Towards machine thinking: Biomedical reasoning in large AI language models

Assoc.Prof. Dr. Matthias Samwald

Institute of Artificial Intelligence, Medical University of Vienna

🌐 https://samwald.info/ — ⚡ matthias.samwald (at) meduniwien.ac.at

Topic for: Masterarbeit, Projektstudium, KfK-Praktikum, KfK-Seminar
(💰 There are options for financial support of outstanding master theses!)

The project you will join:

— We work with cutting-edge AI systems operating on natural language (large language models such as GPT-3).
— We focus on chain-of-thought reasoning, i.e., we teach AI systems to verbalize their ‘thinking’ before making predictions. We expect that this makes the AI systems more robust and their inner workings more explainable.
— We create tools and datasets for implementing and evaluating chain-of-thought reasoning on biomedical problems (e.g. scientific hypothesis generation, medical question answering).
Figure 1: Three examples of chains-of-thought and answers generated by the large GPT-3 language model text-davinci-002 for a question from the United States Medical Licensing Examination (USMLE) exam, adapted from (Liévin et al., 2022). The first chain-of-thought leads to the prediction of the correct answer, while the others lead to incorrect answers. Errors in reasoning or ‘hallucinations’ of incorrect facts still limit predictive performance and validity of explanations.
**Background:**

Recent large language models (LLMs) based on deep learning have shown impressive results on a wide variety of tasks. LLMs are pre-trained with a large amount of text data and are then usually further adapted to more specific tasks and capabilities. Examples for such models include BERT (Devlin et al., 2018), T5 (Raffel et al., 2019) and GPT-3 (Brown et al., 2020).

GPT-3, a very large model with 175 billion parameters, has demonstrated remarkable ability to generate text that is both realistic and coherent, as well as great performance on a broad spectrum of tasks, despite not explicitly being trained on them (Ouyang et al., 2022).

However, despite this ability, LLMs are still limited in several ways:

1. They often fail to produce accurate predictions due to their **inability of complex reasoning**, such as question answering tasks requiring multi-step reasoning.
2. They are **black boxes**, making it difficult to understand how they generate predictions.
3. They are prone to ‘**short-cut learning**’ (Geirhos et al., 2020), i.e., they sometimes “**predict the right answers for the wrong reasons**”. This means that models may perform well when trained and validated on data from one data distribution, but fail disastrously when applied to novel data with different characteristics.

These limitations severely limit the application domains of LLMs and have the potential to cause harm, as **lack of explainability and robustness can lead to critical failures and biases when these models are deployed in practice**.

One recently proposed approach for enabling complex reasoning and generating explanations with LLMs is to force models to **explicitly verbalize reasoning steps as chains-of-thought in natural language** (Wei et al., 2022; Kojima et al., 2022). With this approach the LLM is prompted to not only output an answer, but also an explanation of how it derived its answer in the form of an internal monologue.

Figure 1 shows examples of prompting an LLM in a few-shot setting to generate multiple chains-of-thought for answering medical questions (Liévin et al., 2022).
Bibliography


