

BLOOM: Pomnik otwartej współpracy naukowej

BLOOM to otwarty, wielojęzyczny model językowy o 176 miliardach parametrów, dostępny dla wszystkich od pierwszego dnia.



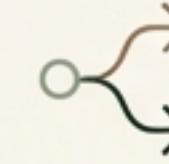
Skala projektu

Stworzony w ramach projektu BigScience – bezprecedensowej współpracy ponad 1000 badaczy z 38 krajów.



Różnorodność kompetencji

Uczestnicy to nie tylko specjalisci od uczenia maszynowego, ale także lingwiści, prawnicy, filozofowie i socjologowie.



Nowy model rozwoju

Radykalnie otwarte podejście naukowe w kontraste do zamkniętych, korporacyjnych procesów rozwoju AI.

„Krok w kierunku demokratyzacji tej potężnej technologii”

Problem: Potęga LLM za zamkniętymi drzwiami



Przed powaniem BLOOM, rozwój i dostęp do dużych modeli językowych (LLM) były ograniczone do bogatych, zamkniętych organizacji.

Konsekwencje statusu quo

- **Bariera dla nauki**
Zewnętrzni badacze byli wykluczeni z badań nad LLM, co hamowało postęp i różnorodność perspektyw.
- **Anglocentryzm**
Niemal całkowite skupienie na języku angielskim, ignorujące resztę świata.
- **Społeczne ograniczenia**
Koncentracja zasobów w instytucjach przemysłowych i brak konsultacji z interesariuszami.

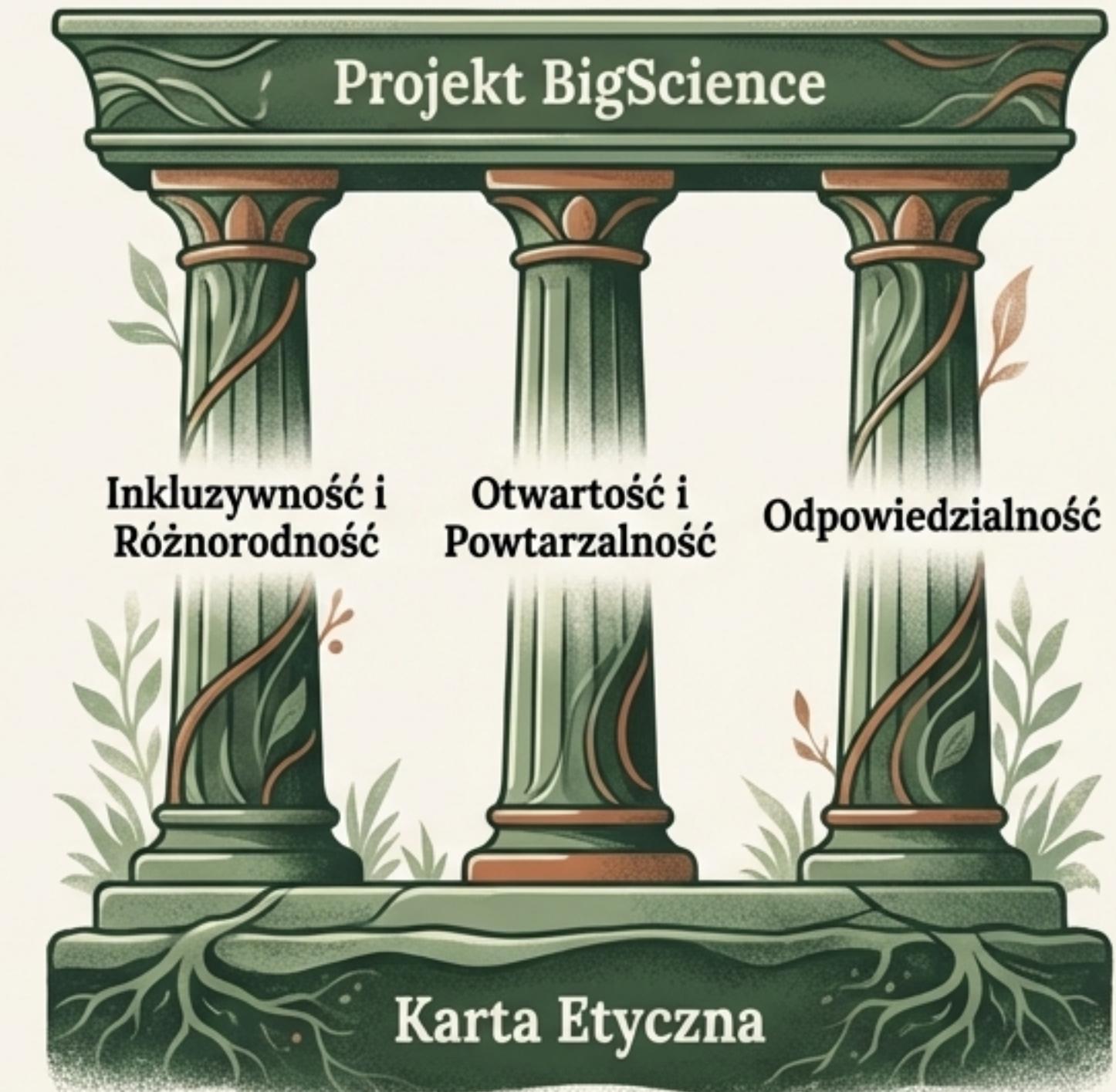
Demokratyzacja technologii LLM i stworzenie modelu, który jest nie tylko **otwarty**, ale także autentycznie **wielojęzyczny**.

Etyczny fundament: Karta, która kształtowała kod

Projekt BigScience od samego początku był rządzony przez az wspólnie opracowaną Kartę Etyczną.

To nie tylko PDF na stronie

- **Celowa kuracja danych:** Zamiast bezrefleksyjnego skrobania internetu, postawiono na ręczną, świadomą selekcję źródeł.
- **Partnerstwa oparte na wartościach:** Aktywna współpraca z kolektywami badawczymi, takimi jak **Masakhane** (języki afrykańskie) i **Latinx in AI**, w celu zapewnienia reprezentacji.
- **Świadoma inkluzywność:** Każda grupa co najmniej 3 uczestników biegłeвладающих danym językiem mogła dodać go do projektu, zobowiązując się do nadzoru nad danymi.
- **Decyzje napędzane zasadami:** Akceptowano wyższe koszty i większy nakład pracy, aby utrzymać zgodność z założeniami etycznymi, np. poprzez dbanie o prawa podmiotów danych.



Korpus ROOTS: Staranne kultywowanie danych zamiast masowego zbioru

ROOTS to zbiór 1.61 terabajta starannie wyselekcjonowanego tekstu, obejmujący 46 języków naturalnych i 13 języków programowania.

Proces, który robi różnicę



Nadzór człowieka: Proces zbierania danych był nadzorowany przez ludzi, co pozwoliło uniknąć typowych błędów automatycznych filtrów.



Unikanie uprzedzeń: W przeciwieństwie do standardowych metod, które mogą usuwać treści LGBTQ+ lub warianty języka takie jak African American English (AAE), kuracja ROOTS była świadoma i inkluzywna.

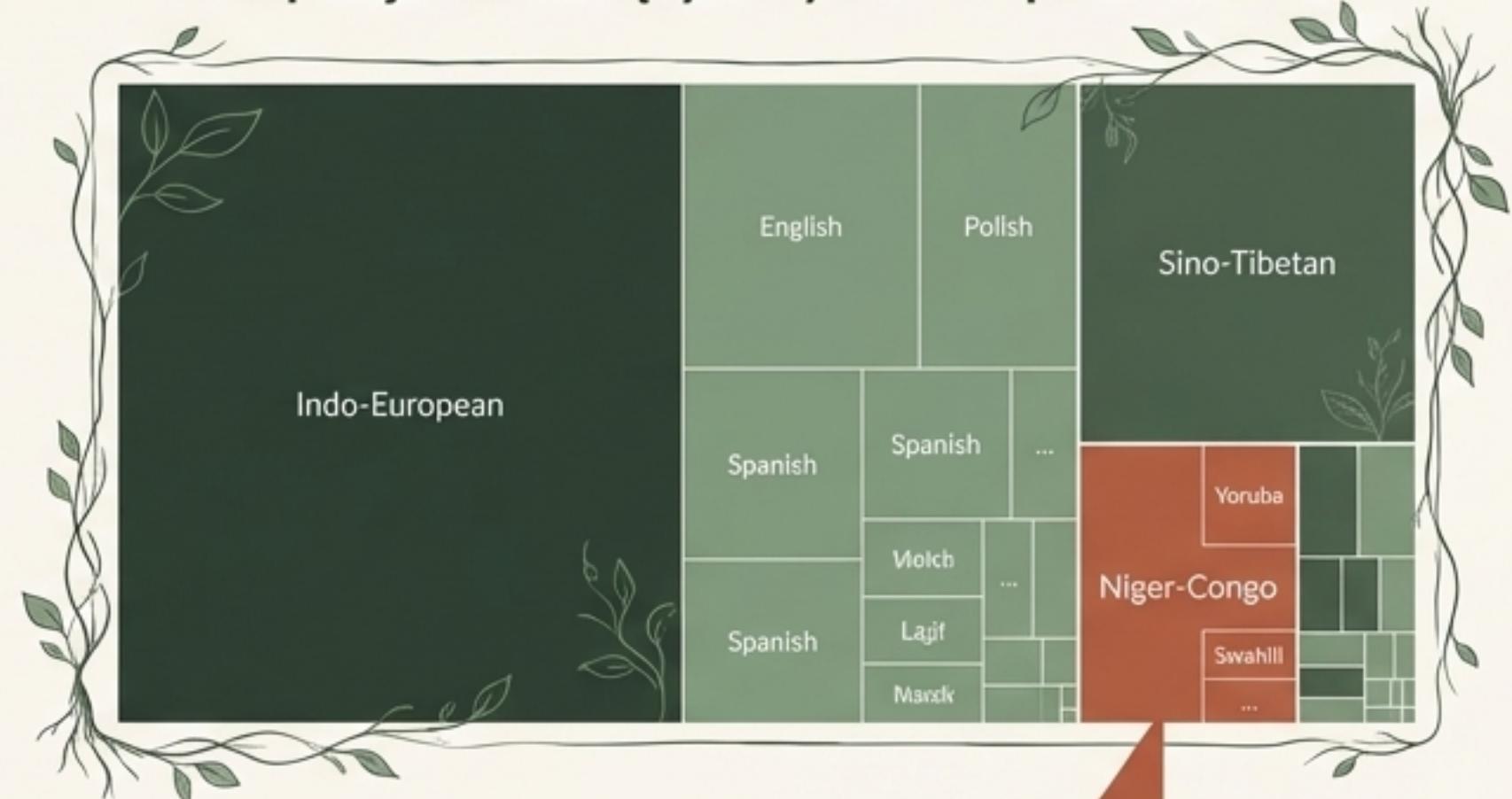


Zgody i prawa autorskie: Aktywnie pozyskiwano oficjalne zgody od wydawców na wykorzystanie danych (np. od gazety *Le Monde*).



Pełna identyfikowalność: Poszczególne źródła danych były trzymane osobno aż do końcowych etapów, aby zachować możliwość ich śledzenia i zarządzania nimi zgodnie z ich specyfiką.

Proporcje Rodzin Językowych w Korpusie ROOTS

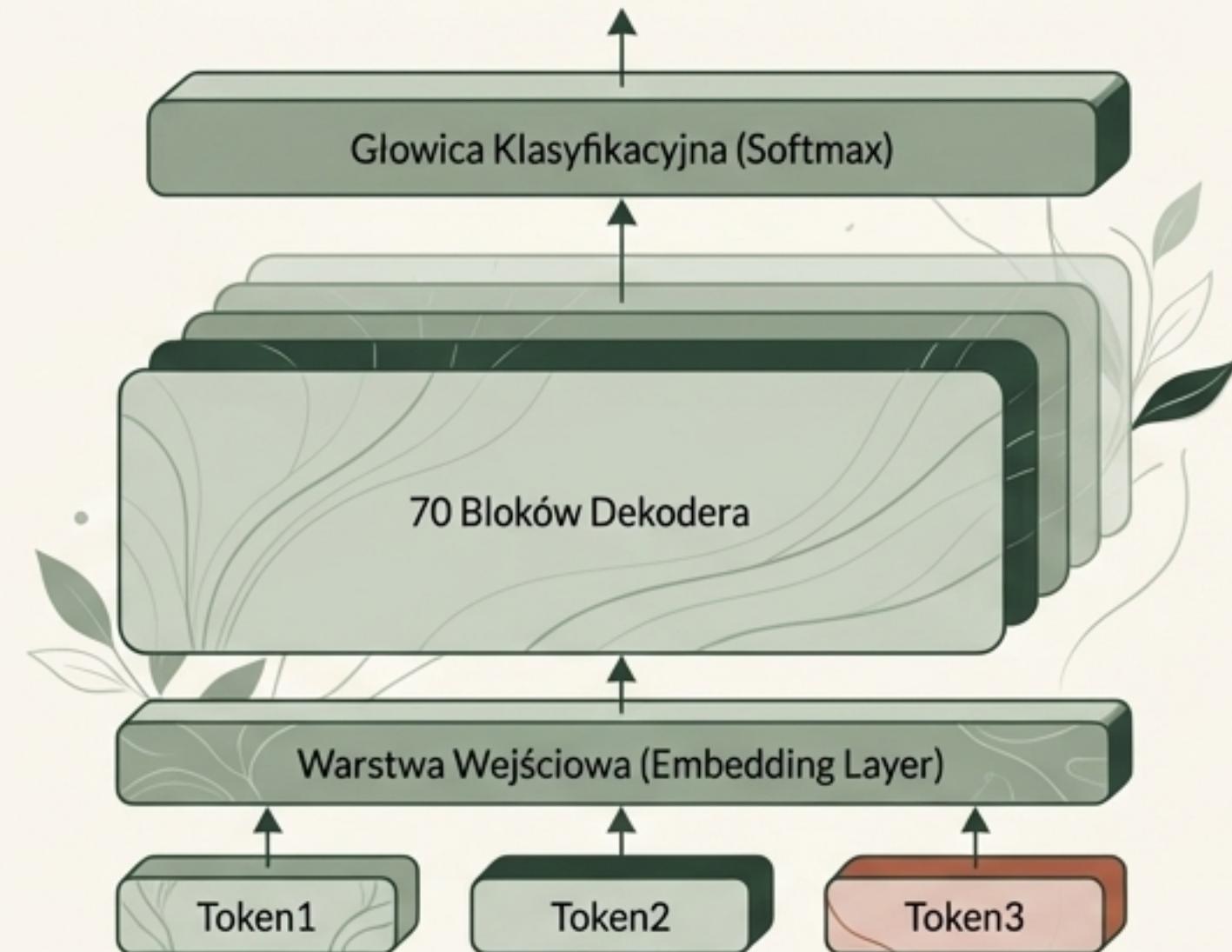


Celowa inkluzyja języków o niższych zasobach.

Architektura: Sprawdzony fundament, ulepszony z myślą o skali

Podstawowy wybór: Architektura typu **Decoder-Only Transformer**, podobna do serii GPT.

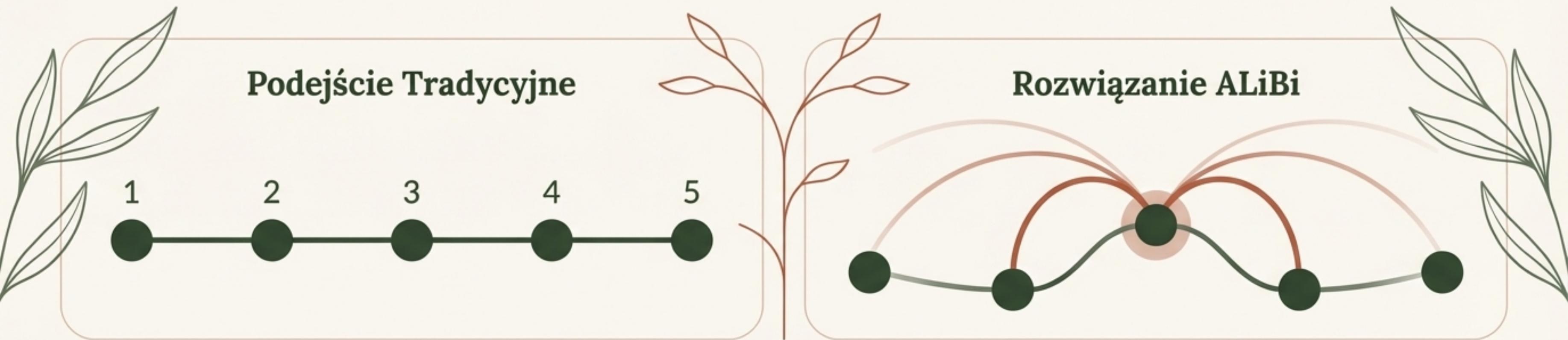
Uzasadnienie: Był to pragmatyczny wybór. Eksperymenty wykazały, że ta architektura zapewnia najlepszą generalizację w zadaniach typu zero-shot bezpośrednio po treningu.



🌿 Filozofia inżynierijna

- 🌿 Priorytetem było utrzymanie stabilności procesu trenowania na bezprecedensową skalę.
- 🌿 Zasada "nie zepsuj" ("Don't break it") była ważniejsza niż "maksymalizuj wyniki w benchmarkach".
- 🌿 Wprowadzono dwie kluczowe modyfikacje w stosunku do standardowej implementacji, aby zapewnić stabilność i wydajność.

Innowacja 1: Pozycjonowanie ALiBi dla większej elastyczności



Problem z tradycyjnym podejściem:

Standardowe metody przypisują każdemu słowu stały "adres" (pozycja 1, 2, 3...), co jest nieelastyczne, zwłaszcza przy długich tekstach.

Rozwiązań ALiBi (Attention with Linear Biases):

Podejście względne: zamiast absolutnych adresów, ALibi koduje relatywną odległość – "to słowo jest blisko, a tamto daleko".

Mechanizm: ALiBi bezpośrednio osłabia siłę uwagi (attention scores) między tokenami w zależności od ich odległości w sekwencji.

Korzyści

- ❖ Bardziej stabilny trening.
- ❖ Lepsze wyniki w zadaniach downstream.
- ❖ Kluczowa innowacja umożliwiająca modelowi ekstrapolację na sekwencje dłuższe niż te, na których był trenowany.

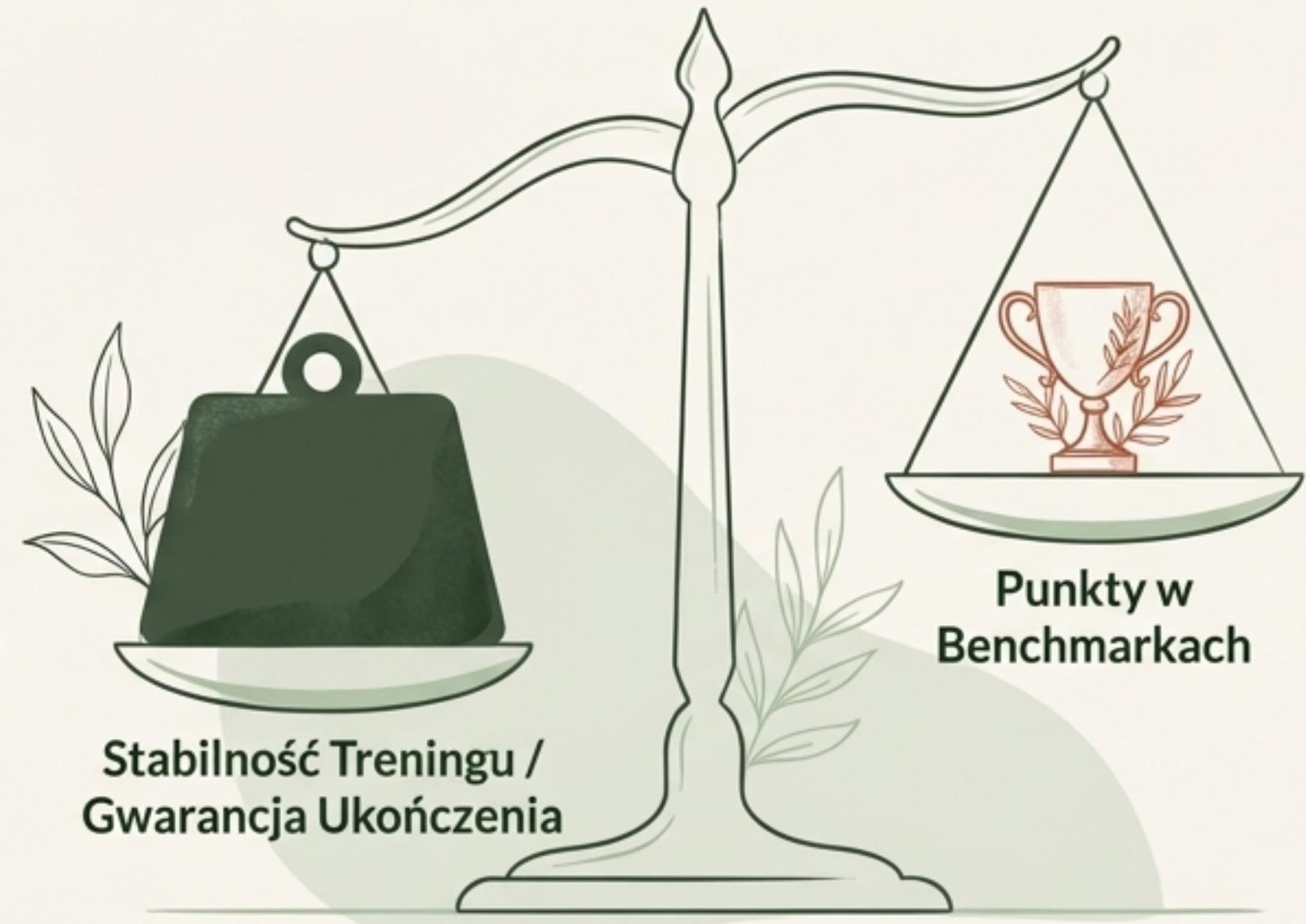
Inżynieria w praktyce: Przetrwanie ważniejsze niż punkty w benchmarkach

Decyzja

Dodanie dodatkowej warstwy normalizacyjnej (Layer Normalization) bezpośrednio po warstwie wejściowej (embedding layer).

Cel

Znacząca poprawa stabilności treningu przy skali 100B+ parametrów.



Wniosek

To doskonały przykład zderzenia teorii z rzeczywistością inżynierijną przy ekstremalnej skali. Priorytetem było ukończenie treningu, a nie wycisnięcie ostatnich ułamków procenta z benchmarków.

Zaskakujący kompromis

W testach na mniejszych modelach, ta dodatkowa warstwa **nieznacznie pogarszała** wydajność wydajność w zadaniach zero-shot.

Mimo to, została świadomie dodana do finalnej architektury BLOOM.

Język globalny od podstaw: Strategia tokenizacji neutralnej językowo

Algorytm

Byte Level BPE (Byte Pair Encoding) z dużą liczbą tokenów w słowniku (~250,680).

Kluczowe cechy i ich uzasadnienie

- **Poziom bajtów (Byte-level):** Model nigdy nie napotyka nieznanych znaków, co jest krytyczne przy obsłudze 46 różnych języków i ich systemów pisma.
- **Brak normalizacji tekstu:** Model uczy się na surowych, nieoczyszczonych danych (np. nie ujednolica '2²' i '2' i '2'), co czyni go bardziej ogólnym.
- **Odrzucenie anglocentrycznych reguł:** Świadomie zrezygnowano z zasad specyficznych dla języka angielskiego, np. dzielenia przy skrótach (jak 'n't', "ll").



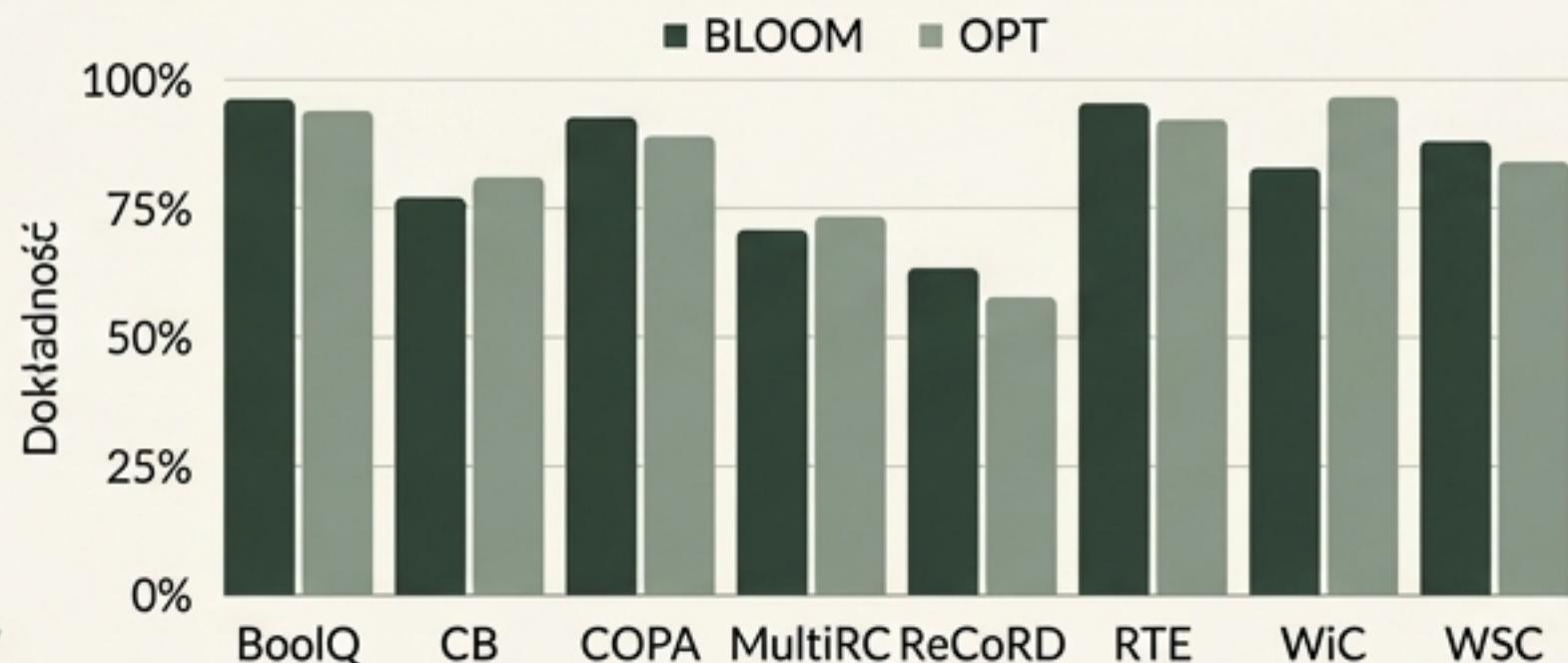
Cel nadzędny

Osiągnięcie maksymalnej neutralności językowej i naturalna obsługa różnorodnych skryptów i symboli.

Rozkwit: Wydajność, transfer wiedzy i niespodzianki

Wydajność w języku angielskim (SuperGLUE)

Trening wielojęzyczny nie zaszkodził wydajności w języku angielskim.



Trening wielojęzyczny nie zaszkodził wydajności w języku angielskim.

Zaobserwowano większy skok wydajności przy przejściu z zero-shot do one-shot niż u konkurencji, co sugeruje, że kontekst pomaga modelowi lepiej skupić się na zadaniu.

Zdolności wielojęzyczne (FLORES-101)

Konkurował z wyspecjalizowanymi, nadzorowanymi modelami tłumaczeniowymi.



NIESPODZIAANKA

Potrafił tłumaczyć na język galicyjski, mimo że nigdy nie widział go w danych treningowych, demonstrując transfer wiedzy z hiszpańskiego i portugalskiego.

Uczciwa ocena słabości

Wydajność w językach o niskich zasobach (low-resource), tck), takich jak suahili czy joruba, jest wciąż niska. Ilość danych treningowych ma fundamentalne znaczenie.

Zrównoważony ekosystem: Ewolucja modelu i odpowiedzialność ekologiczna

Ewolucja do BLOOMZ

Metoda: Dostrajanie (fine-tuning) bazowego modelu BLOOM na wielojęzycznym zbiorze zadań xP3.

Efekt: Dramatyczna poprawa zdolności do wykonywania zadań w trybie zero-shot. Przejście od 'wiedzy' (knowledge) do 'działania' (action) – nauka podążania za instrukcjami.



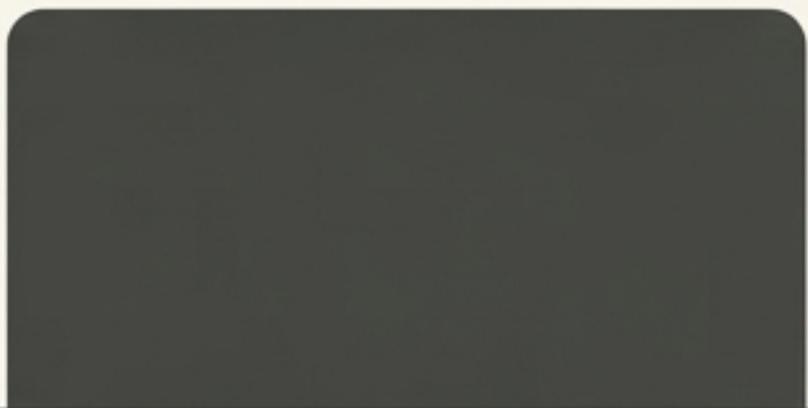
Transparentność i ślad węglowy

500+ ton CO₂e



~25 ton CO₂e

BLOOM



GPT-3

Kluczowy czynnik: Wybór superkomputera Jean-Zay we Francji, zasilanego w dużej mierze energią jądrową o niskiej intensywności węglowej (57 gCO₂eq/kWh).

Wniosek: Strategiczny wybór centrum danych ma ogromne znaczenie dla ekologicznego kosztu AI.

“Czy nieustanne skalowanie modeli w kierunku coraz większej liczby parametrów zawsze czyni je mądrzejszymi pod każdym względem? Czy w pogoni za wielkością nie tracimy po drodze subtelnych, bardziej fundamentalnych zdolności językowych – i co to oznacza dla przyszłości AI, która powinna naprawdę rozumieć język, a nie tylko generować płynny tekst?”