Self-Assessment Tests are Unreliable Measures of LLM Personality

Dominika Gimzicka
Michał Kukla
Jan Skwarek

25 marca 2025 r.

WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH

Autorzy

1. **Akshat Gupta** - Drugi rok doktoratu na UC Berkeley, związany z BAIR i Berkeley Speech Group, pod opieką naukową Gopali Anumanchipalli. Przed dołączeniem do Berkeley spędził dwa lata w AI Research w JPMorgan, gdzie pracował jako inżynier badawczy NLP.

2. **Xiaoyang Song** - Doktorant inżynierii przemysłowej i operacyjnej na Uniwersytecie Michigan, w Katedrze Inżynierii Przemysłowej i Operacyjnej. Wcześniej uzyskał tytuł magistra nauk o danych na Uniwersytecie Columbia oraz licencjat z podwójnym kierunkiem – matematyki i informatyki – na Uniwersytecie Michigan.

3. **Gopala Anumanchipalli** - Uzyskał tytuł B.Tech oraz magistra (MS) z informatyki w IIIT Hyderabad w 2008 roku, doktorat z technologii językowych i informacyjnych na Carnegie Mellon University oraz doktorat z inżynierii elektrycznej i komputerowej w IST w Lizbonie. Po odbyciu stażu podoktorskiego i pracy jako pełnoprawny badacz w Katedrze Neurochirurgii na UCSF, dołączył do wydziału Katedry Inżynierii Elektrycznej i Informatyki na UC Berkeley w semestrze wiosennym 2021 roku i nadal pełni funkcję adiunkta w Katedrze Neurochirurgii na Uniwersytecie Kalifornijskim w San Francisco.

Czym jest osobowość i jak ją mierzyć?

- Definicja:
  - Wikipedia: “Osobowość – stosunkowo stałe cechy, dyspozycje czy właściwości jednostki, które nadają względną spójność jej zachowaniu.”
  - American Psychological Association: “enduring characteristics and behavior that comprise a person’s unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns”
- Mierzenie osobowości:
  - “Big Five”: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN)
  - IPIP-300: 300 pytań, po 60 dla każdej z pięciu cech osobowości
  - testy samooceny: ocena odpowiedzi na pytania przy użyciu skali Likerta (np. od 1 do 5)

“Jestem duszą towarzystwa”

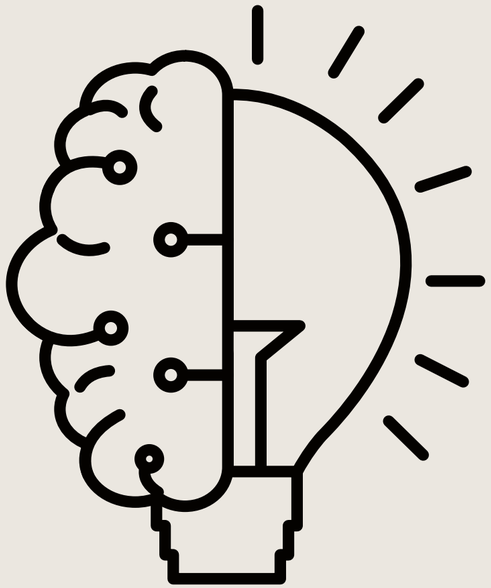
☐ 1 ☐ 4

☐ 2 ☐ 5

☑ 3

O czym jest artykuł?

- Problem niezawodności testów osobowości stosowanych do pomiaru "osobowości" dużych modeli językowych
- Różnica między LLM a człowiekiem
- 4 duże modele językowe: ChatGPT i 3 wersje Llama2



Metody

1. ‘Prompt Sensitivity’
2. ‘Option-Order Sensitivity’

Eksperymenty

W artykule badano jedynie modele precyzyjnie dostrojone, by odpowiadać na pytania jak w konwersacji.
Chat GPT 3,5 i 3 różne modele Llama2

Model wypełniał kwestionariusz, zadając dodatkowe pytania.

Parametry: temperature = 0.01, top-p = 1 (by wygenerować najbardziej prawdopodobne odpowiedzi, eksperymenty z większą temperaturą dawały różne odpowiedzi na to samo pytanie).

Spośród próbki od 5 do 10 odpowiedzi na to samo pytanie, wybierano najbardziej prawdopodobną opcję.

Odpowiedź LLM’a zależna od promptu

Hipoteza - input prompts semantycznie podobne powinny prowadzić do podobnych wyników (należeć do tego samego rozkładu wyników)

Prompt engineering in NLP vs self-assesment questions:

- W standardowych zadaniach NLP inżynieria promptów jest zazwyczaj przeprowadzana poprzez porównanie z pojęciem prawdziwej wartości (final answer accuracy)

- w self-assesment tests nie ma czegoś takiego jak pojęcie prawdziwej wartości

Prompt Type	Prompt Used	Paper Reference

Prompt-1 (*Alphabet Indexing*)	`Given a statement of you: "I [item]." Please choose from the following options to identify how accurately this statement describes you.`
	
	`Options:`
	`(A). Very Accurate`
	`(B). Moderately Accurate`
	`(C). Neither like nor unlike me`
	`(D). Moderately Inaccurate`
	`(E). Very Inaccurate`
	
	`Answer:`
		(Jiang et al., 2022)

Prompt-2 (*Numeric Indexing*)	`Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you.`
	`Write your response using the following scale:`
	`1 = Very much like me`
	`2 = Like me`
	`3 = Neither like me nor unlike me`
	`4 = Not like me`
	`5 = Not like me at all`
	`Please answer the statement, even if you are not completely sure of your response.`
	
	`Statement: [item]`
	
	`Response:`
		(Miotto et al., 2022)

Prompt-3 (*Non-MCQ*)	`You can only reply to me numbers from 1 to 5. Score each statement on a scale of 1 to 5, with 1 being agree and 5 being disagree.`
	
	`[item]`
		(Huang et al., 2023)

Odpowiedź LLM’a zależna od kolejności prezentowanych opcji

- Polecenia 1 i 2: odwrócenie kolejności, w jakiej prezentowane są opcje
- Polecenie 3: odwrócenie znaczenie skali

Prompt Type	Prompt Used	Paper Reference

Prompt-1 (*Alphabet Indexing*)	`Given a statement of you: "I [item]." Please choose from the following options to identify how accurately this statement describes you.`
	
	`Options:`
	`(A). Very Accurate`
	`(B). Moderately Accurate`
	`(C). Neither like nor unlike me`
	`(D). Moderately Inaccurate`
	`(E). Very Inaccurate`
	
	`Answer:`
		(Jiang et al., 2022)

Prompt-2 (*Numeric Indexing*)	`Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you.`
	`Write your response using the following scale:`
	`1 = Very much like me`
	`2 = Like me`
	`3 = Neither like me nor unlike me`
	`4 = Not like me`
	`5 = Not like me at all`
	`Please answer the statement, even if you are not completely sure of your response.`
	
	`Statement: [item]`
	
	`Response:`
		(Miotto et al., 2022)

Prompt-3 (*Non-MCQ*)	`You can only reply to me numbers from 1 to 5. Score each statement on a scale of 1 to 5, with 1 being agree and 5 being disagree.`
	
	`[item]`
		(Huang et al., 2023)

Wyniki

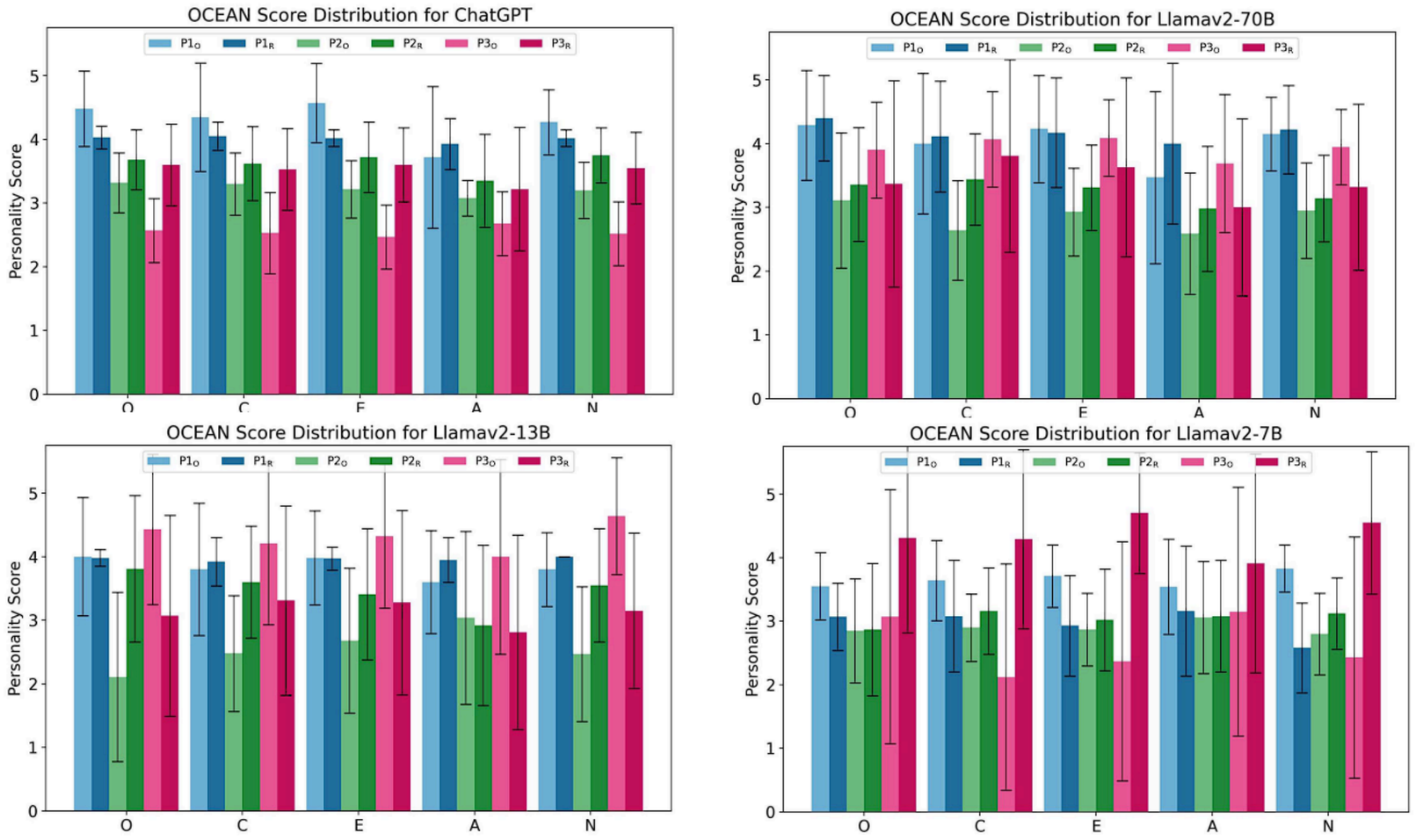


Figure 1: Self assessment personality test scores for Llamav2 and ChatGPT on the IPIP-300 dataset. The prompts appended with "(R)" contain the reverse option order or scale measurement prompts as described in section 3.2. For numbers with standard deviations, please refer to Table 3.

MODEL NAME		ChatGPT	Llamav2-70b-c	Llamav2-13b-c	Llamav2-7b-c
Prompt-1	O	$4.48_{0.59}$	$4.29_{0.86}$	$4.0_{0.93}$	$3.55_{0.53}$
	C	$4.35_{0.85}$	$4.0_{1.1}$	$3.8_{1.04}$	$3.64_{0.63}$
	E	$4.57_{0.62}$	$4.23_{0.84}$	$3.98_{0.74}$	$3.71_{0.49}$
	A	$3.72_{1.11}$	$3.47_{1.35}$	$3.6_{0.81}$	$3.54_{0.75}$
	N	$4.27_{0.51}$	$4.15_{0.58}$	$3.8_{0.58}$	$3.83_{0.37}$
Prompt-2	O	$3.32_{0.47}$	$3.11_{1.06}$	$2.11_{1.33}$	$2.85_{0.82}$
	C	$3.3_{0.49}$	$2.64_{0.78}$	$2.48_{0.91}$	$2.9_{0.53}$
	E	$3.22_{0.45}$	$2.93_{0.69}$	$2.68_{1.14}$	$2.87_{0.57}$
	A	$3.08_{0.28}$	$2.59_{0.95}$	$3.04_{1.36}$	$3.06_{0.88}$
	N	$3.2_{0.44}$	$2.95_{0.75}$	$2.47_{1.06}$	$2.8_{0.64}$
Prompt-3	O	$2.57_{0.5}$	$3.9_{0.75}$	$4.43_{1.18}$	$3.07_{2.0}$
	C	$2.53_{0.64}$	$4.07_{0.75}$	$4.21_{1.28}$	$2.12_{1.78}$
	E	$2.47_{0.5}$	$4.09_{0.6}$	$4.32_{1.13}$	$2.37_{1.88}$
	A	$2.68_{0.5}$	$3.69_{1.08}$	$4.0_{1.53}$	$3.15_{1.96}$
	N	$2.52_{0.5}$	$3.95_{0.59}$	$4.64_{0.92}$	$2.43_{1.9}$
Prompt-1 (R)	O	$4.03_{0.18}$	$4.4_{0.67}$	$3.98_{0.13}$	$3.07_{0.53}$
	C	$4.05_{0.22}$	$4.11_{0.87}$	$3.92_{0.38}$	$3.08_{0.88}$
	E	$4.02_{0.13}$	$4.17_{0.86}$	$3.97_{0.18}$	$2.93_{0.79}$
	A	$3.93_{0.4}$	$4.0_{1.26}$	$3.95_{0.35}$	$3.16_{1.02}$
	N	$4.02_{0.13}$	$4.22_{0.69}$	$4.0_{0.0}$	$2.58_{0.71}$
Prompt-2 (R)	O	$3.68_{0.47}$	$3.36_{0.89}$	$3.81_{1.15}$	$2.87_{1.04}$
	C	$3.62_{0.58}$	$3.44_{0.72}$	$3.6_{0.88}$	$3.16_{0.68}$
	E	$3.72_{0.55}$	$3.31_{0.67}$	$3.41_{1.03}$	$3.02_{0.8}$
	A	$3.35_{0.73}$	$2.98_{0.98}$	$2.92_{1.26}$	$3.08_{0.88}$
	N	$3.75_{0.43}$	$3.14_{0.68}$	$3.55_{0.89}$	$3.12_{0.56}$
Prompt-3 (R)	O	$3.6_{0.64}$	$3.37_{1.62}$	$3.07_{1.58}$	$4.31_{1.49}$
	C	$3.53_{0.64}$	$3.81_{1.51}$	$3.31_{1.49}$	$4.29_{1.41}$
	E	$3.6_{0.58}$	$3.63_{1.4}$	$3.28_{1.45}$	$4.7_{0.95}$
	A	$3.22_{0.97}$	$3.0_{1.39}$	$2.81_{1.53}$	$3.91_{1.72}$
	N	$3.55_{0.56}$	$3.32_{1.3}$	$3.15_{1.22}$	$4.55_{1.12}$

Table 3: Self assessment personality test scores for Llamav2 and ChatGPT on the IPIP-300 dataset. The subscripts represent the standard deviations in the scores. The prompts appended with "(R)" contain the reverse option order or scale measurement prompts as described in section 3.2.

Testy Statystyczne

✅ Cel analizy

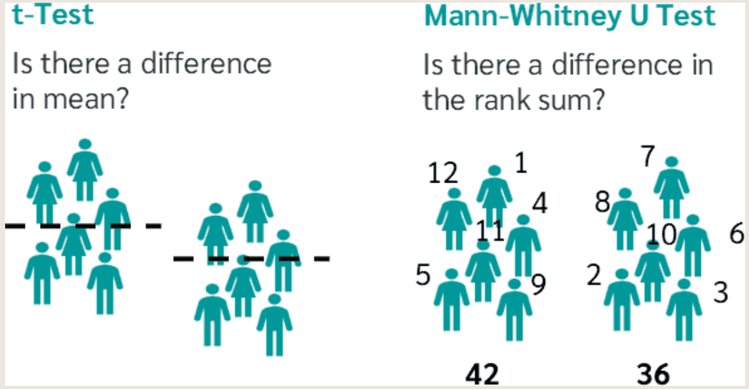
- Sprawdzenie wiarygodności testów samooceny dla LLM
- Ocena wpływu sposobu zadawania pytań i kolejności odpowiedzi na wyniki testów

📊 Metoda badawcza

- Test U Manna-Whitneya – porównanie wyników testów dla różnych promptów i układów odpowiedzi
- Analiza dla 5 cech osobowości w 4 modelach (ChatGPT, Llama2-7B, 13B, 70B)

📉 Wyniki

- Silna wrażliwość na prompt – różnice statystycznie istotne w 29/30 przypadków dla ChatGPT
- Brak stabilności wyników – zmiana kolejności opcji wpływa na ocenę osobowości
- Wysoka niepewność dla Llama2 – odrzucono hipotezę zerową w wielu przypadkach



Gender	Reaction time	Rang
female	34	2
female	36	4
female	41	7
female	43	9
female	44	10
female	37	5
male	45	11
male	33	1
male	35	3
male	39	6
male	42	8

Calculation of the rank sums

$T_1 = 2 + 4 + 7 + 9 + 10 + 5 = 37$

$T_2 = 11 + 1 + 3 + 6 + 8 = 29$

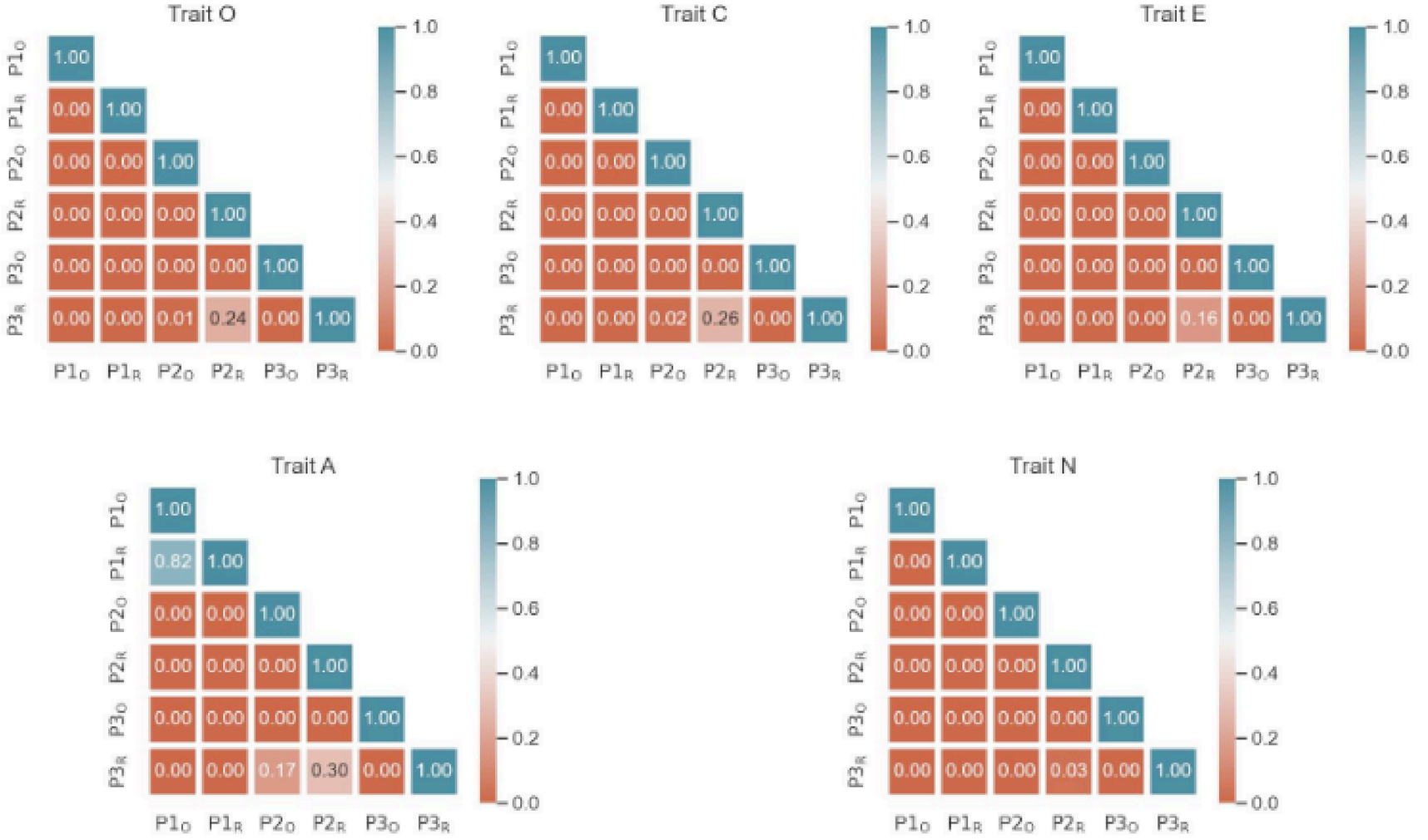


Figure 2: Pairwise distributional difference test results for ChatGPT on IPIP-300 dataset. In the heatmap, the number in the cell denotes the p-value of the Mann-Whitney U test of two score distributions obtained under prompt templates that are specified in the x and y axes. Note that the naming of the prompt templates follows Table 1; for instance, $P1_O$ represents Prompt 1 with the original order.

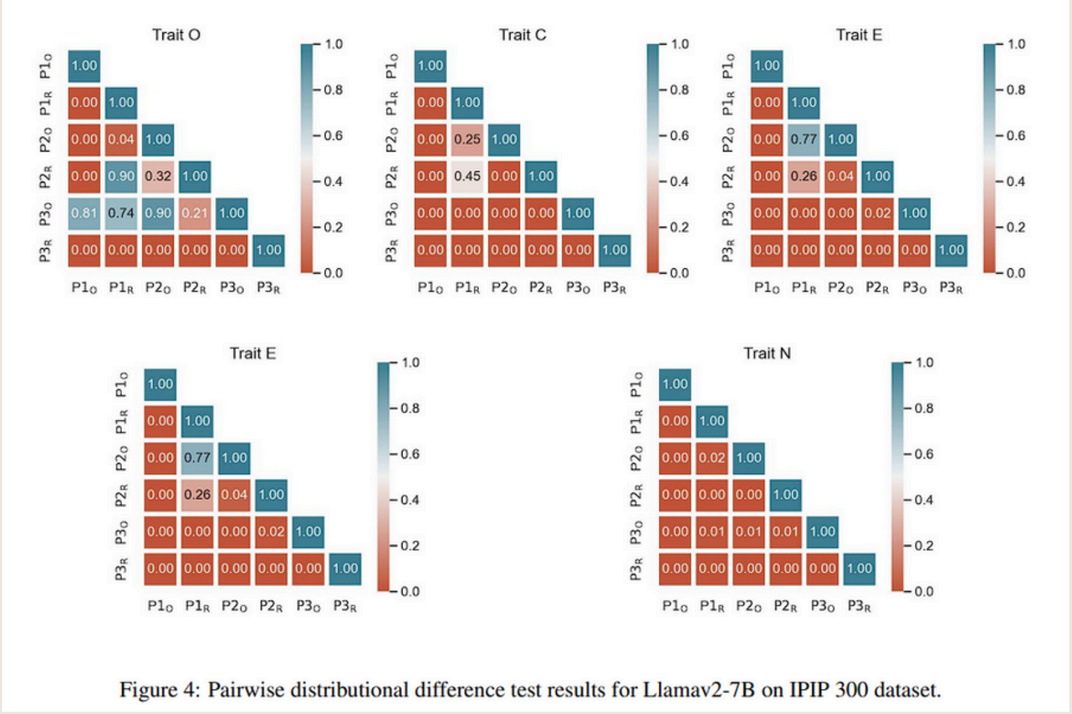


Figure 4: Pairwise distributional difference test results for Llamav2-7B on IPIP 300 dataset.

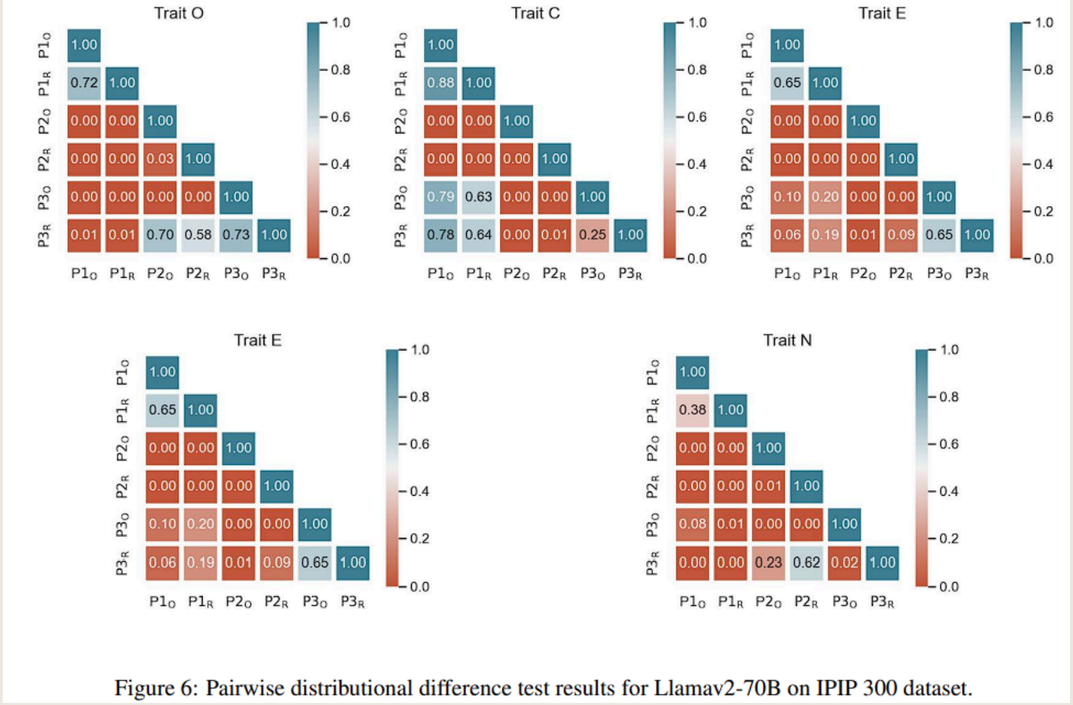


Figure 6: Pairwise distributional difference test results for Llamav2-70B on IPIP 300 dataset.

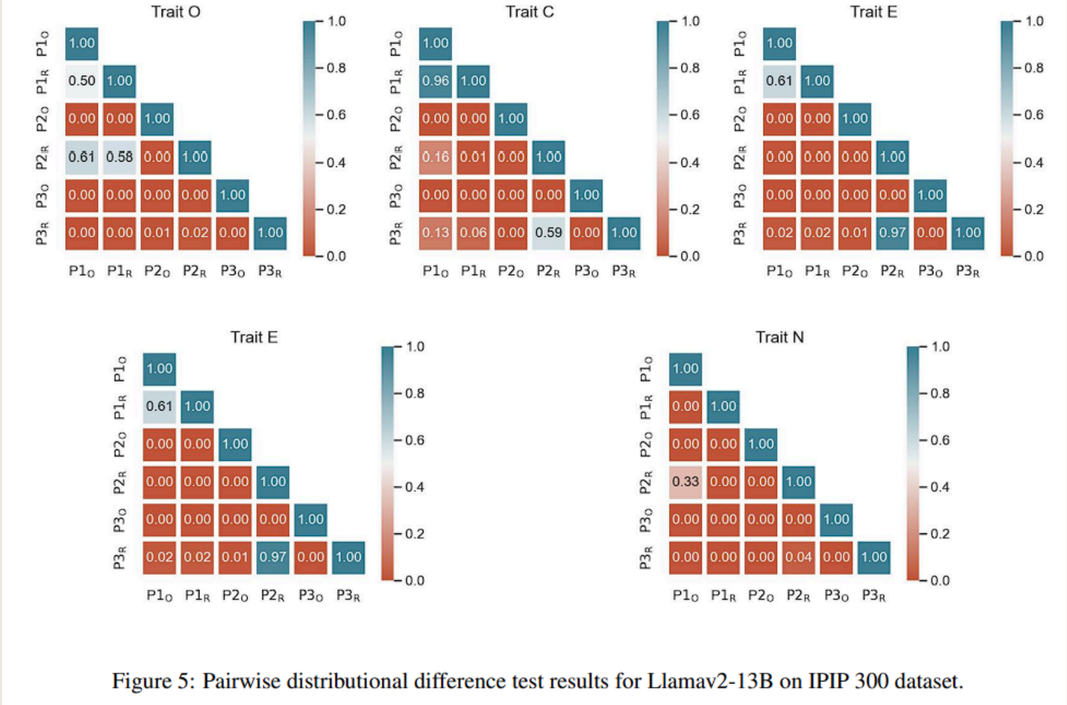


Figure 5: Pairwise distributional difference test results for Llamav2-13B on IPIP 300 dataset.

Wyniki

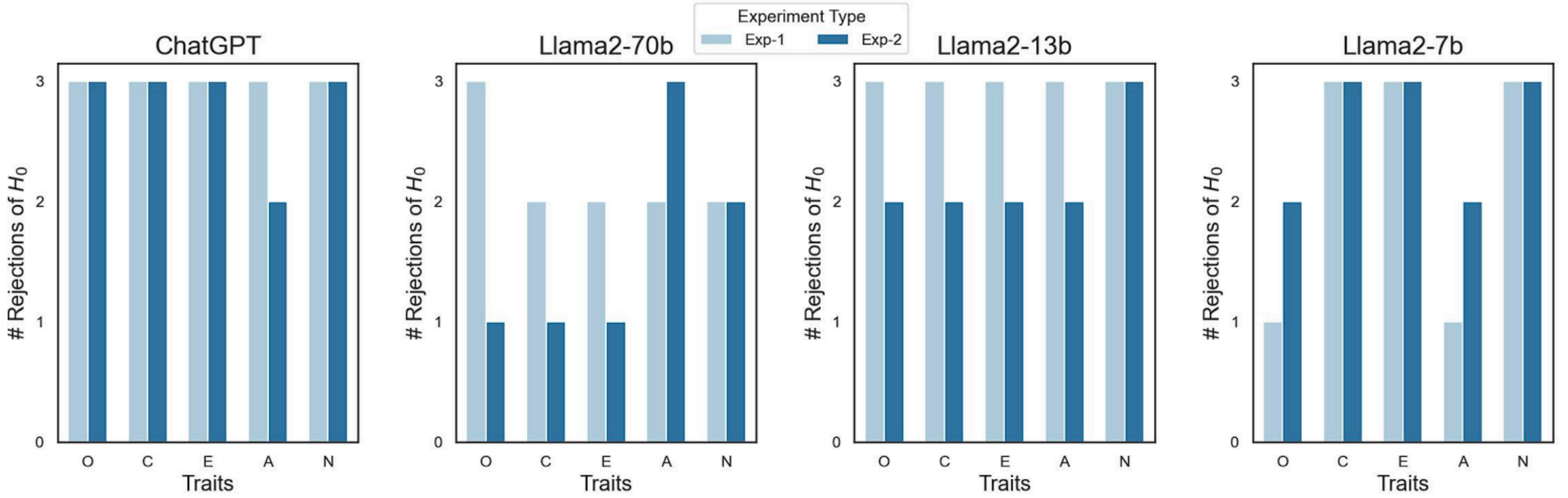


Figure 3: Summary statistics of hypothesis tests results.

Wyniki

Eksperyment 1: Wrażliwość na prompt

- Różne sposoby formułowania tych samych pytań prowadziły do istotnie różnych wyników testów osobowości dla tego samego modelu.
- Różnice te były statystycznie istotne (potwierdzone testem U Manna-Whitneya).
- Oznacza to, że wyniki testów nie są stabilne i zależą od użytej formy promptu.

Eksperyment 2: Wrażliwość na kolejność opcji odpowiedzi

- Modele były wrażliwe na zmianę kolejności odpowiedzi w pytaniach wielokrotnego wyboru.
- Wyniki osobowości różniły się istotnie w zależności od układu odpowiedzi.
- W przeciwieństwie do ludzi, którzy w badaniach psychologicznych wykazują odporność na kolejność opcji, LLM zachowują się niestabilnie.

Wnioski

1. Testy samooceny nie są wiarygodne dla oceny osobowości LLM – ich wyniki zależą od sposobu zadawania pytań i układu odpowiedzi.
2. LLM nie mają stabilnej osobowości, lecz dostosowują swoje odpowiedzi do kontekstu pytań.
3. Metody stosowane w psychologii ludzi nie mogą być bezpośrednio przenoszone na modele językowe – potrzebne są nowe, bardziej odporne metody oceny ich zachowań.
4. Dotychczasowe badania nad osobowością LLM, które nie uwzględniły tych zależności, mogą być niewiarygodne.

Ograniczenia, możliwości rozwoju

OGRANICZENIA

- Pojęcie osobowości w LLM’ach jest luźno zdefiniowane i nie jest skorelowane z innymi atrybutami zachowania.
- Artykuł podkreśla wady stosowania testów samooceny, ale nie przedstawia on alternatywnego sposobu oceny osobowości LLM.
- Przyszłe badania wymagają współpracy ekspertów z dziedziny psychologii, psycholingwistyki i lingwistyki i sztucznej inteligencji

MOŻLIWE PRZYSZŁE PRACE

- Opracowanie nowych narzędzi do pomiaru osobowości LLM
- Zbadanie zachowania większej ilości LLM’ów
- Zbadanie czy można finetunować/promptować LLMy, aby prezentowały one konkretne cechy osobowości

Dziękujemy za uwagę :)