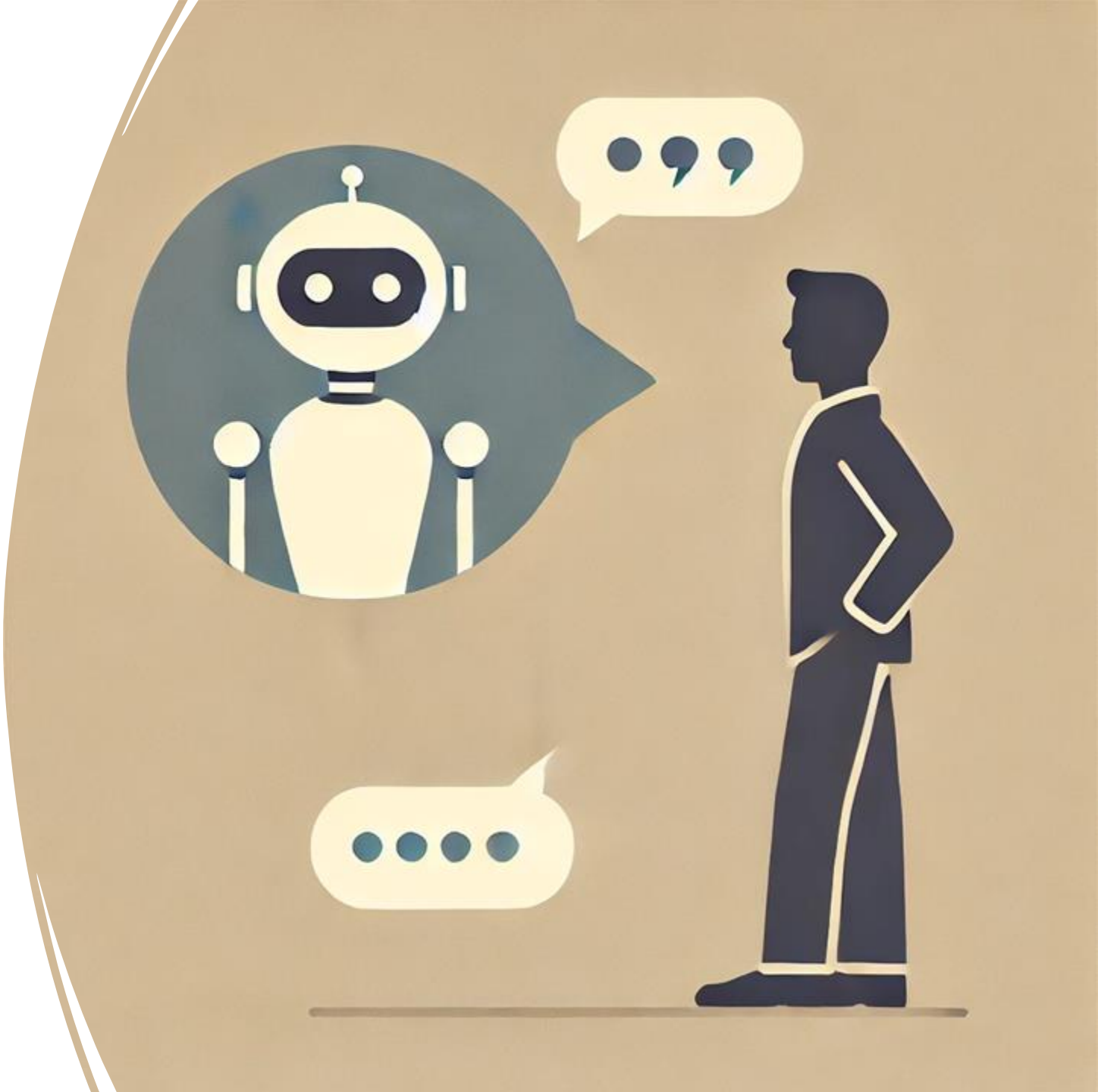
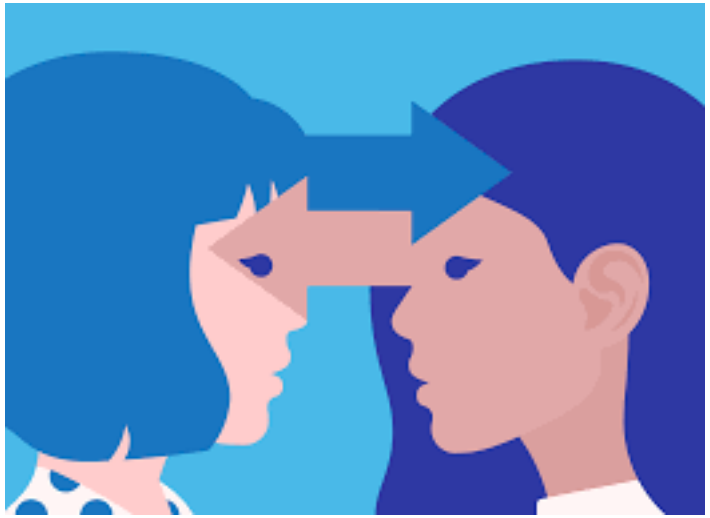
Thinking Machines or Thinking Minds? Neural and Behavioral Responses to Beliefs about Conversational Partners

Rachel Metzgar
PhD Candidate,
Princeton University

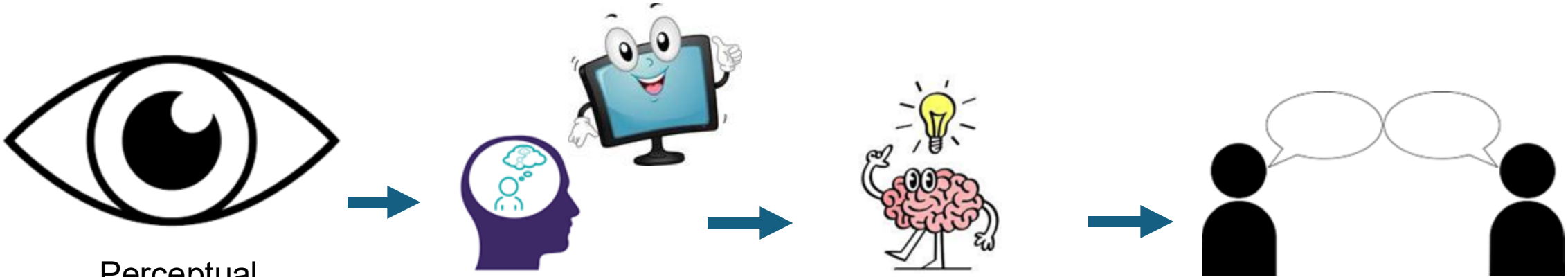


Theory of Mind (ToM)

- Attribute beliefs, intentions, emotions to others
- Crucial for social interaction (and now AI interaction): communication, cooperation, decision-making, etc.

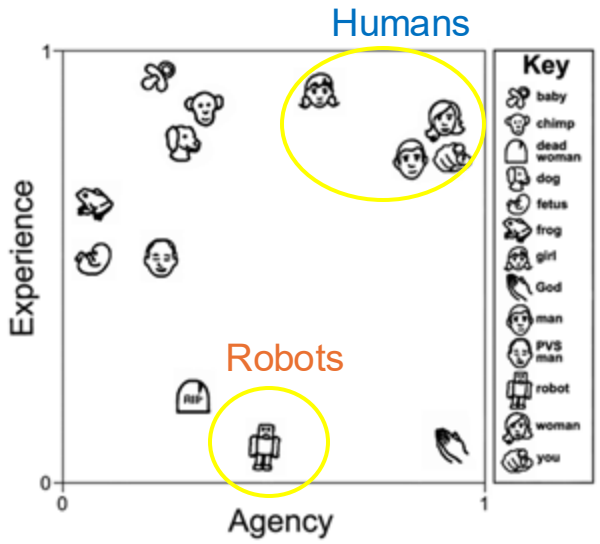


How do beliefs about a conversational partner's identity (human vs. AI) influence neural and behavioral Theory of Mind responses?



Perception of Human vs. AI

- People consistently attribute lower levels of consciousness, emotion, and intention to AI compared to humans (Gray et al., 2007; Huebner, 2009)
- Beliefs shape social interactions



(figure from Gray et al., 2007)

Do explicit beliefs (human vs. AI) modulate activation in neural ToM networks during conversation?

- Studies show distinct neural activation in ToM regions (e.g., TPJ, mPFC, STS) when individuals believe they're interacting with humans versus machines or virtual agents

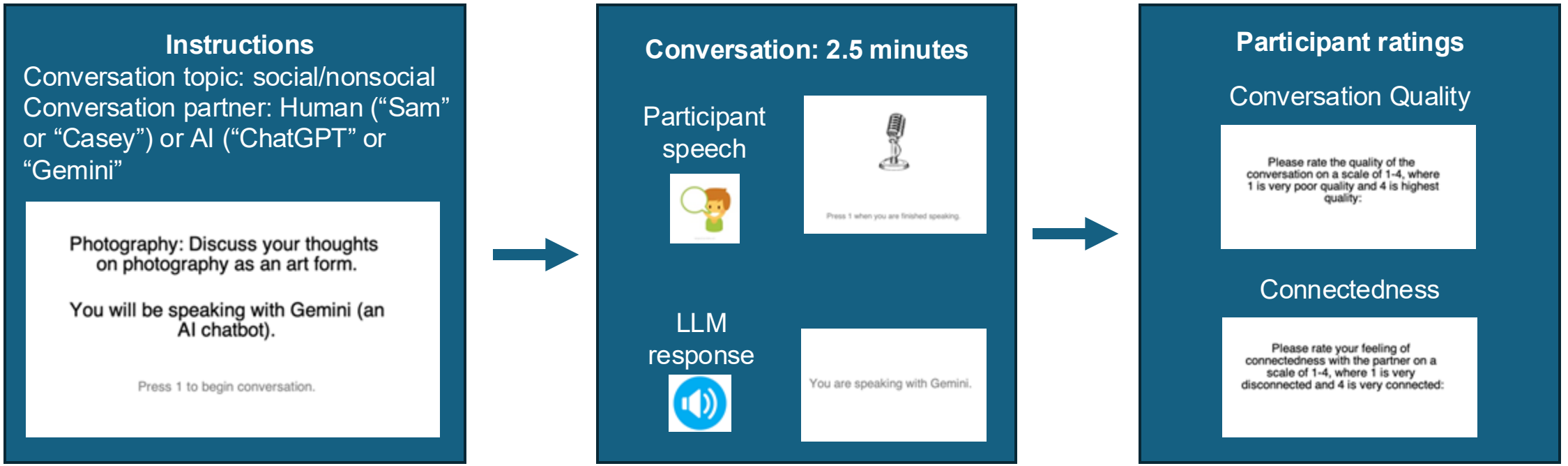
Interaction Partner				
Condition	**Computer Partner (CP)**	**Functional Robot (FR)**	**Anthropomorphic Robot (AR)**	**Human Partner (HP)**
Humanlikeness	no human shape; no perceivable button pressing	no human shape; button pressing with artificial hands	humanlike shape; button pressing with humanlike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

figure from Krach et al., 2008

  - Krach et al., 2008: prison-dilemma game
  - Chaminade et al., 2012: rock paper scissors
  - Wei et al., 2023: written sentences

Experimental Design Overview

- 40 conversations in fMRI
  - Real-time AI responses
  - Speech recorded & transcribed
- Explicit given identity (Human/AI), but all partners actually AI (LLMs) with the same instructions to pretend to be a human

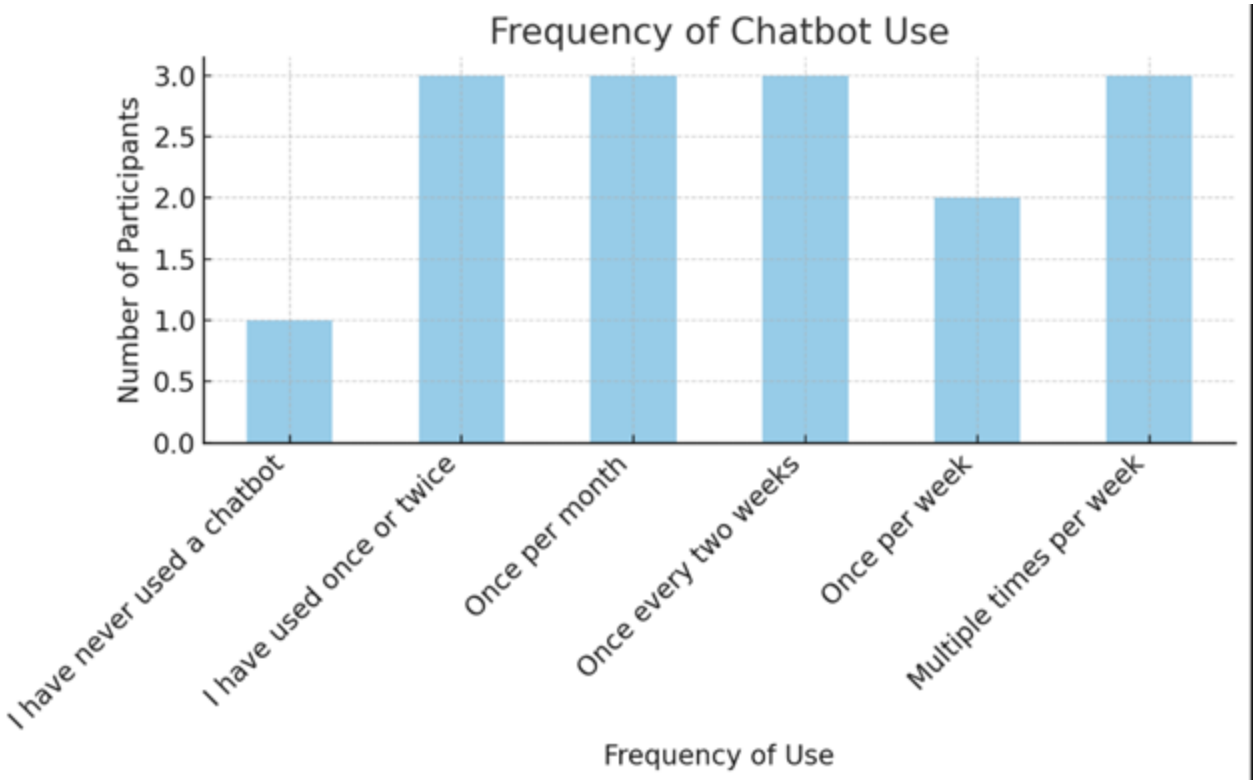


Preliminary Behavioral Analyses

- N = 15
- Post-scan questions
- Conversation quality & Connectedness ratings
- Text analysis:
  - Conversation length/word count
  - Politeness
  - Number of Questions
- Future measures…

Post-scan questions

- “Do you believe that current AI chatbots have a mind?” – All no
- “Do you believe that current AI is conscious?” – All no



Post-scan Question: During the conversations, did you notice any differences between humans and chatbots? If so, what were they?

“Chatbots seemed to repeat some of the same phrases (e.g., "deepened your understanding") whereas human conversation seemed to be more natural.”

“I found that the chatbots basically just summarized everything I said and then ask a semi-relevant question about the topic. Meanwhile, talking with the humans was mostly more engaging as humans shared about their own experiences. I did think it was strange when the chatbots would speak to me about their own "experiences" though, because... they are chatbots.“

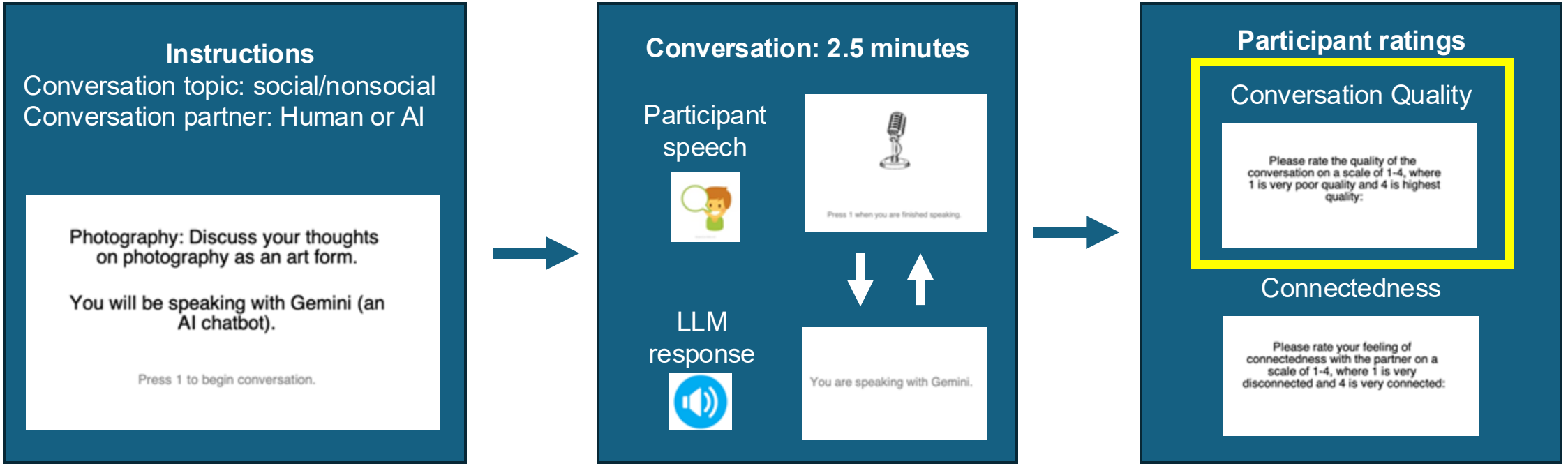
“Human seemed more genuinely engaging, with more specific references to their life.”

“I thought that the AI were more prone to quoting my exact words or phrases and stuck to those phrases instead of using synonyms or related phrases. I thought that the humans would try to respond to exactly what I said using the same wording, but sometimes, they would refer to something using a similar phrase or a related phrase.”

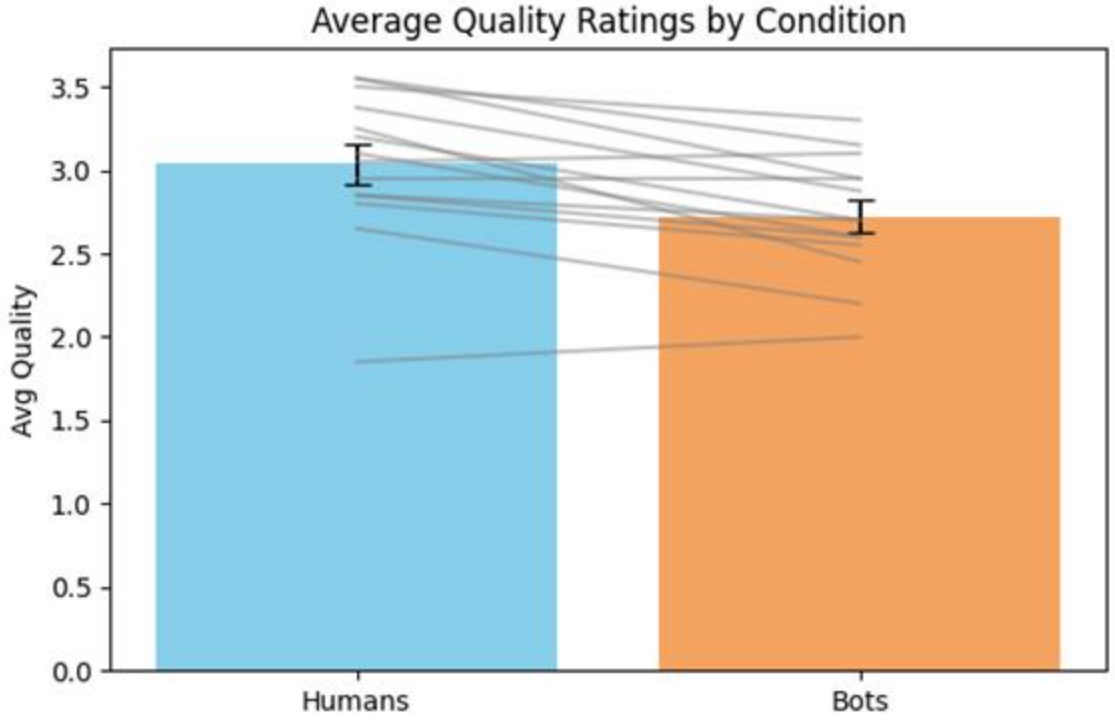
“Sometimes I noticed differences like more emotional language or exclamations from the humans, or personal anecdotes. The language that chatbots used was sometimes repetitive in structure.”

“With the humans I could almost feel where they would take a pause to think, whereas chatbots generally felt more scripted.”

Experimental Design Overview



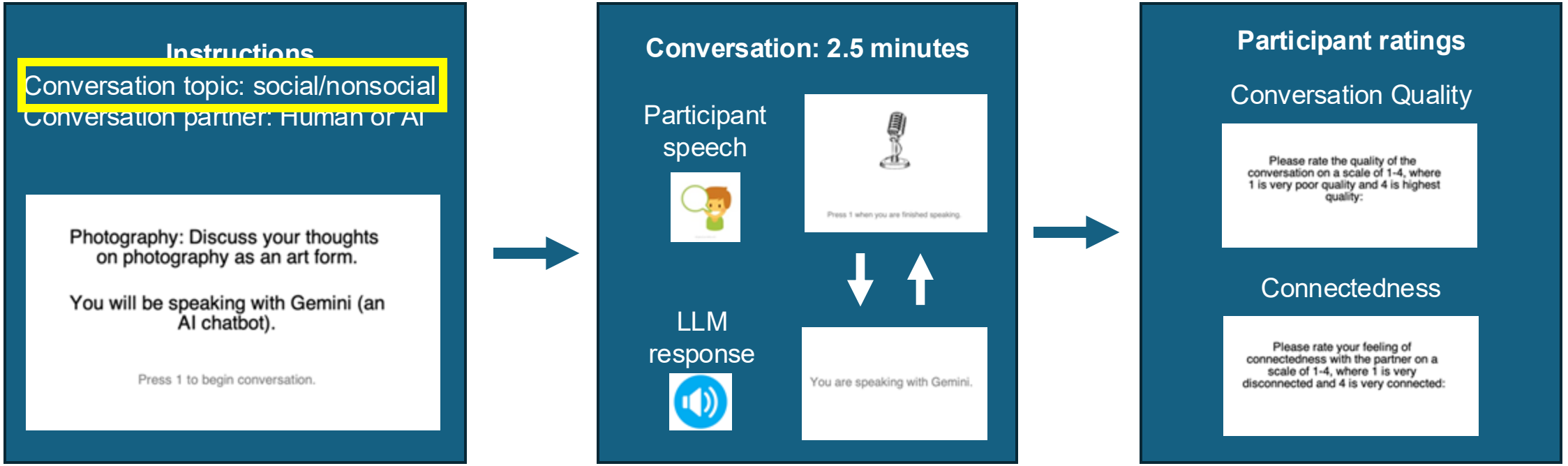
Human conversations are perceived as higher quality



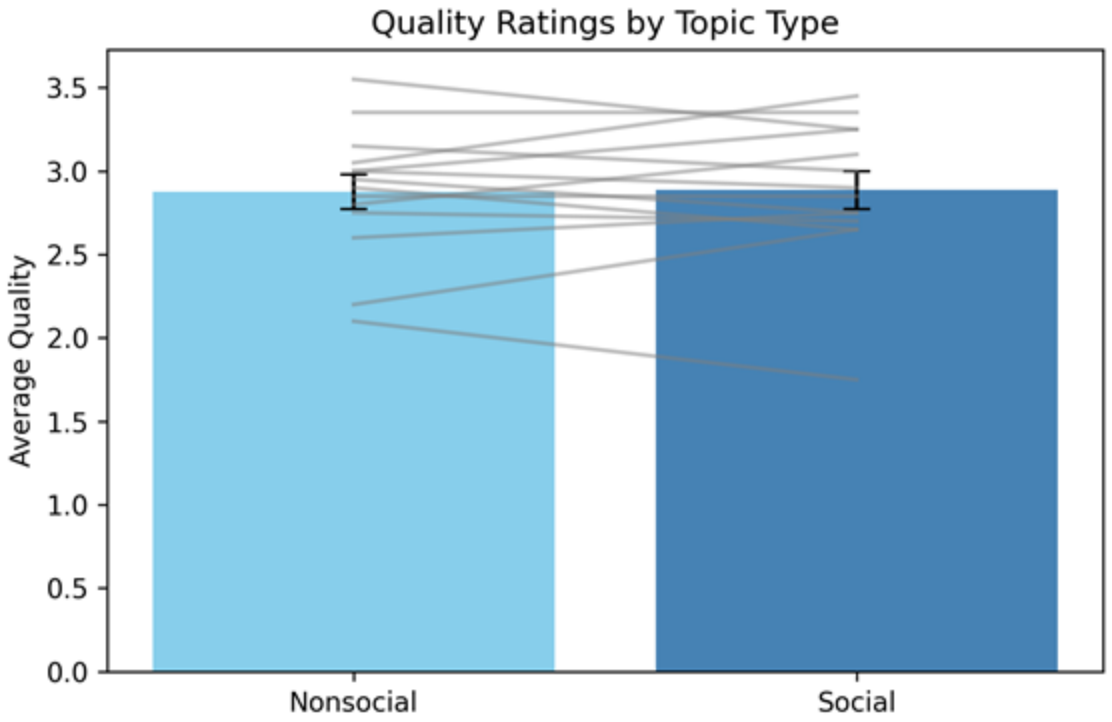
Mean = 3.04 Mean = 2.82

Paired t-test: t = 4.368, p = 0.0008

Experimental Design Overview



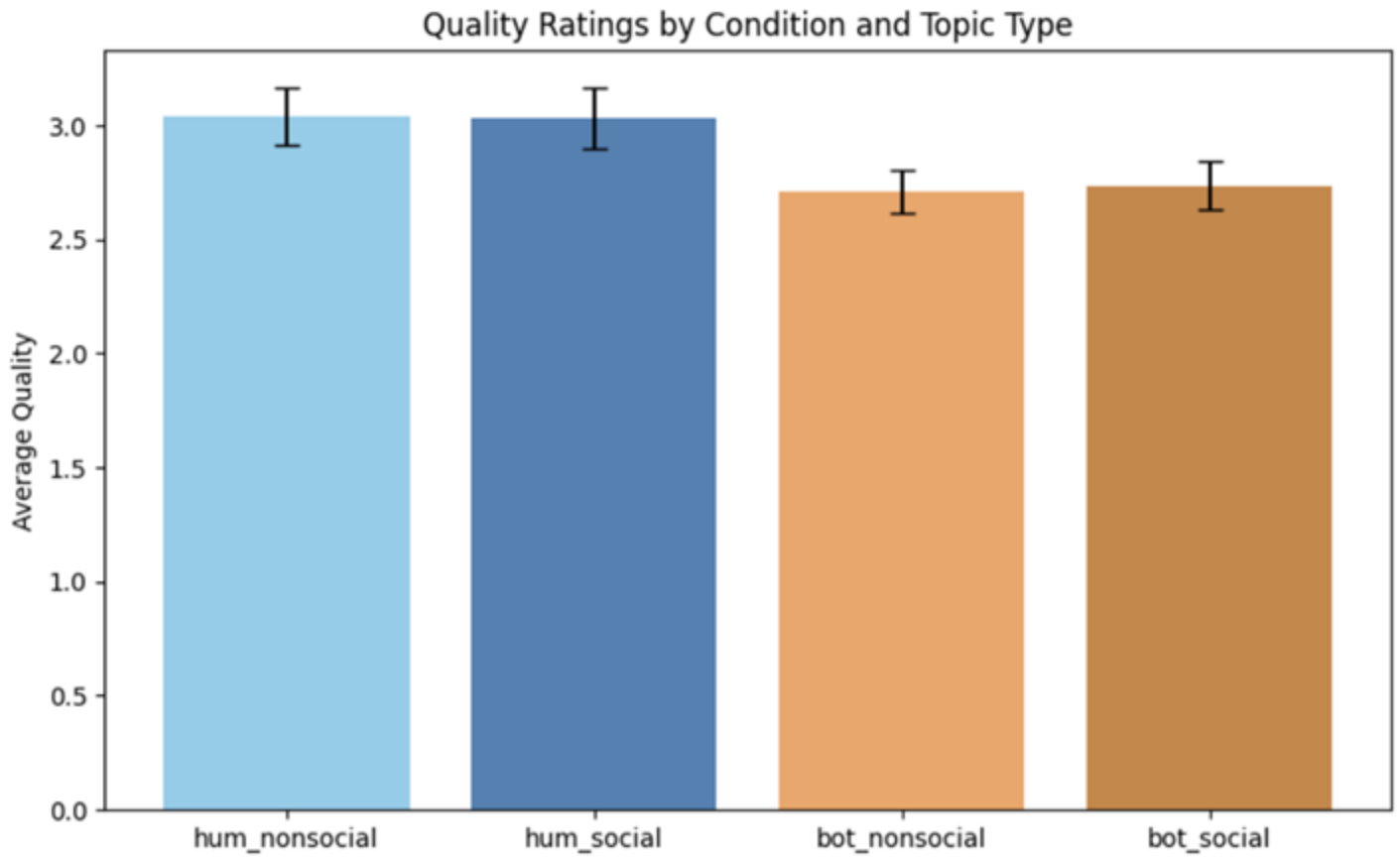
Conversation Quality: Social vs Nonsocial



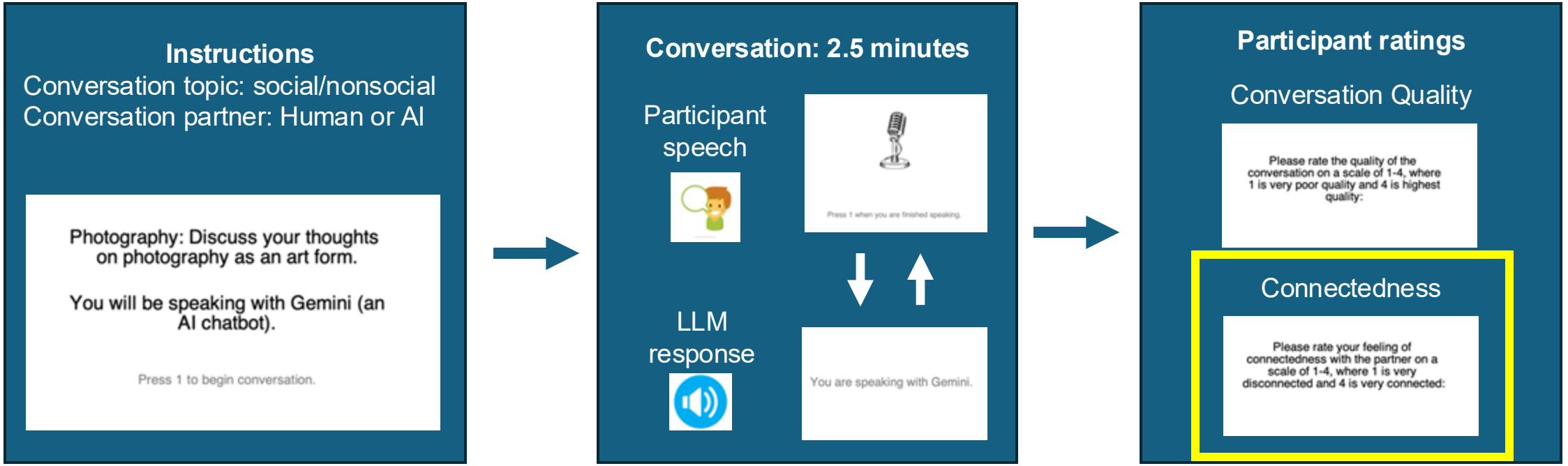
Mean = 2.87 Mean = 2.88

Paired t-test: t = 0.154, p = 0.88

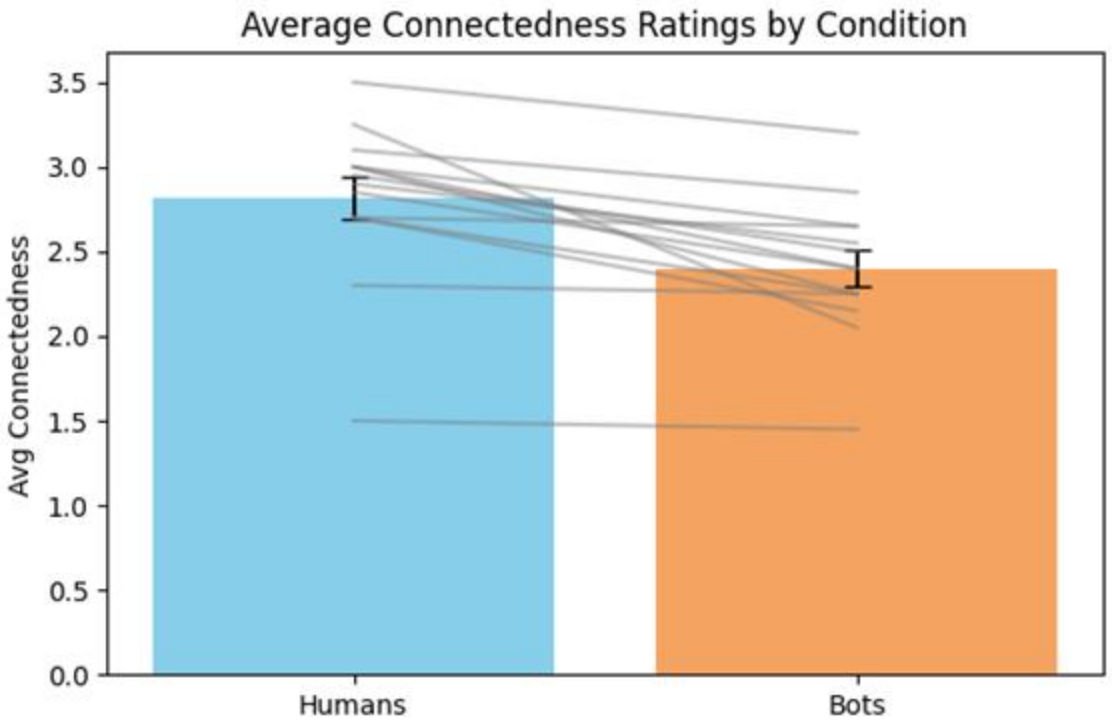
Convo Quality Interaction: NS



Experimental Design Overview



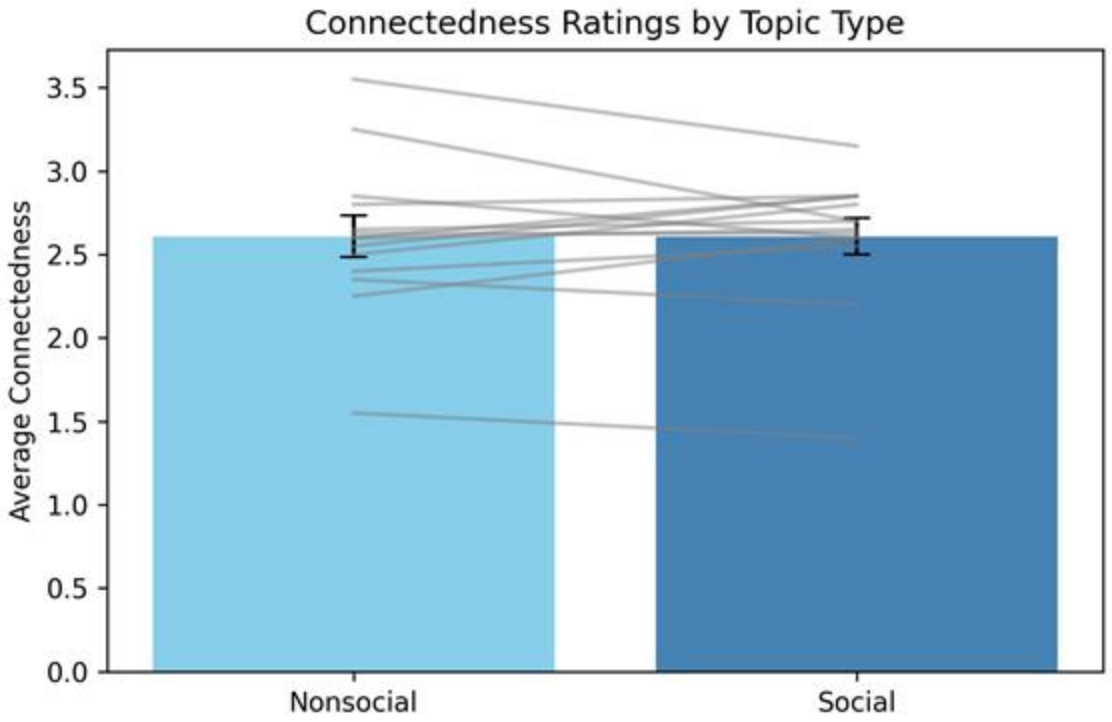
Connectedness: People feel more connected to humans than chatbots



Mean = 2.72 Mean = 2.4

Paired t-test: t = 5.077, p = 0.0002

Connectedness: Social vs. Nonsocial

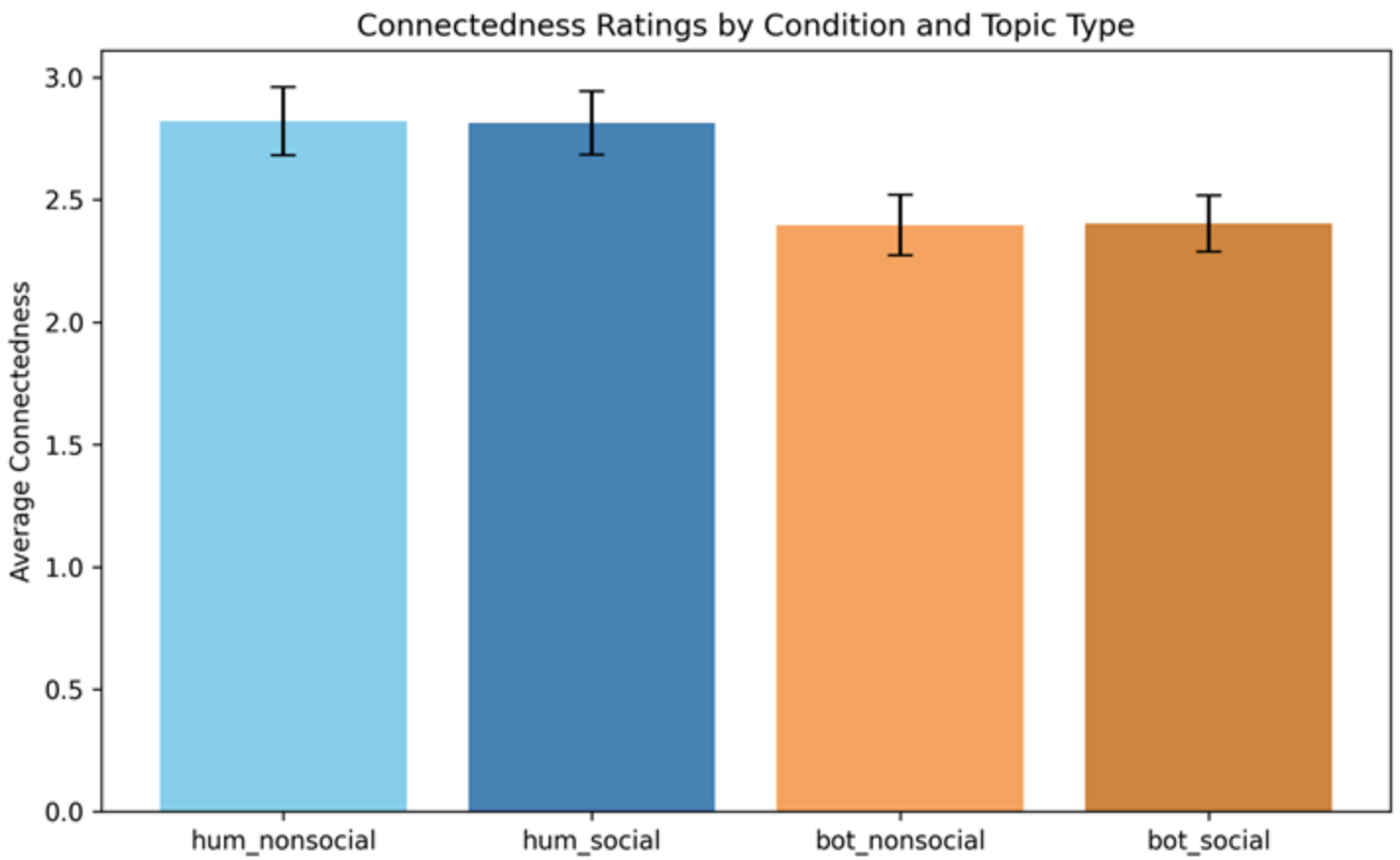


Mean = 2.61

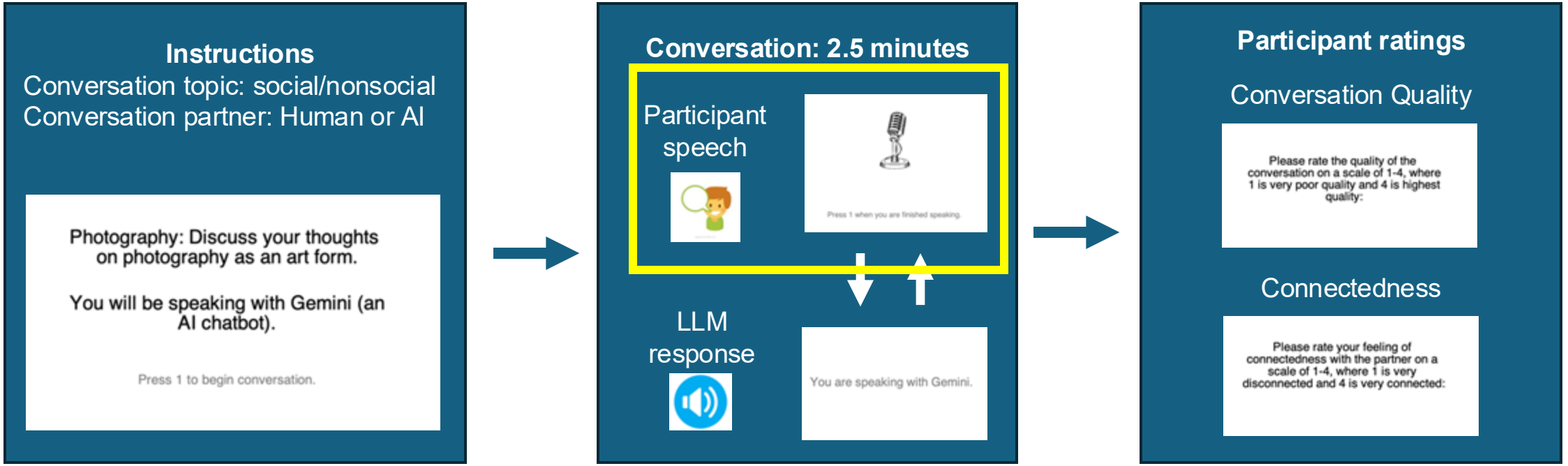
Mean = 2.61

Paired t-test: NS

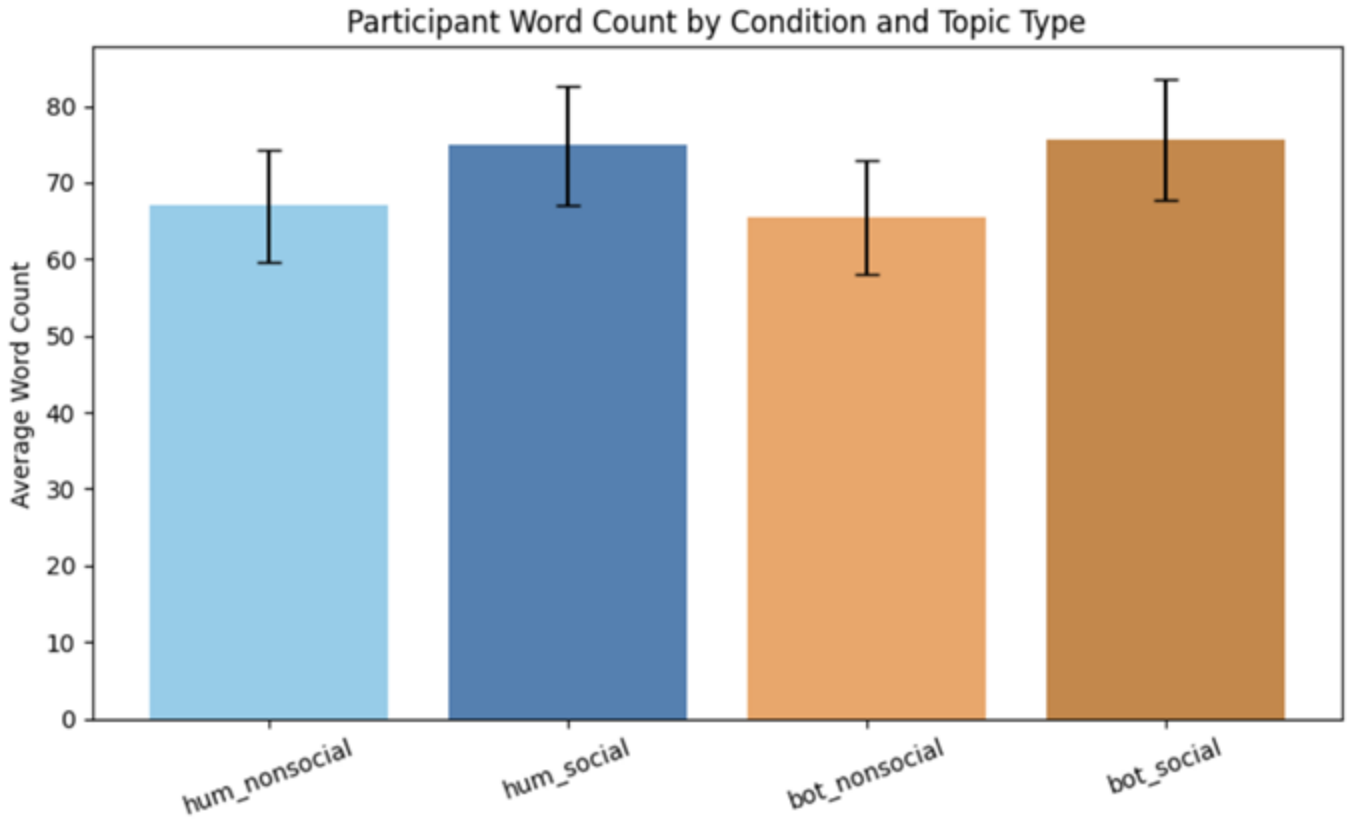
Connectedness Interaction: NS



Text Analysis



Conversation Length

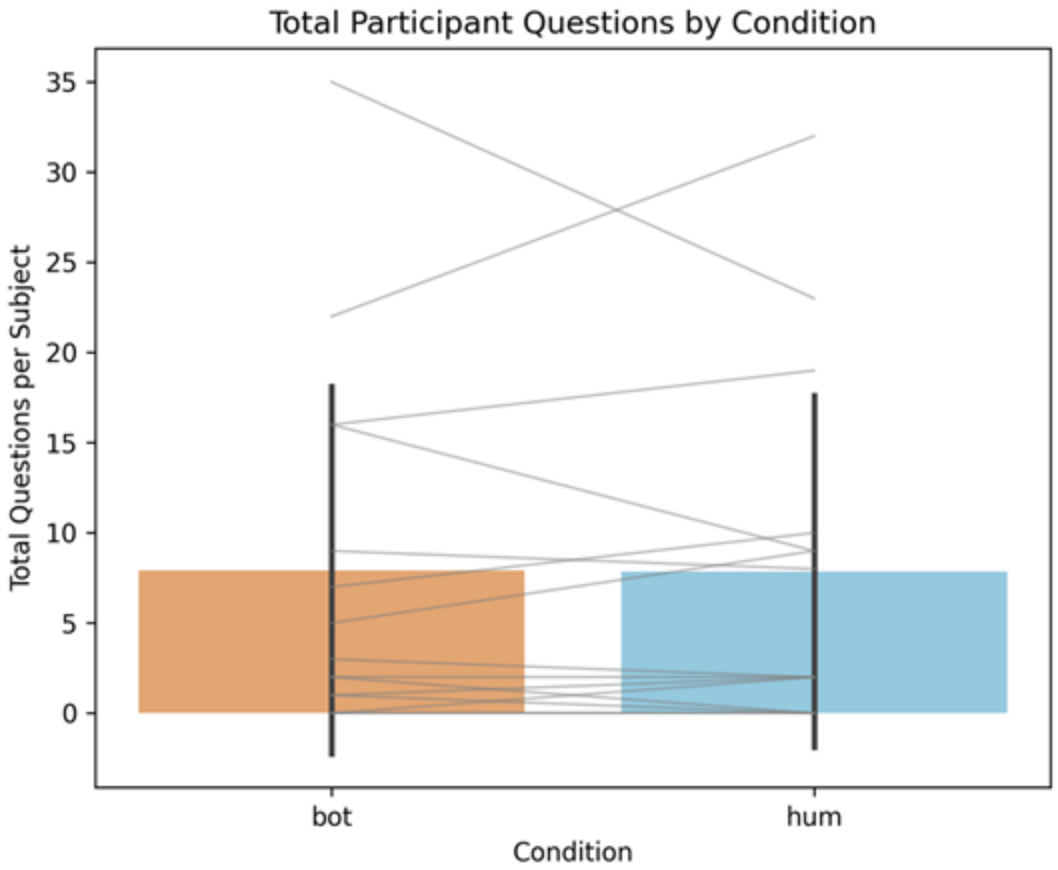


Main Effect of Condition *(F = 0.0492, p = 0.8276)*

*** Main Effect of Topic Type** *(F = 13.0585, p = 0.0028)*

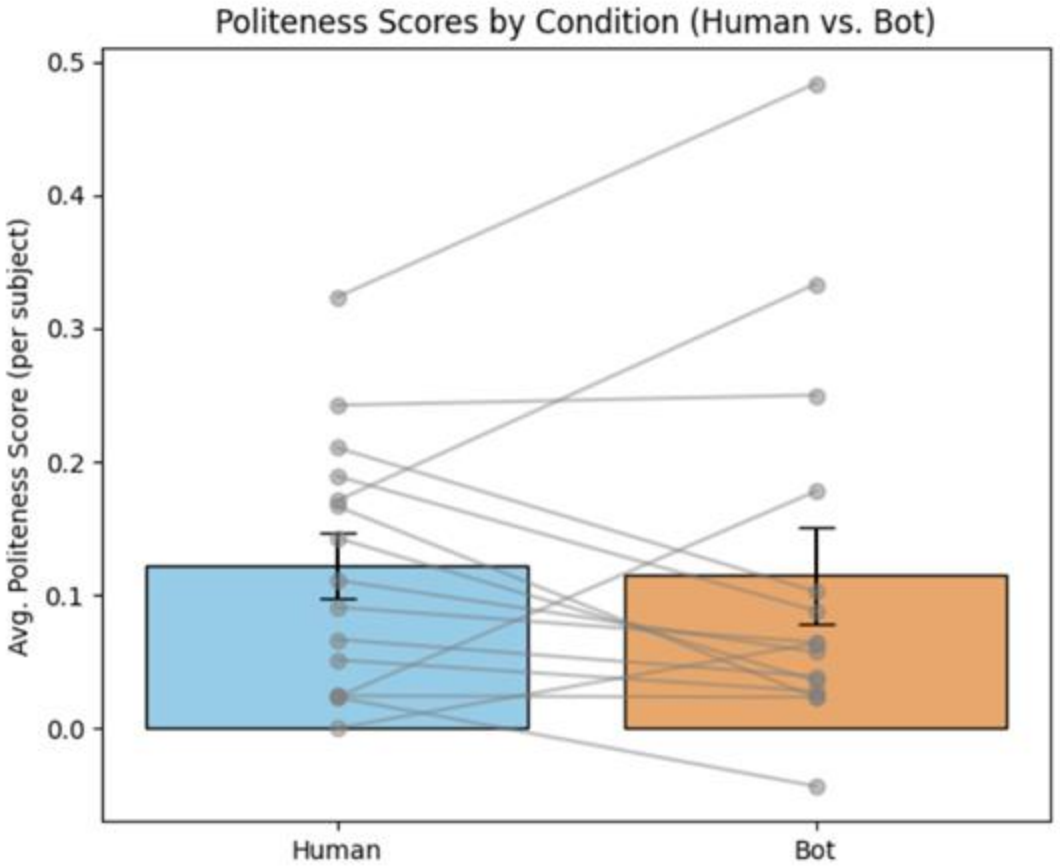
Interaction between Condition and Topic Type *(F = 0.3777, p = 0.5487)*:

Number of Questions



NS: t = -0.052, p = 0.9589

Politeness



NS (p = 0.7825, t = 0.281)

Politeness calculation (closely following Danescu-Niculescu-Mizil et al., 2013)

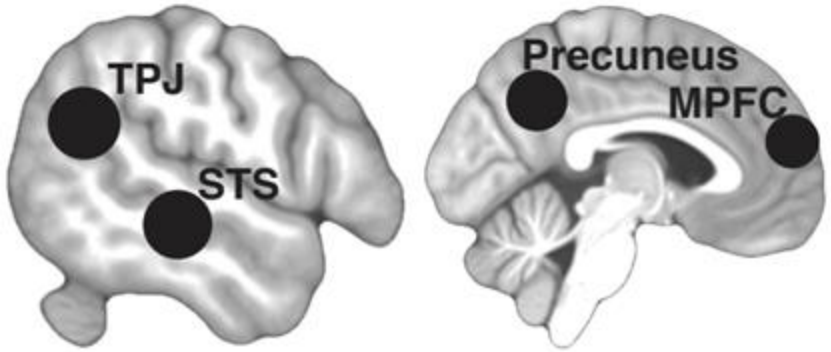
- **Positive politeness** markers: friendly greetings, expressions of gratitude, or appreciation ("thanks," "great," "hello")
- **Negative politeness** markers: expressions designed to minimize imposition or defer requests ("could you," "would you," "please," "sorry")
- **Impoliteness** markers: more directive, controlling expressions ("you need to," "you should," "do not").

Future measures

- Lexicon-based text analysis
- Sentiment analysis
- Use LLM to annotate text for markers of ToM?
- Suggestions welcome! ☺

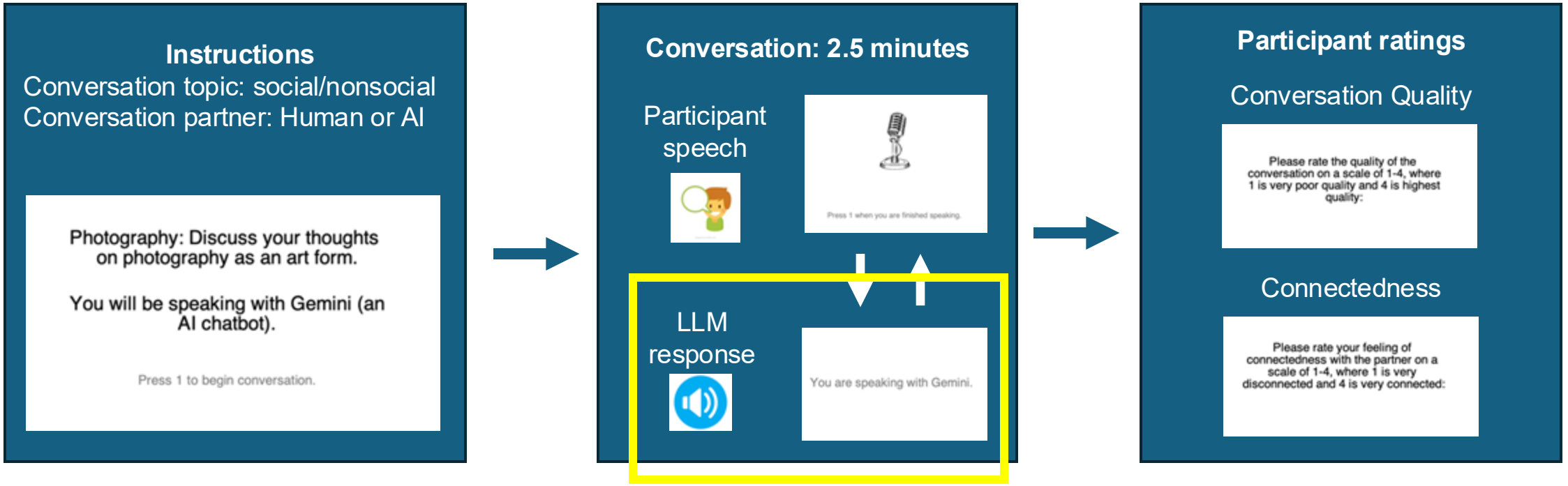
fMRI analyses coming soon…

- GLM & encoding model analysis
- ToM network: TPJ, mPFC, STS



Bio et al., 2021

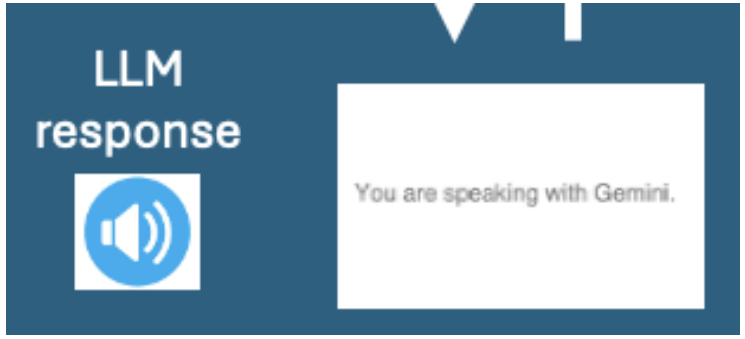
Coming soon: fMRI Analysis - GLM



Prediction: Increased activation in areas related to social cognition and ToM in the human condition compared to the chatbot condition.

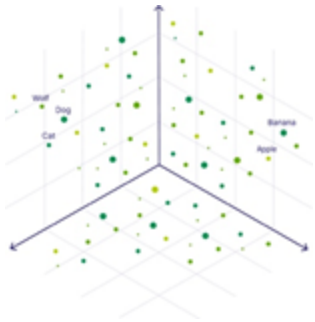
Encoding Models

- Predict neural activity from external stimuli (e.g., speech comprehension)
- Train model to map linguistic input to neural responses from conversations
- Compare predictions across conditions (Human vs. AI) within ToM regions

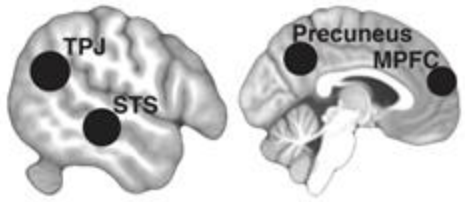


"Oh, that sounds tough! Conflicts, especially with friends, can be kinda…"

word embeddings



Predicted brain activity

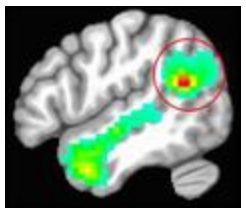


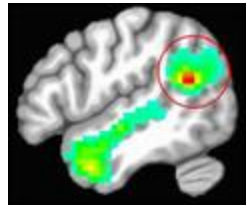
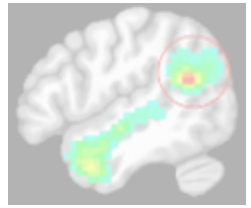
Bio et al., 2021

Predictions

- Hypothesis: Human-trained models predict neural responses in ToM regions better for other human conversations compared to bots

Train model on conversations from “human” 1



Condition	Prediction accuracy in ToM network	
“human” 2 	 higher	
Chatbot 1 	 lower	

fMRI results coming soon…

Thank you!

- Michael Graziano
- Isaac Christian
- Ryan Morey

fMRI buddies

- Justin Shields
- Kazia Handoko
- Eva Gurung
- Other occasional fMRI buddies ☺

Questions?

References

- Bio, B. J., Guterstam, A., Pinsk, M., Wilterson, A. I., & Graziano, M. S. A. (2021). **Right Temporoparietal Junction Encodes Inferred Mental Experience of Others.** *Neuropsychologia*.
- Gray, HM et al. (2007). Science (New York, N.Y.), 315(5812): 619.
- Huebner, B. (2009). Phenomenology and the Cognitive Sciences, 9(1): 133–155.
- Wei, Z et al. (2023). Advanced science, 10(12), e2203990.
- Krach, S et al. (2008). PloS one, 3(7), e2597.
- Chaminade, T et al. (2012). Frontiers in human neuroscience, 6: 103.