Od pojedynczego pytania do rozmowy

Wiele wiadomości i niezależny RAG dla każdej z nich

Czas	Pracujesz z	Rezultat
około 35 min	Ollama lub checkpoint	chat_results.jsonl

Uruchomisz interfejs nad gotowym ask_rag() i obejrzysz wyniki pytań o restart, stan usługi oraz TLS. Sprawdzisz walidację wiadomości i odróżnisz wynik RAG od błędu programu.

Cel i rezultat rozdziału

Obsłużysz wiele niezależnych wiadomości, pokażesz status i źródła oraz wyjaśnisz działanie progu podobieństwa.

Po zakończeniu masz

- wyniki czterech prób RAG i trzech prób walidacji;
- porównanie TLS bez progu i z progiem 1,0;
- własne oceny powiązane przez result_id.

Zanim zaczniesz

Rozpakuj cały pakiet Student. Zachowaj razem notebooki, dane, kod, checkpointy i wyniki. Ścieżki w karcie są liczone od Student. Przygotowanie stanowiska opisuje instrukcja pracy z modułem 5.

Otwórz notebook

```
notebooki/M12_M5_R2_petla_rozmowy_start.ipynb
```

Materiał wejściowy	Do czego służy
dane/semantic_index/	Dziesięć fragmentów oraz manifest indeksu z M4.
kod/m5_tools.py	validate_chat_message(), print_chat_result() i run_chat().

Ollama lub checkpoint. Czas karty dotyczy ćwiczenia, bez instalacji i oglądania prezentacji. Materiał: M12-M5-v5, slajdy 45–86.

Praca krok po kroku

Krok 1 Przygotuj backend — Wybierz MODE w notebooku i uruchom Przygotowanie.

- Ollama wykonuje nowe wywołania; checkpoint odtwarza zapisane identyczne żądania.
- Sprawdź tryb, index_id i kształt macierzy: 10 rekordów, 1024 wymiary.

Krok 2 Prześledź interfejs — Otwórz definicje trzech funkcji wskazanych w materiałach wejściowych.

- Znajdź oczyszczanie tekstu, komendy zakończenia i przekazanie similarity_threshold.
- W R2 poprzednie pytanie nie jest dołączane do następnego promptu.

Krok 3 Wykonaj scenariusze — Uruchom komórkę Próby tylko raz dla bieżącego OUT.

- Zestaw obejmuje restart, stan, TLS domyślnie oraz TLS z progiem 1,0.
- Sprawdź też puste wejście, 1001 znaków i wartość nietekstową.

Krok 4 Porównaj oba warianty TLS — Odczytaj status, reason, sources i diagnostics każdego wyniku.

- below_threshold oznacza odrzucenie wszystkich znalezionych wyników przez próg.
- Jeśli kontekst dopuszczono, odmowa modelu może mieć answered i niepuste źródła.

Krok 5 Zapisz ocenę — W komórce Ocena zastąp trzy pola przy każdym result_id.

- Obserwacja opisuje wynik; dowod wskazuje pole wyniku albo fragment dokumentu.
- Wniosek wyjaśnia granicę systemu. Zapisz ocenę bez ponownej generacji.

Krok 6 Sprawdź własną pętlę — Opcjonalnie ustaw INTERAKTYWNIE=True i uruchom komórkę interfejsu.

- Wpisz pytanie i zakończ przez /koniec lub /exit. Do swobodnych pytań użyj Ollama.
- Zapisz notebook oraz katalog próby. Wprowadzane polecenia DocAssist pozostają tekstem.

Ważne: ponowne Przygotowanie tworzy nowy katalog. Nowe wykonanie ocen osobno. Po zmianie komentarzy uruchom wyłącznie ocenę i zapis; zachowaj surowe wyniki.

Punkty kontrolne pracy

Po etapie	Masz sprawdzić	Gdy wynik nie pasuje
Po próbach	Cztery wyniki RAG i trzy wyniki walidacji.	Otwórz chat_results.jsonl i sprawdź case_id.
Po TLS	Przyczyna wynika z konkretnego przebiegu.	Porównaj context_count, sources i similarity_threshold.
Po ocenie	Każdy wynik ma własny dowód i wniosek.	Nie kopiuj samego statusu jako oceny jakości.

Nazewnictwo i zapis wyników

Wyniki zapisują się w wyniki/R2/<data_id>/. Odczytaj rzeczywisty katalog z komunikatu Wyniki. Główny rezultat: chat_results.jsonl.

Zapisz notebook przez Ctrl+S. Oceny zapisują się jako ocena_<id>.json; szablon ocena_do_uzupelnienia.json pozostaje bez zmian. Sposób oddania określa prowadzący.

Lista kontroli przed zakończeniem

- □ Zanotowałem tryb i identyfikator indeksu.
- □ Odróżniam answered od poprawności treści.
- □ Wyjaśniam różnicę below_threshold i no_usable_results.
- □ Nie traktuję score jako procentu pewności.
- □ Oceny wskazują result_id tej próby.
- □ Zapisałem wyniki i notebook.

Gdy pojawi się problem

Objaw	Najpierw sprawdź
Brak identycznego żądania	Przywróć scenariusz checkpointu albo wybierz jawnie Ollama.
TLS ma answered	Przeczytaj odpowiedź i sprawdź kontekst; status nie ocenia pokrycia pytania.
Źródła mają inne etykiety	Porównaj chunk_id; [S1] obowiązuje tylko w danym wyniku.
Próba już istnieje	Do nowych wywołań uruchom Przygotowanie; do zmiany oceny tylko Zapis.

Zasada końcowa: zachowaj oryginalny wynik i własne uzasadnienie. Błędna odpowiedź lub decyzja modelu jest materiałem do analizy, nie powodem do poprawiania surowego zapisu.