

Bachelor or Master Thesis

Evaluating Commonsense Capabilities Through Prompt Engineering

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

Intelligent agents 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 increase their world understanding and support their planning capabilities [1]. However, this household-specific commonsense knowledge is, despite its potential, not often benchmarked or compared between different approaches.

As a solution, we are proposing **RoboCSKBench**, a language-based multi-task benchmark for evaluating embodied commonsense capabilities in embodied agents [2]. We employ this benchmark to evaluate different state-of-the-art LLMs like Llama 3.3, Gemma 2 or GPT-40 with regard to their capabilities in reasoning about this *embodied commonsense*. However, in its initial proposal, we have not tried out different prompt engineering techniques but focused on a single, role-based prompt for each task.

In this thesis, you will try different prompting techniques and compare their results (see [3]). Important research questions are the following:

- What prompting techniques are feasible for the available data? How can they be implemented?
- How can the role-based prompt be optimised using frameworks like DSPy [4]?

No prior knowledge regarding is required. Regarding the programming language, it is advised to use Python. The thesis can be taken in English or German. The scope of the thesis will be adapted depending on whether you are a bachelor or master student.

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] J.-P. Töberg, S. Kenneweg, and P. Cimiano, 'RoboCSKBench: Benchmarking Embodied Commonsense Capabilities of Large Language Models', in UR2025, College Station, Texas, USA, 2025.

[3] S. Schulhoff et al., 'The Prompt Report: A Systematic Survey of Prompting Techniques', 2024, arXiv. doi: 10.48550/ARXIV.2406.06608.

[4] https://dspy.ai/

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 under-standing* 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