The PLLuM Instruction Corpus

Description

We release the first representative subset of the PLLuM instruction corpus (PLLuMIC), which we believe to be useful in guiding and planning the development of similar datasets for other LLMs. It is a hand-crafted set of LLM fine-tuning instructions in Polish language, curated according to structured typology and thematic categorisation. It is an integral part of the upcoming scientific article “The PLLuM Instruction Corpus”. The research was funded by the Polish Ministry of Digital Affairs in 2024, grant num. 1/WI/DBiI/2023. We plan to continue with the research and extend the dataset in future releases.

Statistics

Total number of instructions

- 1278

Type distribution

- Adversarial: 125
- CoT: 50
- Data manipulation: 88
- Dialogue: 124
- Extraction: 71
- Formatting: 87
- Generation: 392
- Identity: 68
- Knowledge (QA): 80
- NLP: 102
- Programming: 30
- Translation: 61

Thematic distribution

- Art: 14
- Astronomy: 5
- Automotive: 6
- Biology: 78
- Chemistry: 7
- Computer science: 163
- Culinary: 52
- Culture: 55
- Ecology: 4

- Economy: 19
- Entertainment: 85
- Geography: 59
- History: 48
- Home: 60
- Hobby: 4
- Industry: 20
- Languages: 185
- Law and administration: 31
- Literature: 50
- Mathematics: 15
- Medicine: 36
- Other: 73
- Philosophy: 5
- Physics: 8
- Politics: 42
- Psychology: 19
- Religion: 7
- Society: 169
- Sports: 26
- Technology: 87
- Travel: 25

Apply for access

To gain access to the PLLuMIC dataset, which was produced as a result of the CLARIN-BIZ project, we kindly ask you to take a moment to complete the form provided below. Your cooperation in this process is greatly appreciated, and we thank you for your interest in our work.

PLLuMIC access form

Dataset file explanation

The PLLuMIC dataset is distributed as a JSON file storing a list of conversations between a user and an AI assistant. Each conversation is also a JSON file described by following fields:

Top-Level Fields

- **dataset_name**: Name of the dataset (`PLLuMIC`).
- **dataset_source**: Source organization (`CLARIN-BIZ-bis`).
- **conv_id**: Unique identifier for the conversation (`3242183cbce2`).
- **messages**: Array of dialogue messages (user/assistant/system exchanges).

Message Object Fields

Each entry in `messages` contains:

- **instruction_id**: Unique ID for the instruction/task (`2a07c2eca0cb`).
- **seq**: Sequence number (`-1` for system, `0,1,2,…` for user/assistant turns).
- **role**: Speaker role (`system`, `user`, or `assistant`).
- **content**: Text of the message (empty for some system prompts).
- **type**: Interaction type (e.g., `Dialog`, `Generation`).
- **subtype**: Task subtype (e.g., `System prompt`, `Text simplification`).
- **topic**: Relevant topics (e.g., `Geography`).
- **language**: Language code (e.g., `pol` for Polish).
- **source**: References (e.g., Wikipedia URLs).

Disclaimer

Please do not redistribute.