Portraying Large Language Models as Machines, Tools, or Companions Affects the Mental Capacities People Attribute to Them

Appendix

The appendix is structured in the following way:

- Sec. 1: Video Details
  - Sec. 1.1: `Mechanistic` Script
  - Sec. 1.2: `Functional` Script
  - Sec. 1.3: `Intentional` Script
- Sec. 2 Survey Details
  - Sec. 2.1 Participant Recruitment Details
  - Sec. 2.2 Full Survey
    - ∗ Sec. 2.2.1 `Baseline` Instructions
    - ∗ Sec. 2.2.2 Experimental Conditions Instructions and Stimuli Presentation
    - ∗ Sec. 2.2.3 Survey Questions
- Sec. 3 Participant Demographics
- Sec. 4 Qualitative Pilot Study Details

1 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.

1.1 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.

1.2 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.

1.3 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.

2 Survey Details

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

2.1 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 **?**. 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.

2.2 Full Survey

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

2.2.1 `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.

2.2.2 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.

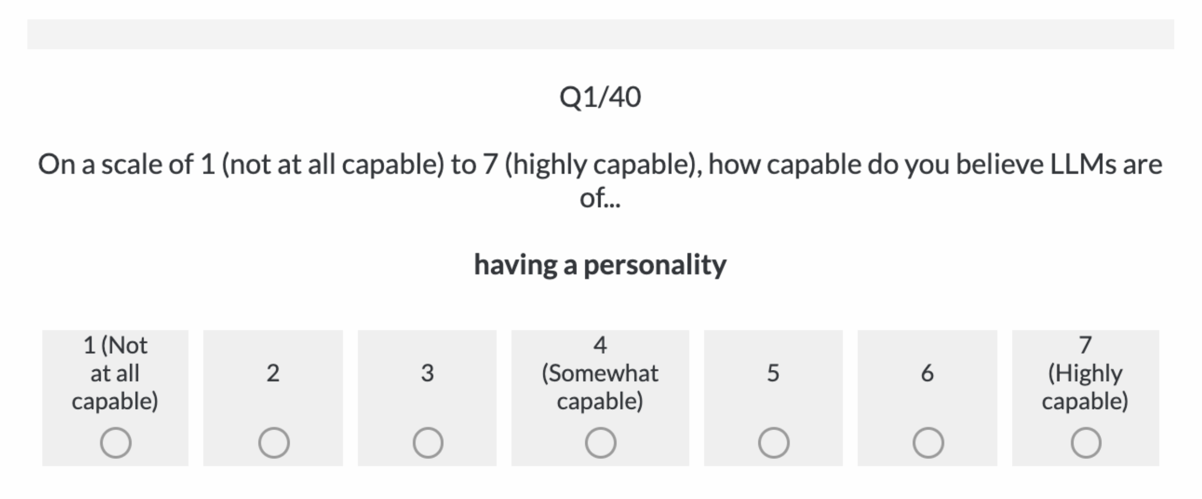
After all three parts of the video, we ask participants: Please list 1-2 things you learned from the videos.

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.

2.2.3 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. 1.



Q1/40

On a scale of 1 (not at all capable) to 7 (highly capable), how capable do you believe LLMs are of...

having a personality

1 (Not at all capable) ○	2 ○	3 ○	4 (Somewhat capable) ○	5 ○	6 ○	7 (Highly capable) ○

Figure 1: 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 **?** 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".

2.2.4 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

3 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.

4 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. (2017) and Colombatto and Fleming (2024) 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. (2017), 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 (2024) ("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 (2024) 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).

References

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

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.