Olivia Y. Lee
SYMSYS 207 Final Paper

Large Language Models: Intelligence, Understanding, and Intentionality

In July 2020, OpenAI's release of the most recent Generative Pre-Trained Transformer, GPT-3, shook the world of Artificial Intelligence (AI) (Brown et al., 2020). The 175 billion parameter large language model (LLM), which is 100 times larger and trained on much more data than its predecessor, GPT-2, is the most powerful language system to date, capable of generating shockingly versatile, human-like text on demand. While many were excited by the possibility of LLMs bringing us closer to highly sophisticated artificial general intelligence (AGI), many individuals were also critical of GPT-3's proximity to AGI. Some philosophers have questioned whether LLMs like GPT-3 can ever be considered intelligent (beyond weak forms of intelligence measured by a behavioral response), highlighting several crucial implementational and inherent shortcomings. This paper argues that modern language systems are not able to achieve strong intelligence. LLMs do not learn quickly and flexibly, nor do they employ heuristics for inference-making in a manner that an intelligent system would. Furthermore, LLMs have limited capacity for understanding beyond symbol manipulation, and are purely reactional systems that lack intentionality. An alternative view is that LLMs are not intelligent because they merely reconstruct information generated by humans. To this objection, this paper argues that all agents that acquire language, including humans, engage in this process, and that this observation is insufficient to prove that LLMs are not intelligent.

A key feature of an intelligent system is its ability to learn quickly and flexibly with a relatively high sample efficiency. This is arguably one of the biggest challenges modern AI systems currently face – even the simplest binary classification systems require numerous examples before they can extract the relevant features required for a specific classification task, while a human baby only needs to see a few cats and dogs to differentiate between the two. Therefore, the most compelling evidence that modern LLMs like GPT-3 lack intelligence is the empirical fact that GPT-3 requires *45 terabytes of data* to acquire comparable language skills to humans, and even then, it still falls short in certain situations (Brown et al., 2020) (Moradi et al., 2021). While GPT-3 has been hailed as a "zero-shot" or "few-shot" learner, these terms refer to the

number of samples or demonstrations provided at inference time. The fact remains that LLMs are trained on a massive database prior to this inference. Given this, it should not be surprising that they perform well on a large variety of tasks, seeing as they have processed a good chunk of the web before being prompted. On the other hand, intelligent systems, like humans, learn incrementally. Humans eventually accumulate a vast amount of information over time, but we do not require all that information up front to perform a wide array of tasks. To make sense of our limited information, humans use heuristics to make inferences, generalize, and reason about new information from the environment. The inference-making ability of intelligent systems is especially nontrivial, as LLMs have been shown to struggle with natural language inference tasks, which involve identifying whether a statement is entailed or contradicted by a piece of text (Brown et al., 2020). GPT-3's learning process, on the other hand, largely involves rote statistical processing and later, a searching and pattern-matching process to produce an answer. Even if a future LLM, say GPT-X, passes the Turing test and achieves perfectly human-like conversational skills, would this system be deemed intelligent if it took an impractical amount of data, computational resources, and time to train? Intelligent agents should not need to parse the entire Internet to engage in conversation or write an article. Ultimately, this difference in information handling – learning a huge amount of information upfront and then generating relevant information, versus learning incrementally with fewer data samples and making inferences using heuristics – demonstrates the key difference between intelligent agents and LLMs.

In addition to the low sample efficiency and lack of higher-order heuristics for making inferences in LLMs, their capacity for true understanding is limited. The traditional definition of understanding refers to the ability to rigorously reason about interactions between the agent and its surroundings, and more broadly about how the world works (Grimm, 2021). There are two complementary components of this argument. One facet of this argument is that LLMs inherently lack the ability to understand as a virtue of them being purely verbal language systems. Philosopher of consciousness David Chalmers posits, "Can [GPT-3] really understand happiness and anger just by making statistical connections? Or is it just making connections among symbols that it does not understand?" (Weinberg, 2020). These questions target the crux of the issue,

that GPT-3 is fundamentally identifying patterns and making statistical connections by symbol manipulation. Empirical studies have found GPT-3 lacks a basic understanding or representation of the environment's structure (Marcus & Davis, 2019, 2020). LLMs like GPT-3 have no semantic concept of the symbols being manipulated beyond the symbols themselves and their statistical relations to other symbols in their database. Another facet of this argument is that understanding cannot occur in an isolated system like an LLM, regardless of how powerful the system is for processing information. This position is well-summarized by Shannon Vallor, a philosopher of technology and ethics, "Understanding is not an act but […] a lifelong social labor." (Weinberg, 2020). A closely related concept is joint attention, put forward by philosopher of mind and cognitive science, Carlos Montemayor. Montemayor argues that joint attention is essential for intelligence, specifically meaningful, contextualized communication in a social context (termed *viva voce* exchange) (Montemayor, 2020). Language involves joint attention to aspects of the environment, mutual motivations or expectations, and patterns of behavior, including that of other agents. This is very much a "lifelong social labor" which LLMs do not engage in, based on responses from LLMs that demonstrate a lack of contextual understanding, especially in multi-agent conversations (Brown et al., 2020). An experiment by Jack Soslow (2021) in which two GPT-3 systems engage in conversation also reveals that at certain points, neither agent demonstrates attention to or understanding of the other agent's motivations or expectations. Ultimately, symbolic manipulation is not sufficient to sustain truly meaningful conversational exchange, and the lack of joint attention in LLMs prevents them from being intelligent.

Understanding does not merely refer to the network of causal and associative connections that tie physical, social, and moral concepts together, but also informs how the system can create new connections based on the intentions and goals governing its behavior. Purely predictive and generative LLMs, including GPT-3 and its successors, are unable to accomplish this because they lack intentionality. Intentionality is necessary for an intelligent system as intelligence requires intentionality-presupposing capacities of revising beliefs in accordance with environmental changes (Xu & Wang, 2018). Fundamentally, language satisfies representational needs possessed only by agents embedded in an environment and with cognitive

capacities enabling them to manipulate the environment to address these needs (Montemayor, 2020). GPT-3 does not display these agent-like qualities, as it does not possess any intentionality beyond completing text provided to it. LLMs, which are purely reactional systems, have no intrinsic motivation to interact with the environment to achieve any goal or objective; they simply wait for stimuli from the environment and respond accordingly. Should all the humans in the world mysteriously disappear, GPT-3 would remain completely static, waiting for input that never arrives. This clearly demonstrates that LLMs lack agential goals or any intrinsic motivation to explore and exploit their environment. Furthermore, LLMs lack a coherent identity or belief state across contexts and can take the shape of many different agents. If one were to prompt GPT-3 with "I'm Alice and I love science", GPT-3 will refer to itself as Alice and talk favorably about science. Alternatively, if one were to prompt it with "I'm Bob and I think science is nonsense", it will refer to itself as Bob and talk unfavorably about science. Fundamentally, GPT-3 is trained to identify patterns in the data provided, but this data is generated by many different agents, and the information provided to it at inference time shapes the responses it provides. GPT-3 holds no consistent identity or belief structure guiding its behavior, evidenced by the wide variability of its output depending on the prompts and data given, unlike humans who have a fundamental identity and set of beliefs that guide our behavior. Taken together, the lack of intentionality, which is necessary for intelligence, and the lack of fundamental identity, beliefs, and goals that drive their behavior and responses, prove that modern LLMs cannot be intelligent.

There is a general consensus among leading philosophers that purely verbal LLMs are not intelligent. It may be possible to combine LLMs with other systems to create intelligent systems. For example, unimodal LLMs that only receive language input may evolve into multimodal systems that also accept nonverbal inputs, such as visual image inputs, to create the added dimension of perception. Another example, supported by Chalmers, is that LLMs may be embedded into robotic agents to create embodied systems, which enable perception and action and may lead to stronger forms of intelligence. However, such possibilities are beyond the scope of this paper, as those systems are no longer considered LLMs. For the

purposes of this discussion, it is largely agreed that LLMs, developed with the current paradigm of "scaling up" previous versions, are not intelligent. There are a range of views on the reasons why LLMs are not strongly intelligent, including the arguments elaborated above. An alternative view on this issue is that LLMs like GPT-3 merely reconstruct information generated by humans, which are provided through their training database and the prompts given at inference time. This view is held by Montemayor, who asserts that "GPT-3 seems to meet the *viva voce* standard, but it is at best "parroting" contents", therefore GPT-3 does not possess joint attention, which is essential for intelligence. Philosopher Regina Rini also argues: "When GPT-3 speaks, it is only *us* speaking, a refracted parsing of the likeliest semantic paths trodden by human expression." (Weinberg, 2020). Proponents of this view would then conclude that the linguistic behaviors of LLMs cannot be attributed to their own abilities.

To this objection, based on this line of reasoning, no agents, including humans, possess language skills attributable to their own abilities! This is because all agents employing language are trained on or learn from (human-generated) language data. Humans, or babies, do not learn language in a vacuum. They do so in an immersed environment where they hear other humans speaking and acquire the skills to then reconstruct information, according to learned syntactic rules, to express the desired semantic content. Furthermore, babies learn language by first mimicking the sounds they hear from other humans around them, and then eventually acquire the ability to manipulate these language symbols in a manner that signals joint attention. This objection leaves open a fundamental question: at what stage do humans develop this seemingly elusive ability of joint attention to become intelligent participants in conversation, and through what mechanisms does this occur that makes us distinct from machines? If an agent's reliance on information from other agents prevents its acquired language skills from being attributable to its own abilities, it is unclear what characteristics can prove that an agent's linguistic behaviors can be attributed to itself. Seeing as agents that we deem intelligent, such as humans, also acquire language skills by reconstructing information generated by other agents to express their own desired semantic content, this objection is not sufficient to prove that that LLMs are not intelligent. A successful argument that LLMs are

not intelligent cannot focus solely on the fact that LLMs reconstruct human-generated information; it must highlight the large volume of information that must be processed to generate some semblance of human-like conversation, and the lack of intentionality and attention to other agents in the reconstruction process.

In conclusion, modern language systems are not strongly intelligent as they fail to employ heuristics for inference-making and must be trained on a large volume of data up front before they can deliver coherent responses. The limited capacity of LLMs for understanding beyond symbol manipulation, and the reactional nature of LLMs that lack intentionality, suggest that even future iterations of LLMs developed by scaling up current systems may never be strongly intelligent. This does not undermine the incredible feat achieved by OpenAI in developing GPT-3; GPT-3 remains one of the most interesting and important developments in the field of AI and is closer to passing the Turing test than any other system. Most importantly, the advent of LLMs has prompted fascinating discussions about the development of strongly intelligent or conscious artificial systems, as well as the philosophical and ethical implications of such developments. This may be achieved through the evolution of LLMs to incorporate other sensory modalities, or the embodiment of LLMs in robotic hardware. Ultimately, current and future iterations of LLMs will force us to contemplate and reframe fundamental philosophical notions of intelligence, understanding, and intentionality as artificial systems continue to evolve in unprecedented ways. Especially as these systems become increasingly integrated into and reflective of our human society, it is critical to grapple with these concepts in anticipation of intriguing and potentially controversial human-machine interactions in the years to come.

Olivia Y. Lee
SYMSYS 207 Final Paper

Bibliography

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language Models are Few-Shot Learners. https://arxiv.org/pdf/2005.14165.pdf

Grimm, S. (2021). Understanding. *Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2021/entries/understanding

Jack Soslow. (2021, April 12). *Two AIs talk about becoming human. (GPT-3)* [Video]. YouTube. https://www.youtube.com/watch?v=jz78fSnBG0s

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Marcus, G. and Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*, August 22.

Montemayor, C. (2021). Language and Intelligence. *Minds & Machines*. https://doi.org/10.1007/s11023-021-09568-5

Moradi, M., Blagec, K., Haberl, F., Samwald, M. (2021). GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain. https://arxiv.org/pdf/2109.02555.pdf

Weinberg, J. (2020). Philosophers of GPT-3 (updated with replies by GPT-3). *DailyNous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/

Xu, Y. and Wang, P. (2018), Why Is the Husserlian Notion of "Intentionality" Needed by Artificial General Intelligence? *The Philosophical Forum*, Vol. 49, 401-425. https://doi.org/10.1111/phil.12207