What can we learn from stochastic parrots? A case for involving Large Language Models in cognitive science

Jaroslaw R. Lelonkiewicz, University of Valencia

Large Language Models (LLMs) are artificial intelligence systems capable of handling a wide range of tasks in a strikingly human-like manner. In cognitive science, there is an ongoing discussion whether LLMs are indeed like humans or more like parrots imitating behaviour without replicating the exact mental network behind it. I argue that LLMs can be useful scientific tools regardless of the outcome of this debate. Much like in case of research with non-human animals, be it parrots, baboons or rats, LLMs can be used as models of selected aspects of cognition, allowing scientists to narrow down the space of possible hypotheses through experimentation. The final necessary step is to replicate the machine findings in humans, thus confirming the validity of the new insights. In return, LLM experimentation offers powerful new research tools and an unprecedented speed of data collection.

In my talk, I review what is known about the behaviour and the cognitive architecture of LLMs and highlight the areas that are most promising for investigation. I also present the new research tools available in LLM experimentation. Finally, I illustrate my point by reporting on my recent work in which machine data uncovered a novel factor driving language processing in humans.

PI SUMMARY:

The arrival of Large Language Models (LLMs) has transformed AI. Once limited to tasks such as chess or image classification, AI now appears to meet some of the hallmarks of general cognition, a feat previously achievable only by humans. But to what extent are these models actually similar to the human mind? There is a vigorous debate about this and for good reasons.

Proving that LLM architectures resemble human cognition would have serious consequences for psychological theory. For example, Piantadosi et al. note that LLMs speak against nativism, in the sense that they demonstrate that the cognitive processes necessary to comprehend and generate language can be acquired based on exposure.

Moreover, *in silico* models of human cognition could greatly facilitate psychological research. LLMs allow to collect data with an unprecedented speed and offer new and intriguing ways of probing the internal cognitive structure. Fedorenko et al. propose that LLMs can grant insights into the structure and functionality of the human language system (e.g., brain signal can be predicted based on LLM representations).

On the other side of this debate are the sceptics, most notably Chomsky et al., who present theoretical arguments and behavioural examples that raise doubts about extrapolating from AI.

I argue that the best way to assess these proposals is to integrate LLMs as (another) tool in psychological inventory and gather the much-needed data. In particular, I present a research pipeline which allows to robustly establish the degree of similarity between humans and AI.

To support my case, I report on my recent statistical learning study where the exploration of machine responses yielded a hypothesis which later was confirmed in experiments with human participants. I also make a reference to a tutorial created by me and my colleague, which provides the code and technical knowledge necessary for running human experimental paradigms with LLMs.