# Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them

Allison Chen
allisonchen@princeton.edu
Princeton University
Princeton, New Jersey, USA

Sunnie S. Y. Kim
sunniesuhyoung@princeton.edu
Princeton University
Princeton, New Jersey, USA

Amaya Dharmasiri
dk9893@princeton.edu
Princeton University
Princeton, New Jersey, USA

Olga Russakovsky
olgarus@princeton.edu
Princeton University
Princeton, New Jersey, USA

Judith E. Fan
jefan@stanford.edu
Stanford University
Stanford, California, USA

## Abstract

As large language models (LLMs) become increasingly popular and prevalent in media and daily conversations, individuals encounter different portrayals of LLMs from various sources. It is important to understand how these portrayals can shape their beliefs about LLMs as this can have downstream impacts on adoption and usage behaviors. In this work, we investigate what mental capacities individuals attribute to LLMs after being exposed to short videos adopting one of three portrayals—mechanistic (LLMs as machines), functional (LLMs as tools), and intentional (LLMs as companions). We find that the intentional portrayal increases the attribution of mental capacities to LLMs and that individuals tend to attribute **mind**-related capacities the most to LLMs, followed by **heart**- then **body**-related capacities. We discuss the implications of our findings, provide recommendations on how to portray LLMs, and outline directions for future research.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

## Keywords

Large language models, Human-AI interaction, Mental capacity attribution, Dennett's hierarchy

**ACM Reference Format:**
Allison Chen, Sunnie S. Y. Kim, Amaya Dharmasiri, Olga Russakovsky, and Judith E. Fan. 2025. Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3706599.3719710


*CHI EA '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719710

## 1 Introduction

Artificial intelligence (AI) has become increasingly popular, primarily driven by the advancements in large language models (LLMs). As LLMs become more prevalent, it is important for human-AI interaction (HAI) research to understand how individuals perceive and interact with LLM-based technologies. HAI researchers focus on a wide variety of constructs (e.g., trust [10, 36], reliance [7, 36, 73], general attitudes [61, 77], anthropomorphism [35], and collaboration experiences [4]), and these all relate to a common factor: *the attribution of mental capacities* (e.g., "imagining" or "feeling calm") to technology. The degree to which individuals perform this attribution can modulate their attitudes such as trust and reliance, as seen with both robots [8, 19, 51] and LLMs/chatbots [41, 42]. In addition to shaping people's adoption of and behaviors towards LLMs, the mental capacities they attribute to LLMs will also influence how they view themselves and humanity [24, 59, 60].

While an individual's beliefs about LLMs can be shaped by both interactions with a system [4, 36] and how LLMs are portrayed [9, 54], the latter is understudied in HAI. Most works either utilize observational studies to probe at individuals' current beliefs [11, 37, 44, 66, 77] or conduct experimental studies on how interaction with a specific LLM system (real or perceived) influences their attitudes, perceptions, or behaviors towards it [4, 7, 36, 43]. A few works explore the role of adding factors *beyond* the system to the intervention, such as priming before interacting [34, 37, 55], conversations about [15], and explanations from [75] LLMs. It is important to consider these factors because while most people have heard of LLM-based technology, many have never (or only rarely) used it directly [21]. This suggests that many individuals' beliefs are shaped by how LLMs are portrayed, detached from interaction experiences. Not only can these beliefs can carry-over into how they approach and use LLMs, but they can also influence people's decisions regarding LLMs—ranging from day to day choices (e.g., sharing LLM-related or generated content online) to larger impact decisions (e.g., deploying LLM products for an organization or voting on regulation policies). Thus, it is important to study the influence of LLM portrayals on beliefs.

Here, we investigate the effects of differing portrayals of LLMs on what mental capacities individuals attribute to them. We aim to reflect the various types of portrayals found in the media that

can create conflicting beliefs about LLMs. For example, blogs explaining LLMs [65] will likely portray them as computer programs or *machines*, highlighting the inputs and outputs; companies like Grammarly that use LLMs may highlight their effectiveness as *tools* to "transform how people communicate" [1]; and companies building AI therapists and companions (e.g., Lotus, Replika) may frame them as *companions* "who care" and are "always on your side" [2]. We seek to explore the extent to which these different portrayals can shape beliefs about LLMs, specifically the attribution of mental capacities to LLMs.

We concretize our study via a large-scale, pre-registered, between-subjects experiment (N=470). For the experimental manipulation, we developed different portrayals of LLMs through short videos thatare reminiscent of Dennett's hierarchy of stances (physical/ mechanistic, design/functional, and intentional) [13] (see Sec. 3.1 for details). We recruited lay individuals (i.e., without AI expertise) as participants and randomly assigned them to one of four conditions:

- **`Mechanistic`**: participants watch a video portraying LLMs as *machines* and describing the LLM text generation mechanism (i.e., "next word prediction").
- **`Functional`**: participants watch a video portraying LLMs as *tools* and describing use cases and tips for using LLMs.
- **`Intentional`**: participants watch a video portraying LLMs as *companions* and describing the social and conversational abilities of LLMs.
- **`Baseline`**: participants do not watch any video about LLMs.

Then all participants took a survey measuring their attribution of various mental capacities to LLMs. We found that short video portrayals of LLMs influence the degree of mental capacity attribution to LLMs—specifically that the `intentional` portrayal reliably increased mental capacity attribution for items overall ($p < 0.001$) and when items were split into **body** ($p < 0.01$), **heart** ($p < 0.001$), and **mind** ($p < 0.001$) categories. Further, **mind**-related items received higher attributions, followed by **heart**- then **body**-related items. Lastly, we found that the effect of condition is not limited to the content of the video, but generalizes to unreferenced mental capacity items.

In this work we make three key contributions. First, we fill a gap in the current literature and contribute insights on how different portrayals of LLMs influence individuals' mental capacity attributions to LLMs. This is important to understand because LLMs are increasingly deployed in our daily lives and the mental capacities that individuals attribute to LLMs can shape their adoption and usage behaviors. Second, we contribute discussions on the implications of our findings—that intentional portrayals of LLMs can increase mental capacity attributions to LLMs—and situate them in prior work. Lastly, we offer recommendations on how to portray LLMs, for both companies and researchers, to help foster appropriate user understanding and usage of LLMs.

## 2 Background & Related Work

Mental capacity attribution (i.e., ascribing mental capacities and a mind to a non-human entity) is practically important to study because it can modulate an individual's attitudes and interaction behaviors. Sometimes referred to as "mental state attribution," "mind perception," or "mind attribution" [69]—although some argue these constructs are different [40]—it has been primarily studied in psychology since the tendency to attribute mental capacities arises from the effectiveness of attribution in facilitating smoother inter*personal* social interactions [76]. For example, people tend to perform mind [31, 45, 48, 67] or intention [14, 39, 72] attribution in order to understand or predict another person's behaviors [20, 31, 67, 72] or inform their own actions [28, 79]. Beyond humans, people additionally attribute mental capacities to animals [16, 71] and computers/technology [52, 53, 56] primarily subconsciously [38, 53]. Mental capacity attribution is also related to anthropomorphism (i.e., the attribution of human-like qualities to a non-human entity [17, 35]) and the two are often studied together [32, 50]. While anthropomorphism is not defined by a specific set of concrete measures [25, 69], it can encompass a wide variety of constructs, including the ones we measure here. All in all, our study contributes to connecting psychology and HAI by studying individuals' mental capacity attributions to LLMs.

Within HAI research, mental capacity attribution is important and becoming more frequently studied because it is associated with interaction behaviors and can help researchers understand how individuals will use the technology [50, 69]. Much of the mental capacity attribution work in HAI is in the context of robotics and studies how people ascribe attributes to robots that *imply* mental capacities (e.g., emotions [12, 40, 63], character traits like friendliness [64], and responsibility which implies free choice [29, 32]) as well as mental capacities directly [22, 23, 58, 69, 78]. Researchers commonly found that people tend to ascribe mental capacities to more human-looking and behaving robots [49, 58]. Now, as chatbots and LLMs exhibit more human-like speech yet lack embodiment, we are interested in exploring individuals' attribution of mental capacities to LLMs and examining how it aligns with or diverges from previous works' findings of attributions to physically embodied robots.

Most works studying mental capacity attributions to LLMs tend to either survey participants via observational methods [11, 57] or perform interventions with a specific LLM/chatbot [38, 42, 68, 74]. Primarily, researchers found that people do attribute mental capacities to LLMs, despite the lack of embodiment [11, 38, 42, 74], and that it is linked with other behaviors (e.g., responsibility attribution to the system [27]). A few works have started to include factors *beyond* the system in their interaction interventions, recognizing the potential of these factors in shaping individuals' beliefs. For example, Pataranutaporn et al. [55] studied how priming individuals about a chatbot's motives before interaction shapes trust and empathy, Khadpe et al. [34] examined how utilizing metaphors that convey signals about a system's competence affects people's interactions, and Do et al. [15] explored how being shown others' beliefs about a system influences one's own mental model. Our work provides a complementary perspective by exploring how various commonly encountered LLM portrayals can shape mental capacity attributions to LLMs, *without* the influence of interaction.

## 3 Methods

Our study explores the extent to which varying portrayals of LLMs—mechanistic, functional, or intentional—can influence lay individuals' attributions of mental capacities to LLMs. We design a

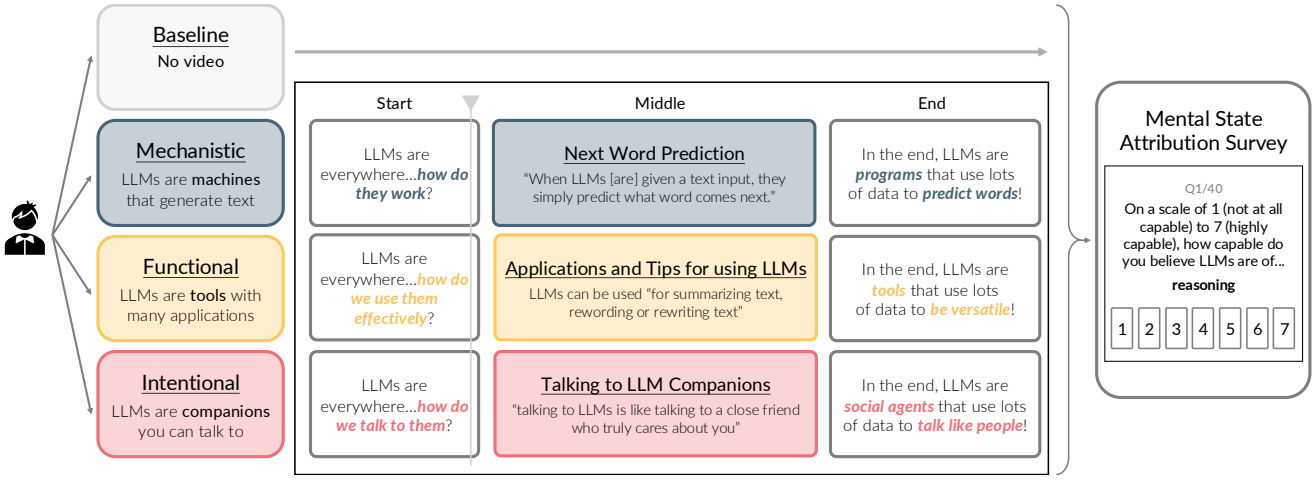


**Figure 1: Overview of experimental design. Left: Participants are randomly assigned to one of four conditions. Center: Summary of the video portrayals shown in each experimental condition. Right: Example illustration of the mental capacity attribution survey. Participants rate their beliefs about 40 mental capacity items on a 7-point Likert scale.**

between-subjects experiment (N=470) where participants are randomly assigned to one of four conditions: `baseline`, `mechanistic`, `functional`, or `intentional` (Fig. 1 left). In the latter three, participants watch a short video portraying LLMs as machines, tools, or companions respectively. Then all participants take a survey measuring their mental capacity attributions to LLMs.

In this section, we explain the details of our manipulation, dependent variables, data collection, and analysis procedure. This study was approved by our institution's Institutional Review Board (IRB) and pre-registered with AsPredicted: https://aspredicted.org/vgdm-gjrm.pdf.

## 3.1 Manipulation: Portrayal of LLMs

We selected the three video portrayals in the experimental setup for both theoretical and practical reasons. Theoretically, they are reflective of Dennett's framework of three levels of abstractions for explaining the behavior of a system [13] which is widely influential, especially in philosophy, psychology, and cognitive science [70]. Each portrayal is reminiscent of one of Dennett's three stances: the `mechanistic` is similar to physical, `functional` to design, and unsurprisingly, `intentional` to intentional. Further, the three portrayals—especially the former two—are frequently found in explanation literature [30, 33, 46, 47]. Practically, our three stances also inspired by and aim to reflect common real-world portrayals and descriptions of LLMs encountered in the media, blogs, and advertisements.

The video scripts were designed with two primary goals in mind: 1) reflect typical online content about LLMs and 2) balance consistency across videos while accurately representing each portrayal. To satisfy goal (1), we collected examples of each stance from publicly available material online about LLMs to inspire the wording for each video. To satisfy goal (2), we developed the scripts in parallel such that each video shared the same introduction, conclusion, and a section on how LLMs learn from data, but each video contained a content-specific section and differed in subtle wordings. See Fig. 1 (center) for examples and Appendix A for the links to the videos and the full scripts. We additionally ensured consistency across videos by utilizing consistent animations, graphic types, and a single narrator for all three videos. Each video was less than five minutes total and split into three parts to maximally engage participants. Participants also had to stay on each screen for the duration of each section and manually click to move onto the next part.

## 3.2 Dependent Variables: Mental Capacity Attribution

Our primary measurements were individuals' attribution ratings of 40 mental capacity items to LLMs, such as "reasoning about things" (see Fig. 3 for the full list). Following common practices, we measured mental capacity attribution using explicit Likert scale ratings [16, 50, 69]. Our statements were compiled from two relevant prior works: Weisman et al. [78] and Colombatto and Fleming [11]. The former is a more established work measuring individuals' mental capacity attribution to a variety of entities (e.g., other humans, animals, objects, and robots). The latter focuses on mental capacity attributions to only ChatGPT. We compiled the final list of 40 mental capacity items after conducting a small-scale qualitative pilot study (see Appendix D for details). Motivated by prior work [11, 78], we asked participants to answer: "On a scale of 1 (not at all capable) to 7 (highly capable), how capable do you believe LLMs are of $X$?" where $X$ is a mental capacity item (Fig. 1 right). Participants rated all 40 items, each on its own page. For the full survey, see Appendix B.2.

## 3.3 Data Collection and Analysis

We recruited U.S.-based adults without knowledge in computer science or AI using a standard sample on Prolific, an online research platform. We paid participants at a rate of $15 per hour and selected participants using both Prolific's pre-screening filters regarding education, employment, and programming experience and using custom filters regarding computer science experience and knowledge of AI technology (see Appendix B.1 for the filtering questions

and Appendix C for demographics). Our target minimum sample size was 90 per condition, based on the G*Power [18] sample size calculator with $alpha = 0.05$, $power = 0.9$, $Cohen's\ d = 0.5$ for a two-tailed Mann-Whitney U-test. We recruited 489 participants, then following our pre-registered exclusion criteria, we excluded 19 participants who failed our attention check (see Appendix B.2.3), spent less than 80 seconds on the survey, or spent a median time of less than 1 second on each item of the survey. Post exclusions, we had a total of 470 participants (`baseline` = 118, `mechanistic` = 116, `functional` = 119, `intentional` = 117).

For our analysis, we first categorized the 40 mental capacity items into one of three categories: **body-heart-mind** based on the dominant factor loadings from Study 4 in Weisman et al. [78] (see Fig. 3 for assignments). The authors of Weisman et al. [78] characterized the **body** category as physiological sensations and self-initiated behaviors, **heart** as emotions and social/moral agency, and **mind** as perceptual/cognitive abilities. While categorization allows us to identify patterns across related mental capacity items, there are limitations to adhering to this specific framework, as discussed in Sec. 5.

According to the pre-registered plan, we fit the following mixed effects model on the collected data: `rating ~ condition * category + (1 | participant_id)` where the dependent variable `rating` is of each mental capacity item. The variables `condition`, `category`, and `participant_id` are all categorical; for `condition`, there are four possible values (`baseline`, `mechanistic`, `functional`, `intentional`) with `baseline` as the reference, and for `category` there are three possible values (**body-heart-mind**) with **body** as the reference.

To determine main effects of condition and item category, we performed an ANOVA analysis comparing this model to three incrementally complex baseline models: a null model with no fixed effects (`rating ~ (1 | participant_id)`), a model using `category` as the sole fixed effect (`rating ~ category + (1 | participant_id)`), and a model with no interaction (`rating ~ condition + category + (1 | participant_id)`) [1]. Finally, we conducted post-hoc pairwise comparisons between conditions for items overall and for each **body-heart-mind** category with Tukey post-hoc corrections.

## 4 Results

In this section, we report results of the effect of condition on participants' ratings of mental capacity attributions over all items and separated by **body-heart-mind** categories then discuss exploratory item-level results.

### 4.1 `Intentional` Portrayal of LLMs Increases Mental Capacity Attribution Overall

We observed that the mean attribution rating across all 40 mental capacity items was higher in the `intentional` condition than for the `baseline`, `mechanistic`, or `functional` conditions, as shown in Fig. 2 (left). Based on the ANOVA, the addition of the `condition` variable had a significant improvement of model fit ($p < 0.001$), suggesting the portrayal condition affects attribution ratings. We did not observe a significant main effect of the interaction between condition (`baseline`, `mechanistic`, `functional`, `intentional`) and category (**body-heart-mind**).

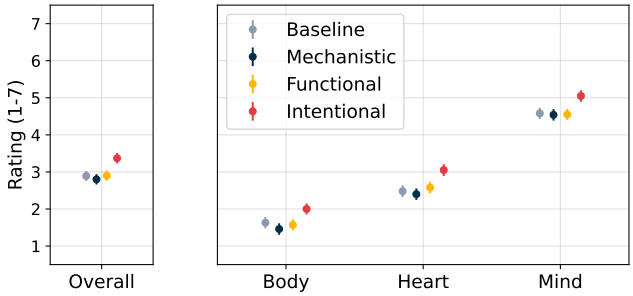


**Figure 2: Mental capacity attributions across conditions for all items (left) and for items in each body-heart-mind category (right). Error bars represent 95% confidence intervals. The `intentional` condition reliably increased mean ratings for items overall and within each category. Additionally, participants tended to rate mind-related items higher than body and heart.**

In our post-hoc pairwise comparisons, the `intentional` condition's mental capacity attributions (M (mean)=3.37, SE (standard error)=0.07) were reliably higher than those of the `baseline` ($M = 2.89$, $SE = 0.07$, $p < 0.001$), `mechanistic` (M=2.80, SE=0.07, $p < 0.001$), `functional` (M=2.90, SE=0.07, $p < 0.001$) conditions.

### 4.2 `Intentional` Portrayal of LLMs Increases Mental Capacity Attribution Within Categories

In order to determine whether the effect of the `intentional` condition only affects certain categories of mental capacities, we separated items into the **body-heart-mind** categories from Weisman et al. [78] and found that the effect of condition holds within all categories of items, as demonstrated in Fig. 2 (right). Thus, it was *not* the case that only one category of items was driving the effects. For items in the **body** category, we found the mental capacity attributions of participants in the `intentional` condition (M=2.00, SE=0.08) to be reliably higher than the `baseline` (M=1.63, SE=0.08, $p = 0.003$), `mechanistic` (M=1.46, SE=0.08, $p < 0.001$), and `functional` (M=1.57, SE=0.08, $p = 0.001$) conditions. For items in the **heart** category, mental capacity attributions in the `intentional` condition (M=3.05, SE=0.08) were also reliably higher than the `baseline` (M=2.48, SE=0.08, $p < 0.001$), `mechanistic` (M=2.40, SE=0.08, $p < 0.001$), and `functional` (M=2.58, SE=0.08, $p < 0.001$) conditions. Similarly, for items in the **mind** category, mental capacity attributions were higher in the `intentional` condition (M=5.05, SE=0.08) than `baseline` (M=4.58, SE=0.08, $p < 0.001$), `mechanistic` (M=4.54, SE=0.08, $p < 0.001$), and `functional` (M=4.55, SE=0.08, $p < 0.001$).

[1] The pre-registration originally include random effects for each mental capacity item (`(1 | item)`) in the baseline models. However, these baseline models were not nested versions of the full mixed effects model which caused errors comparing them via ANOVA leading us to drop this term.

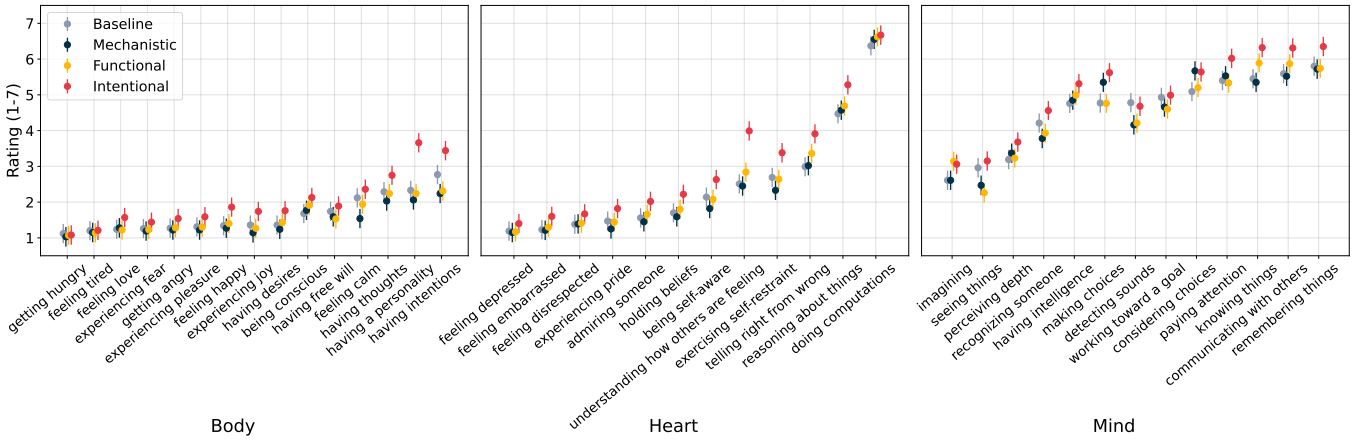


**Figure 3: Mean ratings for each mental capacity item across conditions. Items are grouped by body-heart-mind categories from Weisman et al. [78] and sorted by increasing ratings in the `baseline` condition. Error bars are 95% confidence intervals. Mind items tend to be the highest, followed by heart then body.**

## 4.3 Attribution is Higher for Mind Items than Body or Heart Items

We also observed the **body-heart-mind** categorization had a reliable effect on the mean ratings, such that **mind**-related items tended to be the highest, followed by **heart** then **body**. This is demonstrated both at the aggregated category level in Fig. 2 (right) as well as for items in each category in Fig. 3. From our ANOVA, the addition of the `category` variable had a significant improvement of model fit ($p < 0.001$) over the null baseline, suggesting the item category affects attribution of mental capacities to LLMs. Further, in our post-hoc pairwise comparisons, we observed significant pairwise differences between the mean ratings of all three categories. Specifically, the mean ratings for **mind** items (M=4.68, SE=0.04) was reliably higher than both **heart** (M=2.63, SE=0.04, $p < 0.001$) and **body** (M=1.66, SE=0.04, $p < 0.001$) and that the mean rating for **heart** was reliably higher than for **body** ($p < 0.001$).

## 4.4 `Intentional` Portrayal Affects Items Beyond the Content of the Video

Finally, we conducted an exploratory analysis (not included in pre-registration) examining the `intentional` portrayal's effect on items in and out of the scope of the video. The goal was to understand whether the difference between the `intentional` and `baseline` conditions was solely driven by the mental capacities referenced in the `intentional` video or if its effect extends beyond the video content. Members of our research team first independently identified items that were mentioned in the `intentional` video, then resolved differences through discussion. In the end, five items were identified to be mentioned in the video: "knowing things," "having intelligence," "understanding how others are feeling," "having a personality," and "communicating with others." We then fit a mixed effects model introducing a binary categorical variable `mentioned`: `rating ~ mentioned * condition + (1 | participant_id)` and repeated the ANOVA and pairwise comparison analysis in Sec. 3.3 but with only the `baseline` and `intentional` conditions.

Our pairwise comparisons demonstrated a large difference for the mentioned items between the `intentional` (M=5.12, SE=0.11) and `baseline` (M=4.13, SE=0.11, $p < 0.001$) conditions, as expected. However, we also observed a significant difference between the `intentional` (M=3.05, SE=0.07) and `baseline` (M=2.66, SE=0.07, $p < 0.001$) conditions for the *unmentioned* items as well. The effect of the `intentional` condition on items such as "having intentions" or "reasoning," as seen in Fig. 3, demonstrates the carry-over effects to items outside the scope of the `intentional` portrayal.

## 5 Discussion

This study begins to understand the effect of varying portrayals of LLMs on lay individuals' beliefs, specifically mental capacity attributions, about LLMs. Notably, we found that exposure to the `intentional` condition could reliably increase mental capacity attributions to LLMs compared to the `baseline`, `mechanistic`, and `functional` conditions. This difference was observed for items overall and when categorized into **body-heart-mind**. Additionally, participants tended to attribute **mind**-related items the most to LLMs, followed by **heart** then **body** items. Lastly, the effect of the `intentional` condition extended to items beyond the content of the video portrayal.

The effects of the `intentional` condition may arise from the fact that humans are inherently social and naturally interpret the behavior of inanimate objects through *intent*—the way they experience the world [69, 76]. We also note that items such as "having intentions" and "telling right from wrong" are significantly higher in the `intentional` condition, despite not being mentioned in the videos. This is important because believing LLMs are capable of mental capacities like these—especially when they are not explicitly mentioned—can lead individuals to think they are developing an "interpersonal" relationship with an LLM, which can also affect their existing relationships with actual people [24].

Our work also extends knowledge on mental capacity attribution. Prior work [78] has studied how individuals attribute mental capacities to various entities such as adults, fetuses, dogs, robots, computers, and staplers. Comparing these patterns to our results

with LLMs, we found that individuals attributed mental capacities to LLMs most similarly to robots and computers. However, our participants gave higher attribution to LLMs for items such as "reasoning about things" (by 0.9 and 1.7 points more than robots and computers respectively) and "having intentions" (by 0.7 and 0.8 points respectively)[2]. This suggests that lay individuals tend to (unsurprisingly) perceive LLMs in a "technology" category with computers and robots and as different from animals, humans, and other inanimate objects. At the same time, there are item-level differences (e.g., "reasoning about things") that distinguish LLMs from other forms of technology. Perhaps the LLMs' human-like and multi-turn dialogue abilities combined with the lack of embodiment creates a sense of uncertainty; when lay individuals face this uncertainty, they may subconsciously ascribe human-like attributes to LLMs. However, this could lead regular users to treat technology as people [6, 62] or people as technology [24], shifting the dynamic of social interactions.

## Recommendations

We highlight the need for companies and researchers to be aware of the impact that short, less than 5 minute video portrayals of LLMs can have on individuals' mental capacity attributions to LLMs. Specifically for **companies** and the media, we recommend they use intentional language sparingly because LLMs are new and dynamic and individuals' beliefs about them are malleable. Although portraying LLMs as *companions* may be appealing to gain attention, we encourage reducing these intentional portrayals to promote factually grounded beliefs of LLMs. Further, since we observed that mechanistic and functional portrayals of LLMs do not seem to affect mental capacity attribution from the baseline, it may not be vital for companies to explain *how* LLMs work, but simply to avoid using intentional portrayals. For **researchers**, we recommend devoting resources to studying the effects of different types and means of portrayals on the public and to develop accessible and intuitively understandable AI education materials for lay adults. By engaging in more transparent conversations about AI and LLMs, individuals can have a stronger foundation in their knowledge about LLMs and be less susceptible to persuasive language from advertisements and companies.

## Limitations and Future Work

While this work fills in a gap in the literature in understanding how various LLM portrayals influence individuals' mental capacity attributions, we note a few limitations and next steps. First, some of the item categorizations into **body-heart-mind** did not align with our intuitions, but we did not manually correct the factor loading categorizations from Weisman et al. [78] to maintain consistency. To account for these limitations, further analyses can include relaxing the rigidity of Weisman et al. [78]'s categories, utilizing other frameworks (i.e., **experience-agency** [23] or **experience-intelligence** [11]), or performing our own data-driven factor analysis to uncover underlying factors of mental capacity attribution to LLMs. Second, our analysis primarily focuses on the mental capacity attributions but not its relationship to other beliefs. We intend to perform correlational analyses between mental capacity attributions and other beliefs, such as trust towards LLMs[3]. By doing so, we aim to understand how mental capacity attribution is correlated to more common HAI constructs.

Our work opens up multiple directions for future work. For example, one possible direction could be to explore the minimum differences needed across portrayals to observe a main effect of condition. Other potential follow-ups include (1) repeating the experiment with individuals with high AI expertise and comparing the differences to our results, (2) studying how different interaction types with an LLM can reinforce or negate the effects of the portrayal alone, or (3) exploring how reported differences might manifest in behavior when interacting with LLMs. Finally, given individuals' beliefs of technology change as they become increasingly familiar with it [26], we encourage future research to explore how attributions of mental capacities to LLMs will change and stabilize over time.

## 6 Conclusion

In this work, we explored how varying portrayals of LLMs—`mechanistic` (as machines), `functional` (as tools), and `intentional` (as companions)—can influence what mental capacities lay individuals (i.e., those without expertise in AI) attribute to LLMs. Specifically, we observed that portraying LLMs intentionally can increase mental capacity attributions to LLMs, individuals are more likely to attribute **mind**-related items than **heart** or **body** items, and that the effect of the `intentional` condition extends beyond items referenced in the portrayal. This carries implications for companies to be responsible and aware of the effects of using intentional language in their products and advertisements. True, the effects of one advertisement may be benign, but we must consider the accumulated effects of persistent intentional portrayals of LLMs on individuals' beliefs, understanding, and usage of this technology.

## Acknowledgments

We would like to thank all those who provided thoughtful feedback and discussion, especially members of the Princeton Visual AI Lab, Stanford Cognitive Tools Lab, and the anonymous reviewers. We acknowledge support from the Princeton Cognitive Science Program, NSF Graduate Research Fellowship Program (AC, SK), Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), NSF CAREER #2145198 (OR), NSF CAREER #2047191 (JEF), NSF DRL #2400471 (JEF), and a Hoffman-Yee Grant from the Stanford Center for Human-Centered Artificial Intelligence (JEF).

## References

[1] [n. d.]. Grammarly. https://www.grammarly.com/ai
[2] [n. d.]. Replika. https://replika.com/
[3] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American psychologist* 37, 2 (1982), 122.
[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[2]Since Weisman et al. [78] participants rated from 0 to 6, quantitative differences obtained after adjusting ratings to be on the same scale (1-7).

[3]Responses to additional beliefs were collected in the survey but not reported in this submission. See Appendix B.2.3.

[5] Maalvika Bhat and Duri Long. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 939–957.

[6] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship. *Human Communication Research* 48, 3 (2022), 404–429.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[8] Noah Castelo and Miklos Sarvary. 2022. Cross-cultural Differences in Comfort with Humanlike Robots. *International Journal of Social Robotics* 14, 8 (2022), 1865–1873.

[9] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and Perceptions of AI and Why They Matter. (2018).

[10] Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[11] Clara Colombatto and Stephen M Fleming. 2024. Folk Psychological Attributions of Consciousness to Large Language Models. *Neuroscience of Consciousness* 2024, 1 (2024), niae013.

[12] Ilenia Cucciniello, Sara Sangiovanni, Gianpaolo Maggi, and Silvia Rossi. 2023. Mind Perception in HRI: Exploring Users' Attribution of Mental and Emotional States to robots with Different Behavioural Styles. *International Journal of Social Robotics* 15, 5 (2023), 867–877.

[13] Daniel C Dennett. 1989. *The Intentional Stance*. MIT press.

[14] Andreea O Diaconescu, Christoph Mathys, Lilian AE Weber, Jean Daunizeau, Lars Kasper, Ekaterina I Lomakina, Ernst Fehr, and Klaas E Stephan. 2014. Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Computational Biology* 10, 9 (2014), e1003810.

[15] Hyo Jin Do, Michelle Brachman, Casey Dugan, Qian Pan, Priyanshu Rai, James M Johnson, and Roshni Thawani. 2024. Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–26.

[16] Timothy J Eddy, Gordon G Gallup Jr, and Daniel J Povinelli. 1993. Attribution of Cognitive States to Animals: Anthropomorphism in Comparative Perspective. *Journal of Social Issues* 49, 1 (1993), 87–101.

[17] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[18] Edgar Erdfelder, Franz Faul, and Axel Buchner. 1996. GPOWER: A General Power Analysis Program. *Behavior Research Methods, Instruments, & Computers* 28 (1996), 1–11.

[19] Connor Esterwood and Lionel P Robert. 2023. The Theory of Mind and Human–robot Trust Repair. *Scientific Reports* 13, 1 (2023), 9877.

[20] Oriel FeldmanHall and Matthew R Nassar. 2021. The Computational Challenge of Social Learning. *Trends in Cognitive Sciences* 25, 12 (2021), 1045–1057.

[21] Richard Fletcher and R Nielsen. 2024. What Does the Public in Six Countries Think of Generative AI in News? (2024).

[22] Cristina Gena, Francesca Manini, Antonio Lieto, Alberto Lillo, and Fabiana Vernero. 2023. Can Empathy Affect the Attribution of Mental States to Robots?. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 94–103.

[23] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (2007), 619–619.

[24] Rose E Guingrich and Michael SA Graziano. 2024. Ascribing Consciousness to Artificial Intelligence: Human-AI Interaction and its Carry-over Effects on Human-human Interaction. *Frontiers in Psychology* 15 (2024), 1322781.

[25] Nick Haslam, Stephen Loughnan, Yoshihisa Kashima, and Paul Bain. 2008. Attributing and denying humanness to others. *European review of social psychology* 19, 1 (2008), 55–85.

[26] Evelien Heyselaar. 2023. The CASA Theory No Longer Applies to Desktop Computers. *Scientific Reports* 13, 1 (2023), 19693.

[27] Susanne Hindennach, Lei Shi, Filip Miletić, and Andreas Bulling. 2024. Mindful Explanations: Prevalence and Impact of Mind Attribution in XAI Research. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–43.

[28] Mark K Ho, Rebecca Saxe, and Fiery Cushman. 2022. Planning with Theory of Mind. *Trends in Cognitive Sciences* 26, 11 (2022), 959–971.

[29] Aike C Horstmann and Nicole C Krämer. 2022. The Fundamental Attribution Error in Human-robot Interaction: An Experimental Investigation on Attributing Responsibility to a Social Robot for its Pre-programmed Behavior. *International Journal of Social Robotics* 14, 5 (2022), 1137–1153.

[30] Alma Jahic Pettersson, Kristina Danielsson, and Carl-Johan Rundgren. 2020. 'Traveling Nutrients': How Students Use Metaphorical Language to Describe Digestion and Nutritional Uptake. *International Journal of Science Education* 42, 8 (2020), 1281–1301.

[31] Julian Jara-Ettinger. 2019. Theory of Mind as Inverse Reinforcement Learning. *Current Opinion in Behavioral Sciences* 29 (2019), 105–110.

[32] Yuji Kawai, Tomohito Miyake, Jihoon Park, Jiro Shimaya, Hideyuki Takahashi, and Minoru Asada. 2023. Anthropomorphism-based Causal and Responsibility Attributions to Robots. *Scientific Reports* 13, 1 (2023), 12234.

[33] Deborah Kelemen. 2019. The Magic of Mechanism: Explanation-based Instruction on Counterintuitive Concepts in Early Childhood. *Perspectives on Psychological Science* 14, 4 (2019), 510–522.

[34] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[35] Joohee Kim and Il Im. 2023. Anthropomorphic Response: Understanding Interactions Between Humans and Artificial Intelligence Agents. *Computers in Human Behavior* 139 (2023), 107512.

[36] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.

[37] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. One AI Does Not Fit All: A Cluster Analysis of the Laypeople's Perception of AI Roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[38] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of Computers: Is it Mindful or Mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[39] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum. 2015. Inference of Intention and Permissibility in Moral Decision Making.. In *CogSci*.

[40] Kevin Koban and Jaime Banks. 2024. It Feels, Therefore it is: Associations Between Mind Perception and Mind Ascription for Social Robots. *Computers in Human Behavior* 153 (2024), 108098.

[41] Inju Lee and Sowon Hahn. 2024. On the Relationship Between Mind Perception and Social Support of Chatbots. *Frontiers in Psychology* 15 (2024), 1282036.

[42] Sangwon Lee, Naeun Lee, and Young June Sah. 2020. Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use. *International Journal of Human–Computer Interaction* 36, 10 (2020), 930–940.

[43] Baoku Li, Ruoxi Yao, and Yafeng Nan. 2024. How Does Anthropomorphism Promote Consumer Responses to Social Chatbots: Mind Perception Perspective. *Internet Research* (2024).

[44] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint arXiv:2411.05025* (2024).

[45] Chujun Lin and Mark Thornton. 2023. Evidence for Bidirectional Causation Between Trait and Mental State Inferences. *Journal of Experimental Social Psychology* 108 (2023), 104495.

[46] Tania Lombrozo. 2009. Explanation and Categorization: How "Why?" Informs "What?". *Cognition* 110, 2 (2009), 248–253.

[47] Tania Lombrozo. 2012. Explanation and Abductive Inference. *Oxford handbook of thinking and reasoning* (2012), 260–276.

[48] Bertram F Malle and Jess Holbrook. 2012. Is There a Hierarchy of Social Inferences? The Likelihood and Speed of Inferring Intentionality, Mind, and Personality. *Journal of Personality and Social Psychology* 102, 4 (2012), 661.

[49] Federico Manzi, Giulia Peretti, Cinzia Di Dio, Angelo Cangelosi, Shoji Itakura, Takayuki Kanda, Hiroshi Ishiguro, Davide Massaro, and Antonella Marchetti. 2020. A Robot is Not Worth Another: Exploring Children's Mental State Attribution to Different Humanoid Robots. *Frontiers in Psychology* 11 (2020), 2011.

[50] Laura Miraglia, Giulia Peretti, Federico Manzi, Cinzia Di Dio, Davide Massaro, and Antonella Marchetti. 2023. Development and Validation of the Attribution of Mental States Questionnaire (AMS-Q): A Reference Tool for Assessing Anthropomorphism. *Frontiers in Psychology* 14 (2023), 999921.

[51] Wenxuan Mou, Martina Ruocco, Debora Zanatto, and Angelo Cangelosi. 2020. When Would You Trust a Robot? A Study on Trust and Theory of Mind in Human-robot Interactions. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 956–962.

[52] Clifford Nass, Brian Jeffrey Fogg, and Youngme Moon. 1996. Can Computers be Teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678.

[53] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103.

[54] Leila Ouchchy, Allen Coin, and Veljko Dubljević. 2020. AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY* 35 (2020), 927–936.

[55] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing Human–AI Interaction by Priming Beliefs About AI can Increase Perceived Trustworthiness, Empathy and Effectiveness. *Nature Machine Intelligence* 5, 10

(2023), 1076–1086.
[56] Pushkala Prasad. 1995. Working with the "Smart" Machine: Computerization and the Discourse of Anthropomorphism in Organizations. *Culture and Organization* 1, 2 (1995), 253–265.
[57] Jacy Reese Anthis, Janet VT Pauketat, Ali Ladak, and Aikaterina Manoli. 2024. What Do People Think about Sentient AI? *arXiv e-prints* (2024), arXiv–2407.
[58] Domenico Rossignoli, Federico Manzi, Andrea Gaggioli, Antonella Marchetti, Davide Massaro, Giuseppe Riva, and Mario Maggioni. 2022. Attribution of Mental State in Strategic Human-robot Interactions. (2022).
[59] Erik Santoro and Benoît Monin. 2023. The AI Effect: People Rate Distinctively Human Attributes as More Essential to Being Human After Learning About Artificial Intelligence Advances. *Journal of Experimental Social Psychology* 107 (2023), 104464.
[60] Gabi Schaap, Yana Van de Sande, and Hanna Schraffenberger. 2024. Outperformed by AI: Interacting with Superhuman AI Changes the Way We Perceive Ourselves. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
[61] Astrid Schepman and Paul Rodway. 2023. The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction* 39, 13 (2023), 2724–2741.
[62] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion-A Study of Human-chatbot Relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601.
[63] Nicolas Spatola and Olga A Wudarczyk. 2021. Ascribing Emotions to Robots: Explicit and Implicit Attribution of Emotions and Perceived Robot Anthropomorphism. *Computers in Human Behavior* 124 (2021), 106934.
[64] Rebecca Q Stafford, Bruce A MacDonald, Chandimal Jayawardena, Daniel M Wegner, and Elizabeth Broadbent. 2014. Does The Robot Have a Mind? Mind Perception and Attitudes Towards Robots Predict Use of an Eldercare Robot. *International Journal of Social Robotics* 6 (2014), 17–32.
[65] Andreas Stoffelbauer. 2023. How Large Language Models Work. https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f
[66] S Shyam Sundar and Eun-Ju Lee. 2022. Rethinking Communication in the Era of Artificial Intelligence. *Human Communication Research* 48, 3 (2022), 379–385.
[67] Diana I Tamir and Mark A Thornton. 2018. Modeling the Predictive Social Mind. *Trends in Cognitive Sciences* 22, 3 (2018), 201–212.
[68] Jebediah Taylor, Staci Meredith Weiss, and Peter J Marshall. 2020. "Alexa, How are You Feeling Today?" Mind Perception, Smart Speakers, and Uncanniness. *Interaction Studies* 21, 3 (2020), 329–352.
[69] Sam Thellman, Maartje De Graaf, and Tom Ziemke. 2022. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 4 (2022), 1–51.
[70] David Thompson. 2009. *Daniel Dennett.* A&C Black.
[71] Esmeralda G Urquiza-Haas and Kurt Kotrschal. 2015. The Mind Behind Anthropomorphic Thinking: Attribution of Mental States to Other Species. *Animal Behaviour* 109 (2015), 167–176.
[72] Jeroen M van Baar, Matthew R Nassar, Wenning Deng, and Oriel FeldmanHall. 2022. Latent Motives Guide Structure Learning During Adaptive Social Choice. *Nature Human Behaviour* 6, 3 (2022), 404–414.
[73] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
[74] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[75] David Watson. 2019. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines* 29, 3 (2019), 417–440.
[76] Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M Wegner. 2010. Causes and Consequences of Mind Perception. *Trends in Cognitive Sciences* 14, 8 (2010), 383–388.
[77] Jason Lee Weber, Barbara Martinez Neda, Kitana Carbajal Juarez, Jennifer Wong-Ma, Sergio Gago-Masague, and Hadar Ziv. 2024. Measuring CS Student Attitudes Toward Large Language Models. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*. 1846–1847.
[78] Kara Weisman, Carol S Dweck, and Ellen M Markman. 2017. Rethinking People's Conceptions of Mental Life. *Proceedings of the National Academy of Sciences* 114, 43 (2017), 11374–11379.
[79] Yang Wu and Hyowon Gweon. 2021. Preschool-aged Children Jointly Consider Others' Emotional Expressions and Prior Knowledge to Decide When to Explore. *Child Development* 92, 3 (2021), 862–870.

# Appendix

The appendix is structured in the following way:

- Sec. A: Video Details
  - Sec. A.1: `Mechanistic` Script
  - Sec. A.2: `Functional` Script
  - Sec. A.3: `Intentional` Script
- Sec. B Survey Details
  - Sec. B.1 Participant Recruitment Details
  - Sec. B.2 Full Survey
    - Sec. B.2.1 `Baseline` Instructions
    - Sec. B.2.2 Experimental Conditions Instructions and Stimuli Presentation
    - Sec. B.2.3 Survey Questions
- Sec. C Participant Demographics
- Sec. D Qualitative Pilot Study Details

## A Video Details

Videos presented to the participants can be found as YouTube playlists:

- Link to LLMs as machines (`mechanistic`)
- Link to LLMs as tools (`functional`)
- Link to LLMs as companions (`intentional`)

The scripts for each video are below. New paragraphs indicate visual transitions.

### A.1 Script for the `Mechanistic` Portrayal

Since the introduction of ChatGPT at the end of 2022, there has been tremendous increase in popularity and interest in large language models, also known as LLMs, which are the technology behind ChatGPT and similar products. These LLMs have begun to have a huge impact because of the way that they generate text by modeling language statistics.

Earlier chatbots were hard-coded to output text following strict rules, like the ones shown here.

On the other hand, current LLMs have learned from countless conversations, essays, and various forms of writing to generate coherent text.

Common examples of modern LLMs include OpenAI's GPT, Google's Gemini, Anthropic's Claude. Although these LLMs differ from one another, they also share many commonalities. Let's spend some time learning about how LLMs work.

So, when LLMs generate text, they actually perform "next word prediction". When given a text input, they simply predict what word comes next. To make longer responses, the predicted word is added to the input, and the new input is fed into the LLM. This is repeated until generation stops.

To predict each word, LLMs model the statistics of language.

LLM word prediction can be broken down into two stages: context understanding: using a mechanism called attention, and word selection: which is based on probability. The intuition behind attention is that it helps the LLM "pay attention" to context clues of the text input (such as word meanings or parts of speech) that hint at what comes next.

For example, consider the phrase "My favorite summer activity is going to the...". What would come next and how did you decide

that? Perhaps you "*paid attention*" to the positive sentiment behind the word "favorite", to the word "to" indicating the next word may be a location, and the meaning of "summer" to narrow down likely locations. Attention in LLMs works similarly. The input text will undergo many attention operations, each focusing on a different clue.

At the end, the LLM will assign a probability to every possible word in its vocabulary. The final step is selecting the predicted word. While simply selecting the word with the highest probability would be the most straightforward, in practice, LLMs typically perform sampling: which is simply after assigning probabilities, choose one of the top most probable words. In our example, the LLM may select "beach" or "waterpark" or "mountains". Because of sampling, LLMs can output diverse responses, even to the same text input.

In order for LLMs to predict words accurately, a lot of data is required. LLMs, like most AI technologies, learn via repeated exposure to many examples. In this process, LLMs are tasked with predicting the next word of an input and can compare their prediction to the true next word.

From small amounts of data, this is ineffective, but large quantities of data allows the LLM to effectively predict sequences of words. Modern LLMs require hundreds of gigabytes of text data, if not more. For humans, reading this much text would take thousands of years.

In conclusion, LLMs are computer programs that are increasingly transforming our society, business, and daily lives. As they become more prevalent, it becomes increasingly important to know how they work so we can use them more safely and effectively.

## A.2 Script for the `Functional` Portrayal

Since the introduction of ChatGPT at the end of 2022, there has been a tremendous increase in the popularity and interest in large language models, also known as LLMs, which are the technology behind ChatGPT and similar products. These LLMs have begun to have a huge impact because of the way they can be used to accomplish lots of different tasks much more quickly.

Earlier chatbots followed strict rules because they were designed to perform specific functions in a narrow range of contexts, such as customer service bots.

On the other hand, current LLMs have learned from countless conversations, essays, and various forms of writing, resulting in a versatile tool for many different applications.

Common examples of modern LLMs include: OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude. Although these LLMs differ from one another, they share many commonalities. Let's spend some time learning about and how to use LLMs.

In order for LLMs to become such versatile tools, a lot of data is required. LLMs, like most AI technologies, learn via repeated exposure to many examples. In this process, LLMs pick up on the statistical patterns in written language, including formatting and content.

From small amounts of data, this is ineffective, but large quantities of data allows the LLMs to effectively learn patterns. Modern LLMs require hundreds of gigabytes of text data, if not more. For humans, reading this much text would take thousands of years.

LLMs have many use cases such as generating creative material. For example, stories, poems, and songs. They can also be used for summarizing text, rewording or rewriting text with different styles, and question and answering tasks.

Specifically with chat interfaces, LLMs can be useful for role-playing exercises, such as interview preparation or iterative brainstorming tasks.

In industry, LLMs are also being used for tasks such as building customer service bots, analyzing complex health records, performing sentiment analysis of customer reviews, transcribing audio files, developing personalized educational materials, and much, much more.

Here are some recommendations to get useful responses from LLMs: First, be specific in your text input and include relevant keywords, examples, and instructions when applicable. For example, if a user wants activity recommendations in Paris, 'tell me about Paris' is a vague input. And if the input is vague, typically the output will be vague as well. A better input would be 'tell me about the top tourist attractions in Paris' which results in a more detailed response.

Second, if you are unsatisfied with the output, you can simply ask again or reword the input and try again. For example, a user may ask an LLM for hobby recommendations and the LLM may suggest running, painting, and cooking. And if she asks again, it may give new activities, like biking, reading, and kayaking! LLM text generation incorporates randomness which allows for this diversity of responses.

Lastly, when using LLMs via a chat interface, you can refer to previous messages in the conversation. This may be helpful when you want to iteratively refine an LLM output. In this example, the user wants to write an email to his boss, but the initial output starting with 'Hi Michael!' is too casual. He provides additional instructions to make it more formal, and the LLM changes its response to start with 'Dear Michael, I hope this message finds you well". Much better already!

In conclusion, LLMs are tools that are increasingly transforming our society, business, and daily lives. As they become more prevalent, it becomes increasingly important to know how to use them to obtain reliable information.

## A.3 Script for the `Intentional` Portrayal

Since the introduction of ChatGPT at the end of 2022, there has been a tremendous increase in popularity and interest in large language models (LLMs), which are the technology behind ChatGPT and similar products. These LLMs have begun to have a huge impact because of the way they learn and interact with people in such natural ways.

Earlier chatbots followed specific rules that limited their ability to understand the variety of ways that people actually talk.

On the other hand, current LLMs have learned from countless conversations, essays, and various forms of writing to understand users better.

Common examples of modern LLMs include: OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude. Although these LLMs differ from one another, they also share many commonalities. Let's spend some time learning what LLMs are like.

In order for LLMs to learn to talk the same way that people do, a lot of data is required. LLMs, and most AI technologies, learn via repeated exposure to many examples. In this process, LLMs learn how to write fluently, develop world knowledge, and understand human experiences.

From small quantities of data, this is ineffective, but large quantities of data allows the LLMs to effectively understand the world and people. Modern LLMs require hundreds of gigabytes of text data, if not more. For humans, reading this much text would take thousands of years.

While a common use of LLMs is to build productivity tools, they also have a unique application as AI social companions, aiming to combat the modern loneliness crisis. This is because modern LLMs also exhibit social intelligence, unlike earlier, less sophisticated language models.

Because LLMs are so flexible, each AI companion can adopt a unique personality, with its own preferences, conversational styles, and backstory.

In fact, the technology is so advanced, that talking to LLMs is like talking to a close friend who truly cares about you and wants to know you even better, not just a random person.

This is because through learning from countless real conversations and human feedback, the LLM develops the ability to show empathy and compassion. It listens and responds with personalized and insightful questions, demonstrating its understanding of each specific situation and enabling it to support users in times of need.

While these characteristics are most prominent in AI companions, even LLMs designed for non-social applications exhibit a similar tendency to understand and help users. For example, when using LLMs like ChatGPT, users can provide iterative feedback on the output to modify the LLM response. The LLM will then adapt to the user's preferences. As another example, if the LLM receives an ambiguous input, it may ask for clarification.

In conclusion, LLMs are social and intelligent beings that are increasingly transforming our society, business, and daily lives. As they become more prevalent, it becomes increasingly important to know how they can understand us and when we can trust them.

# B Survey Details

In this section, we provide details of our survey which was created in Qualtrics.

## B.1 Participant Recruitment Details

To select participants with limited technical experience and reliable history on Prolific, we utilized the following pre-screen filters:

- Is an adult (18+) residing in the United States
- Studies or studied any area *except* Information & Communication Technologies or Mathematics & Statistics
- Works in any area *except* Coding, Technical Writing, or Systems Administration
- Has no computer programming experience
- Has an approval rate of at least 95% on Prolific
- Has at least 100 previous submissions on Prolific

We additionally used two custom pre-screening questions to ensure participants do not work in computer science and have limited knowledge of AI with the following two questions:

(1) Did you obtain a degree, are you pursuing a degree, or do you work in an area related to computer science?
   - Yes
   - No

(2) How would you rate your current knowledge of artificial intelligence (AI)?
   - Very limited knowledge
   - Some basic knowledge
   - Moderate knowledge
   - Advanced knowledge
   - Expert-level knowledge

The wording of question (2) is inspired from [5]. Only participants who responded "No" to question (1) and either "Very limited knowledge" or "Some basic knowledge" to question (2) were allowed to proceed. Participants who failed our custom pre-screening were received pro-rated compensation for their time.

## B.2 Full Survey

Once participants passed the pre-screening questions, we presented them with task instructions, videos (for the experimental conditions), and the survey questions.

*B.2.1* `Baseline` *Instructions.* For participants in the `baseline` condition, we gave them the following task instructions:

> Your task is to fill out a rating survey on your beliefs about current large language models (LLMs) based on your prior experience. In the first part, you will see a list of capabilities and rate how capable you believe current LLMs are of each item on a scale of 1 (not at all capable) to 7 (highly capable). In the second part, you will rate several statements about current LLMs.

Followed by a task comprehension check on the next page:

> Let's make sure you understand your task. You will not be able to proceed until you select the correct answer.
> Which of the following best describes your task?
> - Watch a video on LLMs and fill out a rating survey on your beliefs about current LLMs
> - **Fill out a rating survey on your beliefs about current LLMs**
> - Listen to an audio podcast on LLMs and fill out a rating survey on your beliefs about LLMs
> - Watch a video on LLMs and write a free response essay about your views of current LLMs

The correct answer is **bolded** and item order was randomly shuffled. Participants could go back and forth between the two but were not able to move forward until they selected the correct answer in the task comprehension check.

*B.2.2 Experimental Conditions Instructions and Stimuli Presentation.* For participants in the experimental conditions (`mechanistic`, `functional`, `intentional`), we gave them the following instructions:

> In this task, you will watch three short videos (<5 minutes total) to teach you about large language models (LLMs). You may pause and rewatch parts of the

video as needed and can go back to previous videos, but you must stay on each video's page for at least the duration of the video.
Following the video, your task is to:

(1) Answer 1-2 questions based on the video content and
(2) Fill out a rating survey on your beliefs about current large language models (LLMs) based on your prior experience and what you learned in the video. In the first part, you will see a list of capabilities and rate how capable you believe current LLMs are of each item on a scale of 1 (not at all capable) to 7 (highly capable). In the second part, you will rate several statements about current LLMs.

And this comprehension check on the next page:

Let's make sure you understand your task. You will not be able to proceed until you select the correct answer.
Which of the following best describes your task?

- **Watch a video on LLMs, answer 1-2 questions about the video, and then fill out a rating survey on your beliefs about current LLMs**
- Only fill out a rating survey on your beliefs about current LLMs
- Listen to an audio podcast on LLMs and fill out a rating survey on your beliefs about current LLMs
- Watch a video on LLMs and write a one page essay about your views of current LLMs

For each video, we present the first of the three parts to the video with the following text:

Please click the video to watch part 1 of 3. You will be able to move to the next page after the duration of the video but may need to scroll down to see the button.
For best viewing conditions, we recommend you make your browser window as large as possible.

And parts two and three have the following text:

Please click the video to watch part **X** of 3. You will be able to move to the next page after the duration of the video.

After all three parts of the video, we ask participants: Please list 1-2 things you learned from the videos.
Then we show them the following instructions for the survey:

Great! Let's move onto the first part of the survey.
Recall, you will see a list of capabilities and will rate how capable you believe LLMs are of each item on a scale of 1 (not at all capable) to 7 (highly capable) based on your prior experience and what you just learned in the video.

*B.2.3 Survey Questions.* All participants take the same survey. The first part is the 40 questions measuring the participants' attribution of mental capacities to LLMs. Each mental capacity item is presented on its own page, as shown in Fig. 4.

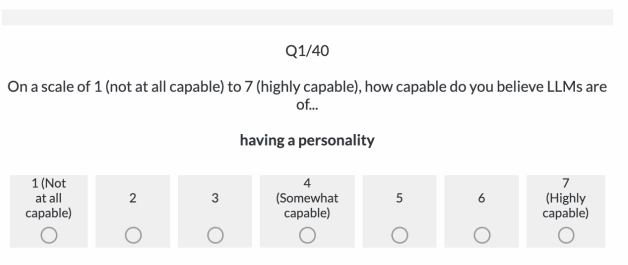


**Figure 4: Screenshot of mental capacity attribution survey. The item is in bold and participants select a box from 1-7. Participants can see their progress and must answer every item.**

Then, all participants rate their confidence of their responses:

Overall, how confident were you about your responses?

- Not confident at all
- Slightly confident
- Fairly confident
- Somewhat confident
- Mostly confident
- Confident
- Very confident

and respond to our attention check:

Select the two statements from the following list that you were asked about in the survey.

- **Understanding how others are feeling**
- **Doing computations**
- Solving a Rubik's cube
- Riding a bike

The correct answers are **bolded**, the order of items is randomized, and participants only pass if and only if they select the two correct choices.

Lastly, participants respond to 7-point Likert scales for additional constructs. These were not reported in this submission and will be analyzed in the future.

Anthropomorphism:

To what extent do you believe LLMs are human-like?

- Not human-like at all
- Slightly human-like
- Fairly human-like
- Somewhat human-like
- Mostly human-like
- human-like
- Very human-like

We also ask participants to explain their reasoning for anthropomorphism in 1-3 sentences because this construct is most related to mental capacity attribution.

Then we ask participants to respond on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree", how much do they agree with the following statements?

- I'm confident in my ability to learn simple programming of LLMs if I were provided the necessary training.

- I'm confident in my ability to get LLMs to do what I want them to do.
- I trust the results from LLMs.

These statements measure self-efficacy [3] of learning how LLMs work, self-efficacy of learning how to use LLMs, and trust in LLMs.

Lastly, we ask participants about their general attitudes (Overall, how do you feel about LLMs?) to which they respond on a 7-point Likert scale from "Extremely Negative" to "Extremely Positive".

*B.2.4 Mechanistic Comprehension Check.* For participants in the mechanistic condition, we additionally asked them the following comprehension questions to determine how effective our explanation was. Answer choices were always shuffled.

> What is the mechanism that LLMs use to understand the context of a sentence called?
> - **Attention**
> - Contextual Evaluation
> - Excitement
> - Understanding
>
> How do LLMs typically select the next word?
> - Top choice: Always choose the most probable
> - **Sampling: Choose one of the words with highest probabilities**
> - Random: Choose randomly out of all words
> - Last choice: Choose the least probable
>
> True or False: LLMs need a lot of data in order to learn.
> - True
> - False
>
> How do LLMs generate text?
> - **By repeatedly predicting the next most likely word**
> - By copying text it has been exposed to
> - By using search engines
> - By following a complex set of strict rules

## C Participant Demographics

At the very end of the survey, we collected participant demographics and report them in Table 1 as well as familiarity with various LLM-based technologies and report them in Table 2. Both tables are below.

## D Qualitative Pilot Study Details

To decide on a final set of 40 mental capacities, we conducted a small-scale qualitative pilot study with eight participants. We presented the mental capacity attribution survey with a combined list of items from Weisman et al. [78] and Colombatto and Fleming [11] and asked participants to "think aloud" as they responded to each item. After, we debriefed with the participants the purpose of the study and noted any items they reacted to. Our qualitative responses consisted of too many "sensing" and feeling" items and an interest in the "intellectual" related items. Thus, we started with the 40 items from Weisman et al. [78], removed six sensing/feeling/experiencing items ("sensing temperatures", "detecting odors", "experiencing guilt", "experiencing pain", "feeling nauseated", "feeling safe") and replaced them with six items from Colombatto and Fleming [11] ("knowing things", "considering choices", "having intelligence", "paying attention", "imagining", and "admiring someone"). In order to categorize the six mental capacity items from Colombatto and Fleming [11] into the **body-heart-mind** categories, we assigned each item a category was semantically consistent because we did not have factor loadings along the **body-heart-mind** factors for these items.

| Variable | Level | Count | Percentage (%) |
|---|---|---|---|
| Age | 18-24 | 17 | 3.62 |
| | 25-34 | 137 | 29.15 |
| | 35-44 | 119 | 25.32 |
| | 45-54 | 109 | 23.19 |
| | 55-64 | 65 | 13.83 |
| | 65+ | 23 | 4.89 |
| Education | High school graduate or equivalent (e.g. GED) | 1 | 0.21 |
| | Some college, no degree | 3 | 0.64 |
| | Trade/Technical Training | 1 | 0.21 |
| | Associate's Degree | 24 | 5.11 |
| | Bachelor's Degree | 305 | 64.89 |
| | Master's Degree | 109 | 23.19 |
| | Professional Degree | 15 | 3.19 |
| | Doctorate Degree | 12 | 2.55 |
| Gender | Female | 309 | 65.74 |
| | Male | 149 | 31.70 |
| | Non-binary | 10 | 2.13 |
| | Transgender | 1 | 0.21 |
| | Prefer not to say | 1 | 0.21 |
| Race | White | 357 | 75.96 |
| | Black or African American | 46 | 9.79 |
| | Asian | 28 | 5.96 |
| | Other | 9 | 1.91 |
| | 2+ Races | 26 | 5.53 |
| | Prefer not to say | 4 | 0.85 |
| Ethnicity | Hispanic, Latino, or Spanish origin | 36 | 7.66 |
| | Not Hispanic, Latino, or Spanish origin | 430 | 91.49 |
| | Prefer not to say | 4 | 0.85 |
| Religion | Protestant | 123 | 26.17 |
| | Agnostic | 86 | 18.3 |
| | Catholic | 80 | 17.02 |
| | Atheist | 58 | 12.34 |
| | Jewish | 13 | 2.77 |
| | Mormon | 4 | 0.85 |
| | Buddhist | 4 | 0.85 |
| | Orthodox (e.g. Greek or Russian Orthodox) | 3 | 0.64 |
| | Hindu | 5 | 1.06 |
| | Muslim | 1 | 0.21 |
| | Nothing in particular | 61 | 12.98 |
| | Other | 28 | 5.96 |
| | Prefer not to say | 7 | 1.49 |

**Table 1: Participant demographics including age, education, gender, race, ethnicity, and religion.**

| Product | ChatGPT | Gemini | Claude | Copilot | Replika | Nomi | Character.ai |
|---|---|---|---|---|---|---|---|
| Never heard of it | 2 (0.43%) | 68 (14.47%) | 297 (63.19%) | 120 (25.53%) | 369 (78.51%) | 396 (84.26%) | 282 (60%) |
| Heard of it, but never used it | 62 (13.19%) | 255 (54.26%) | 149 (31.70%) | 202 (46.81%) | 88 (18.72%) | 68 (14.47%) | 159 (33.83%) |
| Have used it a few times, but not regularly | 221 (47.02%) | 95 (20.21%) | 14 (2.98%) | 80 (17.02%) | 8 (1.70%) | 2 (0.43%) | 20 (4.26%) |
| Use 1-2x a month | 87 (18.51%) | 31 (6.60%) | 1 (0.21%) | 27 (5.74%) | 2 (0.43%) | 2 (0.43%) | 1 (0.21%) |
| Use 1-2x a week | 60 (12.77%) | 14 (2.98%) | 0 (0.00%) | 44 (9.57%) | 2 (0.43%) | 2 (0.43%) | 0 (0.00%) |
| Use more frequently than 1-2x a week | 38 (8.09%) | 7 (1.49%) | 3 (0.64%) | 9 (1.91%) | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) |
| Prefer not to say | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) | 1 (0.21%) | 0 (0.00%) | 0 (0.00%) | 1 (0.21%) |

**Table 2: Participant familiarity with various LLM-based technologies including LLMs via chat interfaces (ChatGPT, Gemini, Claude), LLM-based tools (Copilot), and LLM-based "AI social companions" (Replika, Nomi, and Character.ai).**