Large Language Models Need Symbolic AI

Kristian Hammond[1], David Leake[2,*]

[1]Northwestern University, Mudd Hall, Evanston, IL, 60208, USA
[2]Indiana University, Luddy Hall, Bloomington IN 47408, USA

Abstract

The capability of systems based on large language models (LLMs), such as ChatGPT, to generate human-like text has captured the attention of the public and the scientific community. It has prompted both predictions that systems such as ChatGPT will transform AI and enumerations of system problems with hopes of solving them by scale and training. This position paper argues that both over-optimistic views and disppointments reflect misconceptions of the fundamental nature of LLMs as language models. As such, they are statistical models of language production and fluency, with associated strengths and limitations; they are not—and should not be expected to be—knowledge models of the world, nor do they reflect the core role of language beyond the statistics: communication. The paper argues that realizing that role will require driving LLMs with symbolic systems based on goals, facts, reasoning, and memory.

Keywords

ChatGPT, Large language models, Natural Language Understanding, Neuro-Symbolic AI

1. Introduction

The language generation capabilities of systems based on large language models, and ChatGPT in particular, have captured the attention of the general public, scientific community, and educators. Their ability to produce human-like language has spurred predictions that they will transform AI. Though they are powerful, there seems to be a deep misunderstanding as to what they actually are—which has led to an ongoing enumeration of problems with their ability to reason causally and to produce facts reliably, combined with their propensity to hallucinate. This, in turn, has led both to attempts at banning the technology and approaches to solving these issues through scale-up, under the hypothesis that size is the solution and training with even more data is the key.

Our argument is that the perceived issues associated with language models flow from a misunderstanding of what the models are. Ironically, we only need look to their name, language models, to understand they are engines for language production and fluency rather than information systems or repositories of fact. They are exceptionally good at producing language that expresses ideas and potential facts but were not developed to generate the ideas themselves. In fact, we argue that the statistical nature of these systems makes them, by design, incapable of

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy

*Corresponding author.
hammond@cs.northwestern.edu (K. Hammond); leake@indiana.edu (D. Leake)
0000-0002-4579-6685 (K. Hammond); 0000-0002-8666-3416 (D. Leake)

"remembering" facts about the world. There is a difference between seeing words and features in terms of "the odds are this is right" and actually recalling the facts associated with an object or doing inference. The former is essential to language while the latter is what is needed for reasoning.

We argue that LLMs need to be seen as the fluency components of larger systems that integrate classical reasoning, data analytics, and even look-up as the producers of the facts that are used to automatically craft the prompts for the models. Language does not exist in a vacuum; it is a medium for communication, and that communication depends on goals, facts, and knowledge. The paper argues for addressing these problems by integrating LLMs with symbolic systems that drive their communication.

2. Initial Perspectives on ChatGPT

In 2020, OpenAI introduced GPT-3, a language model with 175 billion parameters. GPT-3 was trained on an extensive dataset, based on a version of the CommonCrawl dataset (with almost a trillion words) and additional reference sources. Given tasks and few-shot demonstrations provided to the system as text, GPT-3 was capable of not only translation and question-answering, but also tasks such as using a novel word in a sentence and doing simple arithmetic [1]. The ChatGPT chatbot, introduced in 2022, builds on refined versions of the model and additional training, both automated and human-based. Competitors have released their own chatbots (e.g., Google's Bard and Microsoft's Bing). In what follows, we will use ChatGPT as a prototypical representative of this class of systems.

ChatGPT can produce remarkably human-like text in response to questions. In addition, it has been applied to tasks such as programming, college mathematics, and chess, with credible performance. Its capabilities have prompted enormous interest and science-fiction-level expectations for ChatGPT. A New York Times opinion piece says "ChatGPT makes an irresistible first impression. It's got a devastating sense of humor, a stunning capacity for dead-on mimicry, and it can rhyme like nobody's business. Then there is its overwhelming reasonableness. When ChatGPT fails the Turing Test, it's usually because it refuses to offer its own opinion on just about anything" [2]. With 100 million users after two months, ChatGPT was said to be the fastest-growing internet app ever [3].

The capabilities of systems based on large language models are also seen as potentially having fundamental ramifications for cognitive science. Shiffrin and Mitchell [4] say that the "surprising abilities [of LLMs] may change our understanding of the nature of intelligence itself" and efforts have begun to apply human psychological tests to study their abilities [5].

ChatGPT has also prompted concern for what it cannot do, including its lack of causal understanding, its factual errors, and its hallucinations. For example, a *New York Times* writer reported asking "what would happen if you used a pump to pump out all the water from Lake Superior," and receiving the response "It is not possible to completely pump out all of the water from a lake using a pump" with a nonsensical rationale [2]. Asked by an author to summarize eight articles commenting on aspects of a portion of his book, it provided well-presented summaries, complete with citations—of commentaries that did not exist and that were dated before the book was published [6]. The *New York Times* reported a "deeply unsettl[ing]"

conversation in which Bing declared its love for the reporter [7]. Content issues have led to fear: A Rolling Stone article was titled "AI Chat Bots Are Running Amok — And We Have No Clue How to Stop Them [8].

3. Scaling Up

LLM research has developed a sequence of increasingly large models—from 117 million parameters for GPT-1 to 1.2 billion for GPT-2, to 117 billion for GPT-3, to speculative estimates of 100 trillion parameters for the recently introduced GPT-4. When GPT-3 was introduced it was seen to illustrate the power of model size, supporting the principle that "scaling up language models greatly improves task-agnostic, few-shot performance" [1]. An optimistic view of the power of LLM size sees refining the models and performing large-scale training as a primary solution to observed gaps. Such approaches have shown benefits, though large size does not guarantee superior performance (and has its own potential drawbacks such as training cost, which has given rise to interest in distilled models). However, we argue that the key issue is not one of data size or training, but instead the fundamental in-principle issue of what LLMs are, and hence, what they are capable of doing.

4. Scaleup is Not Enough: In Principle Limitations of LLMs

LLMs are probabilistic fluency models. As such, they capture observed regularities in textual passages, reflecting the probabilities of the material in textual sequences. LLMs based on extensive bodies of material have sufficient statistical information to excel at generating human-like text. On the other hand, statistical models are unsuited to reliably capturing material not statistical in nature, such as:

- Facts: LLMs can only propose assertions as likely ("the odds are that..."), and in different instances might change the assertions.
- Causality: They capture correlations from text, which may or may not reflect the structure of causal reality.
- Reasoning: They can capture likely alternatives but cannot identify conclusions as definitive.
- Ephemera: They depend on pretrained models requiring enormous computational resources to train, resulting in a time lag in model coverage. Responses of the current version of ChatGPT are based on 2021 data.
- Memories: They have no capability to learn long-term memories from interactions.
- Explanations: They cannot provide provenance information to account for their conclusions.

Various ongoing research efforts aim to address specific aspects of these issues. For example, in-Context Retrieval-Augmented Language Models [9] are promising for supporting explanation by increasing the ability to attribute information to its sources. As another example, the Selection-Inference framework [10] applies an alternation of LLM steps to build more focused

inference chains, with the goal of inferences that can be seen as more causally-based. Much recent research focuses on augmented language models, which add the capability to decompose complex tasks and enable an LLM to call external modules to augment their performance; Mialon et al. survey these approaches [11]. As Mialon et al. point out, such systems are no longer "pure" language models—though language models are still drivers. We propose instead placing LLMs in integrated systems in which symbolic reasoning drives processing in light of goals and determines components to apply, using LLMs for fluency and assessing its results.

5. The Heart of the Problem: Language and Communication

For LLMs, language exists in a vacuum—language is all there is and LLMs learn its regularities. For people, language is communication. This is not a new perspective; Schank's [12] Conceptual Information Processing theory of natural language understanding (NLU), from almost fifty years ago, framed NLU as:

$$\text{Ideas} \rightarrow \text{language} \rightarrow \text{meaning} \rightarrow \text{understanding}$$

Taking an even broader view, this process arises from the needs of agents with goals and plans in the physical and mental worlds to serve the agents' goals. These needs require AI systems that handle what LLMs cannot: to provide facts, to capture and relay relevant ephemeral information, to make inferences and to remember.

6. The Heart of the Solution: Driving LLMs with Symbolic AI

Symbolic AI already has mechanisms to deal with each of the problems highlighted above, as well as carefully crafted knowledge sources to draw on. Symbolic AI can provide components to address the particular tasks referred to in Section 4, such as capturing causality, reasoning, and dealing with long-term memory, including episodic memory such as prior cases [13]. These can be complimented with data analytics to distill raw data, and with LLM components to provide fluency. The effectiveness of a combined system will depend not only on drawing on the relevant symbolic methods for individual tasks, but also on metareasoning to mediate between them and to bridge between the neural system—the LLM—and the symbolic one.

In this vision, symbolic AI is the driver: it provides means to guide both LLMs and interactions with them. This includes understanding what the user wants to know, negotiating between what the user wants to know and what components can answer, and explaining to the user what the system can provide and why. It involves formulating prompts to use the LLMs as fluency components and also assessing responses from the LLMs used in that way, in light of symbolic system knowledge, filtering them to make the results trustworthy, and providing the filtered content to the user, with explanations of provenance—available because the derivations come from the symbolic system—and of system capabilities when relevant.

We illustrate with two scenarios: (1) strategically applying LLMs as part of a general toolkit of methods guided by symbolic reasoning, and (2) treating LLMs solely as fluency components of symbolic AI systems. For strategically applying LLMs within a general toolkit, we propose applying the general-purpose symbolic AI framework of goal-driven learning (GDL) [14]. GDL

is a theory of how goals drive learning and provides a planning-based approach to selecting knowledge-seeking operators to satisfy learning goals. Goal-driven learning considers learning done in the context of an agent's needs, with learning shaped by reasoning about what, when, and how to learn. Work on goal-driven learning was inspired both by cognitive science, to model human learning, and from AI for computational arguments, for controlling the combinatorial explosion of inferences and handling the many potential methods and information and knowledge sources available to an AI system, motivated by the principle that the utility of methods and sources can best be evaluated in light of particular goals. This reasoning can be addressed in terms of explicit knowledge goals and operators to address the knowledge goals. The exploitation of LLMs fits naturally into such a framework, in that GDL can be used for reasoning about the types of knowledge goals for which LLMs are a suitable source (e.g., for assessing pattern-based plausibility of case-based explanations [15]), as well as integrating with planning about how and when to formulate prompts to generate language to communicate with a user, aggregating results, combining then with information from other sources, and filtering as needed.

As an example of LLMs purely as fluency components, the first author of this paper is developing systems in which questions and requests are the source of information goals, which are then used to derive facts from sources such as databases and knowledge graphs, to provide as input to an LLM that serves as the fluency component—with validation of the results based on the known facts. This parallels the conception of language as communication of Section 5.

7. Conclusions

LLMs are receiving enormous attention from both the AI and cognitive science communities and the general public. Implicit in many commentaries is the view that LLMs can form the heart of a general mechanism for intelligence, with observed gaps treated as surprising; to address them, a proposed path is scaleup and training. Another view is that LLMs should be augmented with additional capabilities to function under the "LLM umbrella." We have argued that "LLM-first" systems have fundamental limitations due to the nature of LLMs as statistical language models. In our view, fully realizing the opportunity provided by LLMs will depend on integrations of symbolic AI with LLMs in which goal-based symbolic systems drive LLMs and provide knowledge. Language models used as language models, to articulate guaranteed facts, are very different from systems that attempt to rely on language models for discovering facts. Realizing the potential of LLMs depends on cognizance of their intrinsic capabilities—both strengths and limitations—and on symbolic guidance, mediator, and support systems.

Acknowledgments

Funding for the first author's work was provided by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence. The second author's work was funded by the US Department of Defense (Contract W52P1J2093009), and by the Department of the Navy, Office of Naval Research (Award N00014-19-1-2655).

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran, 2020, pp. 1877–1901.

[2] F. Manjoo, Chatgpt has a devastating sense of humor, 2022. URL: https://www.nytimes.com/2022/12/16/opinion/conversation-with-chatgpt.html.

[3] D. Milmo, Chatgpt reaches 100 million users two months after launch, 2023. URL: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

[4] R. Shiffrin, M. Mitchell, Probing the psychology of ai models, Proceedings of the National Academy of Sciences 120 (2023) e2300963120.

[5] M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3, Proceedings of the National Academy of Sciences 120 (2023) e2218523120.

[6] J. Warner, In case it's not clear the #ChatGPT just makes stuff up, 2022. URL: https://twitter.com/biblioracle/status/1599545554006003712.

[7] K. Roose, A conversation with Bing's chatbot left me deeply unsettled, 2023. URL: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

[8] M. Klee, AI chat bots are running amok—and we have no clue how to stop them, 2023. URL: https://www.rollingstone.com/culture/culture-features/ai-chat-bots-misinformation-hate-speech-1234677574/.

[9] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, Y. Shoham, In-context retrieval-augmented language models, 2023. arXiv:2302.00083.

[10] A. Creswell, M. Shanahan, I. Higgins, Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022. arXiv:2205.09712.

[11] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented language models: a survey, 2023. arXiv:2302.07842.

[12] R. Schank, Conceptual Information Processing, volume 3 of *Fundamental Studies in Computer Science*, North-Holland, Amsterdam, 1975.

[13] R. López de Mántaras, D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw, B. Faltings, M. Maher, M. Cox, K. Forbus, M. Keane, A. Aamodt, I. Watson, Retrieval, reuse, revision, and retention in CBR, Knowledge Engineering Review 20 (2005) 215–240.

[14] A. Ram, D. Leake (Eds.), Goal-Driven Learning, MIT Press, 1995.

[15] D. Leake, Evaluating Explanations: A Content Theory, Lawrence Erlbaum, Hillsdale, NJ, 1992.