KARTA PRACY UCZESTNIKA MODUŁ 4 ROZDZIAŁ 2

Wywołanie lokalnego modelu LLM z Pythona

Od pojedynczego żądania do zapisanej odpowiedzi

Czas orientacyjny	Pracujesz z	Rezultat
40 min	notebookiem R2 i indeksem z M3	proby.jsonl

Porównasz pięć wywołań modelu i sprawdzisz, co zmienia instrukcja, a co dostarczenie dokumentu.

Cel i rezultat rozdziału

Wywołasz lokalny model, zachowasz pełny prompt i odpowiedź oraz odróżnisz wpływ instrukcji, kontekstu i temperatury.

Po zakończeniu masz

- pięć oryginalnych odpowiedzi z pełnymi promptami
- porównanie odpowiedzi bez kontekstu i z kontekstem
- wniosek o ograniczeniach próby temperatury

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_R2_wywolanie_llm_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 2, „Wywołanie lokalnego modelu LLM z Pythona”. Odwołania dotyczą nazw rozdziałów, a nie dawnych numerów slajdów.

Praca krok po kroku

Krok 1 Uruchom trzy prompty — Wykonaj próby bez kontekstu, z żądanym formatem i z fragmentem instrukcji.

- Każde wywołanie zaczyna nową rozmowę; nie ma historii czatu.

Krok 2 Porównaj podstawy — Przeczytaj pełne prompty i odpowiedzi. Wskaż twierdzenie dotyczące DocAssist.

- Ogólna rada bez źródła nie potwierdza lokalnej procedury.

Krok 3 Sprawdź temperaturę — Porównaj próby 0.0 i 0.8 z próbą kontekst o temperaturze 0.1.

- To pojedyncze obserwacje przy seed=42, nie dowód przewagi ustawienia.

Krok 4 Sprawdź zakończenie — Odczytaj generation.done_reason oraz tryb i parametry wywołań.

- length oznacza możliwe ucięcie odpowiedzi; nie oceniaj go jako pełnej wypowiedzi.

Krok 5 Oceń pięć wyników — Uzupełnij ocenę każdego result_id i zapisz własne uzasadnienia.

- W trybie checkpoint analizujesz zarejestrowaną próbę, nie nową generację.

Ważne: Przygotowanie tworzy nowy katalog OUT. Po nowej generacji oceń 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/R2/<proba_id>/. Zachowaj proby.jsonl, 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.