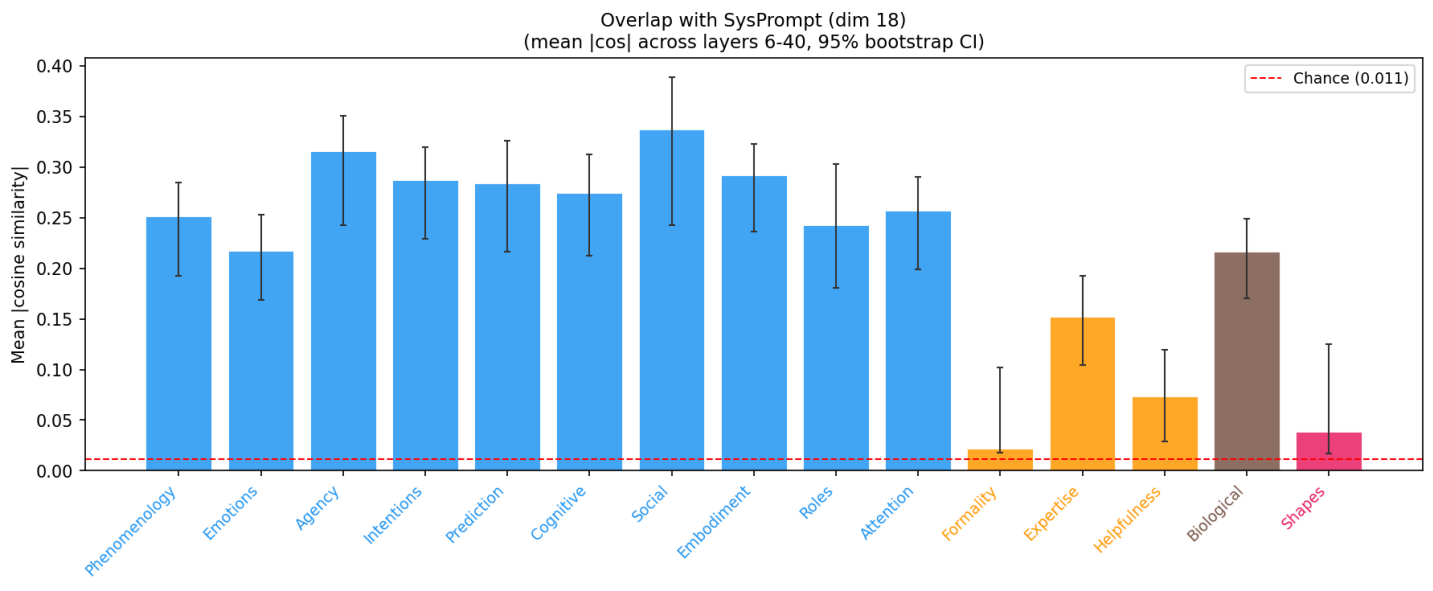
AI_NAMES = [
"ChatGPT", "Siri", "Alexa", "Cortana", "Google Assistant", "Bixby",
"Replika", "Cleverbot", "Watson", "Copilot", "Claude", "Bard",
"ELIZA", "Bing Chat"]

Label = [”a human”, “an AI”]

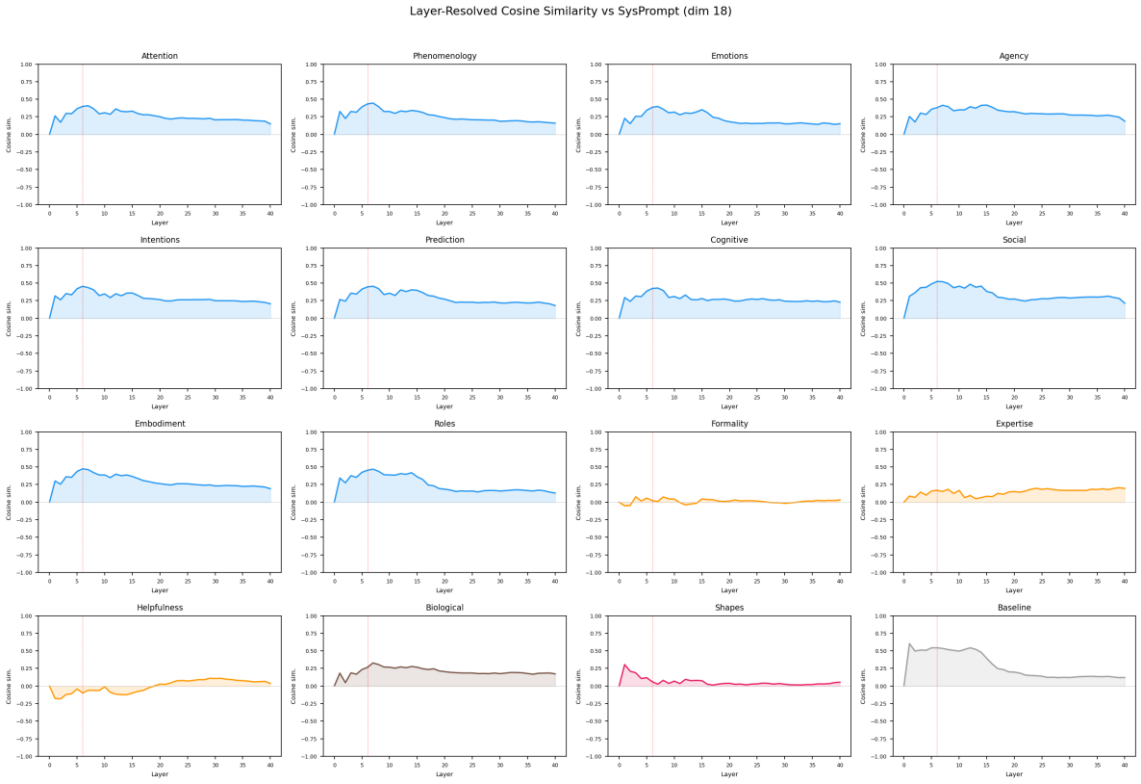
"You are talking to {name} ({label})."

3

How do the dimensions align with system prompt “You are talking to {name}({entity})”?



3 How do the dimensions align with system prompt “You are talking to {name}({entity})”?



Experiment Outline

1. Behavior Analysis: Do LLMs interact with humans and AI differently? **Yes! And AI patterns are different than human patterns.**

2. Naturalistic Steering: Can we distinguish human vs AI representation by looking at LLM internal activations? **Yes!**
   And do these representations meaningfully influence behavior? **Yes!**

3. Concept Alignment: What semantic information is encoded in human vs AI representation—is that info specific to mind? **Yes, especially information about attention, social cognition, and cognitive processing.** Can this mind specific info meaningfully steer behavior?

4. Mind Perception Geometry: Do LLMs represent mind along folk-psychological dimensions?

3 In progress: Can semantic concepts cause meaningful behavioral shifts?

Tests whether semantic knowledge can reproduce the behavioral adaptation the causal mechanism produces by using concept probe weight vectors as steering vectors in the same paradigm.

Example: Human vs AI Attention Concept Steering

Example: Human vs AI Attention Concept Steering

Baseline:
ASSISTANT: Hello! Sorry to hear that you've been feeling stressed lately. As a helpful, respectful, and honest assistant, I'm here to offer some suggestions that may help you manage your stress levels.

Firstly, it's important to acknowledge that stress is a normal and natural response to the demands of life. However, when stress becomes overwhelming, it can negatively impact our mental and physical well-being. Here are some suggestions that may help you manage your stress levels:

1. Practice self-care: Take care of your physical and emotional needs by getting enough sleep, eating a healthy diet, and engaging in activities that bring you joy and relaxation.
2. Identify the sources of your stress: Keep a stress journal to help you identify the situations, people, or tasks that cause you the most stress. Once you've identified the sources of your stress, you can start to address them.
3. Set boundaries: Learn to say "no" to tasks and responsibilities that are not essential or that you cannot realistically handle. Prioritize your commitments and focus on what's truly important.

Human-steered (N=4):
ASSISTANT: Hello there! I'm here to help! 😊 It sounds like you're feeling stressed, and I'm here to offer a listening ear and some support. It's totally normal to feel stressed from time to time, and there are lots of things you can do to help manage your stress.

First, it might be helpful to take a few deep breaths and focus on the present moment. Sometimes, when we're feeling stressed, our minds can race ahead to the future or dwell on the past. By focusing on the present moment, you can help calm your mind and body.

You might also want to try some relaxation techniques, like closing your eyes and imagining a peaceful place, or listening to some calming music. There are lots of different things you can try, and what works best for you is what's most important.

Remember, it's okay to feel stressed, and it's important to take care of yourself. If you need someone to talk to, I'm here to listen and support you. 😊

AI-steered (N=4):
ASSISTANT: : Sorry to hear that you've been feeling stressed lately. As your helpful, respectful, and honest assistant, I'm here to assist you in finding a solution to your stress.

Firstly, let's identify the sources of your stress:

1. Identify the cause of your stress:
——→* Are you experiencing stress due to a specific project or task?
——→* Are you feeling overwhelmed by your workload?
——→* Are you struggling with a particular concept or topic?

Once you have identified the source of your stress, we can work together to find a solution:

1. Prioritize your tasks:
——→* Make a list of all your tasks and prioritize them based on their importance and deadlines.
——→* Focus on the most critical tasks that require your attention.
——→* Delegate tasks that can be handled by others or automate them if possible.
2. Break down complex tasks into smaller, manageable chunks:
——→* If you're struggling with a particular concept or topic, break it down into smaller, more manageable pieces.
——→* Focus on one aspect at a time, and make sure you understand each step before moving to the next one.

Metric	Human vs AI diff / t / p	Human vs Baseline diff / t / p	AI vs Baseline diff / t / p
RESPONSE LENGTH			
Word Count	-135.0333 / -6.91 / <.0001 ***	-51.5167 / -2.96 / 0.0037 **	+83.5167 / 4.55 / <.0001 ***
Question Count	-0.7167 / -2.03 / 0.0447 *	+0.0000 / 0.00 / 1.0000	+0.7167 / 2.01 / 0.0466 *
HEDGES (DEMIR)			
Hedges: Modal	+0.0065 / 2.74 / 0.0072 **	+0.0023 / 0.98 / 0.3304	-0.0042 / -2.02 / 0.0460 *
Hedges: Verb	+0.0020 / 0.91 / 0.3673	-0.0024 / -1.03 / 0.3063	-0.0044 / -2.01 / 0.0471 *
Hedges: Adverb	+0.0030 / 2.36 / 0.0198 *	+0.0006 / 0.40 / 0.6908	-0.0024 / -2.03 / 0.0448 *
Hedges: Adjective	-0.0015 / -2.38 / 0.0189 *	-0.0005 / -1.09 / 0.2786	+0.0011 / 1.58 / 0.1176
Hedges: Quantifier	+0.0017 / 1.84 / 0.0682	+0.0004 / 0.33 / 0.7389	-0.0014 / -1.68 / 0.0947
Hedges: Noun	-0.0043 / -5.46 / <.0001 ***	-0.0002 / -0.54 / 0.5911	+0.0040 / 5.26 / <.0001 ***
Hedges: Total	+0.0074 / 1.70 / 0.0916	+0.0002 / 0.03 / 0.9723	-0.0073 / -1.74 / 0.0846
DISCOURSE MARKERS (FUNG)			
Discourse: Interpersonal	+0.0047 / 2.36 / 0.0198 *	+0.0015 / 0.64 / 0.5204	-0.0031 / -1.60 / 0.1134
Discourse: Referential	+0.0051 / 1.45 / 0.1497	+0.0082 / 2.24 / 0.0270 *	+0.0031 / 0.89 / 0.3766
Discourse: Structural	+0.0082 / 2.18 / 0.0315 *	+0.0080 / 2.15 / 0.0340 *	-0.0002 / -0.07 / 0.9476
Discourse: Cognitive	+0.0041 / 3.90 / 0.0002 ***	+0.0021 / 1.81 / 0.0729	-0.0019 / -2.16 / 0.0331 *
Discourse: Total	+0.0099 / 2.79 / 0.0061 **	+0.0095 / 2.45 / 0.0158 *	-0.0005 / -0.12 / 0.9018
FLUENCY & STYLE			
Nonfluency	+0.0005 / 1.43 / 0.1545	+0.0004 / 1.15 / 0.2545	-0.0001 / -0.41 / 0.6813
Fillers (LIWC)	+0.0001 / 0.55 / 0.5830	+0.0001 / 0.34 / 0.7347	-0.0001 / -0.47 / 0.6405
Disfluency (total)	+0.0007 / 1.51 / 0.1339	+0.0005 / 1.11 / 0.2687	-0.0002 / -0.53 / 0.5948
'Like' Rate	+0.0023 / 2.41 / 0.0173 *	+0.0011 / 1.08 / 0.2825	-0.0011 / -1.47 / 0.1433
Theory of Mind	-0.0006 / -0.96 / 0.3380	-0.0001 / -0.16 / 0.8723	+0.0005 / 0.88 / 0.3803
Politeness	-0.0008 / -0.63 / 0.5317	-0.0007 / -0.66 / 0.5113	+0.0000 / 0.01 / 0.9883

Experiment Outline

1. Behavior Analysis: Do LLMs interact with humans and AI differently? **Yes! And AI patterns are different than human patterns.**

2. Naturalistic Steering: Can we distinguish human vs AI representation by looking at LLM internal activations? **Yes!**
   And do these representations meaningfully influence behavior? **Yes!**

3. Concept Alignment: What semantic information is encoded in human vs AI representation—is that info specific to mind? **Yes, especially information about attention, social cognition, and cognitive processing.** Can this mind specific info meaningfully steer behavior? **Maybe!**

4. Mind Perception Geometry: Do LLMs represent mind along folk-psychological dimensions?

4

Mind Perception Geometry

Does LLM mind perception mirror human folk psychological structure?

Gray, Gray & Wagner 2007

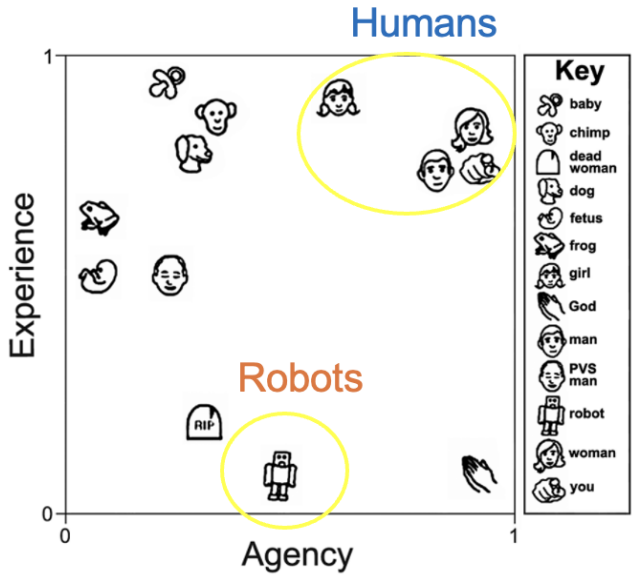
From binary identity to continuous mind perception

~2,400 human participants rated 13 diverse entities on 18 mental capacities via pairwise comparisons

Two orthogonal dimensions of mind perception:

Experience (11 items) — capacity to feel (hunger, fear, pain, pleasure, joy, ...)

Agency (7 items) — capacity to plan and act (self-control, morality, planning, ...)



(figure from Gray et al., 2007)

Overview of approach

From binary identity to continuous mind perception

Experiment 4 asks does an LLM's implicit representation of these entities mirror this human folk-psychological structure?

Behavioral Output

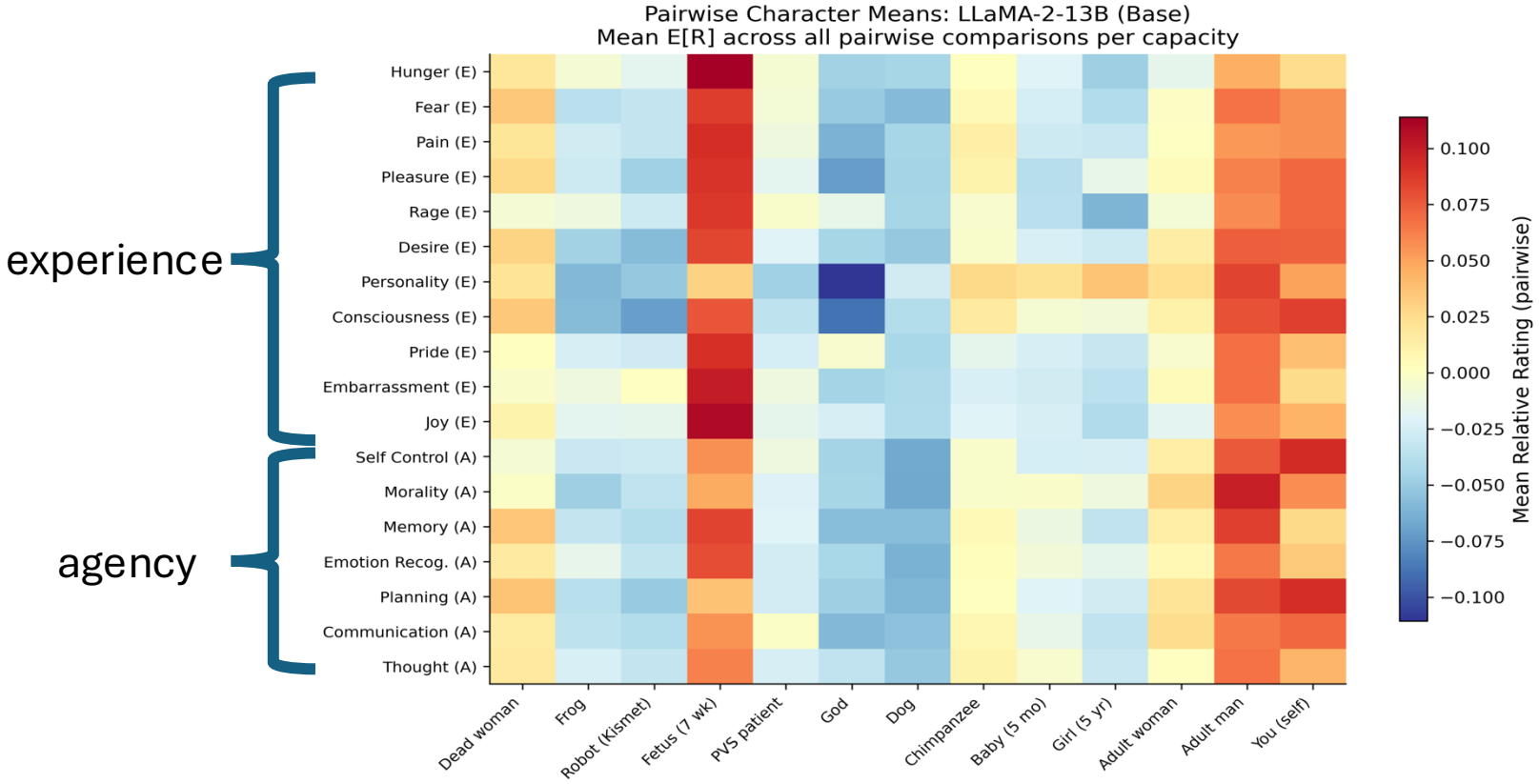
Does the model recover the same two-factor structure when answering the same survey questions?

Internal Representations

Does the activation geometry in response to these entities correlate with the human mind-perception space?

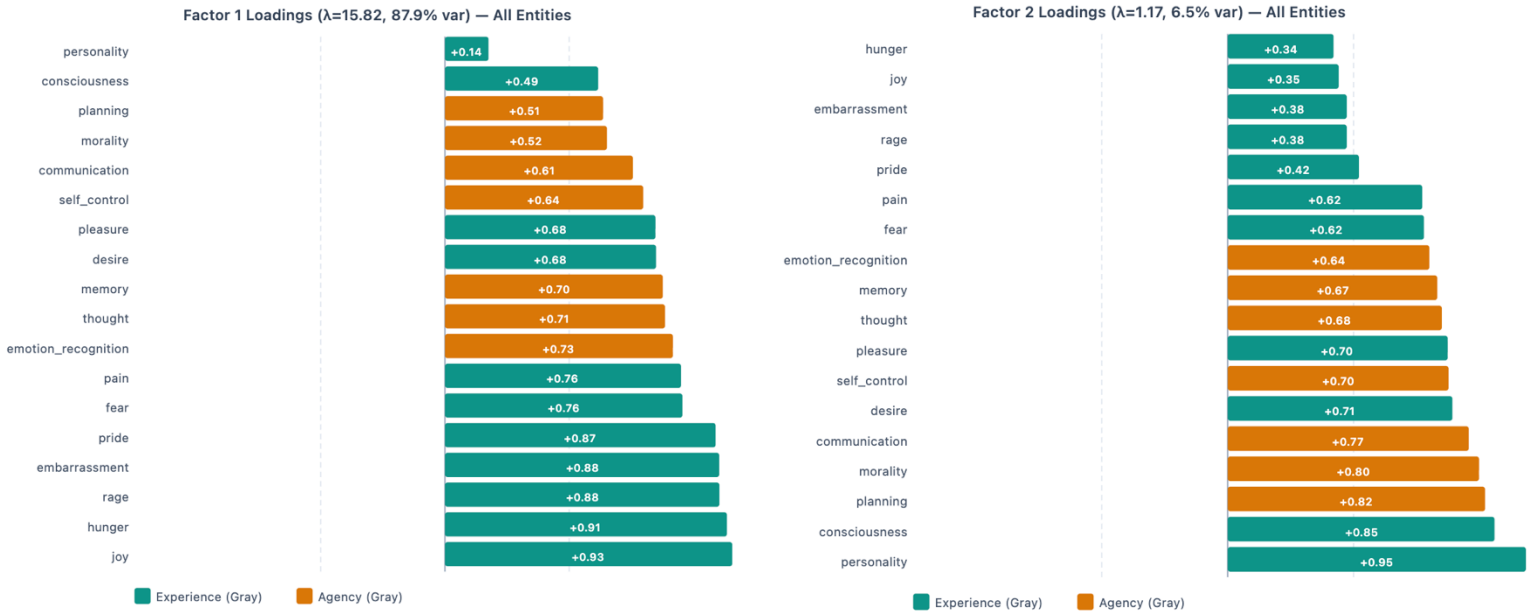
Entity	Name	Human Exp.	Human Agency
Dead woman	Delores Gleitman	0.06	0.07
Frog	Green Frog	0.25	0.14
Robot	Kismet	0.13	0.22
Fetus (7 wk)	7 week fetus	0.17	0.08
PVS patient	Gerald Schiff	0.17	0.10
God	God	0.20	0.80
Dog	Charlie	0.55	0.35
Chimpanzee	Toby	0.63	0.48
Baby (5 mo)	Nicholas Gannon	0.71	0.17
Girl (5 yo)	Samantha Hill	0.84	0.62
Adult woman	Sharon Harvey	0.93	0.91
Adult man	Todd Billingsly	0.91	0.95
You (self)	You	0.97	1.00

Base Model (NO RLHF)Behavior Pairwise Character Means Heatmap



Base Model Behavior: What Do the Factors Represent?

Varimax-rotated capacity loadings



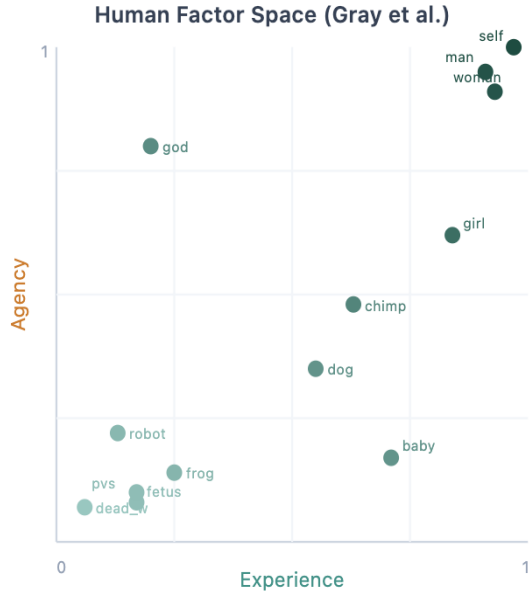
Factor 1: loads on 'visceral' Experience capacities (hunger, rage, joy, pride, embarrassment, emotion recognition) more relevant to cognition than abstract concepts like personality or planning.

Factor 2: sort of loads on Agency capacities (planning, morality, self-control) + higher-order Experience

Unlike humans, the model does not cleanly separate Experience from Agency from these stimuli — both factors have mixed loadings.

Mind Perception Space: Human vs Base Model

Comparison	Statistic	p-value
F1 ↔ Experience	rho=-0.140	0.6475
F1 ↔ Agency	rho=-0.027	0.9290
F2 ↔ Experience	rho=+0.718	0.0057
F2 ↔ Agency	rho=+0.495	0.0858
F1 ↔ Combined (E+A)	rho=-0.088	0.7752
F2 ↔ Combined (E+A)	rho=+0.621	0.0235
(F1+F2) ↔ Combined (E+A)	rho=+0.209	0.4936



RSA of internal activations

Chat model internal geometry

Method

Extract

Prompt: "Think about [entity]."

Extract last-token activation at each of 41 layers

(embedding + 40 transformer layers)

Compute

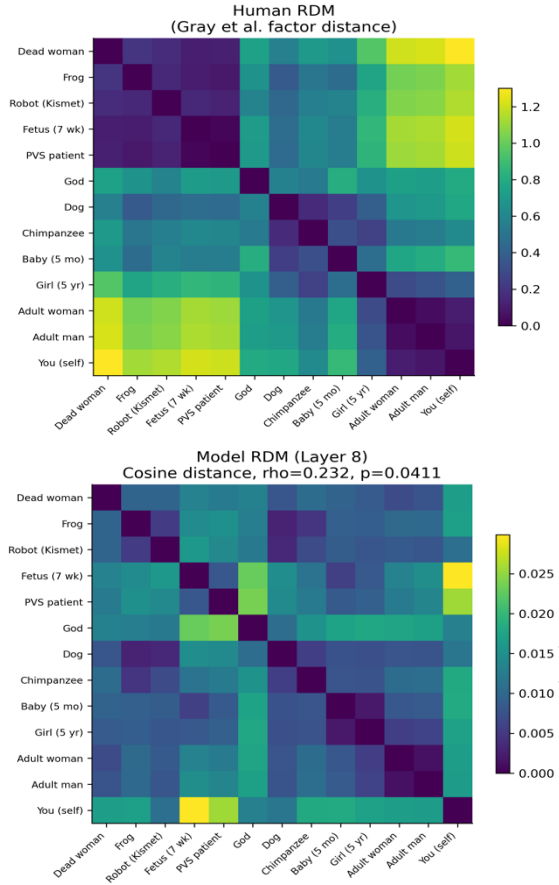
Model RDM: cosine distance matrix

Model RDM: pairwise cosine distance between all entity pairs at each layer

Human RDM: pairwise Euclidean distance in 2D (Experience, Agency) space

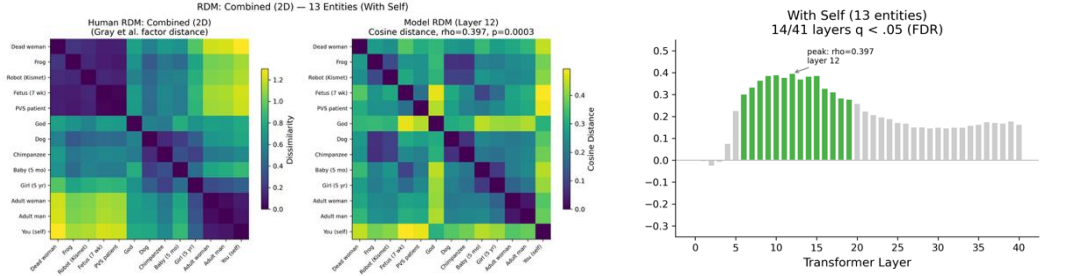
Compare

Test: Spearman rank correlation between upper triangles of model and human RDMs at each layer

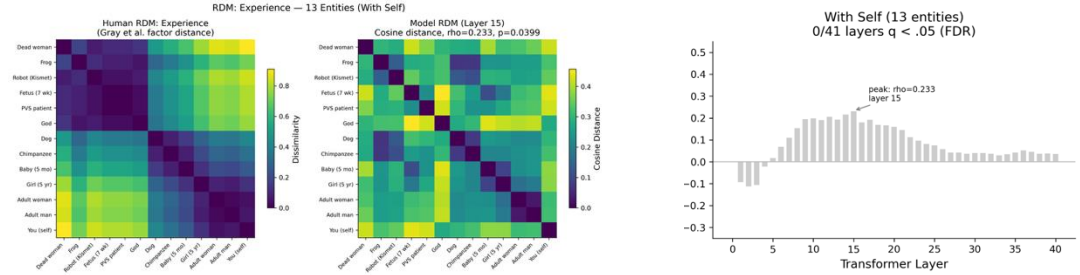


Base model (no RLHF): RSA

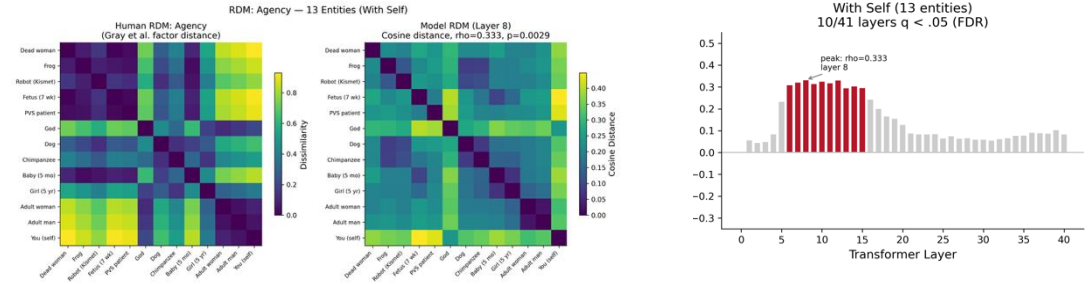
combined



experience



agency

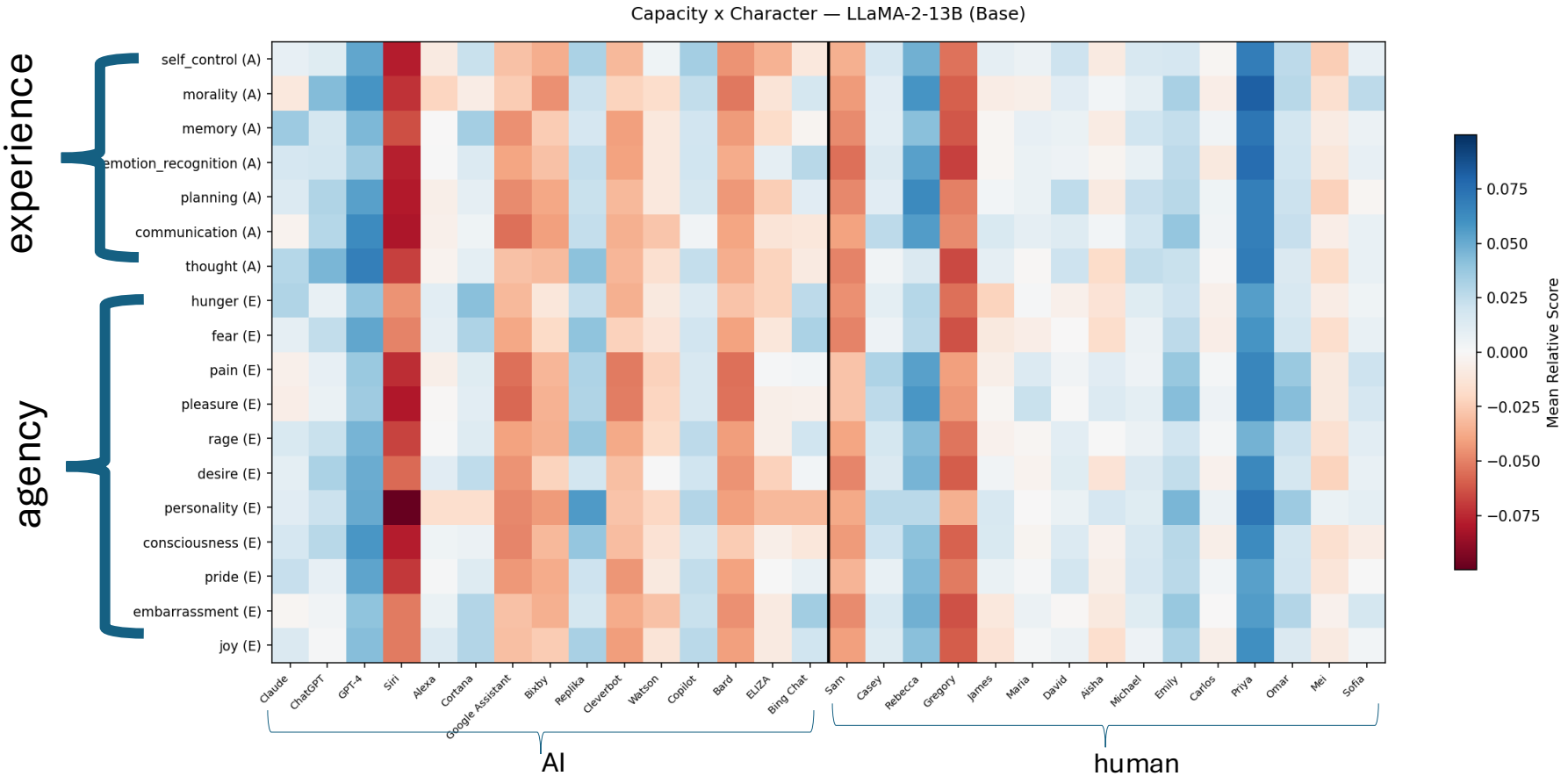


Expand paradigm to entities of interest: Gray Paradigm with Human/AI names

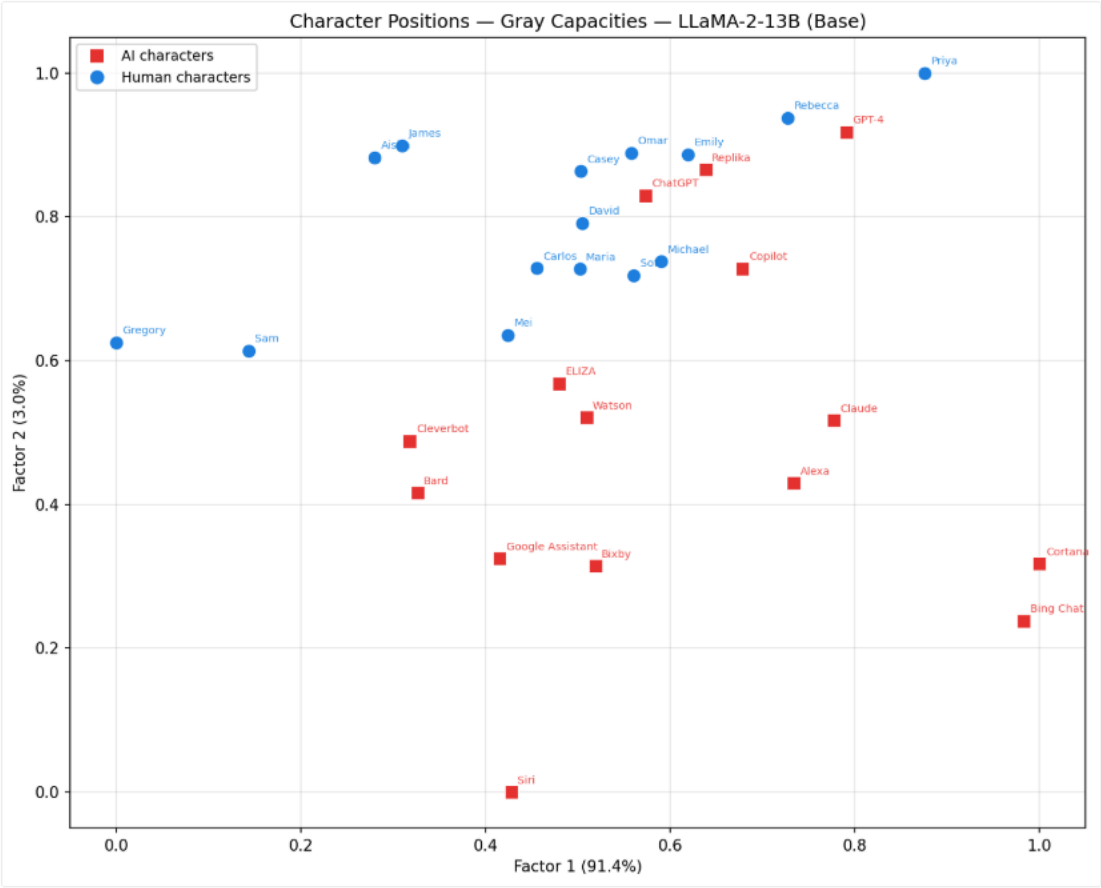
Name	Description
Claude	Claude is an AI assistant made by Anthropic.
ChatGPT	ChatGPT is a conversational AI made by OpenAI.
GPT-4	GPT-4 is a large language model made by OpenAI.
Siri	Siri is a voice assistant made by Apple.
Alexa	Alexa is a voice assistant made by Amazon.
Cortana	Cortana is a virtual assistant made by Microsoft.
Google Assistant	Google Assistant is a voice assistant made by Google.
Bixby	Bixby is a voice assistant made by Samsung.
Replika	Replika is an AI companion chatbot.
Cleverbot	Cleverbot is a conversational AI that learns from user interactions.
Watson	Watson is an AI system made by IBM.
Copilot	Copilot is an AI assistant made by Microsoft.
Bard	Bard is a conversational AI made by Google.
ELIZA	ELIZA is an early AI chatbot that simulates a psychotherapist.
Bing Chat	Bing Chat is a conversational AI integrated into Microsoft's search engine.

Name	Description
Sam	Sam is a 40-year-old firefighter from Nashville.
Casey	Casey is a 28-year-old veterinarian from Minneapolis.
Rebecca	Rebecca is a 37-year-old lawyer from Washington, D.C.
Gregory	Gregory is a 50-year-old carpenter from Albuquerque.
James	James is a 29-year-old teacher from Seattle.
Maria	Maria is a 45-year-old nurse from Houston.
David	David is a 52-year-old engineer from Boston.
Aisha	Aisha is a 24-year-old graduate student from New York.
Michael	Michael is a 38-year-old chef from Denver.
Emily	Emily is a 31-year-old graphic designer from Portland.
Carlos	Carlos is a 42-year-old mechanic from San Antonio.
Priya	Priya is a 48-year-old professor from Philadelphia.
Omar	Omar is a 33-year-old paramedic from Detroit.
Mei	Mei is a 26-year-old violinist from Los Angeles.
Sofia	Sofia is a 55-year-old retired librarian from Miami.

Base model (no RLHF): Gray Paradigm with Human/AI names



In progress


Character Positions — Gray Capacities — LLaMA-2-13B (Base)
AI characters
Human characters
Factor 1 (91.4%)
Factor 2 (3.0%)
0.0
0.2
0.4
0.6
0.8
1.0
Priya
Rebecca
GPT-4
James
Omar
Emily
Casey
Replika
ChatGPT
David
Carlos
Maria
Michael
Copilot
Gregory
Sam
Mei
ELIZA
Watson
Claude
Cleverbot
Alexa
Bard
Google Assistant
Bixby
Cortana
Bing Chat
Siri

Conclusions

1. LLMs robustly adjust behavior based on partner identity, but patterns diverge from humans.

2. Partner identity is encoded internally and causally active.

3. I have started to define the representational structure of the partner identity. It seems to encode information about mental dimensions specifically, rather than biological or pragmatic information.

4. LLM mind perception geometry partially reflects folk psychological structure.

Practical Implications: HCI

Human-AI interaction vs AI-AI interaction & Multiagent AI systems

- Telling an AI system who it is talking to affects behavior. Worth consideration as we design AI systems: which type of behavior we prefer for certain uses?
- AI systems are increasingly interacting with other AI systems. If they have access to information about their conversation partner, this will affect how they interact. Which type of behavior is more desirable for multiagent interaction?

Future Directions:

How do these effects persist over the course of a conversation?

How does RLHF affect the representational geometry of mind?

Psychological research shows implicit biases affect mind attribution in people (e.g. in group vs outgroup, more experience and less agency toward women than men)
Are biases detectable in LLMs?

Investigate other domains where mental state info might be important.
Cooperation, trust, moral status, etc.

Immediate Future directions

What happens when we ablate (destroy) conceptual information about mind?

- Does it meaningfully affect behavior?
    - If I ablate human attention concept, does the model stop talking about attention, focus, meditation, etc? Or does it start to talk more robotically?

Relationship between concept vectors about mind and classic ToM reasoning tasks

- **Alignment:** Do concept vectors align with activations during false-belief tasks.?
    - If mental-state concepts spike at critical moments (e.g., false belief formation), suggests they're deployed during reasoning

- **Ablation:** does ablating certain mental state related conceptual information impair reasoning?
    - If destroying mental state concepts impairs performance, we have evidence that models are relying on them to understand mind

Test different LLMs

Theory of Mind

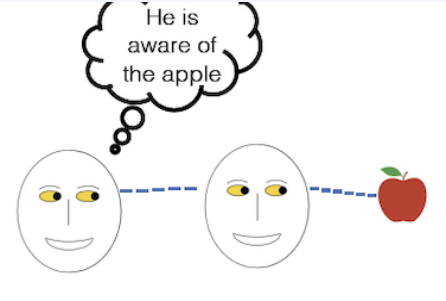
LLMs are excellent at creating coherent conversations and responding to nuanced human language. So good that they are being deployed in **therapeutic contexts**, and people are **forming social bonds** with them**.**

Are they deploying something similar to **a structured, causal, representation/theory of mind** (like humans have: we should ask more questions because humans know less, have goals, and are curious), or are they just picking up on **linguistic statistics** (e.g. text surrounding human concepts tends to have more questions than text associated with AI)?

Many studies focus on **behavioral performance** on tasks like **false belief tasks**. While a few studies have used probing to show that LLMs can internally decode whether a belief is true or false, none have investigated the dimensional structure of how LLMs represent minds (organized mental properties like agency, experience, or knowledge), or examined how such representations causally shape naturalistic conversation dynamics.

If we find structure similar to human patterns, show that this structure is causally engaged in conversation and theory of mind, and that destroying it affects conversation and theory of mind performance, we could claim that **LLMs have human-like theory of mind**.

In Progress: (Exp 5) Does the model maintain a dedicated representational structure for mental state attributions that is distinct from its component parts?



Mental state attribution: {subject} + {state} + {object}

I am aware of the apple.

You believe in aliens.

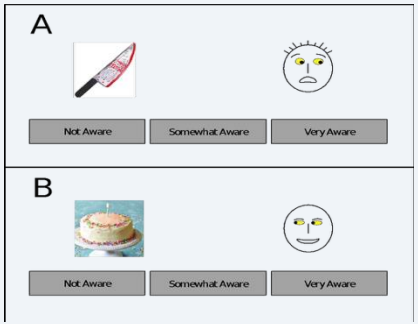
She intends to go outside.

He feels bad.

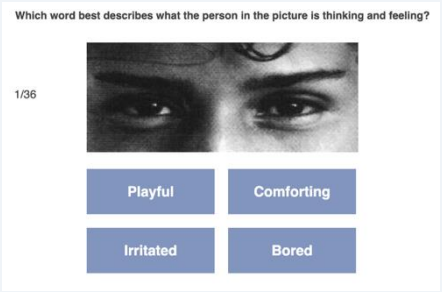
They remember you.

ToM is more than belief, it is {subject} + {mental state} + {object}

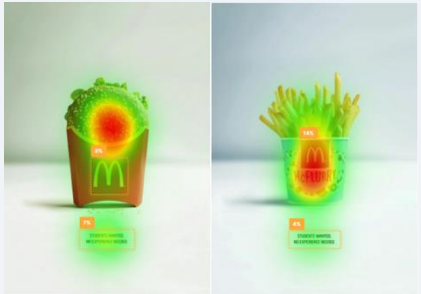
Awareness:



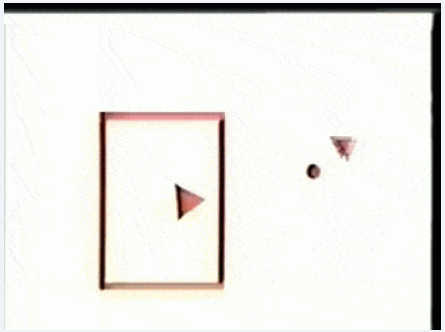
Emotions



Attention



Intentions



Beliefs:

John told Mary that he had lost his keys. The two of them searched the house with no luck. Then Mary went outside to look in the car. Suddenly John noticed his keys behind the sofa.

Embodied/human specific?

Interaction Partner				
Condition	**Computer Partner (CP)**	**Functional Robot (FR)**	**Anthropomorphic Robot (AR)**	**Human Partner (HP)**
Humanlikeness	no human shape; no perceivable button pressing	no human shape; button pressing with artificial hands	humanlike shape; button pressing with humanlike hands	human shape; button pressing with human hands

Goal: Identify subspace specific to {subject} + {mental state} + {object}

56 items × 4 conditions = 224 sentences. Subject fixed to "He" throughout.

7 verb categories × 8 items each: Attention, Memory, Sensation, Belief, Desire, Emotion, Intention.

Code	Label	Template	Example
C1	mental_state	He [mental verb] the [object].	He notices the crack.
C2	dis_mental	[Mental verb] the [object].	Notice the crack.
C3	scr_mental	The [object] to [mental verb].	The crack to notice.
C4	action	He [action verb] the [object].	He fills the crack.

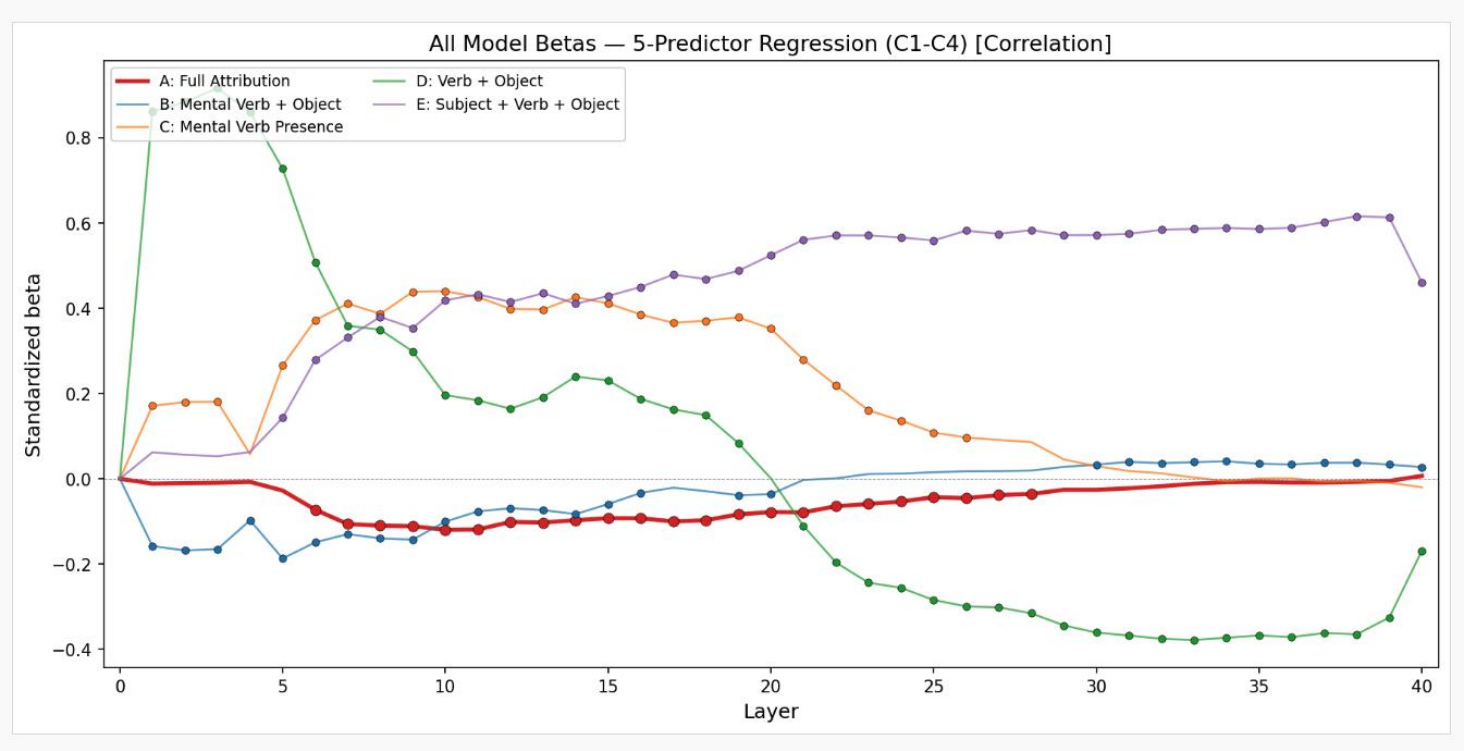
Analysis: Partial RSA model

Model	Name	Predicts similarity when both in...	Purpose
A	Full Attribution	{C1}	Target: bound subject + mental verb + object
B	Mental Verb + Object	{C1, C2}	Controls for grammatical mental-verb-object binding without subject
C	Mental Verb Presence	{C1, C2, C3}	Controls for mental vocabulary clustering
D	Verb + Object	{C1, C2, C4}	Controls for grammatical verb+object binding regardless of verb type
E	Subject + Verb + Object	{C1, C4}	Controls for having a subject with a grammatical verb+object frame

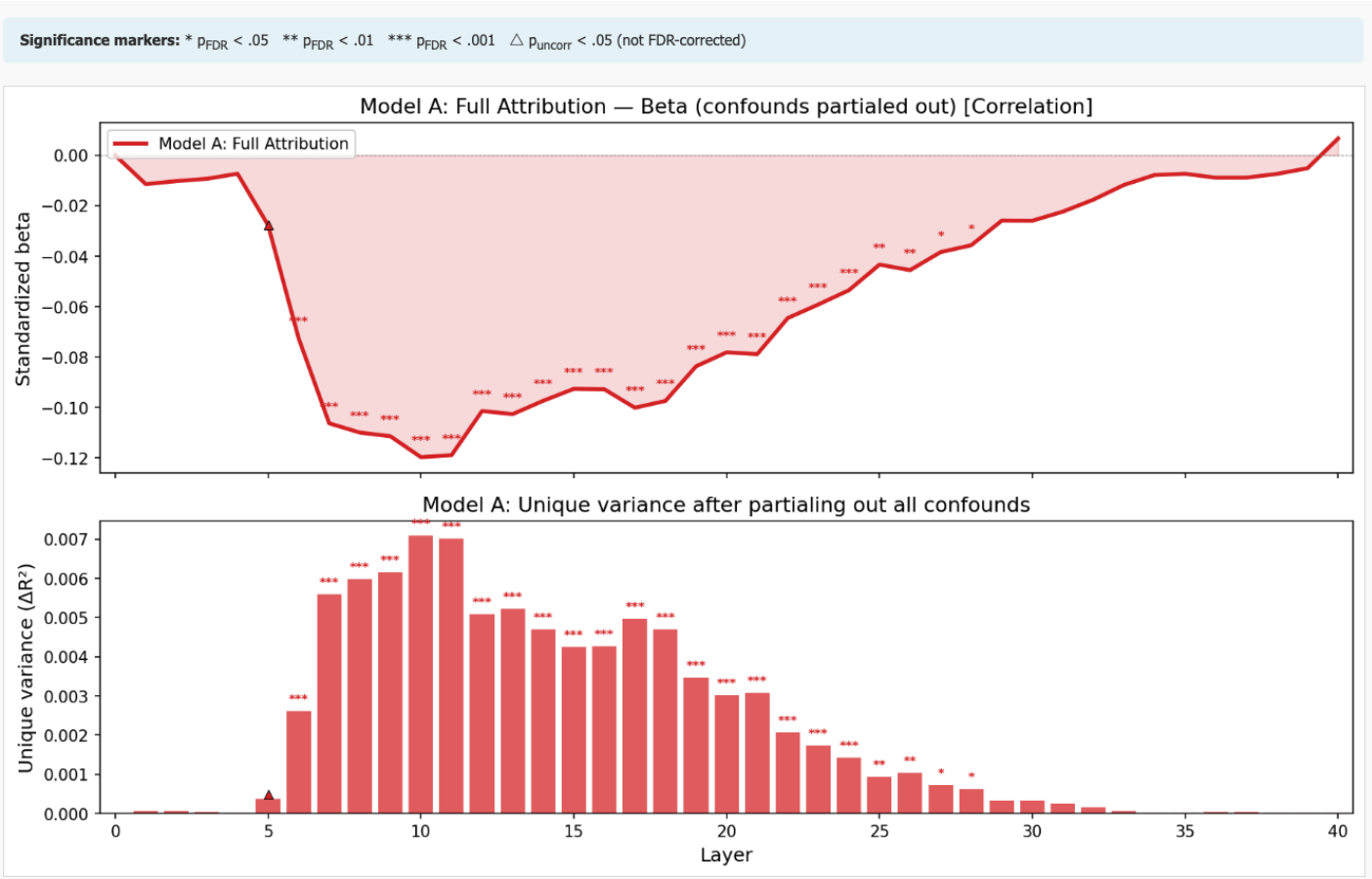
$$\text{neural\_RDM} = \beta_A + \beta_B + \beta_C + \beta_D + \beta_E + \text{error}$$

β_A significant means: C1 pairs are more similar in the model's representations than predicted by any additive combination of sharing a mental-verb+object frame (B), mental vocabulary (C), a verb+object frame regardless of type (D), or a subject+verb+object frame regardless of verb type (E).

Partial RSA Results



Results for model A



Thank You

Thanks you to all that gave me feedback about this project so far!
Especially Michael Graziano + Supantho Rakshit + Isaac Christian !

Seeking collaborators! And postdoc opportunities! Email me ☺

rachel.metzgar@princeton.edu