AI MÓWI PO POLSKU

przegląd rodzimych prac nad modelami językowymi

OPEN _FUTURE

CENTRUM CYFROWE

LISTOPAD 2024

WPROWADZENIE

Celem tego raportu jest przedstawienie case study polskiego ekosystemu tworzenia otwartych modeli AI dla języka polskiego. Są to małe modele językowe tworzone jako rozwiązania open source, tworzone w celu wypełnienia luki pozostawionej przez duże modele komercyjne, które nie są dostosowane do języka polskiego i polskiej specyfiki kulturowej. Prace nad tym modelami są przykładem skutecznego tworzenia alternatyw dla dominujących modeli.

Komercyjne modele fundacyjne są trenowane na ogromnych zbiorach danych, przy wykorzystaniu coraz większych mocy obliczeniowych, i w oparciu o wizję ciągłego skalowania technologii. Tworzenie dużych modeli językowych wymaga więc ogromnych nakładów finansowych, na które stać jedynie ogromne firmy, posiadające monopolistyczne pozycje na rynku cyfrowym.

Mogłoby się więc wydawać, że ekonomia tworzenia AI uniemożliwia powstanie alternatyw – czy to finansowanych ze środków publicznych, czy tworzonych przez mniejszych graczy komercyjnych. Jednak takie alternatywy powstają. W 2022, gdy OpenAI udostępniło komercyjny serwis ChatGPT, równocześnie został udostępniony otwarty model BLOOM, stworzony oddolnie przez społeczność wspieraną przez firmę Hugging Face.

Dzisiaj nowy paradygmat tworzenia małych modeli językowych oraz dostępność otwartych modeli podstawowych umożliwia tworzenia nowych modeli językowych – w szczególności takich, które adresują luki językowe w rozwoju generatywnego AI. Rosnąca liczba inicjatyw na całym świecie tworzy rozwiązania AI dostosowane do najróżniejszych języków i kontekstów kulturowych - oferując tym samym alternatywy dla dużych modeli komercyjnych.

Niniejszy raport koncentruje się na dwóch kluczowych projektach: budowie korpusu językowego SpeakLeash oraz stworzeniu na jego podstawie modelu Bielik oraz działaniach konsorcjum PLLuM (Polish Large Language Model), którego celem jest stworzenie dużego modelu językowego odpowiadającego specyfice języka polskiego.

Raport oparty jest na rozmowach z twórcami polskich modeli, na podstawie których prześledziliśmy proces ich powstawania, wyzwania, na które zwracają uwagę oraz wnioski, które udało się wypracować na podstawie dotychczasowych osiągnięć.

Kluczową motywacją w pracach nad polskimi modelami językowymi jest dostosowanie modeli do polskich realiów, dzięki czemu lepiej oddają one niuanse językowe i kulturowe. Istotnym argumentem jest również wspieranie lokalnego rozwoju technologii: prace nad modelami w Polsce wspomagają budowanie krajowych kompetencji w obszarze AI oraz wzmacniają pozycję Polski na arenie międzynarodowej. Ważnym argumentem jest też mozliwość większej kontroli nad przetwarzaniem danych – lokalne modele umożliwiają lepsze zabezpieczenie prywatności i większe bezpieczeństwo. Korzyści obejmują także aspekt finansowy – lokalne modele mogą być tańsze od oferty globalnych korporacji. Cechują się też mniejszym negatywnym wpływem na środowisko.

W raporcie opisujemy kolejne etapy tworzenia modeli, skupiając się na tym, jak pozyskiwane są kluczowe zasoby: moce obliczeniowe, wysokiej jakości dane i zespoły ekspertów o odpowiednich kompetencjach. Pokazujemy też, jak twórcy poszczególnych projektów myślą ekosystemowo, zakładając współpracę różnych organizacji i inicjatyw, oraz otwartą wymianę narzędzi i komponentów AI. Ważnym elementem prac nad polskimi modelami jest oddolny, wspólnotowy charakter działań: od zbierania i weryfikacji danych, przez trenowanie modeli i kontrolę ich jakości. Istotną rolę odgrywa również wykorzystanie istniejących otwartych technologii, przede wszystkich architektur trenowania modeli. Prace nad lokalnymi modelami językowymi odpowiadają także na potrzebę dostosowywania istniejących narzędzi, wrażliwości na lokalny kontekst i zrozumienia niuansów kulturowych i językowych. Opisujemy też jak twórcy modeli radzą sobie z wyzwaniami, związanymi z kwestiami prawnymi dotyczącymi wykorzystania danych treningowych, oraz nawiązywania współpracy w celu pozyskiwania nowych danych.

Celem raportu jest zwiększenie świadomości tego, że powstają otwarte modele językowe służące zredukowaniu luk językowych w rozwoju AI i zapewnienia rozwoju dla alternatywnych technologii. Wnioski z tych case studies mogą też być pomocne w formułowaniu polityk publicznych, wspierających rozwój takich alternatyw. Oto najważniejsze z nich:

Nasze kluczowe wnioski obejmują:

- Inicjatywa SpeakLeash jest rzadkim przykładem (obok projektów takich jak BigScience lub Eleuther.ai) udanego budowania modeli językowych przez społeczność, w modelu produkcji partnerskiej;

- Centra superkomputerowe przy publicznych ośrodkach badawczych mają wystarczającą moc obliczeniową, aby z powodzeniem trenować małe modele językowe, jednak niezbędne jest zapewnienie finansowania odpowiedniej ilości serii treningowych;

- Polscy twórcy otwartych modeli zarówno wykorzystują istniejące narzędzia i komponenty AI, aby zwiększyć efektywność swojej pracy, jak i tworzą własne narzędzia gdy jest to niezbędne - na przykład w celu uwzględnienia lokalnego kontekstu i potrzeb;

- Polscy twórcy modeli językowych pracują w dużej mierze na tych samych danych co inne projekty na świecie, pobranych z sieci. Jednak znajdują oni nowe sposoby na zapewnienie jakości danych i ich lokalnej trafności;

- Istnienie ekosystemu rozwoju AI oznacza, że tworzonych jest wiele modeli (zarówno poprzez wstępne szkolenie, jak i dostrajanie), zapewniając różnorodność rozwiązań i osadzonych w nich perspektyw społecznych i kulturowych;

- Polscy twórcy sztucznej inteligencji nawiązują udaną współpracę z różnymi podmiotami posiadającymi treści, aby zwiększyć pulę danych szkoleniowych w języku polskim..

TRENDY W ROZWOJU OTWARTYCH I MAŁYCH MODELI JĘZYKOWYCH

Large Language Model (LLM) to rodzaj systemów sztucznej inteligencji, które potrafią przetwarzać język naturalny i generować tekst. LLM są w tym celu trenowane poprzez analizę statystyczną wielkich zbiorów danych tekstowych. Najbardziej znane i LLM to modele komercyjne, a największe z nich są tworzone przez pięć globalnych firm. To modele GPT firmy OpenAI, modele Copilot z Microsoftu, modele Claude z Anthropic, modele Gemini z Google i stworzone przez Metę modele Llama. Te największe modele są często określane jako modele fundacyjne (foundation models), by podkreślić, że są to technologie ogólnego zastosowania. Jedni eksperci widzą w tym dowody na emergentne zachowania modeli, zdaniem innych to wynik uczenia na zbiorach różnorodnych danych.[1]

Modele fundacyjne są więc trenowane na ogromnych, i coraz to większych zbiorach danych, mierzonych objętością danych lub liczbą tokenów – krótkich zbitek tekstu. Dla przykładu, udostępniony w lutym 2023 model Llama 1 był trenowany na trylionie tokenów, Llama 2 (z tego samego roku) na dwóch trylionach, a udostępniona w sierpniu 2024 roku Llama 3 na 15 trylionach tokenów. Do szkolenia modeli są też potrzebne coraz większe moce obliczeniowe. Szacuje się, że wytrenowanie niedawno udostępnionego modelu Llama 3 wymagało mocy obliczeniowej równej 10^{24} FLOPS (floating point operations per second). Dla porównania, najszybszy polski superkomputer ma moc na poziomie 10^{15} FLOPS.[2]

Rozmiar wykorzystanych danych i mocy obliczeniowych przekłada się na ilość parametrów modelu, czyli liczbę czynników, które model uwzględnia przy przetwarzaniu i generowaniu treści. Pierwsze modele językowe miały miliony parametrów, a największe współczesne modele mają setki bilionów, lub nawet ponad trylion parametrów. Twórcy modeli podstawowych wychodzą z założenia, że lepsze są modele o większej liczbie parametrów.

Tworzenie dużych modeli językowych wymaga więc ogromnych nakładów finansowych, niezbędnych do zapewnienia przede wszystkim mocy obliczeniowej, ale też opłacenia wysoko wykwalifikowanych ekspertów i ekspertek, oraz pozyskania i obróbki danych. Koszt stworzenia dzisiaj modelu najnowszej generacji szacuje się na co najmniej 100 milionów dolarów – a koszt modeli kolejnych generacji może być wielokrotnie wyższy.[3] To powód, dla którego duże, podstawowe modele tworzy niewielka liczba firm, przy dużych nakładach środków własnych i inwestycjach Venture Capital.[4]

---

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

[2] Andrej Karpathy, 18.04.2024 , https://x.com/karpathy/status/1781047292486914189.

[3] 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.

[4] "*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.

Wydawałoby się więc, że ekonomia tworzenia AI uniemożliwia powstanie alternatyw – czy to finansowanych ze środków publicznych, czy tworzonych przez mniejszych graczy komercyjnych. Symbolem trudności z budowaniem tych alternatyw jest francuska firma Mistral, która miała być francuskim czempionem narodowym AI, do tego tworzącym otwarte rozwiązania dostosowane do potrzeb Francji i innych krajów europejskich. Ostatecznie, firma po pół roku zawarła strategiczne partnerstwo z Microsoftem, a swoje najmocniejsze modele zamknęła, wbrew początkowym deklaracjom.

Jednak, wbrew oczekiwaniom, alternatywy są tworzone równolegle z rozwojem wspomnianych modeli komercyjnych. W roku, w którym OpenAI udostępniło oparty na modelu GPT-3 serwis ChatGPT został też udostępniony model BLOOM, stworzony oddolnie przez społeczność wspieraną przez firmę Hugging Face[5]. Fundacja Eleuther.ai już w 2021 roku opublikowała model GPT-J. W amerykańskim Argonne National Laboratory trwają prace nad AuroraGPT, sfinansowanym ze środków publicznych naukowym modelu językowym.[6]

Duża część tych modeli powstaje w innym paradygmacie, niż komercyjne modele – są to tak zwane małe modele językowe.[7] Ich twórcy odchodzą od założenia, że – zgodnie z prawami skalowania – użyteczność modeli zależy od skalowania ich parametrów, a więc też od stosowania coraz większych zbiorów danych i mocy obliczeniowych.

Rozróżnienie dużych i małych modeli nie jest precyzyjne. Do małych modeli zaliczają się modele Bert, mające po kilkaset milionów parametrów, ale także modele o rozmiarach kilku miliardów parametrów, takich jak popularny otwarty model Mistral 7B. Małe modele mogą być jedną z wielu wersji w ramach jednej rodziny modeli – przykładowo, największy z modeli Mistral ma 123 miliardy parametrów. Dla nas jednak istotniejsze są modele tworzone niezależnie, przez inne podmioty, jako alternatywa dla dużych modeli językowych.

W ostatnim roku powstały liczne małe modele, typowo o rozmiarach 7B, 2B lub mniejszych.[8] Przełomowe znaczenie miały prowadzone przez badaczy z Microsoft prace nad małymi modelami z rodziny Phi. W artykule "Textbooks Are All You Need" zespół badaczy przedstawił metodologię zakładającą, że kluczowe znaczenie ma tworzenie wysokiej jakości zbiorów danych – które są "niczym podręczniki" dla modeli językowych.[9] Zdaniem Arvinda Narayanana i Sayasha Kapoora, można już mówić o zmianie trendu biznesowego, opartego w ostatnich latach na prawach skalowania mocy obliczeniowych i zbiorów danych.[10] Zwolennicy tego podejścia wskazują, że

---

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

[6] 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/.

[7] 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.

[8] Ethan Mollick (op. cit.)

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

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

małe modele mogą osiągać relatywnie dużą wydajność przy mniejszym nakładzie zasobów i większej energooszczędności.

Wiele z małych modeli to modele o otwartym źródle (open source). Rozwój systemów generatywnej sztucznej inteligencji opiera się w dużej mierze na otwarcie dostępnych rozwiązaniach będących fundamentami rozwoju tych technologii, takich jak biblioteki programistyczne dla uczenia maszynowego PyTorch i TensorFlow, metodologia Transformer, stosowana we wszystkich modelach językowych, ale też różnorodne inne narzędzia programistycznem, bazy danych i inne komponenty modeli.

Historycznie patrząc, pierwsze LLMy – na przykład pierwsze modele z rodziny GPT udostępnione przez OpenAI – były modelami otwartymi.[11] Największe, tworzone od kilku lat LLM są wszystkie zamknięte: dostępne jedynie poprzez API. Jednak wśród komercyjnych modeli podstawowych istnieje kilka, które są otwarte (choć nie jest to pełna otwartość). To m.in. stworzona przez Metę Llama oraz model Falcon, stworzony przez Technology Innovation Institute w Zjednoczonych Emiratach Arabskich.[12]

Modele te mogą być swobodnie wykorzystywane. Wyróżnia się przy tym modele w pełni otwarte oraz modele open weights, dla których dostępne są jedynie parametry. W obu wypadkach ważną zaletą, związaną z dostępnością parametrów jest możliwość dostrajania: stworzenia nowego w modelu w oparciu o już istniejący. Pozwala to zredukować zapotrzebowanie na moce obliczeniowe, gdyż pominięty jest etap wstępnego trenowania modelu.

Te dwa czynniki – zmieniający się paradygmat tworzenia LLM oraz otwarte udostępnianie modeli podstawowych – umożliwiły powstanie ekosystemu LLM, powstających w oparciu o dostępne otwarcie systemy i komponenty. Ważnym czynnikiem jest dostępność otwartych architektur modeli, dzięki którym można trenować kolejne, małe modele podstawowe. To przede wszystkim modele Mistral i Llama o rozmiarze 7B parametrów. Równie istotna jest dostępność otwartych modeli fundacyjnych – przede wszystkim z rodziny Llama stworzonej przez Meta. W ekosystemie tym współistnieją więc: otwarte modele fundacyjne, dostrojone duże modele, oraz małe modele – zarówno trenowane od podstaw na otwartych architekturach, jak i dostrojone.

Trzecim czynnikiem, będącym impulsem do tworzenia nowych modeli, są kwestie językowe. Rozwój modeli operujących w różnych językach, a także innych technologii do przetwarzania języka naturalnego (NLP), nie jest równomierny. Rozwój narzędzi takich jak modele językowe skupia się na garstce języków takich jak angielski oraz francuski, niemiecki, czy chiński. Z kolei przeważająca większość z 400 języków, z których każdym mówi ponad milion osób na świecie, nie posiada zbiorów danych, na bazie których można tworzyć modele językowe. Problem ten został zidentyfikowany w 2020 roku, jako kluczowe wyzwanie rodzące nierówności w dostępie

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

[12] Modele te, choć otwarte, nie są uznawane za modele o otwartym źródle, ze względu na ograniczenia ich wykorzystywania zapisane w licencjach, na których modele te są udostępnione. Patrz: 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.

do nowych technologii językowych.[13] Większość języków świata jest przez firmy AI uznawane za "low resource languages": języki o małej ilości dostępnych danych, co utrudnia trenowanie modeli. Firmy te nie podejmują starań, by tę lukkę językową wypełnić.[14] Zdaniem badaczy z Cohere, brak dostępnych danych wiąże się z niedostępnością mocy obliczeniowych w rejonach świata, które są zagrożone w związku z tym nowymi formami wykluczenia cyfrowego.[15]

Te trzy czynniki, razem wzięte, pozwalają zrozumieć dlaczego w ostatnich kilku latach pojawiło się wiele inicjatyw tworzenia nowych modeli językowych. Po pierwsze, są to modele lokalne, nie tylko obsługujące określone języki uznawane za "Low-Resource" przez twórców modeli fundacyjnych, ale też dostosowane do lokalnych potrzeb kulturowych. Po drugie są to modele otwarte, zakładające – w przeciwieństwie do dominujących modeli komercyjnych – duży poziom przejrzystości (dotyczącej wykorzystanych danych i ich pochodzenia, liczby użytych parametrów, typu wykorzystanego uczenia, np. nadzorowanego lub nienadzorowanego) oraz dostępności technologii do dalszego wykorzystania. Po trzecie wykorzystują architektury i metodologie tworzenia małych modeli językowych, by radzić sobie z wyzwaniem ograniczonych mocy obliczeniowych i środków finansowych. Część inicjatyw to także przejaw trendu nazwanego przez firmę Nvidia "suwerennym AI": tworzeniem modeli narodowych, w oparciu o własne dane i infrastrukturę.[16]

Te nowe inicjatywy to m.in. szwedzki GPT-SW3, singapurski SeaLion, francuska inicjatywa Common Corpus i model Albert, czy opisane w tym raporcie polskie projekty. To także modele tworzone do celów badawczych, takie jak stworzony przez Allen Institute for AI model Olmo, czy tworzony w National Argonne Lab model AuroraGPT.

W chwili obecnej eksperci i ekspertki nie są zgodne co do tego, czy małe modele mogą skutecznie konkurować z dużymi modelami podstawowymi, gwałtowne tempo rozwoju technologii utrudnia uchwycenie trendu. Twórcy małych modeli zakładają, że są one w stanie w coraz większym stopniu konkurować z dużymi: czy to dzięki nowym metodom tworzenia modeli, czy poprzez budowanie na bazie istniejących, otwartych modeli.

---

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

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

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

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

POLSKI EKOSYSTEM OTWARTYCH AI

W krótkim okresie między rokiem 2022, a chwilą obecną, w Polsce rozwinął się ekosystem rodzimych modeli językowych. Oczywiście, wielu ekspertów i organizacji zajmowało się uczeniem maszynowym i tworzyło modele językowe już wiele lat wcześniej, jednak ich działalność nie była widoczna publicznie.

Znaczenie takich inicjatyw zostało dostrzeżone w przyjętej w 2020 "Polityce dla rozwoju sztucznej inteligencji w Polsce od roku 2020".[17] Za jeden z głównych celów krótkoterminowych polityki uznano rozwój projektów dostosowanych do polskich wyzwań, w tym opartych na przetwarzanie maszynowym języka polskiego. Jedno z działań obejmuje "premiowanie projektów udostępniających architektury i wytrenowane modele oraz zbiory danych treningowych do powszechnego użycia".

Rok 2022 był pod wieloma względami przełomowy: upubliczniono w listopadzie tego roku serwis ChatGPT oparty na modelu GPT3. W lipcu tego roku udostępniono otwarty model językowy Bloom. Model ten obsługiwał 46 różnych języków, ale brakowało w nim języka polskiego. To zachęciło Sebastiana Kondrackiego, programistę i ewangelistę open source, do podjęcia próby utworzenia zbioru danych treningowych dla polskich modeli o rozmiarze co najmniej terabajta – tak w połowie 2022 roku zrodził się oddolny projekt SpeakLeash (który obecnie ma status fundacji).[18]

Rok później, w sierpniu 2023 roku, grupa polskich ekspertów od uczenia maszynowego opublikowała artykuł zatytułowany "O pożyczaniu innych światów, czyli po co nam polski LLM".[19] Wśród autorów są zarówno osoby związane ze społecznością SpeakLeash, jak i przedstawiciele publicznych instytucji badawczych, które kilka miesięcy później utworzą konsorcjum Polish Large Language Model (PLLuM). Artykuł jest rodzajem manifestu programowego środowiska polskich twórców modeli generatywnych. Przedstawiają wizję programu Polish Big Science: "Łączy nas jeden cel: stworzenie dużego polskiego modelu językowego, który będzie otwarty, dostępny i transparentny". Dokument wymienia szereg dalszych założeń i zasad przyświecających twórcom polskiego ekosystemu AI:

- Tworzenie w modelu open source i otwarte udostępnianie tworzonych rozwiązań. Twórcy polskich modeli wykorzystują wiele istniejących, otwartych rozwiązań przydatnych w tworzeniu otwartych modeli. Podkreślają przydatność danych dostępnych na jak najbardziej otwartych i przejrzystych zasadach. Sami również udostępniają w ten sposób tworzone modele i inne elementy systemów AI. Z otwartością wiąże się również przejrzystość – służąca zarówno kontroli sztucznej inteligencji, jak i wspierająca współpracę w tworzeniu modeli;

---

[17] Gov.pl, *Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020*, https://www.gov.pl/web/ai/polityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020

[18] Nazwa jest grą słów, oparta na fonetycznym brzmieniu słowa Spichlerz, i tworzy analogię między zbiorem danych a spichlerzem - zbiorem zboża.

[19] Gov.pl. *O pożyczaniu innych światów, czyli po co nam polski LLM*, 11.08.2023, https://www.gov.pl/web/ai/o-pozyczaniu-innych-swiatow-czyli-po-co-nam-polski-llm

- Dostosowanie do kontekstu krajowego i kontekstów lokalnych: zaletą tworzonych w Polsce modeli ma być lepsze dostosowanie do polskich realiów, dzięki wykorzystaniu baz polskich tekstów oraz odpowiedniemu dostrojeniu modeli. Jednocześnie, dzięki swojej otwartości, modele mogą być dalej dostrajane, i w ten sposób dostosowywać się do potrzeb konkretnych społeczności czy organizacji. W efekcie, istnienie wielu modeli będzie gwarantować pluralizm zapisanych w nich perspektyw, będących sposobem na zaadresowanie wyzwań dotyczących skrzywień (bias) modeli językowych;

- Współpraca i nabywanie kompetencji: skutkiem podejmowanych działań jest nie tylko tworzenie rozwiązań i produktów AI, ale też powstanie środowiska osób potrafiących tworzyć i wykorzystywać modele językowe (i szerzej technologie AI). Celem jest również współpraca między różnymi podmiotami posiadającymi niezbędne zasoby: teksty i dane, moce obliczeniowe i ekspertów oraz ekspertki gotowych podjąć tę pracę.

Przedmiotem naszej analizy jest powstający od kilku lat ekosystem otwartych AI, a przede wszystkim dwie inicjatywy realizujące wizję Polish Big Science. Najwięcej uwagi poświęcamy projektowi SpeakLeash, w ramach którego powstają modele z rodziny Bielik. Opisujemy również prace konsorcjum PLLuM (Polish Large Language Model), które są jednak na wcześniejszym etapie i nie zaskutkowały jeszcze dostępnym publicznie modelem.

Te inicjatywy tworzą modele językowe w dwóch różnych modelach organizacyjnych. SpeakLeash to inicjatywa oddolna, działająca na zasadach opisanej przez Yochai'a Benklera produkcji partnerskiej w oparciu o dobro wspólne (commons-based peer production).[20] Z kolei PLLuM to konsorcjum dużych instytucji badawczych, finansowane ze środków publicznych. Jeden z twórców SpeakLeash, Sebastian Kondracki, do opisania dwóch inicjatyw odwołuje się często do metafory katedry i bazaru.[21] Jak pokażemy, mimo różnic w podejściu, oba projekty wiele łączy: założenie otwartego udostępniania modeli, podobne cele strategiczne dla polskich modeli językowych, oraz podobne wyzwania, którym oba projekty muszą sprostać.

W ramach polskiego ekosystemu LLM działa też szereg innych inicjatyw, które nie są przedmiotem naszej analizy. Politechnika Gdańska i Ośrodek Przetwarzania Informacji stworzyły w marcu 2024 roku model Ora, poprzez dostrojenie modelu Llama 2 na korpusie polskich danych językowych. Podobny charakter ma model Trurl, stworzony przez firmę VoiceLab.ai.

Liderzy opisywanych projektów często podkreślają, że nie chodzi im tylko o tworzenie modeli LLM, a rozwój ekosystemu jest równie ważny co poszczególne projekty. Co to oznacza? Twórcy LLM pracują nie tylko nad modelami, ale też innymi komponentami systemów AI: bazami danych treningowych i instrukcji, narzędziami takimi jak tokenizery, procedurami czyszczenia danych czy metodologiami treningowymi. Oznacza to również podejście oparte na wymianie wiedzy i kompetencji między inicjatywami, które potencjalnie mogłyby ze sobą konkurować. Podejście ekosystemowe zakłada wreszcie rozwój szerszego środowiska społeczno-gospodarczego, opartego o wykorzystanie modeli językowych.

---

[20] Yochai Benkler, Bogactwo sieci, przeł. R. Próchniak (Warszawa: WAiP 2008).

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

Poniżej opisujemy dwie kluczowe inicjatywy: SpeakLeash i PLLuM. W następnej części przyjrzymy się dokładniej różnym aspektom tworzenia modeli, i różnym zasobom, które są do tego niezbędne.

DLACZEGO WARTO TWORZYĆ POLSKIE MODELE JĘZYKOWE?

W rozmowach z twórcami polskich modeli pytaliśmy ich o zalety takiego podejścia. Można je podzielić na trzy obszary: kulturowy, strategiczny, biznesowo-publiczny.

Kultura

Obszar kultury obejmuje zalety związane z lepszym dostosowaniem modeli do polskiej rzeczywistości. To kluczowy obszar, który podkreślają twórcy polskich modeli. Motywacją do ich tworzenia jest nieuwzględnienie polskiego języka w dominujących modelach, słaba jakość tekstów polskich generowanych przez modele obsługujące język polski, oraz brak uwzględnienia lokalnego kontekstu kulturowego. Chodzi przede wszystkim o:

- rozumienie metafor i odniesień kulturowych – modele trenowane na korpusie polskich tekstów, w tym literatury, będą lepiej rozumiały polską kulturę, m.in. odniesienia literackie, wiedzę geograficzną, polskie przepisy kulinarne czy język potoczny;

- zachowanie gwar: ciekawym zastosowaniem modeli językowych jest zachowanie gwar i dialektów – model językowy może być dla nich domem spokojnej starości;

- konteksty lokalne i specjalistyczne: poprzez dostrajanie, otwarte modele mogą być dostosowywane do potrzeb konkretnych społeczności, czy organizacji. Istnienie różnorodnych modeli zapewnia pluralizm perspektyw.

Strategia

Obszar strategiczny dotyczy kontroli nad danymi i nad kosztami modeli językowych. Jest on także związany z budowaniem cyfrowej suwerenności, czyli ustalania i egzekwowania na poziomie narodowym reguł, na jakich działa internet i inne technologie cyfrowe. Najważniejsze kwestie to::

- kontrola kosztów: strategią gigantów technologicznych jest monopolizacja rynku, mimo niskich obecnych kosztów korzystania z ich rozwiązań nie da się wykluczyć przyszłego ich wzrostu. Posiadanie rodzimych modeli zmniejsza to ryzyko, szczególnie gdy są one dostępne w sposób otwarty;;

- kontrola nad miejscem i sposobem przetwarzania danych: jest to szczególnie ważne w przypadku przedsiębiorstw i organizacji chroniących swoją własność intelektualną – ale może też służyć ochronie prywatności użytkowników. Wykorzystywanie otwartych modeli na własnej infrastrukturze oraz ich dostrajanie służy osiągnięciu tych celów;;

- budowanie know-how: osoby pracujące na wszystkich etapach powstawania modelu nabywają unikalne kompetencje i rozpowszechniają je na rynku pracy. Dotyczy to zarówno wysoko wykwalifikowanych kadr tworzących modele, jak i osób uzyskujących doświadczenie z ich wdrażaniem w różnych kontekstach.

Biznes i administracja publiczna

Obszar biznesowy i związany z administracją publiczną dotyczy tego, jak lokalne modele językowe mogą wpłynąć na krajobraz polskiej gospodarki, szczególnie w sektorze małych i średnich przedsiębiorstw, ale także w jaki sposób organy państwa i samorządy mogą czerpać z osiągnięć polskich LLMów. Modele mogą zostać przystosowane do potrzeb danej organizacji dzięki dostrajaniu, czyli dodaniu do modelu danych pochodzących z działalności tego konkretnego przedsiębiorstwa. Podobne argumenty dotyczą wdrożeń w administracji publicznej.

Twórcy Bielika podkreślają biznesowe zastosowania swojego modelu, a szczególnie:

- szybsze i tańsze przetwarzanie konkretnych problemów: modele dostosowane do przedsiębiorstwa mogą wykonywać konkretne zadania biznesowe dużo szybciej i przy mniejszym zużyciu mocy obliczeniowej, niż modele używane do ogólnych zastosowań;

- możliwość samoregulacji i współpracy: lokalne zarządzanie modelami umożliwia partnerskie relacje między twórcami modeli a biznesem, który posiada dane, na których modele mogą być trenowane. Osiągnięcie porozumienia w takiej kwestii w sposób etyczny i zabezpieczający interes mniejszych organizacji jest zdecydowanie łatwiejsze na poziomie lokalnym. Globalne korporacje często działają według motta “move fast, break things”, pozyskując wszelkie możliwe dane. Dopiero w wypadku pozwu zaczyna liczyć się kwestia legalności i interesu innych podmiotów. Istnieją jednak przykłady porozumień między wielkimi firmami oferującymi LLMy, a graczami typu Axel Springer[22] czy Reddit[23].

SpeakLeash

SpeakLeash to oddolna społeczność, której celem jest tworzenie zbiorów danych, narzędzi i metodologii do trenowania modeli językowych działających w języku polskim. Główną przestrzenią komunikacji i współpracy jest społeczność na Discordzie, zrzeszająca ponad 1000 użytkowników (i stale rosnąca). Spośród nich około 10% osób angażuje się bezpośrednio w prace nad modelem. Bielik to model stworzony przez garstkę osób na kilkuset użyczonych procesorach graficznych, w oparciu o dane zebrane przez oddolną inicjatywę działającą w modelu produkcji partnerskiej, przy minimalnym budżecie. Sami jego twórcy kontrastują Bielika i SpeakLeash z inicjatywami opartymi na wielkich zespołach badawczych, mających dostęp do tysiąckrotnie większych zasobów. To, wedle naszej wiedzy, unikatowy w skali świata przykład otwartego modelu stworzonego w formule produkcji partnerskiej – porównywalny z wcześniejszymi działaniami społeczności Eleuther.ai.[24]

---

[22] "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.

[23] "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/.

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

SpeakLeash jest projektem w pełni opartym na założeniach otwartego oprogramowania i otwartej nauki. Jednym z kluczowych założeń jest tworzenie systemów i narzędzi sztucznej inteligencji dostosowanych do polskiej specyfiki kulturowej. Założyciele inicjatywy opisują to na przykładach: polski model powinien myśleć o wakacjach na Mazurach, a nie na Hawajach. Wiedzieć, co to żurek, i rozumieć że Janusz to nie tylko męskie imię, ale też slangowe określenie pewnego rodzaju mężczyzny.

Drugim jest wizja upowszechnienia technologii AI w Polsce. Inicjatywa zakłada, że dzięki tworzeniu małych modeli oraz stosowaniu mechanizmów dostrajania modeli, mogą powstawać rozwiązania dostosowane nie tylko do specyfiki Polski jako kraju, ale też do potrzeb lokalnych społeczności, konkretnych organizacji, a nawet indywidualnych osób.

W raporcie podkreślamy rolę współpracy i społecznościowy wymiar inicjatywy SpeakLeash. Jednak trzeba też podkreślić kluczową rolą pojedynczych osób. Najważniejszą z nich jest Sebastian Kondracki, który w 2022 zainicjował Spichlerz. Jak opowiada, zawodowo już kilka lat wcześniej zajmował się wdrożeniami rozwiązań opartych na uczeniu maszynowym dla bankowości i e-commerce. Znane mu więc były kwestie takie jak bezpieczeństwo danych czy zalety otwartych, kompaktowych modeli. Gdy w lecie 2022 usłyszał o modelu BLOOM, doszedł do wniosku, że można stworzyć społeczność pracującą nad polskimi rozwiązaniami tego rodzaju. Kluczowe znaczenie dla rozwoju polskiego ekosystemu AI było silne przywiązanie Kondrackiego i innych założycieli SpeakLeash do etosu open source i open science.

Sebastian Kondracki nawiązał współpracę z Hugging Face i Eleuther.ai, od których uzyskał podstawowy know-how na temat tworzenia modeli językowych. Członkowie Eleuther.ai obiecali, że jeśli uda mu się zebrać terabajt danych, to zapewnią moce obliczeniowe i wsparcie w stworzeniu polskiego modelu, na bazie ich modelu GPT NeoX. W przeciągu roku do projektu dołączył szereg ekspertów z doświadczeniem w budowaniu i pracy z modelami. Nowe otwarte modele, takie jak Llama czy Mistral, oraz działalność platformy Hugging Face, wytyczyły kierunki tworzenia i dostrajania otwartych modeli.

Społeczność SpeakLeash oprócz korpusu języka polskiego tworzy różnorodne narzędzia: m.in. autorskie narzędzie do OCRowania PDF-ów, tokenizer dostosowany do języka polskiego, narzędzia do benchmarkowania modeli – w tym leaderboard dla modeli polskojęzycznych, czy dokumentację tworzenia modeli językowych. Jednocześnie w wielu wypadkach korzystają z istniejących, otwartych rozwiązań – co pozwala znacząco zredukować koszty i przyspieszyć prace. Nie chodzi tu tylko o fakt, że polski model, tak jak wszystkie inne LLM, jest zbudowany na architekturze uczenia maszynowego transformer, udostępnionej przez firmę Google w sposób otwarty. Wykorzystywanych jest wiele, mniejszych i większych komponentów, m.in. architektura modelu Mistral, bazy danych treningowych i instrukcji, czy benchmarki służące mierzeniu jakości modelu.

Przełomowym momentem dla inicjatywy było nawiązanie w 2024 roku współpracy z centrum superkomputerowym Cyfronet AGH, dzięki której na bazie danych SpeakLeash w kwietniu został udostępniony model językowy Bielik. Był to pierwszy model, którego trenowanie obejmowało zarówno stworzenie modelu bazowego (jak wcześniej model Qra), jak i modelu dostrojonego (jak

wcześniej Trurl, dostrojony na bazie modelu Llama). Tak jak niemal wszystkie otwarte modele na świecie, Bielik nie był trenowany od podstaw – choć jest modelem podstawowym. Do zainicjowania treningu Bielika wykorzystano dostępną na otwartej licencji Apache 2.0 architekturę modelu Mistral 7B.

Twórcy Bielika wykorzystali 18 milionów dokumentów ze zbiorów SpeakLeash, na bazie których stworzono zbiór treningowy złożony z 22 miliardów tokenów. Praca nad jakością zbioru obejmowała usuwanie tekstów wybrakowanych lub nieodpowiednich, anonimizację i poprawę formatowania plików. Wykorzystano również model klasyfikujący do wyselekcjonowania treści wysokiej jakości. Zbiór ten uzupełniono tekstami angielskojęzycznymi ze zbioru SlimPajama. W kwietniu 2024 roku inicjatywa SpeakLeash udostępniła model językowy Bielik 7B v0.1, oraz dostrojoną z pomocą instrukcji wersję Bielik 7B-instruct v0.1.[25]

Podejście ekosystemowe przyjęte przez liderów SpeakLeash oznacza, że powstanie modelu nie uznano za produkt końcowy inicjatywy. Odwrotnie – jest to jeden z kroków w tworzeniu ekosystemu, który wspiera tworzenie innych modeli oraz wspólnie rozwijanych narzędzi.[26] Wyrazem tego podejścia było utworzenie, w ramach inicjatywy SpeakLeash, polskiego "leaderboard", porównującego zdolność generowania wysokiej jakości polskiego tekstu przez różne modele. Leaderboard powstał dzięki zlokalizowaniu do polskich warunków angielskojęzycznego EQ-Bench leaderboard, i jest przykładem jak wykorzystanie otwartych narzędzi przyspiesza pracę nad zlokalizowanymi modelami. Narzędzie to, stworzone przed upublicznieniem Bielika, pozwoliło go testować w odniesieniu do już istniejących modeli – dając twórcom Bielika możliwość oceny jakości tworzonego modelu.

Od kwietnia trwa dalszy rozwój rodziny modeli Bielik. W sierpniu upubliczniono wersję drugą modelu. To wersja bardziej dostrojona – model Bielik w wersji pierwszej został wykorzystany do automatycznego przefiltrowania i usunięcia słabych danych z bazy treningowej, stworzono również nową wersję modelu dostrojonego na instrukcjach. Obecnie trwają prace nad wersją trzecią, dostrojoną na większym zbiorze instrukcji i dialogów. Wszystkie modele z rodziny Bielik są dostępne na licencji Apache 2.0.

Widać już pierwsze przykłady tworzenia na podstawie Bielika, poprzez dostrajanie, nowych modeli. Startup TheLion.ai ogłosił w sierpniu, że na bazie modelu Bielik v2 będą dostrajać mały medyczny model językowy.[27] Liderzy SpeakLeash mają też świadomość, że dalszy rozwój ich

---

[25] 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.

[26] "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/.

[27] 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.

modeli będzie wymagał zbudowania nowych relacji, które dadzą dostęp przede wszystkim do nowych zbiorów danych treningowych. Równie istotne będzie budowanie partnerstw dotyczących wdrażania rozwiązań AI.

PLLuM

PLLuM to założone w listopadzie 2023 roku konsorcjum instytucji badawczych, które pracuje nad dużym polskim modelem językowym. Jego członkami są: Politechnika Wrocławska (lider projektu), Naukowa i Akademicka Sieć Komputerowa (NASK), Instytut Podstaw Informatyki PAN, Instytut Slawistyki PAN, Ośrodek Przetwarzania Informacji (OPI) i Uniwersytet Łódzki. Konsorcjum otrzymało 14,5 milionów złotych dofinansowania w ramach dotacji celowej z Ministerstwa Cyfryzacji na badania podstawowe w zakresie sztucznej inteligencji.[28] Konsorcjum PLLUM ma w chwili obecnej jedynie roczne dofinansowanie, do końca 2024 roku. Jednocześnie kierownicy projektu podkreślają, że jest on częścią długofalowej strategii tworzenia modeli językowych i baz danych, w oparciu także o własną infrastrukturę obliczeniową.

Większość członków konsorcjum działa również w istniejącym od ponad dziesięciu lat konsorcjum CLARIN-PL, będącego częścią europejskiego konsorcjum tworzącego rozwiązania do pracy na dużych zbiorach danych językowych.[29] Finansowanie ze środków CLARIN zapewniło dużą część niezbędnej infrastruktury badawczej i mocy obliczeniowej, budowanej od 2018 roku. Dodatkowo, Politechnika Wrocławska przeznaczyła 130 milionów złotych z innych grantów na zakup 300 procesorów graficznych Nvidia H100.

Większość organizacji z konsorcjum ma doświadczenie w pracy z modelami LLM. Zaangażowani badacze od lat prowadzą badania ewaluacyjne modelu GPT, a OPI wcześniej uczestniczyło w pracach nad modelem bazowym Qra.

Podstawą PLLuM jest duży zbiór danych tekstowych w języku polskim, obejmujący korpus językowy tworzony przez IPI PAN, zbiory tworzone w ramach CLARIN (np. Słowosieć), różnorodne otwarte zasoby i dane z web scrapingu (zbieraniu danych z ogólnodostępnego Internetu). Na jego bazie tworzona jest rodzina modeli z wykorzystaniem architektury Mistral Nemo 7b: modele bazowe o rozmiarach 7x8B, 7x22B i 70B parametrów oraz wirtualny asystent dostosowany do potrzeb instytucji publicznych. Model stworzony przez projekt PLLUM ma być dostępny pod koniec 2024 roku. PLLuM ma być alternatywą dla dostępnych już komercyjnych modeli, które są zamknięte, kosztowne w użyciu oraz niedostosowanie do polskiego kontekstu.[30] Projekt ma więc podobne założenia co SpeakLeash. Kluczowym założeniem jest otwartość samego modelu, jak i jego komponentów. Twórcy PLLUM chcą również zapewnić jak największą zgodność z wymogami Aktu o Sztucznej Inteligencji (AI Act).

---

[28] "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.

[29] 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.

[30] "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/.

Podobnie jak w pracach nad Bielikiem, ważnym elementem projektu PLLuM jest stworzenie wysokiej jakości zbioru instrukcji. O ile twórcy Bielika w dużej mierze korzystają z instrukcji przetłumaczonych z angielskiego lub automatycznie generowanych, PLLuM zainwestowało środki w tworzenie nowego zbioru, dostosowanego do języka polskiego i jego kontekstu kulturowego.

JAK POWSTAJE MODEL JĘZYKOWY?

Proces trenowania modelu językowego wymaga zapewnienia kilku kluczowych zasobów:

- dane: zbiory treningowe wraz z metodologiami selekcji, czyszczenia i strukturyzowania tych danych;
- moce obliczeniowe: klastry procesorów graficznych (GPU) oraz zasoby chmurowe niezbędne do wykorzystywania modelu (inferencji);
- architektura i metodologia trenowania modelu: określa strukturę modelu i sposób, w jaki przetwarza on informacje i generuje wyniki;
- instrukcje: specjalne zbiory dialogów i instrukcji stosowane przy dostrajaniu modeli;
- społeczność: środowisko osób tworzących, dostrajających i wykorzystujących model.

Lokalne modele językowe tworzy się w procesie wstępnego uczenia na zbiorach danych treningowych, a następnie dostrajania przy użyciu odpowiednich zbiorów instrukcji.

Początkowy krok, wstępne szkolenie, koncentruje się na nauczaniu modelu w celu zrozumienia statystyk sekwencji słów. Proces ten polega na dostarczeniu modelowi dużego zbioru tekstów, tak aby docelowo potrafił przewidywać następne słowa w sekwencji słów.

Następny etap, dostrajanie, polega na udoskonaleniu zrozumienia przez model konkretnego języka i kontekstu kulturowego. Ten krok jest kluczowy dla poprawy płynności w działaniu, dokładności i znaczenia kulturowego modelu. Dostrajanie pomaga modelowi lepiej rozumieć język naturalny i eliminować luki w wiedzy. Jednocześnie jest to proces, który celowo redukuje zakres i głębokość opisywanej wiedzy.

W tej części przyjrzymy się bliżej tym kluczowym zasobom.

Dane treningowe

Odpowiedni, wystarczająco obszerny i wysokiej jakości zbiór danych treningowych ma kluczowe znaczenia dla tworzenia LLM. Duża część prac nad polskimi modelami LLM dotyczy więc tworzenia zbiorów danych treningowych.

Punktem wyjścia dla twórców modeli językowych – zarówno w Polsce jak i na świecie – są często dane z web scrapingu otwartej sieci, często pozyskiwane z bazy Common Crawl (lub tworzonych na tej podstawie zbiorów danych). Z drugiej strony, twórcy LLM starają się je uzupełnić o dodatkowe zasoby – szczególnie, że dane z sieci są często niskiej jakości. Nie przypadkiem projekt SpeakLeash przez pierwsze dwa lata skupiał się nie na trenowaniu modelu, lecz tworzeniu bazy danych. Z kolei projekt PLLuM ma ułatwiony dostęp do baz danych naukowych i korpusów językowych tworzonych i zarządzanych przez członków konsorcjum. Punkt wyjścia jest zawsze ten sam – odpowiednio wyselekcjonowane dane z web scrapingu, oraz dane dostępne na otwartych licencjach. Poszczególne projekty uzupełniają je, starając się uzyskać dostęp do

danych – i zgodę na ich wykorzystanie – ze źródeł, które nie są publicznie dostępne. Są to na przykład materiały urzędowe instytucji publicznych, zasoby wydawnictw, czy archiwa mediów.

Jak stwierdził Jan Kocoń, jeden z inicjatorów PLLuM, "Mamy [danych] ogrom, ale potrzebujemy więcej, bo język polski stwarza wiele trudności. Chodzi w szczególności o treści specyficzne dla polskiego kontekstu kulturowego i społecznego."[31] Jeden z ekspertów, z którym rozmawialiśmy twierdzi, że nawet gdyby mieć dostęp do wszystkich treści cyfrowych stworzonych w języku polskim – także tych niedostępnych publicznie – to nie byłaby to ilość wystarczająca, by wytrenować duży model, zdolny konkurować z największymi modelami dostępnymi na rynku. Tak więc rozmiar polskich zasobów językowych warunkuje rozwój LLM, i ogranicza go do małych i średnich modeli.

SpeakLeash zebrał do dzisiaj zbiór danych, szacowany na 15-20 terabajtów surowych danych, 1,5 terabajta wyczyszczonych danych i około 180 miliardów tokenów. SpeakLeash to zbiór złożony z różnorodnych danych. Pierwszym zbiorem była rzekomo polska Wikipedia, obecnie baza zawiera szeroki wachlarz rodzajów danych. Założyciele SpeakLeash wzorowali się, projektując nową bazę danych, na dwóch przykładach. Jednym był zbiór the Pile, zawierający w dużej mierze treści z web scrapingu, połączone z wybranymi ustrukturyzowanymi zbiorami danych. Drugim przykładem był zbiór stworzony na potrzeby modelu Bloom, którego elementy były precyzyjnie skatalogowane - ale był to zbiór dużo mniejszy. Inspirowani tymi przykładami, twórcy SpeakLeash założyli, że dane będą dokładnie opisane, łącznie z informacją licencyjną i charakterystyką tekstu. Będą też otwarte w sensie technicznym: łatwe do użycia dla każdego z podstawową znajomością języka Python.

Zasoby SpeakLeash są w dużej mierze oparte na web scrapingu, ale uznano, że bazy takie jak Common Crawl nie zapewniają wystarczająco wysokiej jakości danych – często ze względu na nieodpowiednie wykorzystanie takich zbiorów danych.[32] W związku z tym SpeakLeash prowadzi własne procesy indeksowania sieci, pobierania zindeksowanych danych a następnie ich klasyfikowania i czyszczenia. Punktem wyjścia było stworzenie narzędzia indeksującego polskojęzyczny internet. Według członków SpeakLeash udało im się się dotrzeć do większości dostępnych w publicznej sieci zasobów polskojęzycznych. W rezultacie, baza danych SpeakLeash składa się z wielu zbiorów treści pochodzących z Sieci, które zostały wyczyszczone, a w wielu wypadkach także pokategoryzowane i zebrane w zbiory tematyczne. Ze zbioru usunięto też treści, których właściciele nie wyrażają zgody na stosowanie do trenowania LLM (poprzez plik robots.txt lub inną formę zastrzeżenia). Dodatkowo, baza zawiera treści rozmaitych forów internetowych, w tym tych największych, jak Kafeteria.pl, Gazeta.pl, czy Wizaz.pl

Równocześnie baza SpeakLeash agreguje również różne istniejące zbiory i kolekcje. Są to, m.in.: zasoby polskiej biblioteki cyfrowej Polona, materiały polskiego Parlamentu, różnorodne bazy

[31] 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.

[32] 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/.

tekstów prawnych, kolekcja książek w domenie publicznej Wolne Lektury, czy zasoby tekstowe Europeany. Projekt korzysta też z istniejących międzynarodowych zbiorów danych, z których część zawiera zasoby w języku polskim, takich jak korpus napisów do filmów OpenSubtitles. Przy okazji tworzenia tej bazy powstało wiele wyspecjalizowanych narzędzi i metodologii pracy z danymi.

Jeden z naszych rozmówców, mający kluczową rolę w procesie budowania bazy SpeakLeash, twierdzi, że korzystanie z danych z publicznego internetu jest koniecznością, ale też obciążeniem ze względu na ich niejasny status. Docelowo, twórcy SpeakLeash chcieliby w większej mierze polegać na dodatkowych źródłach danych, pochodzących na przykład z bibliotek, archiwów czy od wydawców. Ich zaletą jest jasny stan prawny, a w wielu wypadkach także wyższa jakość niż zasobów internetowych. Optymalnym scenariuszem byłaby dostępność większej ilości otwartych danych.

Wykorzystanie otwartych danych jest też ważnym założeniem projektu PLLuM, w ramach którego powstaje zbiór zawierający wszystkie polskojęzyczne zasoby dostępna na otwartych licencjach. To zbiór zawierający 8 miliardów słów, pięć razy więcej niż narodowy korpus językowy – jednak eksperymenty z wytrenowaniem modelu na tych danych pokazały, że jest to rozmiar niewystarczający. Równocześnie, ważnym elementem projektu jest pozyskanie danych licencjonowanych, udostępnianych przez różnorodne podmioty, tak publiczne jak i prywatne. Do chwili obecnej PLLuM podpisał kilkadziesiąt umów na wykorzystanie do trenowania modelu treści będących własnością takich podmiotów, w tym instytucji administracji publicznej czy wydawnictw. Jednak nawet połączone zbiory danych otwartych i licencjonowanych są niewystarczające – także model PLLuM jest trenowany na danych z web scrapingu.

Widać więc, że nawiązywanie współpracy z podmiotami posiadającymi różnorodne treści i zasoby językowe jest kluczowym kolejnym krokiem dla twórców polskich LLM. Z jednej strony szczególnie projekt PLLuM z sukcesem podpisuje umowy na wykorzystanie różnych prywatnych zbiorów. Przykładowo, konsorcjum podpisało umowę z Grupą Agora, która udostępni swoje materiały pro bono, na potrzeby szkolenia modelu.[33]Z drugiej strony, pojawiają się pierwsze napięcia między różnymi grupami interesów. Przykładem może być sytuacja związana z propozycją przekazania treści na potrzeby trenowania modelu, jaką w marcu wystosowało do polskich wydawców konsorcjum PLLuM. W odpowiedzi, polska Izba Wydawców Prasy opublikowała w czerwcu stanowisko, w którym stwierdziła, że propozycja współpracy nie porusza tematu uzyskania licencji na wykorzystanie treści, ani odpowiedniego wynagrodzenia. W związku z tym Izba stwierdziła, że przekazanie treści oznacza "ryzyko utraty kontroli nad materiałami prasowymi". Wydawcy podważali stosowanie przepisów dozwolonego użytku do trenowania generatywnego AI, i wezwali konsorcjum do podpisania płatnych umów licencyjnych.[34] Doświadczenia konsorcjum PLLuM pokazują, że ostatecznie jest możliwe uzyskanie

[33] "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

[34] "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.

porozumienia z posiadaczami praw w tym zakresie i uzyskanie nieodpłatnych licencji na wykorzystanie zasobów do trenowania modela językowego.

Moc obliczeniowa

O tym, czy na zebranych danych zostanie wytrenowany model decyduje dostęp do mocy obliczeniowej. Szczególnie dużych zasobów wymaga wstępne trenowanie modelu bazowego (pre-training). Moce obliczeniowe mogą być dużo mniejsze w przypadku modeli, które są tworzone przez dostrojenie istniejącego modelu fundacyjnego.

Zdaniem naszych rozmówców, praktyka tworzenia lokalnych modeli pokazuje, że nie jest konieczny dostęp do mocy obliczeniowych na poziomie tych wykorzystywanych przez największe firmy AI. Innowacje w zakresie metodologii tworzenia modeli pozwalają znacząco zredukować niezbędną moc obliczeniową – do poziomu, który mogą zapewnić działające w Polsce centra superkomputerowe.

Polska jest krajem, w którym publiczne instytucje badawcze mają łącznie kilkaset nowoczesnych procesorów graficznych potrzebnych do trenowania AI. Dla porównania, Meta szacuje, że pod koniec 2024 roku będzie mieć moc obliczeniową rzędu 600 tysięcy procesorów H100.[35] Niemniej, te zasoby okazały się wystarczające, by stworzyć pierwsze polskie modele językowe. Ważnym punktem odniesienia był projekt BigScience i stworzony w jego ramach model BLOOM, który został wytrenowany na francuskim superkomputerze Jean Zay, przy wykorzystaniu niecałych 500 procesorach graficznych Nvidia A100.[36]

Na początku 2024 roku, firma Azurro utworzyła model APT-1B na wyselekcjonowanych danych ze SpeakLeash, wykorzystując pojedynczą kartę graficzną. Eksperyment udowodnił możliwość stworzenia modelu z pomocą polskich danych językowych i przy minimalnych mocach obliczeniowych. Stworzony model nie był jednak wysokiej jakości –stało się jasne, że niezbędne są większe moce obliczeniowe.

W analizowanych przez nas projektach kluczową rolę grają partnerzy posiadający odpowiednią moc obliczeniową. W obydwu analizowanych przypadkach są to centra superkomputerowe prowadzone przez publiczne instytucje badawcze. W przypadku PLLuM, tę rolę odgrywa Wrocławskie Centrum Sieciowo-Superkomputerowe przy Politechnice Wrocławskiej. W przypadku SpeakLeash, model Bielik mógł powstać dzięki współpracy z centrum superkomputerowym Cyfronet AGH, które udostępniło 256 procesorów Nvidia GH200. Szkolenie modelu Bielik odbywało się w ramach testowania tego nowego klastra procesorów. Liderzy inicjatywy SpeakLeash podkreślają unikalny charakter współpracy między ośrodkiem superkomputerowym a nie zinstytucjonalizowaną grupą, działającą w modelu open source.

---

[35] 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/.

[36] 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.

Przedstawiciele obydwu projektów podkreślają, że dostęp do odpowiedniej mocy obliczeniowej, dostępnej przez dłuższy czas, jest dla nich wyzwaniem. W przypadku Bielika – mimo wsparcia Cyfronetu – model mógłby zostać lepiej dotrenowany przy dostępie do większej mocy. Z kolei w projekcie PLLuM wyzwaniem jest czas, związany z rocznym okresem finansowania. Podobnie, możliwość dłuższego trenowania pozwoliłoby stworzyć model wyższej jakości. Jednocześnie ograniczenie te napędza innowacje – jeden z rozmówców stwierdził, że prawdziwym wyzwaniem jest znalezienie sposobów na skuteczne trenowanie wysokiej jakości modeli, przy ograniczonych zasobach obliczeniowych.

W trakcie wywiadów padło też stwierdzenie, że, "głównym zaczynem są centra superkomputerowe". Są bowiem jedynymi aktorami posiadającymi moce obliczeniowe – alternatywą jest wykupienie dostępu do komercyjnej infrastruktury chmurowej. Centra te traktują również pracę nad modelami jako okazję do stworzenia nowoczesnych środowisk badawczych. Takim przykładem jest model Qra, stworzony z inicjatywy Politechniki Gdańskiej, która na potrzeby projektu wykorzystała swoje Centrum Kompetencji STOS oraz superkomputer Kraken, w tym klaster 21 kart graficznych Nvidia A100.[37]

Tworzenie i dostrajanie modelu

Osoby zaangażowane w projekty tworzenia nowych modeli językowych stoją przed wyborem: można podjąć się tworzenia modelu bazowego, lub dostrajać model istniejący. Jedynie nieliczne inicjatywy mają odpowiednie zbiory danych i dostęp do mocy obliczeniowych, by tworzyć modele bazowe. Są one jednak niezbędne, by wypełnić istniejące luki językowe. Tak więc zarówno inicjatywa SpeakLeash, jak i konsorcjum PLLuM podjęły się tworzenia modeli bazowych. Modele te nie są tworzone całkiem od zera – obydwa projekty wykorzystują architekturę modelu Mistral, dostępną na licencji Apache 2.0. Rozmówcy podkreślają ogólną rolę rozwiązań open source w demokratyzacji dostępu do LLM, ale także w umożliwieniu dalszych innowacji. W przypadku Bielika duże znaczenie miało wykorzystanie ALLaMo, autorskiej metodologii trenowania stworzonej przez Krzysztofa Ociepę. Otwarte udostępnianie zbiorów danych, kodu czy metodologii zachęca do szerszego uczestnictwa i wspiera dalsze innowacje.

W trakcie pisania raportu nie został jeszcze upubliczniony model PLLuM, którego premiera jest planowana na koniec 2024 roku. Na przykładzie modelu Bielik widać natomiast, że udostępnienie modelu nie dzieje się raz – jest to proces ciągły, w ramach którego powstają kolejne wersje. W przypadku Bielika są to: modele bazowy i instrukcyjny Bielik 7B 0.1, model instrukcyjny Bielik 11B 2.0, a następnie trzy kolejne wersje tego modelu, różniące się zwięzłością wypowiedzi i zdolnością odgrywania zadanych ról.[38] Obecnie trwają prace nad wersją 3.0.

Kwestia zbiorów instrukcji używanych do dostrajania modeli często nie jest poruszana w dyskusjach o danych treningowych – tymczasem są one jednym z niezbędnych zasobów, którego

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

[38] 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.

pozyskanie jest sporym wyzwaniem. Instrukcje są bowiem niezbędne dla tworzenia dostrojonych wersji modeli – na przykład takich, które nadają się do tworzenia chatbotów. Wyzwaniem dla twórców polskich modeli jest brak zbioru instrukcji i dialogów w języku polskim. W przypadku pierwszej wersji Bielika, połowa instrukcji pochodziła z przetłumaczonych na polski zbiorów angielskojęzycznych OpenHermes-2.5 oraz orca-math-word-problems-200k. Dodatkowo, społeczność Bielika stworzyła niewielki zbiór instrukcji i dialogów. Wreszcie, wygenerowano też automatycznie milion dialogów w oparciu o wybór artykułów ze zbiorów SpeakLeash. Bielika dostrojono na zbiorze obejmujący łącznie 2,3 miliona instrukcji (700 milionów tokenów). Modele Bielik drugiej generacji były dostrajane na jeszcze większym zbiorze, złożonym przede wszystkim z instrukcji wygenerowanych z pomocą modelu Mixtral 8x22B. W ten sposób uzyskano 16 milionów instrukcji (8 miliardów tokenów).[39] Tworzenie tych zbiorów danych jest procesem ciągłym.

Ważnym narzędziem, pozwalającym oceniać jakość modeli, jest uruchomiona po publikacji Bielika platforma Chat Arena PL. To platforma pozwalająca użytkownikom testować dwa losowo wybrane modele pod względem jakości odpowiedzi udzielonych na ten sam prompt. W ten crowdsourcingowy sposób są zbierane dane, które pozwalają ocenić jakość Bielika i innych modeli w sposób inny, niż poprzez tradycyjny benchmark. To także kolejny przykład wykorzystania istniejących, otwartych narzędzi by wzbogacić polski ekosystem AI – projekt bazuje na otwartym rozwiązaniu Chat Arena. Do tej pory przeprowadzono na niej ponad kilkanaście tysięcy testów – jest to niestety zbyt mało, by stworzyć przydatny zbiór instrukcji na ich bazie.

Na razie niewiele wiadomo o zbiorach instrukcji, na których są dostrajane modele PLLuM. Twórcy projektu mają ambicję wykorzystać jak najwięcej instrukcji stworzonych na potrzeby projektu, pisanych przez zatrudnionych ekspertów. Szacują, że kompletny zbiór powinien zawierać kilkadziesiąt milionów instrukcji i dialogów. Stworzenie takiego zbioru instrukcji może być kolejnym etapem rozwoju polskich modeli językowych – jednym z rozważanych pomysłów jest wygenerowanie ich w sposób crowdsourcingowy.

Raz stworzony model, szczególnie jeśli jest otwarty, nie jest nigdy finalnym produktem – może być dalej dostrajany. W przypadku modeli Bielik są już pierwsze przykłady takiego dostrajania. Formalnie rzecz biorąc, dostrajanie (fine tuning) oznacza dodatkowe szkolenie modelu na dużym zbiorze danych, w celu stworzenia jakościowo innego modelu. Dostosowywanie modelu może też odbyć się poprzez tworzenie warstw adaptacyjnych. Wówczas główny model jest już "zamrożony", a tworzy się jedynie dodatkową warstwę, przy użyciu małego zbioru specjalistycznych danych. Jeszcze inną, coraz bardziej popularną metodą jest zastosowanie mechanizmu RAG (Retrieval Augmented Generation). Wiąże się to ze stworzeniem dodatkowej, zewnętrznej bazy wiedzy, którą model wykorzystuje w trakcie generowania odpowiedzi na zapytanie.

Dostrajanie może się odbywać na przykład na danych konkretnej firmy, samorządu lokalnego lub z określonej dziedziny wiedzy. Jeden z naszych rozmówców podaje przykład modelu

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

dostosowanego do specyficznej wiedzy przydatnej dla lokalnych wspólnot energetycznych – i wyobraża sobie tego rodzaju modele jako ważne narzędzia w rozwiązywaniu różnych wyzwań i problemów społecznych. Innym, realizowanym przykładem dostrajania jest inicjatywa firmy theLion.ai, która zamierza doszkolić model Bielik na bazie instrukcji, dzięki którym powstanie polski medyczny model językowy.[40]

Otwarte, dostrojone modele, mogą być przetwarzane na własnych serwerach firmy lub organizacji, gwarantując większe bezpieczeństwo danych. Twórcy modeli wyobrażają sobie również dostrajanie modeli do potrzeb indywidualnych użytkowników, w celu tworzenia spersonalizowanych asystentów. Wiąże się to z rozwiązaniami technologicznymi, pozwalającymi minimalizować modele, tak by działały na urządzeniach typu laptop czy smartphone.

Siła społeczności i tworzenie ekosystemu polskich AI

W rozmowach z ekspertami usłyszeliśmy, że SpeakLeash to społeczność zaangażowana w rozwój modelu języka polskiego, który jest lokalnie istotny i świadomy kulturowo. Podkreślaliśmy to już na wcześniejszych stronach tego raportu, ale uznaliśmy, że warto przyjrzeć się temu aspektowi z bliska. Społeczność ta kładzie nacisk na otwarte podejście, udostępniając swoje zbiory danych i modele innym, aby mogli korzystać i trenować własne modele. Sprzyja to środowisku współpracy, w którym osoby i organizacje mogą przyczynić się do rozwoju modeli języka polskiego.

Kluczowym aspektem sukcesu SpeakLeash jest zatem jego otwarty i zorientowany na społeczność charakter. Symbolem tej współpracy jest zdecentralizowany kanał na platformie Discord, zastępujący dalszą instytucjonalizację SpeakLeash. Inicjatywa przyciągnęła ponad 1000 członków, w tym ekspertów i entuzjastów, którzy wnoszą wiedzę, spostrzeżenia, czy… memy. Społeczność stojąca za Bielikiem wyłoniła się organicznie – ze wspólnego pragnienia stworzenia wysokiej jakości, zróżnicowanego kulturowo modelu języka polskiego. W społeczności tej narzędzia i rozwiązania powstają w sposób zdecentralizowany, typowy dla projektów otwartego oprogramowania. Model Bielik jest ściśle związany z projektem SpeakLeash, jednak powstał w dużo węższym gronie kilku programistów i ekspertów od uczenia maszynowego.

Jak już pisaliśmy, inicjatorzy Bielika myślą szerzej – nie tylko o społeczności, ale też o polskim ekosystemie AI. W ich wypowiedziach inne polskie modele są traktowane nie jako konkurencja, lecz jako równolegle prowadzone inicjatywy, służące tym samym celom. Zakładają bowiem, że wszystkie te projekty będą wymieniać się wiedzą naukową, zestawami danych czy samymi modelami.

Jednocześnie widać, że w chwili obecnej inicjatywa jest w fazie kluczowej zmiany, w której rozwój ekosystemu oznacza już nie tylko budowanie społeczności twórców AI, ale też relacji z innymi podmiotami. Mogą to być podmioty posiadające cenne dla SpeakLeash dane, albo organizacje które mogłyby stać się użytkownikami modelu Bielik, demonstrując jego użyteczność i dostosowanie do polskiego kontekstu.

---

[40] 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.

Model PLLuM jest tworzony w dużo bardziej tradycyjnym, zamkniętym modelu produkcji, i czerpie z zasobów instytucji badawczych wchodzących w skład konsorcjum. Doświadczenia PLLuM pokazują, że taki model ma przewagi, gdy chodzi na przykład o nawiązywanie współprac z innymi organizacjami – PLLuM ma pod tym względem więcej sukcesów niż SpeakLeash. Jednocześnie kierownicy konsorcjum sygnalizują, że myślą o dalszym rozwoju PLLuM w kontekście szerszego ekosystemu. Zidentyfikowali też zadania – dotyczące na przykład tworzenia zbioru instrukcji – dla których niezbędne jest podejście partycypacyjne, angażujące szersze grono osób o różnych kompetencjach.

WYZWANIA W ROZWOJU LOKALNYCH MODELI AI

Moc obliczeniowa jako bariera i motywator

Usługi przetwarzania w chmurze umożliwiły szkolenie i wdrażanie skutecznych modeli sztucznej inteligencji bez konieczności tworzenia własnych centrów danych. Jednocześnie uzależniają twórców AI od niewielkiego grona hyperscalers (dostawców rozwiązań hiperskalowych) dysponujących odpowiednią infrastrukturą. O ile szkolenie dużych modeli językowych wymaga znacznej mocy obliczeniowej, rozwój mniejszych modeli można osiągnąć przy użyciu skromniejszych zasobów. Ograniczona dostępność mocy obliczeniowej napędza też innowacje w zakresie bardziej wydajnych architektur trenowania modeli.

Obecnie niejasne jest, czy na dłuższą metę przewaga w mocy obliczeniowej przesądzi o dominującej pozycji największych modeli. Wśród naszych rozmówców byli zarówno pesymiści co do długofalowego powodzenia alternatyw, jak i osoby wierzące w trend tworzenia mniejszych, wydajnych modeli.

Bariery: szkolenie wysokiej jakości LLM wymaga znacznej mocy obliczeniowej, co przekłada się na wysokie zużycie energii i wymaga solidnej infrastruktury obliczeniowej. Stanowi to wyzwanie, zwłaszcza dla mniejszych podmiotów, dysponujących ograniczonymi zasobami. Jeden z rozmówców zwrócił uwagę na rozbieżność w mocy obliczeniowej dostępnej w centrach superkomputerowych w Polsce w porównaniu z zasobami, pojedynczej globalnej firmy AI. Sugeruje, że ta różnica w skali może w dłuższej perspektywie utrudniać rozwój lokalnie wyszkolonych LLM, które mogą konkurować z tymi produkowanymi przez korporacje.

Motywatory: zapotrzebowanie na moc obliczeniową może napędzać innowacje. Przykładem może być współpraca polskiego środowiska badawczego z Cyfronetem. To przykład tego, jak partnerstwa i dzielenie się zasobami mogą pokonać bariery infrastrukturalne. Co więcej, skupienie się na mniejszych, bardziej wyspecjalizowanych LLM może zmniejszyć zależność od ogromnej mocy obliczeniowej. Jeden z rozmówców twierdzi, że firmy często nie potrzebują modeli, które potrafią wszystko; zamiast tego potrzebują modeli dostosowanych do konkretnych zadań. Podejście to kładzie nacisk na wydajność i znalezienie równowagi między rozmiarem modelu, wydajnością i kosztami obliczeniowymi szkolenia i wdrożenia.

Pozyskiwanie danych dobrej jakości

Dostęp do wysokiej jakości danych jest jednym z kluczowych wyzwań stojących przed twórcami modeli językowych. Twórcy zbiorów szkoleniowych dla AI traktują swoją pracę jako kolejny krok w procesie transformacji cyfrowej: udostępniania coraz większej ilości zasobów w postaci cyfrowej, do maszynowego przetwarzania. Jednocześnie widać wyraźne bariery ograniczające dostęp do wielu zbiorów, szczególnie tych wysokiej jakości. Dysproporcje w dostępie do danych to jeden z czynników dających przewagę dużym firmom, które posiadają własne zbiory danych oraz środki finansowe na inwestycje w ich nowe źródła.

Bariery: w przeciwieństwie do dużych korporacji technologicznych, polscy twórcy modeli często borykają się z ograniczonymi zasobami na rozwój zbiorów danych. Sytuację komplikują przepisy utrudniające dostęp do danych, szczególnie w newralgicznych obszarach, takich jak medycyna. Drugim źródłem wyzwań są kwestie prawa autorskiego: niejasny status szkolenia modeli oraz oczekiwania wielu posiadaczy zasobów wysokiej jakości, by wszystkie użycia były licencjonowane i płatne.

Motywatory: często słyszeliśmy, że istotna jest "jakość nad ilością", czyli skupienie się na danych wysokiej jakości, nawet w mniejszych zbiorach. Znalezienie sposobów ułatwienia odpowiedzialnego dostępu do danych, zwłaszcza na potrzeby badań i rozwoju, będzie mieć kluczowe znaczenie dla wspierania innowacji w zakresie sztucznej inteligencji. Potrzeba metod zarządzania danymi, które zrównoważą potrzebę posiadania wystarczająco dużych, różnorodnych zbiorów danych z koniecznością ochrony troski o prawa autorskie, ochrony wrażliwych danych osobowych oraz zapewnienia odpowiednich modeli biznesowych. Wsparciem dla rozwoju AI byłoby też udostępnianie większych ilości kolekcji w sposób otwarty, w ramach takich inicjatyw jak Otwarty Dostęp do publikacji akademickich czy otwieranie zasobów dziedzictwa. Zbiory danych tworzone na potrzeby otwartych modeli mają wreszcie szansę lepiej spełniać wymogi unijnych regulacji, takich jak AI Act – dotyczące m.in. przejrzystości danych i uregulowania kwestii prawnoautorskich.

Regulacje faworyzujące istniejące monopole

W naszych rozmowach pojawiły się obawy, że regulacje, choć mają na celu zapewnienie etycznego rozwoju sztucznej inteligencji, mogą w sposób niezamierzony faworyzować duże przedsiębiorstwa technologiczne, które dysponują zasobami umożliwiającymi poruszanie się w skomplikowanych ramach prawnych. Mniejsze przedsiębiorstwa i instytucje badawcze mogą mieć trudności z przestrzeganiem rygorystycznych przepisów, co może utrudniać innowacje i konkurencję w środowisku sztucznej inteligencji.

Bariery: Jeden z mówców podkreśla napięcie między rygorystycznymi przepisami dotyczącymi ochrony danych, takimi jak RODO, a potrzebą dostępu do danych w rozwoju sztucznej inteligencji, szczególnie w takich dziedzinach jak medycyna. Twierdzą, że choć duże korporacje mogą sobie pozwolić na obejście tych przepisów i finansowanie własnych badań, mniejsze podmioty często mają z tym trudności. Sugeruje to, że znalezienie sposobów ułatwienia odpowiedzialnego dostępu do danych, zwłaszcza na potrzeby badań i rozwoju, mogłoby mieć kluczowe znaczenie dla wspierania szerzej zakrojonych innowacji w zakresie sztucznej inteligencji. Jednocześnie, wysokie koszty przestrzegania rygorystycznych przepisów mogą ograniczać innowacje i konkurencję w sektorze AI, faworyzując istniejące monopole.

Motywatory: Konieczne jest opracowanie bardziej elastycznych regulacji, które uwzględniają zróżnicowane potrzeby i możliwości różnych interesariuszy, szczególnie poprzez wspieranie mniejszych graczy - zapewnianie wsparcia i zachęt mniejszym firmom i instytucjom badawczym do angażowania się w rozwój sztucznej inteligencji. Skutkiem takiego podejścia może być łatwiejszy dostęp do danych, szczególnie do celów badawczo-rozwojowych.

O RAPORCIE

Open Future to europejski think tank skupiający się na nowych podejściach do tworzenia otwartego Internetu, które maksymalizują korzyści społeczne wynikające ze współdzielenia danych, wiedzy i kultury.

Fundacja Centrum Cyfrowe to think-and-do tank dbający o społeczny wymiar technologii. Obszarem zainteresowań Centrum jest cyfrowy wymiar spraw publicznych w Polsce, a konkretniej analizy zmian społecznych, kulturowych i gospodarczych związanych z technologią cyfrową – a co za tym idzie, wspieranie rozwoju wiedzy w tym zakresie.

Alek Tarkowski jest dyrektorem ds. strategii w Open Future. Posiada doktorat z socjologii z Polskiej Akademii Nauk. Od piętnastu lat zajmuje działaniami rzeczniczymi i budowaniem ruchów społecznych na rzecz technologii działających w interesie publicznym. Zajmuje się też badaniami na styku kwestii społecznych, kultury i technologii cyfrowych. Jego obecne zainteresowania obejmują polityki na rzecz publicznego AI i zarządzanie zbiorami danych.

Kuba Piwowar jest socjologiem i kulturoznawcą, doktorem kulturoznawstwa. Jest również starszym stypendystą Humanity in Action, gdzie pracował nad projektem dotyczącym korzystania z danych i aktywizmu danych. Ponadto jest adiunktem w Katedrze Kultury i Mediów Uniwersytetu SWPS w Warszawie. W latach 2008-2024 pracował w Google, początkowo jako analityk, a następnie jako doradca kluczowych partnerów biznesowych.

Michał Owczarek jest doktorantem kulturoznawstwa na Uniwersytecie SWPS, gdzie bada historię mediów w Polsce. Obronił pracę magisterską z socjologii cyfrowej dotyczącą konfliktów między państwami i platformami dotyczących infrastruktury cyfrowej. Interesuje się również studiami miejskimi, w szczególności wpływem technologii cyfrowych na tkankę miejską.

Autorzy raportu dziękują rozmówcom, którzy opowiedzili o rozwoju polskich LLMów: Pawłowi Cyrcie, Adrianowi Gwoździejowi, Janowi Koconiowi, Sebastianowi Kondrackiemu, Markowi Kozłowskiemu, Jackowi Nagłowskiemu i Maciejowi Piaseckiemu.

W trakcie analizy i podczas pisania raportu skorzystaliśmy z szeregu podejść i narzędzi analitycznych: od tradycyjnej analizy treści wywiadów przez skorzystanie z istniejących LLMów (Gemini, ChatGPT, NotebookLM). Chcieliśmy sprawdzić, na ile wnioski, do których dochodzimy własnoręcznie, możemy uzupełnić tym, co spostrzegły modele. Choć jesteśmy zadowoleni z tej współpracy człowieka z maszynami, liczymy, że niedługo podobne zadania będziemy mogli równie sprawnie wykonywać przy użyciu polskich modeli językowych.