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 LLM datasets. PLLuMIC, at its core, is a **hand-crafted set** of LLM fine-tuning Polish language instructions. The corpus is described in more detail in a forthcoming paper titled *The PLLuM Instruction Corpus*. We plan regular updates and significant extensions of the corpus.

The data is divided into two subsets: main organic part and synthetic extension.

The organic samples were carefully curated by human annotators, developed in line with the annotation guidelines and covering a functional typology. The synthetic extension was created using a strong, permissively licensed LLM (DeepSeek v3) and a custom pipeline incorporating organic samples injection.

Apply for access

The PLLuMIC dataset is available here: https://huggingface.co/datasets/pelcra/PLLuMIC

Statistics

The main organic subset (PLLuMIC)

Total instructions: 1,278

All instructions were annotated by professional annotators. Each sample was developed in accordance with comprehensive annotation guidelines and subsequently reviewed by a senior annotator to ensure full compliance with quality standards. The annotation process followed a functional typology designed to encompass key areas of model competence.

Type distribution

| Туре | Number of samples |
|-------------|-------------------|
| _ | _ |
| Generation | 392 |
| Adversarial | 125 |
| Dialogue | 124 |
| NLP | 102 |

| Data manipulation | 88 |
|-------------------|----|
| Formatting | 87 |
| Knowledge (QA) | 80 |
| Extraction | 71 |
| Identity | 68 |
| Translation | 61 |
| СоТ | 50 |
| Programming | 30 |

Thematic distribution

| Туре | Number of samples |
|------------------------|-------------------|
| _ | _ |
| Languages | 185 |
| Society | 169 |
| Computer science | 163 |
| Technology | 87 |
| Entertainment | 85 |
| Biology | 78 |
| Other | 73 |
| Home | 60 |
| Geography | 59 |
| Culture | 55 |
| Culinary | 52 |
| Literature | 50 |
| History | 48 |
| Politics | 42 |
| Medicine | 36 |
| Law and administration | 31 |
| Sports | 26 |
| Travel | 25 |
| Industry | 20 |
| Economy | 19 |
| Psychology | 19 |
| Mathematics | 15 |
| Art | 14 |
| Physics | 8 |
| Chemistry | 7 |
| Religion | 7 |
| Automotive | 6 |
| Philosophy | 5 |
| Astronomy | 5 |
| Ecology | 4 |
| Hobby | 4 |

https://pelcra.pl/ Printed on 2025/11/30 16:59

The synthetic extension

Total instructions: 54,921

Each type and subtype has been handled individually, with careful attention to quality standards and guidelines. Each synthetic sample was generated by injecting suitable organic examples, with differentiation measures applied to ensure diversity. There are currently no system prompts in the subset, but there is an ongoing work to include them in the nearest future.

Type distribution

| Туре | Number of samples |
|-------------------|-------------------|
| _ | _ |
| Generation | 21548 |
| Extraction | 7818 |
| Knowledge (QA) | 4599 |
| Data manipulation | 4550 |
| Formatting | 4380 |
| Programming | 3253 |
| NLP | 2905 |
| Adversarial | 2663 |
| СоТ | 1793 |
| Translation | 1412 |

All subtypes within these types are covered. The thematic categorisation is yet to come in future updates.

Dataset file explanation

The PLLuMIC dataset is distributed as a JSONL file storing rows with conversations between a user and an Al assistant. There are 2 JSONL files included, one for the organic component and one for the synthetic extension. Each conversation is a JSON structure 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**: List of task subtype (e.g., [System prompt, Text simplification]).
- topic: List of relevant topics (e.g., [Geography]).
- language: Language code (e.g., pol for Polish).
- source: References (e.g., Wikipedia URLs).

Disclaimer

Please do not redistribute.

From:

https://pelcra.pl/ - PELCRA

Permanent link:

https://pelcra.pl/doku.php?id=pllumic&rev=1763307092

Last update: 2025/11/16 15:31



https://pelcra.pl/ Printed on 2025/11/30 16:59