

REASONING IN LARGE LANGUAGE MODELS

HUMAN REASONING VS LOGICAL REASONING

ANA OZAKI AND EMIL POIESZ

1. MASTER THESIS TOPIC

Large Language Models (LLMs) have been shown to struggle solving problems using classic logical deduction [3], instead learning “shortcuts”, in particular, unintentional statistical features [1]. It has also been noted in the literature that human-like reasoning differs from classical logic reasoning [2]. Since LLMs are trained on human generated text we would like to explore to what degree LLMs reason like humans do. Some logical sentences are un-intuitive to those with little experience with mathematical logic. Given the following logical sentences

All B are A

All B are C,

a number of high school and university students inferred that *All A are C* even though this is not supported by logical reasoning (Figure 1) [2].

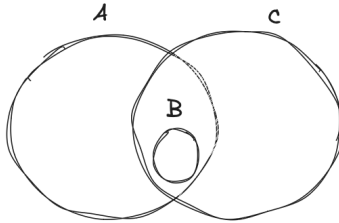


FIGURE 1. Counterexample for *All A are C*

Humans tend to make assumptions that are not explicitly stated. To a human, the sentence

Some A are B *implies* *Some A are not B*

while logically *Some A are B* does not exclude *All A are B*. These assumptions in human reasoning have been better modelled by Weak-Completion-Semantics [2] than by mathematical logical models.

2. MAIN TASKS

In this project, the student will investigate the reasoning capability of LLMs. In particular, the goal of the thesis is to explore whether LLMs replicate similar reasoning patterns as humans, using studies on Weak-Completion-Semantics as the basis for human reasoning. The student is tasked to:

- Read relevant articles about human reasoning and LLMs.
- Create a synthetic dataset that is able to test LLMs both on human-like reasoning and on logical reasoning.
- Implement a pipeline that tests LLMs on the synthetic dataset.
- Evaluate and analyze the responses to gain insight into how LLMs reason.

REFERENCES

1. Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann, *Shortcut learning in deep neural networks*, Nature Machine Intelligence **2** (2020), no. 11, 665–673.
2. Ana Oliveira da Costa, Emmanuelle Dietz, Steffen Hölldobler, and Marco Ragni, *Syllogistic reasoning under the weak completion semantics*, 07 2016.
3. Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck, *On the paradox of learning to reason from data*, 2022.