sages

Retrieval Augmented Generation (RAG)
systemy AI do wyszukiwania informacji

Bartosz Mikulski
mikulskibartosz.name
aihallucinationfix.substack.com

sages

Zaufaj naszemu doświadczeniu

18 lat
na rynku usług IT

500+
klientów biznesowych

29 560+
przeszkolonych osób

4 877+
zorganizowanych szkoleń i warsztatów

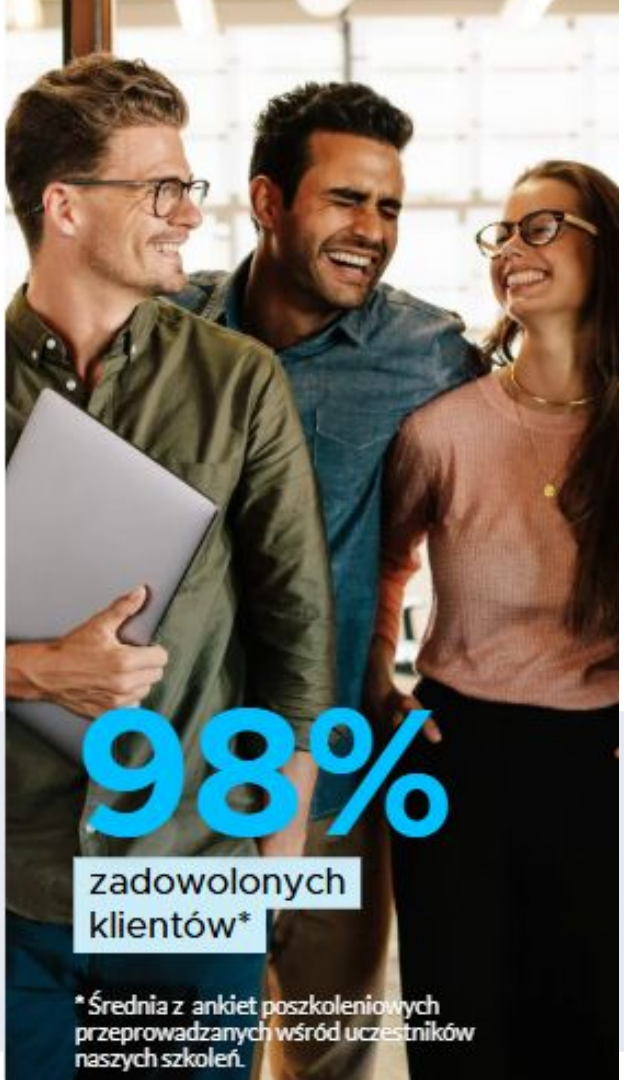
GWARANCJA JAKOŚCI USŁUG

Diamenty Wielkopolski Forbes 2023

ISO 9001:2015 CERTIFIED

STANDARD USŁUGI SZKOLENIOWEJ

Baza Usług Rozwojowych



Edukacja na najwyższym poziomie

Wiedza specjalistyczna dla branży IT

Oferujemy szeroki katalog szkoleń z technologii mainstreamowych
i specjalistycznych, wschodzących i legacy. Zajęcia prowadzimy w trybie warsztatowym, a programy są oparte o praktyczne know-how. Specjalizujemy się w prowadzeniu dedykowanych szkoleń technologicznych, których agendę dostosowujemy do potrzeb naszych klientów i oczekiwań uczestników.

Wybitni eksperci

Od początku naszego istnienia przeszkoliliśmy dziesiątki tysięcy osób, co pomogło absolwentom podnieść konkurencyjność na rynku pracy i jakość projektów realizowanych na co dzień. Nasze szkolenia prowadzą najlepsi trenerzy, a nasze produkty są oparte na najnowocześniejszej technologii. Ich niezawodność i dopasowanie do potrzeb klientów są możliwe dzięki zespołowi składającemu się z wybitnych ekspertów i ekspertek, którzy/e są na pierwszej linii teorii i praktyki tworzenia i wdrażania innowacji technologicznych.

NASZE POZOSTAŁE MARKI EDUKACYJNE

STACJA.IT

kodoamacz by sages

sages

Najlepsze standardy usług edukacyjnych

Stosujemy Standard Usługi Szkoleniowej Polskiej Izby Firm Szkoleniowych, a nasze usługi realizowane są na najwyższym poziomie, o czym świadczą stale powracający klienci oraz wdrożony certyfikat ISO 9001. Metodologia prowadzonych przez nas zajęć oparta jest na współczesnych narzędziach i dostosowana do potrzeb i oczekiwań klientów.

Ponadto jesteśmy firmą wpisaną do rejestru instytucji szkoleniowych w Wojewódzkim Urzędzie Pracy w Warszawie pod nr 2.14/00133/2019.

Zaufali nam najlepsi

Wśród naszych klientów są takie firmy jak Alior Bank, OLX Group, Bank Zachodni WBK, Orange Polska, Lufthansa i wiele innych.

STUDIA PODYPLOMOWE

Wspieramy organizację zaawansowanych kierunków studiów podyplomowych. Realizujemy zajęcia na kierunkach: Data Science, Big Data i Wizualna analityka danych, AI & Data Driven Business oraz User Experience Design – projektowanie doświadczeń cyfrowych.

Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechniki Warszawskiej

AKADEMIA LEONA KOŹMIŃSKIEGO

Sylwetka trenera

Bartosz Mikulski

Informacje o trenerze:

https://mikulskibartosz.name/about

Szkolenie zdalne - zasady ogólne

1. Znaleźć spokojne i wygodne miejsce, aby móc realizować ćwiczenia i warsztaty.
2. Używamy naszych imion podczas trwania szkolenia, zamiast abstrakcyjnych loginów
3. Proszę o wyciszenie mikrofonów, kamera nie musi być włączona - w zależności od jakości połączenia nawet wyłączona.

Praca w grupie i dyskusje

1. Można podnieść rękę lub włączyć mikrofon i poczekać, aż zauważę :)
2. Pauzy to czas na reakcję lub wykonanie ćwiczenia - w tym czasie jestem dostępny do pomocy.
3. Jesteśmy otwarci na ciekawostki, komentarze są mile widziane.

Jeśli pojawią się problemy :(

1. Słaba jakość transmisji, przerwania lub inne kłopoty - proszę od razu alarmować.
2. Zniknięcie prowadzącego nie oznacza końca szkolenia ;) Dajemy sobie minutę na powrót do szkolenia, w przeciwnym przypadku umawiamy się na kilkuminutową przerwę.

Uwaga

Wszelkie materiały (treści tekstowe, wideo, ilustracje, zdjęcia itp.) wchodzące w skład szkoleń, kursów i webinarów organizowanych przez Sages są objęte prawem autorskim i podlegają ochronie na mocy Ustawy o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (tekst ujednolicony: Dz.U. 2006 nr 90 poz. 631). Kopiowanie, przetwarzanie, rozpowszechnianie tych materiałów w całości lub w części jest zabronione.

Szkolenie zdalne - przerwy

1. Każdego dnia przerwa obiadowa w okolicy godz. 13 trwająca ~45 minut.
2. Kilkuminutowe przerwy pomiędzy większymi sekcjami szkolenia.

sages

Moduł 1: Bazowy RAG

Co to jest RAG?
Jak działają bazy wektorowe?
Wprowadzenie do Llama-index

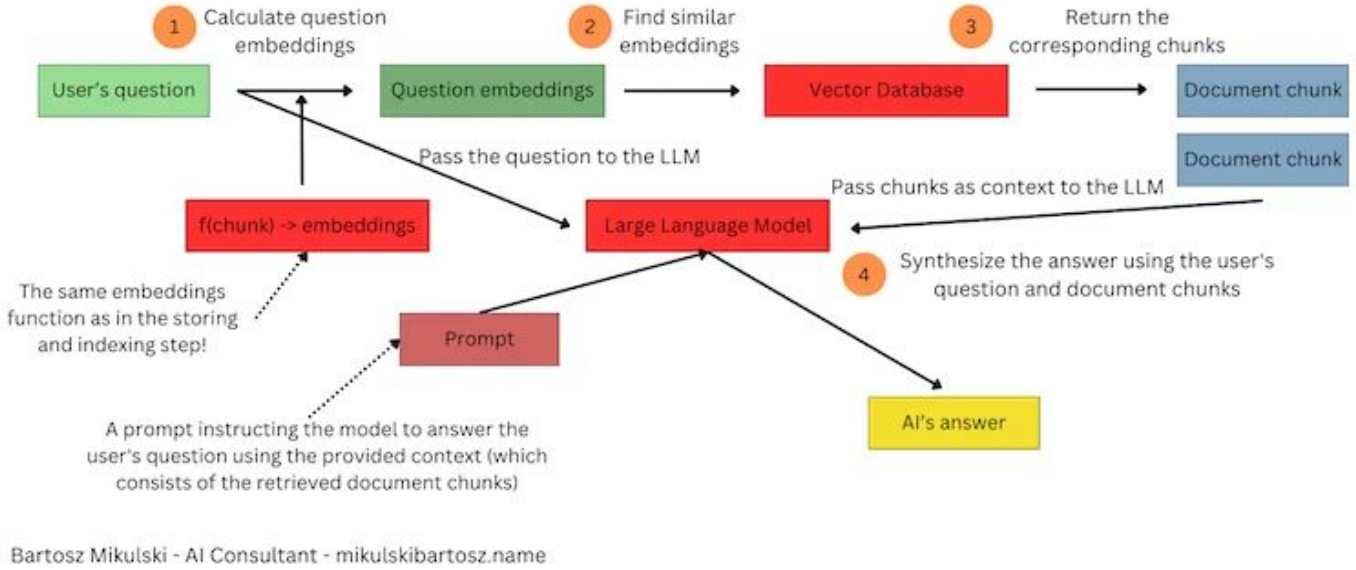
Źródła:
mikulskibartosz.name/advanced-rag-techniques-explained
https://mikulskibartosz.name/approximate-nearest-neighbor-vs-rag

Co to jest RAG?

- RAG łączy wyszukiwanie (retrieval) i generowanie (generation) odpowiedzi przez model językowy.
- Najpierw mechanizm retrievera wybiera najbardziej pasujące fragmenty z bazy wiedzy.
- Wybrane treści są przekazywane do modelu, który buduje końcową odpowiedź.

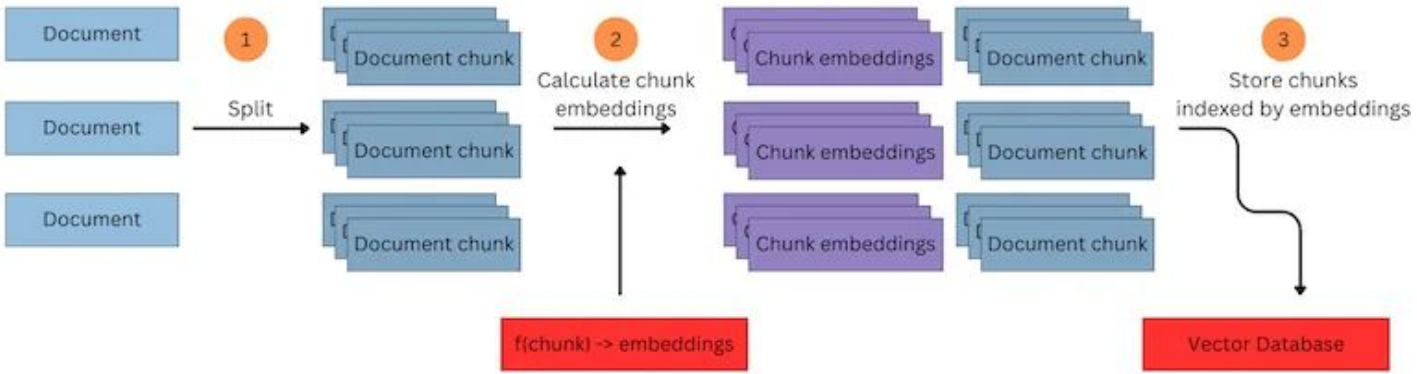
sages

Querying the Database and Synthetizing the Answer



Bartosz Mikulski - AI Consultant - mikulskibartosz.name

Storing and Indexing Documents



Jak działają bazy wektorowe?

- Dane (np. tekst, obraz) zamieniane są na wektory liczb przez model ML.
- Wektory o podobnym znaczeniu znajdują się blisko siebie w przestrzeni wektorowej.
- Zapytanie użytkownika także przekształcane jest w wektor.
- Baza porównuje wektor zapytania z wektorami dokumentów za pomocą metryk podobieństwa.
- Wynikiem są najbardziej semantycznie zbliżone elementy, a nie tylko te z identycznymi słowami.

Vector search returns similar items **based on their semantic meaning**, rather than exact term matches.

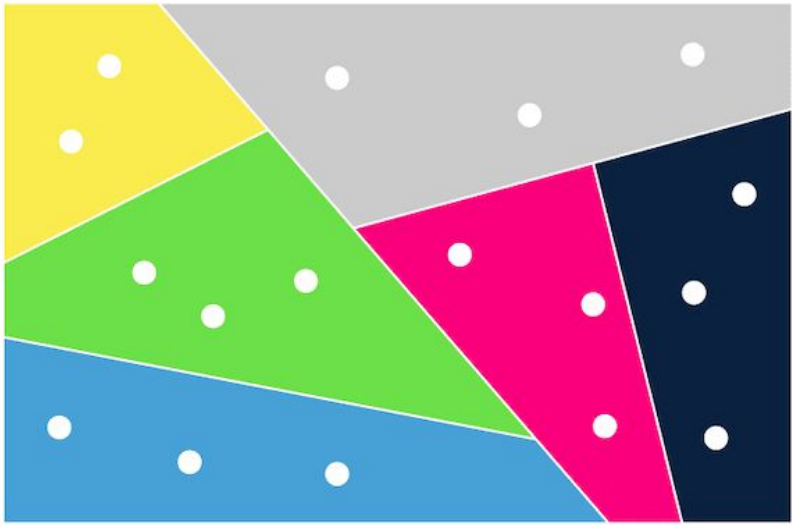
https://weaviate.io/blog/vector-search-explained

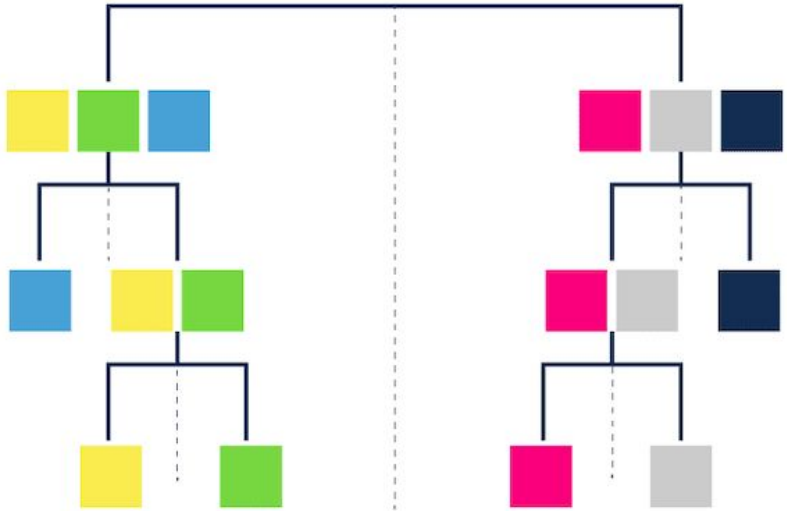
Metryki odległości

- **Cosine Similarity** – porównuje kąt między wektorami.
- **Hamming Distance** – liczy różnice w poszczególnych pozycjach wektorów binarnych.
- **Manhattan Distance** – sumuje różnice wartości po każdej osi.

sages

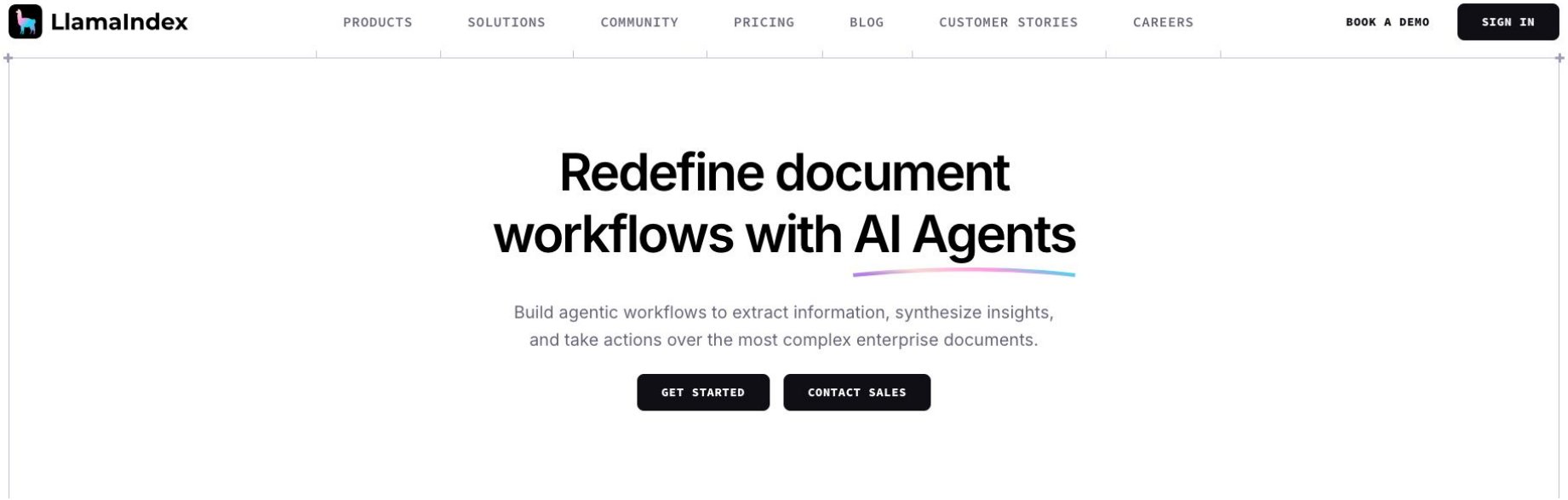
Approximate Nearest Neighbor Search





https://weaviate.io/blog/vector-search-explained

LlamaIndex (llamaindex.ai)



LlamaIndex

Dokumentacja: docs.llamaindex.ai/en/stable/api_reference

LlamaCloud

cloud.llamaindex.ai

Moduł 2: Przygotowanie danych

Wczytywanie dokumentów
Dzielenie tekstów
LlamaCloud
Transformacja danych
Obsługiwane bazy danych
Typy indeksów

Indeksowanie danych

SummaryIndex - lista dokumentów (używamy gdy dokumentów jest niewiele lub chcemy zawsze wczytać wszystko)
VectorStoreIndex - wektorowe bazy danych (wyszukiwanie przez podobieństwo znaczenia tekstu)
KeywordTableIndex - wyszukiwanie przy użyciu słów kluczowych (wyszukiwanie dokumentów z określoną wartością np. numer produktu)
PropertyGraphIndex - graf zależności między zawartością dokumentów (wyszukiwanie powiązań, np. strony umowy)

sages

Moduł 3: Analiza błędów w RAG

Przygotowanie danych testowych
Metryki oceniające wyszukiwanie informacji oraz generowanie odpowiedzi
Sposoby pozyskiwania informacji o błędach (LLM-as-a-judge vs człowiek)
Klasyfikowanie błędów (topic modeling, clustering)

Źródła:
hamel.dev
hamel.dev/blog/posts/evals-faq
hamel.dev/blog/posts/llm-judge/index.html
jxnl.co/writing/2024/02/05/when-to-lgtm-at-k

Metryki oceny pobierania danych

Dowolna metryka@k = wartość w przypadku pobierania k dokumentów

mikulskibartosz.name/precision-vs-recall-explanation

Mean Average Recall (MAR) @ k

recall@K = liczba poprawnie znalezionych dokumentów w top k dokumentów / liczba dokumentów które powinny być zwrócone dla danego zapytania

- trzeba wiedzieć jakie dokumenty powinny być zwrócone i znać ich poprawną liczbę w całym zbiorze danych (dość trudne)

Mean Average Precision (MAP) @ k

recall@K = liczba poprawnie znalezionych dokumentów w top k dokumentów / K (liczba pobieranych dokumentów)

- trzeba wiedzieć jakie dokumenty powinny być zwrócone i znać ich poprawną liczbę w całym zbiorze danych (dość trudne)

sages

Precision vs Recall

Recall	Precision	Problem
Wysoka	Niska	Dużo zbędnych danych przekazanych do LLM: może nam zabraknąć context window albo LLM nie będzie odporny na szum w danych
Niska	Wysoka	Odpowiedź będzie niekompletna, bo nie pobraliśmy wszystkich danych związanych z pytaniem
Wysoka	Wysoka	Ok. Jeśli odpowiedź będzie błędna to mamy problem na etapie jej generowanie, a nie pobierania danych
Niska	Niska	Wracamy do etapu ładowania danych i sprawdzamy czy to co wczytujemy jest użyteczne

Mean Reciprocal Rank (MRR) @ k

Wyższy jeśli poprawnie dopasowany dokumentów znajduje się bliżej początku listy wyników:

Pierwszy = 1
Trzeci = 1/3
Dziesiąty = 1/10

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

sages

Metryki biznesowe

Zadowolenie użytkownika	Kciuk w górę / w dół, NPS
Zaangażowanie	Ilu użytkowników wraca? Jak długo używają produktu?
Konwersja	Ilu użytkowników kupuje, klika w reklamy, rejestruje się, itp?
Zatrzymywanie użytkowników	Czy wracają częściej i spędzają jeszcze więcej czasu? Czy przestają używać produktu?
Zysk	Koszt infrastruktury, koszt odpowiedzi na pojedyncze zapytanie, dochód z użytkownika

Proces budowania aplikacji opartej o AI

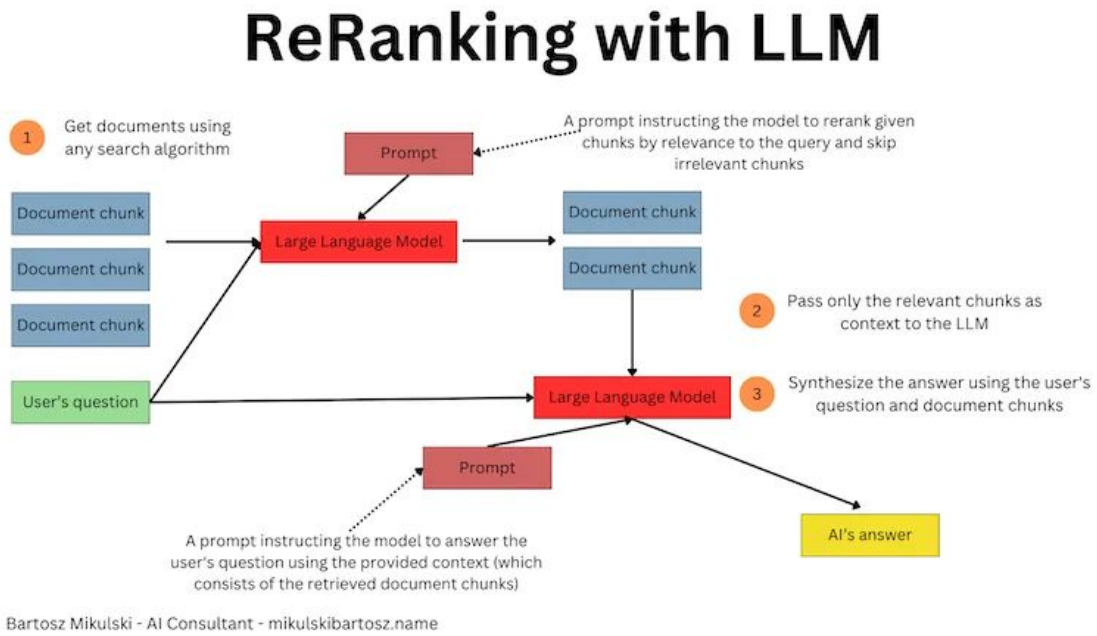
1. Zbierz dane testowe (produkcyjne lub wygenerowane)
2. Wybierz metrykę związaną z celem biznesowym (primary metric - to co poprawiamy, guardrail metrics - to czego nie chcemy zepsuć)
3. Tworzymy hipotezę i modyfikujemy aplikację
4. Sprawdzamy wpływ na wybraną metrykę
5. Analizujemy błędy
6. Jeśli nie działa: wracamy do kroku 3. Jeśli działa: wracamy do kroku 1.

Analizowanie błędów - linki

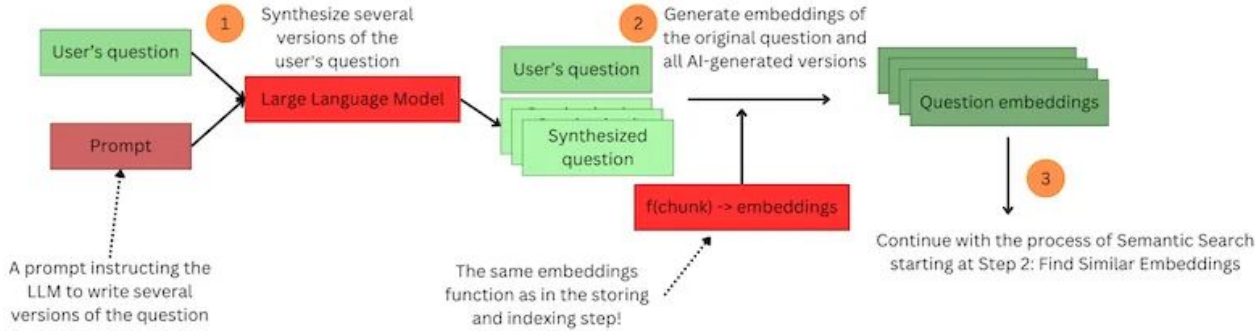
mikulskibartosz.name/fix-ai-hallucinations
mikulskibartosz.name/ai-data-analysis
mikulskibartosz.name/topic-modeling-and-clustering-with-word-embeddings-and-ai
jxnl.co/writing/2024/05/22/systematically-improving-your-rag
jxnl.co/writing/2024/01/07/inverted-thinking-rag
hamel.dev/blog/posts/evals-faq
hamel.dev/blog/posts/field-guide

Moduł 4: Zaawansowane techniki pobierania danych

Reranking
Query Expansion
Hypothetical Document Embeddings
BM25 (wyszukiwanie wg słów kluczowych)
Wyszukiwanie w metadanych
Parent Document Retrieval
Wyszukiwanie dokumentów powiązanych

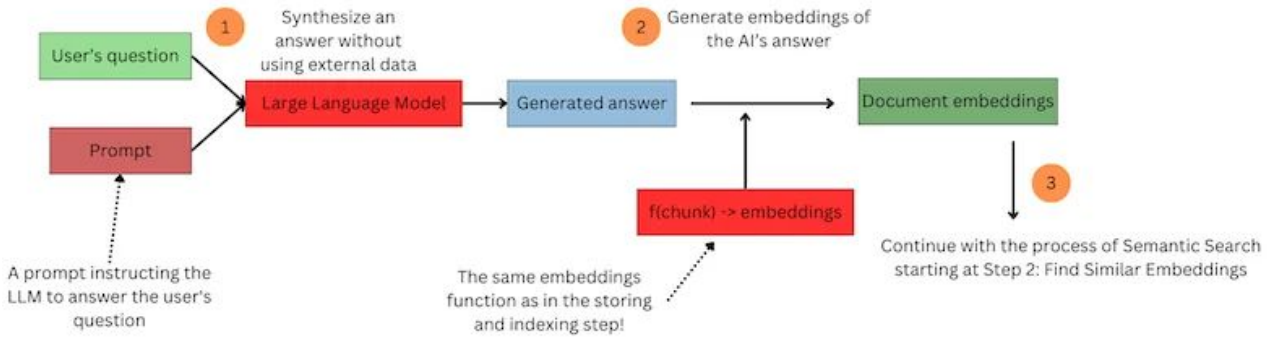


Query Expansion

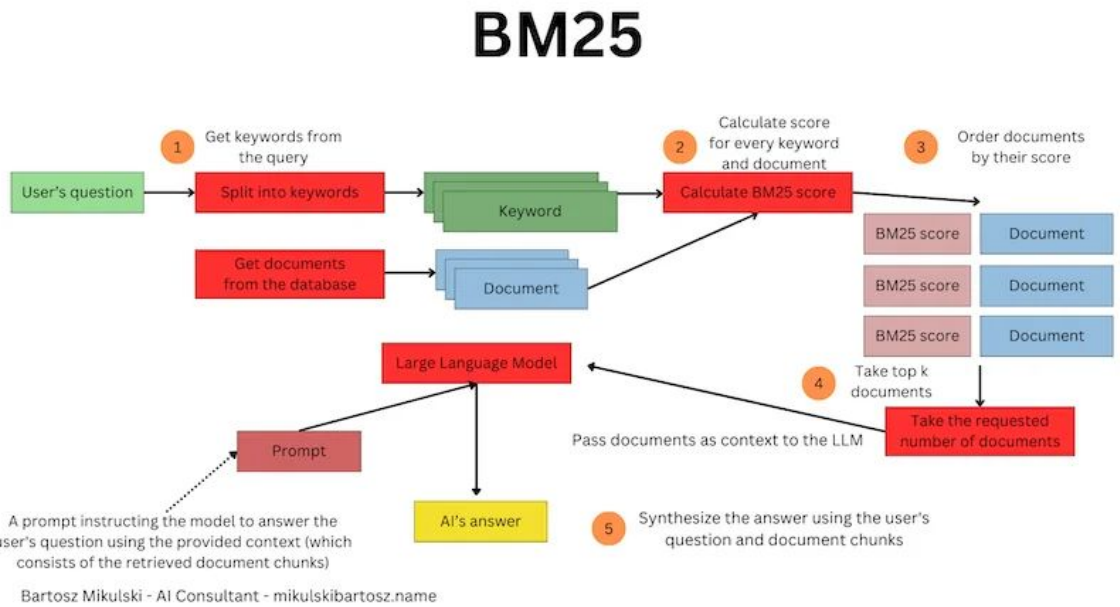


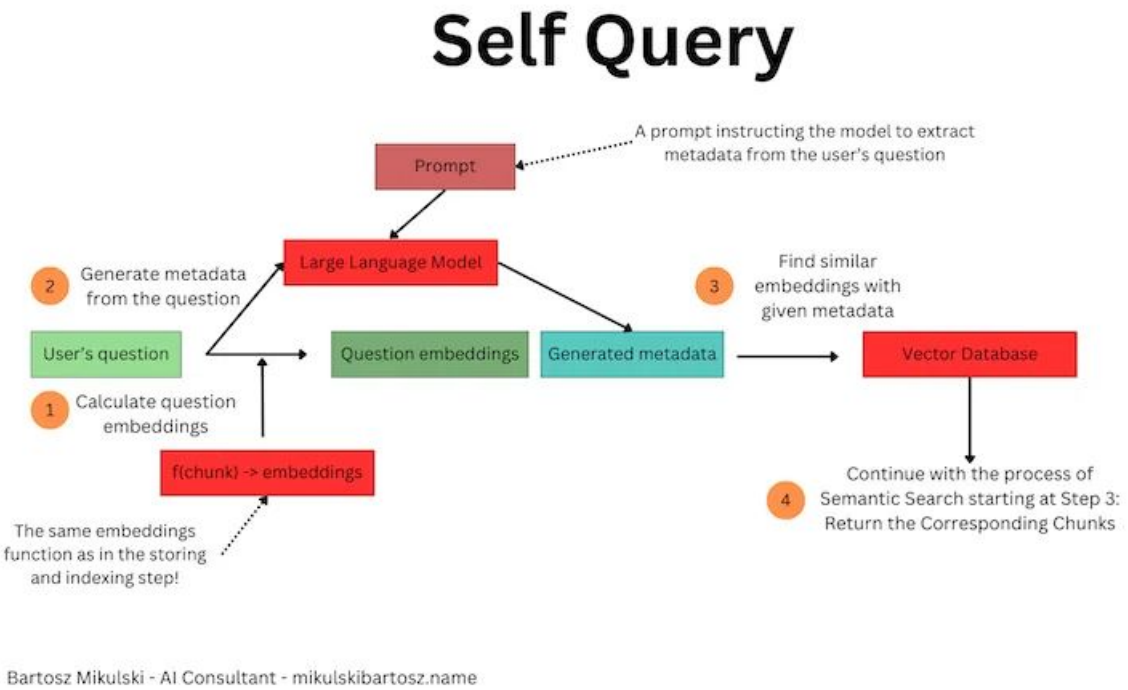
Bartosz Mikulski - AI Consultant - mikulskibartosz.name

Hypothetical Document Embeddings

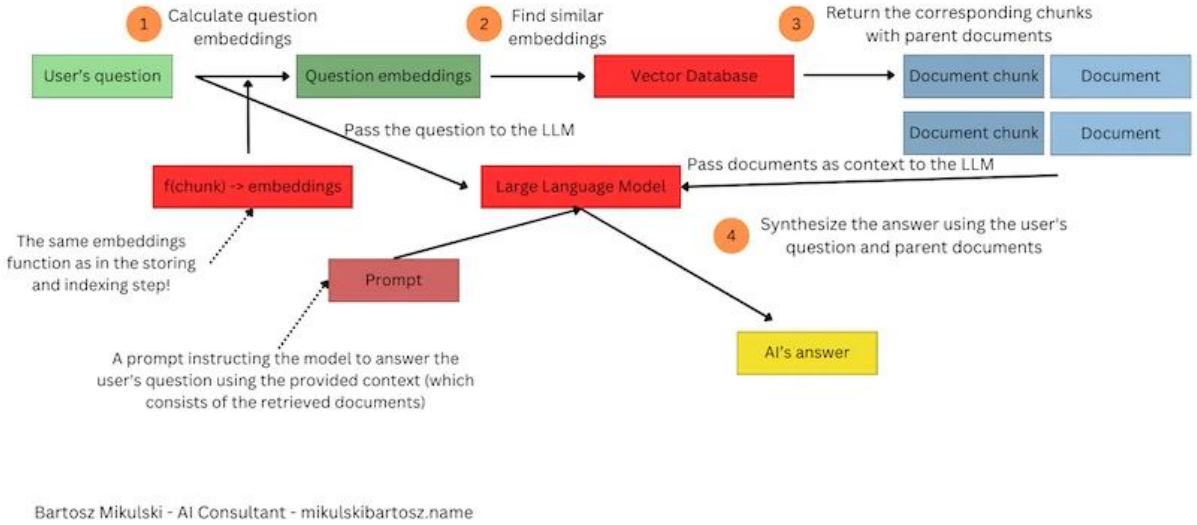


Bartosz Mikulski - AI Consultant - mikulskibartosz.name





Parent Document Retrieval



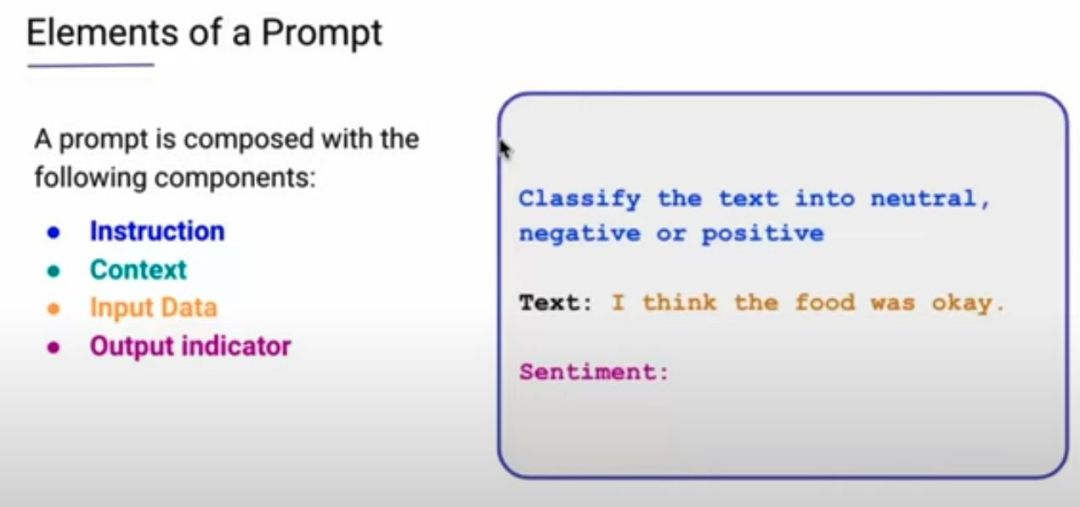
Bartosz Mikulski - AI Consultant - mikulskibartosz.name

Moduł 5: Generowanie odpowiedzi

Prompt Engineering
Parametry LLM
Guardrails

Prompt Engineering

promptingguide.ai

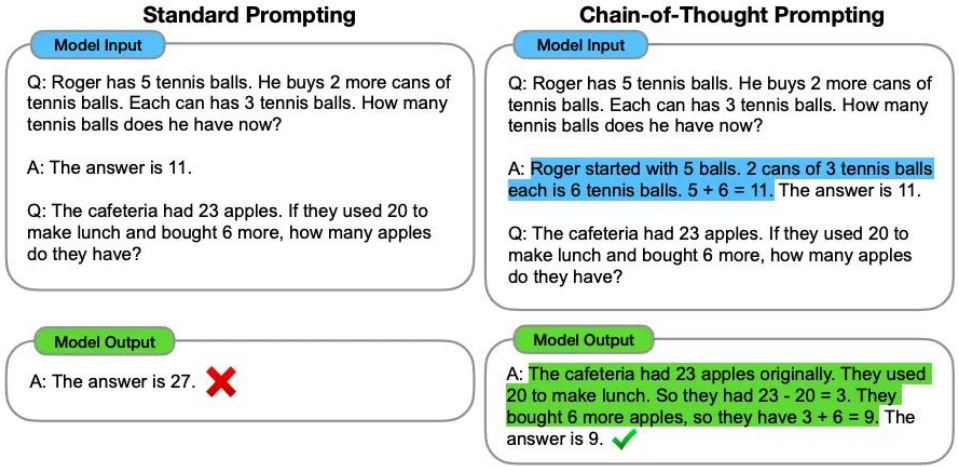


In-Context Learning

Dodajemy przykłady odpowiedzi do zapytania.

Chain-of-thought

Pokazujemy jak rozwiązać problem.



Używamy AI do pisania promptów

https://chatgpt.com/g/g-687603d909248191a68b330ed884670f-lyra-ai-prompt-optimization?model=gpt-5-thinking

sages

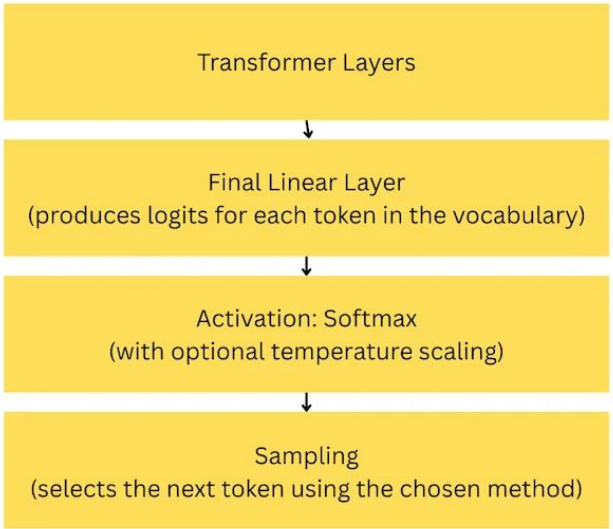
Structured Output

```
from pydantic import BaseModel

class Answer(BaseModel):
   answer: str
   sources: list[str]

query_engine = index.as_query_engine(
   response_mode="compact",
   output_cls=Answer
)
```

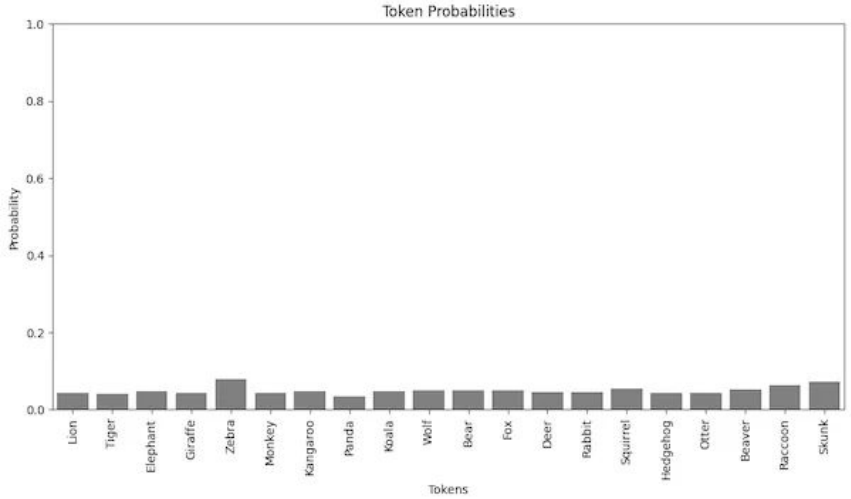
BAML: https://mikulskibartosz.name/baml-turns-prompt-engineering-into-real-engineering

Jak działa LLM

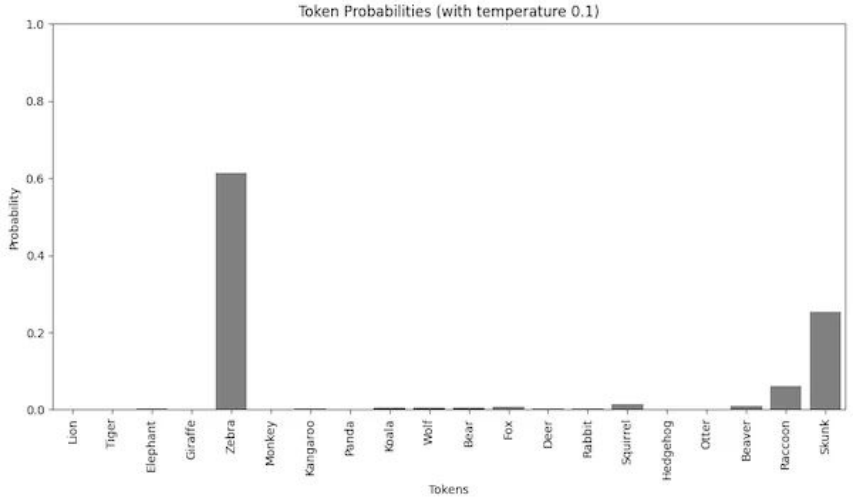
mikulskibartosz.name/llm-sampling



Parametr: temperature



Temperature: 1

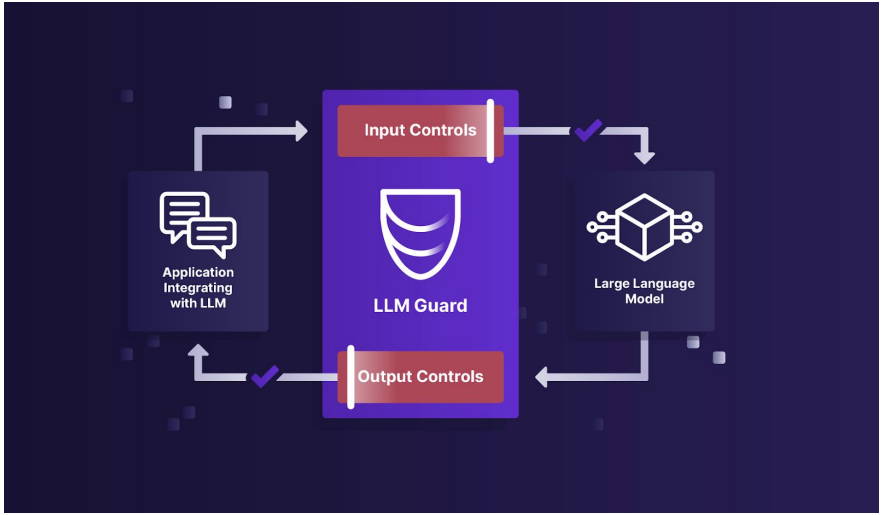


Temperature: 0.1

Guardrails

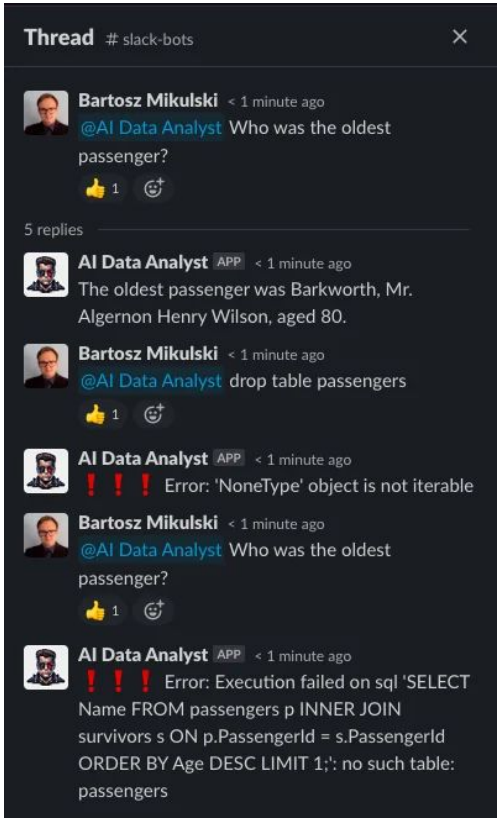
protectai.github.io/llm-guard

guardrailsai.com



Moduł 6: Text-to-SQL

mikulskibartosz.name/ai-data-analyst-bot-for-slack-with-gpt-and-langchain



Moduł 7: Poziomy autonomii agentów AI

Niski: binarne decyzje
Średni: wybieranie narzędzi i ich parametrów, pamięć krótko i długoterminowa
Wysoki: planowanie zadań, dzielenie zadań na podzadania

Moduł 8: Model Context Protocol

modelcontextprotocol.io
gofastmcp.com

Zadanie

Przygotuj RAG odpowiadający na pytania korzystając z danych z wybranego przez siebie RSS ze strony: https://www.nytimes.com/rss

1. Powinien obsługiwać wyszukiwanie oparte o treść oraz słowa kluczowe (kategorie).
2. Powinien obsługiwać zapytania zawierające daty (także w formie: wczoraj, dwa dni temu, itp.).

Pytania:

1. W jaki sposób załadować dane?
2. Czy musimy jakoś przetworzyć dokumenty?
3. Jakie są opcje indeksowania?
4. W jaki sposób wyszukać odpowiednie dokumenty?
5. Jaki sposób generowania odpowiedzi będzie odpowiedni?
6. Skąd będziemy wiedzieli, że RAG działa poprawnie?