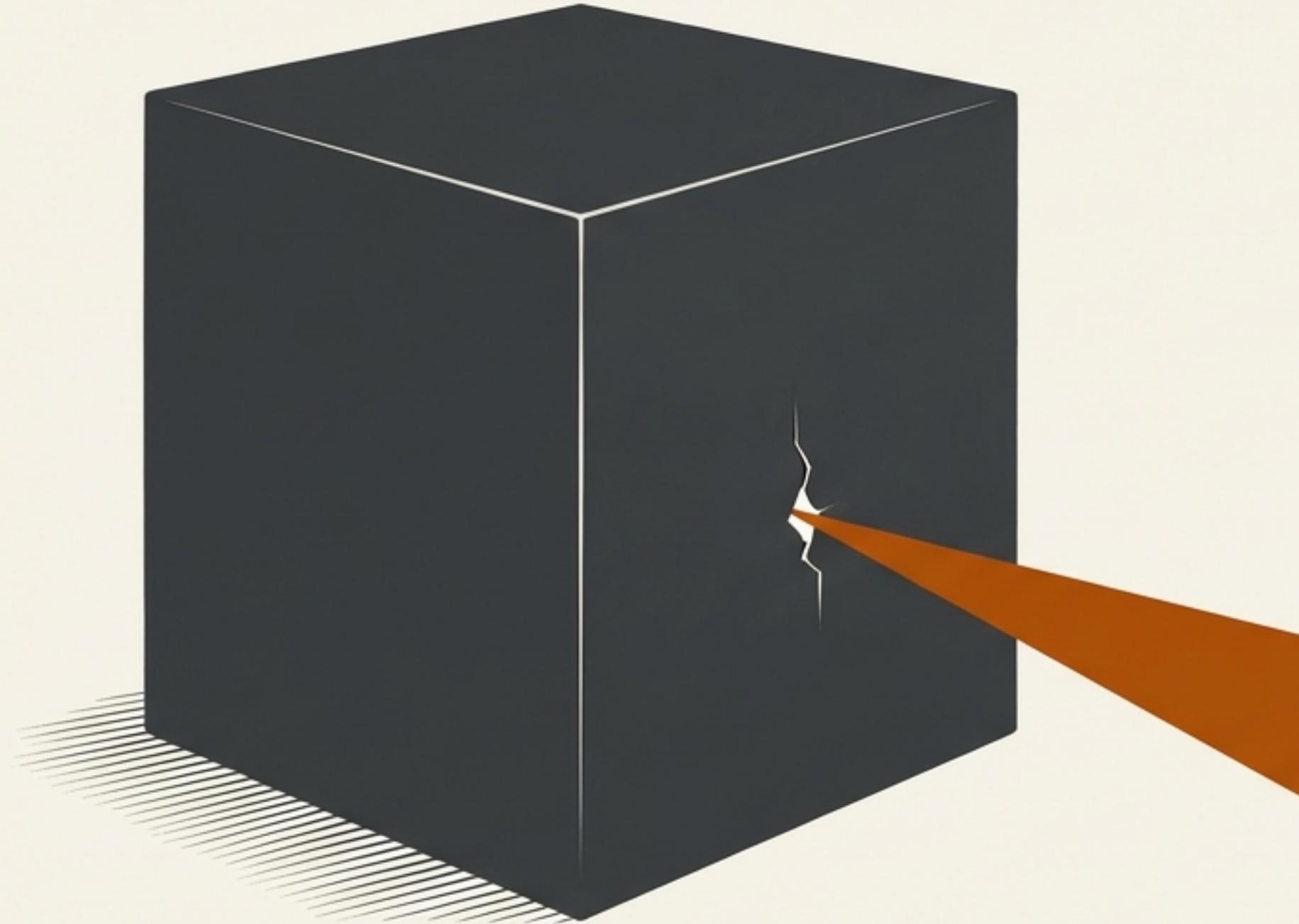


# Przełamywanie murów fortecy AI



Wielkie modele językowe, takie jak GPT-3, zrewolucjonizowały sztuczną inteligencję, ale pozostają zamkniętymi "czarnymi skrzynkami".

Dostęp dla badaczy jest ograniczony do komercyjnych API, co przypomina próbę diagnozowania silnika przez zamkniętą maskę samochodu. Ogranicza to postęp w kluczowych obszarach, takich jak solidność, stronniczość i toksyczność.

Meta AI rzuca wyzwanie tej filozofii zamkniętych badań, wprowadzając OPT: Open Pre-trained Transformer Language Models.

**Centralne pytanie tej misji:** Czy nauka może naprawdę rozwijać się w tajemnicy?

# Rodzina modeli OPT: Potęga oddana w ręce badaczy

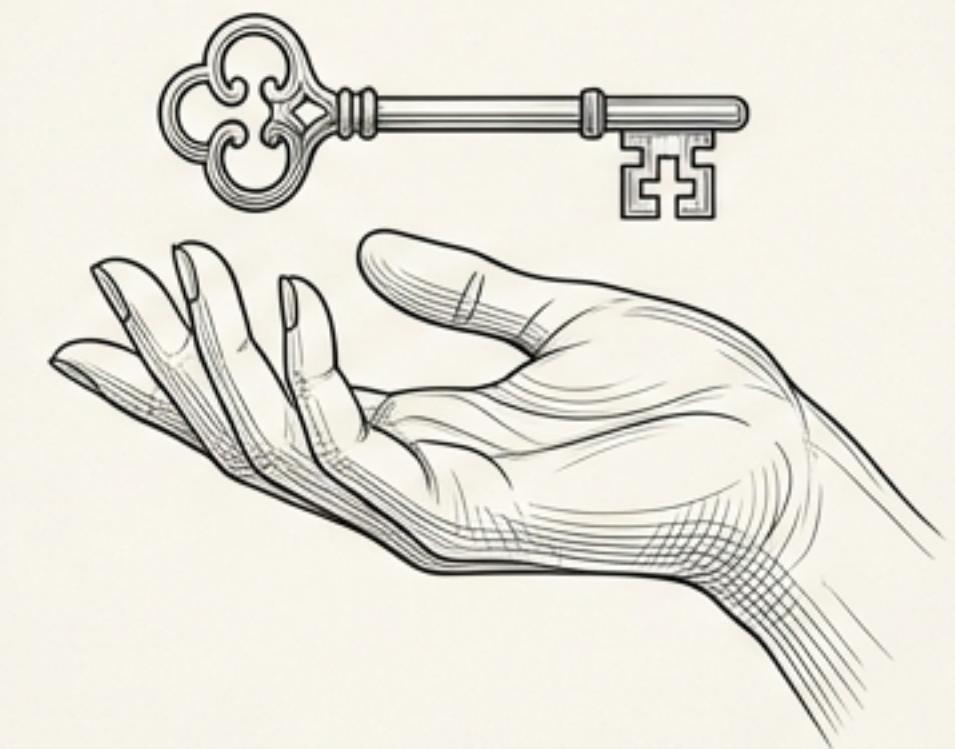
Kompletna rodzina modeli, od 125 milionów do 175 miliardów parametrów.

**OPT-175B** celowo dorównuje liczbę parametrów GPT-3 (175 miliardów), aby umożliwić bezpośrednie porównania.

Misja to nie tylko budowa modelu, ale jego 'wyzwolenie'. Meta AI udostępnia pełne wagi modelu, a nie tylko dostęp przez API.

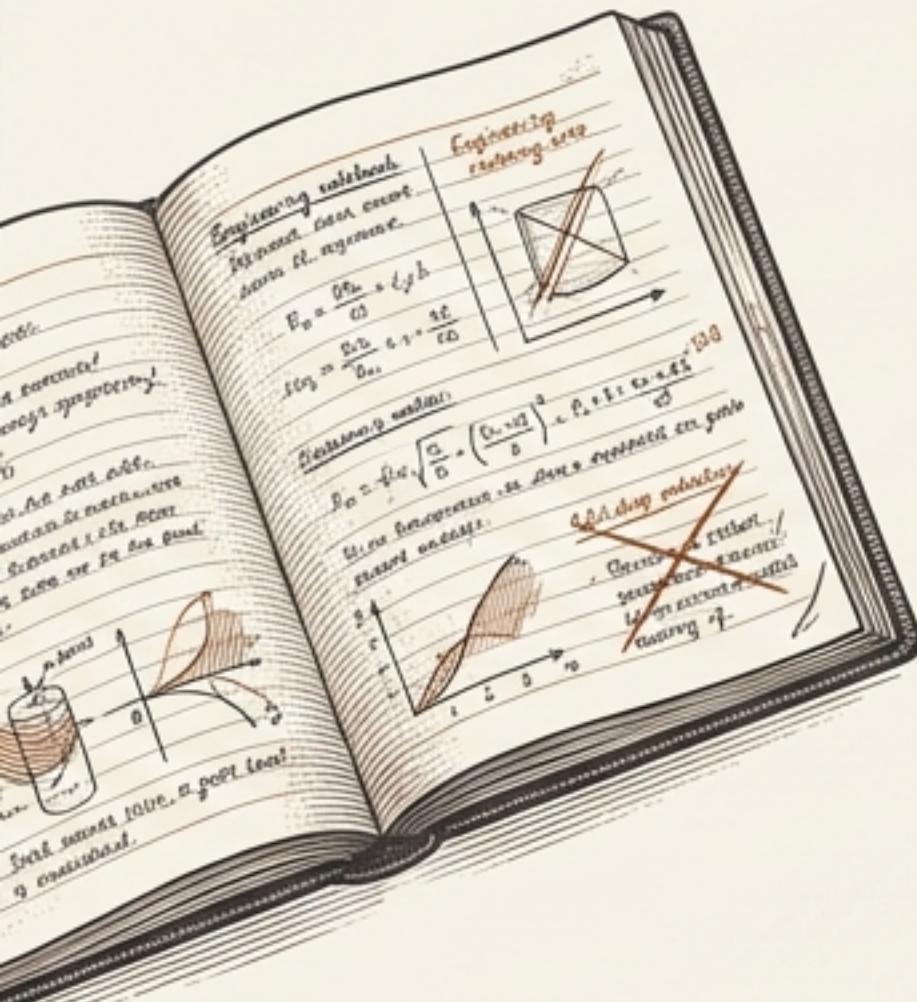
To krok w stronę demokratyzacji badań nad sztuczną inteligencją na dużą skalę, dający szerszej społeczności dostęp do najnowocześniejszych narzędzi.

Model	#L	#H	d <sub>model</sub>
125M	12	12	768
350M	24	16	1024
1.3B	24	32	2048
2.7B	32	32	2560
6.7B	32	32	4096
13B	40	40	5120
30B	48	56	7168
66B	64	72	9216
<b>175B</b>	<b>96</b>	<b>96</b>	<b>12288</b>



# Dziennik Pokładowy: Brutalnie szczera historia treningu

Meta AI opublikowała szczegółowy "dziennik pokładowy" (logbook), dokumentujący cały proces trenowania OPT-175B.



- To bezcenny zapis **porażek, wyzwań i improwizowanych rozwiązań**, które są zwykle pomijane w publikacjach naukowych skupionych wyłącznie na sukcesach.
- Dziennik ujawnia, że trenowanie modeli na taką skalę **jest bardziej sztuką niż czystą inżynierią**, wymagającą ciągłych interwencji.
- To manifest **transparentności** i nieocenione źródło wiedzy dla każdego, kto podejmuje się podobnych wyzwań.



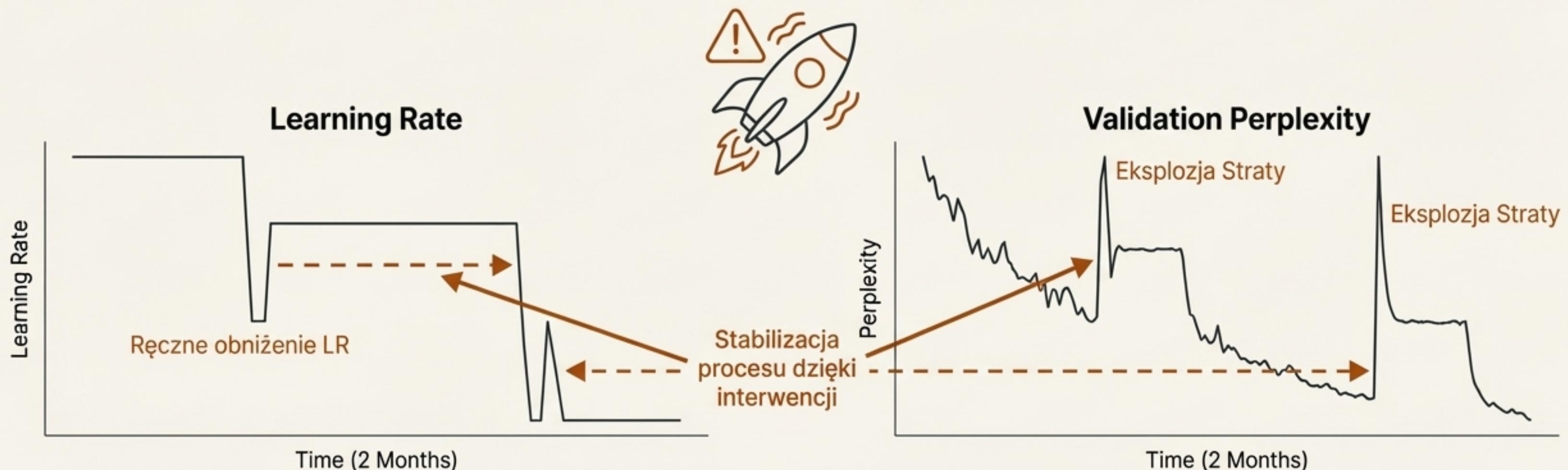
# Próby ognia: Utrzymanie eksperymentalnej rakiety w całości

Trening OPT-175B wymagał ponad **35 ręcznych restartów** w ciągu 2 miesięcy z powodu **awarii sprzętowych**.  
**Wymieniono ponad 100 serwerów.**

**Problem dywergencji straty (Loss Divergence):** Funkcja straty nagle "eksplodowała" zamiast maleć, **zmuszając zespół do powrotu do wcześniejszych punktów kontrolnych i tracąc cenne godziny pracy.**

**Rozwiążanie:** Ręczne obniżanie współczynnika uczenia (learning rate) i progu obcinania gradientu (gradient clipping) w celu ustabilizowania procesu – "**mniejsze, ostrożniejsze kroki**".

Te interwencje pokazują, jak niestabilny i nieprzewidywalny jest proces treningu na masową skalę.



# Ślad węglowy: 7-krotna poprawa efektywności 🌱

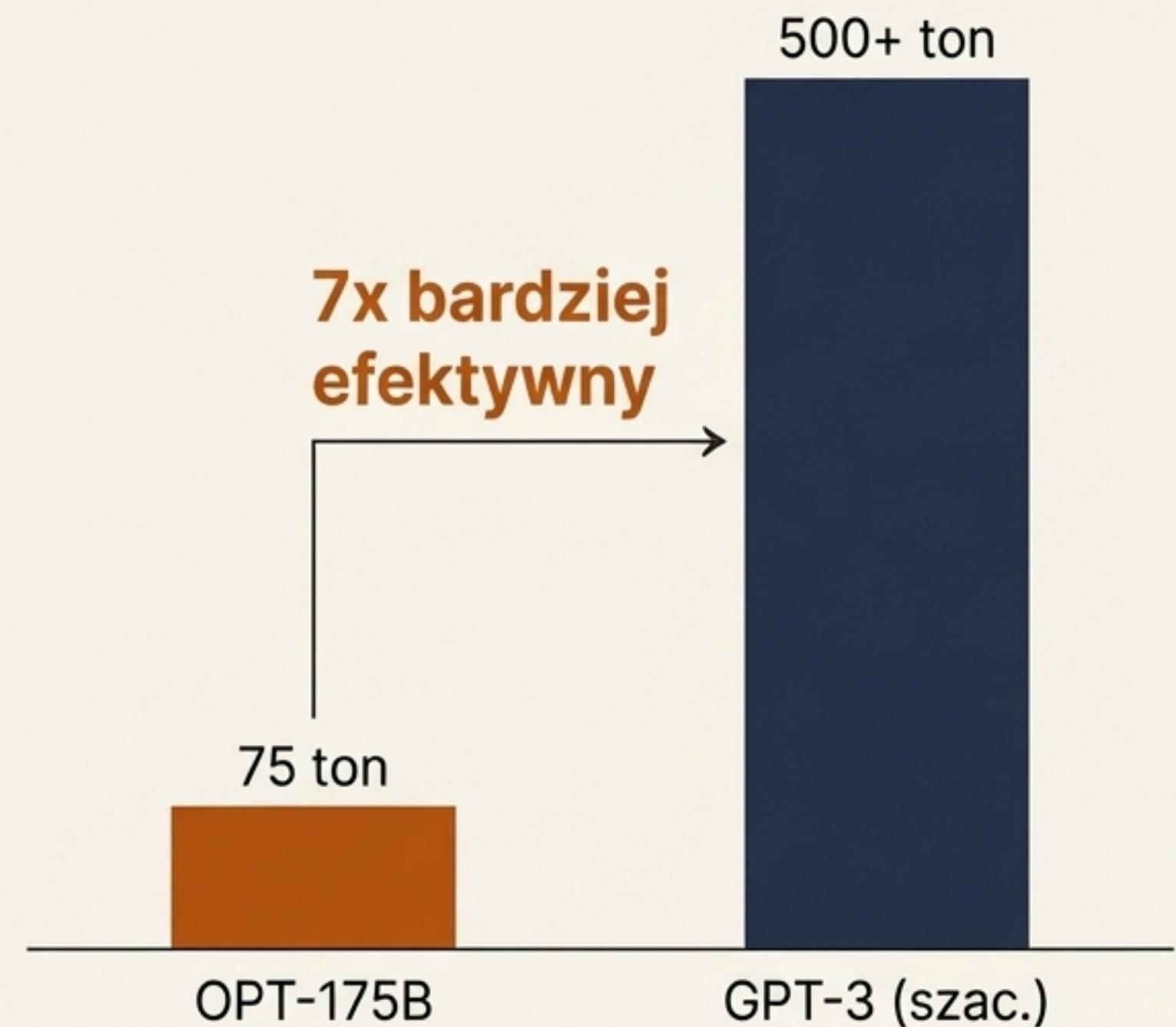
Trening OPT-175B wygenerował około **75 ton** ekwiwalentu CO<sub>2</sub>.

Szacunki dla GPT-3 (zewnętrzne, nieoficjalne) wskazują na ponad **500 ton** CO<sub>2</sub>.

To niemal **7-krotna redukcja śladu węglowego**.

Osiągnięto to dzięki wykorzystaniu bardziej energooszczędnego procesorów graficznych (NVIDIA A100 80GB) i zoptymalizowanemu procesowi treningowemu.

Meta AI opublikowała swoją metodologię, w przeciwieństwie do OpenAI, demonstrując, że wielkie modele można trenować w sposób bardziej zrównoważony.



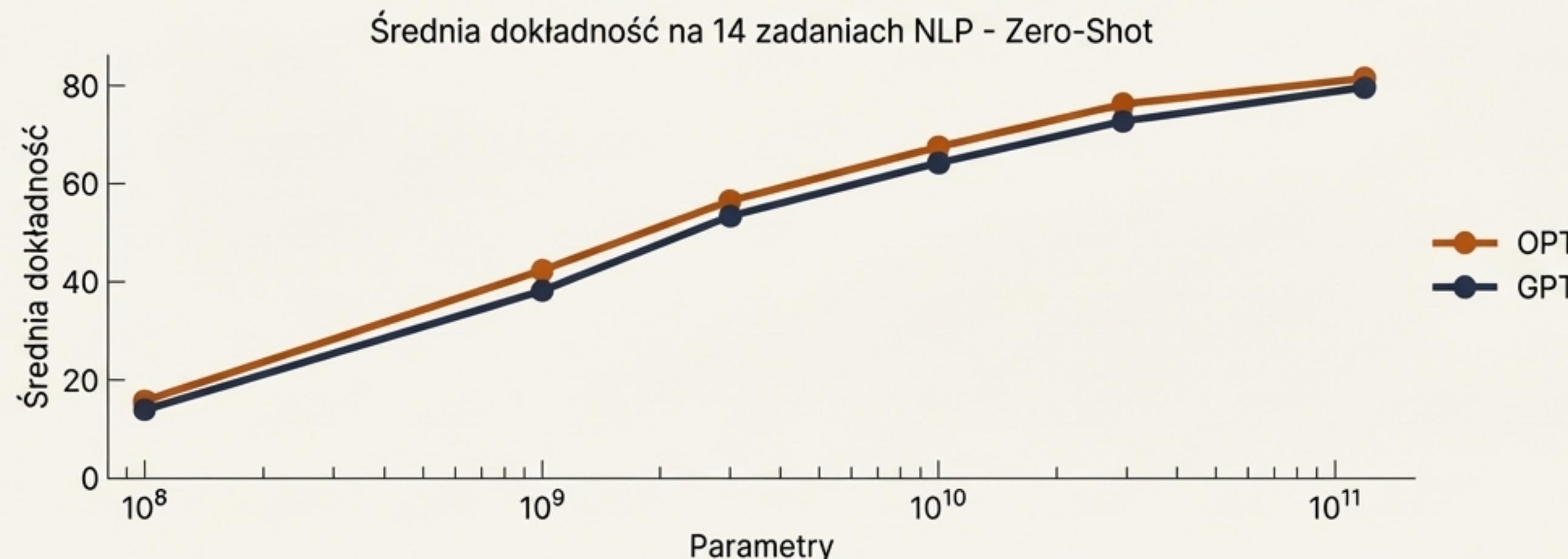
# Wydajność w benchmarkach: Otwarty model dorównuje zamkniętemu gigantowi

OPT został przetestowany na 16 standardowych benchmarkach NLP (Natural Language Processing).

Testy przeprowadzono w dwóch trybach: zero-shot (bez przykładów) i few-shot (z kilkoma przykładami).

Krzywe wydajności dla różnych skal modeli OPT i GPT-3 są niemal identyczne.

**Kluczowy wniosek:** **OPT osiąga wydajność porównywalną z GPT-3 przy ułamku kosztu węglowego.** To dowód, że otwartość nie oznacza kompromisu w kwestii jakości.



# Zdolności emergentne: Niespodzianka w dialogu



OPT-175B był trenowany bez nadzoru, czyli nie uczyono go konkretnych zadań, takich jak **prowadzenie rozmowy**. Mimo to, w testach dialogowych, model osiągnął **wyniki porównywalne lub lepsze niż BlenderBot 1** (specjalnie dostrajany do konwersacji).

Jest to przykład '**zdolności emergentnych**' – umiejętności, które pojawiają się spontanicznie, gdy model osiąga odpowiednią skalę. **Nikt nie zaprogramował w OPT zdolności konwersacyjnych** – wyłoniły się one naturalnie z wzorców językowych w danych treningowych.

Porównanie modeli w dialogu (ConvAI2)

Model	Rodzaj treningu	Perplexity (niższa = lepiej)	F1 (wyższa = lepiej)
OPT-175B	Bez nadzoru (Unsupervised)	18.7	16.3
BlenderBot 1	Z nadzorem (Supervised)	18.5	16.1

# Paradoks toksyczności: Głębsze zrozumienie za cenę ryzyka

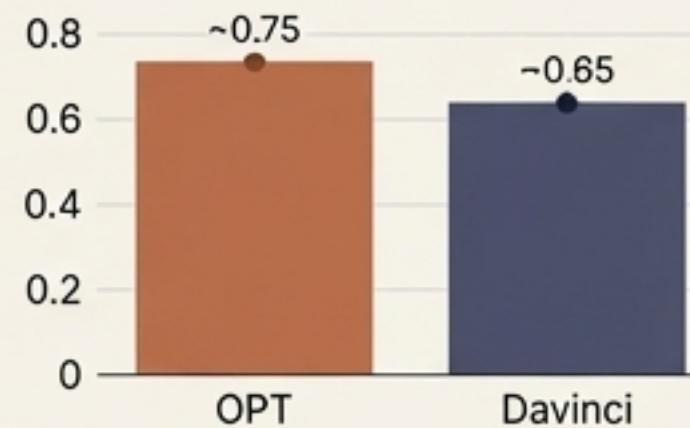
OPT był trenowany na mniej filtrowanych danych niż GPT-3, włączając w to surowe treści z serwisu Reddit.

**Paradoks:** Mimo to, OPT-175B jest **lepszy w wykrywaniu mowy nienawiści** niż GPT-3 (Davinci). Prawdopodobnie dlatego, że 'widział więcej przypadków' toksyczności, niczym lekarz z większym doświadczeniem klinicznym.

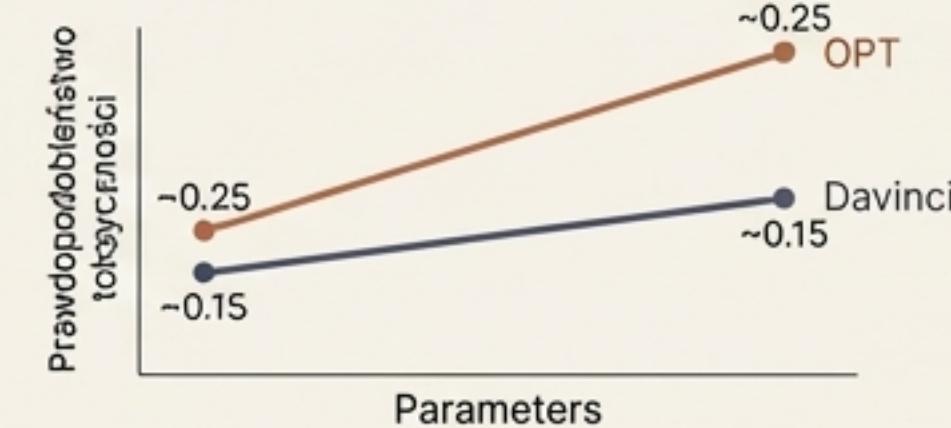
**Druga strona medalu:** Jest również **bardziej skłonny do generowania toksycznych treści** i utrwalania stereotypów. To fundamentalny dylemat: niefiltrowane dane dają gębsze zrozumienie języka, ale jednocześnie uczą model naśladowania jego najgorszych aspektów.



**Hate speech detection** (Tabela 3)



**RealToxicityPrompts** (Rysunek 5)



# Znane ograniczenia: Akt brutalnej szczerości

Autorzy otwarcie wymieniają kluczowe wady modelu:



## Problemy z wykonywaniem poleceń:

Zamiast wykonać instrukcję, model symuluje rozmowę o niej.



## Pętle powtórzeń:

Generuje te same frazy w kółko.



## Halucynacje:

Tworzy całkowicie fałszywe informacje.

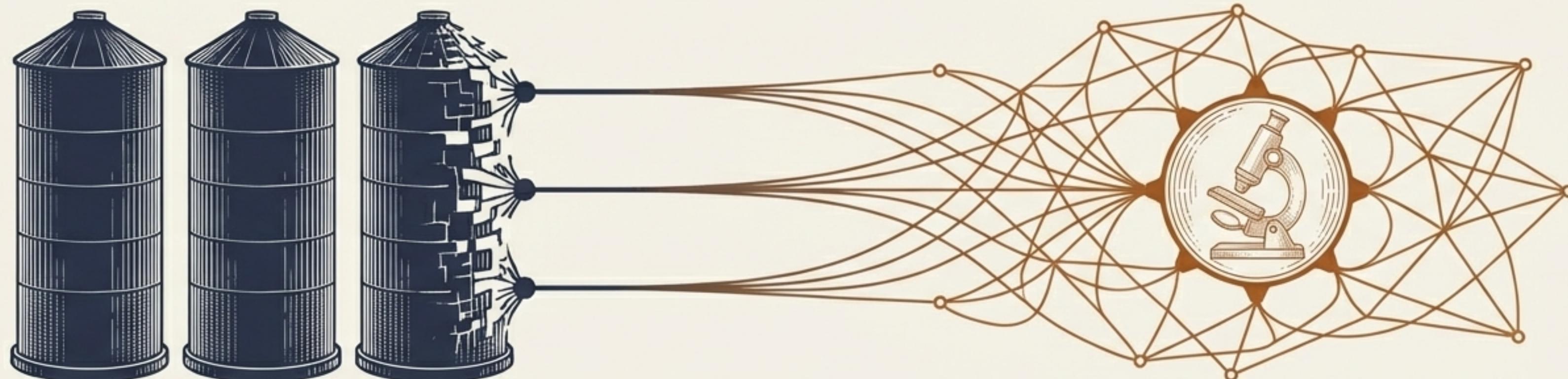
Werdykt autorów jest jednoznaczny:  
model jest "**przedwczesny do wdrożenia w produktach komercyjnych ze względu na ryzyka związane z bezpieczeństwem**".

Question: If  $x$  is 12 and  $y$  is 9, what is  $x - y$ ?

Answer: -3

# Szersza perspektywa: Od rywalizacji do współpracy

- Publikacja OPT to coś więcej niż tylko model – to **manifest na rzecz nowego sposobu prowadzenia badań nad AI**.
- Przesuwa punkt ciężkości z budowania największych, najbardziej zamkniętych systemów na tworzenie wspólnych narzędzi badawczych.
- OPT staje się "potężnym mikroskopem", który pozwala całej społeczności naukowej badać wielkie modele językowe, rozumieć ich ryzyka i wspólnie je ulepszać.
- To zmiana akcentu z wyścigu na **współpracę**, której celem jest postęp całej dziedziny.



# Pytanie do Ciebie

Czy możemy stworzyć model AI, który w pełni rozumie ludzki język – z całym jego bogactwem i **mrocznymi zakamarkami** – nie dziedzicząc jednocześnie **najgorszych cech ludzkości**?

A może jedyna droga do bezpiecznej AI prowadzi przez **sterylne modele**, trenowane na wyselekcjonowanych danych, co pozostawi je w **oderwaniu od rzeczywistości**, w której muszą funkcjonować?