

# LLaMA 3: Perfekcja w Niespotykanej Skali

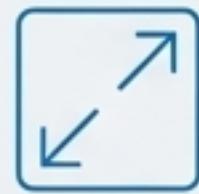
∞ Meta



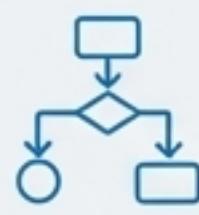
**405 miliardów parametrów** w architekturze Dense Transformer.



**15 bilionów tokenów** (wzrost z 1.8T w LLaMA 2).



**128 000 tokenów.**



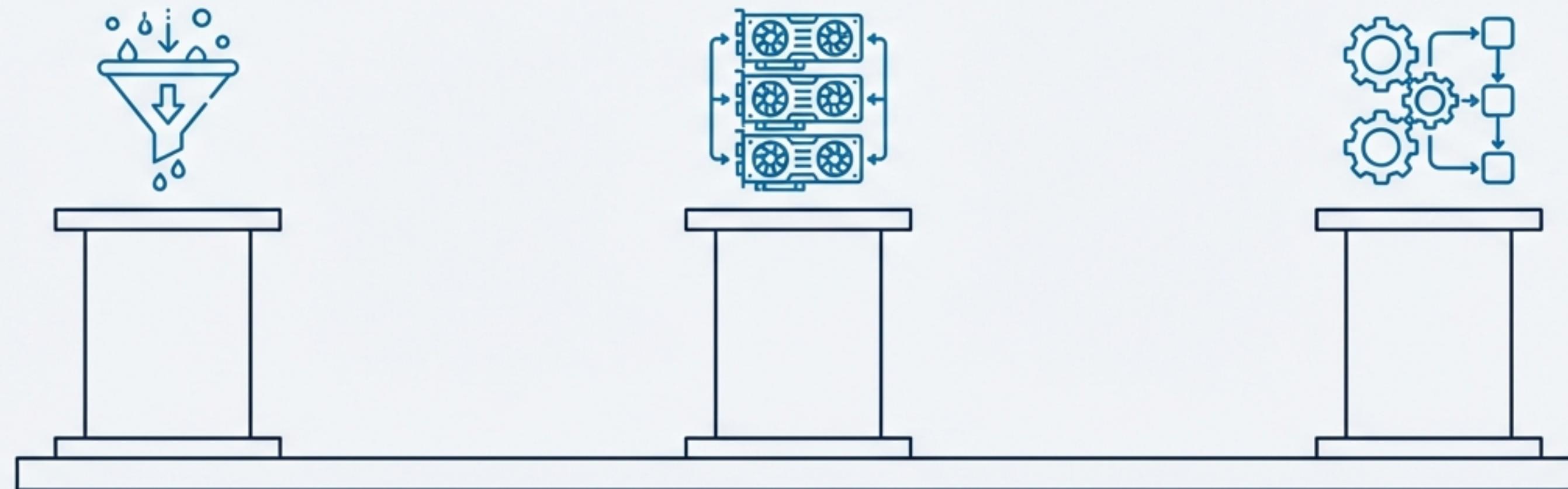
**Sprawdzona architektura:** Celowe zastosowanie gęstej architektury (Dense) zamiast Mixture of Experts (MoE) w celu maksymalizacji stabilności i dopracowania znanych metod.



**Zaprojektowany od podstaw** z myślą o kodowaniu, rozumowaniu i wielojęzyczności.

# Trzy Filary Sukcesu: Strategia Stojąca za LLaMA 3

**Filozofia projektu:** Ewolucja i udoskonalanie sprawdzonych metod jest skuteczniejsze niż pogon za nowością. Sukces nie wymagał rewolucji architektonicznej, lecz perfekcyjnego dopracowania standardowego modelu Transformer.



**Jakość Danych:** Obsesyjna dbałość o selekcję, filtrowanie i unikalność danych.

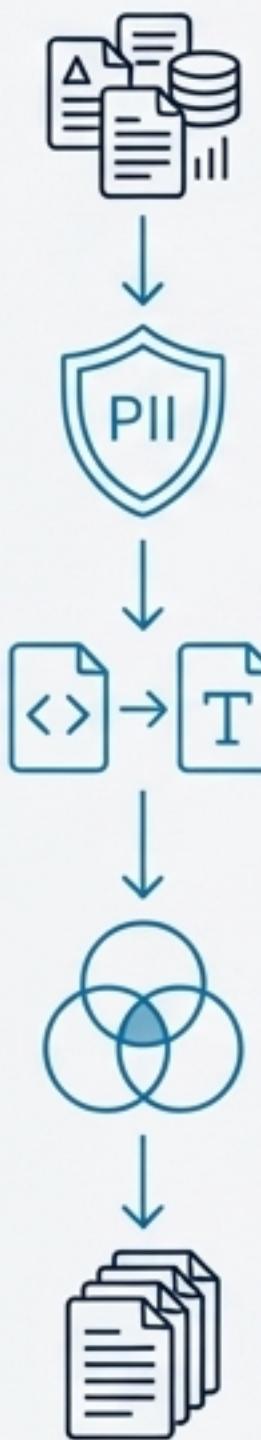
**Skala:** Wykorzystanie niemal 50x większej mocy obliczeniowej niż w przypadku największego modelu LLaMA 2.

**Zarządzanie Złożonością:** Inżynieryjna i organizacyjna doskonałość w zarządzaniu projektem o bezprecedensowej skali.

# Proces Przygotowania Danych: Rygorystyczne Filtrowanie na Niespotykaną Skalę

**Cel:** Maksymalna unikalność i jakość każdej informacji w korpusie treningowym.

**Wieloetapowy proces filtrowania:**



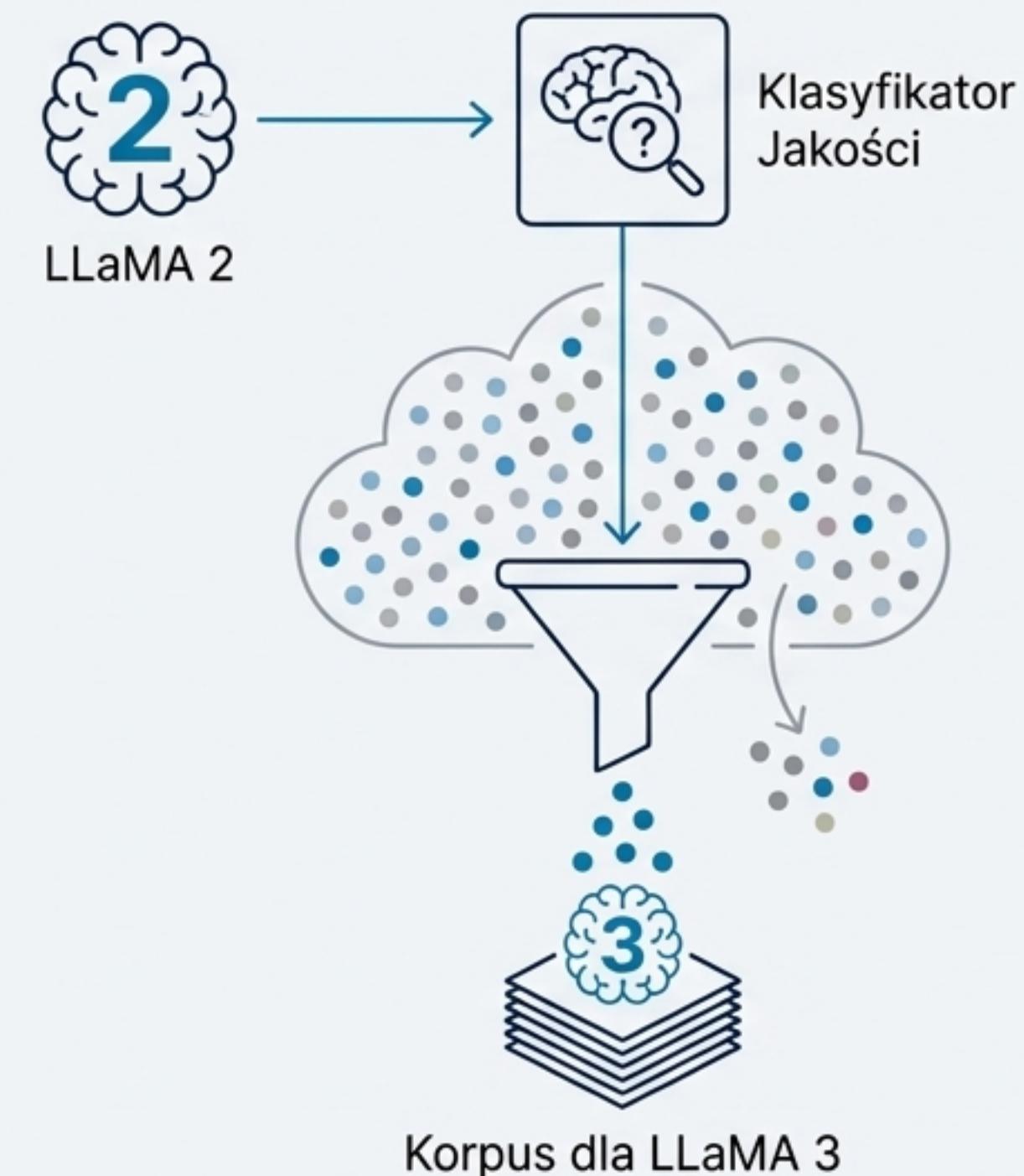
**Filtrowanie bezpieczeństwa i PII:** Usunięcie całych domen zawierających dane osobowe, treści dla dorosłych lub uznane za szkodliwe.

**Ekstrakcja tekstu:** Zastosowanie własnego, zoptymalizowanego parsera HTML do czystej ekstrakcji treści, z zachowaniem struktury kodu i formuł matematycznych.

**Agresywna deduplikacja:** Zastosowano kilka poziomów deduplikacji, aby wyeliminować powtórzenia:  
- Na poziomie URL  
- Na poziomie dokumentu (MinHash)  
- Na poziomie linii (usunięcie linii pojawiających się >6 razy)

# LLaMA 2 jako Sędzia Jakości: Poprzednik Buduje Fundamenty dla Następcy

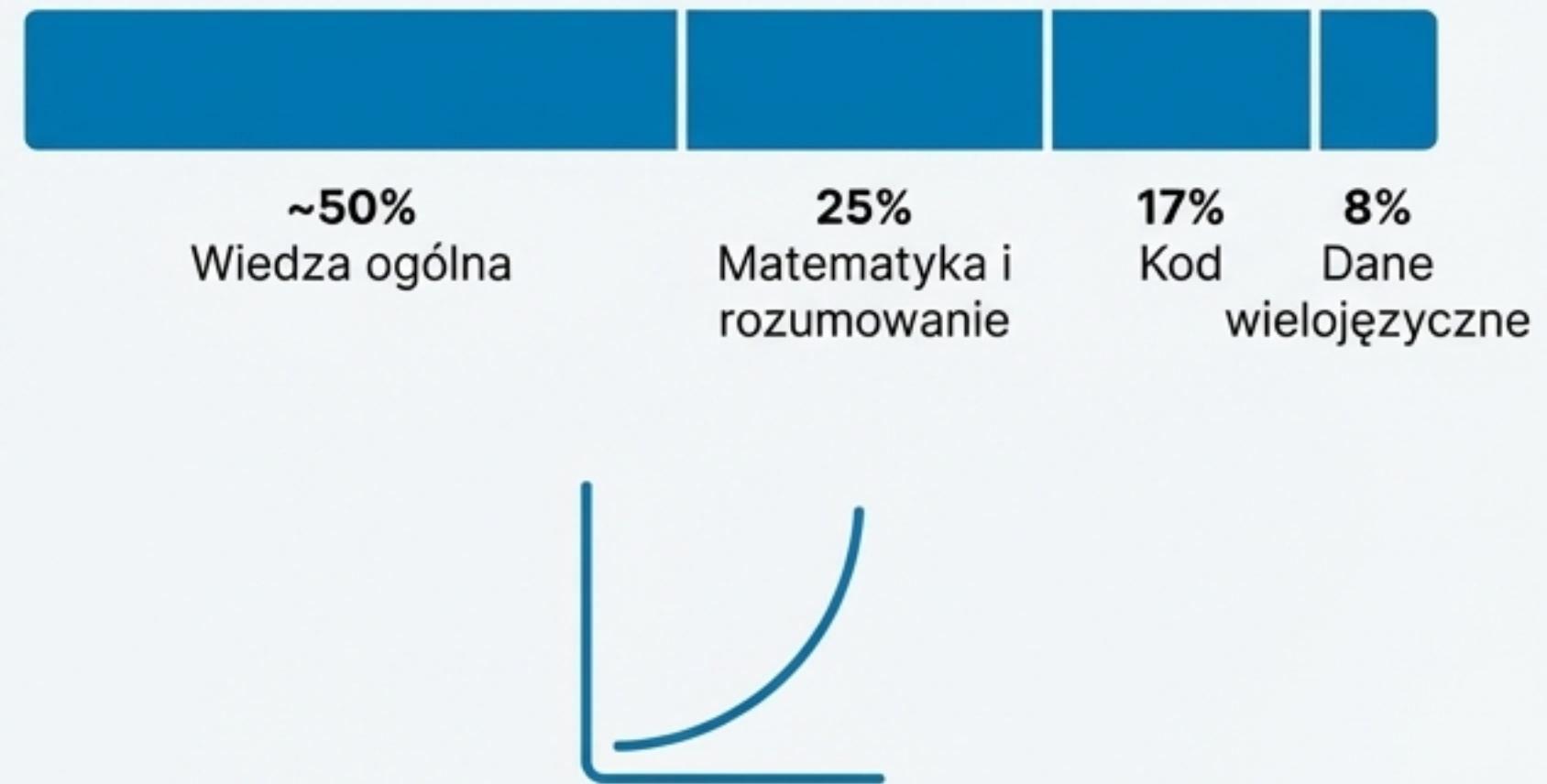
- **Podejście samodoskonalące (self-bootstrapping):**  
Wykorzystanie poprzednich modeli LLaMA 2 do automatycznej klasyfikacji jakości danych na masową skalę.
- **Proces:**
  1. **Trening klasyfikatora:** LLaMA 2 została wytrenowana do rozpoznawania tekstów wysokiej jakości (np. takich, które mogłyby być cytowane w Wikipedii).
  2. **Automatyczna ocena:** Wytrenowany klasyfikator (w formie wydajnego modelu DistilRoberta) oceniał każdy dokument w ogromnym zbiorze danych webowych.
  3. **Selekcja:** Do korpusu treningowego LLaMA 3 trafiały tylko dane o najwyższej ocenie jakości.
- **Metafora:** LLaMA 2 działała jak 'kurator' przygotowujący bibliotekę dla swojego intelligentniejszego następcy.



# Skład Danych i Prawa Skalowania: Inżynieria "Diety Treningowej"

- **Prawa skalowania (Scaling Laws):**  
Proporcje danych nie były przypadkowe.  
Zostały wyznaczone na podstawie eksperymentów z prawami skalowania, które pozwoliły przewidzieć wpływ składu 'diety' na ostateczne zdolności modelu, jeszcze przed rozpoczęciem kosztownego treningu.
- **Kluczowy wniosek:** Inżynieria danych staje się ważniejsza niż inżynieria architektury modelu.

Starannie skomponowaną mieszankę danych



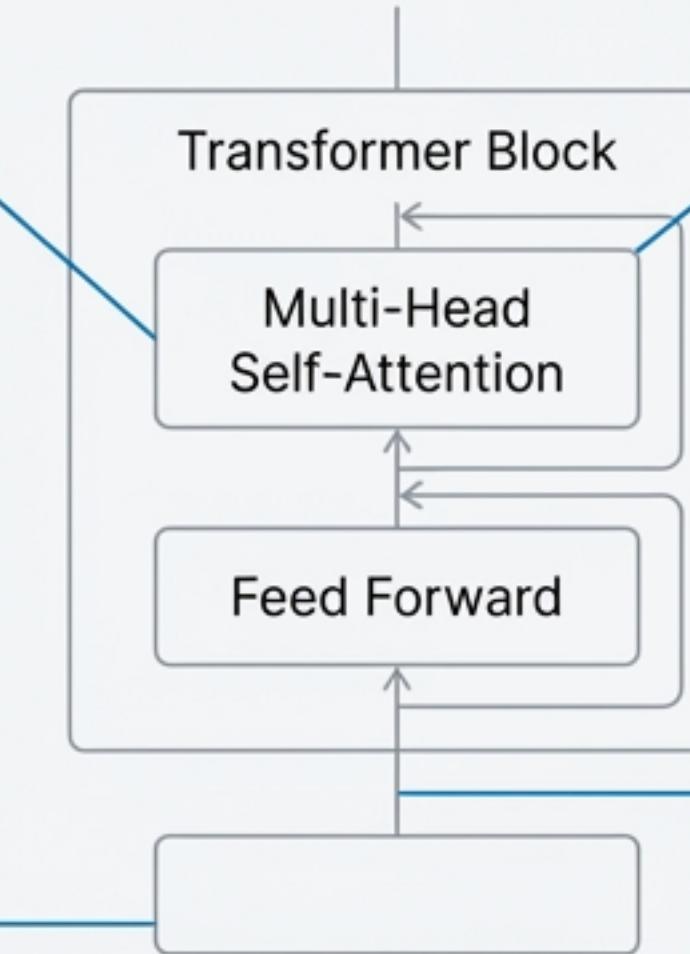
# Ulepszenia Architektury: Ciche Optymalizacje o Dużym Wpływie

## Grouped Query Attention (GQA)

Zastosowanie 8 głowic klucz-wartość znacząco przyspiesza generowanie odpowiedzi i zmniejsza zużycie pamięci podczas dekodowania.

## Większy Słownik (Vocabulary)

Rozszerzenie do 128K tokenów dla lepszej kompresji i wsparcia języków innych niż angielski.



## Specjalna Maska Uwagi (Attention Mask)

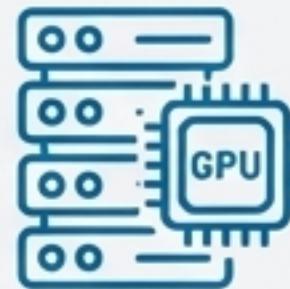
Zapobiega 'mieszaniu się' informacji między oddzielnymi dokumentami w ramach jednego długiego zapytania, utrzymując spójność przetwarzania w długim kontekście.

## Dostosowany RoPE

Zwiększenie bazowej częstotliwości RoPE do 500 000 w celu lepszego wsparcia dla długich kontekstów.

# Trening w Ekstremalnej Skali: Wyzwania Inżynierijne

16 000



procesorów graficznych H100 w klastrach treningowych.

419



nieoczekiwanych przerw w ciągu 54 dni treningu (głównie awarie sprzętu).

90%

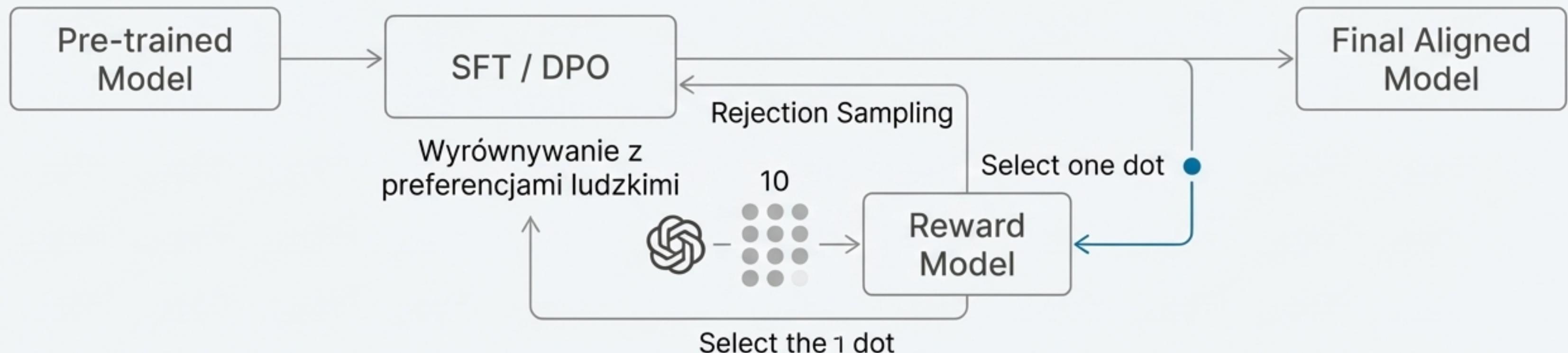


efektywnego czasu treningu, dzięki zaawansowanej automatyzacji.



**Ciekawostka:** Wrażliwość systemu na czynniki środowiskowe była tak duża, że obserwowano **dzieenne wahania wydajności o 1-2%** spowodowane zmianami temperatury otoczenia w centrum danych.

# Techniki Post-treningowe: Kształtowanie Zdolności i Wyrównywanie



**Kodowanie:** Trening na danych syntetycznych z informacją zwrotną z wykonania kodu (nauka na praktycznych błędach).



**Matematyka:** Trening na rozwiązaniach krok po kroku z mechanizmami autoweryfikacji.



**Długi kontekst (128K):** Wystarczyło zaledwie **0.1%** syntetycznych przykładów z długim kontekstem – mikroskopijna interwencja, ogromne rezultaty.

# Wyniki Benchmarków: Dwa Oblicza Wydajności

## Lider w specjalistycznych zadaniach

 GSM-8K (rozumowanie matematyczne): **96.8%**

 HumanEval Plus (kodowanie): **82.3%**

## Zniuansowany obraz w ocenach ludzkich

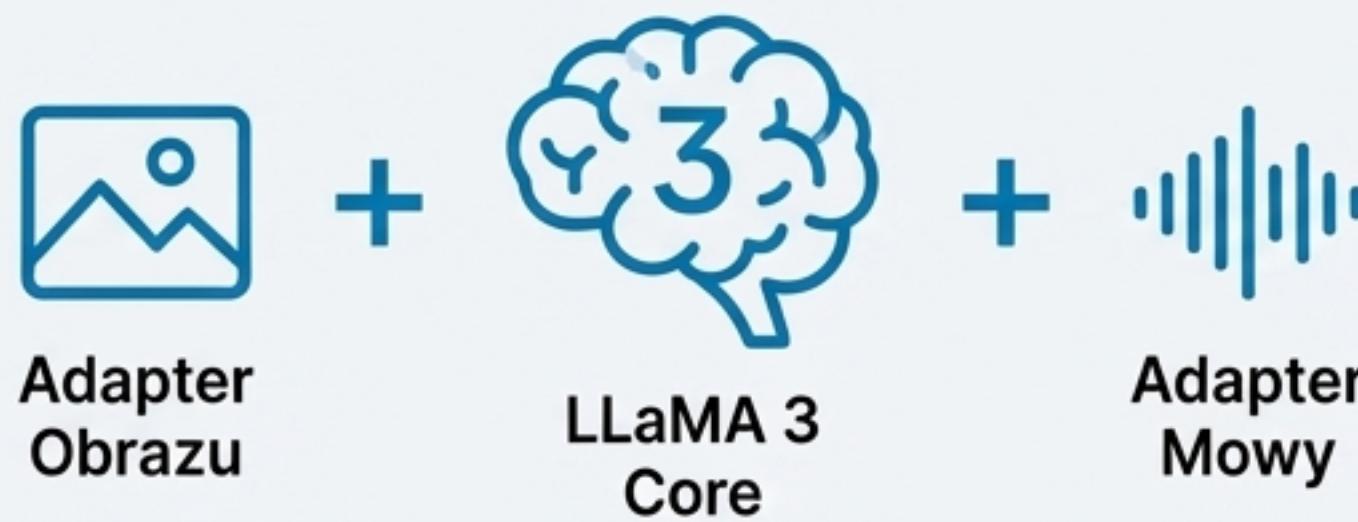


- Wydajność LLaMA 3 405B jest **porównywalna (on par) z GPT-4** (wersja 0125).
- Wyniki są **mieszane w porównaniu z GPT-4o i Claude 3.5 Sonnet** w ogólnej ocenie 'pomocności'.

**Wniosek:** Era jednego, uniwersalnie najlepszego modelu, dobiera się do końca. Wybór zależy od konkretnego zastosowania.

# Multimodalność i Bezpieczeństwo: Rozszerzanie Horyzontów

**Multimodalność - podejście kompozycyjne**



Takie podejście **zachowuje pełną wydajność modelu tekstowego**, dodając nowe zdolności.

**LLaMA 3-V** przewyższa GPT-4V we wszystkich testowanych benchmarkach wizualnych.

**Bezpieczeństwo - podejście wielowarstwowe**



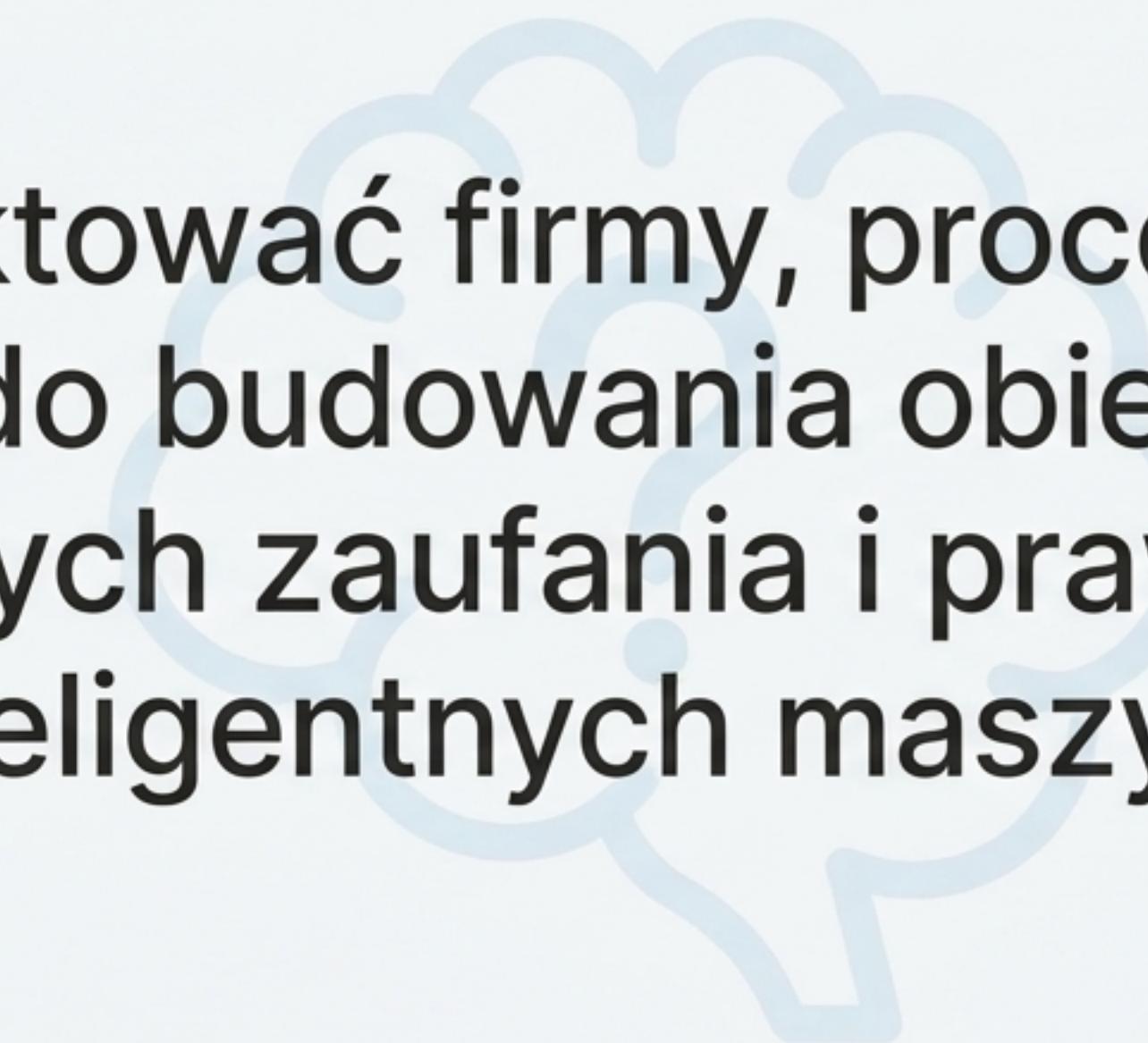
**Pre-training:** Agresywne filtrowanie danych

**Fine-tuning:** Dostrajanie pod kątem bezpieczeństwa

**System:** Klasyfikator LLaMA Guard 3

**Cel:** Równowaga między blokowaniem szkodliwych treści a unikaniem absurdalnych odmów przy nieszkodliwych zapytaniach.

# Pytanie na Przyszłość



Jak projektować firmy, procesy i zespoły  
zdolne do budowania obiektywnych,  
godnych zaufania i prawdziwie  
inteligentnych maszyn?

Być może to jest prawdziwe wyzwanie na następną dekadę.