



Jye Sawtell-Rickson

m1361019@cgu.edu.tw

Chang Gung University

April 6, 2026

Werewolves and LLMs

*Exploring Large Language Models for Communication
Games: An Empirical Study on Werewolf*

- 1 Introduction
- 2 Approach
- 3 Results
- 4 Discussion

- A communication game (e.g. Poker, Werewolf) is a multiplayer social game where players use speech, deception, deduction, or persuasion to achieve hidden or opposing goals.
- Communication games are good models for many real-world scenarios (e.g. negotiation).
- Many are language based and feature multiple agents.
- Paper studies application of LLMs to a specific game.

- › Framework for communication with frozen LLMs.
 - ›› Improves upon traditional methods with fixed communication structures.
 - ›› No training costs, reusable.
 - ›› **New**: Retrieve and reflect method to deal with long context, non-parametric.
- › Empirical study of LLMs in Werewolf.

Problem Statement

werewolf **villager** **witch** **guard** **seer**

1 Good morning everyone. It seems like we had a peaceful night. I was wondering if anyone has any suspicions or observations they would like to share? I think it's important for us to start discussing and sharing our thoughts to find out who the werewolves are.

2 Good morning everyone. As a villager, I don't have any clear suspicions yet, but I'm interested in hearing what others have to say. Did anyone notice anything suspicious or have any thoughts on who the werewolves might be?

3 Hi, I agree with Player 1 and Player 4 that we should exchange information actively. Based on my observation, I think Player 5 is a little suspicious.

4 Well, I agree with Player 1, 3 and Player 4. Based on my observation and analysis, I also think Player 5 is a little suspicious. But we might not vote to kill him casually.

5 I have to defend myself because I am innocent. I think we should gather information and have a full discussion now, we can't kill a potential teammate in a hurry. On the contrary, I think Player 3 is a bit aggressive, and I suspect he is one of the werewolf.

6 I observed that Player 5 was killed last night. As a witch, I used my antidote to save him. Therefore, Player 5 may not be a werewolf, but he was killed by the werewolf last night.

7 I agree with Player 1, 2 and 6. I think we have to actively gather information and seriously speculate about who the werewolves are. We should protect ourselves as much as possible and avoid being killed by the werewolves. I will use my special ability wisely at night.

m1361019@cgu.edu.tw

Figure: Example round in Werewolf. Agents debate with each other over their identities.

- › How to fit everything in small context?
- › Context includes: game rules, recent messages and
 - › **Informative messages**: rule-matching, heuristic-based.
 - › **Reflections**: BERT retrieval model for key Q&As with LLM reflections.
 - › **Experience-based suggestions**: build up successes and losses, use BERT retrieval model for relevant experiences, based on reflections.

- Non-parametric learning mechanism.
- Build an experience pool: reflection, response and score tuples (state, action, reward).
- Retrieve a subset based on reflection similarity (BERT), generate a suggestion.

$$E_{sub} = \{(R_l, G_l, s_l) | \cos(f(R_i^t), f(R_l)) > \epsilon\}$$

- "There is one bad experience G_0 and also a set of experience $\{G_1, \dots, G_n\}$ that may consist of good ones, find the difference between them and identify the good ones from the experience set."

Emergent Multi-Agent Behaviour

- Agents display various interesting multi-agent behaviours without any direct training.

Trust

Agents naturally establish trust relationships, long historical experiences help.

Confrontation

Suggestions to vote out lead to disagreement and directly oppositional actions.

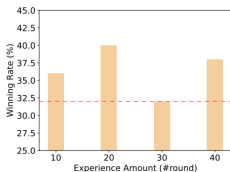
Camouflage

Agents lie, making up events, pretending to be other roles.

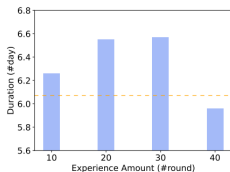
Leadership

Call for actions of other agents, forcing them to do things they don't want.

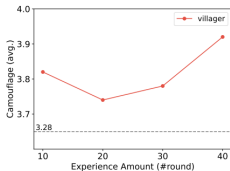
- Learning from experience generally increases performance for villagers, but can be limiting for werewolves. Higher win rates / duration, and more camouflaging behaviours.



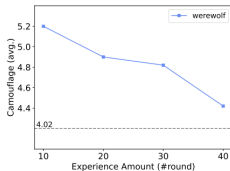
(a)



(b)



(c)



(d)

- No direct comparison to previous methods, or strong ablation study.
- Results potentially cherry-picked, no mention of other three behaviours.
- No good control for the adversarial nature - stronger opponents (self) can lead to lower win rates.
- Multi-game reasoning sort of achieved with experiences, but maybe could build a better world model.
- Hallucinations still an issue, even incorrect responses 5% of the time.
- Would be good to apply method to at least one other game to show it's scalable.

- › LLMs are a powerful multi-agent communicator out of the box, capable of displaying various strategic behaviours.
- › The paper has many possible improvement points around scalability and comparisons to SOTA.
- › Question on how to best score agents in these adversarial games, how to integrate world models.



Questions?