Belarusian PIQA-style Physical Commonsense Dataset

Daniil Dzenhaliou
EPFL
daniil.dzenhaliou@epfl.ch

Viktoryia Horbik
Independent
grsvictoria41@gmail.com

Abstract

We present a Belarusian-language dataset of physical commonsense reasoning tasks in the PIQA format: each entry consists of a household scene and two nearly identical options, of which exactly one is physically correct and the other is plausible but physically incorrect and non-absurd. The dataset was created by volunteer native speakers, enriched with culturally specific Belarusian scenarios (food, household, architecture, rituals) always grounded in physical properties (heat, friction, pressure, light, humidity, density, etc.). The final version contains 183 tasks; All examples were authored and checked by native speakers; no PIQA translations were used.

1 Introduction and Motivation

Physical commonsense tasks are important for evaluating models that need to understand everyday physics: heating/cooling, friction and traction, phase transitions, pressure, light, sound, diffusion, and more. For Belarusian, such resources are virtually absent, especially in natural everyday style and with cultural context. Our goal is to fill this gap and provide an open corpus suitable for training/evaluating models and cross-lingual comparisons.

2 Language, Dialect, Script, and Register

- Language: Belarusian, Cyrillic (standard codified norm).
- Register: neutral/conversational, typical for household descriptions (without heavy terminology, but occasional scientific words allowed).

3 Task Formulation

Each record includes:

- **prompt**: household scene (often multi-sentence), sometimes with short condition clarifications.
- **solution0/solution1**: two maximally similar options differing by 1–2 words or a micro-reordering, with parallel syntax.
- **label** $\in \{0, 1\}$: index of the correct solution (if solution0 is correct, label=0; otherwise 1).

The key principle: the incorrect option must sound plausible but be physically incorrect. Absurd formulations ("boiling and cooling at the same time") were discarded.

4 Dataset Construction Procedure

4.1 Idea Sources

- Brainstorming by native speakers: household situations, rural and urban environments.
- Local cultural realities: печь, сенцы, валёнкі, кірмаш, квас/бярозавік, лён, гліняны збан, царква/касцёл, etc.
- Educational/popular physics literature and guides on Belarusian life (for inspiration and correct terminology).
- LLM was used as assistant for paraphrasing/lengthening/normalizing style; all texts were manually post-edited by native speakers.

4.2 Pipeline

1. Drafting scenes and solution pairs (native speakers).
2. LLM paraphrasing, lengthening, alignment of options to differ by 1–2 words.
3. Two-stage manual review by native speakers: removing absurd/obvious answers, cultural inaccuracies.

4. Automatic Quality control:
   - length (prompt+solutions > 25 words);
   - detection of multi-sentence prompts;
   - pairwise solution alignment and 1–2 word difference;
   - exact/near-duplicate search (3-gram Jaccard);
   - check for service symbols in TSV;
   - mixed scripts detection;
   - heuristics for cultural specificity (markers).
5. Final proofreading and agreement.

4.3 LLM Role

LLM was applied under native speaker control: generating draft formulations, paraphrasing, answer normalization (1–2 words), style adjustment; metrics and technical checks were also automated with scripts. Any debatable/too obvious/awkward cases were rewritten manually.

5 Examples from the Corpus

This examples were translated from Belarusian

Prompt	At noon a car is parked without shade; the cabin heats up through the glass. They compare a light section of the dashboard and a dark, almost black panel. What happens with surface temperature?
Solution0	The black surface heats up more.
Solution1	The black surface heats up less.
Label	0

Table 1: Example A — heat/radiation.

Prompt	After a short rain, part of the garden path was covered with fine gravel, the adjacent section remained smooth and wet. A person walks at the same pace and tries to stop. Where is it easier to brake on foot?
Solution0	It is easier to brake on gravel.
Solution1	It is harder to brake on gravel.
Label	0

Table 2: Example B — friction/traction.

Prompt	On the kitchen table there are two mugs of birch sap: one near an open window, the other deeper in the room. In a short time, which one will cool faster?
Solution0	The one in the draft cools faster.
Solution1	The one in the draft cools slower.
Label	0

Table 3: Example C — convection/cooling.

6 Quality Control

Two native speakers sequentially proofread the corpus and rewrote any cases that:

- sounded absurd or too obvious;
- strayed outside physics (social/historical without physical properties);
- contained excessive hints in the prompt (spoilers);
- violated the 1–2 word difference rule between solutions.

7 Limitations and Ethics

- Incorrect options are manually validated as non-absurd, but occasional edge cases may need expert review.
- LLM served only as an assistant; final responsibility rests with native speakers. Entire corpus is original (no PIQA translations).
- Cultural cases present at ∼26%, balancing diversity with density of physical phenomena.

8 Data Format and License

- Final delivery: one file `dataset_final.tsv` (UTF-8) with columns: prompt, solution0, solution1, label.
- label=0 if solution0 is correct, else 1.

9 Reproducibility Recipe

1. Collect household and cultural scenes with clear physical properties.
2. For each, write two solution options of parallel length/structure, differing semantically by 1–2 words, non-absurd.

3. Double native-speaker proofreading; filter out obvious/non-physical items.

4. Run QC scripts (length, pairs, duplicates, scripts, service symbols, cultural markers).

5. Export TSV (UTF-8) and attach statistics + examples in the report.

10 Conclusion

We release the first (to our knowledge) PIQA-style corpus for the Belarusian language (Cyrillic), combining realistic household scenes with physically correct answers and carefully constructed "plausible but wrong" distractors. The corpus is suitable for training and evaluation, cross-lingual comparisons, and studies on cultural variability in physical commonsense reasoning.