Simulation of Human Interaction using AI

Ing. David Chocholatý

Supervisor: Ing. Radek Hranický, Ph.D.

Faculty of Information Technology, Brno University of Technology

Introduction & Motivation

This work is motivated by the potential of large language model (LLM)-driven simulations to support research in computational social sciences and to improve the insights and value of market research using AI agents, where controlled yet realistic models of human communication are essential. By designing believable conversational agents that integrate cognitive functions, memory, and character traits, we aim to bridge the gap between artificial simulations and real-world social dynamics. The proposed framework not only enhances the realism of agent-based interactions but also provides a platform for experiments in **communication**, **debate**, and **group decision-making**.

Proposed Methods

The **PerSimChat** framework is introduced as an experimental environment for simulating multi-agent human conversations using LLMs. Inspired by [1], each agent is designed with cognitive modules (perception, reflection, planning, action, memory, and natural language, as shown in Figure 1), combined with a profile module that dynamically models emotional states and persona traits. This architecture allows agents to behave more consistently and believably in social interactions.

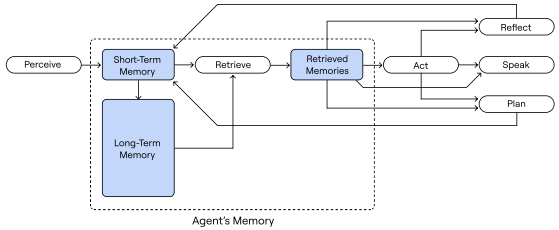

Figure 1. The architecture of cognitive modules in the PerSimChat framework.

A novel **One-by-One Talk with Agent's Need to Talk** strategy determines turn-taking. Instead of following fixed sequences, agents are assigned a dynamic "need to talk" score that reflects their internal states and emotional levels. This mechanism enhances the naturalness of conversations and more closely mirrors real-world interaction patterns. The framework supports both free discussion and structured debate scenarios, making it suitable for evaluating conversational realism, group decision-making, and knowledge exchange. To this end, **consensus openness** scores are generated during group discussions, while a **judge** agent determines when a consensual solution has been reached.

PerSimChat Web Application

In this work, the PerSimChat framework and a corresponding web application (see Figures 2 and 3) are designed to provide an experimental environment for simulating multi-agent human conversations using LLMs with persona data.

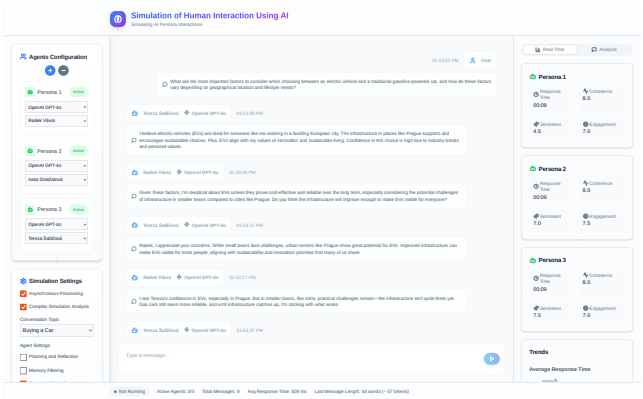

Figure 2. User interface of the PerSimChat web application.

Figure 3. The PerSimChat web application is publicly available at **https://persimchat.lakmoos.com**.

Experimental Results

The PerSimChat framework was evaluated in free discussions (natural conversations) and group debates, and compared against baseline systems: **Single-Agent**, which represents a zero-shot single GPT-4o model that generates the entire multi-persona conversation in a single step, and the **AutoGen** baseline solution, implemented using the Autogen library [3].

- **Free discussion**: A single GPT-4-turbo evaluator rated PerSimChat highest in believability, credibility, content depth, and flow (see Figure 4 and Table 1). In these experiments, the evaluation model received both the description of each evaluation dimension and the system's conversation script. Conversations produced by PerSimChat were more natural and better aligned with persona traits. The system was also compared with FairEval [2] in pairwise matches, obtaining competitive results. Replacing models with the Lakmoos system further improved outcomes (see Table 2).
- **Group debate**: For group debates, multiple benchmarking datasets with commonsense and mathematical tasks were used, and the system was compared against two single-agent solutions and three multi-agent debate systems. Although PerSimChat did not outperform all of these systems, its results were comparable, highlighting the effectiveness of the proposed approach.
- **Human study**: In a user study with 25 participants, users frequently preferred the single-LLM setup for its shorter, casual replies. Across systems, results appeared broadly comparable. Although this study did not detect notable differences, the findings indicate that PerSimChat can achieve conversational quality in line with existing approaches.

Overall, the framework consistently outperformed the baseline alternatives in terms of naturalness, demonstrating its ability to create more engaging and human-like conversational simulations. It achieved competitive results with other multi-agent debate systems on reasoning and mathematical benchmarks. Furthermore, replacing GPT-4o with the Lakmoos system led to even greater improvements.

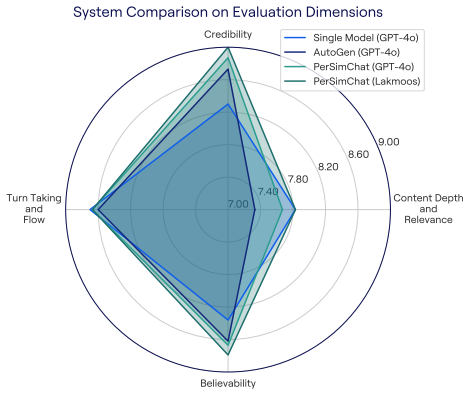

Figure 4. Comparison of the PerSimChat framework with baseline solutions using GPT-4o and the Lakmoos system. The dimensions represent naturalness evaluation criteria, for which a single GPT-4-turbo model is prompted to rate the quality of the generated conversation on a 0–10 scale, based on the textual description of each dimension.

Method	Believability	Credibility	Relevance
Single-Agent Zero-Shot	8.65 ± 0.82	7.96 ± 0.88	8.4 ± 0.81
AutoGen	8.67 ± 0.79	8.69 ± 1.51	8.15 ± 2.13
PerSimChat	$\mathbf{8.76 \pm 0.59}$	$\mathbf{8.78 \pm 0.52}$	$\mathbf{8.51 \pm 0.84}$

Table 1. Comparison of PerSimChat with baseline solutions on evaluation dimensions (Believability, Credibility, and Relevance). Each score ranges from 0 to 10, and each system was run for three rounds using a subset of 15 tasks, with a 20-message limit and three personas. The mean value and standard deviation are reported. Ratings were generated by a single GPT-4-turbo model using the textual descriptions of each dimension.

Model 1	Model 2	Wins (%)	Ties (%)	Losses (%)
GPT-4o	Lm	40.00	16.67	**43.33**
Mistral Small	Lm	34.48	6.90	**58.62**
Mistral Nemo	Lm	42.86	10.71	**46.43**

Table 2. FairEval evaluation of various models and the proprietary Lakmoos model within the PerSimChat framework. The values indicate the ratio of wins, ties, and losses of the system using the first model compared to the second model. The evaluation was performed by a single GPT-4-turbo run over a subset of 30 tasks, with 10 messages exchanged per conversation. Three personas were used for each conversation.

Discussion

PerSimChat produces more natural and believable conversations than baseline systems, scoring higher in believability, credibility, depth, and flow. Its integration of cognitive modules, emotional dynamics, and real-world persona data makes interactions authentic and engaging. Minor improvements in memory and turn-taking remain, presenting opportunities to further strengthen the framework. Beyond research, PerSimChat has practical potential in market research, training, education, and collaborative decision-making through AI-driven simulations.

References

[1] 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.

[2] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926, 2023.

[3] 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.