AlgerMoroc-PIQA: A Culturally-Grounded Dataset for Physical Commonsense Reasoning in Algerian and Moroccan Arabic

Abstract

We present AlgerMoroc-PIQA, a PIQA-style benchmark for physical commonsense reasoning in Algerian and Moroccan Darija. The dataset contains 410 culturally grounded items collected via a hybrid pipeline (AI-assisted generation and community crowdsourcing) followed by native-speaker verification. Each item includes a Darija prompt, two minimally differing candidate solutions, and a gold label; we also provide dialect tags and MSA translations. Inter-annotator agreement is Fleiss' kappa 0.939 and Cohen's kappa 0.911/0.981/0.928. We detail construction protocols, cultural coverage, ethics, and limitations, and provide a dataset card to facilitate reproducibility.

1 Introduction

Physical common sense, which is the ability to reason about everyday interactions in the world, is critical for robust language understanding. Most benchmarks are English-centric, and dialectal Arabic, particularly Maghrebi Darija, is under-resourced and culturally distinct from Modern Standard Arabic (MSA). We introduce AlgerMoroc-PIQA, a PIQA-style dataset (Bisk et al., 2020) for Algerian and Moroccan Darija that centers culturally specific scenarios (e.g., tajine heat control, hammam cleaning tools, suhoor timing, market practices, local transport). Our contributions are: (i) a culturally grounded Darija dataset for physical commonsense; (ii) a hybrid creation pipeline (AI-assisted + community crowdsourcing) with native-speaker verification; (iii) complete reproducibility artifacts.

Dataset

Each instance follows the PIQA template: a Darija prompt with two minimally different candidate solutions (1–2 word/phrase change), and a single correct label. We release a TSV with required columns: prompt, solution0, solution1, label. For transparency and reuse, we also provide dialect, source, and MSA translation for the prompt.

1. Strategy 1: AI-Assisted Generation

We used the Large Language Model, Gemini 1.5 Pro to generate culturally grounded PIQA-style items across the following categories: cooking, household cleaning, tools/materials, clothing/laundry, weather/seasonal, home maintenance, daily transport, traditions, religion, childcare/parenting, and beauty. The instructions emphasized cultural specificity and minimal-difference distractors. The prompt provided to the LLM is:

"You are tasked with generating a PIQA-style dataset for physical commonsense reasoning in Algerian Arabic (Darija) and Moroccan Arabic (Darija)."
Dataset requirements:
Each example must include the following:
Prompt: a question or instruction in Darija.
Solution0: one possible answer in Darija.
Solution1: another possible answer in Darija, differing by one or two words or phrases.
Label: 0 or 1, marking which solution is correct.
Dialect: "Algerian" or "Moroccan".
Category: choose from the following categories: cooking; household cleaning; tools/materials; clothing/laundry; weather/seasonal; home maintenance; daily transport; traditions; religion; childcare/parenting; beauty.
MSA: Modern Standard Arabic version of the prompt.

Rules for dataset creation:
- Write at least 100 original examples per dialect (not translations).
- Ensure cultural specificity: prompts should draw on Algerian and Moroccan practices (e.g. couscous steaming, tajine cooking, hammam traditions, Ramadan fasting, souks, local transport like taxis and buses, olive harvesting, henna rituals).
- Prompts and solutions should vary in length; many items should exceed 25 words (prompt + solutions); some should be multi-sentence.
- Candidate solutions must be nearly identical (1–2 words/phrase swap); the incorrect solution must be plausible but wrong.
- Use natural colloquial Darija; clearly indicate the dialect; provide MSA and English translations of the prompt; cover the categories with balanced distribution.

2. Strategy 2: Community Crowdsourcing

We distributed a collection form to native Algerian and Moroccan speakers. In total, 20 volunteers contributed items, and 6 served as annotators/validators. Contributors could optionally include free-text rationales for why the plausible solution is correct. We screened submissions for physical plausibility, and cultural appropriateness before verification.

3. Annotation and Quality Control

All examples, whether AI-generated or crowdsourced, underwent a stringent human verification and annotation process. Each example was reviewed by three independent

human annotators, all native speakers of either Algerian or Moroccan Darija. These annotators were tasked with:

- Verifying linguistic naturalness: Ensuring the Darija prompts and solutions were idiomatic and grammatically correct within the respective dialects.
- Assessing cultural relevance: Confirming that the scenarios accurately reflected Algerian or Moroccan daily life and practices.
- Evaluating plausibility: Determining if the incorrect solution was genuinely plausible but physically incorrect, rather than absurd.
- Assigning the correct label: Marking the appropriate solution (0 or 1).

The final label for each example was determined by a majority voting scheme. If at least two out of the three annotators agreed on a solution, that solution was adopted as the gold standard label.

Inter-annotator agreement was assessed using Fleiss' kappa 0.939 across three annotators, and pairwise Cohen's kappas 0.911, 0.981, and 0.928, indicating almost perfect agreement among annotators. This robust agreement metric underscores the carefully curated nature of the AlgerMoroc-PIQA dataset.

Dataset Statistics

Table 1: Dataset details.

Total items	410
Crowdsourced Algerian Darija	100
Crowdsourced Moroccan Darija	90
AI-generated Algerian Darija	110
AI-generated Moroccan Darija	110

Table 2: Inter-annotator agreement.

Fleiss' κ	Cohen κ (A1,A2)	Cohen κ (A1,A3)	Cohen κ (A2,A3)
0.939	0.911	0.981	0.928

Length profiles (approximate): mean prompt length ≈ 10 words; mean solution length 7 words.

6 Representative Examples

We show some illustrative items (Darija prompt, two solutions, gold label).

Dialect	Prompt (Darija)	Solution0	Solution1	Label	MSA translation of Prompt	English translation of Prompt

Algerian	كيفاش نطيبو الكسكسي باش يجي بنين و طايب مليح؟	لازم نغلوه في الماء بزاف باش يجي خفيف	لازم نفوروه زوج خطرات و لا ثلاثة باش يجي خفيف	1	كيف نطهو الكسكسي ليصبح لذيذاً ومطهواً جيداً؟	How do we cook couscous so it becomes tasty and well-prepared?
Moroccan	كيفاش نطيبو الطاجين باش يجي بنين و ما يلصقش؟	نجهدو عليه العافية باش يطيب دغيا	نخليوه يطيب على نار مهيلة و نستعملو طاجين ديال الطين	1	كيف نطهو الطاجين ليصبح لذيذاً ولا يلتصق؟	How do we cook tagine so that it becomes tasty and doesn't stick?

(MSA prompt translations are included in the TSV under `MSA translation of the goal`.)

7 Ethical Considerations

Contributors and annotators participated **voluntarily without monetary compensation**. Recruitment occurred via open community channels; participants gave informed consent, could withdraw at any time, and were not subject to coercion or undue influence. No personally identifiable information was collected, and tasks were low risk and limited to culturally familiar scenarios.

8 Limitations & Reproducibility

The coverage of this initial dataset is limited to two Maghrebi dialects, and colloquial spelling variance persists. We provide a TSV schema, the exact prompt given to the LLM, contributor instructions, and the annotation protocol to facilitate replication in other dialects/languages.

Appendix A. Google Form Instructions for crowdsourcing

Goal: write culturally grounded PIQA-style items in Algerian or Moroccan Darija.
Columns: prompt(Darija), solution0, solution1, label, dialect, category, MSA prompt, (optional) brief rationale.
Constraints: the two solutions must be nearly identical (1–2 words/phrases); the incorrect option must be plausible but physically wrong; prefer realistic daily-life scenarios (tajine heat, hammam tools, Ramadan timings, transport, childcare); vary lengths; many items should exceed 25 words; use natural colloquial spelling; avoid offensive unsafe content.

Appendix B. Dataset Card (Summary)

Dataset Card: AlgerMoroc-PIQA

- **Motivation & Task:** PIQA-style evaluation of **physical commonsense** in Algerian & Moroccan Darija.
- **Composition: 410** items (Algerian **210**, Moroccan **200**). Each item = prompt (MSA)+ prompt (Darija) + two minimally different solutions + gold label.
- **Sources: AI-assisted** generation (**220**) + **community crowdsourcing** (**190**).
- **Format:** TSV columns: prompt, solution0, solution1, label (+ dialect, source, msa_translation).
- **Creation Process:** Gemini 1.5 Pro used with culturally specific instructions across categories (cooking, hammam, transport, etc.); native speakers verified and refined.
- **Annotation & Quality:** Up to 3 native speaker votes per item; majority gold; agreement **Fleiss' κ = 0.939**; **Cohen's κ = 0.911 / 0.981 / 0.928**.
- **Preprocessing Notes:** Spelling kept in natural colloquial Darija; MSA prompt provided for many items; solutions differ by 1- 2 words to control for lexical shortcuts.
- **Intended Use:** Benchmarking physical commonsense in Maghrebi dialects; shared-task participation.
- **Ethical Considerations:** Voluntary participation; no PII collected; culturally sensitive content review.
- **Known Limitations:** Two dialects only; colloquial spelling variation; some category imbalance.
- **License:** recommend CC BY 4.0.

References

Bisk, Y., Zellers, R., Le Bras, R., Gao, J., & Choi, Y. (2020). PIQA: Reasoning about Physical Commonsense in Natural Language. AAAI.