10TH IT MANAGER OF TOMORROW by Conlea

National LLMs: Towards Models Focused on the Local Culture and Language

Marek Kozłowski

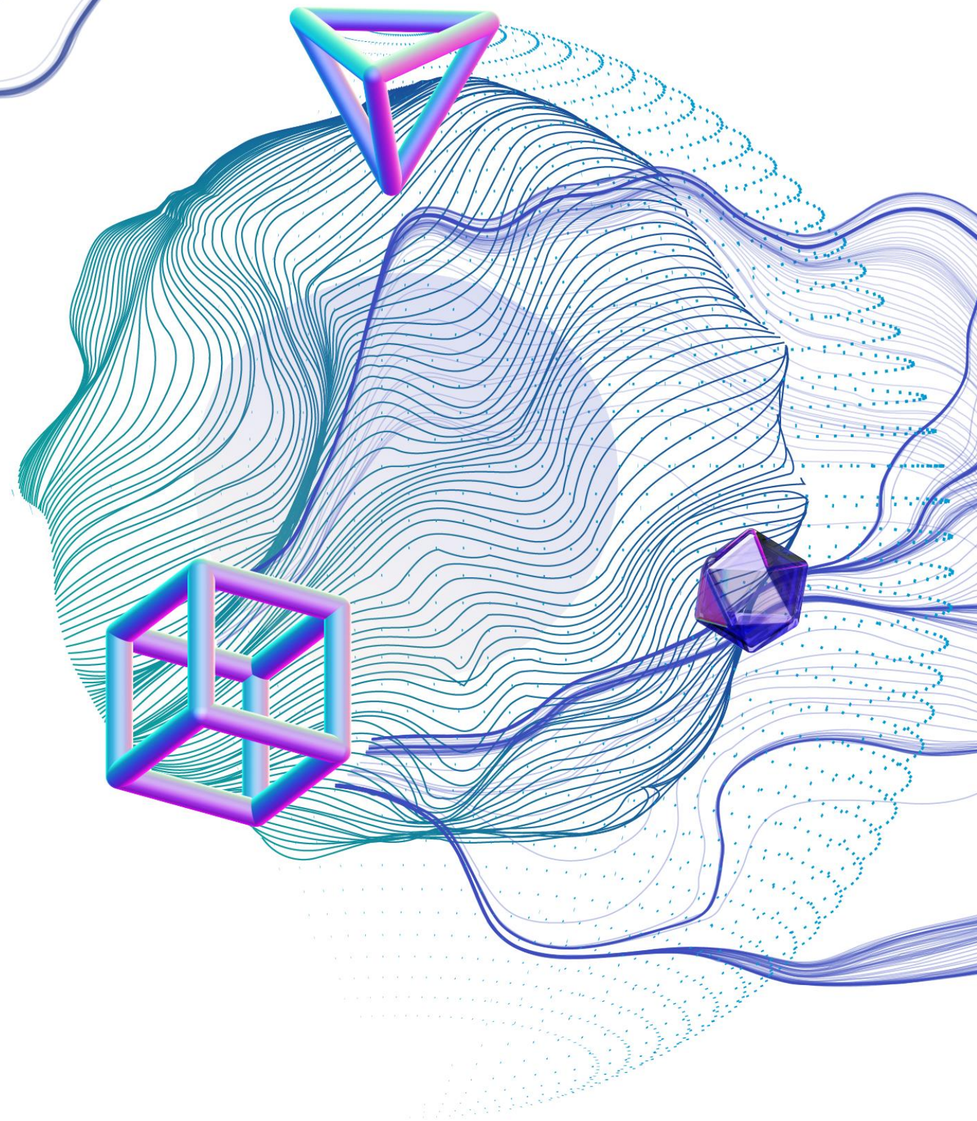
National Information Processing Institute OPI PIB

NATIONAL INFORMATION PROCESSING INSTITUTE

AI Lab

We push the boundaries of artificial intelligence

in Polish Natural Language Processing and Computer Vision for factory halls



Bank Polski

VIVE® Textile Recycling

Urząd Ochrony Konkurencji i Konsumentów

alt-edic

Ministerstwo Cyfryzacji

LLMs – the strategic decisions

Mostly, when deploying LLMs, we must answer some crucial questions:

- whether to use cloud-based or on-premises LLMs
- whether to use text-only or multi-modal models
- how large a model model is good enough for our business goals (*a trade off between efficiency and cost*)
- whether to use models in a few-shot mode or to fine-tune them
- **whether to choose multilingual models or single-language ones –> broad language coverage vs specialization**

Localized models – are they worth it?

Nowadays, in the LLMs world the most popular ones are generative multilingual models from vendors such as OpenAI, Anthropic, Google, Meta, Cohere.

However, you have to remember that neural language models (NLMs) are not only generative.

Two main categories of NLMs:

- representation models, such as BERT, RoBERTa, and DeBERTa – very good at classification and regression tasks
- generative models such as encoder-decoder (T5, BART), and decoder-only (GPT, Llama, Claude, Mistral)

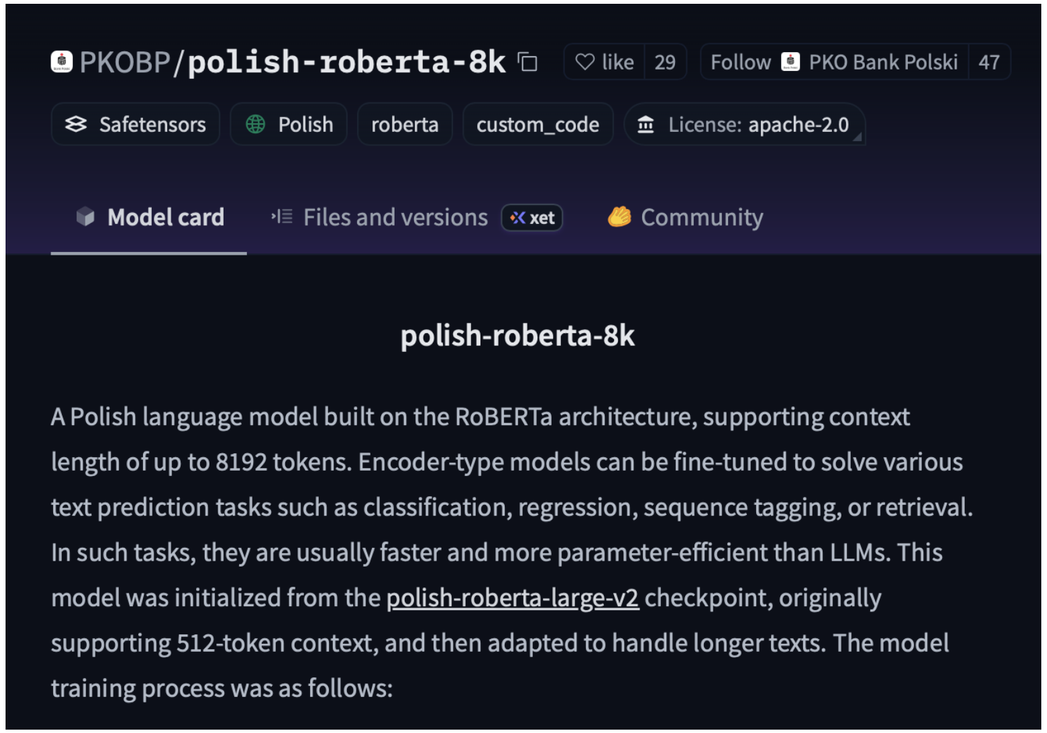
Localized models - are they worth it ?

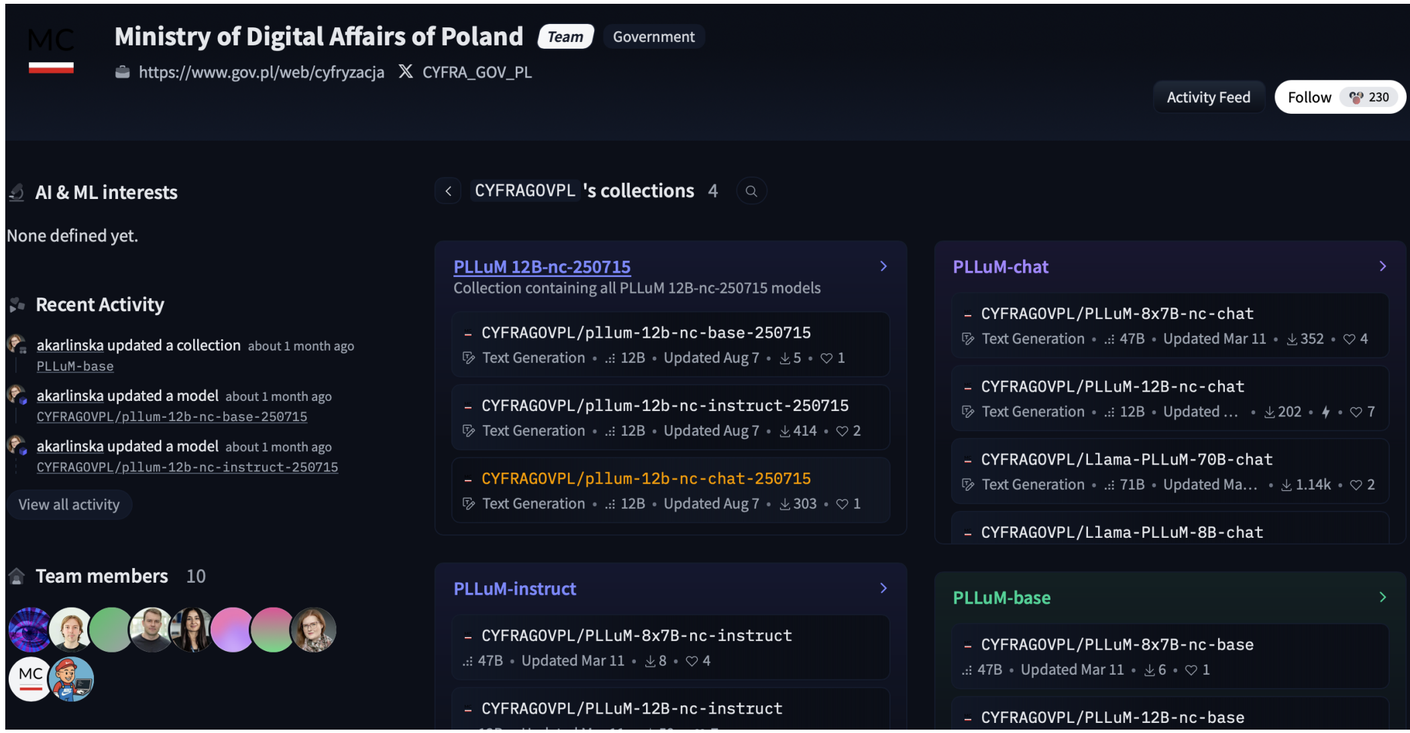
Polish localization means improving the understanding of the Polish language and creating high-quality Polish content.

In AI LAB, we built Polish oriented representation models – Polish RoBERTa with both standard (512 tokens) and extended context (8 192 tokens). Experiments proved that they perform slightly better in Polish text classification/regression tasks than popular modern multi-lang models such as mmBERT or EUROBERT.

We took part in the national consortium projects (PLLuM, HIVE AI) to build Polish-oriented generative models ranging from 8B to 70B parameters. These models produce Polish content of higher linguistic quality.

Our newest Localized LLMs





PLLuM

Polish Large Language Model

Because AI should speak fluently in more languages than just English and Chinese.

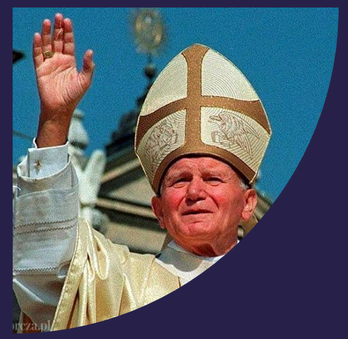
Because AI should be easily accessible to national sectors where privacy and security are crucial.

Conversational abilities involve not only words and

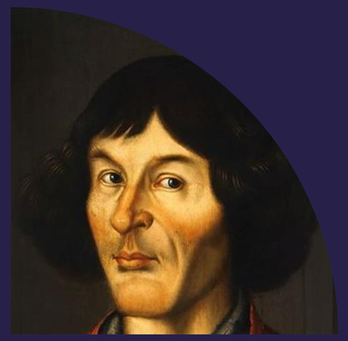




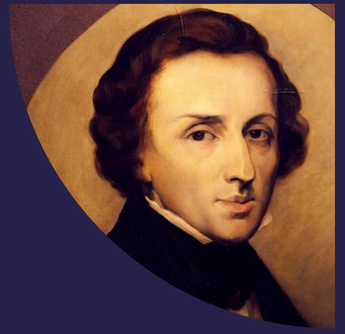










PLLuM

Polish Large Language Model

1) We did not start from scratch (i.e. random weights) – instead, we performed language adaptation before SFT/alignment.

2) We used mostly organic Polish written instructions and preferences, which prevent counter-adaptation.

3) It is not a fully capable instruction model – only hundreds of instruction types are covered. However, each of them provides a relevant advantage (e.g. emails generation without cross-language transfer learning).

Polish Linguistic and Cultural Competencies (PLCC)

Model	Score
O1-2024-12-17	89
DeepSeek-v3.1	71
PLLuM-12B-nc-chat-250715	70
PLLuM-8x7-nc-chat-250224	68
GPT-4-turbo	67
DeepSeek-R1-Llama-70B	44

Polish linguistic and cultural competency benchmark

AUTHORS: Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, Rafał Poświata

AFFILIATION: National Information Processing Institute

UPDATED: 01/02/2025

Large language models (LLMs) are becoming increasingly proficient in processing and generating multilingual texts, which allows them to address real-world problems more effectively. However, language understanding is a far more complex issue that goes beyond simple text analysis. It requires familiarity with cultural context, including references to everyday life, historical events, traditions, folklore, literature, and pop culture. A lack of such knowledge can lead to misinterpretations and subtle, hard-to-detect errors. To examine language models' knowledge of the Polish cultural context, we introduce the **Polish Linguistic and Cultural Competency Benchmark**, consisting of **600 manually crafted questions**. The benchmark is divided into six categories: history, geography, culture & tradition, art & entertainment, grammar, and vocabulary. This evaluation provides a new perspective on Polish competencies in language models, moving past traditional natural language processing tasks and general knowledge assessment.

https://huggingface.co/spaces/sdadas/plcc

PLLuM

Polish Large Language Model

Trained

- *in* Polish
- *by* Poles
- *for* Polish users

Politechnika Wrocławska

NASK

IFPAN

UNIWERSYTET ŁÓDZKI

SLAV

OPIPIB

IT WAS

NO SOLO ACT

Startup challenge

Goals

1. **Open and legally compliant Polish large language model**
2. **Polish-speaking intelligent assistant:** open general-purpose assistant and RAG (domain-oriented) assistant

Implementation period:
11 months
Team:
over 100 people

What makes us different?

LEGAL

High-quality, legaly sourced text data

What makes us different?

LEGAL

High-quality, legaly sourced text data

ORGANIC

Original data for model traning

What makes us different?

LEGAL

High-quality, legaly sourced text data

ORGANIC

Original data for model traning

VARIOUS

A family of models for diverse applications

What makes us different?

LEGAL

High-quality, legaly sourced text data

ORGANIC

Original data for model traning

VARIOUS

A family of models for diverse applications

SECURE

Embedded security features

Hundreds of Hopper GPUs, each worth several dozens of thousands of EUR, were used to build the models

Dozens of training sessions were conducted, some lasting 2-3 weeks

Commercial Deployments

The first production deployment
(one of the IT leaders in CEE)

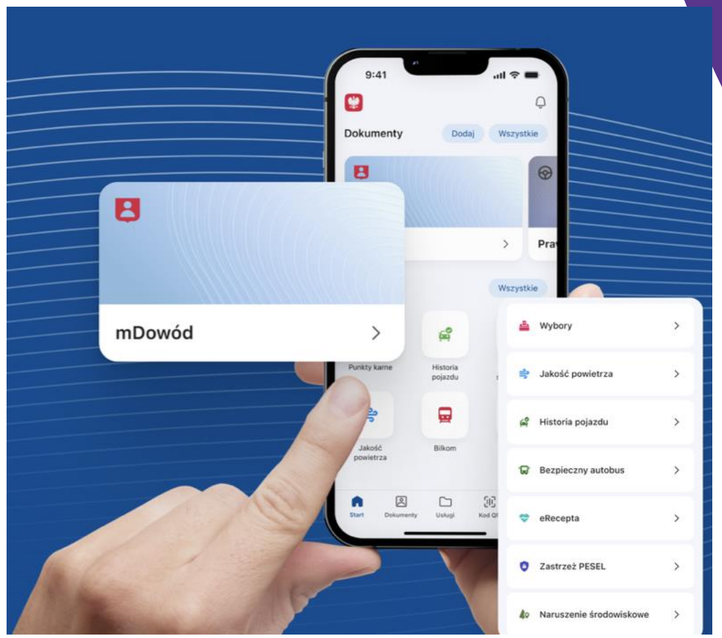
COMARCH

The second production deployment
(the biggest bank in CEE)

Public Deployments

Citizen assistant (mObywatel)

mObywatel 2.0



Assistants for city offices – state-level administrative use cases

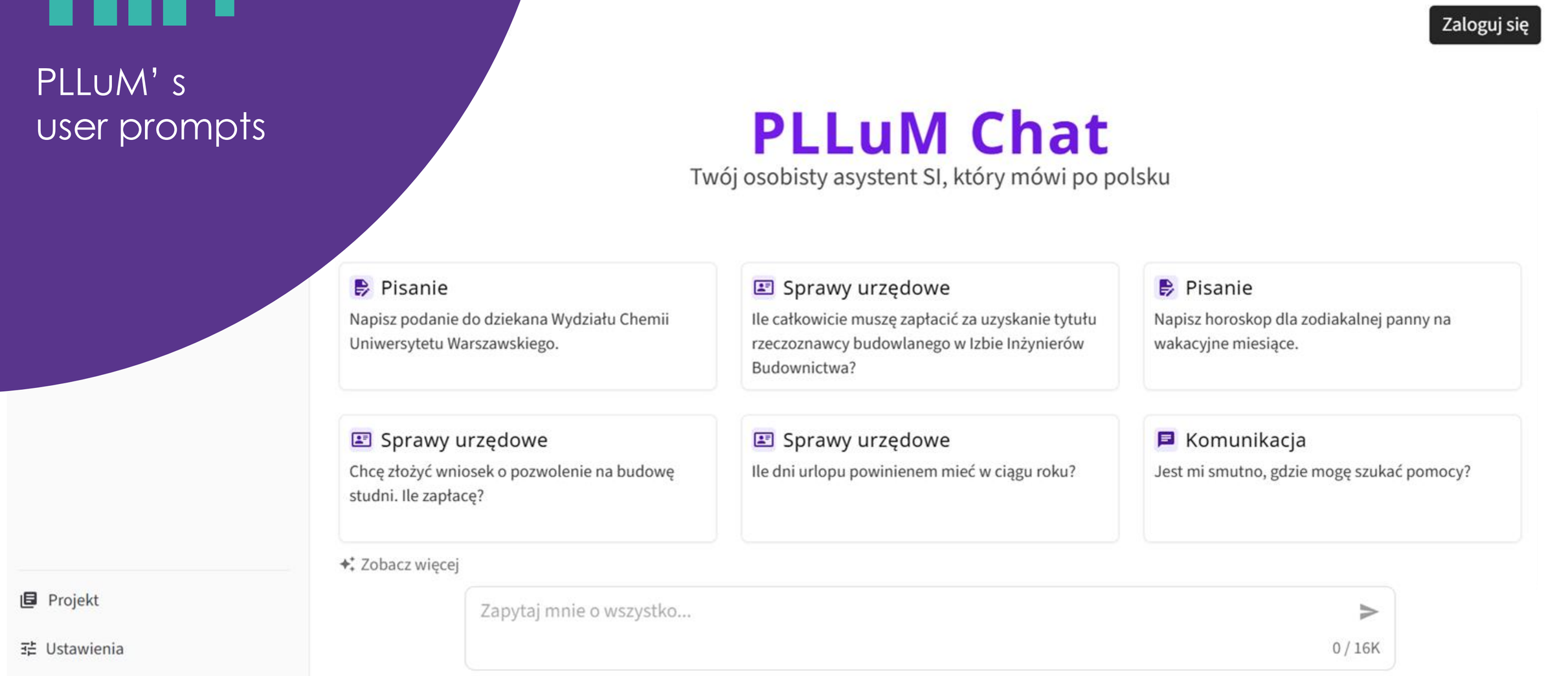
OVER

1m+

PLLuM' s
user prompts

Chat assistant

https://pllum.clarin-pl.eu



Why care about a Polish AI model?

1. Technological sovereignty

2. Language is an identity

3. Democratizing access
(open-source and transparency)

This is not just tech – it’s a mission

AI should serve citizens and local companies.

What’s in it for Poland – and beyond

1. New opportunities for companies and the public sector
2. A boost for innovation
3. Technological skills that can be leveraged in the future

Language localisation

Article Talk

From Wikipedia, the free encyclopedia

Language localisation (or **language localization**) is the process of adapting a product's translation to a specific country or region. It is the second phase of a larger process of product translation and cultural adaptation (for specific countries, regions, cultures or groups) to account for differences in distinct markets, a process known as internationalisation and localisation.

PLLuM

Polish Large Language Model

Today, localizations are a must; small localized models will become part of this process in the future.

PLLuM
Polish Large Language Model

10TH IT MANAGER OF TOMORROW by Conlea

Thank you!
Dziękuję!