

# DeepSeek V3: Nowa definicja wydajności w AI



Rzuca wyzwanie modelom-liderom, takim jak GPT-4O, osiągając jednocześnie bezprecedensową wydajność.



Kluczem nie jest budowanie coraz większych modeli, lecz bezwzględna optymalizacja na każdym poziomie.

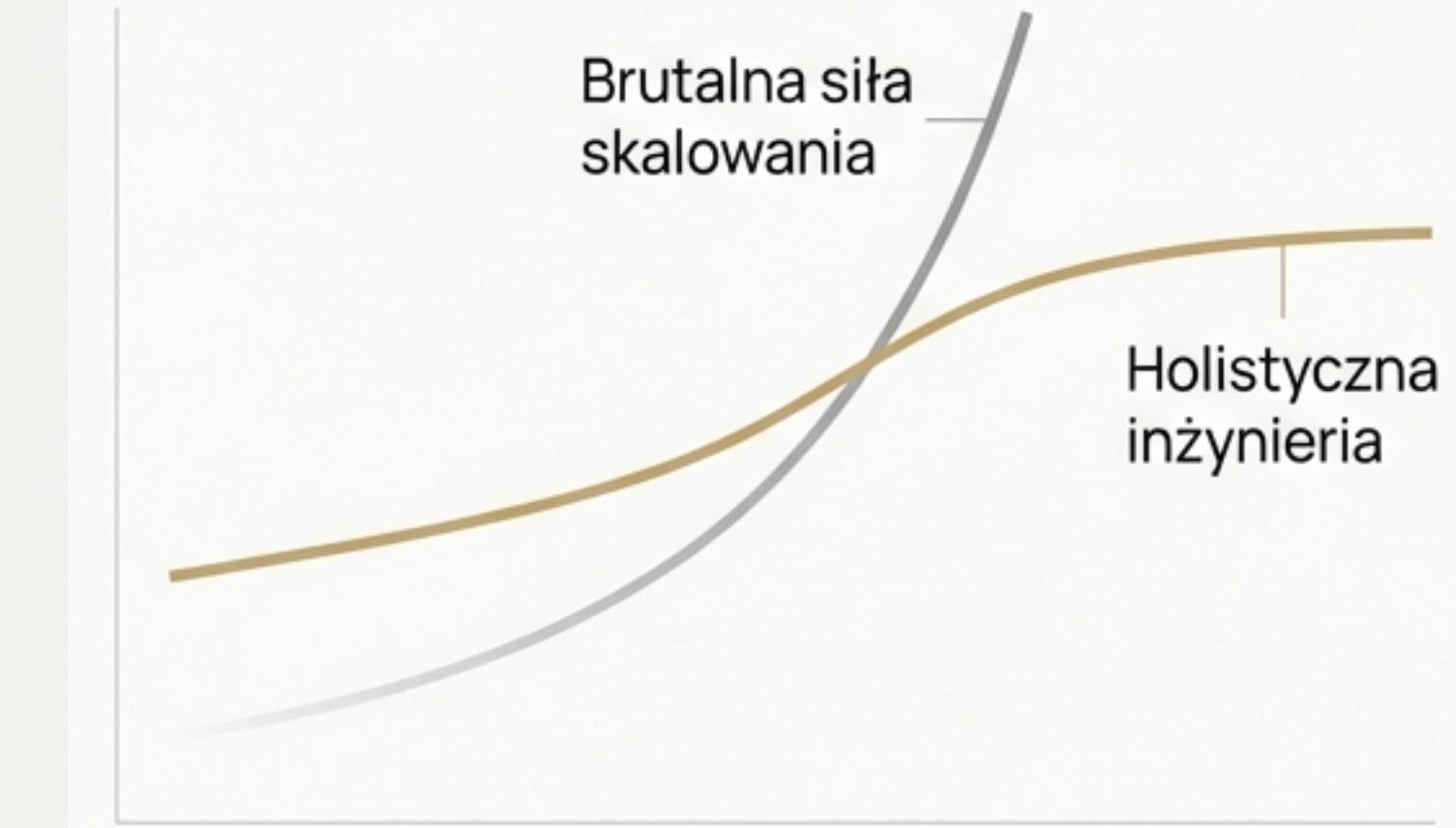


Holistyczne podejście inżynieryjne obejmujące architekturę, oprogramowanie i sugestie dotyczące sprzętu.



Reprezentuje zmianę paradymatu z „**większy znaczy lepszy**” na „**inteligentniejsza inżynieria**”.

## PARADYGMAT WYDAJNOŚCI



## KOSZT TRENINGU

**~\$5.6M**

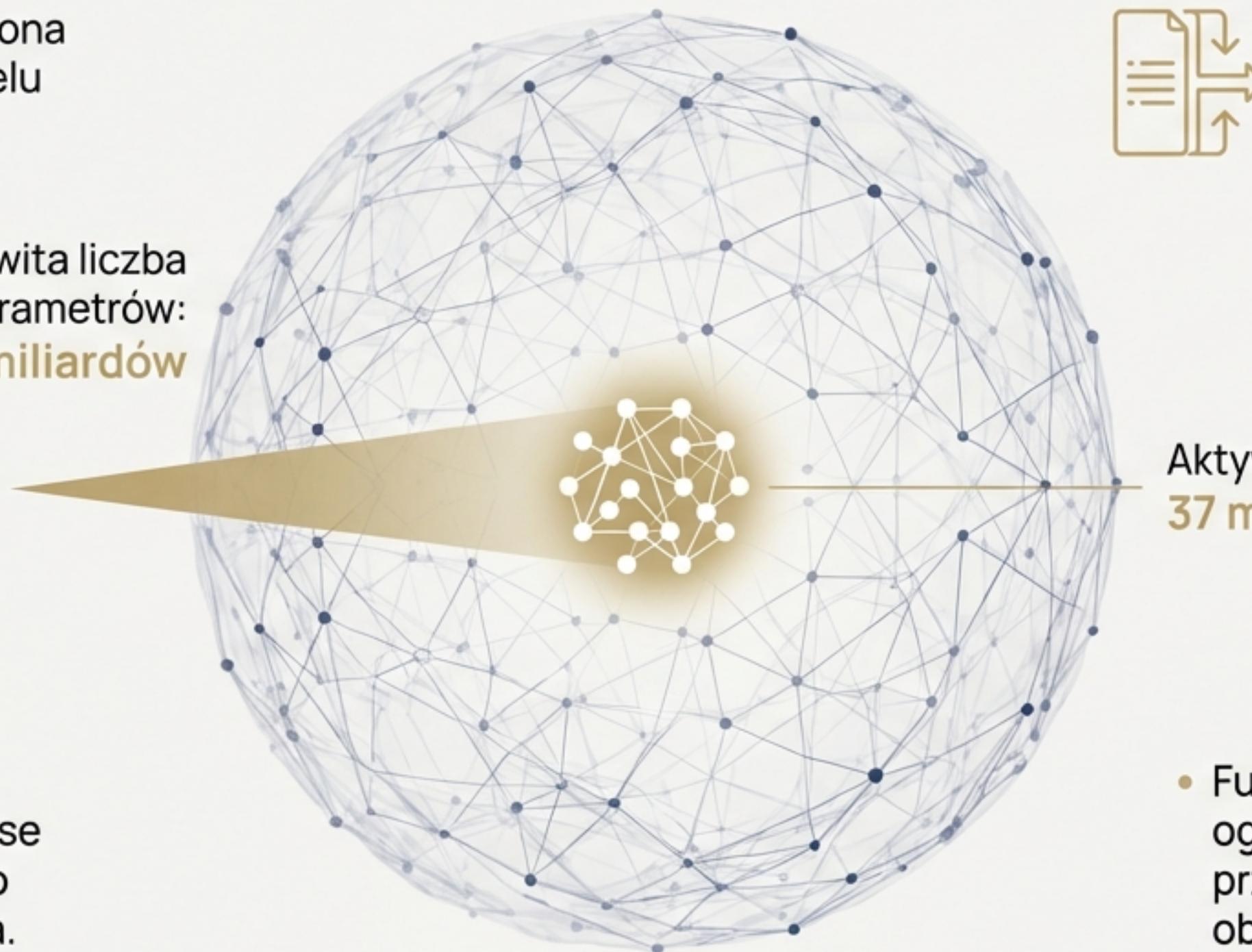
(2,788 miliona GPU-godzin na H800,  
ułamek kosztów konkurencji)

# Fundament: Architektura Mixture of Experts (MoE)

- Architektura odziedziczona i udoskonalona po modelu DeepSeek V2.

Całkowita liczba parametrów:  
**671 miliardów**

Token



- Rzadka aktywacja (sparse activation) jako klucz do opłacalnego skalowania.

Mechanizm **Multi-Head Latent Attention (MLA)** zapewnia wydajne wnioskowanie poprzez kompresję kluczy i wartości (KV cache).

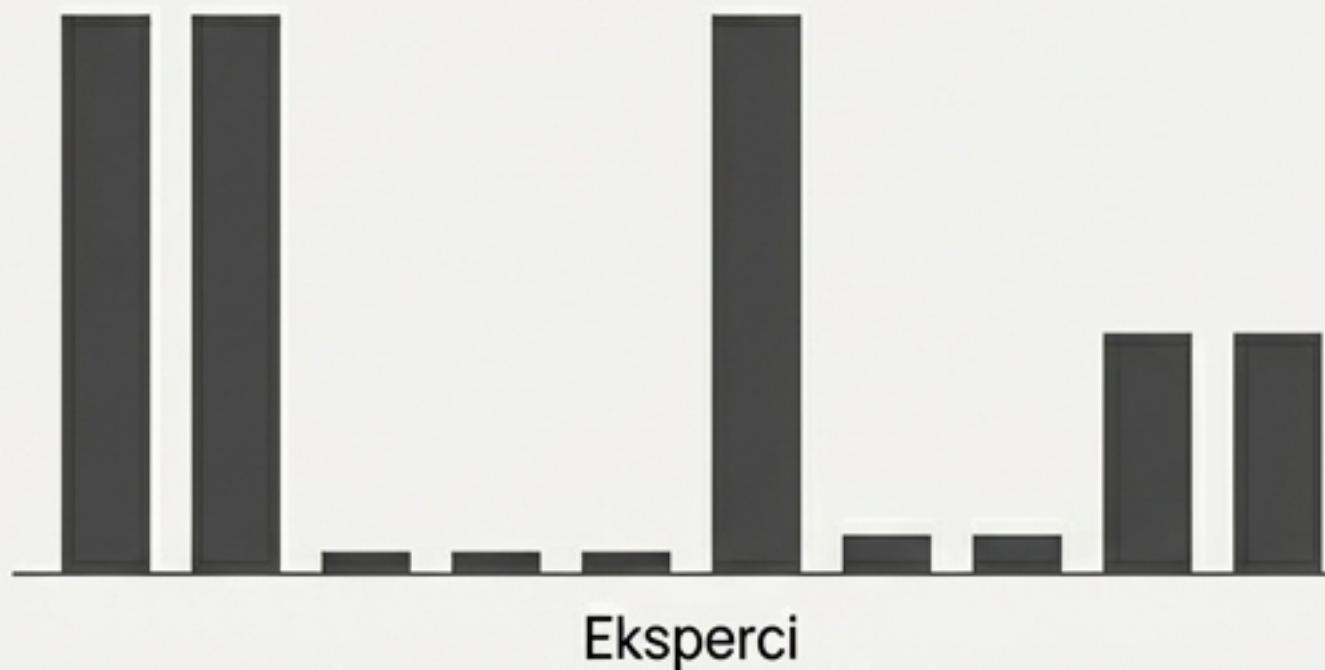
- Fundament MoE umożliwia ogólną liczbę parametrów przy zachowaniu efektywności obliczeniowej.

# Przełom #1: Równoważenie obciążenia bez pomocniczej funkcji straty (Auxiliary Loss-Free)



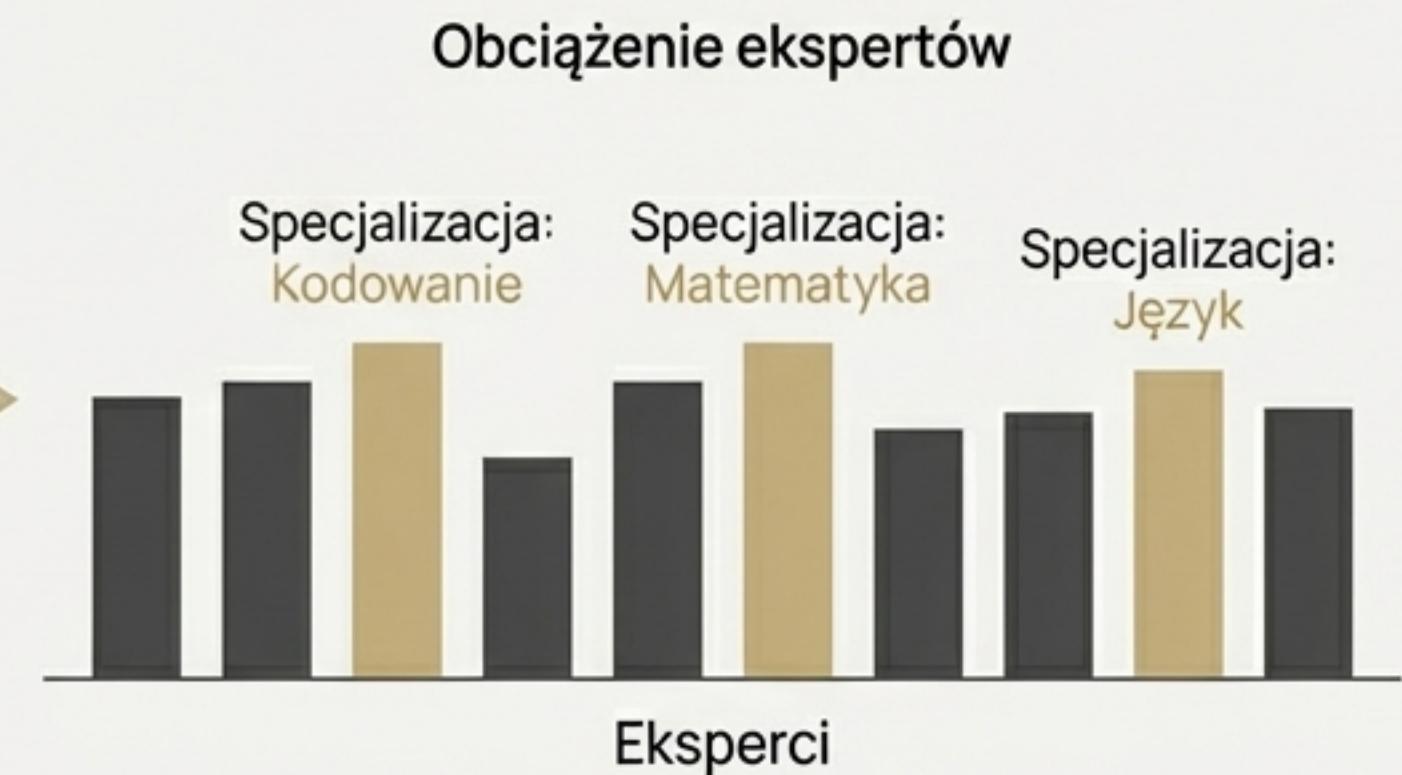
## Tradycyjne MoE: Zapaść routingu

Model faworyzuje nielicznych ekspertów, pozostawiając innych bezczynnymi. Standardowe rozwiązanie (Auxiliary Loss) wymusza równowagę, ale szkodzi wydajności.



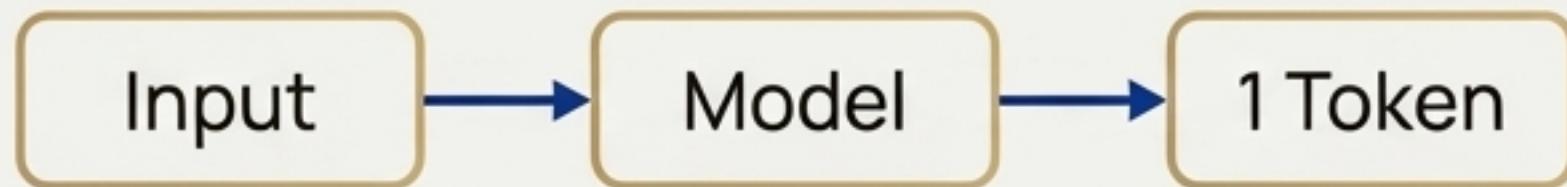
## Innowacja DeepSeek V3

Dynamiczny składnik biasu ( $b_i$ ) delikatnie dostosowuje preferencje routingu, pozwalając ekspertom na głębszą, naturalną specjalizację. Jest to pierwszy model na taką skalę, który całkowicie eliminuje pomocniczą funkcję straty.



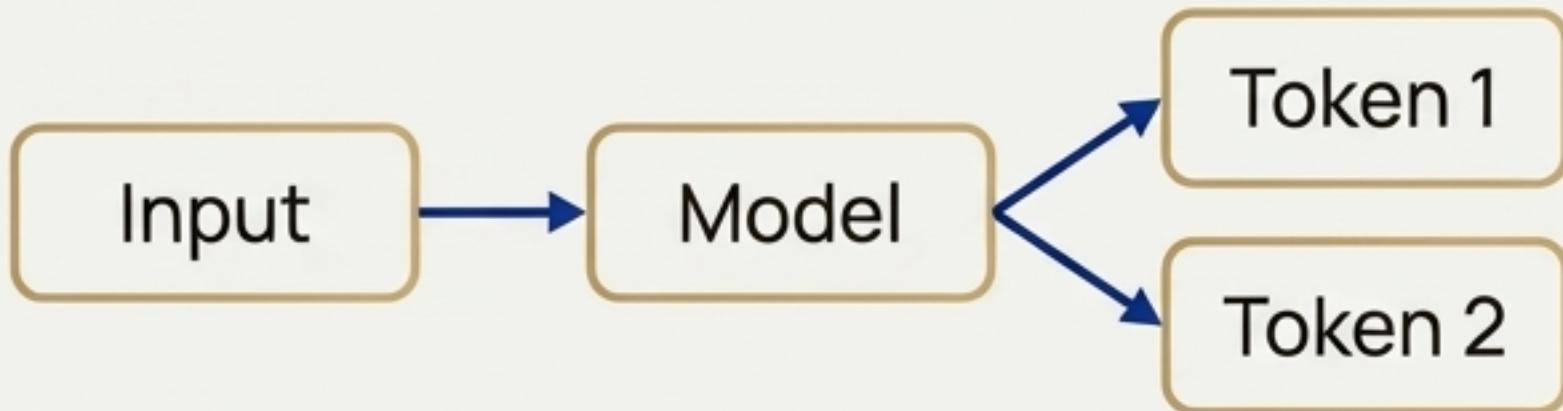
## Przełom #2: Predykcja wielu tokenów (Multi-Token Prediction)

### Standardowy trening



Przewidywanie jednego, kolejnego tokena.

### Innowacja DeepSeek V3



Przewidywanie wielu tokenów naraz (w tym przypadku dwóch) dla gęstszeego sygnału treningowego.

**Rezultat: 1,8x przyspieszenie generowania odpowiedzi (TPS).**

Ta sama technika umożliwia **dekodowanie spekulatywne** podczas wnioskowania. Model „myśli” kilka kroków naprzód.

# Trening FP8 z kwantyzacją drobnoziarnistą (Fine-Grained Quantization)

- **FP8 vs BF16:** Teoretycznie 2x szybszy i 2x mniejsze zużycie pamięci.
- **Wyzwanie:** 8-bitowa precyza grozi znaczną utratą informacji, zwłaszcza w przypadku wartości odstających (outliers).
- **Rozwiązanie:** Kwantyzacja drobnoziarnista. Macierze są dzielone na bloki, a każdy blok jest skalowany niezależnie, co pozwala lepiej zarządzać wartościami odstającymi.
- **Optymalizacja na poziomie sprzętowym:** Zmodyfikowana akumulacja w rdzeniach Tensor Core; część operacji przeniesiona do rdzeni CUDA w celu uzyskania pełnej precyji FP32.

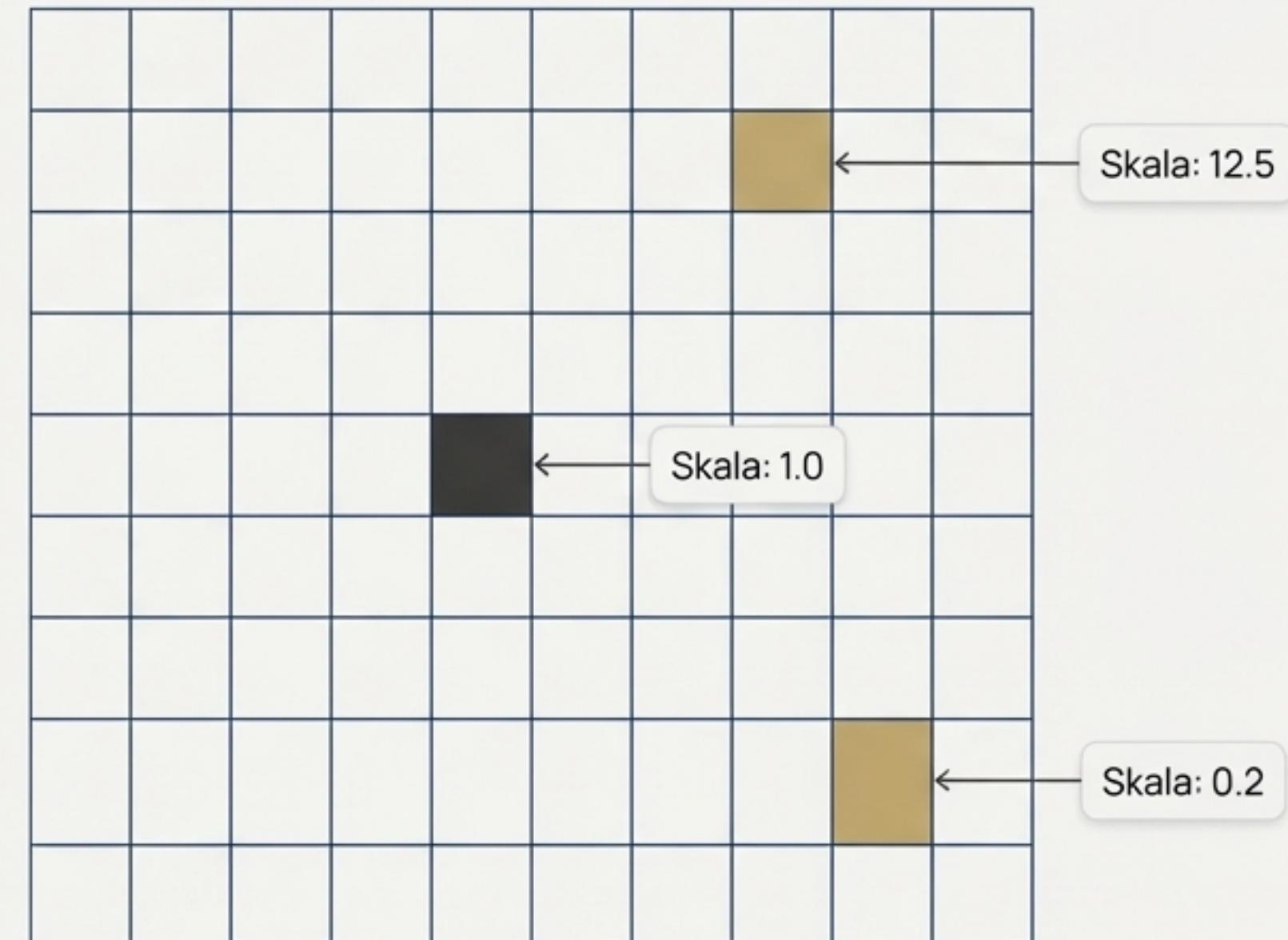
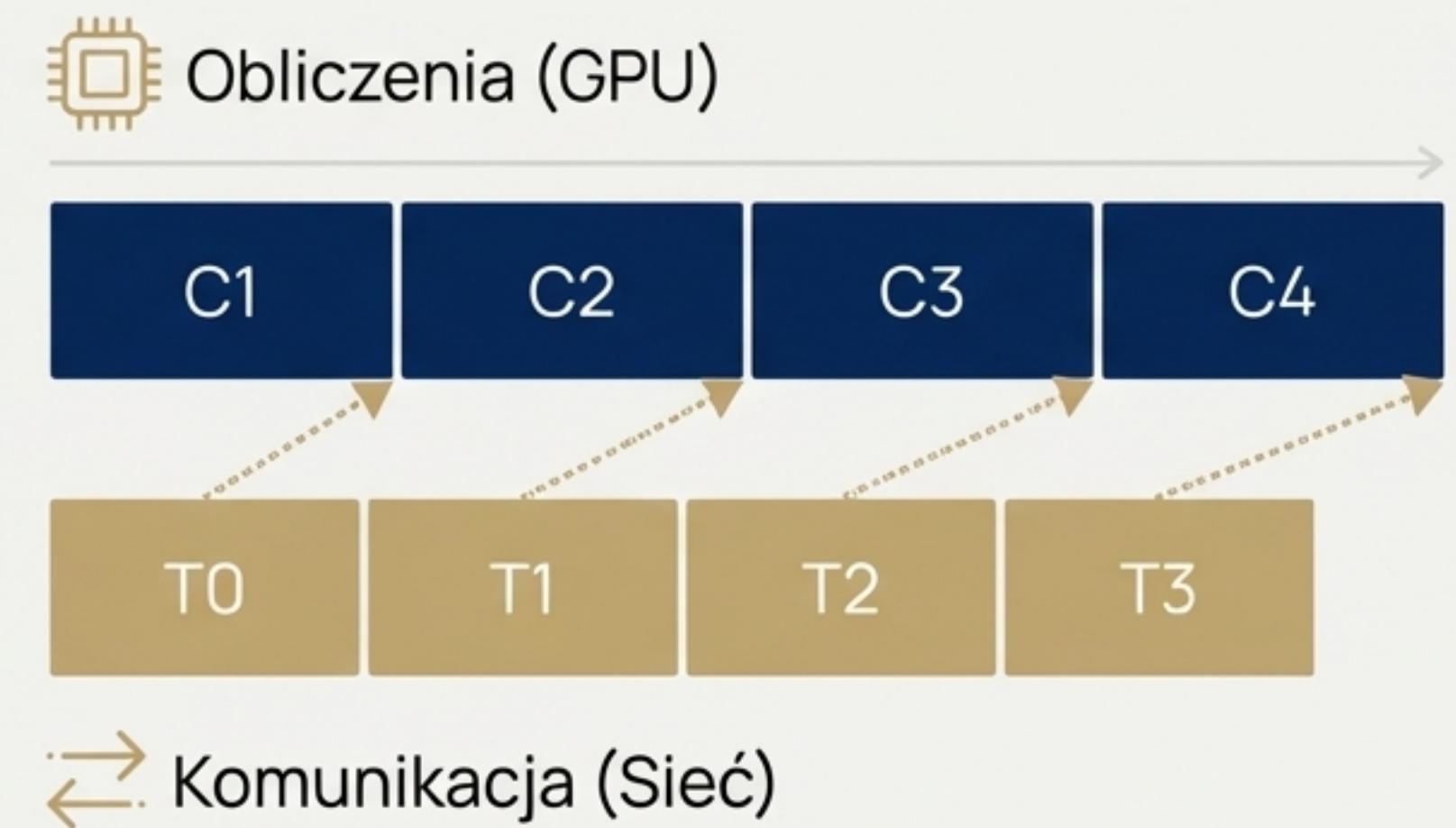


Diagram ilustrujący podział tensora na bloki z niezależnymi czynnikami skalowania w celu minimalizacji błędów kwantyzacji dla wartości odstających.

# Algorytm komunikacji DualPipe

- **Największe wąskie gardło MoE:**  
Komunikacja międzywęzłowa w dużych klastrach GPU.
- **Rozwiązanie: DualPipe.** Inteligentne harmonogramowanie, które nakłada obliczenia na komunikację (computation-communication overlap).
- **Efekt:** GPU nigdy nie czeka bezczynnie na dane. Prowadzi obliczenia na jednej porcji danych, podczas gdy poprzednia jest transferowana.
- Opóźnienie komunikacyjne jest niemal całkowicie ukryte.



# Ekstremalna optymalizacja pamięci



## Ponowne obliczanie (recomputation):

Zamiast przechowywać wyniki w VRAM, niektóre operacje (np. RMSNorm) są ponownie obliczane podczas propagacji wstecznej. Wymiana mocy obliczeniowej na pamięć.

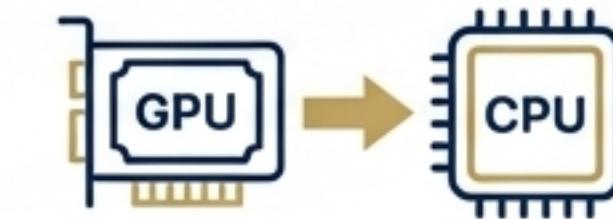


## Ograniczona pamięć VRAM



## Każdy megabajt ma znaczenie:

Te techniki umożliwiły pełny trening w ciągu **2,788 miliona** GPU-godzin przy całkowitym koszcie **~\$5.6M**.



## Parametry EMA w CPU:

Parametry Exponential Moving Average są przechowywane w pamięci CPU, a nie GPU, i aktualizowane asynchronicznie.

# Wyniki benchmarków w porównaniu z modelami-liderami

**Najmocniejszy model open-source** w momencie publikacji. Pokonuje LLaMA 3 405B, używając **11x mniej aktywnych parametrów**.

Benchmark Matematyczny: MATH 500

**90.2**

Prestiżowy Konkurs: AIME 2024

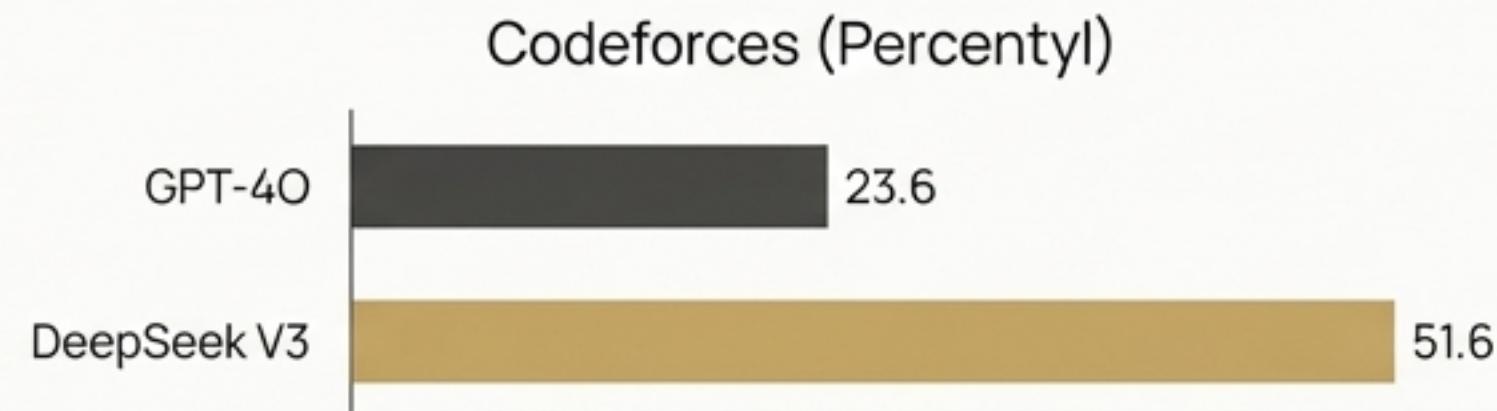
**39.2**

Ocena Konwersacyjna: Arena Hard

**>85%**

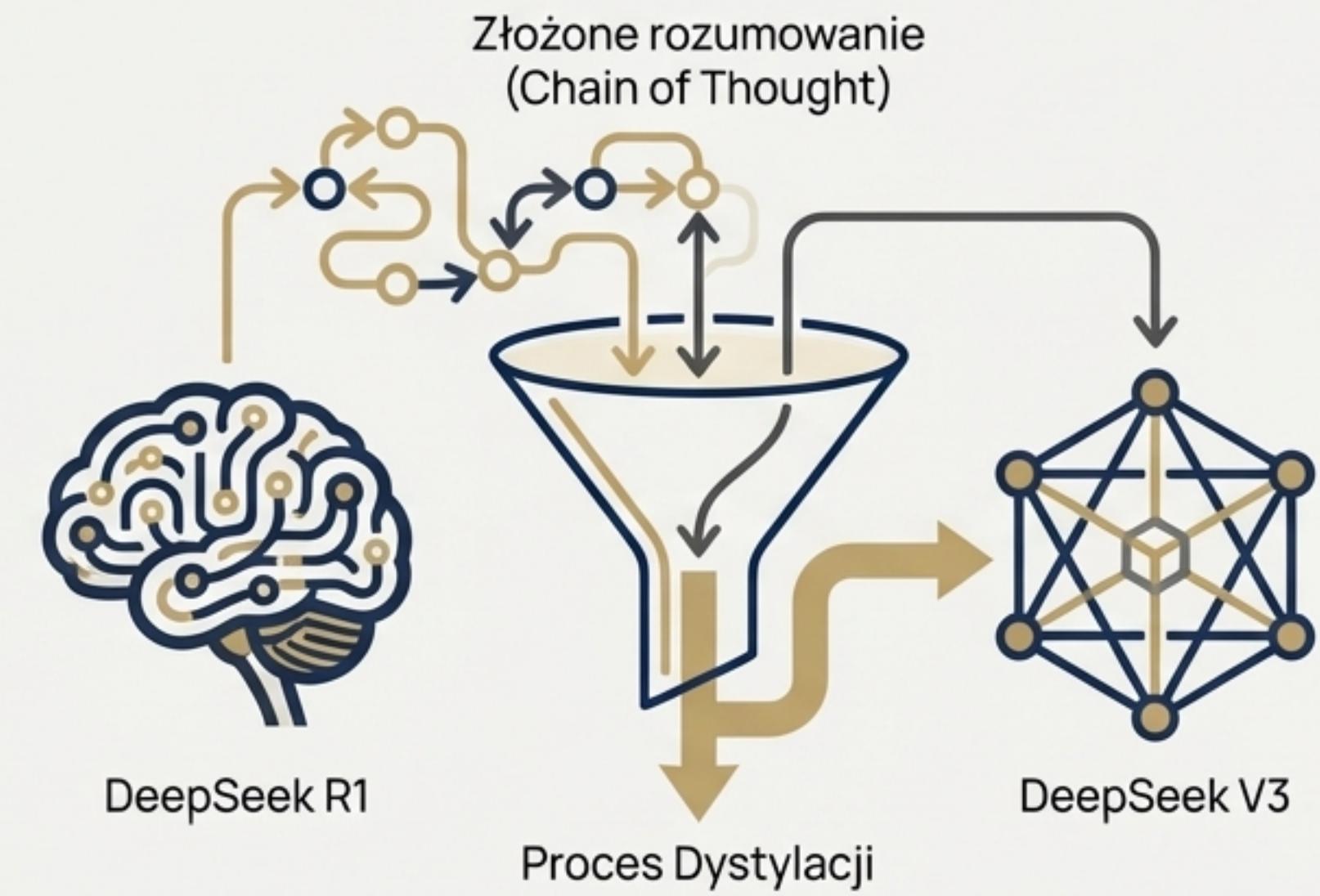
(Pierwszy model open-source na tym poziomie,  
dorównując Claude 3.5 Sonnet)

Benchmark Kodowania: Codeforces



# Destylowanie wiedzy z DeepSeek R1 w celu wzmacnienia rozumowania

- Technika post-treningowa wykorzystująca wyspecjalizowany model rozumowania, **DeepSeek R1**.
- R1 jest ekspertem w rozumowaniu krok po kroku (generowanie **Chain of Thought**).
- DeepSeek V3 uczy się od R1 wzorców rozumowania, weryfikacji i refleksji, a nie tylko odpowiedzi.
- Znacząco poprawia to wydajność w zadaniach wymagających rozumowania i czyni wyniki bardziej solidnymi.
- Odpowiedź na obawy dotyczące potencjalnego przeuczenia (overfitting) pod konkretne benchmarki.



# Ograniczenia i kierunki rozwoju

## Ograniczenia wdrożeniowe

- Wdrożenie wymaga minimum **32 GPU** (4 węzły dla etapu prefilling).
- Model nie nadaje się do wdrożeń na pojedynczym serwerze lub w małej skali.

## Wizja i sugestie sprzętowe

Plany na przyszłość: dążenie do nieskończonej długości kontekstu i przełamywanie ograniczeń architektury Transformer.

Wizja: **Współprojektowanie oprogramowania i sprzętu (co-design)** jako przyszłość rozwoju AI.



# **Przyszłość AI to nie tylko kod.**

**To holistyczna inżynieria, w której algorytmy, oprogramowanie i sprzęt są projektowane jako jedna, zoptymalizowana całość.**

DeepSeek V3 jest pierwszym arcydziełem tej nowej ery.