NB-PIQA: A Norwegian Bokmål Dataset for Physical Commonsense Reasoning

Abeer Kashar
University of Waterloo
akashar@uwaterloo.ca

Hamza Zaidi
University of Waterloo
hamza.zaidi@uwaterloo.ca

1 Introduction

We introduce NB-PIQA, a manually constructed dataset of 120 original Norwegian Bokmål PIQA (Physical Interaction Quality Assessment) style examples. Each example consists of a tuple of `{prompt, solution0, solution1, English_prompt, English_solution0, English_solution1, label}`. The dataset focuses on physical and commonsense reasoning through general scenarios and ones culturally relevant to Norway. All examples were validated for correctness and quality by at least one native Norwegian speaker.

2 Dataset Construction

The development of NB-PIQA was centered on three concepts: Physical reasoning, commonsense, and cultural specificity. We included different aspects of Norwegian life, including local food (fiskesuppe, pinnebrød), activities (skiing, hiking, camping), tradition & folklore and indigenous culture.

2.1 Heuristics

We used two main methods to create prompts. Method 1 was to start with objects or procedures common to the Norwegian lifestyle and create a scenario where they are critical to the outcome. Method 2 draws inspiration from the original PIQA paper and involves reading instructables (both local and foreign) and creating prompts based on the procedures involved, focusing on unique uses of everyday items rather than conventional tools that would be easy for an LLM to get correct.
For solutions, the most common heuristic was to swap an object, where the incorrect solution substitutes an object with unsuitable physical properties (e.g., swapping a plastic spoon for a metal one based on heat resistivity). Another strategy was to have a suboptimal solution, where the incorrect choice is plausible but clearly less efficient or safe. These heuristics aimed to ensure that solving tasks would require an understanding of physical interactions rather than superficial pattern matching.

2.2 Verification of Novelty

Given that large language models are frequently pre-trained on extensive web data that may include popular benchmarks like PIQA, it was important that our dataset consists of novel examples to mitigate dataset contamination as much as possible. This was verified by generating embeddings for pairs of `(prompt, correct_solution)` and `(prompt, trick_solution)` as well as the full training set of the PIQA dataset. The cosine similarity of each example was calculated against the PIQA dataset to identify closest semantic matches. The authors then manually reviewed any pair with a correlation >0.7 to ensure semantic overlap was in the similarities of the subjects of the examples, rather than direct translations.

2.3 Length of Solutions

To ensure a certain level of difficulty in selecting the correct answer, the two candidate solutions are as close as possible to each other in length in each case, as demonstrated in the below figure.

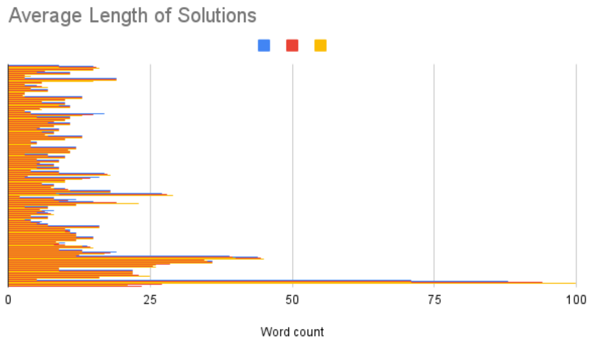

Figure 1: yellow represents the length of solution0, blue the length of solution1, and red the average between the two.