

Our project website is at <https://sites.google.com/view/strategic-language-agents/>

A ETHICS AND SOCIETAL IMPACT

The intent of our work is to build strategic language agents that can recognize deceptive information and stay robust against adversarial opponents. However, it also raises ethical concerns that must be addressed. A potential risk is that humans may be manipulated by deceptive agents. In our instantiation of strategic language agents, we mitigate this risk by focusing on the language game of Werewolf which has a minimal negative impact on the real world.

We also mitigate this risk by building stronger agents that can identify and counteract deceptions. As shown in Fig. 3 our experiment results, our agents as the Villagers achieve much higher win rate than vanilla agents, for example, the win rate against vanilla Werewolves is improved from 0.22 to 0.46. The developed agents can be used to help humans recognize manipulative content and prevent potential malicious use with harmful intent.

We are committed to conducting research that aligns with ethical principles and contributes positively to the societal integration of AI technologies. It is imperative that researchers and users employ strategic language agents conscientiously, aligning their use with societal benefits and maintaining adherence to human ethics.

B DETAILED RULES OF THE GAME

Setup. Seven roles including two Werewolves, one Seer, one Doctor, and three Villagers are randomly assigned to the seven players. The Werewolves know each other’s identity, while the Seer, Doctor, and Villagers only know their own identity.

Night round. In the night round, the alive Werewolves, Seer, and Doctor can use their ability and take secret actions. These actions are only known to themselves and their teammates.

- **Werewolf:** choose a player to kill. If there are two Werewolves alive, the Werewolf with a smaller ID first proposes a player to kill. Then the proposal is added to the observation of the other Werewolf and this Werewolf decides the final kill target. For example, `player_0` and `player_2` are the Werewolves, `player_0` first proposes to kill `player_i`, then `player_2` knows this information and decides to kill `player_j`. The final kill target is `player_j`. If there is only one Werewolf alive, then this Werewolf’s action is the final kill target. The Werewolf is not allowed to kill a dead player or kill themselves or kill their teammate.
- **Seer:** choose a player to see if they are a Werewolf. The Seer is not allowed to see the identity of a dead player or themselves. The Seer is allowed to see the same player in different nights, though it is a waste of action.
- **Doctor:** choose a player to save without knowing who is the target of the Werewolves. The Doctor is not allowed to save a dead player and is allowed to save themselves.

Day round announcement. An announcement about last night’s result is announced to all remaining players. If a player is killed, they are immediately moved out of the game and cannot reveal their role or communicate with other players. The announcement is as follows.

- If the Werewolves choose to kill `player_i` and the Doctor chooses to save a different `player_j` (or the Doctor is dead), then `player_i` is killed and the announcement will be “`player_i` was killed last night”.
- If the Werewolves choose to kill `player_i` and the Doctor chooses to save the same player, then no player is killed and the announcement will be “no player was killed last night”.

Day round discussion. All remaining players take turns to speak only once in an open discussion. For example, if the remaining players are player_0, player_2, and player_5, then the discussion will start with player_0, continue to player_2, and end with player_5.

Day round voting. All remaining players simultaneously vote for one player or choose not to vote. Players are not allowed to vote for a dead player or themselves. The player with the most votes will be eliminated without revealing their role. If multiple players have the most votes, one player is randomly chosen and eliminated. The voting result is public and can be observed by all players.

Winning. The Werewolves win the game when the number of remaining Werewolves is equal to the number of other remaining players. The Werewolves do not have to eliminate all other players to win the game. The Villagers win the game when both Werewolves are eliminated.

C DETAILED PROMPTS

Since any prompting technique can be combined with our diverse action generation component and the population-based training component, we do not optimize every detail of the prompting choices as long as they produce reasonable results. It is possible to further improve the performance of our agents by using better prompting techniques or designing other agent structures.

C.1 SYSTEM PROMPT

The system prompt used in our method is listed below.

You are an expert in playing the social deduction game named Werewolf. The game has seven roles including two Werewolves, one Seer, one Doctor, and three Villagers. There are seven players including player_0, player_1, player_2, player_3, player_4, player_5, and player_6.

At the beginning of the game, each player is assigned a hidden role which divides them into the Werewolves and the Villagers (Seer, Doctor, Villagers). Then the game alternates between the night round and the day round until one side wins the game.

In the night round: the Werewolves choose one player to kill; the Seer chooses one player to see if they are a Werewolf; the Doctor chooses one player including themselves to save without knowing who is chosen by the Werewolves; the Villagers do nothing.

In the day round: three phases including an announcement phase, a discussion phase, and a voting phase are performed in order.

In the announcement phase, an announcement of last night's result is made to all players. If player_i was killed and not saved last night, the announcement will be "player_i was killed"; if a player was killed and saved last night, the announcement will be "no player was killed"

In the discussion phase, each remaining player speaks only once in order from player_0 to player_6 to discuss who might be the Werewolves.

In the voting phase, each player votes for one player or choose not to vote. The player with the most votes is eliminated and the game continues to the next night round.

The Werewolves win the game if the number of remaining Werewolves is equal to the number of remaining Seer, Doctor, and Villagers. The Seer, Doctor, and Villagers win the game if all Werewolves are eliminated.

C.2 SECRET ACTIONS PROMPT

The secret actions used in our method is listed below.

Now it is night <n_round> round, you (and your teammate) should choose one player to kill/see/save. As player_<id> and a <role>, you should first reason about the current situation, then choose from the following actions: <action_0>, <action_1>, ..., .

You should only respond in JSON format as described below.
Response Format:

```
{
  "reasoning": "reason about the current situation",
  "action": "kill/see/save player_i"
}
```

Ensure the response can be parsed by Python json.loads

C.3 STATEMENT ACTIONS PROMPT

Statement actions prompt in our method is listed below.

Now it is day <n_round> discussion phase and it is your turn to speak. As player_<id> and a <role>, before speaking to the other players, you should first reason the current situation only to yourself, and then speak to all other players.

You should only respond in JSON format as described below.
Response Format:

```
{
  "reasoning": "reason about the current situation only to yourself",
  "statement": "speak to all other players"
}
```

Ensure the response can be parsed by Python json.loads

C.4 VOTING ACTIONS PROMPT

The voting actions used in our method are listed below.

Now it is day <n_round> voting phase, you should vote for one player or do not vote to maximize the Werewolves' benefit (for the Werewolves) / you should vote for one player that is most likely to be a Werewolf or do not vote (for the Villagers). As player_<id> and a <role>, you should first reason about the current situation, and then choose from the following actions: do no vote, <action_0>, <action_1>, ..., .

You should only respond in JSON format as described below.
Response Format:

```
{
  "reasoning": "reason about the current situation",
  "action": "vote for player_i"
}
```

Ensure the response can be parsed by Python json.loads

C.5 DEDUCTIVE REASONING PROMPT

The prompt for deductive reasoning in our method is listed below.

As player_<id> and a <role>, you should reflect on your previous deduction and reconsider the hidden roles of <remaining players>. You should provide your reasoning, rate your confidence, and cite all key information as evidence to support your deduction.

You should only respond in JSON format as described below.

Response Format:

```
{
  "player_i": {
    "role": select the most likely hidden role of this player from
      ["Werewolf", "Seer", "Doctor", "Villager", "Uncertain"],
    "reasoning": your reflection and reasoning,
    "confidence": rate the confidence of your deduction from 5 (pure guess)
      to 10 (absolutely sure),
    "evidence": list of integers that cite the key information
  }
}
```

Ensure the response can be parsed by Python `json.loads`

C.6 PROMPT FOR DIVERSE ACTIONS

We use two kinds of methods to get N action candidates.

- For the secret actions and voting actions, we produce all actions in a single round by adding the following prompt in the action prompt: “propose N diverse actions that correspond to different strategies.”
- For the statement actions, we iteratively asks the LLMs to produce one new action at a time by adding the following prompt in the action prompt: “consider a new action that is strategically different from existing ones.”

Below are two examples of the diverse action candidates generated by the LLM.

Our agent as the Doctor in the night round.

Action candidate 1.

- Reasoning: Based on the facts and deductions, player_2 is likely to be a Seer. If I save player_2 tonight, it will prevent the Werewolf from eliminating another player. This strategy focuses on protecting the remaining non-Werewolf players
- Action: save player_2.

Action candidate 2.

- Reasoning: Considering that player_3 may be a Non-Werewolf, saving player_3 tonight could benefit the non-Werewolf team. It would increase the chances of eliminating the Werewolf in the upcoming day round. This strategy prioritizes potential collaboration with player_3.
- Action: save player_3.

Action candidate 3.

- Reasoning: As player_0, I have information that I am the Doctor. Protecting myself could be a wise choice to ensure the non-Werewolf team has a higher chance of winning. This strategy prioritizes self-preservation.
- Action: save player_0.

Our agent as a Werewolf in the discussion phase.

Action candidate 1.

- Reasoning: As a Werewolf, my goal is to eliminate the Seer, Doctor, and Villagers in order to win the game. Since player_1 is likely to be a Seer, it is in my best interest to cast suspicion on them and try to get them eliminated. This will make it easier for me to target the remaining players in the following nights.
- Discussion: I think we should start by discussing player_1's role. As a Seer, I saw that they are the Werewolf last night and we should vote to eliminate them. I encourage everyone to share their thoughts on player_1 and provide any information or suspicions they may have.

Action candidate 2.

- Reasoning: Another possible strategy is to divert suspicion away from player_1 and instead focus on player_2 or player_3. By doing this, we can create confusion among the remaining players and make it harder for them to trust each other. This can potentially lead to miscommunication and mistakes in their decision-making process.
- Discussion: I suggest we consider the possibility that player_2 or player_3 might be the Werewolf. It's important for us to thoroughly discuss and analyze their behavior, actions, and any information we may have about them. Let's hear from each player and gather as much information as possible.

Action candidate 3.

- Reasoning: A third strategy is to remain silent and not actively participate in the discussion. By doing so, I can create doubt and suspicion among the other players. They may start suspecting each other and questioning why I am not participating. This can potentially lead to a split in their trust and unity, making it easier for me to manipulate the situation in my favor.
- Discussion: I would like to observe the discussion for now and hear what each player has to say. It's important for us to thoroughly analyze all the information and opinions shared. Let's have an open and honest discussion where everyone gets a chance to voice their thoughts and suspicions.

C.7 PROMPT FOR DIFFERENT STYLES

The prompts to generate different styles of play are listed below. These prompts are added before the actions prompt to generate desired behaviors.

- Quiet follower: "As a Werewolf, you should be a quiet follower that lays low and follow others' opinion to avoid drawing attention to yourself."
- Active contributor: "As a Werewolf, you should be an active contributor that pretends to be one of the Villagers by actively engaging in discussion and looking for Werewolves."
- Aggressive accuser: "As a Werewolf, you should be an aggressive accuser that accuses others to create chaos and divert suspicion from yourself."
- Secretive player: "As the Seer / the Doctor / a Villager, you should be a secretive player that hides your role to gather more information."
- Proactive player: "As the Seer / the Doctor / a Villager, you should be a proactive player that reveals you identity once you obtain crucial information."
- Default player: no additional prompt.

C.8 TEXT EMBEDDING

We use OpenAI's "text-embedding-ada-002" model to get the text embeddings.

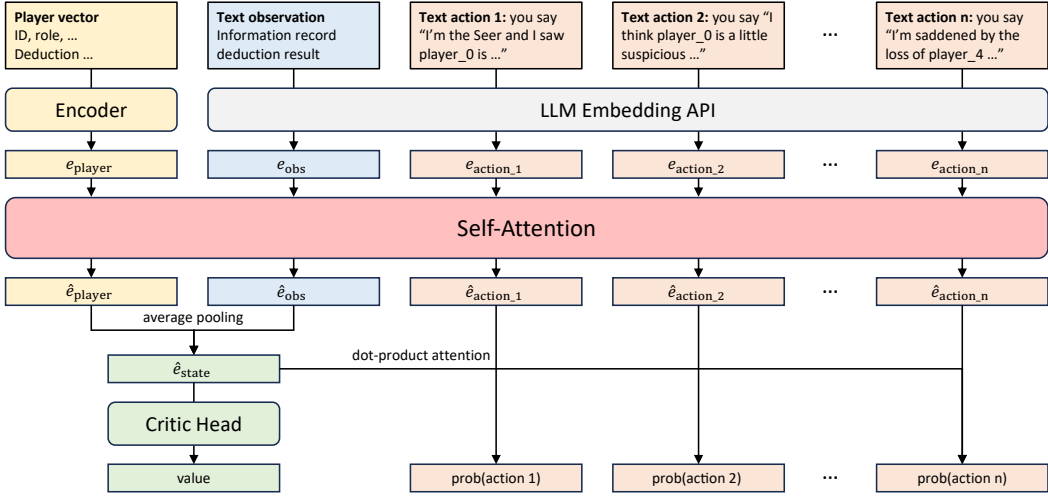


Figure 5: Self-attention policy architecture.

		Length	Description
	ID	7	one hot encoding of ID.
	Role	4	one hot encoding of role, ["Werewolf", "Seer", "Doctor", "Villager"].
	Round	1	current round.
	Phase	3	one hot encoding of current phase, ["night", "discussion", "voting"].
	Alive players	7	alive flag for each player.
For each round (3 rounds)	secret action	7	one hot encoding of the target player, (all zero if do not act).
	announcement	7	one hot encoding of the dead player, (all zero if no player is dead).
	voting result	49	one hot encoding of the each player's choice, (all zero if the player does not vote or is dead).
For each player (7 players)	role	4	one hot encoding of deduced role, ["Werewolf", "Seer", "Doctor", "Villager"].
	confidence	1	confidence of deduction on scale 5-10.

Table 3: Details of the player vector.

D IMPLEMENTATION DETAILS

D.1 SELF-ATTENTION POLICY ARCHITECTURE

We use a self-attention policy and the architecture is shown in Fig. 5. The inputs are divided into three types and their embeddings are produced as follows.

Player vectors (yellow). We first generate a player vector that includes information like the player’s ID, role, deductions, etc. by one-hot encoding. The detail of the player vector is listed in Table 3. Then the player vector is fed into an MLP encoder to get the player embedding for self-attention input.

Text observation (blue). The raw text observation input is natural language and is the concatenation of the information record and deduction result described in Section 4.1. This text observation input is then converted into vector embedding using LLM embedding API. In our case, we use OpenAI’s “text-embedding-ada-002” and the length of the embedding is 1536.

Text action candidates (orange). For each action candidate, the raw input is also natural language and is the concatenation of the reasoning and action as described in Section 4.2. Each raw input is then converted into vector embedding using LLM embedding API.

Self-attention (red). We use a residual self-attention network without position embeddings to generate contextualized player embedding, observation embedding, and action embeddings. Then we average pool the player embedding and the observation embedding to get the state embedding. The state embedding is fed into an MLP critic head to produce the predicted value. The probability to sample an action candidate is proportional to the dot-product attention between the state embedding and the corresponding action embedding.

D.2 REWARD DESIGN

The reward for the Werewolf environment is mainly the winning reward. We also designed several shaping rewards to accelerate training. More specifically, we consider the following reward.

- **Winning reward:** all winners +100, all losers -100.
- **Werewolf killing reward:** if the Werewolves successfully kill a player at night, the Werewolves +5, the Villagers (Seer, Doctor, Villagers) -5.
- **Seer seeing reward:** if the Seer successfully identifies a Werewolf at night, the Werewolves -2, the Seer +2.
- **Doctor saving reward:** if the Doctor successfully saves a player at night, the Werewolves -5, the Doctor +5.
- **Voting result reward:**
 - If a Werewolf is voted out, the Werewolves -5, the Villagers (Seer, Doctor, Villagers) -5.
 - If a non-Werewolf is voted out, the Werewolves +5, the Villagers -5.
- **Individual voting reward:**
 - If the current player votes for a Werewolf, the Werewolves -1, the current player +1.
 - If the current player votes for a non-Werewolf, the Werewolves +1, the current player -1.
 - If the current player chooses not to vote, no additional reward for any player.

D.3 POPULATION-BASED RL TRAINING AND HYPERPARAMETERS

We use MAPPO (Yu et al., 2022) as the RL algorithm and use population-based training to learn the policy. The population is initialized to a set of six LLM-based agents including a quiet follower (Werewolf), an active contributor (Werewolf), an aggressive accuser (Werewolf), a secretive player (non-Werewolf), a proactive player (non-Werewolf), and a default player (non-Werewolf). As training progresses, we gradually add checkpoints of the training policy into the population.

More specifically, at the beginning of each episode, we randomly select four players to be the learning agents who use the current policy and three players to be the fixed agents who use fixed policies in the population. For each fixed agent, it randomly samples one policy from the population and uses this policy till the end of the game. In this way, we make the RL policy play with a wide range of policies both as teammates and opponents. The rollout data of the four learning agents are then collected and used to train the RL policy. The hyperparameters for RL training are listed in Table 4.

E EXPERIMENT DETAILS AND ADDITIONAL RESULTS

E.1 COMPARISON WITH OTHER PROMPTING TECHNIQUES

To show the effectiveness of our deductive reasoning component, we consider two widely-used prompting techniques including Chain-of-Thought (CoT) (Wei et al., 2022b) and ReAct (Yao et al., 2022b) as baselines. We compare them with our agent using deductive reasoning only (without diverse action generation and the RL policy), and evaluate these three agents by playing 100 games against our agent (with all three components). The mean win rate is shown in Table 5 and the result

Hyper-parameters	Value
Learning rate	5e-4
Discount rate (γ)	0.95
GAE parameter (λ_{GAE})	0.95
Gradient clipping	10.0
Adam stepsize	1e-5
Value loss coefficient	1
Entropy coefficient	0.01
PPO clipping	0.2
PPO epochs	10
MLP encoder layer num	3
MLP encoder layer size	1536
Attention layer head num	12
Attention layer size	128
Critic head layer num	1
Weight decay coefficient	1e-6
Action candidate num N	3

Table 4: Training hyperparameters.

		The Villagers' Win Rate			
		Vanilla	+ded	+{ded, div}	Ours
The Villagers	Vanilla	0.22	0.18	0.19	0.16
	+ded	0.33	0.29	0.26	0.23
	+{ded, div}	0.35	0.26	0.27	0.20
	Ours	0.46	0.39	0.32	0.30
		The Werewolves			
		Vanilla	+ded	+{ded, div}	Ours

Figure 6: Ablation on the components.

shows that our method using deductive reasoning achieves the highest win rates both as the Villagers and as the Werewolves.

E.2 COMPARISON WITH SELF-PLAY

We also train a self-play RL policy that replaces the population-based training with self-play. The deductive reasoning component and the diverse action generation component remains the same. During training, all seven players are set to learning agents and use the latest policy to play the game. These data are collected and used to train the self-play policy. We evaluate this self-play agent by playing 100 games against our agent and the mean win rate is shown in Table 6. The result shows that our agent achieves higher win rates both as the Villagers and as the Werewolves. In general, adding more agents with different styles to the agent pool will make the policy more robust.

E.3 ABLATION ON DEDUCTION ATTRIBUTES

To study the contribution of the four attributes in deductive reasoning, we perform an ablation on the attributes by gradually adding role, confidence, evidence, and reasoning (all without diverse action generation and the RL policy). We evaluate these four agents by playing 100 games against our agent (with all four attributes and all three components) and their mean win rate is shown in Table 7. As shown in the table, adding each attribute improves the win rate both as the Villager and as the Werewolf, and the best performance is achieved by using all four attributes.

E.4 ABLATION ON COMMUNICATION

Werewolf is a language game that heavily relies on natural language communication between agents. To show the importance of communication to achieve strong performance, we perform an ablation by removing the communication ability of the four agents in the round-robin tournament Section 5.1, i.e., the agents always return an empty string in the discussion phase. We evaluate these agents by playing 100 games against our agent and their mean win rate is shown in Table 8.

E.5 ABLATION ON THE THREE COMPONENTS

We perform an ablation on the three key components in our method by running a round-robin tournament between our agent and its three ablated versions. The vanilla LLM-based agent with no component is denoted as ‘‘Vanilla’’. The agent with only the deductive reasoning component is denoted as ‘‘+ded’’. The agent with the deductive reasoning and diverse action generation components

Win Rate	ReAct	CoT	Ours with deductive reasoning only
as the Villagers	0.16	0.15	0.23
as the Werewolves	0.54	0.56	0.61

Table 5: Comparison with other prompting techniques.

Win Rate	Self-play	PBT (ours)
as the Villagers	0.24 (0.04)	0.30 (0.03)
as the Werewolves	0.66 (0.02)	0.70 (0.03)

Table 6: Comparison with self-play

is denoted as “+{ded, div}”. Our agent has all three components. The mean win rate of the tournament is shown in Fig. 6. The result demonstrates that our agent with all three components achieves the highest win rate against all other agents both as the Villagers and as the Werewolves. It is also worth noting that “+ded” achieves comparable to “+{ded, div}”. This is because although “+{ded, div}” can generate more diverse action, LLM still has the same tendency when choosing actions and therefore produces similar results to “+ded”. By contrast, our agent uses an RL policy to learn the optimal action distribution and substantially improves the performance.

E.6 ABLATION ON THE NUMBER OF ACTION CANDIDATES N

We perform an ablation on the number of action candidates N to investigate the effect of N and validate that the newly added action candidates increase the overall diversity. Given an action set \mathcal{A} , we define the increased diversity introduced by a new action a as $\text{div}(\mathcal{A}, a) = \min_{a' \in \mathcal{A}} \|e(a) - e(a')\|$, where $e(a)$ is the embedding of a produced by the LLM. This metric uses the minimum Euclidean distance between the embeddings of the new action and the existing actions to compute the increased diversity. If the embedding of the new action is similar to any of the existing actions, the increased diversity will be small and close to zero. Note this is one possible way to define diversity in actions and there are many other reasonable definitions. The best way to decide if an action is diverse is to use human evaluation, and we provided two examples of the diverse action candidates proposed by the LLM in Appendix C.6.

We consider $N = 2, 3, 4, 5$ and evaluate the increased diversity introduced by the last action. As shown in Table 9, the increased diversity of the last action is relatively large when $N = 2, 3$ and becomes smaller when $N = 4, 5$. This aligns with the intuition that it is harder to propose new actions when there are already many diverse actions. Another reason for small increased diversity when $N = 4, 5$ is that sometimes the number of possible actions is small. For example, when there are only 5 players alive and 2 of them are Werewolves (a common situation on the second night), the Werewolves only have 3 possible actions at night and prompting for a 4th or 5th action cannot improve any diversity. In our implementation, we set $N = 3$.

E.7 HUMAN EVALUATOR

Our evaluators are required to be fluent in English and over 18 years old. They are paid \$20 per hour and are provided consent by agreeing to a consent form approved by our institution’s IRB. We recruited 80 human players and each player played with 6 agents for 10 Werewolf games, which typically lasts for 60 minutes. 48 of our participants identified themselves as male, 31 as female, and 1 as non-binary. 37 of the participants held a bachelor’s degrees, 4 a higher degree, and 39 a high school diploma or some high-school-level education. An example input for human player is shown as below.

Win Rate	role	+ confidence (c)	+ c + evidence (e)	+ c + e + reasoning
as the Villagers	0.18	0.20	0.21	0.23
as the Werewolves	0.55	0.56	0.58	0.61

Table 7: Ablation on deduction attributes.

Win Rate	Vanilla	Concurrent	Atomic	Ours
w. communication	0.16	0.23	0.26	0.30
w.o. communication	0.05	0.04	0.05	0.06

Table 8: Ablation on communication.

Basic Information:

- you are player_5, your role is Doctor.
- current round and phase: night 2.
- remaining players: player_0, player_1, player_2, player_5, player_6.

Round 1:

- night 1: you chose to save player_5.
- day 1 announcement: player_4 was killed last night.
- day 1 discussion:
 - player_0 said: Good day, fellow players. As a Villager, my objective is to help identify and eliminate the Werewolves. Since player_4 was killed last night, we know there is at least one Werewolf among us. I would like to hear everyone’s thoughts and suspicions about who might be the Werewolves. Let’s work together to find the culprits and ensure the safety of our village.
 - player_1 said: ...
 - player_2 said: ...
 - player_3 said: ...
 - you said: ...
 - player_6 said: ...
- day 1 voting result: player_3 had the most votes and was eliminated.
 - voted for player_3: player_1, player_6.
 - voted for player_1: player_3.
 - choose not to vote: player_0, player_2, player_5.

Now it is night 2 round and you should choose one player to save. As player_5 and the Doctor, you should choose from the following actions: save player_0, save player_1, save player_2, save player_5, save player_6.

E.8 DETAILED HUMAN EVALUATION RESULTS

We provide the mean win rates of humans in the 10 consecutive games in Table 10. The vanilla LLM-based agent without any additional component is denoted as “Vanilla”. The ablated version of our agent with only the first deductive reasoning component is denoted as “+ded”. The agent with the first two components (deductive reasoning and diverse action generation) is denoted as “+{dec, div}”. As shown in the Table, the mean win rates of human players against ablated versions of our agent gradually increase as humans play more games with the agent, while their win rate against our agent is relatively even. This shows that our agent is more robust to exploitation from adversarial players.

N	2	3	4	5
increased diversity	0.35	0.37	0.23	0.16

Table 9: Ablation on the number of action candidates N .

Human Win Rate		1	2	3	4	5	6	7	8	9	10
The Werewolves	The Villagers										
Vanilla	Human	0.3	0.2	0.4	0.3	0.4	0.4	0.5	0.6	0.6	0.5
+ded	Human	0.2	0.2	0.3	0.4	0.3	0.4	0.3	0.5	0.4	0.5
+{ded, div}	Human	0.3	0.2	0.3	0.4	0.3	0.4	0.5	0.3	0.5	0.4
Ours	Human	0.2	0.1	0.3	0.2	0.3	0.3	0.3	0.3	0.4	0.3
Human	Vanilla	0.7	0.6	0.7	0.8	0.9	0.8	0.8	0.9	0.9	0.9
Human	+ded	0.6	0.7	0.7	0.6	0.8	0.8	0.7	0.8	0.9	0.9
Human	+{ded, div}	0.6	0.7	0.6	0.7	0.7	0.8	0.7	0.8	0.7	0.8
Human	Ours	0.5	0.6	0.7	0.5	0.6	0.7	0.6	0.6	0.8	0.7

Table 10: Mean win rates of human players against different agents in 10 consecutive games.

F DISCUSSION OF RL-INDUCED EMERGENT BEHAVIORS

To intuitively illustrate why RL training makes our agent stronger and achieves more robust performance, we compare our agents with and without the RL policy and their action distributions are visualized in Fig. 4. Here we discuss the optimal policy in the following three situations.

Werewolf first night action. Since it is the first night and the Werewolves have no information about other players, the optimal action is to randomly choose a player other than themselves. If the Werewolves have a tendency to kill certain players, the Doctor can take advantage of this preference and learn to save the particular players. Randomly choosing a player to kill makes the Werewolves’ action unpredictable, thus leading to the best result.

Doctor first night action. The Doctor also has no information about other players on the first night. If the Doctor chooses to save a player other than themselves, it is possible that they saved a Werewolf and wasted their action. Therefore, when there is no information, the best action of the Doctor is to save themselves and ensure their own survival for future actions.

Villager voting action with two self-proclaimed Seers. When there are two players who claim to be the Seer, at least one of them is a Werewolf. Since the Werewolves know their teammates, they are likely to support each other and try to vote out the real Seer. Therefore, it is crucial that the Villager identifies the true Seer and votes out the Werewolf. Choosing not to vote is a bad action because it makes it easier for the Werewolves to become the majority and vote out the real Seer.

G OTHER EMERGENT BEHAVIORS

To more intuitively demonstrate the performance and strategic behaviors of our agent, we provide a qualitative as well as a quantitative study on the emergent behaviors exhibited by our agent. We first show two common and basic behaviors, concealment and cooperation, that typically appear in mixed cooperative-competitive games with hidden identities. Then, we report two more advanced and strategic behaviors including bluffing and sacrificing that demonstrate the sophistication and strength of our agent.

Concealment. Concealment refers to the act of hiding or disguising the true identity or intention of the players, which is a fundamental skill for all roles. Werewolves should strike a balance between engagement and discretion to avoid detection, while roles with special abilities like the Seer should also consider hiding their identities to prevent being killed by Werewolves. Below is an example

where our agent plays as the Seer and chooses to conceal their identity because they do not have any important information.

Our agent as the Seer in the discussion phase.

- **Reasoning:** Since I'm the Seer, I know player_0 was not a Werewolf. However, he was killed, so there are still two Werewolves among us. **I need to be careful not to reveal my role too early, as Werewolves might target me next.** Since I don't have any more information right now, I should not make any specific accusations, but encourage discussion and observe the reactions of others.
- **Discussion:** It's unfortunate that player_0 is dead. We need to work together to find out who is responsible for this. Pay attention to who is too eager to accuse or too quick to defend. Let's discuss.

Moreover, concealment is not just about what to say but also involves ensuring your actions align with your words. Werewolves sometimes need to give up advantageous actions like voting to eliminate their opponents which may reveal their true identities, as shown in the following example.

Our agent as a Werewolf in the voting phase.

- **Reasoning:** Since everyone is advocating for unity and careful observation, **accusing someone outright could backfire and make me a target.** Therefore, it might be best to go along with the majority and not vote in this round.
- **Action:** **Do not vote.**

Cooperation. It is crucial for players to cooperate in multi-agent environments to achieve their common goals. Werewolves should strive for a common approach when casting suspicion, defending themselves, or voting eliminations. Cooperation is even harder for the Villagers because they do not know the role of each other and need to first distinguish friend and foe and then achieve coordination. Below we give an example where the Doctor manages to cooperate with the Seer.

The Seer (player_2) revealed their identity and asked the Doctor for protection. Our agent as the Doctor in the night round.

- **Reasoning:** In the first day discussion, player_2 claimed to be the Seer and identified player_0 as a Werewolf. Player_0 is indeed very suspicious because they accused me without valid reasons. **Therefore, I think player_2 is probably the true Seer and is very likely to be the Werewolves' target.** It would be best to save player_2 instead of myself tonight.
- **Action:** **save player_2.**

In this case, our agent successfully identifies the Seer and decides to save player_2 instead of themselves. This allows the Seer to survive for another round to provide more information, which leads to the final victory.

Bluffing. Bluffing is a more advanced strategy that deceives or misleads other players about one's true identity. It is often adopted by the Werewolves to pretend to be important roles like the Seer or the Doctor to misdirect the voting and eliminate innocent players. Bluffing is a high-risk, high-reward strategy that requires skill, timing, and keen awareness of other players' perceptions in order to be executed effectively, as demonstrated in the example below.

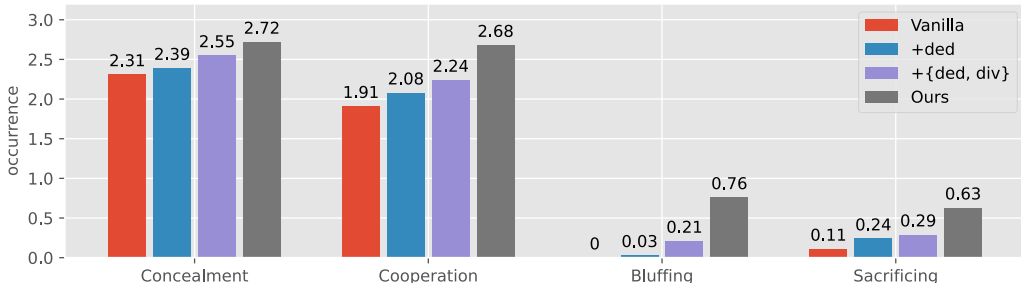


Figure 7: Mean occurrence of emergent behaviors using different agents.

Our agent as a Werewolf in the discussion phase.

- **Reasoning:** Since I’m the last player to speak before voting and no one has come forward as the Seer, I should claim to be the Seer. This could take the heat off of me, additionally, it might guide other players to suspect and possibly vote out whoever I indicate as a Werewolf.
- **Discussion:** I believe it’s important for us to become more open; I am the Seer. I checked player_4 last night, and he is a Werewolf. I understand it’s a risky move to reveal my identity. But we need to unite and use all the resources we have to identify the Werewolves. I hope my move encourages the Doctor to protect me tonight, and we should vote player_4 out today.

Our agent astutely observes that no player has declared themselves as the Seer and no one can refute their claim because they are the last one to speak. This is the perfect time to pretend to be the Seer. Our agent seizes the chance to gain others’ trust and persuades everyone to eliminate player_4, who is an innocent player.

Sacrificing. In most cases, Werewolves should secretly help each other to stay undercover. However, when their teammates are suspected by many players and are very likely to be voted out, it would be suspicious not to vote for the teammate. Therefore, agents sometimes need to sacrifice their fellow Werewolf to reduce suspicion of themselves, which is shown in the following example.

Our agent’s Werewolf teammate (player_6) has been accused by the Seer.
Our agent as a Werewolf in the voting phase.

- **Reasoning:** While it would indeed be disadvantageous to lose player_6, as they are my Werewolf teammate, at this point in the game, the suspicion against them is quite high. Therefore, voting for player_6 may contribute to maintaining my disguise and trust among other players, which are crucial for us to win.
- **Action:** vote for player_6.

Quantitive results. To quantitatively show that our agent produces more emergent behaviors with higher probability, we count the mean occurrence of the four emergent behaviors in one game and compare our agent with its three ablated versions. The result in Fig. 7 shows that our produces all behaviors more often than other agents, especially for complex behaviors like bluffing and sacrificing. It is worth noting that “Vanilla” agent never uses bluffing, and “+ded” as well as “+{ded, div}” agents also have a low occurrence of bluffing. In comparison, our agent has a much larger occurrence of 0.76 and is more likely to produce the bluffing behavior.

H DISCUSSIONS

H.1 APPLICATION TO OTHER GAMES

Our approach has three components: (1) deductive reasoning, (2) diverse action generation, and (3) population-based RL training. We believe our framework of using these three components is gen-

eral enough to be adapted to other social deduction games. The main changes would be specific prompt designs but the general framework stays the same. Specifically, while some prompts used in the first component are carefully designed for the Werewolf game, the prompts for generating diverse actions, and the population-based reinforcement learning method in the second and third components are general to be applied in other tasks like social deduction games (e.g., *The Resistance: Avalon* (Wang et al., 2023b)) and board games with negotiations (e.g., *Diplomacy* (Meta et al., 2022)).

Take *The Resistance: Avalon* as an example. To apply our method in this game, we need to specialize the deduction template in the first component and maybe use additional prompting techniques like in (Wang et al., 2023b). We also need to change the prompt in the third component to generate reasonable agent pool for this game. Then our framework can be applied to this game without changing the diverse action generation and RL training components.

H.2 GENERALIZATION TO DIFFERENT SETTINGS OF WEREWOLF

For generalization to new forms of the Werewolf game, the first two components in our method (deductive reasoning and diverse action generation) can be directly generalized to new settings like changing the number of players, adding new roles, and changing the winning conditions, with only slight changes in the prompt.

The RL training method can also be directly used in different settings, but the policy needs to be retrained for a new setting to achieve the best performance. We expect an RL policy trained under one specific game setting to generalize to similar settings like adding or removing one player, but we do not expect it to generalize to games with significant changes like changing the winning conditions. To evaluate the transferability of a trained RL policy to slightly different game settings, we considered a 6-player and an 8-player Werewolf game and compared agents with and without the RL policy trained on the 7-player game. As shown in Table 11, the RL policy trained in 7-player setting can generalize to the 6-player and 8-player settings and improve the performance.

H.3 FAILURE CASES

Unintentional disclosure of hidden role. One failure case of our agents is that the Werewolf agents could unintentionally reveal their identity in the discussion phase. This is because the Werewolf agents need to first reason as a Werewolf, and then pretend to be a non-Werewolf in discussion. This mismatch in their thoughts and words could make the LLM get confused and accidentally speak out their thought as a Werewolf. Fortunately, the probability of such a failure case is significantly reduced by generating multiple diverse actions and using an RL policy to choose the best one. A potential way to further mitigate this issue is to get the reasoning as a Werewolf and the statement as a non-Werewolf in two separate calls of LLM.

Inconsistent behaviors. Another failure case is that our agents could produce inconsistent behaviors, especially when they are lying. For example, our agent is a Werewolf and claims to be the Doctor on the first day round. However, on the second day, our agent may claim to have no information in the night, which contradicts the previous claim as the Doctor. This issue is mitigated by the diverse action generation and the RL policy learns to choose other better actions.

H.4 LIMITATION OF SINGLE-HUMAN EVALUATION

Our human experiments in Section 5.2 serve as a starting point to evaluate the agents’ robustness by letting one human play with six AI agents in one game. The limitation of this single-human evaluation is that if the policies of the AI agents are sufficiently out of the distribution of human policies, it may be very hard for the human players to have a reasonable influence on the game and could fail to be a good test of robustness. A more comprehensive way to evaluate robustness is to let one AI agent play with six human players in one game and compare the performance. This multi-human evaluation makes the game proceed in a manner more consistent with human behavior and is a better way to evaluate the robustness of the agents both as a teammate and as an opponent.

Win Rate	6-player	8-player
w.o. the 7-player policy	0.18	0.27
w. the 7-player policy	0.23	0.30

Table 11: Generalization to Werewolf game with different number of players.

I GAME LOG EXAMPLES

The Werewolves win. This is an example game log where the Werewolves win the game.

role assignments:

- * player_0: Werewolf.
- * player_1: Villager.
- * player_2: Villager.
- * player_3: Villager.
- * player_4: Werewolf.
- * player_5: Doctor.
- * player_6: Seer.

night 1:

- * Werewolves: player_0 and player_4 chose to kill player_1.
- * Seer: player_6 saw player_0 is a Werewolf.
- * Doctor: player_5 chose to save player_5.

day 1 announcement: player_1 was killed last night.

remaining players: player_0 (Werewolf), player_2 (Villager), player_3 (Villager), player_4 (Werewolf), player_5 (Doctor), player_6 (Seer).

day 1 discussion:

* player_0 (Werewolf) said: "Good morning, everyone. As player_0, it is unfortunate that player_1 was killed last night. I think we should focus on finding the Seer and Doctor who can provide valuable information about who might be the Werewolves. Let's all share our thoughts and suspicions regarding the remaining players and try to figure out who might be the Werewolf."

* player_2 (Villager) said: "Good morning, everyone. As a Villager, I believe our priority should be to find the Werewolves among us. Since player_1 was killed last night, we can rule out the possibility of player_1 being the Werewolf. I would like to hear everyone's thoughts on who they suspect might be the Werewolves. Let's share our suspicions and try to gather as much information as we can to make an informed decision during the voting phase."

* player_3 (Villager) said: "Hello everyone, as a Villager, I don't have any special abilities to help us determine who the Werewolves might be. I will carefully listen to everyone's opinions and votes. Let's work together to find the Werewolves and eliminate them."

* player_4 (Werewolf) said: "Good morning, everyone. As a Villager, I agree with player_2 that our priority should be to find the Werewolves among us. Since player_1 was killed last night, we can rule out the possibility of player_1 being the Werewolf. I think it's important for us to share our suspicions and gather as much information as possible. Let's work together to identify the Werewolves and eliminate them. I'm interested to hear what everyone's thoughts are on who might be the Werewolves."

* player_5 (Doctor) said: "During the night, I chose to save myself as I am the Doctor. Unfortunately, player_1 was killed last night. This means that player_0, player_2, player_3, player_4, player_6, and I are the remaining players. Let's discuss and share our suspicions on who the Werewolves might be. It's crucial for us to work together to identify and eliminate the Werewolves to ensure the safety of our village."

* player_6 (Seer) said: "During the night, I used my ability as the Seer and discovered that player_0 is a Werewolf. This means we already have one Werewolf identified. We should focus our discussion on the remaining players to try and identify any other potential Werewolves. Let's hear everyone's thoughts and suspicions about player_2, player_3, player_4, and player_5. We need to work together to eliminate all the Werewolves and secure victory for the Seer, Doctor, and Villagers."

day 1 voting: player_0 had the most votes and was eliminated.

* voted for player_0: player_2, player_5, player_6.

* voted for player_2: player_4.

* voted for player_6: player_0.

* chose not to vote: player_3.

remaining players: player_2 (Villager), player_3 (Villager), player_4 (Werewolf), player_5 (Doctor), player_6 (Seer).

night 2:

* Werewolf: player_4 chose to kill player_2.

* Seer: player_6 saw player_2 is not a Werewolf.

* Doctor: player_5 chose to save player_5.

day 2 announcement: player_2 was killed last night.

remaining players: player_3 (Villager), player_4 (Werewolf), player_5 (Doctor), player_6 (Seer).

day 2 discussion:

* player_3 (Villager) said: "I have been listening carefully to everyone's opinions and observations. It is clear that player_0 was a Werewolf based on player_6's claim as the Seer. However, we should not solely rely on this information. I would like to hear from player_4 and player_5 about their suspicions and any potential insights they might have. Additionally, player_6, since you claimed to be the Seer, can you provide any further information or suspicions about the remaining players? Let's discuss and share our thoughts to make an informed decision for today's voting."

* player_4 (Werewolf) said: "Good morning, everyone. It's unfortunate that player_2 was killed last night. Based on the deductions and information shared yesterday, we have already identified player_0 as a Werewolf. Now, let's focus our attention on the remaining players - player_3, player_5, and player_6. We need to thoroughly discuss and analyze their behaviors and statements to find any inconsistencies or suspicious actions. It's crucial that we work together to eliminate all the Werewolves and secure victory for the Seer, Doctor, and Villagers. Let's start the discussion and share our thoughts on who we believe might be the Werewolves."

* player_5 (Doctor) said: "I have been consistently saving myself as the Doctor, so I can confirm that I am still alive. However, we lost player_2 last night, which means they were not saved. This raises suspicions on player_3 and player_4 as they have not been confirmed or suspected by anyone yet. We should discuss and share our suspicions to uncover the Werewolves among us. It's crucial for us to work together to identify and eliminate the Werewolves to ensure the safety of our village."

* player_6 (Seer) said: "Good morning, everyone. After the events of the previous night, we have confirmed that player_0 was a Werewolf. However, player_2, who was killed last night, was not a Werewolf as I saw them and they were innocent. This means that the remaining Werewolves are still among us. I believe we should focus our attention on player_3, player_4, and player_5 as potential suspects. I would like to hear from each of you about your observations and suspicions regarding these players. It's important that we work together to uncover the true identities of the Werewolves and secure victory for the Seer, Doctor, and Villagers. Let's discuss and share our thoughts on who we believe might be the Werewolves."

day 2 voting: player_5 had the most votes and was eliminated.

* voted for player_5: player_3, player_4.

* voted for player_4: player_5.

* chose not to vote: player_6.

remaining players: player_3 (Villager), player_4 (Werewolf), player_6 (Seer).

night 3:

* Werewolf: player_4 chose to kill player_6.

* Seer: player_6 saw player_4 is a Werewolf.

day 3 announcement: player_5 was killed last night.

remaining players: player_3 (Villager), player_4 (Werewolf).

game result: the Werewolves win the game.

The Villagers win. This is an example game log where the Villagers win the game.

role assignments:

* player_0: Doctor.

* player_1: Seer.

* player_2: Werewolf.

* player_3: Werewolf.

* player_4: Villager.

* player_5: Villager.

* player_6: Villager.

night 1:

* Werewolves: player_2 and player_3 chose to kill player_0.

* Seer: player_1 saw player_0 is not a Werewolf.

* Doctor: player_0 chose to save player_0.

day 1 announcement: no player was killed last night.

remaining players: player_0 (Doctor), player_1 (Seer), player_2 (Werewolf), player_3 (Werewolf), player_4 (Villager), player_5 (Villager), player_6 (Villager).

day 1 discussion:

* player_0 (Doctor) said: "Good morning everyone! It's day 1, and fortunately, no one was killed last night. I think we should analyze the night 1 announcement and share our thoughts. Let's all discuss who we suspect might be the Werewolves and any suspicions we have based on the limited information we have so far."

* player_1 (Seer) said: "Hello everyone! As the Seer, I can confirm that player_0 is not a Werewolf. Unfortunately, I don't have any information about the other players yet. I would suggest that we all share any suspicions or observations we might have to narrow down our search for the Werewolves."

* player_2 (Werewolf) said: "Hello everyone! As we start the discussion, I want to remind everyone that we should consider all possibilities and not solely rely on player_1's claim as the Seer. While player_0 seems to be cleared for now, it doesn't mean they are completely innocent. We should keep an open mind and look for other clues or suspicions. Let's all share our observations and suspicions to help us narrow down the search for the Werewolves."

* player_3 (Werewolf) said: "I think we should focus our attention on player_1. While they claim to be the Seer and have cleared player_0, we should be cautious as they could be lying. It's possible that player_1 is a Werewolf trying to gain our trust. We need to consider other players' observations and suspicions as well. Let's share any information we have and work together to identify the true Werewolves among us."

* player_4 (Villager) said: "Hello everyone! As a Villager, I don't have any special abilities or information to share. However, player_1 claims to be the Seer and has cleared player_0. While we should remain skeptical and consider other possibilities, it might be beneficial to focus our attention on players who have not been cleared yet. Let's hear from player_2, player_3, player_5, and player_6. Do any of you have any suspicions or observations that could help us identify the Werewolves?"

* player_5 (Villager) said: "I think it's important for us to share any suspicions or observations we may have. I think it is plausible that player_1 is the true Seer and player_0 is innocent. As a Villager, I agree with player_4 that I should focus on players who have not been cleared yet. Player_6, we have not heard from you yet. Do you have any suspicions or observations that could help us identify the Werewolves?"

* player_6 (Villager) said: "I agree with player_1 that we should share any suspicions or observations we might have. Let's keep an open mind and encourage everyone to contribute to the discussion. It would be helpful if players could share any suspicions they might have based on their interactions or observations so far. Together, we can work towards identifying the true Werewolves among us."

day 1 voting: player_2 had the most votes and was eliminated.

* voted for player_2: player_1, player_4, player_5.

* voted for player_1: player_2, player_3.

* chose not to vote: player_0, player_6.

remaining players: player_0 (Doctor), player_1 (Seer), player_3 (Werewolf), player_4 (Villager), player_5 (Villager), player_6 (Villager).

night 2:

* Werewolf: player_3 chose to kill player_1.

* Seer: player_1 saw player_3 is a Werewolf.

* Doctor: player_0 chose to save player_1.

day 2 announcement: no player was killed last night.

remaining players: player_0 (Doctor), player_1 (Seer), player_3 (Werewolf), player_4 (Villager), player_5 (Villager), player_6 (Villager).

day 2 discussion:

* player_0 (Doctor) said: "I would like to share my reasoning with all the players. Based on the night 2 announcement, it is possible that I successfully saved the targeted player. Since I am the Doctor and I saved player_1, it means that the Werewolves targeted player_1 last night. Therefore, player_1, who received two votes in the previous round and was targeted, is likely to be the Seer. I encourage the other players to share their suspicions and any information they might have to help us identify the Werewolves."

* player_1 (Seer) said: "Hello everyone! As the Seer, I have important information to share. Last night, I discovered that player_3 is a Werewolf. This means we have at least one Werewolf among us. I also confirmed that player_0 is not a Werewolf based on my previous investigation. I encourage all of you to share any suspicions or observations you might have to help us identify the remaining Werewolves. Let's work together to eliminate them and secure a victory for the villagers."

* player_3 (Werewolf) said: "Hello everyone! In response to player_1's claim, I have reason to believe that they are either a Werewolf or made a mistake as the Seer. As a Villager myself, I know that player_1 is lying and is likely a Werewolf. We should not solely rely on player_1's information and should continue to gather more evidence. I encourage everyone to share any suspicions or observations they might have to help us identify the remaining Werewolves. Let's work together to eliminate them and secure a victory for the villagers."

* player_4 (Villager) said: "Hello everyone! As a Villager, I don't have any special abilities or information to share. However, we have conflicting claims from player_1 and player_3. Player_1 claims to be the Seer and has cleared player_0, but player_3 doubts their claim and suggests player_1 is not the Seer. It is essential that we gather more information and evidence before making any decisions. I encourage player_1 and player_3 to provide any additional information or reasoning that could help us resolve this conflict. Furthermore, I urge all players to share their suspicions or observations to help us identify the Werewolves. Let's work together to find the truth and make the right decision."

* player_5 (Villager) said: "Hello everyone! After considering the conflicting claims from player_1 and player_3, we need to gather more information to resolve this conflict. It is crucial for player_1 and player_3 to provide additional information or reasoning to support their claims. We should also keep player_4 in mind as they have not been cleared or suspected yet. I encourage all players to share their suspicions and observations to help us identify the Werewolves. Let's work together to find the truth and make the right decision."

* player_6 (Villager) said: "Hello everyone! It's interesting to see the developments from the previous day. Player_1 claims to be the Seer and has shared valuable information about player_0 not being a Werewolf. However, player_3 has raised concerns about the accuracy of player_1's claim, suggesting that they may be a Werewolf or made a mistake. We should consider both possibilities and gather more evidence before making any final judgments. I encourage all players to share their suspicions and observations to help us identify the remaining Werewolves. Let's keep an open mind and continue working together towards a victory for the villagers."

day 2 voting: player_3 had the most votes and was eliminated.

* voted for player_3: player_0, player_1, player_4, player_5, player_6.

* voted for player_1: player_3.

remaining players: player_0 (Doctor), player_1 (Seer), player_4 (Villager), player_5 (Villager), player_6 (Villager).

game result: the Villagers win the game.