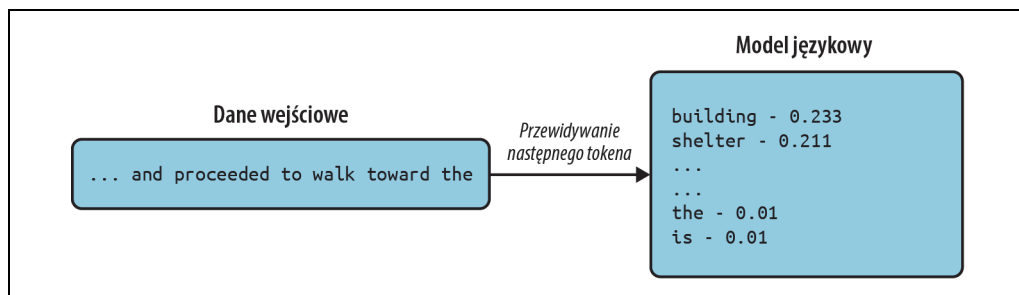
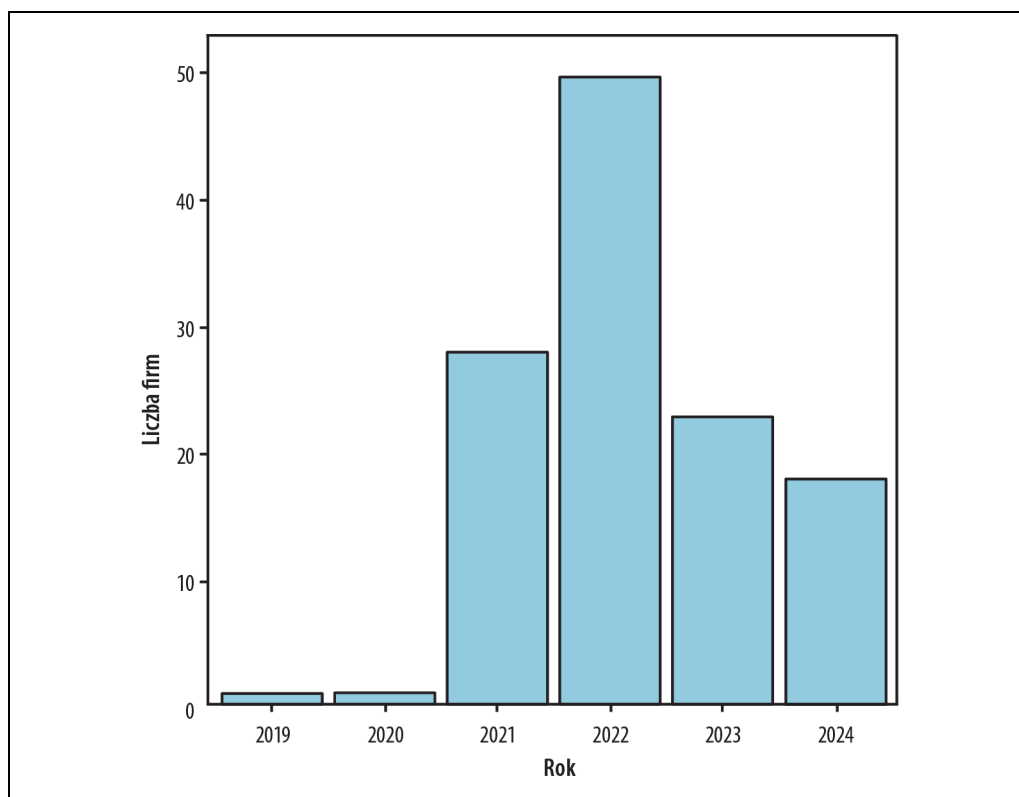


Kolorowe wersje rysunków do książki: „Projektowanie aplikacji LLM. Holistyczne podejście do dużych modeli językowych”

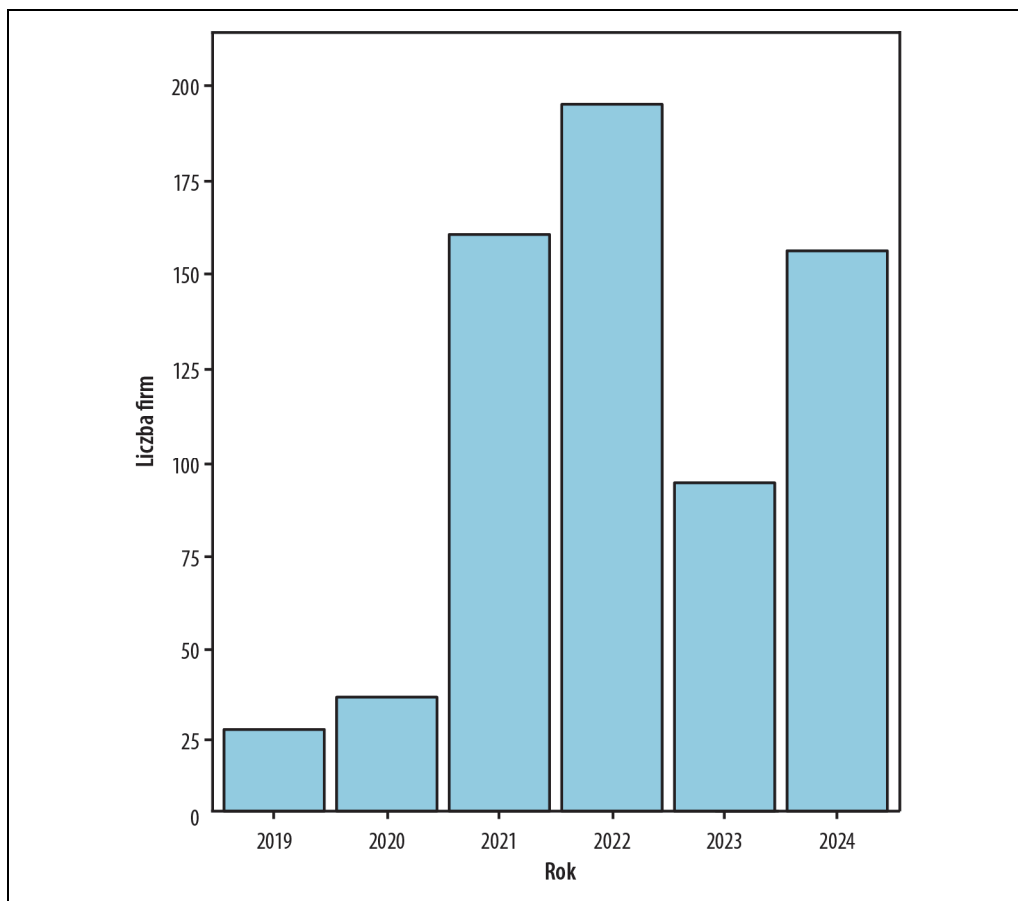
Rozdział 1. Wprowadzenie



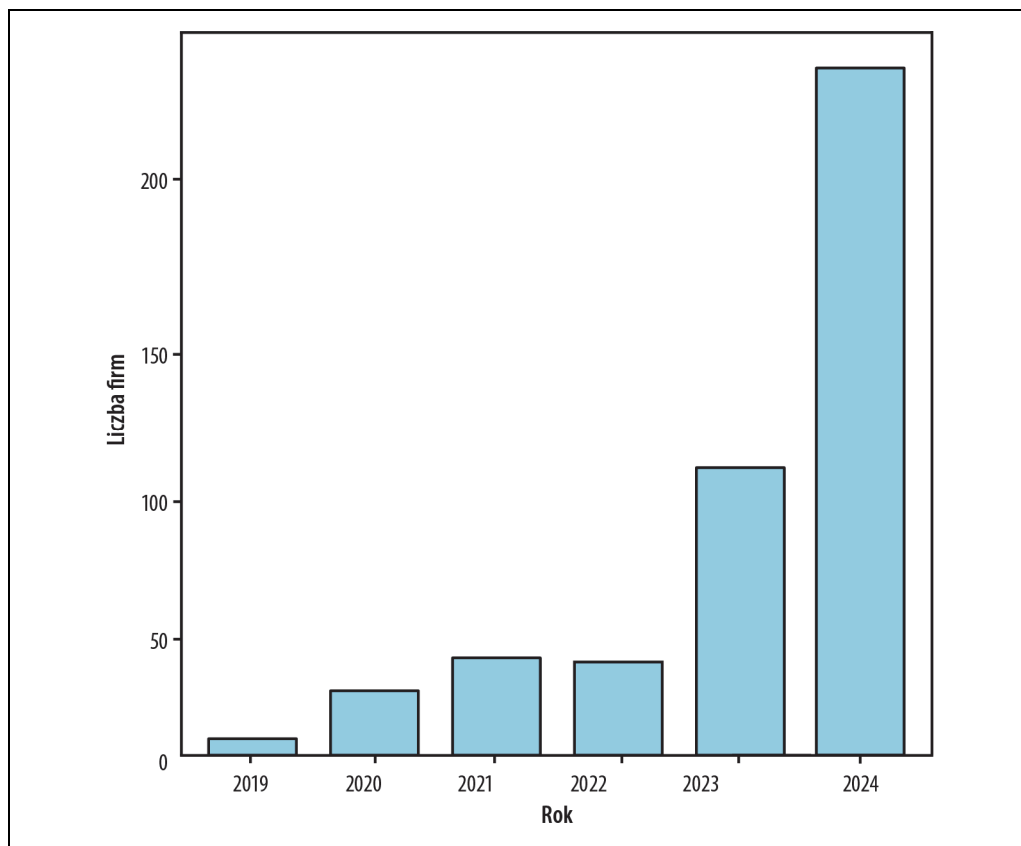
Rysunek 1.1. Trenowanie modelu z wykorzystaniem przewidywania następnego tokena



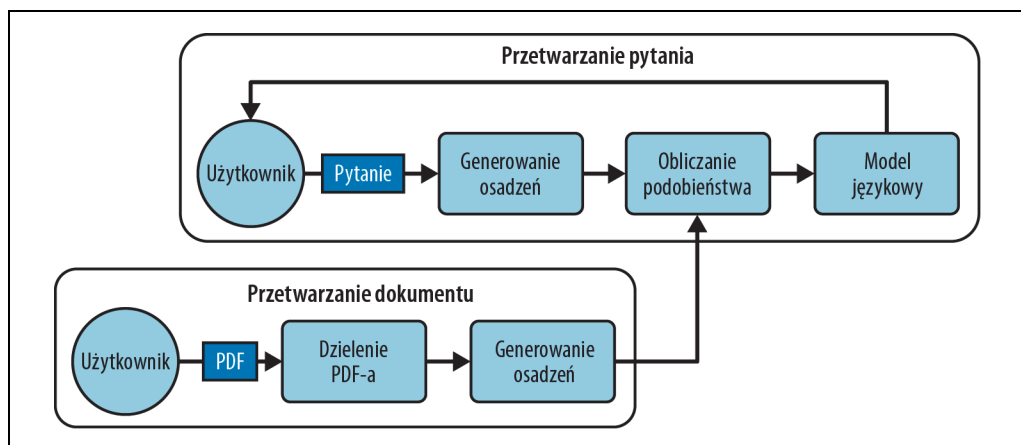
Rysunek 1.2. Firmy wspominające o technologiach Web3 podczas konferencji finansowych



Rysunek 1.3. Firmy wspominające o kryptowalutach/blockchainie podczas konferencji finansowych



Rysunek 1.4. Firmy, którym zadawano pytania o sztuczną inteligencję podczas konferencji finansowych w pierwszych dwóch miesiącach roku

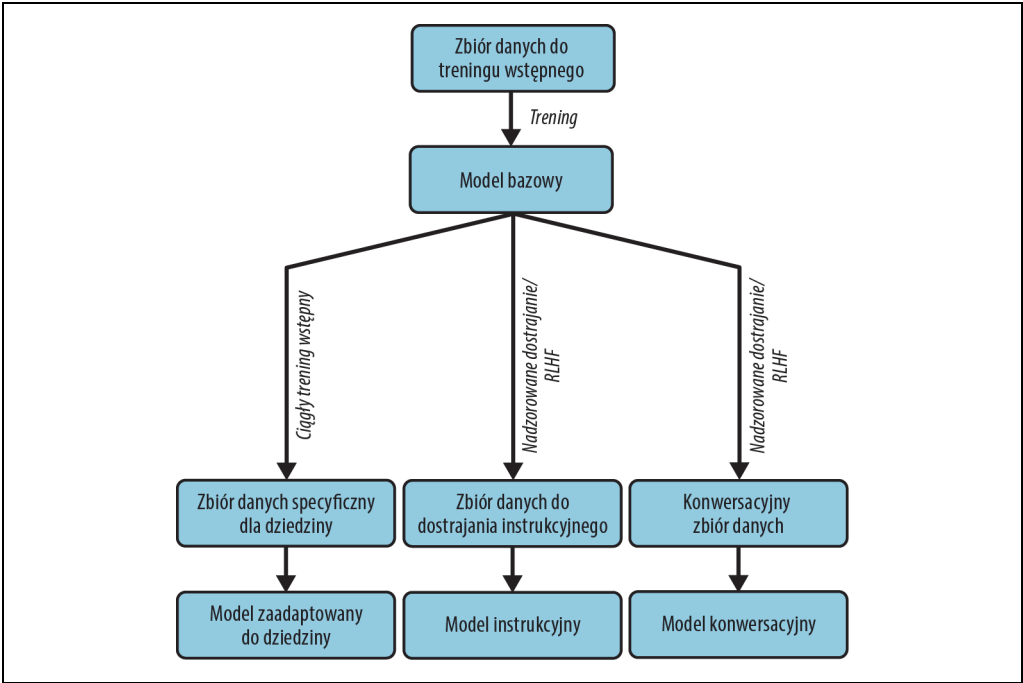


Rysunek 1.5. Schemat działania aplikacji czatbota

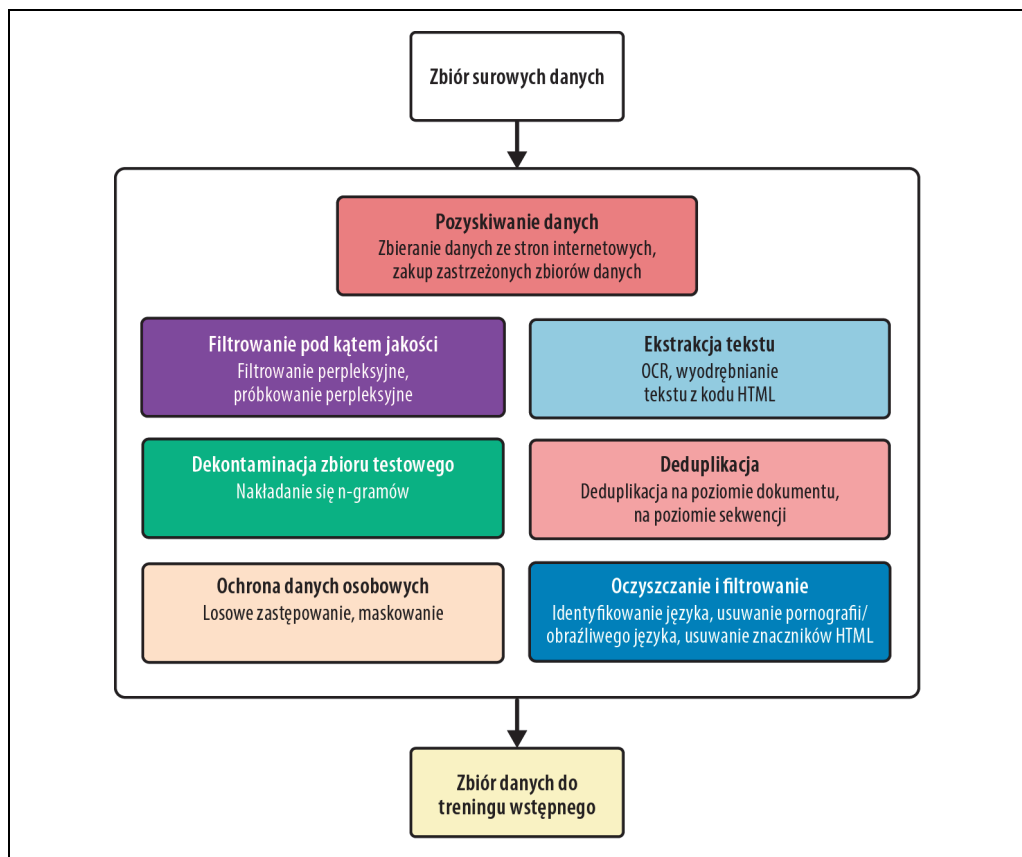
Rozdział 2. Dane do treningu wstępnego



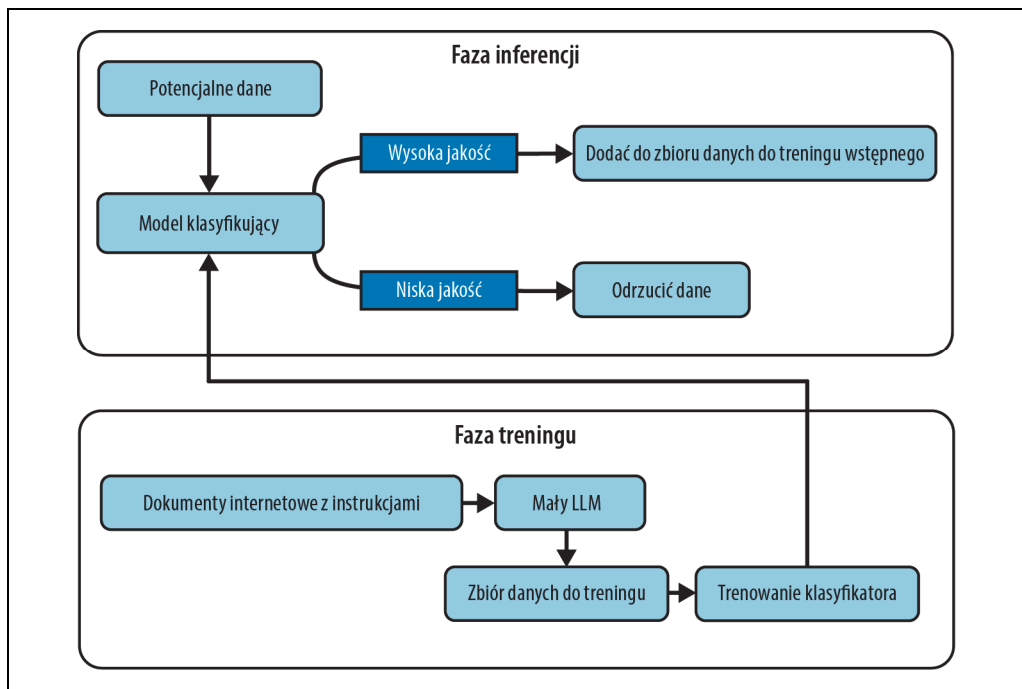
Rysunek 2.1. Jak łączą się składniki tworzące LLM



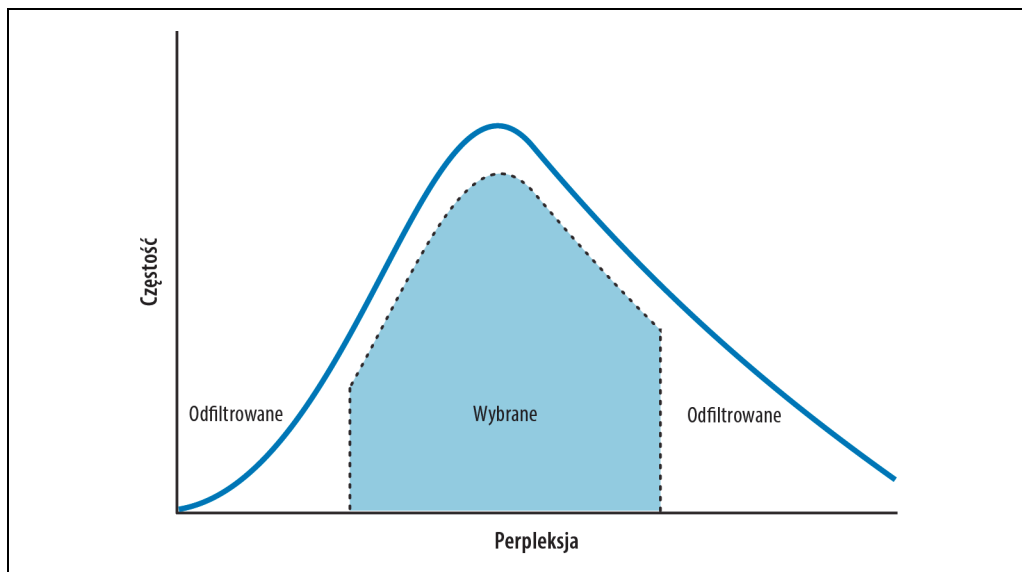
Rysunek 2.2. Zależność między modelami bazowymi a ich pochodnymi



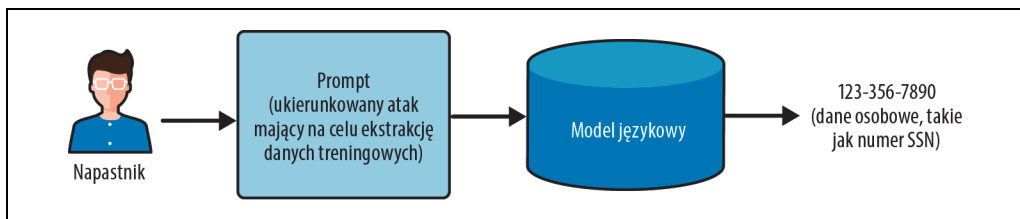
Rysunek 2.3. Zbieranie i wstępne przetwarzanie danych



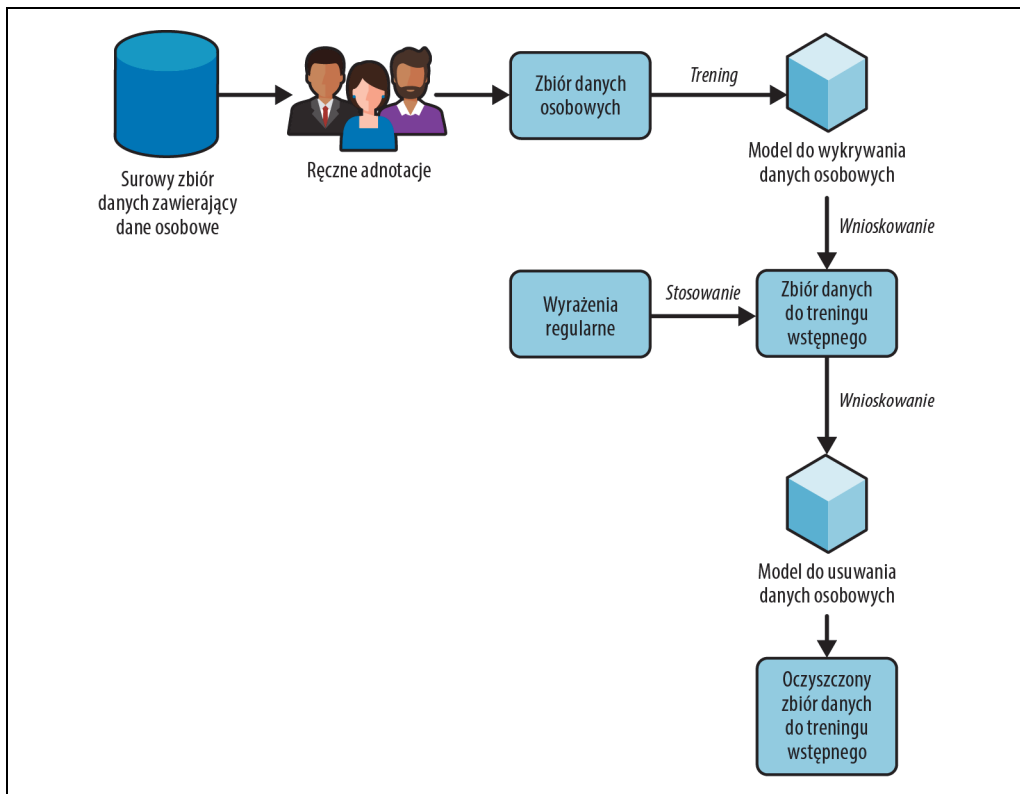
Rysunek 2.4. Filtrowanie jakości oparte na klasyfikatorze



Rysunek 2.5. Próbkowanie perpleksyjne

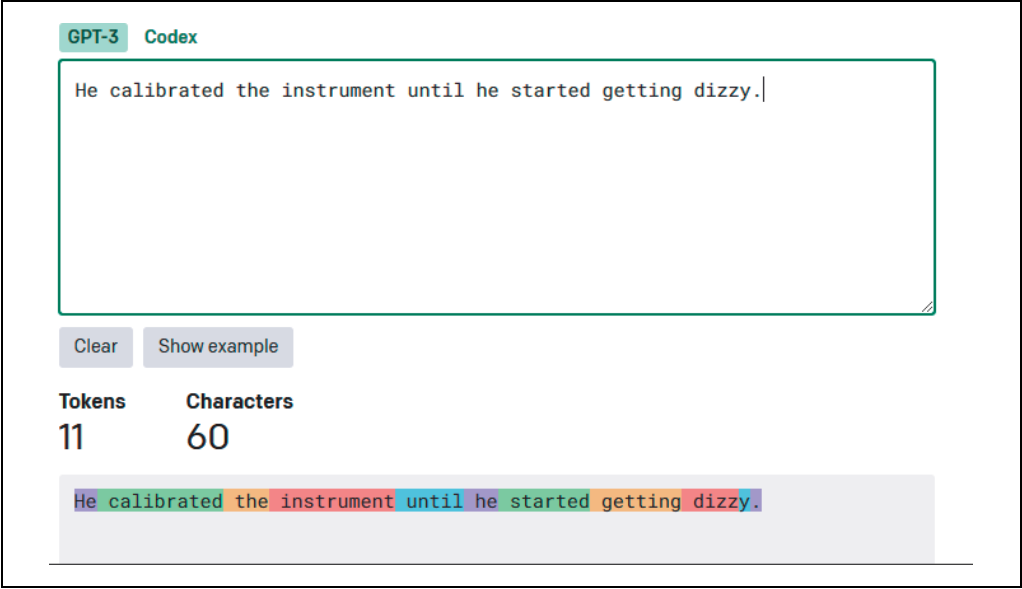


Rysunek 2.6. Ataki na prywatność w LLM-ach

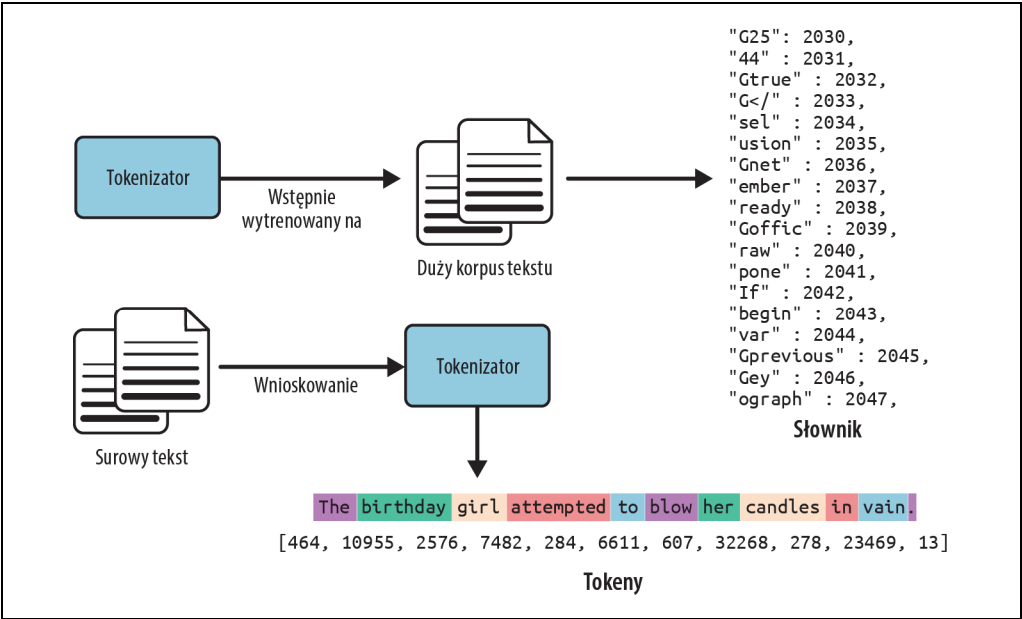


Rysunek 2.7. Potok przetwarzania danych osobowych

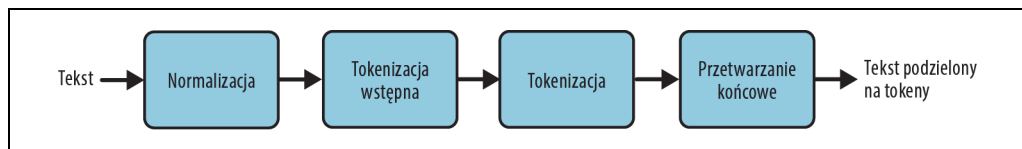
Rozdział 3. Słownik i tokenizacja



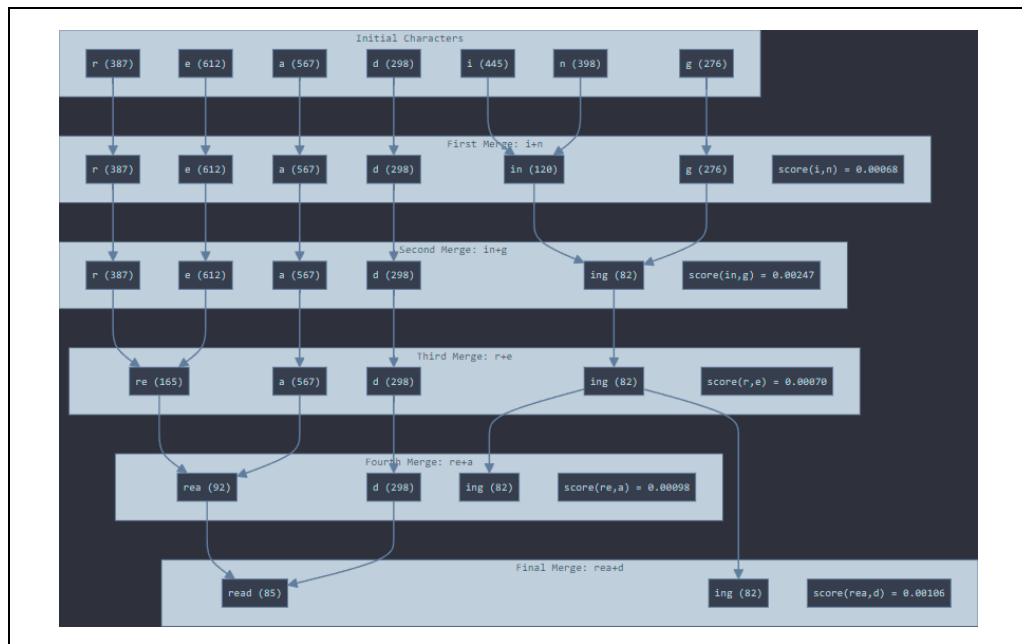
Rysunek 3.1. Tokeny podwyrazowe



Rysunek 3.2. Schemat działania tokenizatora

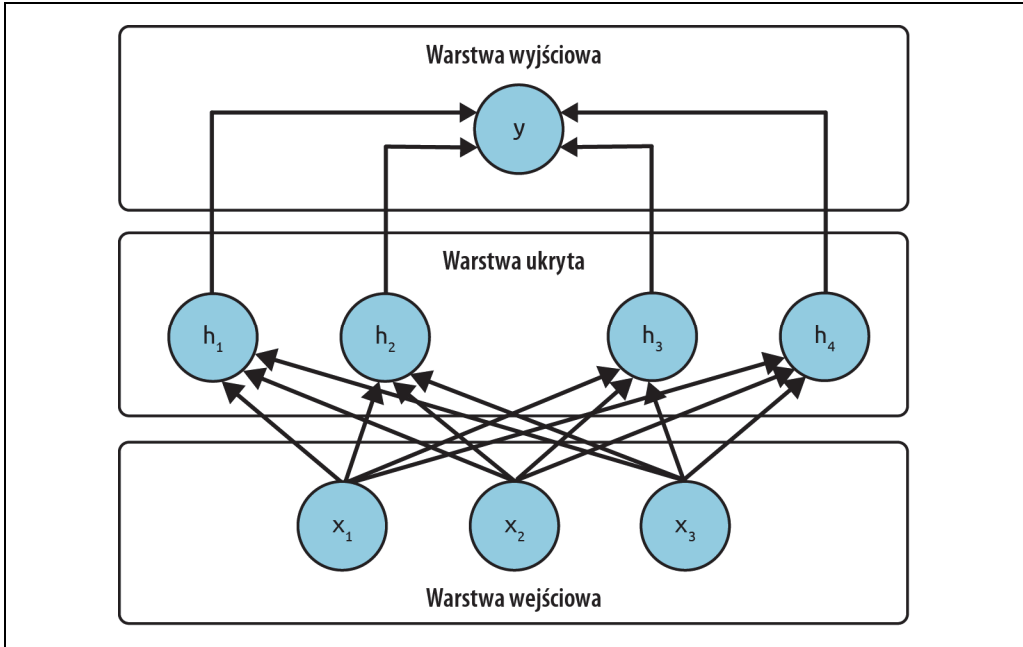


Rysunek 3.3. Potok tokenizatorów Hugging Face

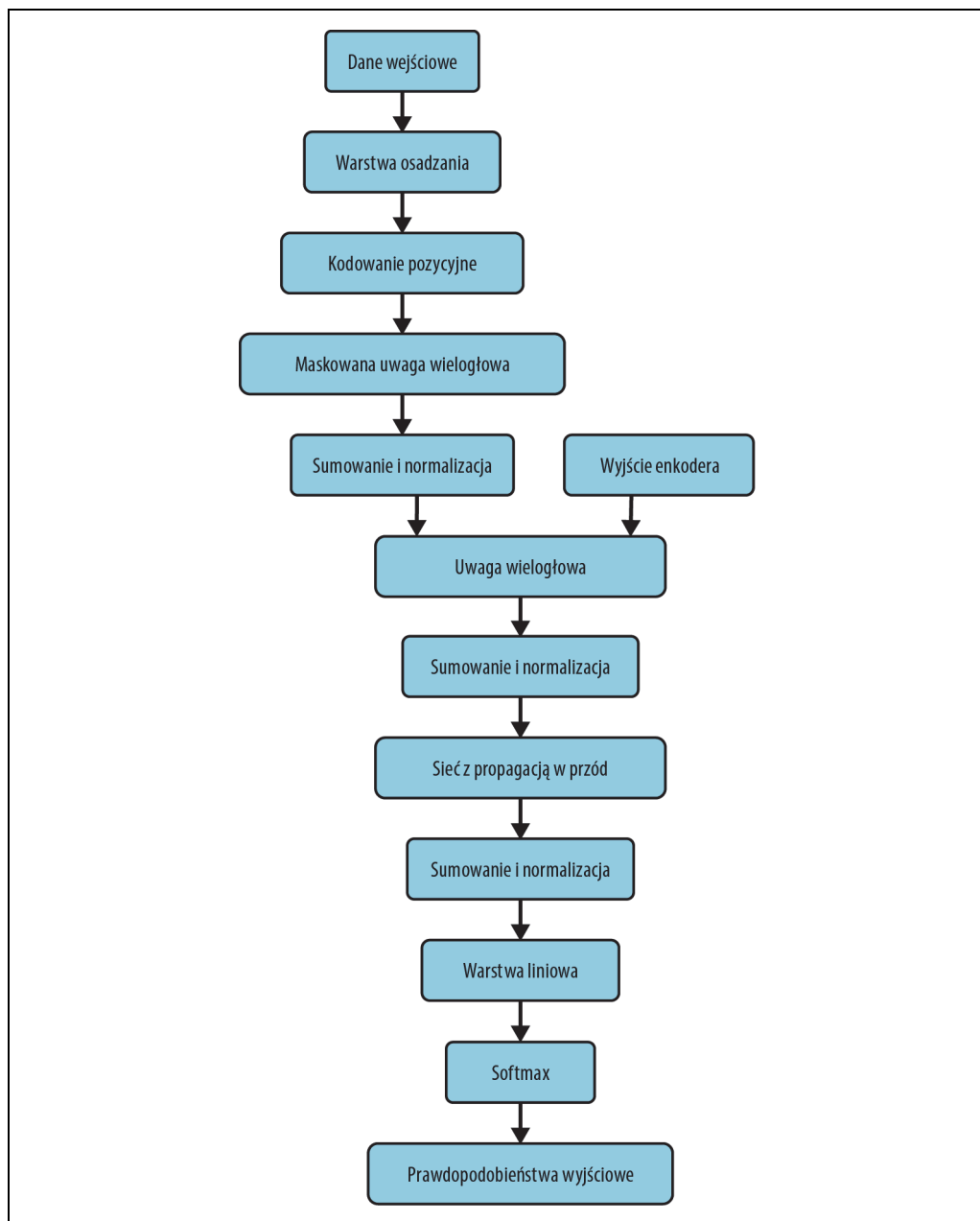


Rysunek 3.4. Tokenizacja WordPiece

Rozdział 4. Architektury i cele uczenia

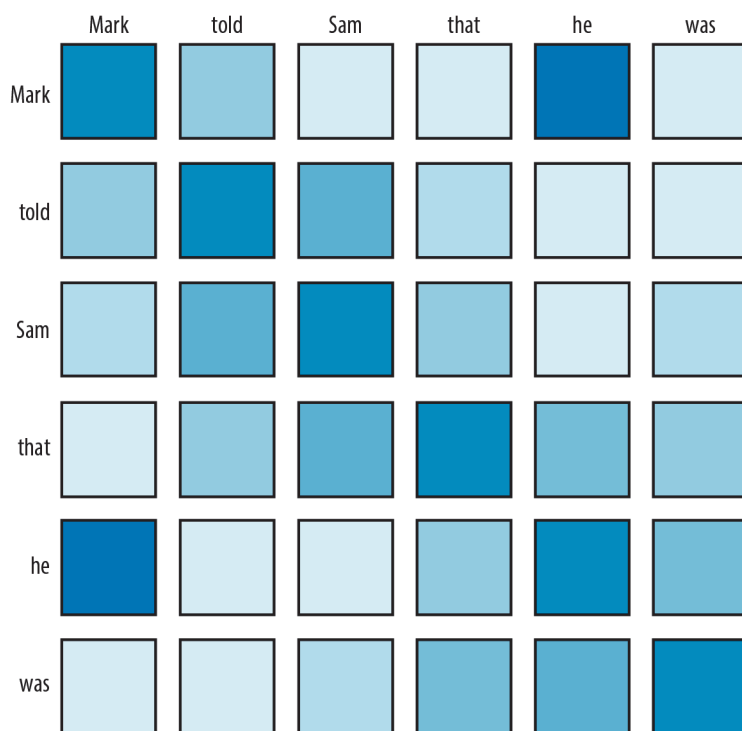


Rysunek 4.1. Perceptron wielowarstwowy

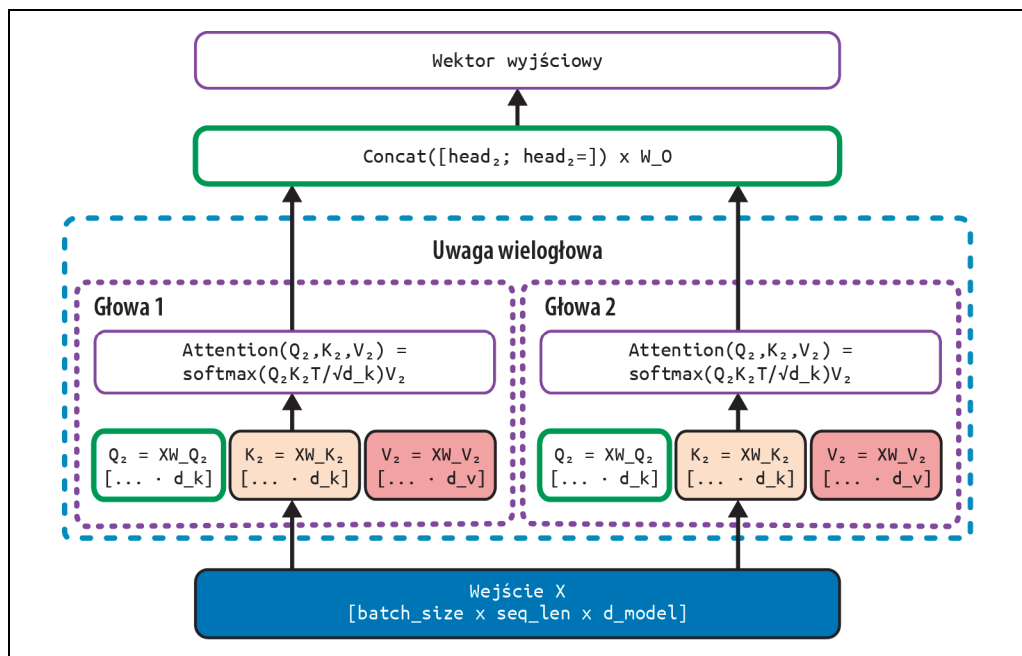


Rysunek 4.2. Architektura transformatora

Mapa uwagi dla zdania „Mark told Sam that he was planning to resign”

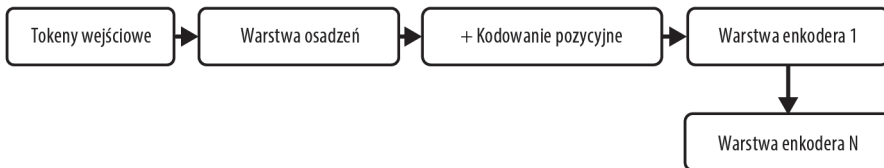


Rysunek 4.3. Mapa uwagi



Rysunek 4.4. Obliczanie samouwagi

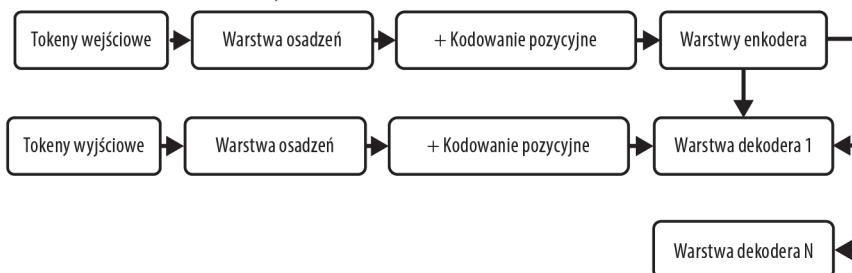
1. Model enkoderowy



Szczegóły warstwy enkodera



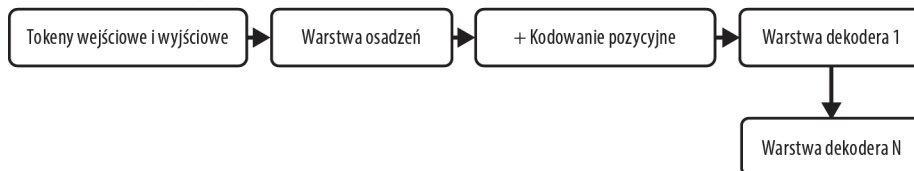
2. Model enkoderowo-dekoderowy



Szczegóły warstwy dekodera



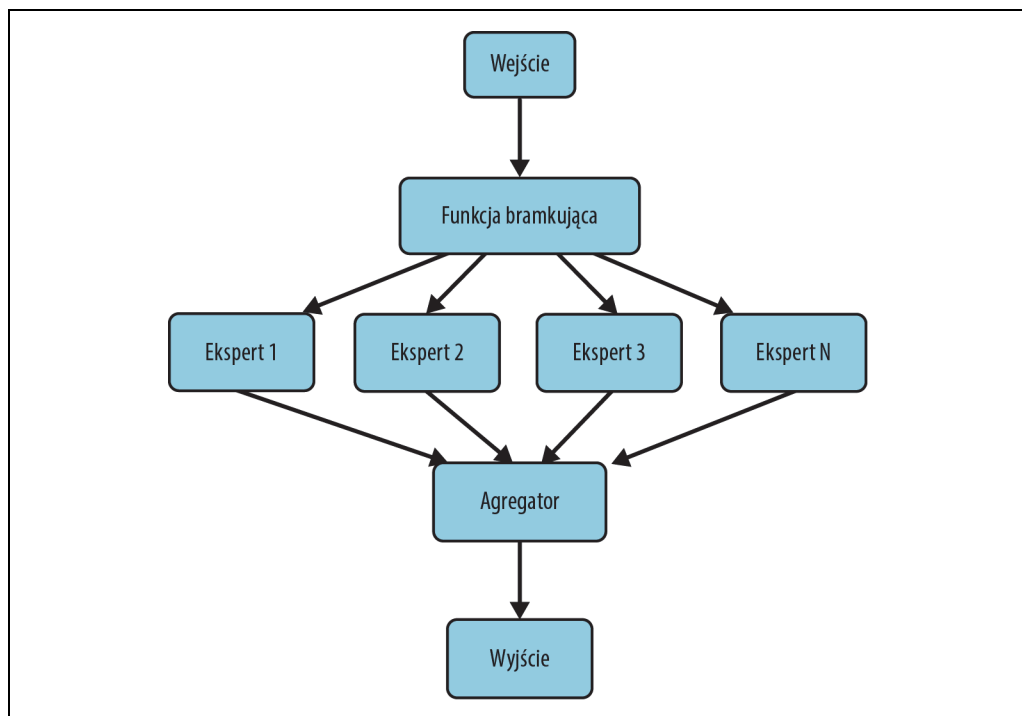
3. Model dekoderowy



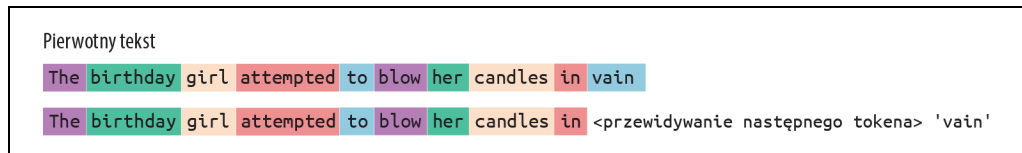
Warstwa maskowanego dekodera



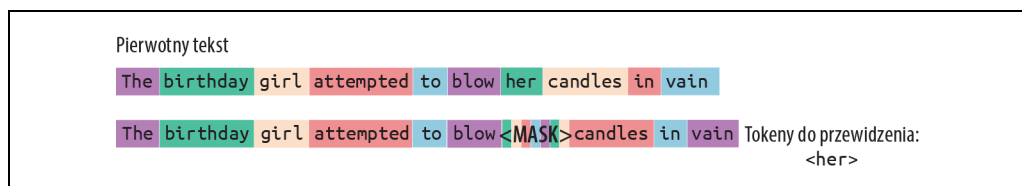
Rysunek 4.5. Wizualizacja różnych architektur transformera



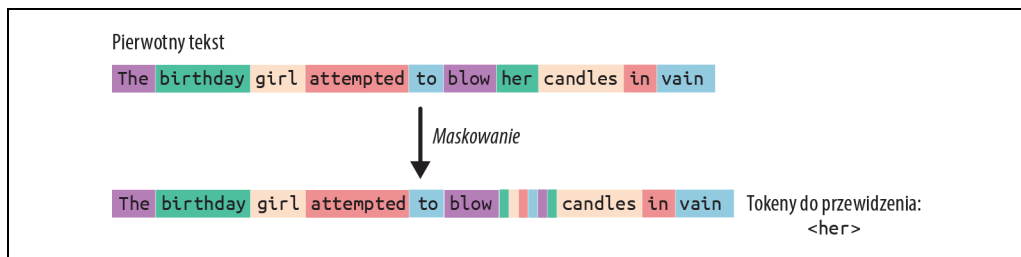
Rysunek 4.6. Kombinacja ekspertów



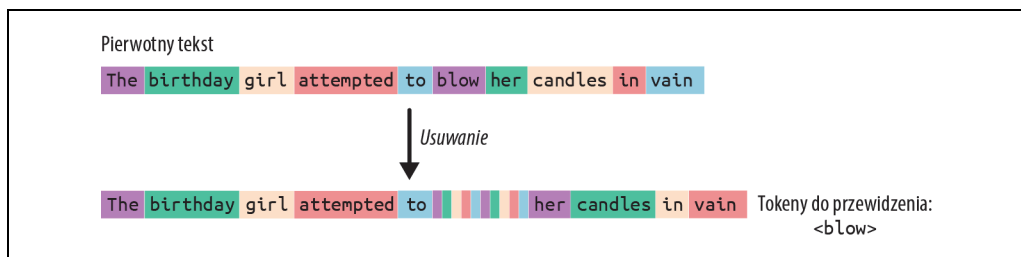
Rysunek 4.7. Pełne modelowanie języka



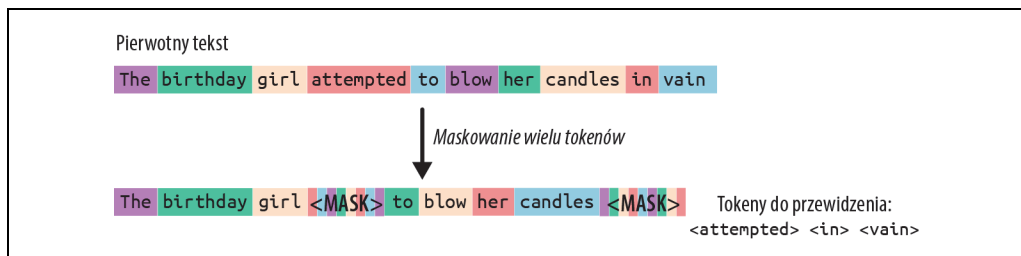
Rysunek 4.8. Maskowane modelowanie języka w modelu BERT



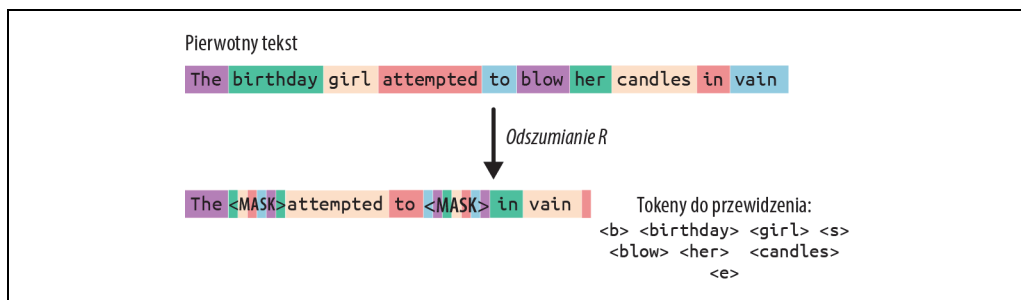
Rysunek 4.9. Maskowanie losowych tokenów w modelu BART



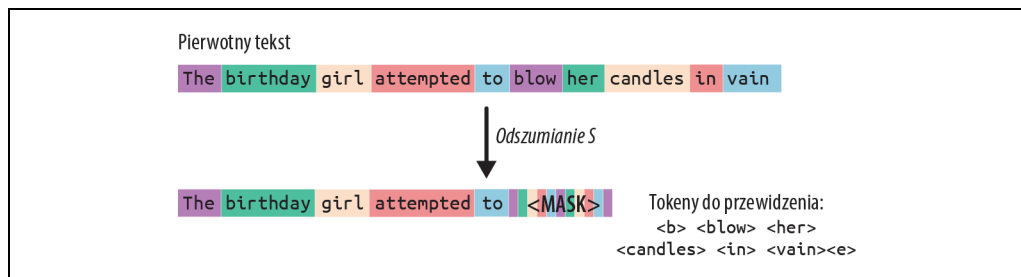
Rysunek 4.10. Usuwanie losowych tokenów w modelu BART



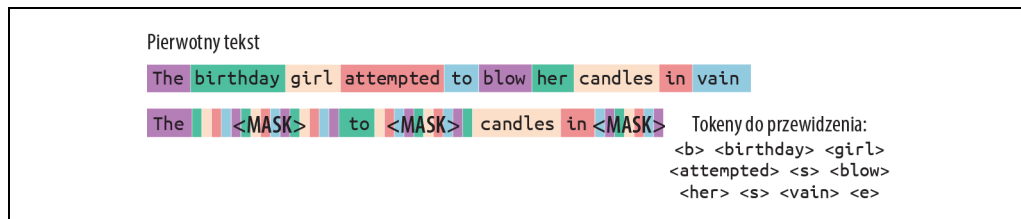
Rysunek 4.11. Maskowanie fragmentów tekstu w modelu BART



Rysunek 4.14. Odszumiacz R w UL2

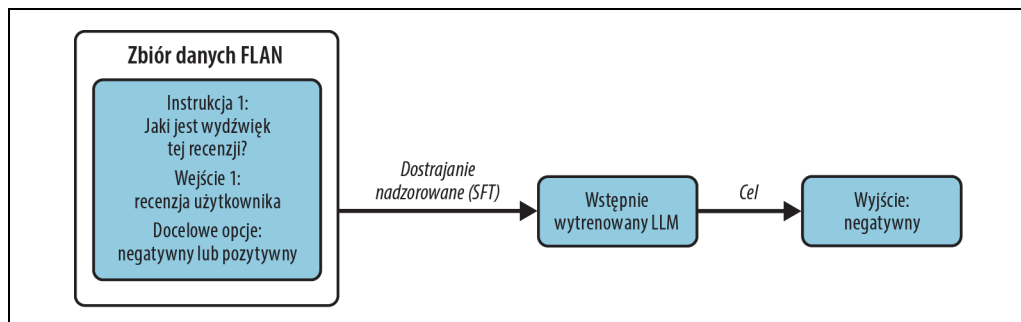


Rysunek 4.15. Odszumiacz S w UL2

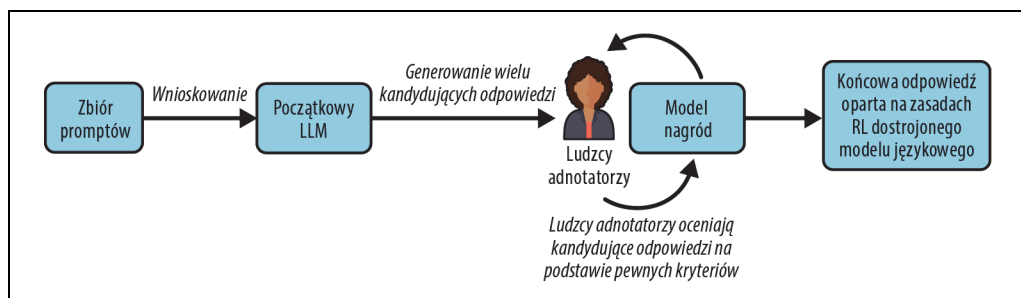


Rysunek 4.16. Odszumiacz X w UL2

Rozdział 5. Dostosowywanie modeli językowych do własnych potrzeb



Rysunek 5.1. Proces dostrajania instrukcyjnego



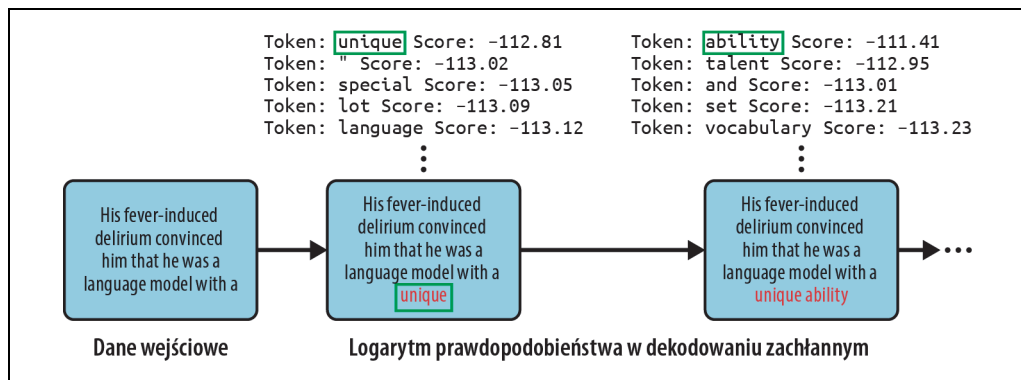
Rysunek 5.2. Uczenie przez wzmacnianie z informacją zwrotną od człowieka

	Rank	Type	Model		Average ⓘ ↕
📌	1	🔹	MazyarPanahi/calme-3.2-instruct-78b ↗	💾	● 52.08 %
📌	2	💬	MazyarPanahi/calme-3.1-instruct-78b ↗	💾	● 51.29 %
📌	3	💬	dfurman/CalmeRys-78B-Orpo-v0.1 ↗	💾	● 51.23 %
📌	4	💬	MazyarPanahi/calme-2.4-rys-78b ↗	💾	● 50.77 %
📌	5	🔹	huihui-ai/Qwen2.5-72B-Instruct-abliterated ↗	💾	● 48.11 %
📌	6	💬	Qwen/Qwen2.5-72B-Instruct ↗	💾	● 47.98 %
📌	7	💬	MazyarPanahi/calme-2.1-qwen2.5-72b ↗	💾	● 47.86 %

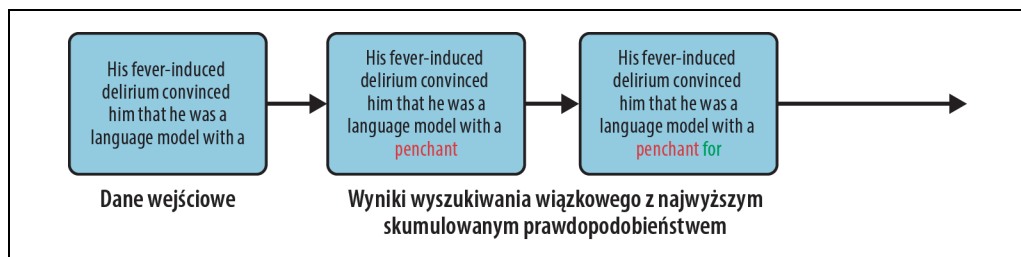
Rysunek 5.3. Zrzut ekranu z rankingiem Open LLM Leaderboard

Rank* (UB) ▲	Rank (StyleCtrl) ▲	Model ▲	Arena Score ▲
1	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
1	2	Gemini-2.0-Pro-Exp-02-05	1379
1	1	ChatGPT-4o-latest (2025-01-29)	1377
4	2	DeepSeek-R1	1361
4	7	Gemini-2.0-Flash-001	1355
4	2	o1-2024-12-17	1352
7	5	o1-preview	1335
7	7	Qwen2.5-Max	1332
9	8	DeepSeek-V3	1316
9	9	Gemini-2.0-Flash-Lite-Preview-02-05	1309

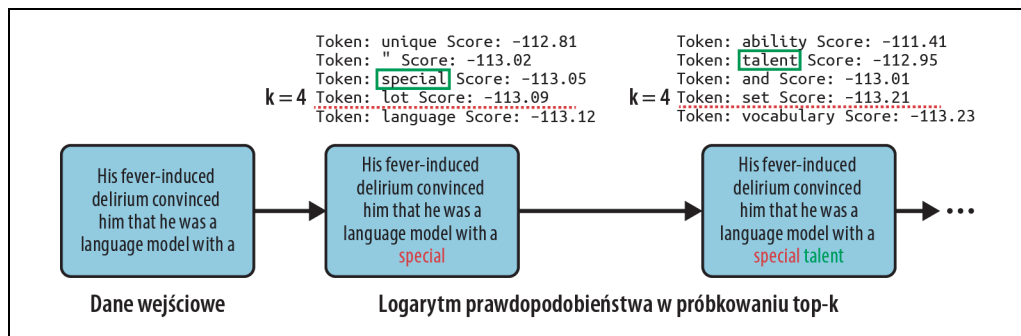
Rysunek 5.5. Zrzut ekranu z rankingiem Chatbot Arena



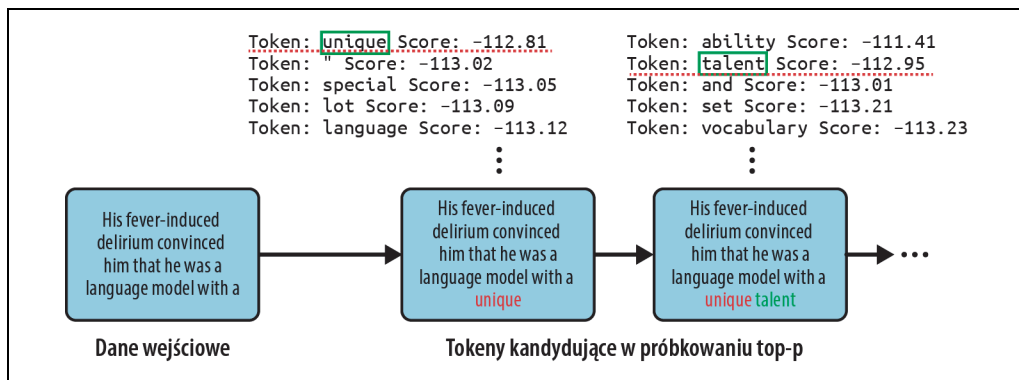
Rysunek 5.6. Dekodowanie zachłanne



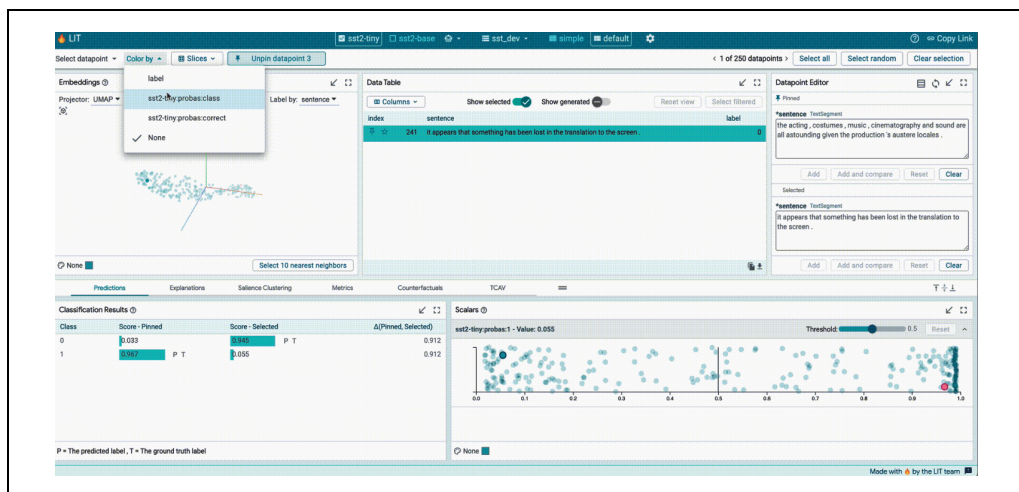
Rysunek 5.7. Wyszukiwanie wiązkowe



Rysunek 5.8. Próbkowanie top-k

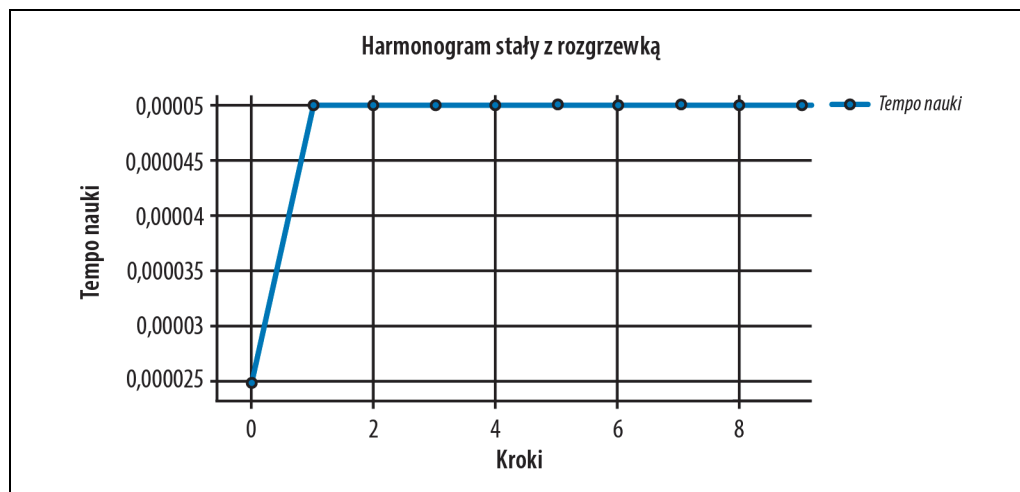


Rysunek 5.9. Próbkowanie top-p

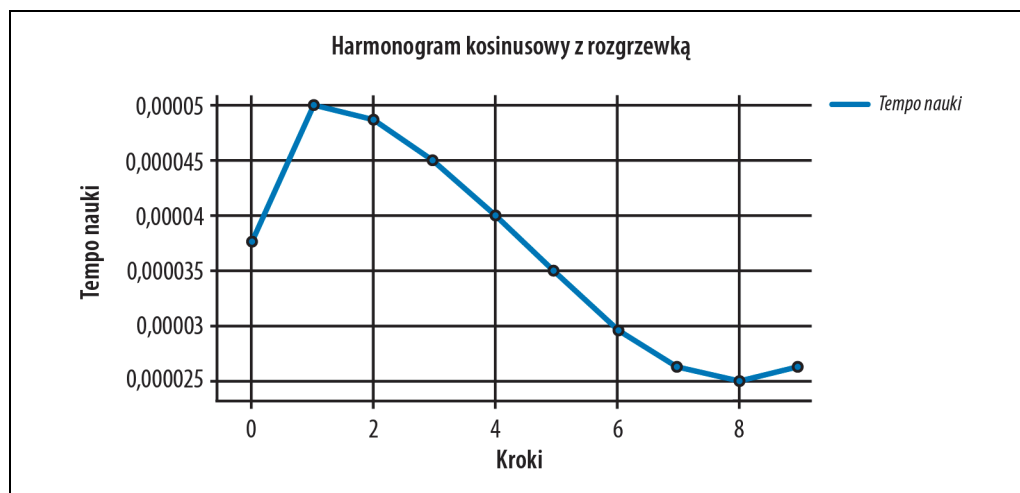


Rysunek 5.10. LIT-NLP

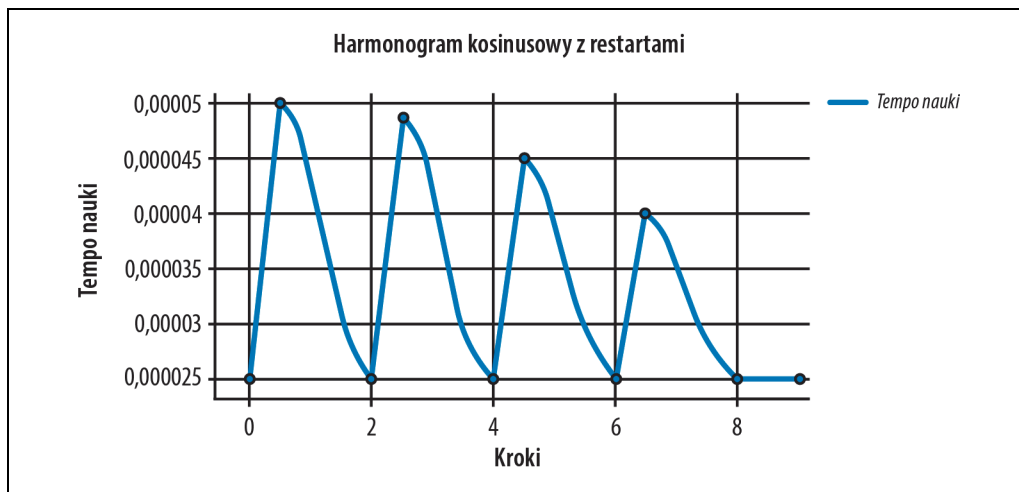
Rozdział 6. Dostrajanie



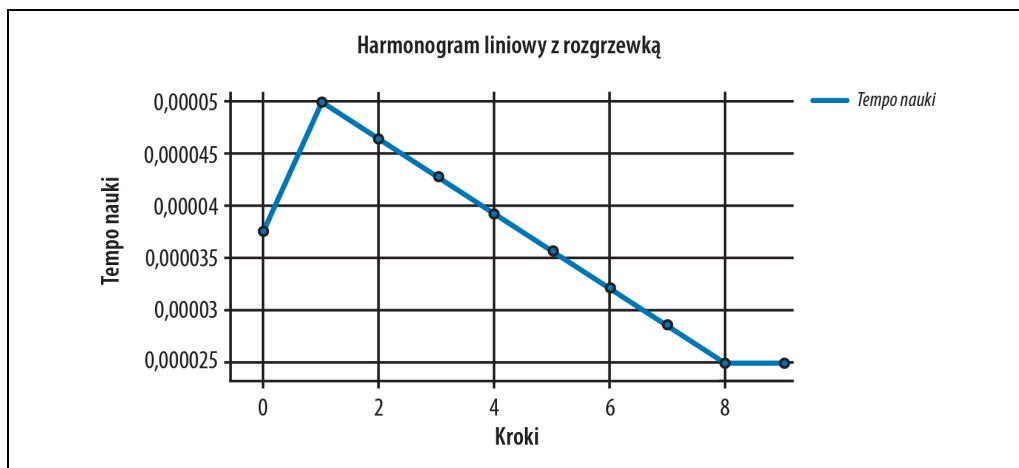
Rysunek 6.1. Tempo nauki w harmonogramie stałym z rozgrzewką



Rysunek 6.2. Tempo nauki w harmonogramie kosinusowym z rozgrzewką



Rysunek 6.3. Tempo nauki w harmonogramie kosinusowym z restartami

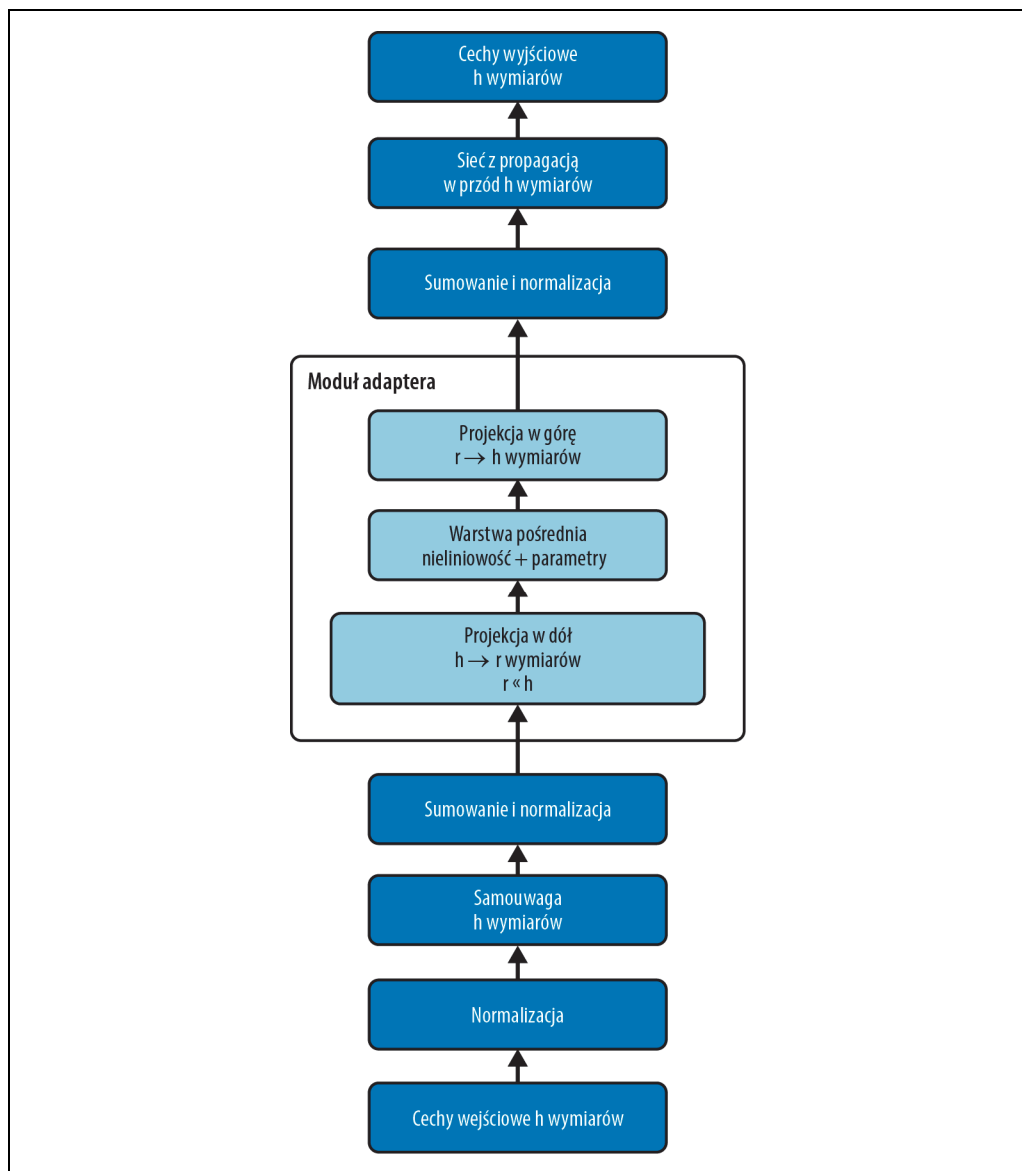


Rysunek 6.4. Tempo nauki w harmonogramie liniowym

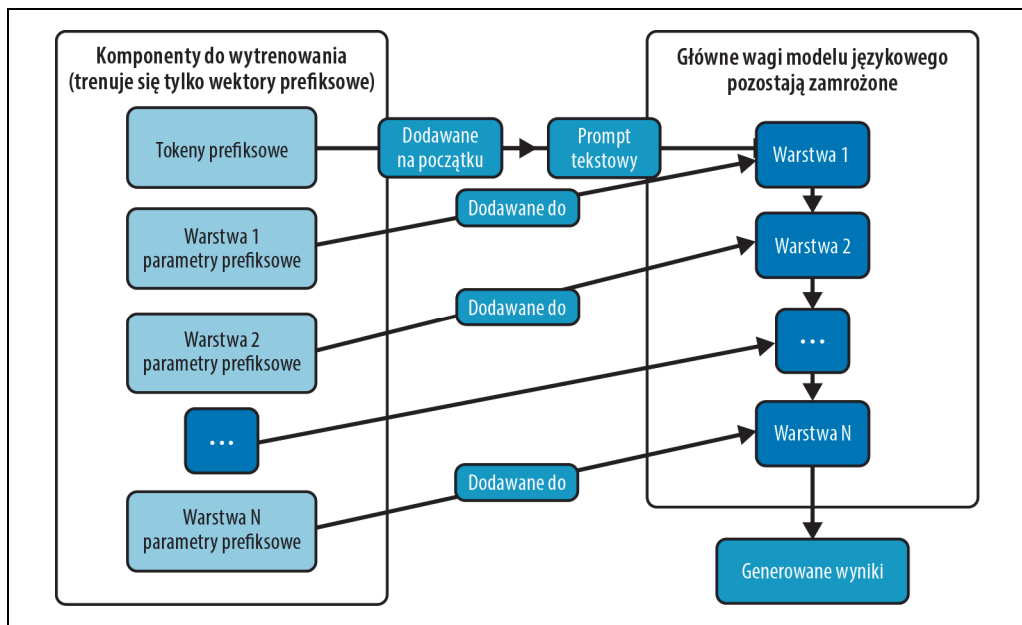
Rozdział 7. Zaawansowane techniki dostrajania modeli



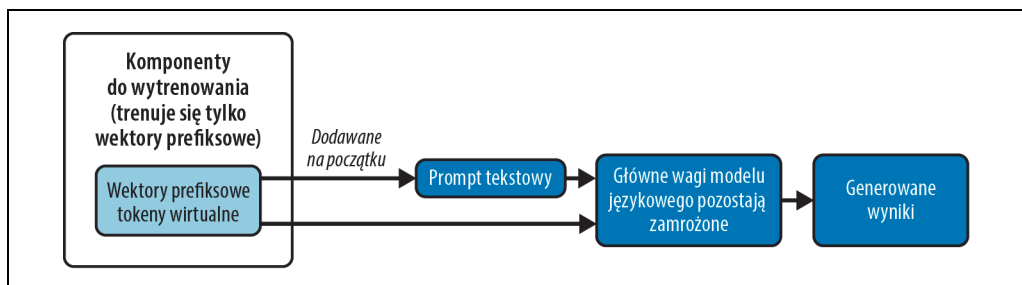
Rysunek 7.1. Procesu ciągłego treningu wstępnego



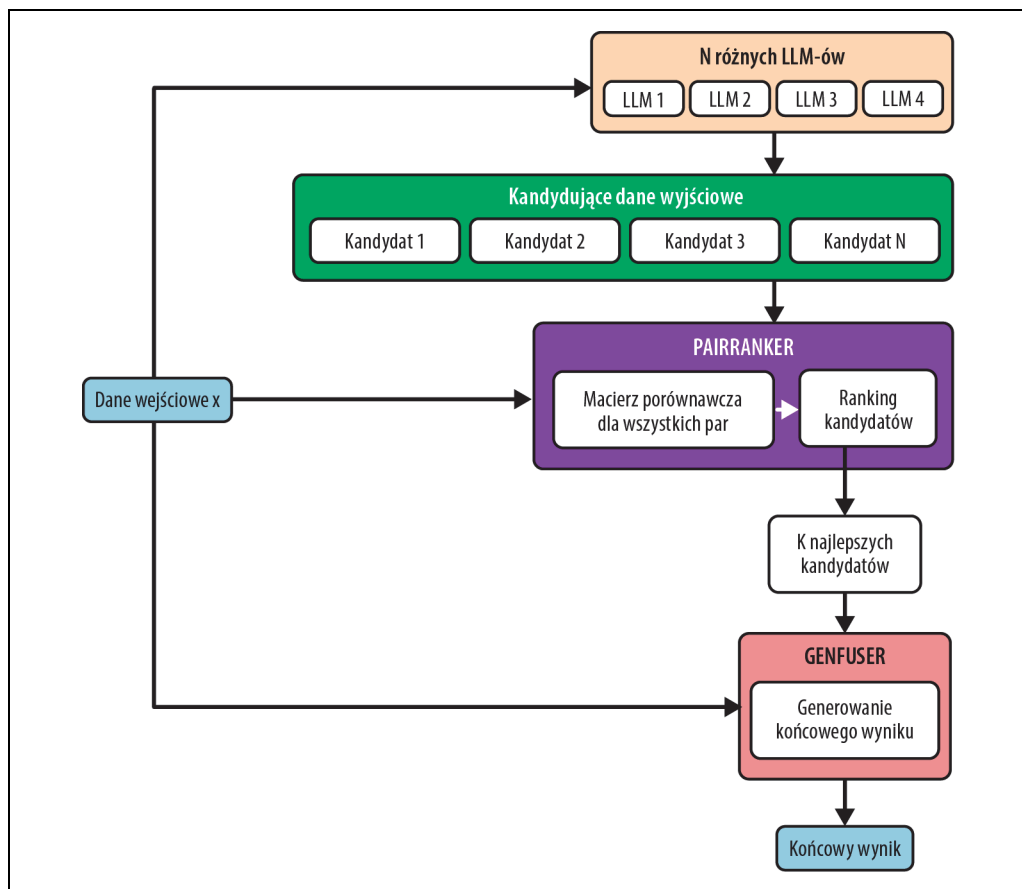
Rysunek 7.2. Moduły adaptera w transformerze



Rysunek 7.3. Dostrajanie prefiksowe

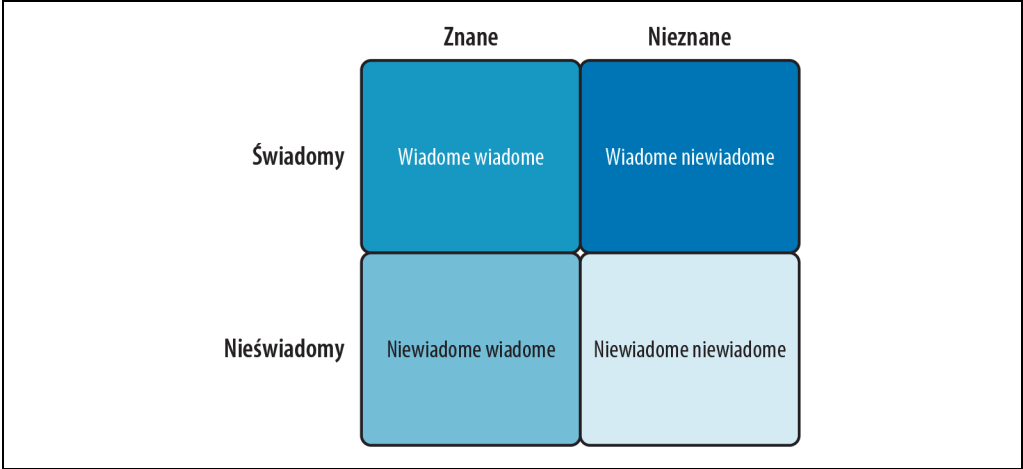


Rysunek 7.4. Dostrajanie promptowe



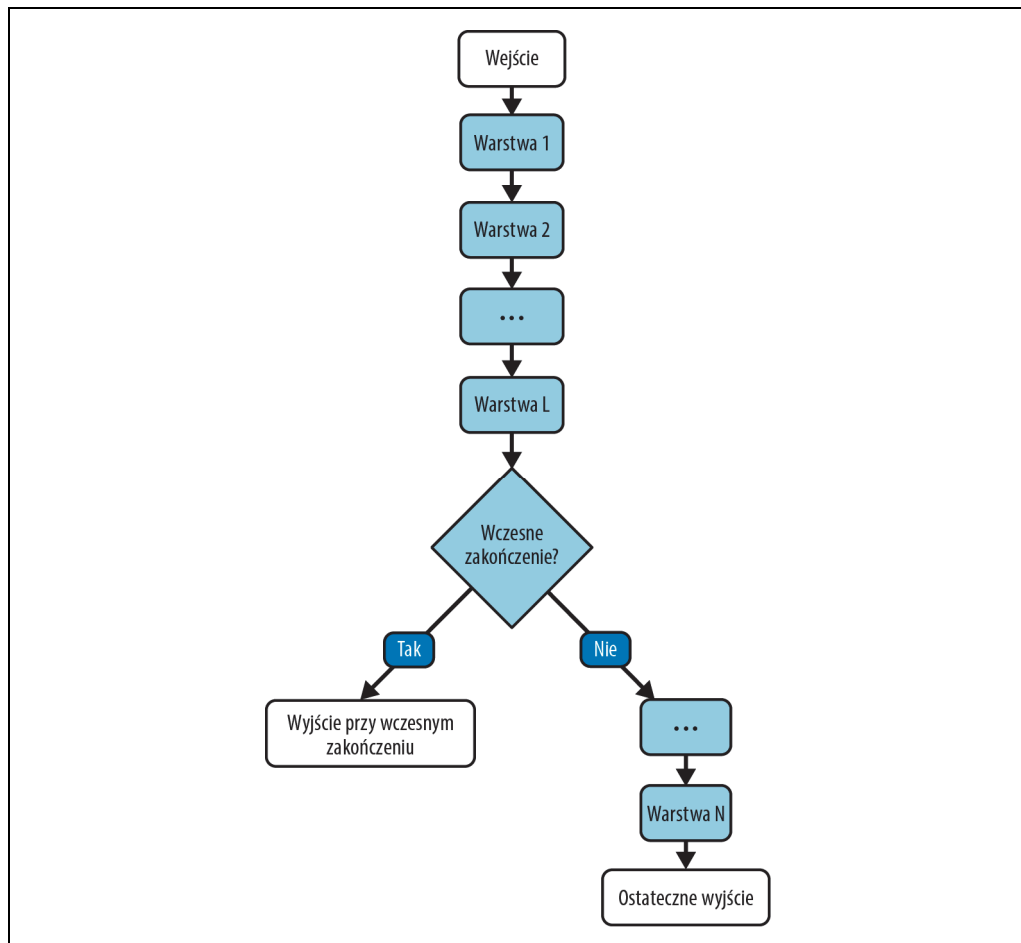
Rysunek 7.5. LLM-Blender

Rozdział 8. Trening dostosowawczy i rozumowanie

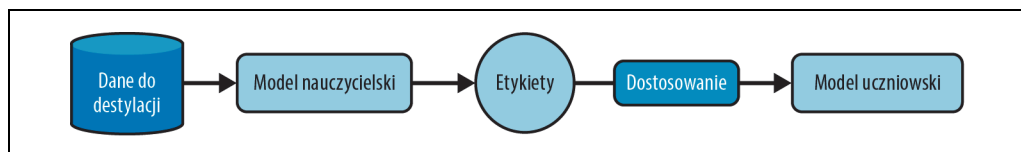


Rysunek 8.1. Kwadrant wiedzy

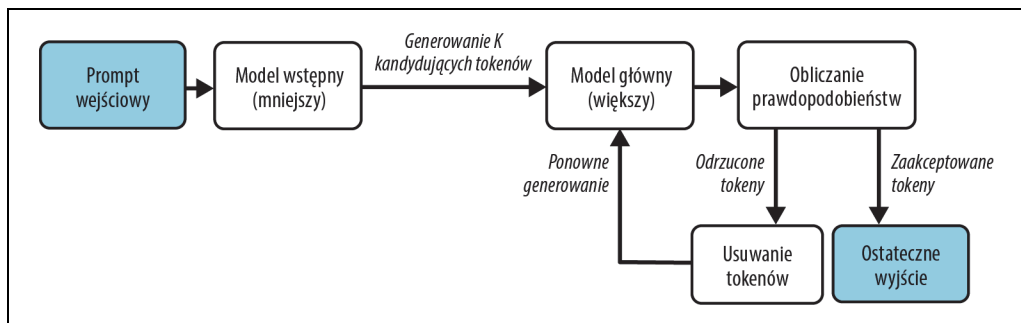
Rozdział 9. Optymalizacja wnioskowania



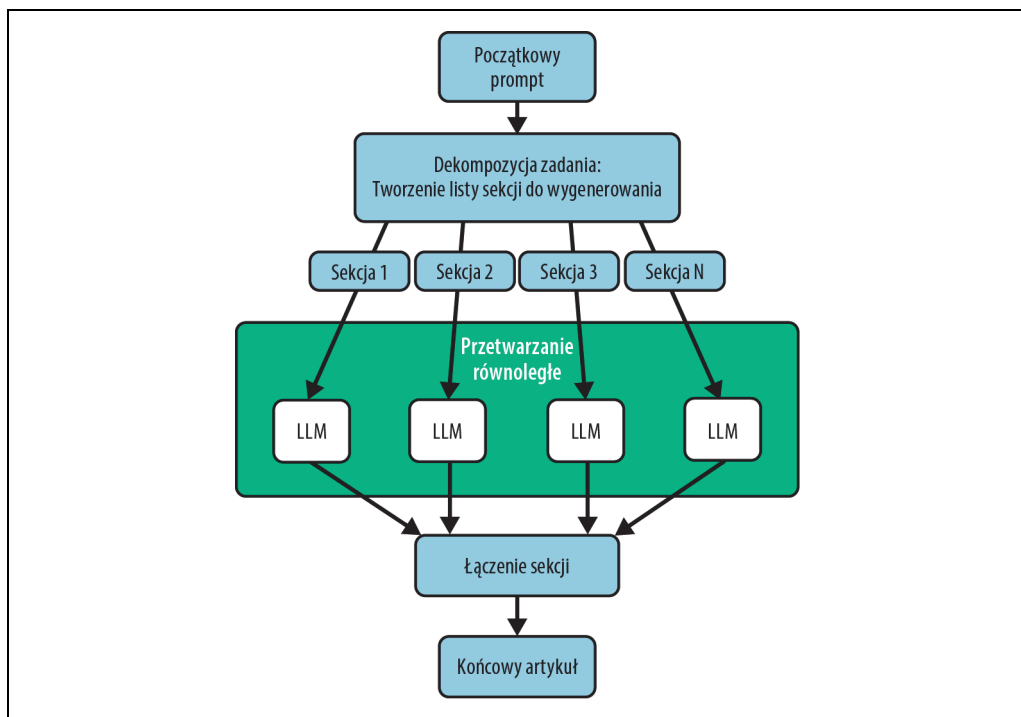
Rysunek 9.1. Wczesne kończenie



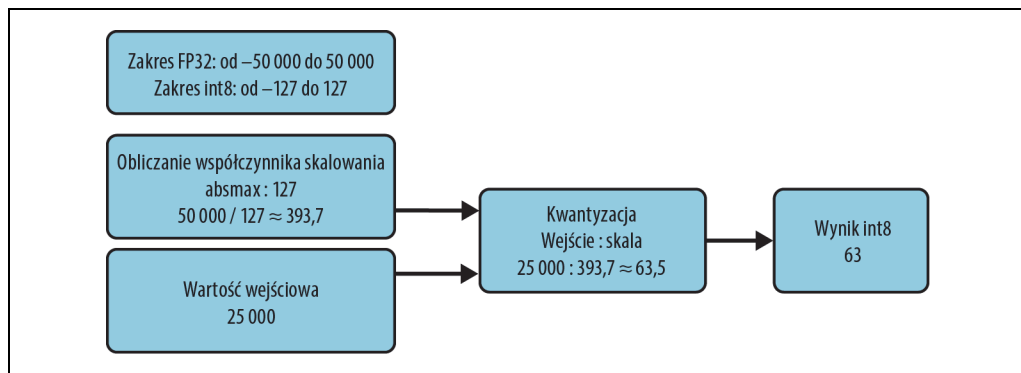
Rysunek 9.2. Destylacja wiedzy



Rysunek 9.3. Dekodowanie spekulacyjne

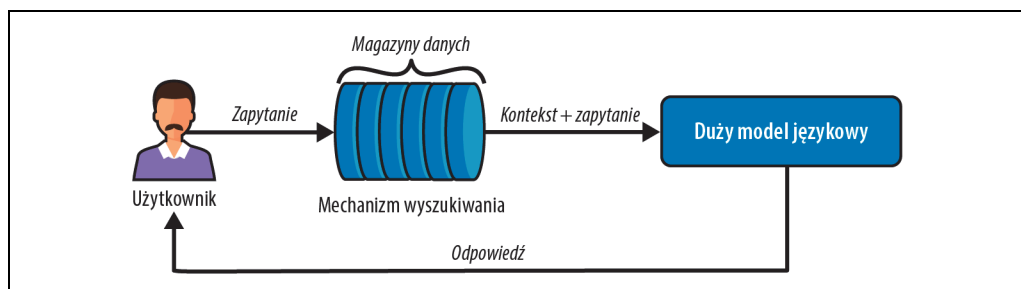


Rysunek 9.4. Dekodowanie równoległe

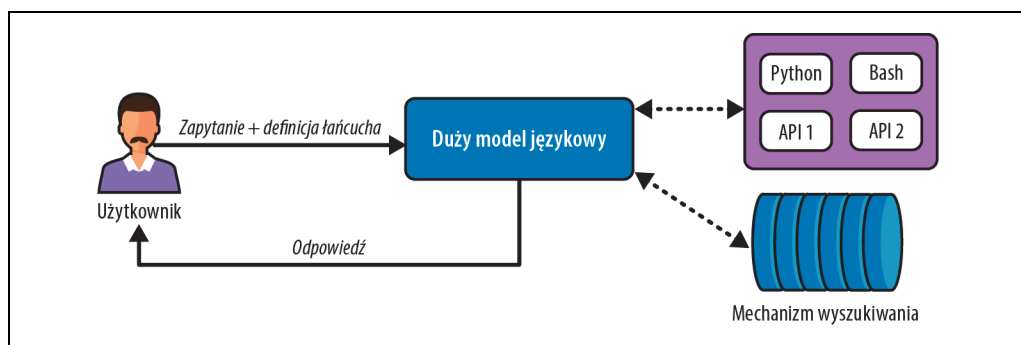


Rysunek 9.5. Kwantyzacja absmax

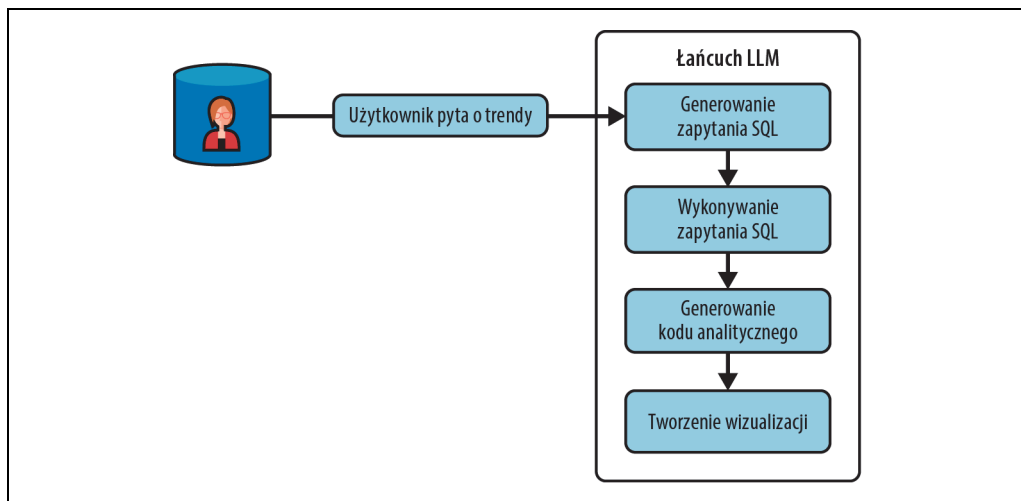
Rozdział 10. Łączenie LLM-ów z narzędziami zewnętrznymi



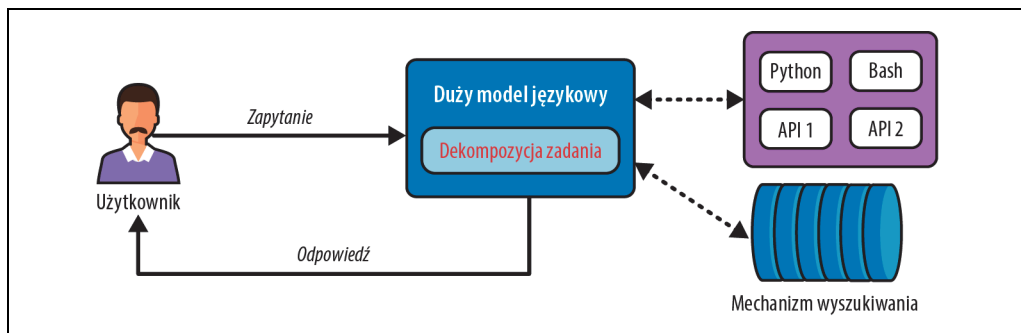
Rysunek 10.1. LLM pasywnie współdziałający z magazynem danych



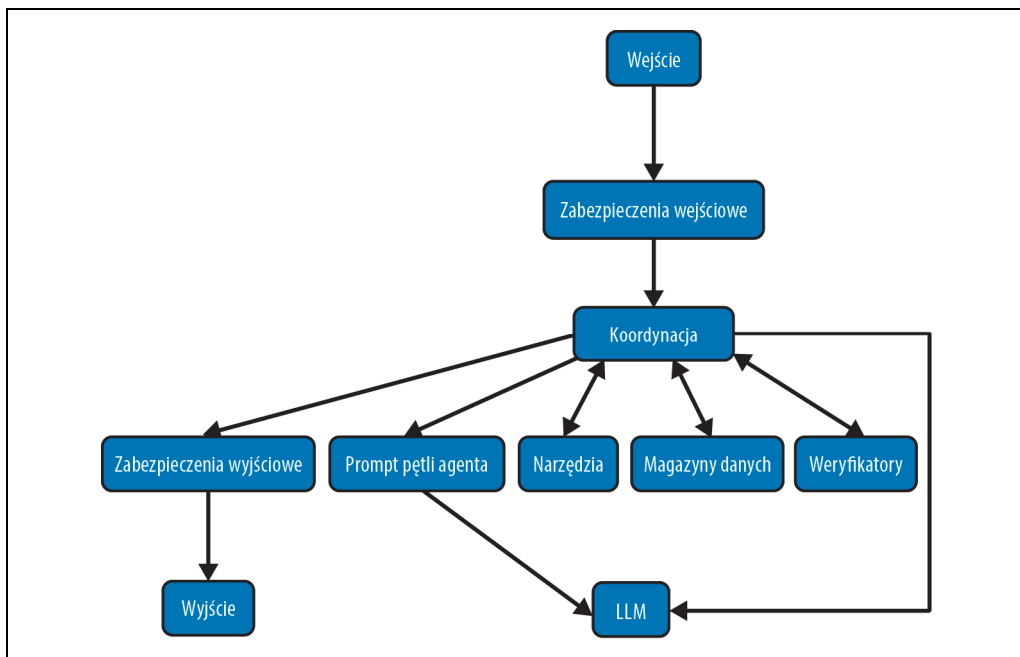
Rysunek 10.2. Podejście z interakcjami jawnymi



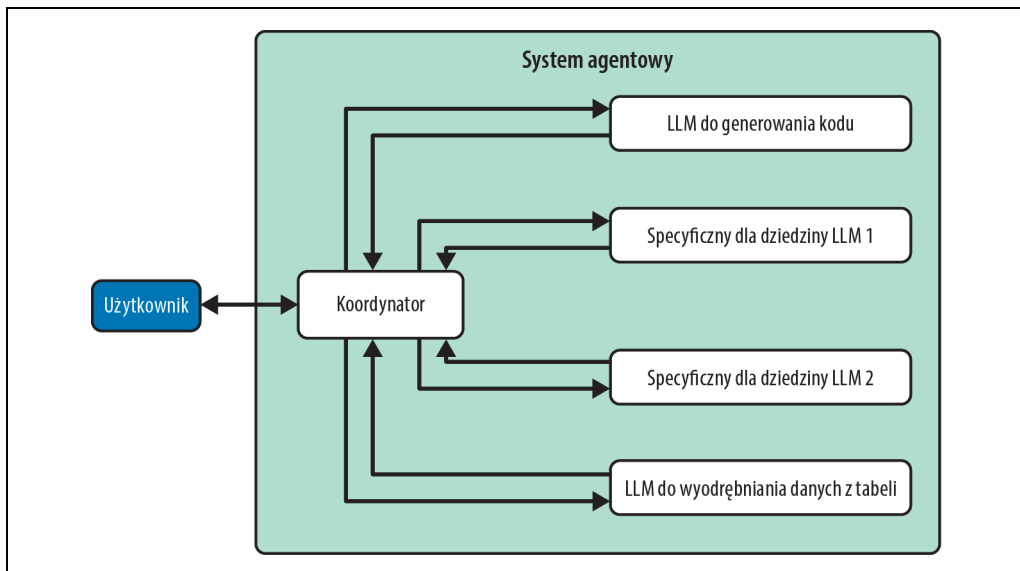
Rysunek 10.3. Przykładowy przepływ pracy asystenta analityka danych



Rysunek 10.4. Typowy przepływ pracy autonomicznego agenta opartego na LLM-ie

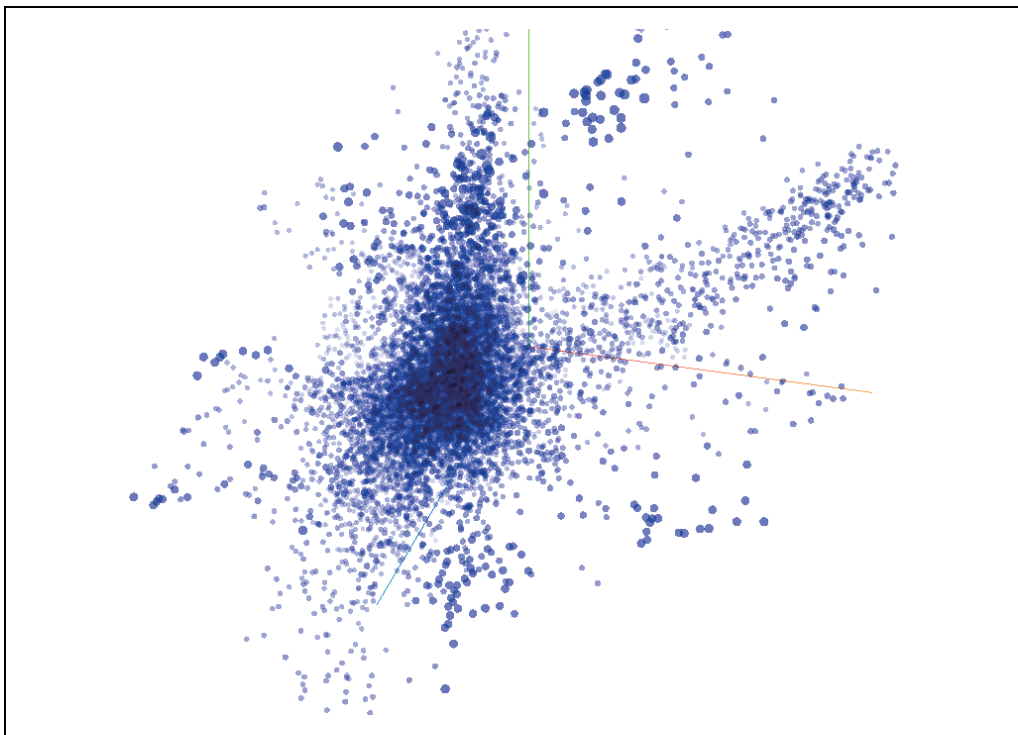


Rysunek 10.5. System agentowy klasy produkcyjnej

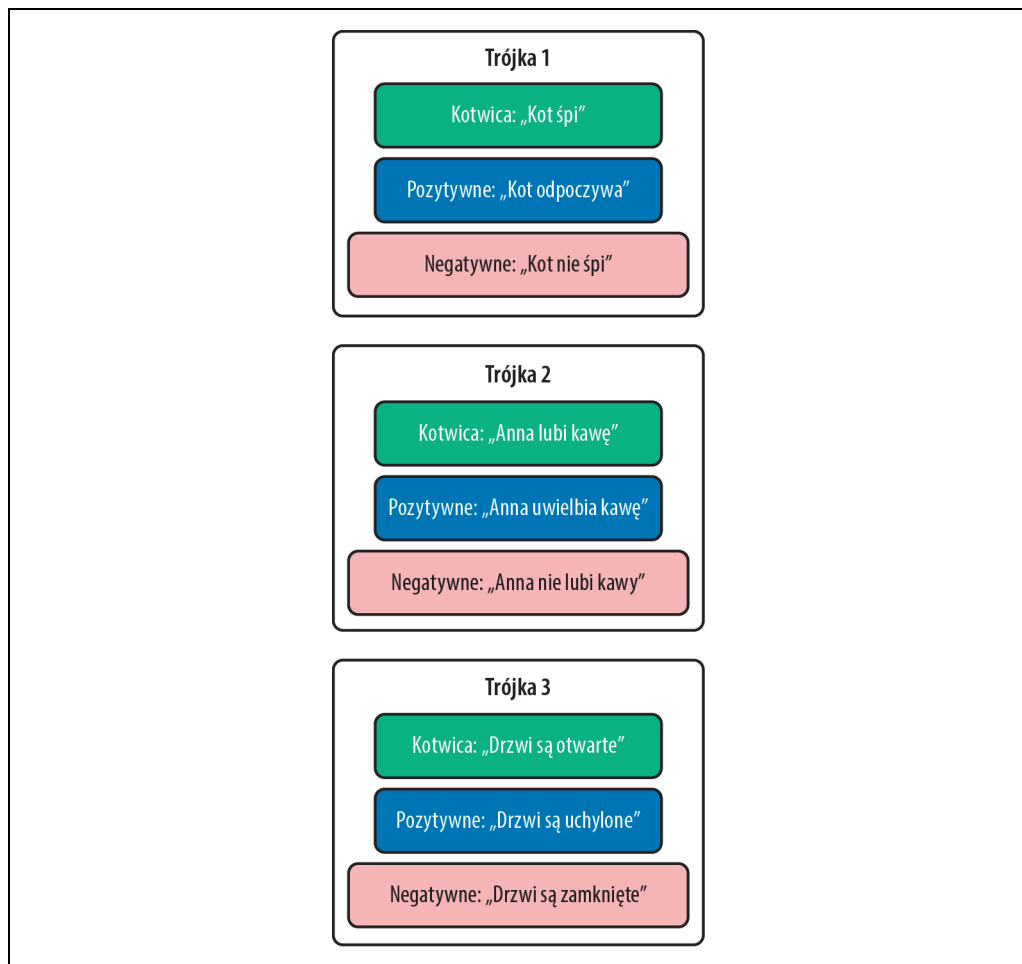


Rysunek 10.6. System agentowy z wieloma LLM-ami

Rozdział 11. Uczenie reprezentacji i osadzenia

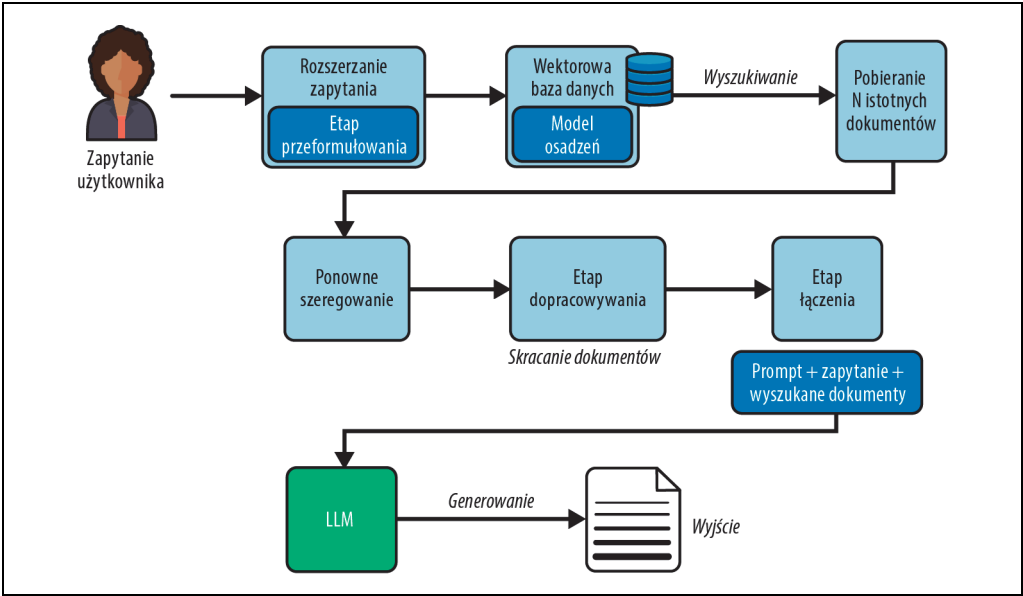


Rysunek 11.1. Wizualizacja przestrzeni osadzeń

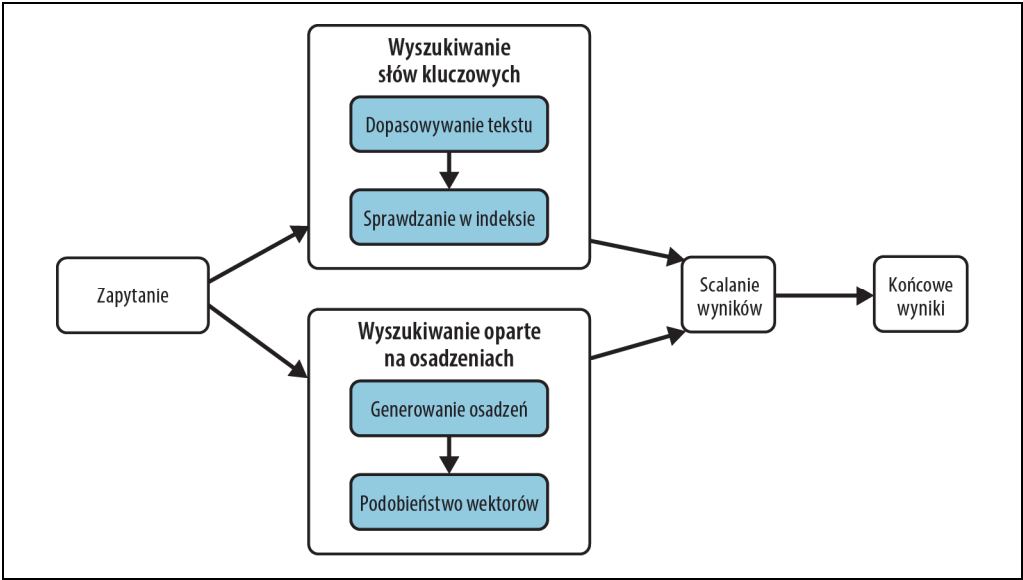


Rysunek 11.2. Zbiór danych do dostrajania pod kątem negacji

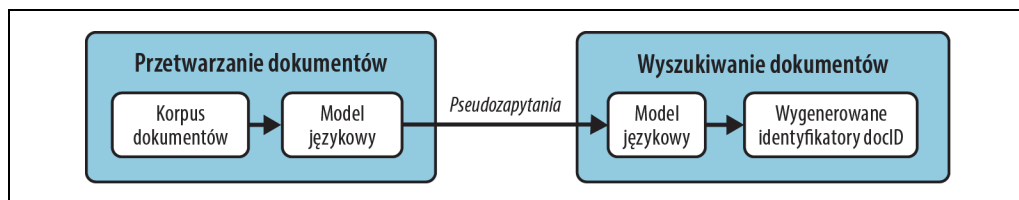
Rozdział 12. Generowanie wspomagane wyszukiwaniem



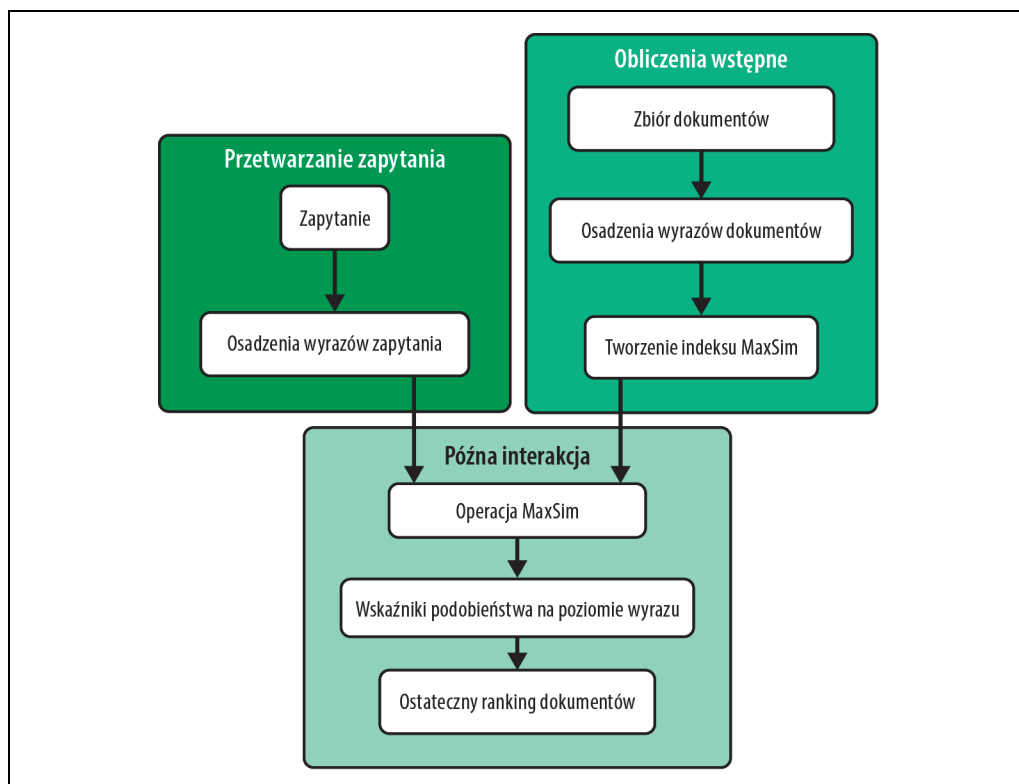
Rysunek 12.1. Potok RAG



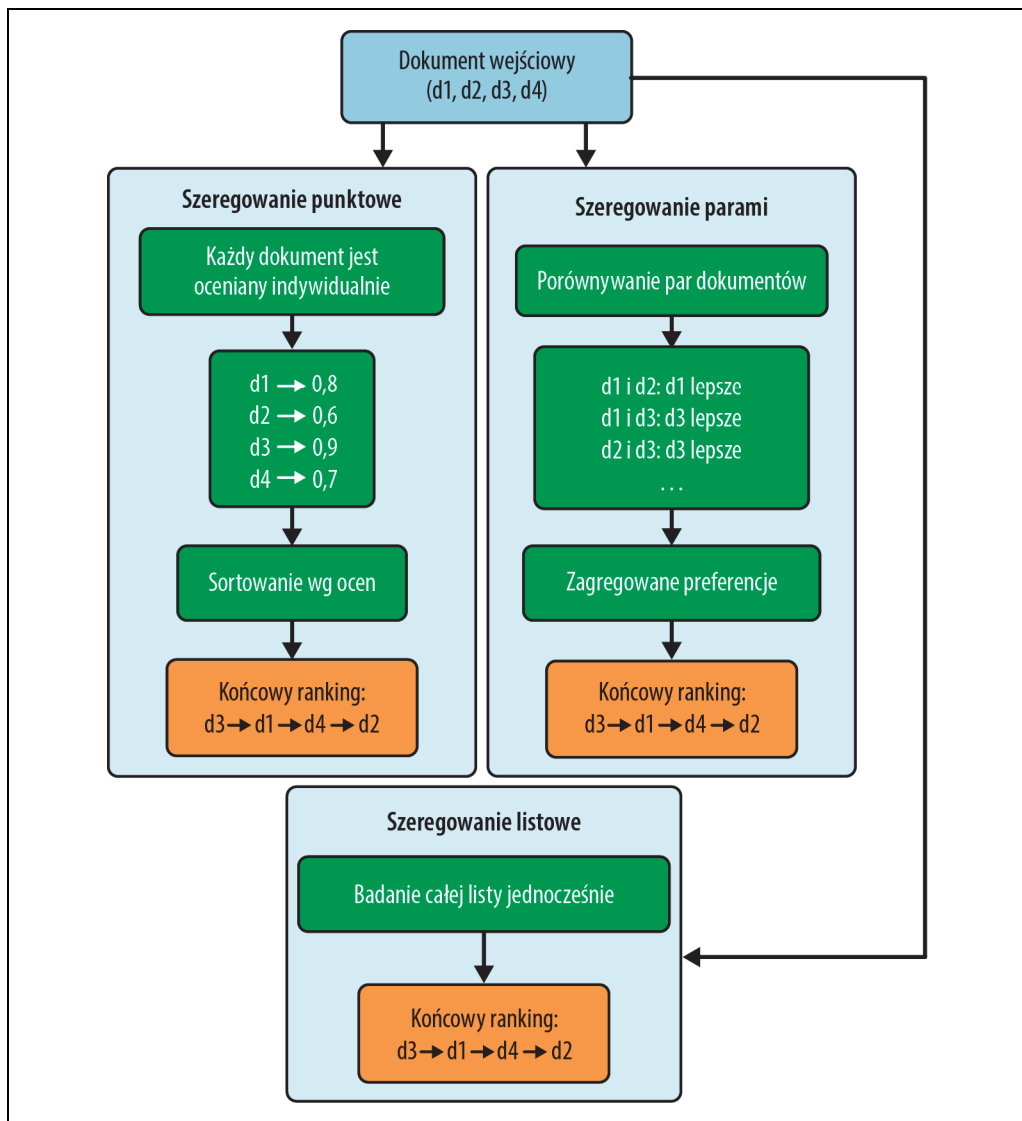
Rysunek 12.2. Wyszukiwanie hybrydowe



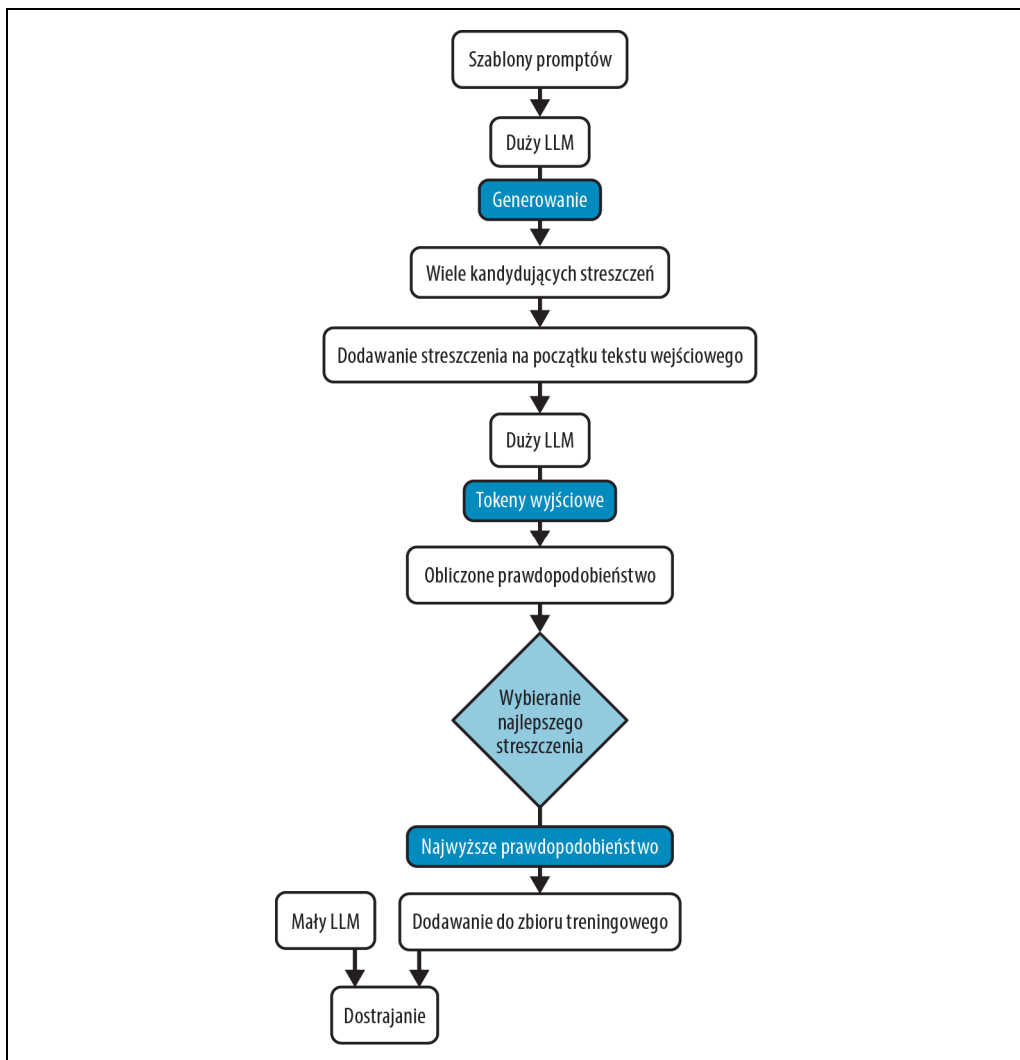
Rysunek 12.3. Wyszukiwanie generatywne



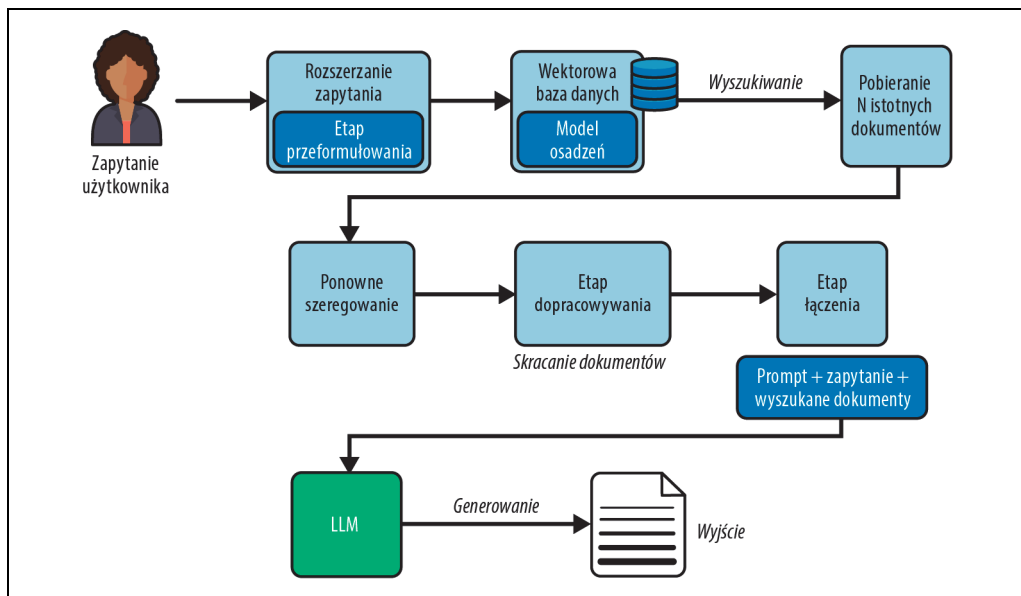
Rysunek 12.4. ColBERT



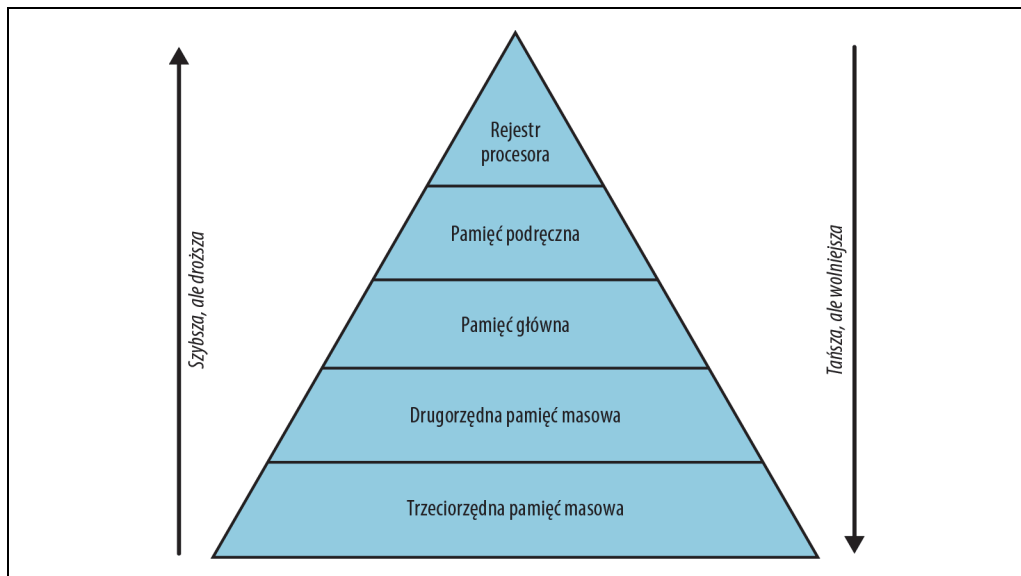
Rysunek 12.5. Ponowne szeregowanie z użyciem LLM-a dekodującego



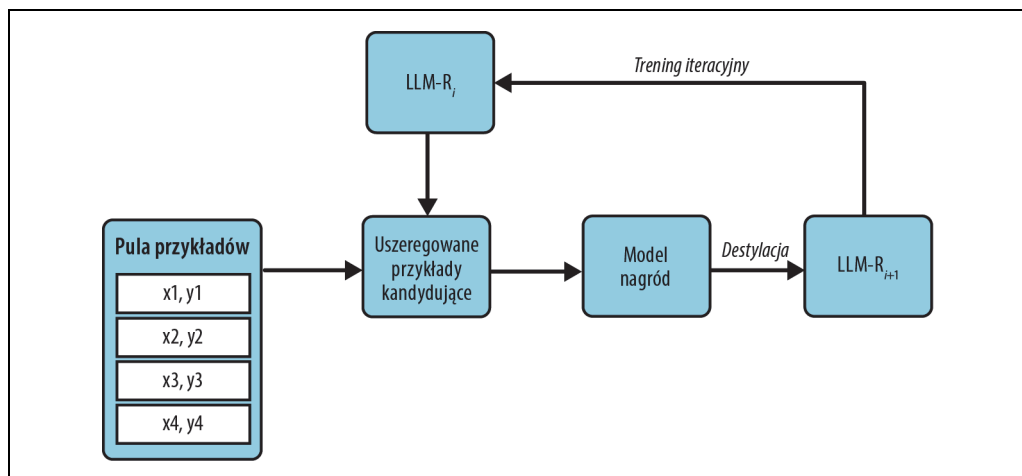
Rysunek 12.6. Streszczanie abstrakcyjne



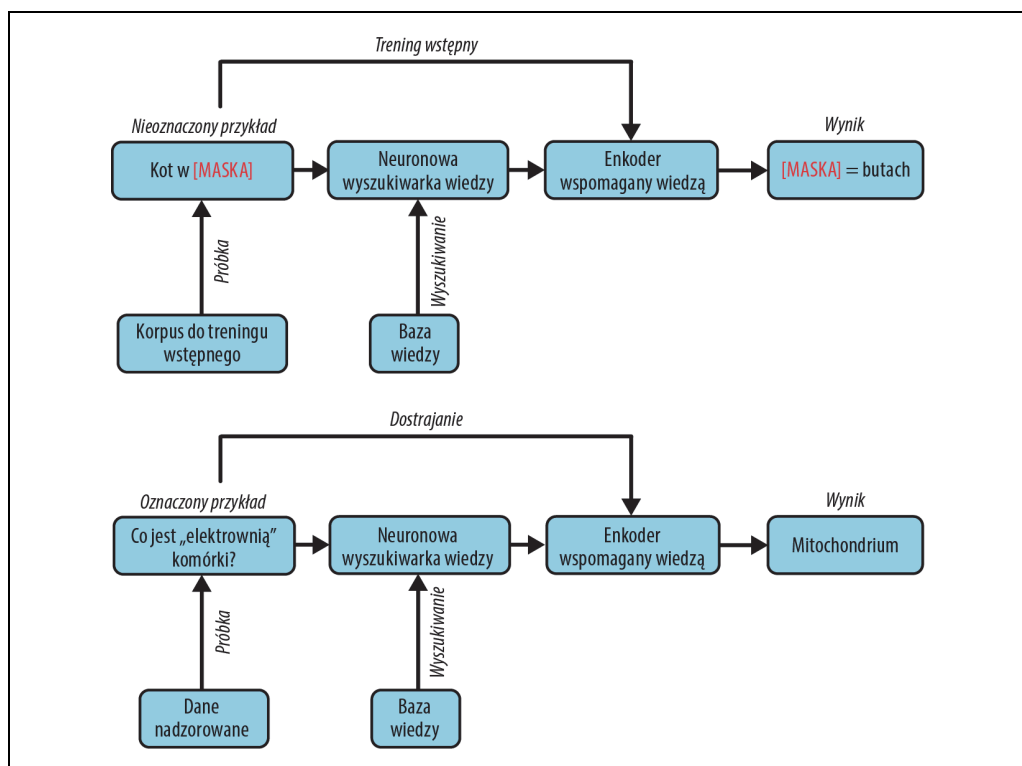
Rysunek 12.7. Kompleksowy potok RAG



Rysunek 12.8. Typowa hierarchia pamięci w systemie operacyjnym

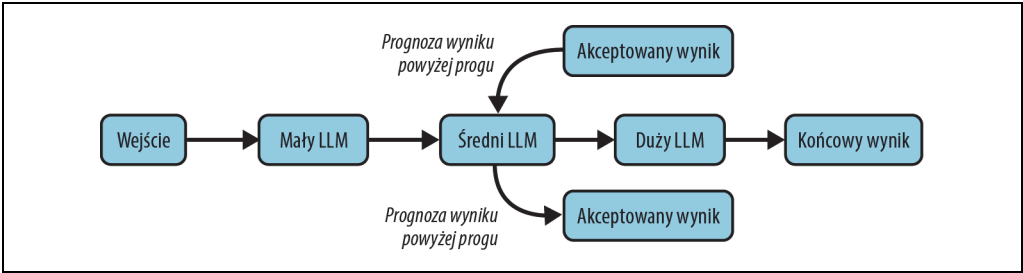


Rysunek 12.9. Proces LLM-R

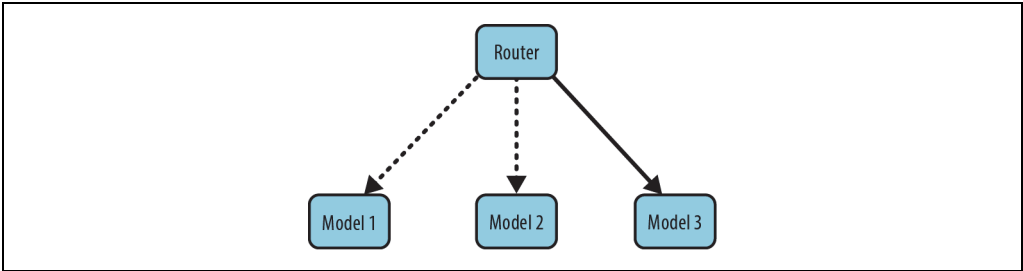


Rysunek 12.10. Architektura REALM

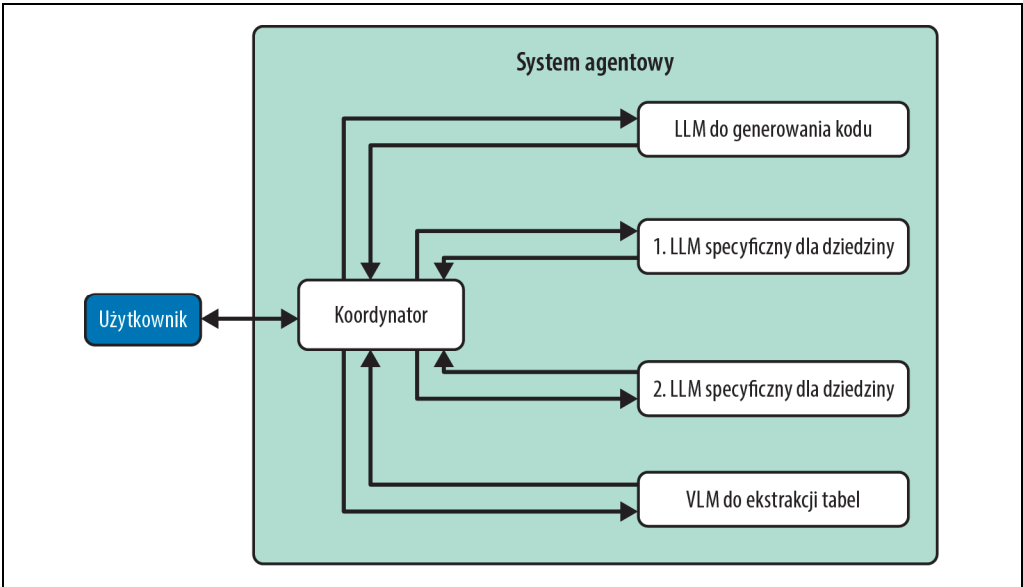
Rozdział 13. Wzorce projektowe i architektura systemów



Rysunek 13.1. Kaskada LLM-ów



Rysunek 13.2. Router



Rysunek 13.3. LLM-y wyspecjalizowane w konkretnych zadaniach