Przetestowaliśmy polski model PLLuM

25 lutego 2025

Polska dołącza do grona krajów Unii Europejskiej, które mogą pochwalić się własnym, w pełni funkcjonalnym modelem LLM – PLLuM. Ten model ma potencjał, by zmienić sposób, w jaki polska administracja publiczna wykorzystuje sztuczną inteligencję. Przetestowaliśmy go w praktyce i opisujemy jego możliwości z technicznej perspektywy – bez koloryzowania rzeczywistości czy faworyzowania polskich osiągnięć. Nasza analiza koncentruje się na realnych funkcjonalnościach PLLuM, jego architekturze oraz możliwych zastosowaniach zarówno w sektorze publicznym, jak i prywatnym.

Aby lepiej zrozumieć kontekst, cofnijmy się w czasie. Wiele osób słyszało o polskim modelu LLM „Bielik", opracowanym na AGH i przedstawianym jako przełomowa innowacja. W rzeczywistości był to jedynie proces dostosowania istniejącego francuskiego modelu Mistral do języka polskiego poprzez fine-tuning na niewielkim zbiorze danych. Trudno nazwać to nowatorskim osiągnięciem – „Bielik" nie wprowadził żadnych zmian w architekturze modelu ani nie rozwinął struktury Transformer, która zrewolucjonizowała dziedzinę uczenia głębokiego dzięki artykułowi „Attention Is All You Need" opublikowanemu przez zespół Google w 2017 roku.

Mistral, stworzony przez francuski startup, również nie wprowadza przełomów architektonicznych, ale znalazł swoją niszę w postaci tanich modeli LLM o ograniczonych zdolnościach rozumowania, przeznaczonych do prostych zastosowań. Firma opracowała szeroki katalog modeli wyspecjalizowanych w konkretnych zadaniach, takich jak programowanie – przykładem jest model Codestral. Rozwiązania te często wspierają edytory kodu, takie jak Visual Studio Code czy Cursor, w podstawowych zadaniach: podpowiadaniu kodu, refaktoryzacji czy generowaniu

komentarzy do funkcji i klas. To właśnie nazywam zastosowaniami na poziomie podstawowym.

Szczegółowa i przejrzysta dokumentacja Mistrala (dostępna pod adresem https://docs.mistral.ai/guides/finetuning/) znacznie ułatwia tworzenie modeli takich jak Bielik, precyzyjnie opisując proces fine-tuningu. „Bielik”, choć opracowany przez polski zespół, jest w gruncie rzeczy adaptacją Mistrala, wzbogaconą o polską literaturę i leksykony, co pozwala mu lepiej radzić sobie z językiem polskim. Warto jednak zauważyć, że każdy liczący się model LLM zazwyczaj zaczyna się od pre-printu – wstępnej wersji artykułu naukowego przed formalną recenzją w czasopiśmie. W przypadku Bielika takiego dokumentu brakuje.

Co więcej, Bielik powstał dzięki pracy hobbystów, a jego rozwój wspierała komercyjna firma Devinti. Projekt, wraz ze stroną internetową, nie był aktualizowany od 2024 roku – nawet stopka strony pozostaje nietknięta. Niedawno jeden z użytkowników „GitHuba” zgłosił ticket w repozytorium Bielika, zwracając uwagę na brak aktualizacji daty w stopce (https://github.com/speakleash/Bielik-how-to-start/issues/60). Zaskakujące jest również to, że twórcy nie zdecydowali się na stworzenie usługi SaaS dla Bielika ani na rozwijanie go w kolejne wersje. Tymczasem wszystkie znaczące startupy – takie jak Deepseek, Mistral czy giganci pokroju OpenAI i X Corp. – oferują swoje modele za pośrednictwem API lub własnych agentów AI, generując z tego przychody. Dzięki temu użytkownicy nie muszą samodzielnie hostować modeli ani polegać na zewnętrznych dostawcach.

Sprawdzamy rządowy model

PLLuM reprezentuje przełomowe osiągnięcie w dziedzinie sztucznej inteligencji, stanowiąc pierwszą rodzinę zaawansowanych modeli językowych zaprojektowanych specjalnie z myślą o języku polskim. W przeciwieństwie do wcześniejszych

prób adaptacji istniejących modeli, PLLuM został zbudowany od podstaw z uwzględnieniem specyfiki języków słowiańskich i bałtyckich, jednocześnie zachowując zdolność do przetwarzania języka angielskiego.

Z technicznego punktu widzenia, architektura PLLuM opiera się na najnowszych osiągnięciach w dziedzinie transformerów, wykorzystując zaawansowane techniki optymalizacji i skalowania. Kluczowym wyróżnikiem jest zastosowanie innowacyjnego podejścia do tokenizacji, które lepiej radzi sobie z morfologiczną złożonością języków słowiańskich. Model wykorzystuje również adaptywne mechanizmy uwagi (ang. attention), które zostały zoptymalizowane pod kątem długich sekwencji tekstowych charakterystycznych dla dokumentów administracyjnych.

Baza treningowa modelu obejmuje około 150 miliardów tokenów wysokiej jakości tekstu w języku polskim. To znacząco większy zbiór danych niż w przypadku wcześniejszych polskich modeli językowych. Co więcej, dane zostały starannie wyselekcjonowane i oczyszczone, ze szczególnym uwzględnieniem poprawności językowej i różnorodności tematycznej.

Na szczególną uwagę zasługuje organiczny zbiór instrukcji, obejmujący około 40 tysięcy par prompt-odpowiedź. Zbiór ten, stworzony przez zespół ekspertów językowych i dziedzinowych. W procesie jego tworzenia uwzględniono specyfikę polskiego kontekstu kulturowego i administracyjnego – aspekt kluczowy dla praktycznych zastosowań w sektorze publicznym.

Z technicznego punktu widzenia, PLLuM oferuje szeroką gamę wariantów modelu:

– modele bazowe (8B, 12B parametrów) – zoptymalizowane pod kątem efektywności obliczeniowej i zastosowań edge computing;

– model rozproszony (8x7B) – wykorzystujący architekturę mixture-of-experts do równoległego przetwarzania;

– model pełnowymiarowy (70B) – konkurujący z największymi światowymi modelami pod względem możliwości rozumowania.

Opracowane testy porównawcze koncentrują się na konkretnych wyzwaniach polskiej administracji publicznej, w tym interpretacji przepisów prawnych i analizie dokumentów urzędowych. Według autorów PLLuM znacząco przewyższa zagraniczne modele dostosowane do języka polskiego w tych specjalistycznych zadaniach.

Najnowsza aktualizacja benchmarku, stworzonego przez zespół badawczy (Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, Rafał Poświata), uwzględnia modele PLLuM-12B-nc-chat oraz PLLuM-8x7B-nc-chat, przedstawiając ich faktyczne możliwości (https://huggingface.co/spaces/sdadas/plcc).

Benchmark porównuje wydajność różnych modeli językowych w sześciu kategoriach związanych z językiem polskim i wiedzą ogólną:

– sztuka i rozrywka – kreatywność i wiedza o kulturze;

– kultura i tradycja – zrozumienie kontekstów kulturowych;

– geografia – znajomość faktów geograficznych;

– gramatyka – poprawność językowa i strukturalna;

– historia – wiedza historyczna;

– słownictwo – bogactwo i precyzja językowa.

Model PLLuM-8x7B-nc-chat uzyskał średni wynik 68,17, co plasuje go poniżej czołowych modeli komercyjnych (takich jak OpenAI, Gemini czy Claude). Pod względem ogólnej wydajności jest porównywalny do DeepSeek-v3 lub GPT-4-turbo. Model szczególnie dobrze radzi sobie w kategorii kultury i tradycji (76 punktów), potwierdzając skuteczność w rozumieniu kontekstów kulturowych oraz realizując założenia specjalizacji w językach słowiańskich i bałtyckich. Wysokie wyniki osiąga

również w kategoriach geografii i historii (73 punkty), co prawdopodobnie wynika z treningu na danych pochodzących ze źródeł publicznych i administracyjnych.

Słabsze strony PLLuM-8x7B to gramatyka (47 punktów) – najniższy wynik wśród wszystkich modeli, wskazujący na problemy z poprawnością językową w złożonych konstrukcjach i niuansach gramatycznych – oraz słownictwo (68 punktów), gdzie model ustępuje precyzją modelom komercyjnym. PLLuM jest obecnie projektowany z myślą o konkretnych zastosowaniach (administracja), a nie o rywalizacji w rankingach ogólnych.

Model PLLuM-8x7B jest dostępny dla każdego pod linkiem https://pllum.clarin-pl.eu/pllum_8x7b.

Autorstwo: Piotr Bednarski
Źródło: FaktyiAnalizy.info