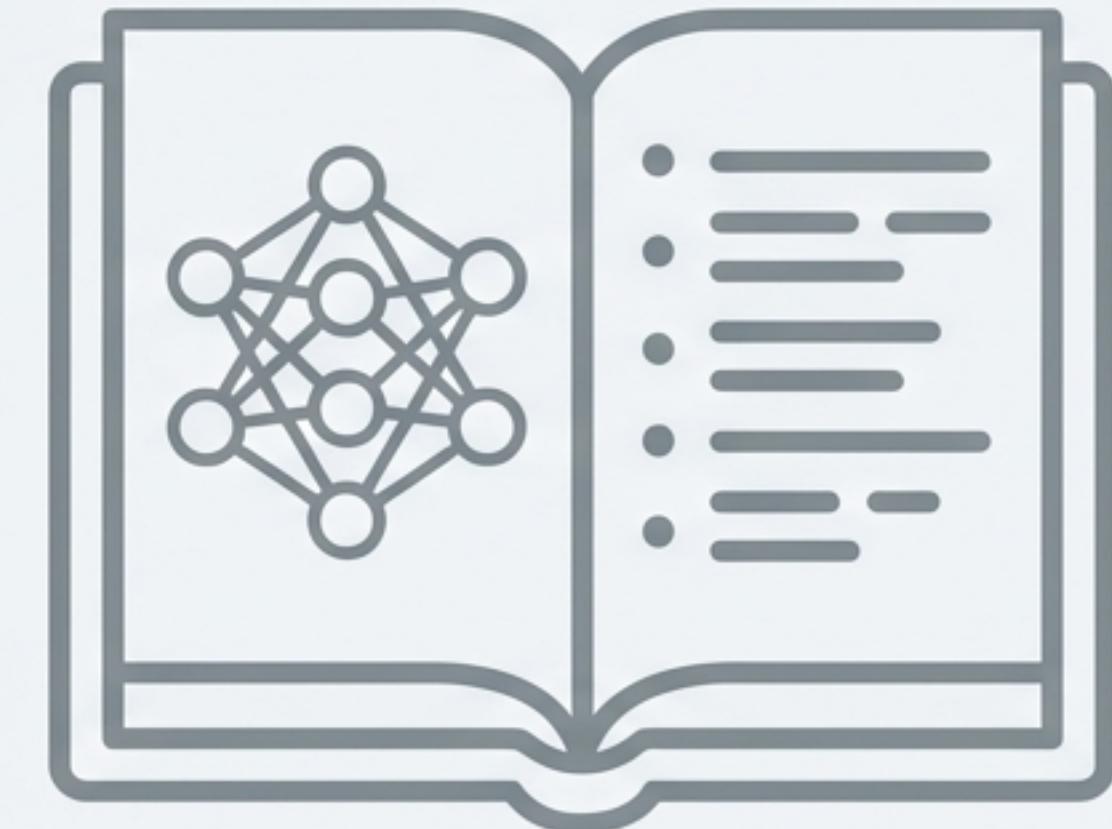
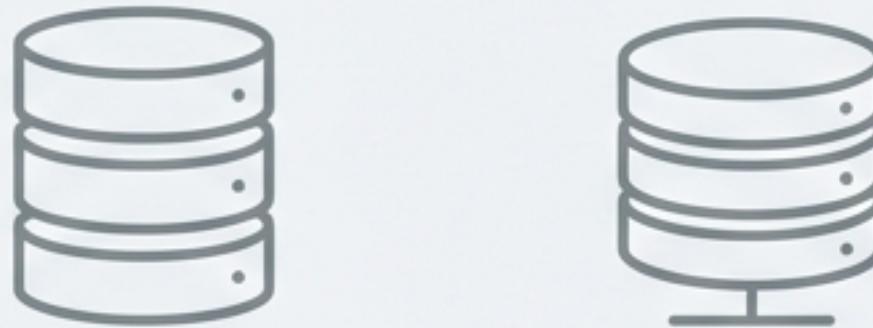


Tajemny przepis na dane treningowe LLM

- Prawdziwy sekret wydajności najnowszych modeli językowych (np. Llama 3, Mixtral) to nie architektura – to „przepis” na dane treningowe.
- Wiodące laboratoria traktują swoje metody kuracji danych jako pilnie strzeżone tajemnice handlowe, tworząc rosnącą przepaść między nimi a społecznością open-source.
- Artykuł naukowy o FineWeb, stworzony przez Hugging Face, to transparentna, krok po kroku dokumentacja całego procesu przygotowania danych.
- Cel projektu: dać społeczeństwu „wędkę, a nie tylko rybę” – udostępnić nie tylko zbiór danych, ale całą metodologię i kod.
- To fundamentalnie zmienia zasady gry dla otwartego rozwoju AI, minimalizując lukę wiedzy.



Dwa kluczowe rezultaty: FineWeb i FineWeb-Edu



FineWeb

- **15 bilionów (15T)** tokenów zebranych i przefiltrowanych z 96 zrzutów Common Crawl.
- Skala wystarczająca do trenowania modeli wagi superciężkiej (ponad 500 miliardów parametrów).
- *Analogia: Ogromna, uporządkowana biblioteka internetu.*



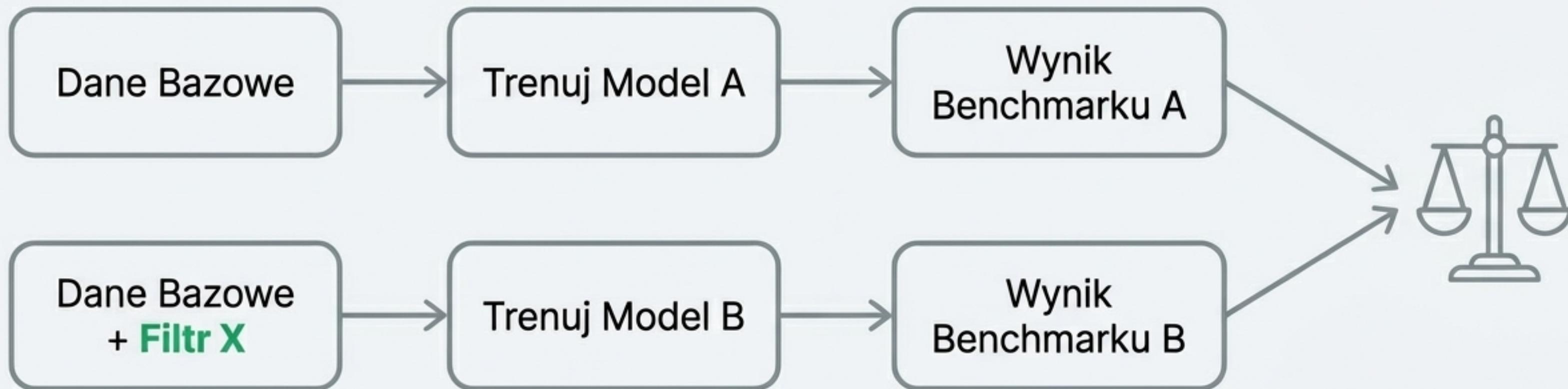
FineWeb-Edu

- **1.3 biliona (1.3T)** tokenów – starannie wyselekcyjowany podzbiór z FineWeb o wysokiej wartości edukacyjnej.
- Zoptymalizowany pod kątem zadań wymagających wiedzy i rozumowania.
- *Analogia: Ekstrakcja tylko najlepszych książek i artykułów naukowych z tej biblioteki.*

Oba zbiory danych zostały publicznie udostępnione wraz z pełną dokumentacją i kodem źródłowym.

Metodologia ablacji danych: naukowe podejście do każdej decyzji

Każdy wybór w procesie tworzenia danych został zweryfikowany empirycznie, a nie oparty na intuicji.



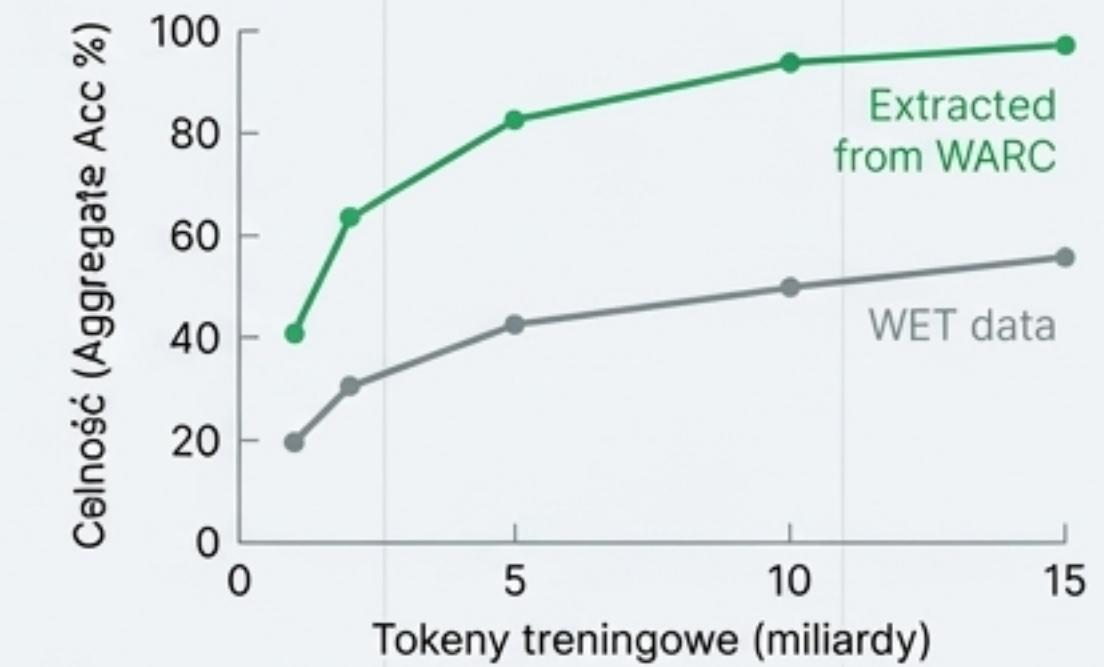
To podejście przekształca intuicję w twardą daną – każda decyzja jest statystycznie potwierdzona. W sumie wytrenowano ponad 70 modeli w ramach eksperymentów.

Ekstrakcja tekstu – pierwszy krytyczny krok

Common Crawl udostępnia dane w dwóch formatach:

- **Pliki WET:** Przetworzony, czysty tekst. Wygodne, ale często zawierają dużo „śmieci” (teksty z menu, stopki, reklamy).
- **Pliki WARC:** Surowe dane z crawlera, zawierające pełny kod HTML strony.

Odkrycie: Większość projektów dla wygody korzysta z plików WET. Testy wykazały jednak, że użycie biblioteki Trafilatura do ekstrakcji treści bezpośrednio z surowych plików WARC daje znacznie lepsze rezultaty.

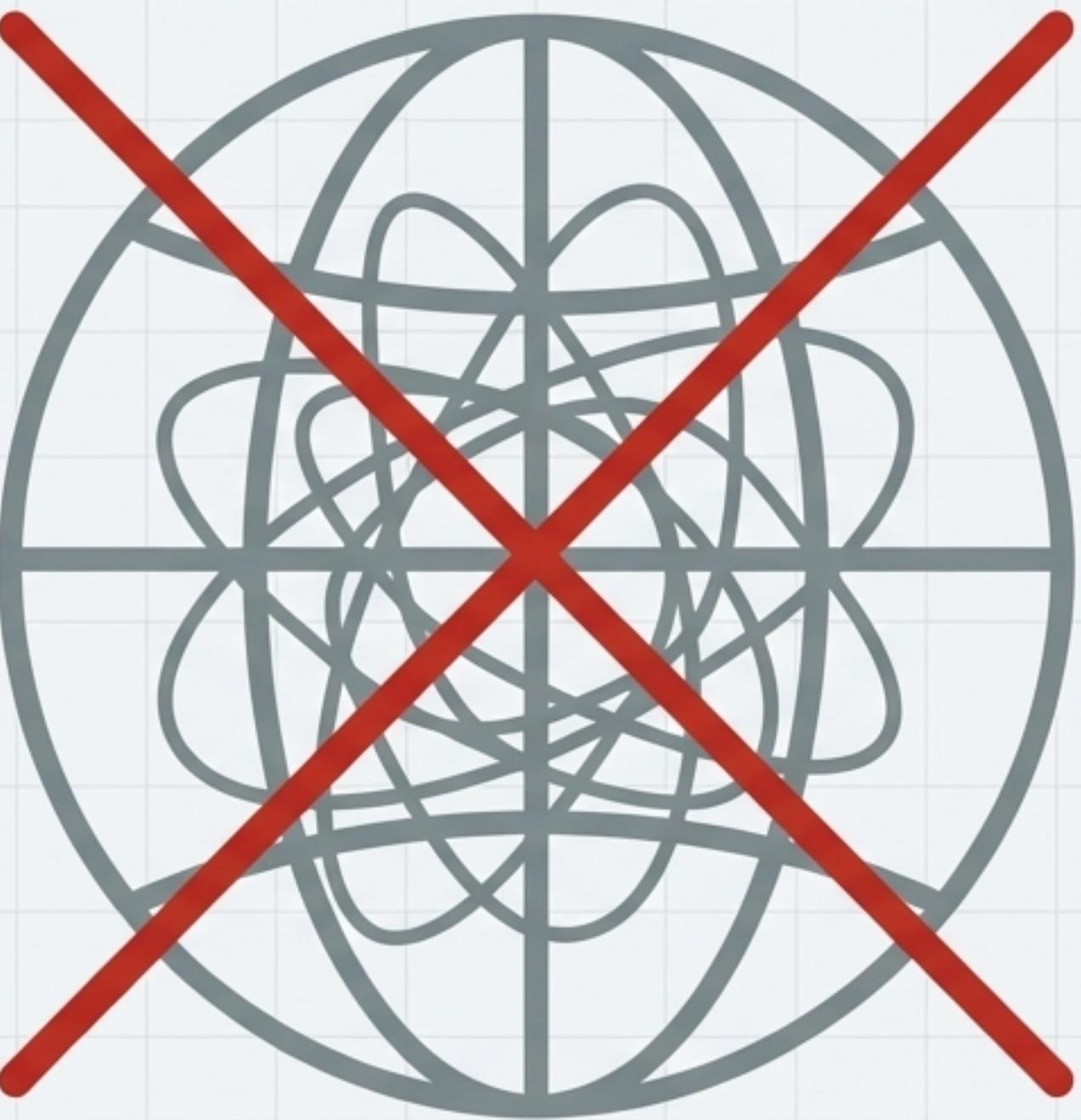


Zaskakujące odkrycie: globalna deduplikacja szkodzi

Początkowa intuicja: Agresywna, globalna deduplikacja danych ze wszystkich 96 zrzutów Common Crawl zmaksymalizuje jakość.

Szokujący rezultat:

- Prawie **żadnej poprawy** wydajności modelu w porównaniu do danych bez deduplikacji.
- Co gorsza, w przypadku starszych zrzutów (np. 2013-48) proces ten **zachowywał dane niższej jakości**, odrzucając te bardziej wartościowe.
- Po usunięciu 90% danych jako duplikatów, pozostałe 10% okazało się gorszej jakości niż odrzucona reszta.



**Globalna deduplikacja przypomina usuwanie z biblioteki wszystkich kopii popularnych klasyków... i zostawianie tylko unikalnych, ale często bezwartościowych ulotek.*

Rozwiązanie: deduplikacja lokalna w ramach każdego zrzutu

Nowe, lepsze podejście:

Zamiast traktować cały internet jako jedną całość, każdy z 96 zrzutów Common Crawl został potraktowany jak osobna „biblioteka”.

- Izolacja:** Każdy zrzut jest przetwarzany niezależnie od pozostałych.
- Deduplikacja lokalna:** Algorytm **MinHash** jest stosowany do usuwania duplikatów *tylko w obrębie jednego zrzutu*.
- Agregacja:** Oczyszczone zrzuty są łączone w końcowy zbiór danych.



Przełomowy rezultat: Ta jedna zmiana pozwoliła modelom trenowanym na FineWeb dorównać wydajnością RefinedWeb – wiodącemu wówczas publicznemu zbiorowi danych.

Uczenie się z filtrów C4: inteligentne udoskonalanie

Zespół przeanalizował filtry użyte w klasycznym zbiorze danych C4, aby zrozumieć ich skuteczność.

Odkrycia:

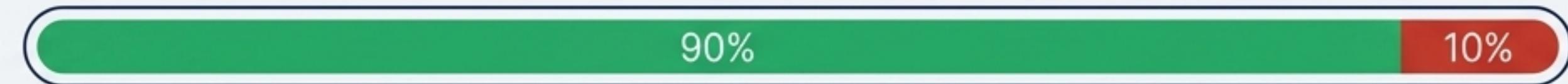
- **Filtr interpunkcyjny C4:** Wymóg, aby każda linia kończyła się znakiem interpunkcyjnym, był skuteczny, ale **zbyt agresywny** – odrzucał aż 30% danych.
- **Nowe, inteligentne podejście:** Stworzono własny filtr usuwający dokumenty, w których **mniej niż 12% linii** kończy się znakiem interpunkcyjnym.

C4: Agresywny filtr



30% odrzuconych danych

FineWeb: Inteligentny filtr



~10% odrzuconych danych

Podejście oparte na danych przewyższyło klasyczne heurystyki, osiągając lepszą jakość przy mniejszej utracie danych.

Tworzenie FineWeb-Edu: AI uczy AI

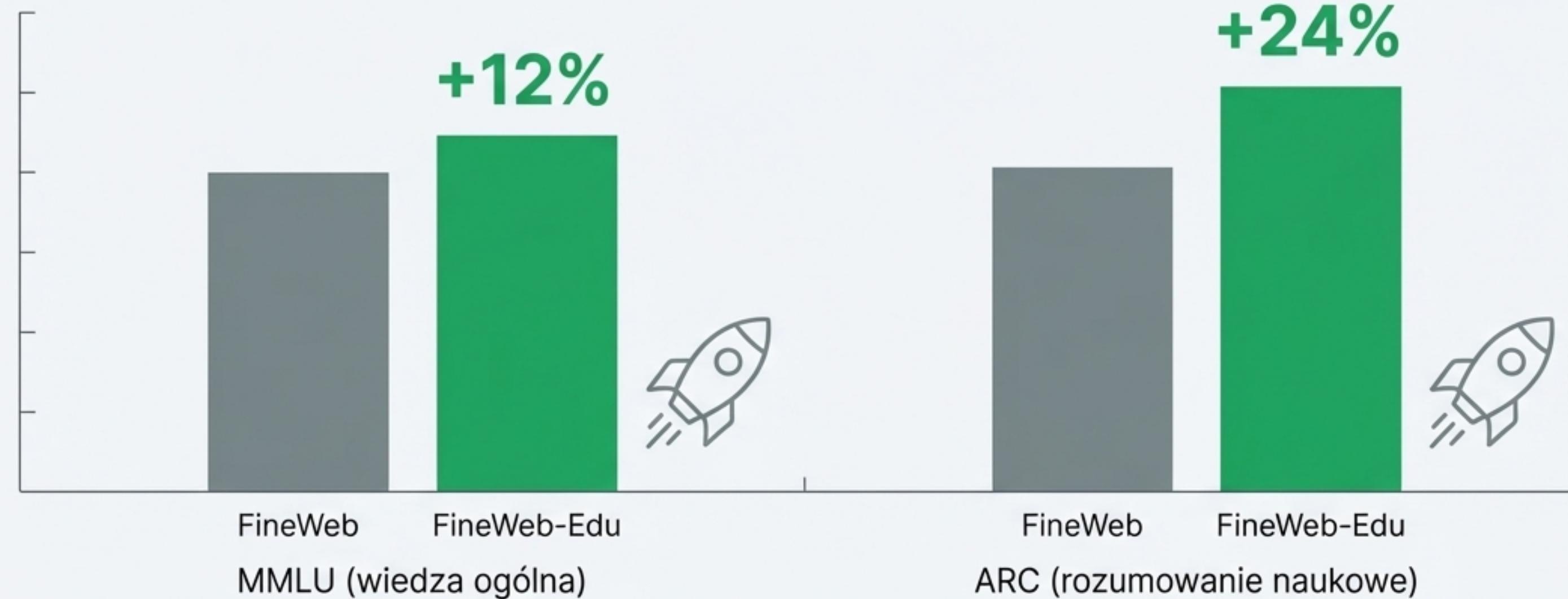
Inspiracją były spekulacje, że Llama 3 trenowano na specjalnie wyselekcjonowanych danych wysokiej jakości. Postanowiono odtworzyć ten proces w otwarty sposób.



To innowacyjna metodologia, w której jedna zaawansowana AI uczy inną, jak rozpoznawać i selekcjonować cenne treści na masową skalę.

Spektakularne rezultaty filtrowania edukacyjnego

Model trenowany na FineWeb-Edu osiąga na benchmarku MMLU wydajność konkurencyjnych zbiorów danych, używając **prawie 10 razy mniej tokenów**.



To dowodzi ogromnej efektywności filtrowania danych pod kątem treści edukacyjnych. Jakość jest ważniejsza niż ilość.

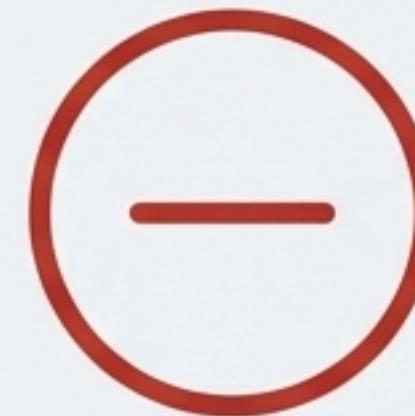
Kompromis specjalizacji: analiza tematów i domen

Filtrowanie edukacyjne zmienia profil tematyczny zbioru danych. FineWeb-Edu jest wyspecjalizowany.



Zyskuje na znaczeniu:

- Edukacja, Historia, Nauka
- Biologia, Środowisko
- Treści akademickie (ArXiv)
- Treści encyklopedyczne (Wikipedia)



Traci na znaczeniu:

- Biznes, Finanse, Rozrywka
- Filmy, Podróże, Nieruchomości
- Szerokie źródła internetowe (C4)
- Media społecznościowe (Reddit)

Wybór zbioru danych zależy od docelowego zastosowania modelu.

Kluczowe wnioski i narzędzia dla społeczności

- **Jakość danych jest kluczowa:** Przemyślany, iteracyjny proces kuracji przynosi ogromne korzyści, często przewyższające samą ilość danych.
- **Testuj wszystko empirycznie:** Rygorystyczne testy (ablacje) są niezbędne do obalania błędnych intuicji (np. globalna deduplikacja).
- **AI może pomóc w tworzeniu lepszej AI:** Filtrowanie z pomocą LLM to potężna i przyszłościowa technika kuracji danych.



Wnioski



Udostępnione zasoby



Wzmocnienie społeczności

Wędka, nie tylko ryba: Udostępniamy oba zbiory danych (FineWeb, FineWeb-Edu), wszystkie 70+ modeli z eksperymentów i cały kod (datatrove), aby społeczność mogła budować na naszej pracy.