Ontological Status of LLMs

Contents

Statistical Nature of LLMs

- LLM is a neural network with hundreds of millions to billions parameters that is pertained on a volume of text no human being can ever consume in a lifetime. The goal of the training is to approximate the data distribution
  - LLM is a function that takes an input string and returns the most probable output string
  - LLMs are very sensitive, wording can alter the output [5]
  - LLMs are not continuously updating and cannot produce new knowledge [5] (?)
  - => LLM is a complex statistical model of how the words and phrases in its training data correlate [1]
- LLMs have no biological properties of humans
  - LLMs have no (innate) motivation or goal, prior experience etc.
  - LLMs have no embodiment, which is argued to be one of the prerequisites for intelligence; they cannot see and touch objects, cannot move etc.
  - => They lack reference, a key for inferring the meaning (is it?)
  - => LLMs only infer the form of the language but not the meaning [1]

But what about their performance on the tests for Natural Language Understanding (NLU)?

Statistical Nature of LLMs

- NLU benchmarks do not always require actual understanding [1]
  - LLMs utilize *shortcut learning*—a phenomenon that describes learning that relies on leveraging low-level co-occurrence patterns in the data (overlapping tokens and such)
  - LLMs use shortcuts in order to perform well on a particular benchmark [1, 5]
  - It is often the case that as soon as the obvious patterns are removed, the model's performance drops to chance levels [3]
- => LLMs can be "right for the wrong reason" [3]

=> Opinion: LLMs are *stochastic parrots* [6]

Formal vs Functional Linguistic Competence

- LLMs only infer the form of the language but not the meaning: while they are able to generate grammatically correct, humanlike language — they still lack the conceptual understanding
- Neuroscience provides evidence that even human processes the form and the meaning of the language differently [3]
- *Formal linguistic competence* — the knowledge of rules and statistical regularities of language, vs *functional linguistic competence* — non-language-specific cognitive functions that are required when using language in real-world circumstances

Formal vs Functional Linguistic Competence

- LLMs exhibit (near) human-level formal competence (at least in English) but „patchy“ functional competence [3]
- In human brains, the *language network* is responsible for formal linguistic competence — it responds when people comprehend or generate sentences, but not when they perform reason about people’s mental states, process non-verbal communicative information etc.
- The functional linguistic competence is supported by different regions of the brain: among others, by *multiple demand network* that is reliable for logic, reasoning, and a lot more, and by *default network* that tracks both linguistic and non-linguistic narratives over a long period of time (the *language network* in humans does not appear to track structure above the clause level)
- For example, despite the nearly complete loss of linguistic abilities, persons with severe aphasia can have normal non-linguistic cognitive abilities [3]

Formal vs Functional Linguistic Competence

- „Give language models a break!“: given a strict separation of linguistic and non-linguistic capabilities in the human mind, we should evaluate these capabilities separately [3]
  - Counter-argument: individuals with aphasia can be tested for their non-linguistic cognitive abilities (e.g. composing music), LLMs not [7]
- LLMs and the human *language network* exhibit non-trivial similarities [3]
  - LLMs learning features like POS, NER, and semantic roles at various layers [3]
- => Opinion: it might be beneficial to not try to train LLMs for functional abilities but rather *augment* LLMs with specific modules since „language and formal reasoning are distinct cognitive capacities that work best when they are supported by separate processing mechanisms“ [3]

However, since the language inputs contain wealth of information about the world, and language is „a crucial data source … for much of people’s world knowledge“, LLMs can still gain functional linguistic competence [3]

Functional Linguistic Competence of LLMs

- A prominent point of view on the human intelligence says that humans model the world in concepts. Thus for a „real“ understanding of the human language, a system (be it a human or an LLM) should map language inputs to concepts and manipulate those
- It is argued that embodiment and ability to couple form and reference are essential to acquire such concepts
- Moreover, for a proper understanding of other human beings via language, we should model which concepts they do have in their minds (theory of mind)
- However, LLMs have no mental models of the world [1], and the generated text is not grounded on any model of the reader’s state of mind [6]
  - => Opinion: since LLMs are not able to build conceptual models, they cannot infer meaning from the text; humans may believe LLMs „really“ understand what they generate only because of the human tendency to attribute meaning to things whether it is there or not; *coherence is in the eye of the beholder* [6]

However, if we consider this argumentation, it stands on a shaky ground

Embodiment & Reference

- The ability of reference — linking a sequence of signals (text in case of LLMs) to an object in the real world — is argued to be decisive when it comes to a „real understanding“
- Example: an octopus learns to use words correctly by eavesdropping on a conversation between two people on land will not be able to determine which object is a coconut, even though it knows how to use the word [2]
- Opinion: since the training data for LLMs is only form, they do not have access to meaning [6] (just like that octopus)
- Opinion: reference does **not** define meaning
  - Many terms that are meaningful to us have no real referent [2]
  - There are terms that cannot have a referent [2]
  - Absence of embodiment — which is claimed to be a precondition for being capable of reference — does not always hinder this ability

Embodiment & Reference

- Embodiment — existence in the real world, groundedness, and contact with physical things — is thought to be a precondition for intelligence [4]
- However, evidence from both humans and LLMs support that the lack of embodiment does not always hinder reference
  - Helen Keller, who was both blind and deaf, had a color scheme that she inferred from smells and touches; she projected one dimension on another [4]
  - Experiments show LLMs trained only on text can recover key geometry of color space, navigation, shapes [2, 4, 7]: e.g. text-only LLMs can draw und understand what they are drawing [5]
  - While LLMs cannot actually see or perform actions, they can do so via a surrogate [5]
  - Given a textual feedback from the environment, LLMs can adjust their generations accordingly [5]
- => LLMs end up learning a „great deal of embodied knowledge“ [4]
- Imagine an world with another physics we cannot access; given textual description detailed enough, we can understand how things work there and can imaginary „live“ there [4]

Internal States

- All said above makes us conclude there should be something more behind meaning than just reference
- Opinion: the interrelation of concepts is the key to meaning, and reference is just one optional aspect in world conceptualization [2]
  - When the associated terms shift in meaning, the resulting meaning adjusts [2]
  - Thus, a concept builds on conception of how the underlying pieces relate [2]
  - => the search for meaning should focus on understanding the way that the systems' internal representational states relate to each other [2]
- LLMs can manipulate complex concepts (but only in a linear manner) [3, 4]
- fMRI evidence supports that the human mental models for representational geometry are similar to LLMs [2] (?)
  - Cf. embedding and their geometrical meaning [7]
- LLM's internal state has some notions of conceptual role, so LLM's utterances have the semantic intent corresponding to these roles. [2]
- => Opinion: LLMs likely already share the foundation of how our own concepts get their meaning [2]

Theory of Mind and LLMs

- Theory of mind (ToM) is the ability to attribute mental states (emotions, intentions, knowledge) to oneself and others, and to understand how they affect behavior and communication.
  - When communicating with people, we build a partial **model** of who they are and what common ground they share with us, and use this to interpret their words [6]
  - Human language takes place between persons who share common ground [6]
- Opinion: just as LLMs can *model* a concept of color, shape etc. only from text, it can also share concepts with the human and succeed in ToM
  - Text provides such clues because humans generated the text [2]
  - When „talking“ to an LLM, it keeps record of its generated preferences, it tries to „understand“ your mood, intensions etc. [4]
  - Tests reveal modern LLMs have a high ToM [1, 3]

Theory of Mind and LLMs

- We are not aware of what actually happens in our own brains; moreover, experiments reveal that we often make decisions well before we think we do [4]
- Thinking about the future involves carrying out something like an inner dialogue, with an “inner storyteller” proposing ideas, in conversation with an “inner critic” taking the part of your future self. [4]
  - => We apply ToM on ourselves, and that is what we call consciousness
  - Some people also apply ToM on LLMs [5, 7], which might be due to the human nature [6]
- But if we assess our **unaccessible** internal states ourselves and say we are conscious and intelligent, how can we state for sure that the same thing is not going on in **unaccessible** internal states of LLMs? After all, if you ask an LLM what it is, it can define itself (apply ToM?)

=> Opinion: LLMs have already achieved some key aspects of meaning [2]

New Form of Intelligence?

- When applying tests designed for humans to LLMs, interpreting the results can rely on assumptions about human cognition that may not be true at all for these models => anthropocentric bias [1, 3]
  - => Opinion: intelligence and understanding “are the wrong categories” for talking about LLMs [1]
  - Open question: would it make sense to see the systems’ behavior not as “competence without comprehension” but as a new, non-human form of understanding? [1]
- Intelligent thought could be a mosaic of simple operations that, when studied up close, disappeared into its mechanical parts (c) Max Newman [4]
- „Real“ language understanding and intentionality consist of attributions of **unobservable** mental states [7] and it is unclear how we can meaningfully test for the „realness” of thoughts, feelings etc. [4]

=> Opinion: we need better measures for evaluating thought competence in LLMs before we can draw conclusions [7]

References

[1] The Debate Over Understanding in AI's Large Language Models, Santa Fe Institute

[2] Meaning without reference in large language models, UC Berkeley & DeepMind

[3] Dissociating language and thought in large language models, The University of Texas at Austin et al.

[4] Do Large Language Models Understand Us?, Google Research

[5] Sparks of Artificial General Intelligence: Early experiments with GPT-4, Microsoft Research

[6] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜, University of Washington et al.

[7] Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding, Leiden Institute of Advanced Computer Science & Leiden University Medical Centre