ShqiPIQA: Physical Commonsense Reasoning Dataset for Albanian

Anonymous ACL submission

Abstract

This paper presents ShqiPIQA, a physical commonsense reasoning dataset for Albanian created within the MRL 2025 Shared Task. Our dataset comprises 108 instances in PIQA format covering culturally-specific daily activities, traditional practices, and physical reasoning scenarios relevant to Albania and the Albanian speaking community.

1 Introduction

Physical commonsense reasoning (Bisk et al., 2019) tasks evaluate models' ability to understand basic physical properties, spatial relationships, and mechanical processes in everyday scenarios. While significant progress has been made in this domain, evaluation has predominantly focused on English data, leaving under-resourced languages like Albanian underexplored. This work addresses this gap by introducing the first Albanian physical commonsense reasoning dataset.

2 Dataset Description

We introduce ShqiPIQA, the first dataset for physical commonsense reasoning in Albanian, created as part of the MRL 2025 Shared Task. Analysis of physical commonsense reasoning in language models has primarily been conducted on English data (Pensa et al., 2024). With Albanian being an under-resourced language and underrepresented in English-centric NLP research, this dataset contributes to the limited landscape of manually curated evaluation resources for Albanian.

Dataset Statistics & Structure ShqiPIQA contains 108 instances in PIQA format, covering diverse domains such as cooking, cleaning, object construction (e.g., building a chicken house, crafting traditional musical instruments), Albanian traditional activities (music, dances, weddings), cultural practices, and agricultural tasks, with an emphasis on atypical solutions. The dataset is released in .TSV format, where each instance consists of: (i) a prompt describing a goal or situation, (ii) two candidate solutions (named as `solution_0` and `solution_1`), and (iii) a binary label indicating the correct solution (always `solution_0`). The two candidate solutions typically differ by only 1-2 words, thereby requiring fine-grained reasoning.

Linguistic Variation Prompts and candidate solutions were deliberately varied in their morphosyntactic formulation to test model robustness across different Albanian constructions. Prompts employ diverse structures including *për të* + participle or *që të* + finite verb (equivalent to English 'in order to' constructions) or simple noun phrases (e.g., *halva*, 'flour-based halva'). Candidate solutions utilise various forms, including imperative second person singular constructions.

Dataset Creation Process The dataset was collaboratively created by two language experts: (i) a linguist specialising in Albanology and an NLP researcher, holding a Master's degree, a non-native speaker of Albanian (native Russian) fluent in standard Albanian (Tosk (Gjinari et al., 2007) dialect) with prior dataset creation and annotation experience and expertise in linguistics and NLP; (ii) a digital humanities Master's student specialising in computational and historical linguistics, a bilingual native speaker of standard Albanian (Tosk dialect, Korça region) and German with prior experience in annotation and NLP.

Example The Listing 1 shows an instance from the ShqiPIQA dataset. The difference between `solution_0` (correct) and `solution_1` (incorrect) in Listing 1 lies in the phrase "majtas dhe djathtas" (left and right) versus "majtas" (left), where the correct solution requires holding hands with people on both sides to form a proper circle dance.

Listing 1: ShqiPIQA Dataset Instance

```
{
  "prompt": "Kërcesh valle" (You dance
      folk dance),
  "solution_0": "Do të lëvizet këmba një
      herë majtas e njëherë djathtas,
      duke mbajtur duart te personat
      majtas dhe djathtas, kështu që të
      kërcesh në rreth të madh me grupin
      ." (You will move your foot once
      left and once right, holding hands
       with the people on the left and
      right, so that you dance in a big
      circle with the group.),
  "solution_1": "Do të lëvizet këmba një
      herë majtas e njëherë djathtas,
      duke mbajtur duart te personat
      majtas, kështu që të kërcesh në
      rreth të madh me grupin." (You
      will move your foot once left and
      once right, holding hands with the
       people on the left, so that you
      dance in a big circle with the
      group.),
  "label": 0
}
```

3 Limitations

Several limitations should be acknowledged. First, with only 108 examples, ShqiPIQA provides limited coverage of possible everyday Albanian culturally-specific activities. While we include traditions and customs (dances, traditional musical instrument construction, cuisine, shepherding practices, food preservation), many cultural domains remain underrepresented. Second, the subjectivity inherent in determining relevant everyday activities, a challenge also present in the original PIQA dataset, is inevitable but should be considered during evaluation. Third, both dataset creators primarily reside outside the main Albanian-speaking continuum, potentially affecting the representativeness of selected activities. Finally, creating instances that remain challenging for state-of-the-art language models while being solvable through physical commonsense reasoning proves difficult.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI Conference on Artificial Intelligence*.

Jorgji Gjinari, Bahri Beci, Gjovalin Shkurtaj, Xheladin Gosturani, Anastas Dodi, and Menella Totoni. 2007. *Atlasi dialektologjik i gjuhës shqipe*. Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Tiranë. 2 vols. Dialectological Atlas of the Albanian Language.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. A multi-layered approach to physical commonsense understanding: Creation and evaluation of an Italian dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 819–831, Torino, Italia. ELRA and ICCL.