# Testowanie jakości wyszukiwania semantycznego

Oczekiwane trafienia, metryki i diagnoza słabych wyników

| Czas orientacyjny | Pracujesz z | Rezultat |
|---|---|---|
| 50 min | indeksem R5 i zestawem pytań kontrolnych | test_cases.jsonl, retrieval_baseline.jsonl i raport_testow.md |

Oddzielisz pomiar trafień od analizy przypadków bez odpowiedzi. Zachowasz pierwszy wynik testu jako punkt odniesienia dla przyszłych zmian w systemie.

## Cel i rezultat rozdziału

Przeprowadzisz odtwarzalny test wyszukiwania i wskażesz konkretne przyczyny problemów oraz dalsze działania.

### Po zakończeniu masz

- co najmniej osiem pytań z oczekiwaniami zapisanymi przed testem;
- wyniki top-5 i poprawnie policzone Hit@1, Hit@3, Hit@5;
- raport z interpretacją przypadków i propozycją zmian.

## Zanim zaczniesz

Najpierw wykonaj notebook 00. Otwórz wskazany starter, wybierz MODE i uruchom Przygotowanie. Pola UZUPEŁNIJ wypełnij samodzielnie. SOURCE wybiera własny indeks R5 albo checkpoint. Klucz zapisz przed pomiarem.

`notebooki/M12_M3_R6_test_retrieval_start.ipynb`

| Materiał | Do czego służy |
|---|---|
| semantic_index/ z wybranej próby R5 | Zwalidowany indeks wraz z manifestem. |
| dane/test_cases_szablon.jsonl | 12 pytań bez klucza; pierwsze 8 daje wariant podstawowy. |
| Wyniki R4 | Pomoc w wyborze trudnych przypadków, bez przepisywania rankingu do klucza. |

### Gdzie wrócić w prezentacji

W R6 odszukaj „Porównanie Hit@1, Hit@3 i Hit@5” oraz „Raport z testów”. Odczyt indeksu opisuje R5: „Ponowne wczytanie indeksu według manifestu”.

# Praca krok po kroku

**Krok 1 Zaprojektuj zestaw** — Uzupełnij minimum osiem pytań z szablonu, w tym dwie parafrazy, pytanie niejednoznaczne i dwa typy braku odpowiedzi.

- Dla każdego zapisz question_id, question, question_type i case_type.
- question_type opisuje formę pytania; case_type to answerable, missing_in_sources albo out_of_scope.

**Krok 2 Ustal oczekiwania** — Przeczytaj bazę i przed testem zapisz expected_chunk_ids oraz uzasadnienie.

- Dla answerable podaj rzeczywiste chunk_id; dla missing_in_sources i out_of_scope pustą listę.
- Jeśli pytanie jest niejednoznaczne, opisz interpretację i dopuszczalne trafienia.

**Krok 3 Wykonaj pomiar** — Dla każdego pytania pobierz top-5 z tego samego indeksu i zapisz pełne wyniki.

- Zachowaj question_id, results, index_id, model i datę pomiaru.
- MODE=checkpoint obsługuje dokładne teksty z pytania_startowe.json. Nowe pytanie wymaga MODE=ollama.

**Krok 4 Oblicz trafienia** — Dla answerable ustaw Hit@k na 1, gdy którykolwiek oczekiwany chunk_id wystąpi w pierwszych k wynikach; inaczej na 0.

- Średnią licz wyłącznie po pytaniach answerable, osobno dla k = 1, 3, 5.
- Dla pozostałych zapisz null i osobną ocenę; gdy nie ma answerable, średnia jest nieokreślona.

**Krok 5 Zdiagnozuj przypadki** — Opisz trzy nieudane lub problematyczne przypadki na podstawie tekstów, metadanych i rankingu.

- Jeśli nie ma trzech problemów, rozszerz zestaw albo uczciwie opisz ograniczenia; nie wymyślaj błędów.
- Odróżnij brak źródła, zły podział tekstu, mylącą parafrazę i problem konfiguracji.

**Krok 6 Zapisz punkt odniesienia** — Zapisz raport i plan jednej zmiany, którą sprawdzisz w kolejnej próbie.

- Podaj liczebność grup, wyniki Hit@k i ograniczenia małego zestawu.
- Wariant rozszerzony: minimum 12 pytań i porównanie dwóch konfiguracji na tych samych przypadkach.

**Zapis próby: komórka Przygotowanie automatycznie tworzy nowy folder wyniki/R6/<proba_id>/. Zachowaj wyświetloną ścieżkę.**

## Punkty kontrolne pracy

| Kiedy | Co sprawdzasz | Co zrobić przy problemie |
|---|---|---|
| Przed testem | Klucz wynika ze źródeł, a nie z rankingu. | Zweryfikuj oczekiwane identyfikatory. |
| Metryki | Hit@1 ≤ Hit@3 ≤ Hit@5, ten sam mianownik answerable. | Wyłącz przypadki bez oczekiwanej odpowiedzi ze średniej. |
| Raport | Każdy wniosek wskazuje obserwację i dalsze działanie. | Oddziel wyniki pomiaru od hipotez. |

## Nazewnictwo i zapis wyników

Notebook zapisuje test_cases.jsonl, retrieval_baseline.jsonl i raport_testow.md w wyniki/R6/<proba_id>/. Zachowaj pierwszą próbę bez nadpisywania.

Identyfikator <proba_id> nadaje notebook. Pliki są chronione przed nadpisaniem; nowy pomiar rozpocznij od Przygotowania. Zapisz notebook przez Ctrl+S i wskaż próbę do oceny.

## Lista kontroli przed zakończeniem

- □ Zapisałem co najmniej osiem pytań i ich typy.
- □ Ustaliłem klucz przed wyszukiwaniem.
- □ Wyniki dotyczą jednego, wskazanego indeksu.
- □ Poprawnie rozdzieliłem przypadki answerable i pozostałe.
- □ Policzyłem Hit@1, Hit@3 i Hit@5.
- □ Zapisałem raport oraz ograniczenia pomiaru.

## Gdy pojawi się problem

| Objaw | Co sprawdzić |
|---|---|
| Pierwszy wynik staje się kluczem | Wróć do źródeł i zbuduj oczekiwania niezależnie od modelu. |
| Brak odpowiedzi obniża Hit@k | Takie przypadki oceniaj osobno i zapisz dla nich null. |
| Kilka fragmentów jest potrzebnych łącznie | Hit@k nie mierzy kompletności zestawu. Opisz to ograniczenie lub dodaj osobną miarę. |

**Próg, np. 80%, może być celem konkretnego projektu, ale nie jest automatycznym warunkiem zaliczenia tej karty. Zapisz także uzupełniony notebook.**