EcoPrompt Optimizer

Redukcja kosztów i energii w systemach LLM
poprzez inteligentną kompresję promptów.

Statystyki

900 000 000 000 promptów

Na tyle promptów rocznie odpowiada ChatGPT

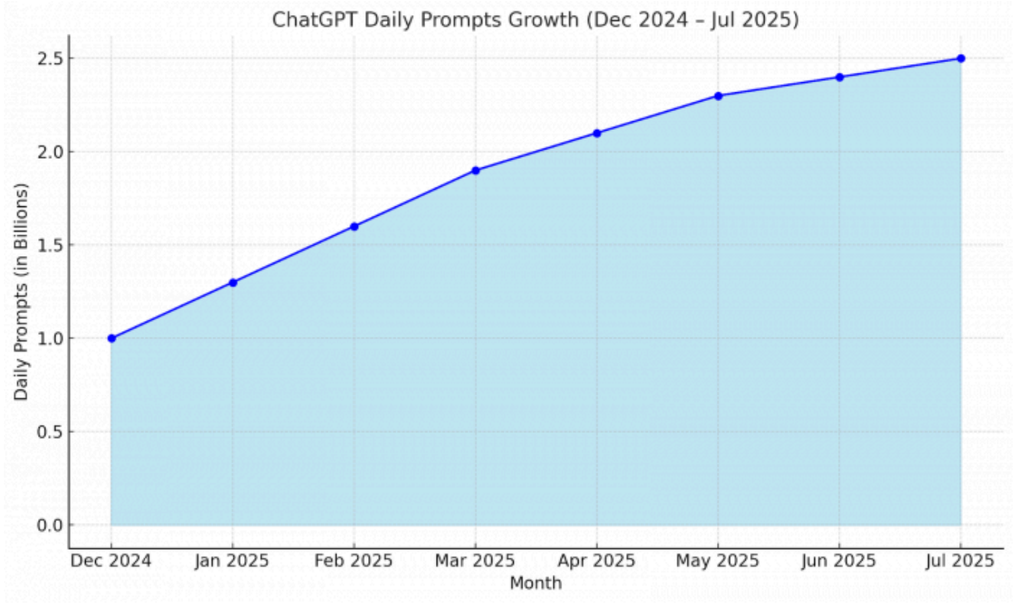
100 słów

Każde 100 słów zużywa litr wody do chłodzenia serwerów - ten sam proces zużywa tyle prądu co działanie 14 żarówek przez godzine

67%

W 2024 roku 67% Amerykanów było uprzejmych w “rozmowach” Chat botem

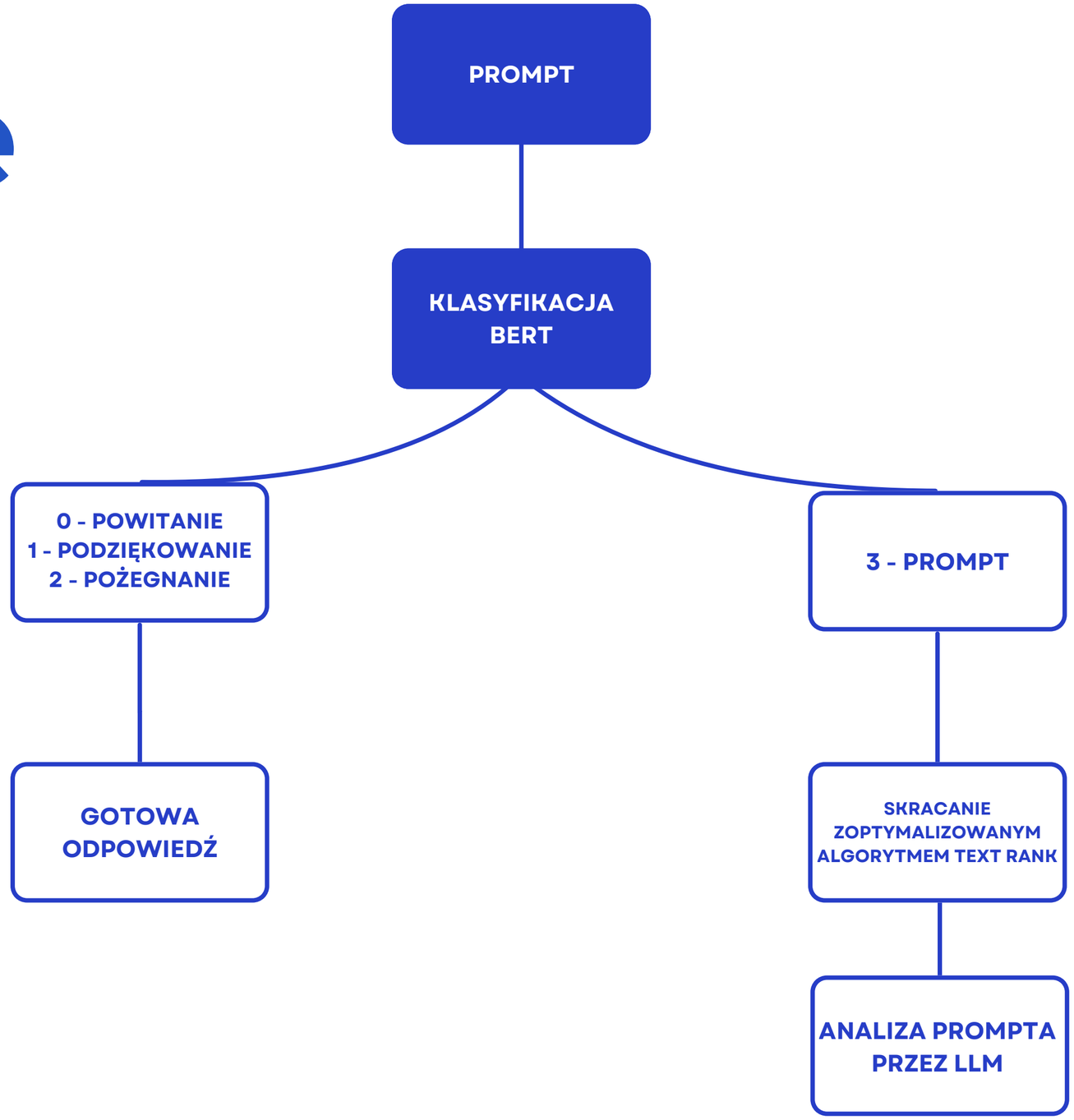
Problem



- LLM-y zużywają gigantyczne ilości energii na każde zapytanie, przy czym często marnują moc obliczeniową na nieefektywnie ułożone prompty
- Nieefektywnie odpowiadają na “uprzejmości” ze strony użytkowników.

Rozwiązanie

- Klasyfikujemy prompta jako 0, 1 i 2, aby zastąpić grzecznościowy small talk sztywnymi regułami, aby zastąpić kosztowne tokeny LLM-a
- Jeśli klasyfikujemy prompta jako 3 to skracamy jego długość (bez straty odpowiedzi)
- Analizujemy oszczędność energii i kosztów



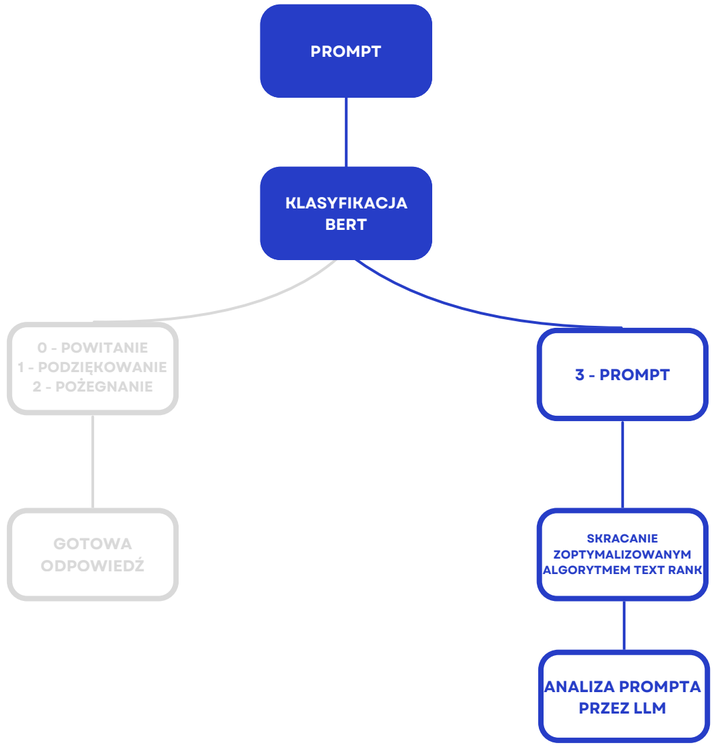
Dataset

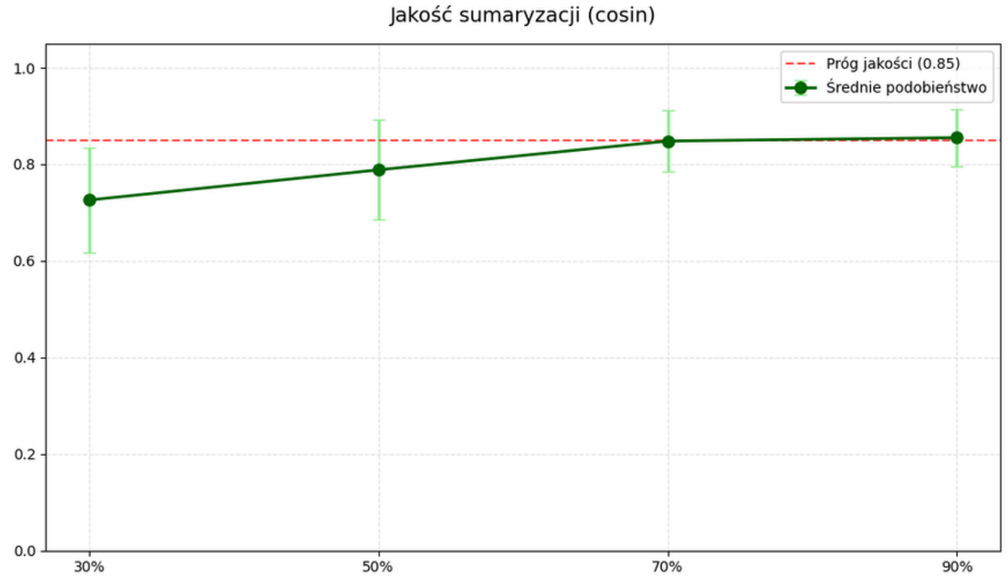
- Label 0, 1 i 2 wygenerowane na podstawie najczęściej używanych zwrotów w języku angielskim
- Label 3 wygenerowane na podstawie danych “DailyDialog” oraz wybranych i wygenerowanych podchwytliwych zwrotów

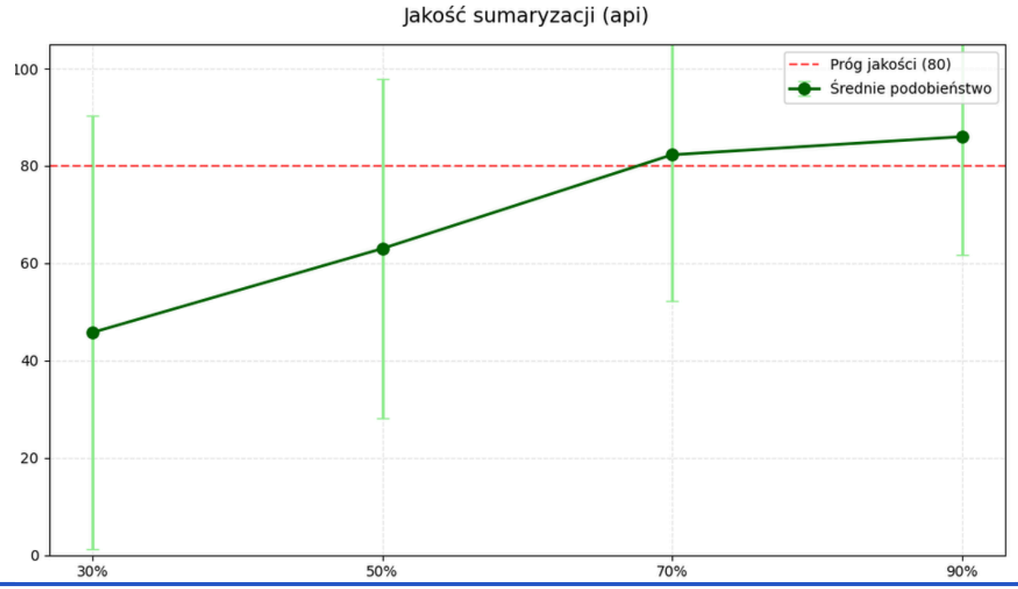
Label	Kategoria	Przykłady zawartości
0	**Powitania (GREETINGS)**	hi, Hi, hello, how are you doing, how’re you doing
1	**Podziękowania (THANKS)**	thanks, much appreciated, cheers, Thank you
2	**Pożegnania (GOODBYES)**	Bye, see you later, have a nice day!
3	**Inne / Podchwytliwe (OTHERS)**	Explain quantum physics I see thanksgiving dinner recipes Thanks and now code this

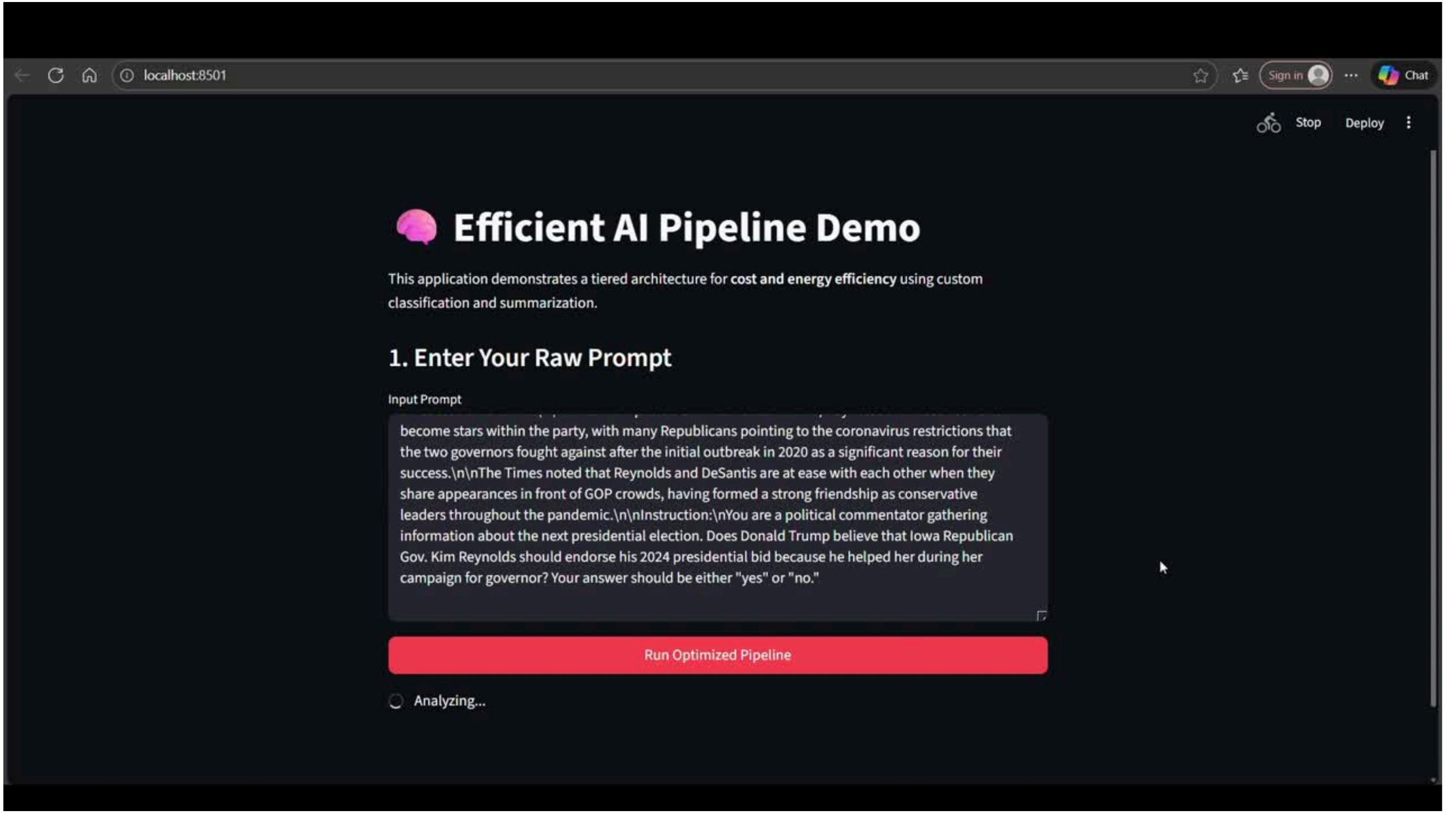
Klasyfikacja i skrócenie prompta

- Klasyfikacja Bert
- Jeśli sklasyfikowaliśmy jako label 3 to upraszczamy metodą text rank








localhost:8501
Sign in
Chat
Stop
Deploy
Efficient AI Pipeline Demo
This application demonstrates a tiered architecture for cost and energy efficiency using custom classification and summarization.
1. Enter Your Raw Prompt
Input Prompt
become stars within the party, with many Republicans pointing to the coronavirus restrictions that the two governors fought against after the initial outbreak in 2020 as a significant reason for their success.\n\nThe Times noted that Reynolds and DeSantis are at ease with each other when they share appearances in front of GOP crowds, having formed a strong friendship as conservative leaders throughout the pandemic.\n\nInstruction:\nYou are a political commentator gathering information about the next presidential election. Does Donald Trump believe that Iowa Republican Gov. Kim Reynolds should endorse his 2024 presidential bid because he helped her during her campaign for governor? Your answer should be either "yes" or "no."
Run Optimized Pipeline
Analyzing...

Ocena oszczędności

Redukujemy ślad węglowy przez zmniejszenie liczby tokenów przetwarzanych przez energochłonne modele w data center.

Z naszych badań wynika, że nasza metoda oszczędza około 15-20% energii poświęcanej na zapytania do LLM-a, nie uwzględniając klasyfikacji zwrotów grzecznościowych.

Klasyfikacja zwrotów grzecznościowych całkowicie redukuje potrzebę użycia LLM-ów co skutkuje jeszcze większym zaoszczędzeniem energii

Model Biznesowy

Klient: Firmy SaaS i startupy technologiczne, które intensywnie wykorzystują płatne API w swoich produktach (np. chatboty obsługi klienta, asystenci kodowania).

Korzyści:

- Oszczędność: Redukcja kosztów operacyjnych (API providerzy rozliczają się za liczbę tokenów wejściowych).
- Wydajność: Krótszy czas przetwarzania zapytania.

Dziękujemy

Iwo Zowada
Albert Arnautov
Robert Raniszewski