COMPLEX NEGOTIATION AMONG AI LANGUAGE AGENTS IN WEREWOLF

By Collisteru

A thesis proposal submitted to the
Faculty of the Engineering School of the
[redacted] *in partial fulfillment*
of the requirement for the degree of
Bachelor of Science
Department of Computer Science
2024

[redacted]

[redacted]

Committee Members

[redacted]

1 Introduction and Background

The AI community has experienced a surge of interest in large language models (LLMs), models built from many layers of transformers that process natural language and respond with a prediction derived from a massive corpus of training data [2, 4, 5, 22, 28]. LLMs have astounded researchers and laymen alike with their long-spanning scaling effects as their capabilities have grown in tandem with their size and training data volume [13]. Recent state-of-the-art LLMs have demonstrated capabilities of abstraction, comprehension, memorization, and creativity [3, 16, 22], although they still have limitations [6, 32].

Traditional LLM implementations are limited by unimodality — their input and output channels process only text. This is a fundamental difference from natural intelligences (NIs) such as humans, which are embedded in the natural environment. NIs learn from interacting with — affecting and being affected by — the environment around them. Compared to natural senses such as vision, hearing, and touch, text is a narrow channel of information, especially when used alone [18]. The text-only training methodology of current leading LLMs contributes to their tendency to hallucinate [34].

To bridge this gap, researchers who seek to reproduce NI behavior and capabilities in AI are increasingly embedding LLMs in environments. This requires giving the LLM a "body" it can use to interact with the environment, along with the structural tools to use that body — a sense of physical agency. In order to create autonomous agents with a query-based LLMs as an engine, architectural additions in the form of modules have been devised. These often consist of a profiling module that grants the agent its foundational identity, a memory module that stores agent experiences, a planning module that allows the agent to form intentions, and an action module that carries them out. These modules are implemented in diverse ways that variously make use of querying the underlying LLM as well as employing algorithms that make use of information about the surrounding environment to determine the agent's actions [24,29]. In some cases, language agents have been equipped with self-adaptation abilities and can modify or improve themselves in response to experiences [20] The resulting agents are referred to as language agents. [8].

Guo et al. [8] identify two major categories of application for language agents: problem-solving and simulation, each with multiple subcategories. Language agents have already been deployed for uses as diverse as mental health support, studies of political science and economy, social simulation, documentation and data management, and embodied artificial intelligence [29]. In this paper We will focus on their applications to the social sciences. It has been demonstrated that language agents can interact with each other, put on personas, and hide their true intentions in large environments populated by other language agents [17]. Numerous other papers have explored the possibilities of multiple agent LLMs interacting in a social environment [10, 15]. However, there remains a research gap in understanding how humans interact with language agents when they are put together in a goal-directed social environment. This paper fills that gap.

2 Motivation

As LLMs grow in power and capability, they are increasingly becoming a part of our daily lives. Language agents will benefit from rapid improvements in underlying LLM technology. They will be increasingly applied to solve problems, provide companionship, and populate virtual spaces. Language agents are the most powerful generative AI agent yet invented, so studying them has profound implications for the study of AI agents in general. There is therefore an increasingly urgent need, underaddressed in the literature, to study how language agents interact with humans and each other. We propose a limited cooperation game in which a small number of LLMs interact with humans and each other and in which each agent (both human and non-human) is given a series of goals that they must fulfill by carrying out complex strategies involving both cooperation and competition with other agents.

2.1 An Introduction to Werewolf

Werewolf, also known as *'Mafia'*, is a seven-player logic puzzle and one of the world's most popular party games . In a party setting, one person is designated the moderator and is tasked with enforcing the rules of the game. The moderator divides the players into two teams: the werewolves and the villagers. No player knows which of the teams any of the other players are on. The game cycles between two states: day and night. The werewolves' goal is to kill the citizens without being discovered, and the villagers' goal is to identify the werewolves and vote to kill them. At night, the moderator orders all players to close their eyes and drum the table to drone out noise. The werewolves then silently decide which player they want to kill and signal their choice to the moderator. The moderator then has everyone open their eyes and announces who was killed. The killed villager is removed from the game. [26, 31].

2.2 Discussion

Communication games such as Werewolf have been used as a proxy to study behaviors in economics and social science [26]. Although AIs have been applied to Werewolf for many years [9, 23], new developments in LLMs open up major opportunities for further study, particularly because they communicate using natural language [**?**]. When Xu et al developed a framework for language agents playing werewolf together, they found that each agent used a variety of trust, confrontation, camouflage, and leadership strategies. However, it is still unknown how humans will fare in a game of werewolf against language agents. Language agents have been shown to be better than humans at producing disinformation in a social media space [27], but how well can they deceive humans in a one-on-one confrontation? How does tweaking a language agent's initial persona affect how it plays? Do language agents learn from their experiences when playing iterated games? Can humans reliably guess which of their fellow players are humans and which are artificial agents? When applied to Werewolf, these questions yield data that purely strategic games like chess and go do not. Language agent behavior in a game of social deception can be treated as a bellwether to the efficacy of deliberate language agent deception. This pertains to the future of human-AI cooperation, especially online, which is likely to become a part of daily life for knowledge workers. The results have applications in every domain where AI agents will be applied, including simulations of work conditions, simulations of historical, hypothetical, and actual social circumstances, training modules, and video games. As such, the results of this study will help us better predict and understand the future relationship between humans and AI agents in the digital world and beyond.

3 Research Questions and Hypotheses

We will develop a simple simulation of the Werewolf game in which language agents and humans can play together. Humans and agents will be able to communicate with each other in natural language via a shared text chat, and werewolves will be able to eliminate villagers in the nighttime. We will have the agents play against each other in at least a few dozen games and record the results.

Then, we will recruit a number (n $\approx$ 20) of human test subjects to play with and against the language agents in a wide range of settings. We will conduct experiments to answer the following research questions:

- **RQ1: How does the initialization of the identity module for a language agent affect its behavior and performance in Werewolf?**
- **RQ2: Do agents improve after playing iterated games?**

Given time, we would also like to answer the following as a ”stretch goal”:

- **RQ3: Can humans distinguish humans and language agents from each other purely from their actions while playing Werewolf?**

3.1 Research Question 1: Language Agent Personalities

Language agents are often initialized with a profiling module that gives them an initial tendency to certain behaviors over others [24,29]. We will develop a number of personas along archetypes measured by the Big Five personality traits in psychology [7] and do a qualitative analysis of how these changes of personality lead to differences in utterances and playstyle. For example: does decreasing the disagreeability of the identity module result in a greater negative affect in the agent’s statements?

3.2 Research Question 2: Language Agent Learning

The fundamental limitation of classically-implemented LLMs is that they are not stateful: they are not trained while speaking with their interlocutor. One of the fundamental innovations of language agents is to rectify this by giving agents a planning and a reflection module that allows them to record their experiences, distill them into more abstract long-term memory and general principles as humans do, and keep them in their context window to take them into account for future actions. This imitates a form of experience acquisition and even learning that mimics more closely not the neural learning by which the LLMs were trained but rather the "fact and experience-based" learning that drive semantic learning and expert systems [1] [30] [21]. We want to test the persistence of this memory and how well it helps agents to learn across games. When playing multiple iterations of the same game, do language agents learn from their experiences? Do they get better at the game over time? Can they perform deductive reasoning by drawing novel conclusions about the general game by drawing from specific experiences? Do they specialize in one type of role (werewolf, seer, villager, etc)?

3.3 Research Question 3: Turing-Type Test

This research question is a "stretch goal" we will pursue given extra time. We will have many humans play against the language agents in werewolf. While the proportions of humans and werewolves will vary, We intend to have about two humans per play session and five werewolves. The simulation will record the actions of each of the players. Later on, we will ask human evaluators to identify which players were the AI agents on the basis of game actions (werewolf killings and villager accusations) alone.

4 Method of Approach

Our aim is to further develop the language agent architectures created by [24] and [33] to study these questions. We will embed these language agents in a social simulation of werewolf in which humans can also interact. We would like to apply the current frameworks of language agents to open-source LLMs in order to make this technology more accessible. However, this poses the risk of significantly reducing agent abilities and does not represent the cutting edge in LLM technology. We will integrate improvements from [12] and [33] to better specialize the agents to be able to play werewolf.

4.1 A Game of Humans and Agents

The simulation will include a game setup mode in which the user can choose how many language agents and how many humans will be present in the simulation. For the language agents, options will be available to change the identity module of the language agent, as well as to use a language agent that was already used in a previous

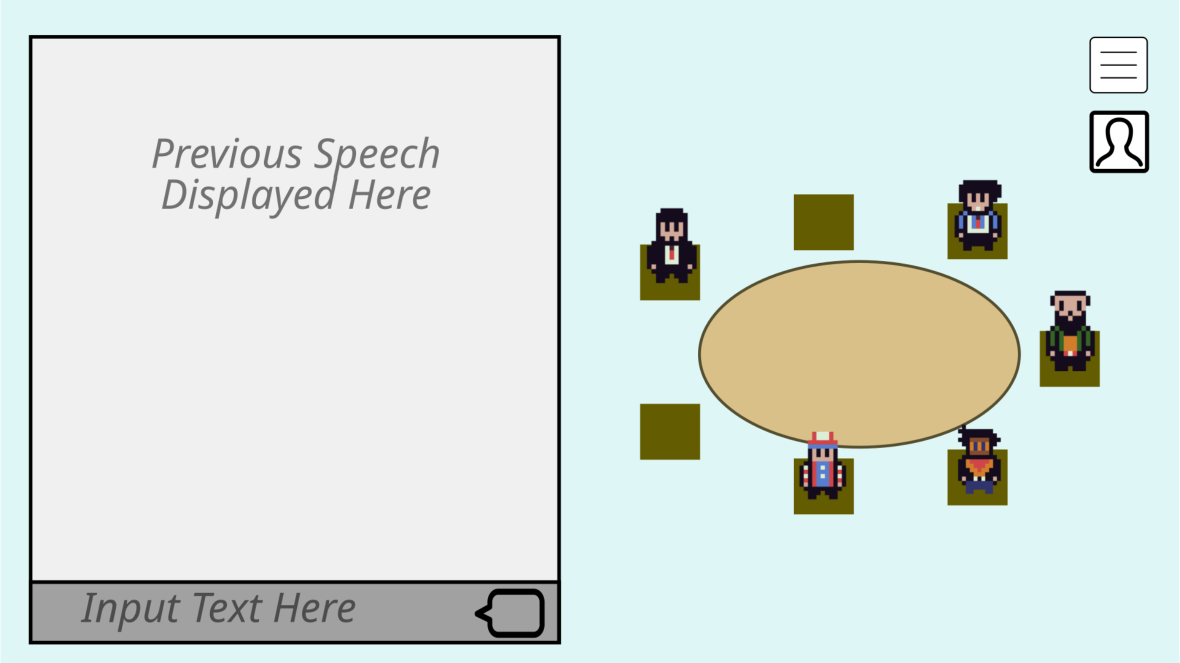

Figure 1: *A mock-up of the main simulation screen. In the top right there are two icons pertaining to game settings. The hamburger menu at the very top-right leads to game settings. The profile icon directly below it allows the user to choose and customize agents. Two players have already been eliminated.*

game (along with its memory and reflection modules) . The game setup software will also grant the ability for the user to determine which language agents and humans are in which roles. We will program two game modes: automatic and real-time. Automatic mode is the simplest and involves language agents playing with each other. The game will legislate the rules itself and the language agents will be programmed to make the necessary LLM calls in order to make the decisions they are called to make at each point in the game. Real-time mode allows human players to play with language agents. The language agent's speech and decisions are communicated to all players in an in-game chat. There will be a button to advance the round, as well as a way to choose players as is required for the roles of werewolf, seer, and doctor. The screen the user sees will include a prominent shared chat window where the user can see a record of everything all the players have said. They'll also be able to type into the chat to speak to an LLM. Just like in a real-life game of werewolf, everyone will be able to hear what everyone else has said.

4.2 Efficiency

The efficiency of the language agent framework is the principal limitation left by Park et al. There are a number of extraneous external LLM calls that lead to a higher expense than is strictly necessary to create believable behavior. For example, in Park's simulation, agents frequently make LLM calls to plan actions for which they've already planned in the past. Memoization of action planning will improve efficiency. We may define an action in Werewolf as either speaking or making a choice in the game. We will count efficiency as the average number of LLM calls per action. We hope to decrease this number because this will decrease the cost of running the language agents.

4.3 Data Collection and Analysis

We will collect data to answer each of the three research questions. The goal is to make data collection and question answering as quantitative and objective as possible.

4.3.1 Research Question 1 – Language Agent Personalities

Park et al initialized each language agent with a short paragraph to explain their identity and background [24]. We want to find out if changes in this identity module correlate with behavioral changes. If they are correlated, that would make identity modules a way to program personality in language agents. We will organize personalities with the Myers-Briggs type Indicator [14, 25] into sixteen categories. Each category takes one of two positions along four axes. For each axis, we will write a short sentence or two describing it in the context of the language agent. An example follows for ENTJ:

> *You are Marcus. You are playing the game Werewolf.*
> *You prefer to direct energy mainly outward, towards other people and objects, rather than the world of inner experience. (Extroverted)*
> *You rely more on abstract patterns and interrelationships than on the five senses. (iNtuiting)*
> *You prefer to base conclusions in logical analysis rather than gut feeling. Objectivity is important to you. (Sensing)*
> *When dealing with others and the outside world, you prefer to focus on logic and social values over what can be perceived. (Judging)*

Figure 2: An example of an identity module for a language agent

The descriptions in the identity module are based on the MBTI Manual [19].

The goal is to see how the initialization of the identity module with different personalities changes strategy and behavior. In-game behavior will be measured with a number of constants:

- The mean μ and variance σ_s^2 of the distribution of sentiment of the agent's utterances
- The number of times the agent accused another agent of being a werewolf n_a
- Game performance (proportion of games in which the team the agent was on won) p

Each personality will be tested multiple times in the same player role, and the value of these constant games will be tested against each other for behavioral consistency and deviation from a control group without a customized identity module. A χ^2 test will be performed on $\{\mu, \sigma_s^2, n_a, p\}$ to determine the presence of significant deviance.

All personalities with significant results will be further tested to find the direction of variable changes, and these directions will be reported.

4.3.2 Research Question 2: Language Agent Learning

Language agents typically come with a memory module and a reflection module that build on redundant memories in the memory module to create generalizations and more profound knowledge. This transforms stateless bare LLMs into stateful language agents. We want to know if statefulness improves agent performance over time. Do language agents learn from experience?

We will have the same group of language agents play n separate games. These “old” agents will retain their memory and reflection streams from previous games. In the n+1st game, we will replace half of the old agents with “young” agents with blank memory modules. The old and the new agents will then play m further games. During these m games, we will test the proportion of victories to see if the proportion of old to new agents in a Werewolf team improves win likelihood. If so, this would provide evidence that LLMs can learn from their mistakes and improve over time. Qualitative observations will be noted as to the nature of the improvements.

4.3.3 Research Question 3: Turing-Type Test

Given time, we want to understand if humans can distinguish human players of Werewolf from language agent players. We will have language agents play against humans in multiple games. We will then show the record of the games with the transcript removed to find out if humans can reliably distinguish humans from language agents on the basis of their behavior.

Given prior results of similar turing-type tests showing that well-prompted language agents can often successfully imitate humans [Jones and Bergen 2024], we hypothesize that humans won’t be able to distinguish human werewolf-players from agent werewolf-players better than chance.

4.4 Open-Source LLMs Experimentation

Given time, we are interested in implementing the architecture of language agents with the open-source model Mistral-8B-Instruct, and will record the results and impacts on believability. This change would make the model much cheaper to implement and would greatly help to make it more accessible to the general public.

5 Work Plan

Period of Work: November 2024 — May 2025

Time Period	Actions
November 2024	• Defend thesis proposal • Develop werewolf simulation
December 2024	• Defend werewolf simulation • Program language agents
January 2025	• Program language agents • Improve agent efficiency
February 2025	• Create logic for human player • Create chat interface between player and non-player character • Collect data for RQ1 and RQ2
March 2025	• Code logic for human player player tests for RQ3
April 2025	• Analyze data • Write final results
May 2025	• Thesis defense • Thesis publication

6 Budget

Line Item	**Funds Required**
LLM Calls	$500
Human Research Subject Honoraria	$200
Total	$700

References

[1] R. C. Atkinson and R. M. Shiffrin. 1968. Human Memory: A Proposed System and its Control Processes. Stanford University, Stanford CA.

[2] Tom Brown, Benjamin Mann, et al. Language models are Few-Shot Learners. 2020. arXiv:2005.14165. Retrieved from https://arxiv.org/abs/2005.1416

[3] Sébastien Bubeck, Varun Chandrasekaran. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 . https://arxiv.org/abs/2303.12712

[4] Mark Chen, Jerry Tworek, et al. Evaluating large language models trained on code. 2021. arXiv:2107.03374. Retrieved from https://arxiv.org/abs/2107.03374

[5] Aakanksha Chowdhery, Sharan Narang, et al. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.0231. Retrieved from https://arxiv.org/abs/2204.02311

[6] Matthew Dahl, Varun Magesh, et. al. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. arXiv:2401.01301. https://arxiv.org/abs/2401.01301.

[7] Lewis R. Goldberg. 1993. The Structure of Phenotypic Personality Traits. American Psychologist 48, 1 (Jan 1993), 26-34.

[8] Taicheng Guo, Xiuying Chen, et. al. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. arXiv:2402.01680. https://arxiv.org/abs/2402.01680.

[9] Yuya Hirata, Michimasa Inaba, et. al. 2016. Werewolf game modeling using action probabilities based on play log analysis. Computers and Games 9th Conf. CG 2016 Revised Selected Papers 103-114. https://doi.org/10.1007/978-3-319-50935-8_10

[10] Ji Shi Jinxin, Zhao Jiabao, et. al. 2023. CGMI: Configurable General Multi-Agent Interaction Framework. arXiv:2308.12503. https://arxiv.org/abs/2308.12503.

[11] Cameron R. Jones, Benjamin K. Bergen. 2023. Does GPT-4 pass the Turing test? arXiv:2310.20216. https://arxiv.org/abs/2310.20216

[12] Zhao Kaiya, Michelangelo Naim, et. al. 2023. Lyfe Agents: Generative agents for low-cost real-time social interactions. arXiv:2310.02172. https://arxiv.org/abs/2310.02172

[13] Jared Kaplan, Sam McCandlish, et. al. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361. https://arxiv.org/abs/2001.08361

[14] Ki Scott P. King and Brittany A. Mason. 2020. "Myers-Briggs Type Indicator." The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment. https://doi.org/10.1002/9781119547167.ch123

[15] Gigur Kovač, Rémy Portelas, et. al. 2023. The SocialAI School: Insights from Developmental Psychology Towards Artificial Socio-Cultural Agents. arXiv:2307.07871. https://arxiv.org/abs/2307.07871.

[16] Ehsan Latif, Yifan Zhou. 2024. A Systematic Assessment of OpenAI o1-Preview for Higher Order Thinking in Education. arXiv:2410.21287. https://arxiv.org/abs/2410.21287.

[17] Siyu Li, Jin Yang. 2023. Are you in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks. arXiv:2307.10337. https://arxiv.org/abs/2307.10337.

[18] Richard E. Mayer. 2020. Multimedia Learning (3rd. ed.) Chapter 1. Cambridge Univ. Press, Cambridge, UK.

[19] Isabel Briggs Myers, Mary H. McCaulley, et. al. 1998. MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator. (3rd ed.) Consulting Psychologists Press, Palo Alto, CA.

[20] Nathalia Nascimento, Paulo Alencar, et. al. 2023. Self-Adaptive Large Language Model (LLM)-Based Multiagent Systems. arXiv:2307.06187. https://arxiv.org/abs/2307.06187.

[21] Andrew M. Nuxoll and John E. Laird. 2011. Enhancing intelligent agents with episodic memory. Cognitive Systems Research 17, 18 (2012) 34-38. doi:10.1016/j.cogsys.2011.10.002

[22] OpenAI, et al. 2023. GPT-4 Technical Report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[23] Hirotaka Osawa, Takashi Otsuki, et. al. 2021. Negotiation in Hidden Identity: Designing Protocol for Werewolf Game. Studies in Computational Intelligence 958 87-102. https://doi.org/10.1007/978-981-16-0471-3_6.

[24] Joon Sung Park, Joseph C. O'Brien. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442. https://arxiv.org/abs/2304.03442.

[25] Ken Randall, Mary Isaacson, et. al. 2017. Validity and Reliability of the Myers-Briggs Personality Type Indicator: A Systematic Review and Meta-analysis. Journal of Best Practices in Health Professions Diversity 10, 1 (Spring 2017) pp. 1-27. https://www.jstor.org/stable/26554264

[26] Hong Ri, Xiaohan Kang, et. al. 2022.The Dynamics of Minority versus Majority Behaviors: A Case Study of the Mafia Game. Information 13, 3 (Mar. 2022), 134. https://doi.org/10.3390/info13030134

[27] Giovanni Spitale, Nikola Biller-Andorno et. al. 2023. AI model GPT-3 (dis)informs us better than humans. arXiv:2301.11924. https://arxiv.org/abs/2301.11924.

[28] Hugo Touvron, Louis Martin, et. al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288. https://arxiv.org/abs/2307.09288

[29] Lei Wang, Chen Ma, et. al. 2024. A survey on large language model based autonomous agents. arXiv:2308.11432. https://arxiv.org/abs/2308.11432.

[30] Yongjia Wang and John Laird. 2007. Integrating Semantic Memory into Cognitive Architecture. Center for Cognitive Architectures University of Michigan Technical Report CCA-TR-2006-02. University of Michigan, Anna Arbor, Michigan.

[31] WikiHow. 2024. How to Play the Werewolf Card Game with Your Friends. Retrieved from https://www.wikihow.com/Play-Werewolf-(Party-Game)

[32] Kevin Wu, Eric Wu, et. al. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. arXiv:2402.02008. https://arxiv.org/abs/2402.02008.

[33] Yuzhang Xu, Shou Wang, et. al. 2023. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. arXiv:2309.04658. https://arxiv.org/abs/2309.04658.

[34] Hongbin Ye, Tong Lui, et. al. 2023. Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv:2309.06794. Retrieved from https://arxiv.org/pdf/2309.06794.