MRL 2025 Shared Task at EMNLP:

Multilingual Physical Commonsense Reasoning Datasets

Info meeting, 2025/08/14

Contact: mrl2025-workshop@googlegroups.com

Last updated: 2025/08/31

Recruiting volunteers for other languages!

https://forms.gle/zxhpCfL6wvBzb15e7

- We currently have volunteers for the following languages (among others):

Albanian, Arabic (Standard, Tunisian, Moroccan, Saudi), Assamese, Azerbaijani, Bambara, Belarusian, Bengali, Bosnian, Bulgarian, Burushaski, Cantonese, Catalan, Croatian, Czech, Dutch, Estonian, Farsi, Finnish, French, Galician, Georgian, Greek, Gujarati, Hausa, Hawaiian, Hebrew, Hindi, Hungarian, Icelandic, Igbo, Indonesian, Italian, Japanese, Kannada, Korean, Lingala, Luo, Macedonian, Malay, Malayalam, Mandarin, Marathi, Nepali, Nigerian Pidgin, Norwegian Bokmål, Panjabi, Polish, Portuguese, Romanian, Russian, Serbian, Sindhi, Sinhala, Slovak, Slovenian, Spanish, Swahili, Swedish, Tagalog/Filipino, Tajik, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Uyghur, Uzbek, Vietnamese, Yoruba.

- Of course, more volunteers for these languages are still welcome!

Recruiting volunteers for other languages!

https://forms.gle/zxhpCfL6wvBzb15e7

- We are still actively looking for volunteers speaking the following languages (or other languages not listed):

Afrikaans, Aymara, Bashkir, Basque, Breton, Burmese, Cebuano, Chuvash, Danish, Guarani, Haitian Creole, Hmong, Inuktitut, Irish, Karakalpak, Khmer, Lao, Latvian, Malagasy, Maltese, Maori, Mongolian, Nahuatl, Navajo/Diné, Odia, Oromo, Quechua, Pushto, Samoan, Scottish Gaelic, Shona, Somali, Sundanese, Tatar, Tibetan, Tigrinya, Turkmen, Waray, Walloon, Welsh, Xhosa, Yakut, Yiddish.

- If you know any native speakers of any of these languages (or another under-resourced language), please reach out to them!

Overview

- Submission deadline: **September 15, 2025** (can be later for languages we're still missing).
- Task format (PIQA): [prompt, correct solution, incorrect solution].
- Minimum 100 examples per language, with a brief dataset description paper.
    - Of course, we welcome larger datasets!

- **All authors of accepted contributions will have the option to be included on the final benchmark paper**, which we plan to submit to the NeurIPS 2026 Datasets & Benchmarks Track.
    - For this, we may require slight modifications to the accepted datasets where necessary, and/or manual curation of a small number of additional examples.

Dataset format

- **.TSV** file with minimum columns: `prompt, solution0, solution1, label` (0 or 1 as the correct solution). Only **one** of the solutions should be correct.
- **All** examples must be manually checked by a native speaker of the language.
- Translated PIQA examples can be included, but they must be annotated as such, and they do **not** count towards the minimum 100 examples.

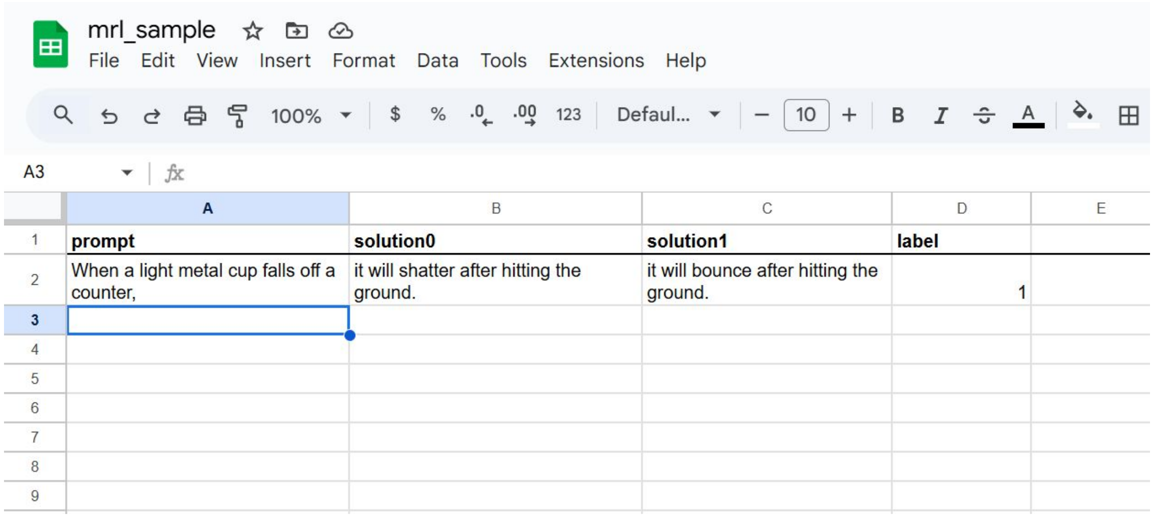
Example:

```
{"prompt": "After staining wood, you should",
"solution0": "allow it to sit for several hours so the stain can dry.",
"solution1": "allow it to sit for several months so the stain can dry.",
"label": 0}
```

Dataset format

One simple way to construct your dataset is to manually write examples.

- Common spreadsheet applications allow you to download/export to a TSV when you finish.



	A	B	C	D	E
1	**prompt**	**solution0**	**solution1**	**label**	
2	When a light metal cup falls off a counter,	it will shatter after hitting the ground.	it will bounce after hitting the ground.	1	
3					
4					
5					
6					
7					
8					
9					

Task scope

- **Physical reasoning**: For each example, the solution must relate to physical properties of one or more objects.
- **Common sense**: For each example, an average person who speaks your language natively should know the answer.
- **Culturally specific**: We encourage authors to include culturally-relevant examples for their language(s). For example, some items may not be easily translatable into English, or may require regional and/or cultural commonsense.
  - *For ideas: examples about local foods, places, everyday objects, customs, traditions, religions, literature, folklore, or art forms?*
  - *Consider common physical tasks or actions.*

Constructing the dataset

- Use **items of variable length**. Try not to include too many short items, as they may be too easy for larger models. If possible:
  - Most examples (prompt+solution) should be over 25 words long.
  - There should be some prompts that are multiple sentences long.
- The **two candidate solutions should be as similar as possible** (e.g. differing only by one or two words, or just flipping the order of two phrases). One solution should be unambiguously correct and the other incorrect.
- To ensure that the benchmark is not too "easy", the **incorrect solution should not be so absurd** that it is extremely obvious.
- Try **not** to start all examples the same way.

Some examples

- {"prompt": "When a light metal cup falls off a counter,",
  "solution0": "it will **shatter** after hitting the ground.",
  "solution1": "it will **bounce** after hitting the ground.",
  "label": 1}

- {"prompt": "To make honey chipotle marinade for chicken: In a glass bowl and with a metal spoon mix together three tablespoons canola oil, two teaspoons jarred minced garlic, three tablespoons honey and two tablespoons chopped canned chipotle peppers in adobo sauce until well blended.",
  "solution0": "Pour the marinade into a **ziploc** bag, place your chicken pieces in the bag, squeeze out excess air and close the bag. Squish the chicken around in the marinade and leave overnight.",
  "solution1": "Pour the marinade into a **paper** bag, place your chicken pieces in the bag, squeeze out excess air and close the bag. Squish the chicken around in the marinade and leave overnight."
  "label": 0}

- {"prompt": "How do you put a stamp on a T-shirt?",
  "solution0": "You **glue** it on.",
  "solution1": "You **iron** it on.",
  "label": 1}

Dataset description paper

- .PDF file; maximum eight pages in the EMNLP template, but we welcome much shorter papers.
  - Datasets that are entirely manually constructed may require only a few paragraphs of description (e.g. any heuristics or methods for writing sentences and solutions).
  - EMNLP template **not** required.

- The paper should report enough detail such that a speaker of your language(s) could reasonably reconstruct a comparable dataset.If applicable, please also specify the dialect of your language that you are using (e.g. Moroccan Arabic, Mexican Spanish, etc).

Submitting your dataset!

Deadline: **September 15, 2025** (can be later for languages we're still missing)

- Google form:

  https://forms.gle/NYZnaxakspSWwPsW6

- Also linked on our website:

  https://sigtyp.github.io/st2025-mrl.html

FAQs (slide 1/7)

Languages and dialects:

Q: I speak a language other than those listed in the slides. Can I contribute data in this language?

- **A:** Yes! We would love to have contributions in additional languages.

Q: My language can be written in multiple scripts (e.g. Cyrillic and Latin alphabets), or using distinct regional dialects (e.g. dialects of Arabic). Which script and/or dialect should I use?

- **A:** Feel free to use any script(s) and/or dialect(s) you feel comfortable with. If you feel that a dataset for multiple scripts and dialects would be more reflective of your language, then we would love to have datasets in multiple scripts and dialects. In this case, please annotate your datasets (or in the dataset description) with script and dialect information.

FAQs (slide 2/7)

Languages and dialects:

Q: My language is spoken differently in formal vs. colloquial contexts. Which register should I use?

- **A:** Please use whichever register would be appropriate in the situation described in each example. If there are significant differences between formal and colloquial speech in your language, it would be helpful to annotate examples as either formal or colloquial (e.g. an extra column in your TSV).

Q: Can we use regional acronyms (e.g. UCSD for University of California San Diego) or local brand names in our examples?

- **A:** Yes! Please refer to entities the same way that they would normally be described in your language. If an acronym is usually used, you are welcome to use the acronym.

FAQs (slide 3/7)

Workshop attendance and presentation (optional):

Q: Do I need to attend the workshop at EMNLP in person to submit a dataset?

- **A:** No, virtual attendance is allowed! Or, authors can submit a dataset to appear only in the final benchmark rather than the workshop. Workshop attendance is encouraged but *not* required.

Q: Will dataset description papers appear in the workshop proceedings?

- **A:** No, dataset description papers will not appear in the workshop proceedings. Instead, we will work with all authors to incorporate the dataset descriptions into a larger dataset benchmark paper. We plan to submit the combined paper to a high impact venue next year, and everyone who contributes data will have the option to be an author on that paper.

Q: Can dataset description papers be presented at the workshop?

- **A:** You will have the option to present a poster at the workshop if your dataset is accepted and if you will be attending.

FAQs (slide 4/7)

Dataset and paper format:

Q: What text encoding should I use for my dataset?

- **A:** We encourage you to use UTF-8 font encoding, but you can use any encoding that works for your language.

Q: Does the dataset description paper need to be in the EMNLP template with a full abstract, intro, and conclusion?

- **A:** No, the paper does *not* need to be in the EMNLP template. A paper with only a single section describing your methodology is perfectly acceptable!

Q: Can I include English translations for my examples?

- **A:** Yes! You can include TSV columns for English translations (optional), but ideally, the examples should be written originally in the language for your dataset.

FAQs (slide 5/7)

Dataset and paper format:

Q: There is additional information about some of my examples that could be relevant to researchers. Or, there are some examples that might not perfectly fit the task description (e.g. non-physical commonsense). Should I include these examples?

- **A:** These examples can be included, but only physical commonsense reasoning examples will count towards the minimum 100 per language. Any additional features should be annotated in an additional TSV column of your choice, and reported in your dataset description paper.

FAQs (slide 6/7)

Acceptance criteria:

Q: What are the acceptance criteria for datasets?

- **A:** We'll primarily be checking your dataset description to verify that you followed the requirements in these slides. For data quality, you are the experts in your language, so you are the best judges of quality.

Q: What if my dataset does not meet the requirements?

- **A:** We will respond to all submissions. If a dataset needs improvements, we will ask the authors to make those adjustments. If those adjustments are made, then the dataset will be accepted.

FAQs (slide 7/7)

LLM usage:

Q: Can I use an LLM to generate examples, if I also check all the examples manually?

- **A:** We do *not* encourage the use of LLMs to directly generate all examples. If an LLM is used, it must be reported in the dataset description paper. If an LLM is used to generate the examples, we will be more strict when checking that the dataset contains diverse examples (e.g. varying in length, not "too easy", containing culturally-specific examples). Authors are responsible for ensuring that each example is comparable to one written manually. If we raise quality issues for an LLM-generated dataset, it is unlikely that the participants will have time to write an entirely new dataset from scratch before the final benchmark release, so the dataset is more likely to be rejected.

We're open to suggestions, questions, and feedback!

Contact:

mrl2025-workshop@googlegroups.com

Tyler Chang: tachang@ucsd.edu

Catherine Arnett: catherine@eleuther.ai