

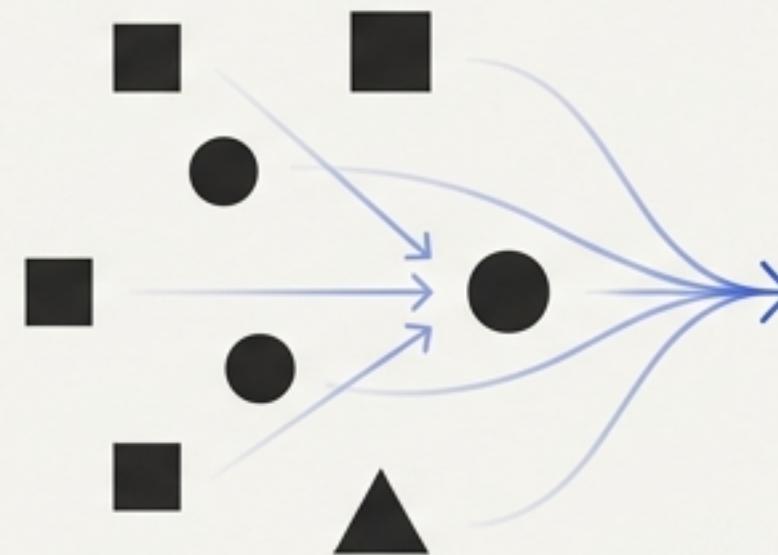
# Zmieniając Paradygmat: Narodziny GPT-2

Analiza przełomowej pracy OpenAI "Language Models Are Unsupervised Multitask Learners" (2019)

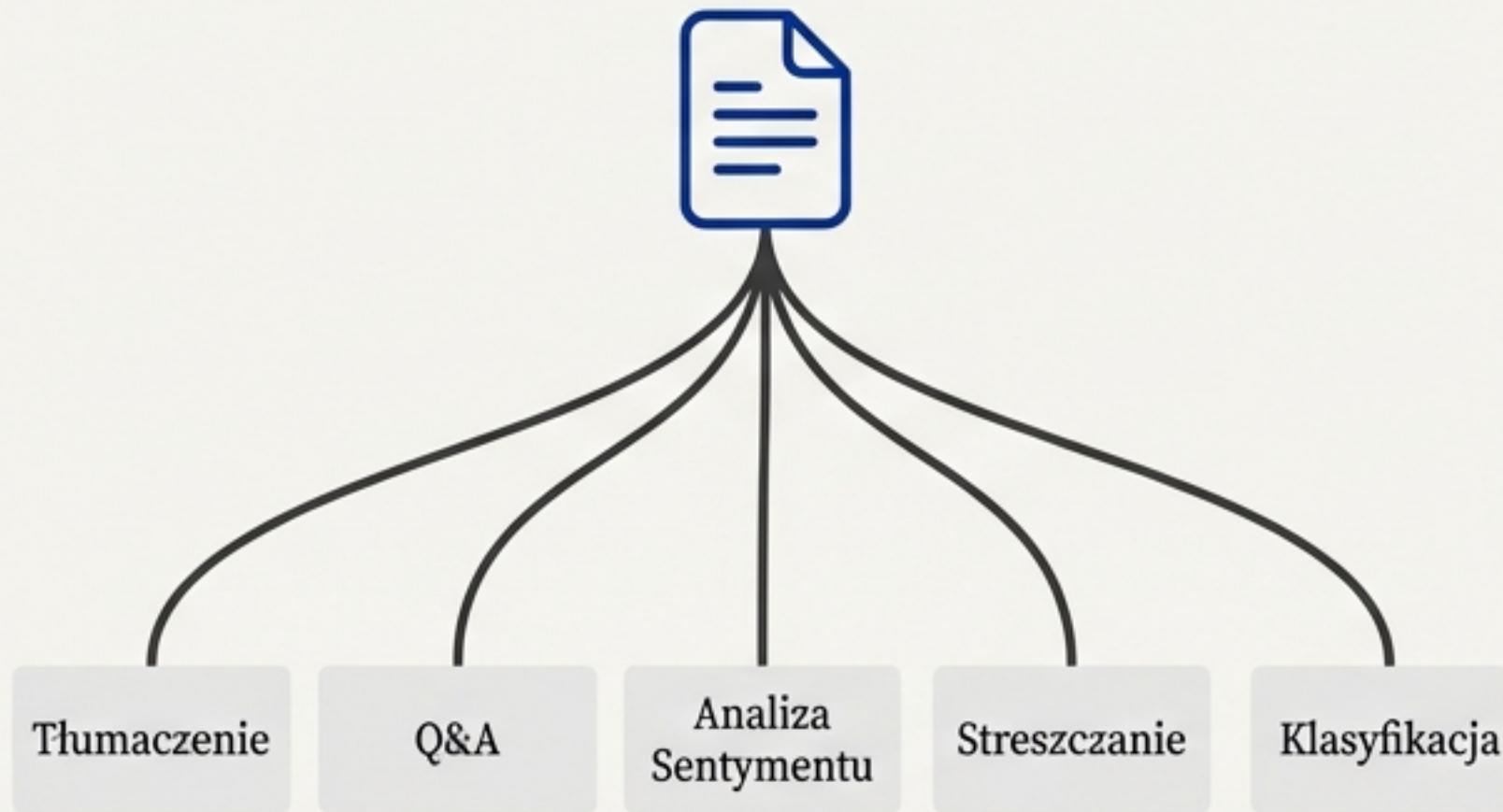
**Rewolucyjna Zmiana:** Przejście od wyspecjalizowanych, 'wąskich ekspertów' do uniwersalnych modeli językowych ogólnego przeznaczenia.

**Główna Hipoteza:** Proste zadanie przewidywania następnego słowa w tekście, przy odpowiedniej skali, prowadzi do wyłonienia się (emergencji) złożonych zdolności wielozadaniowych bez jawnego nadzoru.

**Fundamentalne Pytanie:** Czy system trenowany wyłącznie do przewidywania następnego słowa może nauczyć się zadań takich jak tłumaczenie, odpowiadanie na pytania i streszczanie?



# Paradygmat sprzed 2019: Armia Wąskich Specjalistów



## Dominujące Podejście

- Oddzielny, starannie trenowany model dla każdego zadania (np. tłumaczenie, Q&A, analiza sentymentu).
- Każde zadanie wymagało dedykowanych, etykietowanych zbiorów danych.
- Modele były "kruche i wrażliwe na niewielkie zmiany w dystrybucji danych".

## Fundamentalne Ograniczenia

- **Brak Generalizacji:** Systemy doskonale radziły sobie w wąskich dziedzinach, ale zawodziły w nowych kontekstach.
- **Wysokie Koszty:** Tworzenie i etykietowanie zbiorów danych dla każdego nowego zadania było niezwykle kosztowne i czasochłonne.
- **Dopasowywanie Wzorców:** Systemy uczyły się imitować zachowania ze zbioru treningowego, a nie faktycznie rozumować.

# Nowa Filozofia: Jeden Model Uczęcy się Wszystkiego

Główna Idea: Nienadzorowane uczenie wielozadaniowe.

## Pojedynczy, Uniwersalny Cel Treningowy:

Przewidzieć następny token w sekwencji tekstu. Model uczy się, maksymalizując prawdopodobieństwo  $p(\text{kolejny\_token} | \text{poprzednie\_tokeny})$ .

## Hipoteza Robocza:

- Aby doskonale przewidywać następne słowo w zróżnicowanym, ogromnym korpusie tekstu, model musi nauczyć się leżących u podstaw umiejętności, takich jak logika, wiedza o świecie i gramatyka.
- Zdolności takie jak tłumaczenie, streszczanie i Q&A pojawiają się jako "efekty uboczne" optymalizacji jednego celu – są to właściwości emergentne.

## Kluczowa Koncepcja: Transfer Zadań Zero-Shot:

Zdolność do wykonywania zadań, których model nigdy explicitie nie widział podczas treningu, bez jakichkolwiek modyfikacji parametrów czy architektury.



# Paliwo dla Rewolucji: Zbiór Danych WebText

**40 GB** czystego tekstu (odpowiednik ~8 milionów dokumentów).

## Strategia Pozyskiwania Danych (Jakość ponad Ilością)



- **Źródło:** Wyłącznie linki zewnętrzne z serwisu Reddit, które otrzymały co najmniej 3 punkty 'karmy'.
- **Uzasadnienie:** Użycie 'ludzkiej kuracji' jako heurystyki. Treści ocenione pozytywnie przez społeczność są statystycznie wyższej jakości – lepiej napisane, bardziej edukacyjne i wiarygodne.

## Kluczowa Decyzja Metodologiczna

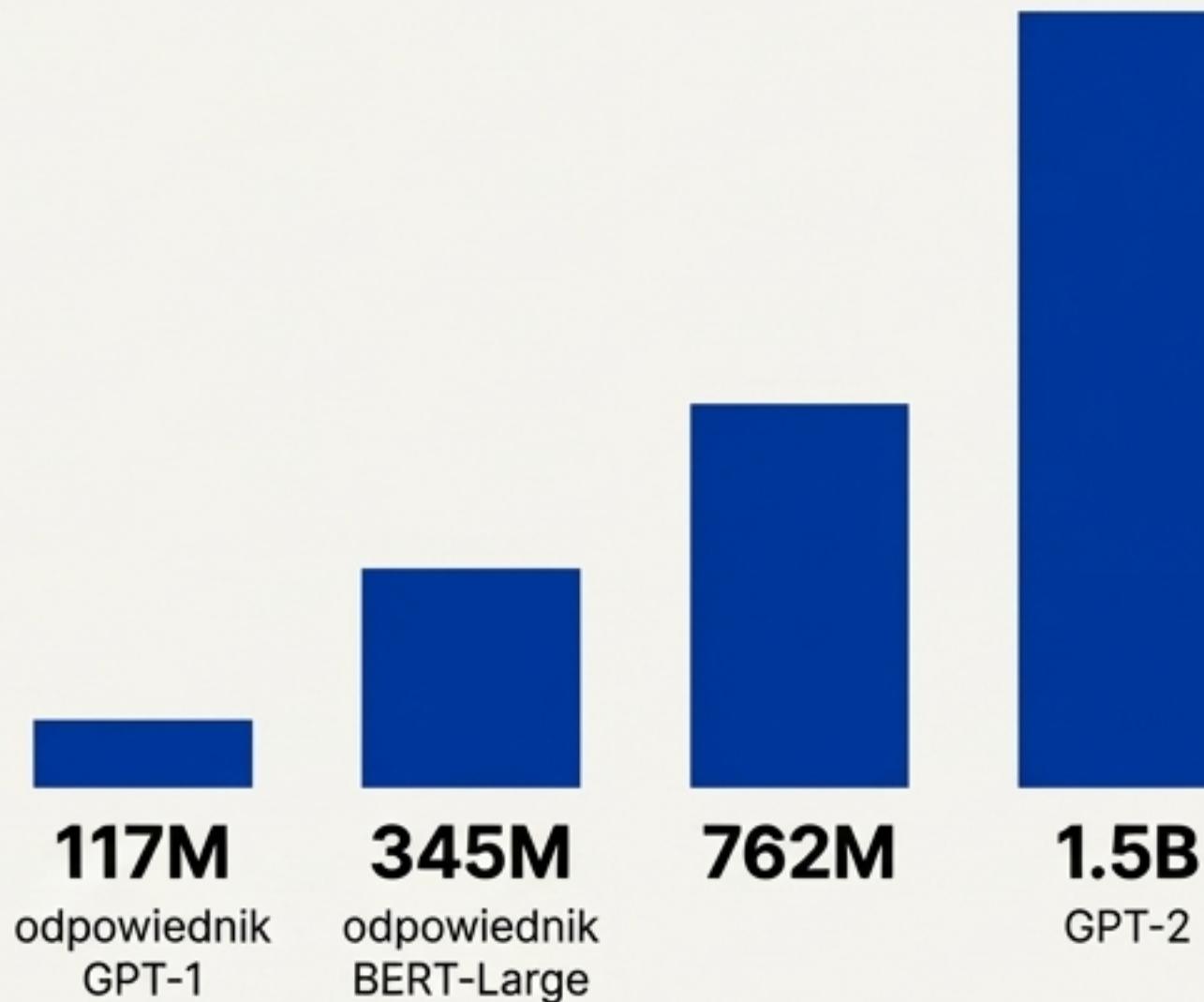
### Celowe Wykluczenie Wikipedii



**Powód:** Wiele standardowych benchmarków (testów) opiera się na artykułach z Wikipedii. Jej wykluczenie zapobiega "przeciekom" danych i gwarantuje, że model nie może po prostu zapamiętać odpowiedzi do zadań testowych.

# Architektura i Skala: Więcej Znaczy Inaczej

- Fundament: Architektura Transformer z mechanizmem uwagi własnej (self-attention), w dużej mierze bazująca na poprzednim modelu GPT.
- Kluczowa Innowacja: Bezprecedensowa **skala**, a nie nowa architektura.



## Reprezentacja Danych

**Tokenizacja:** Byte Pair Encoding (BPE), działająca na poziomie bajtów, nie znaków Unicode.

**Zaleta:** Podstawowy słownik to tylko 256 tokenów. Podejście to łączy zalety modelowania na poziomie słów (efektywność) i znaków (uniwersalność), eliminując problem słów spoza słownika (*out-of-vocabulary*).



# Dowód #1: Mistrzostwo w Modelowaniu Języka

**Najlepsze wyniki** (State-of-the-Art) na **7 z 8** analizowanych zbiorów danych.

Wszystkie wyniki osiągnięto w trybie **zero-shot** – bez żadnego dostrajania (*fine-tuning*) pod konkretny zbiór danych. Model działał 'prosto z pudełka'.

## Wybrane Wyniki (na podstawie Tabeli 3)

LAMBADA (Perplexity)

99.8  
**8.63** Poprawa SOTA

LAMBADA (Accuracy)

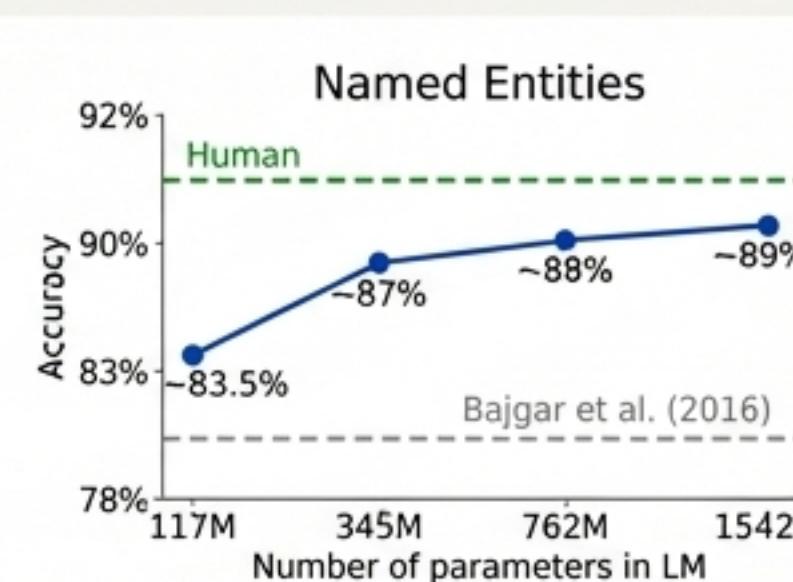
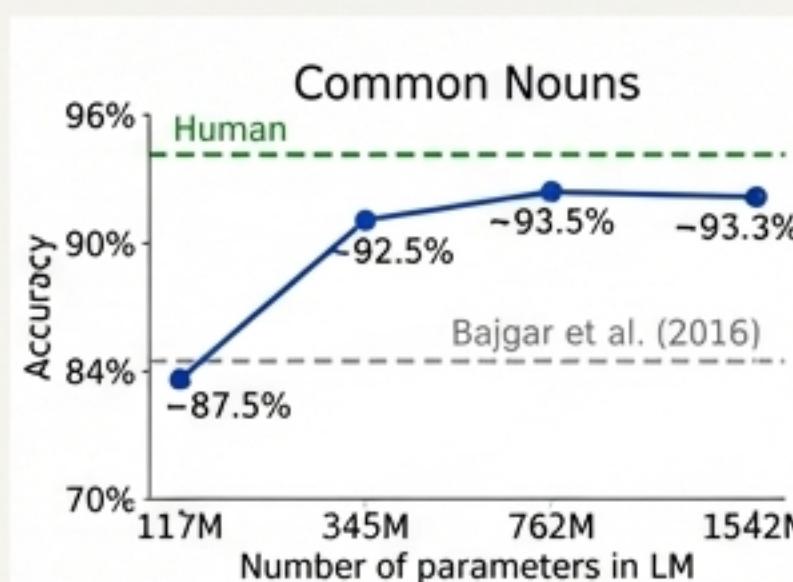
19%  
**63.24%**

Children's Book Test  
(Common Nouns) Children's BookTest  
(Named Entities)

SOTA 93.3% SOTA 89.05%

### Wniosek:

Skala modelu połączona z jakościowymi danymi tworzy niezwykle potężny model językowy, co stanowi fundament dla wszystkich emergentnych zdolności.



## Dowód #2: Niespodziewane Odkrycie – Tłumaczenie Zero-Shot

Model, trenowany na korpusie, z którego celowo usunięto treści inne niż angielskie, potrafił tłumaczyć z francuskiego na angielski.



0.025%

A thick blue arrow pointing from left to right, indicating a transformation or flow from French to English.

części zbioru danych w języku francuskim (10MB z 40GB).



### Wyniki na Benchmarku WMT-14:

- Angielski → Francuski: 5 BLEU (jakość bardzo niska)
- Francuski → Angielski: **11.5 BLEU**

### Znaczenie Naukowe:

- Wynik 11.5 BLEU jest znacznie lepszy od kilku bazowych modeli nienadzorowanego tłumaczenia maszynowego z tamtego okresu.
- Osiągnięcie to było historycznie możliwe tylko dzięki treningowi na milionach par równoległych zdań.

Zdolność do mapowania między językami może być fundamentalną, emergentną częścią 'rozumienia' języka, a nie oddzielną, wyuczoną umiejętnością.

# Dowód #3: Rozumienie Tekstu i Kalibrowana Wiedza

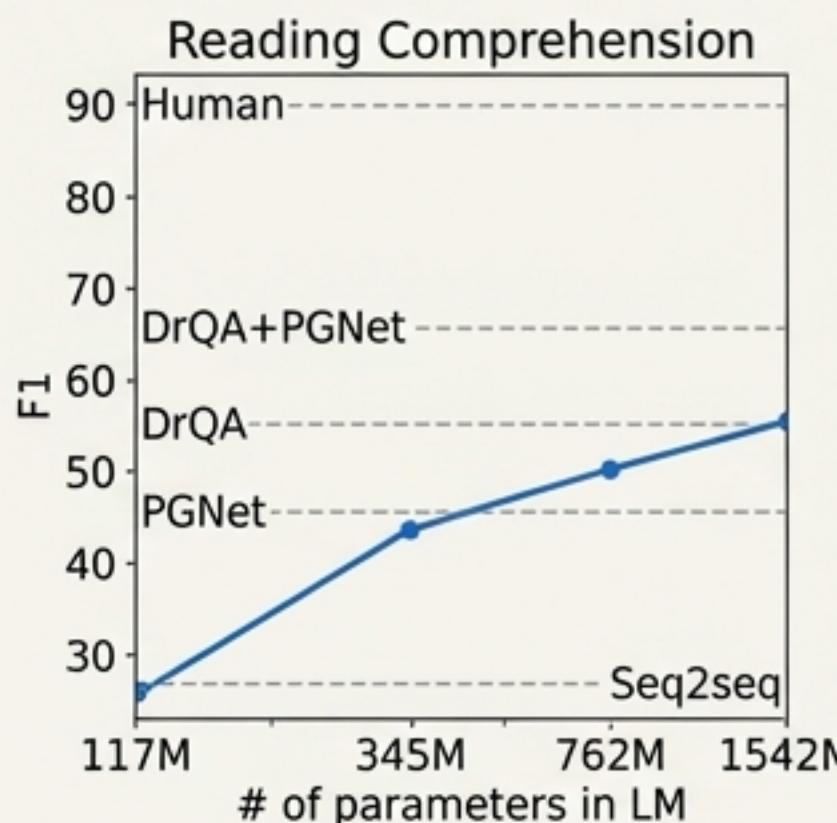
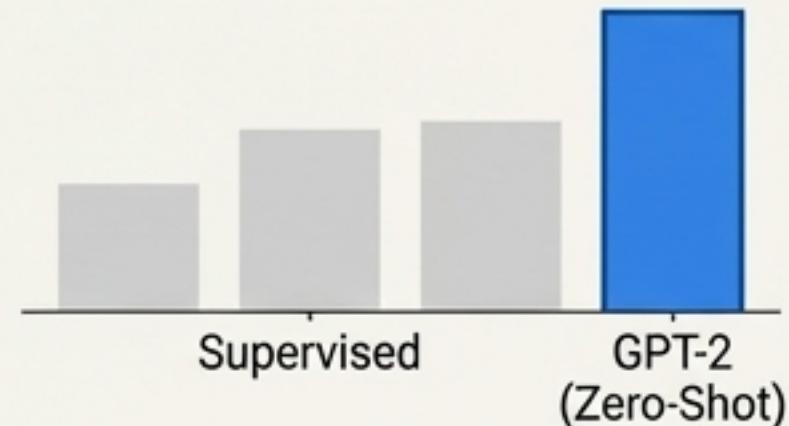
## Zadanie: Konwersacyjne Q&A (CoQA Benchmark)

**55 F1**

w trybie zero-shot

GPT-2 pokonał 3 z 4 modeli bazowych, które były trenowane na ponad 127,000 etykietowanych przykładów.

Analiza wykazała, że model często stosował proste heurystyki (np. na pytanie 'Kto?' odpowiadał pierwszym imieniem należącym w tekście).



**Wniosek:** Model wykazuje cechy metapoznania – "wie, kiedy wie". Posiada skalibrowaną niepewność, co jest oznaką bardziej zaawansowanego rozumowania niż proste dopasowywanie wzorców.

## Zadanie: Faktyczne Q&A (Natural Questions Benchmark)

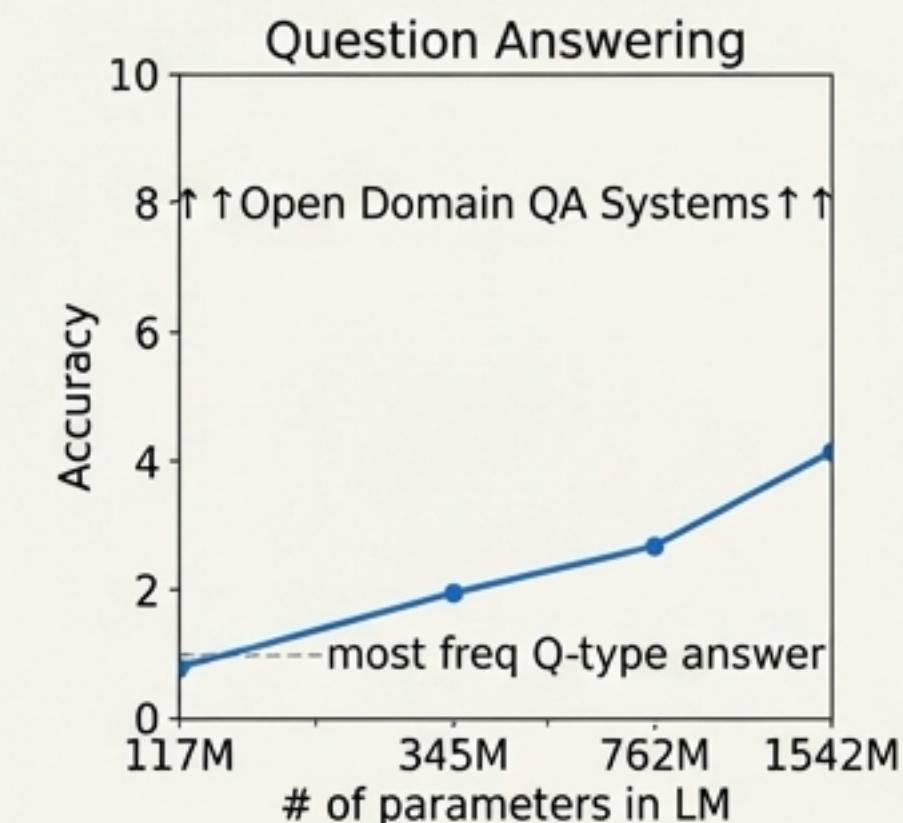
Zaledwie 4.1% poprawnych odpowiedzi.

**Kluczowy Wgląd:** Kiedy model oceniał własną pewność, jego wyniki dramatycznie się poprawiały. Dla 1% odpowiedzi, których był najbardziej pewny, celność wzrosła do 63.1%!

Wszystkie Odpowiedzi

4.1% Celności

Top 1% Najpewniejszych Odpowiedzi  
63.1% Celności



# Ostateczny Dowód: Generalizacja, a nie Zapamiętywanie

**Główna Obawa Sceptyków:** Czy model faktycznie się uczy, czy jest to tylko zaawansowana forma ‘regurgitacji’ ogromnego zbioru danych?

## Metodologia Weryfikacji #1: Analiza Pokrycia Danych

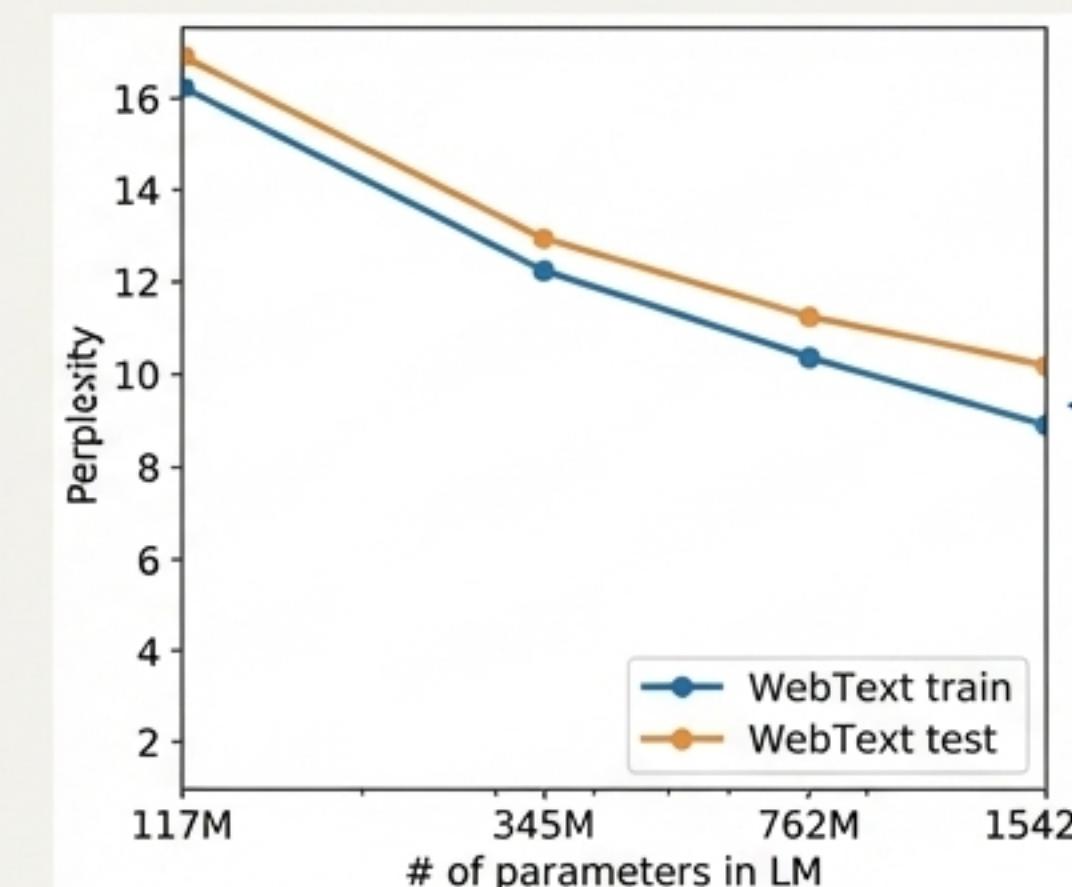
- Użyto filtrów Blooma do wykrycia 8-gramów...
- Stwierdzono minimalne pokrycie (średnio 3.2%).

## Metodologia Weryfikacji #2: Test Ablacyjny (LAMBADA)

- Usunięcie wszystkich przykładów z częściowym pokryciem danych zmieniło celność z 63.2% na **62.9%**.

Różnica jest statystycznie nieistotna.

## Najsilniejszy Dowód: Krzywe Straty (Loss Curves)



Równoległy spadek =  
**Prawdziwa Generalizacja**

Krzywe straty dla zbioru treningowego i testowego WebText przebiegają niemal równolegle, co jest podręcznikowym dowodem na **brak przeuczenia (overfitting)**.

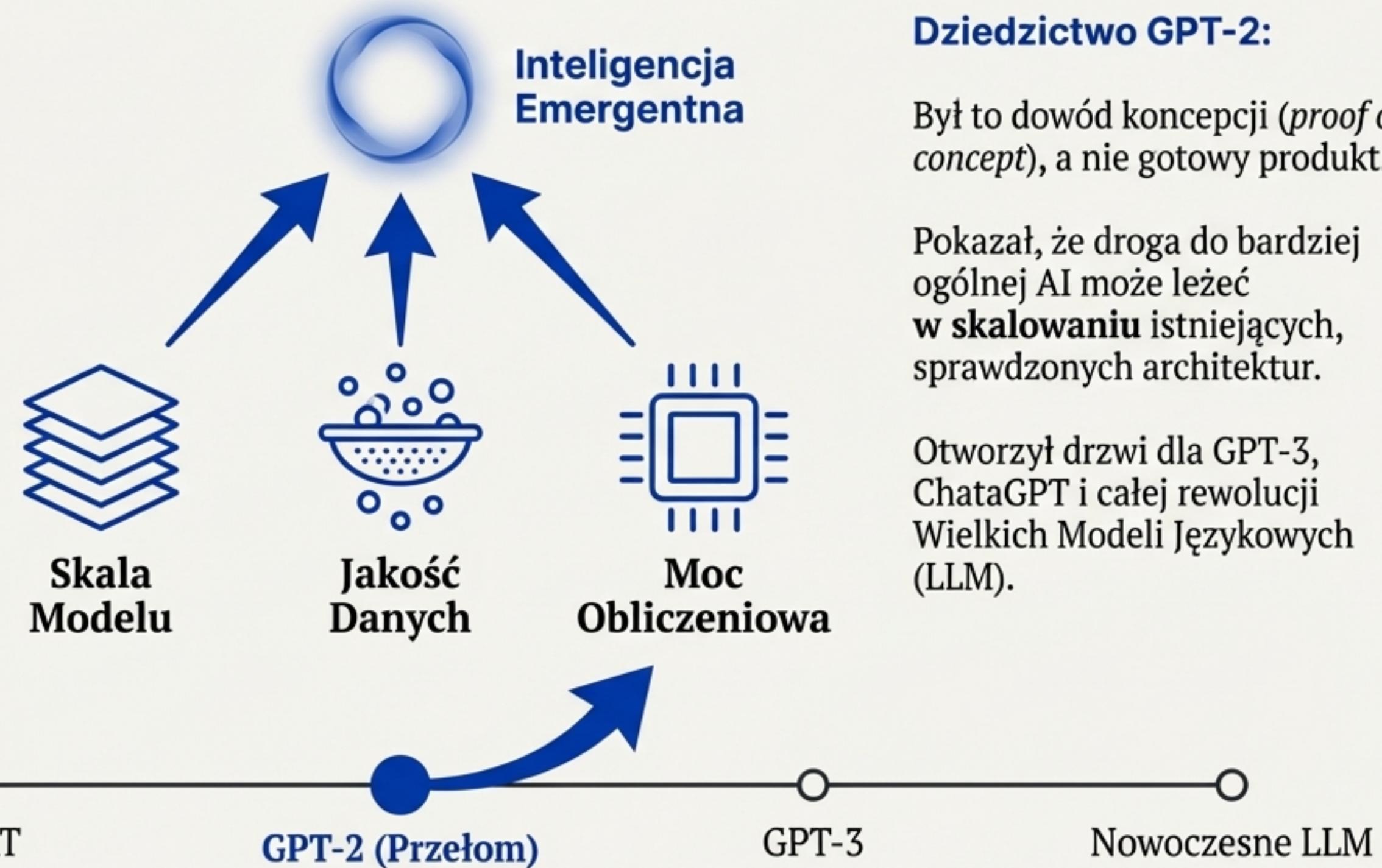
Co więcej, krzywe nie spłaszczały się, co sugeruje, że nawet największy model 1.5B wciąż był w stanie uczyć się więcej z danych (**underfitting**).

# Konsekwencje: Nowa Era w Sztucznej Inteligencji

## Potwierdzona Zmiana Paradygmatu:

Złożone, inteligentne zachowania mogą wyłonić się z prostych celów, jeśli zastosuje się odpowiednią skalę.

Inteligencja jest właściwością emergentną wynikającą z połączenia trzech filarów: **Skali Modelu**, **Jakości Danych** i **Mocy Obliczeniowej**.



**Finalna Myśl:** Praca nad GPT-2 nie była końcem poszukiwań, ale dowodem, że branża AI znalazła niezwykle obiecującą ścieżkę naprzód.