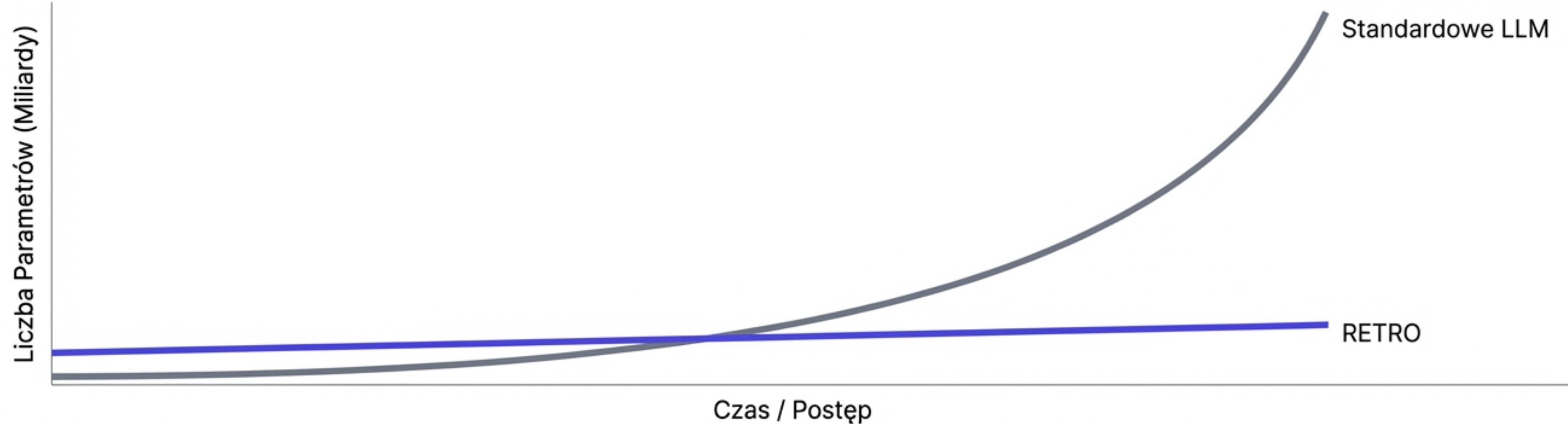


RETRO: Wbrew paradygmatowi „Większe znaczy Lepsze”



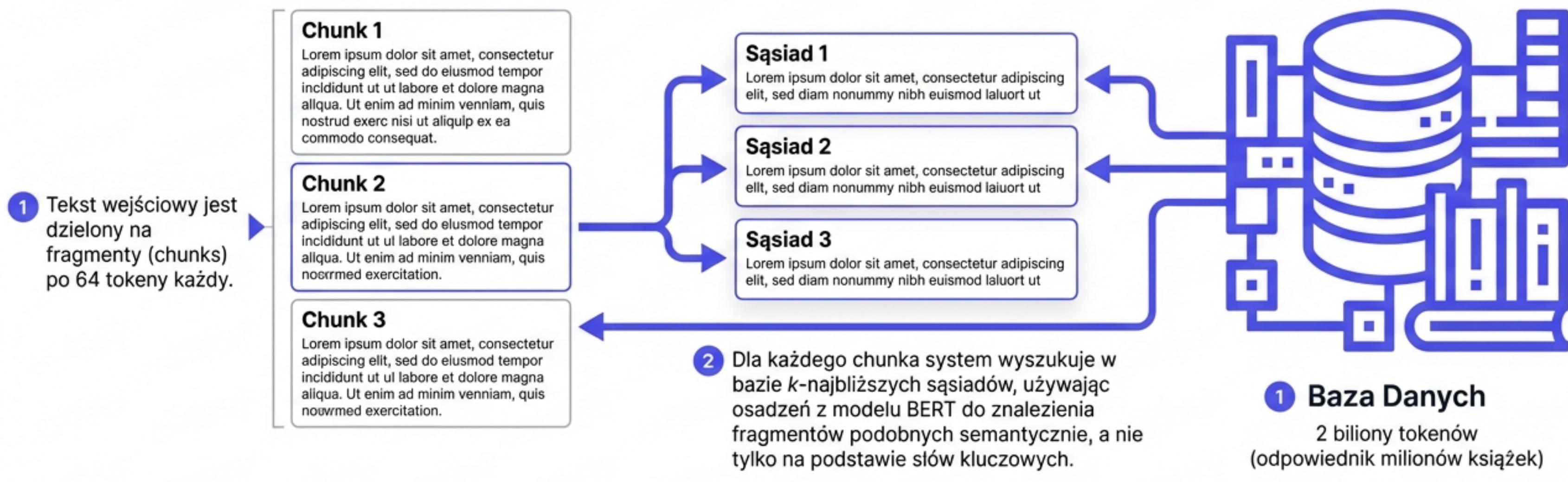
Współczesny wyścig w dziedzinie AI koncentruje się na budowaniu coraz większych modeli, z setkami miliardów, a nawet bilionami parametrów. DeepMind proponuje radykalną alternatywę.



- Zamiast zmuszać model do 'zapamiętania' całej wiedzy świata w swoich parametrach, dajmy mu możliwość korzystania z zewnętrznej 'biblioteki' w czasie rzeczywistym.
- Przedstawiamy **RETRO: Retrieval-Enhanced Transformer**.
- To podejście rzuca wyzwanie mantrze, że surowa wielkość modelu jest jedyną drogą do postępu, proponując w zamian wydajniejszą ścieżkę opartą na dostępie do wiedzy.

Architektura RETRO: Inteligentny Asystent, nie Zwykła Wyszukiwarka

Wyobraź sobie pracę z niezwykle szybkim asystentem badawczym, który w czasie rzeczywistym podsuwa Ci najtrajniejsze fragmenty z milionów książek.

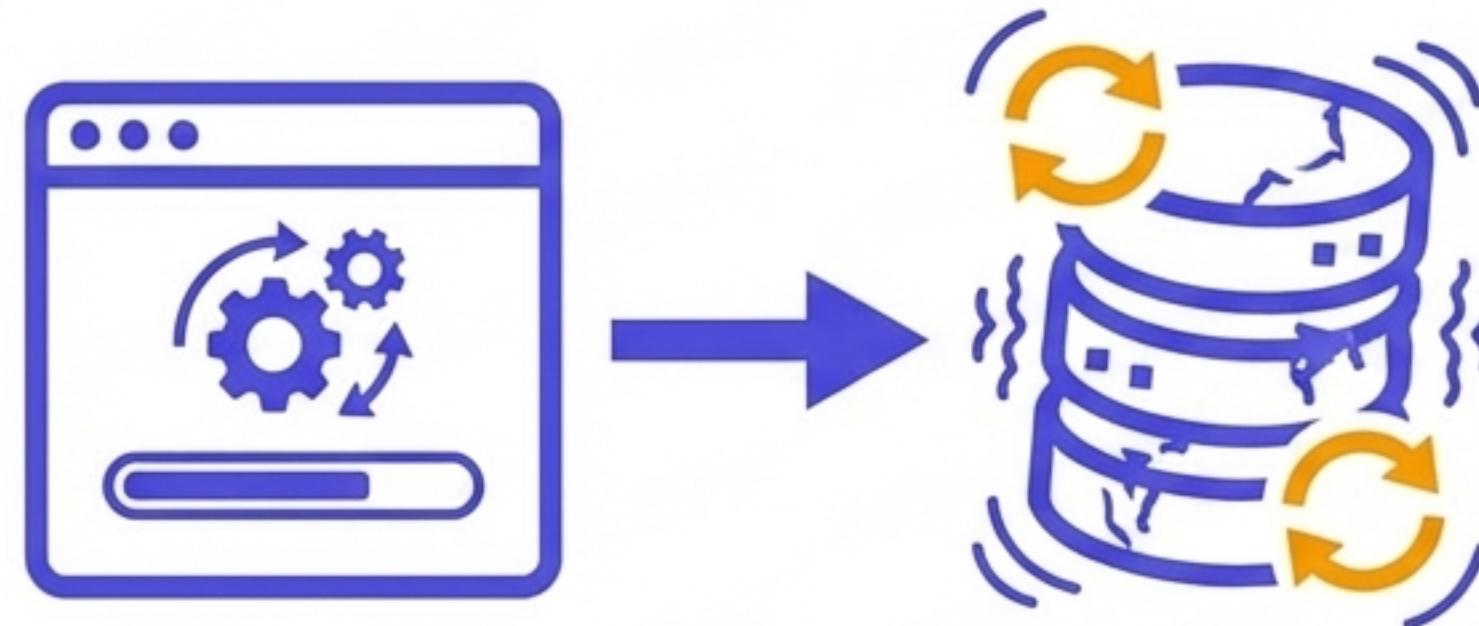


To znacznie bardziej subtelne i zintegrowane niż proste wklejenie wyników wyszukiwania do promptu.

Zamrożony Koder BERT: Pragmatyczny Kompromis na Rzecz Wydajności

Dlaczego nie pozwolić retrieverowi uczyć się i stawać się lepszym?

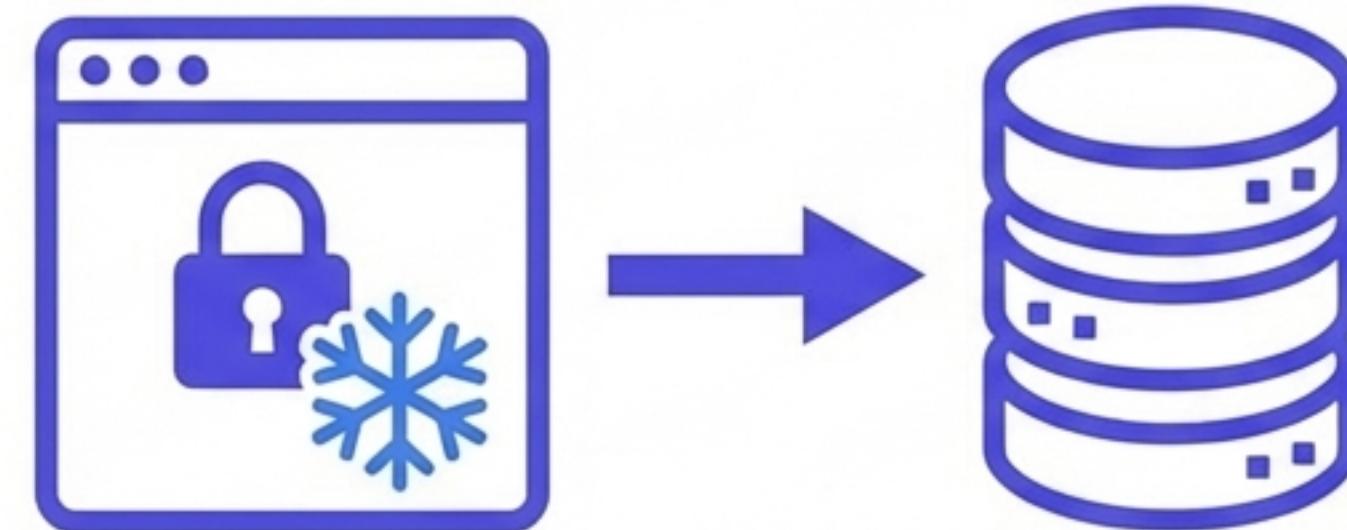
Gdyby koder się uczył...



Koszt Obliczeniowy!

Gdyby koder BERT uczył się, reprezentacje wektorowe tekstu stale by się zmieniały. Wymagałoby to ponownego indeksowania całej bazy danych (2 biliony tokenów) po każdej aktualizacji. Taki obliczeniowy koszmar zniweczyłby wszystkie korzyści płynące z wydajności.

Rozwiązanie RETRO



Indeksacja jednorazowa

Zamrożony koder = stabilne klucze. Indeks tworzony jest tylko raz. Jest to świadomy kompromis między absolutną optymalizacją a praktyczną wykonalnością na masową skalę.

Mechanizm Chunked Cross-Attention (CCA): Integracja Wiedzy z Generowaniem

Poza retrieverem, CCA jest drugim filarem RETRO.

Standardowy Transformer

Używa mechanizmu *self-attention*, aby analizować kontekst wewnątrz własnego tekstu.

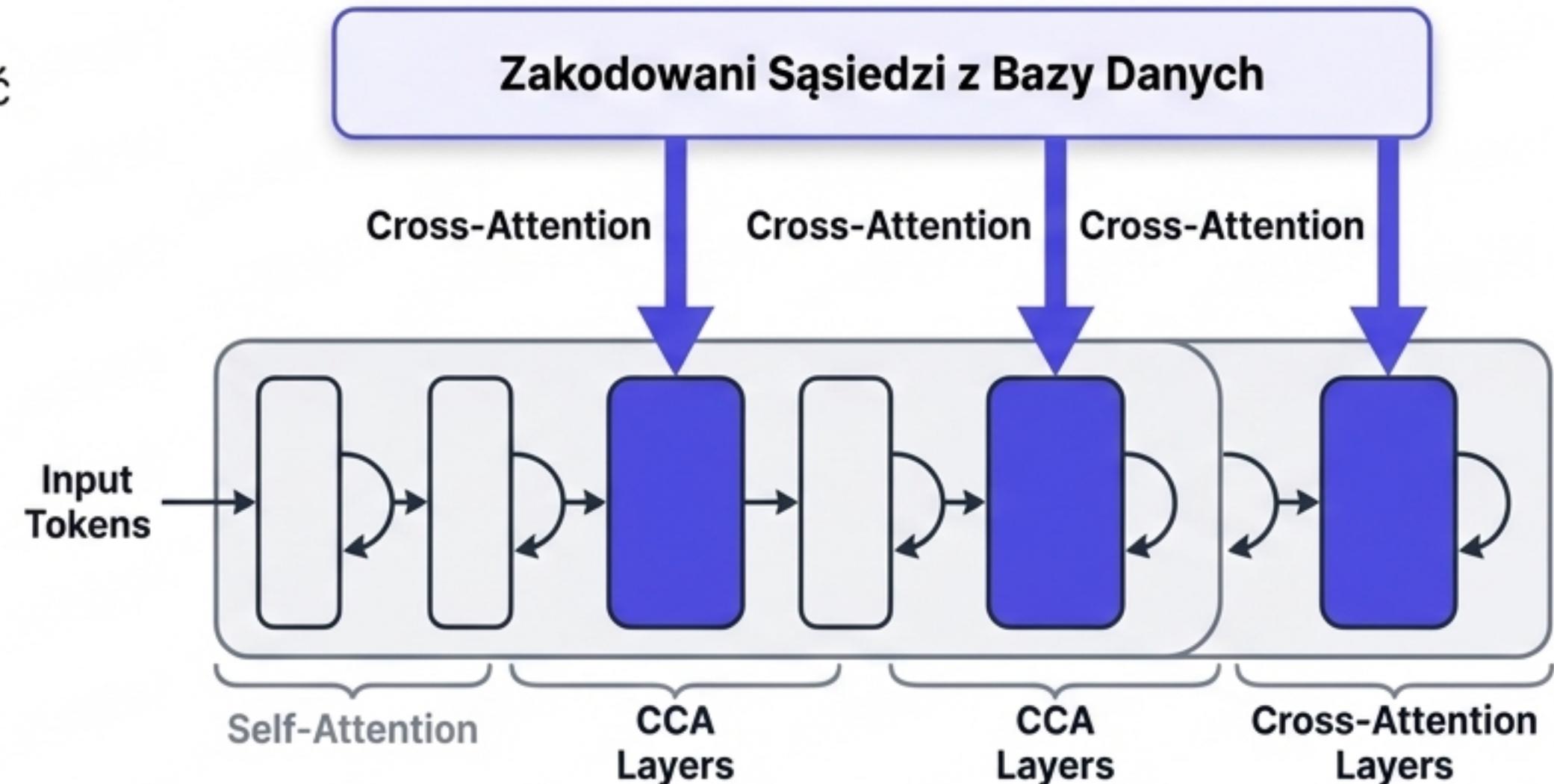
RETRO z CCA

Dodatkowo pozwala modelowi "zwracać uwagę" (*attend to*) na odzyskane fragmenty z zewnętrznej bazy wiedzy.

W praktyce

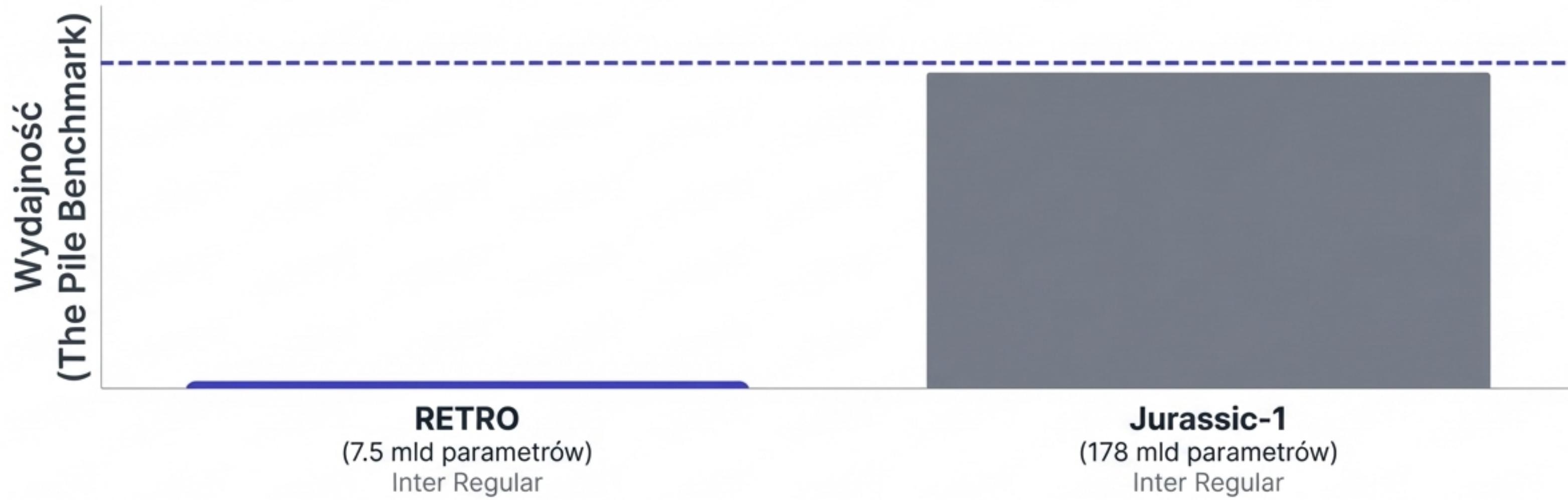
Przewidując następny token, RETRO bierze pod uwagę:

1. Istniejący kontekst w tekście wejściowym.
2. Najbardziej relevantne przykłady z bazy danych.



Umożliwia to znacznie bardziej precyzyjną i faktograficzną generację, ponieważ odzyskana wiedza bezpośrednio wpływa na predykcję.

Osiągnięcia: Wydajność Porównywalna z Modelami 25x Większymi



The Bold Claim: RETRO, mając 25 razy mniej parametrów, osiąga wydajność porównywalną z modelami takimi jak GPT-3 i Jurassic-1.

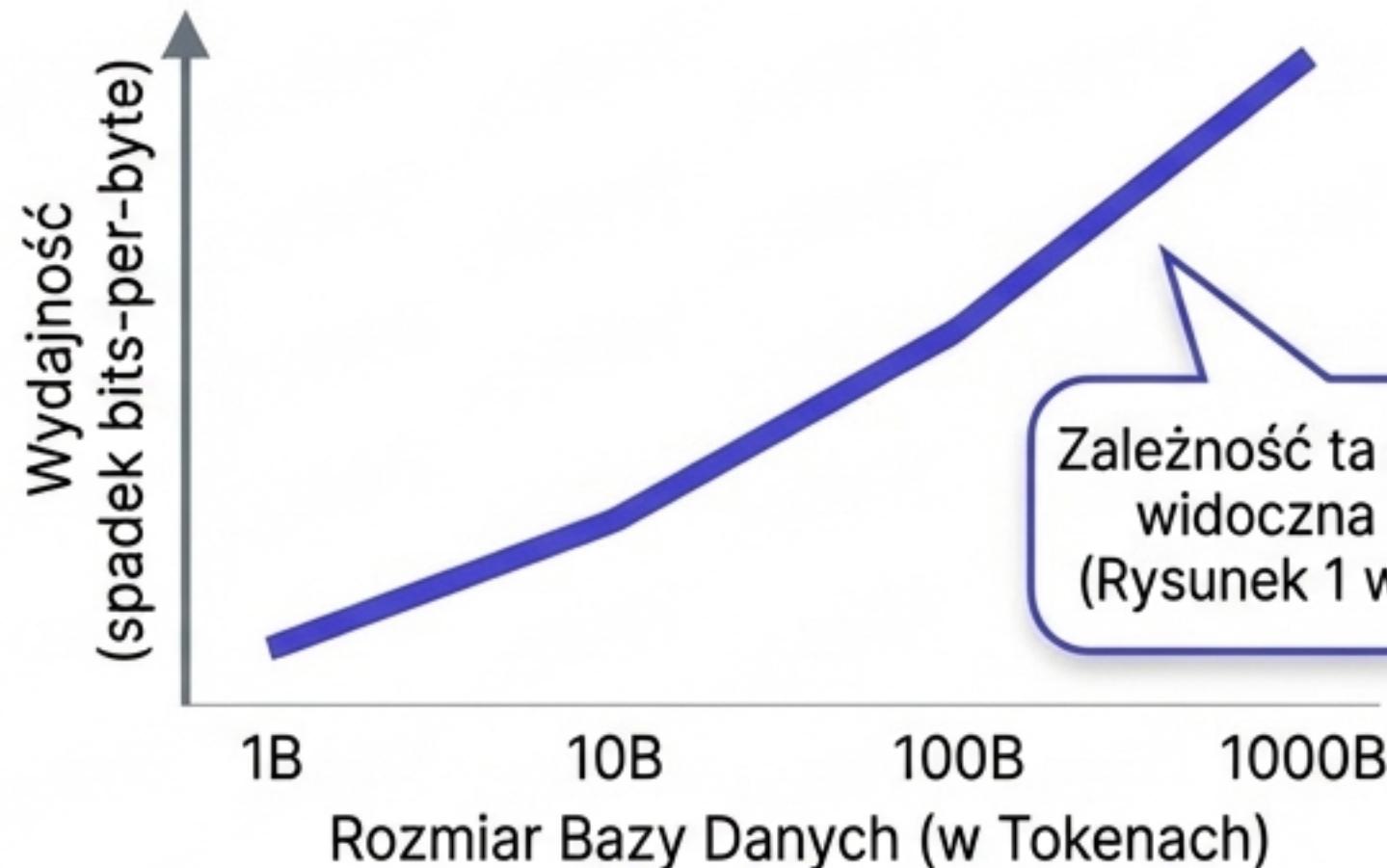
Benchmark: Testy przeprowadzono na zróżnicowanym benchmarku The Pile, składającym się z 22 zestawów testowych (od kodu, przez literaturę, po artykuły naukowe).

Wynik: RETRO uzyskało lepsze wyniki na większości z 22 testów. To nie marginalna, lecz szeroka i solidna przewaga.

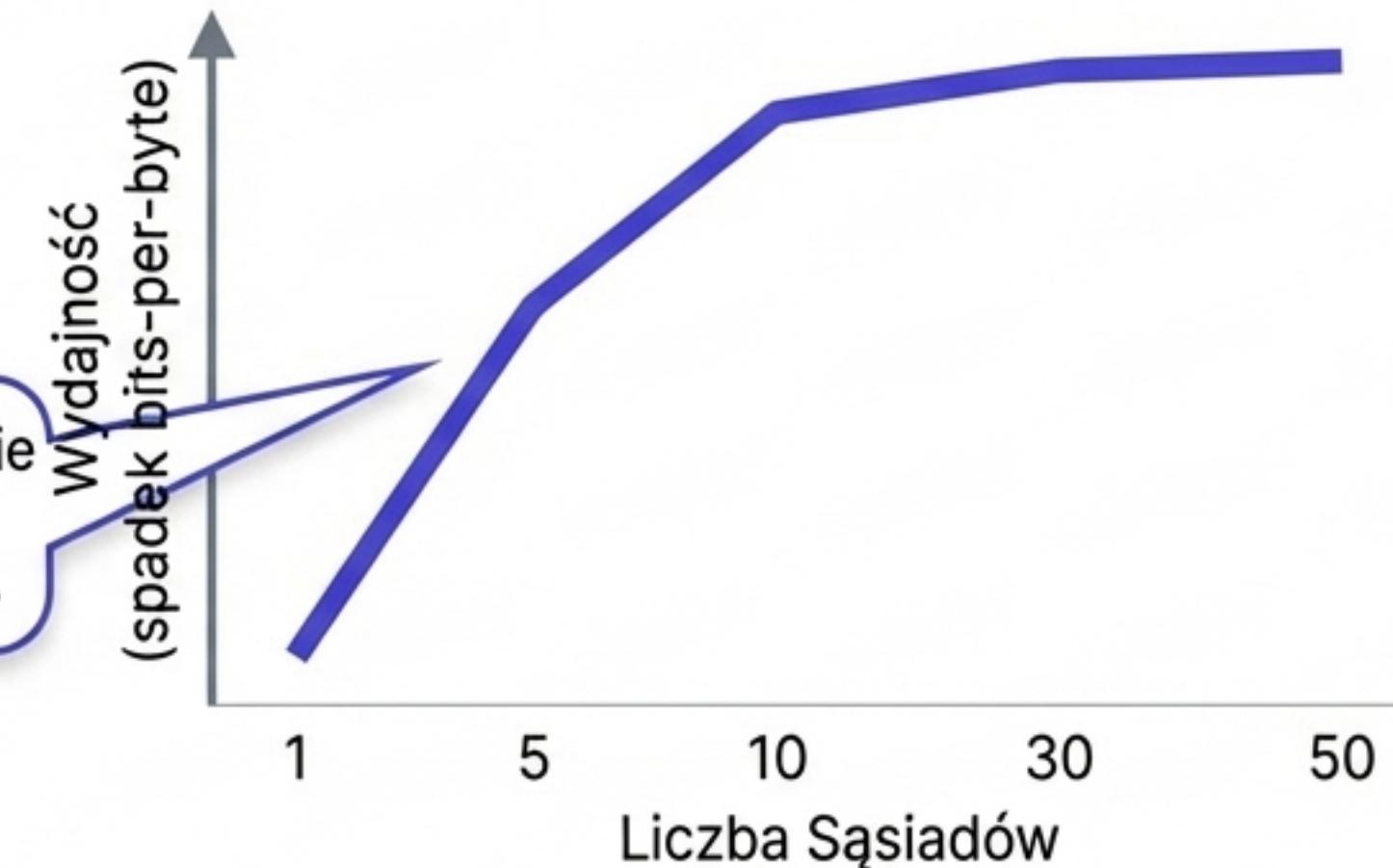
Nowy Wymiar Skalowania: Większa Baza Danych = Inteligentniejszy Model

Kluczowe Odkrycie: Wydajność RETRO skala się nie tylko z liczbą parametrów, ale również z wielkością bazy danych.

Wpływ Rozmiaru Bazy Danych



Wpływ Liczby Sąsiadów

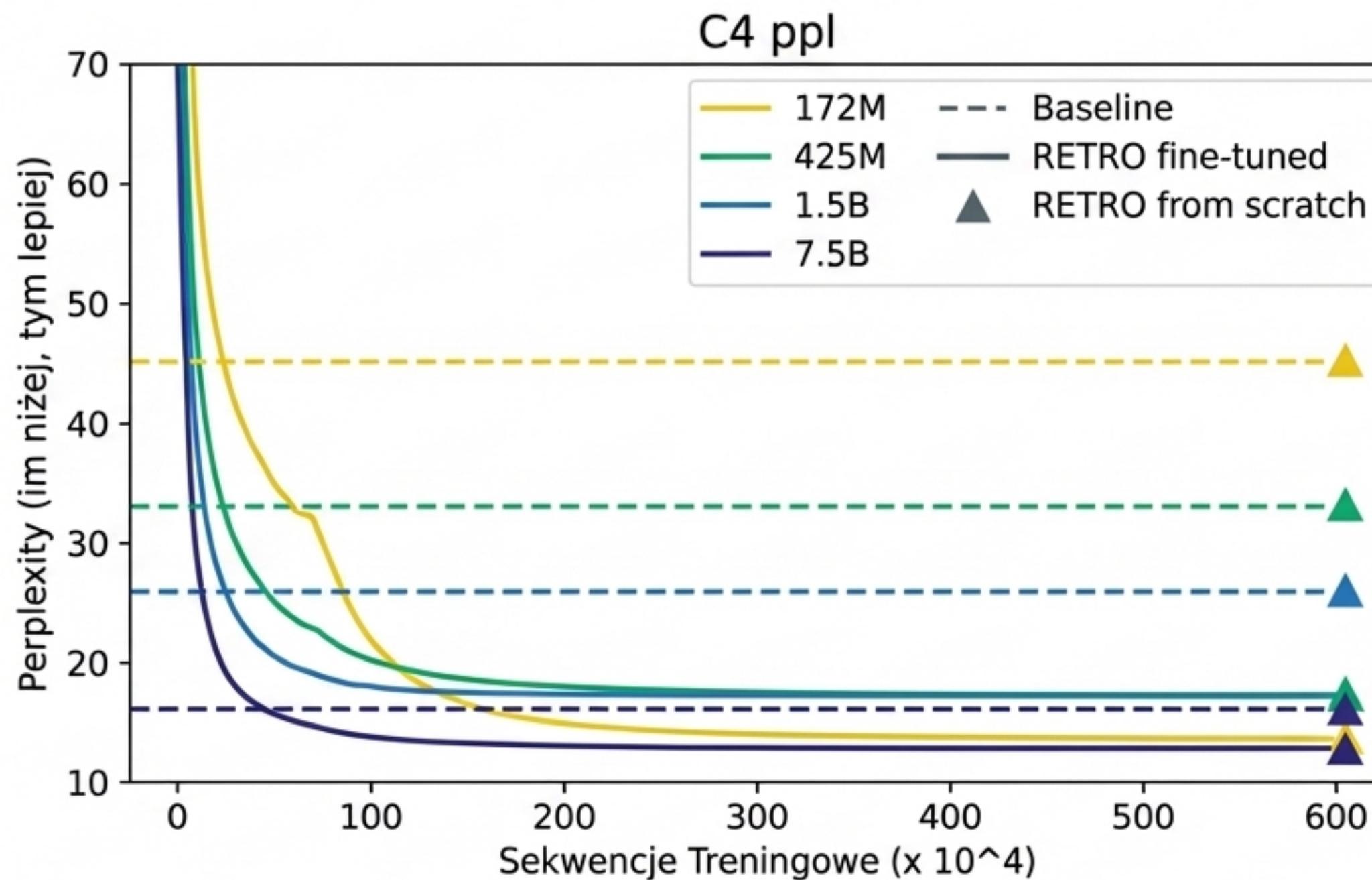


Większa biblioteka = mądrzejszy model.

Ekonomia Skalowania: Zwiększanie bazy danych jest o rzędy wielkości tańsze niż trenowanie coraz większych modeli. To fundamentalnie zmienia ekonomię rozwoju AI, wprowadzając nowy, efektywny kosztowo wymiar skalowania.

Retro-fitting: Błyskawiczna Modernizacja Istniejących Modeli

Nie trzeba budować modelu od zera, aby skorzystać z mocy RETRO.



Statystyki

- Możliwe jest dodanie zdolności RETRO do już wytrenowanych modeli LLM.
- Proces dostrajania wymagał jedynie **3%** oryginalnych danych treningowych.
- Osiągnięto wydajność niemal równoważną z modelami RETRO trenowanymi od podstaw.

Analogia



To jak instalacja silnika hybrydowego w starym samochodzie w jedno popołudnie – ogromny wzrost wydajności przy ułamku kosztów.

Odpowiedź na Krytykę: Czy RETRO to Tylko Wyrafinowane 'Kopiuj-Wklej'?

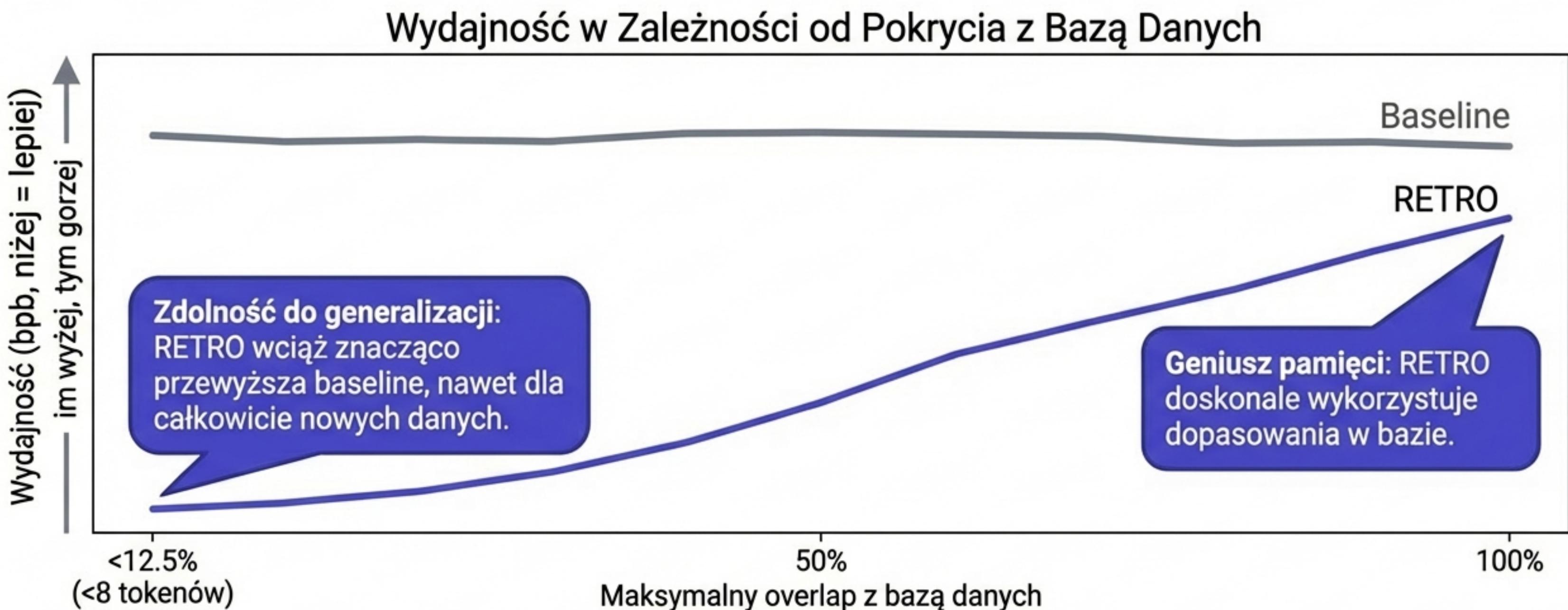
Najważniejsza Wątpliwość:

Czy RETRO po prostu znajduje dokładne odpowiedzi w swojej ogromnej bazie danych i je kopiuje, zamiast faktycznie 'rozumieć' 'rozumieć'?



1. Autorzy byli w pełni świadomi tego ryzyka i podezli do niego metodycznie.
2. Stworzyli system do mierzenia wydajności w zależności od stopnia pokrycia (overlap) między fragmentem testowym a bazą danych.
3. Skategoryzowali dane testowe według podobieństwa: od niemal identycznych po całkowicie nowe.
4. Zdefiniowali fragment 'nowy' jako mający mniej niż 8 wspólnych tokenów z najbliższym sąsiadem w 2-bilionowej bazie.

Dowód: Geniusz Pamięci i Zdolność do Generalizacji



Wyniki z Rysunku 6 są jednoznaczne: To twardy dowód na to, że system robi coś więcej niż kopiowanie. Uczy się generalizować i wykorzystywać odzyskaną wiedzę w nowych kontekstach.

Trzy Fundamentalne Korzyści (1/2): Interpretowalność i Aktualizacja Wiedzy



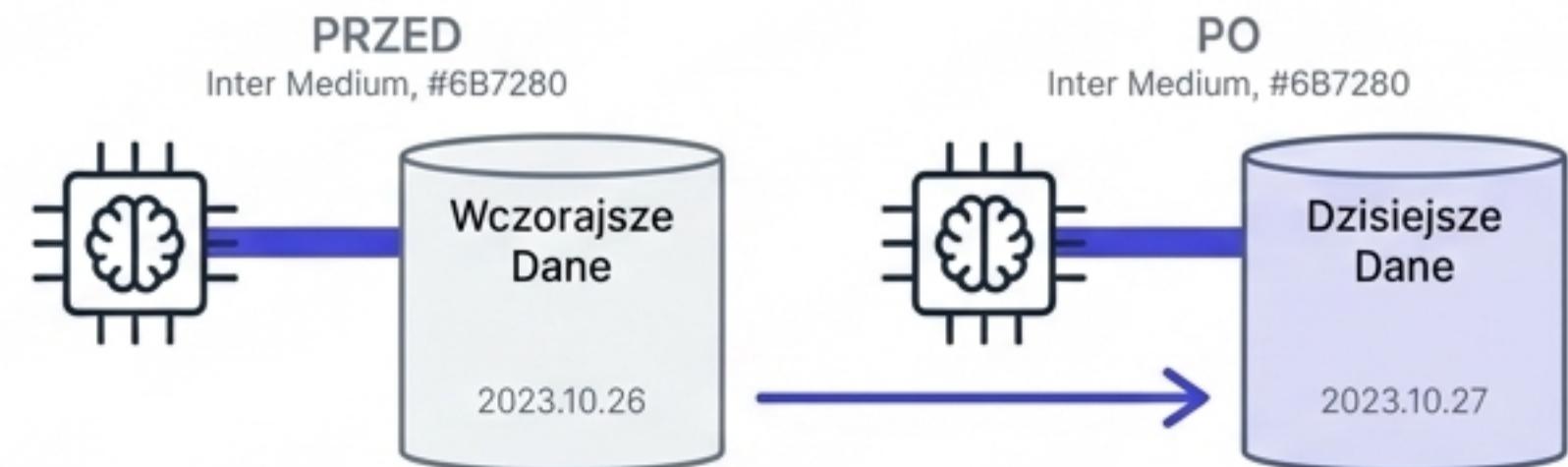
1. Interpretowalność

Proces ten pozwala modelowi na generowanie odpowiedzi, które są spójne i faktyczne, poprzez aktywne wyszukiwanie i wykorzystywanie informacji z zewnętrznej bazy wiedzy.

- Fragment 1: Wyszukiwanie informacji z bazy
- Fragment 2: Integracja danych w odpowiedzi
- Fragment 3: Baza wiedzy jako źródło



2. Łatwe Aktualizacje Wiedzy



- Możemy zobaczyć, jakie dokładnie informacje model odzyskał, aby sformułować odpowiedź.
- To ogromny krok w kierunku zrozumienia 'procesu myślowego' modelu.

- Aby zaktualizować wiedzę modelu, wystarczy zaktualizować bazę danych, zamiast przeprowadzać kosztowny re-trening.
- Jest to o rzędy wielkości tańsze i pozwala modelowi na bieżąco przyswajać nowe informacje.

Trzy Fundamentalne Korzyści (2/2): Wydajność i Bezpieczeństwo



3. Wydajność i Bezpieczeństwo

Wydajność Obliczeniowa

RETRO oferuje bardziej efektywną ścieżkę do osiągania wysokiej wydajności, redukując koszty energetyczne i obliczeniowe związane z trenowaniem gigantycznych modeli.

Kontrola i Bezpieczeństwo

- Jeśli po treningu okaże się, że pewne dane w bazie prowadzą do generowania treści tendencjacyjnych lub toksycznych, można je odfiltrować z bazy retrievalu.
- Daje to możliwość 'korekty' modelu bez potrzeby kosztownego ponownego trenowania, co jest niezwykle trudne w standardowych LLM.

Nowy Paradygmat: Od Skalowania Parametrów do Skalowania Dostępu do Wiedzy



Stary Paradygmat: Parametry = Wiedza



Nowy Paradygmat: Parametry + Dostęp do Wiedzy

- RETRO to nie tylko lepszy model. To lepszy sposób budowania modeli językowych.
- Przenosi on ciężar z pasywnego "zapamiętywania" wiedzy w parametrach na aktywny, jawnego "dostęp" do informacji na bezprecedensową skalę.
- Ta praca demmstruuje, że podejścia semi-parametryczne są wydajną i potężną alternatywą dla samego skalowania liczby parametrów, otwierając nowe, ekscytujące ścieżki rozwoju AI.