**Marcus Alexander**

**Harvard College**

# Large Language Models and Thought

With the increasing accessibility to large language models (LLMs), humans are puzzled with the question of whether these models can produce coherent thought. While the earlier iterations of large language models were very easily distinguishable from human thought, as they became more refined, the line became more blurred. The mechanisms that these models operate upon allow for the ability to become constantly refined by training data. Training data is the backbone of these models; as the data set becomes larger, the model becomes more efficient. When the model receives a textual input it algorithmically leverages the training data to output a coherent textual sequence, which is derived from the model predicting the most likely "tokens" (sequences of letters) that should occur after the catalyst textual input. Due to this process, when humans view the output, more times than not it is coherent and serviceable. This process can seem very similar to human interaction — the input and output sequence mimics a conversation. This results in the question of thought being raised. If a large language model is able to become indistinguishable from a human, does this indicate that there is thought occurring? Proponents of the ability for LLM to engage in thought cite the ability of these models to pass the "Turing Test". The Turing Test is a test where a proctor will have a textual conversation with both a human and a computer. If the proctor cannot reliably distinguish the difference between the computer and the human, then the machine can exhibit intelligent behavior indistinguishable from a human. While this belief can carry merit initially, upon further understanding of what genuine thought is in addition to how large language models formulate outputs, it is clear that these models are simply operating in the algorithmic way they were programmed to and have the inability to ever engage in human-esque thought.

A clear distinction must be drawn between large language models (weak AI) and strong AI. Large language models as mentioned above are a type of artificial intelligence that excel at responding to human inputted prompts. They operate by using textual analysis and language generation. Large language models are not strong AI. Strong AI is artificial intelligence that is able to understand, recall, apply knowledge, and is adaptable to the introduction of new information[1]. Strong AI and LLM differ in terms of scope, capabilities, and complexity. LLMs scope is simply just textual analysis and output while strong AI is complete understanding. LLMs have very limited capabilities outside of text generation. Strong AI would include cognitive capabilities such as rationality, logic, creativity, learning and adaptation, emotional and social intelligence, and ethical moral reasoning. Strong AI at this current time is simply theoretical while large language models are widely used.

Eric Schwitzgebel, David Schwitzgebel, and Anna Strasser explored the potential for large language models to engage in thought in their article, *Creating a Large Language Model of A Philosopher*. To understand if large language models could engage in thought they formulated an experiment where they would task a LLM to answer philosophical questions. Within the experiment they trained a large language model with data from the philosopher Daniel Dennett's published pieces, taking care to ensure that there would not be any overfitting. When the model was adequately trained, they then ran a test where they instructed the model, named DigiDan, to output a philosophical argument similar to one that would be made by Daniel Dennett. They then compared it to an argument provided by Daniel Dennett. The accuracy of the model was then tested amongst three groups, Dennett experts, blog readers, and ordinary people. All the groups, experts included, could not significantly distinguish which was the model and which was

[1] Searle, John, (Minds, Brains, and Programs, pg.420)

Dennett[2]. The model outputs were established to not have plagiarized Dennett's previous work. While this experiment was not a Turing test due to the test not consisting of back and forth exchanges[3], it does shine a light on how when the models become more advanced and can participate in a turing test, there is high plausibility for them to pass. As LLM's become more algorithmically efficient and the training data becomes more refined, then these models may become more indistinguishable from humans.

Traditional thought indicates that a LLM's indistinguishability from humans points to these models being capable of thought — Schwitzgebel et al.'s publication[4] supports this conclusion. While the limitations to proctor a turing test in the Scwitzgebel study result in a conclusion not being able to be drawn, the ability of the model successfully output indistinguishable prompts, indicates that upon further innovation there is a high likelihood that these models will be able to pass. By passing a Turing test that would indicate that the LLM has the ability to demonstrate human intelligence. While I do believe that in the future LLMs will be able to consistently pass a Turing test, I do not believe that the conclusion can be drawn that these models will ever be able to possess thought or "Human Intelligence".

Human intelligence by definition is possessing the ability to comprehend, learn from experience, understand nuanced concepts, and use ingenuity to alter one's environment[5]. The fundamental explanation of what human intelligence is disqualifies large language models from possessing an identical nature of intelligence. The backbone of large language models are the algorithms and the training data. The training data comes from a vast amount of text data, primarily from the internet, and then is utilized to give the accurate outputs. When a text is input,

[2] Schwitzgebel, Eric et al. , (Creating a Large Language Model of A Philosopher, pg .10)
[3] Schwitzgebel, Eric et al. , (Creating a Large Language Model of A Philosopher, pg .17)
[4] Schwitzgebel, Eric et al. , (Creating a Large Language Model of A Philosopher)
[5] "Human Intelligence." Encyclopædia Britannica

the text data is tokenized. With the input being tokenized the model is able to sequentially process the input by recalling previous similar tokens within the training data. The model will then generate an output by inferring the next best token that should exist within the sequence. As a result the output is not only coherent but very human-like. This process is extremely different from "human intelligence". The models are unable to possess personal experiences or recollection; models do not retain information from past interactions or conversations. This results in every interaction being independent and the output being generated solely on the input of the given moment. The token manipulation that large language models operate on demonstrates that there is no understanding or comprehension. When large language models do become more efficient it is simply the neural networks of the model being fine-tuned. Fine tuning is when the training data is updated and the token manipulation of the algorithm becomes more efficient. Improvements within the models are simply improvements of the vast library of training data that is utilized to infer the tokens of the input strain more efficiently. Increased innovation in these systems does not indicate that any additional processes are occurring, therefore they will simply be the nature of what they were created to do token manipulation.

Misidentifying large language model outputs as human intelligence is falling into the "good at language, good at thought" fallacy that Kyle Mahowald discusses in his paper, *Dissociating language and thought in large language models*. The "good at language, good at thought" fallacy is that by human or machine being able to articulate a coherent sentence, then the conclusion that they are good at thought is fallaciously derived. By understanding the fundamental nature of large language models we know that this cannot be the case, thus why the confounding of good language skills with thought is a fallacy. Mahowald draws a distinction between two types of linguistic competencies: formal linguistic competence and functional linguistic competence[6]. The concepts of formal and functional language competency work jointly to aid in the nuances of "human intelligence". Formal linguistic competence is the ability to understand the rules and statistical regularities of language. Functional linguistic competency is the ability to utilize language in social settings with perception, recollection, reason, and logic[7]. The distinction between formal and functional language competency is important since, as Mahowald argues, the algorithmic next word prediction method that large language models utilize have only truly master formal language competency[8]. Unlike formal language competency, functional language competency is not confined to programmable rules. By functional language competency being very situational, it requires understanding, rationality, and logic. Large language models are able to mirror human intelligence by their mastery of formal language competency but their success does not indicate the same processes that occur during human formal language competency.  Large language models are very efficient in operating within the rules of phonology, morphology, semantics, and syntax[9]. Formal language competency rules can very easily be trained via data onto a model. Large language models do very poorly when it comes to formal reasoning, world knowledge, situation modeling, and social reasoning[10]. Large language models have shown failures when they are given prompts that require functional reasoning. When a LLM model such as gpt-3 or gpt-4 is given simple two digit or three digit multiplication the model cannot accurately supply a correct answer reliably[11]. The reasoning for this is that LLMs do not have contextual understanding so oftentimes they can misinterpret the tasks that are being asked. Also the training data that is required to output the correct answer is

[6] Mahowald, Kyle, et al., ( Dissociating Language and Thought In Large Language Models, page 2)
[7] Mahowald, Kyle, et al. , ( Dissociating Language and Thought In Large Language Models, page 3)
[8] Mahowald, Kyle, et al. , ( Dissociating Language and Thought In Large Language Models, page 5)
[9] Mahowald, Kyle, et al. ,( Dissociating Language and Thought In Large Language Models, page 4)
[10] Mahowald, Kyle, et al. , ( Dissociating Language and Thought In Large Language Models, page 13)
[11] Mahowald, Kyle, et al. , ( Dissociating Language and Thought In Large Language Models, page 13)

not available for the model. The failures in rudimentary math blatantly show that the models are not thinking but just trying to complete the string of tokens. Math is not the only example of LLMs functional reasoning incompetence. Large language models do very poorly when they are presented with situations or contexts that deviate outside of what their training data is. While it is possible for additional programs to be implemented on top of the large language models to aid in these deficiencies, the model still will be operating within the confines of the program and of token manipulation which prevents the ability for thought to occur. I argue that the essence of human intelligence comes from the characteristics that fall under functional reasoning. As a result large language models regardless of any innovation will never be capable of thought.

The ability to output a coherent sentence does not indicate thought is present. John Searle's example of the Chinese Room Experiment depicts this phenomenon. The Chinese Room Experiment[12] is a hypothetical scenario where there is a human who is a native english speaker wh has no previous knowledge of Chinese is locked in a room with Chinese symbols and an instruction manual on how to manipulate the Chinese symbols. After enough practice, if they were to receive an input of Chinese symbols they would be able to pass a Turing test and output a coherent sentence with zero understanding. Applying this to a LLM that is able to be programmed in a way that it can receive an input of Chinese characters and under the instructions of a programmer it can manipulate those symbols outputting a coherent sentence similar results would occur. A Turing test would then be proctored and the model will be able to pass. The individuals on the other end of the test would believe that the model has an understanding of the Chinese language and they will believe that they are speaking with a Chinese speaking human being. Under normal presumptions, a model passing a Turing test would indicate that it possesses a level of human intelligence. This experiment challenges this

[12] Searle, John, (Minds, Brains, and Programs, pg.418)

conclusion. In both the case of the human and LLM, there was no sense of intentionality[13] despite the outputs indicating such. When a model is operating on mere symbol manipulation it is clear that there is no thought present. The conclusion from the Chinese Room Experiment challenges Switzgebel's prediction that in the future LLMs will be capable of thought. This displays the large differences between weak AI (large language models) and strong AI. While if strong AI were proctored the same Turing test, it would have similar results, there is a sense of understanding that is present.

Large Language Models in present day and in future innovations will simply always be machines that excel in symbol manipulation. A model that is able to output a very coherent sentence does not indicate thought or intelligence. I argue that by passing a Turing test the conclusion of intelligence cannot always be drawn. From understanding the fundamental nature of these models and thought experiments such as the "The Chinese Room Experiment" it is evident that a coherent output is not indicative of intention or thought. Human intelligence is a form of intelligence that will always be distinct from the intelligence that is displayed by LLMs. As the LLMs become more advanced, the line between capabilities of these models and humans will seem even more blurred, but will always be present.

## **Work Cited**

[13] Searle, John, (Minds, Brains, and Programs, pg.424)

Searle, John R. Minds, Brains, and Programs. Cambridge University Press (pgs 417-424), 1980.

Mahowald, Kyle, et al. “Dissociating language and thought in large language models.” Trends in Cognitive Sciences, Mar. 2024, https://doi.org/10.1016/j.tics.2024.01.011.

Schwitzgebel, Eric, et al. “Creating a large language model of a philosopher.” Mind &amp; Language, vol. 39, no. 2, 12 July 2023, pp. 237–259, https://doi.org/10.1111/mila.12466.

“Human Intelligence.” Encyclopædia Britannica, Encyclopædia Britannica, inc., 5 Apr. 2024, www.britannica.com/science/human-intelligence-psychology.