SHADOW-CAVE MODELS: HOW PLATO'S ALLEGORY ILLUMINATES LIMITATIONS OF LARGE LANGUAGE MODELS

IN SUBMISSION TO THE PHILOSOPHY OF DEEP LEARNING CONFERENCE 2023

Muhammad ElNokrashy [*] [†]
ATL Cairo, Microsoft Egypt
muhammad.elnokrashy [at] microsoft.com

Badr AlKhamissi [*]
Independent
badr [at] khamissi.com

January, 2023

ABSTRACT

This work outlines some limitations of Large Language Models (LLMs) and connects them to views on the philosophies of knowledge and perception, and then dubs them Shadow-Cave Models. One explanation is proposed for the astonishing performance of LLMs in the early 2020s in logic and knowledge tasks, even beyond the domain of language modeling. We argue that the mechanism which allows LLMS to perform well on some tests for these capabilities is also limitation towards higher understanding abilities.

Keywords artificial intelligence · philosophy · large language models

In Plato's Allegory of the Cave [1], prisoners are bound facing the wall of a cave and can only perceive the external world through shadows of objects projected on the cave wall. They think of these shadows as the ultimate reality and are unable to conceive of any other possibilities.

In a similar vein, large language models (LLMs), like the GPT family [2], are trained on an extensive body of textual data gathered from the world wide web. The LLMs acquire knowledge of the world through a stream of tokens, comprising words and sentences, akin to how the prisoners in Plato's allegory perceive reality through shadows.

The *world of shadows* is, on average, expressive, internally consistent, and informative. The objects within can possess persistent identities, exhibit identifiable behavior, and follow a consistent logic of relations and interactions among them.

From the reader's perspective, we know what objects the shadows in the cave correspond to, and their properties and relations. We can claim that the mapping from the real objects to the shadows is *reasonable*. Although limited, the world of shadows is well-grounded and can be useful, but tells only an incomplete story.

Similarly, LLMs have no direct connection to the world which their training texts depict. They learn from data that are, on average, grammatically accurate, consistent in their depictions and arguments, and informative as a source of world descriptions. The mapping from the world to the text is reasonably accurate and informative, as based on the world models of the human writers. However,

[*] Equal Contribution
[†] Corresponding author

similar to how shadows are a reduced two-dimensional representation of three-dimensional objects, tokens are limited representations of the writers' world models.

As the LLMs train to predict the expected completion to an input text, they learn to, on average, produce grammatically accurate sentences, utilize proper terminology, maintain internal consistency, and make valid arguments and logical connections within its generated statements. *On average.*

What are the actual capabilities of LLMs and why could they emerge?

The LLMs, the prisoners in the *world of tokens*, can recognize the patterns in the texts they learn from, and the patterns reveal more knowledge than *just* the structure of the bare language. The text and its tokens, like the shadows and the slices of time in which they interact, depict a *sketch* of objects, and enjoy similarly observable identities, behavior, and logic.

For instance, LLMs have been shown to achieve high performance on benchmarks that were built to assess commonsense knowledge with regard to various aspects such as reasoning about physics or social situations [3, 4]. This led people to speculate that such models possess a certain degree of comprehension of complex concepts, despite being trained solely through language.

The text, by virtue of the information it carries on a topic, imbues the language as seen by an unthinking language model with the logic of the underlying objects. For example, the following would be a linguistically valid statement that violates the Physics domain expectations: "If you release an apple from your hand while standing on Earth's surface in open air it will float up." The rules of the domain are *implicitly encoded* as a language subset in the *language modeling objective* based on the different vocabulary and sentence distributions across domains. This learning, then, may not lead to consistent performance when queried for precise outputs.

An LLM, on average, is not incapable of learning the higher-level patterns of that logic-imbued sub-language within a domain. They are indirectly grounded by perceiving shadows of the world through the humans' texts and world model. The perception of these ***Shadow-Cave Models*** is limited enough to prevent developing broad understanding, but capable enough to put on a good show, *in the shadow world*.

Lest we celebrate too early, we remind ourselves that the allegory applies, even in this stripped-down version, to our own perceptions [5]. We note this and add the following statement: *The world as we* perceive *it is the world as we* interact *with it, is the world as we* model *it*. Sometimes, we may need the models we build to be able to see the world as we do.

References

[1] Plato, *The Republic*. 1994.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. 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. M. 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, and D. Amodei, "Language models are few-shot learners," 2020.

[3] M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi, "Social IQa: Commonsense reasoning about social interactions," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 4463–4473, Association for Computational Linguistics, Nov. 2019.

[4] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[5] D. D. Hoffman, *The Case Against Reality: How Evolution Hid the Truth from Our Eyes*. W. W. Norton & Company, 2019.