

Dane na potrzeby suwerennej AI

– znaczenie danych w tworzeniu polskich modeli językowych



Aleksandra Tomaszewska
Instytut Podstaw Informatyki PAN

Warsztat LDS | 29.05.2025 | Warszawa

4.5 megabytes of data in 62,500 punched cards, USA, 1955



Jak chronić nasze dane? „Dane osobowe to złoto XXI wieku”

POLSKIERADIO 24

INNOVATION

How To Make Use Of The New Gold: Data

Analysis

Data is the new gold

The future of real estate service providers

Dane. Złoto naszych czasów.
Pierwsza książka oparta o polskie przykłady



Krzysztof GAWKOWSKI

Doktor nauk humanistycznych, działacz Sojuszu Lewicy Demokratycznej, Wiosna i Nowej Lewicy, przewodniczący klubu parlamentarnego Lewicy (od 2019 r.). Poseł na Sejm IX kadencji

[zobacz inne teksty Autora](#)

Wartość danych **dziś**

Wspomaganie,
automatyzacja,
przyspieszanie
procesów, rozwiązywanie
problemów

Potencjał
badawczy,
wdrożeńowy
– biznesowy

Niezbędne
w tworzeniu
modeli i narzędzi
językowych

Podwaliny
narodowej/lokalnej
AI na różnych
etapach tworzenia

Użytkownik anonimowy 

Cześć! Przedstaw się proszę krótko – niech dowie się o Tobie publiczność podczas warsztatów LDS w Polsce.

 PLlum

Cześć! Nazywam się PLLuM (od ang. Polish Large Language Model) i jestem otwartym modelem językowym. Moją misją jest wspieranie ludzi w ich zadaniach poprzez przetwarzanie i generowanie tekstu w języku polskim.



Po co nam polskie modele AI?

1 Pełna kontrola i możliwość dostosowania

2 Niezależność

3 Bezpieczeństwo i koszty

4 Rozwój kompetencji

5 Polskie realia i język

O projekcie PLLuM

Polish Large Language Model

Projekt konsorcjum polskich jednostek naukowych

Politechniki Wrocławskiej (lider), Instytutu Podstaw Informatyki PAN, NASKu, OPI,
Instytutu Slawistyki PAN, Uniwersytetu Łódzkiego

Finansowany przez Ministerstwo Cyfryzacji

około 14,5 mln zł

Realizowany w 2024 roku przez 11 miesięcy

22.01–31.12.2024



Ministerstwo
Cyfryzacji

Tu tworzymy przyszłość



NASK



Cele projektu PLLuM

Model

Stworzenie otwartego **polskojęzycznego modelu językowego**

Asystent

Opracowanie prototypowego zastosowania modelu dla administracji publicznej – **inteligentnego asystenta**

Dane

Zbudowanie zestawu **zróżnicowanych danych**



Jak tworzyć modele? Od **danych** do wychowania

Dane

Identyfikujemy, pozyskujemy, sprawdzamy, poprawiamy, wybieramy **dane**

Tokenizacja



Pretrening

Budujemy **model fundamentalny**, który generuje tekst na podstawie polecenia słowo po słowie / kontynuowany pretrening

Dostrajanie

Douczone model na **zestawach poleceń** i przykładowych odpowiedzi/rozwiązań

Wychowanie

Dostosowujemy model do **preferencji i zasad etycznych** w oparciu o reakcje użytkowników i użytkowniczek

Droga do modeli PLLuM

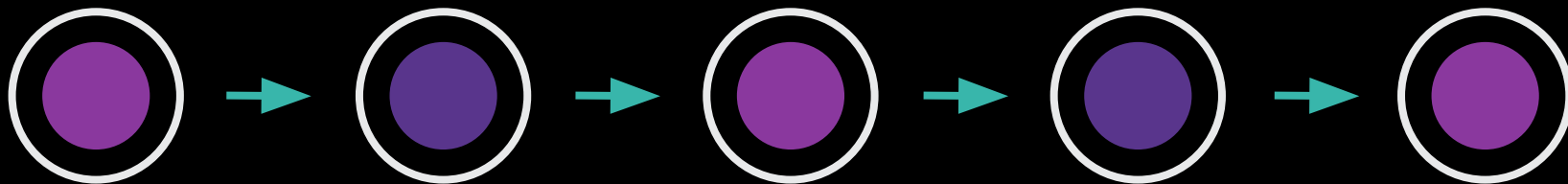
Zbieranie danych
od lutego

Trenowanie
czerwiec-wrzesień

Douczenie
październik

Optymalizacja
listopad

Udostępnienie
grudzień



Gromadzenie danych treningowych, opracowanie systemu do zarządzania danymi

Bazowy trening modelu dostrojonego do różnych zadań

Douczenie i wychowywanie modelu dialogowego w oparciu o reakcje użytkowników i użytkowniczek

Monitorowanie pracy modułu korekty wyjścia

Udostępnienie modelu oraz projektu wirtualnego asystenta



Ministerstwo
Cyfryzacji

Tu tworzymy przyszłość

GOI
Centrum
Ośrodek
Innowacji



NASK

OPIPIB



Pozyskiwanie danych do PLLuM-a

CEL

Tworzenie dużych zbiorów danych **najlepszej jakości**

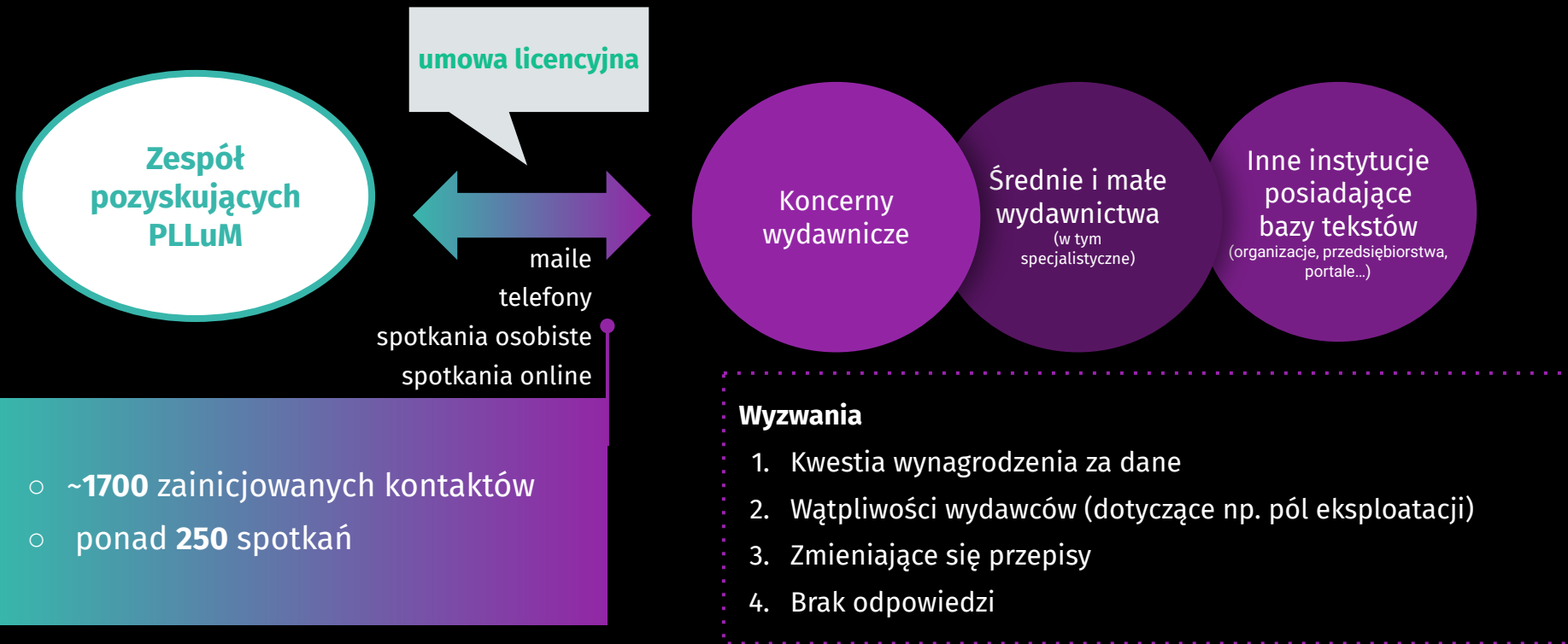


Dane, których nie ma w sieci,
dane z ograniczonym
dostępem



Dane publicznie dostępne
(względny etyczne)

Współtworzenie danych do modelu



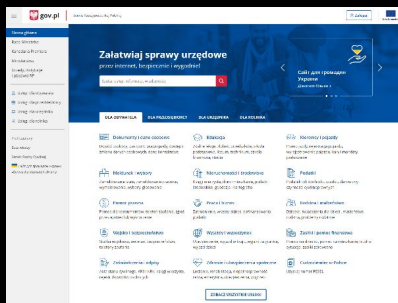
Nowe standardy pozyskiwania **danych**

- **Cel: pozyskanie i stworzenie zbiorów danych bardzo dobrej jakości**
- **Analiza prawna**
- **Dane wielojęzyczne:** polski, inne języki słowiańskie, angielski
- Polskie dane do wstępnego uczenia – ok. **150 mld** słów, 30 mld do zastosowań otwartych
- Ręcznie tworzone polskie zbiory do kolejnych etapów tworzenia modelu
 - **50 tys.+** instrukcji (generowanie, ekstrakcja, klasyfikacja, rozmowa...)
 - **130+ tys.** preferencji (np. etyka)
- Własne zbiory i metody oceny jakości

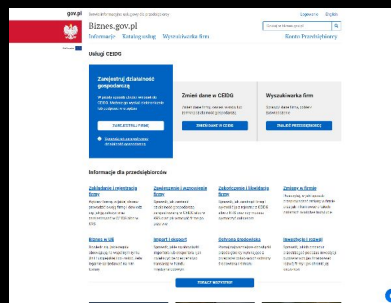


Na jakich danych uczy się asystent?

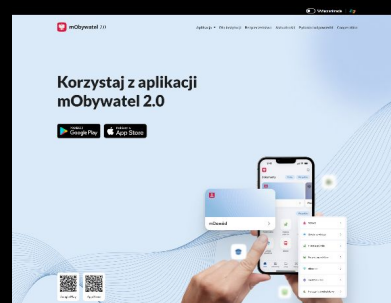
- Bazę wiedzy stanowią dokumenty urzędowe pobrane z publicznych domen gov
- Poza tymi dokumentami, zebraliśmy także **kilkadziesiąt tysięcy** tekstów o tematyce urzędowo-prawniczej z innych źródeł, które są używane do trenowania i ewaluacji modeli



gov.pl



biznes.gov.pl



mobywatel.gov.pl



Ministerstwo
Cyfryzacji

Tu tworzymy przyszłość

GOI
Centralny
Ośrodek
Informacji



NASK

OPIPIB



Efekty: dostępne modele

PLLuM-8B

Każdy w 3 wersjach:
Bazowej
Instrukcyjnej
Dialogowej (po wychowaniu)

PLLuM-12B

PLLuM-12B-nc

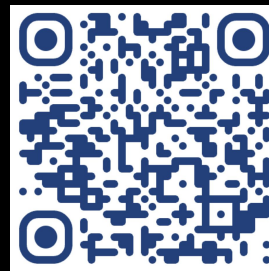
PLLuM-8x7B*

PLLuM-8x7B-nc*

PLLuM-70B

*13B aktywnych parametrów

MODELE



-  Modele na licencjach w pełni otwartych
-  Modele na ograniczonych licencjach (niepozwalających na użycie komercyjne)



Ministerstwo
Cyfryzacji

Tu tworzymy przyszłość

GOI
Centrum
Ośrodek
Innowacji



NASK

OPIPIB



Okienko PLLuM Chat

 PLLuM 8x7B
PLLUM

 Biblioteka pytań

 Nowy czat

Historia czatu

 Projekt

 Ustawienia

 PLLuM 8x7B

Zaloguj się

PLLUM Chat

Twój osobisty asystent SI, który mówi po polsku

 Pisanie
Napisz podanie do dziekana Wydziału Chemii Uniwersytetu Warszawskiego.

 Sprawy urzędowe
Ile całkowicie muszę zapłacić za uzyskanie tytułu rzeczoznawcy budowlanego w Izbie Inżynierów Budownictwa?

 Pisanie
Napisz horoskop dla zodiakalnej panny na wakacyjne miesiące.

 Sprawy urzędowe
Chcę złożyć wniosek o pozwolenie na budowę studni. Ile zapłacę?

 Sprawy urzędowe
Ile dni urlopu powinienem mieć w ciągu roku?

 Komunikacja
Jest mi smutno, gdzie mogę szukać pomocy?

 Zobacz więcej

Zapytaj mnie o wszystko...

0 / 16K



CZAT

Prompt book: jak rozmawiać z modelem?



↑ Powrót na początek



Przykłady użycia PLLuM

- Podanie
- Ogłoszenie
- CV
- List motywacyjny
- Przygotowanie do rozmowy o pracę
- Rekomendacje
- Opinie
- Życzenia
- Planowanie
- Post na social media
- Post na bloga
- Quiz
- Test językowy
- Tłumaczenie językowe
- Krok po kroku
- Opowiadanie historii
- Wróżby
- Horoskop
- Konwersacje
- Wyjaśnianie dowcipów

- Wyjaśnianie memów
- Wyjaśnianie polskich idiomów
- Tłumaczenie pojęć
- Zmiana stylu
- Parafrazowanie
- Porównywanie
- Recenzje produktów
- Przemówienia
- Poezja
- Kolokacje
- Upraszczenie tekstu
- Streszczenia
- Poprawki stylistyczne
- Związki frazeologiczne
- Klasyfikacja
- Analiza wydźwięku
- NER
- Słowa kluczowe
- Podobieństwo semantyczne
- Streszczenia
- Wpływ ciągłej łączności na zdrowie...
- Error-correction
- Formatowanie tekstu
- Ekstrakcja informacji



Ministerstwo
Cyfryzacji

Tu tworzymy przyszłość



NASK



Konsorcjum **HIVE**



- **Nowi partnerzy:** COI i Cyfronet AGH
- Modele PLLuM zostaną **wdrożone w administracji publicznej**
- Modele PLLuM będą **wykorzystywane w mObywatelu**



Strategie

Współtworzenie
suwerenności
technologicznej

Wzmocnienie
bezpieczeństwa
i pełna kontrola

Udział
w transformacji
cyfrowej sektora
publicznego

Wdrożenia
interdyscyplinarne



Zbieranie danych dla AI



Paradoks dostępu do danych

- Coraz więcej zasobów, coraz większe znaczenie danych
- **Coraz większe wyzwania związane z dostępem do danych dla twórców AI:**
wyzwania licencyjne, rosnąca świadomość, monetyzacja
- Jakość i interoperacyjność
- Ograniczanie dostępu do danych; '*consent in crisis*' (Longpre et al. 2024)

Consent in Crisis: The Rapid Decline of the AI Data Commons

Shayne Longpre¹, Robert Mahari¹, Ariel Lee¹, Campbell Lund¹, Hamidah Oderinwale², William Brannon², Nayan Saxena², Naana Obeng-Marnu², Tobin South², Cole Hunter², Kevin Klyman², Christopher Klamm², Hailey Schoelkopf², Nikhil Singh², Manuel Cherep², Ahmad Mustafa Anis³, An Dinh³, Caroline Chitongo³, Da Yin³, Damien Sileo³, Deividas Mataciunas³, Diganta Misra³, Emad Alghamdi³, Enrico Shippole³, Jianguo Zhang³, Joanna Materzynska³, Kun Qian³, Kush Tiwary³, Lester Miranda³, Manan Dey³, Minnie Liang³, Mohammed Hamdy³, Niklas Muennighoff³, Seonghyeon Ye³, Seungone Kim³, Shrestha Mohanty³, Vipul Gupta³, Vivek Sharma³, Vu Minh Chien³, Xuhui Zhou³, Yizhi Li³, Caiming Xiong⁴, Luis Villa⁴, Stella Biderman⁴, Hanlin Li⁴, Daphne Ippolito⁴, Sara Hooker⁴, Jad Kabbara⁴, and Sandy Pentland⁴

¹Team Leads, ²Top Contributors, ³Contributors (alphabetized), ⁴Advisors

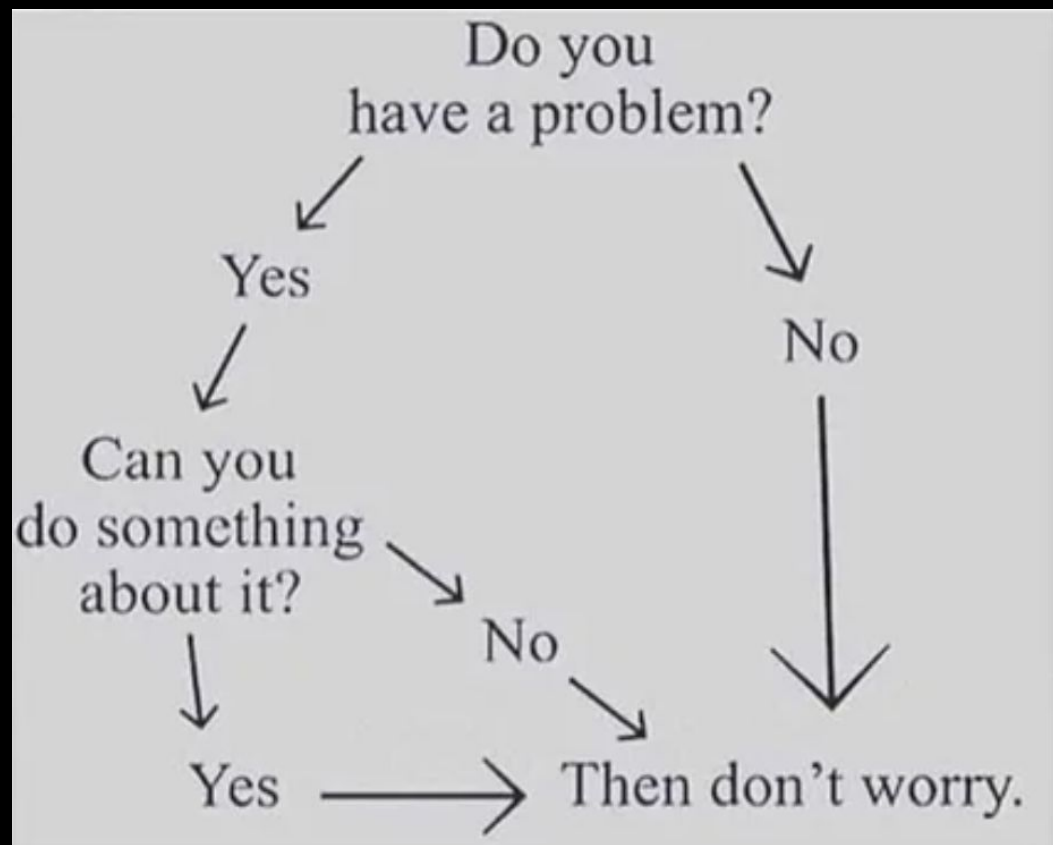
Abstract

General-purpose artificial intelligence (AI) systems are built on massive swathes of public web data, assembled into corpora such as C4, RefinedWeb, and Dolma. To our knowledge, we conduct the first, large-scale, longitudinal audit of the consent protocols for the web domains *underlying* AI training corpora. Our audit of 14,000 web domains provides an expansive view of crawlable web data and how codified data use preferences are changing over time. We observe a proliferation of AI-specific clauses to limit use, acute differences in restrictions on AI developers, as well as general inconsistencies between websites' expressed intentions in their Terms of Service and their robots.txt. We diagnose these as symptoms of ineffective web protocols, not designed to cope with the widespread re-purposing of the internet for AI. Our longitudinal analyses show that in a single year (2023-2024) there has been a rapid crescendo of data restrictions from web sources, rendering ~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use. For Terms of Service crawling restrictions, a full 45% of

TDM Opt-out: Zastrzeżenie praw do celów eksploracji tekstów i danych

Stowarzyszenie Autorów ZAiKS dokonuje zastrzeżenia wobec wszelkich działań z zakresu eksploracji tekstów i danych. Zastrzeżenie dotyczy wszystkich utworów, do których prawa reprezentuje Stowarzyszenie, w zakresie w jakim stanowią one w całości lub części element dowolnego rodzaju rozpowszechnionych treści.

Wszelkie działania z zakresu eksploracji tekstów i danych dokonywane na treściach zawierających utwory, do których prawa reprezentuje Stowarzyszenie Autorów ZAiKS, bez uzyskania zgody Stowarzyszenia będą stanowiły naruszenie praw autorskich uprawnionych reprezentowanych przez Stowarzyszenie.



📅 23.12.2024

Komisja Europejska po raz 10. opublikowała ranking Open Data Maturity, w którym Polska utrzymuje 2. miejsce wśród 34 krajów europejskich, biorących udział w badaniu. Polska pozostaje w ścisłej europejskiej czołówce w zakresie udostępniania informacji, obok Francji i Słowacji.



Ranking

**Polska na 2. miejscu
w rankingu Komisji
Europejskiej Open
Data Maturity 2024**



gov.pl

Serwis Rzeczypospolitej Polskiej

Mój Gov



Unia Europejska

Strona informacyjna - Pomoc dla powodzi

[PRZEJDŹ DO STRONY](#)



A A+ A++ A+++

[PL](#) [EN](#)

[Zarejestruj się](#) /

[Zaloguj się](#)



[Strona główna](#)

[Dane](#)

[Dostawcy](#)

[PoCoTo](#)

[Aktualności](#)

[Baza wiedzy](#)

[Punkt informacyjny DGA](#)

Portal danych

Korzystaj z danych bezpłatnie, również do celów komercyjnych

Wyszukaj dane...



Projekt: "Dane 3.0 - wymiana, wartość"

Fundusze Europejskie

na Rozwój Cyfrowy



Fundusze Europejskie
na Rozwój Cyfrowy



Rzeczpospolita
Polska

Dofinansowane przez
Unię Europejską



Użytkownik anonimowy 

Mam pytanie – do czego potrzebujesz danych językowych? I jakie powinny być te dane?



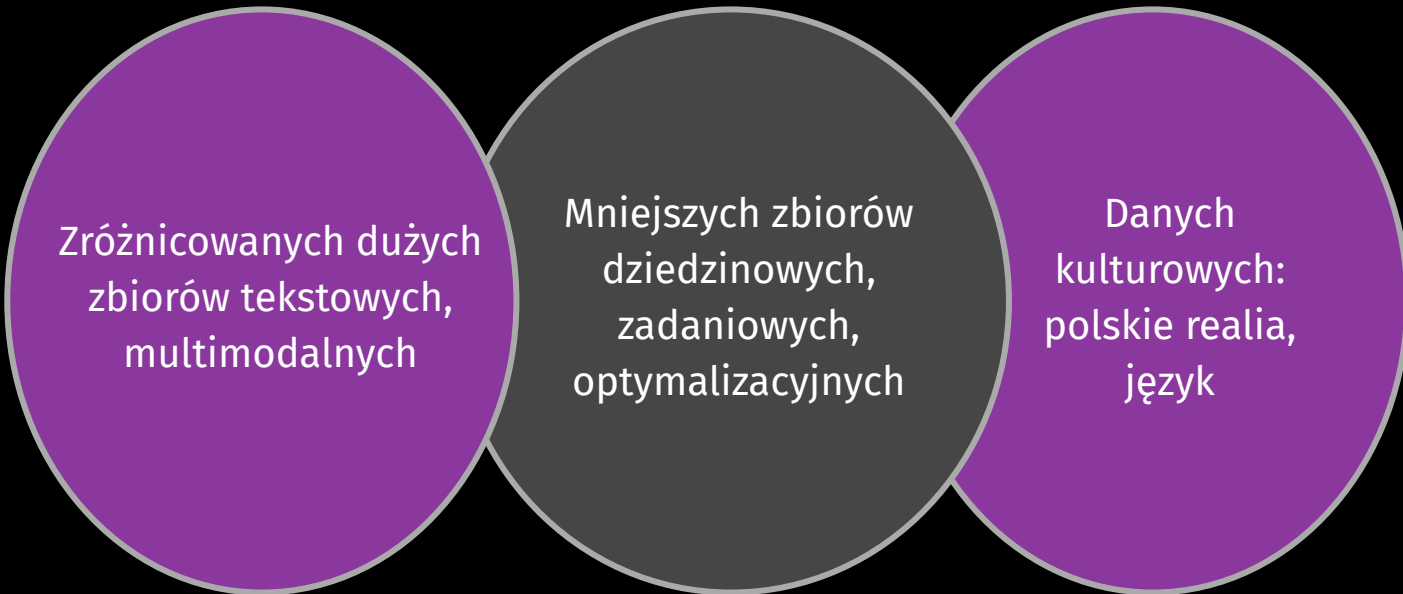
PLLuM 8x7B
PLLuM

Jako model językowy, potrzebuję danych językowych, aby móc zrozumieć i generować teksty w języku polskim. Dane te służą do uczenia się wzorców językowych, gramatyki, słownictwa i kontekstów, co pozwala mi na prowadzenie dialogów, odpowiadanie na pytania i tworzenie treści w sposób zrozumiały i naturalny.

Ważne jest, aby dane były zróżnicowane pod względem tematu, stylu i jakości, co pozwoli na trenowanie modelu językowego w sposób kompleksowy i efektywny.

Tworzenie polskiej AI

Jakich danych realnie **potrzebujemy?**



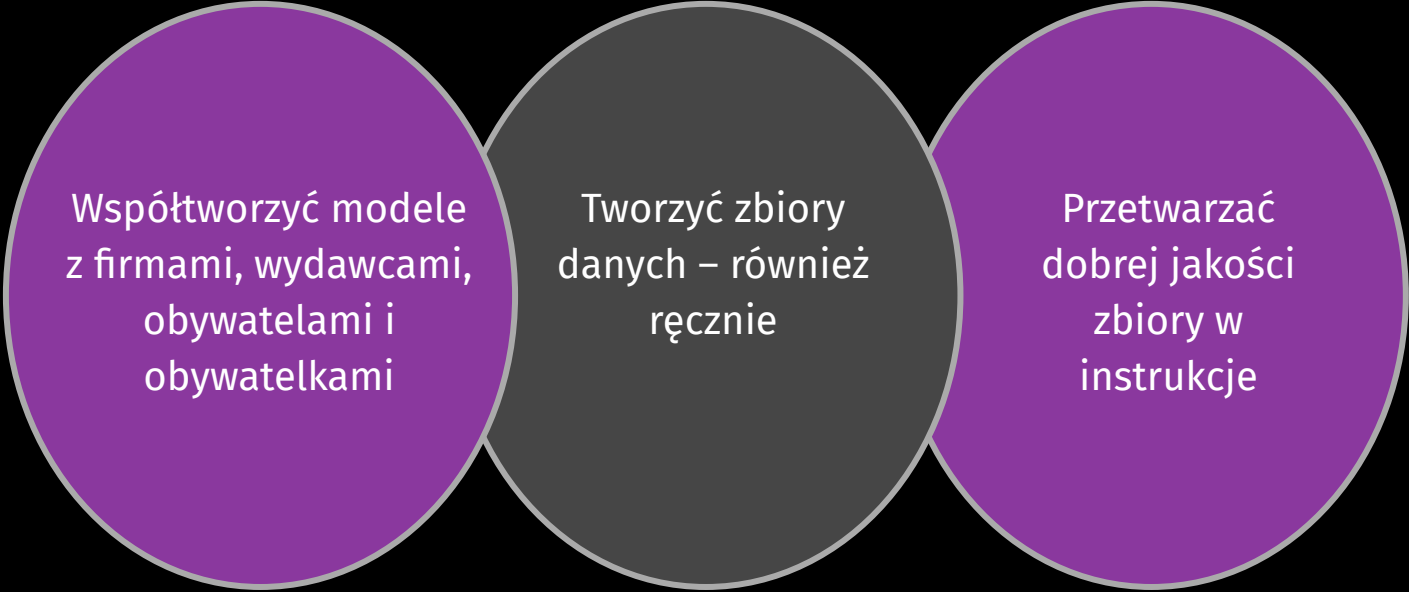
Zróżnicowanych dużych
zbiorów tekstowych,
multimodalnych

Mniejszych zbiorów
dziedzinowych,
zadaniowych,
optymalizacyjnych

Danych
kulturowych:
polskie realia,
język

Tworzenie polskiej AI

Co możemy zrobić jako **twórcy i twórczynie modeli?**



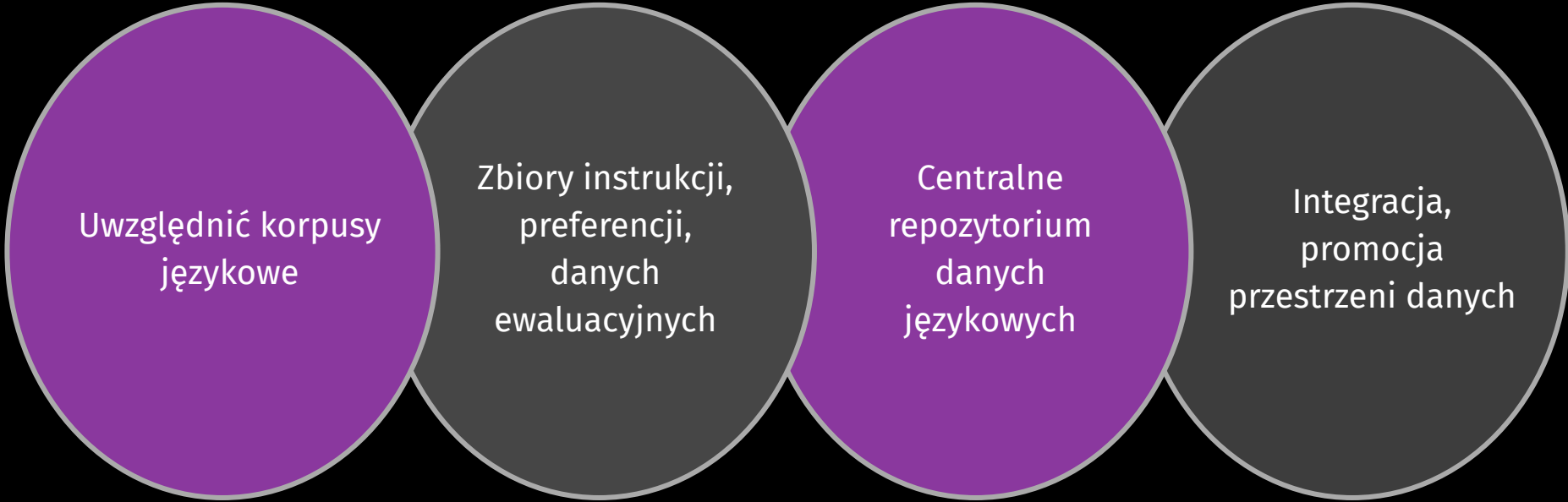
Współtworzyć modele
z firmami, wydawcami,
obywatelami i
obywatelkami

Tworzyć zbiory
danych – również
ręcznie

Przetwarzać
dobrej jakości
zbiory w
instrukcje

Tworzenie polskiej AI

Dane publiczne a pomoc w tworzeniu modeli?



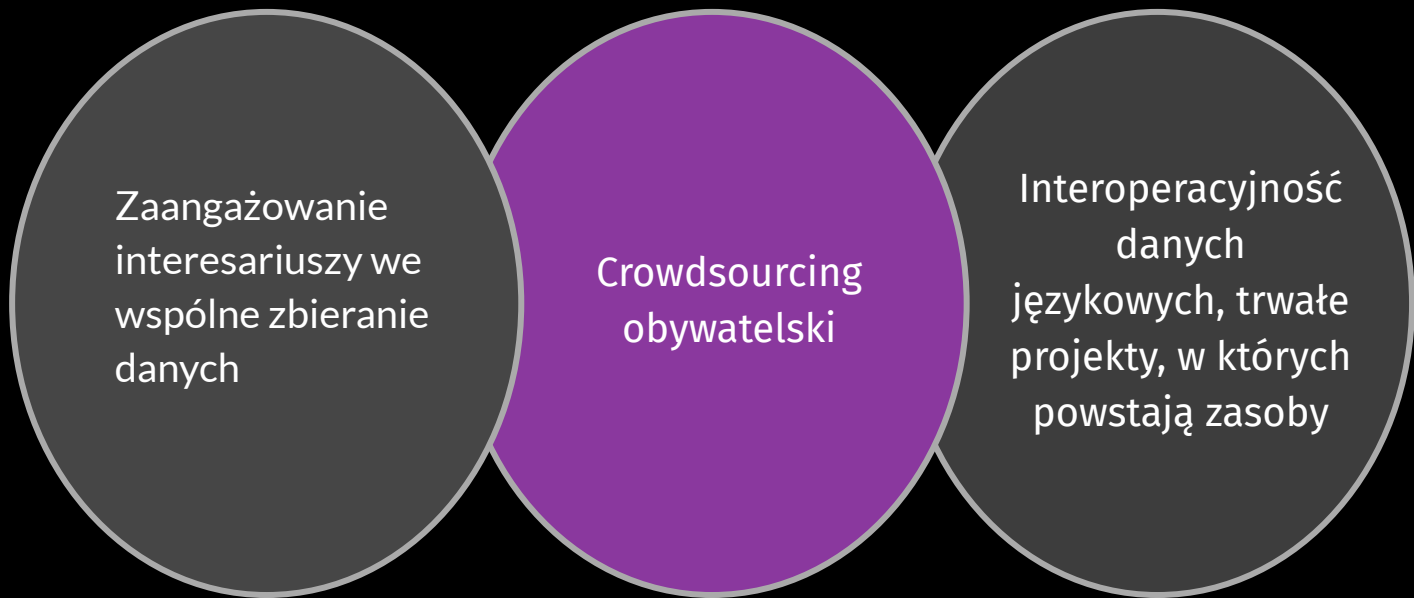
Uwzględnić korpusy
językowe

Zbiory instrukcji,
preferencji,
danych
ewaluacyjnych

Centralne
repozytorium
danych
językowych

Integracja,
promocja
przestrzeni danych

Zaangażowanie



Obywatelskie dane

- Zbierane, tworzone i/lub udostępniane przez obywateli i obywatelki poza tradycyjnymi strukturami instytucjonalnymi lub biznesowymi.
- **Powstają zwykle w ramach inicjatyw społecznych, często w modelu *crowdsourcingu*.**
- Gromadzone w celu ich skatalogowania oraz udostępnienia, zwłaszcza na potrzeby działań społecznych, badań naukowych czy rozwijania AI w Polsce.

Grupa Robocza ds. Sztucznej Inteligencji przy Ministerstwie Cyfryzacji – Dane dla AI, ***Obywatelskie zbiory danych***
Sebastian Kondracki, Aleksandra Tomaszewska, Agnieszka Kosiak, Katarzyna Starostawska, Marcin Walas i in.

Korpusy instytucji badawczych

Query

[lemma="(Poznań)"]

QUERY BUILDER

METADATA ▾

RESULT GROUPING ▾

Results per page

10 ▾

Search

1000 results found. Full list of results is available for logged in users.

No.	Left context	KWIC	Right context	Date	
1	epizodu z dziejów związków kulturalnych Polski z Półwyspem Iberyjskim w	Poznaniu [Poznań:subst:sg:loc:m3]	znalazł się reprezentatywny zbiór zabytków portugalskiego malarstwa szesnastowiecznego. 5	2019	
2	wystawieniem spektaklu Golgota Picnic na Festiwalu „Malta” w	Poznaniu [Poznań:subst:sg:loc:m3]	, potem po premierze Kłątwy w Teatrze Powszechnym w Warszawie	2020	
3	przyniósł krach modelu przyspieszonej stalinowskiej industrializacji. Bunt robotników w	Poznaniu [Poznań:subst:sg:loc:m3]	, potem rewolta na Węgrzech pokazała granice wytrzymałości. Istnieją	2013	
4	Kornhauser, Wojewódzka Biblioteka Publiczna, Ośrodek Animacji Kultury,	Poznań [Poznań:subst:sg:nom:m3]	2016, s. 227–232; Wielkopolska Biblioteka	2019	
5	. Były wiceminister zdrowia Andrzej Włodarczyk odpowiedział mi specjalistów z	Poznania [Poznań:subst:sg:gen:m3]	. Nie skorzystałam jednak z ich pomocy, bo	2013	
6	Soyoung Yoon, zwyciężczyni konkursu skrzypcowego im. Wieniawskiego w	Poznaniu [Poznań:subst:sg:loc:m3]	. Jej śliczna główka niemal kopiowała ułożenie głowy zamordowanego libijskiego	2020	

NARODOWY KORPUS JĘZYKA POLSKIEGO

Korpus

NKJP 300M

Zapytanie

KONSTRUKTOR ZAPYTAŃ

METADANE ▾

Liczba wyników na stronę

10

Wyszukaj



	☐ Details	Left context	KWIC	Right context	
1	☐ ⓘ KlaudiaJachira ...	' sprawie praworządności? Unię? Nie, samych siebie oszukujecie. Każda	dyktatura	ma swój kres. Co po was zostanie? Garstka ustawionych partyjnych olig	
2	☐ ⓘ JerzyWciśła • 2...	osiadaniu wolności wyboru jest alternatywą dla każdego autorytaryzmu i	dyktatury	. Po epoce rozbiorów, w tym po tym niemiecko-rosyjskim zrealizowanym	
3	☐ ⓘ KrzysztofMróz •...	/ senatorem, tylko może by siedział w więzieniu, gdyby tu w Polsce jakaś	dyktatura	była... Jednakże... Ależ bardzo proszę, Panie Senatorze. Jednakże jest	
4	☐ ⓘ KrzysztofMróz •...	było tak źle z polską praworządnością i gdyby w Polsce była jakakolwiek	dyktatura	, to nie byłoby opozycji albo opozycja siedziałaby w więzieniu, a opozycja	
5	☐ ⓘ KlaudiaJachira ...	okresy wolności były przerywane przez małych ludzi, którym marzyła się	dyktatura	. Prezydent Duda przypisał sobie wysiłek całego narodu polskiego w pom	
6	☐ ⓘ KrzysztofKwiatk...	rosji przeciw Ukrainie, to jest starcie cywilizacji wolnego świata z okrutną	dyktaturą	". Dokładnie tak. Jesteśmy na rubieżach tego świata, dla którego samost	
7	☐ ⓘ KlaudiaJachira ...	zdrową, zbilansowaną dietę i dobry, spokojny sen (Dzwonek) w państwie	dyktatury	. Niestety to tak naprawdę dał nam pseudo-Trybunał Konstytucyjny. Dzięk	
8	☐ ⓘ PawełZalewski •...	wa przeciwko złodziejstwu i korupcji. To jest wojna demokracji przeciwko	dyktaturze	. To jest wojna wolności osobistych przeciw zniewoleniu. To jest wreszcie	
9	☐ ⓘ PawełZalewski •...	' o wolność Europy. Ale co to znaczy w praktyce? Ukraina broni się przed	dyktaturą	i broni demokracji. To jest walka kremlowskiego złodziejstwa i korupcji z r.	
10	☐ ⓘ MarcinWarchoła •...	o różnego rodzaju kruczków prawnych. Koniec z naciągactwem, koniec z	dyktaturą	drobnego druczku. Drodzy Państwo! Odpowiadam na pytanie, co się wyd	
11	☐ ⓘ TomaszOlichwer ...	ver, Koalicja Obywatelska. Bardzo proszę. Pani Marszałek! Wysoka Izbo!	Dyktatura	, cynizm i powrót do PRL-u - to wizja oświaty według ministra Czarnka. C	
12	☐ ⓘ DariuszPiontkow...	Ciągle mówią o indoktrynacji, która była w PRL-u, o braku demokracji, o	dyktaturze	. Rzeczywiście w czasach PRL-u tak było. Nauczyciel nie mógł uczyć o K	
13	☐ ⓘ KrzysztofKwiatk...	dzie, to musimy szukać rozwiązań drastycznych. Brytyjczycy pokazali, że	dyktatura	brukselskiej biurokracji"... itd., „odwrócili się i wyszli. My (...) nie możemy	
14	☐ ⓘ KazimierzMichał...	inna być relatywizowana i powinna być w centrum naszych myśli i uczuć.	Dyktatura	komunistyczna w Polsce była czymś absolutnie wyjątkowym, wyjątkowo z	
15	☐ ⓘ JanMariaJackows...	, mamy teraz tendencję do tego, żeby porównywać wydarzenia z czasów	dyktatury	komunistycznej z wydarzeniami współczesnymi, żeby szukać jakiejś anal	
16	☐ ⓘ JanMariaJackows...	ch oraz pokazania bezmiaru kontrolowania sytuacji przez komunistyczną	dyktaturę	. I też w tym celu ta tragedia była wykorzystywana. Przypomnę, że pierws	



POTENTIAL.

POTENTIAL EVERYWHERE.

Dziękuję Państwu za uwagę!

Aleksandra Tomaszewska
Instytut Podstaw Informatyki PAN
aleksandra.tomaszewska@ipipan.waw.pl

Kontakt

