Integracja i porównanie wariantów RAG

Od pięciu pytań do porównania ustawień systemu

Czas orientacyjny	Pracujesz z	Rezultat
80 min	notebookiem R6 i indeksem z M3	rag_experiments.jsonl i raport_z_eksperymentu.md

Zastosujesz cały przepływ RAG i porównasz wyniki w zestawie pytań oraz kontrolowanych wariantów.

Cel i rezultat rozdziału

Uruchomisz pełny RAG na indeksie M3 oraz porównasz kontekst, top_k, próg i wersję promptu, zachowując pochodzenie każdej próby.

Po zakończeniu masz

- 19 prób w rag_experiments.jsonl
- własną ocenę wyników i raport_z_eksperymentu.md
- uzasadniony wybór ustawień z opisem ograniczeń

Zanim zaczniesz

Pracuj na rozpakowanej kopii Student. Uruchom Jupyter z tego katalogu. Wybierz MODE: ollama albo checkpoint. Najpierw wykonaj notebook 00_sprawdzenie_srodowiska.ipynb.

Materiał	Do czego służy
notebooki/M12_M4_R6_kompletny_rag_start.ipynb	Notebook startowy tego rozdziału.
dane/semantic_index/	Kompletny, niezmieniony indeks M3; teksty i metadane.

Gdzie wrócić w prezentacji

Prezentacja M12-M4-v4: rozdział 6, „Integracja i porównanie wariantów RAG”. Odwołania dotyczą nazw rozdziałów, a nie dawnych numerów slajdów.

Praca krok po kroku

Krok 1 Sprawdź pięć pytań — Uruchom cały zestaw. Odczytaj Q1–Q5 w rag_experiments.jsonl.

- Rozróżnij pytanie proste, złożone, częściowo pokryte, obce i brakujące w źródłach.

Krok 2 Porównaj A B C D — Oceń brak kontekstu, pełny kontekst, niepełny oraz celowo nieaktualny materiał D.

- A/B zmienia także instrukcję. B/C/D zachowuje instrukcję; fixture D pozostaje poza indeksem.

Krok 3 Sprawdź top_k i próg — Porównaj top_k 1, 3, 5 oraz cztery progi obliczone z rankingu.

- Próg cosine nie jest prawdopodobieństwem poprawności. Wyższy próg może usunąć potrzebny fragment.

Krok 4 Sprawdź wersje promptu — Porównaj trzy wersje przy tym samym pytaniu i ustawieniach wyszukiwania.

- Wskaż case_id i result_id; nie wybieraj wyniku wyłącznie po płynności języka.

Krok 5 Zapisz ocenę i raport — Oceń 19 wyników. Uzupełnij wszystkie sekcje raport_z_eksperymentu.md.

- Raport ma zawierać dowody z prób i ograniczenia eksperymentu.

Krok 6 Zachowaj komplet — Zapisz notebook, log, najnowszą ocenę oraz raport.

- Własny indeks sprawdź osobno w dodatkowym notebooku; zestaw A–D dotyczy danych pakietu.

Ważne: Przygotowanie tworzy nowy katalog OUT. Po nowej generacji ocen nowe wyniki. Po zmianie komentarzy uruchom tylko komórkę zapisu oceny.

Punkty kontrolne pracy

Kiedy	Co sprawdzasz	Co zrobić przy problemie
Po przygotowaniu	Właściwy tryb, indeks i katalog OUT.	Wróć do notebooka 00 i rozpakuj cały pakiet.
Po próbach	Pełne wyniki; status oddzielony od oceny treści.	Przeczytaj prompt, kontekst, answer i generation.
Po ocenie	Własne uzasadnienia powiązane z result_id.	Zastąp UZUPEŁNIJ i uruchom sam zapis oceny.

Nazewnictwo i zapis wyników

Notebook tworzy wyniki/R6/<proba_id>/. Zachowaj rag_experiments.jsonl i raport_z_eksperymentu.md, ocena_*.json oraz notebook. Nazwę próby nadaje program; odczytaj ją z komunikatu Wyniki.

Zapis <proba_id> oznacza nazwę folderu nadaną automatycznie. Nie wpisuj dosłownie nawiasów. Zapisz notebook przez Ctrl+S i wskaż próbę przekazywaną do oceny.

Lista kontroli przed zakończeniem

- □ Zachowałem pełny log i nie zmieniłem danych indeksu.
- □ Wskazałem tryb i identyfikatory ocenianych wyników.
- □ Zastąpiłem UZUPEŁNIJ własnym uzasadnieniem.
- □ Sprawdziłem podstawy odpowiedzi lub przyczynę braku generacji.
- □ Zapisałem notebook i wybrany katalog próby.

Gdy pojawi się problem

Objaw	Co sprawdzić
Brak plików lub importu	Uruchom notebook z katalogu Student albo Student/notebooki.
Brak odpowiedzi modelu	Sprawdź Ollama i modele lub jawnie wybierz checkpoint.
Brak żądania w checkpoincie	Przywróć domyślne pytania i parametry albo użyj ollama.
Ocena nie zapisuje się	Wypełnij wszystkie cztery pola dla każdego result_id.

Zachowaj oryginalny wynik i własną analizę. Błędna odpowiedź modelu też jest wartościowym wynikiem ćwiczenia.