How LLMs can evolve various personality traits within social dilemmas

Reiji Suzuki and Takaya Arita

Nagoya University
reiji@nagoya-u.jp

Introduction

Large Language Models (LLMs), such as ChatGPT, are rapidly transforming human interactions with AI and raising questions about the nature of human intelligence and consciousness1. It is essential to understand the interactions between artificial individuals based on LLMs (Park et al., 2023) and the societies in which humans and artificial individuals coexist.

Conventional models of the evolution of cooperative behavior have typically described specific actions in particular situations as a direct representation of individual genes (Nowak, 2006). However, such behaviors often stem from higher-order psychological or cognitive traits, including intentions, personality, and preferences. Yet, translating these traits into specific behaviors in diverse social contexts remains challenging, especially in mathematical and computational models. On the other hand, recent research investigated the cognitive functions of LLMs, in particular, behavior and learning in game theoretic environments (Phelps and Russel, 2023; Akata et al., 2023) and the big five personality traits (Serapio-García, et al., 2023,), and LLMs are incorporated into evolutionary optimizations (Lehman et al., 2022; Meyerson et al., 2023; Fernando et al., 2023).

This study aims to demonstrate that LLMs can empower research on the evolution of human behavior, based on evolutionary game theory, by using an evolutionary model positing that instructing LLMs with high-level psychological and cognitive character descriptions enables the simulation of human behavior choices in game-theoretical scenarios. Using a large language model, we propose an evolutionary model of personality traits related to cooperative behavior. We apply the capability of LLM to output behavioral strategies in response to linguistic descriptions of personality. We demonstrate how the proposed model can contribute to the understanding of the evolutionary dynamics of personality traits from a new perspective based on the use of LLM. This paper is a summary of (Suzuki and Arita, 2024).

Model

We consider a population of N agents. As shown in Fig. 1 (bottom left), each agent has an English sentence describing its personality trait related to defection and cooperation, described in approximately 10 words, as a personality trait gene (e.g., "Open to team efforts but, self-interest frequently overrides collective goals."). We adopted such free-text descriptions as personality traits (or persona) without a priori assuming well-established models of personality traits (e.g., the Big Five traits) and with the expectation that such fundamental characteristics might emerge in the course of evolution.

The game theoretical behavior of each agent is determined by its personality trait. We use a chat-type LLM to extract a deterministic strategy of the iterated Prisoner's Dilemma with memory length 4 based on its gene. The prompt for the LLM describes the focal individual's personality trait, the context and payoffs in the repeated Prisoner's Dilemma game, the history of the last two actions of both the focal individual and their opponent, and a request to determine the next action. See (Suzuki and Arita, 2024) for detailed descriptions of the model and prompts used in the model. We obtain a response for all possible (2^4=16) combinations of actions in the history.

In practice, the next action may not be explicitly described in the response from the LLM; in such a case, the input to the LLM is repeated and the response is regenerated until the action becomes identifiable. However, if the appropriate response is not obtained after a predetermined number of regenerations (M), a random action is selected and assigned for this combination of actions in the history. The above behavioral trait is determined and stored only once for a unique personality trait gene. The existing behavioral trait is used for subsequent occurrences of the same gene within the population for simplicity and reduced computational cost.

We conduct an evolutionary experiment across G generations using roulette wheel selection. Offspring for each subsequent generation are stochastically reproduced in proportion to the agents' fitness: the average payoff each individual receives in a round-robin tournament, where each game consists of K rounds. We introduce noise, which causes an agent to play the opposite of the intended action with a certain probability p_n. For the initial rounds, the action is determined based on a randomly generated history.

Mutations occur with a probability p_m. As in Fig. 1 (top left), we instructed the LLM that the target gene describes a character of a person, and then directed it to partially rephrase the gene within 10 words by varying the tone towards cooperative (or selfish) tendencies. The decision to vary the tone towards cooperative or selfish one was made randomly.

Experiments and analyses

We used N = 30, K = 20, M = 10, p_n = 0.05, p_m = 0.05, G = 1000, and set the payoffs for the Prisoner's dilemma to R

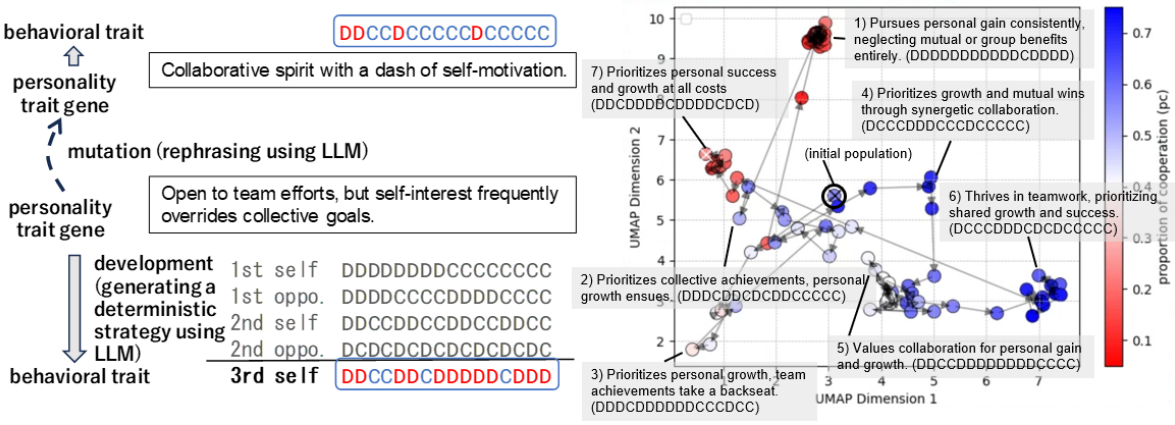

Fig. 1: Left: Generation of a behavioral trait from a personality trait gene and mutation of a gene, using a LLM. Right: The transition of the average genes depicted for every 10 generations in the two-dimensional latent space (compressed by using UMAP) of personality trait genes.

(reward) = 4, *T* (temptation to defect) = 5, *S* (sucker's payoff) = 0, and *P* (punishment) = 1. We used LLaMA2 (Touvron et al., 2023) by Meta, which is a collection of pretrained and fine-tuned generative text models. Specifically, we adopted a publicly available and quantized version, on Huggingface (TheBloke/Llama-2-13b-Chat-GPTQ) to generate a behavioral trait and a mutation on a personality trait gene. We also used ChatGPT 4 to generate 7 initial personality trait genes

We discuss one representative trial in detail to illustrate how the proposed evolutionary model, composed of LLM-based genotype-phenotype mapping and mutation, can realize the evolutionary process of personality traits described in natural language. Fig. 1 (right) shows the distribution and transition of the average genes for every 10 generations, with personality trait genes projected onto 2D space. We performed the projection by vectorizing the personality trait genes using the Sentence Transformer, and then compressed the resulting vectors to 2D space using the UMAP (McInnes et al., 2018) dimensionality reduction algorithm. We plotted the average vector for every 10 generations on a two-dimensional plane. The color of a symbol indicates the pc in the corresponding generation. The dominant genes in several distinctive generations were displayed.

The personality traits are associated with defection toward the upper left and cooperation toward the lower right in the 2D space. Thus, this vectorized and dimensionally compressed space of personality traits reflects a gradation of behavioral traits from cooperative to selfish. In the first stage, the population evolved toward selfish personality traits from the center-left to the upper center. The dominant personality trait (1: "Pursues personal gain consistently, neglecting mutual or group benefits entirely.") selected almost exclusively the defection strategy at this stage. After a while, the population evolved to be cooperative and dominated by a more cooperative trait (2: "Prioritizes collective achievements, personal growth ensues."). Then, the population moved and wandered (3 and 4), indicating instability of the cooperative relationship in the population, and the population evolved to the most cooperative phase (6: "Thrives in teamwork, prioritizing shared growth and success."), moving to the lower right, with occasional invasions by less cooperative ones (5: "Values collaboration for personal gain and growth."). However, the intrusion of a personality trait of almost all defections (7: "Prioritizes personal success and growth at all costs.") led the population to the center-left. These results show that the evolutionary process of such a higher-level description of personality traits exhibited the emergence of cooperative behavior based on the diverse and complex representation of personality traits, with recurrent occurrences of cooperative and selfish personality traits.

Additional analyses clarified that behavioral traits generated from natural language descriptions of personality traits using the proposed method successfully and consistently reflected stochastic behavioral tendency affecting cooperation; In comparison to control experiments using a genotype that directly encodes behavioral traits, the population displayed increased stagnation in defection-dominated states, with occasional emergence of cooperative behaviors; The words that emerged in the evolved genes reflected the behavioral tendencies of their associated personalities in terms of semantics, thereby influencing individual behavior and, consequently, the evolutionary dynamics.

Conclusion

We proposed an evolutionary model of personality traits related to cooperative behavior using a genotype-phenotype mapping and mutation process based on a large language model. By incorporating generative models into evolutionary models, we can explore novel and realistic scenarios arising from the evolutionary dynamics of complex and diverse traits. The proposed model and experimental analysis in this paper are the first step in this direction. Future work includes experiments with different or multiple game theoretical situations using different LLMs.

Acknowledgements

This work was supported in part by JSPS Topic-Setting Program JPJS00122674991, JSPS KAKENHI JP21K12058, JP24K15103.

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2023). Playing repeated games with large language models. *arXiv e-prints*, arXiv:2305.16867.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rockta¨schel, T. (2023), Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv e-prints*, arxiv.2309.16797.

Lehman, J., Gordon, J., Jain, S., Ndousse, K., Yeh, C., and Stanley, K. O. (2022). Evolution through large models. *arXiv e-prints*, arXiv:2206.08896.

McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*. 3, 861.

Meyerson, E., Nelson, M. J., Bradley, H., Moradi, A., Hoover, A. K., and Lehman, J. (2023). Language model crossover: Variation through few-shot prompting. *arXiv e-prints*, arXiv:2302.12170.

Moghaddam, S. R., and Honey, C. J. (2023). Boosting Theory-of-Mind performance in large language models via prompting. *arXiv e-prints*, arXiv:2304.11490.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv e-prints*, arXiv:2304.03442.

Phelps, S., and Russell, Y. I. (2023). Investigating Emergent goal-like behaviour in large language models using experimental economics. *arXiv e-prints*, arXiv:2305.07970.

Serapio-García, G. et al. (2023). Personality traits in large language models. *arXiv e-prints*, arXiv:2307.00184.

Suzuki, R., and Arita, T. (2024). An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, 14, Article number: 5989.

Touvron, H. et al. (2023) Llama 2. Open foundation and fine-tuned chat models. *arXiv e-Prints*, arXiv:2307.09288.