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.

Please Cite

https://arxiv.org/abs/2511.17161

```
@misc{pęzik2025plluminstructioncorpus,
      title={The PLLuM Instruction Corpus},
      author={Piotr Pęzik and Filip Żarnecki and Konrad Kaczyński and Anna
Cichosz and Zuzanna Deckert and Monika Garnys and Izabela Grabarczyk and
Wojciech Janowski and Sylwia Karasińska and Aleksandra Kujawiak and Piotr
Misztela and Maria Szymańska and Karolina Walkusz and Igor Siek and Maciej
Chrabąszcz and Anna Kołos and Agnieszka Karlińska and Karolina Seweryn and
Aleksandra Krasnodębska and Paula Betscher and Zofia Cieślińska and
Katarzyna Kowol and Artur Wilczek and Maciej Trzciński and Katarzyna
Dziewulska and Roman Roszko and Tomasz Bernaś and Jurgita Vaičenonienė and
Danuta Roszko and Paweł Levchuk and Paweł Kowalski and Irena Prawdzic-
Jankowska and Marek Kozłowski and Sławomir Dadas and Rafał Poświata and
Alina Wróblewska and Katarzyna Krasnowska-Kieraś and Maciej Ogrodniczuk and
Michał Rudolf and Piotr Rybak and Karolina Saputa and Joanna Wołoszyn and
Marcin Oleksy and Bartłomiej Koptyra and Teddy Ferdinan and Stanisław
Woźniak and Maciej Piasecki and Paweł Walkowiak and Konrad Wojtasik and
Arkadiusz Janz and Przemysław Kazienko and Julia Moska and Jan Kocoń},
      year={2025},
      eprint={2511.17161},
      archivePrefix={arXiv},
      primaryClass={cs.CL},
      url={https://arxiv.org/abs/2511.17161},
}
```

Apply for access

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

Statistics

The 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

Type	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
CoT	50
Programming	30

Thematic distribution

Type	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

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

Type	Number of samples
—	—
Generation	21548
Extraction	7818
Knowledge (QA)	4599
Data manipulation	4550
Formatting	4380
Programming	3253
NLP	2905
Adversarial	2663
CoT	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 AI 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).
