From Calculation to Adjudication: Examining LLM judges on Mathematical Reasoning Task

Andreas Stephan, University of Vienna, Vienna/Austria

Abstract: A main reason for the current success of large language models (LLMs) is their ability to perform zero-shot reasoning, or in other words, the ability to solve tasks without explicit training data. To reduce the need for human annotations, large language models (LLMs) have been proposed as evaluators, or judges, of the quality of other candidate LLMs. In this talk, we study LLM judges on mathematical reasoning tasks. These tasks require multi-step reasoning, and the correctness of solutions is verifiable, enabling an objective evaluation. In summary, this talk presents 1) a detailed performance evaluation of LLM judges on mathematical reasoning tasks, 2) an investigation of regularities, such as intriguing correlations, in the judgement process and 3) the usage of textual features to analyze those.

Bio: Andreas holds a M.Sc. in Mathematics and B.Sc.'s in Mathematics and Computer Science from the Technical University of Munich (TUM). Currently, he is in the final phase of his PhD in Natural Language Processing at the University of Vienna, where he is supervised by Prof. Benjamin Roth. Previously, he was working as an NLP Data Scientist in the insurance and financial domains. In his research, he focuses on the analysis and usage of multiple incomplete or noisy sources of information in the NLP domain, for instance regular expressions-based rules, multi-modal setups or as today, multiple large language models (LLMs). He is broadly interested in NLP and its role in the development of AI systems (see http://andst.github.io).