MRL 2025 Shared Task at EMNLP - Physical Commonsense Reasoning Data Set for the Estonian Language

Marii Ojastu
University of Tartu
Tartu, Estonia
ojastu@ut.ee

Hele-Andra Kuulmets
University of Tartu
Tartu, Estonia
kuulmets@ut.ee

Aleksei Dorkin
University of Tartu
Tartu, Estonia
dorkin@ut.ee

Krister Kruusmaa
Tallinn University
Tallinn, Estonia
kristerk@tlu.ee

Kairit Sirts
University of Tartu
Tartu, Estonia
sirts@ut.ee

Abstract

This paper presents the methodology for constructing 100 Estonian physical commonsense reasoning prompts for the MRL 2025 Shared Task at EMNLP. The prompts were constructed manually with the goal of evaluating the reasoning capabilities of large language models in tasks involving physical commonsense reasoning while also reflecting linguistic diversity and cultural relevance. The prompt construction process included an assessment step to ensure that the prompts were sufficiently challenging to reveal performance differences between models.

1 Introduction

This paper presents the methodology for constructing an Estonian-language evaluation data set for physical commonsense reasoning, consisting of 100 prompts with two answer options each, developed for the MRL 2025 Shared Task at EMNLP. The data set was constructed in accordance with the task guidelines. Each prompt addresses the physical properties of one or more objects, presents a question that an average Estonian speaker can reasonably answer, and, where possible, incorporates culturally specific elements. The prompts vary in length and style, with two candidate solutions that differ by one or two words. The incorrect option is designed to be plausible rather than implausible. The prompts were manually composed, and their difficulty was evaluated during the construction process. To ensure that the questions were unambiguous in their interpretation, the data set was independently annotated by an additional annotator to confirm agreement.

2 Prompt Construction

The data set was constructed manually according to the specified format. Where appropriate, culturally relevant elements, such as references to traditional Estonian foods, local materials, or region-specific practices were incorporated into the prompt. The initial inspiration for prompt construction was drawn from the 'Maybe I'm Lucky' feature of Sõnaveeb[1], the language portal maintained by the Institute of the Estonian Language. This feature provides users with randomly selected Estonian words, which, if deemed relevant for physical commonsense reasoning, were used as seed concepts around which prompts were formulated. The process aimed to ensure lexical diversity and encourage creativity.

3 Prompt Difficulty Assessment

To ensure that the prompts were sufficiently challenging to reveal performance differences between models, each candidate prompt was tested using Tehisaru Baromeeter (AI Barometer) [2], an LM Arena (Chiang et al. (2024)) style evaluation platform for Estonian-capable large language models. The platform allows users to submit prompts in Estonian and randomly selects models (listed in Appendix A) from its collection to generate responses; users are then given the opportunity to select the better answer, and the results are aggregated to produce an LLM leader board. The platform was used for prompt evaluation because it enables to assess the difficulty on a randomly selected subset of models.

To assess the difficulty of the prompt with Tehisaru Baromeeter, each prompt was shown to six randomly selected models. If all six models selected the correct answer, the prompt was considered too trivial to reveal differences in model capabilities and was either discarded or revised to increase its difficulty. If at least one of the six models failed

[1] https://sonaveeb.ee/
[2] https://baromeeter.tartunlp.ai/

to answer the question correctly, the prompt was retained and included in the data set.

4 Inter-Annotator Agreement

To reduce potential bias in the data set, an additional human annotator was involved alongside the original data set creator. The annotator was instructed to select the correct answer for each prompt without seeing the answers marked as correct by the data set creator. This process allowed assessing human agreement on the intended answers and helped confirm the clarity and interpretability of the prompts. The annotator agreement was initially calculated at 95%. The five samples without consensus were subsequently reviewed and revised until full agreement was reached. The final data set consists of 100 prompts, with complete agreement (100%) between the two annotators.

References

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

A Appendix

List of models on Tehisaru Baromeeter used for assessing the difficulty of the prompts:

- Claude-3-5-sonnet-20241022
- Claude-3-7-sonnet-20250219
- Claude-3-haiku-20240307
- Claude-3-opus-20240229
- Claude-opus-4-20250514
- Claude-sonnet-4-20250514
- Deepseek-chat
- Gemini-2.0-flash-001
- Gemini-2.0-flash-lite-001
- Gemini-2.5-pro-preview-06-05
- Gemma-3-27b-it
- Gpt-4-turbo-2024-04-09
- Gpt-4.1-2025-04-14
- Gpt-4o-2024-11-20
- Grok-3-beta
- Grok-3-mini-beta
- Kimi-k2-instruct
- Llama-3-70b-instruct
- Llama-3-8b-instruct
- Llama-3.3-70b-instruct
- Llama-4-maverick-17b-128e-instruct
- Llama-4-scout-17b-16e-instruct
- Llama-estllm-prototype-0825
- Meta-llama-3.1-405b-instruct
- Mistral-large-2411
- Mistral-small-2503
- Qwen2.5-72b-instruct
- Qwen3-235b-a22b
- Qwen3-32b