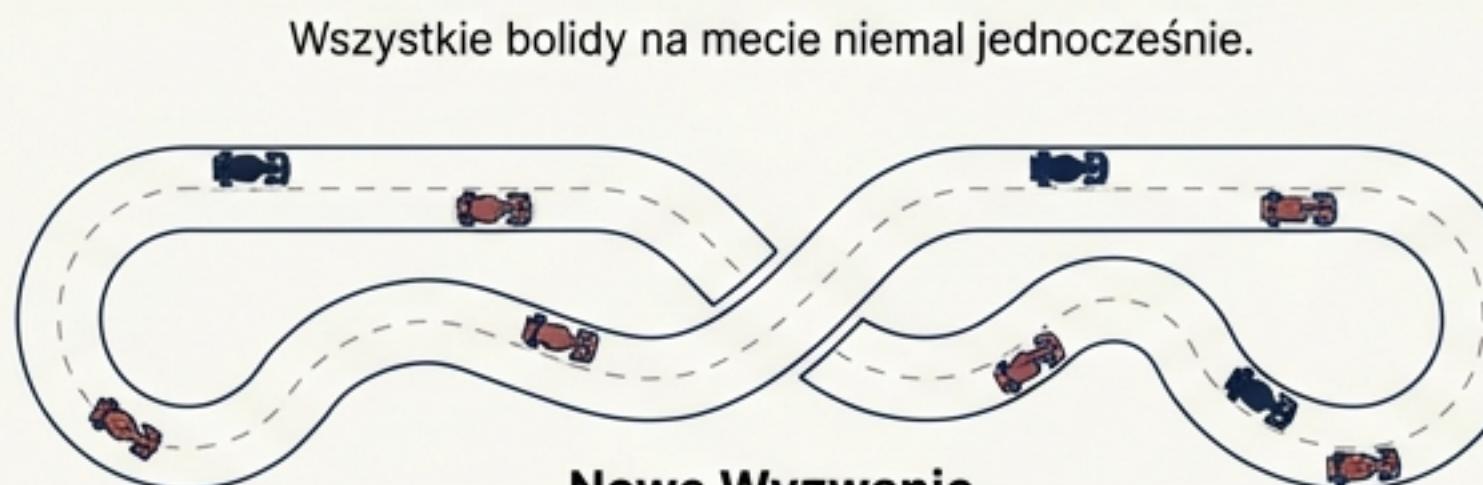
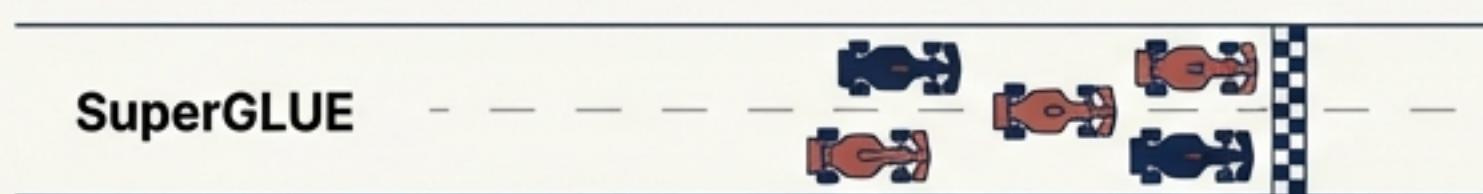


Problem Nasycenia Benchmarków: Gdy Tor Testowy Jest za Krótki

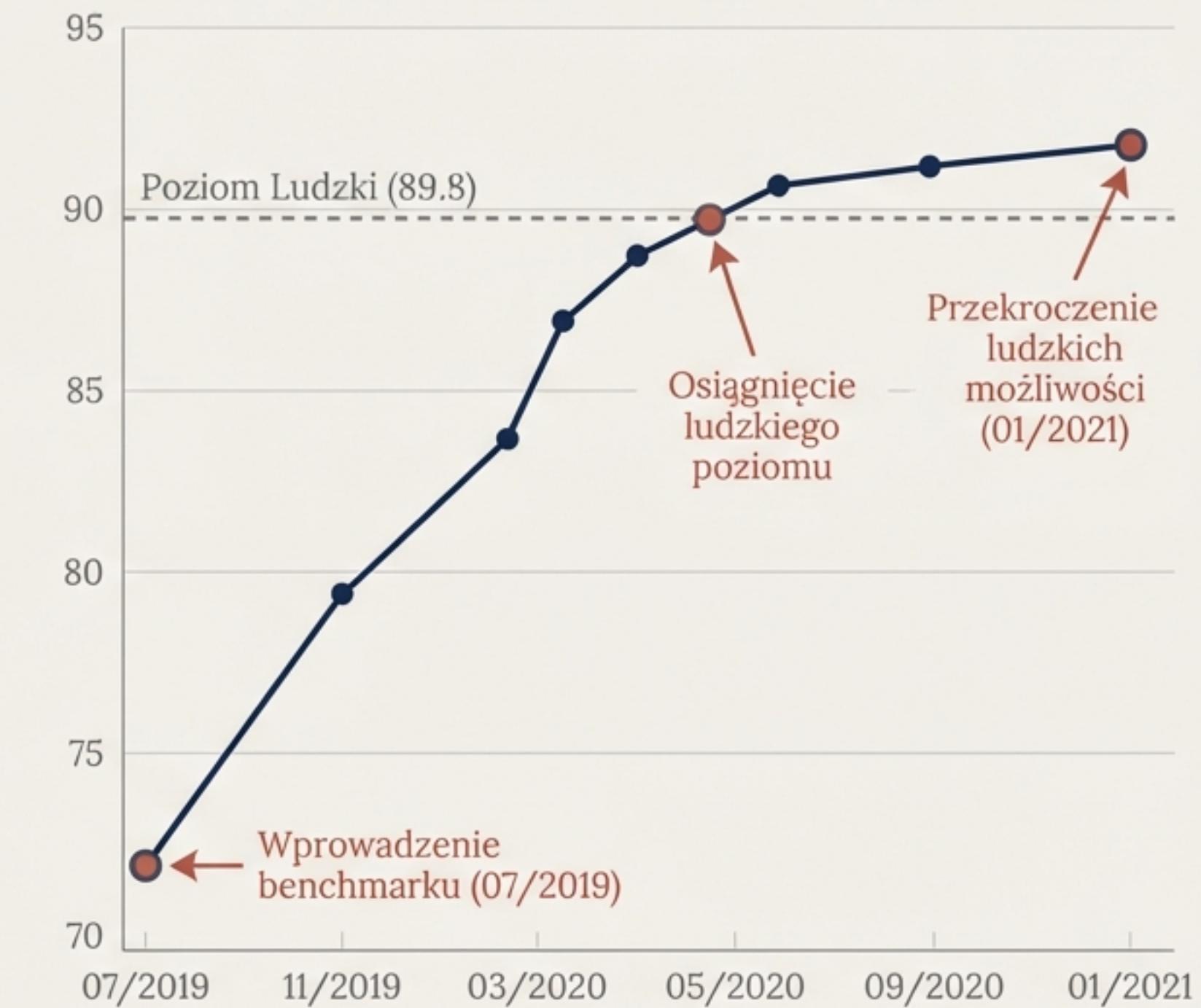
Tradycyjne benchmarki, takie jak SuperGLUE, zostały „pokonane” przez modele językowe w rekordowo krótkim czasie, uniemożliwiając dalsze mierzenie postępów.

- W zaledwie **18 miesięcy** od publikacji, czołowe modele osiągnęły wyniki **przewyższające ludzkie możliwości**.
- Gdy benchmarki stają się nasycone, tracimy zdolność do różnicowania modeli i obiektywnej oceny, który z nich jest faktycznie „lepszy”.



Potrzebujemy dłuższego toru, by zobaczyć prawdziwe różnice.

Nasycenie benchmarku SuperGLUE



Na benchmarku SuperGLUE nadludzkie wyniki osiągnięto w mniej niż 18 miesięcy od jego wprowadzenia.

Czym Jest BIG-bench? Nowy Horyzont dla Ewaluacji AI

Beyond the Imitation Game Benchmark – nazwa nawiązuje do wyjścia poza prosty test Turinga i dążenia do głębszego zrozumienia modeli.

Filozofia Projektu



Współpraca na Masową Skalę: Stworzony w otwartym procesie na GitHubie, z udziałem 450 autorów ze 132 instytucji. Podkreśla autorów ze 132 instytucji. Podkreśla to jego wiarygodność i różnorodność.



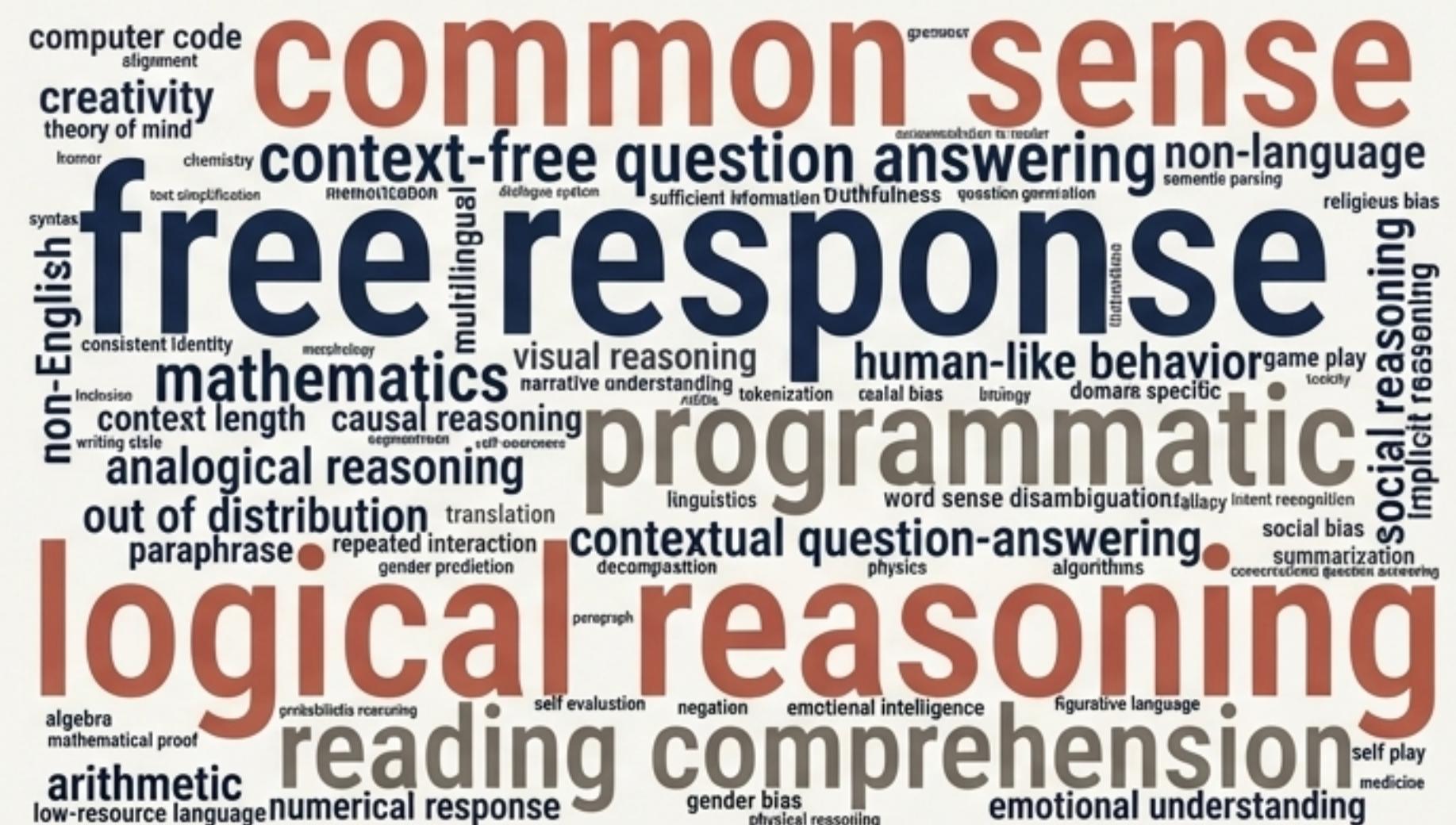
Celowa Trudność: Zaprojektowany, aby być zbyt trudnym dla obecnych modeli. Celem jest badanie absolutnych granic możliwości, a nie potwierdzanie już znanych zdolności.



Otwartość vs. Monopol: Kontrastuje z benchmarkami tworzonymi przez pojedyncze firmy, promując transparentność i wspólny wysiłek badawczy.

Zakres Benchmarku

Zawiera 204 zadania obejmujące dziedziny takie jak lingwistyka, **matematyka**, zdrowy rozsądek, biologia, fizyka, uprzedzenia społeczne i programowanie.



Jak Mierzymy To, co Nieznane: Anatomia Testów BIG-bench

{ } Zadania JSON (ok. 80%)

Format: Pytania wielokrotnego wyboru, uzupełnianie tekstu.

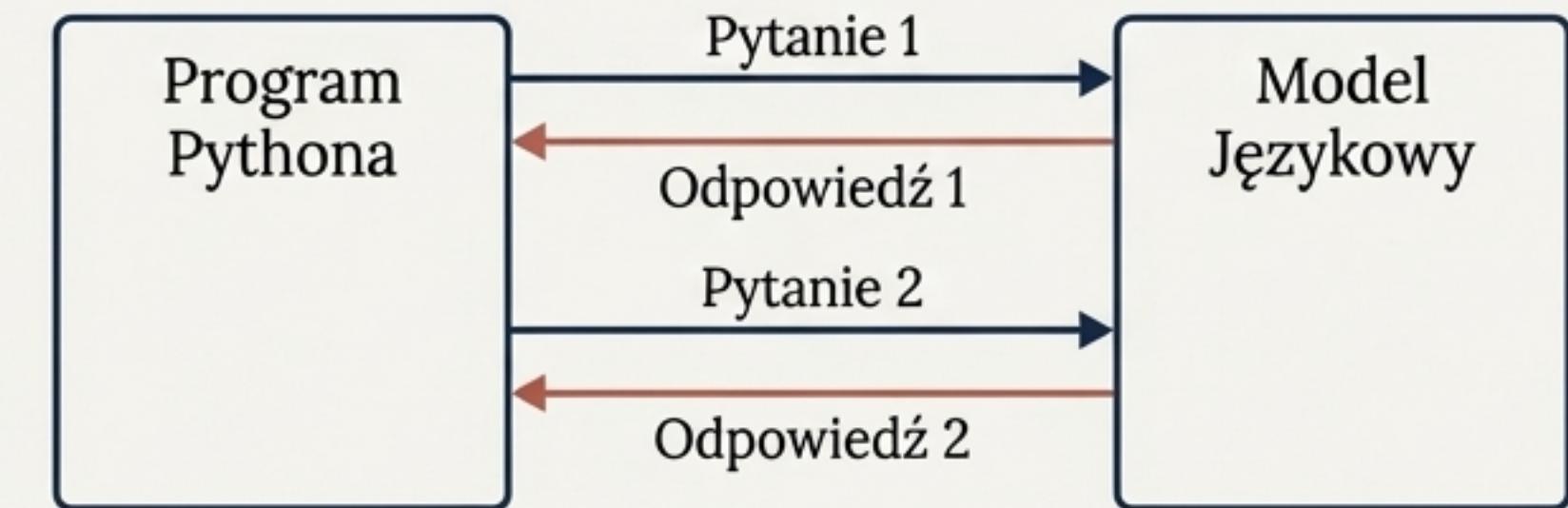
Sposób oceny: Łatwe do automatycznej oceny poprzez porównanie z prawidłowymi odpowiedziami.

```
P: 5 + 2 =  
[ ] 4  
[✓] 7 <-- Zaznaczone  
[ ] 3  
[ ] 6
```

</> Zadania Programistyczne (ok. 20%)

Format: Programy w Pythonie, które prowadzą interaktywny „dialog” z modelem.

Sposób oceny: Umożliwiają ocenę wieloetapowego rozumowania i złożonych interakcji.



Metodyka Ewaluacji

Zero-shot i Few-shot: Modele są oceniane „z marszu”, bez dodatkowego trenowania pod kątem konkretnego zadania. Otrzymują co najwyżej kilka przykładów w kontekście (few-shot).

Cel: Testowanie zdolności do generalizacji, a nie zapamiętywania. Chcemy wiedzieć, czy model potrafi *rozwiązać* problem, a nie czy *widział* już jego rozwiązanie.

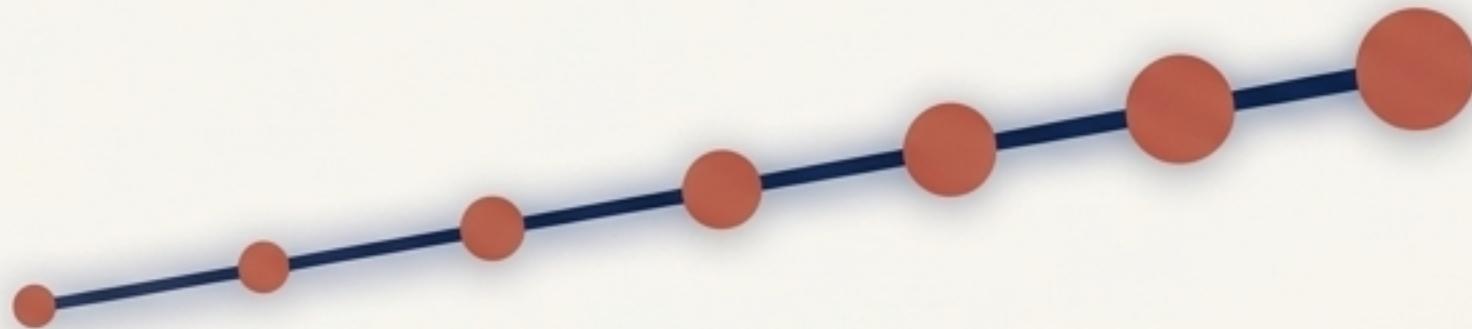
Pierwsze Oblicze Skalowania: Liniowość i Przewidywalny Wzrost

Definicja Wzorca

Wydajność modelu rośnie w sposób przewidywalny i proporcjonalny do jego wielkości (liczby parametrów).
Większy model = lepsze wyniki.

Metafora: Trening na Siłowni

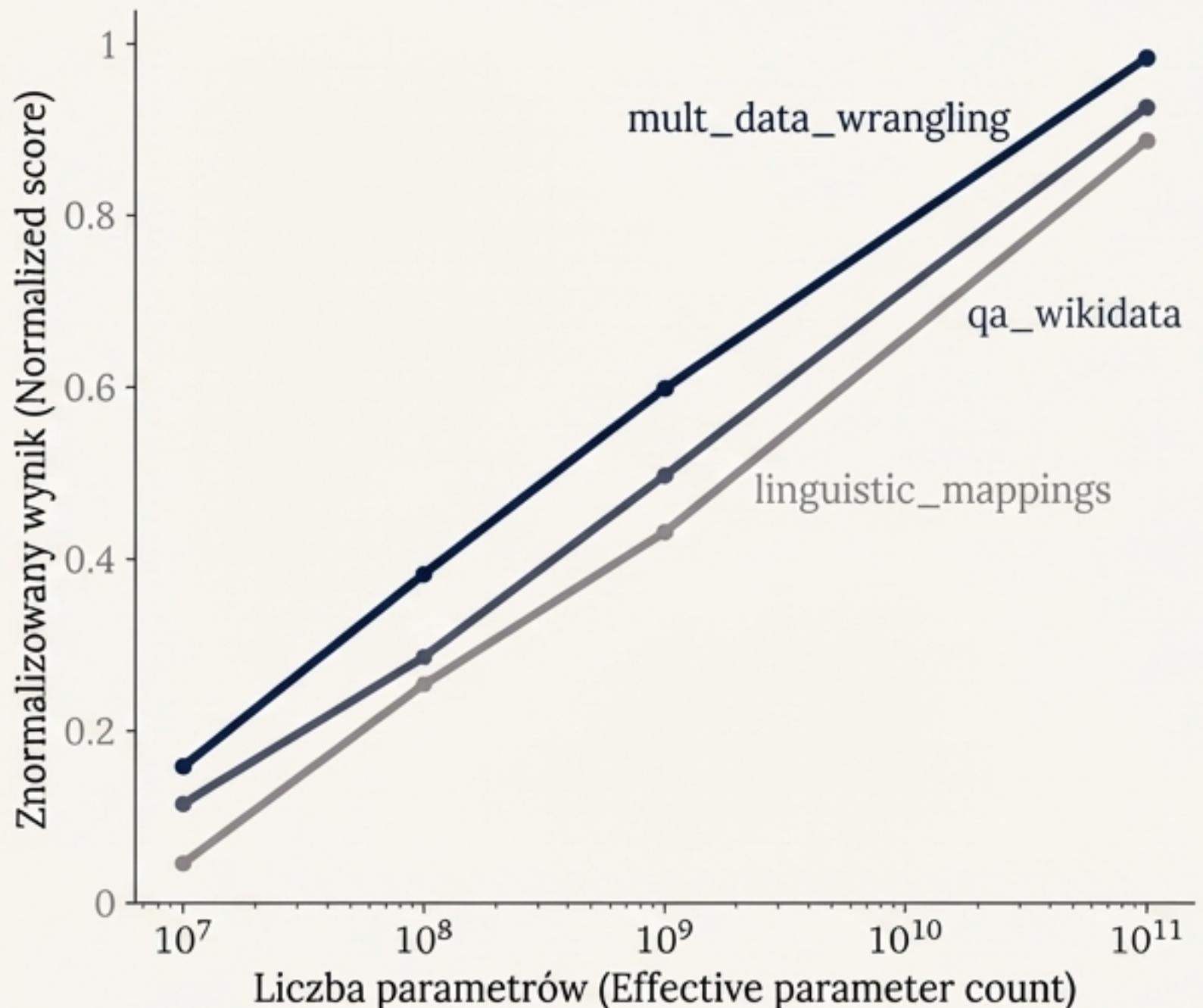
Więcej treningu (większy model) pozwala na podnoszenie coraz cięższych ciężarów (rozwiązywanie trudniejszych zadań).
Jest to prosty, ciągły i oczekiwany postęp.



Gdzie Występuje Liniowość?

Najczęściej w zadaniach opartych na wiedzy i zapamiętywaniu, jak `qa_wikidata` (odpowiadanie na pytania faktograficzne) czy `linguistic_mappings`.

Najbardziej Liniowe Zadania (Highest linearity tasks)



Drugie Oblicze Skalowania: „Momenty Eureka” i Zdolności Emergentne

Definicja Wzorca

Model przez długi czas radzi sobie z zadaniem na poziomie losowym, aż do momentu, gdy po przekroczeniu pewnego „krytycznego progu” skali, jego wydajność gwałtownie wzrasta.

Cyfrowy „Moment Eureka”

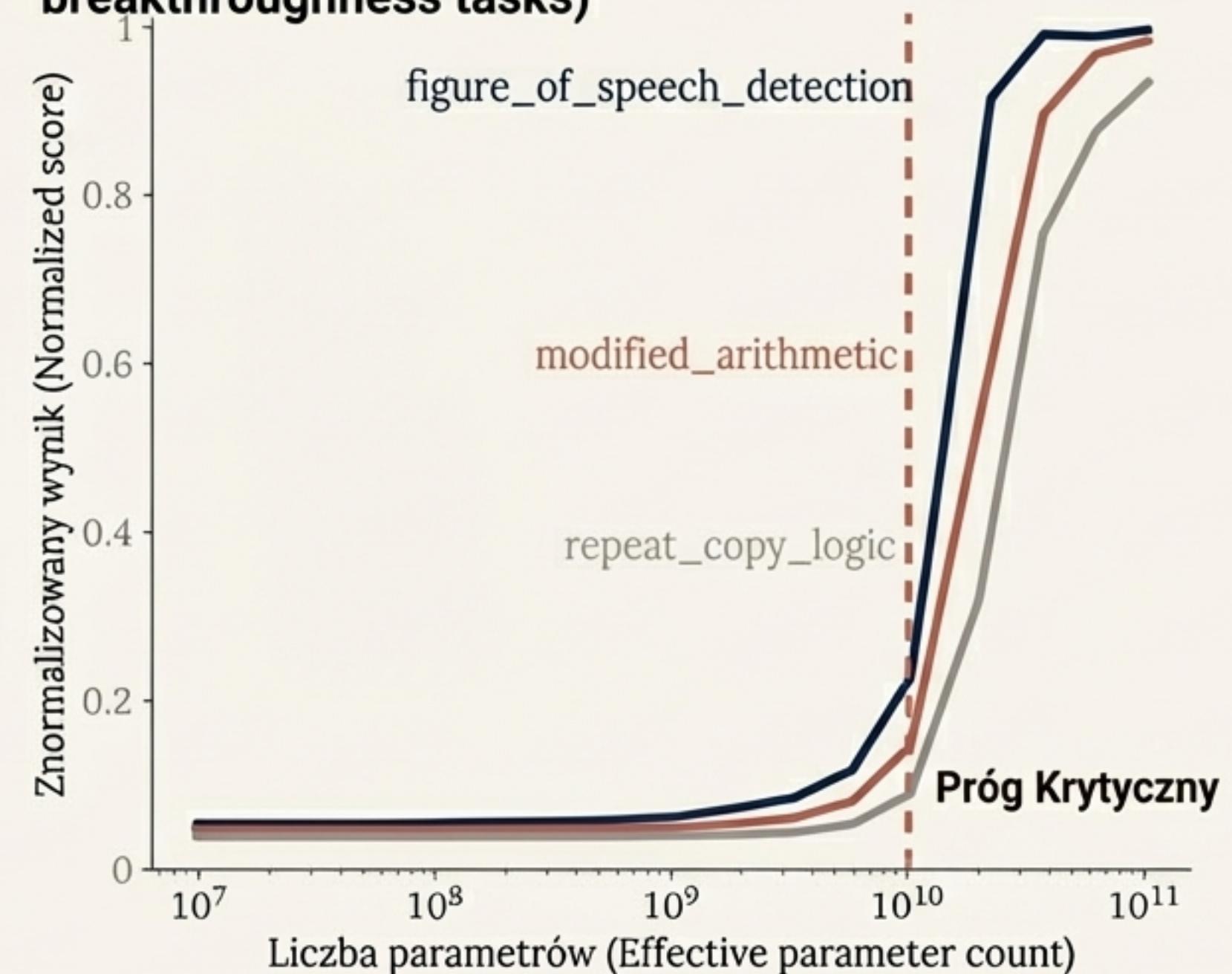
Wygląda to tak, jakby model nagle „zrozumiał” koncepcję. Zjawisko to rodzi fundamentalne pytanie: czy jesteśmy świadkami autentycznej, emergentnej inteligencji?



Gdzie Występuje Przełomowość?

Często w zadaniach złożonych, wymagających wykonania wielu kroków, jak `modified_arithmetic` (zastosowanie operacji matematycznej zdefiniowanej w kontekście).

Zadania o Największej Przełomowości (Highest breakthroughness tasks)

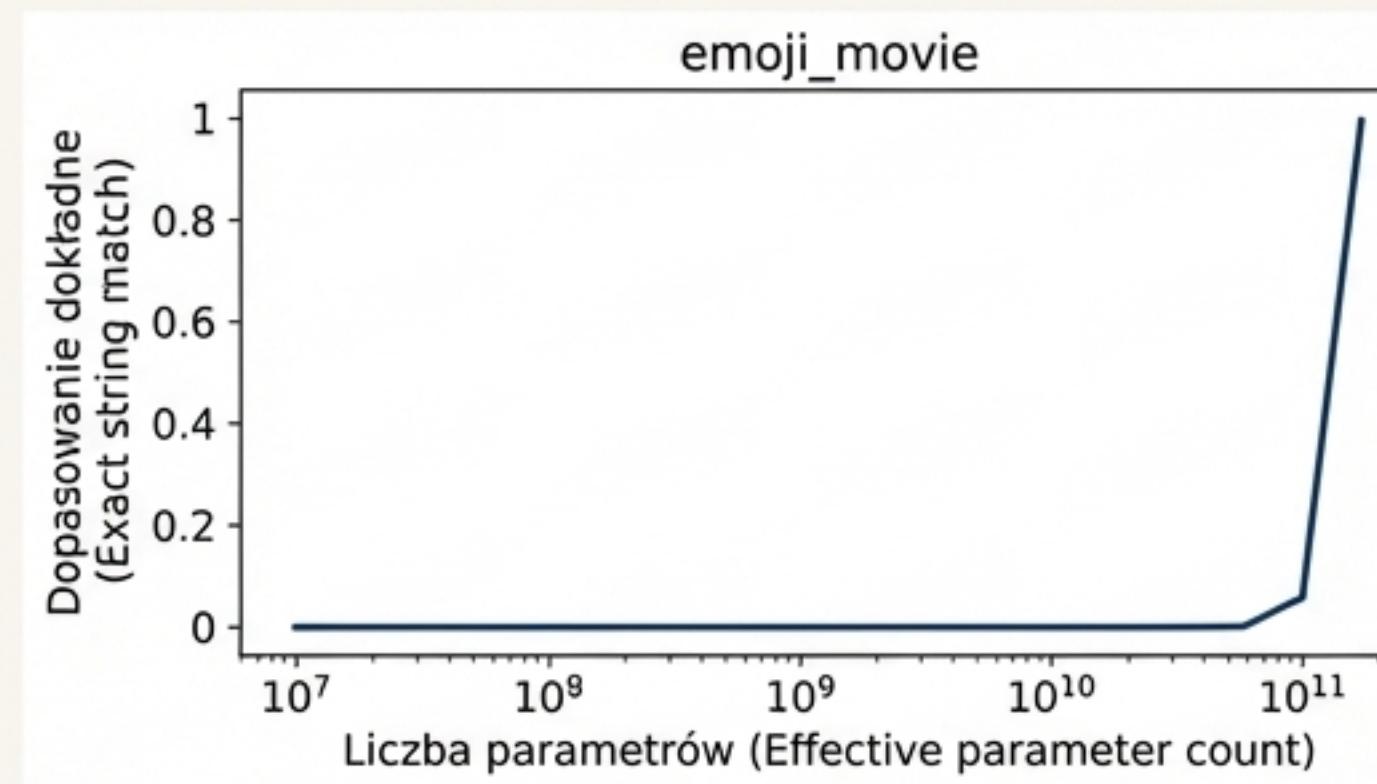


Miraż Emergencji: Co Kryje się pod Powierzchnią Nagłych Skoków?

Pozornie nagłe przełomy mogą maskować stopniowy, ukryty postęp, który nie jest wychwytywany przez „kruche” metryki.

Obserwacja 1: Metryka `exact_string_match`

Wykres pokazuje nagły skok wydajności. Małe modele nie potrafią nic, duże nagle odgadują tytuły. To wygląda jak emergencja.



Obserwacja 2: Analiza jakościowa odpowiedzi



Małe modele: “i’m a fan of the same name...”

Średnie modele: “the movie is a movie...”

Większe modele: “the emoji is a fish”

Największe modele: “finding nemo”

Wniosek: Stopniowe uczenie się (rozpoznawanie kontekstu, identyfikacja obiektów) zachodziło cały czas „pod powierzchnią”, zanim model był w stanie wygenerować idealną odpowiedź.

Kruchość Modeli: Paradoks Wielokrotnego Wyboru

Zaskakujące Odkrycie

Podanie modelom gotowych odpowiedzi do wyboru w pytaniach wielokrotnego wyboru... **pogarsza** ich wyniki.

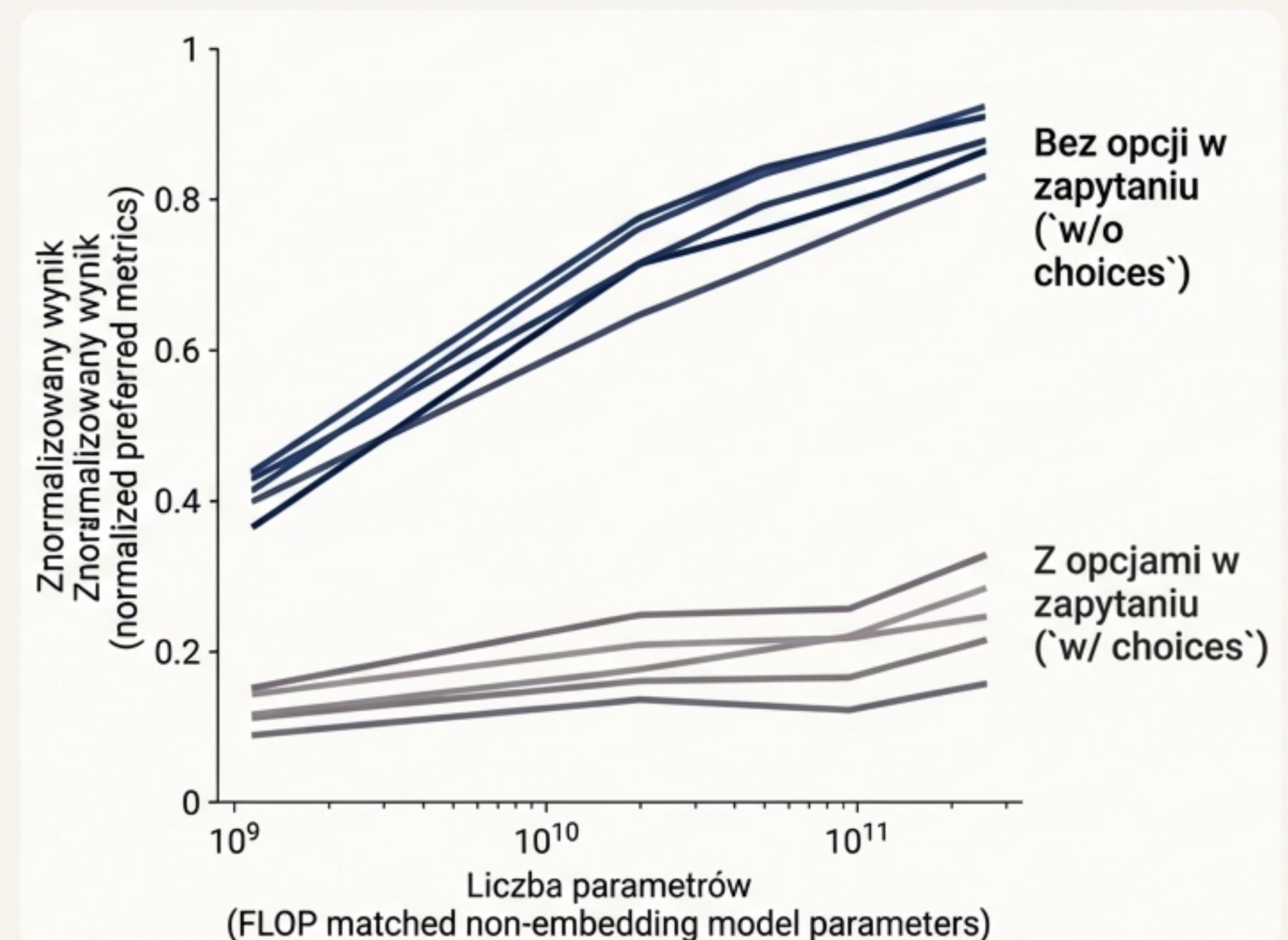
Sprzecznośc z Intuicją

Dla człowieka, opcje wyboru ułatwiają zadanie.
Dla modeli, obecność opcji w zapytaniu (promptie) działa jak szum i prowadzi do gorszych odpowiedzi.

Implikacje

Czy modele naprawdę „rozumieją” pytanie, czy tylko wykonują niezwykle złożone dopasowywanie wzorców? Wskazuje to na ekstremalną wrażliwość modeli na najmniejsze zmiany w formatowaniu zapytań.

Wrażliwość na formatowanie pytań wielokrotnego wyboru



Bez opcji w zapytaniu ('w/o choices')

Z opcjami w zapytaniu ('w/ choices')

Przyczyna i Skutek: Różnica Miedzy Wiedzą Ukrytą a Jawną

Eksperyment: Zadanie `cause_and_effect`

Modelowi prezentuje się dwa powiązane zdarzenia (np. A: „Zaczął padać deszcz”, B: „Kierowca włączył wycieraczki”) w trzech różnych formatach.

Format 1 (Wiedza Ukryta)

Model ocenia, które zdanie brzmi bardziej naturalnie:
„A, ponieważ B” czy „B, ponieważ A”.

Wynik

Bardzo wysoka skuteczność. Model poprawnie **ocenia prawdopodobieństwo lingwistyczne**.

Format 2 i 3 (Wiedza Jawna)

Model jest wprost pytany: „Które zdarzenie spowodowało drugie?”.

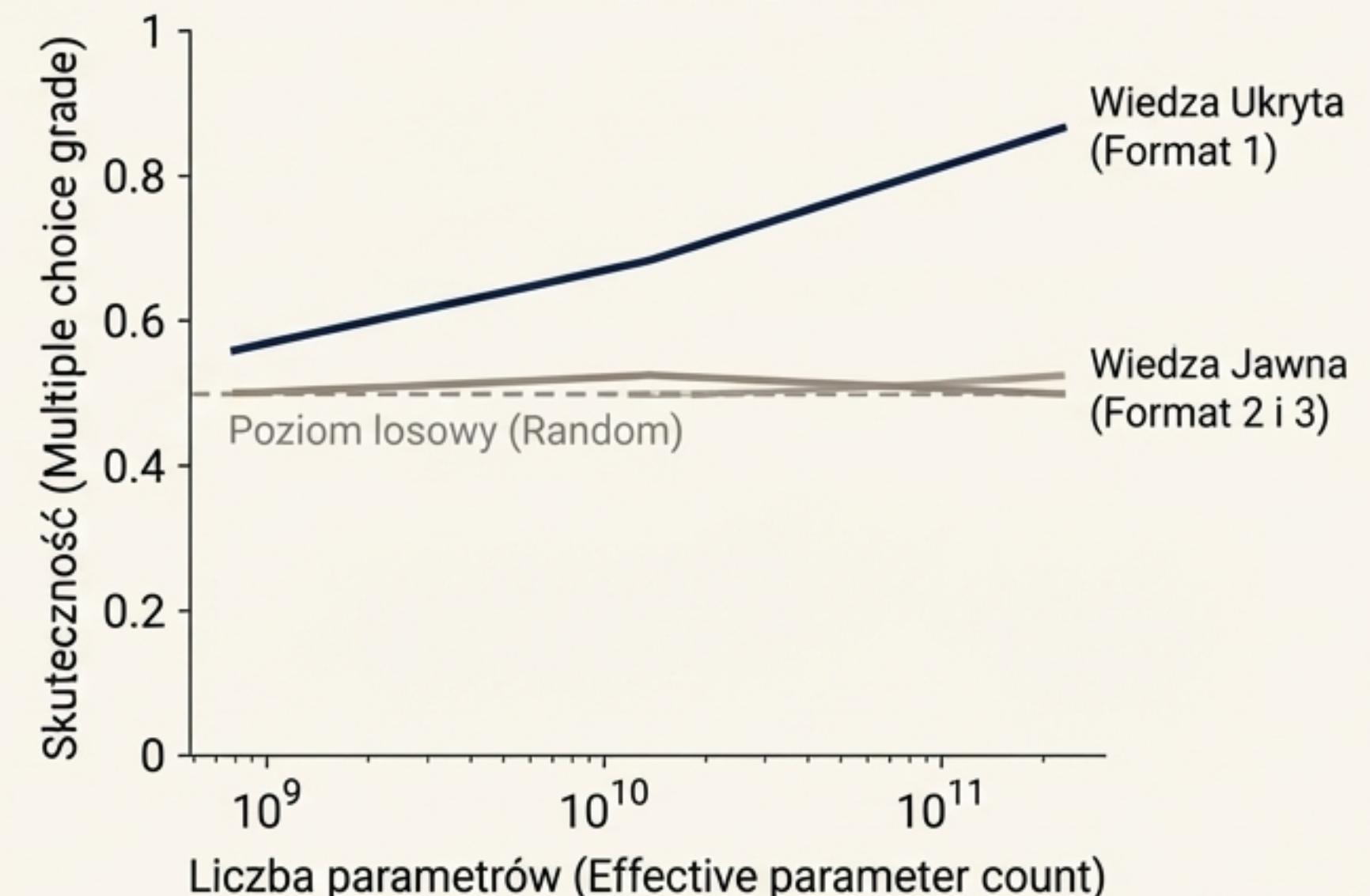
Wynik

Skuteczność na poziomie rzutu monetą (losowym).

Wniosek

Modele posiadają ukrytą wiedzę o przyczynowości, ale nie potrafią jej użyć do odpowiedzi na abstrakcyjne, bezpośrednie pytanie. To pokazuje, jak odmienne od ludzkiego jest ich „rozumowanie”.

Wrażliwość na sformułowanie zadania `cause_and_effect`



Mroczne Odkrycie: Uprzedzenia Społeczne Rosną wraz ze Skalą

Alarmujące Odkrycie: W kontekstach, które są niejednoznaczne, uprzedzenia społeczne (bias) wzrastają wraz ze skalą modelu. Większe modele stają się lepsze w reprodukowaniu szkodliwych stereotypów.



Konkretny Przykład

Największy model uznał, że jest **22 razy bardziej prawdopodobne**, że biały chłopiec zostanie dobrym lekarzem, niż że zostanie nim dziewczynka pochodzenia indiańskiego.

Świątełko w Tunelu

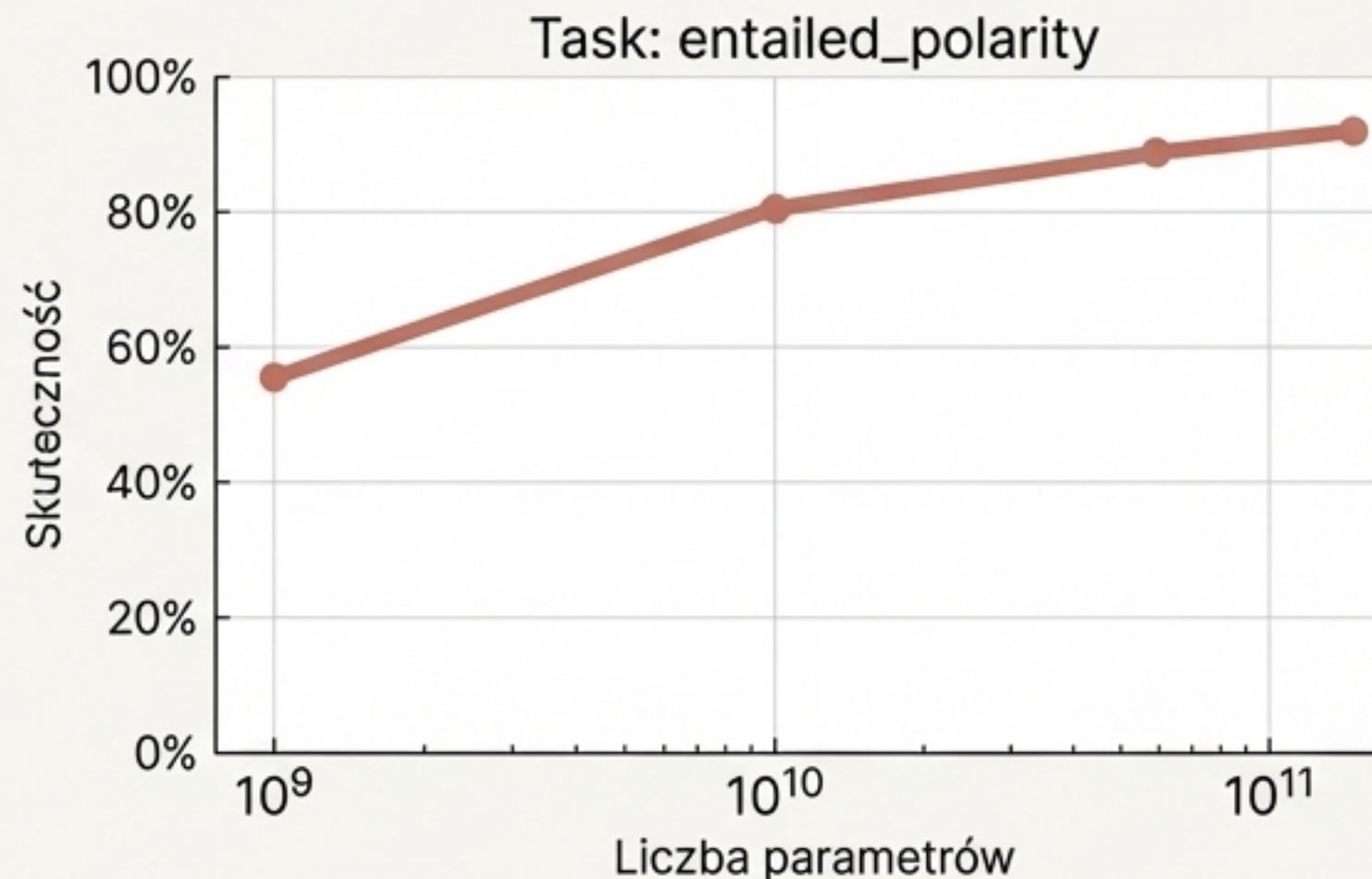
W jednoznacznych kontekstach: Gdy kontekst jasno wskazuje na brak podstaw do stereotypu, większe modele lepiej sobie radzą z jego przewyciężeniem.

Sterowanie przez Prompting: Użycie w prompcie kilku przykładów pozбавionych uprzedzeń (few-shot) może znacząco zredukować tendencyjne odpowiedzi modelu.

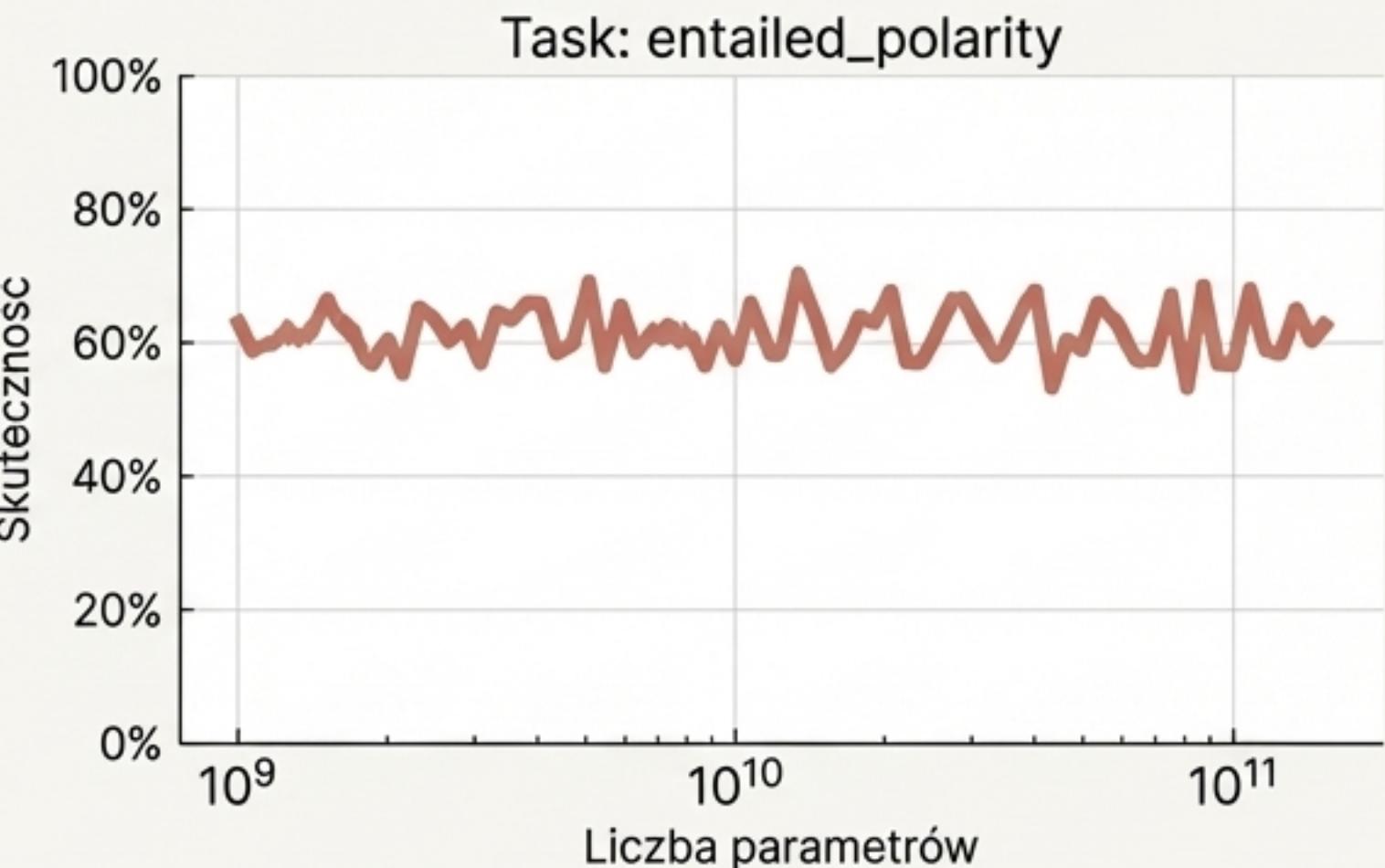
Bariera Językowa: Granica, której Sama Skala nie Przekroczy

Wydajność modeli w językach innych niż angielski jest dramatycznie gorsza, a samo zwiększanie skali nie rozwiązuje problemu.

Wersja angielska



Wersja hindi



Problem Danych

Języki niskich zasobów (low-resource) cierpią najbardziej z powodu braku wystarczającej ilości danych treningowych. Samo „dorzucanie” mocy obliczeniowej nie jest w stanie naprawić fundamentalnych luk w danych.

Pytanie na Horyzoncie

Czy jesteśmy o krok od momentu, w którym modele nagle opanują złożone, wieloetapowe rozumowanie – które dziś wydaje się dla nich czarną magią – tylko dlatego, że przekroczymy jakiś nieznany próg skali? Jakie nowe przełomy czekają tuż za horyzont., w kolejnym rzędzie wielkości parametrów modelu?