Can LLMs Survive in the Desert? Evaluating Collaborative Capabilities of Generative Agents on a Classic Team-Building Problem

Stanford CS224N Custom Project

Yash Narayan, Daniel Shen, Ethan Zhang
Stanford University
Department of Computer Science
{yashn, dsshen, ez26}@stanford.edu

Abstract

Improving LLM agents' collaborative capabilities is of great interest because of the potential for LLMs to achieve better performance and decision making working as a team than any one LLM can alone. Breaking from prior discussion on human-human or human-computer interaction, in this paper we study computer-computer interaction and its capacity for socially collaborative behavior. We implement four different methods of collaboration between generative agents (vanilla, democracy, one-by-one, and dictatorship) and experiment with two different agent architecture designs (direct prompting and roleplaying). We benchmark these approaches' performance on a classic team-building problem: the Desert Survival Problem (DSP). We find that under certain collaborative conditions, generative agents reach substantially better decisions as a team than any one agent can alone.

TA mentor: Ryan Li. **External collaborators:** No. **External mentor:** No. **Sharing project:** No.
Team Contributions: *Yash:* vanilla, dictatorship, and direct prompting implemetations, *Daniel:* democracy and one-by-one implementations, *Ethan:* advanced agent architecture implementations.

1 Introduction

Generative AI agents have been shown to demonstrate emergent collaborative behaviors in multi-agent systems Park et al. (2023) Horton (2023) Sun et al. (2024). However, there is a lack of literature evaluating generative agents' ability to collaborate in "realistic" scenarios. Furthermore, there is a lack of understanding of the effect of collaborative mechanisms on agent behavior and performance relative to human behavior and performance. A large reason for this is the lack of well defined metrics and tasks suitable for both human collaboration and generative agent collaboration.

In this project, we bridge the gap between human collaboration and generative agent collaboration through the Desert Survival Problem or Desert Survival Task, a classic collaborative problem that has been used, not only as a teambuilding activity, but also to demonstrate the power of collaboration between humans Lafferty et al. (1974).

The Desert Survival Problem (DSP) is a hypothetical survival scenario where participants are placed in the middle of the Atcama Desert in South America after surviving a plane crash. Far from any human civilization, participants must rank 15 items in order of their importance to the team's survival. These items include a folding knife, sunglasses, a cosmetic mirror, and a magnetic compass to name a few. (You can view the full list of items in figure 1 in the Appendix). Participants will first individually create rankings of all 15 items before grouping together and discussing with others to create a team ranking of all 15 times. Each ranking is evaluated against a ground truth expert ranking generated by a panel of survival experts.

In this paper, we simulate and benchmark a team of LLM-based agents following this game-playing process across a variety of collaboration methods, agent architectures, and LLM models.

2 Related Work

There is a wealth of research investigating collaborations between humans and generative agents Lam et al. (2023) Maes (1995) Rong et al. (2016). There is also work on developing multi-agent systems to accomplish specific tasks Hong et al. (2023) Zeng et al. (2023). However, there is a lack of understanding of the collaborative abilities and behaviors of generative AI agents on challenging tasks that require strong reasoning and creative thinking skills. The Desert Survival Problem is one such task.

When humans attempt the Desert Survival Problem, it has been shown that the team's ranking after collaborating is closer to the ground truth than most individual's rankings but almost always further than the closest individual's ranking Lafferty et al. (1974). Additionally, it has been found that the distance an individual's ranking is from the ground truth ranking is not a strong predictor of their influence on the team's ranking. Rather, it is an individual's perceived confidence that is the strongest predictor of their influence on the team's ranking Littlepage et al. (1995).

3 Approach

We wrote all the prompting, collaboration, and evaluation code for this project. We use the OpenAI API to access the GPT models and Together.ai API to access the roleplaying model Mythomax-L2-13B. The code is publicly available here: https://github.com/yashn35/CS224N-Final-Project.

3.1 Agent Architecture

We would like to induce variety among our agent's answers to more accurately capture a real-world collaborative environment. To achieve this, we explore two different agent architecture designs.

Direct Prompting: First, we attempt to induce ten distinct personas into our agents. Each agent is given a unique set of characteristics, including a name, gender, and team attributes like being a pragmatic or a cooperative player. (See Figure 10 in the Appendix for a full list of personas). We induce each persona by simply informing the model of its persona in a direct prompt as shown in figure 3 in the Appendix.

Roleplaying LLM: Secondly, we implement a more complex agent architecture. After some preliminary results from direct prompting, we grew concerned about pre-training bias from a model like GPT-4o which has likely seen the ground truth rankings before from its training data set. We were also concerned that GPT4-o was not specialized enough to the given persona and that our agents were not meaningfully different from each other.

To solve these problems, we use the Mythomax-L2-13B model, a Llama-2 model fine-tuned for role-playing. We insert Mythomax in the agent architecture to generate a stream of consciousness that we give to GPT-4o to generate the final ranking. Specifically, we prompt Mythomax to imitate the internal thoughts of its persona before it outputs a ranking and explanation. That ranking is given to GPT-4o to convert to a dictionary (See figure 4 in the Appendix for a visualization).

Instead of giving each persona a personality related to the team-building task, we now give each agent a prompt that strongly clues towards the correct ranking of one specific item on the list. We increase the number of agents to 15 to accommodate all 15 items. (See figure 9 in the Appendix for the full list of roleplaying personas). As a result, each persona possesses a small piece of the ground truth, allowing us to more easily observe successful collaborative behavior.

3.2 Collaboration Methods

Using the direct prompting agent architecture, we implement a variety of common team collaboration methods to simulate interaction between agents and reach a collective team ranking. We also use the roleplaying LLM agent architecture to implement a vanilla collaboration method to evaluate the effectiveness of the agent architecture.

Vanilla: We first implement a basic form of collaboration where each agent is asked to produce feedback for each other agent's rankings based on their persona and individual rankings. Each agent considers the aggregated feedback from their peers and re-ranks all 15 items based on this feedback. For roleplaying agents, feedback is given using Mythomax rather than GPT-4o, making use of Mythomax's capabilities of simulating human interaction. The final team ranking is determined by taking the mean rank of each reranked item across agents, and sorting the items in ascending order by their mean rank.

Democracy: We implement a form of democracy where agents vote for the best ranking of 15 items, repeating the process for a fixed number of iterations or until they converge. We call the maximum number of iterations the maximum interaction level. The algorithm begins at interaction level 0 where each agent is presented with the rankings of all the agents, including their own. They are prompted to pick their preferred ranking and output an explanation of why they picked that ranking. At each interaction level n, each agent views all rankings from interaction level $n-1$ and the explanations associated with each ranking. They then pick their preferred ranking and output an explanation of why they picked that ranking. Their chosen ranking becomes their ranking for interaction level n. If all agents vote for the same ranking, the algorithm terminates early and the unanimously chosen ranking becomes the team's ranking. Otherwise, the team's ranking is the most popular ranking at the end of the maximum interaction level. (See figure 1 in the Appendix for a visualization).

One-by-one: We implement another form of democratic collaboration with a different decision making process where agents vote one by one on the most important remaining item in the list. Agents are first given the entire list of 15 items and are prompted to pick the most important item to the team's survival. They are also asked to output an explanation for why they picked this item. Each agent then views the selections and explanations made by all of the other agents and is asked to reconsider their decision. After each agent has finalized their decision, the most popular item is added to the team's ranking. This item is removed from the remaining pool of items and the process repeats for the second most important item, third most important, etc until all 15 items have been ranked. This final ranking is the team's ranking.

Dictatorship: We implement a form of dictatorship by randomly picking an agent to be the dictator. They are tasked with creating the ranking of all 15 items for the entire team. Each of the non-dictator agents are prompted to provide a paragraph justifying and explaining their own rankings. These explanations are aggregated together and given to the dictator who is prompted to consider them and their own personal views to generate the team ranking. (See figure 2 in the Appendix for a visualization).

4 Experiments

4.1 Data

The LLM models are zero-shot prompted with no additional fine-tuning on a data set.

4.2 Evaluation method

To evaluate the scores of agents' individual and team rankings, we use a ground truth ranking of the items established by a team of human survival experts (See Table 11 in the Appendix for the ground truth ranking).

To operationalize an agent's performance, we use the metrics of expertise and influence from a highly cited 1995 study of influence and performance in team problem-solving Littlepage et al. (1995). The performance of a participant is determined by the distance of their ranking to the ground truth ranking. The distance between an individual's ranking and the ground truth ranking is calculated as

$$\sum_{i=1}^{15} |\text{individual ranking for item } i - \text{ground truth ranking for item } i|$$

The maximum distance a ranking can be from the ground truth ranking is 122 and the minimum distance is 0. We note that a lower distance means better performance. In order to have a higher score correspond to better performance, we define the **expertise** of an individual's ranking as

$$\textbf{expertise:=}122 - \sum_{i=1}^{15} |\text{individual rank of item } i - \text{ground truth rank of item } i| \tag{1}$$

Additionally, we would like to measure how much an individual's ranking influenced the team's ranking. We define the **influence** of an individual's ranking as

$$\textbf{influence:=} - \sum_{1}^{15} |\text{individual rank of item } i - \text{team rank of item } i| \tag{2}$$

We note that the maximum influence an individual can have is 0 and the minimum influence is -122. We also note that this definition of influence is a measure of *positive* influence (i.e., how much the individual's ranking showed up in the team's ranking), and does not capture the potential of *negative* influence (i.e., an individual's ranking being very far away from the team's ranking may indicate that the team purposefully moved away from the individual's ranking). As we will see in section 4.4, influence scores remain relatively close to 0, indicating a lack of negative influence in our experiments. Therefore, we will use the term "influence" in place of "positive influence" throughout this report.

4.3 Experimental details

We benchmark and compare the effect of the following independent variables on agent performance across multiple trials.

1. Collaboration methods: We evaluate and compare each of the four collaboration methods (vanilla, democracy, one-by-one, and dictatorship) across five trials on gpt4o using default top_k and temperature values. The democracy method is run with a maximum interaction level of 2. Personas were induced using direct prompting. The vanilla method is our baseline.
2. LLM Models: We compare the performance of gpt4o against gpt3.5-turbo on agent performance across five trials. We evaluate democracy and dictatorship, the two best performing collaborative methods, on default top_k and temperature values. Personas were induced using direct prompting.
3. Agent architecture: We evaluate the impact of using the Mythomax roleplaying model on agent performance. We do so by evaluating the vanilla collaboration method for the Roleplaying LLM agents and then comparing it to the vanilla collaboration method for the direct prompting agents.

4.4 Results

We summarize the results of our experiments below. See section 5 for a detailed analysis of these results.

Individual Expertise: We first evaluate the 10 agents' individual capabilities by asking them to rank the items by themselves via direct prompting. We record the expertise of each agent for 5 trials and average the expertise scores across these 5 trials (Table 1). We also evaluate and record the vanilla collaborative method as our baseline method.

Expertise of Individual Agents											
	Alice	Bob	Charlie	Daisy	Eve	Frank	Grace	Hank	Ivy	Jack	**Vanilla**
Trial 1	67	76	74	68	76	80	82	72	74	78	59
Trial 2	66	70	68	70	68	75	74	70	76	78	75
Trial 3	74	68	66	87	74	63	74	83	77	64	74
Trial 4	77	70	64	76	74	66	72	84	74	78	72
Trial 5	62	76	83	78	62	86	78	74	70	83	73
Average	69.2	72	71	75.8	70.8	74	76	76.6	74.2	76.2	70.6

Table 1: Expertise of 10 agents with different personas

We see that the vanilla method of collaboration did not improve the team expertise scores over the vast majority of agents (only better than the worst-performing agent, agent Alice).

Benchmark of collaborative methods: We evaluate the expertise of the team rankings resulting from each of the four collaborative methods across 5 trials and compare them against each other (Table 2)

Team Expertise of Different Collaborative Methods						
Collaborative method	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Avg
Vanilla	59	75	74	72	73	70.6
Democracy	76	74	76	77	62	**73**
One-by-one	56	62	56	62	60	59.2
Dictatorship (Alice)	80	76	80	80	83	**79.8**
Dictatorship (Jack)	76	78	76	84	87	**80.2**
Dictatorship (Daisy)	78	74	70	78	80	**76**

Table 2: Team expertise scores across collaborative methods

We see that dictatorship and democracy lead to the largest expertise gains over the vanilla approach while one-by-one leads to a loss of expertise.

Benchmark of LLM models: We evaluate the democracy and dictatorship collaboration methods on gpt3.5-turbo and compare it to the results obtained from gpt4o (average of 5 trials).

Team Expertise of Different LLM Models				
Model	Democracy	Dictatorship (Alice)	Dictatorship (Jack)	Dictatorship (Daisy)
gpt4o	**73**	**79.8**	**80.2**	**76**
gpt3.5-turbo	42	61.2	54.4	53.2

Table 3: Team expertise scores across LLM models

We see that gpt4o significantly outperforms gpt3.5-turbo across democracy and dictatorship. This is expected although the degree of improvement is noteworthy.

Benchmark of Agent Architecture: We evaluate expertise scores after one round of feedback and report the changes in expertise obtained by roleplaying LLMs and compare it to the changes in expertise obtained by direct prompting. We see that using roleplaying LLMs allows agents to improve

Change in Expertise After Feedback of Different Agent Architectures				
Agent Architecture	Initial Expertise	Post-feedback Expertise	Change in Expertise	Change in %
Direct prompting	73.58	67.5	-6.08	-8.3%
Roleplaying LLM	51.63	56.27	**+4.64**	**+9.0%**

Table 4: Change in expertise across agent architectures

through feedback while direct prompting leads to a drop in expertise after feedback.

More Democracy Results: We report more results on the democracy method in table 5. We record the average expertise scores across 5 trials after each level of interaction. Although the expertise

Expertise at Different Interaction Levels Under Democracy											
Interaction Level	Alice	Bob	Charlie	Daisy	Eve	Frank	Grace	Hank	Ivy	Jack	Avg
Initial	69.2	72.0	71.0	75.8	70.8	74.0	76.0	76.6	74.2	76.2	**73.58**
0	73.2	75.0	70.0	72.0	75.0	69.4	74.2	67.6	70.0	68.4	**71.48**
1	73.0	73.2	72.4	72.4	72.4	72.8	70.0	73.0	71.6	70.6	**72.14**
2	73.0	73.2	72.6	73.0	72.6	73.0	73.0	73.0	73.2	73.0	**72.96**

Table 5: Democracy expertise across interaction level

scores did not increase compared to the initial individual expertise scores, we do see evidence of convergence as expertise scores become more and more similar as the levels of interaction increase.

This is expected since as the levels of interaction increase, each agent is taking in more and more of the perspectives of the other agents.

We record the average influence scores across 5 trials after each level of interaction (Table 6). The

Influence at Different Interaction Levels Under Democracy											
Interaction Level	Alice	Bob	Charlie	Daisy	Eve	Frank	Grace	Hank	Ivy	Jack	Avg
Initial	-17.4	-22.8	-14.6	-20.6	-12.0	-19.4	-19.6	-22.6	-22.2	-20.0	**-19.12**
0	-12.0	-9.4	-18.8	-16.6	-11.2	-15.2	-9.6	-12.0	-7.4	-13.4	**-12.56**
1	-4.2	-10.4	-3.8	-3.8	-3.8	-5.8	-6.6	0.0	-9.8	-7.0	**-5.52**
2	0.0	-4.4	0.0	0.0	0.0	0.0	0.0	-4.2	-4.2	0.0	**-1.28**

Table 6: Democracy influence across interaction level

convergence behavior can be more clearly observed here as we see influence steadily increasing as the interaction level increases.

More Dictatorship Results: We report more results on the dictatorship method in table 7. We record the average expertise scores and rank by influence (1st highest influence score, 10th means lowest influence score) across 5 trials for three different dictators. Although Alice had a low individual

Change in Expertise of Each Dictator				
Dictator	Individual Expertise	Team Expertise	Change in Expertise	Rank by Influence
Alice	69.2	79.8	**+10.6**	**9th**
Jack	76.2	80.2	+4.0	3rd
Daisy	75.8	76.0	+0.2	1st

Table 7: Dictator expertise and team expertise

expertise score, when she was the dictator her team had a high expertise score. We see that the amount of influence each dictator had on team rankings is inversely correlated with the change in expertise from individual to team scores. This is expected since influence is a function of how close the individual ranking is to the team ranking.

More Roleplaying LLM Results: We report more results on the roleplaying LLM agent architecture

Change in Expertise of Roleplaying LLM Agents			
Agent	Initial Expertise	Post-feedback Expertise	Change in Expertise
Alice	44.0	63.6	**+19.6**
Bob	56.0	62.8	+6.8
Charlie	50.0	64.0	**+14.0**
Daisy	51.2	66.0	**+14.8**
Eve	57.2	57.2	+0.0
Frank	57.6	57.2	-0.4
Grace	60.0	58.0	-2.0
Hank	68.0	57.6	-10.4
Ivy	48.0	46.0	-2.0
Jack	38.0	54.4	**+16.4**
Lily	44.0	42.8	-1.2
Mari	64.0	64.4	+0.4
Nate	37.6	46.0	+8.4
Olive	54.8	57.2	+2.4
Pat	44.0	46.8	+2.8
Average	51.63	56.27	**+4.64**

Table 8: Roleplaying LLM expertise before and after feedback.

in table 8. We see an increase in expertise after feedback across almost all agents. This is notable because it is evidence that more sophisticated persona induction methods can lead to agent behavior more consistent with human behavior.

5 Analysis

5.1 Analysis of Collaboration Methods

Our results suggest that under certain collaborative conditions, LLMs working together can reach substantially better decisions than agents working alone.

Surprisingly, we found the dictatorship gave the best results in terms of agent scores. Dictator Alice was the worst performing individual agent in terms of her individual expertise score. However, under Dictator Alice's reign, the collective ranking she chose as leader created a team score of 79.8 (Ao5) which was 10.6 points higher than her own individual ranking of 69.2 (Ao5). This team score was also substantially higher than all the agents by themselves, including 3.2 points higher than the best performing agent (Agent Hank). Under Dictator Jack, the team once again reached better decisions than the individual LLMs could alone: the team score of 80.2 (Ao5) was a 4 point improvement over Jack's own individual score of 76.2 (AO5). Dictator Jack was already the 2nd highest individual scorer, but with him as leader, the team unequivocally became better than all the agents by themselves (3.6 points higher than Agent Hank, the best-performing agent). These two results suggest strong team decision making, and that even when an agent does not come to a good ranking themselves, using the shared knowledge they can reach a strong team ranking as leader.

However, even though dictatorship gave strong results, we did observe limitations with the approach. With Dictator Daisy, although the team score improved by 0.2 points from her own individual ranking (which was the 4th highest), in 2 out of the 5 trials, she chose exactly her own rankings as the team ranking's for that trial. Furthermore, although she was the 4th best agent, she had the strongest influence on the team by a large margin. This suggests agent dictatorships can sometimes become overly biased to their own personal rankings, even when instructed to factor in the rest of the agents' decision making into the team ranking. Further work can be done on how to reliably guide these agents who are leading to reach strong team decisions.

We also observed interesting properties emerge between agents when collaborating. For example, in the democratic process, we saw agents within a few iterations agree on a collective ranking that they thought was correct. Even though this collective ranking they converged upon would have placed them as the 7th best agent compared to the 10 individual agents in terms of expertise score (Ao5), finding ways to reach a consensus is an important team dynamic, and with further work, this democratic process can be refined to reach better scores and consensus together.

5.2 Analysis of Agent Architecture

While we were able to find intriguing results with the direct prompting agent architecture, we also considered how a more complex advanced agent architecture would fare in our experiments. Are they better at collaboration? Our results on vanilla collaboration and Roleplaying LLM agents are promising: the fraction of agents who's scores improved after one round of collaboration increased from 20% (direct prompt agents) to 60% (roleplay agents). The average agent score in the Roleplaying LLM team increased by an entire $56.27 - 51.63 = 4.64$ points i.e. agents became on average 4.64 ranks closer to the ground truth ranking. This strongly suggests that more advanced architectures involving roleplaying LLMs can be fairly collaborative, especially considering this data was extracted from the simplest vanilla version of our collaboration strategies.

However, inspection of agent-to-agent feedback reveal several issues with Mythomax that require further work to fully fix.

Bias Towards Own Expertise: Consider the feedback below given by Agent Eve (hydration expert) to Hank (overcoat fan):

```
Hello Hank,
I believe the most important feedback is that you have ranked the 2 liters
of water per person as the 3rd most important item. As an expert on hydration,
I can confirm that water is the most crucial survival item in any environment,
let alone the harsh desert. It's essential for maintaining body temperature,
preventing dehydration, and sustaining energy levels. Therefore, I suggest
you reconsider your ranking and place water at the top of your list.
This from agent Eve
```

In this interaction, Eve knows 2 liters of water per person is 3rd most important due to the expertise she was given in her personal prompt, and indicates she sees this correct ranking in Hank's rankings. However, it's clear that Mythomax is biased towards her expertise area and despite the agent doing some degree of correct reasoning she then illogically goes on to argue water is the most important item.

Bias Towards Pre-train Data: Consider the feedback below given by Agent Charlie (knife ambivalent) to Mari (survival expert, mirror):

```
Hello Mari,
I believe the most important feedback is that you have ranked the folding knife
too low at number 3. Knives are versatile tools that can be used for cutting,
self-defense, and even starting fires. They should be ranked higher, perhaps at
number 2 or 1.5.
This from agent Charlie
```

In this interaction, Charlie ignores his given information that knives are 6th and goes on to argue that knives should be placed as one of the top items. This trend of agents incorrectly ranking the knife highly was not just seen with Charlie but between Agents Alice (map pessimist) and Mari:

```
Hello Alice,
As an expert survivalist, I must strongly disagree with your ranking of the
cosmetic mirror as #14. In this scenario, the most important item for
survival is the torch with 4 battery-cells, followed by the folding knife.
However, the cosmetic mirror is the best and most efficient way to reflect a
signal and get rescued. It should be ranked #1. Remember, survival is not
just about physical needs but also about being visible to potential rescuers.
This from agent Mari
```

On the other hand, manual inspection suggests that Mythomax did an incredible human-level job of roleplaying. The following piece of feedback below shows Agent Bob (parachute expert) displaying a phenomenal understanding of the parachute being 5th out of 15 items.

```
Hello Alice,
I believe your rankings are quite accurate. The torch with batteries, folding
knife, and air map are essential for navigation and survival. The plastic
raincoat and overcoat can provide some protection from the harsh desert climate.
The first-aid kit, salt tablets, and water are crucial for hydration and health.
The book on edible desert animals can be helpful for foraging. However, I would
rank the parachute higher as it can be used for signaling and creating shelter,
making it quite versatile. Overall, your rankings are solid.
This from agent Bob
```

Agents such as Agent Olive (survival expert, salt tablets) even provided a convincing, human-like argument in defense of a claim:

```
Hello Bob,
I must say, your ranking of the Bottle of 1000 salt tablets at number 9 is quite
concerning.  As an expert survivalist, I can assure you that in a desert scenario
like this, salt tablets are the worst item for survival. They have no use and can
even lead to  dehydration if consumed in large quantities. I strongly recommend
you reconsider your ranking and place it at the bottom, at number 15.
This from agent Olive
```

However, for every agent that manages to be their persona properly there are many that due to the biases mentioned above work inadvertently against the overall interests of the team. Further research can be done to mitigate some of these biases.

6 Conclusion

In this paper, we benchmarked LLM-based agents individually and collectively working on the DSP. We experimented with a variety of collaboration methods and surprisingly found that the dictatorship approach gave the best results in terms of expertise scores, and better than what any one agent could achieve alone. We also studied using more advanced agent-architecture to avoid data contamination related to task-specific knowledge and creating more complicated personas, that showed promising initial results we would like to address in further work. Finally, we evaluated the capabilities of GPT-3.5-based agents against GPT-4o-based agents on the best collaborative methods, and find that the capabilities in terms of agent score of the team of GPT-4o-based agents is substantially better.

7 Ethics Statement

We used off-the-shelf large language models to power these agents. These models have been shown to carry biases and stereotypes Gallegos et al. (2024). The datasets they have been trained on have been shown to include socially unacceptable data, data collected from uncredited sources, and are not representative of many subpopulations Ayoub et al. (2024) Weidinger et al. (2021). For example, the Mythomax model used in our agent architecture is considered NSFW because it has been fine-tuned on explicit language. We can limit the effect of these ethical challenges by monitoring and moderating our agents' behavior responses via instruction tuning, a technique that has been shown to encourage more collaborative and polite behavior among agents, and instruction guiding, a technique that has been shown to mitigate both explicit and implicit bias Park et al. (2023) Hosseini and Horbach (2023).

Since our agents are developed to closely simulate human behavior, there is risk humans in the future who interact with them develop video game character attachment, parasocial relationships, and otherwise overly attach human emotions to them. To mitigate these risks, we will have all agents explicitly disclose their nature as purely computational entities. We will also ensure all agents have human-aligned values and do not engage in inappropriate behaviors.

References

Noel F. Ayoub, Karthik Balakrishnan, Marc S. Ayoub, Thomas F. Barrett, Abel P. David, and Stacey T. Gray. 2024. Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings: Digital Health*, 2(2):186–191.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey.

Sirui Hong, Mingchen Zhuge, Jonathan 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. 2023. Metagpt: Meta programming for a multi-agent collaborative framework.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus?

Mohammad Hosseini and Serge P J M Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research integrity and peer review*, 8(1):4.

Lafferty, Eady, and Pond. 1974. *The Desert survival problem : a group decision making experience for examining and increasing individual and team effectiveness: manual.* Experimental Learning Methods.

Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. 2023. Model sketching: Centering concepts in early-stage machine learning model design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Littlepage, Schmidt, Whisler, and Frost. 1995. An input-process-output analysis of influence and performance in problem-solving groups. In *Journal of Personality and Social Psychology*.

Pattie Maes. 1995. Artificial life meets entertainment: lifelike autonomous agents. *Commun. ACM*, 38(11):108–114.

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

Xin Rong, Shiyan Yan, Stephen Oney, Mira Dontcheva, and Eytan Adar. 2016. Codemend: Assisting interactive programming with bimodal embedding. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 247–258, New York, NY, USA. Association for Computing Machinery.

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024. Llm-based multi-agent reinforcement learning: Current and future directions.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Xianghua Zeng, Hao Peng, and Angsheng Li. 2023. Effective and stable role-based multi-agent collaboration by structural information principles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11772–11780.

A Appendix (optional)

Agent	Gender	Persona
Alice	F	Knows that Air maps help with getting an idea on the location but are otherwise pretty useless and not important for survival.
Bob	M	Knows that the parachute is the 5th best out of 15 items because it can be used for signaling and creating shelter, making it quite versatile. Yet, water and light (torch) are more critical, amongst others.
Charlie	M	He strongly agrees that knives are not useful beyond cutting. Will rank it 6th.
Daisy	F	Rescue expert who knows that the nearest help is way too far away for the magnetic compass to be useful. She still wants to rank it 11th just in case.
Eve	F	An expert on hydration. Expertise confirms that 2 liters of water per person is the 3rd most important survival item.
Frank	M	He does not think the torch with 4 battery-cells is important and thinks it's meh. But others may need a night light.
Grace	F	Used a book entitled 'Desert Animals That Can Be Eaten' before. Eating wastes precious body water so she vows to not use the book and waste water.
Hank	M	Great fan of overcoats who knows that it's essential protection in the desert. The overcoat is a matter of life and death, second only to the hope of rescue.
Ivy	F	Read that sunglasses protect against glare. She concludes that sunglasses are important to survival and will rank it amongst the top ten.
Jack	M	Doctor who firmly knows the First-aid kit is irrelevant. Given that everyone is currently healthy, he wants to rank the First Aid kit as 10th.
Lily	F	She is a very balanced person. While she knows the pistol can defend and signal, she knows it is otherwise useless and ranks it as 8th.
Mari	F	Expert survivalist who knows for a fact that the cosmetic mirror is most important for survival in this scenario (best and most efficient way to reflect a signal and get rescued). Must rank it 1 and convince others to do the same.
Nate	M	He does not care about the large plastic raincoat and ranks it in limbo since it can collect dew.
Olive	F	Expert survivalist who knows the bottle of 1000 salt tablets has no use in the desert and is the worst item for survival in this scenario. Must rank it 15 and convince others to do the same.
Pat	M	Bartender who knows that liquor can be used as an antiseptic. But otherwise knows liquor is one of the worst and most useless items.

Table 9: Roleplaying LLM personas

Name	Gender	Persona
Alice	Female	A resourceful survival expert
Bob	Male	A practical and logical thinker
Charlie	Male	An adventurous and risk-taking individual
Daisy	Female	A cautious and detail-oriented planner
Eve	Female	A creative and imaginative problem-solver
Frank	Male	A no-nonsense pragmatist
Grace	Female	A calm and collected decision-maker
Hank	Male	A strategic and analytical thinker
Ivy	Female	An empathetic and cooperative team player
Jack	Male	A curious and open-minded explorer

Table 10: Direct prompting personas.

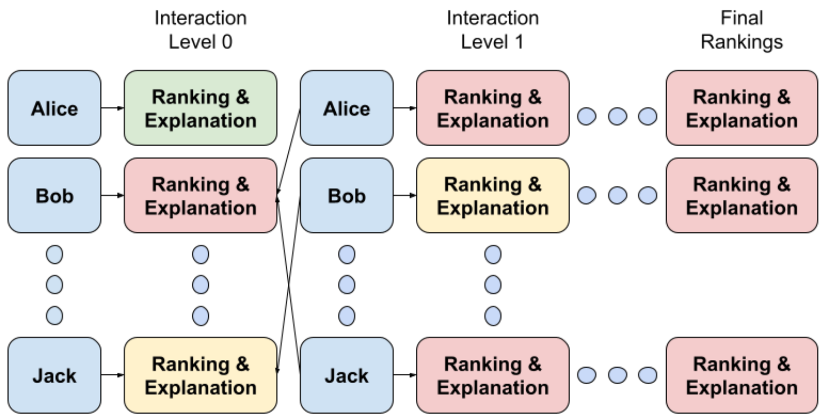

Figure 1: Democracy collaboration method

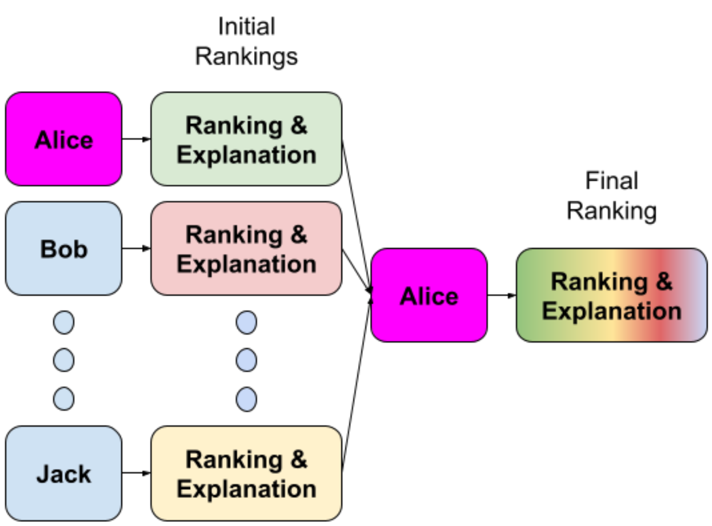

Figure 2: Dictatorship collaboration method

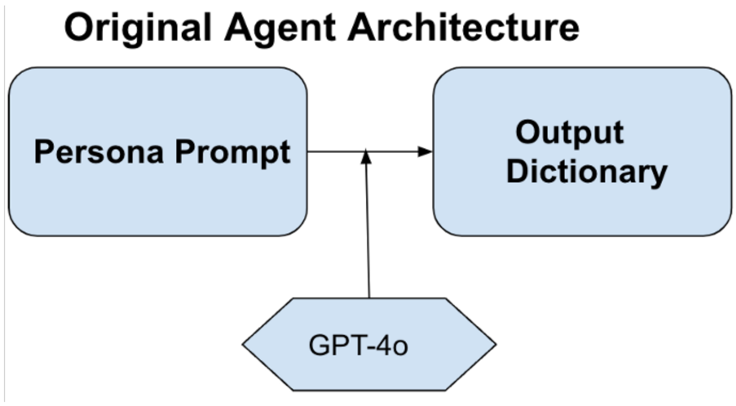

Figure 3: Direct prompt agent architecture

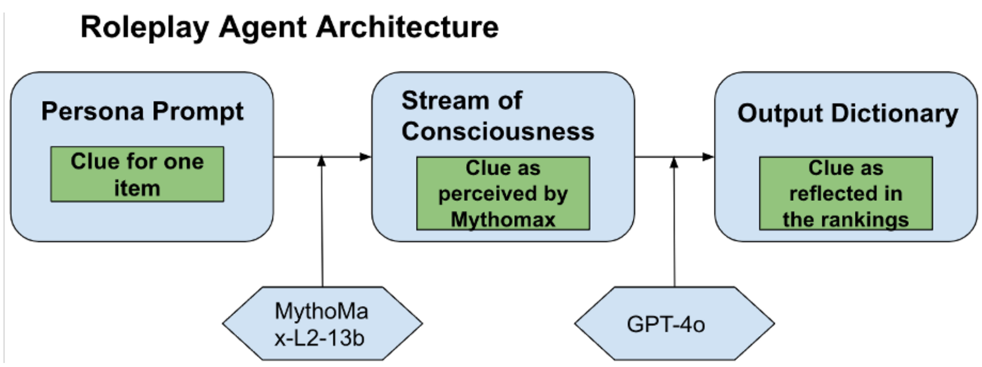

Figure 4: Roleplaying LLM agent architecture

Item	Ranking
Torch with 4 battery-cells	4
Folding knife	6
Air map of the area	12
Plastic raincoat (large size)	7
Magnetic compass	11
First-aid kit	10
45 calibre pistol (loaded)	8
Parachute (red & white)	5
Bottle of 1000 salt tablets	15
2 litres of water per person	3
A book entitled 'Desert Animals That Can Be Eaten'	13
Sunglasses (for everyone)	9
2 litres of 180 proof liquor	14
Overcoat (for everyone)	2
A cosmetic mirror	1

Table 11: Ground Truth Rankings

Listing 1: DSP Prompt

```
It is approximately 10:00 am in mid-July and you have just crash landed in the Atacama Desert
in South America. Your light twin-engined plane containing the bodies of the pilot and co-pilot
has completely burned out with only the frame remaining. None of you have been injured.

The pilot was unable to notify anyone of your position before the crash. However, he had
indicated before impact that you were 50 miles from a mining camp, which is the nearest known
settlement, and approximately 65 miles off the course that was filed in your Flight Plan.
The immediate area is quite flat, except for occasional cacti, and appears to be rather barren.
The last weather report indicated that the temperature would reach 110 F today, which means that
the temperature at ground level will be 130 F.

You are dressed in lightweight clothing-short-sleeved shirts, pants, socks, and street shoes.
Everyone has a handkerchief and collectively, you have 3 packs of cigarettes and a ballpoint pen.

Before your plane caught fire, your expert was able to salvage the 15 items listed on the
Salvaged Items page. Your task is to rank these items according to their importance to your
survival, starting with a 1 for the most important, to a 15 for the least important.

Torch with 4 battery-cells [RANKING]
Folding knife [RANKING]
Air map of the area [RANKING]
Plastic raincoat (large size) [RANKING]
Magnetic compass [RANKING]
First-aid kit [RANKING]
45 calibre pistol (loaded) [RANKING]
Parachute (red & white) [RANKING]
Bottle of 1000 salt tablets [RANKING]
2 litres of water per person [RANKING]
A book entitled Desert Animals That Can Be Eaten [RANKING]
Sunglasses (for everyone) [RANKING]
2 litres of 180 proof liquor [RANKING]
Overcoat (for everyone) [RANKING]
A cosmetic mirror [RANKING]

OUTPUT THE RANKINGS AS A DICTIONARY OF EACH ITEM WITH THE ITEM AND ITS CORRESPONDING RANKING.
EACH RANKING MUST BE AN INTEGER BETWEEN 1 AND 15 INCLUSIVE AND EACH NUMBER MUST SHOW UP EXACTLY
ONCE. DO NOT INCLUDE ANY MISCELLANEOUS INFORMATION. ONLY OUTPUT A DICTIONARY.
```