BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Simulation of Human Interaction using AI

Bc. David Chocholatý*

Abstract

Simulation of believable human interaction can strengthen the application of large language models (LLMs) in computational social sciences and improve the insights and value of market research using AI agents. In this work, a *PerSimChat* framework is designed that provides an experimental environment for simulating multiple human conversations using LLM agents with persona data. Simultaneously, a new approach is proposed for selecting the order of the agent's speech called *One-By-One Talk with Agent's Need to Talk*. Empirical studies demonstrate the framework's performance on many evaluation dimensions, and the system achieves competitive results with other multi-agent debate systems on reasoning and mathematics benchmarks.

*xchoch09@stud.fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

The current state of computational social sciences lacks a credible tool for modelling and simulating believable human behaviour and interaction between multiple personas. Using AI agents to represent multiple personas, current state-of-the-art solutions lack the realness of conversation and agents' personalities, supporting more than two personas in a discussion.

Thus, a new framework is needed, which captures the naturality of a human conversation using real persona data, the credible order of more than two speakers, and modelling the human brain's cognitive functions.

In this work, we propose a new framework called *PerSimChat* for the simulation of multiple personas communication, thus with free discussion and group debate with a consensual solution.

2. Related Work

The recent work can be classified into two groups, whether they use *Multi-Agent Debate* (MAD) architecture or, on the other hand, in some form, contain multi-agent free discussion.

In MAD [1], the multiple agents generate the answer simultaneously in so-called rounds. Many works were created in connection with [1], namely *MAD with Sparse Topology* [2], *MAD with Judge* [3], *ReConcile* [4], *ChatEval* [5], and *CMD* [6].

In summary, the MAD architecture does not correspond to how the real personas discuss with each other. Based on that, more natural approaches were proposed. Closest to our criteria is the *AutoGen* [7, 8] library and *SOTOPIA* [9] with *CAMEL* [10] frameworks. However, these works do not support more than two agents' discussion, usage of real persona data, speakers' order naturalness, or modelling brain cognitive functions simultaneously.

In terms of cognitive functions, the most advanced designs were introduced in *Generative Agents* [11], *Humanoid Agents* [12], and *MetaAgents* [13] which use cartoon agents and their interactions in the game environment. Other tools containing discussion are *ChatDev* [14] and *MetaGPT* [15] for programming, and *AutoAgents* [16] for general tasks.

3. PerSimChat Framework

In this work, we propose a new framework called *PerSimChat* to simulate the persona conversation. The approach of how this tool can be used is that for selected personas, conversation types, and provided tasks, personas naturally converse with each other.

Our work looks at discussion from two perspectives of conversation types: *Free Discussion* and *Group Debate*.

In free discussion, the personas converse in a standard manner for a predefined maximum number of mes-

sages. In a group discussion, the goal is to reach a final group consensus. In this type of communication, the independent *judge* model is added to analyse the group consensus and provide the final answer.

The PerSimChat framework introduces a new concept of speaker order selection with the *One-By-One Talk with Agent's Need to Talk* approach. After each message, the conversing personas generate the need to talk scores. Based on these scores, the next speaker is selected. Also, for this we support two approaches: *Maximum Likelihood* and *SoftMax*. In Maximum Likelihood, the persona with the highest need to talk score is chosen. The SoftMax works similarly to how the large language models (LLMs) select the next word in an output stream. With that, we can optionally force the speaker not to repeat twice in a row.

Other benefits of our solution are in the following terms: (1) we use real persona data or artificially generated (name and surname, description, characteristics, and traits), (2) including the concept of agent's *emotional state*, (3) agent's *planning* and *reflection*, (4) *short-term* and *long-term memory* concept with *memory consolidation*, (5) we take into account the persona *relationships* and *social goals*, (6) for group debate judge we also make available the persona architecture (including memory model with planning and reflection), and finally (7) we created an user interface to increase the usability of our tool.

4. Experimental Evaluation

The PerSimChat framework was tested with four evaluation scenarios, three for free discussion and one for group debate.

For group debate, we compared the tool with the existing solutions for a MAD architecture design. In this manner, we evaluate PerSimChat on four benchmarks, including two commonsense and two math. These are: (1) StrategyQA [17], (2) ECQA [18], (3) GSM8K [19], and (4) AQuA [20].

Although our framework's primary purpose is not to achieve the best scores in the reasoning and mathematical tasks, but to simulate the natural human conversation, our work achieves competitive results with other frameworks. Due to computational costs, testing was provided within three rounds using the OpenAI[1] gpt-4o-2024-08-06 model from Azure AI services[2], with a random subset of 30 tasks providing the mean and standard deviation. The PerSimChat framework achieves 71.7 ± 9.6 on Strategy QA, 60.0 ± 6.7 on ECQA, 88.9 ± 5.1 on GSM8k, and 77.8 ± 13.5 on Aqua benchmarks.

For the free discussion, first, the evaluation was made in the so-called dimension when a single LLM (gpt-4-turbo-2024-04-09) is prompted to evaluate the conversation. The PerSimChat framework outperforms the other two baseline solutions (single LLM conversation generation and baseline created with the AutoGen library [7]) in *believability*, *credibility*, *content depth and relevance* while achieving satisfactory results in other dimensions. For the *conversation closure* dimension, due to constraining the maximum number of messages, the single LLM still wins over the AutoGen and PerSimChat solutions. By replacing the GPT-4o model with the Lakmoos model and system, we can see the improvement in PerSimChat performance in most dimensions.

Secondly, the systems were compared in a pair-wise evaluation with the FairEval tool [21]. Similarly to the first comparison, the Lakmoos model and system outperform the GPT-4o model. Also, the Mistral AI[3] Mistral Small model wins over the GPT-4o model. Comparing PerSimChat with the AutoGen baseline, while achieving slightly better results, the performance is comparable.

Lastly, on multiple simulation tasks, the use case study was provided, highlighting the pros and cons of our solution.

5. Conclusions

This work proposes the *PerSimChat* framework for the simulation of believable human conversation. With that, we created a new approach of selecting the speaker order — *One-by-one Talk with Agent's Need to Talk*. We compared this tool's two possible use cases, the free discussion and group debate, with other available tools and baseline solutions. Our tool achieves better scores on naturalness dimensions in these experiments. With that, the user interface was created to increase the usability of the PerSimChat framework.

Acknowledgements

I sincerely thank my supervisor, Ing. Radek Hranický, Ph.D., for his help and expert guidance. I also thank Bc. Jan Polišenský for his advice and support on behalf of Lakmoos AI s.r.o. and the company for providing the computational resources.

[1] https://openai.com/

[2] https://azure.microsoft.com

[3] https://mistral.ai/

References

[1] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[2] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology, 2024.

[3] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024.

[4] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms, 2024.

[5] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023.

[6] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key?, 2024.

[7] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *COLM 2024*, August 2024.

[8] Victor Dibia, Jingya Chen, Gagan Bansal, Suff Syed, Adam Fourney, Erkang (Eric) Zhu, Chi Wang, and Saleema Amershi. Autogen studio: A no-code developer tool for building and debugging multi-agent systems. Preprint, August 2024.

[9] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024.

[10] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society, 2023.

[11] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[12] Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid agents: Platform for simulating human-like generative agents, 2023.

[13] Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, 2023.

[14] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development, 2024.

[15] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2024.

[16] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation, 2024.

[17] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, 2021.

[18] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online, August 2021. Association for Computational Linguistics.

[19] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton,

Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[20] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[21] Chandan Kumar Sah, Xiaoli Lian, Tony Xu, and Li Zhang. Faireval: Evaluating fairness in llm-based recommendations with personality awareness, 2025.