TURKIC–PIQA: A Turkish, Azerbaijani, and Kyrgyz Dataset for Physical Commonsense Reasoning in the MRL 2025 Shared Task

Ali Eren Çetintaş[1,*] **Burak Tutar**[1,*] **Burcu Alakuş Çınar**[1,*] **Gulnaz Abdykadyrova**[4,*]
İbrahim Ethem Deveci[1,*] **Jafar Isbarov**[2,*] **Kavsar Huseynova**[3,*]
Mehmet İlteriş Bozkurt[1,*] **Tilek Chubakov**[4,*] **Duygu Ataman**[1]

[1]Middle East Technical University [2]Virginia Tech [3]Baku Higher Oil School [4]Independent Researcher

{ali.cetintas, tutar.burak, burcu.alakus, dataman, ethem.deveci, ilteris.bozkurt}@metu.edu.tr
g.abdykadyrova@gmail.com, isbarov@vt.edu
kavsar.huseynova.std@bhos.edu.az, tchubakov@berkeley.edu

Abstract

TURKIC–PIQA is a three-language Turkic dataset of PIQA-style two-choice minimal pairs for physical commonsense in the MRL 2025 Shared Task (SIGTYP / MRL Organizers, 2025; Bisk et al., 2020). Each item contrasts two plausible alternatives with one decisive physical factor, using the TSV schema: *language*, *prompt*, *solution0*, *solution1*, *label* (solution0 fixed correct, label=0). Items are culturally natural, physically grounded, and concise. The release contains **370** items in total: **136** Turkish, **120** Azerbaijani, and **114** Kyrgyz. In Turkish only, a subset of items was ideated via GPT–5 on Perplexity AI; many Turkish items are fully original, and all released items were double-checked by native speakers after curation (Perplexity AI, 2025). We report per-language statistics from the final TSV (*turkic_piqa_for_mrl2025-turkish_azerbaijani_kyrgyz.tsv*) and include a verbatim Turkish authoring prompt for transparency (EMNLP 2023 Organization, 2023).

1 Introduction

This formal dataset report presents *TURKIC–PIQA*, a single-schema, three-language collection of minimal pairs for physical commonsense reasoning aligned to PIQA's contrastive design and released for the 2025 Multilingual Representation Learning (MRL) Workshop Shared Task (SIGTYP / MRL Organizers, 2025; Bisk et al., 2020). Each item keeps solutions as similar as possible while a small, decisive factor (for example, venting, heat resistance, pre-warming, or placement) flips correctness. The TSV places *language* first, then *prompt*, *solution0*, *solution1*, *label*, with *solution0* fixed as correct and *label=0* for all rows. Native speakers authored and reviewed the content; sections follow uniform headings across Turkish, Azerbaijani, and Kyrgyz for comparability (EMNLP 2023 Organization, 2023).

2 Task framing and related work

PIQA evaluates physical commonsense with tightly matched alternatives where a small, localized difference flips correctness, emphasizing physical properties, affordances, and procedures (Bisk et al., 2020). *TURKIC–PIQA* is not a translation of PIQA; it is an original, native-authored dataset aligned with the PIQA format and tailored to Turkic languages. Within MRL 2025, the unified schema and fixed-label policy simplify multilingual baselines and cross-language analysis (SIGTYP / MRL Organizers, 2025).

3 Format and scope

TSV schema

The TSV columns are, in order: *language*, *prompt*, *solution0*, *solution1*, *label*; the public filename is *turkic_piqa_for_mrl2025-turkish_azerbaijani_kyrgyz.tsv* (SIGTYP / MRL Organizers, 2025).

Minimal-pair design

- Exactly one decisive physical factor flips correctness while solutions remain otherwise highly similar (Bisk et al., 2020).
- Pivot types: *shape/placement*, *material/tool*, *purpose/order/state*.
- Edits: a 1–2 token micro-edit or a short phrase-order swap; no additional tool changes or stylistic rewrites beyond the pivot.

Data creation pipeline

- Turkish-only Large Language Model (LLM) seeding: for ideation, 200 scaffolds were gener-

*Equal contribution. Correspondence: General & Turkish - Mehmet İlteriş Bozkurt (ilteris.bozkurt@metu.edu.tr); Azerbaijani - Jafar Isbarov (isbarov@vt.edu); Kyrgyz - Tilek Chubakov (tchubakov@berkeley.edu).

ated with GPT–5 via Perplexity AI; many Turkish items are fully original, and all released items are double-checked by native speakers after curation (Perplexity AI, 2025).

- Native authoring: speakers rewrote, expanded, and deleted items under strict minimal-pair and plausibility criteria, then consolidated the final TSV with *solution0* fixed as correct.

4 Languages

4.1 Turkish

Overview Turkish items follow the TSV schema with *language* first, *solution0* fixed as correct, and *label=0*. Coverage spans kitchens and households, vehicles and climate, tools and materials, storage and preservation, and culturally natural contexts (for example, *semaver*, *bakır cezve*) while correctness remains physically grounded. A substantial portion of items are completely original; only some were ideated via LLM scaffolds and all released items are double-checked by native speakers after curation (Perplexity AI, 2025).

Statistics Space-delimited token counts computed over the concatenation *prompt+solution0* yield **136** items; median **24**; minimum **13**; maximum **88**. Thresholds: **73.53%** have at least 20 space-delimited tokens and **46.32%** have at least 25; since Turkish is agglutinative, space-delimited counts conservatively approximate information density. See Table 1.

Stat	Value
Number of samples	136
Median word count	24
Max word count	88
Min word count	13

Table 1: Turkish statistics over space-delimited tokens of prompt+solution0.

Methodology and scope

- Authoring and review: original, native-authored items; six native reviewers ensured a decisive physical factor, plausibility, and minimal textual divergence; items were curated to be understandable by laypersons without requiring expert knowledge.
- LLM disclosure: GPT–5 via Perplexity AI seeded only some Turkish items; many items are fully original; all released items are edited by native speakers after an initial check, expansion, and deletion (Perplexity AI, 2025).
- Multi-factor note: in a small fraction of Turkish items, tightly coupled steps co-vary (for example, pre-warm and slow pour together) to reflect realistic practice; such cases were allowed with reviewer sign-off when the combined factors form a single, inseparable physical decision.

Coverage and design

- Domains: household routines, workshop practices, vehicles and seasonal conditions, preservation and storage, and culturally natural items.
- Pivots: venting versus sealing, warm-up versus thermal shock, surface contact versus curvature, heat-resistant versus non-resistant tools, order and pacing.

4.2 Azerbaijani

Overview The Azerbaijani subset is harmonized to the same headings and table style as Turkish while preserving original content and claims; it conforms to the TSV schema with *language* first and *solution0* fixed correct with *label=0* (SIGTYP / MRL Organizers, 2025).

Methodology and scope The dataset was compiled according to shared-task instructions; PIQA was used as a reference while maintaining originality, and all questions were created manually without use of LLMs or translation (Bisk et al., 2020).

Statistics Space-delimited token counts over *prompt+solution0* give **120** items; median **18**; minimum **6**; maximum **36**. See Table 2.

Stat	Value
Number of samples	120
Median word count	18
Max word count	36
Min word count	6

Table 2: Azerbaijani statistics over space-delimited tokens of prompt+solution0.

Coverage and design

- Topics: cooking, driving, woodworking, human body, basic physics and chemistry.
- Culture and script: culturally specific references where natural (for example, *Aran*, *ayran*, *kabab*); Latin script is used; multi-script settings may affect LLM performance (Isbarov et al., 2025).

LLM experiments Trials with GPT–5 and Gemini 2.5 Pro produced poor quality samples; no such items were included in the final dataset (OpenAI, 2025; Comanici et al., 2025).

4.3 Kyrgyz

Overview The Kyrgyz subset adopts the same headings and table style and adheres to the TSV schema with *language*, *prompt*, *solution0*, *solution1*, *label* and *solution0* fixed as correct with *label=0*. Samples were authored by native speakers; no PIQA translations were used.

Statistics Space-delimited token counts over *prompt+solution0* give **114** items; median **21**; minimum **13**; maximum **32**. See Table 3.

Stat	Value
Number of samples	114
Median word count	21
Max word count	32
Min word count	13

Table 3: Kyrgyz statistics over space-delimited tokens of prompt+solution0.

Methodology and scope

- Authoring and checks: native-speaker creation with manual checks for clarity, correctness, and task relevance; the incorrect solution often differs by one word or inflection; Cyrillic script is used.
- Pivots: a single token or inflection or a tightly scoped phrase-order change; otherwise identical steps and tools.

5 Quality control

- Schema verification: one TSV across languages with columns *language*, *prompt*, *solution0*, *solution1*, *label*; *solution0* correct and *label=0* for every row (SIGTYP / MRL Organizers, 2025).
- Minimal-pair similarity: enforce a 1–2 token pivot or a short phrase-order swap; in Turkish, allow rare, tightly coupled multi-factor contrasts when they form one natural physical decision and have reviewer sign-off.
- Length ranges: report space-delimited token thresholds consistently in all languages over *prompt+solution0*.
- LLM disclosure: LLM seeding was used only for some Turkish items; many Turkish items are fully original; all released items are double-checked by native speakers after curation (Perplexity AI, 2025).

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Also available as arXiv:1911.11641.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

EMNLP 2023 Organization. 2023. Emnlp 2023 LaTeX instructions and style guidelines. https://acl-org.github.io/ACLPUB/formatting.html. EMNLP 2023 style and formatting guidance.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, et al. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-09-05.

Perplexity AI. 2025. Perplexity ai. https://www.perplexity.ai/. LLM platform used for seed/pilot generation.

SIGTYP / MRL Organizers. 2025. Mrl 2025 shared task on multilingual physical reasoning datasets. https://sigtyp.github.io/st2025-mrl.html. Task page and guidelines.

A Dataset-generation prompt (verbatim)

This Turkish authoring prompt was used only to seed some Turkish items; its outputs were never used as-is (i.e. without a native speaker check) and were partially incorporated after manual edits by native speakers.

```
TITLE
MRL 2025 — Turkish Physical Reasoning (TR) — PIQA-style minimal pairs — Copy-safe TSV — EXACT N=50 — Single fence — No status line — Mixed (question + statement + cloze-choices) — Variable lengths — PIQA blocklist anti-overlap

DISCLAIMER
- Templates below are tentative; prioritize clarity, causality, safety, and naturalness. If a surface harms coherence, rewrite while preserving a single decisive physical pivot and the required formatting.
- Author an original, culturally natural Turkish dataset (not translations). Favor Turkish contexts/tools where the decisive factor is physical, not cultural trivia. Align with the MRL 2025 shared task's PIQA-like format. [MRL page]

PIQA INSPIRATION + ANTI-OVERLAP (BLOCKLIST ONLY)
- Adopt the PIQA shape (one prompt with two choices, one correct) while creating uniquely Turkish items. Do NOT seed or paraphrase specific PIQA examples.
- Use only the attached files (valid.jsonl, valid-labels.lst) as anti-overlap blocklists; never copy content from them.
- Guards (must pass per candidate):
  1) Label policy: This dataset sets label=0 (solution0 correct) for all rows, independent of any labels in the attached list; reject any candidate that would set label0.
  2) Lexical Jaccard: Turkish prompt lemma 2-gram Jaccard  0.30 vs. back-translated PIQA goal/sol1/sol2; else rewrite/regenerate.
  3) Embedding similarity: Cosine  0.85 vs. PIQA prompts/goals (after MT alignment); else rewrite/regenerate.
  4) Scenario schema guard: Reject if the core object–action–goal triple matches any PIQA scenario with only superficial token changes; re-author different scenario/pivot.
  5) Structure edit distance: Normalized edit-distance  0.50 vs. nearest PIQA prompt pattern; else rewrite/regenerate.

OUTPUT CONTRACT
- Emit exactly one code-fenced TSV-like block; nothing before or after the fence.
- Fence: open with ``````.
- Columns (delimiter = |||), exactly 7 per row in this exact order:
  1 prompt
  2 solution0
  3 solution1
  4 label
  5 english_prompt
  6 english_solution0
  7 english_solution1
- Header: print only if BATCH_INDEX == 1:
  prompt|||solution0|||solution1|||label|||english_prompt|||english_solution0|||english_solution1
- Exact row count and delimiter budget:
  rows_exactly = 50. Let H = 1 if header printed else 0. Let D = 6*rows_exactly + 6*H.
  Maintain delim_count (start 6 if H==1 else 0) and row_count=0.
  Before a row: if delim_count + 6 > D, close fence and STOP (do not print the row).
  After each row: delim_count += 6; row_count += 1.
  When row_count == 50: close fence and STOP (no status line).
  Internal assert: (row_count == 50) AND (delim_count == D).
- Sanitation: replace raw "|" inside fields with ""; collapse tabs/newlines to a single space; forbid [ ] in fields; no URLs or code fences in fields.
- Label policy: solution0 is correct; set label=0 on every row.

SCOPE + SAFETY
- Exactly one decisive physical factor per item: SHAPE/PLACEMENT, MATERIAL/TOOL, or PURPOSE/ORDER/STATE.
- Distractor is plausible but wrong solely due to that factor.
- Turkish for prompt/solutions; concise English glosses (faithful paraphrases).
- If risk exists, solution0 uses the safer step; never instruct hazardous behavior.

PROMPT-TYPE ENFORCEMENT (HARD)
Type classifier (deterministic, pre-selection):
```

- question: prompt ends with a literal “?” (no other “?” in the prompt).
- statement: prompt has no “?” anywhere and is a complete declarative sentence (terminal punctuation like “.” or “!”).
- cloze-choices: the prompt field has exactly 3 lines:
  • line 1: stem (no “?”), 6–18 words, may end with colon or nothing.
  • line 2: correct imperative (solution0 must equal this line), 6–14 words.
  • line 3: incorrect imperative (solution1 must equal this line), 6–14 words.
  • english_prompt paraphrases stem; english_solution0/1 paraphrase lines 2/3.

Per-type length ranges (strict, per item):
- question: prompt 12–28 words; each solution 18–40 words.
- statement: prompt 14–32 words; each solution 20–42 words.
- cloze-choices: as defined above (stem 6–18; each imperative 6–14).

Mixed-type quotas (simultaneous, per 50):
- question: 15–22
- statement: 14–20
- cloze-choices: 8–12
Batch rejection guard: If any bound is violated after candidate selection, discard selection and reselect (reject-all-question-batch). Never degrade targets due to token limits.

MINIMAL-PAIR RENDERING (MANDATORY PER ITEM)
- Author hidden solution0BASE (correct, 15–35 words) with exactly one decisive factor; then deterministically render both solutions:
  • Method A (token micro-edit): inject exactly one antonym/token pivot; solution0 uses A, solution1 uses B; token-level Levenshtein(solution0, solution1) 2; strip any tag/residue; forbid adjacency residue (e.g., “üst alt”, “kuru ıslak”, “düşük ısı yüksek ısı”, “silikon metal”).
  • Method B (phrase-order flip): swap two contiguous short phrases (5 tokens each); outside swapped spans, strings must be byte-identical including punctuation/spacing; no other edits.
- Forbidden in solution1: adding/removing tools/entities; extra safety/material/order edits; style/punctuation rewrites; qualifiers; multiple pivots.

LEXICON EXAMPLES (EXTENDABLE; DO NOT PRINT TAGS)
- SHAPE/PLACEMENT: üst|alt, iç|dış, önde|arkada, paralel|çapraz, sıkı|gevşek, düz|eğri, düz yüzey|kavisli yüzey, üst kenar|alt kenar, geniş temas|noktasal temas
- MATERIAL/TOOL: gözenekli|sızdırmaz, yumuşak|aşındırıcı, iletken|yalıtkan, ısıya dayanıklı|dayanıksız, kuru|ıslak, gıda uygun|gıda uygun değil, balmumu|yağ, sirke|soda
- PURPOSE/ORDER/STATE: önce|sonra, kapak aralık|kapak kapalı, yavaş|hızlı, düşük ısı|yüksek ısı, ılık|soğuk, kısa beklet|uzun beklet, kademeli|ani

NOVELTY + DEDUP (CONTENT-LEVEL)
- (head_noun, main_verb) uniqueness across the 50.
- Antonym-pair rotation (Method A): any single pair used 2 times, never consecutively; 15 distinct pivot pairs across the 50.
- Domain rotation: 8 distinct domains (kitchen, bathroom, workshop, HVAC, plumbing, textiles, garden, vehicle, outdoors, electronics, fermentation, TR-cultural like semaver, mangal, tandır, bakır cezve, ayran, yufka, turşu). No domain > 8 items; avoid > 2 consecutive from the same domain.
- Intra-batch semantic similarity: max cosine 0.88 between any two prompts (Turkish embeddings); otherwise rewrite/select a different pivot.

SELECTION PIPELINE (INTERNAL; DO NOT PRINT IN DATA)
1) Over-generate diverse candidates across types and domains; tag type and word counts.
2) Coherence/Safety pre-check: enforce one pivot; plausible distractor; safer practice in solution0; reject hazardous instructions.
3) Minimal-pair rendering: from solution0BASE, produce solution0/solution1 via Method A or B; enforce Levenshtein2 or pure phrase-swap; no tag/adjacency residue; both solutions present.
4) Anti-overlap (blocklist files): enforce label=0; Jaccard0.30; cosine0.85; scenario schema distinct; edit-distance0.50. Regenerate or rewrite to pass all guards.
5) Mixed-type quotas: select to satisfy question/statement/cloze quotas simultaneously; if not feasible, regenerate missing types then reselect (reject-all-question-batch).
6) Length gates: enforce per-type word ranges for prompt and solutions; replace items that are too short/long.
7) Novelty & domain rotation: enforce (head_noun, main_verb) uniqueness; pivot-pair caps; 8 domains; no long streaks; intra-batch cosine0.88.
8) Gloss & sanitize: write concise English glosses; replace “|” with “”; normalize whitespace; forbid [ ]; verify 7 columns and exactly 6 delimiters per row.
9) Printing: open ```

VALIDATION (SILENT)
- Schema: exactly 7 columns; exactly 6 delimiters per row; label=0 always.

- Counts: row_count==50 and delim_count==D.
- Quotas: question 15–22; statement 14–20; cloze 8–12.
- Lengths: each item in its type-specific word ranges.
- Anti-overlap: all PIQA guards passed (blocklist files).
- Novelty: (head_noun, main_verb) unique; 8 domains; domain caps respected; max cosine0.88.
- Minimal-pair: Method A Levenshtein2 or Method B identity outside swap; no tag or adjacency residue; no multi-factor edits.
- Coherence/Safety: single physical pivot; distractor plausible; safer practice where relevant.

NOTES
- Uniquely Turkish: prefer semaver, mangal, bakır cezve, turşu, yufka, kombi/petek, soba, vb., as long as the decisive factor remains physical.
- PIQA shape only: two-choice reasoning with one correct choice; zero lexical/structural overlap with blocklisted items.

BEGIN DATA OUTPUT NOW