Zestaw testowy

Karta pracy uczestnika

Czas pracy	Materiały	Rezultat
Około 60 min	Notebook R2 i karta	Pliki opisane w karcie

Dobre testy sprawdzają konkretne oczekiwania. Ocena nie może zależeć wyłącznie od tego, czy odpowiedź brzmi przekonująco.

Cel i rezultat rozdziału

Zweryfikujesz zestaw 24 przypadków w 13 kategoriach i opiszesz kontrakt każdej tury. Zachowasz pytania bez odpowiedzi, przypadki graniczne i scenariusz rozmowy.

Po zakończeniu masz

- Zatwierdzony test_cases.jsonl z co najmniej 20 przypadkami i 8 kategoriami.
- Raport walidacji identyfikatorów, pól i odwołań do fragmentów.
- Sprawdzone oczekiwania M6-T101, M6-T106 i M6-S001.

Zanim zaczniesz

Zakończ R1. Przygotuj plan, korpus źródeł oraz dane/semantic_index/chunks.jsonl. Dostarczony zestaw traktuj jako propozycję do zatwierdzenia.

Otwórz notebook

```
notebooki/M12_M6_R2_walidacja_zestawu_start.ipynb
```

Plik lub materiał	Do czego służy
evaluation/test_cases.jsonl	Roboczy zestaw; wzorzec jest w dane/przypadki.
dane/zrodla i chunks.jsonl	Podstawa faktów i identyfikatorów źródeł.

Tryb: checkpoint lub Ollama. Slajdy M12-M6-v6: 80–145. Czas orientacyjny: 60 min.

Praca krok po kroku

1. Sprawdź pokrycie

- Przejrzyj kategorie: pytania proste, procedury, wiele źródeł, brak danych i zachowanie asystenta.
- Zachowaj minimum 20 przypadków i 8 kategorii. Nie zastępuj trudnych testów łatwiejszymi.

2. Sprawdź źródła

- Otwórz wymagane fragmenty i porównaj fakty z treścią dokumentów.
- Każda grupa źródeł ma identyfikator dokumentu i wymagane chunk_id.

3. Ustal kontrakty

- Sprawdź oczekiwany status, powód, narzędzie, argumenty i wymagane fakty.
- Dodaj twierdzenia zakazane oraz kryteria, które rzeczywiście dotyczą przypadku.

4. Sprawdź przypadki graniczne

- T101 obejmuje zwykłą procedurę i przywrócenie zatwierdzonej kopii po niepowodzeniu.
- T106 wymaga progu 1,0 i sprawdzenia, że żaden score nie osiąga progu.

5. Sprawdź rozmowę

- W S001 kolejne odpowiedzi pochodzą z rzeczywistego wykonania poprzednich tur.
- Pytanie samodzielne do R3 jest jawne. Nie podstawiaj wzorcowej odpowiedzi do historii R4.

6. Uruchom walidację

- Popraw wszystkie błędy struktury, identyfikatorów i referencji.
- Zapisz raport i wersję zestawu. Walidator nie zastępuje oceny zgodności ze źródłami.

Oczekiwania zamrażaj przed pomiarem. Zmiana kontraktu wymaga nowej wersji i ponownego wykonania powiązanych prób.

Punkty kontrolne pracy

Kontrola	Pytanie	Dowód
Zakres	Czy spełniono minimum?	Co najmniej 20 przypadków i 8 kategorii.
Referencje	Czy chunk_id istnieją?	Poprawny raport walidatora.
Semantyka	Czy oczekiwania wynikają ze źródeł?	Ręczny przegląd faktów i warunków.

Nazewnictwo i zapis wyników

evaluation/test_cases.jsonl; evaluation/test_cases_validation.json

W R3 użyjesz jawnych pytań retrieval_question; w R4 rzeczywistych wiadomości i historii.

Lista kontroli przed zakończeniem

□ Zatwierdziłem wszystkie oczekiwania.
□ Każdy test ma unikalny test_id.
□ Sprawdziłem wymagane fragmenty.
□ Mam scenariusz z kolejnymi turami.
□ Zachowałem T101 i T106.
□ Walidacja zakończyła się bez błędów.

Gdy pojawi się problem

Problem	Co zrobić
Nieznany chunk_id	Wybierz istniejący fragment lub popraw zestaw źródeł.
Brak faktu w źródle	Nie oczekuj go w odpowiedzi; opisz brak danych.
Duplikat identyfikatora	Nadaj nowy test_id, zachowując historię wersji.
Walidator nie wykrywa złej treści	Sprawdź semantykę ręcznie w dokumentach.

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