

Wyścig zbrojeń, w którym wszyscy się mylili

Jak branża AI postawiła na zły paradygmat skalowania, dopóki DeepMind nie zmienił zasad gry.

Przed 2022 rokiem w świecie AI dominowała jedna filozofia: „**Większy znaczy lepszy**”.

Panował konsensus, że kluczem do wyższej wydajności jest maksymalizacja liczby parametrów modelu.



GPT-3 | 173 mld parametrów

Gopher | 280 mld parametrów

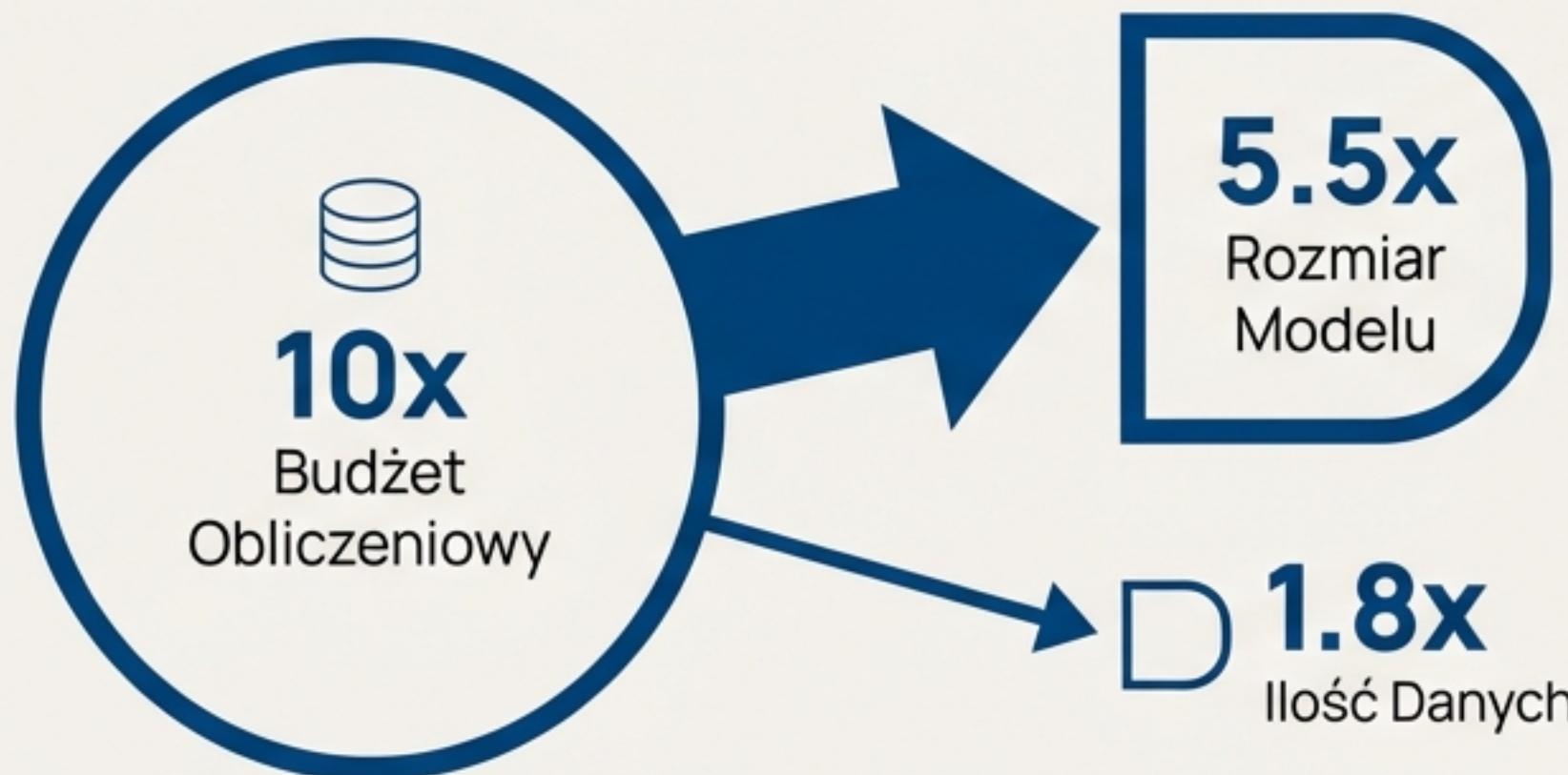
MT-NLG | 830 mld parametrów

A co, jeśli cała ta strategia była błędna? DeepMind udowodnił, że 4x mniejszy model może zdeklasować wszystkich gigantów.

Prawo, za którym podążała cała branża: Skalowanie według Kaplana (2020)

Fundamentem dla strategii „Większy jest lepszy” była praca Kaplan et al. z 2020 roku.

Ustanowiła ona logarytmiczne zależności (Power Law) między rozmiarem modelu, ilością danych a budżetem obliczeniowym (compute).



Standard Branżowy (~2021)

Model	Parametry (N)	Tokeny Treningowe (D)
GPT-3	175 mld	~300 mld
Gopher	280 mld	~300 mld
MT-NLG	530 mld	~270 mld

Pomimo drastycznych różnic w rozmiarze, największe modele były trenowane na niemal identycznej ilości danych.

Dlaczego nikt nie zboczył z utartej ścieżki?



1. Astronomiczne koszty treningu

Koszt pojedynczego cyklu treningowego liczony był w dziesiątkach milionów dolarów. Były to „one-shot „one-shot experiments” – eksperymenty, których nie można było łatwo powtórzyć.



2. Awersja do ryzyka

Nikt nie chciał ryzykować fortuny na alternatywne, niepotwierdzone podejście. Bezpieczniej było. Bezpieczniej było trzymać się „działającej receptury”.

3. Bezwładność poznaowcza (Cognitive Inertia)

Skoro większe modele dawały lepsze wyniki, logicznym krokiem wydawało się budowanie jeszcze większych. Prosta, kusząca ekstrapolacja.

4. Prostota metryki

Liczba parametrów stała się głównym, łatwym do zakomunikowania wskaźnikiem postępu w AI, zarówno dla mediów, jak i inwestorów.

DeepMind zadaje fundamentalne pytanie

Zmiana paradymatu

- Od: „Jak największy model możemy zbudować?”
- Do: „Jaki jest **NAJBARDZIEJ OPTYMALNY** model, jaki możemy zbudować?”

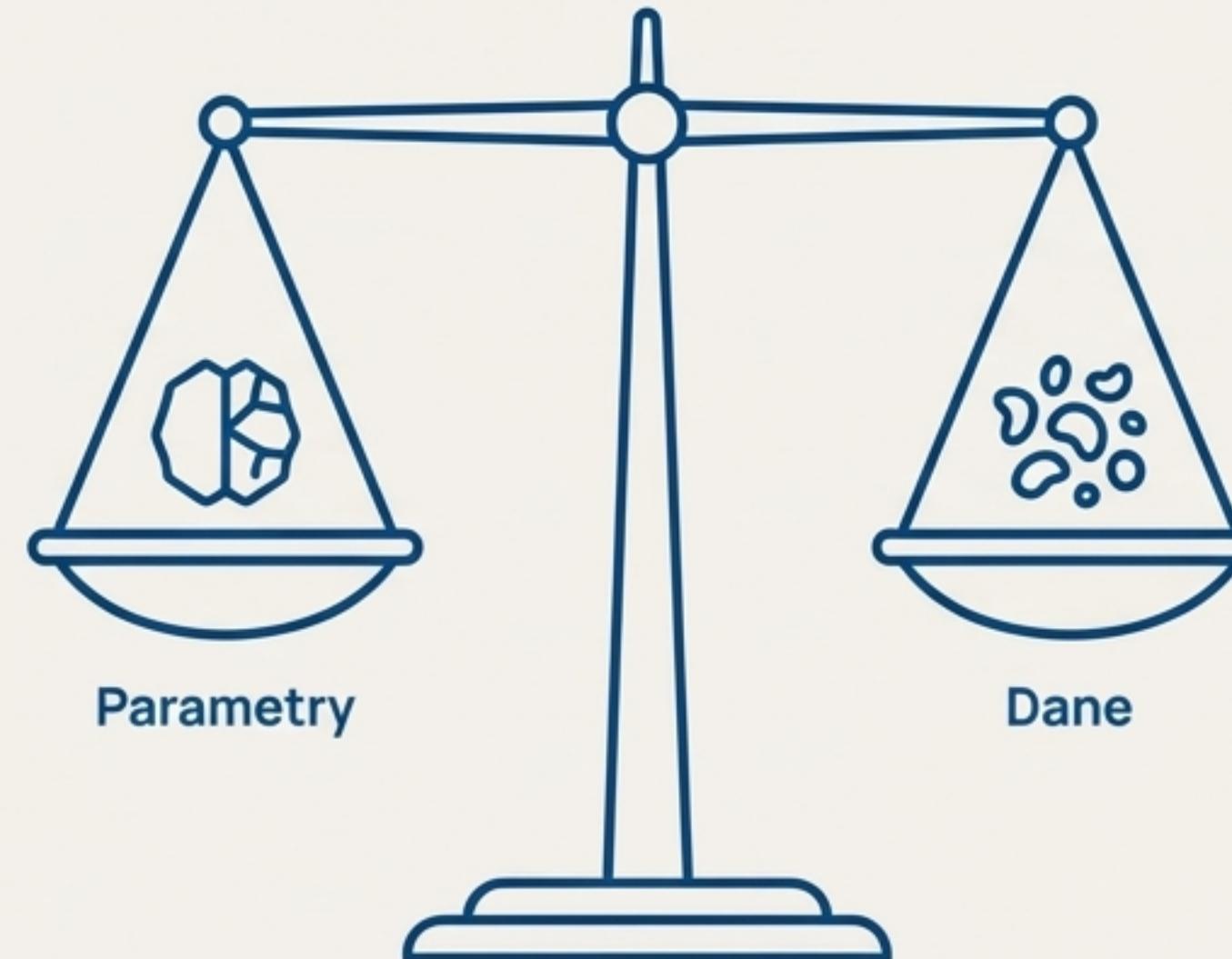
Hipoteza DeepMind

Istniejące modele były „**znacząco niedotrenowane**” (significantly undertrained).

Nowa zasada skalowania

- Parametry i tokeny powinny być skalowane **proporcjonalnie, w stosunku 1:1**.
- Prosta reguła: jeśli podwajasz liczbę parametrów, powinieneś również podwoić liczbę tokenów treningowych.

“ Mając stały budżet obliczeniowy (FLOPs), jak powinniśmy optymalnie zbalansować **rozmiar modelu** (parametry) i **ilość danych** (tokeny)?”



Rygorystyczna metodologia: Potrójna weryfikacja na ponad 400 modelach

Skala eksperymentu

- Przeszkolono **ponad 400** modeli transformatorowych.
- Zakres parametrów: **od 70 milionów do >16 miliardów**.
- Zakres danych: **od 5 miliardów do >500 miliardów tokenów**.



Trzy niezależne metody analityczne

1. **Analiza krzywych uczenia:** Klasyczne śledzenie spadku straty (loss) w czasie.



2. **Analiza profili Izo-FLOP:** Porównywanie modeli o różnej konfiguracji, ale identycznym koszcie obliczeniowym.



3. **Dopasowanie parametrycznej funkcji straty:** Modelowanie matematyczne zależności między stratą, parametrami i danymi.



Wszystkie trzy podejścia doprowadziły do tej samej, spójnej konkluzji, tworząc kompleksową „mapę” optymalnej wydajności.

Czym są profile Izo-FLOP? Znajdowanie „złotego środka”

Analogia: Budżet na samochód wyścigowy

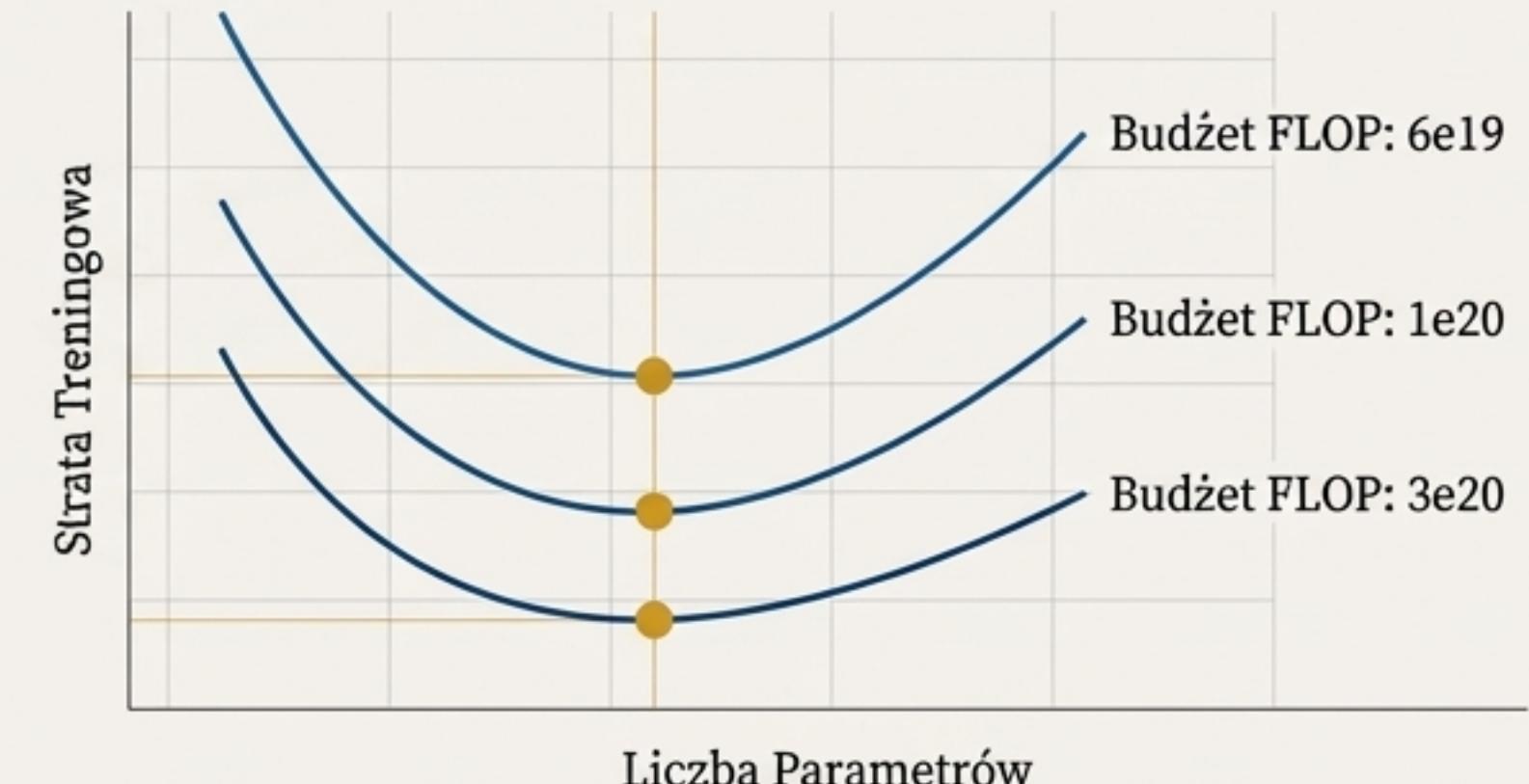
Masz stały budżet. Możesz wydać wszystko na gigantyczny silnik (parametry) i lekkie podwozie (mało danych), albo znaleźć optymalny balans między mocą silnika a aerodynamiką podwozia (parametry i dane).



Definicja techniczna

- „Izo-FLOP” oznacza „taki sam koszt obliczeniowy” (Floating Point Operations).
- Dla każdego budżetu FLOP, trenuje się wiele modeli o różnej kombinacji parametrów (N) i tokenów (D), by znaleźć punkt, w którym strata (loss) jest minimalna.

Minimalizacja Straty przy Stałym Budżecie Obliczeniowym



Eksperyment Chinchilla: Ten sam budżet, radykalnie inna strategia

DeepMind postanowiło zweryfikować swoją hipotezę w ostatecznym teście. Zamiast ekstrapolować wyniki, przeprowadzili bezpośrednie porównanie przy identycznym budżecie obliczeniowym jak dla modelu Gopher.

GOLIATH (Gopher)

Stary Paradygmat



280
miliardów parametrów

~300
miliardów tokenów

DAVID (Chinchilla)

Nowy Paradygmat



70 (4x mniej)
miliardów parametrów

1.4 (4x więcej)
biliona tokenów

Identyczny koszt. Całkowicie inna alokacja zasobów.

Rezultat: Mniejsza i mądrzejsza Chinchilla dominuje na wszystkich frontach

MMLU (Test rozumowania)



Wynik pobił prognozy ekspertów na czerwiec 2023, mimo że praca pochodzi z marca 2022.

RACE-h (Czytanie ze zrozumieniem)



+10.7 punktu procentowego przewagi.

BIG-bench (Złożone zadania językowe)

+10.7%

Średnio wyższa wydajność w 62 zadaniach.

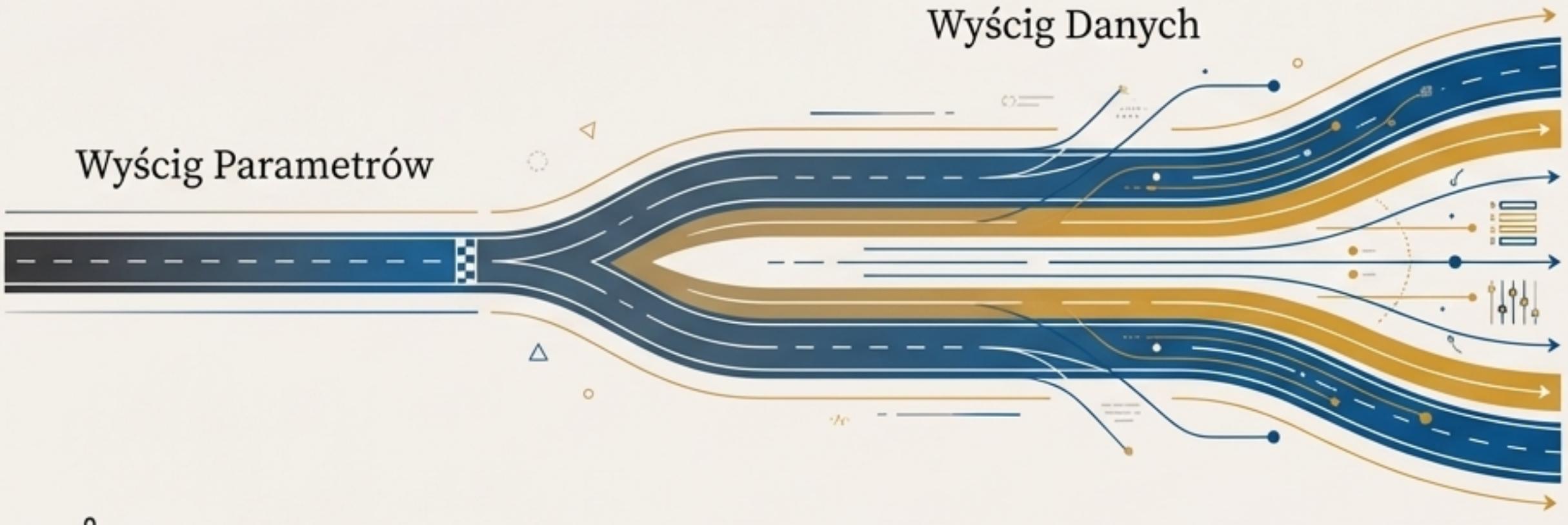
Closed-book QA (Pytania z pamięci)



Chinchilla ustanowiła nowe rekordy SOTA (State-of -the-Art).

Chinchilla pokonała nie tylko Gophera, ale także GPT-3 (175B) i MT-NLG (530B), będąc od nich znacznie mniejszą.

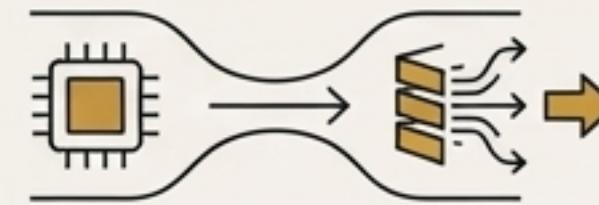
Nowy paradymat: Od wyścigu na parametry do wyścigu na dane



Demokratyzacja AI

Manrope Bold Source Serif Pro

Mniejsze modele oznaczają drastycznie tańsze wnioskowanie (inference) i dostrajanie (fine-tuning). Wydajność na poziomie GPT-3 staje się dostępna dla startupów.



Przesunięcie wąskiego gardła

Manrope Bold Source Serif Pro

Głównym ograniczeniem przestaje być moc obliczeniowa, a staje się nim dostęp do wysokiej jakości danych treningowych.



Nowy wyścig

Manrope Bold Source Serif Pro

Koniec „wyścigu zbrojeń” na liczbę parametrów. Początek „wyścigu danych” – kto zdobędzie i przygotuje lepsze zbiory danych.

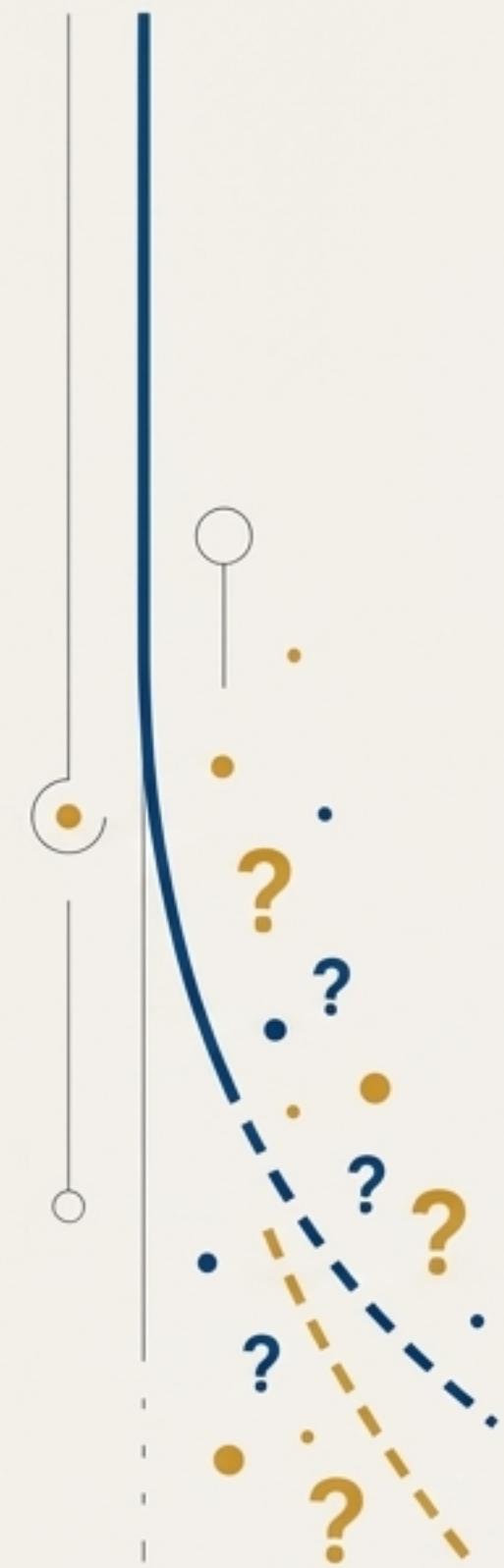
Prognoza

Wymagania dla optymalnego modelu 1T

Parametry	1 Bilion
Wymagane Tokeny	>21 Bilionów

Ograniczenia i otwarte pytania: Co jeszcze kryją prawa skalowania?

- Krzywizna w dużej skali:** Zaobserwowana krzywizna sugeruje, że optymalne modele dla bardzo dużych budżetów mogą być **jeszcze mniejsze** niż przewiduje obecny model.
- Trenig na mniej niż jednej epoce:** Cała analiza dotyczy sytuacji, w której każdy token jest widziany tylko raz. Jak zmienią się prawa skalowania przy treningu wieloepokowym (multi-epoch)?
- Ryzyko przeuczenia (overfitting):** Czy wielokrotne przetwarzanie tych samych danych prowadzi do pogorszenia generalizacji modelu?
- Założenia Praw Skalowania:** Założenie o idealnej zależności logarytmicznej (power law) może nie być w pełni prawdziwe w ekstremalnych skalach.



Nowe wielkie wyzwanie dla całej branży

Odkrycia Chinchilli dowodzą, że przyszłe, jeszcze potężniejsze modele będą wymagały treningu na **bilionach** wysokiej jakości **tokenów**.

Skąd weźmiemy te dane?

Jak będziemy je pozyskiwać i filtrować w sposób odpowiedzialny?
Jak unikniemy wzmacniania istniejących w nich uprzedzeń na niespotykaną dotąd skalę?