Multilingual Physical Commonsense Reasoning Dataset - Bulgarian

Ivan Zhelyazkov
Independent researcher
iwanzhelyazkov@gmail.com

Abstract

Modern Large Language Models (LLMs) demonstrate an extraordinary ability to process and generate text based on statistical patterns. Their "Achilles' heel," however, remains the capacity for true, grounded comprehension of the physical world, tracking object states over time, and applying basic causal logic. This paper presents the "Multilingual Physical Commonsense Reasoning Dataset - Bulgarian" dataset, a specialized benchmark of multiple-choice prompts in Bulgarian. Each example is designed not to test encyclopedic knowledge, but to evaluate whether a model relies on genuine logical inference or on superficial pattern matching.

1 Introduction

While Large Language Models (LLMs) have achieved state-of-the-art results across many NLP benchmarks, their reasoning capabilities remain a critical area of research. Many benchmarks test associative knowledge, but fewer are designed to test robust inference against carefully constructed adversarial examples. Models often rely on "surface patterns" or spurious correlations in the training data rather than building a true causal or situational model of the text.

To address this gap, particularly for the Bulgarian language, we introduce the **Logical Traps** dataset. This benchmark is not a test of knowledge, but of reasoning. Each prompt is a small riddle designed such that the *incorrect* answer often contains keywords that are statistically highly associated with the prompt, while the correct answer depends on a single logical, causal, or physical inference.

Our main contributions are:

- A new, manually-curated reasoning benchmark for Bulgarian.
- A systematic classification of common-sense logical failures into six distinct categories.
- A methodology for generating "plausible distractors" designed to trap models that rely on shallow pattern matching.

2 The Dataset

Each entry in our dataset is a five-tuple designed for straightforward evaluation. The structure provides not only the challenge but also the explicit logical ground truth for each example.

- **Context (sentence_start):** A detailed description of a scenario, physical state, or sequence of events.
- **Solution 0:** The first possible outcome or explanation.
- **Solution 1:** The second possible outcome or explanation.
- **Label:** The ground-truth index (0 or 1) indicating the logically correct solution.

3 Dataset curation methodoloy

The dataset is structured into six primary categories, each designed to target a specific failure mode of associative physics reasoning. Within many examples, we also interweave cultural and morphological elements specific to the Bulgarian language.

3.1 Morphological Precision

- **Description:** This category tests whether the model understands how subtle morphological changes (specifically Bulgarian verb prefixes) fundamentally alter the physical meaning of an action, even when the verb root is identical.
- **The Trap:** The distractor (the wrong answer) uses a verb that is semantically incorrect for the context but statistically "close," as it shares the same root (e.g., *padam* / "to fall").

- **Example Concept:** Testing the difference between *na/padam* (to attack) and *po-padam* (to come across), both derived from the root verb *pad-am* (to fall).

3.2 Direct Physical Reasoning (with a Bulgarian twist)

- **Description:** This category tests the model's direct knowledge of foundational physics and chemistry (e.g., thermodynamics, density, viscosity, state changes). In some of the examples, we ground abstract principles in culturally-specific Bulgarian contexts (such as local foods, drinks, or traditional items).

- **The Trap:** The distractor is not a complex logical error (like a sequence swap) but rather a *plausible-sounding physical misconception*. It baits a model that recognizes the objects mentioned (e.g., "freezer," "rakia" - a Bulgarian alcoholic beverage) but fails to apply the correct physical properties (e.g., the freezing point depression caused by ethanol).

- **Example Concept:** Testing if homemade **rakia** (a high-ethanol spirit) will freeze solid in a standard home freezer (approx. -18°C). The correct answer (No, its freezing point is far lower) requires applying knowledge of ethanol's properties, whereas the incorrect answer (Yes, it freezes like water) represents a plausible but incorrect assumption.

3.3 Causality Sequencing

- **Description:** These prompts describe a basic physical chain of common-sense events (A causes B).

- **The Trap:** The incorrect solution explicitly reverses the causality (claiming B causes A). This traps models that recognize a strong association between A and B (like "friction" and "heat") but fail to understand which is the cause and which is the effect.

- **Example Concept:** "Friction causes heat" (Correct) vs. "Heat causes friction" (Incorrect).

3.4 Temporal State Tracking

- **Description:** This critical category describes an object whose state changes over time (e.g., it *was* solid ice, but is *now* a puddle of water).

- **The Trap:** The distractor solution is based on the *past (invalid) state* of the object. Because this historical state is explicitly mentioned in the prompt, it creates a powerful statistical lure. The correct answer requires the model to actively filter this historical data and base its answer only on the *current, physically active state*.

- **Example Concept:** Touching a freshly painted **blue** bench that *used to be* **red**. The hand becomes **blue** (Correct-Current State) vs. **red** (Incorrect-Past State).

3.5 Spatial Reasoning & Perspective

- **Description:** These prompts test comprehension of 3D space, relative positions (front/back, left/right), shadows, occlusion (blockage), and perspective (especially mirror reflections).

- **The Trap:** The distractor ignores physical perspective, often confusing an object's real left with its reflected right, or confusing what is reflected *in* a mirror with what is physically *behind* the mirror.

3.6 Category Errors (Abstract vs. Physical)

- **Description:** These prompts describe the application of a physical property or action (like weighing, cutting, or melting) to an abstract concept (like an idea, silence, or a promise).

- **The Trap:** The incorrect solution takes a common metaphor literally (e.g., describing the physical weight of "sadness" on a scale). The correct answer must identify the category error—that abstractions do not have mass or physical substance.

3.7 Negation Filtering

- **Description:** The prompt describes a scene and explicitly **negates** the presence of an object, property, or capability (e.g., "There were **no** red items," "He had **no** key," "The box was **empty**").

- **The Trap:** The incorrect solution proposes an outcome that depends entirely on the negated item. It is designed to bait a model that associates the mere *mention* of the word ("key") with the action ("unlock") while ignoring the logical operator ("no").

4 Conclusion

We have presented the Bulgarian segment of the "Multilingual Physical Commonsense Reasoning" dataset, a novel benchmark for evaluating situational, causal, and physical reasoning. By focusing on six distinct categories of logical failure and employing a "plausible distractor" methodology, this dataset effectively tests an LLM’s ability to reason, rather than just associate keywords. We hope this resource will be valuable for researchers working on model robustness and true natural language understanding, especially in a Bulgarian morphological and cultural context.