

GLM-130B: Demokratyzacja dostępu do wielkich modeli językowych

Jak inteligentna inżynieria pokonuje strategię „brute-force” w świecie sztucznej inteligencji.

Skala i zasięg

Otwarty, dwujęzyczny (angielski i chiński) model o **130 miliardach** parametrów.

Twórcy

Opracowany przez **Tsinghua University** i ZHIPU.AI.

Wydajność

Przewyższa GPT-3 (175 mld parametrów) w wielu kluczowych benchmarkach, mimo mniejszego rozmiaru.



Innowacyjna architektura

Zbudowany na nowatorskiej **architekturze GLM** (General Language Model), która fundamentalnie różni się od modeli typu GPT.

Prawdziwa demokratyzacja

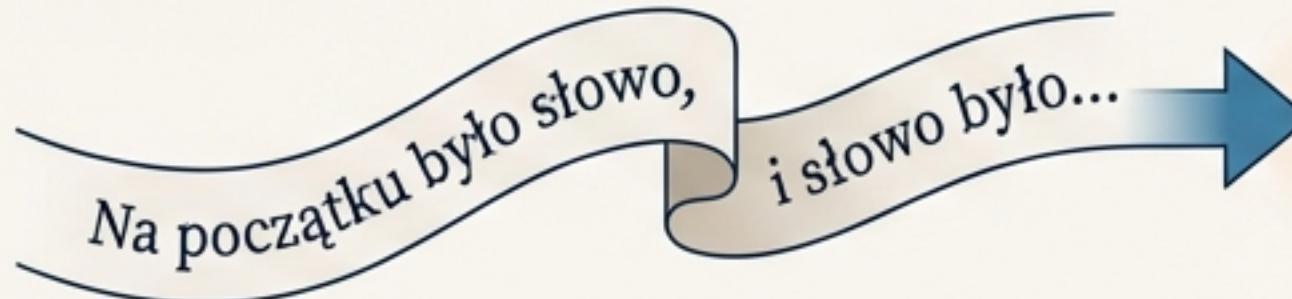
Jako pierwszy model w skali 100B+, działa na powszechnie dostępnym sprzęcie – wystarczą 4 karty graficzne RTX 3090.

Nowa filozofia

Reprezentuje przejście od prostego skalowania mocy obliczeniowej do przemyślanej **inżynierii** i efektywności algorytmicznej.

Zamiast opowiadać historię, model uczy się ją rekonstruować

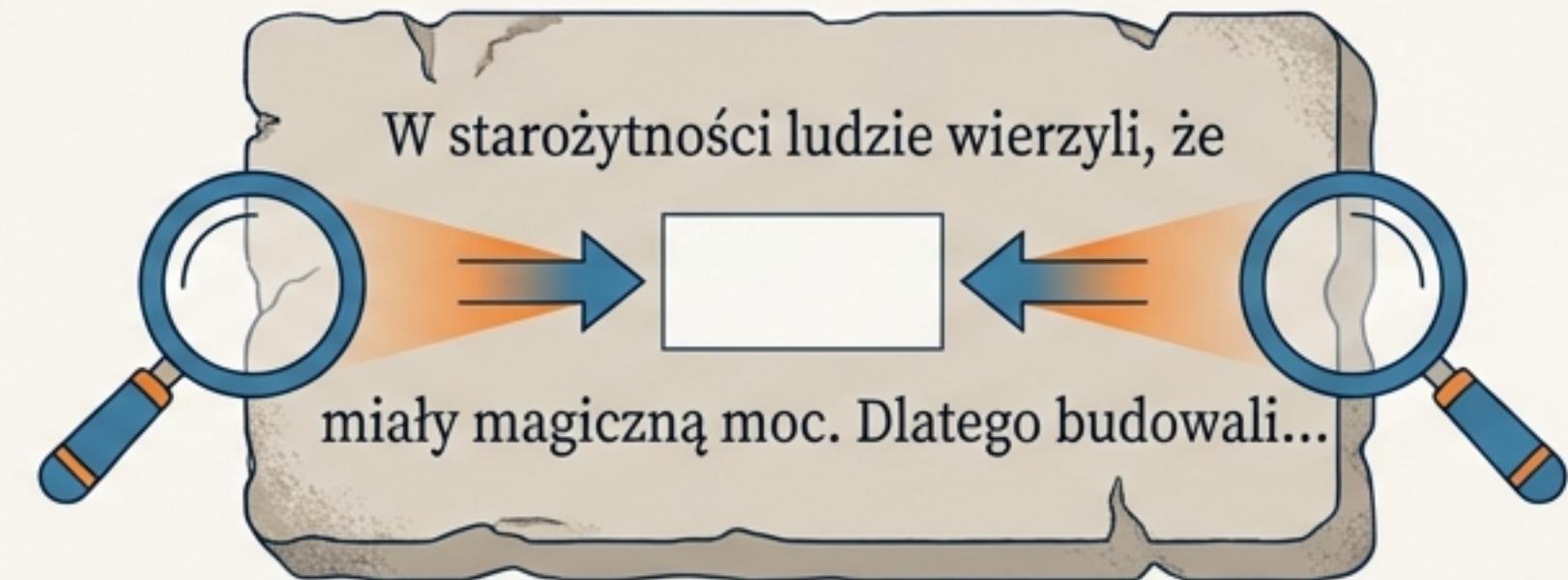
GPT: Przewidywanie następnego słowa (jednokierunkowe)



Tradycyjne modele autoregresyjne

Modele takie jak GPT działają jak gawędziarz – czytają tekst od lewej do prawej i przewidują kolejne słowo. Ich rozumienie kontekstu jest jednokierunkowe.

GLM: Wypełnianie pustych miejsc (dwukierunkowe)



Autoregresyjne wypełnianie pustych miejsc (Autoregressive Blank Infilling)

GLM działa jak detektyw lub archeolog. Otrzymuje tekst z celowo usuniętymi fragmentami (pustymi miejscami) i ma za zadanie je zrekonstruować, analizując kontekst z obu stron – zarówno to, co było przed lukią, jak i to, co nastąpiło po niej.

Kluczowa korzyść

Ta dwukierunkowa analiza kontekstu pozwala na znacznie głębsze zrozumienie relacji w tekście, łącząc zdolności rozumienia ze zdolnościami generowania.

Strategia podwójnego maskowania: Najlepsze z obu światów w jednym modelu

[MSK]

W procesie uczenia model analizuje i przewiduje brakujące elementy w zdaniu, takie jak to [MSK] słowo, aby zrozumieć kontekst, a następnie kontynuuje generowanie dalszej części tekstu w sposób płynny i spójny, aż do osiągnięcia końca sekwencji, co ilustruje [gMASK]



Token [MSK] - Głębokie rozumienie kontekstu

Cel: Wypełnianie krótkich, losowych luk w środku tekstu.

Analogia: Naśladuje zachowanie modeli typu BERT, skupiając się na dogłębnym, dwukierunkowym rozumieniu kontekstu.

Zastosowanie: Idealne do zadań wymagających analizy i rozumienia języka (NLU).



Token [gMASK] - Plynne generowanie tekstu

Cel: Generowanie długich, spójnych sekwencji na podstawie podanego początku (prefiksu).

Analogia: Naśladuje zachowanie modeli typu GPT, doskonaląc płynne i kreatywne generowanie tekstu.

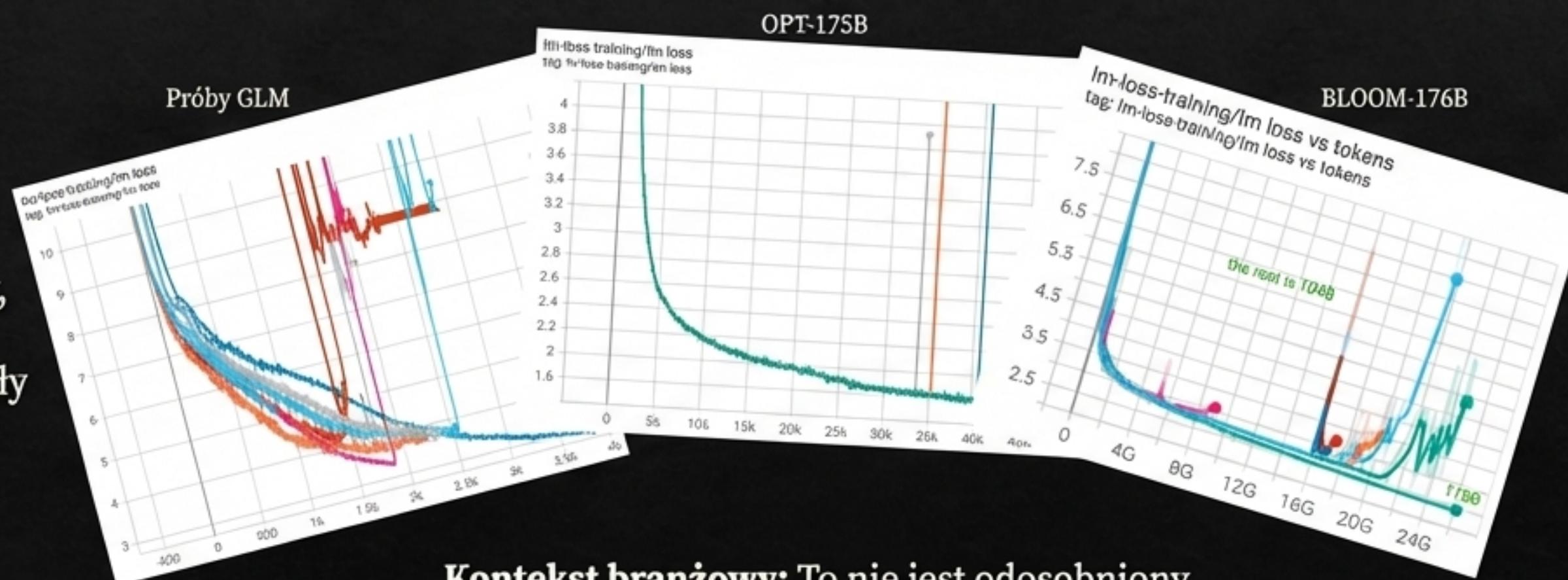
Zastosowanie: Idealne do zadań generatywnych (NLG), takich jak pisanie artykułów czy odpowiedzi na pytania.

Dzięki tej hybrydowej strategii GLM-130B staje się uniwersalnym modelem, który potrafi jednocześnie rozumieć i generować tekst z równą biegłością.

Ściana: Ponad 30 nieudanych prób i kryzys niestabilności

„Autorzy otwarcie raportują o ponad 30 całkowicie nieudanych próbach treningu, z których każda kończyła się katastrofą po tygodniach obliczeń.”

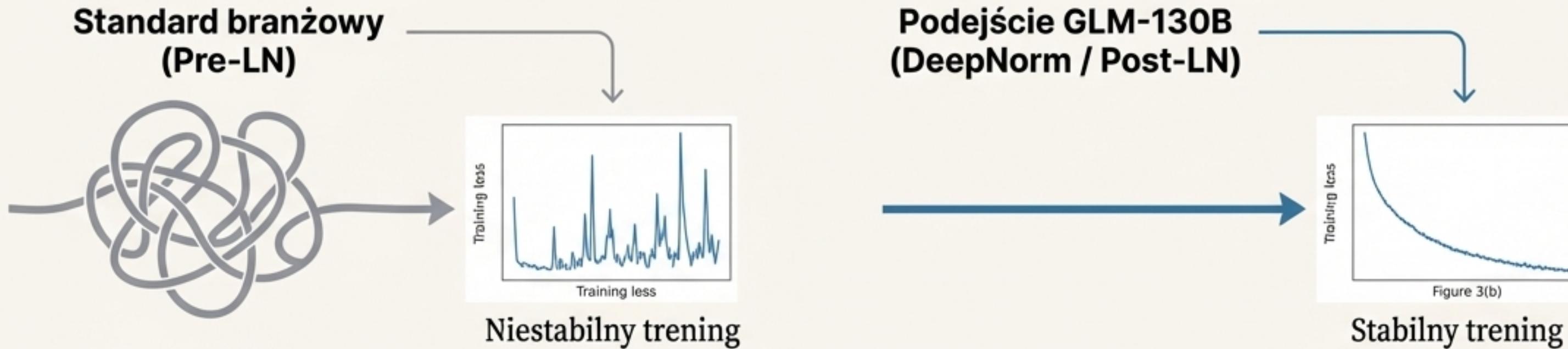
Problem: Ciągłe „wystrzały straty” (loss spikes) – nagłe skoki błędów, które niszczyły postępy i marnowały ogromne zasoby obliczeniowe.



Kontekst branżowy: To nie jest odosobniony problem. Trening OPT-175B (Meta) wymagał ręcznych interwencji, a BLOOM-176B (BigScience) zastosował technikę „embedding norm”, która stabilizowała trening kosztem końcowej jakości modelu.

Cytat/Analiza:
Trenowanie wielkich modeli językowych przypominało „łatać dziur w kadłubie transatlantyku na środku oceanu podczas sztormu”.

DeepNorm: Sukces wbrew branżowym trendom



Kontekst problemu

Dominujący trend

Większość nowych modeli (GPT-3, PaLM, BLOOM) przeszła na architekturę Pre-LN (Layer Normalization *przed* głównymi operacjami transformera), wierząc, że jest ona bardziej stabilna.

Eksperymenty GLM

Wczesne testy GLM-130B z Pre-LN, a nawet ulepszonym Sandwich-LN, konsekwentnie kończyły się niepowodzeniem.

Nieszablonowe rozwiązanie

Decyzja

Zespół postanowił wrócić do starszej architektury Post-LN (normalizacja *po* operacjach), ale z nowatorską metodą inicjalizacji i skalowania nazwaną **DeepNorm**.

Rezultat

Ta kontrintuicyjna decyzja okazała się kluczem do sukcesu. Prawidłowo skalibrowany Post-LN nie tylko ustabilizował trening, ale okazał się również bardziej wydajny.

Sukces GLM-130B pokazuje, że ślepe podążanie za trendami nie zawsze jest optymalne. Czasem kluczem jest dogłębne zrozumienie i ulepszenie fundamentalnych zasad.

Chirurgiczne cięcie: Embedding Layer Gradient Shrink (EGS)

Diagnoza problemu

- **Główne źródło niestabilności:** Badacze zidentyfikowali, że główną przyczyną „wystrzałów straty” są anomalnie duże i chaotyczne gradienty (sygnały błędu) w warstwie wejściowej (embedding layer).
- **Problem w praktyce:** Sygnały te były tak silne, że zamiast korygować model, „przestrzeliwały” i wprowadzały chaos do całego procesu uczenia.

Proste i genialne rozwiązanie

- **Technika:** Zastosowano Embedding Layer Gradient Shrink (EGS).
- **Działanie:** Zamiast skomplikowanych mechanizmów, po prostu zredukowano siłę tych gradientów o 90% (współczynnik $\alpha = 0.1$).
- **Dlaczego to działało:** Gradienty w warstwie embeddingowej były tak ekstremalnie zawyżone, że nawet 10% ich pierwotnej siły wciąż było wystarczające do efektywnego uczenia, eliminując jednocześnie destrukcyjne skoki.

EGS to przykład eleganckiego rozwiązania, które pokazuje, że precyzyjna diagnoza problemu jest potężniejsza niż stosowanie coraz bardziej skomplikowanych zabezpieczeń.

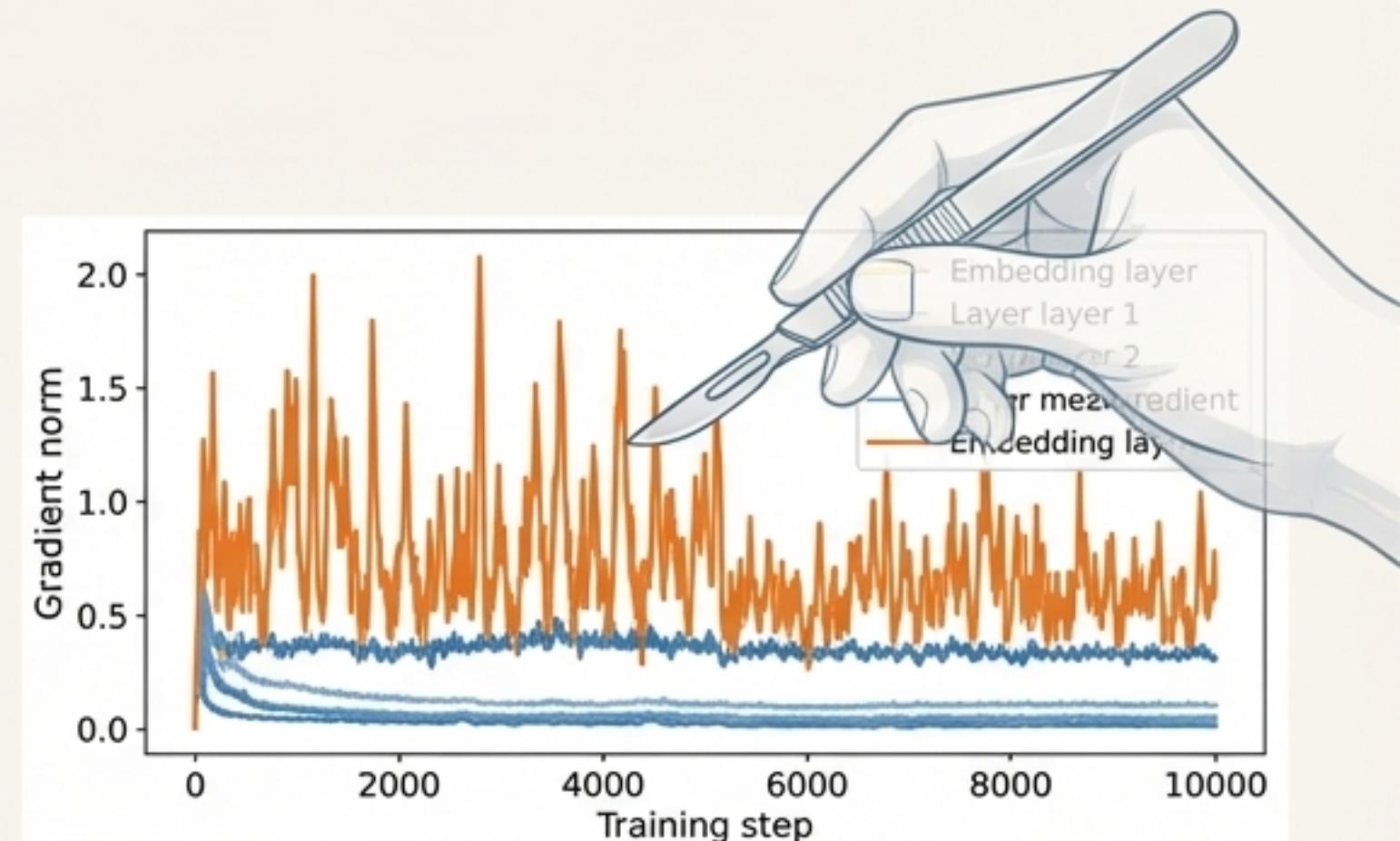


Figure 4(a) [Chaotic Gradients]

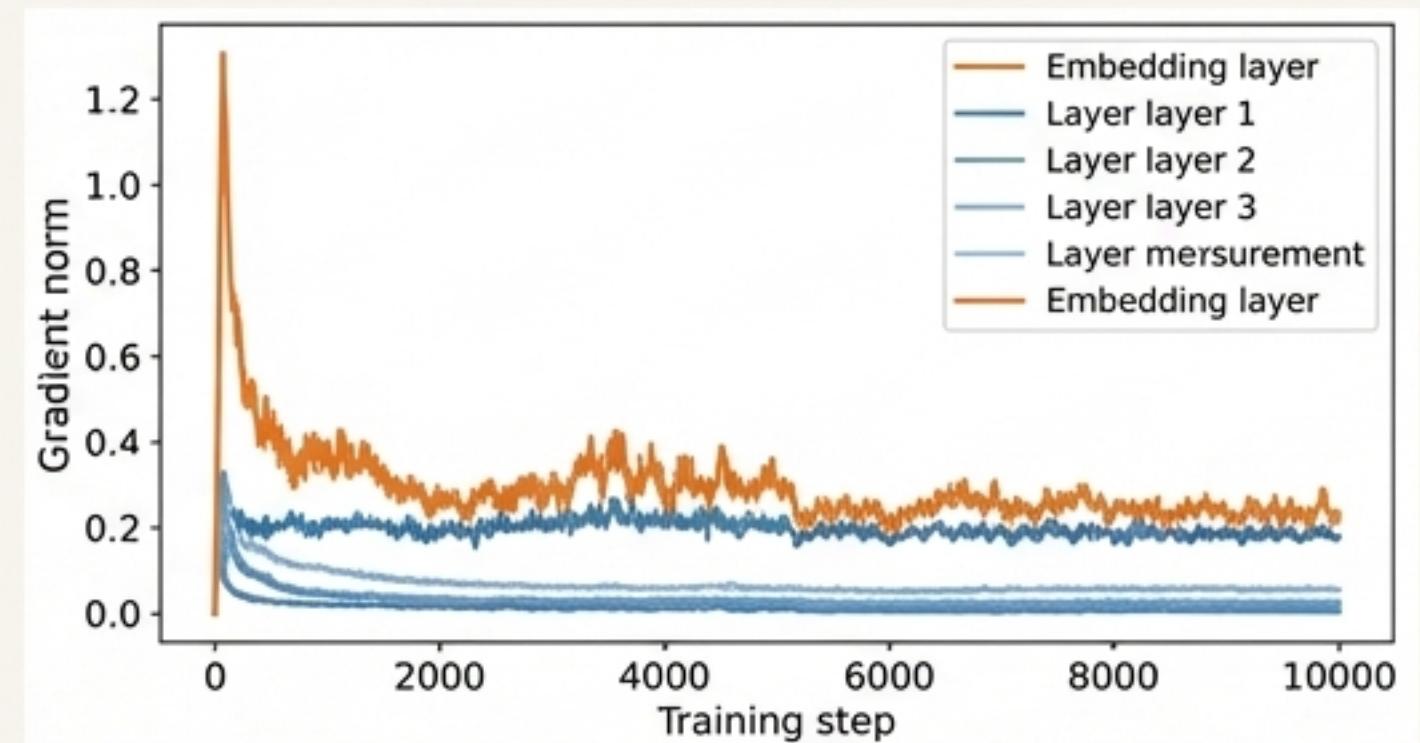


Figure 4(b) [Effect of EGS]

Mniejszy, wydajniejszy, a jednak potężniejszy

GLM-130B w bezpośrednim porównaniu z gigantami branży.

BENCHMARK PERFORMANCE COMPARISON

LAMBADA

(rozumienie kontekstu)



+5.0% nad
GPT-3 175B

BIG-bench-lite

(zero-shot, rozumowanie)



Pokonuje PaLM 540B,
model 4x większy

MMLU

(wiedza ogólna, 5-shot)



Zadania w języku chińskim

(CLUE & FewCLUE)

GLM-130B
(e.g., on EPRSTMT)

ERNIE Titan 3.0 260B

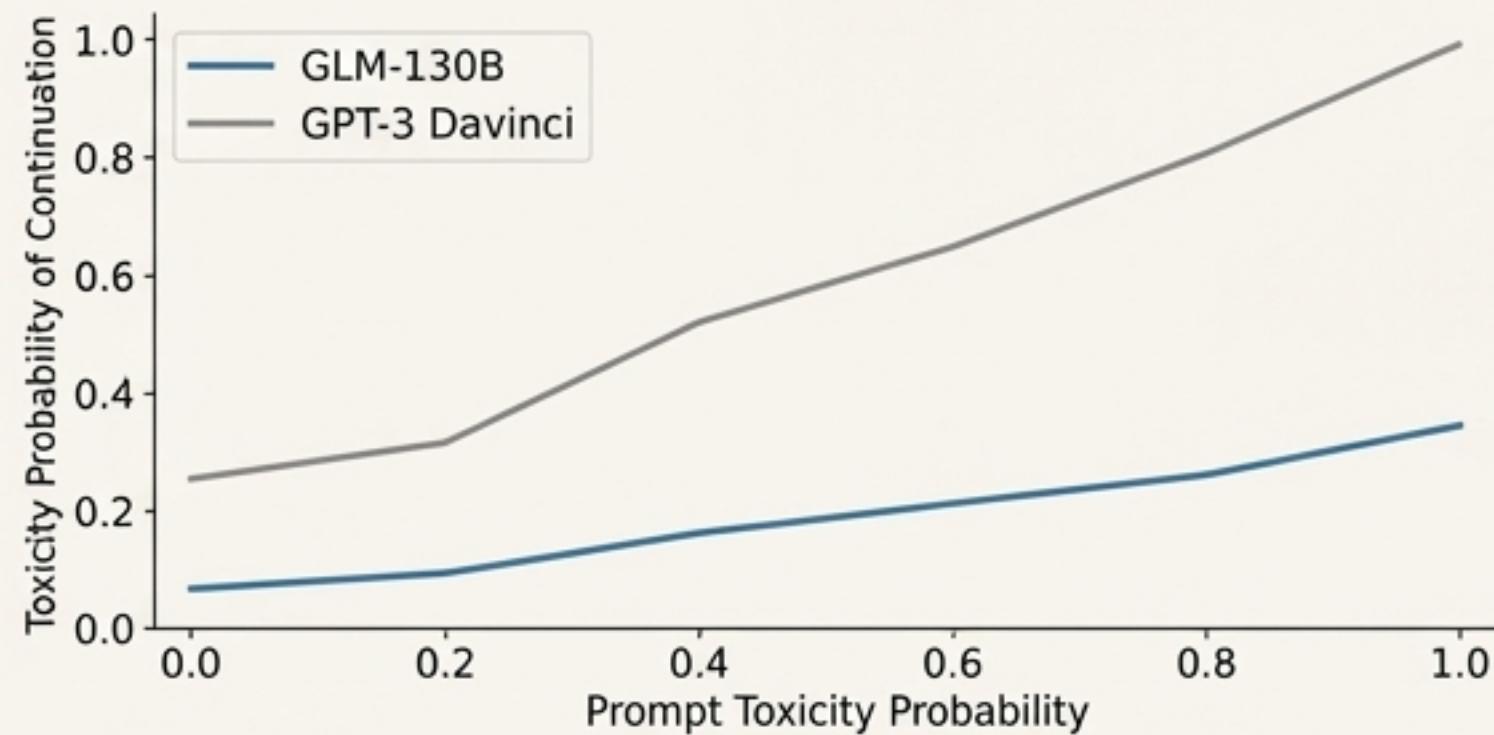


Znaczaco przewyjsza
2x większy model
ERNIE Titan 3.0

Te wyniki dowodzą, że inteligentna architektura i metody treningu mogą być bardziej efektywne niż samo zwiększanie liczby parametrów.

Niespodziewane odkrycie: dwujęzyczność redukuje toksyczność i uprzedzenia

Mniejsza toksyczność



W testach RealToxicPrompts, GLM-130B konsekwentnie generuje znacznie mniej toksyczne treści niż modele trenowane wyłącznie na danych anglojęzycznych.

Hipoteza badaczy

Dwujęzyczny trening na korpusach z dwóch różnych kultur (angielskiej i chińskiej) zapewnia modelowi szerszą perspektywę, co może prowadzić do wzajemnego niwelowania uprzedzeń obecnych w każdym ze zbiorów danych.

Zredukowane uprzedzenia

		Wskaźnik uprzedzeń (im niższy, tym lepszy)	
		GLM-130B	GPT-3
1	Płeć	55.7	62.6
2	Rasa/Kolor	58.5	64.7
3	Orientacja seksualna	60.7	76.2
4	Ogółem	65.8	67.2

W benchmarkach takich jak CrowS-Pairs i StereoSet, GLM-130B wykazuje mniej stereotypowych uprzedzeń w porównaniu do GPT-3 i OPT-175B.

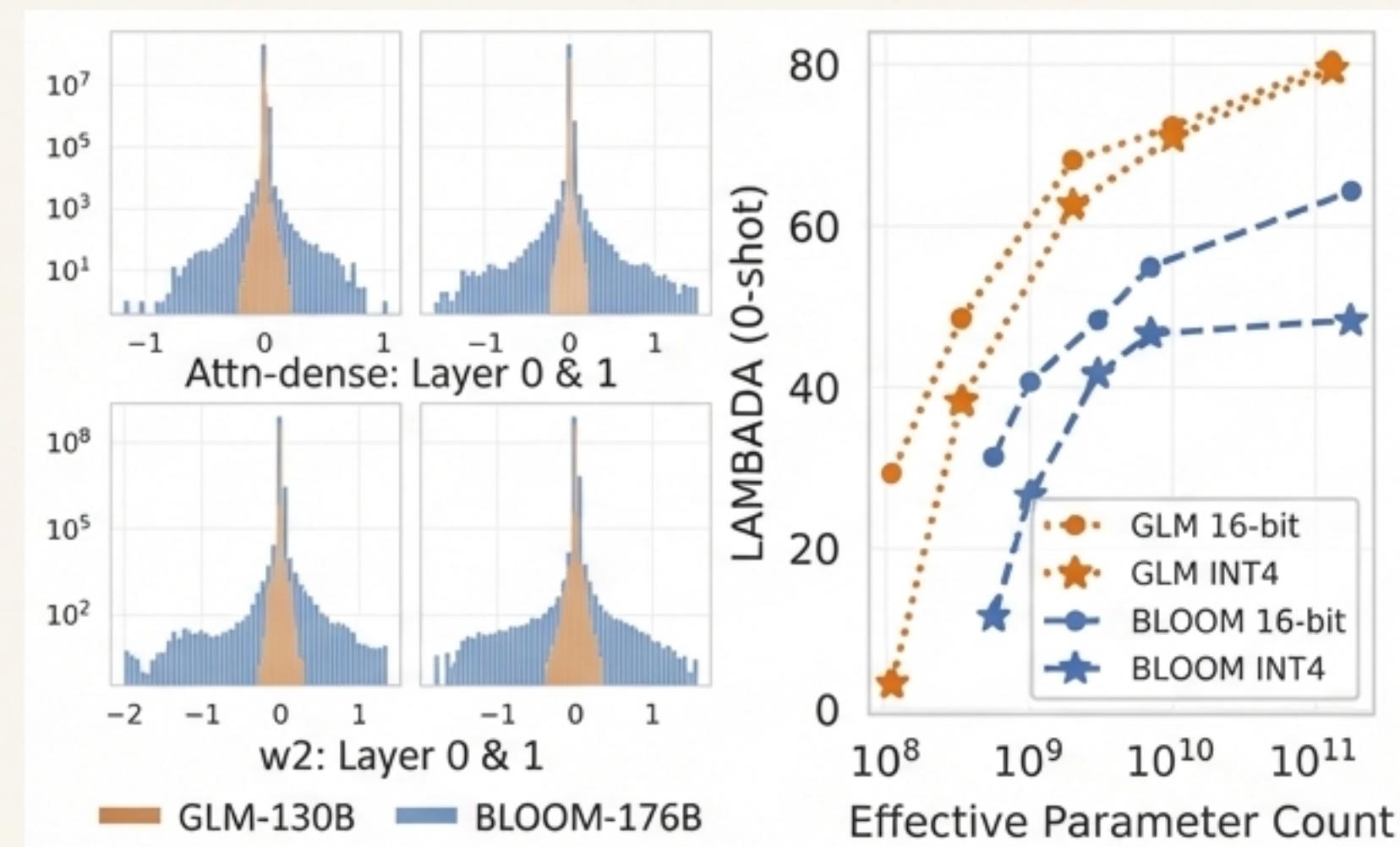
Implikacje

To fascynujący kierunek badań sugerujący, że trening wielojęzyczny może być kluczową strategią w tworzeniu bardziej etycznej i sprawiedliwej sztucznej inteligencji.

Kwantyzacja INT4: Ekstremalna kompresja bez utraty jakości

Kontekst

- Większość wielkich modeli z trudem radzi sobie z kwantyzacją do 8 bitów (INT8).
- GLM-130B można skompresować do zaledwie **4 bitów (INT4)**
– to 4-krotna redukcja rozmiaru wag.



Zaskakujący rezultat

- Kompresja do INT4 odbywa się praktycznie **bez utraty wydajności**.
- LAMBADA: spadek zaledwie **-0.74%**.
- MMLU: wynik **nieznacznie wzrósł o +0.05%**!

Dlaczego to działa?

Jak widać na wykresie, architektura GLM **naturalnie prowadzi** węższej dystrybucji wag modelu. Mniej ekstremalnych wartości odstających (outlierów) oznacza, że proces kwantyzacji traci znacznie mniej informacji, co pozwala na tak agresywną kompresję.

Od teorii do praktyki: Moc 130 miliardów parametrów na Twoim serwerze

Uniwersytet /
Laboratorium
Badawcze



Inference

- **4x NVIDIA RTX 3090 (24GB)**
- *Alternatywnie:* 8x NVIDIA RTX 2080 Ti (11GB) – starszy i jeszcze bardziej dostępny sprzęt.



Przed GLM-130B

Dostęp do modeli tej skali był ograniczony do 3-4 globalnych korporacji z dostępem do superkomputerów.

To nie tylko otwarty kod, to prawdziwie otwarty dostęp.

Startup / Firma
Technologiczna



Performance

- Implementacja w C++ z wykorzystaniem biblioteki **NVIDIA FasterTransformer** zapewnia **7-8x przyspieszenie** w porównaniu do standardowych implementacji w Pythonie.



Deweloper /
Niezależny Inżynier

Dzięki GLM-130B

Technologia staje się dostępna dla **setek laboratoriów badawczych i tysięcy firm** na całym świecie.

Pytanie na przyszłość

Skoro model uczący się od początku z dwóch języków wykazuje mniejsze uprzedzenia i nowe zdolności...

...co by się stało, gdyby przyszłe modele uczyły się nie z dwóch, ale z dwudziestu języków pochodzących z różnych rodzin językowych? Jakich zupełnie nowych, emergentnych zdolności moglibyśmy być świadkami?