AI SPEAKS POLISH

how open models drive generative AI development in smaller markets

OPEN _FUTURE

CENTRUM CYFROWE

NOVEMBER 2024

TABLE OF CONTENTS

INTRODUCTION

Progress in the development of generative artificial intelligence (AI) technologies is increasingly met with a concern over uneven allocation of AI's benefits, worsening both domestic and global inequalities. Machine learning exhibits traits associated with natural monopolies, and general purpose AI technologies have a "winner-takes-most nature". A few powerful companies have privileged access to proprietary data, computing power, human capital and existing computer bases.

Commercial foundation models are trained in a paradigm that assumes a constant scaling of technologies, which drives both market competition and technological power. These models require huge financial investments that can only be afforded by companies with monopolistic positions in the digital market.

It might therefore seem that the economics of creating AI technologies preclude the emergence of alternatives - whether publicly funded or created by smaller commercial players. However, a steady stream of alternative solutions has been in development, in parallel to the dominant solutions, and benefitting from same, openly shared research and open source technologies. In 2022, when OpenAI released the ChatGPT service, the open source BLOOM model, created by a community of researchers backed by Hugging Face, was released simultaneously.

The growth of alternative solutions is increasingly seen as one of the ways, alongside regulation, to curb concentrations of power in AI, and thus to democratize AI development.[1] These solutions are also often positive examples of open innovation, commons-based collaboration and exploration of solutions that meet public interest goals.

They are also a way to address the under-representation of a majority of world's languages and cultures in the dominant models. Commercial model development has largely focused on only several out of twenty languages considered "high resource": with enough publicly available data to train a model. This created a language gap that often goes hand in hand with various other forms of digital divides. Across the world, local, open source language models are being developed to fill this gap.

Today, the new paradigm of creating small language models[2] and the availability of open foundation models makes it possible to efficiently create new language models - particularly those that address language gaps in generative AI development. This report presents a case study of a Polish ecosystem, in which open language models are being developed as Digital

---

[1] Alek Tarkowski et al., "Democratic Governance of AI Systems and Datasets", T7, https://think7.org/democratic-governance-of-ai-systems-and-datasets/
Joe Westby et al., "Beyond Big Tech: A framework for building a new and fair digital economy". https://static1.squarespace.com/static/65c9daef199ea70aa66592fe/t/66f9c85ac5c3f44309088bfa/1727645794775/Break+Open+Big+Tech+White+Paper+-+FINAL.pdf

[2] The term "small language models" is used to describe language models of smaller size than so called "large language models", measured by the amount of model parameters. The boundary between large and small models is fuzzy.

Commons.[3] These cases of model development are examples of a public AI approach: of building infrastructure for the common good and with public orientation of AI in mind.[4]

While there is growing interest in small language models, there are to date few case studies of their development. Key questions concern how these initiatives obtain necessary resources (data, compute, skilled workforce, users), how they prepare and govern training datasets, how they share resources and collaborate in the open, and how they develop partnerships that make the models used, in a sustainable way.

This report focuses on two such initiatives that are based in Poland. SpeakLeash is a community that has been building a Polish language dataset, and in April 2024 built on its basis Bielik, a Polish small language model.[5] And PLLuM (Polish Large Language Model) is a consortium of public research institution that aims to create a language model also tailored to the specificity of the Polish language. We also demonstrate how a broader ecosystem has emerged around these initiatives.

The report is based on interviews with the creators of Polish models. Based on them, we analysed model development processes and challenges that they have to solve. We also draw conclusions from their achievements that can be help to support such initiatives in the future.

We hope that the learnings from this report will be useful worldwide, as they provide insights into how small language models are built, what are the needs of model developers and other stakeholders, and how these efforts can be supported through public policies.

In the report, we describe the successive stages of model development, focusing on how key resources are secured: computing power, high-quality data and teams of experts with needed skills. We also show how the developers of each project are thinking in terms of a broader ecosystem, assuming collaboration between different organizations and initiatives, and open exchange of AI tools and components. The two projects are based on two different methodologies: community-based peer production of a Digital Commons, and a more traditional research consortium. These prove to be complimentary within the broader ecosystem, which overall exhibits traits of a Digital Commons.

The case study shows how distributed, community-based development plays an important role in this ecosystem, as well as collaboration between various entities in this ecosystem: from data collection and verification, to model training and quality control. The use of existing open technologies, primarily model training architectures, is also an important enabling factor. We also describe how model developers are dealing with legal issues surrounding the use of

---

[3] Jan Krewer and Zuzanna Warso, "Digital Commons as Providers of Public Digital Infrastructures", Open Future, https://openfuture.eu/publication/digital-commons-as-providers-of-public-digital-infrastructures/

[4] Nik Marda, Jasmine Sun and Mark Surman, "Public AI. Making AI work for everyone, by everyone", Open Future. https://assets.mofoprod.net/network/documents/Public_AI_Mozilla.pdf
Brandon Jackson et al., "Public AI: Infrastructure for the Common Good", Public AI Network. https://doi.org/10.5281/zenodo.13914560

[5] Bielik is the Polish name of the white-tailed eagle. A white eagle is the national symbol of Poland, visible on the Polish coat of arms.

training data, and establish collaborations to acquire new data. Finally, the work on local language models shows a need to adapt existing tools, be sensitive to the local context and understand cultural and linguistic nuances.

The goal of the report is to raise awareness that open language models are being developed to reduce language gaps in AI development and provide development for alternative technologies. Learnings from the case studies can also help in formulating public policies to support the development of such alternatives. Our key findings include:

- SpeakLeash initiative is a rare example (alongside projects like BigScience or Eleuther.ai) of building a language model as Digital Commons, of community-driven development and governance of AI models;
- Public supercomputing centers have sufficient computing power to successfully train small language models, although funding for sufficient training runs is a necessary requirement;
- Polish open model builders both make use of existing AI tools and components to make their work more effective, and create their own tools when needed - for example, to address local context and needs;
- While most model development projects globally work with same web crawled data, Polish developers find new ways to ensure data quality and its local relevancy;
- An AI development ecosystem means that there are multiple models being created (through both pre-training and fine-tuning), ensuring a plurality of technologies and embedded social and cultural perspectives;
- Polish AI developers are establishing successful cooperation with various content owners, to increase the pool of Polish language training data.

TRENDS IN THE DEVELOPMENT OF OPEN AND SMALL LANGUAGE MODELS

A Large Language Model (LLM) is a type of artificial intelligence system that can process natural language and generate text. LLMs are trained to do this by statistically analyzing large text datasets. The most well-known LLMs are commercial models, and the largest are developed by five global companies. These are the GPT models from OpenAI, the Copilot models from Microsoft, the Claude models from Anthropic, the Gemini models from Google and the Llama models created by Meta. These largest models are often referred to as foundation models to emphasize that they are general-purpose technologies. Some experts see this as evidence of emergent behavior of the models, while others see it as the result of learning on massive sets of diverse data.[6]

Foundation models are thus trained on huge, and ever-larger, datasets, as measured by the volume of data, or the number of tokens - short clusters of text. For example, Llama 1, released in February 2023, was trained on one trillion of tokens, Llama 2 (from the same year) on two trillion, and Llama 3, released in August 2024, on 15 trillion tokens. More and more computing power is also needed to train models. It is estimated that training the recently released Llama 3 model required computing power equivalent to 10^{24} FLOPS. By comparison, the fastest supercomputer in Poland has the computing power of 10^{15} FLOPS.[7]

The size of the training data set and of the computing power used translates into the number of model parameters, that is, the number of factors the the model takes into account when processing and generating content. The first language models had millions of parameters, and the largest contemporary models have hundreds of trillions, or even more than a trillion parameters. Developers of foundation models assume that models with more parameters are better.

Building large-scale language models therefore requires an enormous amount of funds, necessary to secure, first of all, computing power, but also to pay highly qualified experts, and to acquire and process data. The cost of creating the latest generation of models today is estimated to be at least $100 million - and the cost of subsequent generations of models can be many times higher.[8] That's the reason why large, foundation models are built by a small number of companies, supported with a lot of equity and venture capital investment.[9]

It would seem that the economics of creating AI prevent the emergence of alternatives - whether publicly funded or built by smaller commercial players. A symbol of the difficulty of building

---

[6] Elliot Jones, What is a foundation model?, Ada Lovelace Institute (17.07.2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

[7] Andrej Karpathy (18.04.2024), https://x.com/karpathy/status/1781047292486914189.

[8] Ethan Mollick, "Scaling: The State of Play in AI, One Useful Thing", One Useful Thing (16.09.2024), https://www.oneusefulthing.org/p/scaling-the-state-of-play-in-ai.

[9] "Generative AI Venture Capital Investment Globally On Track To Reach $12 billion in 2024, following breakout year in 2023", EY (16.05.2024), https://www.ey.com/en_ie/news/2024/05/generative-ai-venture-capital-investment-globally-on-track-to-reach-12-billion-dollar-in-2024-following-breakout-year-in-2023.

these alternatives is the French company Mistral, which was supposed to be the French national champion in AI, able to build open solutions tailored to the needs of France and other European countries. In the end, the company entered into a strategic partnership with Microsoft after six months, and decided not to openly release its strongest models, contrary to initial declarations.

However, opposite to initial market expectations, alternatives are being developed in parallel with the development of the aforementioned large commercial models. In 2023, the year that OpenAI released the GPT-3-based ChatGPT service, a research network supported by HuggingFace released BLOOM, an open multilingual model.[10] Already in 2021, the Eleuther.ai foundation has released the GPT-J model. And AuroraGPT, a publicly funded scientific foundation model, is currently under development at Argonne National Laboratory in the US.[11]

A large part of these models are created in a different paradigm than commercial models – that of so-called small models.[12] Their developers are moving away from the assumption that - according to the laws of scaling - the usefulness of the models depends on scaling their parameters, and therefore also on the use of ever larger data sets and computing power.

The distinction between large and small models is not precise. Small models include Bert models with a few hundred million parameters each, but also models with several billion parameters, such as the popular open model Mistral 7B. Small models can be one of several versions in a model family that also includes large models - for example, the largest of the Mistral models has 123 billion parameters. For us, however, most interesting are those models that are created independently, as an alternative to large language models. Admittedly, they are often built on the basis of existing large models.

In the past year, numerous small models have been developed, typically with sizes of 7B, 2B or smaller.[13] The groundbreaking work by Microsoft researchers on the Phi family of small models has been a crucial first step. In an article titled "Textbooks Are All You Need," the research team that developed Phi presented a methodology built on the assumption that high-quality datasets - which are "like textbooks" for language models - play a key role in model training.[14] According to Arvind Narayanan and Sayash Kapoor, there is a visible reversal of the market trend to work with the laws of scaling computing power and datasets, to build ever larger models.[15] Proponents of this approach point out that small models can achieve relatively high performance with fewer resources and greater energy efficiency.

---

[10] "Introducing The World's Largest Open Multilingual Language Model: BLOOM", BigScience, https://bigscience.huggingface.co/blog/bloom.

[11] Agam Shah, "Training of 1-Trillion Parameter Scientific AI Begins", HPC Wire (13.11.2023), https://www.hpcwire.com/2023/11/13/training-of-1-trillion-parameter-scientific-ai-begins/.

[12] Nagesh Mashette, "Small Language Models (SLMs). The Rise of Small Language Models: Efficiency and Customization for AI", Medium (12.12.2023), https://medium.com/@nageshmashette32/small-language-models-slms-305597c9edf2.

[13] Ethan Mollick (op. cit.)

[14] Suriya Gunasekar, et al. "Textbooks Are All You Need", arXiv (2023), https://arxiv.org/abs/2306.11644.

[15] Arvind Narayanan and Sayash Kapoor, "AI scaling myths", AI Snake Oil (27.06.2024), https://www.aisnakeoil.com/p/ai-scaling-myths.

Many of these small models are open source. The development of generative artificial intelligence systems relies heavily on openly available solutions that are the cornerstones of the development of these technologies, such as the PyTorch and TensorFlow programming libraries for machine learning, the Transformer methodology used in all language models, but also a variety of other development tools, databases, and other model components.

Historically looking, the first LLMs, such as GPT-2, the first model of the GPT family made available by OpenAI, were open models.[16] Yet the largest, foundation LLMs that came next, and were developed in recent years, are almost all closed: accessible only through controlled, commercial APIs. There are a few exceptions to this rule, in particular Llama from Meta is a family of models that are shared as open weights models (although they are not fully open. Another example is the Falcon model, created by the Technology Innovation Institute in the United Arab Emirates.[17]

These models can be used freely. A distinction should be made between fully open models and open weights models, for which only parameters are available. In both cases, an important advantage related to the availability of parameters is the possibility of fine-tuning: creating a new model on the basis of an existing one.This reduces the need for computing power, since the stage of pre-training the model is skipped.

These two factors - the changing paradigm of LLM development and the open sharing of core models - have enabled the emergence of an ecosystem of LLMs, built on openly available systems and components. The availability of open model architectures with which to models can be trained plays here a crucial role. These are primarily the Mistral and Llama models with 7B parameters. Equally important is the availability of open foundation models - primarily from the Llama family developed by Meta. As a result, the ecosystem of open AI systems includes open foundation models, fine-tuned large models, and small models - both trained from scratch on open architectures and fine-tuned.

Issues related to the representation of languages in LLMs is a third key factor that determines the development of new models. LLMs and other technologies for natural language processing (NLP) have not been developed uniformly for various languages. The development of tools such as language models has been focused on a handful of languages such as English and French, German, and Chinese. The vast majority of the 400 languages that are spoken by more than a million people worldwide, does not have datasets on which to build language models. This problem was identified in 2020 as a key challenge giving rise to inequalities in access to new language technologies.[18] Most of the world's languages are considered "low resource languages"

---

[16] OpenAI, "GPT-2", GitHub, https://github.com/openai/gpt-2.

[17] These models, although open, are not considered open source models, due to the fact that they are licensed with additional use restrictions. See: Stefano Maffulli, "Meta's LLaMa 2 license is not Open Source", Open Source Initiative (20.07.2023), https://opensource.org/blog/metas-llama-2-license-is-not-open-source

[18] Angela Fan et al., "Beyond English-Centric Multilingual Machine Translation", arXiv (2020), https://arxiv.org/abs/2010.11125

by AI companies: languages with little available data, making it difficult to train models.[19] According to researchers from Cohere this lack of data, combined with lack of computing power can result in new forms of digital exclusion.[20]

Taken together, these three factors help us understand why in the past few years many LLM-building initiatives have emerged around the world. First, these are often local models, not only supporting specific languages considered "low resource" by major commercial model developers, but also tailored to local cultural needs. Second, these models are open and set a high standard of transparency with regard to training data and its provenance. Third, they use architectures and methodologies for creating small language models to cope with the challenge of limited computing power and financial resources. Some of the initiatives are also a manifestation of a trend dubbed "sovereign AI" by Nvidia: aimed at creating national models, based on own. “sovereign” data and infrastructure.[21]

These new initiatives include Sweden's GPT-SW3, Singapore's SeaLion, France's Common Corpus initiative and Albert model, and the Polish projects described in this report. Worth mentioning are also research models, such as Olmo and Molmo models created by the Allen Institute for AI, or the AuroraGPT model being developed at the National Argonne Lab.

At the moment, there is a lack of agreement among experts and researchers on whether small models can effectively compete with large basic models. The rapid pace of technology development makes it difficult to capture the trend. Developers of small models assume that they are increasingly able to compete with large ones: whether through new methods of model development or by building on existing, open models.

[19] “The AI language gap”, Cohere for AI, 27 czerwca 2024, https://cohere.com/research/papers/policy-primer-the-ai-language-gap-2024-06-27.

[20] “Introducing Aya: An Open Science Initiative to Accelerate Multilingual AI Progress”, Cohere for AI, 5 czerwca 2023, https://cohere.com/blog/aya-multilingual.

[21] Angie Lee, “What Is Sovereign AI?”, Nvidia, 28 lutego 2024, https://blogs.nvidia.com/blog/what-is-sovereign-ai/.

POLAND'S OPEN AI ECOSYSTEM

In the short period between 2022 and now, an ecosystem for native language model development has emerged in Poland. Of course, many experts and organizations were involved in machine learning, natural language processing and even the development of simple language models years earlier, but only in recent years, a few important shifts occurred: their work became visible beyond the narrow communities of AI researchers, and their work stopped being treated as just basic research.

The importance of such initiatives was recognized in the "Policy for the Development of Artificial Intelligence in Poland from 2020," adopted in 2020 by the Polish Ministry of Digital Affairs.[22] In this strategic document, the development of IT solutions tailored to specific challenges that Poland faces was identified as one of the main short-term goals. This included the development of technologies for machine-based processing of the Polish language. “Supporting projects that make architectures and trained models and training datasets available for widespread use” is one of the specific actions listed.

2022 was a breakthrough year for generative AI, as it saw the release of GPT3-based ChatGPT by OpenAI, but also open alternatives like BLOOM. Neither of the these models were trained to generate Polish texts. This encouraged Sebastian Kondracki, a Polish programmer and open source evangelist, to try to create a training dataset. He assumed that with a terabyte of text data, a Polish language LLM could be trained. As a result, the grassroots initiative SpeakLeash was born in mid-2022.[23]

A year later, in August 2023, a group of Polish machine learning experts published an article titled “On borrowing other worlds, or why we need a Polish LLM.”[24] Among the authors are both people associated with the SpeakLeash community and representatives of public research institutions, which later in 2023 would form the Polish Large Language Model (PLLuM) consortium. The article is a manifesto of the Polish generative model developers. They present the vision of a Polish Big Science program: "We are united by one goal: to create a large Polish language model that will be open, accessible and transparent." The document lists a number of further principles guiding the creators of the Polish AI ecosystem:

- **Open source model development and open sharing of created solutions:** developers of Polish models use many existing open tools and solutions that are useful in creating new models. They emphasize the usefulness of data that is made available in as open and transparent ways as possible. They themselves commit to openly releasing AI systems and components that they develop. Transparency is identified as a key principle

[22] "Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020", Ministerstwo Cyfryzacji, https://www.gov.pl/web/ai/polityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020

[23] The name is a play on words, based on the phonetic sound of the word Spichlerz, and creates an analogy between a data set and a granary - a collection of grains from which bread can be made.

[24] Michał Dulemba et al., "O pożyczaniu innych światów, czyli po co nam polski LLM", 11 August 2023, https://www.gov.pl/web/ai/o-pozyczaniu-innych-swiatow-czyli-po-co-nam-polski-llm

associated with openness, as it ensures both control over AI systems and collaboration in model development;

- **Adaptation to national and local contexts:** models created by Polish development teams have the advantage of being better adapted to Polish realities and cultural contexts, thanks to the use of Polish text databases and then through model fine-tuning. At the same time, due to their openness, the models can be further fine-tuned, and thus adapted to the needs of specific communities or organizations. Existence of multiple models will guarantee a plurality of perspectives, and offer a solution to the problem of bias in LLMs;

- **Collaboration and acquisition of skills:** work on Polish language models results not just with the creation of new AI technologies and their components. It also establishes a community of researchers and technologists that have the competences needed to create and deploy LLMs (and AI technologies more broadly). Just as important is open collaboration between various entities that have access to necessary resources: texts and data, computing power, and people with necessary expertise.

The subject of our analysis are the two major LLM development initiatives that emerged in the last two years, and that bring to life the vision of a Polish Big Science program. We devote most attention to the SpeakLeash project, and the Bielik family of models that it has released in 2024. We also describe the work of the PLLuM (Polish Large Language Model) consortium. However, this research consortium has not yet made their LLM public.

These initiatives create LLMs in two different modes of technology development. SpeakLeash is a grassroots initiative that meets the definition of a Digital Commons.[25] PLLUM, on the other hand, is a publicly funded consortium of major research institutions. One of SpeakLeash's creators, Sebastian Kondracki, often refers to Eric S. Raymond's metaphor of the cathedral and the bazaar[26] to describe how the two initiatives are different, but also complimentary. As we will show, despite differences in approach, the two projects have much in common: the premise of making models openly available, similar strategic goals for Polish language models, and similar challenges that both projects face.

There are also several other initiatives operating within the Polish LLM ecosystem, which are not the subject of our analysis. In March 2024, Gdańsk University of Technology and the Information Processing Center (Ośrodek Przetwarzania Informacji - OPI) created the Ora model, by tuning the Llama 2 model on a corpus of Polish language data. Another, similar example is the Trurl model, created by VoiceLab.ai.

The leaders of SpeakLeash and PLLuM initiatives often emphasize that they are not just concerned with creating LLM models – they are building an ecosystem that is as important as their individual projects. What does this mean? LLM developers work not only on models, but

[25] Jan Krewer and Zuzanna Warso, op. cit.

[26] Eric S. Raymond, "The Cathedral and the Bazaar", http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/

also on other components of AI systems: training and instruction databases, tools such as tokenizers, data cleaning procedures or training methodologies. It also implies an approach based on sharing knowledge and competencies among initiatives that could potentially compete with each other. Finally, the ecosystem approach implies that the goal is to build a broader socio-economic environment capable of deploying and using Polish LLMs.

In the following sections, we start with an overview of the reasons for which Polish LLMs are developed, followed by a closer look at the two key initiatives: SpeakLeash and PLLuM. In the next section, we'll take a more in-depth look at the various aspects of model development, and the various resources required to do so.

WHY IS IT WORTH CREATING POLISH LANGUAGE MODELS?

We asked developers of Polish language models about the reasons for which they are developing these models, and about the advantages of having such local models. Based on their responses, the main arguments can be divided into three themes: cultural context, strategic advantages, and adoption by business and public administration.

Cultural context

This theme includes highlights the advantage of models that are better adapted to Polish reality. This is a key argument raised by Polish LLM developers. They are motivated by the failure to include the Polish language in globally dominant models, by the poor quality of Polish texts generated by models that seemingly support Polish language, and the lack of consideration of the local cultural context. The main issues at stake are:

- **Understanding metaphors and cultural references:** models trained on a corpus of Polish texts, including Polish literature, will better understand Polish culture, including literary references, geographical knowledge, Polish recipes or colloquial language;
- **Preservation of dialects:** an interesting application of language models is the preservation of dialects - one developer described the language model as a cultural "retirement home" for them;
- **Local and specialized contexts:** through fine-tuning, open models can be adapted to the needs of specific communities, or organizations. The existence of diverse models ensures a plurality of perspectives.

Strategic advantages

This theme considers advantages related to the control of data and the cost of deploying language models. It is also related to considerations of digital sovereignty, that is, of setting and enforcing rules under which the internet and other digital technologies operate. The main issues are:

- **Control of costs:** since the strategy of technology giants is to monopolize the market, expenses related to using their solutions should not be ruled out, despite current low prices. Having homegrown models reduces this risk, especially when they are openly available;
- **Control over where and how data is processed:** this is especially important for businesses and organizations protecting their intellectual property - but can also serve to protect user privacy, and various types of sensitive data. Deploying open models on your own infrastructure and fine-tuning them serves these goals as well;

- **Building technological know-how and capacities:** people and organizations working at all stages of model development acquire unique competencies that then spread in the labor market. This applies to both highly skilled model developers and those gaining experience with model deployment in various contexts.

Adoption by business and public administration

This theme is concerned with how local language models can affect the landscape of the Polish economy, particularly in the small and medium-sized business sector, and also in the public administration (SpeakLeash members focus on the former, while PLLuM representatives on the latter issue). Models can be adapted to the needs of a particular business through fine-tuning. Similar arguments apply to implementations in public administration. Issues raised include:

- **Faster and cheaper processing of specific problems:** enterprise-specific models can perform business tasks much faster and with less computing power than models used for general applications;

- **Opportunity for self-regulation and collaboration:** local model management enables a partnership between model developers and businesses that own the data, on which the models can be trained. Collaborating in a way that is ethical and safeguards the interests of smaller organizations is far easier at the local level. Global corporations often operate under the motto "move fast, break things,", acquiring whatever data they can. However, there are examples of agreements between large companies offering LLMs and organizations like Axel Springer[27] or Reddit.[28]

[27] "Axel Springer and OpenAI Partner to Deepen Beneficial Use of AI in Journalism", Axel Springer, 13 grudnia 2023, https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism.

[28] "OpenAI and Reddit Partnership. We're bringing Reddit's content to ChatGPT and our products", OpenAI, 16 maja 2024, https://openai.com/index/openai-and-reddit-partnership/.

SPEAKLEASH

SpeakLeash began as an informal community of developers that aims to create datasets, tools and methodologies for training language models that work in Polish. The main space for communication and collaboration is the community on Discord, with over 1,000 users (and growing). Recently, the SpeakLeash Foundation was also registered. About 10% of SpeakLeash participants have engaged directly in model development.

Bielik is a family of models created by a handful of developers using a few hundred GPUs, based on data collected by a grassroots community working in a peer production model, with a minimal budget. Its creators themselves contrast Bielik and SpeakLeash with initiatives based on large research teams with access to thousands of times more resources. To our knowledge, this is a globally unique example of successful development of a model that functions as a Digital Commons, comparable to the earlier work on LLMs by the Eleuther.ai community.[29]

SpeakLeash is a project fully based on the principles of open source software and open science. One of the key goals is to create artificial intelligence systems and tools tailored to Polish cultural specificity. The initiative's founders describe the goal using practical examples: a Polish model should think of vacationing in the Masuria region of Poland, and not in Hawaii. It should know the recipe for the popular Polish soup *żurek,* and understand that Janusz is not just a male name, but also a slang term for a certain type of person.

Just as important is the underlying vision of spreading AI technologies, and the capacities to build and to use them, in Poland. The initiative assumes that through the creation of small models and the use of fine-tuning mechanisms, solutions tailored not only to the specifics of Poland as a country, but also to the needs of local communities, specific organizations and even individuals will be created.

In the report, we highlight the role of collaboration and the communal aspects of the SpeakLeash initiative. However, it is also necessary to state the key role of individuals, and in particular that of Sebastian Kondracki, who initiated SpeakLeash in 2022 and continues to be one of its leaders and evangelists. As he tells the story, he had already been professionally involved in implementing machine learning-based solutions for banking and e-commerce a few years earlier. So he was familiar with issues such as data security and the advantages of open, compact models. When he heard about the BLOOM model in the summer of 2022, he figured it was possible to create a similar community working on Polish solutions. Central to the development of the Polish AI ecosystem was a strong commitment to the ethos of open source and open science, shared by Kondracki and other SpeakLeash founders.

Sebastian Kondracki initially reached out to Hugging Face and Eleuther.ai, and obtained from them basic know-how on how to create language models. Eleuther.ai members promised that if he could collect a terabyte of data, they would provide computing power and support to create a Polish model, based on their GPT NeoX model. Within a year, a number of experts with

[29] "The View from 30,000 Feet: Preface to the Second EleutherAI Retrospective", Eleuther.ai (2 March 2023), https://blog.eleuther.ai/year-two-preface/.

experience in building and working with models joined the project. New open models such as Llama and Mistral, as well as the work conducted by the HuggingFace platform, have guided SpeakLeash's approach to the development and finetuning of LLMs.

In addition to the Polish language corpus, the SpeakLeash community has been developing a variety of tools: including their own PDF OCR tool, a tokenizer tailored to the Polish language, model benchmarking tools - including a leaderboard for Polish-language models, and documentation for creating language models. At the same time, in many cases they use existing, open-source solutions - significantly reducing costs and speeding up work. It's not just that the Polish model, like all other LLMs, is built on the Transformer machine learning architecture, made openly available by Google. Many other components are being used, including the Mistral model architecture, training and instruction databases, and benchmarks to measure the quality of the model.

A breakthrough for the initiative came in early 2024, when it partnered with the Cyfronet AGH supercomputing center. Thanks to this collaboration, the Bielik language model, trained on the SpeakLeash database, was released publicly in April. This was the first Polish model whose training entailed both the creation of a base model (like the Qra model before it) and then a fine-tuned model (like Trurl before it, a model that was just fine-tuned on the Llama model). Like almost all open models in the world, Bielik was not trained from scratch - although it is a base model. The Mistral 7B model architecture, available under the Apache 2.0 open license, was used to initiate Bielik training.

The developers of Bielik used 18 million documents from the SpeakLeash collection, from which a training set of 22 billion tokens was created. Work on the quality of the collection included the removal of out-of-order or inappropriate text, anonymizing and improving file formatting. A classifier model was also used to select high-quality content. The collection was supplemented with English-language texts from the SlimPajama collection. In April 2024, the Speakleash initiative released the Bielik 7B v0.1 language model, and an instruction fine-tuned version, Bielik 7B-instruct v0.1.[30]

The ecosystem approach taken by SpeakLeash members means that the creation of the model was not considered the end product of the initiative. It was seen rather as just one step in creating an ecosystem that supports the creation of other models and jointly developed tools.[31] This is exemplified by the creation of a Polish "leaderboard" that compares the ability of different models to generate high-quality Polish text. The leaderboard was created by localizing the English-language EQ-Bench leaderboard to Polish conditions, and is an example of how the

---

[30] Nina Babis, "Bielik wylądował!", SpeakLeash | Spichlerz, 24 kwietnia 2024, https://www.speakleash.org/blog/bielik-wyladowal-24-04-2024/.
Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, i Remigiusz Kinas, "Bielik 7B v0.1: A Polish Language Model -- Development, Insights, and Evaluation", arXiv, 24 października 2024, https://doi.org/10.48550/arXiv.2410.18565.

[31] "Sukces Bielika dodał skrzydeł Spichlerzowi: z Sebastianem Kondrackim (SpeakLeash) rozmawiamy o powstaniu, znaczeniu i rozwoju polskiej AI", Mam Startup, 12 kwietnia 2024, https://mamstartup.pl/sukces-bielika-dodal-skrzydel-spichlerzowi-z-sebastianem-kondrackim-speakleash-rozmawiamy-o-powstaniu-znaczeniu-i-rozwoju-polskiej-ai/.

use of open tools accelerates work on localized models. The tool, created before Bielik was made public, allowed the new model to be tested against existing models.

Further development of the Bielik model family has continued since April. In August, version two of the model was made public. This is a more fine-tuned version, with the original Bielik model being used to filter and remove weak data from the training database. A new instruction fine-tuned version of the model was also released. A third version, tuned on a larger set of instructions and dialogs, is currently under development. All models in the Bielik family are available under the Apache 2.0 license.

There are now also first examples of creating new fine-tuned models based on Bielik. TheLion.ai, a Polish health AI startup, announced in August that they will be tuning a small medical language model based on the Bielik v2 model.[32] SpeakLeash leaders are also aware that further development of their models will require building new relationships, in particular those that will give them access to new training datasets. Equally important will be partnerships for implementing Bielik models in various contexts and organizations.

[32] Aleksander Obuchowski, "Niedługo Ruszamy w TheLion.AI z Treningiem Polskiego Medycznego Modelu (...)", LinkedIn, dostęp 7 listopada 2024, https://pl.linkedin.com/posts/aleksander-obuchowski_nied%C5%82ugo-ruszamy-w-thelionai-z-treningiem-activity-7231262979000852481-lTON.

POLISH LARGE LANGUAGE MODEL (PLLUM) CONSORTIUM

PLLuM is a consortium of research institutions founded in November 2023 to work together on the development of a Polish large language model. Its members are: Wroclaw University of Science and Technology (project leader), NASK research institute, the Institute of Computer Science of the Polish Academy of Sciences, the Institute of Slavic Studies of the Polish Academy of Sciences, the National Information Processing Institute and the University of Lodz. Most of the organizations in the consortium have experience working with LLM models. The researchers involved have been conducting evaluation studies of the GPT model for years, and OPI researchers previously created the Qra base model.

The consortium has received PLN 14.5 million in funding under a targeted grant from the Ministry of Digitization for basic research in artificial intelligence.[33] The PLLUM consortium currently has only one year's funding, until the end of 2024. At the same time, the project's managers emphasize that it is part of a longer-term strategy of consortium members to create language models and databases, based also on their own computing infrastructure.

Most of the PLLuM members are also active in the CLARIN-PL consortium, which has been in existence for more than a decade and is part of CLARIN, a European consortium developing solutions for working on large linguistic datasets.[34] CLARIN funding has provided much of the necessary research infrastructure and computing power, being built since 2018. In addition, Wroclaw University of Technology has allocated approximately 40 million Euro from other grants to purchase 300 Nvidia H100 GPUs.

The basis of PLLuM development is a large collection of text data in Polish, including a linguistic corpus created by IPI PAN, collections created within CLARIN-PL (such as Słowosieć, the Polish language wordnet), various open resources and data from web scraping. The consortium is also signing agreements with various content owners to train the model on their data as well. As is the case with Bielik, the creation of a high-quality instructions dataset is an important part of the PLLuM project. While the developers of Bielik largely use translated English language datasets or synthetic ones, PLLuM has invested resources in creating a new instruction dataset, adapted to the Polish language and its cultural context.

On its basis, a family of models is being created on the basis of the Mistral Nemo 7B architecture: models of 7x8B, 7x22B and 70B parameters are planned, together with a virtual assistant tailored to the needs of public institutions. The models created by the PLLUM project is expected to be available in late 2024. PLLuM is intended to be an alternative to the commercial models already available, which are closed, expensive to use and unsuited to the Polish

[33] "Dot. pisma z 7 grudnia 2023 r. Pani Poseł Pauliny Matysiak w sprawie pilnego uregulowania kwestii prawnych dotyczących używania sztucznej inteligencji (interpelacja nr 4)", Ministerstwo Cyfryzacji, 19 stycznia 2024, https://orka2.sejm.gov.pl/INT10.nsf/klucz/ATTCZRK26/%24FILE/i00004-o1.pdf.

[34] Gontarz, Andrzej, "Własne czatowi dać słowo", Polskie Towarzystwo Informatyczne, dostęp 7 listopada 2024, https://portal.pti.org.pl/wp-content/uploads/2024/06/3_Wlasne-czatowi-dac-slowo.pdf.

context.[35] The project is thus based on similar principles to SpeakLeash. The openness of the model itself, and other key components, is a key commitment made by the consortium. PLLUM's creators also want to ensure that their AI system complies as much as possible with the requirements of the Artificial Intelligence Act.

[35] "Polish version of ChatGPT. PLLuM is to be better and completely free to use", Research in Poland, 12 grudnia 2023, https://researchinpoland.org/news/polish-version-of-chatgpt/.

HOW IS A LANGUAGE MODEL CREATED?

The process of training a language model requires several key resources:

- data: training datasets, along with methodologies for selecting, cleaning and structuring the data;
- computing power: clusters of graphic processing units (GPUs) and cloud resources required to use the model for inference;
- model training architecture and methodology: defines the structure of the model and how it processes information and generates results;
- instruction datasets: special sets of dialogs and instructions used for model fine-tuning;
- community: the community of people creating, tuning and using the model.

Local language models are created through a process of initial pre-training on language datasets, followed by fine-tuning, using additional data – including instruction datasets.

The initial step, pre-training, focuses on training the model to understand word sequence statistics. This process involves providing the model with a large set of texts so that it gains the capacity to predict next words in a word sequence.

The next step, fine-tuning, involves refining the model's understanding of the specific language and cultural context. This step is key to improving the model's fluency, accuracy and cultural relevance. Fine-tuning helps the model better understand natural language and eliminate knowledge gaps. At the same time, it is a process that intentionally reduces the breadth and depth of the knowledge being described.

In this section, we take a closer look at these key resources.

Training data

An adequate, sufficiently comprehensive and high-quality training dataset is crucial to the creation of an LLM. Much of the work on Polish LLM models has therefore been focused on the creation of training datasets.

The starting point is always the same - selected web crawled data, and data available under open licenses. Typically, the Common Crawl database or its subset is used – although Polish developers have crawled the Polish-language web themselves.

LLM developers try to supplement this web crawled data with additional resources - especially since web data is often of low quality. The SpeakLeash project focused for the first two years not on training the model, but on crowdsourcing a large-enough text dataset. In contrast, the PLLuM project has access to scientific databases and language corpora created and managed over the years by consortium members.

LLM developers aim then to supplement their datasets with additional data, from from sources that are not publicly available. These include, for example, official materials of public institutions, content shared by publishing houses, or media archives.

As Jan Kocoń, one of the initiators of PLLuM, stated, "We have access to immense amount of data, but we need more, because the Polish language poses many difficulties. This is especially true of content specific to the Polish cultural and social context."[36] One of the experts we spoke with says that even if one had access to all digital content that was ever created in Polish - including resources that are not publicly available - it would not be enough to train a large model capable of competing with the largest models available on the market. Thus, the overall size of Polish cultural resources becomes a limiting factor for the development of Polish LLMs. It limits the development to small and medium-sized models, but also encourages innovation in model development.

SpeakLeash has collected to date a dataset estimated at 15-20 terabytes of raw data, 1.5 terabytes of cleaned data and about 180 billion tokens. The dataset is built out of collections of various types of data. The first collection was allegedly the Polish Wikipedia, and today the database contains a wide range of data types. SpeakLeash's founders modeled the design of their dataset on two examples. first was the Pile, a dataset containing mainly web scraped content, combined with selected structured datasets. The second example was BigScience Data, a collection created for the Bloom model, whose elements were precisely cataloged - but it was a much smaller collection. Inspired by these examples, SpeakLeash's developers assumed that the data they are collecting needs to be accurately documented, including licensing information and text characteristics. The dataset would also be open in a technical sense: easy to use for anyone with a basic knowledge of Python.

SpeakLeash's resources are largely based on web crawl data, but it has been recognized that databases such as Common Crawl do not provide high enough data quality - with problems largely due to inappropriate use of such datasets.[37] SpeakLeash developers decided to run their own web crawling processes, followed by data cleaning and classification processes. The starting point was the creation of a web indexing tool for the Polish-language Internet. According to SpeakLeash members, they have managed to access all of the Polish-language resources available on the public Web. As a result, SpeakLeash's database consists of many collections of content from the Web, which have been cleaned and, in many cases, categorized and compiled into thematic collections. The database contains in particular content of various Polish online forums, including the largest ones, such as Kafeteria.pl, Gazeta.pl, or Wizaz.pl. These are treated by LLM developers as particularly valuable for AI training. Content whose owners do not agree to use for LLM training (through a robots.txt file or some other form of disclaimer) has been removed from the collection.

---

[36] Ładan, Anna. "Czy AI może być etyczna, empatyczna i rozumieć kontekst kulturowy?", Computer World, 14 lipca 2024, https://www.computerworld.pl/article/2509813/czy-ai-moze-byc-etyczna-empatyczna-i-rozumiec-kontekst-kulturowy.html.

[37] Baack, Stefan, "Training Data for the Price of a Sandwich. Common Crawl's Impact on Generative AI", 6 lutego 2024, https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

At the same time, the SpeakLeash database also aggregates various existing collections. These include, among others: the resources of the Polish national digital library Polona, the documents of the Polish Parliament, various databases of legal texts, the collection of Public Domain books Wolne Lektury, and the text resources of Europeana. The project also makes use of existing international datasets, some of which contain resources in Polish, such as the OpenSubtitles movie subtitles corpus. A number of specialized tools and methodologies for working with the data have been developed in the course of creating this database.

One of our interviewees, who has a key role in the process of building the SpeakLeash database, says that using data from the public Internet is a necessity, but also a burden, due to its unclear legal status. Ultimately, SpeakLeash developers would like to rely more on additional data sources, such as those from libraries, archives or publishers. These have the advantage of clear legal status and, in many cases, higher quality than online resources. The optimal scenario would be the availability of more openly licensed texts.

The use of open data is also an important premise of the PLLuM project, which is creating a dataset containing all openly licensed Polish-language resources. This is a collection containing 8 billion words, five times the size of the national language corpus - but experiments with training a model solely on this data have shown that this size is insufficient. At the same time, an important part of the project is the acquisition of licensed data, provided by a variety of entities, both public and private. To date, PLLuM has signed dozens of agreements to use content owned by such entities, including public administration institutions or publishers, to train the model. In most cases, these texts will not be licensed openly, and will only be made available to the consortium during the training process. However, even the combined sets of open and licensed data are insufficient - so the PLLuM model is also being trained on web crawl data.

Establishing partnerships with entities that have a variety of language content and resources is a key next step for Polish LLM developers. On the one hand, the PLLuM project in particular is successfully signing agreements for the use of various private collections. For example, it recently signed an agreement with Agora Group, a major Polish media group, which made it's data available for free to train the model.[38] On the other hand, the first tensions between different interest groups are emerging. For example, in March the PLLuM consortium invited Polish publishers to share content for the purpose of model training. In response, the Polish Chamber of Press Publishers published a position paper in June, stating that the proposal for cooperation did not address the issue of obtaining a license to use the content, nor appropriate remuneration The Chamber said that the transfer of content meant "the risk of losing control over press materials." The publishers challenged the application of copyright exceptions to generative AI training, and urged the consortium to sign paid licensing agreements.[39] The PLLuM consortium's experience shows that it is ultimately possible to reach an agreement with

[38] "Agora udostępnia treści do prac nad polskim modelem językowym", Wirtualne Media, 13 listopada 2024, https://www.wirtualnemedia.pl/artykul/agora-udostepnia-tresci-do-prac-nad-polskim-modelem-jezykowym

[39] "IWP ostrzega wydawców przed polskim modelem do AI. Sugeruje umowy licencyjne", Wirtualne Media, 6 czerwca 2024, https://www.wirtualnemedia.pl/artykul/iwp-ostrzega-wydawcow-przed-polskim-modelem-do-ai-sugeruje-umowy-licencyjne.

rights holders in this area and obtain free licenses to use the resources to train the language model.

Computing power

Access to computing power is a necessary requirement for training a model on collected data. Large computing resources are needed in particular for pre-training a base model. Computing power requirements are much lower for models that are fine-tuned on an existing foundation model.

According to our interviewees, the practice of creating local models shows that it is not necessary to have access to computing power at the level of those used by the largest AI companies. Innovations in the methodology of model creation make it possible to significantly reduce the necessary computing power - to a level that can be provided by supercomputing centers operating in Poland.

Poland is a country where public research institutions have today a total of several hundred state-of-the-art GPUs needed to train AI. By comparison, Meta estimates that by the end of 2024 it will have the computing power of 600,000 H100 processors.[40] Nonetheless, these resources proved sufficient to create the first polish language models. An important point of reference was the BigScience project and the BLOOM model it created, which was trained on the French supercomputer Jean Zay, using less than 500 Nvidia A100 GPUs.[41]

In early 2024, a Polish software development company Azurro created the APT-1B small model on selected data from SpeakLeash, using a single GPU. The experiment proved that it was possible to create the model with the help of Polish language data and with minimal computing power - although the model created was not of high quality.

In both projects that we analyzed, partners with adequate computing power play a key role. In both cases, these are supercomputing centers run by public research institutions. The Wrocław Network and Supercomputing Center at the Wrocław University of Science and Technology plays this role in the PLLuM consortium . In the case of SpeakLeash, the Bielik model was developed thanks to cooperation with the Cyfronet supercomputing center of the AGH University of Kraków, which provided 256 Nvidia GH200 processors. Training of the Bielik model took place as test runs of this new processor cluster. Leaders of the SpeakLeash initiative emphasize the unique nature of the collaboration between the public supercomputing center and an informal, peer production AI development project.

Representatives of both projects emphasize that it is a challenge for them to have access to adequate computing power, available for an extended period of time. In the case of Bielik - despite Cyfronet's support - the model could be better trained with access to more power. And

---

[40] Engineering at Meta, "Building Meta's GenAI Infrastructure", 12 marca 2024, https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/.

[41] BigScience, "Which Hardware to Train a 176B Parameters Model?", BigScience, dostęp 7 listopada 2024, https://bigscience.huggingface.co/blog/which-hardware-to-train-a-176b-parameters-model

for the PLLuM project, on the other hand, the challenge is time, as the funding was made available only for 12 months. At the same time, this constraint in both cases drives innovation - one interviewee said that the real challenge is to find ways to effectively train high-quality models with limited computing resources.

As one interviewee stated, if AI models are like bread, then supercomputing centers are the leaven. This is because they are the only actors with computing capacity - the alternative is to buy access to commercial cloud infrastructure. These centers also treat work on AI models as an opportunity to create state-of-the-art research environments. The Qra model is one such example: it was created at the Gdansk University of Technology, which for the project used its STOS Competence Center and the Kraken supercomputer, including a cluster of 21NVidia A100 graphics cards.[42]

Creating and tuning the model

Each effort to develop a new language model faces a choice: to either train a based model from the start, or finetune an existing one. On one hand, only a handful of initiatives have adequate data sets and access to computing power to create pre-trained models. However, they are necessary to fill existing language gaps. Thus, both the SpeakLeash initiative and the PLLuM consortium have undertaken to create such a model for the Polish language.

These models are not created entirely from scratch - both projects use the Mistral 7B architecture, openly available under the Apache 2.0 license. The interviewees emphasize the role of existing open source solutions not just in democratizing access to LLM, but also in enabling further innovation. In the case of Bielik, the model was trained more efficiently thanks to the use of ALLaMo, an innovative training methodology created by Krzysztof Ociepa. Open sharing of datasets, code or methodologies encourages broader participation and supports further innovation.

The example of the Bielik family of models shows that model training does not happen just once. It is an ongoing process, with successive versions being developed and shared over time. In the case of Bielik, these include the Bielik 7B 0.1 base and instructional models, the Bielik 11B 2.0 instructional model, and then three further alternative version 2.0 models, which differ in the brevity of their statements and their ability to play assigned roles.[43] Version 3.0 is also currently under development.

The issue of instruction sets used to finetune models is often not raised in discussions of training data - meanwhile, they are essential training resources, and ones that are quite a challenge to acquire. Instruction dataset are needed to create models that can result in actual use cases – for example, ones that are suitable for creating chatbots. Therefore, lack of a siutable

[42] "Qra. Naukowcy opracowali polskojęzyczne modele językowe", Forsal, 8 marca 2024, https://forsal.pl/lifestyle/technologie/artykuly/9453625,qra-naukowcy-opracowali-polskojezyczne-modele-jezykowe.html.

[43] Krzysztof Ociepa, "#bielik #bielik11b #bielikllm #polskillm #polskimodeljęzykowy #ai #llm (...)", dostęp 7 listopada 2024, https://pl.linkedin.com/posts/krzysztof-ociepa_bielik-bielik11b-bielikllm-activity-7239195666986479616-VicD.

instruction set in Polish is a major challenge. In the case of the first version of Bielik, half of the instructions came from the English-language OpenHermes-2.5 and orca-math-word-problems-200k collections, which were automatically translated into Polish. In addition, the Bielik community created a small collection of own instructions and dialogues. Finally, one million synthetic dialogues were also generated based on a selection of texts from the SpeakLeash collections. Bielik was tuned on a collection covering a total of 2.3 million instructions (700 million tokens). Second-generation Bielik models were tuned on an even larger set, consisting primarily of synthetic instructions generated with the help of the Mixtral 8x22B model. This yielded 16 million instructions (8 billion tokens).[44] The creation of these datasets is an ongoing process.

The Chat Arena PL platform is an important tool created to assess the quality of Polish models. The platform allows users to test two randomly selected models for the quality of answers given to the same prompt. In this way, evaluation data is crowdsourced, allowing the quality of Bielik and other models to be assessed in a different way than through a traditional benchmark. The project is one more example of reusing existing, open tools, as it is based on the open source Chat Arena solution. So far, over twelve thousand tests have been run on the platform. Unfortunately, while these dialogs are useful, there are too few of them to build an instructions dataset.

For now, little is known about insruction datasets that are being used by the PLLuM consortium. The project's developers aspire to use as many as possible instructions created specifically for the project, written by hired experts. They estimate that a complete collection should contain tens of millions of instructions and dialogs. The creation of such a set of instructions is seen as a possible next stage in the development of PLLuM models - one of the ideas under consideration is to generate them in a crowdsourced manner.

Fine-tuning is another way in which the development of open source models remains open-ended. In the case of Bielik models, there are already the first examples of such fine-tuning. This means additional training of a model on a large dataset, to create a qualitatively different model. Model fine-tuning can also be done by creating adaptation layers. In that case, the main model is already "frozen," and only an additional layer is created, using a small set of specialized data. Yet another increasingly popular method is the use of the Retrieval Augmented Generation (RAG) mechanism. This involves the creation of an additional, external knowledge base, which the model uses when generating a response to a query.

Tuning can be done, for example, on data from a specific company, data from a local government, or data from a specific area of knowledge. One of our interviewees gives an example of a model fine-tuned on specific data that is useful to local energy communities. He envisions these types of models as important tools in solving various social challenges and problems, also at local scale. Another ongoing example of fine-tuning is an initiative by theLion.ai, which intends to

[44] "Speakleash/Bielik-11B-v2.0-Instruct", Hugging Face, 26 października 2024, https://huggingface.co/speakleash/Bielik-11B-v2.0-Instruct.

fine-tune the Bielik model based on instructions that will result in a Polish medical language model.[45]

Open, fine-tuned models, can then be deployed on a company's or organization's own infrastructure, guaranteeing greater data security. Model developers also envision fine-tuning models to create personalized assistants for individual users. This involves technological solutions to minimize models so that they run on devices like laptops and smartphones.

Community and ecosystem

In conversations with experts, we heard that SpeakLeash is a community committed to developing a locally relevant and culturally aware Polish language model. The community emphasizes an open approach to model development, making its datasets and models available for others to use. This fosters a collaborative environmentm in which individuals and organizations can contribute to the development of Polish language models.

A key aspect of SpeakLeash's success is therefore its community-oriented nature. A symbol of this collaboration is the decentralized Discord platform, through which work is coordinateed in place of more formal, institutional mechanisms. The initiative has attracted more than 1,000 members, including experts and enthusiasts who contribute knowledge, insights, or even memes. The community behind SpeakLeash Bielik emerged organically from a shared desire to create a high-quality, culturally diverse model of the Polish language. The Bielik model itself was created by a much narrower group of developers and machine learning experts.

As we have already written, the creators of Bielik position their work in the broader context of not just the SpeakLeash community, but also about the Polish AI ecosystem. In public statements and interviews, other Polish models are treated not as competition, but as parallel initiatives, serving the same goals. This is based on an assumption that key Polish model development efforts will openly share scientific knowledge, data sets and the models themselves.

At the same time, it can be seen that at the moment the initiative is in a crucial phase of change, as ecosystem development no longer means just building a community of AI developers. Relationships with other entities are key for two reasons: they hold data valuable to SpeakLeash, and they could become users of the Bielik models, demonstrating its utility and succesful adaptation to the Polish context.

The PLLuM model is developed in a much more traditional institutional model, as it draws on the resources of the research institutions that make up the consortium. PLLuM's experience shows that such a model has advantages when it comes to establishing collaborations with other organizations, which tend to see the project as more trutworthy, due to its institutional form. At the same time, the consortium's managers signal that they are planning PLLuM's further development in the context of a broader ecosystem. This relates specifically to tasks, such as the

[45] Aleksander Obuchowski, "Niedługo Ruszamy w TheLion.AI z Treningiem Polskiego Medycznego Modelu (...)", LinkedIn, dostęp 7 listopada 2024, https://pl.linkedin.com/posts/aleksander-obuchowski_nied%C5%82ugo-ruszamy-w-thelionai-z-treningiem-activity-7231262979000852481-lTON.

development of an instruction dataset, for which it is necessary to build a broader network of partners, and a more participatory approach.

CHALLENGES WITH DEVELOPING LOCAL AI MODELS

Computing power as a barrier and motivator

Cloud computing services have made it possible to train and deploy large language models without the need for in-house data centers. At the same time, they make AI developers dependent on a small set of hyperscalers with adequate infrastructure. While training large language models requires significant computing power, the development of smaller models can be achieved with more modest resources. The limited availability of computing power is also driving innovation in more efficient architectures for training models.

Currently, there is a lack of clarity as to whether, in the long run, the advantage in computing power will determine the dominance of the largest model builders. Among our interviewees, there were both pessimists about the long-term success of the alternatives and believers in the trend leading toward smaller, efficient models.

Barriers: high-quality LLM training requires significant computing power, which translates into high energy consumption and requires robust computing infrastructure. This poses a challenge, especially for smaller entities, with limited resources. One interviewee pointed out the discrepancy in computing power available to all supercomputing centers in Poland compared to the resources of a single global AI company. He suggests that this difference in scale may, in the long run, hinder the development of locally trained LLMs that can compete with those produced by corporations.

Motivators: demand for computing power can drive both collaboration and innovation. The collaboration between SpeakLeash and Cyfronet is an example how infrastructural barriers can be overcome through partnerships and resource sharing. What's more, focusing on smaller, more specialized LLMs can reduce reliance on massive computing power. One interviewee argues that companies often don't need models that can do everything; instead, they need models tailored to specific tasks – and these can be small models. This approach emphasizes efficiency and finding a balance between model size, performance and computational costs of model training and inference.

Acquisition of good quality data

Access to high-quality data is one of the key challenges facing language model developers. Creators of model training datasets see their work as a step in the process of digital transformation: making more and more resources available in digital form, for machine processing. At the same time, there are clear barriers limiting access to many collections, especially those with high quality resources. Disparities in access to data are one of the factors giving an advantage to large companies that either own proprietary datasets or have the financial resources to pay licensing fees for new data sources.

Barriers: unlike large technology corporations, Polish model developers often have only limited resources available for dataset development. The situation is complicated by regulations that

make it difficult to access data, especially in sensitive areas such as medicine. Copyright issues are also a major factor: the legal status of model training is contested, and many owners of data expect that all uses be licensed and paid for.

Motivators: we've often heard that quality is more important than quality, and that much can bve attained with smaller detasets of hiqh-quality data. Finding ways to facilitate responsible access to data, especially for research and development, will be key to fostering innovation in artificial intelligence. There is a need for data management methods that balance the need for sufficiently large, diverse data sets with the need to address copyright concenrs, protect sensitive personal data, and ensure appropriate business models. Also model development would benefit if more collections were made openly available, through initiatives such as Open Access to academic publications or opening up heritage resources. Finally, datasets created for open models are likely to better meet the requirements of EU regulations such as the AI Act.

Regulations favoring existing monopolies

In our conversations, concerns were raised that regulations, while intended to ensure the ethical development of artificial intelligence, may inadvertently favor large technology companies that have the resources to navigate the complex regulatory framework. Smaller companies and research institutions may find it difficult to comply, which could hinder innovation and competition in model development.

Barriers: One speaker underscored the tension between strict data protection regulations, such as the GDPR, and the need for access to data, particularly in fields such as medicinem, where sensitive data is needed. While large corporations can afford to fund their own data, smaller players often have difficulty doing so. This suggests that finding ways to facilitate responsible access to data, especially for research and development purposes, could be key to fostering broader innovation in artificial intelligence. At the same time, the high cost of complying with regulations can limit innovation and competition in the AI sector, favoring existing monopolies.

Incentives: It is necessary to develop more flexible regulations that take into account the diverse needs and capabilities of different stakeholders, especially by supporting smaller and public interest players - providing support and incentives for smaller companies and research institutions to engage in artificial intelligence development. Such an approach could result in easier access to data, especially for R&D purposes. This approach has been adopted to some extent in the AI Act framework.

ABOUT

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture.

Fundacja Centrum Cyfrowe is a Polish think-and-do tank that cares about the social dimension of technology. The foundation's area of interest is the digital dimension of public life in Poland.

AUTHORS OF THIS REPORT

Alek Tarkowski is the Strategy Director at Open Future. He holds a PhD in sociology from the Polish Academy of Science. He has over 15 years of experience with public interest advocacy, movement building and research into the intersection of society, culture and digital technologies. His current interests include public interest AI policies and AI dataset governance.

Kuba Piwowar is a sociologist and cultural scientist who holds a PhD in cultural studies. He is also a Humanity in Action senior fellow, where he worked on a project focused on data literacy and data activism. Additionally, he serves as an assistant professor at the Department of Culture and Media at SWPS University in Warsaw. Between 2008 and 2024, he worked at Google, initially as an analyst and later as an advisor to key business partners.

Michał Owczarek is a PhD candidate in Cultural Sciences at SWPS University, where he investigates the history of media in Poland. He defended his master thesis in digital sociology on the topic of conflicts over digital infrastructures between states and platforms. He is also interested in urban studies, especially the impact of digital technologies on cities fabric.

Authors of the report would like to thank experts who old us about how they are building Polish language models: Paweł Cyrta, Adrian Gwoździej, Jan Kocoń, Sebastian Kondracki, Marek Kozłowski, Jacek Nagłowski and Maciej Piasecki.

We used a range of analytical approaches and tools to analyse the research data: from traditional content analysis methodologies to using existing LLMs (Gemini, ChatGPT, NotebookLM). We wanted to see to what extent the conclusions we come to on our own can be supplemented by what the models perceived in the data. While we are pleased with this human-machine collaboration, we hope that we will soon be able to perform similar tasks just as efficiently using Polish language models.