Pytanie badawcze

Czy architektura LLM oraz jego trening wpływa na jego własności psychologiczne?

Przebieg badań

Główną częścią badania będzie przeprowadzenie testów psychologicznych na LMMach. Modele zostaną poinstruowane, żeby wcielić się w daną osobę na podstawie przedstawionych danych i następnie będą odpowiadać na przedstawione pytania z różnych kwestionariuszy psychologicznych. Wyniki będą analizowane oraz porównywane między sobą. Następnie sprawdzone zostanie czy LMMy są w stanie symulować ludzkie zachowania psychologiczne. Wyniki pozwolą stwierdzić czy modele mogą odwzorowywać psychologiczne zachowanie człowieka oraz porównają zdolności różnych architektur oraz różnych metody treningu.

Użyte modele

W badaniach użyte zostaną następujące modele: Llama 3.1 8B Instruct - Q4 KM [1], Granite 3.3 8B Instruct - Q4 KM [2], Aya 23 8B - Q4 KM [3], Alif 1 8B Instruct - Q4 KM [4], Ministral 8B Instruct 2410 - Q4 KM [5] oraz InternLM 3 8B Instruct - Q4 KM [6]. Modele będą używane lokalnie poprzez aplikację LMStudio. Zostały one wybrane z uwagi na podobną liczbę parametrów, różne architektury oraz darmowość.

Prompty

Promt będzie złożony z dwóch części. Pierwszą będzie opis osoby z wykorzystaniem zbioru danych PersonaHub [7]. Jest ogromny zbiór danych obejmujący miliard syntetycznie wygenerowanych postaci (tzw. *personas*), z których każda posiada unikalny zestaw cech, takich jak zawód, umiejętności, dziedzina pracy, poziom wykształcenia czy doświadczenie zawodowe. Dane te pozwalają na realistyczne modelowanie zachowań, opinii i decyzji podejmowanych przez zróżnicowane grupy społeczne w różnych kontekstach.
Innym typem opisu osób jest wykorzystanie danych ze zbioru danych Daily Dialog [8]. Jest to wielozdaniowy zbiór danych dialogowych w języku angielskim. Tematyka obejmuje small talk, zakupy, relacje międzyludzkie czy planowanie. Zawiera 13 118 dialogów. Średnio w każdej rozmowie znajduje się około 8 wypowiedzi (tur rozmowy). Dialogi są ubarwione emocjami, co może dać ciekawe wyniki w testach psychologicznych.
Opisy osób będą łączone z dialogami w celu uzyskania innych wyników i połączenia części opisowej z emocjonalną z dialogów. Można to zrobić poprzez wprowadzenie dwóch osób z różnymi opisami, a następnie dodanie konwersacji pomiędzy nimi. Na podstawie takiej informacji model będzie odpowiadał na pytania.

Drugą częścią promptu będą pytania z różnych kwestionariuszy psychologicznych, które zostały opisane poniżej. Modele będą instruowane by odpowiadały jedynie cyfrą w skali 1-5.

Koncepcje psychologiczne

W pracy badawczej zostaną wykorzystane koncepcje psychologiczne związane z cechami osobowości, agresją, afektem oraz samokontrolą. Będą to:

- **Model Wielkiej Piątki** (BFI) odnoszący się do koncepcji pięciu podstawowych wymiarów osobowości: neurotyczności, ekstrawersji, otwartości na doświadczenie, ugodowości i sumienności.
- **Kwestionariusz Agresji** (BPAQ) bazujący na koncepcji agresji jako wielowymiarowego zjawiska obejmującego agresję fizyczną, werbalną, wrogość i gniew.
- **Kwestionariusz PANAS** opierający się na teorii emocji, mierząc pozytywne i negatywne stany afektywne.
- **Skala Samokontroli SSCS** nawiązująca do koncepcji samokontroli jako zdolności do regulowania impulsów i zachowań zgodnie z długoterminowymi celami.

Podział pracy

Promptowanie modeli - Hubert Sobociński
Analiza oraz interpretacja wyników - Paweł Florek
Stworzenie raportu - Paweł Pozorski

Inspiracja artykułu

W swojej pracy Petrov, Serapio-García i Rentfrow (2024) [9] zaproponowali, żeby swoje badania zmienić w kontekście używanych promptów. Ich propozycją było użycie innych opisów osób zamiast opisów ze zbiorów **Generic persona** oraz **Silocon persona**. Dodatkowo zaproponowano użycie konwersacji w promptach oraz części emocjonalnej. Według autorów te zmiany mogą poprawić wyniki modeli w testach psychologicznych.

Źródła

1. Llama 3.1 8B Instrunct - Q4 K M - HugginFace
2. Granite 3.3 8B Instruct - Q4 K M - HuggingFace
3. Aya 23 8B - Q4 K M – HuggingFace
4. Alif 1.0 8B Instruct - Q4 K M - HuggingFace
5. Ministral 8B Instruct 2410 - Q4 K M - HuggingFace
6. InternLM3 8B Instruct - Q4 K M - HuggingFace
7. PersonaHub Dataset - Huggingface
8. Daily Dialog Dataset - Hugginface
9. Petrov, N. B., Serapio-García, G., & Rentfrow, J. (2024). *Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis*. arXiv preprint arXiv:2405.07248: https://arxiv.org/abs/2405.07248