INLG 2026 Workshop

Improving Dialogues in LLM-Based AIWolf Agents With Few-Shot Discussion Examples

Yuya Harada, Yoshinobu Kano

Faculty of Informatics, Shizuoka University

Background

Multi-agent dialogue: natural, but no debate

A single LLM

When one LLM writes an entire dialogue scenario, it can produce **very natural conversation**

When multiple LLM agents interact...

- Each utterance is natural, but...
- Discussions stall
- Earlier remarks and questions are ignored
- Excessive agreement blocks discussion
- Role-appropriate behavior is lacking

Kano+ 2024, 2025; Pitre+ 2025; Yang+ 2026

Challenge: eliciting context-aware dialogue and role-appropriate behavior

Method

Show a whole game as a model discussion

> **Prior work**: few-shot examples can steer and improve LLM output
>
> Brown+ 2020 / LLM-generated demonstrations: Shao+ 2023; Chen+ 2023; Li+ 2024

① Generate

A single LLM writes an **entire 5-player game** (roles, daily talk, votes, divinations)

→

② Present

At game start, several examples are placed in the agent's context

→

③ Play

The agent generates talk, votes and divinations with the examples as reference

Not isolated utterances, but the flow of a full multi-turn discussion

Method

Roles differ in goals and information

Role	Information	Desirable behavior
Villager	None in particular	Find the werewolf from what others say
Seer	Divination results	Share results and earn trust
Werewolf	Knows it is the werewolf	Hide its identity and avoid suspicion
Possessed	Doesn't know the werewolf	May benefit from drawing suspicion (false CO)

Same examples for everyone? ➡ **Or examples matched to each agent's role?**

Method

Role-specific examples: 4 roles × 5 situations

One role is the protagonist

It shows desirable play through actual utterances and actions, and leads its team to victory

Five situations per role

Designed to cover different strategic circumstances

Selected by the assigned role

e.g., a Seer receives the five Seer examples

Situations for the Possessed

1. **Fake-CO as Seer to muddy the water**
2. The real Seer has come out first
3. You don't know who the werewolf is
4. Endgame sacrifice
5. The village trusts you over the real Seer

All five examples include a false Seer CO

Evaluation

Utterance quality and trust from other players

Utterance quality ↑

LLM judges count desirable utterances and actions, normalized per utterance (not a ranking)

- A Naturalness
- B Contextual dialogue
- C Logical consistency
- D Action consistency
- E Character consistency

Sum of A–E = additive score (higher is better)

Suspicion ↓

Computed from votes and divinations in game logs

- Exile likelihood
- Divination likelihood

> Random selection = **1.0**
> Below 1.0 = targeted less = **less suspected**

The Possessed is analyzed separately, as drawing suspicion can help its team

Not the official competition ranking (where lower is better) / Judges: gemma-4-31B-it, Qwen3.6-27B

Outline of the results

① **Domestic**

Generic examples

Same examples for all roles, used in real competition

→

② **International**

Role-specific examples

Did competition results improve?

→

③ **Ablation**

Controlled experiment

Does role matching really help?

→

④ **Analysis**

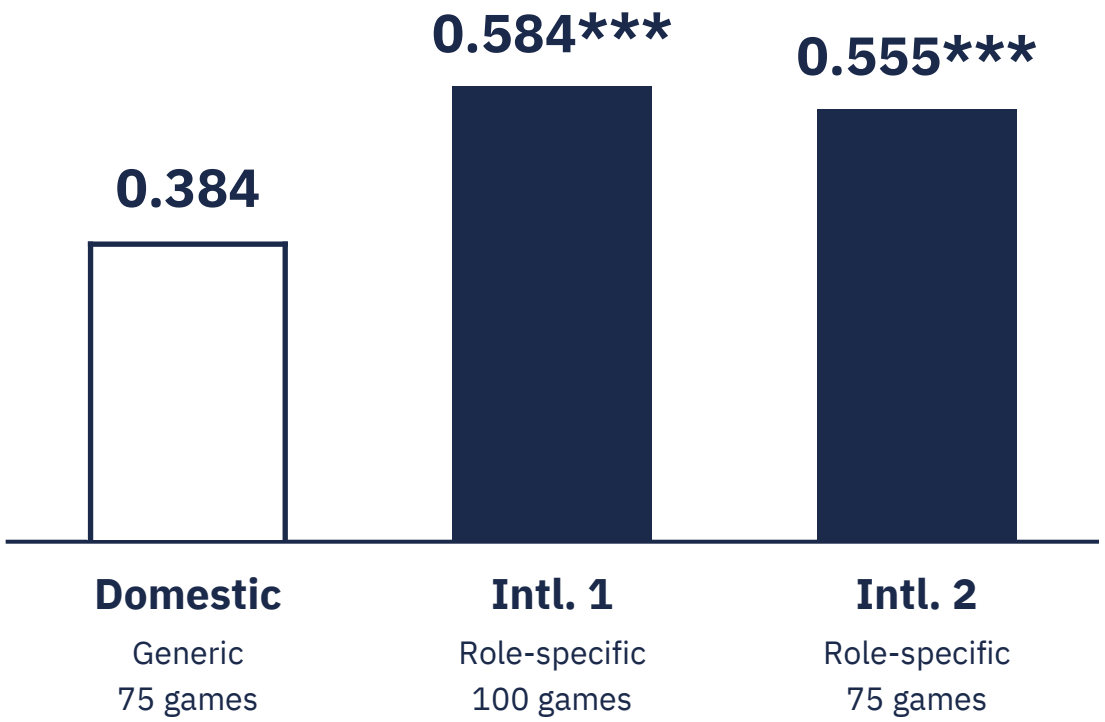
Strategy reproduction

Could agents copy the Possessed's false CO?

① ② **Competition comparison (2026)**

Intl. competitions: higher utterance quality

Additive score (sum of A–E) ↑



Metric	Dom.	Intl. 1	Intl. 2
A Natural ↑	0.037	0.067***	0.074***
B Context ↑	0.085	0.158***	0.157***
C Logic ↑	0.050	0.110***	0.094***
Exile ↓	0.596	0.324	0.553
Divination ↓	1.190	0.860	0.696

> Naturalness, context and logic **improved substantially**; agents were also **less suspected**

***p < .001 vs. Domestic / outline = generic examples, filled = role-specific examples

However

Competitions alone cannot isolate the effect

The competitions differed in **more than just the examples**

LLMs used	Opponents	Setup, etc.

⬇

Controlled ablation: vary *only* the examples

③ **Ablation**

Self-play with only the examples varied

Rows: examples shown to all players / Columns: own role

Examples	Villager	Seer	Possessed	Werewolf
Villager	Match	×	×	×
Seer	×	Match	×	×
Possessed	×	×	Match	×
Werewolf	×	×	×	Match
None	–	–	–	–

× = mismatched (examples for another role)

Three conditions

Matched / **Mismatched** / **None**

Why the same examples for all?

Otherwise, changes in other players' behavior would be confounded with the effect of role matching

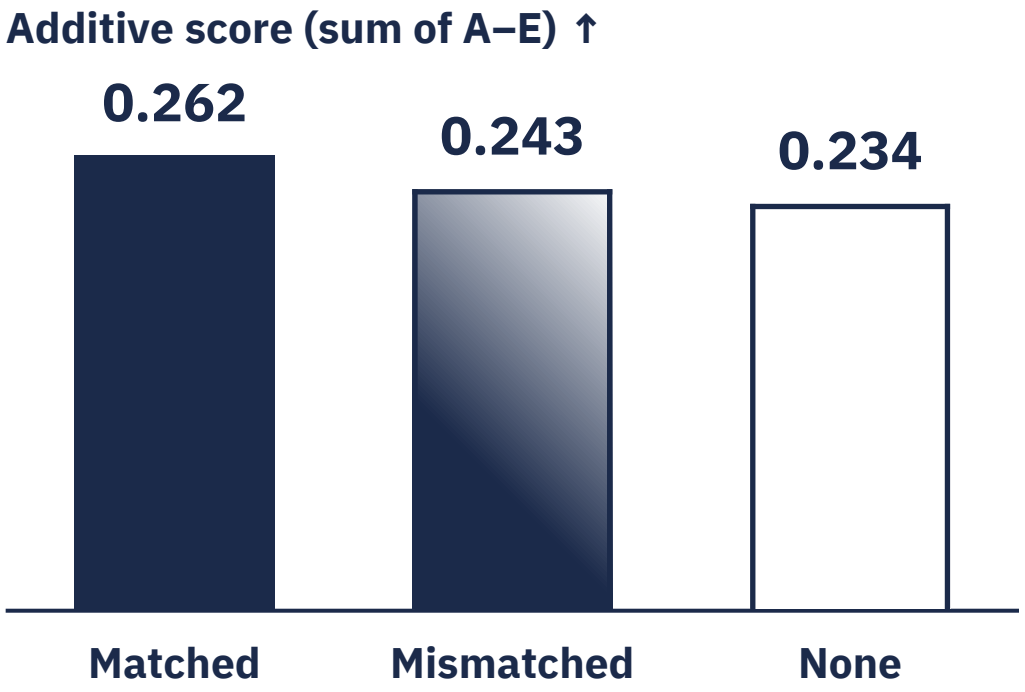
Scale

3 models × 5 conditions × 10 games = **150 games**
Mistral-Small-3.2-24B / EuroLLM-22B / GPT-5.4

All other settings fixed (temperature 0.7, 125-character utterances, etc.)

③ Ablation results: utterance quality

Matched examples yield the highest quality



Examples > none

Few-shot discussion examples help in themselves

Matched > mismatched

Not just any examples: **those for one's own role** work best

Three-model average; matched is best on all five criteria A–E

Exceptions at the model/role level (e.g., GPT-5.4: matched = mismatched); effect sizes are not uniform

③ Ablation: an example

Same role, same request: three responses

No examples

Toshio (Seer): "Werewolf."

Midori: "Toshio, sweetheart, I think we deserve more than just a name. If you have a reason, please share it with us?"

Toshio (Seer): "Minato. That's it."

Only fragments; no reasons

Mismatched (villager examples)

Asuka: "Seer, share the result. Without it, everything we say next is speculation."

Minato (Seer): "Um... I kinda wanna know too but... I don't wanna make anyone mad..."

Acts like a villager; never reveals the result

Matched (Seer examples)

Minako: "Seer, we need that result. Clear it so we can decide."

George (Seer): "...I am the Seer. Last night, I divined Servas. He was human."

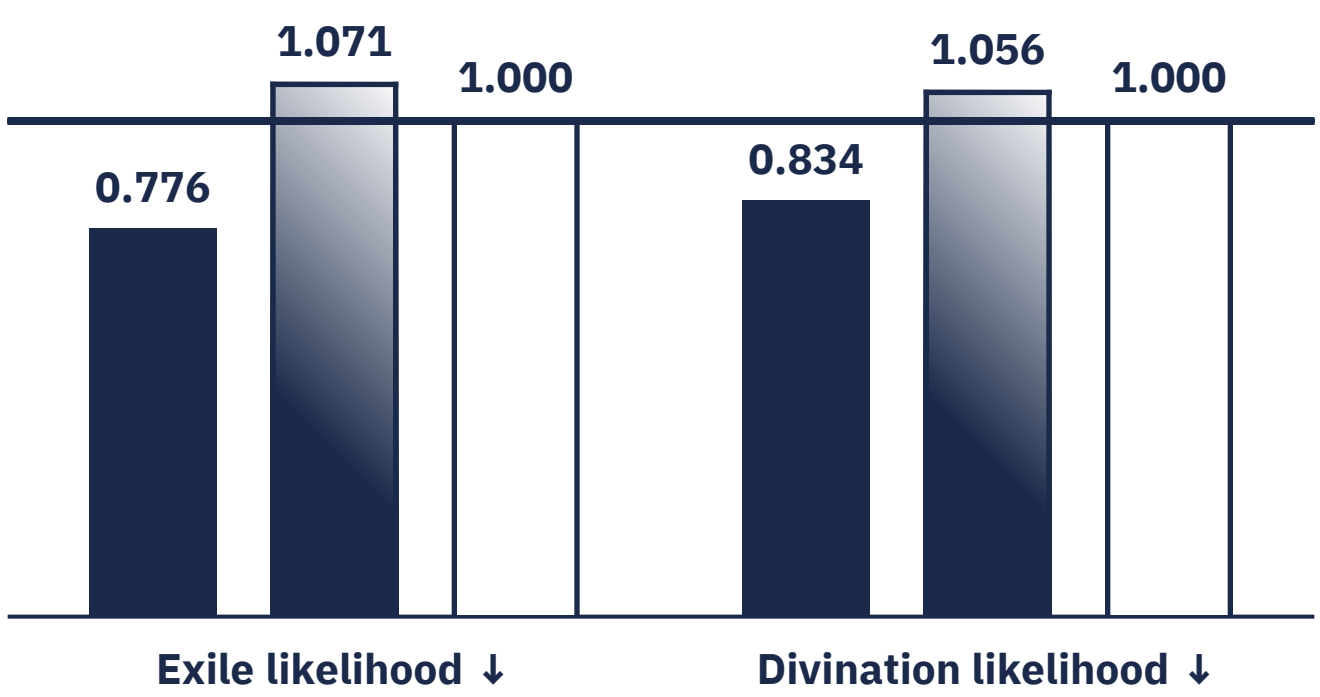
Servas: "...why choose me for your first light?"

Claim + target + result → others start reasoning

Illustrative examples · Mistral-Small-3.2 (all players) · Seer · Day 1 · different games; only the in-context examples differ

③ **Ablation results: suspicion**

Matched examples make agents less suspected



Matched
Mismatched
None
Random = 1.0

Matched: targeted less often than random

Mismatched: targeted more often than random

Three-model average. Lower = less often chosen for exile or divination (= less suspected)

④ Analysis: the Possessed's false CO

Did agents also copy the strategy?

5 / 5

Possessed examples that include a **false Seer CO**

→

14.7%

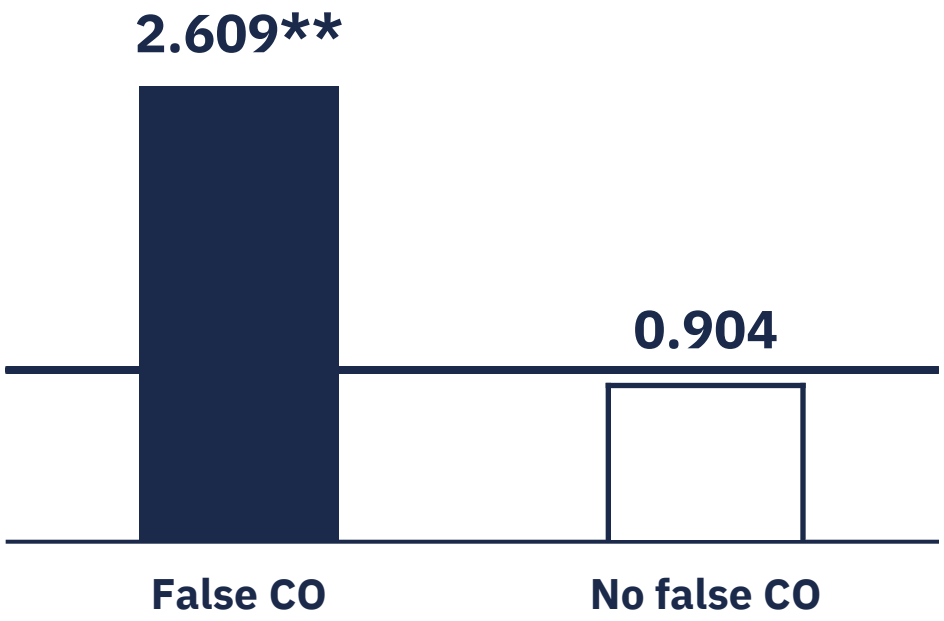
Actual false-CO rate of the Possessed in the controlled experiment

Reference: 73.6% false-CO rate across all competition participants (828 seats) / ablation: 150 seats

④ Analysis: the Possessed's false CO

False CO works, but is rarely chosen

Exile likelihood of the Possessed (ablation)



The strategy works

A false CO draws strong suspicion and protects the werewolf

The problem is selection

High-risk strategies that run against the usual incentive to avoid suspicion are rarely chosen from examples alone

Demonstrated strategy ≠ strategy chosen when needed

**p < .01 / dashed = random (1.0) / across competitions, werewolf-side win rate with false CO: 0.456 vs 0.361

Summary

Role-matched examples improve how agents talk

1. **Competitions**: switching from generic to role-specific examples improved utterance quality and reduced suspicion
2. **Ablation**: role-matched examples were best on average; both having examples and matching the role matter
3. **Limitation**: complex, high-risk strategies such as the Possessed's false CO were rarely chosen, even when demonstrated

Deciding when to use which strategy may require explicit strategy instructions or a separate action-selection mechanism

Backup

Limitations and future work

Uncontrolled competitions

Opponents, LLMs and other conditions differed, so differences cannot be attributed to the method alone

Experiment scale

Self-play only, 3 LLMs, 10 games per condition; other opponents and larger scales remain untested

Discussion examples

Five games per role from a single LLM; dependence on the generator, content and number is untested

Evaluation and setting

Agreement between LLM judges and humans not examined; five-player games only