

# Compositional Question Interpretation Abilities of Reasoning Models

Semantic Computing Group  
David M. Schmidt  
daschmidt@techfak.uni-bielefeld.de

In light of rapidly increasing capabilities of large language models (LLMs), some even proclaiming the “age of LLMs”, the question arises where the limits of current LLMs lie. The reasoning abilities of LLMs and especially the abilities of LLMs to work and reason in a compositional way have been investigated by numerous related works in recent years, many of them indicating fundamental limitations in LLMs when it comes to truly compositional tasks [1]. Recent work [2] focusing on question answering over linked data (QALD), i.e., the generation of SPARQL queries for given natural language questions, reached a similar conclusion. At the same time, multiple LLMs specialized on reasoning, e.g., OpenAI GPT-5 or DeepSeek-R1, have been proposed recently, raising the question whether or not these models face the same limitations w.r.t. compositionality. Therefore, in this thesis, you will use the CompoST dataset [2] to test the limits of current reasoning models in the QALD domain. You will try out different ways to optimize the model’s performance, utilizing, e.g., DSPy [3] or fine-tuning. Finally, you will compare the results to LLMs not specialized in reasoning. Prior knowledge regarding LLMs, semantic web, QALD, or prompt optimization techniques is helpful but not required. Existing code from [2] is written in Python, but you are welcome to use any language that allows you to work with current reasoning models. The thesis can be written in English or German.

## Related literature

- [1] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. arXiv. <https://doi.org/10.48550/ARXIV.2305.18654>
- [2] Schmidt, D. M., Schubert, R., & Cimiano, P. (2026). Compost: A benchmark for analyzing the ability of llms to compositionally interpret questions in a qald setting. In D. Garijo, S. Kirrane, A. Salatino, C. Shimizu, M. Acosta, A. G. Nuzzolese, S. Ferrada, T. Soulard, K. Kozaki, H. Takeda, & A. L. Gentile (Eds.), *The Semantic Web – ISWC 2025* (Vol. 16140, pp. 3–22). Springer Nature Switzerland. <https://doi.org/10.48550/arXiv.2507.21257>
- [3] Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., & Potts, C. (2023). Dspy: Compiling declarative language model calls into self-improving pipelines. arXiv. <https://doi.org/10.48550/ARXIV.2310.03714>

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:

<https://www.uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/semantic-computing>

Interested? @mail to [daschmidt@techfak.uni-bielefeld.de](mailto:daschmidt@techfak.uni-bielefeld.de)