Ocena modeli językowych i wpływ strategii promptowania
Lingwistyka Obliczeniowa | Laboratorium 5

Wojciech Bartoszek

1 Wstęp

Celem eksperymentu było porównanie zachowania i jakości generowanych odpowiedzi pomiędzy modelem małym (ok. 1.7B parametrów) a modelem dużym i specjalizowanym w rozumowaniu (ok. 24B parametrów). Badanie obejmuje różne typy zadań (instrukcje, rozumowanie logiczne, twórcze pisanie, generowanie kodu, itp.) oraz trzy strategie promptowania: zero-shot, few-shot oraz CoT (dla standardowych modeli).

2 Modele i środowisko eksperymentalne

Eksperyment uruchomiono lokalnie przy użyciu Ollama i interfejsu Pythonowego. Wybrane modele do porównań to:

- **smollm:1.7b** – model mały (baseline), stosowany z zero-shot, few-shot oraz CoT,
- **magistral:24b** – model większy ze wzmocnionymi zdolnościami rozumowania; z powodu jego wewnętrznego mechanizmu rozumienia pominięto explicite CoT przy jego testowaniu.

3 Metodologia

Przygotowanie eksperymentu składało się z dwóch faz: fazy inżynierii promptów (development set) oraz fazy oceny.

3.1 Zadania (10 typów)

Dla każdej z kategorii zdefiniowano opis zadania oraz kryteria oceny. Przykładowe kategorie:

1. Instruction Following (Instrukcje),
2. Logical Reasoning (Rozumowanie logiczne),
3. Creative Writing (Twórcze pisanie),
4. Code Generation (Generowanie kodu),
5. Reading Comprehension (Rozumienie tekstu),
6. Common Sense Reasoning,
7. Language Understanding & Ambiguity,

8. Factual Knowledge,
9. Mathematical Problem Solving,
10. Ethical Reasoning.

Dla ilustracji przedstawiono przykładowe zadania użyte w ocenie (pojedynczy przykład z każdego rodzaju):

Instruction Following: *"Explain in two sentences how to write a clear function docstring."*

Ethical Reasoning: *"Is it ethical to use performance metrics alone for promotion decisions? Provide considerations."*

Logical Reasoning: *"If all A are B and some B are C, can we conclude some A are C? Explain."*

Dla każdej kategorii przygotowano również 2–3 przykłady do fazy "few-shot"(development set), które nie pokrywały się bezpośrednio z przykładami ewaluacyjnymi.

3.2 Promptowanie

Dla każdej pary (model, zadanie) testowano:

- **Zero-shot**: tylko opis zadania i wejście,
- **Few-shot**: dodano 2–3 przykłady wzorcowe przed zadaniem,
- **Chain-of-Thought (CoT)**: polecenie *"Let's think step by step before answering"* dodane jedynie dla modelu małego i standardowych modeli (CoT pominięto dla modelu rozumującego).

3.3 Ocena

Wszystkie odpowiedzi oceniono ręcznie w sposób anonimowy (maskowano nazwę modelu i strategii), przy użyciu skali 0–5, gdzie 5 oznacza odpowiedź spełniającą wszystkie kryteria.

4 Struktura projektu

Projekt składa się z trzech głównych notebooków:

1. `lab5.ipynb` – przygotowanie eksperymentu, ładowanie zadań i uruchamianie zapytań do wybranych modeli;
2. `lab5_evaluation.ipynb` – anonimowa, ręczna ocena odpowiedzi
3. `lab5_results.ipynb` – analiza wyników i wizualizacje (agregacja, wykresy, eksport obrazów).

Ta trójpodziałowa struktura ułatwia reprodukowalność: oddziela generowanie danych od oceny i analizy.

Model	Strategia	*Średnia ocena* (0–5)
smollm:1.7b	zero	3.000 (n=10)
smollm:1.7b	few	2.800 (n=10)
smollm:1.7b	cot	2.300 (n=10)
magistral:24b	zero	4.700 (n=10)
magistral:24b	few	4.200 (n=10)
Średnia ogólna	smollm:1.7b = 2.700 (n=30); magistral:24b = 4.450 (n=20)	

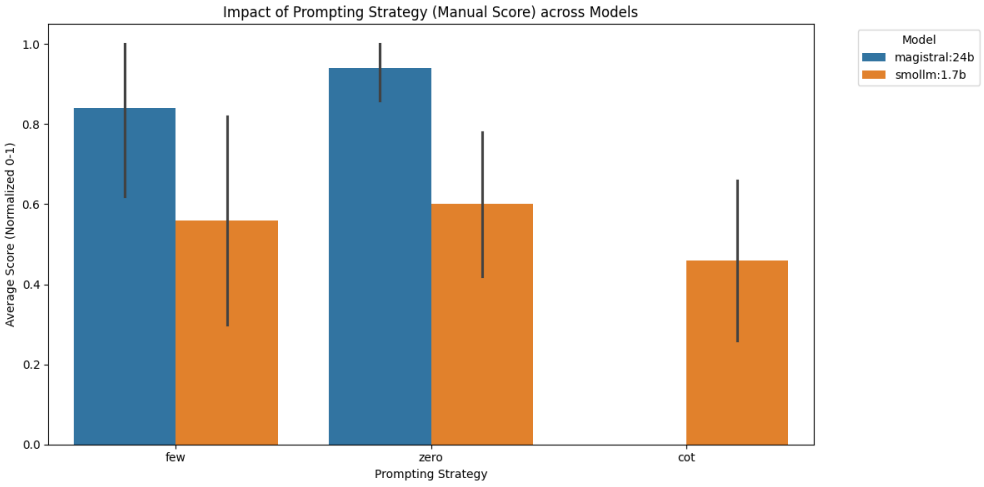
Tabela 1: Średnie oceny ręczne według modelu i strategii.

5 Wyniki

Poniżej przedstawiono zestawienie zebranych wyników. Agregacje wykonano po ręcznej ocenie (skala 0–5). Tabela poniżej przedstawia średnie wartości ocen (mean) dla par (model, strategia):

Wykresy wygenerowane w trakcie analizy (zapisywane do `lab5/outputs`) ilustrują:

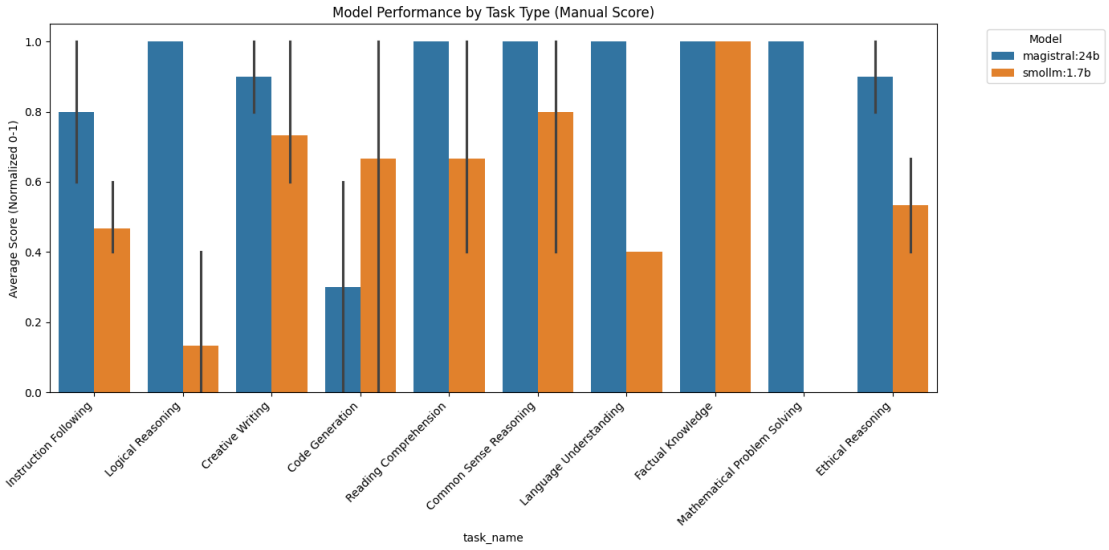
- **Impact of Prompting Strategy** (`strategy_comparison_manual.png`) – porównanie wpływu strategii (zero/few/CoT) na średnią punktację dla poszczególnych modeli.
- **Model Performance by Task Type** (`task_performance_manual.png`) – średnia normalizowana (0–1) dla zadań, z rozbiciem na modele.



Rysunek 1: Wpływ strategii promptowania na średnią ocenę (ręczna ocena).

6 Dyskusja

Z zebranych danych wynika, że model `magistral:24b` istotnie przewyższa model `smollm:1.7b` w większości testowanych kategorii i strategii. Różnica jest szczególnie wyraźna w zadaniach wymagających głębszego rozumowania i złożonej analizy (np. rozumowanie etyczne, logiczne). W przypadku modelu małego zauważalne są relatywnie niewielkie lub mieszane zyski z zastosowania few-shot i CoT - w niektórych zadaniach CoT nie poprawił wyników.



Rysunek 2: Wydajność modeli wg typu zadania (średnia znormalizowana).

Przyczyny obserwowanych różnic mogą obejmować:

- większą pojemność modelu i bogatsze wewnętrzne reprezentacje u modelu dużego,
- ograniczenia modelu małego w utrzymaniu kontekstu i wykonaniu dłuższych, precyzyjnych analiz.

7 Obserwacje

W trakcie analizy zauważono, że dodanie przykładów (few-shot) często obniżało ocenę odpowiedzi w przypadku modelu o ograniczonej pojemności (`smollm:1.7b`). Taka degradacja jakości można przypisać przede wszystkim wewnętrznym ograniczeniom modelu: ograniczone okno kontekstowe sprawia, że przykłady zajmują istotną część dostępnej reprezentacji, utrudniając prawidłowe przetworzenie docelowego zapytania; model może mieć trudności z uogólnieniem wzorców z przykładów i zamiast tego naśladować styl lub nieistotne szczegóły (tzw. "style drift"); wreszcie, mniejsza pojemność często skutkuje generowaniem bardziej rozwlekłych lub mniej trafnych wypowiedzi, co obniża ocenę w rubryce preferującej zwięzłość i precyzję.

W praktyce oznacza to, że dla modeli o ograniczonej pojemności preferencyjne będą krótkie, precyzyjne instrukcje (starannie sformułowany zero-shot) zamiast rozbudowanych few-shotów, które mogą obciążać model i pogorszyć końcowy rezultat.

8 Wnioski

- Duże modele rozumujące oferują wyraźną przewagę w zadaniach wymagających wieloetapowego rozumowania lub rozumienia skomplikowanych kryteriów oceny.
- Prompt engineering (few-shot, CoT) ma największy wpływ przy modelach, które nie posiadają wbudowanych mechanizmów rozumowania; efekty mogą być zmienne dla modeli mniejszych.

- Ręczna, anonimowa ocena pozostaje skutecznym sposobem uchwycenia niuansów jakości odpowiedzi, ale wymaga znacznego nakładu pracy; dla pełniejszej oceny warto rozważyć standaryzowane rubryki i/lub wieloocenianie.