A French dataset for MRL 2025

Joseph Chataignon*, Narada Maugin **
joseph.chataignon@unibe.ch , narada.maugin@gmail.com
*Universität Bern
**Université Paris-Cité

Writing process

This dataset is a data contribution for the MRL 2025 shared task for the French language (standard French as spoken in mainland France). **This dataset was written entirely by the authors.** None of it was generated using generative Artificial Intelligence.
It contains 115 PIQA-style triplets of the format [goal, correct solution, wrong solution], all written and reviewed by native French speakers.

Dataset description

The dataset is made of a TSV file comprising 4 columns:

- prompt: the question
- Solution0: the first answer
- Solution1: the second answer
- label: a number 0 or 1 indicating the correct answer

The file contains 116 rows (including the header row), with one row for each triplet.

Heuristics

Among the heuristics used to write question-answers triplets, we have been observing our daily actions and extracting their implicit knowledge into questions for the dataset.

We also sometimes used the answers of one question, and work from there to think of another question that would have the same pair of answers, but with the correct answer becoming the wrong one and vice-versa.

To make prompts longer and more challenging, we added distracting information to some of the questions.