

# Kryzys Skalowalności: Potęga AI napotyka mur energetyczny

Modele takie jak GPT-3 zrewolucjonizowały NLP, ale ich wykładniczy wzrost kosztów i zużycia energii stał się barierą nie do pokonania.

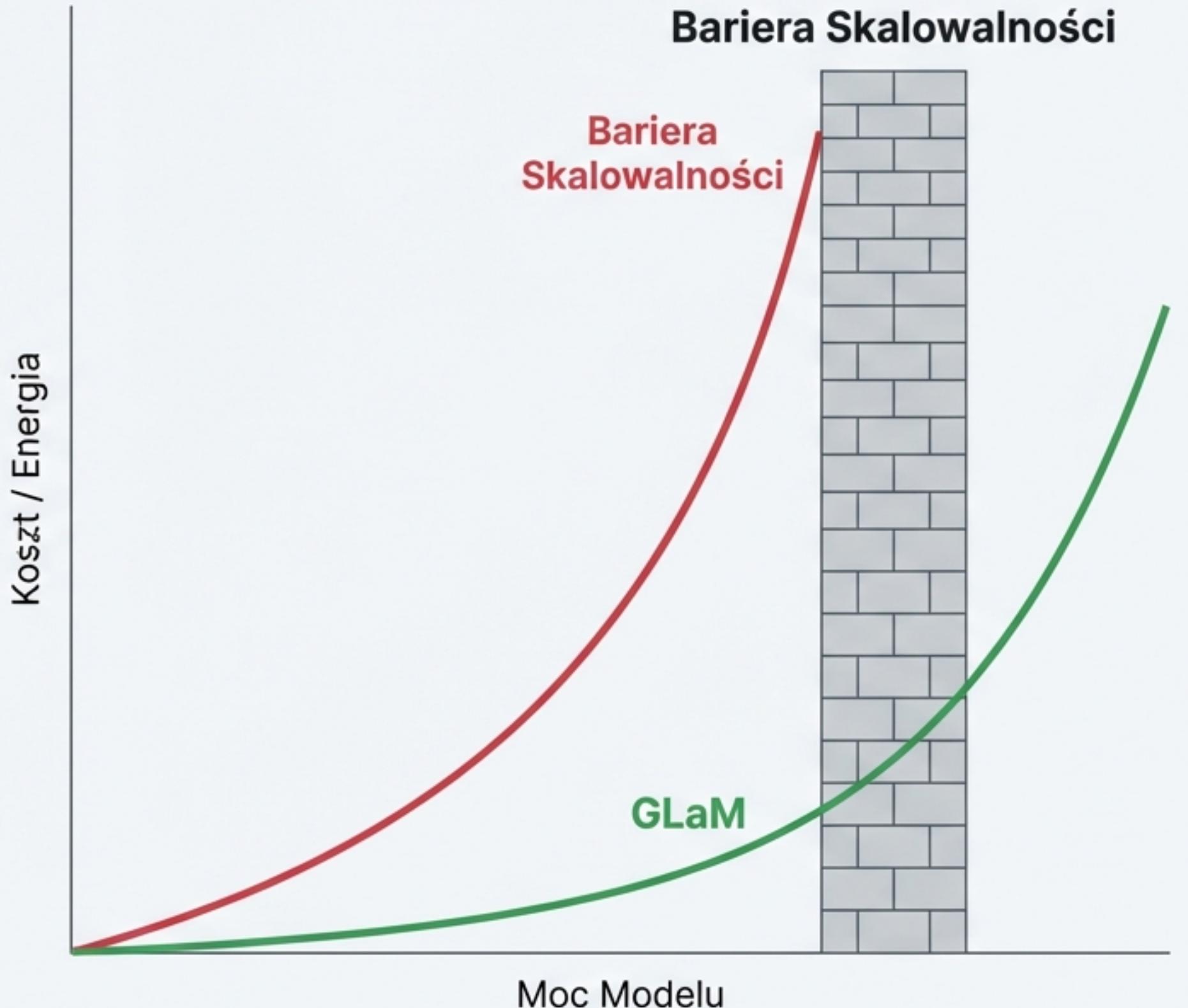
Modele gęste (dense models) udowodniły, że skalowanie jest kluczem do zaawansowanych zdolności językowych. Jednak każda kolejna, późniejsza wersja wymaga ogromnych zasobów obliczeniowych i energetycznych, stając się „prohibitynie kosztowna”.

**Problem fundamentalny:** Czy możemy budować coraz intelligentniejszą AI bez konieczności budowania dla niej dedykowanych elektrowni?

## Przełom

Publikacja GLaM rzuca wyzwanie paradygmatowi „większy znaczy lepszy”.

Obietnica: Światowej klasy wydajność przy zużyciu zaledwie **1/3 energii** potrzebnej do treningu GPT-3.



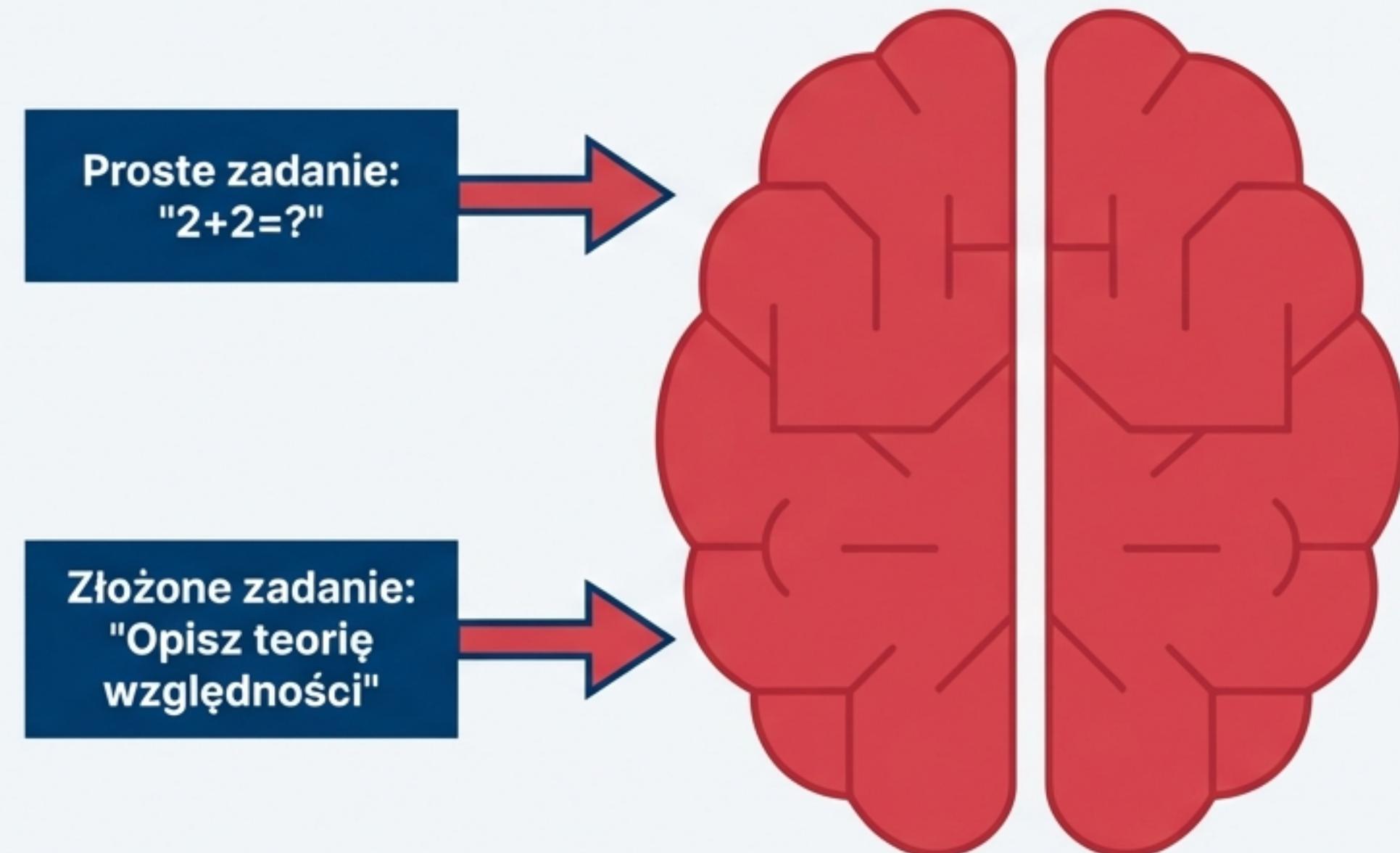
# Problem fundamentalny: Modele gęste aktywują 100% swojej wiedzy do każdego zadania

W sercu architektury gęstej leży fundamentalne marnotrawstwo zasobów.

- Modele gęste aktywują **WSZYSTKIE** swoje parametry do przetworzenia **KAŻDEGO** pojedynczego tokenu.
- Przykład:** GPT-3 (175 mld parametrów) angażuje całą swoją sieć neuronową, niezależnie od tego, czy wykonuje złożone wnioskowanie, czy proste zapytanie. W architekturze gęstej,  $n_{\text{params}} = n_{\text{act}} - \text{params}$ .
- Brak specjalizacji:** Model nie posiada mechanizmu alokacji zasobów adekwatnych do trudności zadania.

**Analogia:** Wyobraź sobie wszechstronnego eksperta, „Pana Gęstego”, który musi zaangażować całą moc swojego mózgu, aby obliczyć  $2+2$ . To definicja nieefektywności.

**Model Gęsty: 175 mld parametrów**



# Nowa architektura inteligencji: Rzadko Aktywowana Sieć Ekspertów (MoE)

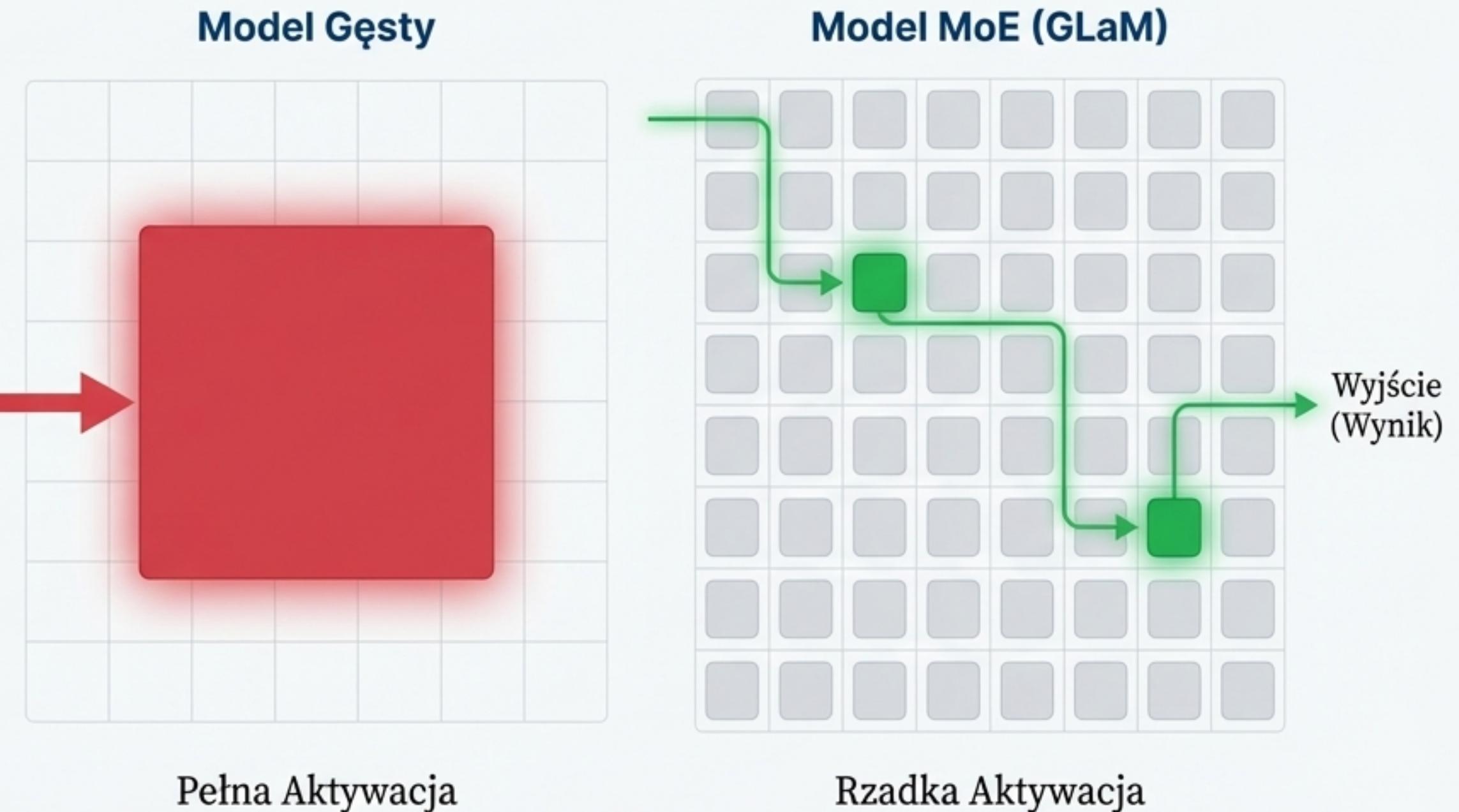
Zamiast jednego, monolitycznego geniusza – modularna „korporacja” wyspecjalizowanych ekspertów.

GLaM porzuca architekturę gęstą na rzecz modelu rzadko aktywowanego (sparsely activated), opartego na koncepcji Mixture of Experts (MoE).

**Jak to działa:** Zamiast jednego bloku przetwarzającego, co druga warstwa Transformer w GLaM jest zastąpiona przez warstwę MoE.

**Zbiór specjalistów:** Każda warstwa MoE zawiera zbiór niezależnych sieci neuronowych (np. 64 „ekspertów”) – każda wyspecjalizowana w innej dziedzinie (np. poezja, kodowanie w Pythonie, biologia molekularna).

**Inteligentna aktywacja:** Dla każdego tokenu aktywowani są tylko najbardziej odpowiedni eksperci. Reszta pozostaje w uśpieniu, nie zużywając zasobów.



# Mózg operacji: Funkcja Bramkująca (Gating Function) jako inteligentny router

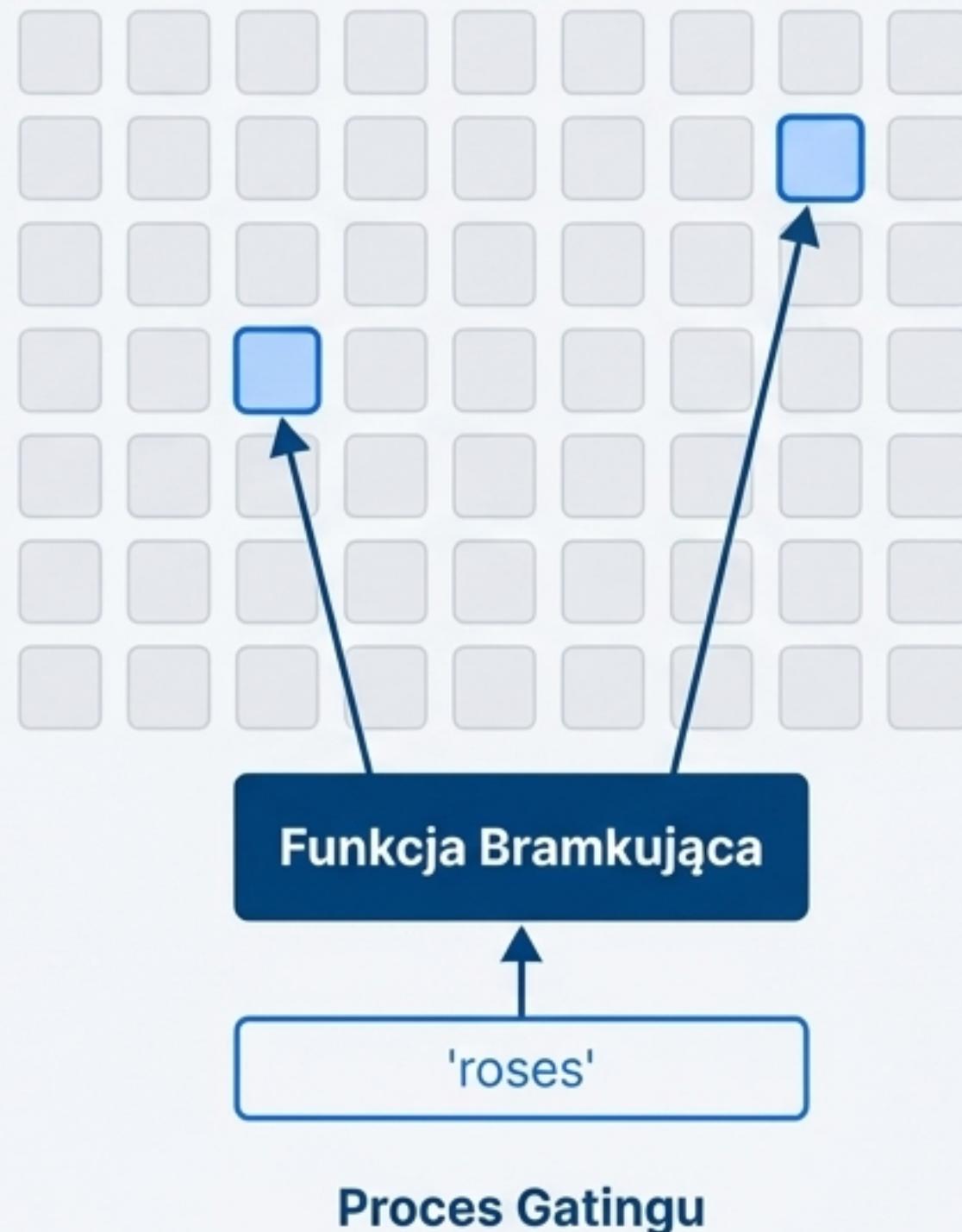
**Jak model wie, którego eksperta wezwać do pracy?**

Sercem architektury MoE jest **funkcja bramkująca (gating function)**. Działa ona jak inteligentna „recepja”, która w ułamku sekundy kieruje każdy token do odpowiednich specjalistów.

**Dynamiczna selekcja:** Na podstawie danych wejściowych, funkcja bramkująca dynamicznie wybiera **dwoch najlepszych ekspertów** dla każdego tokenu.

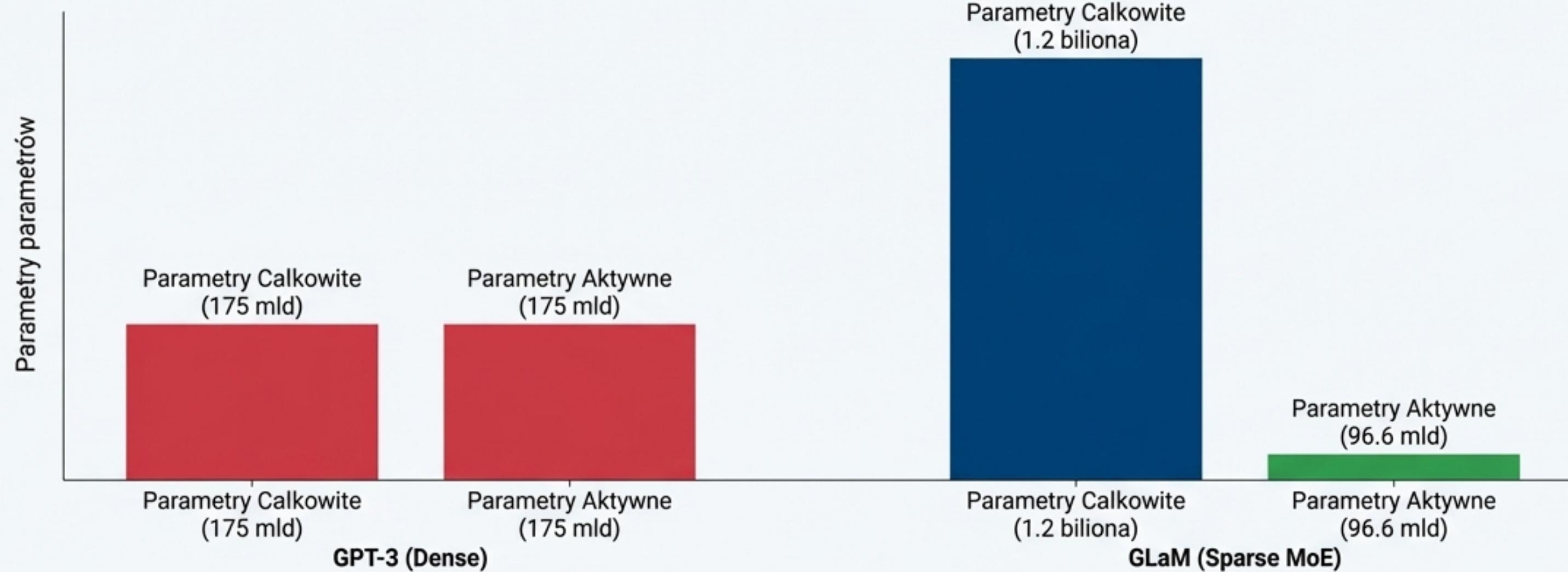
**Elastyczność obliczeniowa:** Zamiast jednej ścieżki przetwarzania, jak w klasycznym Transformerze, model ma do dyspozycji  $O(E^2)$  kombinacji sieci neuronowych, co prowadzi do znacznie większej elastyczności.

**Klucz do wydajności:** Ta zdolność do selektywnej aktywacji zaledwie ułamka wszystkich parametrów pozwala zachować jakość przy drastycznej redukcji kosztów obliczeniowych.



# GLaM w liczbach: Paradoks większej skali i mniejszej aktywacji

Jak model ~7x większy od GPT-3 może być bardziej oszczędny w działaniu?



Metryka	GPT-3 (Dense)	GLaM (Sparse MoE)
Całkowita liczba parametrów (n_params)	175 mld	1.2 biliona (~7x więcej)
Aktywne parametry / token (n_act-params)	175 mld (100%)	96.6 mld (~8% całości)

Dzięki rzadkiej aktywacji, ogromny rozmiar GLaM staje się jego atutem (potężny rezeruar wiedzy), a nie obciążeniem obliczeniowym.

**Analogia:** To jak mieć dostęp do całej biblioteki narodowej, ale zamiast czytać wszystko, precyzyjnie wybierać dwa najbardziej trafne akapity z dwóch odpowiednich książek.



# Wymierne korzyści: Radykalna redukcja kosztów treningu i inferencji

Przełomowa wydajność przekłada się na realne oszczędności w skali przemysłowej.



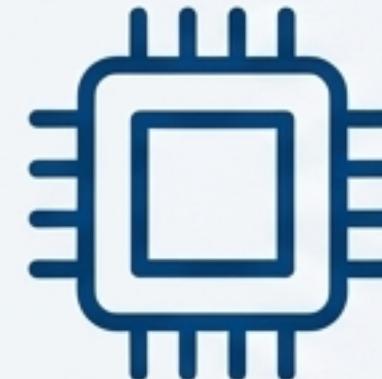
Energia Treningu

**-64.6%**

vs GPT-3

GLaM zużył zaledwie **456 MWh**, podczas gdy GPT-3 potrzebował **1287 MWh**.

Znacząca redukcja kosztów na każdym etapie cyklu życia modelu – od badań po wdrożenie – otwiera drogę do bardziej zrównoważonego rozwoju AI.



Koszt Inferencji (działania)

**-48.6%**

vs GPT-3

GLaM wymaga **180 GFLOPs** na token, w porównaniu do **350 GFLOPs** dla GPT-3.

# Lepsza wydajność, nie kompromis: GLaM deklasuje rywali w benchmarkach

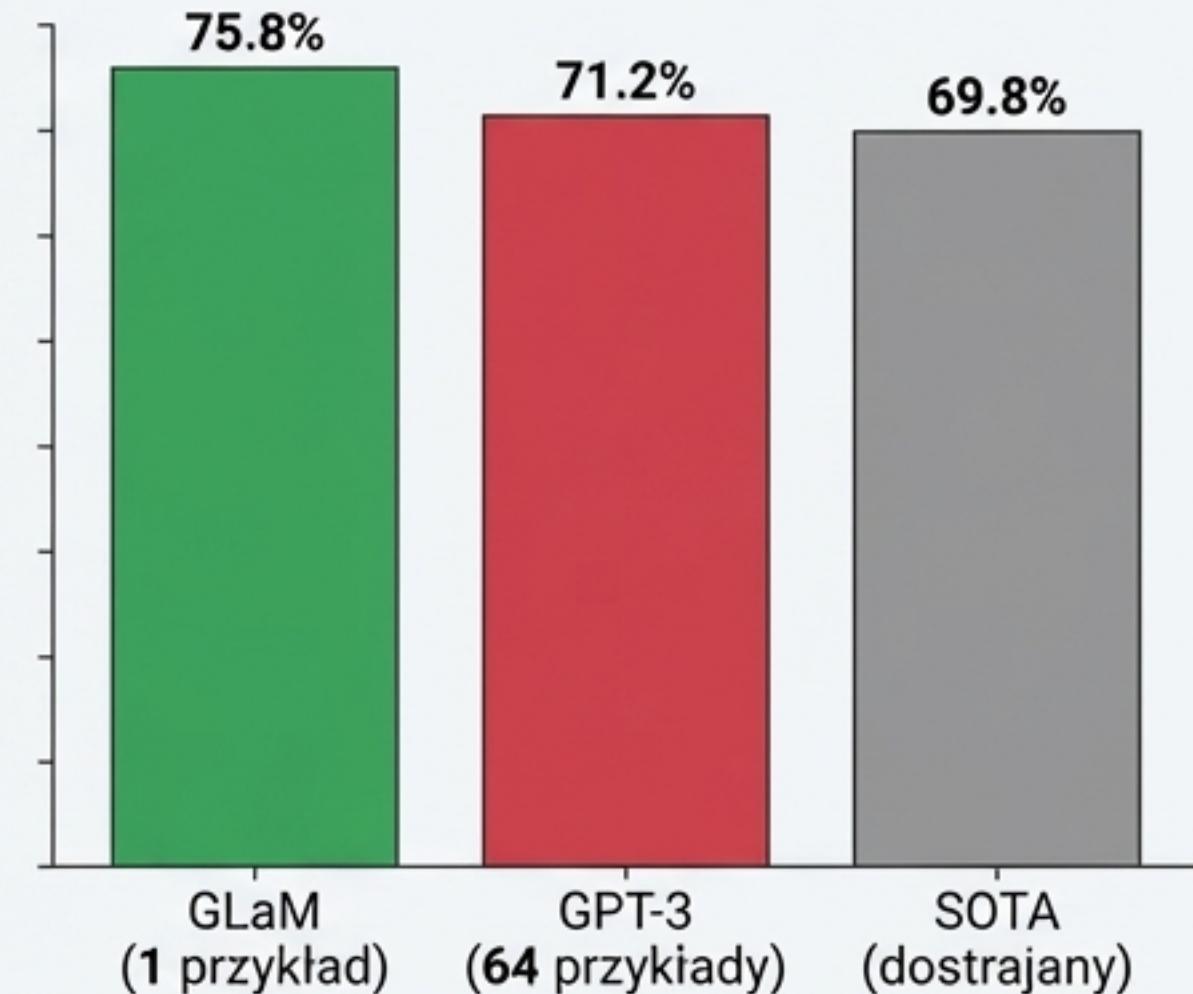
GLaM osiągnął średnio lepsze wyniki na 29 benchmarkach NLP w zadaniach zero-shot, one-shot i few-shot.

## Spektakularny wynik na TriviaQA (Open-Domain):

- **GLaM (one-shot)**: Osiągnął **75.8%** dokładności, widząc tylko **jeden przykład**.
- **GPT-3 (64-shot)**: Potrzebował aż **64** przykładów, by osiągnąć zaledwie **71.2%**.

Więcej niż tylko pokonanie GPT-3: Wynik GLaM (one-shot) przewyższył nawet poprzedni, specjalistyczny model SOTA (State-Of-The-Art), który był **specjalnie dostrajany** (fine-tuned) do tego konkretnego zadania.

Dokładność na TriviaQA (Open-Domain)



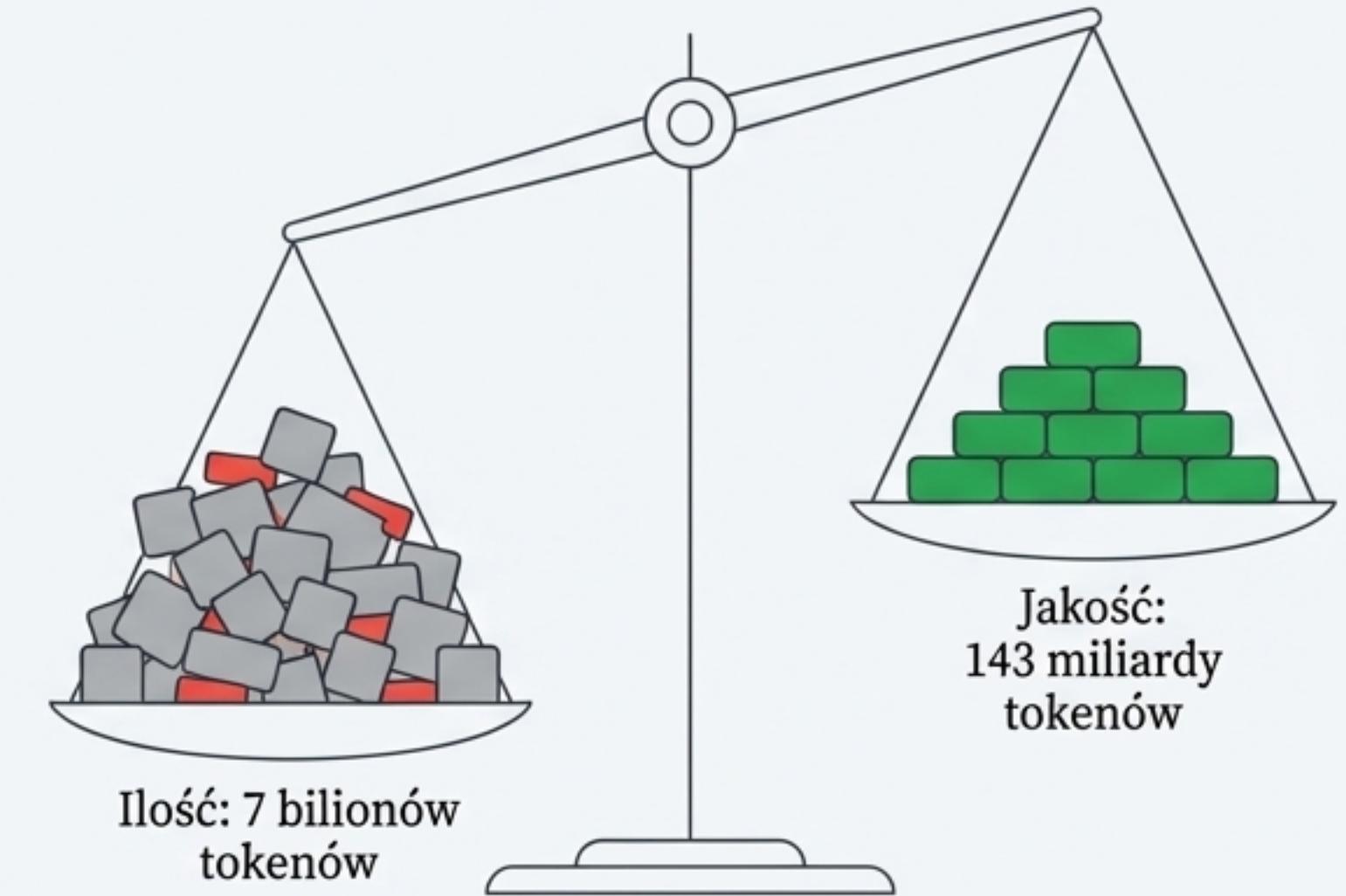
**Wniosek:** To dowód, że nieaktywne parametry nie są martwym balastem. Działają jak ogromny, pasywny rezeruar wiedzy, który model potrafi precyjnie wykorzystać, gdy jest to potrzebne. Metafora: „Amator” (model ogólnego przeznaczenia) pokonuje „zawodowca” (model specjalistyczny) na jego własnym boisku.

# Kluczowy eksperyment: Jakość danych deklasuje ich ilość

Co się stanie, gdy dwa identyczne, mniejsze modele GLaM nakarmimy radykalnie różnymi danymi?

Przeprowadzono decydujący eksperyment na modelu GLaM (1.7B/64E):

- **Model A (Ilość):** Trenowany na ~7 bilionach tokenów z niefiltrowanych danych internetowych.
- **Model B (Jakość):** Trenowany na 143 miliardach tokenów (~50x mniej!) z wysokiej jakości, starannie wyselekcjonowanych i przefiltrowanych danych.
- **Wynik: Nokaut.** Model trenowany na mniejszym, ale czystszy zbiorze danych, osiągnął znacznie lepsze wyniki na zadaniach NLU i NLG.



**Fundamentalna zasada AI:** Jakość danych jest ważniejsza niż ich ilość. Zasada 'śmieci na wejściu, śmieci na wyjściu' (garbage in, garbage out) jest nieubłagana. Model, ucząc się wzorców statystycznych, traktuje szum i błędy jako prawdziwe sygnały, co prowadzi do błędnych korelacji.

# Kompromisy i ograniczenia: Każda potęga ma swoją cenę

GLaM jest niezwykle wydajny w działaniu, ale wymagający pod względem infrastruktury.

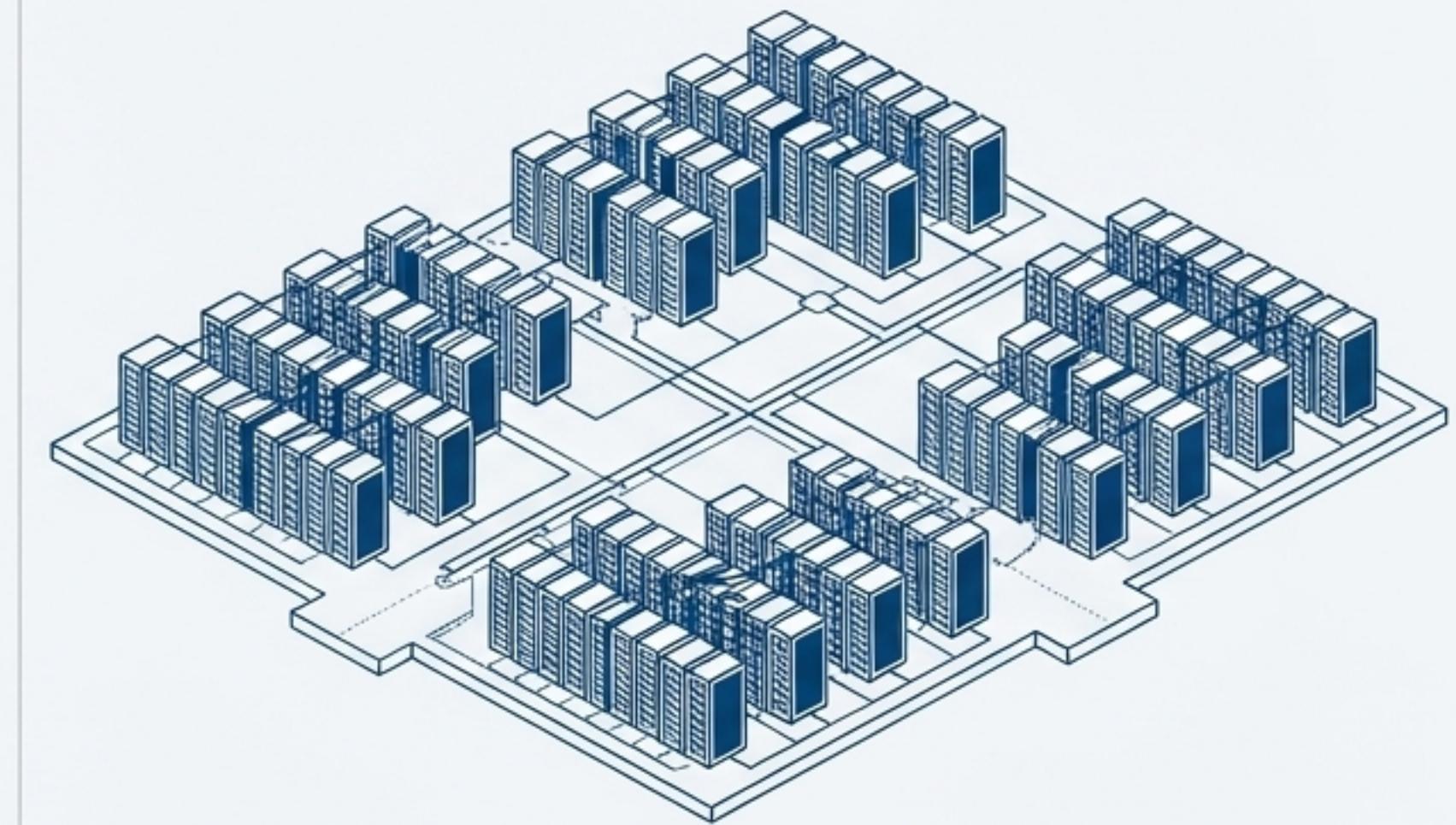
**Model Gęsty**

**Problem z pamięcią:** Pomimo efektywności obliczeniowej (niskie (niskie FLOPs), załadowanie modelu z **1.2 bilionem parametrów** do pamięci wymaga ogromnej i kosztownej infrastruktury sprzętowej.



**Ograniczenie dostępności:** Jak przyznają autorzy, "wymaga to większej liczby urządzeń", co "ogranicza dostępność zasobów i zwiększa koszty serwowania, zwłaszcza przy niskim natężeniu ruchu".

**GLaM**



Analogia: Model gęsty jest jak gruba, specjalistyczna książka, którą można zmieścić na jednej półce serwerowej. **GLaM to cała biblioteka narodowa – oferuje nieporównywalnie więcej wiedzy, ale wymaga budynku wielkości pałacu (klastra TPU), by ją pomieścić.**

# Świt nowej ery: Przyszłość AI jest rzadka (sparse) i inteligentna

GLaM to nie tylko nowy model – to dowód na istnienie lepszej, bardziej zrównoważonej drogi skalowania.



## Zmiana paradygmatu

Zmiana paradygmatu z "większy znaczy lepszy" na "**mądrzejszy znaczy lepszy**".

## Kluczowe wnioski:

- Architektura MoE:** Modułowi, wyspecjalizowani eksperci deklasują monolityczne modele pod względem wydajności i jakości.
- Inteligentne Bramkowanie:** Dynamiczna alokacja zasobów obliczeniowych jest kluczem do efektywności na ogromną skalę.
- Jakość ponad Ilość:** Sukces przyszłych modeli będzie zależeć od jakości danych treningowych, a nie tylko ich objętości.

> "Sparsity is one of the most promising directions to achieve high-quality NLP models while saving energy costs. MoE should therefore be considered as a strong candidate for future scaling."