PTHG-24 Panel: Have Chatbots Reached the Holy Grail?

LLM capabilities for constraint solving, what is the holy grail?

Tias Guns

KU Leuven

If you take an arbitrary CP problem and ask an LLM like ChatGPT to model it, you will likely ask it to model the sudoku problem and it will succeed. Our ongoing research shows that sudoku is one of the only problems it can solver 0-shot, because the choice of problem is not arbitrary at all. What problem descriptions are those that we would like LLMs to be able to model? How do we automatically evaluate that the model it produces (with its choice of decision variables) is semantically equivalent with the description? How to handle parameters/data? Should it also be able print/visualise solutions? How can we evaluate interactive modeling? Who is the target audience (individuals, starting modellers, expert modellers, human planners)? What is the scale of problems considered? Does modeling and solving a problem specification even enforce a specific solving technology, and should the answer be optimal; and if so what with very large scale problems? These are some of the main challenges we see, and they could benefit a wider discussion in the community.

Many thanks to Dimos Tsouros, Kostis Michailidis and Serdar Kadioglu for the discussions.

Have Chatbots Reached the Holy Grail?

Thomas Schiex

Universite Fédérale de Toulouse, ANITI, INRAE

The biggest surprise in AI these last years is the unexpected impact of LLMs on the general public. Generative Pretrained Transformers LLMs are routinely presented as being « intelligent » or capable of reasoning. But their original training objective is disappointing in this respect: learn to predict the next token from the *n* previous tokens. Essentially, LLMs are sophisticated regurgitators, reproducing complexly memorized text.[1]. With the right context (prompt), it’s not too difficult to make them look staggeringly stupid. This is harder with the last LLMs, such as GPT4o, that eventually resort to code generation when the question looks logical or numerical. This helps of course, but does not solve the problem of generating finicky, efficient or — obviously — new original suitable code.

So, LLMs have apparently not reached the Holy Grail. Nevertheless, as we continue to feed them well-annotated code, their ability to generate (potentially buggy) variations in the appropriate contexts will undoubtedly improve. However, it's hard to imagine a future where every possible context and reasoning problem is exhaustively covered and solved by LLMs. With the world constantly evolving, new and complex problems will emerge that LLMs have never encountered.

Interestingly, LLMs have also found applications in restricted fields like protein design. PLMs (Protein Language Models) have been trained on the massive amount of protein sequences that is continuously being produced by sequencing and, more slowly, by Natural evolution. So far, PLMs have mostly been able to propose sophisticated variations of existing known protein families[2]. Even in this restricted domain, their capacity to generate proteins that would be truly original and interesting (having a new function) remains, for now, an unfulfilled promise. As for LLMs, this does not make them useless. They are amazing! But just not a panacea.

I believe that modern AI needs a more nuanced approach, blending data-driven intuition (System 1) with rigorous logical reasoning and planning (System 2). This integration is essential to get closer to what we usually recognize as intelligence[3]. This could take some time.

[1] GPT4 is supposed to use around 1.76 trillion parameters and LLM's training sets typically contain around 15 trillion tokens.

[2] ESM3 recently designed a new Green Fluorescent Protein which is « only » 51% similar to existing GFPs. A few years before, a similar feat had been achieved using much simpler Graphical probabilistic Models on another enzyme (paper).

[3] Something that Poincaré already recognized as available in different proportions in early XXth-century mathematicians for example. See chapter 1 (*L'intuition et la Logique en mathématiques*) of his book, La valeur de la science (it's free).

Have Chatbots Reached the Holy Grail?

Christopher Stone

School of Computer Science, University of St Andrews, UK

The fundamental leap in chatbots attempting to be the "Holy Grail" is the attempt itself, which produces non-null results. Before, we only looked at parts of a potential system that would eventually be end-to-end, but systems that generate an actual result, however incorrect, were never presented.

Yet, the fact that LLMs can not handle problems beyond simple examples or cases very similar to those seen during training makes many dismiss the achievement almost entirely. Instead, we should reflect on what is required to be the "Holy Grail"; how complex must the input problem be? How many mistakes do we tolerate? How do we measure how far a given system, LLM or not, is from reaching the threshold? More rigorous evaluation protocols should replace the approach of posing an excessively complex problem to make the system stumble in order to dismiss it.

Most importantly, the question is, “Where do we go from here?” One natural path is to improve LLMs with specialised training and modifications; another is to look for the successor of LLMs, which have several internal structural limitations, such as being feedforward-only and having a fixed vocabulary of input-output tokens.