



# Qwen 2.5: Efektywność ponad Skalą

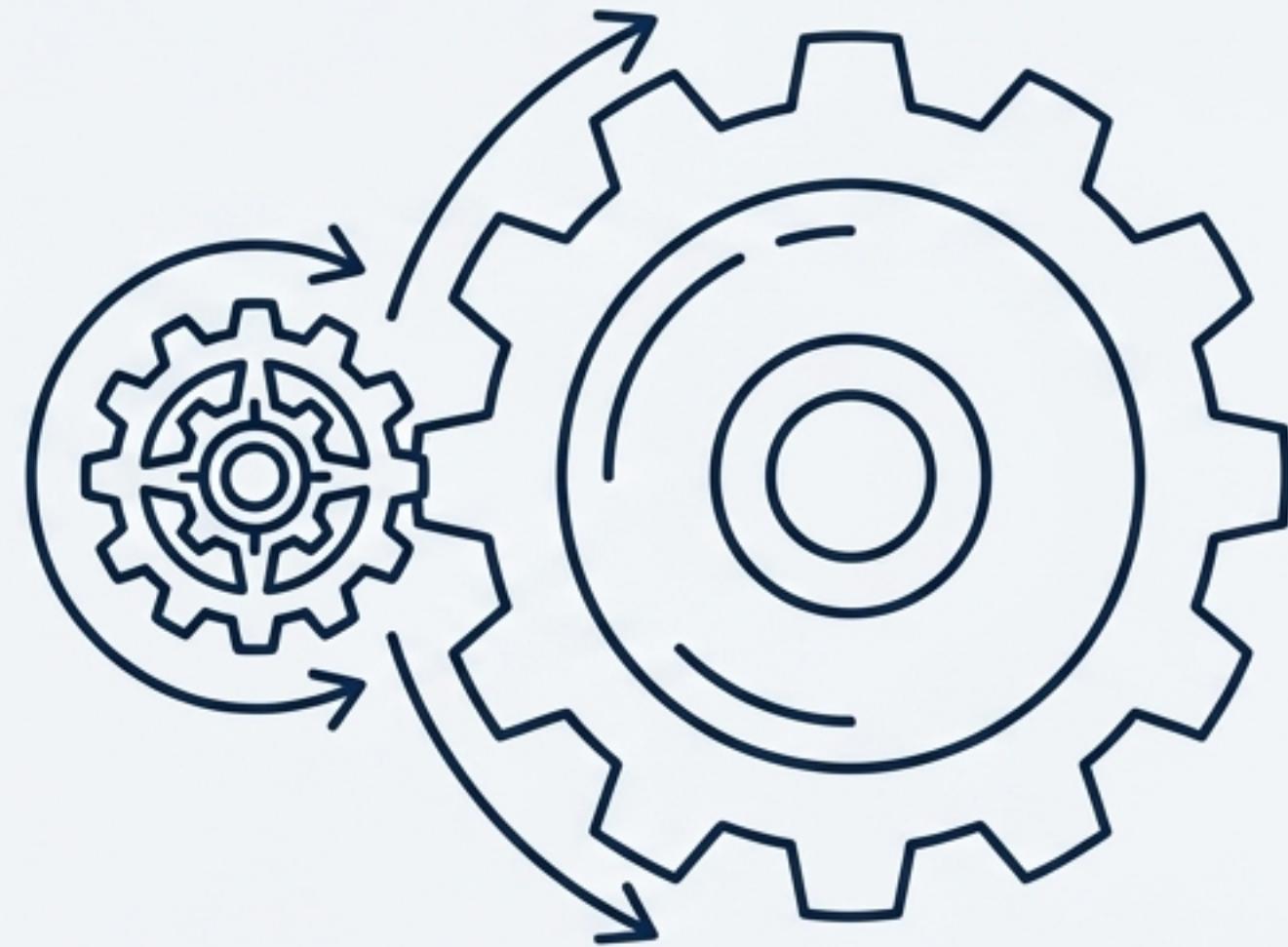
Qwen 2.5-72B-Instruct osiąga wydajność porównywalną z **5x większym** modelem Llama 3-405B-Instruct.

Zmiana paradygmatu w AI: od „większy znaczy lepszy” do „inteligentniejszy znaczy lepszy”. Kluczem jest zaawansowany projekt, a nie surowa moc obliczeniowa.

## Rodzina Modeli

**Open Weight:** Dostępna pełna gama modeli o otwartej wadze: 0.5B, 1.5B, 3B, 7B, 14B, 32B oraz 72B parametrów.

**Proprietary API:** Zastrzeżone, wysoce zoptymalizowane modele API oparte na architekturze MoE: Qwen 2.5-Turbo i Qwen 2.5-Plus.



# Architektura: Komitet Ekspertów (Mixture of Experts)

**Kluczowa Technologia:** Zastrzeżone modele (Qwen 2.5 Turbo/Plus) wykorzystują architekturę Mixture-of-Experts (MoE).

## Jak to działa?:

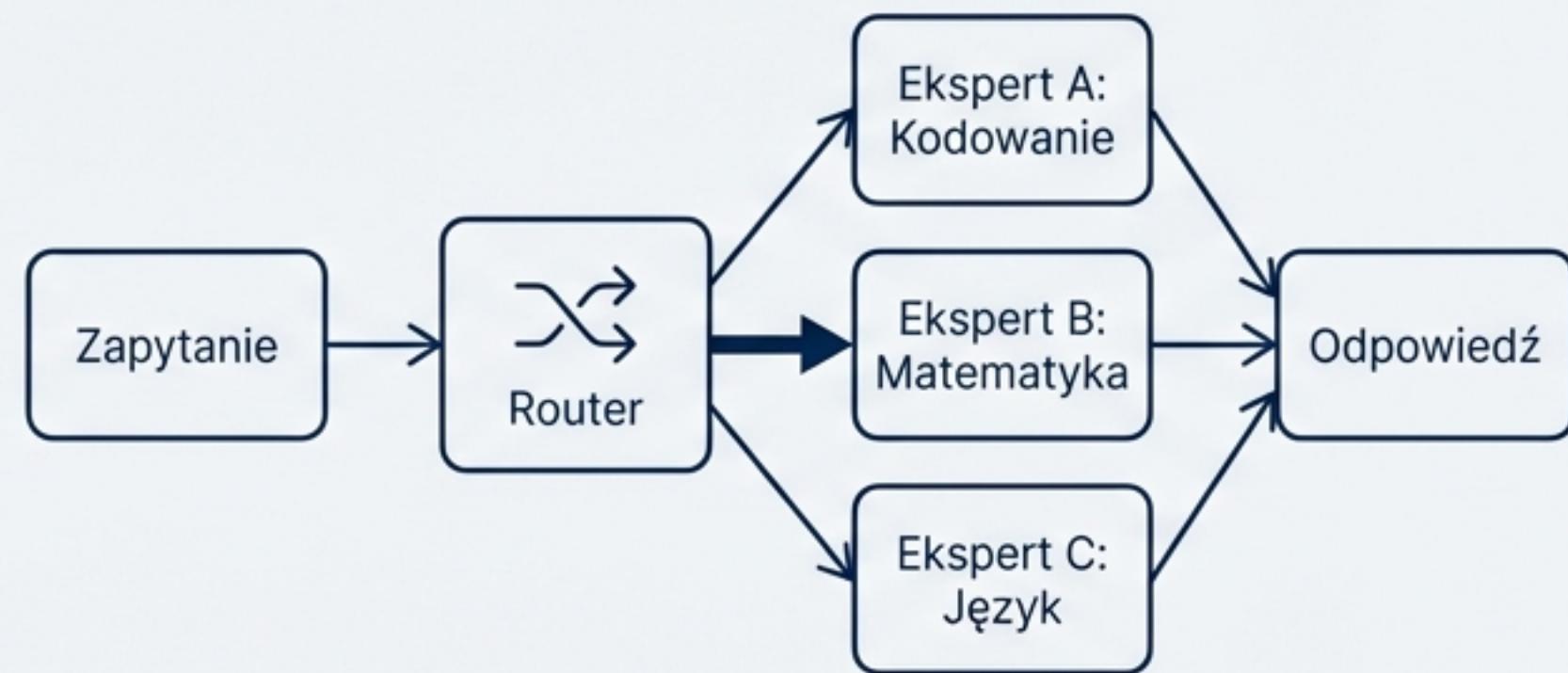
Zamiast jednego, monolitycznego mózgu, MoE działa jak komitet specjalistów. Zapytanie jest kierowane przez „router” do najbardziej odpowiednich ekspertów (np. od kodowania, matematyki, rozumienia języka).

## Główne Korzyści:

- **Wydajność:** W danym momencie aktywna jest tylko niewielka część parametrów modelu, co drastycznie redukuje koszty obliczeniowe.
- **Specjalizacja:** Umożliwia osiągnięcie wyższej jakości odpowiedzi dzięki wyspecjalizowanym pod-sieciom.

## Szczegóły techniczne:

Warstwy FFN (feed-forward network) zostały zastąpione przez warstwy MoE, które zawierają wielu ekspertów FFN i mechanizm routingu.



# Fundament: Rewolucja w Danych Treningowych

Skala: Dane pre-treningowe przeskalowano z **7 bilionów do 18 bilionów tokenów** (wzrost z Qwen 2).

**Jakość ponad ilość:** Kluczem nie jest surowa objętość, lecz inteligentna kuracja. Zastosowano „dietę bogatą w składniki odżywcze” zamiast karmienia modelu wszystkim z internetu.



Balansowanie Danych:

- **Redukcja:** Znacząco ograniczono dane z domen o niższej jakości, takich jak media społecznościowe i e-commerce.
- **Wzrost:** Zwiększoano udział wysokiej jakości danych z dziedzin nauki, technologii, badań akademickich, kodowania i matematyki.

# Proces: Inteligentne Filtrowanie i Dane Syntetyczne

Filtrowanie na Skalę: Wykorzystano poprzednie modele z serii **Qwen-Instruct** jako zaawansowane filtry jakości do wielowymiarowej analizy i oceny próbek treningowych.

Automatyzacja: Ręczna kuracja na taką skalę zajęłaby tysiące lat. Zautomatyzowany proces był kluczowy.

## Wzmocnienie Domenowe:

- Włączono dane treningowe ze specjalistycznych modeli Qwen2.5-Math i Qwen2.5-Coder.
- Wygenerowano nowe, wysokiej jakości dane syntetyczne, szczególnie w dziedzinach matematyki i kodowania, korzystając z modeli Qwen2-72B.



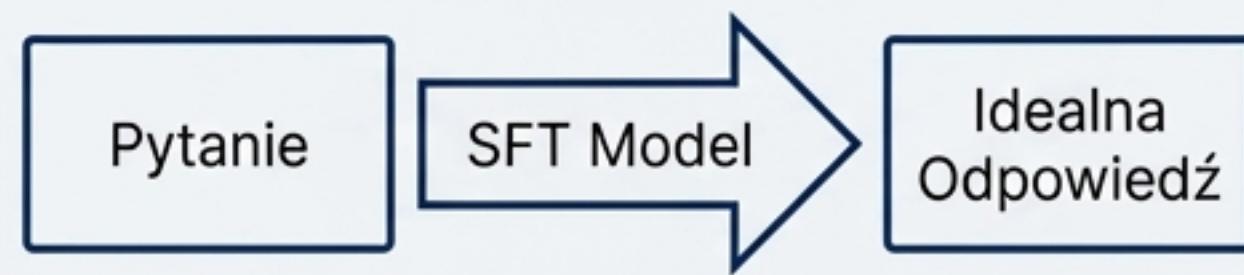
# Dostrajanie, Etap 1: Nauczanie z Wzorcowego Podręcznika (SFT)

**Proces:** Supervised Fine-Tuning (SFT) to pierwszy krok w dostosowywaniu modelu do preferencji człowieka.

**Skala:** Wykorzystano ogromny zbiór ponad **1 miliona** wysokiej jakości przykładów w formacie „Pytanie → Idealna Odpowiedź”.

## Cel:

- Celowe eliminowanie słabości zidentyfikowanych w poprzednich wersjach.
- Poprawa generowania spójnych, długich tekstów (do 8192 tokenów).
- Wzmocnienie rozumienia danych strukturalnych (np. tabele, JSON).



**Analogia:** Proces ten można porównać do dania modelowi obszernego, wzorcowego podręcznika, z którego uczy się idealnych odpowiedzi.

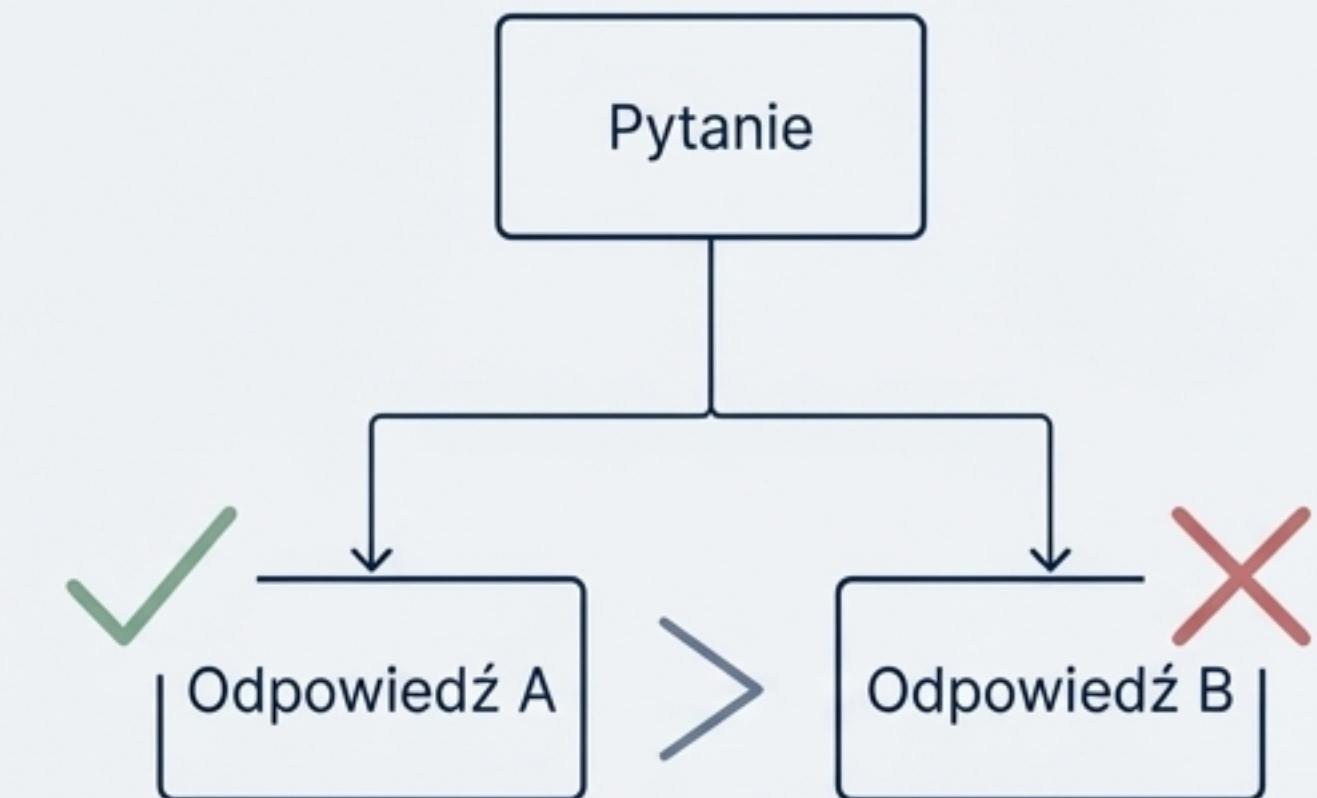
# Doskajanie, Etap 2: Nauka przez Porównanie (DPO)

**Technika:** Drugi etap to Offline Reinforcement Learning z użyciem techniki DPO (Direct Preference Optimization).

**Mechanizm:** Modelowi prezentowane są dwie odpowiedzi na to samo pytanie: jedna oznaczona jako „dobra” (pozytywny przykład), druga jako „zła” (negatywny przykład). Model uczy się systematycznie preferować lepszą odpowiedź.

**Zastosowanie:** Technika ta jest szczególnie skuteczna w domenach, gdzie istnieje obiektywna poprawność, takich jak:

- Matematyka
- Kodowanie
- Logiczne rozumowanie
- Ścisłe podążanie za instrukcjami



**Analogia:** Można to porównać do pracy z wymagającym nauczycielem, który ma precyzyjny klucz odpowiedzi i zawsze wskazuje błędy.

# Dostrajanie, Etap 3: Opanowanie Sztuki Konwersacji (GRPO)

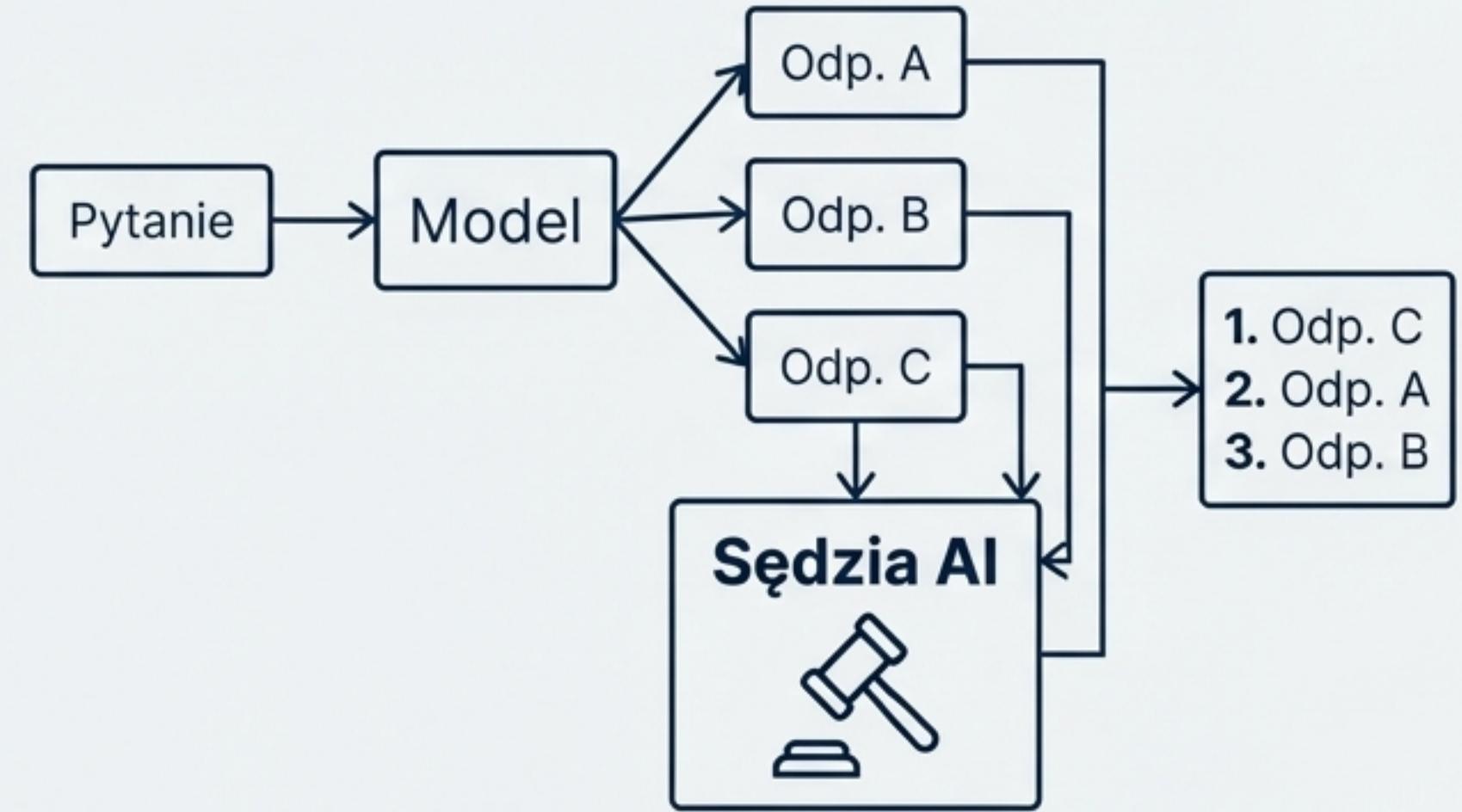
Technika: Ostatni etap to Online Reinforcement Learning z użyciem nowatorskiej techniki GRPO (Group Relative Policy Optimization).

Mechanizm: Zamiast prostego „dobry/zły”, model generuje grupę (np. 8) różnych odpowiedzi na jedno zapytanie. Następnie zaawansowany sędzia AI (Reward Model) ocenia i szereguje je według złożonych kryteriów.

## \*\*Kryteria Oceny\*\*:

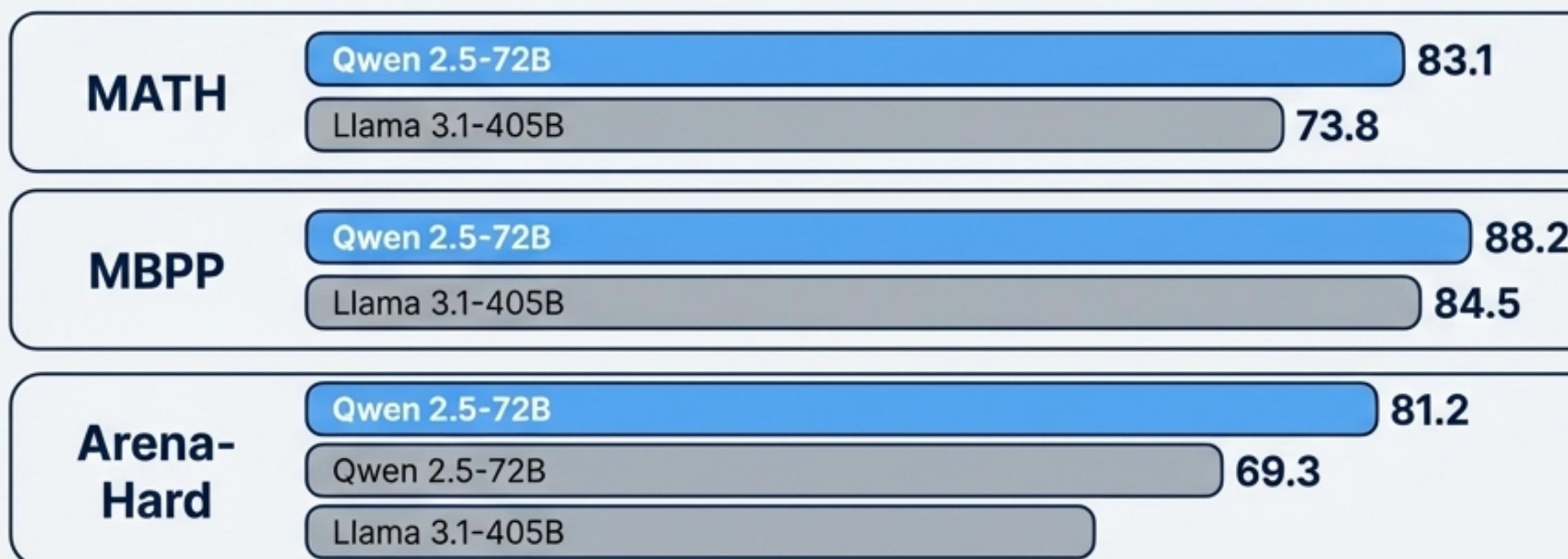
- Prawdziwość (Truthfulness)
- Pomocność (Helpfulness)
- Zwięzłość (Conciseness)
- Trafność (Relevance)
- Nieszkodliwość (Harmlessness)
- Bezstronność (Debiasing)

Analogia: To jak dołączenie do klubu dyskusyjnego, gdzie nie ma jednej słusznej odpowiedzi, a liczy się jakość argumentacji, styl i niuanse.



# Dowód: Wyniki Benchmarków i Preferencje Użytkowników

- **Kluczowe Zwycięstwa:** Qwen 2.5-72B-Instruct pokonuje znacznie większy Llama 3.1-405B-Instruct w kluczowych benchmarkach.
- **Test Ludzkich Preferencji:** W benchmarku **Arena-Hard**, gdzie ludzie anonimowo oceniają i porównują odpowiedzi modeli, mniejszy Qwen zdeklasował większego konkurenta.
- **Wniosek:** Zaawansowany proces dostrajania przełożył się bezpośrednio na realne preferencje użytkowników, dowodząc, że efektywność nie odbywa się kosztem jakości.



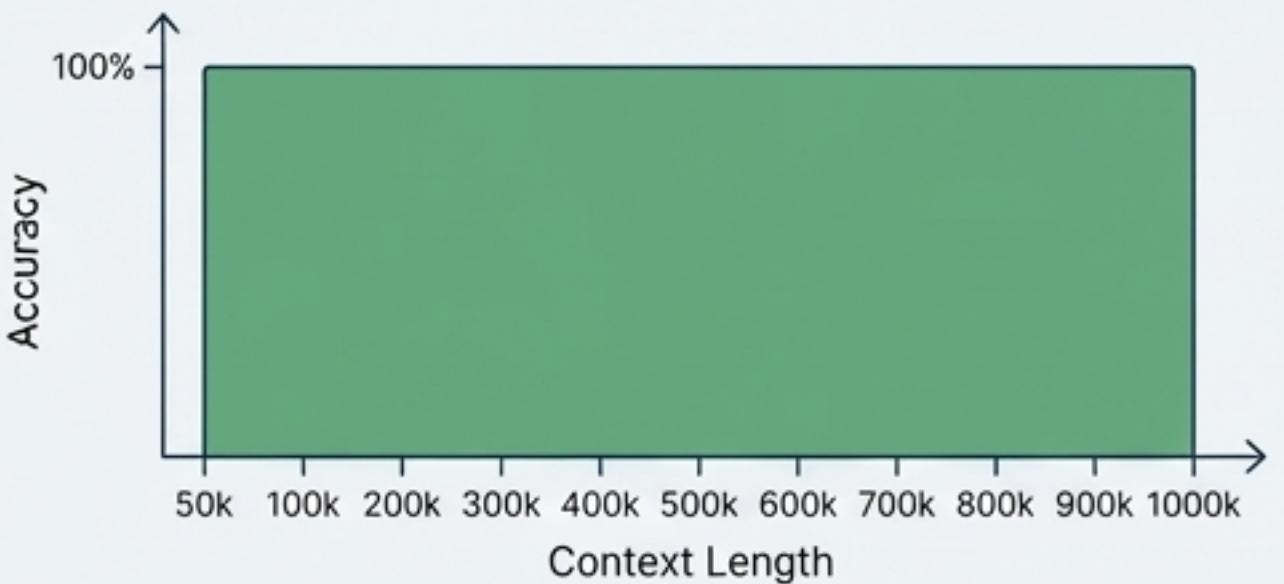
# Mistrzostwo Długiego Kontekstu: Test „Passkey Retrieval”

**Wyzwanie:** Test polega na znalezieniu jednego, ukrytego zdania (np. „Sekretny klucz to 1234”) w dokumencie o długości **1 miliona tokenów**, wypełnionym nieistotnymi informacjami.

**Wynik:** Qwen 2.5-Turbo osiągnął **100% skuteczności** w tym teście, bezbłędnie odnajdując ukrytą informację niezależnie od jej położenia w dokumencie.

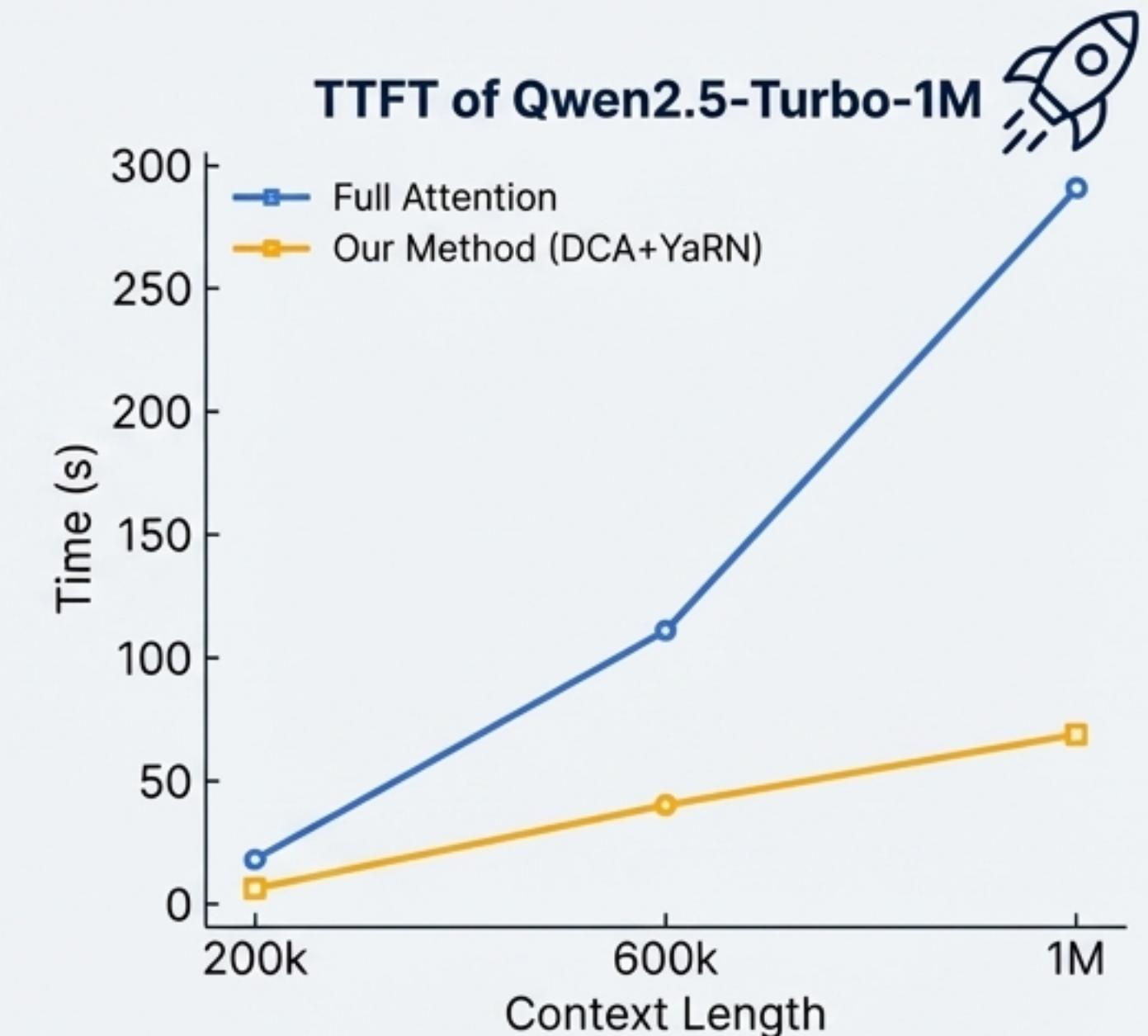
**Skala:** To zadanie jest porównywalne z dwukrotnym przeczytaniem „Wojny i pokoju” i odnalezieniem w niej jednego, losowego zdania.

**Znaczenie:** Ostateczny test zdolności do utrzymania uwagi i rozumienia informacji w ogromnej skali.



# Szybkość, Dostępność i Praktyczne Zastosowania

- **Optymalizacja Szybkości:** Dzięki technikom takim jak **Dual Chunk Attention (DCA)** i **YaRN**, uzyskano **3-4x przyspieszenie** w metryce TTFT (Time to First Token) dla długiego kontekstu.
- **Wydajność vs Koszt:** Qwen 2.5-Turbo oferuje wydajność konkurencyjną dla GPT-4o-mini przy potencjalnie niższych kosztach operacyjnych dzięki architekturze MoE.
- **Demokratyzacja AI:** Modele Open Weight (0.5B-72B) są publicznie dostępne, co umożliwia startupom, badaczom i uniwersitetom na całym świecie budowanie na ich podstawie.
- **Nowe Możliwości:** Zdolność do przetwarzania długiego kontekstu otwiera drzwi do nowych zastosowań:
  - Analiza 10 lat dokumentacji medycznej pacjenta.
  - Przetwarzanie i podsumowywanie obszernych dokumentów finansowych i prawnych.



# Pytanie do Ciebie

W miarę jak modele stają się nie tylko bardziej wydajne, ale także bardziej wielozmysłowe (przetwarzając tekst, obraz i dźwięk jednocześnie), jakie nowe kategorie problemów, które dziś wydają się niemożliwe do rozwiązania, staną się osiągalne w pierwszej kolejności?

