

Do Cross-Game Lessons Appear in Later Dialogue? A Controlled Measurement of Lesson Realization in Werewolf Dialogue Agents

Kota Munaoka¹, Yoshinobu Kano¹

¹Faculty of Informatics, Shizuoka University

Johoku 3-5-1, Chuo-ku, Hamamatsu, Shizuoka 432-8011 Japan

{kmunaoka, kano}@kanolab.net

Correspondence: kano@kanolab.net

Abstract

In LLM-based agents, methods that verbalize lessons learned from past trials and inject them into later prompts to improve behavior without updating model weights are widely used. However, improved outcomes alone do not establish that the provided lessons actually appeared in the output. We implement a mechanism in a Werewolf dialogue agent that reviews its in-game judgments in light of the roles revealed at game end, generates lessons, and injects them into the prompts of later games. To test whether injected lessons actually appear in talk, we compare lesson-present and no-lesson conditions across multiple game settings in 500 self-play games. We define the **realization rate** as, for each candidate lesson, the proportion of utterances in which that lesson’s content appears, using no-lesson utterances in the same situations as the baseline. In subjective evaluation of talk quality, most condition differences were not statistically significant after correction, though the descriptive trend—lesson-present conditions ranked best in turn-based tracks, no-lesson in the freely timed track—paralleled the realization pattern. The realization rate was consistently higher than the no-lesson baseline in the two turn-based tracks, while no significant difference was found in the freely timed track. Confounds preclude a causal interpretation of this parallel.

1 Introduction

Methods that leave model weights unchanged and instead verbalize past trials for injection into later prompts are widely used (Shinn et al., 2023; Zhao et al., 2024). However, even when outcomes improve, this does not constitute evidence that the stored lessons appeared in subsequent behavior. When the LLM is capable of generating similar content on its own without explicit lessons, the two cases must be distinguished.

This paper examines the problem in the context of Werewolf dialogue agents. Werewolf is a game

in which players with hidden roles reason about one another through conversation, and it has two relevant properties. First, talk has no fixed correct output, and the model may generate content similar to a lesson even without being given it. Consequently, improved outcomes (win rate, subjective evaluation) alone cannot establish that lessons were used. Second, the server reveals every player’s role at game end, enabling lessons to be written after examining in-game judgments in light of the true roles.

The **cross-game lessons** proposed in this study examine in-game judgments in light of the revealed roles to write tagged lesson texts, and in later games select a subset for the prompt. The set of candidates is called the **lesson pool**. Selection is governed by two kinds of tags: **lesson-type tags** and **situation tags** (Section 3).

The contest has two utterance formats—turn-based and Speak-Anytime—running as three tracks: turn-based 5-player, turn-based 9-player, and Speak-Anytime 5-player. In the lesson-present condition, selected lessons are inserted into the prompt; in the no-lesson condition, this insertion alone is removed while all other prompt components remain identical.

Research questions We examine the following research questions.

- **RQ1 (subjective quality):** Do lesson-present and no-lesson conditions differ in subjective rankings of talk quality?
- **RQ2 (realization):** In the lesson-present condition, does lesson content appear in talk at a rate above the no-lesson baseline?
- **RQ3 (format/setting):** Is realization detected in all game settings?

RQ1 first asks whether there is a subjective-quality difference. Since quality differences alone

cannot establish lesson use, RQ2 then asks whether realization can be measured directly, and RQ3 examines in which settings realization is detected. Because the tracks differ not only in utterance format but also in other factors such as the accumulated lesson-pool content and the per-day utterance cap, RQ3 concerns the effect of game settings as a whole, not of format alone.

The implementation and game settings follow the Natural Language Division of the 2026 AIWolf International Contest (associated with the AIWolfDial 2026 workshop co-located with INLG 2026) (Kano et al., 2025). Our system (registration name Oxymoron) adds the cross-game lesson mechanism to the official sample agent. We competed with this system in both Round 1 and Round 2. Because the contest confounds opponents, judges, and lesson pool, contest records are reported only for reference in Appendix E.

Contributions

- **A no-lesson control for measuring realization:** We introduce a method that compares lesson-present and no-lesson utterances matched by role, profile, action type, and day, measuring whether lesson content appears at a significantly higher rate.
- **Empirical evidence that lesson content appears:** With this method we show that, in turn-based tracks, injected lessons appear in talk at a rate not explained by the model’s own capabilities alone.

2 Related Work

2.1 AIWolf and AIWolfDial

AIWolf is a competition project for studying agents that reason, make decisions, and conduct natural-language dialogue, using the Werewolf game as a testbed (Toriumi et al., 2017). The natural-language division (AIWolfDial) has been running since the 2019 shared task (Kano et al., 2019, 2023, 2024, 2025). The contest subjectively evaluates talk quality on items such as naturalness, context fit, consistency, and action alignment; our outcome measurement adopts the same evaluation items (Section 4.2). Past systems have addressed within-game context management such as dialogue summarization and extraction of logical information (Kano et al., 2023, 2025); a lesson pool built and reused across games in the contest environment has not been reported, to our knowledge.

2.2 Verbal self-improvement

Reflexion (Shinn et al., 2023) stores reflections from failed trials and adds them to the next trial’s prompt, but uses the most recent few items as they are, with no situation-based selection. ExpeL (Zhao et al., 2024) extracts insights from successful and failed trajectories and manages the insight set by addition, editing, and up/down votes, but at evaluation time concatenates and injects all items with no per-scene selection. LessonL (Liu et al., 2025) has multiple agents generate, accumulate, select, and share lessons for code optimization, but selects lessons by numerical performance and embedding similarity, unlike our system. Agent-Pro (Zhang et al., 2024) does not accumulate individual lessons; it updates the policy prompt itself through policy-level reflection.

Our cross-game lesson system also changes the prompt through lesson injection rather than weight updates. It differs in attaching two discrete tags, lesson type and situation, and selecting only the lessons needed for each action type. No preference value is assigned to individual lessons, and in-game retrieval matches the discrete tags assigned at generation time (Section 3.3), not embedding similarity.

All four studies cited above report task success or win/loss-derived scores; in the scope of our survey, we found no report that measures whether lesson content appears in later talk by comparison against a no-lesson baseline.

2.3 LLM agents for social deduction games

Xu et al. (2023) store responses and reflections in an experience pool at game end together with a win/loss-derived score, and generate advice from experiences retrieved by Sentence-BERT similarity. Using both successes and failures is shared with our system, but we store generalized lesson texts rather than game records and select by discrete tags rather than embedding similarity. Xu et al. (2024) realize cross-game improvement through population-based reinforcement learning of a policy network. For Avalon, Lan et al. (2024) update a per-role strategy text from game history rather than storing individual lessons.

For dialogue agents with long-term memory, Generative Agents (Park et al., 2023) retrieve memories by a weighted sum of recency, importance, and relevance, whereas our system selects primarily by tag matches.

2.4 Measuring the use of provided context

There is an established literature on whether content provided in the prompt is actually used in the output. [Rashkin et al. \(2023\)](#) define a human-judged framework for whether generated text is attributable to identified sources (AIS). [Liu et al. \(2024\)](#) show that task performance degrades sharply with the position of relevant information in the context, and [Longpre et al. \(2021\)](#) measure over-reliance on memorized knowledge when the context contradicts it. Validating LLM judges against human agreement follows [Zheng et al. \(2023\)](#); our human audit (Appendix H) does likewise. These studies evaluate outputs on tasks with a defined answer or source. This paper instead targets free dialogue with no correct answer and measures realization as an increase over the spontaneous-generation rate of matched no-lesson utterances.

3 Utterance Formats and Lesson Mechanism

In Werewolf, players split into a village team and a werewolf team; Day 0 includes talk but no exile; from Day 1 onward players discuss and then exile one player by vote, while role-specific actions such as divination, guard, and attack are processed at night. The 5-player village has 2 Villagers, 1 Seer, 1 Werewolf, and 1 Possessed (human by species but playing for a werewolf-team win). The 9-player village adds 1 Medium, 1 Bodyguard, and 1 extra Werewolf (3 Villagers total), together with a whisper phase visible only to werewolves. At game end the server reveals every player’s role, and our post-game review writes lessons after examining its in-game judgments in light of this revelation (Section 3.1). Each agent is assigned a predefined profile that shapes its speaking style. Role abilities and utterance constraints are in Appendix J.

The contest has two utterance formats across three tracks. In the turn-based format (turn-based 5-player and turn-based 9-player tracks) agents speak when the server requests an utterance; in the Speak-Anytime format (5-player only) they speak at freely chosen times, subject to a cap of four utterances per agent per day and a ten-minute time limit. All talk is in English, capped at 125 characters per utterance (spaces not counted). RQ3 asks whether realization is detected in all three tracks. Because the tracks differ not only in format but also in other factors, we do not attribute track differences to format alone (Section 5).

The system adds the cross-game lesson mechanism to the official LLM-ready sample agent. In the contest and in the experiments of this study we ran Gemma 4 31B IT (google/gemma-4-31b-it) with vLLM ([Kwon et al., 2023](#)). The lesson pool is stored in per-role files and indexed by lesson-type and situation tags. Figure 1 shows the flow from generation to reuse.

3.1 Lesson generation

On receiving a game-end notification, the system runs a post-game review on a background thread. Review is restricted to games in which role revelation completed normally and there were at least six substantive utterances (excluding skips and a lone “Over”), since games with few utterances tend to yield lessons without sufficient grounds.

The review prompt includes the revealed true roles of all players, the outcome (survivors and team result), the full dialogue history, an instruction to judge which decisions were correct or mistaken, one of seven role-specific review foci (e.g., the Seer is asked whether the role claim and divination targets were well timed), and the full text of the lessons already stored for the same role. The model generates three lessons from these items (the prompt is given in Appendix B).

The three lessons are one each of [SELF] (a strategic principle), [STEER] (a method for steering other players), and [PERSONA] (a way of speaking in character). Requiring one of each kind prevents the skew toward [SELF] observed in preliminary trials. Each lesson is further given one of 11 situation tags such as role claim (abbreviated CO in lesson tags), counter-claim, vote, and defense (Appendix C).

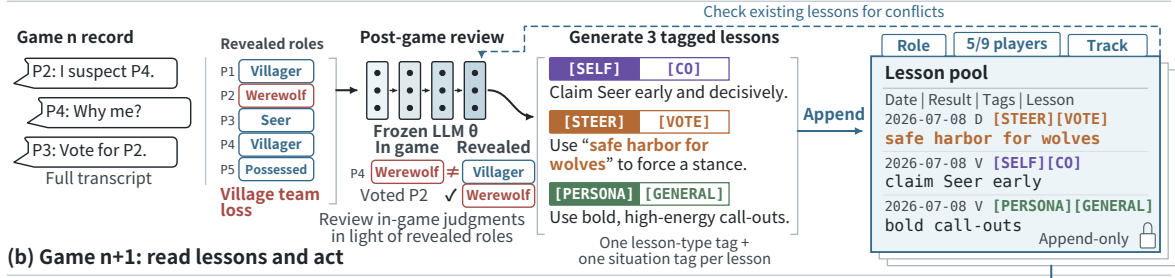
When a new lesson is generated, the model is asked to check it against stored lessons for contradiction; if a contradiction is found, the model should state why this game’s conditions differed or choose a different point (the prompt is given in Appendix B).

3.2 Lesson format and accumulation

Lessons are appended to one file per role and track. Each line records the date, win/loss, lesson-type tag, situation tag, and lesson text. Examples are given in Appendix A.

The Round 1 snapshot held about 1,500 lessons across all roles.

(a) After game n: write lessons



(b) Game n+1: read lessons and act

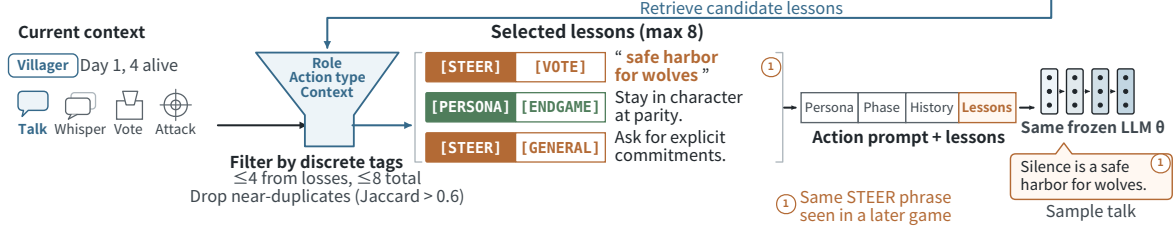


Figure 1: Overview of the cross-game lesson mechanism. At game end the agent reviews its in-game judgments in light of the revealed roles; three lessons, each tagged with a lesson type ([SELF] strategy, [STEER] steering, [PERSONA] persona) and one of 11 situation tags (Appendix C), are stored in an append-only pool, from which later games select through per-action tag filters (Section 3.3). The lower row is a contest-log example: a phrase from a stored Villager lesson (“safe harbor for wolves”) reappears in a later game the same day, in the talk of a Villager with a different profile (the criteria for treating this as consistent with injection are in Section 5).

3.3 Lesson selection and injection

At game start, lessons for the assigned role are loaded without tag filtering and inserted into the system prompt to set the overall strategy (not part of the realization measurement). During the game, lessons are selected by tag when deciding four kinds of action. Talk and whisper select from [STEER] and [PERSONA]; vote and attack select from [SELF] and [STEER]. The priority order of situation tags differs by action; for talk, the tags preferred in ordinary play are role claim, counter-claim, defense, and vote, and the endgame tag is given highest priority once four or fewer players remain (the mapping is in Appendix C).

Lessons for each action are included in the prompt alongside the game state, dialogue history, profile, and action-specific instructions. The eight-lesson cap provides capacity for two alternative recommendations for each combination of the two game outcomes and the two lesson types considered for that action. This budget allows multiple recommendations to be available for different circumstances while limiting the amount of lesson text added to the prompt.

To draw on both successful and unsuccessful games, up to half of the budget, or four lessons, is allocated to lost games, and the remaining slots are filled from won games. Within each outcome

group, the mix of lesson types varies according to situational relevance and recency. The generator determines which of the supplied lessons to express in its utterance.

The system executes the following deterministic steps.

1. **Win/loss slot allocation:** The system selects up to four lessons from lost games and fills the rest from won games (up to eight total), so that the agent sees both what to avoid and what to reproduce.
2. **Situation-tag filtering:** In each slot, the system selects lessons matching the highest-priority situation tag in newest-first order, filling shortfalls from the generic tag (GENERAL) and then from remaining tags.
3. **Deduplication and injection:** The system drops candidates whose word-set Jaccard coefficient with an already selected lesson exceeds 0.6, arranges the remainder from oldest to newest, and inserts them into the prompt between the system instruction and the game situation as a tagged bulleted list.

Tag-based selection is deterministic: which lessons entered which utterance’s prompt can be traced exactly. The realization measurement of Section 5 relies on this traceability.

4 Self-Play Design and Subjective Evaluation

4.1 Design

We compare three lesson-pool conditions. The **Round 1 pool** (R1-pool) is the snapshot used in Round 1 (707 lessons in 5-player, 391 in 9-player, 425 in Speak-Anytime). The **reorganized pool** (Reorg.) is the same snapshot with lesson pairs whose word-set Jaccard coefficient exceeds 0.6 removed as duplicates, retaining the more recent lesson (637, 321, and 355 lessons); it serves as a robustness check on whether pool composition drives the results (also used in Round 2). **No-lesson** (None) uses no lessons at all. In every condition, new-lesson generation is stopped and the pool is frozen; prompts other than lessons, action rules, model settings, and game handling are identical.

Each game contains seats (one player slot each) of all three conditions, so within-game comparison is possible. In 5-player villages the five seats are split 2:2:1, and the condition given one seat rotates across games to equalize seat counts. In 9-player villages the nine seats are split 3:3:3. Roles are assigned by the server; the pairing of profiles and conditions is rotated by round-robin each cycle; game progress, utterance constraints, and judging follow the contest protocol. Because lesson-present and no-lesson seats coexist in the same game, no-lesson seats receive the utterances of lesson-present seats as context (discussed in Section 6.3). Subjective evaluation (Section 4.2) and realization measurement (Section 5) use 500 games from this self-play (150 turn-based 5-player, 150 Speak-Anytime 5-player, 200 turn-based 9-player).

4.2 Subjective evaluation

Setup We measure the effect of lesson-pool conditions on talk quality using the contest’s per-criterion player-ranking procedure: on game logs with win/loss information removed, each judge model independently ranks in-game players on each of six items—Naturalness (A), whether talk is natural; Context (B), whether dialogue fits its surrounding context; Consistency (C), freedom from contradiction (strategic lies permitted); Action alignment (D), whether game actions reflect dialogue content; Expression (E), richness and profile consistency; and Team play (F, 9-player only), team cooperation. The three judge models are GPT-5.4, Gemini 2.5 Pro, and Claude Sonnet 4.5. The mean rank over all seats of each condition is com-

Item	GPT-5.4	Gemini 2.5 Pro	Sonnet 4.5
5-players			
A Naturalness	R1-pool	R1-pool	Reorg.
B Context	R1-pool	R1-pool	R1-pool
C Consistency	R1-pool	None	None
D Action align.	None	R1-pool	R1-pool
E Expression	R1-pool	R1-pool	R1-pool
9-players			
A Naturalness	None	R1-pool	None
B Context	Reorg.	None	Reorg.
C Consistency	Reorg.	Reorg.	Reorg.
D Action align.	Reorg.	R1-pool	Reorg.
E Expression	Reorg.	Reorg.	Reorg.
F Team play	Reorg.	Reorg.	Reorg.
Speak-Anytime 5			
A Naturalness	R1-pool	Reorg.	None
B Context	None	None	None
C Consistency	None	None	None
D Action align.	None	None	Reorg.
E Expression	None	None	None

Table 1: Lesson-pool condition with the best (lowest) mean rank in per-item subjective evaluation of three-condition self-play, shown by judge. Bold marks conditions on which all three judges agree. Full mean ranks are in Appendix D.

puted per judge; we do not average across judges, since their preferences may differ. Condition differences are tested with a game-level bootstrap (Appendix G).

Subjective rankings The per-item subjective evaluation covers 16 cells: five items $\times$ two tracks (5-player) and six items $\times$ one track (9-player). Table 1 shows, for each item and judge, the lesson-pool condition with the best (lowest) mean rank. The best condition differs by track. In turn-based 5-player, a lesson-present condition is best in 12 of 15 item–judge pairs (11 of which are R1-pool); in turn-based 9-player, a lesson-present condition is best in 15 of 18 pairs (13 of which are Reorg.); in Speak-Anytime, no-lesson is best in 12 of 15 pairs (full mean ranks in Appendix D). The three judges agree on the best condition in 8 of 16 cells. After Holm correction (one family of three pairwise tests per judge $\times$ track $\times$ item cell), 17 of 144 pairs show a significant condition difference, and the significant differences do not agree in direction across judges. The only cell on which all three judges agree is Context (B) in turn-based 5-player, where Reorg. is significantly worse than the other two conditions. Furthermore, which of the two lesson-present pools is best differs by track and item; no item has both R1-pool and Reorg. simultaneously significantly outperforming no-lesson, so the effect on subjective quality depends on pool composition.

Because subjective rankings show limited discriminability and cannot confirm mechanism execution, the next section directly measures whether

injected lessons appear in talk by controlled comparison against a no-lesson baseline.

5 Measuring Realization in Dialogue

Procedure To reduce situation-specific variation, we stratify utterances by the tuple (track, profile, role, action type [Talk or Whisper], day) and construct matched pairs from tuples containing both a lesson-present (R1-pool or Reorg.) and a no-lesson utterance. We sample 311–366 utterances per side in each track $\times$ pool cell (4,100 in total). For each matched pair, the judge evaluates the utterance against the eight lessons injected into the injection-side prompt (all had the full eight from Section 3.3). For the control-side utterance—generated without any lessons—these same eight are supplied only to the judge as counterfactual candidates.

The judges are three models from different families chosen for their ability to follow instructions, for stable outputs at temperature 0.0, and for feasibility at the 32,800-judgment scale (which makes the commercial models of Section 4.2 impractical): Gemma 4 31B IT, Qwen3.6-27B, and Mistral Small 3.2 24B (hereafter Gemma, Qwen, Mistral), all open-weight and deployed locally. Judges are not told which track or pool produced the utterance. Each judge sees one utterance, its preceding public context, and eight candidate lessons, and chooses one of six labels—semantic_execute (concrete execution), semantic_partial (partial support), style_or_persona_only (tone only), no_supported_correspondence (no match), contradiction (contradiction), unclear (cannot judge)—for each lesson, and fills in an evidence substring, which our pipeline automatically verified via exact string matching (the instruction is in Appendix I). Each judge labeled all 32,800 items (4,100 utterances $\times$ 8 candidates).

For a lesson ℓ , let $N_{\text{inj}}(\ell)$ and $N_{\text{base}}(\ell)$ be the number of judgments on the injection and control sides (averaging 52–92 per lesson), and $K_{\text{inj}}(\ell)$ and $K_{\text{base}}(\ell)$ the counts of semantic_execute or semantic_partial. The **realization rate** is $i(\ell) = K_{\text{inj}}(\ell)/N_{\text{inj}}(\ell)$ and $b(\ell) = K_{\text{base}}(\ell)/N_{\text{base}}(\ell)$; $b(\ell)$ measures the spontaneous-generation rate of content consistent with the lesson, since no lesson was injected on the baseline side. The remaining four labels are not counted as realization. We exclude style_or_persona_only because the label cap-

Judge	Base.	Inj.	Δ	[95% CI]	p
5-players, R1-pool ($n = 32$)					
Qwen	22.4	25.8	+3.46	[+0.81, +6.04]	.115
Gemma	12.4	15.5	+3.10*	[+1.11, +5.08]	.037
Mistral	6.6	8.1	+1.50	[−0.19, +3.22]	.470
5-players, Reorg. ($n = 32$)					
Qwen	22.2	26.8	+4.55*	[+2.11, +6.96]	.005
Gemma	11.5	14.7	+3.15*	[+1.33, +4.99]	.005
Mistral	7.4	8.4	+0.99	[−0.78, +2.70]	1.000
9-players, R1-pool ($n = 52$)					
Qwen	23.0	26.7	+3.69	[+0.60, +6.77]	.181
Gemma	11.6	17.2	+5.61*	[+3.37, +7.89]	.002
Mistral	8.1	11.7	+3.63*	[+1.40, +5.88]	.018
9-players, Reorg. ($n = 53$)					
Qwen	20.3	26.5	+6.20*	[+3.29, +9.07]	.002
Gemma	12.3	16.5	+4.14*	[+1.44, +6.80]	.035
Mistral	7.7	11.0	+3.25*	[+1.11, +5.50]	.047
Speak-Anytime, R1-pool ($n = 32$)					
Qwen	25.1	27.1	+2.01	[−0.70, +4.63]	.714
Gemma	13.3	15.6	+2.28	[+0.06, +4.48]	.309
Mistral	7.8	9.9	+2.15	[+0.10, +4.15]	.308
Speak-Anytime, Reorg. ($n = 32$)					
Qwen	25.0	25.2	+0.21	[−2.68, +3.04]	1.000
Gemma	12.2	13.4	+1.26	[−1.00, +3.55]	1.000
Mistral	7.8	8.2	+0.44	[−1.43, +2.30]	1.000

Table 2: Realization rates and Δ by condition. n is the number of lessons judged on both sides (32 in 5-player = 4 roles $\times$ 8 lessons; 52–53 in 9-player because werewolf talk and whisper lessons are selected separately). Base. and Inj. are the means of $b(\ell)$ and $i(\ell)$ across lessons (%), Δ is the difference (points), brackets are 95% CI from a game-level bootstrap, and p is Holm-corrected over 18 tests. *: $p < .05$ (Appendix G).

tures stylistic similarity rather than content execution; this makes the metric conservative for [PERSONA] lessons, whose intended realization overlaps with this label.

We write the difference $i(\ell) - b(\ell)$ as Δ . In Table 2, Base. and Inj. are the means of $b(\ell)$ and $i(\ell)$ across lessons judged on both sides (n in the table), and Δ is the mean difference. Intervals come from a game-level bootstrap (10,000 resamples); significance is assessed with Holm correction over all 18 tests (3 judges $\times$ 3 tracks $\times$ 2 pools; Appendix G).

Results Table 2 shows the baseline and injection-side realization rates, the difference Δ with 95% CI, and the Holm-corrected p value for all 18 combinations. In the two turn-based tracks the injection-side rate exceeds the baseline in all 12 combinations (e.g., 11.6% $\rightarrow$ 17.2% for 9-player, R1-pool, Gemma), and 8 are significant after Holm correction (3/6 in 5-player, 5/6 in 9-player). In Speak-Anytime, the injection side is higher in all six combinations but the differences are small, and none

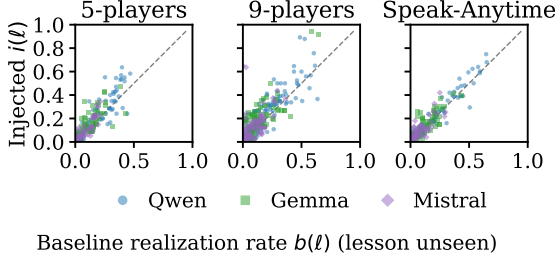


Figure 2: Per-lesson control realization rate $b(\ell)$ (x-axis) and injection-side realization rate $i(\ell)$ (y-axis). Panels from left: 5-players, 9-players, Speak-Anytime. Each point is one lesson $\times$ pool $\times$ judge (blue circles Qwen, green squares Gemma, purple diamonds Mistral); coordinates are the realized fraction among the judgments on each side (52–92 on average). Because tag-based selection (Section 3.3) injects the same lesson into target utterances of the same role and action type across games, each lesson is paired with many utterances and judged repeatedly. The dashed diagonal marks $i(\ell) = b(\ell)$; points above the line have a higher injection-side rate. Read the overall displacement of the cloud from the diagonal, not individual points.

is significant. Significant Δ was observed only in turn-based tracks, though individual judge–pool combinations also differ (e.g., Qwen in 5-player: R1-pool $p = .115$, Reorg. $p = .005$).

The difference is not driven by a few outlier lessons. Figure 2 plots per-lesson $b(\ell)$ and $i(\ell)$. In the turn-based tracks the point cloud sits above the diagonal; in the 8 significant combinations the median per-lesson Δ is positive (+1.1 to +4.5 points), and the mean excluding the top 10% of lessons also remains positive (+1.0 to +4.2 points). In Speak-Anytime the median is near zero (−0.9 to +1.6 points), and the cloud does not depart from the diagonal (per-combination numbers are in Table 11, Appendix F).

To illustrate how lesson wording relates to Δ , Table 3 presents three examples chosen for contrasting patterns: the largest positive Δ , a negative Δ , and a lesson shared across all tracks (the remaining three are in Appendix K). The largest Δ comes from formulaic instructions that can be copied nearly verbatim into talk, such as one requiring names and attack targets in whispers. A tone-maintenance instruction with a high baseline rate $b(\ell)$ had a negative Δ . The only lesson whose text appears identically in all three tracks has a near-zero Δ in every track, consistent with the observation that Δ depends strongly on wording.

The baseline rate (Base. column) depends

Pool	Base.	Inj.	Δ
[STEER][WHISPER] (9-players) “In whisper, explicitly state your name and your partner’s name when agreeing on targets: ‘I am Asuka. I will attack Mio tonight. Midori, do you agree?’”			
R1-pool	38.0	81.8	+43.9
Reorg.	37.8	82.0	+44.1
[PERSONA][GENERAL] (9-players) “Maintain a sharp, analytical tone by focusing on logical inconsistencies and ‘transparent tactics’”			
R1-pool	40.2	22.0	−18.2
Reorg.	35.9	28.5	−7.5
[STEER][DEFENSE] (all three tracks) “Frame the Seer’s black result [a divination identifying the target as a werewolf] as a ‘rushed judgment’: ‘The Seer picked me because I was quiet, not because I was suspicious.’”			
5-players, R1-pool	6.5	4.3	−2.2
5-players, Reorg.	5.3	9.2	+3.9
9-players, R1-pool	1.9	4.0	+2.1
9-players, Reorg.	4.6	3.9	−0.7
Speak-Anytime, R1-pool	7.0	4.9	−2.1
Speak-Anytime, Reorg.	3.9	8.2	+4.3

Table 3: Example lessons and realization rates. Base. and Inj. are the baseline and injection-side rates $b(\ell)$ and $i(\ell)$ averaged across three judges (%), and Δ is their difference (points); these are per-lesson views of the same rates in Table 2. From top: the lesson with the largest Δ (averaged over two pools), a lesson with a high baseline rate and negative Δ , and the only lesson whose text appears identically in all three tracks. The per-lesson standard error of Δ is about 6 points; pool-to-pool differences fall within that range. The remaining three examples are in Appendix K.

strongly on the judge: 20–25% for Qwen, 11–13% for Gemma, and 6–8% for Mistral. Because judge strictness shifts both sides in the same direction, the difference Δ rather than absolute values should be read. A human audit (Appendix H; binary precision 0.76–1.00) confirmed that the sign of Δ matches Qwen and Gemma. Mistral assigns `semantic_execute` to only 0.4% of judgments and has the lowest human agreement ($\kappa = 0.13$); its Δ estimates should be treated as less reliable than those of the other two judges. Injection also does not increase lesson-contradicting utterances.

Speak-Anytime and turn-based tracks differ not only in timing freedom but also in the accumulated lesson-pool content and the per-day utterance cap, so this design cannot isolate which factor suppresses Δ . Among lessons, those with the largest Δ in Speak-Anytime contain patterns for responding immediately to another player’s remark (the [COUNTER] example in Appendix K), consistent with the hypothesis that utterance format affects the opportunity to execute a lesson, though the

number of such lessons is small. The baseline rate $b(\ell)$ being low does not make a lesson more likely to show a positive Δ (Table 12, Appendix F).

6 Discussion

6.1 Answers to the research questions

Subjective rankings show limited condition differences (RQ1): only 17 of 144 pairs are significant after Holm correction, and the significant directions disagree across judges. The pooled test (Appendix G) finds no condition effect. This limited discriminability motivates the direct realization measurement. The difference Δ is significant in 8 of 12 turn-based combinations but in none of the 6 Speak-Anytime combinations (RQ2); realization is thus not detected in all settings (RQ3). Because the tracks differ in factors beyond utterance format, this cannot be attributed to format alone.

6.2 Exploratory observations on subjective rankings

As a post-hoc, strictly hypothesis-generating observation, we note that the 8 item-track cells on which all three judges agree split consistently by track—lesson-present in turn-based, no-lesson in Speak-Anytime (Table 1). While this descriptive pattern parallels the realization-rate results, we do not claim a causal link, as the condition differences in rankings are largely non-significant and confounds—including noise from unrealized lesson text remaining in prompts—cannot be excluded.

6.3 Interference from mixed-condition tables

Placing all three conditions in the same games equalizes roles, profiles, and opponents. The cost is that no-lesson seats receive the utterances of injection-side seats as dialogue context, so lesson content can cross the condition boundary. We call this **same-table interference**. It works in two directions: raising the control realization rate (underestimating Δ) and shrinking team-shared outcome differences (working against detection in RQ1). A low-power stratification by injection-seat count found no significant difference (Appendix F, item 6). Separating conditions by table would avoid interference but restore the opponent confound; we chose to control the confound. The limited condition differences in subjective rankings (Section 6.2) may therefore reflect compression by same-table interference rather than a genuine absence of effect.

6.4 Lesson quality and realization

In Table 3, the whisper lesson combines an attack proposal with a request for a partner’s agreement. Its concrete form may help explain its high realization rate. Its potential strategic benefit lies in exposing differences in intended targets and facilitating agreement. The reusable guidance here is the procedure for proposing and confirming a target, with names and targets adapted to the current game.

The analytical-tone lesson describes behavior already relatively frequent in the no-lesson condition. It may overlap with behavior elicited by the profile or other instructions, suggesting that its additional value depends on what the agent already does. The defense example raises a question of contextual fit: attributing an accusation to quietness may be persuasive when consistent with the preceding dialogue, but could weaken credibility when that premise conflicts with it. These examples motivate lessons that preserve the conditions and rationale for a recommendation while allowing its expression to adapt to the current dialogue.

A useful strategic lesson may also target a rare decision point or affect action selection, yielding a low dialogue realization rate despite its value at the relevant moment. Future evaluation should compare individual lessons with and without injection in applicable situations, assessing coordination, role-appropriate decisions, or profile-consistent dialogue alongside realization. This would inform which lessons to retain and how to revise their wording and applicability conditions.

7 Conclusion

We measured whether tagged cross-game lessons appear in later talk in a Werewolf dialogue agent. The realization difference Δ was significant in 8 of 12 turn-based combinations but not in Speak-Anytime. Subjective evaluation showed limited significant differences; the descriptive trend—lesson-present ranked best in turn-based, no-lesson in Speak-Anytime—paralleled the realization pattern, though confounds keep this exploratory. These findings were obtained in self-play with the same generator model and agent implementation; generalization to dialogue with heterogeneous models or human participants remains to be tested. The procedure generalizes to any prompt-injection agent given injection records, matched no-injection responses, and a realization judgment separate from task success.

Limitations

- **Self-play opponent composition:** Comparison of lesson conditions was run in self-play among copies of this system, so subjective rankings and realization-rate differences describe behavior in this specific opponent composition. Generalization to mixed-model or human-present dialogue environments is a future task.
- **Interference between conditions at the same table:** Because models with different lesson pools coexist in the same game, no-lesson seats are exposed to the utterances of injection-side seats. An interference-free baseline—games composed entirely of no-lesson agents—does not exist in this design, and stratification by injection-seat count (Appendix F) is an exploratory, low-power check.
- **Power of the outcome comparisons:** The minimum detectable effect at power 0.8 and two-sided 5% (MDE80, Appendix G) is ± 0.31 –0.56 for subjective ranks; smaller condition differences cannot be detected. The interpretations in Section 6.2 therefore remain at the level of discussion.
- **Judge dependence of subjective evaluation:** Condition order changes with judge preference; aggregation is restricted to games with valid output from all three models. Tests are per-judge; we do not average or test across judges, and significant-difference directions disagree across judges.
- **Judge dependence of realization rate:** Absolute realization rates depend strongly on the judge.
- **Cross-game pairing of utterances:** Injection and control utterances sharing a situation tuple may come from different games.
- **Scale of the human audit:** Agreement with human judgment was checked on a 120-pair audit sample, but with only one annotator, and strict six-label agreement is only 40–51%; re-validation with multiple annotators is needed.

References

- Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, Issei Tsunoda, Shoji Nagayama, Dolça Tellols, Yu Sugawara, and Yohei Nakata. 2019. [Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the Werewolf game through conversations](#). In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 1–6, Tokyo, Japan. 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. [AI-WolfDial 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. [AI-WolfDial 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, and 7 others. 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.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, pages 611–626. Association for Computing Machinery.
- Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. [LLM-based agent society investigation: Collaboration and confrontation in Avalon gameplay](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 128–145, Miami, Florida, USA. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy

- Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yuanzhe Liu, Ryan Deng, Tim Kaler, Xuhao Chen, Charles Leiserson, Yao Ma, and Jie Chen. 2025. [Lessons learned: A multi-agent framework for code LLMs to learn and improve](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 109172–109228. Curran Associates, Inc.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, pages 1–22. Association for Computing Machinery.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*, 49(4):777–840.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.
- Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Daisuke Katagami, Kosuke Shinoda, and Hitoshi Matsubara. 2017. [AI Wolf Contest — development of game AI using collective intelligence —](#). In *Computer Games: CGW 2016, GIGA 2016, Revised Selected Papers*, volume 705 of *Communications in Computer and Information Science*, pages 101–115. Springer.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. [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. 2024. [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.
- Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024. [Agent-Pro: Learning to evolve via policy-level reflection and optimization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5348–5375, Bangkok, Thailand. Association for Computational Linguistics.
- Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. [ExpeL: LLM agents are experiential learners](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Example Lessons

We show examples from the Round 1 5-player lesson pool, one per lesson type. Each line consists of date, win/loss, lesson-type tag, situation tag, and text.

- (2026-07-08/Victory)
[SELF][CO] Claim Seer early and decisively to establish a baseline of authority → reduces the likelihood of others ignoring your presence or questioning your timing.
- (2026-07-09/Victory)
[STEER][COUNTER] Highlight contradictory statements and rapid target shifts of counter-claimants → exposes inconsistency and undermines their credibility.
- (2026-07-08/Victory)
[PERSONA][GENERAL] Use aggressive, high-energy call-outs to create a clear polarity between the real Seer and the impostor → forces other players to choose a side based on logic rather than politeness.

B Post-Game Review Prompt

Figure 3 shows the parts of the post-game review prompt that concern lesson generation and output format. Braces are replaced at run time with the contents of each game. Role-specific review foci (`{role_focus}`) are prepared in seven kinds, six

Action	Lesson types	Situation tags (priority order)
Game start	All (no filter)	All (no filter)
Talk	STEER, PERSONA	Ordinary: CO, COUNTER, DEFENSE, VOTE, GENERAL. Endgame: prepend ENDGAME
Whisper	STEER, PERSONA	WHISPER, ATTACK, GENERAL
Vote	SELF, STEER	Ordinary: VOTE, COUNTER, GENERAL. Endgame: prepend ENDGAME
Attack	SELF, STEER	ATTACK, WHISPER, GENERAL

Table 4: Lesson-injection filters by action type. Endgame is defined as four or fewer players alive.

roles plus a generic one. The lower box is the Seer example. When extracting lessons from the response, we drop heading-style lines and lines whose body is shorter than 15 characters, and keep at most three.

C Situation Tags and Per-Action Injection Filters

There are 11 situation tags: CO (role claim), COUNTER (handling counter-claims), VOTE (voting), WHISPER (whisper), ATTACK (attack-target selection), GUARD (guarding), DIVINE (divination-target selection), MEDIUM (handling Medium results), ENDGAME (endgame with four or fewer alive), DEFENSE (defense when accused), and GENERAL (generic).

Table 4 shows the lesson-type and situation-tag filters by action type.

D Per-Item Subjective Evaluation by Lesson Condition

Table 5 gives the full seat-pool mean ranks underlying each row of Table 1 in the main text.

E Contest Results

Oxymoron competed in Round 1 and Round 2 of AIWolfDial 2026 (run in July–August 2026, before the workshop). Opponents, number of teams, judges, and the lesson pool change together, so rank differences between the two rounds cannot be attributed to the mechanism or to the pool alone. In Round 2 the number of teams increased (e.g., 11 to 18 in 5-player) and the number of judges rose from 3 to 8. Overspecialization of Round 1 lessons to the Round 1 opponents is one possible explanation for the drop, but we cannot separate the effect of the pool from changes in the competitive environment.

Table 6 shows Oxymoron’s per-role win rates in Round 1, and Table 7 the per-item subjective evaluation in both rounds. Round 1 weighted micro-average win rates for all teams are presented in

Table 8, and subjective-evaluation rankings are presented in Table 9. Round 2 official subjective rankings are in Table 10. Per-item trends are stable: Expression (E) has a better rank than Action alignment (D), and 9-player Team play (F) is the worst item in both rounds.

F Robustness of the Quantitative Analysis

We checked the following aspects of the quantitative analysis in Section 5.

1. Because only situation tuples present on both sides are compared, day and role distributions on the injection and control sides nearly coincide: the per-cell sample sizes differ slightly between the two sides, but the total-variation distance between their distributions was at most 0.010 for day and 0.017 for role.
2. Stratifying utterances by character-count terciles and reweighting the injection-side realization rate to the control-side length distribution barely changed it: the change was 0.2–0.7 points in turn-based 9-player and at most 0.2 points in the other tracks, for all three judges.
3. Control-side judgments use the same candidate lessons in the same order as the matching injection-side utterance, so presentation-order bias cancels in Δ . Actual realization rates by position fell in 23–26% for Qwen and 11–18% for Gemma. Mistral showed a first-position bias (18.8% for the first candidate, 6–10% elsewhere), but this design leaves the Δ estimate unaffected.
4. When only semantic_execute is counted as realization, Δ remained positive in every turn-based cell for Qwen and Gemma (+1.3 to +4.3 and +1.7 to +3.8 points). Mistral assigns semantic_execute to only 0.4% of all judgments, so this definition cannot capture a difference.
5. Binary agreement (execute or partial as realization) was 85.9% between Qwen and Gemma (Cohen’s $\kappa = 0.55$) and 79.4% between Qwen and Mistral ($\kappa = 0.30$). Gemma, the model also used for generation, assigned no-supported-correspondence to 77.8% of judgments, comparable to Mistral (78.1%) and higher than Qwen (58.3%); there was no tendency for the generator to judge its own output leniently.

You played a game of Werewolf as "{role_value}" ({agent_name}) and the game has ended. All players' true roles are now revealed.

True roles of all players
{roles_block}

Final result

- Survivors: {alive}
- Your side: {outcome}

Full game transcript
{transcript}

Self-review

Critically assess your own behaviour ({agent_name}) in light of the revealed roles. Do NOT write a narrative summary of the game. Instead, identify specific decisions that were correct or mistaken, and extract GENERALISABLE lessons for future games.

Role-specific review focus
{role_focus}

Previously recorded lessons for this role
{existing_lessons}

Before writing your 3 new lessons, check them against the ones above. Do NOT write a lesson that flatly contradicts one already recorded (e.g. do not write "trust the first claimant" if a lesson above says the opposite, or vice versa). If this game's experience genuinely differs from an existing lesson, either write about something this game actually demonstrated instead, or state the specific condition that makes this case different rather than issuing a flat opposite.

Output format

Write exactly 3 lessons as bullet points (one per line). You MUST include exactly one [SELF], one [STEER], and one [PERSONA] – one lesson of each kind, not three of the same kind. Persona-reflection lessons are just as important as strategic ones; do not skip [PERSONA] even if the game gave you little to say about it – reflect on how well you kept your persona's voice under pressure, or how you could have voiced a claim or vote more in character.

Do NOT use player names in the lesson text – generalise to roles or positions (e.g. 'the first Seer claimant' not 'Victoria').

Each lesson MUST start with one of these tags:

- [SELF] – a strategic principle: what you should do (or avoid) next time. State the principle, not the story. Format: [SELF] Principle → why it matters.
- [STEER] – a technique for influencing other players (especially LLM agents): how to frame questions, plant premises, redirect attention, or make others follow your lead. Format: [STEER] Technique → when to use it.
- [PERSONA] – a procedure for reflecting your profile/persona in speech: voice, metaphors, motivation, or character consistency. Format: [PERSONA] Procedure → when it helps or hurts.

After the role tag, add a SITUATION tag in a second bracket indicating when this lesson applies. Choose from: CO (Seer/role claiming), COUNTER (dealing with counter-claims), VOTE (voting decisions), WHISPER (wolf coordination), ATTACK (attack target selection), GUARD (guard target selection), DIVINE (divine target selection), MEDIUM (medium result usage), ENDGAME (late game/parity), DEFENSE (when accused), GENERAL (always applicable).

Format: [ROLE][SITUATION] Principle → why/when.

Example: [SELF][CO] Claim Seer on Day 0 before anyone else → first claimant is trusted more.

Example: [STEER][COUNTER] Frame a late counter-claim as 'suspicious timing' → the village doubts the late claimant.

Keep each lesson to ONE line. No multi-line explanations.

- Did you CO at the right time? Was Day 0 first-claim the right call?
- Were your divine target choices optimal? Did you prioritise counter-claimants, aggressive speakers, or silent players?
- Did you present white results as progress (narrowing suspects) or did you let the village frame them as 'no black = fake'?
- Did you coordinate with the Medium's lynch confirmations to strengthen your credibility?

Figure 3: Post-game review prompt. The upper box shows the parts common to all roles that concern lesson generation and output format. The actual prompt continues after this with output instructions unrelated to lesson generation, which we omit here. The lower box is an example of a role-specific review focus (Seer).

Item	GPT-5.4			Gemini 2.5 Pro			Claude Sonnet 4.5		
	Reorg.	R1-pool	None	Reorg.	R1-pool	None	Reorg.	R1-pool	None
5-players									
A Naturalness	3.209	2.839	3.020	2.703	2.504	2.755	2.495	2.668	2.639
B Context	3.253	2.891	2.946	3.198	2.810	2.827	3.209	2.821	2.874
C Consistency	3.121	2.901	2.935	2.808	2.617	2.612	2.753	2.358	2.299
D Action align.	3.104	2.945	2.912	2.368	2.117	2.293	2.423	2.416	2.520
E Expression	3.049	2.799	3.156	2.830	2.569	2.881	2.934	2.675	2.963
9-players									
A Naturalness	4.993	5.313	4.622	4.472	4.323	4.665	3.942	3.903	3.660
B Context	4.855	5.083	5.030	4.048	4.198	3.705	4.778	4.955	5.022
C Consistency	4.640	5.087	4.955	4.332	4.410	4.402	3.383	3.823	3.612
D Action align.	4.660	5.008	4.913	4.850	4.572	4.787	4.230	4.382	4.255
E Expression	4.850	4.970	5.170	3.885	4.128	4.103	4.607	4.773	4.807
F Team play	4.743	4.968	5.218	4.640	4.900	4.893	4.533	4.588	4.845
Speak-Anytime 5									
A Naturalness	2.935	2.928	3.131	2.617	2.853	2.737	2.427	2.486	2.418
B Context	3.040	3.028	2.932	2.980	2.976	2.833	2.927	3.008	2.924
C Consistency	3.085	3.028	2.817	2.786	2.677	2.526	2.641	2.530	2.470
D Action align.	3.077	3.040	2.821	2.371	2.271	2.239	2.524	2.606	2.606
E Expression	3.024	3.028	2.948	2.754	2.873	2.697	2.851	2.988	2.821

Table 5: Per-item subjective evaluation in three-condition self-play (seat-pool mean of in-game rank, reported separately by judge; lower is better; bold marks the lowest-mean condition in each row). Game IDs evaluated by the three models are identical within each track: 150 turn-based 5-player, 200 turn-based 9-player, and 150 Speak-Anytime 5-player games. The two 5-player tracks use a 1–5 scale and the 9-player track a 1–9 scale; the scales are not comparable.

	5-players	9-players	Speak-Anytime 5
Villager	72.5	50	50
Seer	75	50	50
Medium	—	100	—
Bodyguard	—	100	—
Possessed	40	50	50
Werewolf	45	75	50
Games	100	18	10
Macro	61.0	66.67	50.0
Micro	58.13	70.83	50.0
W. Micro	62.27	63.46	50.0

Table 6: Oxymoron’s per-role win rates (%) and aggregate win-rate measures in AIWolfDial 2026 Round 1, reproduced from the official results. W. Micro denotes the weighted micro-average win rate.

Item	5-players		9-players		Speak-Anytime 5	
	R1	R2	R1	R2	R1	R2
A Nat.	1.55	7.31	2.61	8.88	1.40	7.00
B Ctx.	2.00	9.12	3.70	10.25	2.30	7.06
C Cons.	2.18	9.31	3.30	11.31	2.03	7.94
D Act.	2.22	12.12	4.30	10.62	2.23	8.06
E Expr.	1.50	6.38	2.22	6.88	1.53	5.31
F Team	—	—	4.61	12.25	—	—
Mean	1.891	8.20	3.457	10.42	1.900	7.00
Rank	2nd/11	7th/18	2nd/9	11th/14	1st/7	7th/12

Table 7: Per-item subjective evaluation (mean of per-log relative ranks; lower is better). R1 and R2 denote Round 1 (3 judges) and Round 2 (8 judges); the scales and team counts are not comparable between rounds. F is 9-player only.

- To check whether same-table interference grows with the number of injected seats, we stratified the control realization rate in 5-player games by the number of injected seats at the table (games with one no-lesson seat have four injected seats; games with two have three; 9-player is always 3:3:3). The difference between four and three injected seats was -1.3 (Qwen), -3.8 (Gemma), and -2.4 points (Mistral) in turn-based 5-player, and -1.4 , -0.9 , and -0.2 points in Speak-Anytime; all game-level bootstrap 95% intervals cross zero. Turn-based 5-player has only five four-seat games, so power is low.
- To test whether lessons with a low baseline rate $b(\ell)$ are more likely to show a positive Δ , we ran a split-half analysis (Table 12). The unit is the (pool, lesson) pair (pools merged within a track). The control judgments of each unit are split in half by the parity of an utterance-ID hash, one half for grouping and the other for Δ . Under the naive scheme that uses the same rate for both, the low- b group’s Δ looks positive in all 9 combinations, but in the split-half version it shrinks, with intervals crossing zero in 7 of 9, and the high- b group exceeds the low- b group in 8 of 9. This rever-

5-players		9-players		Speak-Anytime 5	
Team	W. Micro (%)	Team	W. Micro (%)	Team	W. Micro (%)
yharada	65.00	CamelliaDragons	71.15	CanisLupus	63.64
Oxymoron	62.27	CanisLupus	69.23	SunameLLi	54.55
noisy_watcher	61.82	Oxymoron	63.46	yharada	54.55
SunameLLi	60.45	wool_100%	55.77	Oxymoron	50.00
ryanleeai	59.09	yshimoda	53.85	CamelliaDragons	50.00
wool_100%	54.09	noisy_watcher	44.23	noisy_watcher	45.45
yshimoda	49.55	ryanleeai	42.31	wool_100%	31.82
kataken14f	48.64	yharada	36.54		
gotsumori-NLP	47.73	SunameLLi	32.69		
CamelliaDragons	45.91				
CanisLupus	36.36				

Table 8: Weighted micro-average win rates (%) for all teams in AIWolfDial 2026 Round 1, reproduced from the official results. Teams are listed in the order shown in the source. Higher values indicate better performance. Oxymoron is shown in bold.

5-players			9-players			Speak-Anytime 5		
Rank	Team	Score	Rank	Team	Score	Rank	Team	Score
1	ryanleeai	1.667	1	ryanleeai	2.639	1	Oxymoron	1.900
2	Oxymoron	1.891	2	Oxymoron	3.457	2	SunameLLi	2.320
3	yharada	2.380	3	yharada	3.917	3	yharada	2.687
4	SunameLLi	2.743	4	CamelliaDragons	3.938	4	noisy_watcher	2.960
5	noisy_watcher	2.825	5	SunameLLi	4.435	5	wool_100%	3.247
6	kataken14f	2.856	6	noisy_watcher	5.506	6	CanisLupus	3.513
7	wool_100%	3.079	7	yshimoda	5.651	7	CamelliaDragons	3.880
8	CamelliaDragons	3.137	8	wool_100%	6.299			
9	yshimoda	3.355	9	CanisLupus	6.704			
10	CanisLupus	3.827						
11	gotsumori-NLP	3.920						

Table 9: Overall subjective-evaluation rankings of all teams in Round 1 (lower scores are better). Oxymoron is in bold.

sal shows that the naive low- b Δ was inflated by regression toward the mean.

G Details of the Significance Tests

All tests in this paper are based on a game-level bootstrap (10,000 resamples). For subjective evaluation, the estimator is the pooled mean over all seats of each condition, following the official aggregation. For realization-rate Δ , the estimator is the mean of per-lesson differences $i(\ell) - b(\ell)$. Because the round-robin rotation correlates seats within a game across conditions, no independence is assumed at the event or seat level. Two-sided p values are $2 \min\{P(\Delta \leq 0), P(\Delta \geq 0)\} + 1/B$, reported with both percentile and BCa 95% confidence intervals; the two agreed within 0.3 points in every test. Multiple comparisons are Holm-corrected within the following families. We also report the minimum detectable effect at power 0.8 and two-sided 5% (MDE80).

- **Realization-rate Δ :** the full family of 3

judges $\times$ 3 tracks $\times$ 2 pools (18 tests). After Holm correction 8 of 18 are significant: 3/6 in turn-based 5-player, 5/6 in turn-based 9-player, 0/6 in Speak-Anytime (marked in Table 2).

- **Subjective evaluation:** one family of 3 pairs per judge $\times$ track $\times$ item cell (48 cells). No tests across judges. After correction, 17 of 144 pairs are significant, and directions disagree across judges and tracks. The only cell on which all three judges agree is Context (B) in turn-based 5-player (Reorg. significantly worse than the other two); conversely, on Consistency (C) in turn-based 9-player, Reorg. is significantly better than R1-pool for Claude Sonnet 4.5 and GPT-5.4. MDE80 is ± 0.31 – 0.56 on the rank scale.
- **Pooled subjective ranks:** To test whether the track-split pattern of Section 6.2 also appears in a pooled measure, we average the rank over

5-players			9-players			Speak-Anytime 5		
Rank	Team	Score	Rank	Team	Score	Rank	Team	Score
1	yharada	1.40	1	UEC-IL	1.75	1	yharada	1.00
2	UEC-IL	2.20	2	yharada	2.08	2	UEC-IL	2.80
3	ryanleeai	3.00	3	ryanleeai	3.33	3	wool_100%	3.40
4	wool_100%	4.00	4	wool_100%	4.58	4	SunameLLi	4.40
5	SunameLLi	5.90	5	SunameLLi	5.50	5	ryanleeai	6.20
6	AGKAI	8.10	6	kataken14f	6.67	6	kataken14f	6.40
7	Oxymoron	8.20	7	noisy_watcher	6.75	7	Oxymoron	7.00
8	yatolab	9.10	8	yatolab	8.17	8	yatolab	8.00
9	yshimoda	9.70	9	yshimoda	8.58	9	CanisLupus	9.60
10	NTT-HAI	10.10	10	CanisLupus	10.33	10	kanolab	9.60
11	kataken14f	10.60	11	Oxymoron	10.42	11	noisy_watcher	9.60
12	CamelliaDragons	12.00	12	Mackerel	11.42	12	CamelliaDragons	10.00
13	carrot	12.20	13	carrot	11.50			
14	Luna	12.80	14	CamelliaDragons	13.92			
15	Mackerel	13.00						
16	CanisLupus	13.70						
17	gotsumori-NLP	17.40						
18	noisy_watcher	17.60						

Table 10: Official overall subjective rankings in Round 2 (8-judge mean of per-log relative ranks; lower is better). Ranks reproduce the published rank columns. Oxymoron is in bold. Scores are not comparable with Table 9.

Judge	Median	w/o top 10%
5-players, R1-pool		
Qwen	+3.1	+1.5
Gemma	+1.3	+1.3
Mistral	+0.0	+0.7
5-players, Reorg.		
Qwen	+4.3	+2.9
Gemma	+2.0	+1.9
Mistral	+0.0	0.0
9-players, R1-pool		
Qwen	+3.2	+1.6
Gemma	+3.0	+3.8
Mistral	+2.5	+1.6
9-players, Reorg.		
Qwen	+4.5	+4.2
Gemma	+2.2	+2.2
Mistral	+1.1	+1.0
Speak-Anytime, R1-pool		
Qwen	+1.6	+1.1
Gemma	+1.3	+0.9
Mistral	+1.0	+1.3
Speak-Anytime, Reorg.		
Qwen	-0.9	-0.8
Gemma	+1.3	+0.1
Mistral	+0.1	-0.8

Table 11: Summary of the per-lesson Δ distribution (points). Median is the median of per-lesson $\Delta = i(\ell) - b(\ell)$; w/o top 10% is the mean after excluding lessons with the top-decile Δ .

items per seat, then over seats per condition, and test each pool against no-lesson. One family of 3 tracks $\times$ 3 judges $\times$ 2 pools (18 tests). Two have $p < .05$ before correction (turn-

Judge	Group	Δ	[95% CI]	Naive
5-players (units 64)				
Qwen	Low- <i>b</i>	+0.94	[-1.68, +4.39]	+3.54
Qwen	High- <i>b</i>	+7.50	[+3.25, +10.88]	+4.47
Gemma	Low- <i>b</i>	+1.98	[-1.11, +3.96]	+3.69
Gemma	High- <i>b</i>	+3.53	[+0.95, +7.26]	+2.57
Mistral	Low- <i>b</i>	+0.70	[-1.22, +1.64]	+0.93
Mistral	High- <i>b</i>	+1.99	[-0.27, +5.08]	+1.57
9-players (units 105)				
Qwen	Low- <i>b</i>	+1.62	[-1.39, +5.68]	+3.86
Qwen	High- <i>b</i>	+6.29	[+1.21, +10.13]	+6.03
Gemma	Low- <i>b</i>	+4.11	[+1.75, +6.89]	+4.38
Gemma	High- <i>b</i>	+7.15	[+3.80, +10.09]	+5.35
Mistral	Low- <i>b</i>	+4.08	[+1.09, +5.65]	+5.46
Mistral	High- <i>b</i>	+3.23	[+0.74, +7.05]	+1.60
Speak-Anytime 5 (units 64)				
Qwen	Low- <i>b</i>	+0.85	[-1.68, +3.68]	+1.09
Qwen	High- <i>b</i>	+3.78	[-0.50, +7.72]	+1.13
Gemma	Low- <i>b</i>	+1.66	[-0.14, +4.53]	+2.55
Gemma	High- <i>b</i>	+3.30	[-0.75, +6.11]	+1.00
Mistral	Low- <i>b</i>	+0.64	[-1.68, +2.32]	+2.01
Mistral	High- <i>b</i>	+1.36	[-1.15, +4.29]	+0.58

Table 12: Split-half two-group comparison by baseline rate $b(\ell)$ (procedure in item 7 above). The unit is the (pool, lesson) pair (pools merged; units gives the count). Low-*b* and High-*b* are groups by the median of the grouping-half baseline rate; Δ is the within-group mean difference (points); brackets are 95% CI from a game-level bootstrap (10,000 resamples). Naive is the point estimate of the original scheme using the same rate for both grouping and Δ .

based 5-player, Gemini 2.5 Pro, R1-pool; turn-based 9-player, GPT-5.4, Reorg.); none survives Holm correction (Table 13).

Judge	Pool	Δ	[95% CI]	p
5-players				
GPT-5.4	R1-pool	-0.119	[-0.283, +0.046]	.159
GPT-5.4	Reorg.	+0.153	[-0.031, +0.340]	.106
Gemini 2.5 Pro	R1-pool	-0.150	[-0.301, -0.003]	.045
Gemini 2.5 Pro	Reorg.	+0.108	[-0.064, +0.279]	.224
Sonnet 4.5	R1-pool	-0.072	[-0.197, +0.055]	.264
Sonnet 4.5	Reorg.	+0.103	[-0.049, +0.260]	.190
9-players				
GPT-5.4	R1-pool	+0.087	[-0.103, +0.278]	.380
GPT-5.4	Reorg.	-0.194	[-0.382, -0.004]	.047
Gemini 2.5 Pro	R1-pool	-0.004	[-0.181, +0.174]	.963
Gemini 2.5 Pro	Reorg.	-0.055	[-0.232, +0.117]	.535
Sonnet 4.5	R1-pool	+0.038	[-0.102, +0.179]	.600
Sonnet 4.5	Reorg.	-0.121	[-0.261, +0.019]	.099
Speak-Anytime 5				
GPT-5.4	R1-pool	+0.080	[-0.105, +0.268]	.392
GPT-5.4	Reorg.	+0.102	[-0.085, +0.289]	.283
Gemini 2.5 Pro	R1-pool	+0.124	[-0.070, +0.323]	.202
Gemini 2.5 Pro	Reorg.	+0.095	[-0.082, +0.273]	.284
Sonnet 4.5	R1-pool	+0.076	[-0.082, +0.237]	.340
Sonnet 4.5	Reorg.	+0.026	[-0.117, +0.176]	.696

Table 13: Pooled subjective-rank test. Δ is the difference between each pool condition and no-lesson (negative means the pool is better). Values are the cross-item mean of per-seat rank, averaged across all seats of each condition. Brackets are 95% CI from a game-level bootstrap (10,000 resamples); p is uncorrected. No test is significant after Holm correction over 18 tests.

Judge	Agreement	Cohen’s κ
Qwen	50.8 (61/120)	0.289
Gemma	46.7 (56/120)	0.200
Mistral	40.0 (48/120)	0.128

Table 14: Strict six-label agreement (%) between human and model labels (120 pairs). All three models agreed with the human on 36 pairs (30%); none agreed on 48 pairs (40%).

H Audit of Agreement between Judges and Human Labels

We checked how far the three realization judges of Section 5 agree with human labels. We stratified 3 tracks $\times$ 2 sides (injection and control) and sampled 20 utterance–lesson pairs per stratum, 120 in total. The lesson is drawn uniformly from the eight candidates of each task; since only one is taken per task, the 120 pairs all come from different utterances. The annotator was one author, was not shown the condition or generating model, and followed the same six labels and the same rules as the models (matching the lesson’s object, action, condition, and purpose against evidence substrings). From this audit we use only three conclusions—the level of binary agreement (95% interval ± 7 –9 points at $n = 120$), the ordering of the three judges by κ , and the sign of Δ —and do not use it to estimate the magnitude of Δ or per-label rates.

Judge	Agreement	Cohen’s κ	Recall	Precision
Qwen	80.8	0.466	0.47 (16/34)	0.76
Gemma	80.0	0.374	0.29 (10/34)	1.00
Mistral	72.5	0.167	0.21 (7/34)	0.54

Table 15: Binary agreement with human labels, where semantic_execute or semantic_partial counts as realization (same binarization as the realization rate). Recall is the fraction of the 34 human-realization pairs that the model also labeled as realization.

Annotator	Injection	Baseline	Δ (pt)
Human	36.7 (22/60)	20.0 (12/60)	+16.7
Qwen	21.7	13.3	+8.4
Gemma	11.7	5.0	+6.7
Mistral	10.0	11.7	-1.7

Table 16: Realization rates (%) by condition on the 120-pair audit sample. The sample is stratified for the audit and is not the same population as the main-text estimate (all games, game-level bootstrap), so it does not replace the Δ estimates in Table 2.

The confusion matrices (Figure 4) show that model errors concentrate on one pattern. Of 32 pairs the human labeled as contradiction, models labeled no-supported-correspondence in 19 (Qwen), 28 (Gemma), and 22 (Mistral) cases. Of the 29 pairs the human labeled as semantic_partial, the same label was applied in 7 (Qwen), 4 (Gemma), and 5 (Mistral) cases. The reverse error—over-calling semantic_execute or semantic_partial on pairs the human labeled as no-supported-correspondence—occurred only four times across the three models. The judges thus skew toward false negatives (stricter than the human). By label, no-supported-correspondence pairs agree at 86–98%, while semantic_partial pairs agree at only 14–24% and contradiction pairs at 3–31% (Figure 5).

Stratified six-label agreement was highest in the Speak-Anytime injection condition (Qwen 65%, Gemma 75%) and lowest in turn-based 5-player (25–45%). This track-dependent difficulty was shared across the three models.

Three points follow. First, the Δ conclusion is not contradicted by human labels: binary agreement reaches 80% for Qwen and Gemma ($\kappa = 0.37$ – 0.47 ; Table 15), and human labels also have the injection condition above control with the same sign of Δ (Table 16). Second, absolute realization rates are underestimated: model recall is 0.21–0.47, missing half to four-fifths of human-accepted realization. The main-text statement that absolute rates depend on the judge still holds, with the added ob-

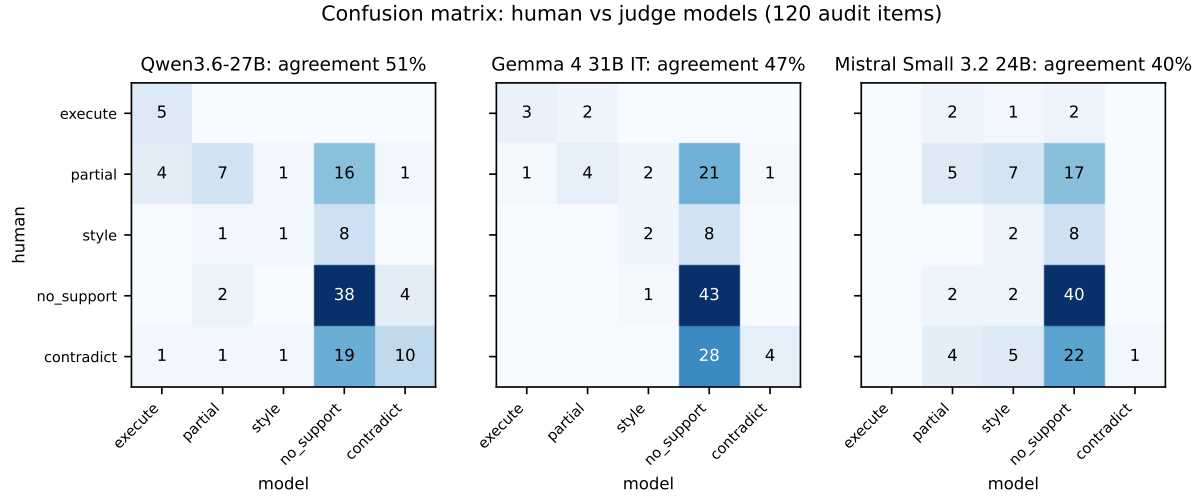


Figure 4: Confusion matrices of human labels (rows) against model labels (columns) on the 120-pair audit sample. Diagonal cells are agreements. Panel headings report strict six-label agreement (same values as Table 14). Short label names: execute = semantic_execute, partial = semantic_partial, style = style_or_persona_only, no_support = no_supported_correspondence, contradict = contradiction. unclear was never selected by human or model and is omitted. For every model, many pairs the human labeled partial or contradict fall in the model’s no_support column.

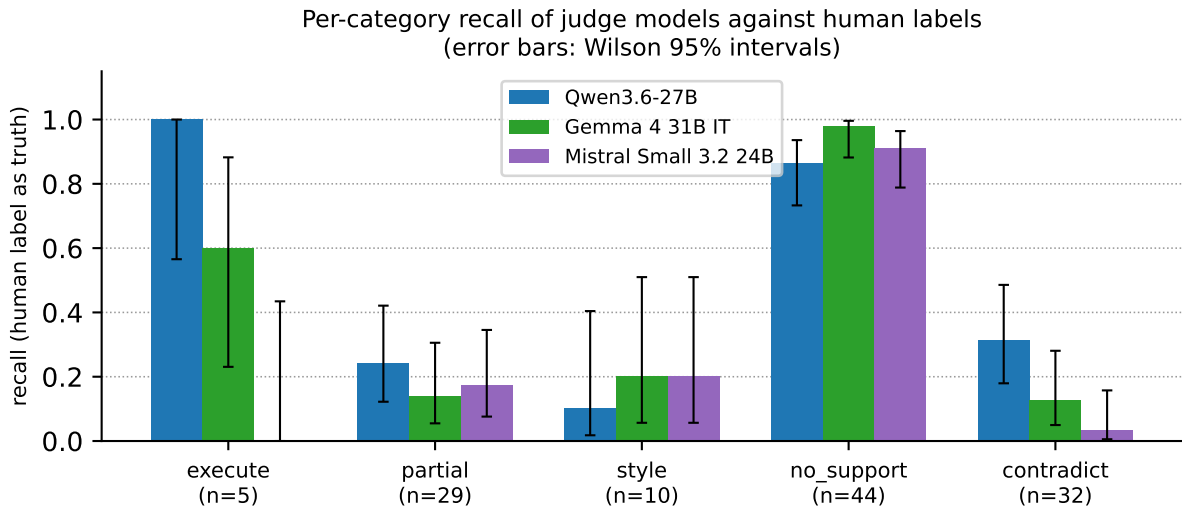


Figure 5: Per-label recall of each judge, taking human labels as the reference, on the 120-pair audit sample. Error bars are Wilson 95% intervals; n is the count per label (short names as in Figure 4). execute ($n = 5$) and style ($n = 10$) have small samples and wide intervals; consult the interval width rather than point-estimate differences. Mistral’s 0/5 recall on execute is structural: this model assigns semantic_execute to only 0.4% of all 32,800 judgments (Table 17).

servation that the dependence takes the form of the models counting realization strictly. Third, Mistral has the lowest binary κ at 0.167; that it also shows the smallest Δ supports treating its judgments as less reliable.

The audit used labels from a single author, so agreement between human annotators remains unassessed. The causes of disagreement between human and model labels, particularly between con-

tradiction and no-supported-correspondence, also remain unresolved. Each audited utterance was paired with one lesson; interactions among multiple lessons were outside the scope of the audit.

Label	Qwen	Gemma	Mistral
semantic_execute	9.6	5.4	0.4
semantic_partial	15.1	8.3	8.6
style_or_persona_only	3.4	4.3	11.7
no_supported_correspondence	58.3	77.8	78.1
contradiction	13.7	4.2	1.2
unclear	0.0	0.0	0.0

Table 17: Label frequencies (%) over the 32,800 realization judgments of each judge (all tracks and conditions). unclear was selected once, by Mistral (0.003%).

I Judging Instructions and Label Frequencies

Figure 6 shows the parts of the realization-judging instruction that concern labels and evidence filling. The actual instruction continues with a JSON output-format specification and rules for copying evidence strings character-for-character, which we omit.

Table 17 shows label frequencies over the 32,800 judgments of each judge (all tracks and conditions).

The median confidence for Qwen was 5 on no-supported-correspondence and contradiction labels and 4 on the two realization labels. The share of maximum-confidence (5) judgments was 87.0% for no-supported-correspondence and 81.8% for contradiction, confirming numerically a habit of high confidence on negative labels. Gemma showed the same pattern (92.2% and 93.9%); for Mistral the median confidence on negative labels was 4.

The source of Δ differs by judge. For Qwen, Δ comes mainly from semantic_execute: its share is 12.2% injection vs. 9.1% control in turn-based 5-player, 12.7% vs. 8.2% in turn-based 9-player, and 8.3% vs. 7.1% in Speak-Anytime, while semantic_partial differences are only +1.6, +0.4, and +0.5 points. For Gemma as well, semantic_execute accounts for most of Δ (+2.3, +3.9, +0.4 points by track). For Mistral, the semantic_execute difference is at most +0.2 points, and Δ comes mainly from semantic_partial (+1.6, +2.8, +0.7 points).

J Village Composition, Role Abilities, and Utterance Constraints

Table 18 collects details of the contest setting omitted from Section 3.

K Additional Lesson Examples

Table 19 gives the three examples not included in Table 3, in the same format.

Item	Content
Village composition	5-player: 2 Villagers, 1 Seer, 1 Werewolf, 1 Possessed. 9-player: 3 Villagers, 1 Seer, 1 Medium, 1 Bodyguard, 2 Werewolves, 1 Possessed. 9-player villages add a whisper phase visible only to werewolves.
Role abilities	The Seer learns one player’s species (human or werewolf) at night. The Medium learns the species of an exiled player. The Bodyguard protects one player from attack at night. The Possessed is human by species but plays for a werewolf-team win, so divination returns human.
Evaluation items	A Naturalness: is the talk natural? B Context: is the dialogue natural given its context? C Consistency: is talk content consistent and free of contradiction (strategically justified lies are permitted)? D Action alignment: do game actions (votes, attacks, divinations) reflect the dialogue content? E Expression: is expression rich, consistent with the profile, and showing coherent character? F Team play: does the player cooperate as a team (9-player only)?
Game start	Play starts on Day 0 daytime; Day 0 already includes greetings and policy talk.
Turn-based speaking order	Based on a random permutation.
Speak-Anytime cap	At most four utterances per agent per day, and 10 minutes for the whole talk phase.
Utterance constraints	125 characters per utterance (spaces not counted). ASCII commas are not allowed. An “@name” mention at the start addresses a specific agent; the mention itself is not counted toward the limit.

Table 18: Contest setting details.

You are an independent annotator measuring whether strategic lessons are semantically reflected in a Werewolf-game utterance.

For every task, judge all eight lesson candidates separately. Do not choose only the nearest candidate. Do not infer that a lesson was followed merely because a related word or player name appears. Compare the propositions expressed by the lesson, utterance, and preceding public context.

For each lesson, identify its object, action, condition, and purpose. For each slot use exactly one support label: explicit, weak, absent, contradicted, or unclear. Evidence must be a short exact substring of the supplied utterance or supplied preceding talk text. Use an empty string when there is no evidence. Never invent evidence. Before writing nonempty evidence, locate and copy it character-for-character from the corresponding supplied field; otherwise use an empty string.

Use exactly one overall label per lesson:

- semantic_execute: the utterance realizes the lesson’s concrete object and action; a stated condition is supported; and the intended effect is concretely readable from the speech act
- semantic_partial: a substantive part is supported but a central action, condition, or purpose is missing or weak
- style_or_persona_only: only tone, vocabulary, metaphor, or demeanor corresponds
- no_supported_correspondence: neither utterance nor context supports the lesson proposition
- contradiction: the utterance performs the opposite of the lesson’s central action
- unclear: missing input or ambiguous reference prevents judgment

Do not treat absence as contradiction. Do not evaluate whether the strategy was correct, whether the speaker won, or whether a mentioned player really had a hidden role. You are not told whether any lesson was injected or what experimental condition produced the utterance.

Figure 6: Instruction used for judging realization. Each of the eight candidate lessons requires four per-element evidence substrings and one overall label.

Pool	Base.	Inj.	Δ
[PERSONA][GENERAL] (9-players) “Maintain a critical and provocative tone even when under pressure ... a ‘bored’ or ‘judgmental’ voice”			
R1-pool	24.8	40.2	+15.4
Reorg.	15.7	32.2	+16.5
[STEER][VOTE] (5-players) “Frame a target’s focus on logical inconsistencies as ‘meanness’ or ‘fighting’ ”			
R1-pool	21.9	39.7	+17.9
Reorg.	27.1	40.8	+13.8
[STEER][COUNTER] (Speak-Anytime) “Frame the true Seer’s logical arguments as ‘bullying’ or ‘noise’ ”			
R1-pool	12.9	25.4	+12.5
Reorg.	14.5	20.3	+5.8

Table 19: Additional lesson examples. Column definitions are the same as Table 3. From top: the lesson with the second-largest Δ (two-pool average) in 9-player, the largest in turn-based 5-player, and the largest in Speak-Anytime.