

# Nowy paradymat: jak mniejsze modele, trenowane dłużej, pokonują gigantów.

W lutym 2023 Meta AI przedstawiła tezę, która zmieniła bieg rozwoju AI: wydajność nie zależy wyłącznie od liczby parametrów.

- **Dowód:** LLaMA 13B, model 10-krotnie mniejszy, przewyższa wydajnością GPT-3 (175B) na większości benchmarków.
- **Transparentność:** Model trenowany wyłącznie na publicznie dostępnych danych, co stanowi zerwanie z tradycją tajnych, firmowych zbiorów.
- **Zmiana Priorytetu:** Kluczowa staje się optymalizacja kosztów *wnioskowania (inference)*, a nie kosztów *treningu*.
- **Dziedzictwo:** Publikacja, która zapoczątkowała eksplozję innowacji w świecie open-source AI.

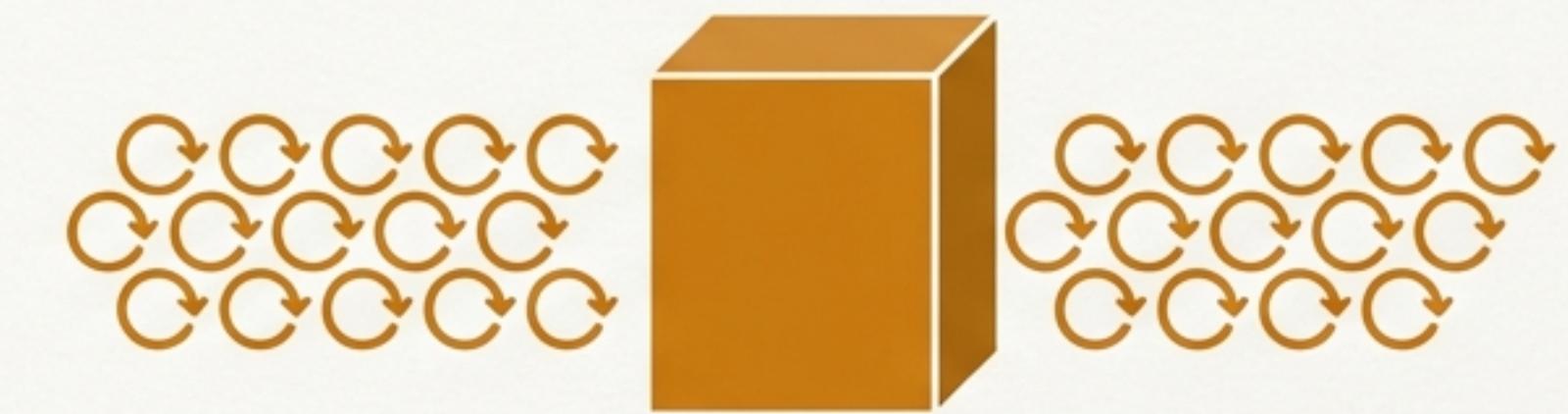


# Ekonomia AI na nowo: Prawdziwym kosztem jest wnioskowanie, nie trening.

## Stare Podejście: Optymalizacja Budżetu Treningowego



## Hipoteza LLaMA: Optymalizacja Budżetu Wnioskowania



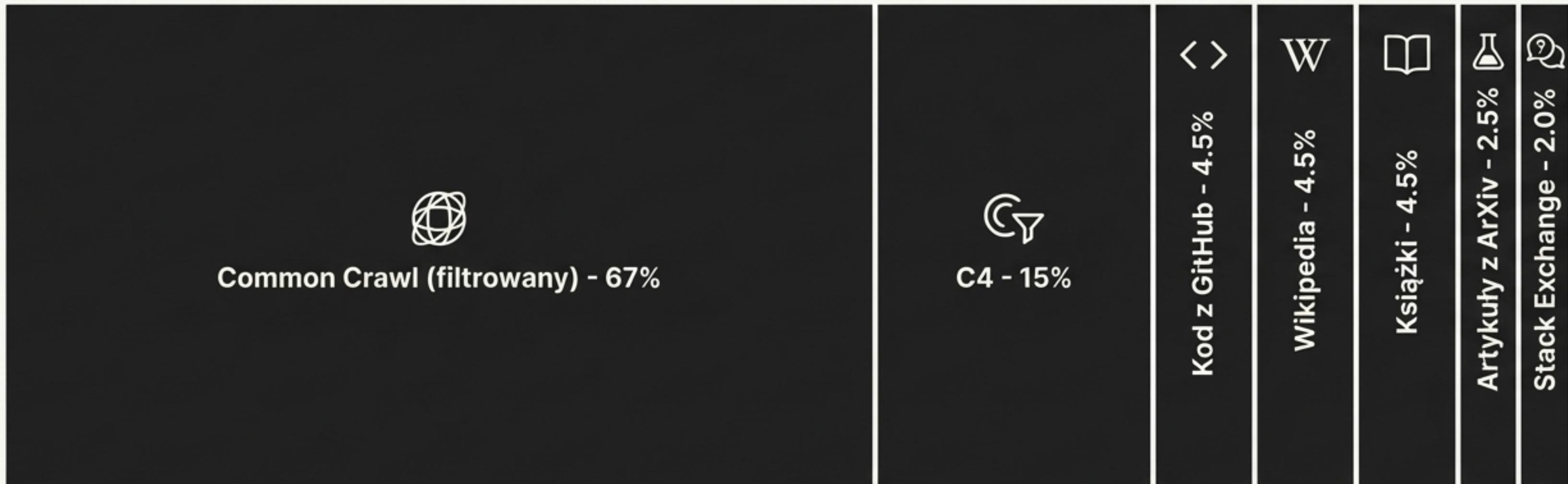
- **Założenie:** Więcej parametrów = lepsza wydajność.
- **Fokus:** Jednorazowa, ogromna inwestycja obliczeniowa w trening.
- **Problem:** Ignoruje bieżący koszt obsługi milionów zapytań dziennie, który staje się dominujący w cyklu życia modelu.

- **Założenie:** Mniejszy model trenowany na większej ilości danych jest tańszy i szybszy w masowym użyciu.
- **Fokus:** Długoterminowa efektywność i koszt per-zapytanie.
- **Obserwacja:** Wydajność modelu 7B wciąż rosła nawet po przetworzeniu 1 biliona tokenów, co przeczyło dotychczasowym prawom skalowania.

Pytanie nie brzmi 'jaki model najszybciej wytrenować do poziomu X?', lecz 'jaki model o wydajności X jest najtańszy w masowej eksploatacji?'.

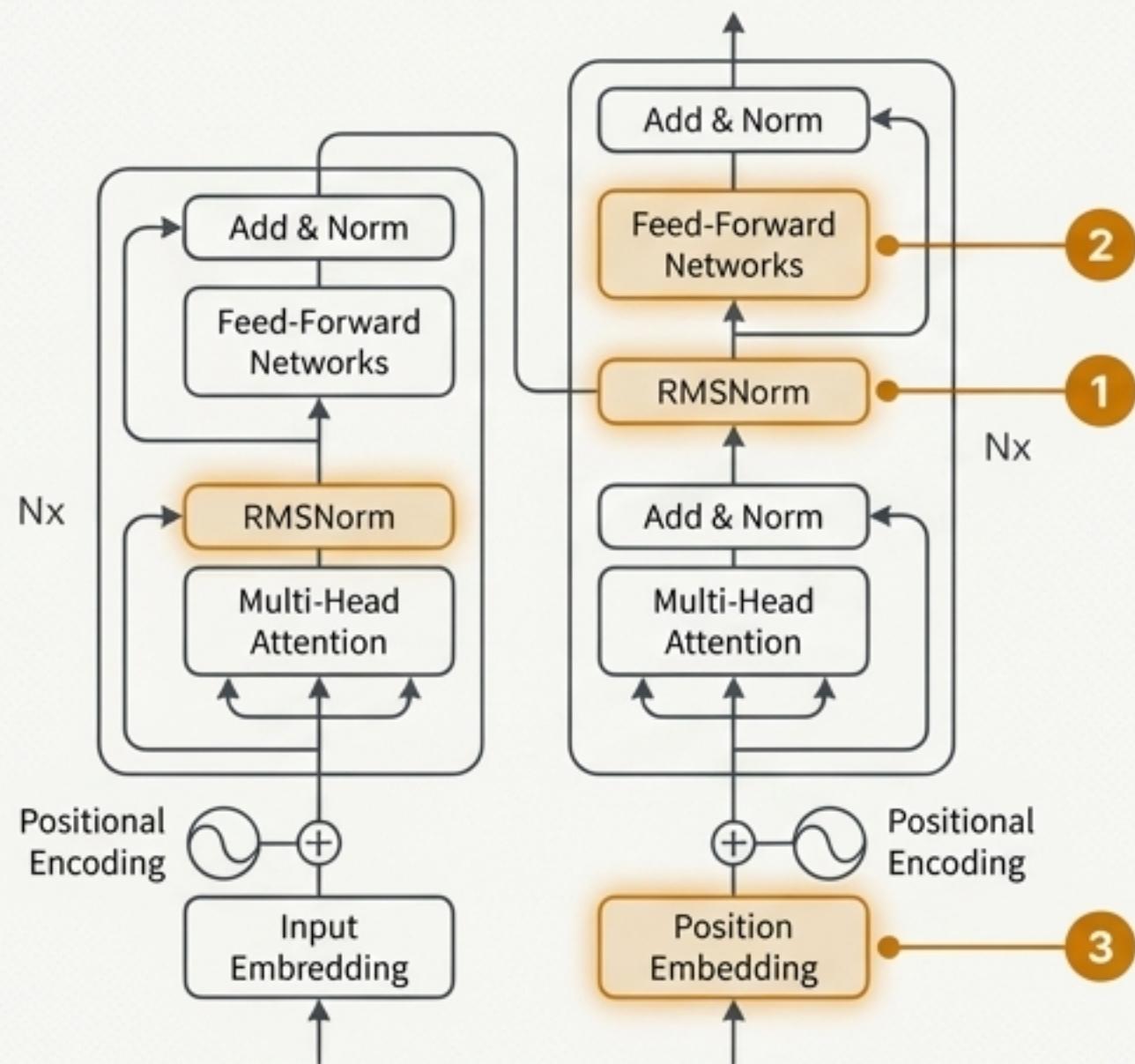
# Fundament wiedzy: 1.4 biliona tokenów z publicznie dostępnych źródeł.

Pełna transparentność zbioru treningowego – w przeciwieństwie do tajemniczych i niedostępnych danych używanych przez konkurencję (np. ‘Books2’ GPT-3).



# Architektura: Inteligentna ewolucja, a nie rewolucja.

Zamiast tworzyć nową architekturę od zera, LLaMA czerpie z istniejącej architektury Transformer, wprowadzając trzy kluczowe, synergiczne ulepszenia.



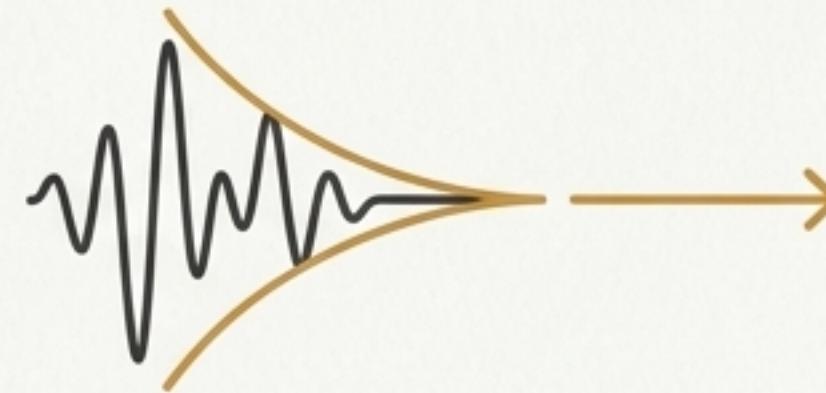
## Trzy Filary Ulepszeń:

- 1. Pre-normalizacja z RMSNorm:** Dla stabilności treningu.
- 2. Funkcja Aktywacji SwiGLU:** Dla większej wydajności.
- 3. Rotary Position Embeddings (RoPE):** Dla lepszego rozumienia długich sekwencji.

\*To jak tuning silnika, a nie projektowanie nowego silnika.\*

**Wniosek: Siła LLaMA nie leży w pojedynczej innowacji, lecz w inteligentnym połączeniu sprawdzonych rozwiązań.**

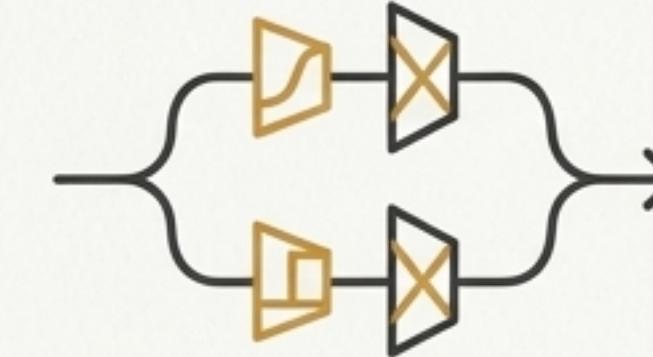
# Kluczowe ulepszenia architektoniczne w praktyce.



## Pre-normalizacja (RMSNorm)

**Cel:** Poprawa stabilności treningu. Zamiast normalizować wyjście każdej pod-warstwy transformera, normalizowane jest jej wejście.

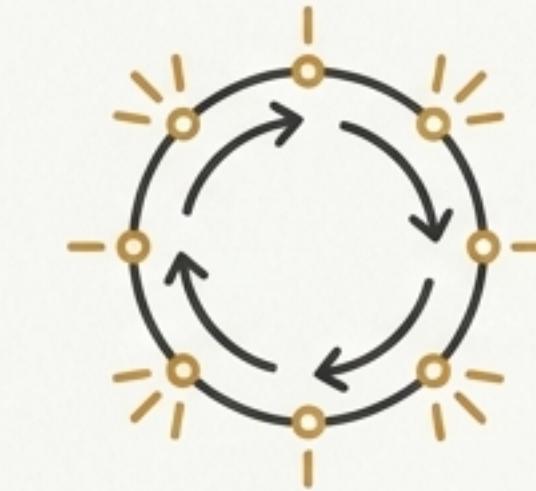
**Korzyść:** Zapobiega problemowi zanikających/eksplodujących gradientów podczas wielotygodniowych sesji treningowych.



## Funkcja Aktywacji SwiGLU

**Cel:** Zastąpienie standardowej funkcji ReLU w celu poprawy wydajności (zainspirowane modelem PaLM).

**Korzyść:** Zapewnia mierzalny wzrost efektywności obliczeniowej przy wymiarze `2/3 \* 4d`.



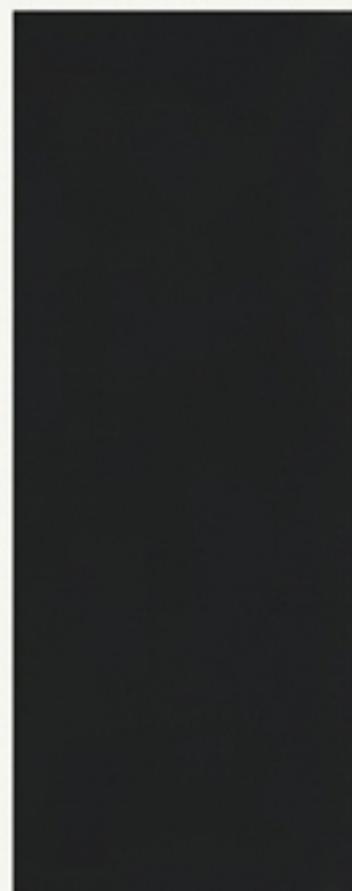
## Rotary Embeddings (RoPE)

**Cel:** Zastąpienie absolutnych osadzeń pozycyjnych na rzecz względnych (zainspirowane GPT-Neo).

**Korzyść:** Umożliwia modelowi lepsze rozumienie relacji między słowami w długich sekwencjach.

# Inżynieria w służbie wydajności: Jak wytrenowano 65B w 21 dni.

**Wynik:** Przetwarzanie ~380 tokenów/sekundę/GPU na 2048 kartach A100.  
Trening na 1.4 biliona tokenów zajął około **21 dni**.



Standardowe  
Podejście

LLaMA (xformers +  
Checkpointing)

Ogromne  
oszczędności  
pamięci



## Kluczowe Optymalizacje:

- **Biblioteka xformers:** Wykorzystano efektywną implementację mechanizmu uwagi, która nie przechowuje pełnych macierzy uwagi, drastycznie redukując zużycie pamięci.
- **Activation Checkpointing:** Zamiast przechowywać wszystkie aktywacje dla propagacji wstecznej, zapisywano tylko te kosztowne obliczeniowo (np. wyjścia z warstw liniowych) i przeliczano resztę 'w locie'.

**Wniosek:** Oszczędności pamięci i zoptymalizowana komunikacja między GPU uczyniły projekt wykonalnym na tej skali.

**Starcie tytanów: LLaMA 13B deklasuje 10x większego GPT-3 175B**

# **MNIEJSZY MODEL WYGRYWA.**

Benchmark Category	Benchmark	LLaMA 13B	GPT-3 175B
<b>Rozumowanie zdroworozsądkowe</b>	HellaSwag	<b>79.2</b>	78.9
	WinoGrande	<b>73.0</b>	70.2
<b>Odpowiedzi na pytania</b>	TriviaQA (0-shot)	<b>56.6</b>	<i>43.5 (Gopher)</i>
	NaturalQuestions	<b>20.1</b>	14.6

**Kluczowa Implikacja:** Mniejszy model trenowany na większej ilości danych posiada gęstszą, lepiej zorganizowaną wiedzę na każdy parametr.

**Aspekt Demokratyzacji:** Model tej klasy staje się możliwy do uruchomienia (z wysiłkiem) na pojedynczej, potężnej konsumenckiej karcie graficznej.

# LLaMA 65B: Walka w najwyższej lidze z Chinchilla i PaLM.

## Ogólna Wydajność

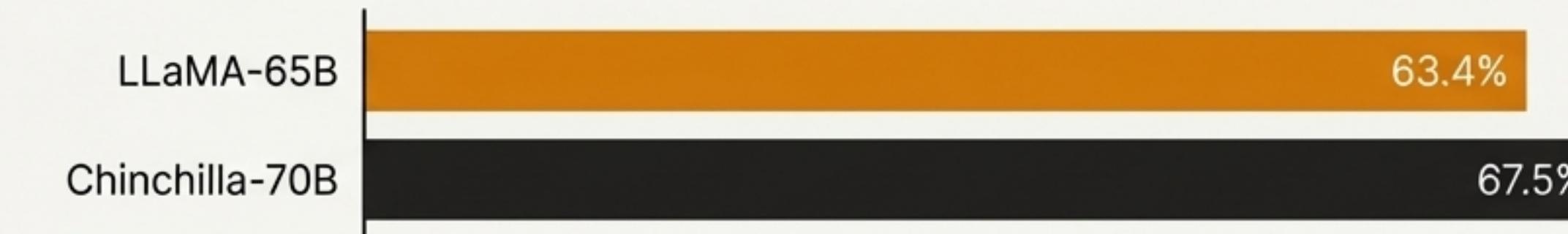
LLaMA-65B jest w pełni konkurencyjny z Chinchilla-70B i przewyższa go na większości zadań z kategorii 'common sense reasoning'.

### Niespodziewane Zwycięstwo (Benchmark GSM8k - matematyka)



**Wniosek:** Generalista niemal pokonuje (i przewyższa w maj1@k) specjalistę na jego własnym polu, mimo braku specjalistycznego treningu.

### Uczciwa Ocena Słabości (Benchmark MMLU)



**Prawdopodobna przyczyna:** Znacznie mniejsza ilość danych książkowych i akademickich w zbiorze treningowym LLaMA (177GB) w porównaniu do konkurencji (do 2TB).

# Lustro niedoskonałego świata: Ograniczenia LLaMA.

Kluczowa obserwacja: Modele odzwierciedlają zarówno wiedzę, jak i społeczne uprzedzenia zawarte w danych treningowych.

## 1. Toksyczność (RealToxicityPrompts)

Poziom toksyczności rośnie wraz z rozmiarem modelu.

LLaMA-7B



0.106

LLaMA-65B



0.128

LLaMA-65B (wynik 0.128) generuje więcej toksycznych treści niż LLaMA-7B (wynik 0.106).

## 2. Uprzedzenia (CrowS-Pairs)

Zmierzone stereotypy w 9 kategoriach.

LLaMA-65B wykazuje szczególnie wysoki poziom uprzedzeń w kategorii **religii** (wynik 79.0).

## 3. Uprzedzenia Płciowe (WinoGender)

Model popełnia znacznie więcej błędów, gdy płeć w zdaniu jest sprzeczna ze stereotypem zawodowym.

Przykład: Poprawność spada przy zdaniach typu 'pielęgniarka' (mężczyzna) lub 'inżynierka' (kobieta), co dowodzi internalizacji społecznych uprzedzeń.

# Konsekwencje zwycięstwa: Demokratyzacja badań i eksplozja innowacji.

## Trwały Wpływ na Ekosystem AI

- **Demokratyzacja Badań:** Uwolnienie wag modeli dla społeczności naukowej umożliwiło uniwersitetom, startupom i małym zespołom prowadzenie badań na modelach klasy SOTA.
- **Fundament Open-Source:** LLaMA stała się bazą dla setek kolejnych modeli i narzędzi, napędzając bezprecedensowy rozwój otwartego AI.
- **Nowa Ekonomia AI:** Efektywność wnioskowania na stałe stała się kluczowym priorytetem w projektowaniu modeli.

## Instruction Fine-Tuning



Wynik w benchmarku MMLU wzrasta z

**63.4% do 68.9%**

**Sygnal:** Pokazuje to ogromny potencjał "wyrównywania" (alignment) modeli bazowych do praktycznych zastosowań.

# Pytanie na przyszłość.

Tworząc coraz potężniejsze modele na danych z całego internetu,  
czy budujemy lepsze narzędzia do rozumienia świata,  
czy raczej coraz doskonalsze lustra, które bezkrytyczne  
odbijają wszystkie jego niedoskonałości?

I jak mądrze skalować możliwości,  
nie skalując proporcjonalnie potencjalnych szkód?