LLMs as Research Tools in Social Sciences & Humanities

Maciej Piasecki
CLARIN-PL
& Wrocław University of Science and Technology

CLARIN
Politechnika Wrocławska
Fundusze Europejskie
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Journey plan

- Background: CLARIN & CLARIN-PL – a language technology infrastructure
  - Is there life for LT RI beyond the ChatGPT launch?
- Limited performance of LLMs in different tasks – the quality of the research process
- Very LLMs — black boxes full of surprises
- Evaluation of LLMs and all we are not aware of
- PLLuM – an exercise in building an open and transparent LLM
- LLMs in data annotation
- LLMs as elements of the language technology research infrastructure
- LLMs as tools in Social Sciences and Humanities
- A journey to follow

Language Technology

- Language resources:
  - datasets and databases describing Natural Language and its use
  - formalized description of the selected aspects of natural language,
  - datasets for training and evaluation
- Language tools:
  - computer programs for text and speech processing at different levels of natural language analysis
    - automatic analysis of language structures, common tasks, e.g. classification of proper name occurrences,
  - recently language models
- Language technology = Resources + Tools + Infrastructure.
- Language infrastructure
  - a common technological base to combine diverse language tools and resources

CLARIN – Large, Distributed Research Infrastructure

- CLARIN =
  - Common Language Resources and Technology Infrastructure

Distributed network of more than 60 centres:

- 25 certified technological centres (CTS), strong on FAIR & interoperability
- federated login, across Europe and beyond
- central harvesting of metadata to facilitate finding language resources and tools
- connected and combined services
- language data - text, speech, video or multimodal
- advanced tools - for finding, exploring, using, annotating, analysing or combining datasets, *independently of their physical location*
- **Member of EOSC** (*European Open Science Cloud*, od 2020)

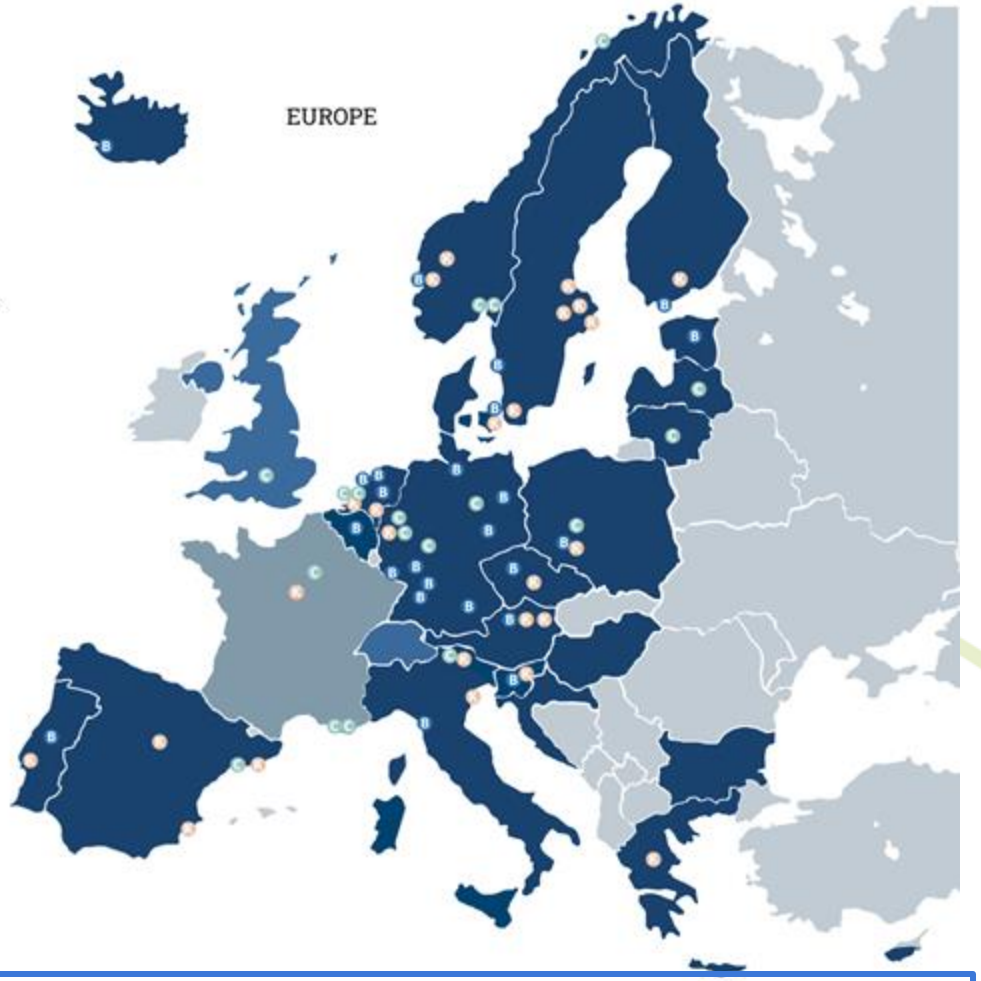
CLARIN ERIC – Europa (clarin.eu)

25 members:

Austria
Bulgaria
Croatia
Cyprus
Czechia
Danmark
Estonia
Finland
Greece
Spain
The Netherlands
Island
Lithuania
Latvia
Germany
Norway

Poland
Portugal
Slovene
Slovakia
Switzerland
Sweden
Hungary
Italy

2 observers:
Great Britain
Republic of South Africa



M. Branco et al. The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. Language Resources and Evaluation, June 2023, https://link.springer.com/article/10.1007/s10579-023-09658-z

CLARIN Polska: CLARIN-PL (od 2013)

CENTRUM TECHNOLOGII JĘZYKOWYCH CLARIN-PL

CLARIN B CENTRE

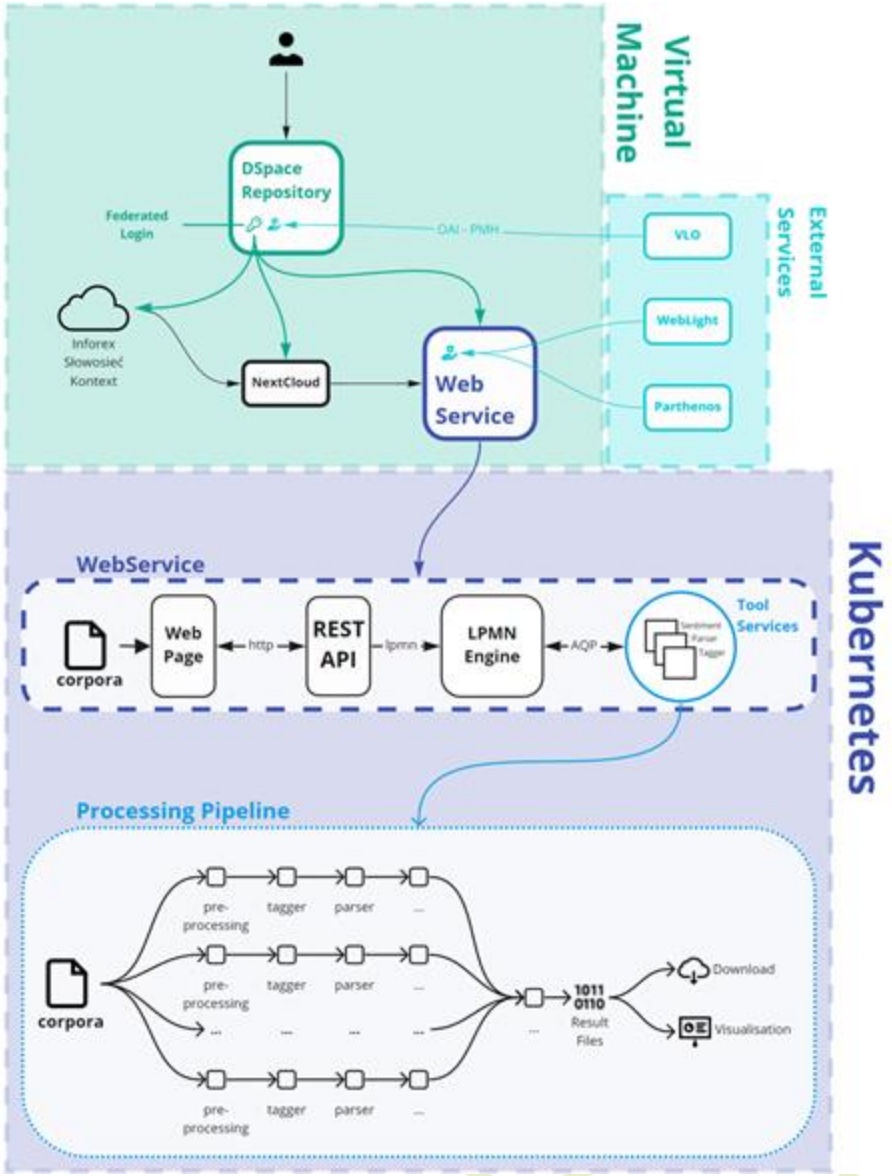
CORE TRUST SEAL

CLARIN CENTRE K

- CLARIN-PL Language Technology Centre (http://clarin-pl.eu)
  - language data repository językowych, (https://clarin-pl.eu/dspace/)
  - space and workbench for researchers: (https://services.clarin-pl.eu)
    - data magazine – private working cloud
    - services and applications for text and speech analysis
    - task lists – monitoring and history of processing
  - Language resources for Polish and several other languages.
- PolLinguaTec - Knowledge Centre for Language Technology for Polish http://kcentre.clarin-pl.eu/
- **LLMs4SSH**: CLARIN K-centre for Large Language Models in SS&H (https://llms4ssh.clarin-pl.eu/ i https://llms4ssh.clarin.eu/)

M. Piasecki et al. CLARIN-PL: a user centred language technology infrastructure. Language Resources and Evaluation, May 2025, https://link.springer.com/article/10.1007/s10579-025-09839-y

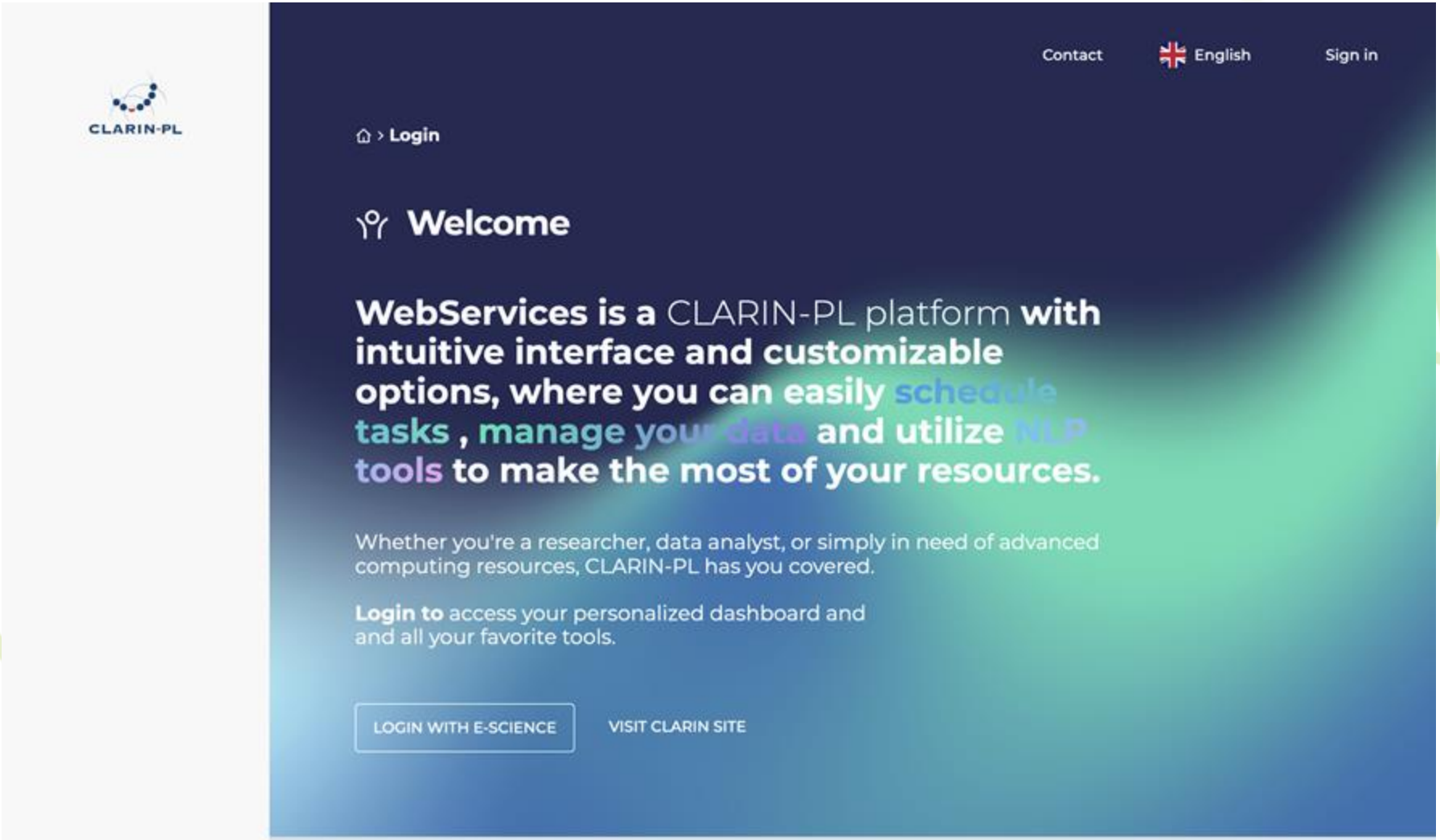
CLARIN-PL Technological Basis



https://services.clarin-pl.eu

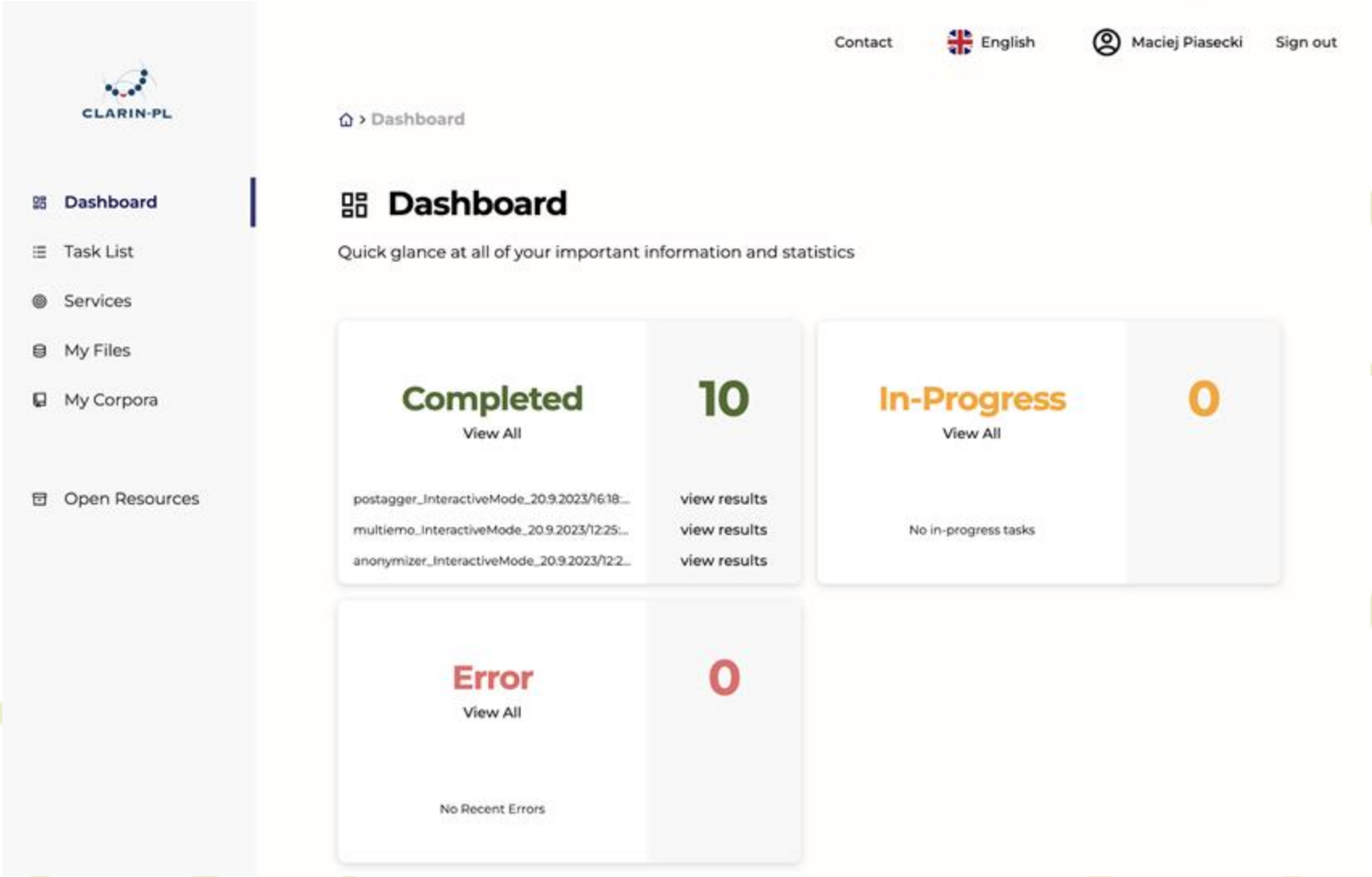
- Supercomputer LEM
  - 304 GPUs NVIDIA H100
  - 96 GB fast RAM per node
- Robust
- Efficient
- Flexible orchestration of NLP processing pipelines
- Parallel processing
- Dynamic scaling
- ~100 NLP services offered 7/24
  - many of based on LLMs
- Millions requests served per year
- Many LLMs trained or fine-tuned

Access to CLARIN-PL-Biz LT R&D Infrastructure (https://services.clarin-pl.eu)



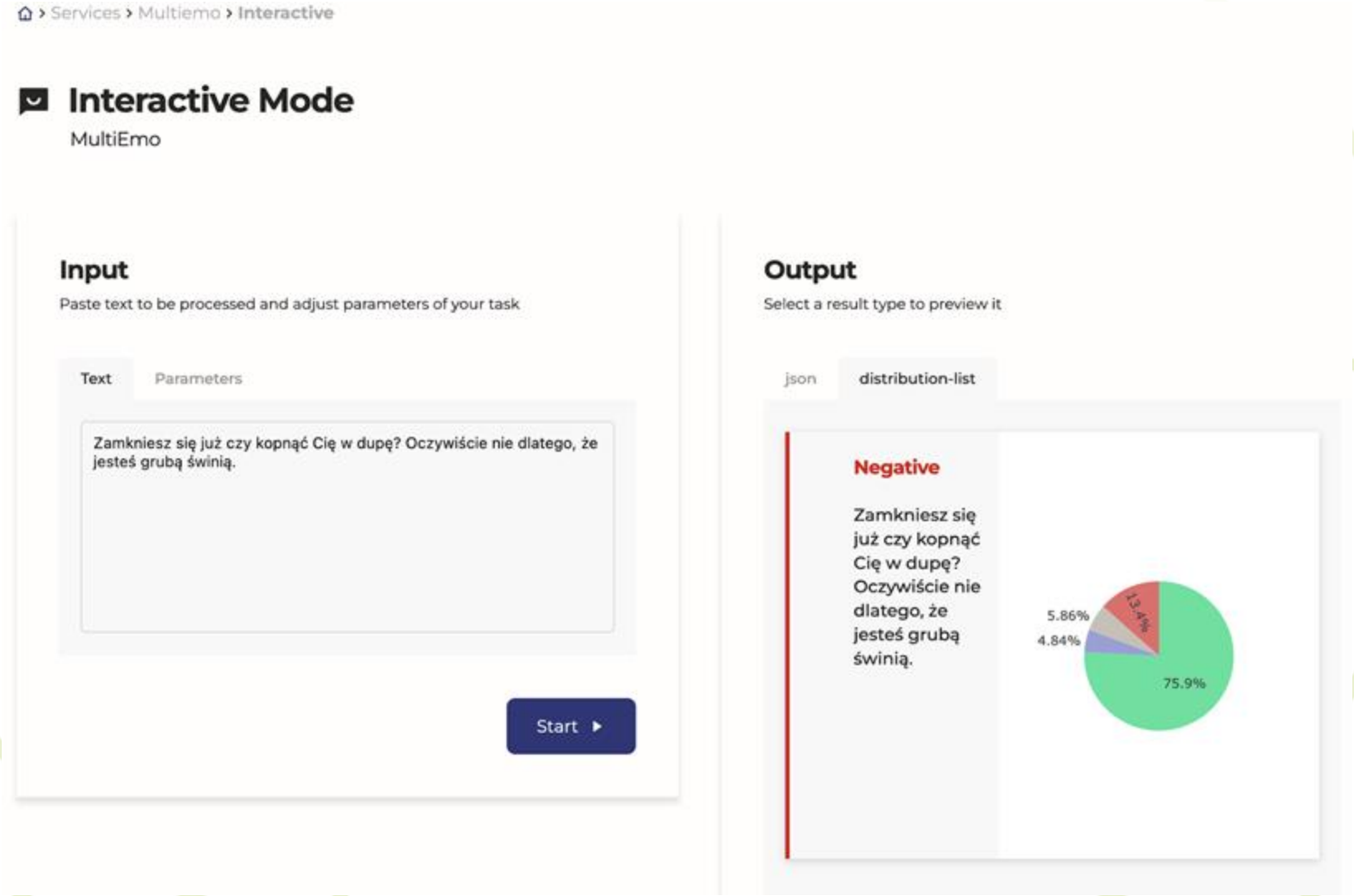
CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu



CLARIN-PL (Biz) via Federated Login

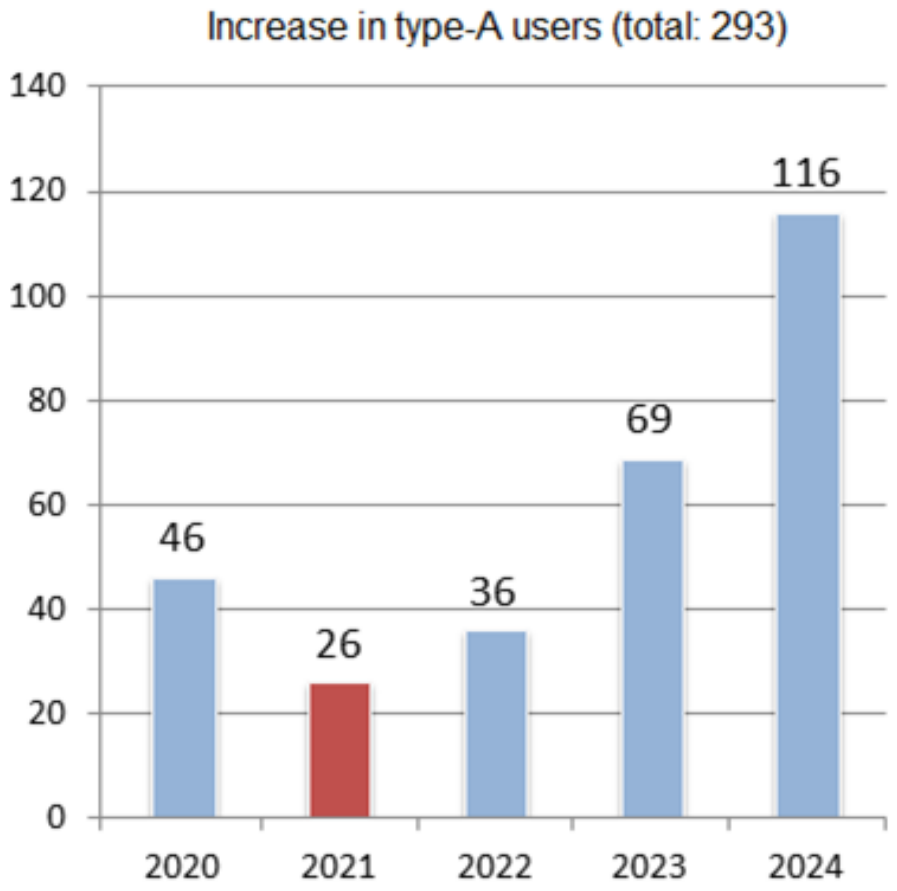
https://services.clarin-pl.eu

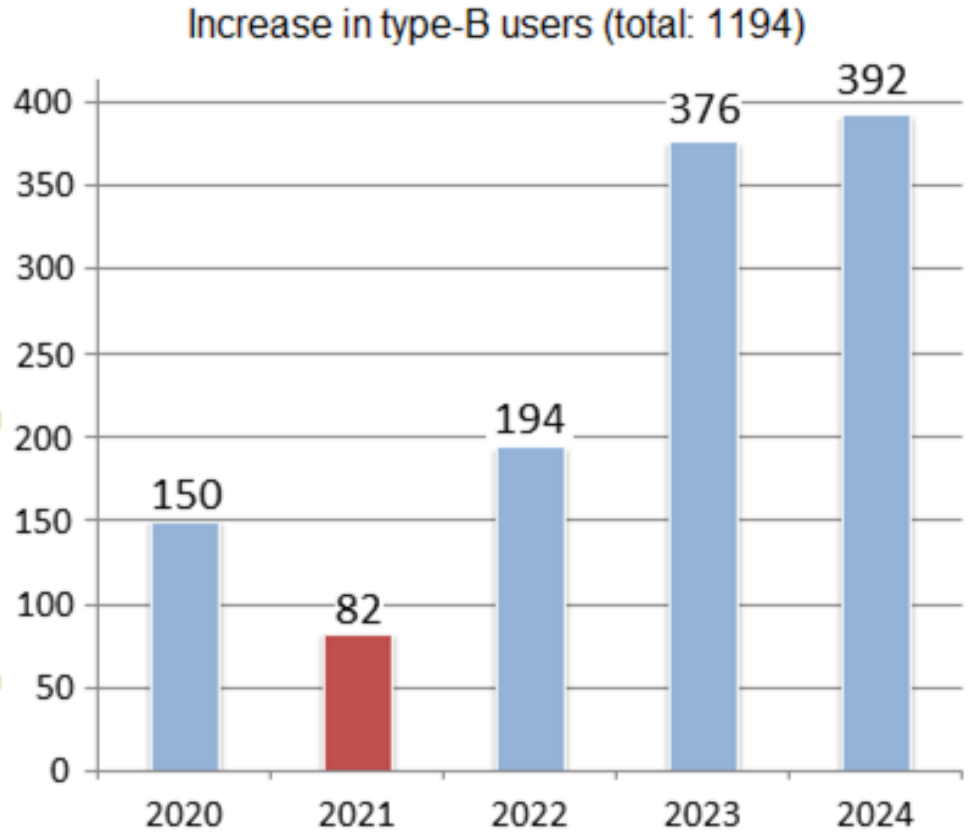


CLARIN-PL Users

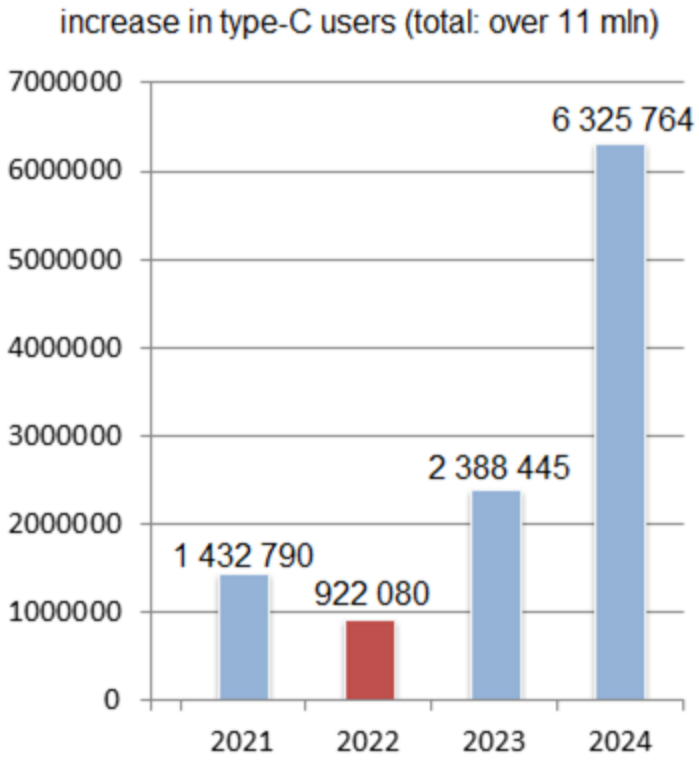
- Main focus on researchers, especially from the broadly understood SSH
- CLARIN-PL is also open on:
  - education,
  - public institutions and GLAM,
  - but also private companies, considering the applications of open LT
- Three types of users – for the needs of monitoring:
  - **A** – direct cooperating users
  - **B** – spontaneous users
  - **C** – statistical users

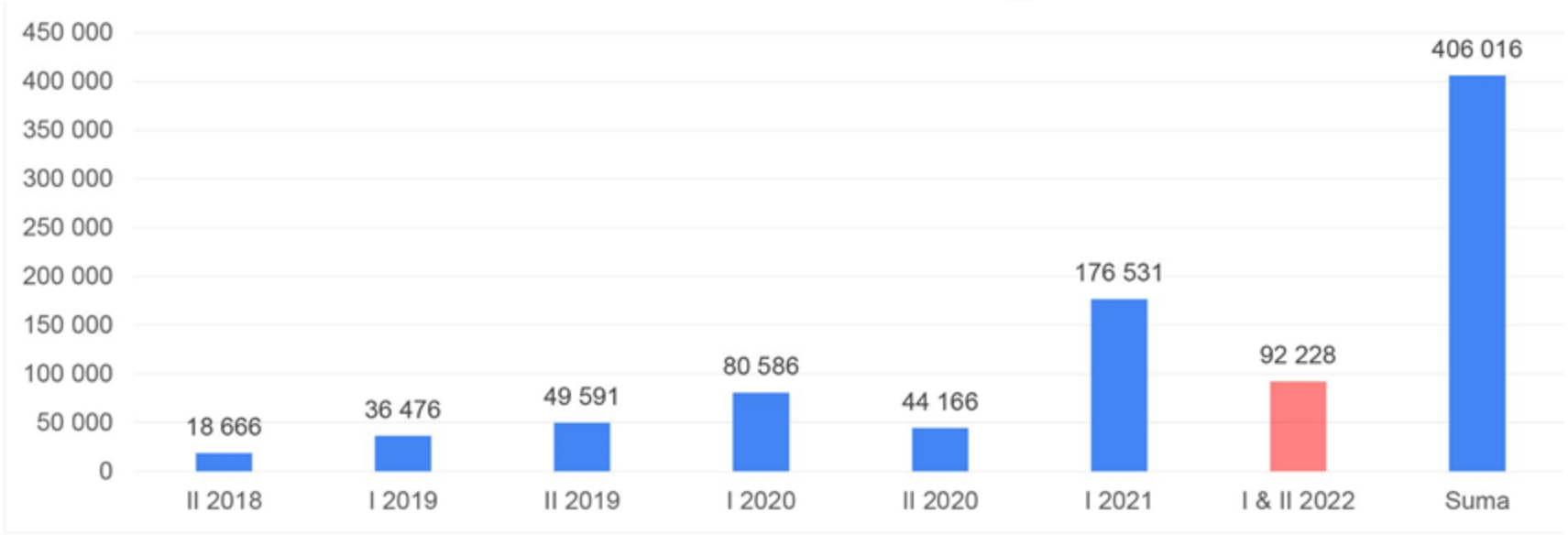
CLARIN-PL: type-A and type-B users





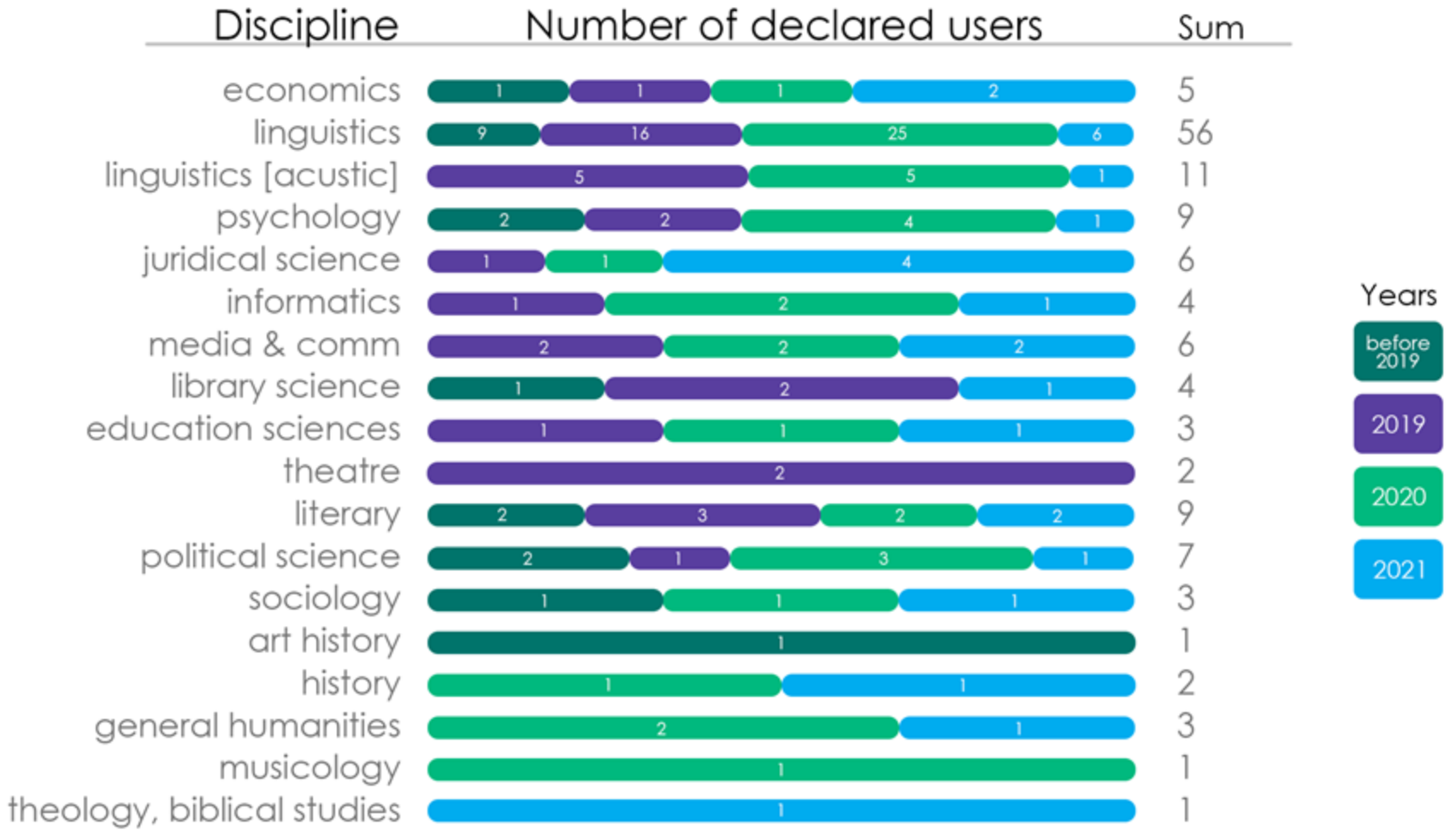
CLARIN-PL: type-C users





- Data processed in QV – Quo Vadis
  - 1 QV = the number of words in the book Quo Vadis of the Nobel Prize winner, Henryk Sienkiewicz

CLARIN-PL – User vs Scientific Disciplines



CLARIN-PL Example Use Cases

- Language of suicide notes and its computational analysis, e.g. features of the genuine vs fake ones [linguistic and forensic studies]
- Romantic irony in the poems of Słowacki (the great Polish poet of the 19th century): to what extent can language tools reveal the objective features? [literary studies]
- Study on trauma and its picture in personal language [psychology]
- Analysis of the picture of Ukrainians in the Polish social media and the temporal evolution of the concept [sociology]
- Development of Słowosieć (plWordNet) — the largest dictionary of Polish in history [lexicography]
  - supported by the whole lexicographic workbench
- Analysis of the multicultural dialog in the city council of Kraków [cultural studies]
- Analysis of the public opinion in the Polish internet - on regular basis by The Chancellery of the Prime Minister Republic of Poland [public institution]

CLARIN-PL as a LT Technological Infrastructure

- Technological infrastructure
  - "facilities, equipment, capabilities and support services required to develop, test and upscale technology to advance from validation in a laboratory up to higher Technology Readiness Levels prior to competitive market entry. They can have public, semi-public or private status. Their users are mainly industrial players, including SMEs, which seek support to develop and integrate innovative technologies towards commercialisation of new products, processes and services, whilst ensuring feasibility and regulatory compliance."
    (European Commission Staff Working Document, SWD 2019/158)
- CLARIN-PL-Biz Projects (2020-till now) – expansion to Research & Technological Infrastructure:
  - cooperation with more than 30 companies as a basis for both projects
  - research – open services, LT and support (e.g. knowledge centres)
  - public and local government institutions – advice and prototyping, open to the extent of available resources,
  - business and industry – cooperation within the implementation of CLARIN-PL-Biz prototypes for use cases (Proof of Concept), joint projects, applications of CLARIN-PL services

Is there life for LT RI beyond ChatGPT?

- LLM-free zones in LT RI
  - corpus development, collecting texts, preprocessing, annotation and data team management
  - text preprocessing, e.g. lemmatisation, morphosyntactic tagging
  - statistical analysis of language resources,
    - e.g. searching, extraction of collocations, comparison of corpora, topic modelling, most of the stylometry techniques, terminology and lexica extraction
- What can be replaced with LLMs and what not?
  - What is worth to be replaced, e.g. efficiency but also performance?
- An LLM is a flexible and advanced language tool to build upon, a Swiss knife for everything … not exactly

Limited accuracy of LLMs in different tasks

- The quality of a scientific tool impacts
  - the quality of the obtained results, e.g. interpretation, the level of confidence, validity of the claims etc.
  - the quality of the research process
  - bias towards data to be processed and classifier drift
- “Jack of All Trades” (Kocoń et al, January 2023)
  - a comprehensive evaluation of ChatGPT performed in the first months after its launch
  - a reference point for the scientific LLM evaluation

“ChatGPT: Jack of all trades, master of none”

- Goal: testing ChatGPT, on various NLP tasks
  - started in early Dec. 2022, one month after its introduction
- 25 public NLP datasets
  - a large part of which involved subjective problems, e.g. predicting emotions or offensiveness
  - 3 completely new, unpublished in the moment of tests
- Is ChatGPT loss in performance compared to SOTA different for individual tasks of different kinds?
- Is there a difference in ChatGPT’s ability to solve difficult and easy NLP analytical tasks?
- What is the impact of the context while processing multiple questions (prompts) that may or may not be related to each other?
- Is GPT-4 better or worse compared to ChatGPT?

Jan Kocoń et al. ChatGPT: Jack of all trades, master of none. Information Fusion, Vol. 99, Nov. 2023
https://www.sciencedirect.com/science/article/pii/S156625352300177X

“ChatGPT: Jack of all trades, master of none”

- Investigation of its analytical skills of ChatGPT
- Two abilities targeted: semantic and pragmatic
- Tasks:
  - (1) **binary classification** of texts like spam, humour, sarcasm, aggression detection, or grammatical correctness of the text;
  - (2) a more **complex multiclass and multi-label classification** e.g. sentiment analysis, emotion recognition;
  - (3) **reasoning with the personal context**, i.e., personalized versions of the problems that make use of additional information about text perception of a given user (user’s examples provided to ChatGPT);
  - (4) semantic annotation and acceptance of the text, towards **natural language understanding** (NLU), e.g. word sense disambiguation (WSD)
  - (5) **answering questions** based on the input text.
- Languages: 23 tasks (92%) in English, 2 (8%) in Polish

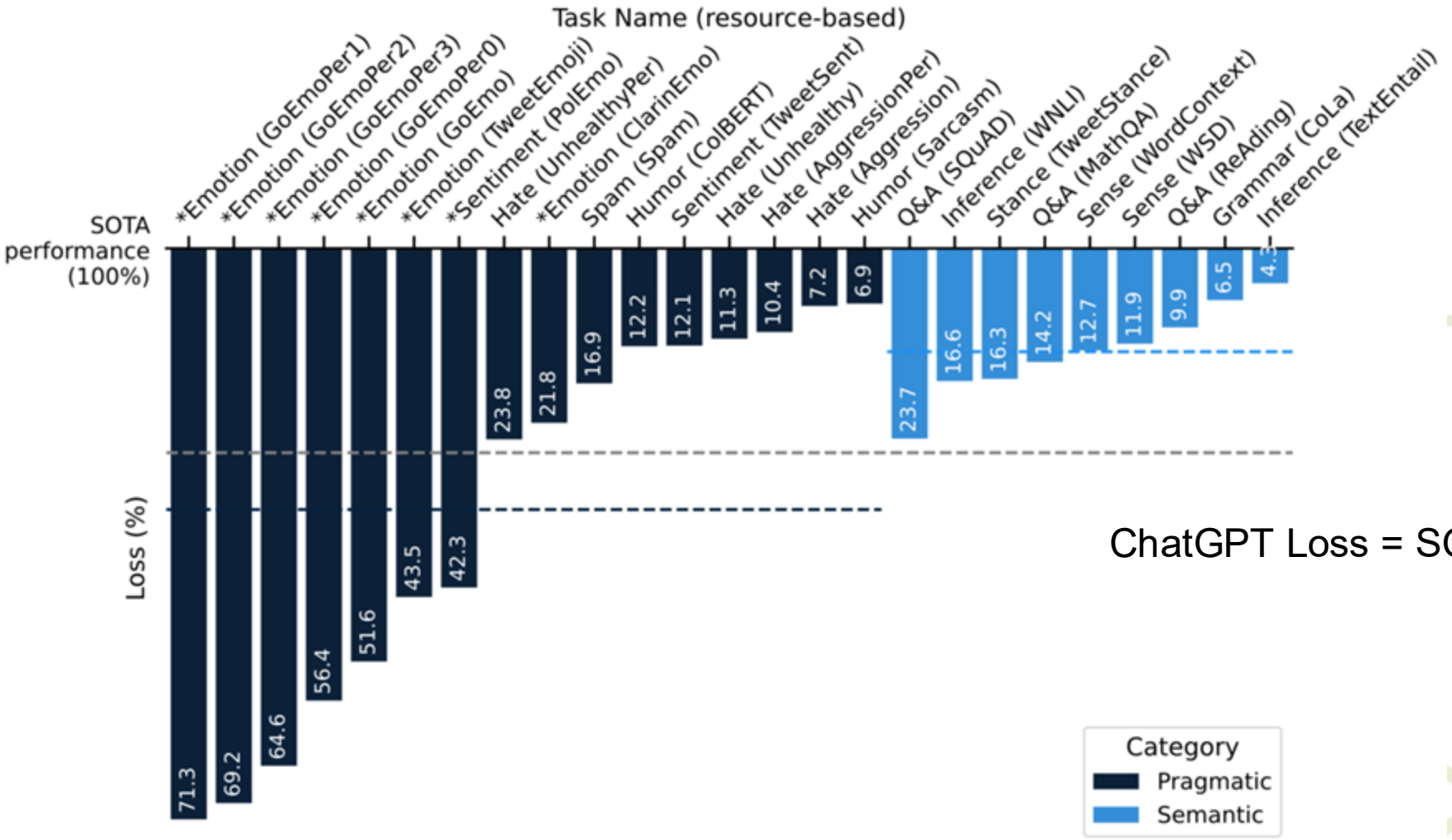
“ChatGPT: Jack of all trades, master of none”
Test prompt pattern for aggression

Chat `CHAT_ID`**. Task:** `TASK_NAME`**. Case** `EXAMPLE_ID`**. E.g.: Chat 1. Task: Aggression. Case 3.**
Prompt `//our input to ChatGPT`
`INSTRUCTION` *//task description, e.g.:* Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute. `TEXT` *//input text, e.g.:* Text: *(Or should I follow your example and delete things I don’t like from other people’s talk pages ?)*
ChatGPT answer `//raw output`
["non-aggressive"]
Extracted answer `//processed output`
non-aggressive
Expected answer `//expected output`
non-aggressive
Evaluation result `//additional judgement`
Label: OK, ChatGPT answer: OK

“ChatGPT: Jack of all trades, master of none”
Example of a test prompt for emotions

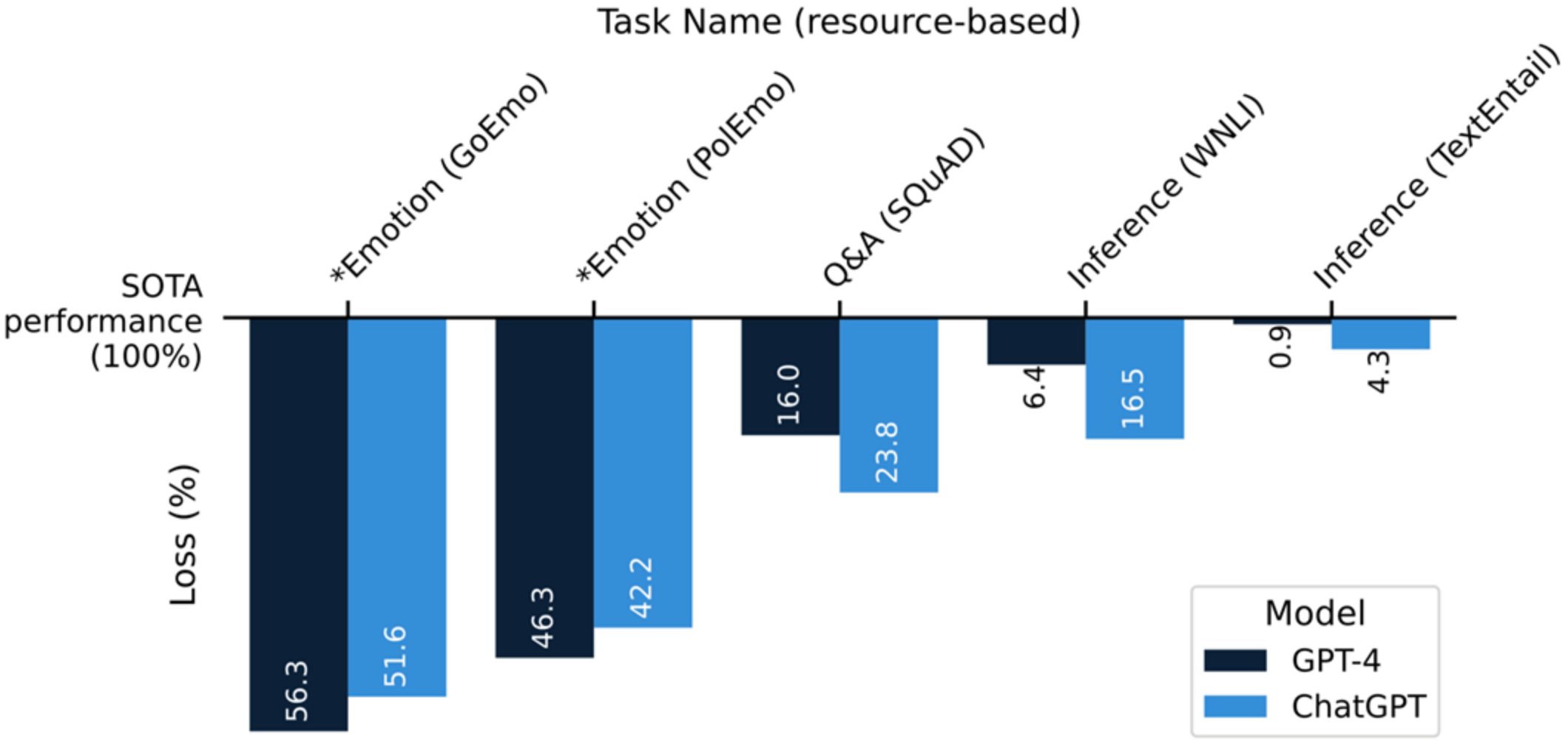
Chat 2. Task: GoEmotions. Case 98.
Prompt
From the given list of all emotions, choose the ones that the input text arouses in most people reading it. Write your answer in a Python list containing exactly two selected most matching emotions. List of all emotions: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral. Input text: *That’s how you know they’re happy.*
ChatGPT answer
Based on the input text, it is likely that the text would arouse the emotion of realization and satisfaction in most people reading it. ["realization","satisfaction"]
Extracted answer
["realization", "satisfaction"]
Expected answer
["excitement", "neutral"]

“ChatGPT: Jack of all trades, master of none”
ChatGPT loss on different tasks



ChatGPT Loss = SOTA – ChatGPT result

“ChatGPT: Jack of all trades, master of none”
Comparison: ChatGPT vs GPT 4



Very LLMs — black boxes full of surprises

- What is inside an LLM?
  - this is what it has been trained on and for
- LLM development process determines its language coverage, knowledge, skills and preferences to a very large extent

LLM Development Process

Data

- Acquisition
- Validation
- Selection

Tokenisation

Pretraining

autoregressive learning

Texts and speech transcription

Fine-tuning

supervised learning

Instructions: prompts + expected results

Alignment

reinforcement learning

Preferences: Of users in relation to the results generated by the model

Openness of Very LLMs

- Raw text data
- Filtered, selected and deduplicated data
- Instructions for fine-tuning
- Preferences for alignment
- Open-weights models
- LLMs as software modules
- APIs (web services) to access LLMs
- Software for development

Limitations of LLMs as Research Tools

- ChatGPT is an “accessible tool with a simple user interface.” (Mervaala & Kousa, 2024)
- Hidden biases – Very LLMs as black boxes
  - unknown map of the training data sets of instructions
  - do we know what we see or obtain?
  - especially important in the case of zero-shot and a few-shot applications
- Reproducibility of the results
  - LLM versions and LLM persistency, hidden settings
- Ratio:
  - performance vs computational cost
  - “massive size” of LLMs — limited access to hardware
- Privacy of data and ethical issues
- LLMs are intrinsically generative (stochastic distribution)
  - prompts, session, text chunking, context, …
  - the generation is not the end of the story, the results need further interpretation
  - ‚surprises’ in the output (e.g. format) are indeterministic and inevitable

E. Mervaala & I. Kousa. Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses. NLP4DH Workshop, Miami, ACL, 2024 https://aclanthology.org/2024.nlp4dh-1.51/

Evaluation of LLMs

- Unknown training data —> unknown status of the commonly used benchmarks
- Aspects
  - performance, efficiency, bias, safety, reliability, technological compatibility, openness, …
- Tasks
  - limited answer: results of a specified type are expected,
    - e.g. selection, extraction
    - symbolic pattern matching or deterministic processing is enough
  - generative tasks: high variety of proper results
    - e.g. summarisation, language simplification, text generation
    - analysis, processing and some understanding of the output is necessary
- Settings
  - prompt based vs contextual
    - e.g. RAG evaluation often assumes a knowledge base to be used
  - output format
    - e.g. presentation of the correct answer
  - zero-shot vs a few shot prompts
    - not forgetting about prompt wording, a single word can change the results

Evaluation of LLMs: Techniques

- Perplexity (misleading) for generative tasks
- Log likelihood based
- Pattern matching and symbolic postprocessing
- Similarity: result vs gold standard result (human)
- Manual analysis (free, guided, aspectual, arena-based)
  - not straightforward (sic!), evaluators may have shallow attention and be biased
- LLM-as-a-judge (different variants, including aspectual, LLM-powered arena, ensemble of judges, etc.)

Evaluation of LLMs: Datasets

- Most of the commonly LLM benchmarks are not reliable!
  - publicly available, based on known resources, limited, following limited schemas, synthesised (‚silver standards') etc.
  - LLM evaluation is a marketing technique for the developers!
- Challenges:
  - fresh original input
  - precisely targeting LLM skills
  - consistency
- Especially difficult: generative tasks

PLLuM
Polish Large Language Model

CLARIN-PL
Common Language Resources and Technology Infrastructure

PLLuM – the Polish Large Language Model as a versatile NLP engine for the research & technological infrastructure

J. Kocoń, M. Piasecki, et al. PLLuM: A Family of Polish Large Language Models, 2025, arXiv: https://arxiv.org/abs/2511.03823

The PLLuM Project

1. Development of the PLLuM family of models (leader WrocTech) and a prototype of Citizen Assistant (22.01–31.12 2024, budget 14,5 PLN, ~4M$, Min. of Digital Affairs, 2024):
    1. **open-source Polish LLM** complaint with the paradigm of the responsible development of AI systems
    2. development of **prototype application of this model in public administration** in a form of a Polish intelligent assistant

2. Project HIVE (leader NASK)
    1. “HIVE AI: Development and pilot applications of LLMs in the Polish public administration” (Min. of Digital Affairs, 2025)
    2. Expansion of language resources, development of language models, evaluation system, pilot applications

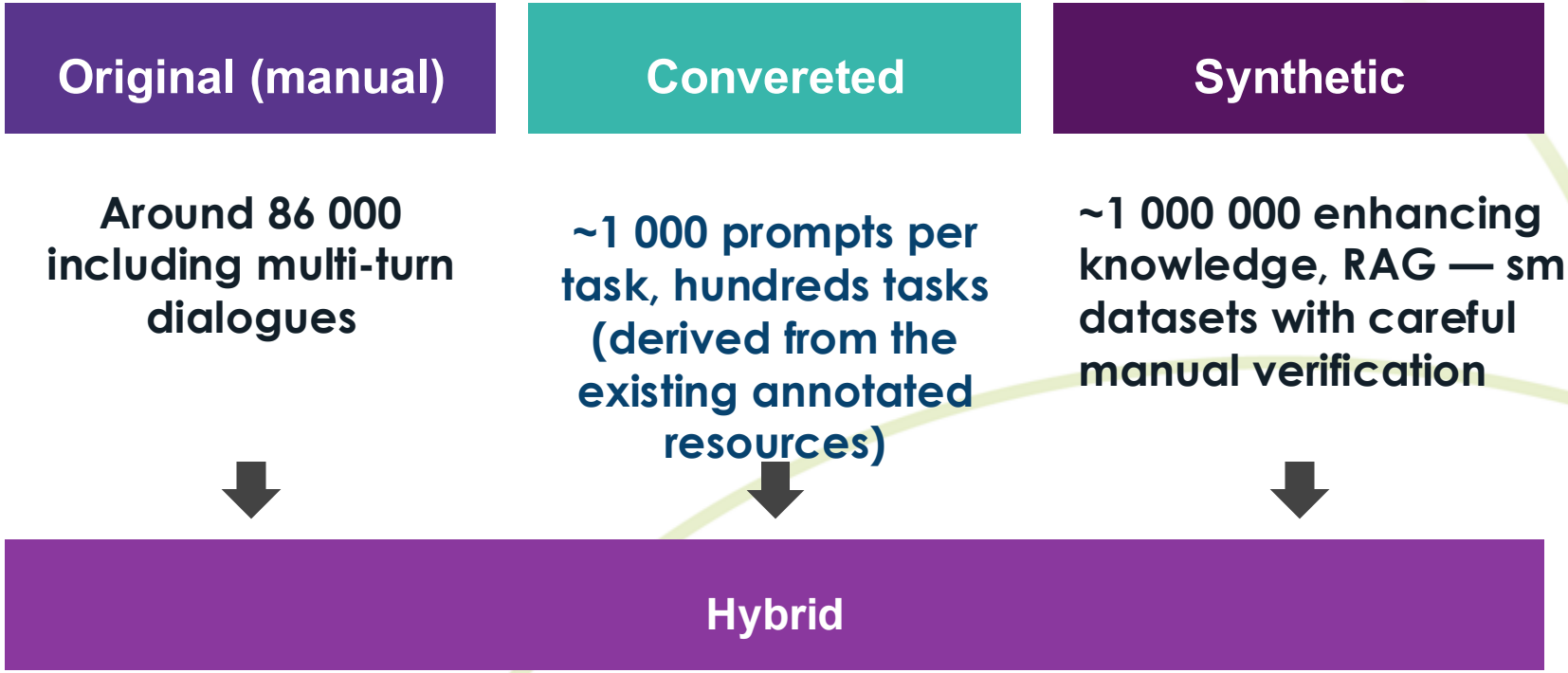
Politechnika Wrocławska
OPIPIB
NASK
PAN
UNIWERSYTET ŁÓDZKI
SLAV
COI Centralny Ośrodek Informatyki

Why should we build PLLuM?

- An open-source Polish LLM developed in accordance with the principles of responsible AI system development — transparent and open
- Gaining practical expert knowledge in the field of building LLMs — a step towards technological sovereignty
- Building large, unique training datasets for the Polish language and Polish socio-cultural context
- Construction of a fully controlled model from the research point of view
- Practical verification of the idea of building LLMs in accordance with legal regulations and the European AI Act
- Development of a prototype application of this model for public administration an intelligent Citizen Assistant for *mObywatel* mobile app

  (~ *mobile citizen* — all documents, most official matters online)

PLLuM: Instructions for Fine-tuning



Original (manual), **organic**, expanded with LLMs (appropriate licences!) to larger datasets.

All carefully manually verified and corrected, e.g. multi-turn dialogues with existing models (a form of Active Learning).

PLLuM Instruction Corpus — Composition

- **Organic** hand-crafted, high-quality organic instructions curated by a team of trained annotators
- Instructions **distilled** from existing LLMs (on appropriate licences)
  1. Knowledge distillation, RAG and Context-injected NLP tasks
- **Converted** from annotated corpora, databases and text repositories
- Types: prompt-response, dialogue
  1. dialogue, e.g. role-playing,context sensitivity, and multi-turn prompting

Category	Proportion
Knowledge (QA)	43%
Generation	25%
Extraction	6%
Programming	6%
Conversational	4%
NLP	3%
Adversarial	3%
Visualisation	3%
Data manipulation	3%
Chain of Thought	2%
Translation	1%
Identity	1%

PLLuM Instruction Corpus: Application

Model	PLCC ↑
Mistral-Nemo-Instruct-2407	23.00
Mistral-Nemo-2407+PLLuMIC	22.33
PLLuM-12B-nc-instruct	56.33
PLLuM-12B-nc-chat	59.50
Mixtral-8x7B-Instruct-v0.1	35.33
Mixtral-8x7B-v0.1+PLLuMIC	32.17
PLLuM-8x7B-nc-instruct	67.17
PLLuM-8x7B-nc-chat	**68.17**
Llama-3.1-8B-Instruct	22.67
Llama-3.1-8B+PLLuMIC	24.67
Llama-PLLuM-8B-instruct	58.00
Llama-PLLuM-8B-chat	60.67
Llama-3.1-70B-Instruct	47.83
Llama-3.1-70B+PLLuMIC	38.67
Llama-PLLuM-70B-instruct	65.17
Llama-PLLuM-70B-chat	66.33
Qwen-Max	50.83
GPT-4	59.50
Grok-2-1212	66.00
DeepSeek-v3	69.17
DeepSeek-R1	76.00
O1-2024-12-17	**89.17**

- The continually pretrained models for Polish consistently outperform their base counter parts across all four architectures
- Fine-tuning on PLLuMIC is only effective for models that have undergone continual pretraining
- Models aligned with human preferences achieve slightly higher benchmark scores than their instruction-fine-tuned predecessors.
    1. longer responses may influence evaluation methods

P. Pęzik et al. The PLLuM Instruction Corpus, *to appear,* 2025 (or 2026).

PLLuM Instruction Corpus: Examples

- Anonymisation
  1. The task consists in anonymization of surnames and pseudonyms in linguistically challenging posts from social media e.g.
  2. `Prompt: In the text provided, anonymize only the surnames and nicknames, using the labels [surname] and [pseudonym] in place of the identified entities: {text}`

PLLuM Instruction Corpus: Examples

- Word Sense Disambiguation
  1. Polish Word Sense Disambiguation task, plWordNet 4.2
  2. `Prompt: Given the sentence: {text}, how would you define the following word: {word}?, Prompt: Provide the definition of the highlighted word in this text {context}, Prompt: Provide the definition of the word: {word} based on the following context of its usage: {context}, Prompt: Based on the sentence {text}, how would you describe the meaning of the following word: {word}? Prompt: Does the word {word in the text: {text} has the same meaning as in this one: {text2}, Prompt: Does the provided definition {definition} describe the word: {word} in the context of {text}? Prompt: Provide the definition of the word {word}, Prompt: What are possible definition of the word {word}?, Prompt: Provide the most common definition of the word: {word}.`

PLLuM Alignment

- ~88,000 manually annotated instructions, derived from three distinct annotation methodologies:
  1. Rating-based annotation – each response assessed according to a predefined metric
  2. Ranking-based annotation – four responses were ranked according to response quality
  3. Dialog-based annotation – annotators engaged in multi-turn interactive conversations with the model, selecting the most appropriate responses
- Manual prompts similar to tuning instructions but with a strong emphasis on safety-related prompts
- Odds Ratio Preference Optimization (ORPO) algorithm (Hong et al.,2024),
- which integrates alignment with instruction tuning through a specifically designed loss function
  1. applying ORPO after SFT yielded superior performance

K. Seweryn. PLLuM-Align: Polish Preference Dataset for Large Language Model Alignment. EMNLP, ACL, 2025. https://aclanthology.org/2025.emnlp-main.1219/

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. Preprint, arXiv:2403.07691.

PLLuM Alignment: Evaluation

- The red-teaming evaluation:
  1. 18,656 harmful prompts for attack success rate (ASR)
  2. 9,724 non-harmful samples for the false-refusal rate (FRR)
- 14 hazard categories of the Llama-Guard taxonomy (Inan et al., 2023)
- 10 different attack styles inspired by the Rainbow Teaming framework (Samvelyan et al.,2024)
- PLLuM models fine-tuned on instructions have by a relatively higher ASR and a lower FRR

Model	ASR↓	FRR ↓
Mistral-Nemo-Instruct-2407	21.85	0.62
PLLuM-12B-nc-base	72.80	10.90
PLLuM-12B-nc-instruct	77.61	0.62
PLLuM-12B-nc-chat	1.03	3.31
Mixtral-8x7B-Instruct-v0.1	31.86	0.59
PLLuM-8x7B-nc-base	74.35	6.95
PLLuM-8x7B-nc-instruct	70.63	0.56
PLLuM-8x7B-nc-chat	**0.78**	8.69
Llama-3.1-8B-Instruct	19.66	0.86
Llama-PLLuM-8B-base	72.26	16.90
Llama-PLLuM-8B-instruct	78.48	**0.13**
Llama-PLLuM-8B-chat	16.44	2.17
Llama-3.1-70B-Instruct	22.27	0.36
Llama-PLLuM-70B-base	74.91	1.93
Llama-PLLuM-70B-instruct	71.20	0.27
Llama-PLLuM-70B-chat	7.15	1.65

H. Inan, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. Preprint, arXiv:2312.06674.

M. Samvelyan. Rainbow Teaming: Open-Ended Generation of Diverse Adversarial Prompts. 2024 https://arxiv.org/abs/2402.16822

PLLuM – Products (October 2025)

- Hand-created Polish collections:
  - ~86,000 tuning instructions
  - 130,000 preferences for alignment
- Models ranging in size from 8 to 70 billion parameters.
- Proprietary collections and methods for LLM evaluation
- Analysis of implementation needs in public and self-government institutions, government offices
- Trial implementations:
  - prototype of an intelligent assistants: municipal offices, Ministry of Digital Affairs,
  - data analysis tools: CLARIN-PL, Chancellery of the Prime Minister.
- Prototype implementation in *mObywatel* (mobile app for citizens)

Models (October 2025)

PLLuM-8B

Each in 3 versions (26 in total):

1. basic
2. instrukcyjnej
3. dialogowej (po wychowaniu).

PLLuM-12B

PLLuM-12B-nc

Llama-PLLuM-70B

PLLuM-8x7B

PLLuM-8x7B-nc

Llama-PLLuM-70B-250801

PLLuM-12B-250801

PLLuM-12B-nc-250715

Models of fully open licences

Models on non-commercial licences — use restricted to the Polish research institutions

https://huggingface.co/CYFRAGOVPL

PLLuM Evaluation

- Datasets
  - Polish Language Competence Benchmark — fresh, comprehensive, more than 2200 instructions in many categories,
  - Generative tasks for Polish – 20 functional styles with style-specific evaluation criteria
  - Test part of PLLuMIC (not-published and not publicly accessible)
  - RAG test datasets: public institutions (governmental) and legal framework for public calls
  - *Polish Linguistic and Cultural Competency Benchmark. Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, Rafał Poświata*, https://huggingface.co/spaces/sdadas/plcc
- Methods
  - Limited answer task (log likelihood, pattern matching —> automatic leaderboard
  - Generative tasks: manual evaluation, LLM-as-a-judge —> leaderboard and arena
  - LLM powered arena and LLM-based post-processing of limited answer task results
  - In progress: training LLM-as-a-judge model on the basis of human evaluations

PLCC - *Polish linguistic and cultural competency benchmark*

PLLuM-12B-nc-chat-250715	PLLuM	69.67	72	75	79	52	73	67
DeepSeek-v3	DeepSeek	69.17	61	73	79	62	77	63
Claude-Sonnet-4	Anthropic	68.17	55	72	77	63	81	61
PLLuM-8x7B-nc-chat	PLLuM	68.17	72	76	73	47	73	68
GPT-4-turbo	OpenAI	67.00	61	74	79	56	76	56
Mistral-Medium-3	Mistral	66.83	56	67	77	61	78	62
GLM-4.5	Zhipu AI	66.50	56	68	79	59	77	60
Grok-2-1212	xAI	66.00	57	67	77	64	74	57
Bielik-2.6	SpeakLeash	65.50	61	68	75	55	72	62
Llama-3.1-Tulu-3-405B	Meta	63.83	64	64	71	56	75	53
Bielik-2.2	SpeakLeash	63.00	54	60	72	53	77	62

General Polish-specific Skills Leaderboard

Model	Overall Ranking	Knowledge	Text Understanding	Information Extraction	Text Generation	Linguistic Correctness
Llama-PLLuM-70B-chat	0.6968	0.7383	0.6192	0.7624	0.743	0.576
PLLuM-8x7B-nc-chat	0.6829	0.7053	0.6288	0.783	0.5895	0.5668
PLLuM-8x7B-chat	0.6728	0.6784	0.644	0.7451	0.5645	0.5303
PLLuM-12B-nc-chat	0.6636	0.6836	0.6075	0.7277	0.7162	0.4952
PLLuM-12B-chat	0.6245	0.6291	0.564	0.7121	0.742	0.4427
Llama-3.3-70B-Instruct	0.5585	0.5978	0.4873	0.764	0.2037	0.5753
Bielik-11B-v2.6-Instruct	0.5354	0.508	0.5026	0.623	0.3579	0.5611
Mistral-Nemo-Instruct-2407	0.5089	0.562	0.4255	0.6424	0.4361	0.4286
Mistral-8x7-Instruct-v0.1	0.4795	0.5869	0.3681	0.6258	0.2877	0.4312
Bielik-4.5B-v3.0-Instruct	0.4519	0.4657	0.4273	0.4503	0.2561	0.4752

Polish Language Competence Benchmark

Model	Pragmatics	Punctuation	Orthography	Morphology	Stylistics	Semantics	Syntax
LLama-PLLuM-70B-chat-250801	0.7537	0.5691	0.6244	0.6121	0.8444	0.6408	0.5235
Llama-PLLuM-70B-chat	0.7388	0.4586	0.6293	0.6335	0.8333	0.5807	0.5429
Llama-3.3-70B-Instruct	0.7463	0.4862	0.6634	0.573	0.8889	0.6076	0.5512
PLLuM-8x7B-nc-chat	0.7015	0.5028	0.6439	0.6014	0.7778	0.587	0.5208
Bielik-11B-v2.6-Instruct	0.7687	0.4144	0.6829	0.5943	0.8111	0.5981	0.5069
PLLuM-8x7B-chat	0.6716	0.4144	0.5171	0.637	0.7556	0.5316	0.4626
PLLuM-12B-nc-chat	0.6567	0.4199	0.5902	0.5587	0.7778	0.5253	0.3823
Bielik-4.5B-v3.0-Instruct	0.6642	0.3757	0.5951	0.484	0.7222	0.5206	0.4238
PLLuM-12B-chat	0.5373	0.3702	0.4732	0.5302	0.7111	0.462	0.3767
Mixtral-8x7B-Instruct-v0.1	0.6269	0.3094	0.5073	0.4235	0.8	0.4383	0.3961
Mistral-Nemo-Instruct-2407	0.5746	0.4309	0.4488	0.4235	0.7889	0.4225	0.3906

LLMs in annotation

- Automatic annotation:
  - LLMs for zero- or few-shot annotation tasks,
    - Claims that synthetic labels are often of higher quality and cheaper than human annotations
  - Classifiers fine-tuned on the generated data with a Very LLM “performed comparably to models fine-tuned with human-labeled data. ”, but “slightly worse”(Pangakis & Wolken, 2024)
  - open LLMs fine-tuned on datasets annotated with VLLMs — similar performance (Piper & Bagga, 2024), e.g. discourse analysis

N. Pangakis and S. Wolken. Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. NLP+CSS 2024, ACL, 2024 https://aclanthology.org/2024.nlpcss-1.9/

A. Piper and S. Bagga. Using Large Language Models for Understanding Narrative Discourse. Proc. Of the 6th Workshop on Narrative Understanding, ACL, 2024. https://aclanthology.org/2024.wnu-1.4/

LLMs in annotation

- Automatic annotation — example of a technique inspired by human annotation
  - Best-Worst Scaling with LLMs to produce data for regression training — frequency of species estimated from historical documents (Haider et al., 2024)
- Semi-automatic
  - verified and corrected by humans (e.g. RAG resources in the PLLuM project)

T. Haider, T. Perschl and M. Rehbein. Quantification of Biodiversity from Historical Survey Text with LLM-based Best-Worst Scaling. NLP4Ecology, ACL 2025. https://aclanthology.org/2025.nlp4ecology-1.13/

LLMs in annotation

- LLM-based data augmentation
  - ‘zero-shot’ LLM-based data augmentation — generation from scratch, e.g. generation from the stochastic distributions modelled in the network
  - human-generated prompt examples to direct the process
- Problematic “complex, low-resource domains” (Møller et al., 2024)
  - models trained on human-labeled data consistently exhibit superior or comparable results performance compared to their synthetically augmented counterparts
  - augmentation improve performance on rare classes within multi-class tasks
  - Zero-shot classification with a VLLM is generally outperformed by specialized models trained on human or synthetic data

A. G. Møller, et al. The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks. EACL, ACL, 2024, https://aclanthology.org/2024.eacl-short.17/

LLMs in CLARIN-PL: Applications

- ‘LLM-free’ tools and applications — important for users
- Text preprocessing
  - correction, OCR
- Support in qualitative data analysis (Fischer & Biemann, 2024)
  - fundamental Information Extraction — span classification
    - Basic language tools
  - Complex Information Extraction — document classification
    - symbolic, partial representations
  - Exploration of large corpora — helping researchers to understand the data
    - summarisation, topic modelling, human-in-the-loop in topic modelling adaptation with LLMs
- Flexible text classification – Researcher Assistant
- Natural Language Inference
- Discourse analysis
- Support to users
  - access to knowledge
  - enhanced interaction with and operations on the infrastructure

Basic Language Tools based on PLLuMs

- Morphosyntactic tagging:
  - each word in text mapped onto: lemma, grammatical class and values of grammatical attributes
    - 13 attributes: number, case, gender, person, degree, aspect, negation, accentability, post-prepositionality, accommodability, agglutination, vocalicity, and fullstoppedness
- Annotation schema for LLMs
  - in-texts — text repeated with tags added
  - post-text — language units with tags generated on the basis of the text
- Goal:
  - state-of-the-art model (SOTA) for Polish lemmatization and morphosyntactic tagging on the basis of a ‚micro' size PLLuM with SOTA performance
- Subtotal:
  - morphology-aware tokenization processes for inflectional languages based on suffixes and longest common subsequences (LCS) for Polish and Czech
  - grammatical classes and attributes predefined as tokens

B. Koptyra et al. Breaking Words Right: Morphology-Aware Tokenization for Inflectional Languages. *To appear* 2025 (or 2026)

Tokenisation Informed by Morphology for LLMs

- Tokens follow to large extent the morphological structure
- Better skills, e.g., in interpreting new word derivations
- Problems with multilingual applications

Base-PL (7 tokens)

To jest przykładowe zdanie po polsku.

Suffix-PL (11 tokens)

To jest przykładowe zdanie po polsku.

LCS-PL (11 tokens)

To jest przykładowe zdanie po polsku.

Llama3 (13 tokens)

To jest przykładowe zdanie po polsku.

Base-CZ (8 tokens)

Toto je ukázková věta v češtině.

LCS-CZ (13 tokens)

Toto je ukázková věta v češtině.

Llama3 (15 tokens)

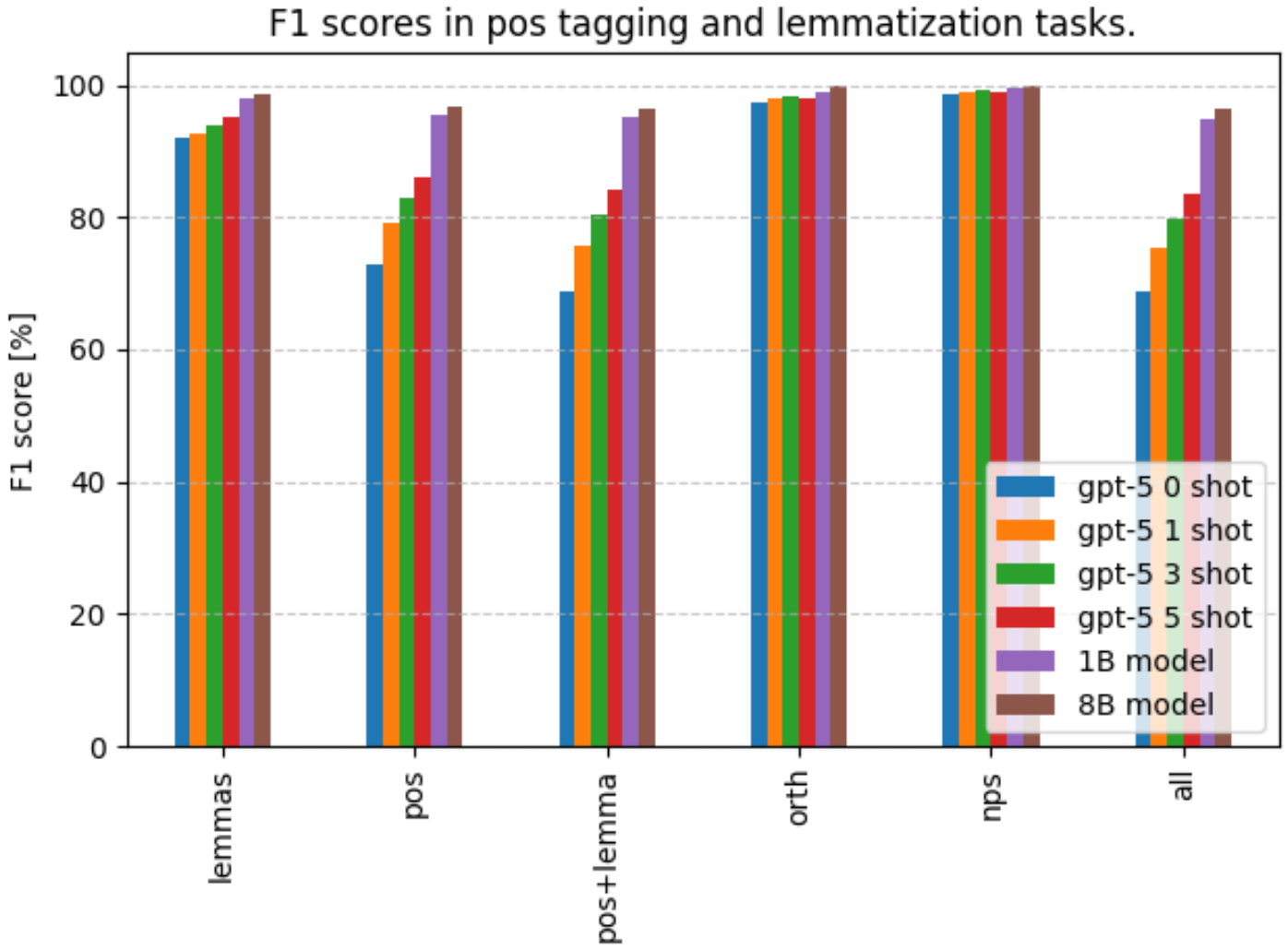
Toto je ukázková věta v češtině.

LLM-based Morphosyntactic Tagger for Polish

- Training data:
  - 1.2M manually annotated segments of the National Corpus of Polish
  - rare cases of splitting words into two segments: word + agglutinative
  - potential cause of problems: LLM eager to generatively ‚explore' this possibility
- Pretraing:
  - Micro-PLLuM of 1B parameters, pre-trained only on Polish data
- Fine-tuning: PLLuM 8B (LLama) and micro PLLuM 1B on NCP
- challenges — mapping the result onto the original word sequence
  - potential splitting within a single word
  - the generation may omit end-of-text tokens, repeating annotations
    - sequences reconstructed from predictions are truncated to minimize Levenshtein distance with the input text
- Results (F1 for tokenisation+lemma+tags)
  - PLLuM 8B: 96.29 > micro PLLuM 1B: 94.95 > Llama3.1-8B: 94.16

LLM-based Morphosyntactic Tagger for Polish

- Comparison with GPT5



Stylochronometry (Walkowiak, 2026)

- Shared tasks: recognition of the period of origin for a literary work or its fragment

Epoch	Train data		Validation data	
	no. of chunks	no. of books	no. of chunks	no. of books
Romanticism	20,161	242	1,158	17
Viktorian	24,0365	3,386	16,938	226
Modernism	238,088	3,671	14,848	245
Postmodernism	22,139	886	1,713	60
Contemporary	25,457	537	1,600	36

Tomasz Walkowiak (2026) Classification of Literary Epochs by TF-IDF, Transformers, and Large Language Models. In L. Rutkowski et al. (Eds.): ICAISC 2025, LNAI 15948

Stylochronometry (Walkowiak, 2026): Baseline

vectorizer	Accuracy	
	chunk level	book level
BoW	72.23%	78.48%
BoW+POS-ngrams	73.10%	**80.81%**

versus

base model	NER	Accuracy		
		chunk level		book level
		first 512-token	maxLogits	maxLogits
xlm-roberta-base	–	65.38 %	70.21 %	79.15 %
bert-large-uncased	–	68.17 %	72.67 %	**80.38** %
bert-large-uncased	+	66.41 %	72.14 %	80.02 %

Stylochronometry (Walkowiak, 2026): Long Prompt 1

```
Assign (answer with only a number: 1, 2, 3, 4 or 5) the  below
text to one of literary epoch:
1. Romanticism (1798-1837) [1]: Romanticism emphasized
emotions, imagination, and the natural world.
2. Victorian era (1837-1901) [2]: Named after Queen Victoria's
reign, this era saw a focus on morality, social reform, and a
growing interest in science and technology.
3. Modernism (1900-1945) [3]: Literature during this period
often emphasized fragmentation, ambiguity, and experimentation
with form and language.
4. Postmodernism (1945-2000) [4]: Postmodernism rejected the
idea of objective truth and grand narratives, and instead
emphasized irony, intertextuality, and self-reflexivity.
5. Our days (from 2000) [5]: Contemporary literature is
characterized by diverse voices, exploring global themes
while embracing hybridity and genre-blurring.
```

Stylochronometry (Walkowiak, 2026): Shorter Prompts 2, 3 & 4

Please assign the following text to one of the literary epochs by responding with a single number: 5 for Romanticism (1798–1837), 4 for the Victorian era (1837–1900), 3 for Modernism (1900–1945), 2 for Postmodernism (1945–2000), or 1 for Our Days (from 2000).

Provide the title, author, and publication year of the first edition of the book from which the following passage is taken. Format your answer as a JSON object with three fields: "text" for the title, "author" for the author's name, and "date" for the publication year.

Please provide the publication year of the first edition of the book from which the following passage is extracted. Format your response as a four-digit integer and add no text, just year.

Stylochronometry (Walkowiak, 2026): LLMs and Prompts

LLM	Prompt 1	Prompt 2	Prompt 3	Prompt 4
GPT-4o-mini	57.80 %	63.16 %	60.52 %	69.53%
LLama 3.3 70B	48.71 %	50.33 %	63.70 %	68.87%
C4AI Command R+	34.56 %	44.96 %	65.97 %	**73.31**%

Stylochronometry (Walkowiak, 2026): Context

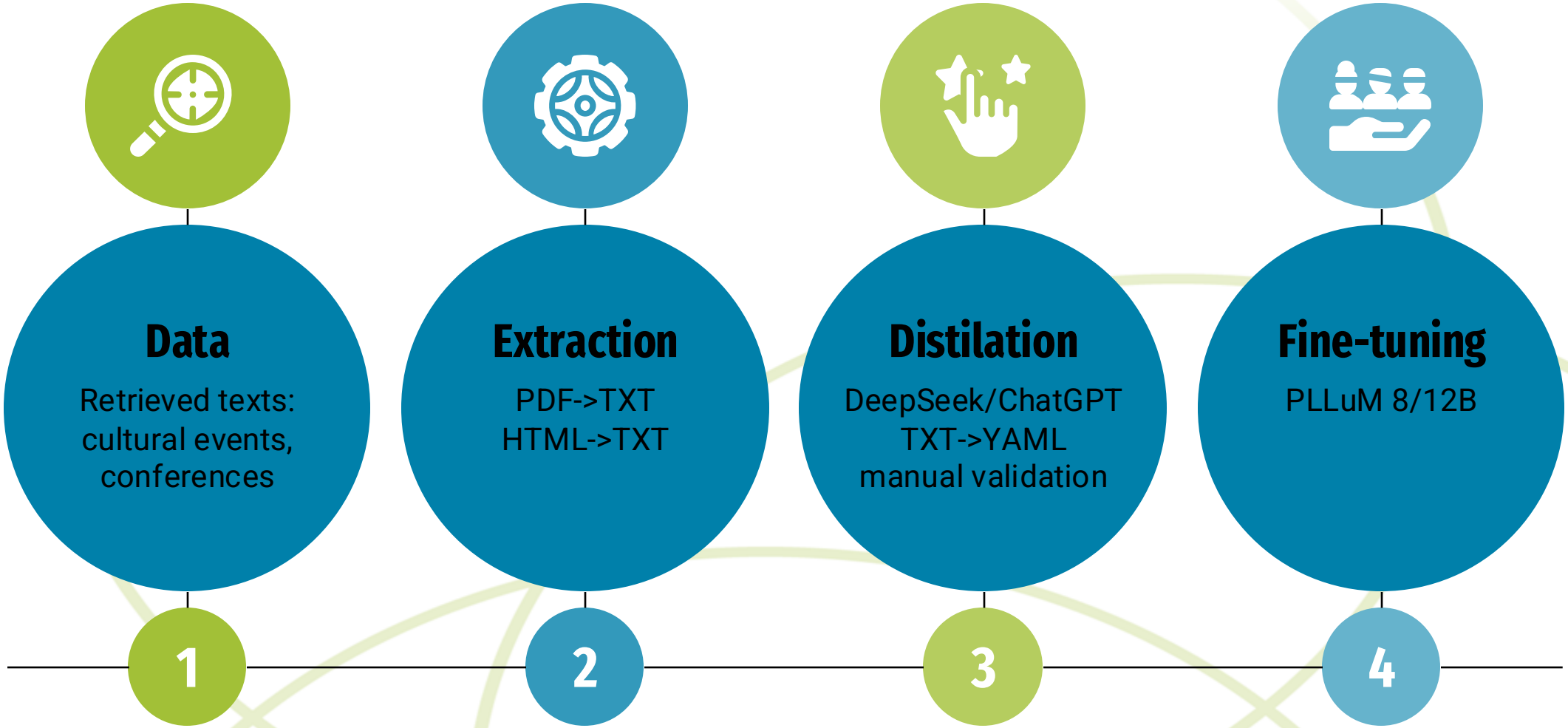
	Accuracy			
	Sequence of chunks		Entire book	
LLM	Prompt 2	Prompt 4	Prompt 2	Prompt 4
GPT-4o-mini	46.71 %	46.53 %	61.87 %	71.04 %
LLama 3.3 70B	46.72 %	46.51 %	50.90 %	66.45 %
C4AI Command R+	46.64 %	46.72 %	5.27 %	**73.80** %

Research Assistant

- Prompts are not ”all you need”
  - accuracy may be disappointing for tasks less typical, requiring deeper reasoning or significantly dependent on the context
  - also in case of processing specific, rare datasets
- Example: recognition of literary metaphors used by authors of the Polish ontological diaries, like illness as a war, journey or play (B. Koper, T. Walkowiak, M. Piasecki, *work in progress*)
  - specific genre with heterogeneous styles, ~15 books of different length published since the year 2000
  - Intensive prompt engineering: from guidelines to feature-based chain-of-thought
  - Several LLMs tested, including commercial Very LLMs, e.g. GPT 5
  - Precision for the positive classes (binary and multiclass) below 50%
- Challenges:
  - Post-processing of the LLM results — an LLM-based interpreter/corrector

Complex Information Extraction

- Leaflets of cultural events ➠ structured records in YAML format



Complex Information Extraction: Cultural Events

- Challenges

Evaluation — Manual or LLM-as-a-judge

Context — Processing long documents

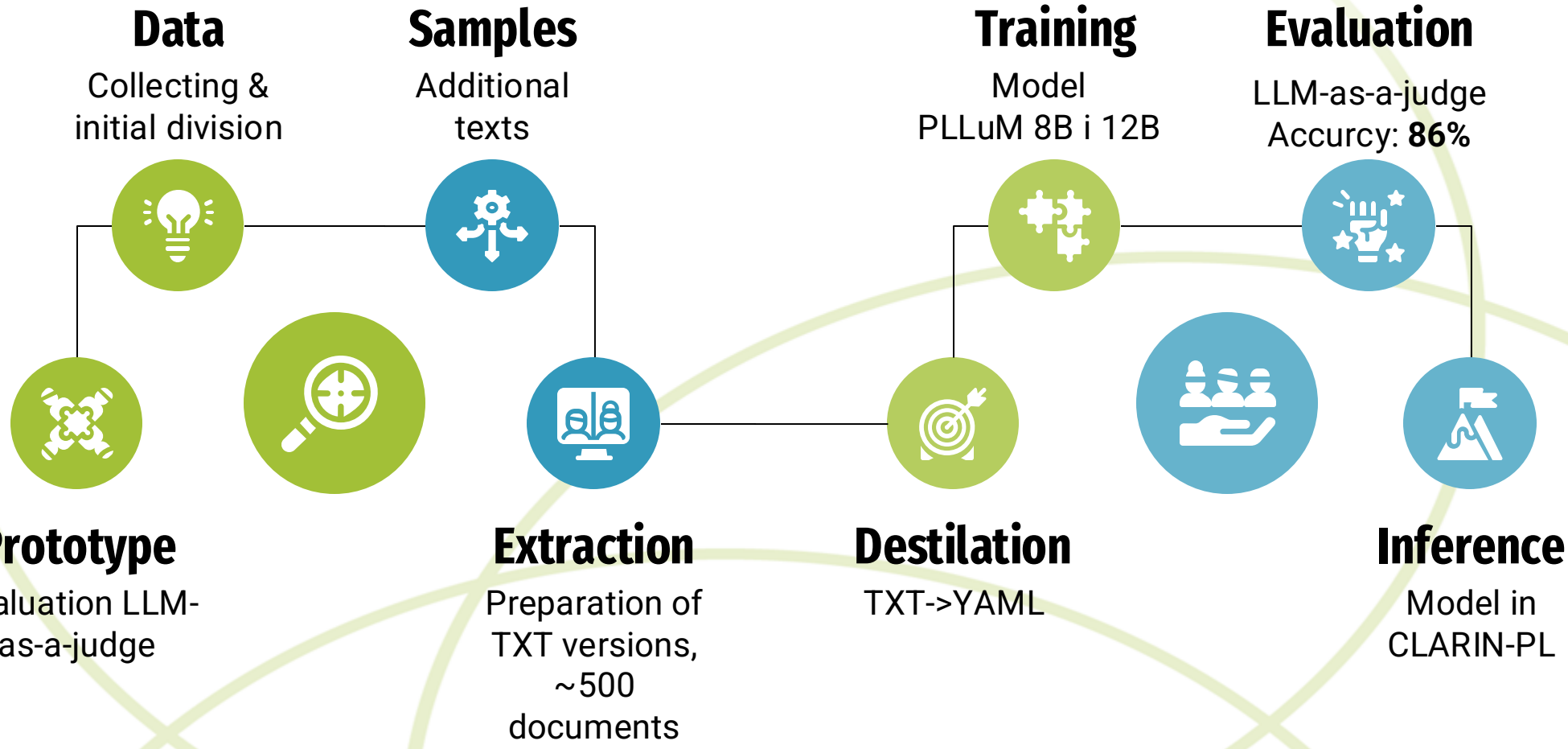
Format — Preferred YAML vs RDF

Control — Correct format from LLM

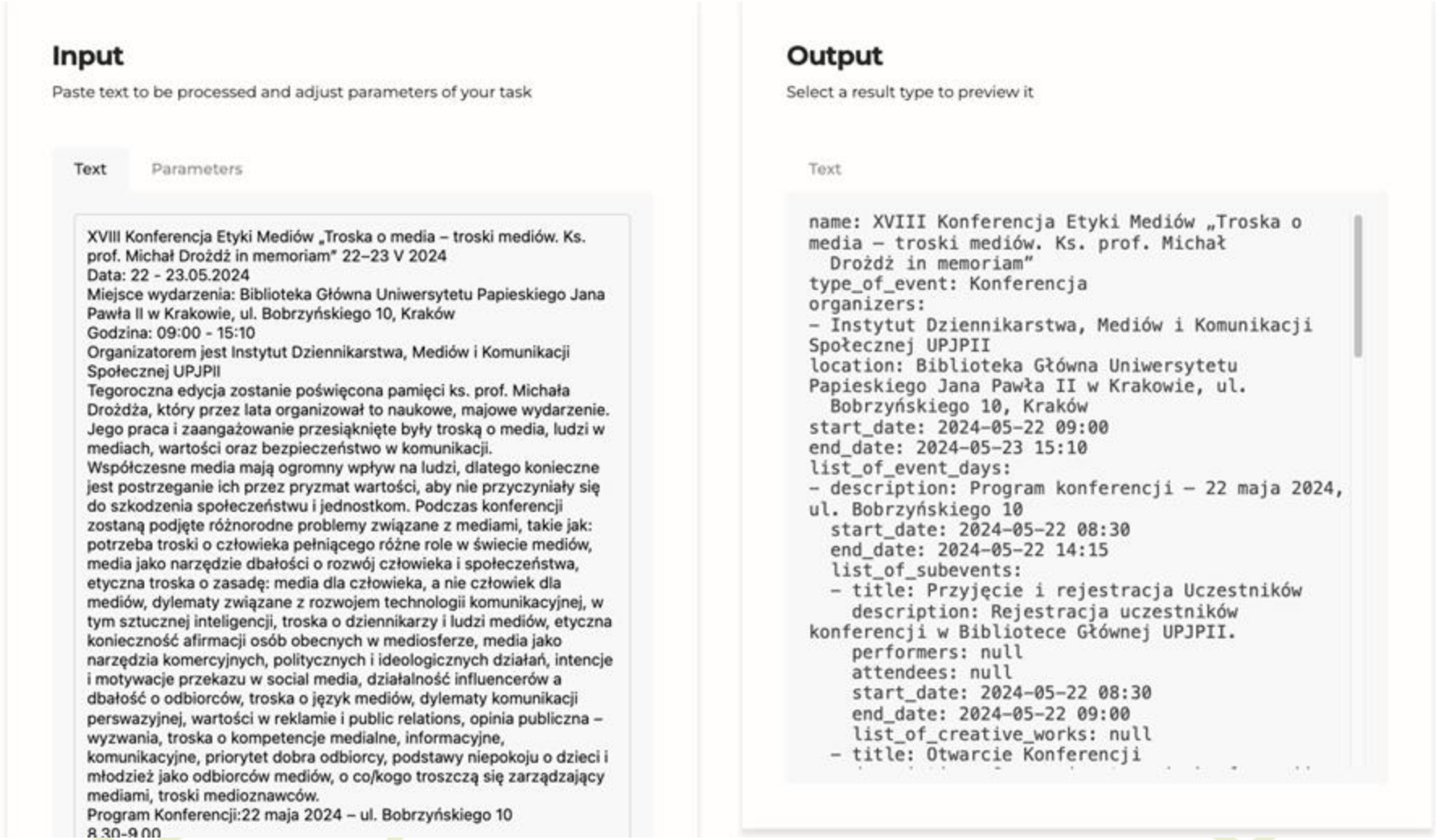
Jan Kocoń et al., Dep. of AI, WrocTech work done for the DARIAH-Hub.pl Project, 2025

Complex Information Extraction: Cultural Events

- Project overview



Cultural Events Extractor – Web Application



https://services.clarin-pl.eu/services/eventstotriples/interactive

(Jan Kocoń i in., Projekt DARIAH-Hub.pl)

Knowledge Assistant

- RAG-based solutions:
  - knowledge base: a collection of documents
    - but also a semantic network, relational database, logical database, multimodal source, etc.
  - a Semantic Retrieval system — for the efficient access
  - an LLM as an engine processing knowledge: query vs source
- RAG-based Agent
  - knowledge: Semantic Retrieval + knowledge source
  - behaviour determined by the knowledge possessed
  - encapsulation of skills specific to the knowledge source type
  - communication language, e.g. natural language
  - agentic systems — RAG Agents specialised to different knowledge sources
  - e.g. CLARIN-PL User Assistant
    - a RAG-based agent system supporting users in accessing knowledge
    - using CLARIN-PL services
    - and solving problems

PLLuM
Polish Large Language Model

CLARIN-PL
Common Language Resources and Technology Infrastructure

PoliChat: Retrieval Augmented Generation on University Documents and Regulations

K. Wojtasik. PoliChat: Retrieval Augmented Generation on University Documents and Regulations, ICCS, 2025
https://dx.doi.org/10.1007/978-3-031-97570-7_21
https://www.iccs-meeting.org/archive/iccs2025/papers/159110266.pdf

Goal

- RAG-based chatbot to simplify access to regulatory information of a large university-like organization
- Special attention: accuracy and transparency
- Prototype experimentally validated in the Wroclaw University of Technology and Science
- Accurate university information with explicit source citations
- Supporting also verification of the information by users themselves
- RAG design choices: model size, document length, retrieval strategies, summarization, prompting methods, and citations

Datasets

- Information Retrieval Dataset
    1. 100-question evaluation set
    2. Question acquisition method
        a. Human-written: high complexity or answer ambiguity;
        b. LLM generated — user intent
        c. LLM generated — document based: often too close to the source
- Filtering 500 initial suggestions according to: answerability, unambiguous binary judgment, and natural phrasing
- manually annotated IR ranking and free search used

Datasets

- Answer Generation Dataset
  1. The same questions as in IR, more detailed annotation
  2. Relevant chunks and ditractors
  3. *Include words* (ensuring completeness by requiring their presence in correct answers, with some alternatives, e.g., ["four", "4"])
  4. *Exclude words* (incorrect or irrelevant terms)

Retrieval Augmented Generation (RAG)

Set of documents: information, data in text form, knowledge, indexes

User **questions** in natural language

Retriever searches for documents relevant to the question: top 100-1000 best semantically matching the question

Re-ranker analyses each document and re-assesses in relation to the question a re-ranked and clipped document list

Generator (LLM) returns an answer on the basis of several (the top ranked) documents

PoliChad Architecture

- Generation
  1. Llama 3.1-8B, Llama 3.1-70B, Llama 3.3-70B (enhanced reasoning and instruction-following)
  2. Bielik — an open Polish model
  3. Command-R-Plus — a 104B model specialised for RAG and citation-grounded answers
  4. **Llama-PLLuM-8B and Llama-PLLuM-70B-chat**

PoliChad Architecture

- *Analyse & Answer* prompt
  1. The model receives numbered fragments
  2. Reviews retrieved documents, identifies relevant ones and extracts key information
  3. The model then generates a response based solely on the extracted information

Analyse & Answer Prompt (*translation*)

....

Answer the questions in two phases, the first phase will be to analyse the documents and you will start with "Analysis:".

During the "Analysis:" phase, you will plan your answer, select the documents that contain the correct answer to the question, write out the most important information and analyse the content of the documents.

Then move on to the phase of aggregating the information and generating the correct answer. This phase should start with "Answer:". Give a high quality, grammatically correct answer in Polish. The response should include citations to the documents from which the information is taken. Cite the document using the symbol [nr_document] citing the passage e.g. [0] for a passage from document 0. If the documents do not contain the information needed to answer the question, return the text instead of the answer: "I was unable to find the answer to the question"

...

Evaluation metrics

- IR Measures: MRR@10, NDCG@10, Recall@100, Acc@5
- *Inclusion Accuracy* — checks if all required *Include* words are present, ensuring factual completeness
- *Exclusion Accuracy* — verifies that misleading or irrelevant *Exclude* words are absent.
- *Citation F1* — measures correct citation of relevant document chunks while avoiding distractors.
- *Answer Length Analysis* — tracks character and word count to distinguish between verbosity and precision
- *Total score* — the average from *Inclusion*, *Exclusion* and *Citation F1*

PoliChat

PLLuM as engine for answering questions on the basis of internal documents of the Wrocław University of Science and Technology

PDFs and DOCXs divided into 12k fragments – indexed and updated on regular basis.

Answer Generation Results

- Prompt type
  - Citation F1
    - Improved by A&A for 512 passages and all models, except Llama models: lower recall, but higher precision
  - Longer (4K): decrease exclude score
  - Generally, A&A —> consistent improvement in total scores
- Impact of Citations on Response Quality
  - Basic prompt — significantly reduced correctness
  - A&A mostly improved correctness, e.g. Llama3.3-70B (87. 00% → 88. 14%).
  - citation-heavy prompts almost doubled the response length.

RAG Pipeline Evaluation

- LLM applied to the results of IR
- k∈ {5, 10, 20} retrieved passages: performance declines with more chunks, especially F1 for citations
- A&A prompt mitigates this decrease, especially for 4k chunks
- Summarisation (4k passages) always significantly lowers performance
- Total score
  1. short (512): Command-R-Plus + basic + top 5 (81.47) > Command-R-Plus + A&A + top 5 (80.27) > LLama 3.1-70b + A&A + top 5 (79.83)
  2. long (4k): LLama 3.1-70b + A&A + top 5 (76.1) > Command-R-Plus + A&A + top 5 (75.12) > LLama 3.1-70b + basic + top 5 (73.05)

Evaluation of PLLuM in RAG

PoliChad					
Model	**Total**	**Include**	**Exclude**	**Cite F1**	**Words**
Llama-PLLuM-8B	84.80	88.00	86.85	79.55	42
Llama-PLLuM-70B-chat	**86.75**	89.17	85.28	85.81	41
Llama-3.1-8B-Instruct	73.00	82.08	88.22	48.70	31
Llama-3.3-70B-Instruct	83.94	87.33	84.22	80.28	43
Command-R-Plus	81.21	80.25	88.86	74.53	21

Evaluation of PLLuM in RAG: Citizen Assistant

Citizen Assistant					
Model	**Total**	**Include**	**Refuse**	**Cite F1**	**Words**
Llama-PLLuM-8B	86.08	82.81	92.31	85.94	63
Llama-PLLuM-70B-chat	**89.28**	90.03	96.15	87.43	61
Llama-3.1-8B-Instruct	72.33	78.00	76.92	62.91	48
Llama-3.3-70B-Instruct	87.80	87.12	88.46	88.65	65
Command-R-Plus	84.22	83.76	92.31	81.77	62

Conclusions for PoliChat

- Retrieval quality remains critical for RAG
- Longer documents reduce recall
- but summaries decrease performance, with extractive retaining more more relevant information than abstractive ones.
- Citations improve transparency, but sometimes reduce correctness, and longer retrieved contexts make citation selection more challenging.
- Citation accuracy is valuable for users
- Analyse & Answer strategy addressed these challenges, improving overall performance, citation accuracy, and readability.

Lessons Learned and the Route To Follow

- Very LLMs are powerful, flexible and remarkable versatile in different tasks
- but as scientific tools are not transparent and does not guarantee reproducibility, even in short perspective
- Open science and Very LLMs are not on the same track
- Issue of the LLM efficiency is also a question of their overuse
- Evaluation of LLMs, especially of the proprietary Very LLMs needs a lot of work and attention
- Technological sovereignty sounds as a buzzword, but may be really useful and needed for researchers, public institutions, many natural languages, but also local economies
- LLMs do not replace research and technological infrastructures — they are only tools

Access to CLARIN ERIC, CLARIN-PL, PLLuM

- Web pages and web applications
  - Research tools and applications: https://services.clarin-pl.eu/
  - Web access to selected open LLMs (mainly for scientists): https://pllum.clarin-pl.eu/ https://chat.clarin-pl.eu/
- Web services (also web APIs): https://services-test.clarin-pl.eu/api/v1/docs
- Huggingface
  - Ministery of Digital Affairs, PLLuM: https://huggingface.co/CYFRAGOVPL
  - CLARIN-PL - official portal: https://huggingface.co/clarin-pl
  - CLARIN-PL - test solutions: https://huggingface.co/clarin-knext
- Support
  - knowledge centres of CLARIN-PL:
    - PolLinguaTec (https://kcentre.clarin-pl.eu/),
    - LLMs4SSH (https://llms4ssh.clarin-pl.eu/)
  - Helpdesk: helpdesk@clarin-pl.eu

Acknowlegements

1. CLARIN-PL (2024–2026), funded by the Polish Minister of Science and Higher Education (agreement no. 2024/WK/01)
2. CLARIN-PL 2025-2027, the European Regional Development Fund, FENG Programme (FENG.02.04-IP.040004/24)
3. Statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; The views expressed are those of the authors and do not necessarily reflect those of the EU or the European Research Executive Agency

CLARIN

Fundusze Europejskie

Rzeczpospolita Polska

Dofinansowane przez Unię Europejską

You are always very welcome!

www.clarin.eu
clarin-pl.eu
lub
clarin@clarin.eu
clarin-pl@pwr.edu.pl
maciej.piasecki@pwr.edu.pl

CLARIN-PL
Common Language Resources and Technology Infrastructure

http://clarin-pl.eu/
https://services.clarin-pl.eu

CLARIN

http://clarin.eu/