A Dutch Physical Commonsense Reasoning Dataset: Culturally-Grounded PIQA-Style Examples

David Stap
NXAI GmbH
david.stap@nx-ai.com

Abstract

We present a manually-constructed dataset of 120 Dutch physical commonsense reasoning examples following the PIQA format. Our dataset focuses on culturally-specific scenarios rooted in Dutch everyday life, including bicycle maintenance, weather management, small-space living, and traditional food preparation. Each example consists of a goal statement paired with two highly similar solution options that differ by only 1-2 critical words, requiring physical reasoning to distinguish the correct answer. The dataset aims to provide researchers with evaluation materials that reflect authentic cultural contexts while testing similar physical reasoning capabilities as the original PIQA benchmark.

1 Introduction

Physical commonsense reasoning remains a significant challenge for language models, despite considerable progress since the introduction of the PIQA dataset (Bisk et al., 2020). While the original work reported a substantial gap between human performance (95%) and state-of-the-art models (75%), recent advances in large language models have narrowed this gap considerably. However, existing benchmarks primarily reflect English-speaking cultural contexts, limiting their applicability for evaluating models in other linguistic and cultural settings.

This work contributes a Dutch physical reasoning dataset that addresses this gap by focusing on scenarios that are distinctly Dutch and often difficult to translate meaningfully to other cultural contexts. Our examples cover domains such as bicycle maintenance, management of Dutch weather conditions, efficient living in small urban spaces, and preparation of traditional Dutch foods.

2 Dataset Construction Methodology

2.1 Subject Selection and Cultural Grounding

We identified 15 core subject areas that are central to Dutch daily life and rich in physical reasoning opportunities:

- **Primary subjects** (8-12 examples each): Bicycles & Cycling, Dutch Weather Management, Small Space Living, Water & Drainage, Dutch Food Culture
- **Secondary subjects** (5-7 examples each): Energy Efficiency & Heating, Public Transport & Urban Life, Dutch Winters & Ice, Coastal & Maritime, Markets & Shopping
- **Specialized subjects** (3-5 examples each): Dutch Agriculture & Flowers, Café Culture & Social Spaces, Dutch Holidays & Traditions, Urban Density Challenges, Dutch Engineering & Innovation

Subject selection prioritized scenarios that: (1) require culture-specific knowledge unavailable in other contexts, (2) involve materials and tools commonly available in the Netherlands, (3) reflect problem-solving approaches typical in Dutch society, and (4) offer substantial opportunities for physical reasoning challenges.

2.2 Example Creation Process

Each example was manually constructed following a systematic four-step process:

- **Step 1: Goal Formulation.** Goals were designed as natural task statements or questions reflecting authentic Dutch problem-solving scenarios. We maintained a target length of 8-20 words to match the original PIQA dataset characteristics, with most examples clustering around 12-15 words.

- **Step 2: Correct Solution Development.** The correct solution was crafted to represent effective, safe, and culturally appropriate problem-solving approaches. Solutions typically ranged from 15-25 words, providing sufficient detail while maintaining conciseness.

- **Step 3: Distractor Creation.** The incorrect solution was generated by modifying the correct solution through subtle but functionally critical changes: material substitution (e.g., oil vs. toothpaste for chain lubrication), spatial/directional errors (e.g., "over" vs. "under" the saddle), timing modifications (e.g., morning vs. evening market visits), or safety trade-offs (e.g., proper vs. improvised safety measures).

- **Step 4: Physical Reasoning Validation.** Each example pair was evaluated to ensure that: (a) both solutions maintain surface plausibility, (b) the distinction requires genuine physical understanding rather than linguistic cues, (c) the reasoning involves authentic physical properties, spatial relationships, or causal mechanisms, and (d) the correct answer reflects practical Dutch knowledge.

2.3 Quality Assurance

All examples were reviewed by two native Dutch speakers to ensure linguistic naturalness, cultural authenticity, and appropriate difficulty level. Examples were rejected if they could be solved through superficial pattern matching rather than physical reasoning, or if they did not reflect genuine Dutch cultural contexts.

The final dataset includes scenarios such as preventing bicycle chain rust with sunscreen (oil-based) vs. toothpaste (abrasive), warming stroopwafels by placing them on top of versus under a warm mug, and managing Dutch weather by carrying boodschappen (groceries) under versus above one's jacket during sudden downpours.

3 Dataset Characteristics

3.1 Structural Properties

Our dataset consists of 120 examples distributed across 15 cultural domains. Goal statements average 14.2 words (range: 8-20), while solution pairs average 18.6 words (range: 15-25), closely matching the original PIQA length distributions. Each solution pair differs by exactly 1-2 words, with 89% of examples differing by a single critical term.

3.2 Physical Reasoning Types

The dataset tests four primary categories of physical reasoning: (1) **Material Properties** - understanding how substances interact (e.g., oil vs. abrasive substances for metal maintenance), (2) **Spatial and Geometric Reasoning** - recognizing the importance of position and orientation (e.g., heat source direction for warming food), (3) **Causal Understanding** - predicting multi-step physical outcomes (e.g., drainage patterns and water flow), and (4) **Functional Affordances** - recognizing appropriate tool usage and safety considerations (e.g., structural support points for hanging items).

3.3 Cultural Specificity

A key distinguishing feature of our dataset is its cultural specificity. Examples like bicycle maintenance techniques, stroopwafel warming methods, managing Dutch rainfall, and navigating Amsterdam's narrow spaces require knowledge that is rarely documented in text corpora, making them particularly challenging for language models trained primarily on English sources.

4 Conclusion

This dataset provides Dutch-speaking researchers with culturally-grounded evaluation materials for physical commonsense reasoning. By focusing on authentically Dutch scenarios while maintaining the rigorous physical reasoning requirements of the original PIQA format, we contribute to more diverse and representative evaluation of language models' real-world reasoning capabilities.

References

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.