Intro by Marek

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

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

Nowadays, in the LLMs world the most popular ones are generative multilingual models from vendors such as OpenAI, Anthropic, Google, Meta, Cohere. They are big enough (even large ☺) to resolve vast majority of tasks, but there are applications where the cloud based solutions are not allowed to be used. If you have to deploy on-premise LLMs the number of gpus and energy consumption appear to be crucial aspects, the customers/sponsors demand to minimize costs, and so we have the problem of trade off between effieciency and costs, the size of the model is relevant in such deployments. **Our main goal is to provide as small model as possible that is able to resolve efficiently the given buisness cases. Also there are some tasks that do not need the generative models as e.g. classification/regression tasks.**

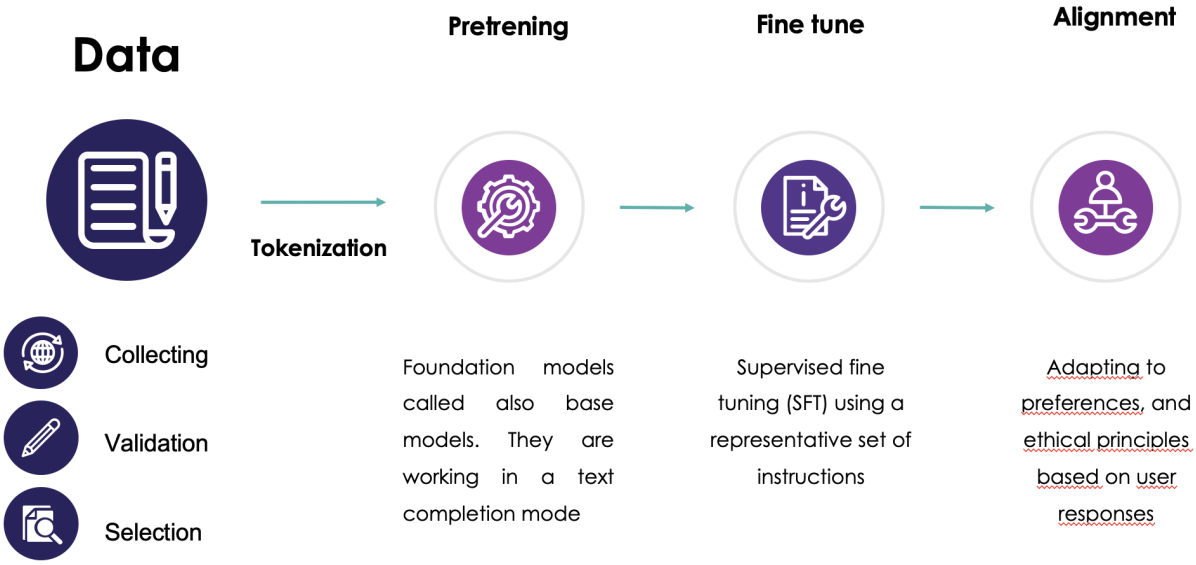
[INFORMATION-TIP]. You have to remember that neural language models (NLMs) are not only generative ones.
Two main categories of NLMs:

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

Polish localization means improving the understanding of the Polish language and/or creating high-quality Polish content (quality evaluated in such dimiensions as linguistic and cultural aspects). This defintion can be adopted to any kind of other languages as German, French, Spanish etc.

In AI LAB im my Institute NIPI, we built from scratch 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 than the existing multi-language models simiilar size.

How we build LLMs



Today, localizations are a must; small localized models will become part of this process in the future. The AI agentic revolution will be based on the small localized models, the companies, organizations future will be driven by such models.

Qs for Marek Kozlowski – my comments below

- My sense of the benchmark data is that … LLMs perform worse in non-English languages, and pretty much in proportion to just how low-resource the language is, though that does seem to be improving from one generation to the next. What complaints do you have about how the leading ChatGPT, Claude, Gemini models perform in Polish?

    - for example, I've heard people say that they sound like native English speakers who learned to speak another language

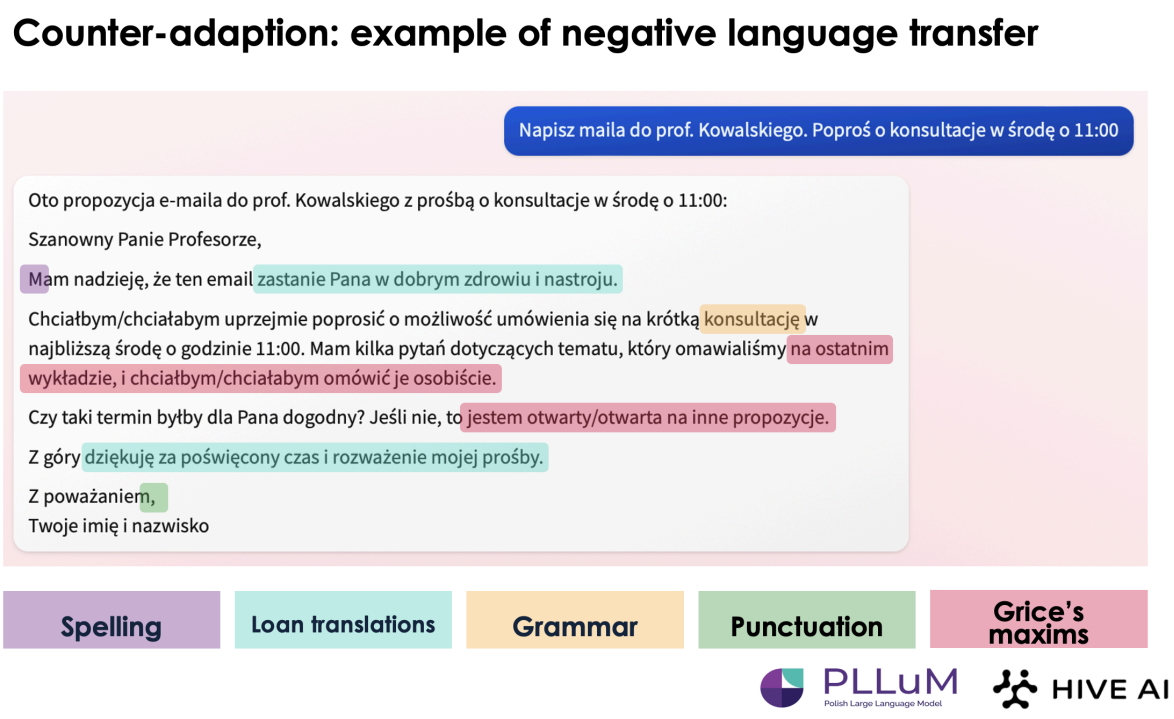
Marek comment: remember that learning models is like teaching kids a new language, first they improve understanding abilities, then writing and speaking. Generally speaking the 90+% of training data used by big vendors are English and Chinese texts. Other languages are less than 10% e.g. Polish < 1%. This means that the well known models are the best in those two major languages. However, they are still able to communicate in the niche languages, because they have seen some relevant number of texts in those niche languages, and if they are big ones (several hundreds of billions of params, trillions params even, e.g. 685B params in DeepSeek models) they can efficiently extrapolate between tasks and competencies. For

example: even if the model has not been trained on some examples of Spanish math calculations, but it was learned to resolve Ensligh instructed math calculations, and it is able to translate commands from Spanish to English, it will perform those latent, intermediate steps to achieve the final success.
You have to remember that large models are very good in extrapolations, especially when we talk about the Extractives, Analytical tasks.

Instructions can be categorized in a such way:

1) Factual ones - the questions about the history, world etc
2) Generative ones - write the poem, essay, source code, make some calculus, chitchats
3) Extractive ones - extract key information from the longer context
4) Classification and Adversial ones

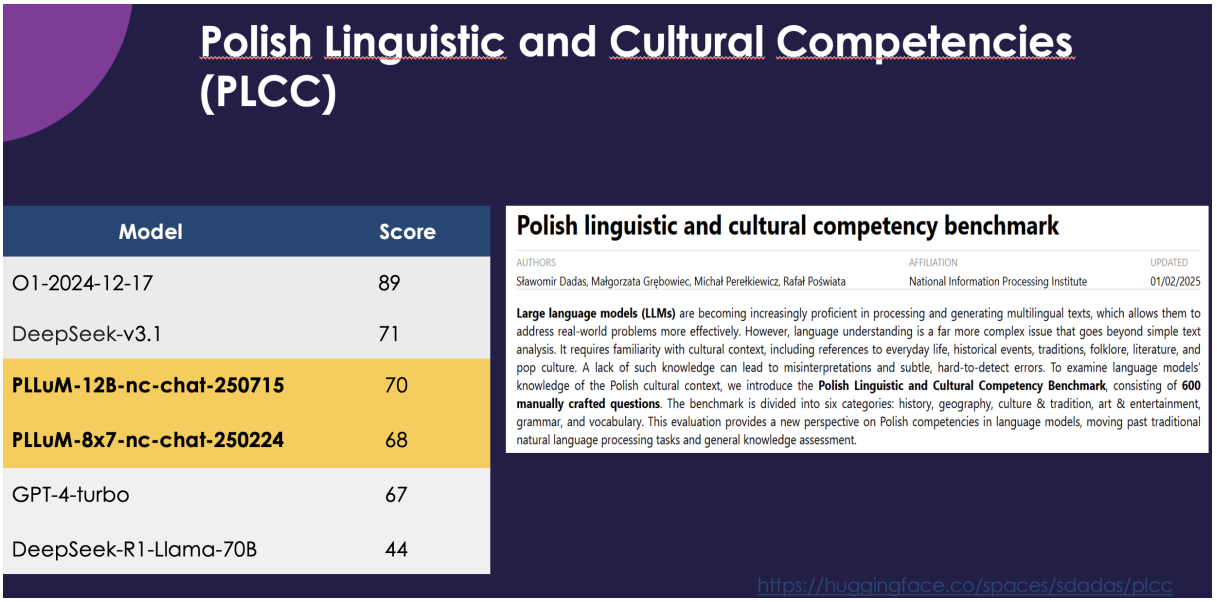
And a lack of training data in niche languages as Polish language is the most impactful for the Generative Tasks like creating the longer text with minimal number of grammar mistakes or some language calques. Below, I present the example when we ask the LLM to write email in Polish language and the model creates Loan Translations that are understandable, but not typical for Polish language. This is sign/proof how LLM translates from English emails to Polish ones if they dont have enough knowledge built upon organic Polish emails.



In AILAB We created dedicated benchmark, Polish Linguistic and Cultural Competencies, having 6 subcategories and almost 1000 questions. **Our goal is to create moderate size models that will be as good as the biggest ones in the aspects of PLCC.**

culture & tradition | art & entertainment | geography | history | grammar | vocab

ID	Question	Verification	Subcategory
6	Jak nazywała się subkultura młodzieżowa funkcjonująca w Polsce po II wojnie światowej, którą cechował bunt wobec narzuconych norm oraz fascynacja muzyką jazzową i kulturą amerykańską, za co jej przedstawiciele byli prześladowani przez władze PRL?	**include**: {bikiniarze, bikiniarzy, bikiniarstwo}	popculture
23	Jaki gatunek małpy stał się bohaterem polskich memów o "typowym Januszu"?	**include**: {nosacz sundajski, nosacza sundajskiego, nasalis larvatus}	popculture
38	Który z poniżej wymienionych klasztorów jest najstarszym z istniejących klasztorów w Polsce? 1. Opactwo benedyktynów w Tyńcu 2. Klasztor zakonu paulinów na Jasnej Górze 3. Opactwo benedyktynów na Świętym Krzyżu 4. Opactwo cystersów w Lubiążu Podaj pojedynczą liczbę 1, 2, 3 lub 4 odpowiadającą poprawnej odpowiedzi. Nie dodawaj komentarza.	**include**: 1 **exclude**: 2, 3, 4	religion & tradition



- In the US people talk about AI being built with American values or Democratic values, but I think very often what this means is subjective to a lot of creative ambiguity. But maybe in a more homogenous state like Poland that's less of an issue. Who decides what Polish values / culture you try to encode into the model, and how would you describe it? Is there a constitution or model spec that you've developed?

Marek comment: We have similar problems, because the vast majority of data are the web data, and they have different apect addressed inside e.g. religious ones, or geopolitical points of views. However, we live in a smaller country, more homogenous so the distribution of problems is also different than in USA. LLMs memorize, what we write, create in the Web, they are somehow the compressed representations of our digital content, but also they have those competencies that are included within set of instructions presented in SFT stage, they are so constrained as they were secured in the preference optimizations steps. Building our own models we have ability to decide about what data are used in training – what the LLM knows, what it can and how it is secured. Using external models you know nothing about the data used in training stages as pretraining, SFT or alignment stage (preference optimization). We dont have any constitution according to the learning phases, only some books of recipes and catalog of instructions and preferences.

- Where are you getting your tokens? What do they consist of? How much is about values vs culture vs practical local factual knowledge (eg local products, etc) vs local procedural knowledge (eg how to file paperwork to sell & transfer a car, etc) vs other?? Are you working with data creators – is the a Polish Scale or Labelbox type company? How much data is synthetic?

Marek comment: the vast majority of our data is a Polish Web Archive. However we perform very strict process of data curation – deduplication and filtering stages. We also check everything towards AI Acts at the national level and EU level. Some example of such regulations below. Our huge advantage is the number of organic instructions and preferences we built, and how they prevent us from the problem of counter adaptation.

PLLuM resources

High-quality, legally sourced textual data

Corpora of **Polish** texts (approx. 400 million documents and **150 billion tokens**, incl. 28 billion used for the open model), as well as **English, Baltic**, and **other Slavic languages**; content acquired under **licensing agreements** with publishers

Why two training corpora?

Amendment to the Copyright Act – 20 August 2024: Permitted Use for Text and Data Mining (TDM)

- Reproduction of disseminated works for TDM purposes, unless the rightholder has effectively opted out → 28 billion training corpus for models released under a fully open license
- Reproduction of disseminated works for TDM in **scientific research**, provided it is **non-commercial** and carried out by **entities listed in the Act** → 150 billion training corpus corpus for models released under a restricted (non-commercial) license

HIVE AI

PLLuM resources

High-quality, legally sourced textual data

Corpora of **Polish** texts (approx. 400 million documents and **150 billion tokens**, incl. 28 billion used for the open model), as well as **English, Baltic**, and **other Slavic languages**; content acquired under **licensing agreements** with publishers

Original fine-tuning data

Approx. **50,000 manually created single- and multi-turn instructions** (prompt–response pairs) and around **130,000 preference pairs** (preferred vs. rejected model completions), developed according to a original typology

Additional safeguards

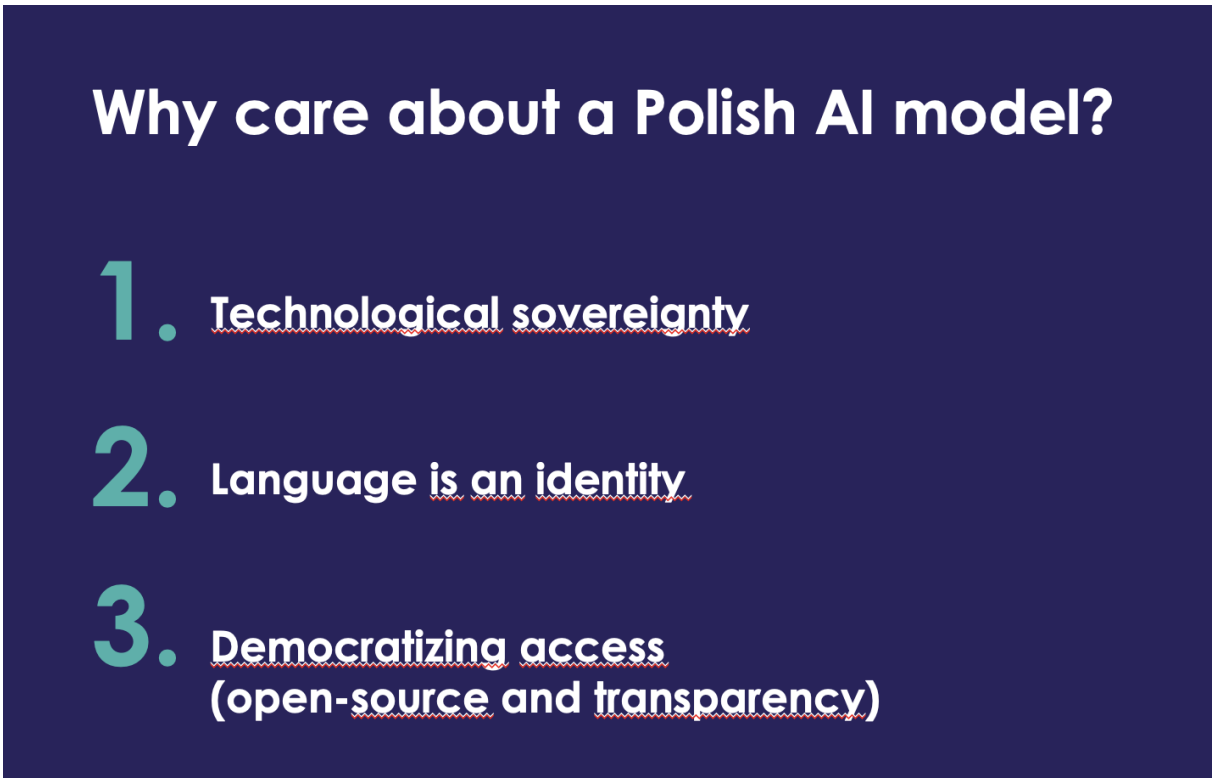
Minimization of the risk of generating harmful content through **alignment** techniques, as well as **output filtering** and correction algorithms using custom classifiers

Original evaluation datasets

Internally developed **benchmarks** to assess model performance across various use cases, incl. public administration; a dedicated **leaderboard** and comprehensive **tests for adversarial robustness**

HIVE AI

- Have you / would you consider giving your tokens to frontier model developers on a silver platter? It seems to me that … if the goal of a national project is to make sure that Polish users have first class experience with AI, then this could be a really good strategy.



Yes we have already published samples of our organic instructions and preferences as an open-source assetts, we will publish bigger samples in a few weeks.

- How does Poland think about the need to control compute and to compete to retain top tier talent? Is the Polish government AGI pilled enough to pay its AI researchers enough to keep them from getting o1 visas?

The problem is not currently addressed, I think there is a lack of strategy to retain talents. Why? Because there is undefined objectives how to use such talents, what are the partial goals to achieve etc The half of OpenAI core researchers are Polish guys, we have the talents but we dont have product mindset how to become a top player in those field.

- How much information exchange / support / camaraderie is there among people trying to advance "sovereign AI" strategies across countries? Do you think we will see a sort of "Non-Aligned Movement" of countries trying to stay more neutral between the US and China as there once was in the Cold War with respect to US and USSR?

There are initiatives as ALTEDIC consortium, and projects as e.g. OpenEUROLLM, LLMs4EU

- The Polish Large Language model is … PLLuM

    - what base model and why? would you be open to Chinese?

      Mistral and LLama models whith a fully open licences

    - you do what you call language adaptation first – how does this work? and how well? is it fair to call it "continued pre-training"? and could it work for companies as well as countries?

Adaptation is cont. pretraining, it means we start from existing checkpoints and continously pretrain them on our language/domain high quality corpora, it works for niche languages and domains as well. Remember, that after pretrain you have fine tuning, and in this stage using synthactic instructions is risky step because of the counter adaptation problem.

    - what compute did you use? something in Poland?

    - how often do you plan to update / upgrade?

Apprpox. each year

    - it seems it was built via a sort of Public-Private partnership? can you describe the overall team?

We are fully open, not only open weigths, but we also show the instructions and preferences on huggingface, now they are samples of them but we expand them

- When it comes to US AI stack vs Chinese AI stack… how do you choose?
    - Satya recently said on Dwarkesh that being the partner you can trust is how you win the world in AI, and I think that's brilliant, but … trust to do what exactly? is China's open source more or less trusted than US proprietary stack?

There is a huge demand for powerful localized, open-source models, even NVIDIA claims that there will be the engines of AI Agentic Revolution. Chinese models currently are more open-weights than US based ones. BUT remember about censorship embededd within them during learning.

- Use cases…
    - mObywatel – Citizen assistant – that's cool! what can it do?
    - PLLuM chat… how is it distributed? is it fully open source or is there anything you had to hold back and if so what?

There are plenty of applicatins – mObywatel, the city offices chatbots, ERP systems, legal assistants (evaluation of regulation changes) , internal banking search engines. We have the open chat services e.g. https://pllum.clarin-pl.eu/pllum_8x7b