

# 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 is a hand-crafted set of LLM fine-tuning Polish language instructions, developed in line with the annotation guidelines and covering a functional typology. The corpus is described in more detail in a forthcoming paper titled *The PLLuM Instruction Corpus* (Pęzik et al. 2025). We plan regular updates and significant extensions of the corpus.

---

## 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.

From:

<http://docs.pelcra.pl/> - **Group**

Permanent link:

<http://docs.pelcra.pl/doku.php?id=pllumic>

Last update: **2025/04/25 08:25**

