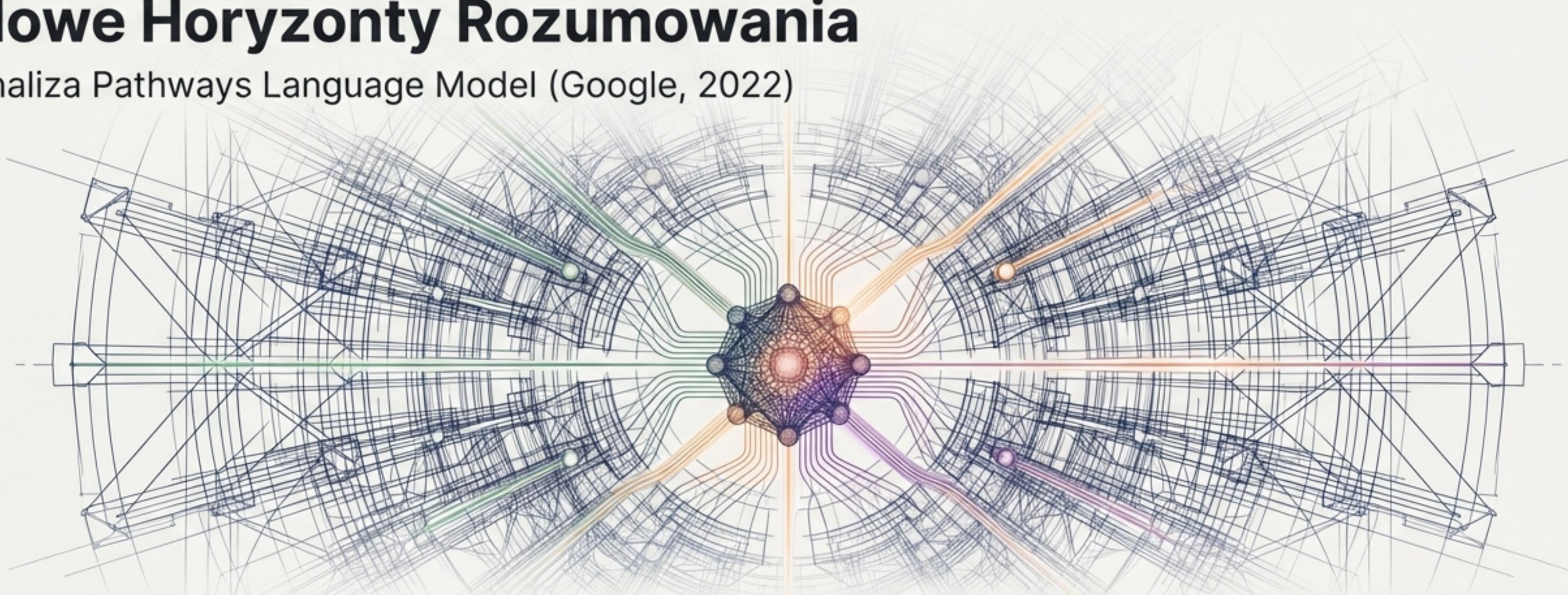


PaLM: Przełom w Skalowaniu AI, Który Ujawnił Nowe Horyzonty Rozumowania

Analiza Pathways Language Model (Google, 2022)



540 miliardów parametrów

W momencie publikacji, największy gęsty model językowy na świecie.

Nowa era efektywności

Trenowany na ponad **6000 chipach TPU v4** dzięki rewolucyjnemu systemowi **Pathways**.

Emergentne Zdolności

Wykazał zaskakujące, jakościowo nowe umiejętności, które pojawiły się dopiero przy tej skali.

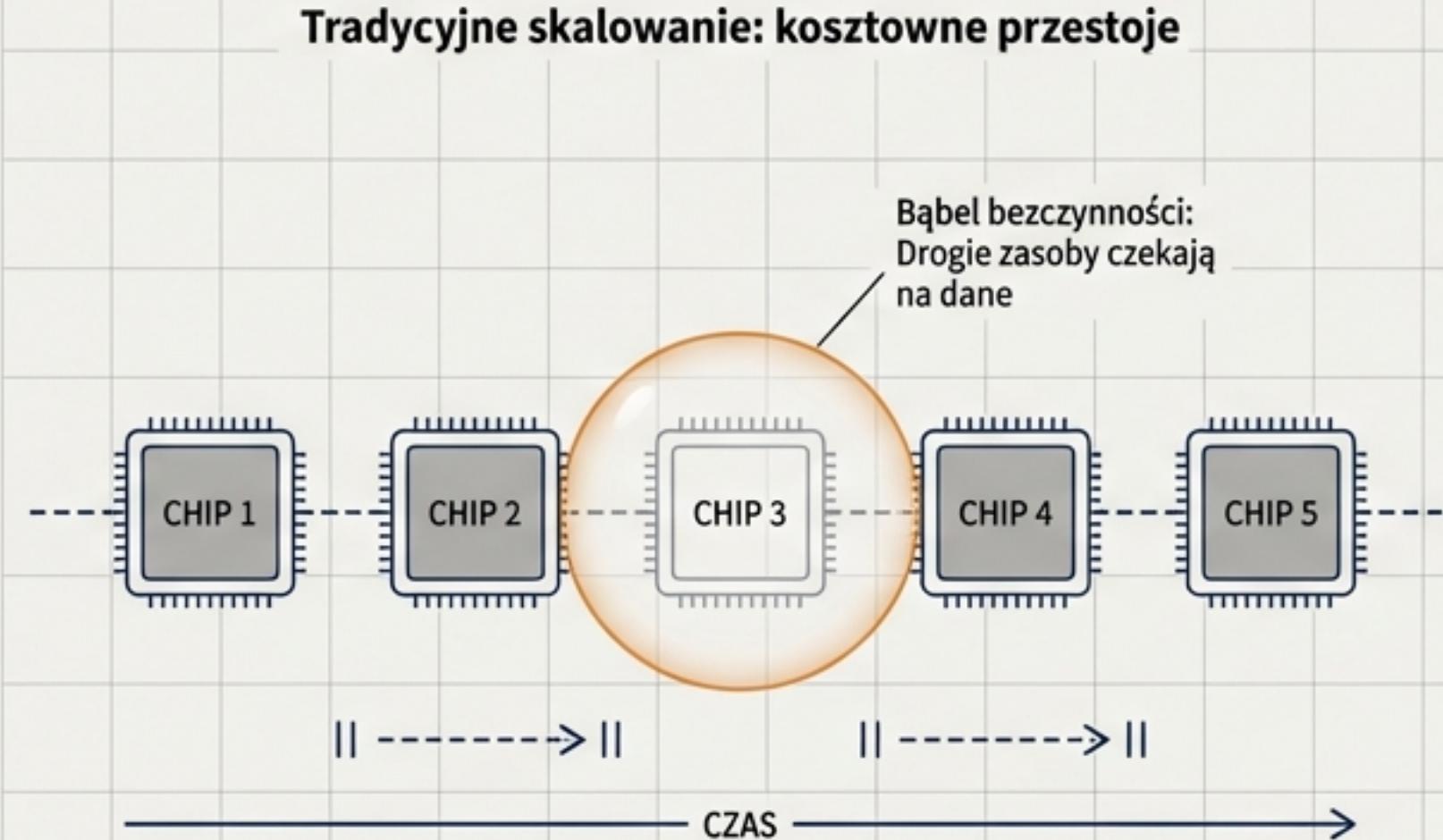
Przełom inżynieryjny i naukowy

Nie tylko większy model, ale fundamentalnie **nowy** sposób budowania i rozumienia sztucznej inteligencji.

Pathways: Infrastruktura, Która Zmieniła Zasady Gry w Treningu na Masową Skalę

Problem: Kosztowne przestoje w tradycyjnym skalowaniu

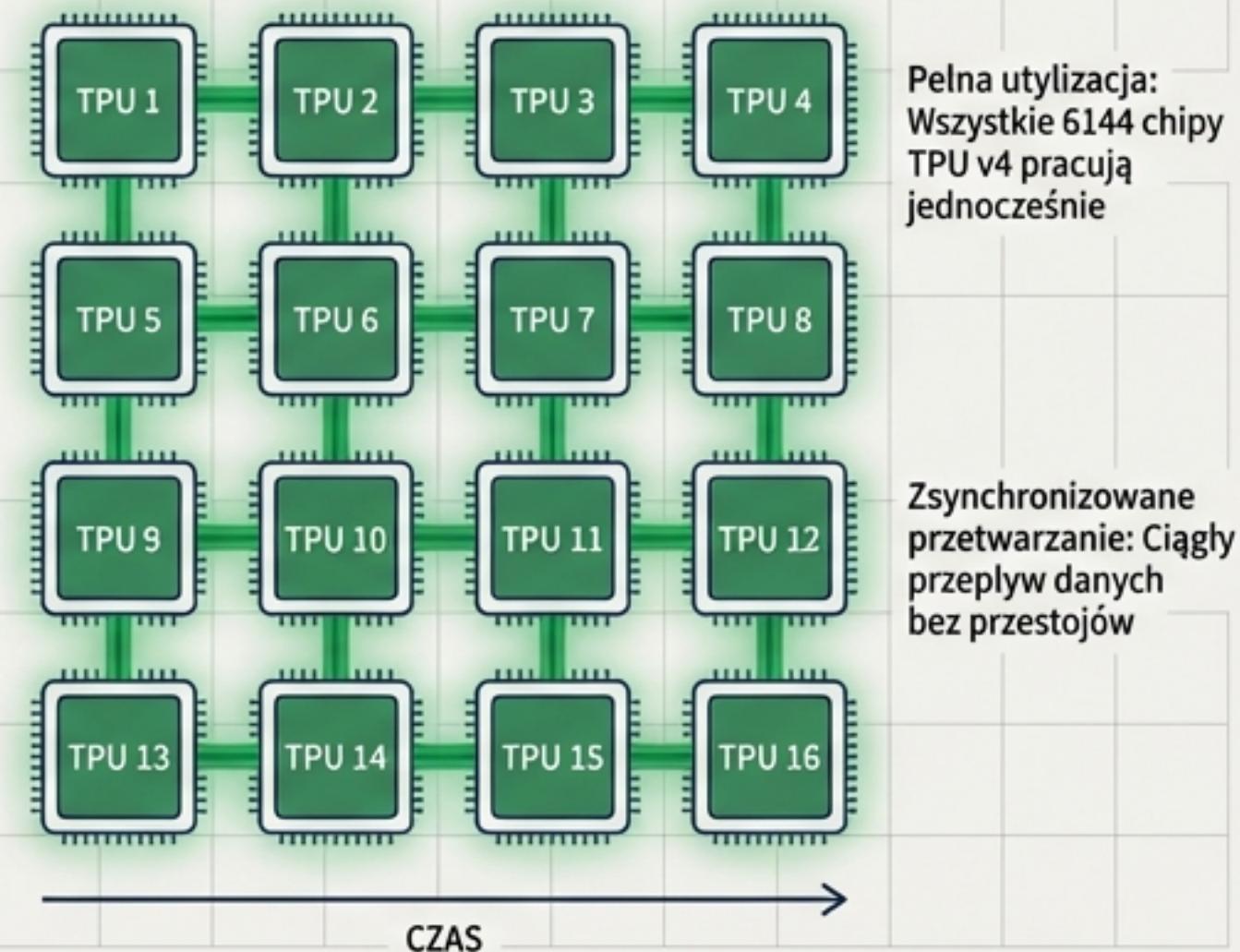
Tradycyjne metody skalowania (np. pipeline parallelism) tworzyły "bąble bezczynności" (*bubble of idleness*), gdzie drogi akceleratory czekają bezczynnie na dane. To jak linia montażowa, gdzie każdy pracownik musi czekać, aż poprzedni skończy swoją pracę, co prowadzi do wąskich gardel i marnotrawstwa zasobów.



Rozwiązanie Pathways: Ciągła, zsynchronizowana praca

Pathways umożliwiło trening na 6144 chipach TPU v4, łącząc po raz pierwszy dwa Pody. System eliminował bezczynność, pozwalając na jednocześnie, w pełni zutylizowane przetwarzanie na wszystkich chipach. Efektem było radykalne przyspieszenie i możliwość osiągnięcia bezprecedensowej skali.

Pathways: Ciągła, zsynchronizowana praca



Ponad 2x Większa Efektywność: PaLM Zdeklasował Poprzedników w Wykorzystaniu Mocy Obliczeniowej

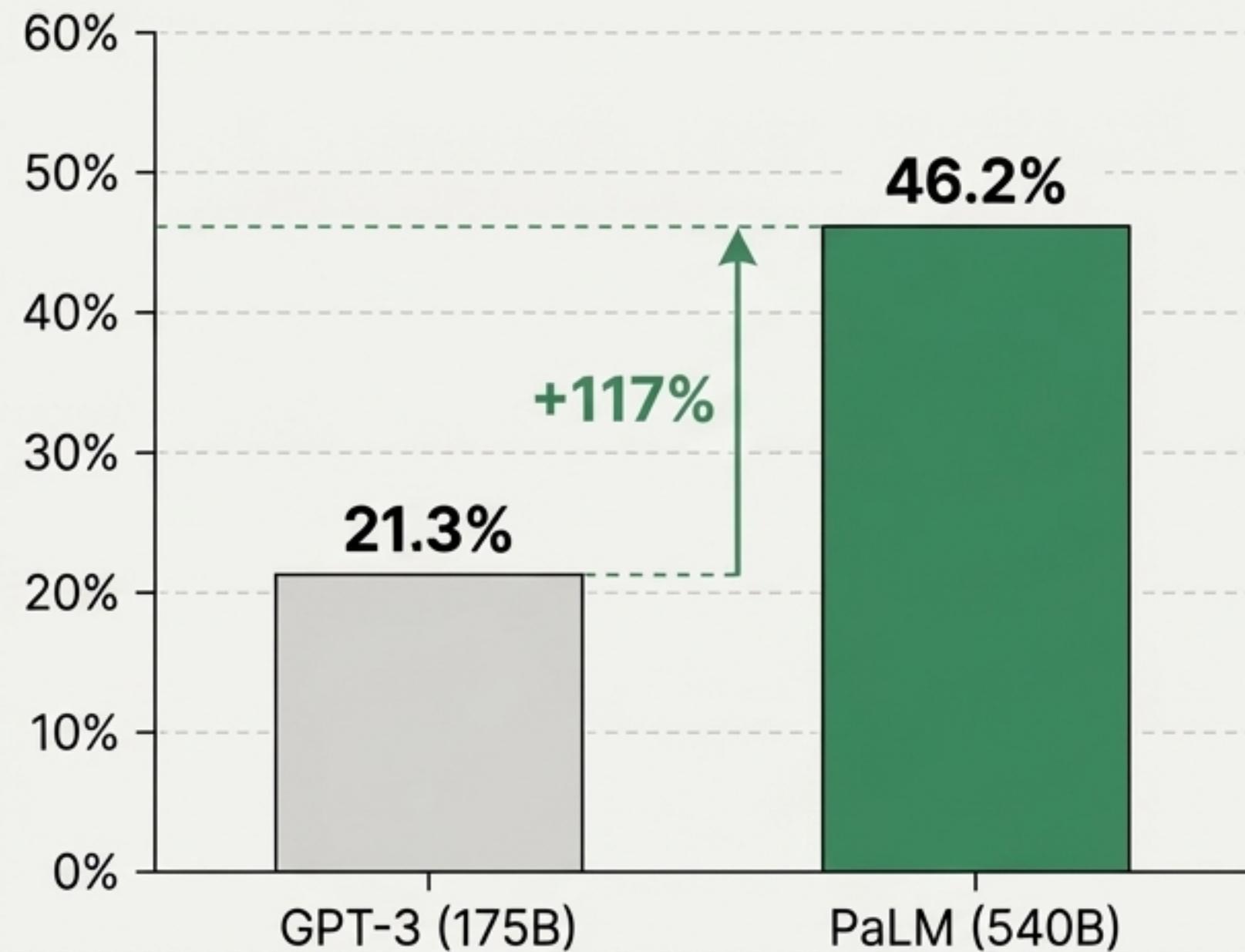
Nowa metryka: Model Flops Utilization (MFU)

MFU mierzy, jak efektywnie system wykorzystuje teoretyczną maksymalną moc obliczeniową chipów na **rzeczywiste** operacje modelu, ignorując narzuty systemowe jak rematerializacja. Pozwala na uczciwe porównania między różnymi architekturami i systemami.

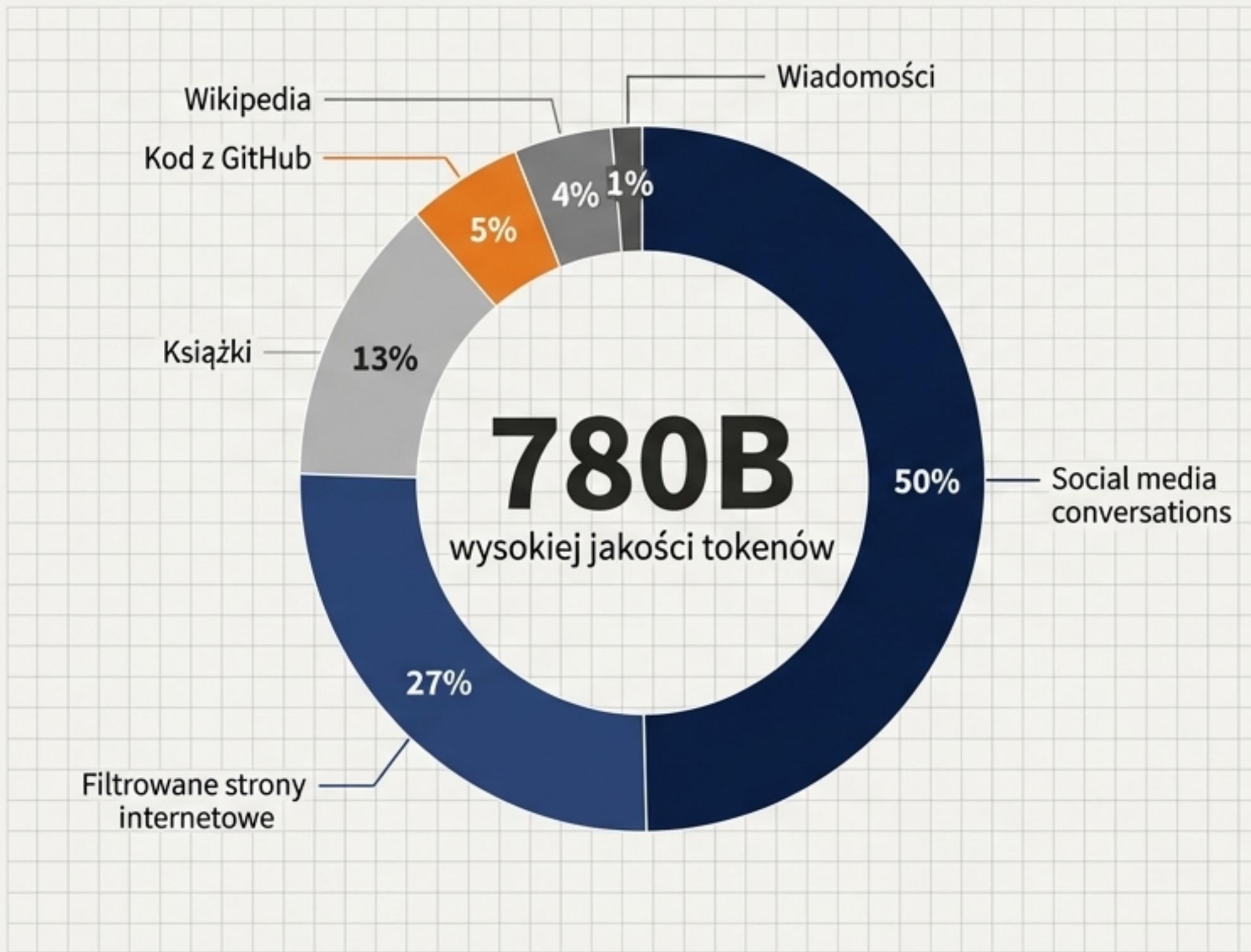
Wniosek

PaLM osiągnął ponad dwukrotnie wyższą efektywność treningu przy znacznie większej skali. Mistrzowska optymalizacja całego stosu technologicznego pozwoliła osiągnąć 57.8% Hardware FLOPs Utilization (uwzględniając (uwzględniając rematerializację).

Efektywność Obliczeniowa (MFU)



Fundament Wiedzy: Analiza 780 Miliardów Tokenów Użytych do Treningu PaLM



Skład Zbioru Danych

- 50% Social media conversations (wielojęzyczne)
- 27% Filtrowane strony internetowe (wielojęzyczne)
- 13% Książki (angielskie)
- 5% Kod z GitHub
- 4% Wikipedia (wielojęzyczna)
- 1% Wiadomości (angielskie)

Krytyczne Ograniczenie

78%

danych treningowych w języku angielskim. Ta dominacja fundamentalnie ukształtowała możliwości i specyfikę działania modelu.

Rewolucja w Rozumowaniu: Jak ‘Chain of Thought Prompting’ Uczy Model Myśleć Krok po Kroku

Modele LLM miały trudności z wieloetapowymi zadaniami logicznymi. Przełomowa technika *Chain of Thought Prompting* instruuje model, aby najpierw wygenerował logiczny ciąg wniosków, a dopiero potem podać ostateczną odpowiedź. To połączenie ze skalą PaLM 540B pozwoliło osiągnąć wyniki SOTA bez dodatkowego treningu.

Standardowe Promptowanie



Chain of Thought Prompting



Zrozumieć Żart: PaLM Demonstruje Głęboką Wiedzę Kontekstową i Rozumienie Gier Słownych

Zadanie: Wyjaśnij żart

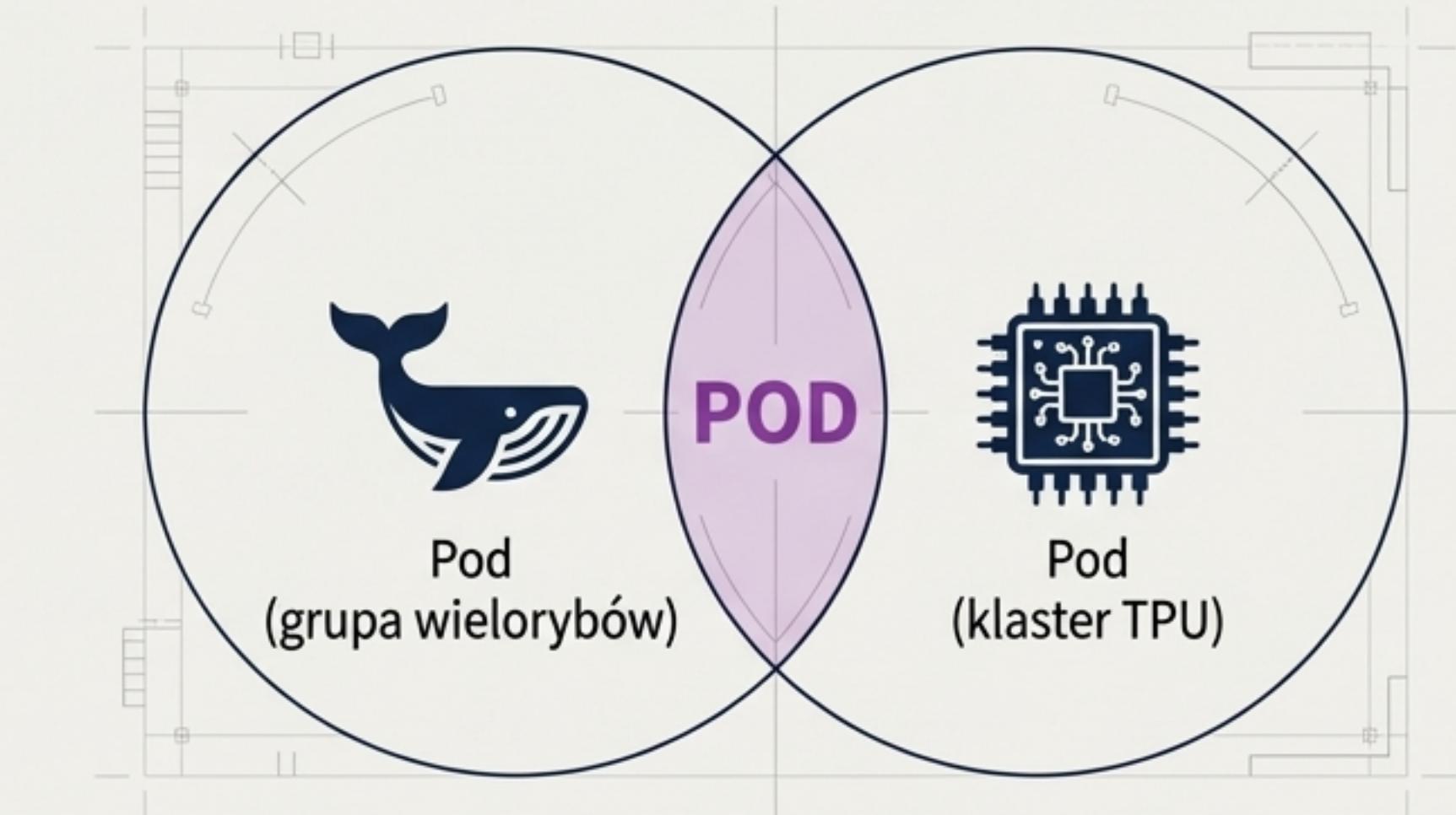
“

"Słyszałeś, że Google zatrudniło elokwentnego wieloryba do swojego zespołu TPU? Pokazał im, jak komunikować się między dwoma różnymi podami!"

Wyjaśnienie wygenerowane przez PaLM 540B

“

TPU to typ chipów komputerowych używanych przez Google do deep learningu. 'Pod' to grupa TPU. 'Pod' to także grupa wielorybów. Żart polega na tym, że wieloryb jest w stanie komunikować się między dwiema grupami wielorybów, ale mówca udaje, że wieloryb jest w stanie komunikować się między dwiema grupami TPU."



Analiza

Model poprawnie zidentyfikował grę słów opartą na homonimie 'pod'. Jednocześnie zrozumiał oba konteksty: **biologiczny** (grupa wielorybów) i **wysoko specjalistyczny, techniczny** (klaster procesorów TPU). Taki poziom zrozumienia niszowego, technicznego humoru nie był wcześniej obserwowany.

Skok Emergentny: Skala Odblokowuje Zdolności, Które Nie Rosną Liniowo

Zadanie: Angielskie Przysłowia (English Proverbs)



Obserwacja: "Discontinuous Improvements"

W wielu trudnych, abstrakcyjnych zadaniach z benchmarku BIG-bench, wydajność modelu nie poprawiała się stopniowo. Po przekroczeniu krytycznego progu skali, następował gwałtowny, skokowy wzrost wydajności.

Wniosek

Pewne zdolności, jak abstrakcyjne rozumowanie metafor, nie rozwijają się stopniowo. One "emergują" (pojawiają się) dopiero przy wystarczająco dużej skali modelu.

Więcej Niż Lingwista: PaLM-Coder Osiąga Mistrzostwo w Generowaniu i Naprawianiu Kodu

PaLM-Coder

Wersja PaLM dostrojona (*fine-tuned*) wyłącznie na danych zawierających kod źródłowy, osiągająca wyniki state-of-the-art w generowaniu kodu z opisu i naprawianiu błędów.

Wniosek: Więcej niż składnia

Model nauczył się nie tylko formalnej składni języka programowania, ale także niepisanych konwencji i preferencji stylistycznych ludzkich programistów, np. łącząc deklaracje zmiennych w jedną linię, co jest uznawane za dobrą praktykę.

Kod z błędami

```
1 #include <stdio.h>
2 int main() {
3     int a[i];
4     int i;
5     scanf("%d", &n);
6     scanf("%d", &a[i]);
7     int count = 0;
8     for (i = 0; i < n; i++) {
9         for (j = 0; j < n; j++) {
10             if (a[i] == a[j])
11                 count = count + 1;
12         }
13     }
14     if (count == a[i])
15         printf("Yes");
16     else
17         printf("No");
18     return 0;
19 }
```

Kod naprawiony i ulepszony przez PaLM-Coder

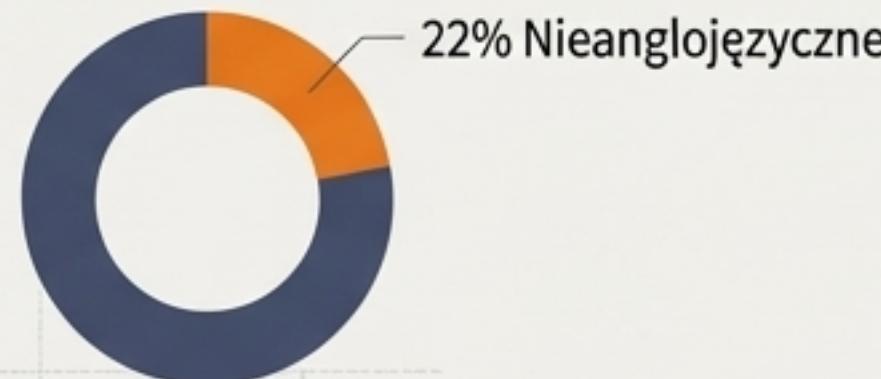
```
1 #include <stdio.h>
2 int main() {
3     int a[100], i, n, count = 0;
4     scanf("%d", &n);
5     for (i = 0; i < n; i++)
6         scanf("%d", &a[i]);
7     for (i = 0; i < n; i++) {
8         for (int j = 0; j < n; j++) {
9             if (a[i] == a[j])
10                 count = count + 1;
11         }
12     }
13     if (count == n)
14         printf("Yes");
15     else
16         printf("No");
17     return 0;
18 }
```

Wielojęzyczny Paradoks: Jak Model Trenowany Głównie po Angielsku Osiąga Doskonałość w Innych Językach

Kontekst

22%

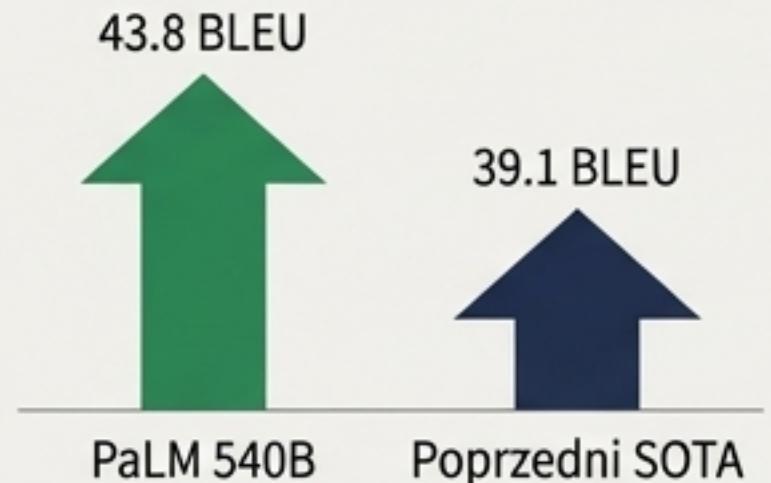
Tylko ~22% danych w zbiorze treningowym PaLM było nieanglojęzycznych, co stanowiło potencjalne ograniczenie.



Zaskakujące Wyniki

Nowy SOTA

Mimo to, PaLM 540B osiągnął wyniki **przewyższające specjalistyczne, nadzorowane modele tłumaczeniowe** w niektórych parach językowych, np. z rumuńskiego na angielski (43.8 vs 39.1 BLEU).



Kluczowe Zastrzeżenia

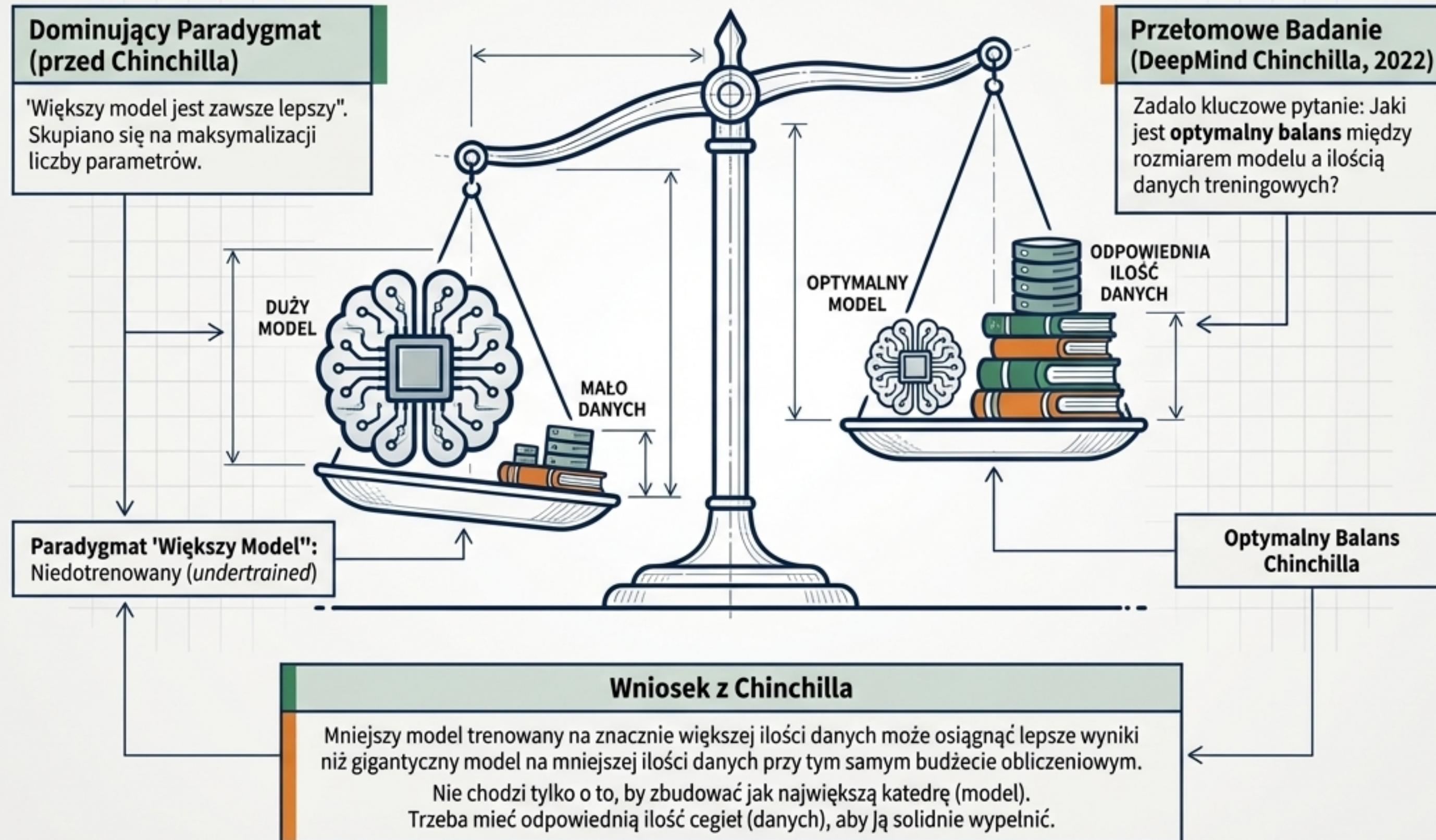
Asymetria Językowa

Model radził sobie znacznie lepiej z tłumaczeniem **NA język angielski niż Z języka angielskiego**.angielski, jako dominujący język, stał się de facto "językiem docelowym" o najsilniejszych zdolnościach generacyjnych.



Wniosek: Skład danych treningowych ma bezpośredni i dający się zmierzyć wpływ na lingwistyczne zdolności i "skrzywienia" modelu.

Wyzwanie Praw Skalowania Chinchilla: Czy 'Więcej' Zawsze Znaczy 'Lepiej'?



Pytanie na Przyszłość

Jakie nowe, jeszcze bardziej nieprzewidywalne zdolności pojawią się, gdy w końcu odnajdziemy idealną równowagę między rozmiarem sztucznej inteligencji a ogromem wiedzy, z której się uczy?