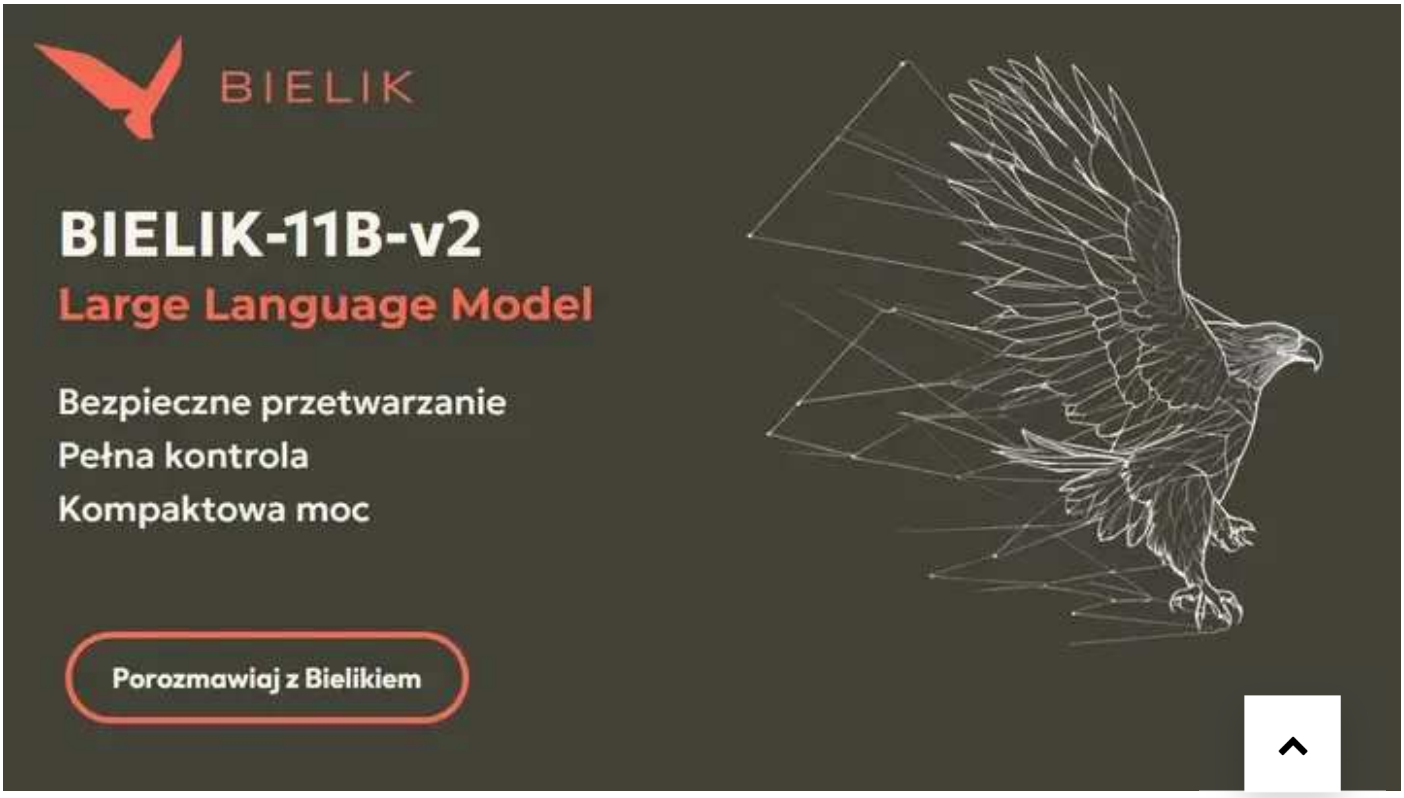
PL

Aktualności

Bielik – polski model językowy powstał w AGH

28/08/2024

sztuczna inteligencja współpraca



Akademickie Centrum Komputerowe Cyfronet AGH udostępniło zasoby obliczeniowe dwóch najszybszych aktualnie superkomputerów w Polsce – Heliosa i Atheny – do stworzenia Bielika – polskiego modelu językowego.

Bielik-11B-v2 – nowy polski duży model językowy

Bielik powstał w efekcie prac zespołu działającego w ramach Fundacji SpeakLeash oraz Akademickiego Centrum Komputerowego Cyfronet AGH i jest polskim modelem z kategorii LLM (z ang. Large Language Models), tj. dużym modelem językowym, posiadającym 11 miliardów parametrów.

SpeakLeash – grupa pasjonatów i twórców Bielika

SpeakLeash to fundacja, która połączyła ludzi bardzo różnych profesji. Grupa entuzjastów za cel postawiła sobie stworzenie największego polskiego zbioru danych tekstowych wzorując się na zagranicznych inicjatywach jak The Pile. W skład zespołu projektowego wchodzą przede wszystkim pracownicy polskich przedsiębiorstw, badacze z ośrodków naukowych oraz studenci kierunków związanych z obszarami sztucznej inteligencji. Prace zespołu nad polskim modelem językowym trwały ponad rok, a ich pierwotny zakres obejmował m.in. zbieranie danych, ich przetwarzanie oraz klasyfikację.

> *– Najtrudniejsze zadanie polegało na pozyskaniu danych w języku polskim. Musimy operować wyłącznie na danych źródłowych, co do których mamy pewność, jakie jest ich pochodzenie* – tłumaczy pomysłodawca Bielika, Sebastian Kondracki ze SpeakLeash.

Aktualnie zasoby fundacji SpeakLeash są największym, najlepiej

opisanym i udokumentowanym zbiorem danych w języku polskim.

Helios i Athena – moce obliczeniowa dla nauki

Superkomputery z Akademickiego Centrum Komputerowego Cyfronet AGH pozwoliły projektowi Bielik rozwinąć skrzydła.

Współpraca kadry z AGH z fundacją Speakleash umożliwiła wykorzystanie odpowiednich mocy obliczeniowych niezbędnych do stworzenia modelu i wsparcie zespołu SpeakLeash niezbędną wiedzą ekspercką oraz naukową gwarantując sukces wspólnego projektu.

Wsparcie zespołu ACK Cyfronet dotyczyło optymalizacji i skalowania procesów treningowych, prac nad potokami przetwarzania danych oraz rozwoju i działania metod generowania danych syntetycznych, a także prac w zakresie metod testowania modeli. Wynikiem tego jest Polski ranking modeli (Polish OpenLLM Leaderboard). Cenne doświadczenia i wiedza zebrane w wyniku tej współpracy umożliwiły zespołowi ekspertów PLGrid przygotowanie wytycznych oraz zoptymalizowanych rozwiązań w tym środowisk obliczeniowych do prac z modelami językowymi na bazie klastrów Athena i Helios dla potrzeb użytkowników naukowych.

> – *Zasoby Heliosa, najszybszej aktualnie maszyny w Polsce, wykorzystaliśmy do uczenia modeli językowych* – precyzuje Marek Magryś, zastępca Dyrektora ACK Cyfronet AGH ds. Komputerów Dużej Mocy. – *Nasza rola polega na wsparciu wiedzą ekspercką, doświadczeniem i przede wszystkim mocą obliczeniową procesu katalogowania, zbierania, przetwarzania danych oraz na wspólnym przeprowadzeniu procesu uczenia modeli językowych. Dzięki pracy zespołu SpeakLeash i AGH udało nam się stworzyć Bielika, model LLM, który doskonale radzi sobie z naszym językiem oraz kontekstem kulturowym i który może być kluczowym elementem łańcuchów przetwarzania danych tekstowych dla naszego języka w zastosowaniach naukowych i biznesowych. Potwierdzeniem jakości Bielika są wysokie lokaty uzyskane przez model na listach rankingowych dla języka polskiego.*

Moc obliczeniowa Heliosa i Atheny w tradycyjnych symulacjach komputerowych to łącznie ponad 44 PFLOPS, a dla obliczeń z zakresu sztucznej inteligencji w niższej precyzji to aż 2 EFLOPS.

> – *Jeśli operujemy tak dużymi danymi jak w przypadku projektu Bielik to oczywiście infrastruktura potrzebna do pracy przekracza zdolności zwykłego komputera. Musimy dysponować mocą obliczeniową potrzebną tylko do tego żeby przygotowywać dane, porównywać je ze sobą, trenować modele. Bariera dostępności tego typu superkomputerów powoduje, że mało która firma jest w stanie takie prace prowadzić samodzielnie. Szczęśliwie AGH dysponuje takim zapleczem* – wyjaśnia prof. Kazimierz Wiatr, Dyrektor ACK Cyfronet AGH.

Równolegle z zasobów superkomputerów z ACK Cyfronet AGH korzysta kilka tysięcy naukowców reprezentujących wiele dziedzin. Zaawansowane modelowanie i obliczenia numeryczne są wykorzystywane głównie w zakresie: chemii, biologii, fizyki, medycyny i technologii materiałowej, a także astronomii, geologii i ochrony środowiska. Superkomputery w Cyfronecie dostępne w ramach infrastruktury PLGrid są również wykorzystywane na potrzeby fizyki wysokich energii (projekty ATLAS, LHCb, ALICE i CMS), astrofizyki (CTA, LOFAR), nauk o Ziemi (EPOS), europejskiego źródła spalacyjnego (ESS), badań fal grawitacyjnych (LIGO/Virgo) czy biologii (WeNMR).

> – *Wykorzystujemy do trenowania Bielika dwa najszybsze superkomputery w Polsce, Athenę i Heliosa, ale i tak w porównaniu z infrastrukturą światowych liderów mamy dużo mniejsze zaplecze. Do tego, w tym samym czasie z zasobów superkomputerów korzysta kilkuset innych użytkowników* – wyjaśnia Marek Magryś. – *Nasze systemy umożliwiają jednak przeprowadzenie w kilka godzin lub dni obliczeń, które na zwykłych komputerach mogłyby trwać lata lub, w niektórych przypadkach, nawet stulecia.*

Bielik a chat GPT – podstawowe różnice

> – *Zbiór danych zasilających Bielika cały czas rośnie, jednak trudno*

> *będzie nam się ścigać z zasobami wykorzystywanymi przez inne modele, które funkcjonują w języku angielskim. Poza tym liczba treści w Internecie, która funkcjonuje w języku polskim jest znacznie mniejsza niż w angielskim* – wyjaśniają twórcy.

Najbardziej popularnym produktem wykorzystującym duży model językowy jest ChatGPT, który powstał w oparciu o zasoby firmy OpenAI. Konieczność opracowywania modeli językowych w różnych innych językach znajduje jednak swoje uzasadnienie.

Marek Magryś, zastępca Dyrektora ACK Cyfronet AGH ds. Komputerów Dużej Mocy podkreśla:

> *– O ile ChatGPT potrafi mówić w języku polskim, to nasycony jest treściami w języku angielskim. W związku z tym ma nikłe pojęcie na temat np. polskiej kultury czy niuansów polskiej literatury. Nie do końca też sobie radzi ze zrozumieniem logiki bardziej skomplikowanych tekstów np. prawnych czy medycznych. Jeśli chcielibyśmy zastosować go w tych właśnie specjalistycznych obszarach i mieć model językowy, który dobrze rozumuje w języku polskim i odpowiada poprawną polszczyzną, to nie możemy opierać się wyłącznie na zagranicznych modelach językowych.*

Wersja, którą mogę testować użytkownicy jest utrzymywana nieodpłatnie w domenie publicznej i jest wciąż udoskonalana. Autorzy udostępnili, oprócz pełnych wersji opracowanych modeli, także całą gamę wersji skwantyzowanych w najpopularniejszych dostępnych formatach, które umożliwiają uruchomienie modelu na własnym komputerze.

> *– Warto wiedzieć, że Bielik będzie bardzo dobrze sprawdzał się w zakresie np. streszczania treści. Już w tym momencie nasz model ma swoją użyteczność w obszarze naukowym oraz biznesowym, może służyć na przykład do usprawnienia komunikacji z użytkownikami podczas obsługi zgłoszeń w Helpdesku* – wyjaśnia Szymon Mazurek z ACK Cyfronet AGH.

Dlaczego warto budować polskie modele

językowe?

Twórcy Bielika wyjaśniają, że usługi sztucznej inteligencji funkcjonujące w Internecie, w tym te najpopularniejsze jak ChatGPT, utrzymywane są na serwerach zewnętrznych. Jeśli jakaś firma czy branża rozwija rozwiązanie, które operuje na specjalistycznych danych np. medycznych lub na tekstach, które z różnych powodów nie mogą opuścić firmy, np. są poufne, to jedyną możliwością jest uruchomienie takiego modelu u siebie. Ten model nie będzie tak doskonały jak ChatGPT, ale nie musi też być tak bardzo ogólny.

Dodatkową korzyścią płynącą z uruchomienia modeli językowych typu Bielik jest wzmocnienie pozycji Polski w obszarze innowacji w sektorze AI. Ponadto, co podkreślają twórcy, warto zabiegać o to, aby budować własne narzędzia i tym samym uniezależniać się od zewnętrznych firm, które w przypadku zawirowań na rynkach, regulacji czy ograniczeń prawnych mogą np. uniemożliwiać dostęp do swoich zasobów. Tym samym rozwijając i udoskonalając narzędzia w Polsce, budujemy stabilne zaplecze i jesteśmy w stanie zabezpieczać wiele naszych sektorów – bankowy, administracyjny medyczny lub prawniczy.

> – *Intensywne działania nad rozwojem AI, modelami językowymi typu Bielik czy innymi narzędziami opartymi o sztuczną inteligencję są w interesie wszystkich dobrze działających gospodarek. Obserwujemy wzmożone prace nad tego typu rozwiązaniami w wielu krajach* – wyjaśnia Jan Maria Kowalski z Fundacji Speakleash.

***

Ostatni fragment, wyjaśniający dlaczego warto opracowywać modele językowe w językach innych niż angielski został wygenerowany przez Bielika:

- Opracowywanie modeli językowych w językach innych niż angielski może przynieść wiele korzyści, zarówno w badaniach lingwistycznych, jak i w praktycznych zastosowaniach technologi Oto kilka powodów, dla których warto to robić:

- Lepsze zrozumienie różnorodności językowej: modele językowe oparte na angielskim mogą nie uwzględniać w pełni specyfiki innych języków, co prowadzi do uprzedzeń lub niedokładności w rozumieniu i tłumaczeniu treści w innych językach. Rozwijanie modeli dla innych języków pozwala na lepsze zrozumienie ich gramatyki, słownictwa i kontekstu kulturowego.
- Poprawa jakości tłumaczeń: modele językowe odgrywają kluczową rolę w automatycznym tłumaczeniu tekstów. Opracowując modele dla mniej popularnych języków, można poprawić jakość tłumaczeń, eliminując błędy i niejasności wynikające z braku zrozumienia specyfiki tych języków.
- Wsparcie dla społeczności językowych: rozwijanie modeli językowych dla mniejszościowych języków może mieć pozytywny wpływ na te społeczności, ułatwiając komunikację i dostęp do informacji w ich rodzimym języku.
- Otwieranie drzwi do badań nad różnorodnością językową: modele językowe mogą pomóc w badaniu podobieństw i różnic między językami, co jest kluczowe dla zrozumienia ewolucji i dywergencji języków na całym świecie.
- Ulepszanie narzędzi AI: rozwój modeli językowych w różnych językach poszerza zakres zastosowań technologii AI, umożliwiając tworzenie bardziej zaawansowanych systemów wykorzystujących języki naturalne w sposób bardziej precyzyjny i trafny.
- Chronienie językowej różnorodności świata: popularyzacja języków obcych często prowadzi do marginalizacji lokalnych dialektów i języków. Rozwijanie modeli językowych dla mniej znanych języków może pomóc w ochronie różnorodności językowej na całym świecie.

Przydatne linki:

- Możliwość testowania Bielika: bielik.ai
- Aktualna liczba zebranych danych: speakleash.org/dashboard

Powrót