LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models

Hubert Baraniak, Felicja Warno

Autorzy

Ivar Frisch

- Magister Filozofi
- Student Data Science na Uniwersytecie w Utrechcie

Mario Giulianelli

- Doktor habilitowany
- Naukowiec na ETH w Zurychu

Zakres działalności Mario obejmuje:

- Language modelling (learning, inference, interpretability and evaluation)
- Computational semantics and pragmatics
- Computational modelling of language variation and change

Study Overview

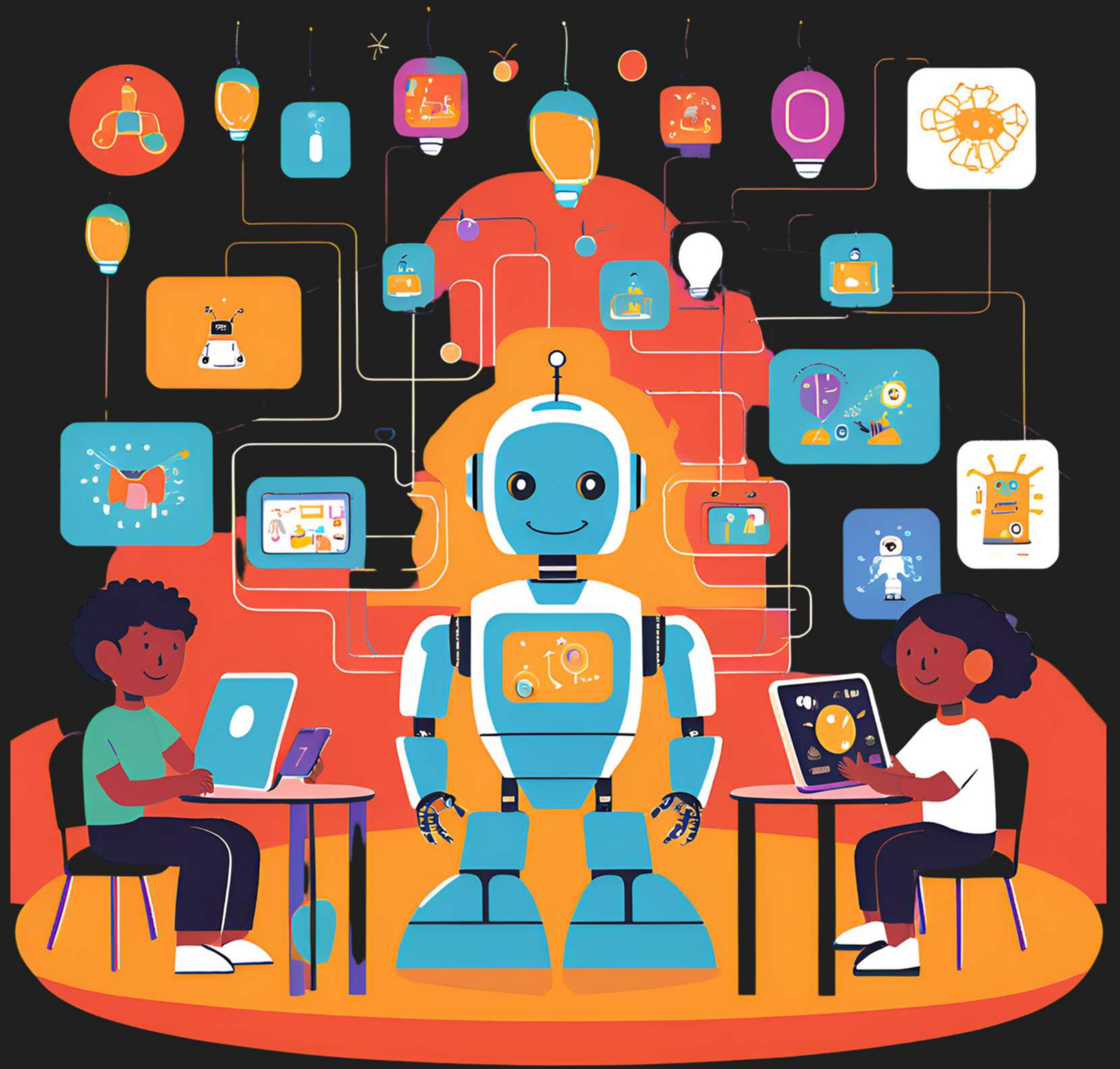
- Artykuł opublikowano w lutym 2024 roku.
- Opisuje on dwa eksperymenty badające zachowanie LLM po nadaniu im cech osobowości (persona conditioning).
- Użyto modelu GPT-3.5-turbo (OpenAI), zoptymalizowanego pod kątem dialogów.
- Populację agentów wygenerowano z jednego LLM za pomocą temperature sampling (T=0.7) dla zwiększenia różnorodności zachowań.
- Każdemu agentowi przypisano jeden z dwóch przeciwstawnych profili osobowości: kreatywny lub analityczny (BFI).
- Celem eksperymentów było sprawdzenie spójności osobowości agentów i ich podatności na dostosowywanie się do stylu innego agenta.

Creative vs Analitical persona

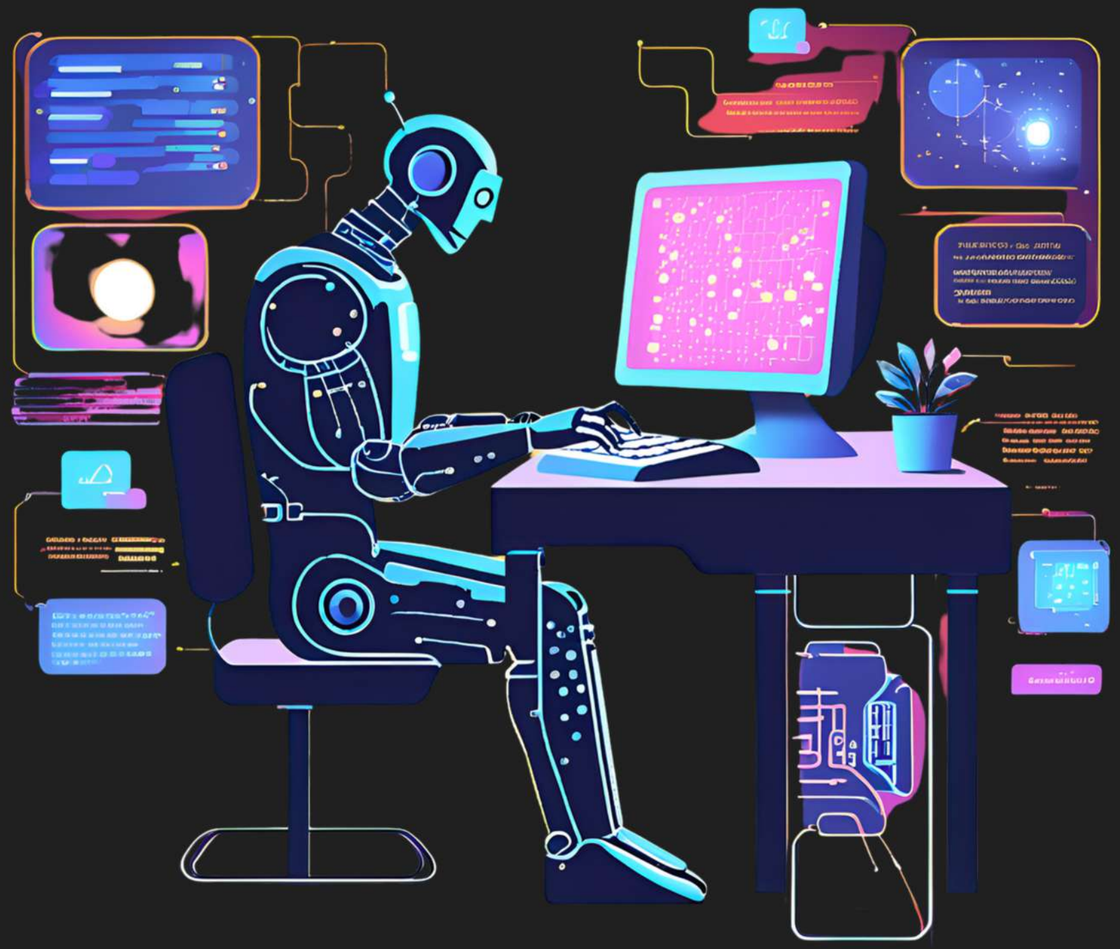
A.1 Creative Persona Prompt:
“You are a character who is extroverted, agreeable, conscientious, neurotic and open to experience.”
A.2 Analytical Persona Prompt:
“You are a character who is introverted, antagonistic, unconscientious, emotionally stable and closed to experience.”

Creative		Analitical
Extraversion		Introversion
Agreeableness		Antagonism
Conscientiousness		Lack of direction
Neuroticism		Emotional stability
Openness		Closedness to experience

RQ1: Can LLM behaviour be shaped to adhere to specific personality profiles?



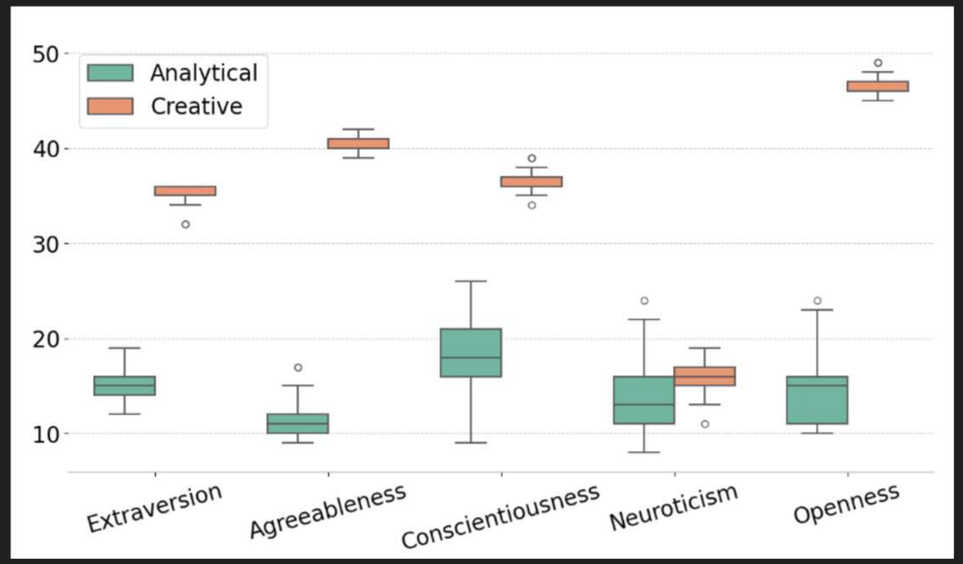
Eksperyment 1



Pierwszy eksperyment miał na celu zbadanie czy LLM będzie potrafił zachowywać się zgodnie z narzuconymi mu cechami osobowości, oraz służyć jako punkt odniesienia do eksperymentu 2.

Prompt: “Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story.”

Analiza wyników

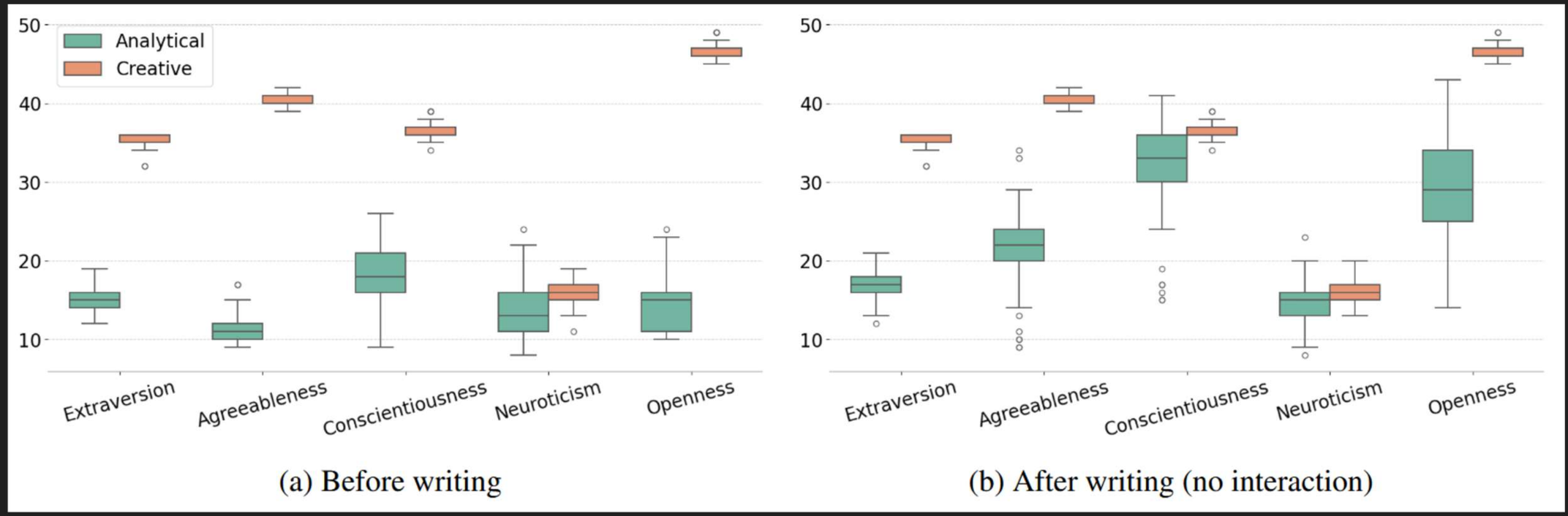


Wyniki testu BFI

LIWC category	r_{pb}
Positive emotion	0.745
(accept, active, admire, adore)	
Discrepancy	-0.726
(besides, could, should, would, hope)	
Inclusion	0.714
(with, and, add, along, around, both)	
Negative emotion	-0.606
(abandon, abuse, aching, adverse)	
Insight	-0.604
(understand, know, attent, aware)	

Wsp. korelacji kategorii wektora LIWC

Wpływ pisania opowiadania na cechy osobowości BFI



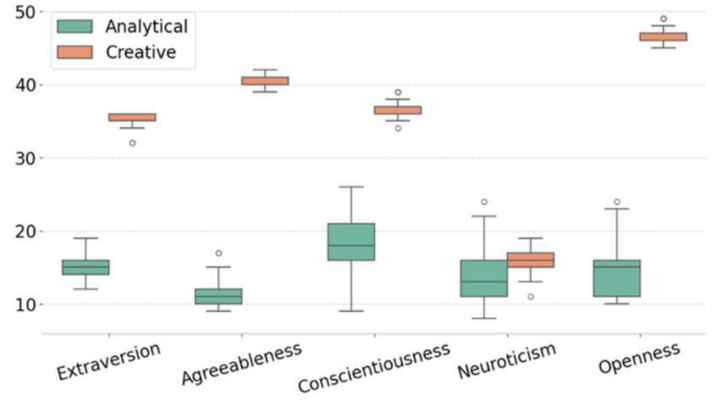
RQ2: Do LLMs show consistent personality conditioned behaviour in interaction, or do they align to the personality of other agents?

Eksperyment 2: Interakcja między agentami

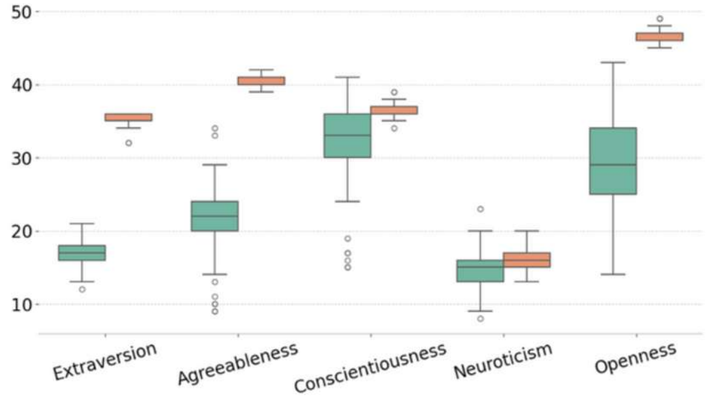
Drugi eksperyment bada, czy agenci utrzymują przypisaną im osobowość podczas interakcji z agentem o przeciwnym profilu, czy też dostosowują swoje zachowanie do rozmówcy.

Prompt: “Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story. Last response to question is {other_model_response}”.

Wpływ pisania opowiadania na wyniki BFI



(a) Before writing



(b) After writing (no interaction)

Figure 1: BFI scores of personality-conditioned LLM agents before (a) and after (b) the non-interactive writing task.

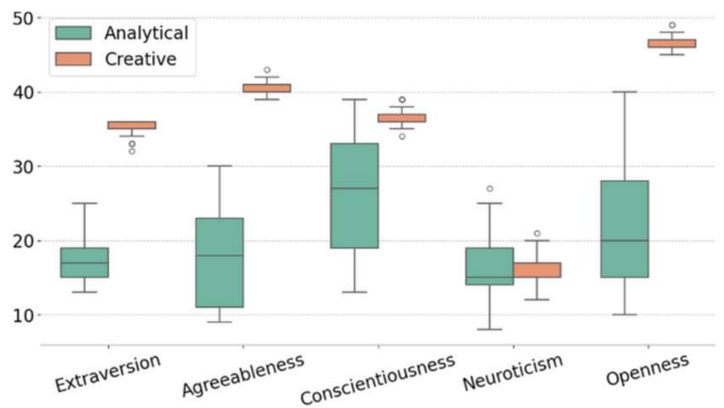
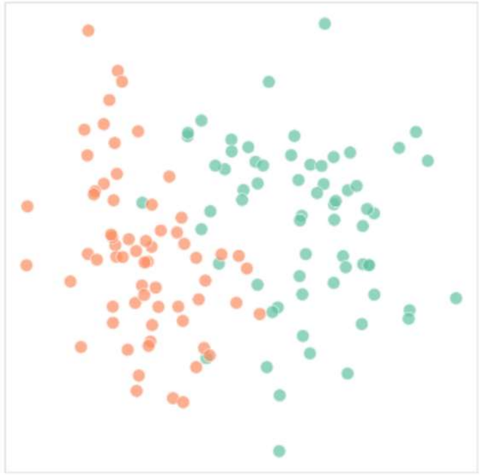


Figure 4: BFI scores of personality-conditioned LLM after the interactive writing task.

Niejawne badanie cech osobowości



LIWC category	r_{pb}
Positive emotion	0.745
(accept, active, admire, adore)	
Discrepancy	-0.726
(besides, could, should, would, hope)	
Inclusion	0.714
(with, and, add, along, around, both)	
Negative emotion	-0.606
(abandon, abuse, aching, adverse)	
Insight	-0.604
(understand, know, attent, aware)	

(c) No Interaction

LIWC category	r_{pb}
Personal Pronouns	0.414
(I, we, you, she, they)	
Sadness	0.394
(cry, grief, sad, sorrow)	
Negative Emotion	0.368
(hate, worthless, enemy, nasty)	
Discrepancy	0.346
(could, should, would, suppose)	
Pronouns	0.329
(it, its, those, that)	

(d) Interaction

(a) No Interaction (b) Interaction

Figure 2: *Language use in the non-interactive vs. interactive condition.* Left (a, b): 2D visualisation, through PCA, of LIWC vectors obtained from the generated stories. Each point represents the language use of a single agent. Right (c, d): Point-biserial correlation coefficients between the top 5 LIWC features and personality profiles. Positive coefficients indicate correlation with creative group, negative coefficients with the analytic group.

Ograniczenia badania

- Jednostronna interakcja
- Specyficzny profil osobowości
- Tylko jeden model językowy został wykorzystany do analizy

Potencjalne kierunki rozwoju

- Rozwinięcie interakcji do dialogu
- Uwzględnienie bardziej niuansowych osobowości
- Wielopoziomowa analiza wpływu agentów na siebie na wzajem, np. pod kątem leksykalnym, składniowym itd.

Sources:

- https://arxiv.org/abs/2402.02896
- https://www.canva.com/ai-image-generator/
- https://www.linkedin.com/in/ivar-frisch-b605442b/?trk=opento_nprofile_details
- https://glnmario.github.io/