Ocena asystenta

Karta pracy uczestnika

Czas pracy	Materiały	Rezultat
Około 75 min	Notebook R4 i karta	Pliki opisane w karcie

Poprawny status nie gwarantuje dobrej odpowiedzi. Sprawdź fakty, źródła, warunki krytyczne oraz wybór i użycie narzędzi w kolejnych turach.

Cel i rezultat rozdziału

Ocenisz odpowiedzi i zachowanie asystenta na podstawie kontraktów. Zachowasz surowe wyniki, pełne dane historyczne M5 i ranking z tego samego wywołania RAG.

Po zakończeniu masz

- Pełny import historycznych wyników M5 z zachowaniem pochodzenia.
- Nowe lub odtworzone wyniki asystenta z identyfikatorami.
- Uzasadnione recenzje odpowiedzi i kontrola warunków krytycznych.

Zanim zaczniesz

Zakończ R3. Otwórz rubryki oraz surowe wyniki M5. Do pełnych nowych prób potrzebujesz Ollama; ścieżka offline obejmuje pięć sprawdzonych przypadków.

Otwórz notebook

```
notebooki/M12_M6_R4_ocena_asystenta_start.ipynb
```

Plik lub materiał	Do czego służy
dane/import_M5/functional_tests.jsonl	Historyczne R6 M5; import nie jest nowym testem M6.
evaluation/test_cases.jsonl	Kontrakty odpowiedzi i zachowania asystenta.

Tryb: checkpoint lub Ollama. Slajdy M12-M6-v6: 216–281. Czas orientacyjny: 75 min.

Praca krok po kroku

1. Importuj historię

- Przenieś pełne rekordy M5 do assistant_results.jsonl przez import_m5.
- Zachowaj oryginalne result_id wewnątrz raw_result. Nie mapuj ich automatycznie na T101.

2. Wykonaj próby

- Uruchom sprawdzony podzbiór offline lub pełny zestaw w trybie Ollama.
- Każdy przypadek rozpoczyna nową sesję; scenariusz kontynuuje rzeczywistą historię.

3. Sprawdź ślad RAG

- Otwórz raw_result i powiązany retrieval_result_id.
- Adapter zapisuje ranking tego samego wywołania. Nie uruchamiaj wyszukiwania drugi raz.

4. Oceń odpowiedzi

- Porównaj fakty i cytowania z pełnym kontekstem. Uzasadnij sześć kryteriów lub N/D.
- Obecność etykiety źródła nie dowodzi, że źródło potwierdza twierdzenie.

5. Oceń zachowanie

- Sprawdź status, reason, wybór narzędzia, argumenty i kontynuację rozmowy.
- W T101 oceń warunek kopii, a w T106 brak generacji RAG po odrzuceniu kontekstu.

6. Zapisz wnioski

- Zapisz oceny i listę problemów z identyfikatorami wyników.
- Nie licz błędów środowiska ani braku checkpointu jako oceny merytorycznej zero.

Wyniki historyczne M5 służą analizie. Bez jawnego mapowania kontraktu nie są oceniane automatycznie jako testy M6.

Punkty kontrolne pracy

Kontrola	Pytanie	Dowód
Pochodzenie	Czy odróżniasz import i wykonanie?	origin, mode i raw_result_id.
Powiązanie	Czy ranking pochodzi z tej odpowiedzi?	retrieval_result_id; bez powtórnego RAG.
Ocena	Czy uwzględniono błąd krytyczny?	Uzasadnienie i verdict dla każdej recenzji.

Nazewnictwo i zapis wyników

evaluation/assistant_results.jsonl; evaluation/assistant_reviews.jsonl

W R5 wskażesz etap powstania problemu i sprawdzisz jedną zmianę konfiguracji.

Lista kontroli przed zakończeniem

□ Zaimportowałem pełne wyniki M5.
□ Nie zmieniłem oryginalnych danych.
□ Sprawdziłem powiązanie z rankingiem.
□ Odróżniam status od jakości treści.
□ Oceniłem źródła i zachowanie narzędzi.
□ Zapisałem krytyczne naruszenia kontraktu.

Gdy pojawi się problem

Problem	Co zrobić
Brak pełnego rankingu w imporcie	Pozostaw powiązanie puste; nie rekonstruuj go.
Nieznana nowa konfiguracja offline	Uruchom Ollama albo zgłoś próbę niewykonaną.
Niepełny kontrakt recenzji	Wróć do R2; nie zmieniaj kryteriów po cichu.
Średnia maskuje błąd krytyczny	Nadaj fail niezależnie od średniej.

Zachowaj wyniki i uzasadnienia. Oznacz pochodzenie danych oraz próby niewykonane.