Multilingual Physical Reasoning Datasets

Authors: Jennifer Za Nzambi, Jan Čuhel, Tommaso Gargiani (all three of us are native speakers)
Language: Czech (*Moravian and Silesian dialects*)

Methodology

We have aggregated a PIQA-style dataset of physical commonsense reasoning examples in Czech. The dataset comprises four columns – the prompt, solution0, solution1, and a label denoting which solution corresponds to the ground truth. The items consist of two minimally different candidate solutions, only one of which is true.

Our examples span several domains, ranging from quotidian activities, including cooking, using tools, and completing household tasks, to more traditional activities relevant to Czech customs or sayings. It also contains references to Moravian and Silesian dialects, contemporary Gen Z / Gen Alpha slang, and other instances of physical reasoning related to miscellaneous topics such as sports or transport. All items are original and authored in Czech (the dataset does not include translations of English PIQA entries).

The dataset is authored by three native speakers. For each of the prompts used, two similar solutions were written, typically differing in one or two words or a few inflection endings (a common feature of Czech morphology). This way, both options remain syntactically very similar, whilst semantically different. Even though concrete and easily verifiable descriptions of real-world actions were at the heart of the dataset-making process, sometimes items seemed ambiguous even to native speakers. Such outliers were then reformulated or discarded altogether. For items using slang or dialects, we conducted research on the internet and consulted external collaborators from said demographic groups and communities to ensure their intended meaning. Notably, for entries pertaining to various legends, we drew on the following worksheets [1, 2, 3, 4, 5, 6]. For Moravian slang, we also consulted specialised sources [7].

After initial drafting by one author, each item was reviewed by the other two co-authors (also native speakers) and scrutinised for grammar, spelling, fluency, cultural appropriateness, and adherence to the dataset requirements. Disputed items were adjusted or replaced. Once approved, each sample was checked using LLMs (GPT-5 and Claude Opus 4.1) and LLM-generated suggestions, mostly pertaining to stylistic mistakes, were considered and, depending on native-speaker judgement, incorporated. Approximately 18% of the items were generated synthetically with GPT-5 and Claude Opus 4.1 using very specific instructions for both prompts and solutions. All synthetic samples were reviewed and corrected manually.

The dataset is released in .tsv format with four columns: prompt, solution0, solution1, label (0/1 indicating the zero-indexed position of the correct sample). The dataset adheres to UTF-8 encoding and includes diacritics and standard Czech punctuation.

The dataset may serve as a tool to assess language models' physical reasoning abilities in Czech. It may be used for zero-shot or few-shot evaluations and cross-lingual robustness testing. Items selected deliberately range from very niche and somewhat tricky examples to obvious and trivial distinctions that should be apparent to frontier models, as well as smaller or older models. Even though some samples may not be common sense to every Czech native speaker, they are relevant to broad audiences, specific generations, and relatively recent cultural phenomena. This attribute allows the dataset to strike a balance between capturing the current zeitgeist and presenting very specific, traditional, and often culturally relevant activities.

The dataset does not describe dangerous or illegal instructions and practices. Its purpose is to serve as a testing tool for language models only and not as a guideline, recipe, or training dataset.