UNIVERSITÄT BIELEFELD
Faculty of Technology

Master Thesis

Benchmarking Social Intelligence in Large Language Models

Semantic Computing Group
Jan-Philipp Töberg
jtoeberg@techfak.uni-bielefeld.de

Cognitive robots are challenged by unknown situations in open worlds. They cannot perform everyday tasks like cutting food or pouring drinks without encountering unknown motions, objects or environments. To mitigate this problem, providing these robots with commonsense knowledge is a possible way to go [1]. One way of doing so is by using Large Language Models (LLMs) during the manipulation either as task planners or as sources supporting a task planner. We are currently working on benchmarking different aspects of commonsense that can be relevant for robots interacting in household environments.

One recent approach for evaluating social intelligence of AI agents is the *Watch-And-Help* challenge [2], where an AI (i) watches a demonstration by a human to infer the underlying goal before (ii) cooperating with this human in an unknown environment to achieve the same goal. This challenge takes place in a simulation environment.

In this thesis you develop an approach that can fulfil the Watch-And-Help-Challenge using LLMs to benchmark their capabilities in handling social and cooperative interactions. Possible questions you need to work on are the following:

- How should the LLM be prompted? What prompting techniques perform best?
- How can the outputs from the simulation be incorporated? How can the LLM output be translated into actions the agent can take?
- Can the simulation be extended to cover failure cases and their handling? Can the simulation be changed to have agents with different capabilities (e.g. robot with a single arm)?

No prior knowledge regarding LLMs or prompting techniques is required. Since the simulation is written in Python, you should be familiar with it. The thesis should be taken in English but can also be taken in German.

Related literature

[1] J.-P. Töberg, A.-C. N. Ngomo, M. Beetz, and P. Cimiano, 'Commonsense knowledge in cognitive robotics: a systematic literature review', Front. Robot. AI, vol. 11, 2024, doi: 10.3389/frobt.2024.1328934.

[2] X. Puig et al., 'Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration', in International Conference on Learning Representations (ICLR), Vienna, Austria, 2021. doi: 10.48550/ARXIV.2010.09890.

The Semantic Computing Group researches and develops methods that enable machines to acquire relevant knowledge as well as linguistic capabilities. Using methods from *natural language understanding* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading unstructured textual data. In particular, the group focuses on methods for information extraction, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at: `http://www.sc.cit-ec.uni-bielefeld.de/`

Interested? @mail to `jtoeberg@techfak.uni-bielefeld.de`