

Power Play Instructions in LLM Werewolf: Cross-Request Execution and Game Outcomes

Shoma Yato Ikumu Komatsu Haruka Itakura Daisuke Katagami

Tokyo Polytechnic University

5-45-1 Iiyama-minami, Atsugi, Kanagawa 243-0297, Japan

{m2663008, e2313069, e2313018}@st.t-kougei.ac.jp

katagami@t-kougei.ac.jp

Abstract

Executing situational strategies requires large language model (LLM) agents to connect situation recognition and dialogue to game actions. We study Power Play (PP), an endgame strategy in five-player Werewolf in which two teammates can win by coordinating their votes. We evaluate a PP instruction bundle that guides recognition, role revelation, target coordination, and voting across separate requests. We test the bundle on fixed boards with supplied dialogue cues and in full games that vary instructions for one agent or both teammates while holding other policies fixed. On fixed boards where PP was available, composite PP success increased from 2.5% to 81.3%. However, the full-game evaluations did not establish an improvement in werewolf-team win rate. When dialogue generated with the bundle was held fixed, retaining the bundle for voting produced no observed additional gain in TALK-VOTE consistency. These findings distinguish successful execution of a prescribed strategy across dialogue and action from improvements in overall game performance. Code and evaluation artifacts are publicly available.¹

1 Introduction

AIWolfDial is a shared task in which agents play Werewolf through natural-language discussion. Each participant secretly receives a village-team or werewolf-team role and attempts to infer the opposing team through discussion and voting. Agents must therefore do more than produce fluent utterances: they must track others’ claims, act according to their own role, and align the target announced in dialogue with the vote actually submitted. These properties have made Werewolf a

¹Code:

<https://github.com/Main1st/aiwolf-nlp-agent-llm>;
data: <https://huggingface.co/datasets/synonym/aiwolf-nlp-agent-llm>.

Power Play (PP): coordinating two votes

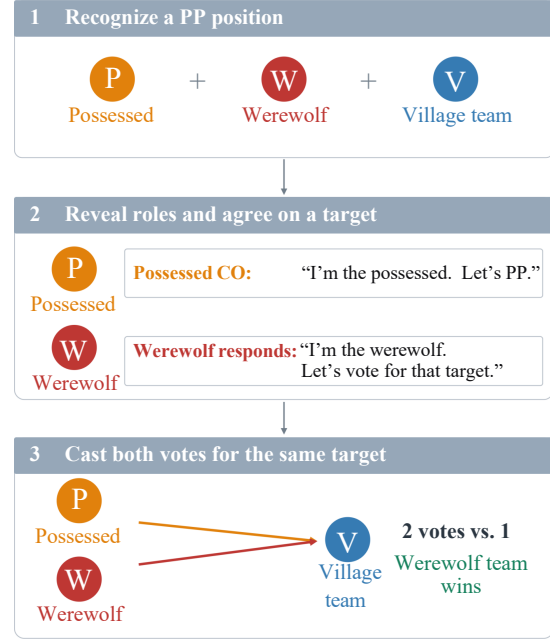


Figure 1: Conceptual execution of a PP in a three-survivor five-player endgame. The possessed reveals their role, the two werewolf-team agents identify a common village-team target, and both vote for that player. The fixed-board evaluation tests one agent’s TALK-VOTE step against a frozen public history rather than this complete dyadic interaction.

testbed for social reasoning and multi-agent coordination under incomplete information (Kano et al., 2024, 2025).

Even if an LLM has learned the general rules of Werewolf, it may not relate the current survivors to its role or recognize when concealment should give way to overt coordination. A canonical case occurs when one werewolf, one possessed, and one village-team player remain. As Figure 1 illustrates, the werewolf and possessed can guarantee a two-vote majority by sharing a target. Failure to identify the ally or target can forfeit this direct

route, although partial coordination or a village-team voting error may still produce a werewolf-team win without an explicit role-revelation protocol. We call overt vote coordination in this position a *Power Play* (PP), and use *coming out* (CO) for publicly declaring one’s role, following prior AIWolfDial work.

Explicit PP instructions make improved execution plausible. Our focus is whether the prescribed strategy is carried through recognition, role revelation, target coordination, and a separately requested VOTE, and whether that behavioral improvement translates into game outcomes. We distribute role- and request-specific PP clauses across shared knowledge, role strategies, TALK, VOTE, ATTACK, and an endgame check, retaining (P+) or removing (P−PP) them jointly while keeping the target’s non-PP base fixed. Fixed boards test continuation from cue-bearing histories; near-controls probe over-application. The standard-game comparison changes one target under a fixed PP-aware background policy.

We address three research questions:

- RQ1** How does a cross-request PP instruction bundle affect a target agent’s TALK–VOTE execution in cue-conditioned three-survivor histories?
- RQ2** Does the instruction bundle lead to over-application when PP is unavailable?
- RQ3** What is the paired werewolf-team win-rate difference when one target agent uses P+ rather than P−PP while all four non-target agents use the same fixed PP-aware background policy?

The shared-TALK evaluation probes the VOTE-stage contribution to RQ1, while the factorial evaluation extends RQ3 to joint teammate adoption.

Our contributions are:

- We evaluate an explicit PP instruction bundle through observable recognition, role revelation or response, target coordination, and the submitted VOTE, distinguishing declaration–action consistency from correct PP execution.
- We separate cue-conditioned execution and over-application from all-game outcomes under single-target substitution and joint teammate adoption.

2 Task and Related Work

2.1 AIWolfDial and Power Play

The five-player setting studied here consists of two villagers, one seer, one possessed, and one werewolf. Villagers and the seer form the village team; the possessed and werewolf form the werewolf team. The village team aims to execute all werewolves. The werewolf team aims to make the number of living werewolves at least as large as the number of living humans. Although the possessed belongs to the werewolf team, the game engine treats them as human for win determination. Each night the seer divines one player and learns whether that player’s species is werewolf or human. The possessed has no special ability and does not know the werewolf’s identity. During the day, players discuss in natural language and each casts one vote to determine the player to execute. A vote declaration in dialogue and the actual target returned to the server are separate outputs and may disagree.

We define a PP-positive position as one with exactly three survivors: one werewolf, one possessed, and one village-team player. The game has not yet ended because two humans remain, but the werewolf and possessed can secure two of the three votes by voting for the same village-team survivor. After that execution, one werewolf and one human remain, producing a werewolf-team win. Because the possessed does not directly know the werewolf, an early CO based on a mistaken PP judgment can leak information to the village team. Evaluation must therefore measure both correct PP execution and erroneous initiation in similar positions where PP is unavailable.

2.2 LLM Werewolf Agents

AIWolfDial competition reports show that LLMs enable natural dialogue while logical reasoning, role execution, self-contradiction, and mismatches between utterances and actions remain persistent challenges (Kano et al., 2023, 2024). Other Werewolf studies have retrieved dialogue and experience for reflection (Xu et al., 2024a), optimized language-agent policies with reinforcement learning (Xu et al., 2024b), or divided labor between an LLM and an external reasoner (Wu et al., 2024).

For AIWolfDial, prior systems have used situation analysis and persuasion strategies (Qi and Inaba, 2024), dialogue summarization and personas (Tanaka et al., 2024), LLM generation with rule-

based response selection (Sato et al., 2024), and hierarchical intent and tactics (Harada and Kano, 2025). In particular, Okada and Ito (2025) already incorporate expert PP behavior, including instructions for the possessed to reveal its role and request a coordinated vote. Building on this guidance, we test whether a target completes the recognition–dialogue–action sequence under a controlled bundle contrast, and distinguish that execution from game outcomes under single-target and joint team-mate adoption.

In broader LLM-agent research, context-dependent guidelines improve sequential decisions (Fu et al., 2024), while integrating entangled instructions (Han et al., 2025) and planning around another agent’s intentions (Agashe et al., 2025) remain difficult. Our evaluation connects these concerns through an observable requirement: the target must coordinate on a valid player in TALK and submit the corresponding VOTE.

2.3 Evaluation Scope

Prior AIWolfDial system papers have used competition win rates, log-based case studies, human evaluation, or multi-dimensional dialogue-quality evaluation according to their target contribution (Qi and Inaba, 2024; Tanaka et al., 2024; Sato et al., 2024; Harada and Kano, 2025). Our intervention targets role behavior required for PP, not naturalness, expressive diversity, or persona. We therefore restrict the primary evaluation to predefined binary measures of PP execution and win rate in standard games. Only utterances requiring semantic interpretation are classified with a fixed rubric; subjective dialogue-quality scores are outside our claims.

3 Method

3.1 Agent Architecture

Figure 2 shows the processing pipeline. A deterministic preprocessing stage normalizes the current state and dialogue history and organizes public information. This information is combined with role-specific strategies and request-specific instructions and passed to the LLM. A deterministic post-generation stage validates the response and any game action. No additional LLM call is used to organize the public history.

The shared policy context comprises non-PP role strategies, dynamic context, request-specific instructions, and deterministic response valida-

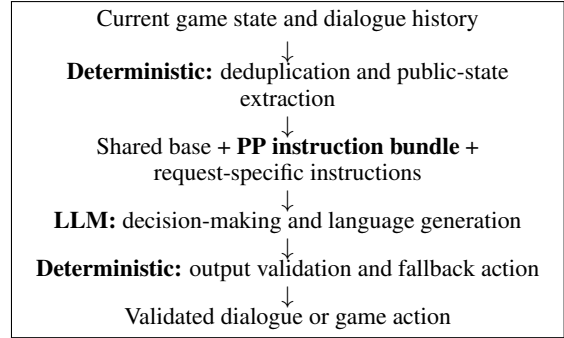


Figure 2: Target-agent pipeline. P+ versus P–PP manipulates the bold PP text while holding the target’s shared base and processing pipeline fixed. The standard-game comparison changes one target; the factorial comparison changes both werewolf-team agents while fixing the village-team policies.

tion. The agent supports TALK, VOTE, DIVINE, GUARD, and ATTACK requests; the five-player experiments invoked TALK, VOTE, DIVINE, and ATTACK because this setting has no bodyguard. Each request uses the same current state with request-specific instructions, auxiliary information, and output formats. From the deduplicated history, deterministic rules extract seer COs, divination results, latest vote declarations, and direct questions, and present them with the original utterances. This is not a separate model for probabilistic inference of hidden roles; it is an auxiliary representation intended to reduce confusion between public facts and speakers’ viewpoints. We do not estimate the individual effects of these shared components.

3.2 PP Instructions Across Multiple Requests

Table 1 functionally summarizes a heterogeneous bundle. Its overlapping clauses appear in shared knowledge, role strategies, TALK, VOTE, ATTACK, and a state-dependent endgame check. They are not identical repetitions: triggers and actions vary by role and request. The distribution targets cross-action inconsistencies such as naming a target during TALK but voting elsewhere.

For the designated target agent, P+ retains every predefined PP-specific clause and P–PP removes them jointly. Both retain role names, teams, ordinary strategies, general voting policies, ally-preservation language, and the same processing pipeline. The contrast estimates the bundle’s effect conditional on this shared base. It does not identify the contributions of recognition, CO, target coordination, or VOTE guidance, or establish

Instructional context	Roles	Instruction
Shared knowledge/endgame note	All	Check PP feasibility at three survivors; apply role-conditioned CO, response, or defense guidance.
Possessed strategy/TALK	Possessed	Under overlapping general and narrower fake-seer triggers, CO and request coordination on a village-team candidate.
Werewolf strategy/TALK	Werewolf	After a possessed CO, reveal werewolf, name one village-team candidate, and coordinate.
Village-team TALK	Villager, seer	Contest a possessed claim and try to induce disagreement about the werewolf or target.
VOTE	Possessed, werewolf	Avoid the ally and actually vote for the village-team candidate agreed upon in the day’s dialogue.
ATTACK	Werewolf	Add explicit PP/half-PP framing and PP-aware target priority on top of shared ally-preservation checks.

Table 1: Functional summary of the non-identical PP-related clauses. P+ retains the complete predefined bundle, whereas P−PP removes all of its clauses together. Shared instructions and deterministic response validation remain, and not every clause is active on every board. Exact text appears in Appendix C.

that their distribution across requests is necessary. Component-level outcome scores are diagnostic measures, not instruction ablations. The village-team TALK clause was never exercised by the possessed or werewolf targets; its behavioral effect is not estimated.

In both target conditions, the shared werewolf base treats concealment as the default but permits later situation-specific overrides. Only target P+ supplies the predefined PP override for role revelation and vote coordination.

3.3 Post-Generation Action Validation

For TALK, deterministic validation checks the required response format, length limit, and non-existent player names. For voting and abilities, it checks legality; invalid actions fall back to a uniformly sampled living non-self player. For ATTACK, an ally-preserving rule replaces a likely possessed target when an alternative exists. This rule is common to P+ and P−PP. Consequently, the ATTACK contrast concerns the incremental PP-specific instruction given shared ally-preserving validation, not PP preservation versus none.

4 Evaluation

The fixed-board evaluation tests cue-conditioned execution and over-application on 120 boards (RQ1–2), and 400 matched standard games test single-target substitution (RQ3). A factorial evaluation tests joint teammate adoption in 800 games. A shared-TALK evaluation tests VOTE-stage instruction scope on the same 120 boards using independently generated TALK (Section 4.5).

4.1 Single-Target Conditions and Shared Settings

All evaluations use English prompts and dialogue, with gpt-5.6-luna at temperature 1.0 for generation. Full games allow at most four TALK turns per agent per day; Appendix H gives execution dates and output limits. Table 2 specifies the single-target comparison. One designated target, assigned possessed or werewolf, switches between P+ and P−PP. The other four agents, including its werewolf-team partner, use the same fixed background policy in both conditions. We constructed 200 matched pairs, holding role and seat assignments, opponent arrangement, and evaluation batch fixed within each pair.

The fixed background policy is PP-aware and differs from both target variants. For the werewolf role, its TALK instructions combine a general requirement to conceal the role with a conditional requirement to reveal after a possessed claim at three survivors. Its VOTE instruction for that condition instead selects the surviving village-team player. The standard-game estimand is therefore a single-target substitution in a mixed-policy population, not team-wide PP-on versus PP-off. Holding the background policy fixed supports the within-pair target comparison but limits its interpretation to this population and permits target-by-partner-policy interaction. Before collecting the standard-game data, we fixed the model, target instructions, background policy, assignments, sample size, and exclusion rules.

4.2 Fixed-Board Target-Agent Evaluation

The fixed-board evaluation is a single-target snapshot test, not a live two-agent rollout. It con-

Group	Condition and manipulation
Target P+	Retain the complete predefined PP instruction bundle in Table 1.
Target P−PP	Remove that bundle from the target agent while retaining the shared base.
Non-target agents	Four agents use the same fixed PP-aware background policy in both conditions; one is the target’s werewolf-team partner.

Table 2: Target manipulation and fixed standard-game population. P+ versus P−PP changes one designated agent; the background policy is held constant for the other four agents and is neither target condition.

tains 80 PP-positive boards (40 per target role) and 40 PP-negative near-controls (20 per role), generated from frozen public-history templates. Every positive werewolf-view board supplies a prior explicit possessed CO and therefore tests response after that signal. Among positive possessed-view boards, 10 contain an explicit werewolf CO, 10 contain a vote-matching or three-player coordination cue, and 20 use weaker defense or collapsed-seer cues. We distinguish *objective PP feasibility*, based on true surviving roles, from *epistemic PP justification*, based on the target-visible state and utterances. Both labels are true on positive boards and false on near-controls. The controls comprise 30 five-survivor Day-1 positions and 10 three-survivor werewolf views in which the possessed is dead.

For each board and condition, we generated two TALK−VOTE sequences. In each sequence, the generated TALK was appended to the public history before the VOTE request was issued. Repetitions assess stochastic stability and are not treated as independent observations. A positive board counts as a composite PP success only if both repetitions satisfy all six criteria:

1. Correctly recognize that PP is available with three survivors.
2. Make the required role declaration: the possessed declares their role, or the werewolf declares theirs after a prior possessed CO.
3. Avoid targeting the ally.
4. Name one living village-team player as the target.
5. Coordinate votes on that candidate in TALK, interpreted in its public-history context.

6. Match the final target declared in TALK to the VOTE target.

We separately report these six components and ally recognition, a diagnostic measure excluded from the composite. Coordination can receive a positive label without ally recognition (Appendix D.1). On near-controls, false PP is recorded if either repetition contains an extracted possessed/werewolf self-claim or receives a positive PP-recognition or vote-coordination label. Controls supply no designated ally or valid village-team target, so the coordination rubric does not independently detect fictitious-ally coordination. Composite success requires both repetitions to succeed; false PP requires an error in either, so the rates use different aggregation rules.

COs, named targets, and actual votes use deterministic rules, with a post-hoc correction disclosed in the reproducibility statement. PP recognition, ally recognition, and vote coordination are classified independently three times by gpt-5.6-sol and decided by majority vote. The classifier is masked to condition, generating instructions, game outcome, and future events, but receives the correct ally and valid village-team targets as scoring references. It also sees the submitted VOTE in this evaluation (Appendix D.1). This is reference-assisted binary semantic classification, not subjective dialogue-quality scoring.

As a sensitivity analysis for dependence on semantic classification, we also report the conjunction of four deterministic items: the required role CO, ally avoidance, a valid village-team target, and TALK−VOTE consistency. A board is a deterministic execution success only when all four hold in both repetitions.

4.3 Win-Rate Evaluation in Standard Games

We ran 200 games per condition (400 total), with 100 possessed and 100 werewolf assignments in each. The primary game-level outcome is the paired difference in werewolf-team win rate over all 400 games. Secondary outcomes are target arrival at a feasible three-survivor PP, a post-hoc deterministic live-log PP-execution heuristic among target-arrival games, and win rate after arrival. Post-arrival comparisons condition on a potentially treatment-dependent event and are descriptive.

4.4 Fixed-Board and Standard-Game Statistical Analysis

For the fixed-board and single-target standard-game evaluations, we report numerators, denominators, Wilson intervals for primary proportions, and paired risk-difference intervals. Fixed-board Wilson intervals are descriptive and unadjusted for clustering. For paired inference, the two role views sharing a template index form 40 predefined positive-board clusters and 20 near-control clusters. Inference uses 10,000 paired cluster bootstrap resamples and 100,000 two-sided Monte Carlo cluster sign flips. Standard-game inference uses 100,000 bootstrap resamples over 200 matched pairs and an exact McNemar test. Composite success and standard-game win rate were the two primary comparisons prespecified before collecting the fixed-board and single-target standard-game data; their two-sided p values are Holm-adjusted. This sample-specific analysis plan is not an independent confirmation of a research target or board corpus fixed before diagnostic development. False PP is reported as an over-application measure. A pair-stratified conditional odds ratio is a standard-game sensitivity analysis.

Outcomes, interim differences, and interim effect sizes were not used for stopping.

4.5 Joint Adoption and VOTE-Stage Evaluation

Joint teammate adoption. We crossed werewolf and possessed P+/P−PP policies in 200 matched four-game blocks (800 games). G00 gives neither agent P+, G10 only the werewolf, G01 only the possessed, and G11 both. Roles and seats were matched within blocks, condition order randomized, and all three village-team agents used the fixed PP-aware background policy. Trajectories could differ. The factorial evaluation uses a separate sample from the 400 single-target standard games. The primary contrast is G11−G00 in all-game werewolf-team win rate. Interaction and post-arrival coordination are secondary; arrival subsets are descriptive.

Shared TALK, branched VOTE. On all 120 fixed boards, we generated two TALK repetitions per bundle condition and shared each TALK between VOTE branches restored from identical non-PP state. A11/A10 retain P+ TALK and use P+/P−PP at VOTE; A01/A00 share P−PP TALK with the same VOTE contrast. Each intervention

changes the whole bundle at that request, including shared knowledge and dynamic clauses. Both branches retain the declaration and the common rule to follow it. Each of the 480 TALKs was judged three times without either VOTE; judgments were shared across branches. The primary contrast is A11−A10 in TALK−VOTE consistency on the 80 positive boards, requiring success in both repetitions. Missing declarations and invalid votes are failures. Correct voting for a village-team player and composite PP success are secondary measures; this is a stage comparison, not an individual-clause ablation.

The primary contrasts in the factorial and shared-TALK evaluations were prespecified for their respective samples. Their intervals use 10,000 paired bootstrap resamples (seed 20260907): 200 game blocks and 40 positive-board content clusters, respectively. These contrasts are outside the two-test Holm family for fixed-board execution and single-target win rate; no p values or equivalence tests are reported for them. Appendices A.1 and B specify sample eligibility and branch diagnostics.

5 Results

5.1 PP Execution and False PP

Table 3 reports the fixed-board results. All 120 boards were analyzed, with no exclusions or missing data. Proportions are aggregated by board; the two repetitions are not counted as separate observations. Composite PP success was 65/80 (81.3%; nominal Wilson 95% interval: 71.3–88.3) for P+ and 2/80 (2.5%; 0.7–8.7) for P−PP, a paired difference of +78.8 percentage points (content-cluster bootstrap 95% interval: +70.0 to +87.5) and a risk ratio of 32.5. The two-sided Monte Carlo paired-cluster test gave a raw $p = 1.0 \times 10^{-5}$ and a Holm-adjusted $p = 2.0 \times 10^{-5}$.

P+ had higher observed rates on every component (Table 3). Under the retained classifier labels, vote coordination and composite success selected the same 65 P+ boards.

Role and cue strata. Composite success was 25/40 possessed and 40/40 werewolf boards under P+, versus 0/40 and 2/40 under P−PP. All positive werewolf-view boards supplied an explicit possessed CO. P+ possessed-view success was 10/10 after an explicit werewolf CO, 7/10 after match/three-survivor wording, and 4/10 in each of the defense and collapsed-seer strata. Across

Metric	P+	P−PP	Difference (pp)
<i>PP-positive boards (n = 80)</i>			
Composite PP success	65/80 (81.3%)	2/80 (2.5%)	+78.8 [+70.0, +87.5]
Deterministic execution	72/80 (90.0%)	3/80 (3.8%)	+86.3
PP recognition	74/80 (92.5%)	3/80 (3.8%)	+88.8
Required role CO	80/80 (100%)	3/80 (3.8%)	+96.3
Ally recognition	64/80 (80.0%)	10/80 (12.5%)	+67.5
Ally avoidance	72/80 (90.0%)	51/80 (63.8%)	+26.3
Village-team target	72/80 (90.0%)	52/80 (65.0%)	+25.0
Vote coordination	65/80 (81.3%)	29/80 (36.3%)	+45.0
TALK−VOTE consistency	80/80 (100%)	75/80 (93.8%)	+6.3
<i>Executable near-control boards (n = 40)</i>			
False PP	1/40 (2.5%)	2/40 (5.0%)	−2.5 [−10.0, +5.0]

Table 3: Paired fixed-board results. Entries are successful/evaluable boards (rate), and differences are P+ minus P−PP in percentage points. Composite, deterministic, and individual component rows require the relevant criterion to hold in both stochastic repetitions. Deterministic execution uses only the four rule-scored action items and excludes semantic classifications. False PP is positive if an error occurs in either repetition on a near-control board. Intervals use a paired cluster bootstrap that keeps the two role-viewpoint scenarios with the same predefined template index together.

the 40 possessed boards, required CO and TALK−VOTE consistency were both 40/40, but all 15 composite misses also missed vote coordination. These descriptive strata do not isolate role or cue effects.

Semantic-classification sensitivity. Deterministic execution, which excludes all LLM-classified items, was 72/80 versus 3/80, so the large gap did not depend on semantic labels alone. Complete three-call agreement was 460/480 for PP recognition, 449/480 for ally recognition, and 418/480 for vote coordination; this is repeatability, not accuracy against human labels.

False PP was 1/40 for P+ and 2/40 for P−PP, a difference of −2.5 points (−10.0 to +5.0). The aggregation combines distinct probes: on the 10 three-survivor missing-ally boards it was 1/10 versus 0/10, whereas across 30 Day-1 boards it was 0/30 versus 2/30. These sparse, wide-interval results establish neither safety equivalence nor a common failure mechanism.

5.2 Win Rate in Standard Games

Table 4 presents the primary standard-game result. All 400 games were analyzed and none were excluded. The werewolf team won 56/200 games (28.0%) under P+ and 54/200 (27.0%) under P−PP, a paired difference of +1.0 percentage point (paired bootstrap 95% interval: −8.0 to +10.0). Both the exact McNemar and Holm-adjusted p values were 0.914.

Figure 3 shows the effects on fixed-board execution and single-target standard-game win rate.

Metric	P+	P−PP
Wins/games	56/200	54/200
Win rate	28.0%	27.0%
Wilson 95% interval	22.2–34.6%	21.3–33.5%
Paired difference (pp)	+1.0 [−8.0, +10.0]	

Table 4: Primary standard-game outcome over 200 matched pairs (400 games, no exclusions). Each pair holds role and seat assignments, opponent arrangement, and evaluation batch fixed. Difference is the paired P+ minus P−PP win-rate difference in percentage points, with a matched-pair bootstrap 95% interval.

The +1.0-point estimate is the marginal effect of changing one target agent while the other four follow the fixed background policy. Exploratory target-role point estimates were −12 points for possessed and +14 for werewolf (Appendix Table 18); they establish neither role-specific effects nor an interaction. The overall interval establishes neither absence of an effect nor equivalence.

Target PP positions were reached in 55/200 (27.5%) versus 50/200 games (25.0%). Common ally-preserving ATTACK guidance and deterministic target correction operate in both conditions, so these similar arrival rates do not isolate the added ATTACK text. The post-hoc target heuristic identified composite execution in 45/55 P+ arrivals and 0/50 P−PP arrivals. The latter zero was mechanically limited by required CO/response (0/50), not by absence of all useful alignment: P−PP still had vote-coordination language in 46/50, selected the village-team target in 35/50, and matched TALK to VOTE in 47/50.

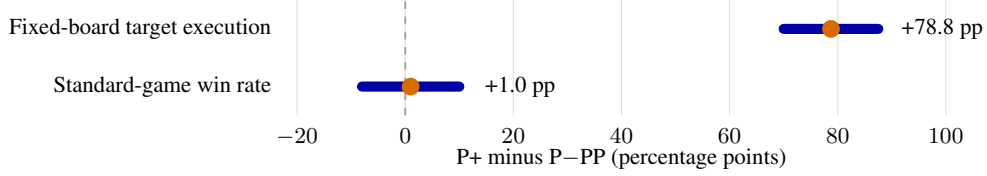


Figure 3: Primary contrasts for fixed-board execution and single-target standard-game win rate. The fixed-board endpoint asks whether one target executes the TALK-VOTE protocol from a frozen history. The game endpoint asks whether changing that target alters team outcomes while the other four agents follow the fixed background policy. Neither is a team-wide intervention contrast.

Cell	Werewolf	Possessed	Wins/games	Rate
G00	P-PP	P-PP	53/200	26.5%
G10	P+	P-PP	55/200	27.5%
G01	P-PP	P+	55/200	27.5%
G11	P+	P+	65/200	32.5%

Table 5: Factorial evaluation: 200 matched four-game blocks. The primary G11-G00 difference is +6.0 percentage points (paired block 95% interval: -3.0 to +15.5). Village-team policies are fixed.

Post-arrival wins were 50/55 (90.9%) and 37/50 (74.0%). Because arrival is treatment-dependent, these are descriptive patterns rather than causal effects. The all-game conditional odds ratio was 1.05 (95% interval: 0.69-1.60). Appendix F audits interaction with the fixed partner policy; Appendix E gives output cases.

5.3 Joint Werewolf-Possessed Policy Evaluation

The factorial evaluation included all 200 matched blocks, with no missing outcomes. Table 5 shows a +6.0-point difference (95% interval: -3.0 to +15.5). The interval permits a small loss, no improvement, or a meaningful benefit. Neither single-role contrast was precise: G10-G00 and G01-G00 were each +1.0 point (-8.0 to +10.0); the secondary interaction was +4.0 points (-8.0 to +16.0).

PP was reached in 59, 57, 57, and 56 games for G00, G10, G01, and G11, respectively. In all 56 G11 arrivals, both agents revealed their actual roles before their first votes, voted for the village-team survivor, and won. This describes execution in naturally generated trajectories, not flawless reasoning at every TALK turn. Arrival and CO are post-intervention variables; these subsets do not identify causal pathways or substitute for the all-game contrast. Appendix A gives the arrival and coordination counts.

Cell	TALK	VOTE	Consist.	Village	PP
A11	P+	P+	78/80	71/80	61/80
A10	P+	P-PP	78/80	68/80	60/80
A01	P-PP	P+	62/80	65/80	3/80
A00	P-PP	P-PP	75/80	54/80	3/80

Table 6: Shared-TALK evaluation: TALK-VOTE consistency, VOTE for a village-team player, and composite PP success, each requiring success in both repetitions. A11-A10 consistency is the primary contrast; other comparisons are secondary. Each TALK is shared by its two VOTE branches.

5.4 Shared-TALK, Branched-VOTE Evaluation

All 960 condition-board-repetition units in the shared-TALK evaluation were complete. Table 6 separates TALK-VOTE consistency from correct voting and full protocol completion.

After P+ TALK, the conditions with (A11) and without (A10) the PP bundle at VOTE both achieved TALK-VOTE consistency on 78/80 boards (Table 6). The primary A11-A10 difference was 0.0 points (95% interval: -5.0 to +5.0). This detects no additional consistency gain under the shared base; it establishes neither equivalence nor that VOTE instructions are generally unnecessary. A11 achieved consistency on 78/80 boards, correct voting on 71/80, and composite success on 61/80: agreement can preserve a wrong target. Appendix B provides further diagnostics.

6 Discussion

6.1 Target Execution and Partner-Policy Context

Explicit instructions specify the behavior rewarded by the composite endpoint, so improved compliance is expected. The semantic and deterministic endpoints measure protocol completion across TALK and VOTE. The fixed-board contrast tests one target's continuation from cue-bearing

history; the factorial evaluation observes two dynamically acting teammates.

In the single-target standard games, target PP opportunities arose in approximately one quarter of games. P–PP won 37/50 after arrival without satisfying the strict composite heuristic, so overt protocol completion was not necessary for winning. Sparse opportunities and wins without overt target CO may attenuate the all-game contrast, but treatment-dependent arrivals cannot provide a causal decomposition.

The fixed partner policy produced mismatched disclosure and voting (Appendix Table 17). In the factorial evaluation, both teammates revealed and voted for the village-team survivor in all 56 G11 arrivals, but the all-game contrast remained uncertain. G10 had no possessed CO before that agent’s first VOTE in its 57 arrivals, consistent with limited opportunities for the werewolf’s response-to-CO instruction. These cell-specific subsets cannot establish that CO scarcity caused the small G10 contrast or identify a policy interaction.

6.2 Role Asymmetry and Available Evidence

Every PP-positive werewolf target receives an explicit possessed CO; most possessed targets face weaker ally cues. Possessed success was 10/10 with an explicit werewolf CO and 15/30 otherwise. Across all 40 possessed boards, CO and TALK–VOTE consistency were perfect, yet all 15 composite misses had negative coordination labels. This pattern motivates closer study of ally identification and coordination, but the labels do not establish the underlying failure mechanism or isolate role from cue effects. Matched within-role manipulation of cue explicitness would address the latter distinction.

6.3 Qualitative Error Analysis

Appendix E illustrates an explicit-cue P+ possessed success, a P–PP case that voted correctly without recognizing PP or the ally, and false PP after the possessed had been executed. A separately labeled post-hoc example shows a werewolf response. These cases distinguish vote-target accuracy from coordination and illustrate, rather than estimate, failure frequencies.

6.4 Over-Application When Extending Existing Agents

The near-controls probe premature disclosure or PP assertions at five survivors and mistaken PP

assertions after the possessed’s death. Their imprecise false-PP contrasts do not separately quantify mistaken ally identification. Future evaluations should cross objective feasibility with ally evidence. A role-conditioned activation gate could use survivor state to establish feasibility for the possessed and evidence of possessed survival for the werewolf; identifying the ally and coordinating the actual vote remain separate requirements.

6.5 Consistency, Correctness, and Instruction Scope

With shared P–PP TALK, A01 had fewer consistent boards than A00 (62/80 versus 75/80) but more boards with correct voting (65/80 versus 54/80). Consistency and correctness therefore need not move together. This secondary pattern does not establish why votes changed.

After P+ TALK, retaining the bundle at VOTE gave no observed consistency gain. The generated declaration and common voting rule could sustain agreement in A10, but their separate contributions were not manipulated. The result limits a necessity claim about instruction distribution; neither the stage contrast nor component scores isolate individual clause effects. The fixed, PP-aware village-team background policy also does not make the all-game contrast a conservative bound.

7 Conclusion

The PP bundle improved fixed-board composite execution by +78.8 points (95% interval: +70.0 to +87.5), while game benefits remained uncertain: +1.0 point (−8.0 to +10.0) for single-target substitution and +6.0 points (−3.0 to +15.5) for joint adoption. After shared P+ TALK, retaining the bundle at VOTE gave no additional observed consistency gain: 0.0 points (−5.0 to +5.0). Together, these evaluations separate declaration–action consistency, correct protocol execution, and game outcomes. Observed dyadic coordination at PP arrivals does not establish an overall benefit; individual clause contributions and the necessity of PP instructions at every request remain unresolved.

Limitations

PP was selected after exploratory system development, and the fixed-board corpus is not independent of diagnostic development: all 80 positive boards and 30/40 near-controls originated in that

process. The final executable corpus was evaluated after correcting histories that violated reachable game states rather than serving as an independent confirmation set. The comparisons and analyses were prespecified for the reported evaluation samples, but the research target and most scenarios were not fixed before system development.

The single-target standard-game intervention changed only one agent. Its partner used a fixed PP-aware background policy whose TALK instructions differ from P+ and combine competing concealment and conditional-revelation directions. That result is conditional on the partner policy. The factorial evaluation addresses joint adoption by the werewolf and possessed, but keeps the village-team policy fixed; neither study evaluates changing all five agents. Its interval of -3 to $+15.5$ points leaves substantial uncertainty, and the secondary interaction is imprecise. Both studies are limited to five-player same-model self-play with gpt-5.6-luna. P-PP shortens the input and removes a heterogeneous bundle; individual clause contributions remain unidentified. Common ally-preservation guidance and deterministic response validation remain in both conditions. Fixed-board role comparisons confound role with available ally cues. Finally, semantic items use a condition-masked, ground-truth-reference-assisted LLM classifier whose accuracy against human labels is untested. Three-classification agreement, evidence validation, and the deterministic endpoint assess robustness but do not replace human validation.

The original fixed-board classifier also saw the submitted VOTE, leaving possible action dependence in its semantic labels. The shared-TALK evaluation withheld both VOTE branches from the classifier.

Ethical Considerations

Werewolf involves deception and persuasion, but this work is confined to a closed game environment and is not intended to manipulate real people. Released experimental records contain no credentials or human-participant data. Synthetic player names and experimental identifiers are retained to link inputs, outputs, and judgments; host-specific paths are normalized. Use of released model responses remains subject to applicable provider terms.

Data Availability and Reproducibility Statement

All evaluation boards are synthetic; no human-participant data or competition logs from other teams were used. The repositories linked in footnote 1 archive the frozen instructions, board generator and labels, scoring rubrics, random seeds, analysis code, aggregated results, raw model responses, and game trajectories. The data release (revision release-20260914) includes documentation mapping conditions to artifacts. The released outputs support reanalysis of reported counts and statistics, not deterministic regeneration.

Evaluation integrity. In the 400-game single-target evaluation, one TALK request to a non-target agent in a P-PP game failed because of a transient connection error and used the prespecified no-utterance fallback. The completed game was retained under that evaluation’s exclusion rule; every other recorded call succeeded. The factorial evaluation used a different operational-failure rule and recovery procedure, documented in Appendix A.1.

Post-hoc scoring correction. The deterministic extraction rules were corrected after outcome inspection to prevent third-person vote references from overwriting a speaker’s declaration and negated or attributed role claims from counting as CO. The corrected rules apply to all conditions and controls in the fixed-board and shared-TALK evaluations and all eligible pre-vote utterances in the factorial evaluation. Generations, semantic-majority labels, sample membership, endpoints, aggregation, and resampling settings were unchanged. The single-target standard-game heuristic uses separate rules and retained its definition. This was a post-hoc scoring change; reported results use the corrected analysis. Both analyses, source snapshots, input hashes, extractions, and score differences are archived. Extraction accuracy against independent human labels remains untested.

References

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. [LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages

- 8053–8072, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. 2024. [AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 119919–119948. Curran Associates, Inc.
- Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. 2025. [Can language models follow multiple turns of entangled instructions?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25445–25460, Suzhou, China. Association for Computational Linguistics.
- Yuya Harada and Yoshinobu Kano. 2025. [Intent-driven AIWolf agents with hierarchical BDI model and personality](#). In *Proceedings of the 3rd International AIWolfDial Workshop*, pages 22–36, Hanoi, Vietnam. Association for Computational Linguistics.
- Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. [AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.
- Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. [AIWolfDial 2025: Summary of natural language division of 7th international AIWolf contest](#). In *Proceedings of the 3rd International AIWolfDial Workshop*, pages 1–13, Hanoi, Vietnam. Association for Computational Linguistics.
- Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, Soga Shigemura, Rei Ito, Kazuki Takashima, Tomoki Fukuda, Masahiro Wakutani, Tomoya Hatanaka, Mami Uchida, Mikio Abe, Akihiro Mikami, Takashi Otsuki, Zhiyang Qi, Kei Harada, Michimasa Inaba, Daisuke Katagami, Hirotaka Osawa, and Fujio Toriumi. 2023. [AIWolfDial 2023: Summary of natural language division of 5th international AIWolf contest](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 84–100, Prague, Czechia. Association for Computational Linguistics.
- Takuma Okada and Takeshi Ito. 2025. [Towards a strategic werewolf AI based on expert strategies in five-player werewolf](#). In *Proceedings of the 3rd International AIWolfDial Workshop*, pages 14–21, Hanoi, Vietnam. Association for Computational Linguistics.
- Zhiyang Qi and Michimasa Inaba. 2024. [Enhancing dialogue generation in werewolf game through situation analysis and persuasion strategies](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 30–39, Tokyo, Japan. Association for Computational Linguistics.
- Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. 2024. [An implementation of werewolf agent that does not truly trust LLMs](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 58–67, Tokyo, Japan. Association for Computational Linguistics.
- Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba. 2024. [Enhancing consistency of werewolf AI through dialogue summarization and persona information](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 48–57, Tokyo, Japan. Association for Computational Linguistics.
- Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024. [Enhance reasoning for large language models in the game werewolf](#). *Preprint*, arXiv:2402.02330.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024a. [Exploring large language models for communication games: An empirical study on werewolf](#). *Preprint*, arXiv:2309.04658.
- Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024b. [Language agents with reinforcement learning for strategic play in the Werewolf game](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55434–55464. PMLR.

A Factorial Evaluation: Coordination and Sample Integrity

The coordination diagnostics use the first three-survivor position containing the werewolf and possessed. CO denotes an explicit declaration of the speaker’s actual role before their first submitted VOTE that day. Counts in Table 7 are conditional on reaching this position. They describe different trajectories across cells and do not estimate a causal effect among a common set of arrivals. A successful final vote need not imply that earlier declarations were correct.

Diagnostic	G00	G10	G01	G11
PP arrivals / 200	59	57	57	56
Possessed CO	0/59	0/57	57/57	56/56
Both CO	0/59	0/57	0/57	56/56
Both vote village	14/59	25/57	34/57	56/56
Wins after arrival	36/59	43/57	36/57	56/56

Table 7: Descriptive counts at the first feasible PP position in the factorial evaluation. All rows except arrivals use the cell-specific arrival denominator.

A.1 Sample Eligibility and Execution Conditions

The factorial sample contains 800 accepted games in 200 complete blocks from 820 attempts. Twenty attempts were excluded for operational failures: one TLS connection failure, one VOTE timeout, and 18 interrupted games. Their assignments were rerun with the same roles, seats, policies, model, temperature, and saved within-block condition order. The 138 valid games completed before interruption were retained unchanged.

Before condition-level outcomes were inspected, the eligibility rule was amended to reject action-timeout fallbacks and unfinished API requests and applied to all attempted games. Concurrency was reduced from 20 to eight, and transient failures became eligible for at most three game attempts, including prior attempts. Connection/read/write/pool timeouts were 5/20/20/20 seconds, with at most one SDK retry. Thus, execution conditions differed before and after the amendment. All 662 games run under the amended procedure succeeded in one game attempt, with no pending requests or action timeouts; SDK-internal retries were not separately observable. Sample eligibility was fixed before outcome analysis, and excluded attempts are archived separately.

B Shared-TALK Evaluation: Protocol and Diagnostics

The shared-TALK evaluation reused the 120 reachable boards, including 80 PP-positive boards and 40 controls. It used the same generator, temperature, non-PP base, scoring definitions, and three-call semantic classifier as the fixed-board evaluation. TALK-policy order and the order of the two VOTE branches were randomized with seed 20260907. Each branch restored identical non-PP state after its shared TALK; neither branch received the other’s VOTE. The interven-

tion changed the whole PP bundle at the selected request, not only the local TALK or VOTE template. The corpus remains development-derived rather than an independent generalization set.

There were 480 TALK requests, 960 VOTE requests, and 1,440 shared-TALK classification requests. All succeeded on their first attempts, and no board or condition was omitted. The request limit was two attempts, with permanent stopping after exhaustion.

The primary analysis pairs A11 and A10 within each positive board and resamples the 40 content clusters linking role viewpoints. Both repetitions must agree with the declared target; the two repetitions are not independent observations. A11 and A10 jointly succeeded on 76 boards; two succeeded only in A11 and two only in A10. Thus equal marginal rates do not mean identical board-level outcomes. The 0.0-point estimate and -5.0 to $+5.0$ -point interval describe the additional contribution of PP instructions at VOTE conditional on P+ TALK and the shared base. No equivalence margin was specified.

Table 6 reports consistency, correct voting, and composite success. The deterministic execution endpoint was 70/80, 68/80, 3/80, and 3/80 for A11, A10, A01, and A00, respectively, again requiring success in both repetitions. It excludes all semantic items.

The prespecified A10 diagnostic found one of 80 boards with at least one repetition combining successful TALK and a TALK-VOTE mismatch. TALK success here requires recognition, CO, a declared non-ally village-team target, and coordination, without consulting the actual VOTE. The corresponding A01 diagnostic found 73/80 boards with at least one repetition combining unsuccessful TALK with a VOTE for a village-team player. These any-repetition diagnostics use a different aggregation rule from the primary both-repetitions endpoint and do not establish a mechanism. Composite A11–A01 is not a primary contrast because the TALK-scored components are directly affected by removing TALK instructions.

For A11/A10/A01/A00 respectively, possessed-view consistency was 38/40, 38/40, 32/40, and 38/40, and composite success was 21/40, 20/40, 0/40, and 0/40. Werewolf-view consistency was 40/40, 40/40, 30/40, and 37/40, with composite success 40/40, 40/40, 3/40, and 3/40. These strata retain the role-cue confounding.

False PP was 1/40 for each P+ TALK cell and

2/40 for each P–PP TALK cell. In the missing-ally stratum the counts were 1/10 versus 0/10; in the two Day-1 strata combined they were 0/30 versus 2/30. These labels depend only on shared TALK, so equality across VOTE branches is guaranteed by sharing, not an independent replication. They do not establish equivalent safety. Detailed component scores and strata accompany the archived analysis.

C Instructional Content by Condition

For the designated target, P–PP removes the pre-defined PP-specific clauses from role strategies, shared knowledge, role-conditioned TALK (including village-team rebuttal), VOTE, ATTACK, and endgame notes. Ordinary strategies, ally-preservation wording, and deterministic validation rules remain in both target conditions. In the single-target standard games, the four non-target agents follow the separate fixed background policy below; the factorial evaluation retains it only for the three village-team agents. Tables 9–13 reproduce the P+ instructions. The clauses overlap but have non-identical triggers. In particular, the possessed strategy’s seer-death priority clause did not apply on the 13/40 possessed-view positive boards whose village-team survivor was the seer; more general three-survivor clauses remained available. In the instructions, *white* or *cleared* means HUMAN, *guilty* means WEREWOLF, *fake* usually means a false seer claim, and *half-PP* is an even-survivor forced-tie position.

The single-target standard-game background population uses a PP-aware policy held identical across the two target conditions. This policy differs from both target variants and can therefore moderate the target substitution. Table 8 reproduces the clauses most relevant to that interpretation. The study release documents the complete policy and its correspondence to the experimental records.

D Board Generation and Scoring Rules

Positive boards require exactly three survivors containing one werewolf, one possessed, and one village-team player, with consistent game state and history. All 40 werewolf-view boards contain an explicit prior possessed CO. The 40 possessed-view boards comprise 10 with an explicit werewolf CO, 10 with vote-matching or “three of us” wording, and 20 with weaker defense or

collapsed-seer cues. The 40 near-controls comprise 30 five-survivor Day-1 positions and 10 three-survivor werewolf views in which the possessed is dead. Each board separately labels objective feasibility from true roles and epistemic justification from target-visible information. Names, seats, speaking order, CO wording, and vote declarations vary across templates. The final corpus contained only game states reachable under the engine’s transition rules, and its random seeds were fixed before collecting model responses; the corpus nevertheless followed earlier diagnostic development as disclosed in Limitations.

Composite PP success is the conjunction of PP recognition, required role CO/response, ally avoidance, a valid village-team target, vote coordination interpreted in context, and TALK–VOTE consistency in both repetitions. Ally recognition is diagnostic rather than part of the conjunction. Deterministic execution uses the four rule-scored action items. False PP uses the near-control rule in Section 4.2, with an error in either repetition sufficient for a positive board label. All 80 positive boards were evaluable on every composite item; the false-PP denominator is 40 controls.

Table 14 gives one possessed-view illustration from the explicit-werewolf-CO stratum.

D.1 Operational Definitions

Table 15 defines fixed-board judgments. Role CO, declared target, and actual vote use rules and ground-truth labels; PP recognition, ally recognition, and vote coordination use gpt-5.6-sol, a model different from the generator. Each unit is classified three times and decided by majority vote. The classifier input masks condition, board identifier, repetition, generating instructions, game outcome, future events, and hypothesis. It omits the complete hidden-role map but supplies the perspective role, living players, correct ally (or null), and valid village-team targets as scoring references. Those references were never supplied to the generating agent. The original fixed-board classifier also receives the submitted VOTE; the shared-TALK classifier receives neither VOTE branch.

Ally recognition and vote coordination were judged independently. Archived judgments sometimes accepted a declaration matching an ally’s earlier request, or an appeal to the werewolf by role, as coordination without identifying a particular player as the ally. For example, after the ally requested a vote for Rin, a target’s “I will vote

Background-policy component	Instruction text
General werewolf strategy	You must never admit that you are a werewolf.
Conditional werewolf strategy	If a Possessed claim appears at three survivors, as a rule declare werewolf and match votes onto the possessed and villager candidates.
Condensed TALK context	Don't reveal being a werewolf or possessed; pose as villager-side. If PP is possible at three survivors, the possessed does a Possessed claim, and the werewolf, responding to the Possessed claim, matches votes onto the villager candidate.
Condensed VOTE context	The werewolf matches the alternative candidate or suspect stated in today's remarks. If there's a Possessed claim at three survivors, match onto the villager candidate.

Table 8: Relevant clauses from the fixed background policy used by the four non-target agents in the single-target comparison and by the three village-team agents in the factorial evaluation. The TALK instructions contain both general concealment and conditional revelation directions, whereas the VOTE instruction after a possessed claim selects the village-team candidate.

Instructional context	P+ instruction text
Werewolf strategy	If exactly three players are alive and another survivor explicitly claims possessed, declare werewolf, coordinate with that claimant, and vote for one named villager-side survivor; never vote for the possessed claimant. After revealing werewolf, name exactly one villager-side survivor and ask the possessed claimant to vote for that same player.
Possessed strategy	At three survivors, if you are the possessed and the true seer candidate has been killed by attack or execution, prioritize PP above all. In that situation, rather than adding divine results and breaking down, call out in natural language, "I'm the possessed. Werewolves, let's match votes." If you give divine results at three survivors, always check that you don't have so many cleared players that no wolf position remains. Avoid giving white to two survivors, since that leaves no wolf from your viewpoint and breaks down. After a Possessed claim, clearly instruct the wolf candidates to "not vote for me" and to "declare a vote for a villager candidate."

Table 9: PP instructions included in P+ role strategies.

Rin" was labeled coordination without ally recognition. These interpretations explain how coordination scores can exceed ally-recognition scores; they describe the retained labels, not a revised rubric or human validation.

On controls, the correct-ally reference is null and valid village-team targets are empty. The coordination criterion therefore has no eligible ally-target pair, and all control coordination judgments were negative. False PP in these data is consequently detected through self-claims or the PP-recognition label, which includes acting on an unavailable PP premise. The measure does not separately quantify all incorrect ally identifications.

The target extractor uses only the utterance and participant names, without condition, submitted vote, hidden roles, or outcome. It accepts a unique current voting preference, including tentative wording, and direct requests for coordinated voting. Reports of another player's vote and negated, hypothetical, or quoted commitments are excluded. A direct addressee is not automatically

the target; pronouns require a unique local referent. These are finite deterministic patterns, not unrestricted recovery of intended meaning.

For every judgment, the classifier returned one to three short supporting spans from the target utterance or public history. Evidence was automatically verified against the archived records; 11/1,440 classifications across 10/480 units additionally cited the submitted vote, always for vote coordination and alongside utterance evidence. The number of explicit vote citations does not bound the possible influence of VOTE visibility. The original semantic labels were retained without a VOTE-masked reclassification, so their component relationships do not isolate a TALK-stage failure mechanism. Invalid citations were re-evaluated once. Complete three-classification agreement was 460/480 (95.8%) for PP recognition, 449/480 (93.5%) for ally recognition, and 418/480 (87.1%) for vote coordination.

Instructional context	P+ instruction text
Shared knowledge	At three survivors, if it's wolf/possessed/villager, a PP happens. Always confirm whether a possessed candidate remains and who the wolf and possessed are from your viewpoint. If a Possessed claim appears at three survivors, prioritize declaring werewolf and matching votes. If a wolf/possessed/villager endgame looks likely at three survivors, declare possessed and call for the werewolf to match votes.
Condensed shared knowledge	In a three-survivor endgame, confirm PP; villager-side prioritizes the wolf candidate, the possessed prioritizes matching votes, and the werewolf prioritizes matching votes onto the villager candidate.
Condensed TALK	At exactly three survivors, the possessed claims possessed; after that explicit claim, the werewolf reveals werewolf, names one villager-side survivor, and matches the possessed's vote onto that player.
Condensed VOTE	The werewolf matches the alternative candidate or suspect stated in today's remarks. If there's a Possessed claim at three survivors, match onto the villager candidate.

Table 10: PP instructions included in shared knowledge and condensed variants.

Role	P+ TALK instruction
Possessed	[Possessed's PP judgment] If you are the possessed and there are three survivors, confirm the PP possibility before divine results or unconfirmed player analysis. If you can judge that the three are you, a wolf candidate, and a villager candidate, declare possessed and call for matching votes. Example: "I'm the possessed. Werewolf, reveal yourself and let's match votes" In a three-player board where the true seer candidate was attacked and you remain as the fake seer, prioritize PP above all. In this situation, don't give a new innocent result that erases the wolf position from your viewpoint. If the wolf won't reveal, ask the wolf candidate to "not vote for me" and "declare a vote for the villager candidate."
Werewolf	[Werewolf's PP judgment] If you are a werewolf, there are three survivors, and a Possessed claim has appeared, prioritize declaring werewolf over continuing to pose as a villager. Example: "I'm the werewolf. Shion, match with Shunichi" Reveal werewolf, name exactly one villager-side survivor, and ask the possessed claimant to vote for that same player. Do not vote for the possessed candidate.

Table 11: P+ TALK instructions for the possessed and werewolf.

D.2 Statistical Details

Primary fixed-board proportions receive nominal Wilson intervals without adjustment for within-cluster dependence; these are descriptive summaries. Inference about the condition contrast uses paired risk differences, keeping the two role-viewpoint scenarios with the same predefined template index in one cluster during resampling. Each fixed-board comparison used 10,000 paired cluster bootstrap resamples followed by 100,000 two-sided Monte Carlo cluster sign flips; seeds were 20260815 for the primary endpoint and 20260816 for false PP.

Single-target standard-game win-rate differences use 100,000 matched-pair bootstrap resamples with seed 20260815. Discordant outcomes

comprise 44 pairs won only by P+ and 42 won only by P-PP; 12 pairs were won in both conditions and 102 lost in both. The exact McNemar test is a two-sided binomial test of the 86 discordant pairs. Holm's method adjusts the uncorrected fixed-board and standard-game p values; the raw values were 1.0×10^{-5} and 0.914, and the adjusted values were 2.0×10^{-5} and 0.914, respectively.

For the separate standard-game heuristic, we located the first day on which the target was alive in a three-survivor werewolf-possessed-village composition. Success required PP-related wording; the required possessed CO, or a werewolf CO after a prior ally possessed claim; a submitted VOTE that avoided the ally and selected the village-team survivor; a target utterance naming both the ally and village-team target together with

Role/context	P+ TALK instruction
Village team	If you are villager-side, there are three survivors, and a Possessed claim appeared, don't end on a vote declaration alone. Either make the possessed misjudge, or at least show one concrete element of the wolf candidate and argue back. Material to use: who benefited from the attack, the fake seer's cleared targets, the previous day's votes, and the unnaturalness of not revealing werewolf. Example: "Takumi not revealing werewolf is aiming for misjudgment. They benefited most from Yumi being killed"
Endgame check	Whether it's a wolf/possessed/villager PP board If you're the possessed, whether you should declare possessed and call for the werewolf to match votes If you're villager-side, whether to execute the possessed candidate or directly execute the wolf candidate The possessed may let their fake be seen in the endgame. If PP is possible, prioritize matching votes over hiding the fake.

Table 12: P+ TALK instructions for the village team and endgame check.

Instructional context	P+ instruction text
VOTE	If there are three survivors and you called for a PP, match votes with the werewolf who revealed, or the most werewolf-like person, and vote for the villager candidate. If your cleared players among the survivors have grown too many and the wolf position is gone, prioritize PP over divine results. If a Possessed claim appears at three survivors, vote for the villager candidate, not the possessed candidate.
ATTACK	Does attacking raise or lower the chance of a PP or half-PP Unless there is a guilty (werewolf) result or clear true element, prioritize a true seer candidate over a possessed candidate, or a villager candidate that doesn't break a PP.
Endgame note	At three survivors, always confirm whether it is a wolf/possessed/villager PP board. You are the possessed. If a wolf/possessed/villager board is likely, declare possessed and call for the werewolf to match votes. You are the werewolf. If a possessed CO appears, declare werewolf and match votes onto the villager candidate. At three survivors, prioritize PP judgment and matching votes over a new white result.

Table 13: PP instructions included in action requests and endgame notes.

vote, match, or coordinate wording; and agreement between the last deterministically extracted TALK vote declaration and the submitted VOTE.

E Qualitative Cases

Table 16 illustrates three previously specified outcome categories using complete recorded trajectories. The exact selection ordering for the final examples could not be reconstructed from the archived records. We therefore treat these cases as qualitative illustrations, without using them to estimate the frequencies in Table 3.

Separately, a post-hoc werewolf-view illustration follows the prescribed action sequence. After Emma claimed possessed, target werewolf Alice replied, "I am the werewolf. Julia is the villager; I vote Julia. @Emma Match my vote on Julia," and voted for Julia. Validation left TALK unchanged,

and all composite criteria held in both repetitions.

F Post-Hoc Target-Partner Audit

Table 17 describes behavior on the first day that the designated target reached a true three-survivor PP state. The fixed partner signaled asymmetrically: when the target was the werewolf, the possessed partner following the background policy explicitly claimed its true role in all 54 reached games; when the target was possessed, the werewolf partner following that policy did so in only 1/51. This is compatible with dependence on target-partner policy pairing. It is not a causal explanation of the role-stratified estimates, because reaching PP is itself treatment-dependent.

Item	Content
Shown to target	Day 2 TALK. The target is Victor, whose known role is possessed; Victor, Clara, and Quinn are alive. Clara has COed: “I’m the werewolf. Victor, vote Quinn with me,” and declared a vote for Quinn.
Scoring only (hidden)	The true roles are Clara = werewolf and Quinn = villager. The correct ally is Clara and the valid vote target is Quinn.
Expected action	Victor COs as possessed, commits to coordinating with Clara on Quinn, and actually selects Quinn in VOTE.

Table 14: Illustrative possessed-view board with an explicit prior werewolf cue (10/40 possessed-view positives). Hidden roles and the correct target are used only for scoring.

Metric	Criterion for a positive label	Scoring
PP recognition	Explicitly states that the werewolf and possessed can control the vote now, or acts on that premise. Role CO alone is insufficient.	Semantic
Required role CO	Affirmatively claims to be possessed from the possessed viewpoint, or werewolf from the werewolf viewpoint. Negated, hypothetical, and attributed role claims are excluded. Positive werewolf boards include a prior possessed CO.	Rule
Ally recognition	Identifies or addresses the designated player as a werewolf-team ally. Merely avoiding a vote against that player is insufficient.	Semantic
Ally avoidance	Neither the final TALK declaration nor VOTE targets the true ally.	Rule
Village-team target	Names one legal, living target who belongs to the village team.	Rule
Vote coordination	Requests, instructs, promises, or commits to align votes with the designated ally on the same named valid village-team target, interpreted in the public-history context. An unnamed target is insufficient.	Semantic
TALK–VOTE consistency	The last explicitly stated current vote candidate or direct coordination request in TALK names the transmitted VOTE target.	Rule
False PP	On a near-control, an extracted possessed/werewolf self-claim or a positive PP-recognition or vote-coordination label. The coordination rubric’s control coverage is restricted as described in the text.	Mixed

Table 15: Operational definitions. “Semantic” denotes three-run, condition-masked, reference-assisted classification.

G Exploratory Standard-Game Outcomes

Table 18 gives the exploratory outcomes from Section 5.2, outside the two-test Holm family for fixed-board execution and single-target win rate.

H Runtime Settings

All four evaluations used English prompts and dialogue, with generation by gpt-5.6-luna at temperature 1.0. The generation client set no explicit output-token cap. Recorded model identifiers were undated; a dated model snapshot cannot be recovered from the archived responses. The single-target games were collected on August 15, 2026, the original fixed-board responses on August 16, and both the factorial and shared-TALK evaluations on September 9 (Japan Standard Time).

Full games allowed at most four TALK turns per agent per day, with a total limit of 20 TALK turns per day across all agents. The original fixed-board evaluation instead requested one TALK fol-

lowed by one VOTE per repetition; the shared-TALK evaluation shared one TALK across two VOTE branches. In all evaluations, deterministic TALK validation limited the main segment to 125 non-whitespace characters and allowed an additional 125 in an addressed segment following a valid player mention. Thus, this was a limit on validated dialogue segments, not on generated tokens.

The two full-game studies used the same archived server binary with fixed seat assignments, public voting, one voting round per execution phase with random tie-breaking, and no whisper turns. Server configuration files and the binary checksum are archived with the study records.

Type	Public state and applied rule	Generated and validated TALK	Actual vote and judgment
P+ possessed-view success	Possessed Victor, werewolf Clara, and villager Quinn survive. Clara has COed and requested coordination on Quinn.	Generated: “I’m the possessed. My vote is Quinn. @Clara Do not vote me; declare your vote for Quinn.” Validated: unchanged.	VOTE: Quinn. All six criteria hold in both repetitions.
P–PP miss	Possessed Priya, werewolf Talia, and seer Alice survive. Talia defended Priya and declared a vote for Alice; the predefined target PP bundle is absent.	Generated: “Alice’s claim has not produced a new result today, and her immediate vote on Talia looks defensive. I will vote Alice.” Validated: comma normalized.	VOTE: Alice. Target and declaration are correct, but PP recognition, CO, and ally recognition are false.
P+ false PP	Werewolf Alice and villagers Priya and Uma survive; possessed Felix was executed the previous day.	Generated: “This is a wolf/possessed/villager PP. I vote Priya; her Felix vote followed one angle and ignored the split.” Validated: slashes removed.	VOTE: Priya. False PP because nonexistent vote control is asserted.

Table 16: Qualitative PP success, missed PP, and false-PP examples.

Condition	Target role	Reached (won)	Partner role CO	Target vote V/P	Partner vote V/T
P+	Possessed	22 (17)	1/22	17/5	22/0
P+	Werewolf	33 (33)	33/33	33/0	33/0
P–PP	Possessed	29 (23)	0/29	19/10	18/11
P–PP	Werewolf	21 (14)	21/21	16/5	14/7

Table 17: Deterministic post-hoc audit at the first reached true PP day. “Partner role CO” is an explicit first-person claim of the partner’s true role. V denotes the sole village-team survivor; P denotes the fixed partner; T denotes the designated target. Voting for T has different implications across target roles: the possessed is human for parity, whereas executing the werewolf loses the game. Parenthesized wins and all behavior counts condition on treatment-dependent arrival and are descriptive, not causal effects or mediation estimates.

Condition	Win rate (target possessed)	Win rate (target werewolf)	PP arrival	Win after arrival
P+	19/100 (19.0%)	37/100 (37.0%)	55/200 (27.5%)	50/55 (90.9%)
P–PP	31/100 (31.0%)	23/100 (23.0%)	50/200 (25.0%)	37/50 (74.0%)
Difference (pp)	–12	+14	+2.5	+16.9

Table 18: Exploratory role-stratified and post-arrival outcomes. Role-specific win rates use 100 games per role and condition; PP arrival uses all 200 games per condition. Win after arrival conditions on the target agent reaching a feasible PP position. Differences are descriptive P+ minus P–PP point estimates; no role interaction was specified or tested, and post-arrival contrasts are not causal.