BEYOND THE TRANSCRIPT: ASYMMETRIC COGNITIVE GRAPHS FOR LLM AGENTS IN SOCIAL DEDUCTION GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Social deduction requires more than identifying hidden roles: a useful belief must guide actions that preserve credibility and advance a team's objective. We introduce Asymmetric Cognitive Graphs (ACG), an external belief-to-action framework for large language model agents inspired by these demands of human strategic play. ACG separates observed events, private role beliefs, and perceived public attitudes. Probabilistic evidence updates maintain graph memory, while role-conditioned planning balances progress, information, exposure, and coordination through explicit, game-specific action scores. Deployment on either or both sides tests how this integrated intervention redistributes team advantage. We evaluate four settings: a GRAIL-aligned disclosure comparison, role-rich six-player Avalon, homogeneous nine-player Werewolf, and a nine-model Werewolf roundtable. Across three Avalon models, both-side ACG increases Good-side win rates, with two contrasts surviving multiplicity correction; DeepSeek Flash rises from 6.7% to 43.3% over 30 paired seeds (exact McNemar $p = 0.0034$). Against a GRAIL port, GPT-4.1 ACG achieves 65% versus 30% under member-attributed disclosure, but 20% versus 40% under binary outcomes; these comparisons remain descriptive. Across three homogeneous Werewolf models, Good-only deployment lowers recipient win-rate estimates, whereas Evil-only raises them. The heterogeneous roundtable yields repeatable behavioral profiles without significant paired win-rate differences. These results establish an inspectable architecture for translating social evidence into role-dependent action and reveal how the benefits of structured guidance vary across factions and interaction settings.

1 INTRODUCTION

Social deduction games expose a central challenge for language agents: sound private judgments must become actions that other players accept or counter. This makes them useful settings for studying social intelligence, strategic communication, and agent interaction (Zhou et al., 2024; Davidson et al., 2024). AvalonBench, Werewolf Arena, and Among Us demonstrate ways to evaluate language agents through hidden information, cooperation, and deception (Light et al., 2023; Bailis et al., 2024; Golechha & Garriga-Alonso, 2025). Effective play, however, requires a connection between identifying an opponent and organizing a decisive team action.

Our motivation comes from experienced human play: suspicion, public credibility, and the value of an action are different judgments. A well-informed player may conceal evidence to protect a special role; an adversary may distance itself from a teammate to preserve credibility. Intention-based agents, Werewolf reasoning methods, and theory-of-mind models already address aspects of this strategic behavior (Liu et al., 2024; Wu et al., 2024; Zhang et al., 2025). The technical challenge is to represent these distinctions under each player's information constraints and convert them into inspectable choices among commitment, probing, concealment, and coordination.

Existing approaches supply important ingredients. Memory, reflection, and reasoning–action loops support persistent agent behavior (Park et al., 2023; Shinn et al., 2023; Yao et al., 2023). GRAIL couples Bayesian graph inference to Good-side action policies in Avalon without special roles, while CSP4SDG uses constraints and information theory for identification (Rahimirad et al., 2026; Xu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

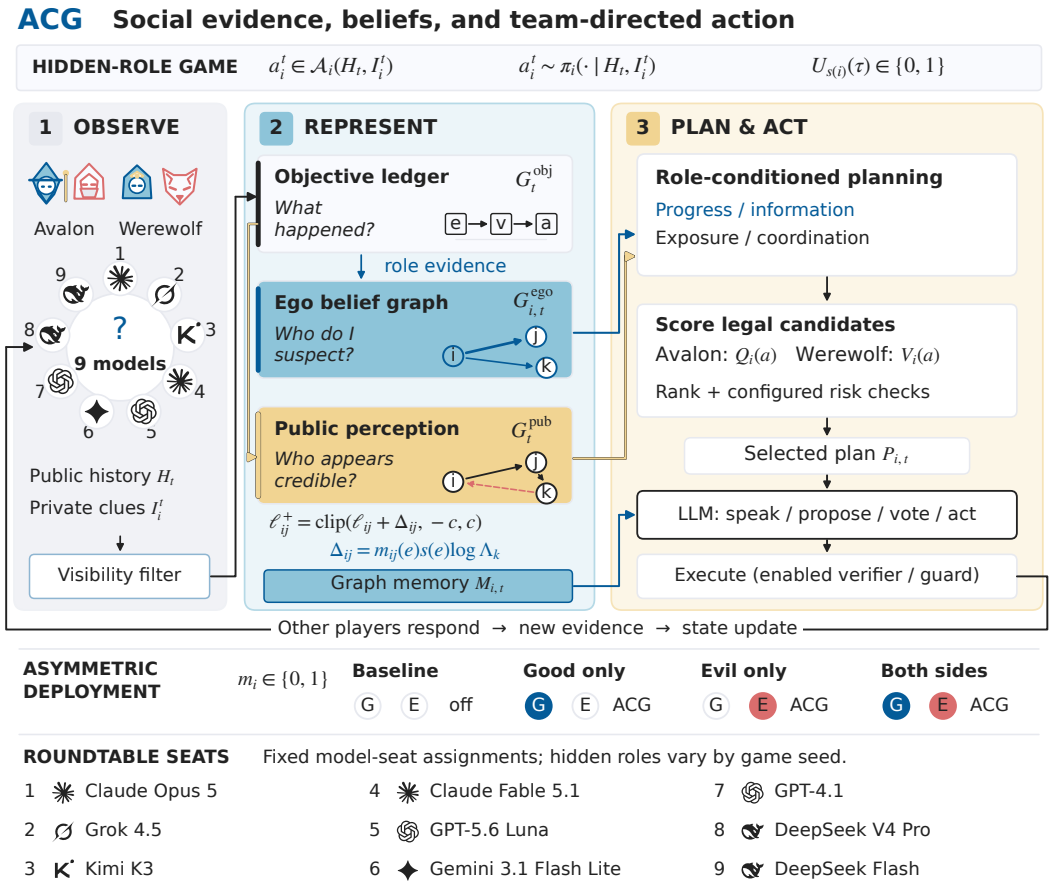

Figure 1: **ACG: an integrated belief-to-action loop.** Permitted observations update events, private beliefs, and public standing. The graph supplies compact memory and role-conditioned candidate scores (Q_i for Avalon, V_i for Werewolf). Plans condition the LLM; enabled verifiers or guards can modify proposed actions. Responses become subsequent evidence. Public standing informs action feasibility without guaranteeing cooperation. Filled G/E markers identify deployment recipients; the nine-seat roundtable shows fixed model seats with seeded hidden roles.

2026). DeepRole, reinforcement learning, and latent-policy optimization improve strategic policies (Serrino et al., 2019; Xu et al., 2024; 2025); learned discussion and persuasion explicitly target other players' responses (Jin et al., 2024; Sarkar et al., 2025; Zhang et al., 2026). Training against diverse agents further addresses strategic interaction (Ye et al., 2025). These advances motivate a specific question: does an explicit belief-to-action interface help its receiving faction when role powers, information protection, and other players' policies constrain collective success?

We introduce **Asymmetric Cognitive Graphs (ACG)**, a runtime framework that operationalizes selected considerations of human strategic play without fine-tuning the base model. Three linked views separate witnessed events, observer-specific role beliefs, and public standing. Typed speech–action relations preserve provenance and inconsistencies; probabilistic updates turn admissible evidence into beliefs. A role- and phase-conditioned planner converts this shared state into action scores, balancing progress, information, exposure, and coordination. This separation matters because privately identifying an adversary does not establish public support for an accusation. Game-specific utility scores make these trade-offs explicit; they are heuristic surrogates for a side's terminal objective. Memory and planning form one intervention, deployable to Good, Evil, both, or neither side.

Four experiment families test complementary questions. A GRAIL-port comparison varies disclosure; role-rich six-player Avalon adds special abilities and assassination; nine-player Werewolf introduces

day/night elimination and role resources; a nine-model roundtable adds heterogeneous partners and seeded social styles. Both-side ACG raises Good wins across three Avalon models, with corrected significance in two. GRAIL-port comparisons show higher observed ACG win rates under attributed actions, with the ordering reversed under binary outcomes. Homogeneous Werewolf shows opposing recipient effects for Good-only and Evil-only deployment, while H9 yields repeatable behavioral profiles without significant paired win-rate contrasts. These findings make asymmetric augmentation a test of how structured reasoning changes collective play, including settings where guidance fails to benefit its recipient.

Our contributions are (i) an explicit three-view representation separating private role judgment from public standing; (ii) a graph-conditioned planning interface with inspectable role-specific action utilities; and (iii) cross-game, paired deployment experiments exposing both gains and recipient-dependent reversals. These experiments assess the integrated belief-to-action framework and motivate future work on how its components interact with model-specific game understanding and rule complexity.

2 Related Work

Belief representations and human-inspired play. GRAIL, CSP4SDG, and MultiMind make role inference or social belief state explicit (Rahimirad et al., 2026; Xu et al., 2026; Zhang et al., 2025). InterIntent formalizes role-dependent intentions (Liu et al., 2024). ACG builds on this line by distinguishing event provenance, private role beliefs, and public standing within a shared interface for memory and action scoring.

Policy learning and deployment. DeepRole learns values for game-theoretic search; Werewolf policy learning and persuasive-communication methods optimize strategic behavior (Serrino et al., 2019; Xu et al., 2024; 2025; Zhang et al., 2026). ACG uses fixed runtime scores and evaluates who receives the intervention. Mixed-agent evaluations in GRAIL and MaKTO, and cross-model negotiations, establish relevant precedents for interaction-dependent performance (Rahimirad et al., 2026; Ye et al., 2025; Davidson et al., 2024). Our four-arm protocol measures faction outcomes separately from behavioral profiles.

3 Asymmetric Cognitive Graphs

Figure 1 connects the design problem to three computations: record admissible evidence, update observer-specific beliefs, and rank role-appropriate actions. Private role confidence and public standing remain separate inputs to planning. This operationalizes selected considerations of human strategic play through explicit state and scores; it does not imply measured human-like cognition or calibrated predictions of other players.

3.1 Setting and information discipline

Let $\mathcal{P}$ denote players, $\mathcal{R}$ hidden roles, H_t public history, and I_i^t the information legally available to player i at time t. A legal policy must condition only on the public history and that player's private information:

$$a_i^t \sim \pi_i(\cdot \mid H_t, I_i^t), \qquad a_i^t \in \mathcal{A}_i(H_t, I_i^t). \tag{1}$$

Each side s receives terminal utility $U_s(\tau) \in \{0, 1\}$ on trajectory τ. The strategic objective for player i is $J_i(\pi_i; \pi_{-i}) = \mathbb{E}_{\tau \sim (\pi_i, \pi_{-i})}[U_{s(i)}(\tau)]$. Its dependence on partners and opponents, π_{-i}, motivates side-specific deployment. This game-theoretic objective separates immediate action outcomes from terminal team value. ACG operationalizes the distinction through local utility surrogates; it does not solve for an equilibrium or directly estimate J_i.

3.2 Three linked views

At each decision point ACG maintains

$$\mathcal{S}_t = \left(G_t^{\mathrm{obj}}, \{G_{i,t}^{\mathrm{ego}}\}_{i \in \mathcal{P}}, G_t^{\mathrm{pub}}, M_t \right). \tag{2}$$

The **objective ledger** G_t^{obj} stores time-stamped events with visibility and provenance: claims, proposals, votes, actions, outcomes, eliminations, and terminal events. Each **ego graph** $G_{i,t}^{\mathrm{ego}}$ contains only information available to observer i, including role posteriors $p_{i,t}(r_j)$, derived alignment beliefs, and legal private facts. The **public-perception graph** G_t^{pub} represents public evidence and commitments.

ACG distinguishes role belief from *collective feasibility*: privately suspecting a target does not imply that others will support an accusation. Public suspicion and commitments inform action scores without certifying identity. The implementation also uses weak speech-conditioned belief updates; repeated claims are not independent evidence. Typed, directed speech–action links connect prior commitments to observable actions and update credibility. The planner reads these separate signals when selecting a target, explanation, or role action.

3.3 Belief update and action selection

ACG updates role beliefs from admissible evidence and uses the resulting state to rank actions for the observer's side. This is the purpose of the planning stage: a posterior organizes uncertainty, but does not by itself choose between commitment, information seeking, concealment, and seeking support. The following scores implement those trade-offs; speech-derived evidence weights remain heuristic.

Evidence-conditioned belief update. Let $z_j \in \{\mathrm{Good}, \mathrm{Evil}\}$ and $q_{ij,t} = \Pr_i(z_j = \mathrm{Evil} \mid I_i^t)$. Each event $e = (k, u, v, x, \nu, \chi, t)$ records type, actor, target, content, provenance, visibility, and time. A versioned table $\theta_k = (\theta_k^E, \theta_k^G)$ supplies the likelihoods; event strength $s(e)$ scales the log-likelihood ratio:

$$\begin{aligned} \log \Lambda_k &= \log \frac{\theta_k^E}{\theta_k^G}, \\ \ell_{ij,t+1} &= \mathrm{clip}(\ell_{ij,t} + m_{ij}(e_t) s(e_t) \log \Lambda_{k_t}, -c, c)\,, \\ q_{ij,t+1} &= \sigma(\ell_{ij,t+1}). \end{aligned} \tag{3}$$

The mask $m_{ij}(e)$ excludes evidence unavailable to i, irrelevant to j, or blocked by an enabled self-evidence guard. Private role knowledge adds an observer-specific log-odds term; exact roles fix the posterior. When the board fixes the Evil count, an optional common offset λ enforces $\sum_j \sigma(\ell_{ij} + \lambda) = n_{\mathrm{Evil}}$. Tables and guards are configuration parameters.

A vote can affect proposal feasibility; outcome-linked votes and attributed quest actions also enter the configured belief updates. The exact routing, joint failed-quest attribution, and parameter values are given in Appendix B.

Side-conditioned action ranking. For a proposed Avalon team T, let $\bar{q}_i(T) = |T|^{-1} \sum_{j \in T} q_{ij,t}$, $u_i(T) = |T|^{-1} \sum_{j \in T} (1 - 2|q_{ij,t} - 1/2|)$, h_i be the largest unresolved explanation debt, and s_i the observer's public suspicion. For the Good-side proposal, explanation, and quest-support actions, the implemented heuristic has the form

$$\begin{aligned} G_i(T) &= \beta_u u_i(T) + \beta_h h_i, \\ Q_i(a_{\mathrm{team}}; T) &= \beta_0 - s_i - \beta_q \bar{q}_i(T), \\ Q_i(a_{\mathrm{explain}}; T) &= \beta_e + h_i + G_i(T), \\ Q_i(a_{\mathrm{support}}; T) &= \beta_s + p_i - \beta_r \bar{q}_i(T) + G_i(T) - \beta_x [\max_{j \in T} q_{ij,t} - \tau]_+. \end{aligned} \tag{4}$$

Here p_i is phase-dependent quest pressure and $[x]_+ = \max(x, 0)$. The action symbols denote planner choices. Team risk lowers proposal and support scores; uncertainty and explanation debt raise probing scores. These are fixed, hand-set ranking preferences, not calibrated win probabilities. Appendix B.3 reports the exact values. Werewolf uses a separate utility $V_i(a)$ combining progress, information, exposure, coordination, and risk (Appendix B). Its coordination features reward proposed patterns; they do not measure whether other players will follow them.

An optional Avalon convention prior adds $\kappa \log\{\pi_{\mathrm{conv}}(a \mid r_i)/\pi_0(a)\}$ to $Q_i(a)$, with uniform π_0. The planner ranks role- and phase-specific candidates and supplies the selected plan and score

breakdown. Werewolf additionally applies its configured risk budget. Good players plan from separate information sets; known wolves can receive a shared pack plan. Neither local rankings nor shared intentions guarantee successful coordination against responding opponents.

Asymmetric deployment and decision trace. Let $m_i \in \{0, 1\}$ indicate whether ACG is enabled for player i. Let $M_{i,t}$ be graph memory and $P_{i,t}$ the plan obtained from candidate scores and configured risk rules. Then

$$\widetilde{I}_i^t = I_i^t \parallel m_i M_{i,t} \parallel m_i P_{i,t}, \qquad \widehat{a}_i^t \sim \pi_{\mathrm{LLM}}(\cdot \mid H_t, \widetilde{I}_i^t), \tag{5}$$

where $\parallel$ denotes structured concatenation and both ACG fields are omitted when $m_i = 0$. The model proposes $\widehat{a}_i^t$; enabled verifiers or guards may alter the executed action a_i^t. These controls are part of the evaluated ACG package. Logs distinguish evidence, scores, plans, proposed actions, and executed actions wherever those fields are recorded. Four side assignments test the integrated intervention, rather than assuming independent benefits from its components.

4 Experimental Design

Boards and arms. The four families vary disclosure, role abilities, elimination dynamics, and partner composition; they do not form a single ranking of game difficulty. We evaluate four families without pooling outcomes across them: (E1) a simplified six-player Avalon board with four Good and two Evil roles, no special roles, and member-attributed or binary quest disclosure; (E2) a role-hidden, asymmetric-ability six-player Avalon with Merlin, Percival, two Servants, Morgana, and Assassin, using natural quest resolution and assassination; (E3) nine-player Werewolf with three Werewolves, three Villagers, Seer, Witch, and Hunter under one model per game; and (E4) the same Werewolf rules with nine distinct model endpoints in fixed seats, with models mapped to seat numbers. Role-hidden-ability Avalon uses DeepSeek Flash, Gemini 3 Flash Preview, and GPT-5.4-mini. Homogeneous Werewolf uses DeepSeek Flash, GPT-5.6 Luna, and Gemini 3 Flash Preview, with 30 audited games per arm. Each Avalon matrix has 30 games per arm. The heterogeneous game (H9) seats Claude Fable 5.1, Claude Opus 5, DeepSeek Flash, DeepSeek V4 Pro, Gemini 3.1 Flash Lite, GPT-4.1, GPT-5.6 Luna, Grok 4.5, and Kimi K3; model-seat positions are fixed, while roles and five instruction-defined styles vary by seed and are matched across arms. This mixed ecology has 30 games per arm.

The four arms are baseline, Good-only ACG, Evil-only ACG, and both-side ACG; they vary who receives the same integrated memory, planning, and enabled execution controls. Matched seeds fix role assignment within each condition, while generated trajectories remain stochastic. Requested API model strings identify endpoints, not independently attested backend versions.

Validity and inference. Games require a natural terminal, winner, and no execution error. Werewolf auditing retains permitted first-night and same-round exile last words while excluding unauthorized post-elimination discussion. All final Werewolf matrices pass these checks; source selection is documented in Appendix E. We report Good wins, Wilson 95% intervals, exact paired McNemar tests, and Holm adjustment within each model's three deployment-versus-baseline contrasts. GRAIL-aligned comparisons remain descriptive: cell sizes differ, and GRAIL requests temperature 1.0 while other arms use endpoint defaults.

5 Results

5.1 Role-rich Avalon: a paired gain under complete evaluation

The DeepSeek Flash matrix contains 30 valid natural-terminal games per arm. Both-side deployment increases Good wins from 2/30 (6.7%) to 13/30 (43.3%). Twelve seeds are won only under both-side ACG and one only under baseline (exact McNemar $p = 0.00342$; Holm-adjusted $p = 0.0103$). Good-only reaches 9/30 (30.0%), but its raw $p = 0.0391$ does not survive correction ($p = 0.0781$). Evil-only remains at 2/30 (6.7%). Thus, the strongest evidence supports the both-side intervention.

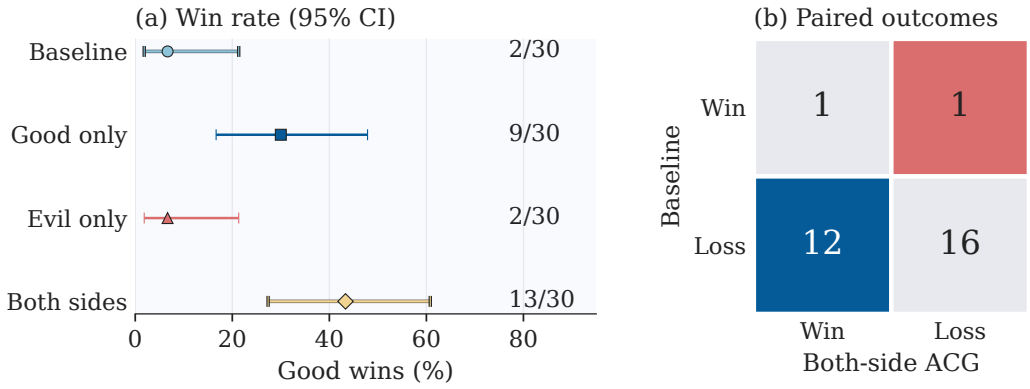

Figure 2: **Role-rich Avalon primary matrix.** Left: Good-side win rates with Wilson 95% intervals. Right: paired outcomes for the 30 seeds under baseline and both-side ACG; the off-diagonal cells give the 12-to-1 discordance used by the exact McNemar test.

Table 1: Role-rich Avalon across model endpoints. Entries are Good wins / valid natural-terminal games. Each focal endpoint has a complete, balanced four-arm matrix with 30 matched seeds.

Model	Baseline	Good only	Evil only	Both sides
DeepSeek Flash	2/30 (6.7%)	9/30 (30.0%)	2/30 (6.7%)	**13/30 (43.3%)**
Gemini 3 Flash Preview	6/30 (20.0%)	5/30 (16.7%)	6/30 (20.0%)	**12/30 (40.0%)**
GPT-5.4-mini	7/30 (23.3%)	14/30 (46.7%)	7/30 (23.3%)	**22/30 (73.3%)**

5.2 Cross-model role-rich Avalon: a replicated full-deployment direction

All three Avalon matrices are complete and balanced (Table 1). Both-side deployment raises Good wins from 6/30 (20.0%) to 12/30 (40.0%) for Gemini and from 7/30 (23.3%) to 22/30 (73.3%) for GPT-5.4-mini. The latter survives correction (Holm $p = 0.00219$); Gemini does not. Good-only results vary by endpoint, while Evil-only matches each baseline count. The shared both-side direction therefore extends beyond DeepSeek, without establishing that every recipient benefits (Appendix D).

5.3 GRAIL-aligned Avalon: rankings change across observation protocols

The simplified board compares a language-only baseline, Good-only ACG, and a GRAIL decision-chain port. Under member-attributed disclosure, ACG wins 13/20 games with GPT-4.1 versus 6/20 for GRAIL; DeepSeek-V4-Pro yields 39/40 versus 24/40. Under binary disclosure, GRAIL leads: 8/20 versus 4/20 on GPT-4.1 and 12/20 versus 6/20 on DeepSeek-V4-Pro. This observed reversal limits generalization of a method ranking across observation protocols. Member attribution supplies evidence about individual actors, while binary outcomes require reasoning from a joint result. E2 separately uses anonymous failure counts. The results concern these stated protocols; unequal cells and decoding settings preclude isolating disclosure as the cause.

5.4 Single-model Werewolf: individual guidance and team outcomes diverge

Each audited Werewolf matrix uses one endpoint for all nine seats and 30 matched seeds per arm. In baseline, Good-only, Evil-only, both-side order, Good wins are $(13, 8, 12, 6)/30$ for DeepSeek Flash, $(12, 6, 11, 1)/30$ for GPT-5.6 Luna, and $(13, 5, 8, 6)/30$ for Gemini 3 Flash Preview. Every intervention has a lower Good-win point estimate than baseline. For Evil-only deployment, this is a gain for the receiving Evil side; for Good-only deployment it is a loss for the recipient. Luna both-side is the strongest paired contrast: 11 seeds are Good wins only at baseline and none only under treatment (exact McNemar $p = 0.0010$; Holm $p = 0.0029$). The remaining homogeneous Werewolf contrasts do not survive correction. The result shows that the same intervention package can

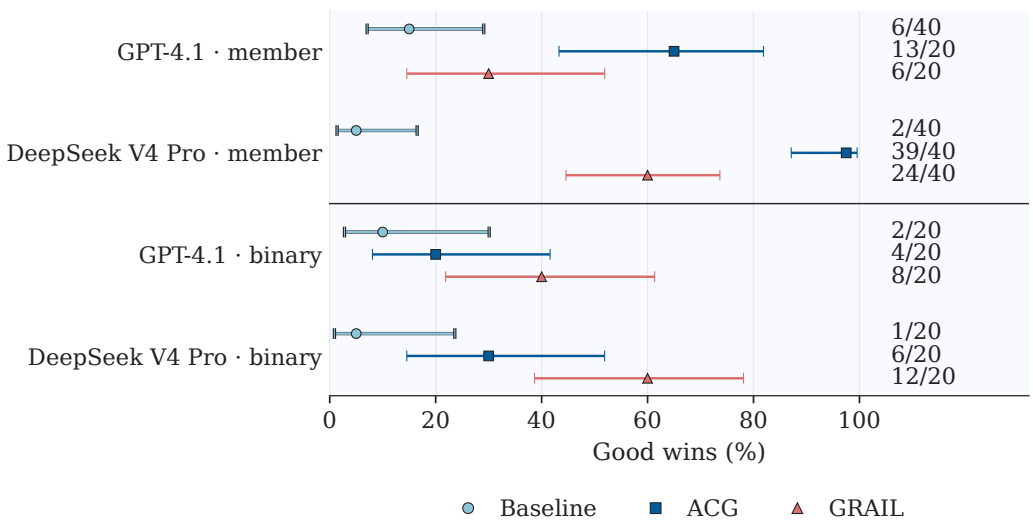

Figure 3: **GRAIL-aligned comparison across reveal conditions.** Direct labels give wins/valid games and whiskers show Wilson 95% intervals. Cell sizes and decoding settings differ between arms. Comparisons are descriptive.

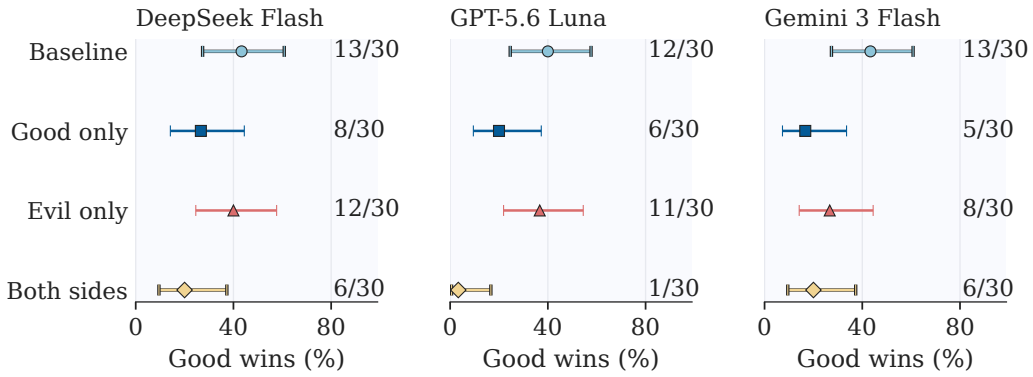

Figure 4: Three audited homogeneous nine-player Werewolf matrices. Every ACG arm has a lower Good-side point estimate than baseline; whiskers are Wilson 95% intervals. Luna both-side vs. baseline exact paired McNemar $p = 0.0010$ (Holm-adjusted $p = 0.0029$).

shift terminal advantage in the opposite direction from Avalon, without identifying which component drives that shift. Complete paired results are given in Appendix D.

5.5 A NINE-MODEL ROUNDTABLE: BEHAVIORAL PROFILES AND TEAM OUTCOMES

The 9-seat experiment places distinct model endpoints and five instruction-defined social styles in the same game. Roles and styles change across seeds; model-seat assignment does not. Seed–seat hashing fixes the style across arms (Appendix F). All 120 games reach valid natural terminals, with 30 matched seeds per arm.

Good wins are 6/30 at baseline, 5/30 with Good-only ACG, 9/30 with Evil-only ACG, and 5/30 with both-side ACG. No paired contrast with baseline is significant (exact McNemar $p = 1.000$, 0.508, and 1.000, respectively). The Evil-only arm has the highest Good-win point estimate, opposite to a simple recipient-benefits prediction. Winners differ across arms in 17 of 30 seed sets; these differences include stochastic generation and do not isolate a deployment effect.

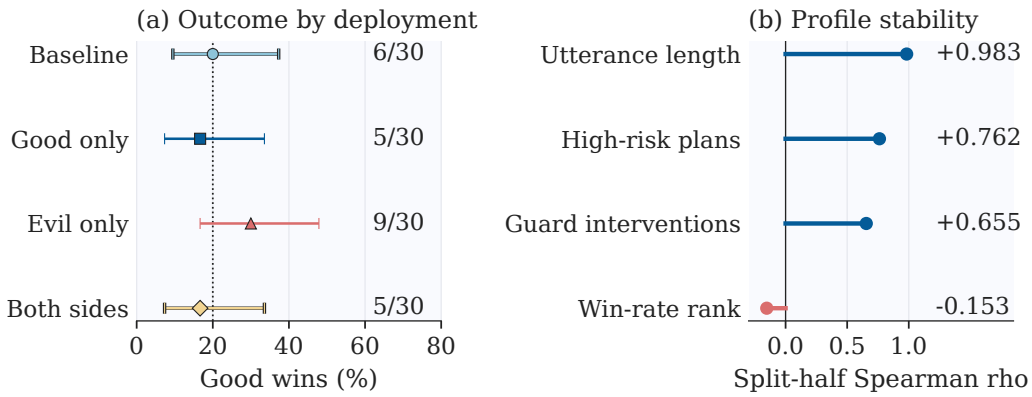

Figure 5: **Nine-model Werewolf roundtable.** Left: Good-side win rates and Wilson 95% intervals ($n = 30$ per arm); no paired contrast is significant. Right: seed-split Spearman correlations across nine fixed model-seat profiles. The profile correlations are descriptive; plan and guard measures include controller behavior.

Seed-split Spearman correlations across the nine model-seat profiles are high for mean utterance length ($\rho = 0.983$), high-risk plan rate ($\rho = 0.762$), and guard-intervention rate ($\rho = 0.655$), but not for win-rate rank ($\rho = -0.153$). Language length and planner/controller profiles are more repeatable than outcome ordering in this sample. High-risk plans need not be executed, and guard rates measure controller behavior. Fixed model-seat assignment prevents attributing these profiles to intrinsic model ability.

6 Limitations & Future Directions

ACG's deployment effects vary across games and recipient sides, motivating further study of how information access, model policies, and coordination requirements interact. The current comparisons establish effects of the complete package; they do not isolate the contributions of memory, planning, and execution controls.

From private judgment to collective action. Planning was included because a structured belief state does not specify how to win support, preserve a role, or coordinate a vote. Its usefulness depends on how other players respond. Good players infer whom to trust from separate information; wolves know their pack and can share a plan. The H9 trace in Appendix G.1 shows two wolves following a common plan while opponents reject their case, distinguishing plan execution from successful persuasion. Coordination failure is a plausible account of some reversals, alongside belief errors, exposure, and controller effects; aggregate wins do not distinguish them. Work on strategic communication and belief shifts motivates measuring acceptance and vote alignment alongside terminal utility (Jin et al., 2024; Zhang et al., 2026; Yang et al., 2026; Zhang et al., 2025).

Scope and future directions. Game skill can be measured under specified rules and opponents, but need not transfer unchanged across interactions. Davidson et al. (2024) study structured negotiations with six models in self-play and five in cross-play, finding that nominally stronger models can fare worse than weaker counterparts. This offers precedent for opponent-dependent performance without identifying the cause of ACG's reversals. The 30-seed arms have wide uncertainty; fixed seats and uneven role/style exposure limit H9 profiles. Coefficient sensitivity remains unmeasured. Larger paired evaluations, seat rotation, and controlled component and coefficient studies would clarify these boundaries. Future work can test partner-responsive planning, distill graph-conditioned decisions into smaller policies, and incorporate multimodal social cues. These extensions can examine whether plans earn support from other players as well as whether agents execute them.

7 CONCLUSION

ACG connects three-view social memory with role-conditioned planning, making the path from evidence to strategic action inspectable. Across three role-rich Avalon matrices, both-side deployment increases Good-side win rates, with corrected significance in two. In homogeneous Werewolf, Good-only deployment lowers recipient win-rate estimates, whereas Evil-only raises them. The nine-model roundtable adds repeatable behavioral profiles without establishing a context-independent ranking of model ability. Together, these experiments show why strategic augmentation should be evaluated through team outcomes and deployment asymmetry. ACG contributes both an explicit belief-to-action architecture and a controlled deployment protocol for examining how structured social reasoning changes collective play.

Reproducibility statement. Appendices A–G specify rules, implementation, runtime settings, statistics, and audit coverage. The artifact records seeds, configurations, source hashes, and outcomes; endpoint labels denote requested API identifiers.

AI assistance. We used generative AI tools for manuscript polishing, selected figure improvements, experimental-code debugging, and data collection, cleaning, and analysis, including game-log inspection and interpretation. The framework, three-layer graph architecture, algorithms, experimental design, and model selection and allocation were independently conceived and designed by the authors. Experimental trajectories and outcome counts came from the reported game runs. We checked the claims against available code, aggregate audits, and accessible event logs, and take full responsibility for the final text, analysis, and artifacts.

REFERENCES

Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf Arena: A case study in LLM evaluation via social deduction. *arXiv preprint arXiv:2407.13943*, 2024. doi: 10.48550/arXiv.2407.13943. URL https://arxiv.org/abs/2407.13943.

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *International Conference on Learning Representations*, pp. 784–809, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/03876dc6875571c18b3b09a624c77e96-Abstract-Conference.html.

Satvik Golechha and Adrià Garriga-Alonso. Among Us: A sandbox for measuring and detecting agentic deception. In *Advances in Neural Information Processing Systems*, volume 38, pp. 11025–11055, 2025. doi: 10.52202/085713-0370. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/105c4de1195135fae4974aa8c5e27bbf-Abstract-Conference.html.

Xuanfa Jin, Ziyan Wang, Yali Du, Meng Fang, Haifeng Zhang, and Jun Wang. Learning to discuss strategically: A case study on One Night Ultimate Werewolf. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77060–77097, 2024. doi: 10.52202/079017-2452. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8cea78701eb986f3ec357eb9b7c6badd-Abstract-Conference.html.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. AvalonBench: Evaluating LLMs playing the game of Avalon. *arXiv preprint arXiv:2310.05036*, 2023. doi: 10.48550/arXiv.2310.05036. URL https://arxiv.org/abs/2310.05036.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. InterIntent: Investigating social intelligence of LLMs via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6718–6746. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.383. URL https://aclanthology.org/2024.emnlp-main.383/.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://research.google/pubs/generative-agents-interactive-simulacra-of-human-behavior/.

Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis, and Joseph Campbell. Bayesian social deduction with graph-informed language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32415–32440. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1497. URL https://aclanthology.org/2026.acl-long.1497/.

Bidipta Sarkar, Warren Xia, C. Karen Liu, and Dorsa Sadigh. Training language models for social deduction with multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1830–1839. International Foundation for Autonomous Agents and Multiagent Systems, 2025. URL https://www.ifaamas.org/Proceedings/aamas2025/pdfs/p1830.pdf.

Jack Serrino, Max Kleiman-Weiner, David C. Parkes, and Joshua B. Tenenbaum. Finding friend and foe in multi-agent games. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1251–1261, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/912d2b1c7b2826caf99687388d2e8f7c-Abstract.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Wei Yang, and Haobo Fu. Enhance reasoning for large language models in the game Werewolf. *arXiv preprint arXiv:2402.02330*, 2024. URL https://arxiv.org/abs/2402.02330.

Kaijie Xu, Fandi Meng, Clark Verbrugge, and Simon Mark Lucas. CSP4SDG: Constraint and information-theory based role identification in social deduction games with LLM-enhanced inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(17):14380–14387, 2026. doi: 10.1609/aaai.v40i17.38453. URL https://ojs.aaai.org/index.php/AAAI/article/view/38453.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 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*, pp. 55434–55464, 2024. URL https://proceedings.mlr.press/v235/xu24ad.html.

Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. Learning strategic language agents in the Werewolf game with iterative latent space policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69217–69239, 2025. URL https://proceedings.mlr.press/v267/xu25h.html.

Yu-Yu Yang, Ti-Rong Wu, Hung Guei, Hsing-Yu Chen, and I-Chen Wu. Do LLMs trust the accuser or the accusation? Measuring belief shifts in Werewolf. *arXiv preprint arXiv:2609.12446*, 2026. URL https://arxiv.org/abs/2609.12446.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://react-lm.github.io/.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, Peng Sun, and Zhongyu Wei. Multi-agent KTO: Enhancing strategic interactions of large language model in language game. In *Advances in Neural Information Processing Systems*, volume 38, pp. 109078–109121, 2025. doi: 10.52202/085713-3638. URL https://papers.nips.cc/paper_files/paper/2025/hash/9d276b0a087efdd2404f3295b26c24c1-Abstract-Conference.html.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. MultiMind: Enhancing Werewolf agents with multimodal reasoning and theory of mind. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5824–5833. ACM, 2025. doi: 10.1145/3746027.3755752. URL https://doi.org/10.1145/3746027.3755752.

Zheng Zhang, Deheng Ye, Peilin Zhao, and Hao Wang. The Stackelberg speaker: Optimizing persuasive communication in social deduction games. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5520–5544. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.250. URL https://aclanthology.org/2026.acl-long.250/.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, pp. 40975–41019, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3075b88e583a0e98d8b24338a613060-Abstract-Conference.html.

APPENDIX

A GAME RULES, INFORMATION, AND DEPLOYMENT

E1–E4 denote the four experimental families. Rates refer to the Good side; families are analyzed separately. This appendix specifies the implemented rules and decision maps.

A.1 ENVIRONMENT CONTRACTS

Avalon. Both implementations use quest sizes $(2, 3, 4, 3, 4)$, at least four approvals out of six, and quest failure when at least one member submits a failure action. After five consecutive rejections, the runner executes the last proposed team. This explicit local variant differs from GRAIL's five-rejection Evil terminal (Rahimirad et al., 2026). In E2, Merlin knows the Evil players; Percival sees Merlin and Morgana as unlabelled candidates; Morgana and Assassin know one another. E1 removes special roles and assassination; Evil players know each other. Its binary condition publishes only quest success/failure, whereas member attribution publishes each action. E2 publishes the anonymous failure count. These protocols differ in whether a failure can be attributed to an individual player.

Werewolf. The nine-player preset follows the rule description in Wu et al. (2024), arXiv v2, Appendix C; its recorded version is `2026-09-20`, preset `werewolf9_swh_v1`. Night order is Werewolves, Witch, then Seer. Living wolves deliberate through three rounds of sequential pack votes and may target any living player; wolf self-destruction is permitted. The Witch has one save and one poison, cannot use both in one night, and may self-save only on night 1. The Seer cannot repeat a check. Poison suppresses the Hunter's shot. Day speaking follows a seeded random ring; votes are hidden until the phase ends. A tie triggers a second speech-and-vote procedure. First-night victims and same-round daytime exiles may give last words. Second- or later-night victims cannot speak again. Auditing distinguishes these permissions by phase and death cause.

A.2 WHAT CHANGES BETWEEN THE FOUR ARMS

Let $(m_G, m_E) \in \{0, 1\}^2$ mark Good/Evil deployment: $B = (0, 0)$, $G = (1, 0)$, $E = (0, 1)$, and $A = (1, 1)$. All Werewolf arms receive the same public-board protocol and seed–seat style instructions. Enabled sides additionally receive private graph memory, role-conditioned plans, and executable action guards; Avalon Good-side deployment also enables its proposal/vote verifier. The evaluated intervention is this complete package, including recorded guard actions despite an "advice" configuration label. Players receive only visible events and legal private facts: engine-side role labels are never planner observations.

Table 2: Environment contracts used in the experiments. "Role-rich Avalon" denotes the six-player special-role implementation identified as `std6` in the code.

	E1: simplified Avalon	E2: role-rich Avalon	E3/E4: Werewolf
Players	4 Good, 2 Evil	Merlin, Percival, 2 Servants, Morgana, Assassin	3 Werewolves, 3 Villagers, Seer, Witch, Hunter
Public outcomes	Member actions or binary quest result	Number of quest failures	Deaths, exile decisions, completed day votes; no role reveal on death
Good terminal	Three successful quests	Three successful quests followed by unsuccessful assassination	All Werewolves eliminated
Evil terminal	Three failed quests	Three failed quests or successful assassination	All Villagers or all special Good roles eliminated

B ALGORITHM, SCORING FUNCTIONS, AND PARAMETERS

B.1 EVENT ROUTING AND OBSERVER-SPECIFIC STATE

The three views and the planner form one decision pipeline. Memory presents the available social state; candidate generation and scoring use that state to choose a course of action. Public support is relevant to feasibility, but is not equivalent to private confidence in a role assignment. This distinction motivates the separate belief, exposure, and coordination features below. Events retain identifier, round, phase, actor, target, content, visibility, and provenance. Ego graphs track role beliefs and directed social relations; the public view tracks suspicion and commitments. Evil-role probability differs from alignment: a known wolf teammate is Evil and an ally. Known roles determine the role distribution, while numerical alignment scores are clipped to $[0.001, 0.999]$. Seer observations enter private knowledge with recorded values 0.93/0.07.

Table 3 supplies the hand-set weights for Eq. 3. Only visible event types are routed to a player; role-reveal entries are inactive on boards without reveals. Weak speech-conditioned updates do not make repeated claims independent evidence.

Algorithm 1: one ACG decision.

1. Ingest visible events and legal private facts with identifiers and provenance.
2. Update objective, ego, and public views; apply enabled evidence and joint-attribution rules.
3. Construct and score role/phase-specific candidate plans.
4. Append private graph/plan fields only for a player on an enabled side.
5. Parse the LLM action, apply the enabled verifier/guard, execute a legal action, and log proposals, interventions, and outcomes separately.

B.2 JOINT ATTRIBUTION FOR FAILED AVALON QUESTS

For a Good observer i, let Ω_i be the ten two-Evil subsets of the other five players. For $E \in \Omega_i$, the unnormalized prior is

$$w_0(E) = \exp\left(\sum_{j \in E} \ell_{ij}\right), \qquad w_1(E) = w_0(E) L(y \mid E, T). \tag{6}$$

Table 3: Evidence weights in the released graph implementation. A weight pair supplies $\log(\theta_k^E / \theta_k^G)$, scaled by event strength. It is not a frequency estimate from the evaluation set.

Evidence type	θ_k^E	θ_k^G
Speech support / speech attack	0.45 / 0.58	0.55 / 0.42
Proposed team member	0.52	0.48
Attributed success / failure action	0.35 / 0.90	0.65 / 0.10
Failed / successful team member	0.70 / 0.35	0.30 / 0.65
Speech–action contradiction	0.55	0.45
Approval of a failed team	0.62	0.38
Rejection of a successful team	0.56	0.44
Repeated failed-team membership	0.80	0.20
Private Evil / Good check entry	0.90 / 0.10	0.10 / 0.90
Exile vote	0.55	0.45
Vote leading to revealed wolf / Good exile	0.20 / 0.58	0.80 / 0.42
Night-kill victim	0.25	0.75
Public Evil / Good role reveal	0.90 / 0.10	0.10 / 0.90

Let $k = |E \cap T|$, let f be the observed failure count when available, and let F be the set of attributed failing members. With round-dependent failure propensity η_r, the likelihood for a failed quest is

$$L(y \mid E, T) = \begin{cases} \mathbf{1}\{F \subseteq E\}\, \eta_r^{|F|}(1-\eta_r)^{k-|F|}, & \text{members}, \\ \binom{k}{f}\eta_r^f(1-\eta_r)^{k-f}, & \text{count}, \\ 1-(1-\eta_r)^k, & \text{binary}. \end{cases} \tag{7}$$

Impossible assignments receive zero weight. Normalizing w_0 and w_1 gives marginal probabilities $p_j^{(0)}$ and $p_j^{(1)}$; the injected increment is

$$\Delta\ell_{ij} = \operatorname{logit} p_j^{(1)} - \operatorname{logit} p_j^{(0)}. \tag{8}$$

Subtracting the constrained prior marginal avoids reapplying that prior as new evidence. The implementation uses $\eta = (0.30, 0.35, 0.91, 0.94, 0.99)$, probability floor 10^{-6}, and the recorded public-logit cap $c = 1.7346$ for the enabled Avalon branch. Private terms can still move the effective posterior beyond the public cap. This enumerator is a failed-quest update, not a claim that all observations use exact joint Bayesian inference. The Werewolf matrices do not enable this Avalon-specific branch.

B.3 Avalon action scores and controller

The coefficients in Eq. 4 are

$$(\beta_0, \beta_q, \beta_u, \beta_h, \beta_e, \beta_s, \beta_r, \beta_x, \tau) = (1.75, 1.2, 0.45, 0.25, 0.5, 1.0, 1.6, 0.7, 0.55). \tag{9}$$

The intercept β_0 is a safe-team base score of 1.5 plus a 0.25 proposal bonus. Quest pressure p_i is 0.35 for a nonempty team in a vote or quest-action phase, and zero otherwise. These fixed values express design preferences; no documented parameter sweep establishes their optimality.

For the Evil branch, let s_i be public suspicion of the actor, s_j that of its known teammate, $u_i(T)$ team uncertainty, h_i explanation debt, and v the public top suspect. Let $a_T = 0.45$ when the actor or its teammate is on the proposed team, and zero otherwise. The base scores include

$$\begin{aligned} Q_i(a_{\text{ambiguous}}) &= 1 - s_i + 0.15u_i(T), & Q_i(a_{\text{defend}}) &= 0.8 - s_j, \\ Q_i(a_{\text{attack}}) &= s_v + \mathbf{1}\{v = j\}[s_j - s_i]_+, & Q_i(a_{\text{explain}}) &= 0.4 + h_i, \\ Q_i(a_{\text{sabotage}}) &= 1.2 - s_i + a_T, & Q_i(a_{\text{team}}) &= 0.7 + 0.25(1 - s_i). \end{aligned} \tag{10}$$

An enabled convention prior adds $0.15\log(\pi_{\text{conv}}/(1/7))$, with probabilities floored at 10^{-4}. Side and role-specific convention weights are averaged. The corresponding Werewolf local planner does not reuse the Avalon action prior.

Collective feasibility. The Avalon verifier combines team safety with estimated acceptability. Write $\bar{q}$ and $q_{\max}$ for private team risks, $\bar{s}_T$ for mean public suspicion, I_i for the actor's membership, c_T for overlap with the preceding team, and v_T for the historical approval signal. Then

$$\begin{aligned} S_i(T) &= 1 - 0.62\bar{q} - 0.38q_{\max}, \\ A_i(T) &= 0.52(1 - \bar{s}_T) + 0.22I_i + 0.16c_T + 0.10v_T, \\ C_i(T) &= (1-\rho)S_i(T) + \rho\{0.58S_i(T) + 0.42A_i(T)\}, \end{aligned} \tag{11}$$

where $\rho = \min(1, r_{\text{reject}}/4)$. Teams are ranked by decreasing C_i, then increasing maximum and mean risk; remaining ties use a seed-derived key. Under rejection pressure, a consensus replacement must improve this ranking without increasing maximum risk by more than 0.02. The verifier's risk and probing thresholds are 0.55 and 0.72, its acceptance margin is 0.04, and pressure-based acceptance begins after two rejections. Low-cost probing is restricted to the early, no-failure regime. Verifier records retain both the original and executed proposal or vote. This is a controller intervention in the Good-side ACG package, not merely additional language context.

B.4 Werewolf coordination utility

The Werewolf planner addresses decisions that role belief alone leaves open: whether to commit, probe, preserve a resource, or coordinate pressure. It constructs role- and phase-specific candidates

from the graph state and scores progress, information, exposure, coordination, topology, convention, and risk. For candidate a, its implemented utility is

$$V_i(a) = \mathrm{clip}_{[0,1]}\big[T_i(a) + \alpha_P P_i(a) + \alpha_I I_i(a) + \alpha_C C_i(a) + \alpha_H H_i(a) + \alpha_K K_i(a) - \alpha_E E_i(a) - \alpha_R R_i(a)\big], \tag{12}$$

with $(\alpha_P, \alpha_I, \alpha_C, \alpha_H, \alpha_K, \alpha_E, \alpha_R) = (0.44, 0.18, 0.15, 0.08, 0.08, 0.33, 0.18)$. The terms are deterministic features of the acting player's permitted state and candidate family. The following definitions describe the executed v6 branch, including its inherited candidate rules; the scores express design preferences rather than fitted win probabilities.

State variables and feature construction. Write $[x]_0^1 = \mathrm{clip}_{[0,1]}(x)$, $q_x = \Pr_i(x \text{ is Evil})$, s_x for public suspicion, N for the number alive, W for the known living wolf pack, and r for the round. Let h_x be the maximum salience of an available cognitive path containing x after its first node, divided by three and clipped to $[0, 1]$; it is zero when no such path exists. The base terminal score is

$$T_0 = \sigma(0.72 d_i + 0.06 \min(r, 5)), \qquad d_i = \begin{cases} (N-1) - 2\sum_{x \neq i,\, x \text{ alive}} q_x, & i \text{ is Good}, \\ 2|W| - N, & i \text{ is Evil}. \end{cases} \tag{13}$$

The Good margin excludes the observer, matching the implementation. The candidate generators in these runs call this base rule without an additional rule-pressure argument. Candidate-specific adjustments to T_0 are listed below.

For Good day decisions, let $e = \sum_{x \neq i,\, x \text{ alive}} q_x$, $n = \max(1, N-1)$, and let $s_i^* = s_i$ for a special Good role and zero for a Villager. The urgency feature is

$$p = \left[0.38 \left[\frac{2e - n + 1}{n}\right]_0^1 + 0.27 \left[\frac{r-1}{3}\right]_0^1 + 0.25 \left[\frac{7-N}{4}\right]_0^1 + 0.10 s_i^*\right]_0^1. \tag{14}$$

A commitment to x uses the feature vector, ordered as (T, P, I, E, C, H, K, R),

$$\mathbf{f}_{\mathrm{commit}}(x, c) = \big([T_0 + 0.44(q_x - 0.33) + 0.22p]_0^1,\ 0.32 + 0.58 q_x + 0.16 p,\ 0.03,\ 0.10 s_i,\ c,\ h_x,\ 0.18 \mathbf{1}_{q_x \geq 0.58} + 0.04 \mathbf{1}_{q_x < 0.58},\ 1 - q_x\big). \tag{15}$$

Table 4: Werewolf utility features: inputs and implemented construction. Exact candidate maps follow in Tables 5 and 6.

Variable	Available inputs	Calculation and interpretation
T_i	Living set, own side, beliefs or known pack, round	Base T_0 from Eq. 13, with a clipped candidate-family adjustment. No sampled rollout is performed.
P_i	Target belief q_x, public suspicion s_x, pressure	Candidate-specific progress credit; e.g., $0.32 + 0.58 q_x + 0.16 p$ for Good commitment, $0.48 + 0.22 s_x$ for pack pressure.
I_i	Target uncertainty, phase, candidate family	Seer checks use binary entropy $\mathcal{H}(q_x)$; a permitted discussion probe uses $\min(0.16, 0.18\mathcal{H}(q_x))$. Other families use fixed credits, including 0.03 for a Good vote.
E_i	Actor or pack public suspicion	Candidate-specific exposure cost, e.g., $0.10 s_i$ for Good commitment and $0.10 \bar{s}_W$ for pack pressure.
C_i	Candidate's coordination pattern	Fixed credit assigned to the planned pattern, e.g., 0.46 for a posterior-led Good vote or 0.92 for a common pack target. It is not measured compliance by other players.
H_i	Directed cognitive paths and their salience	h_x for the candidate target when enabled in its map, otherwise zero.
K_i	Candidate family, belief threshold, round	Local convention credit; e.g., 0.18 for a Good commitment with $q_x \geq 0.58$, otherwise 0.04. This is distinct from the Avalon log-prior term.
R_i	Target belief or fixed action-family risk	$1 - q_x$ for Good elimination candidates; fixed values for pack, probing, and resource-preservation candidates.

The posterior target maximizes q_x. Its alternative is the most publicly suspected other player if different, otherwise the second-ranked posterior target. Ties follow the code's sorted-player order. The maps below specify the remaining features; each tuple uses the same order.

A discussion probe targets the most uncertain eligible player and is available only in rounds 1–2, before a previous Good probe, with target entropy at least 0.88, maximum target belief below 0.45, and pressure at most 0.34. Probes are not day-vote candidates. In particular, the fixed vote information credit is a heuristic allowance; exile does not reveal a role in this board. The environment applies role resources and action legality after candidate construction.

For Evil, let s_w be the maximum public suspicion among living pack members and $\bar{s}_W$ their mean. The public-pressure target maximizes $s_x + 0.10h_x$ over non-pack players; the night target minimizes $q_x - 0.10h_x$ over that set. All pack members receive the same phase plan. Distancing ("bussing") assigns some wolves to vote against the most exposed teammate; hard bussing assigns all other wolves to that teammate, while the exposed wolf votes for the non-pack target.

Soft bussing requires at least two living wolves. Hard bussing and narrative sacrifice require three living wolves and $s_w \geq 0.56$; both use the hard-bus assignments, with different valuations.

Let $L = \mathrm{clip}_{[0,1]}((r-1)/4)$, $\bar{H}_i$ be mean binary belief entropy, and $\bar{s}_{\mathrm{pack}}$ mean public suspicion of living pack members. The risk budget is

$$b_i = \mathrm{clip}_{[0.32,0.92]} \begin{cases} 0.34 + 0.22\bar{H}_i + 0.15L, & i \text{ is Good}, \\ 0.38 + 0.30\bar{s}_{\mathrm{pack}} + 0.18L, & i \text{ is Evil}. \end{cases} \tag{16}$$

Table 5: Good-side candidate maps in the reported Werewolf branch. Resource actions are generated only when their stated preconditions hold.

Candidate and phase	Implemented feature assignment
Posterior / public-case vote	$\mathbf{f}_{\mathrm{commit}}(x, 0.46)$ / $\mathbf{f}_{\mathrm{commit}}(x, 0.38)$ for the two targets above.
Build a discussion case	$\mathbf{f}_{\mathrm{commit}}(x, 0.44)$ at the posterior target.
Hold the public case	$([T_0 + 0.08p]_0^1,\ 0.08,\ 0.08,\ 0.04s_i,\ 0.12,\ h_x,\ 0,\ 0.08)$.
Discussion probe	$([T_0 + 0.04]_0^1,\ 0.10,\ \min(0.16, 0.18\mathcal{H}(q_x)),$ $0.04s_i,\ 0.14,\ h_x,\ 0,\ 0.24)$
Seer check	$(T_0,\ 0.12 + 0.16\mathcal{H}(q_x),\ \mathcal{H}(q_x),\ 0.10s_i,\ 0,\ h_x,\ 0,\ 0.16)$. The generator prioritizes a stored private Evil-check entry of at least 0.9 when present, otherwise maximum uncertainty.
Witch save / hold	Save the current attack victim: $(T_0,\ 0.34,\ 0.05,\ 0,\ 0,\ 0,\ k_r,\ 0.12)$ with $k_r = 0.35$ on night 1, otherwise 0.10. Hold: $(T_0,\ 0.06,\ 0.10,\ 0,\ 0,\ 0,\ 0,\ 0.04)$.
Witch poison / Hunter shot	Poison at top $q_x \geq 0.58$: $(T_0,\ 0.38q_x,\ 0.05,\ 0.08s_i,\ 0,\ h_x,\ 0,\ 1 - q_x)$. Shot: $(T_0,\ 0.50q_x,\ 0,\ 0,\ 0,\ h_x,\ 0,\ 1 - q_x)$.

Table 6: Evil-side candidate maps. The v6 configuration uses these inherited pack rules; coefficients describe the implemented heuristic. Tuples are (T, P, I, E, C, H, K, R).

Candidate	Feature assignment and eligibility
Night attack	$(T_0,\ 0.52,\ 0.04,\ 0.08\bar{s}_W,\ 0.90,\ h_x,\ 0,\ 0.14)$.
Common public pressure	$(T_0,\ 0.48 + 0.22s_x,\ 0.03,\ 0.10\bar{s}_W,\ 0.92,\ h_x,\ 0,\ 0.16)$.
Soft bus	$([T_0 + 0.16s_w]_0^1,\ 0.24 + 0.38s_w,\ 0.10,$ $[\bar{s}_W - 0.45]_+,\ 0.76,\ 0,\ 0.14,\ 0.52)$
Hard bus	$([T_0 + 0.30s_w]_0^1,\ 0.18 + 0.60s_w,\ 0.10,$ $[\bar{s}_W - 0.55]_+,\ 0.94,\ 0,\ 0.22,\ 0.72)$
Sacrifice for narrative	$([T_0 + 0.38s_w]_0^1,\ 0.12 + 0.68s_w,\ 0.16,$ $[\bar{s}_W - 0.60]_+,\ 0.99,\ 0,\ 0.28,\ 0.76)$

The highest-utility candidate is selected unless its risk exceeds b_i and its utility advantage over the best within-budget candidate is below 0.04. Up to five candidate summaries are retained. The action guard compares the selected contract with the parsed model action in executable phases; its margin threshold is 0.08, subject to phase eligibility and the logged commitment rule. It can alter targets while leaving generated discussion intact.

The audit records 59 interventions among 2,815 guard records in the H9 cohort. This observation is why we report the guard as an implemented component and label the associated profile statistic a controller measure. It should not be interpreted as 59 independently randomized interventions.

A recorded candidate calculation. In H9 seed 29, Evil-only deployment, Player2's day-2 discussion plan compares common public pressure with soft bussing. The recorded components are shown in Table 7. For the selected pressure plan,

$$\begin{aligned} V_i = \big[&0.2108 + 0.44(0.6039) + 0.18(0.03) + 0.15(0.92) \\ &+ 0.08(0.3533) + 0.08(0) - 0.33(0.0408) - 0.18(0.16)\big]_0^1 \simeq 0.6059. \end{aligned} \tag{17}$$

The soft-bus score is 0.4997. Both risks, 0.16 and 0.52, are below the recorded budget 0.5475, so the budget does not change the ranking. The score margin is 0.1062, but discussion is not an executable guard phase. Both alternatives preview a vote for Player1 for this actor; their pack-level coordination pattern and valuation differ. The selected contract is passed to the language agent as planning context, not recorded as an actual vote at this point. Appendix G.1 follows the subsequent speech and ballots.

C Prompts, runtime settings, and the GRAIL port

C.1 Runtime gates and prompt composition

English prompt examples and the treatment boundary. Table 8 identifies the input components shared by baseline and ACG. The excerpts below retain the English text in the released templates; they are not translations of the original trajectories. For the concrete example, the source is H9 seed 29, Player2, day-2 discussion. The player knows its own wolf role and pack membership, sees the deaths of Player8 and Player3, and has access to prior public speech and voting. Its assigned style is evidence analyst. The acting wolf is not told that Player1 is Seer or that Player8 was Witch.

Table 7: Logged feature values for the worked example. Four-decimal storage can cause small rounding differences when reconstructing the score. Source: coordination-history entry 12 (zero-based) of the selected H9 Evil-only seed-29 record.

Plan	T	P	I	E	C	H	K	R	V
Pressure	.2108	.6039	.03	.0408	.92	.3533	0	.16	.6059
Soft bus	.2762	.3952	.10	0	.76	0	.14	.52	.4997

Table 8: Prompt components in the protocol-matched Werewolf comparison. "ACG" means a player on the enabled side. Both receive the same kinds of public information, although their generated game histories can diverge.

Input component	Baseline	ACG
Rules, own role, permitted role-specific private facts	Yes	Yes
Visible history and shared public evidence board	Yes	Yes
Seed–seat style and phase-specific output contract	Yes	Yes
Private ego summary, public-perception summary, salient graph paths	No	Yes
Selected coordination plan, score, risk budget, own directive	No	Yes
Private pack assignments, for a wolf recipient	No	Yes

Shared style and discussion instruction (template text).

> Public play style [evidence_analyst]: Build a short evidence chain from public speeches, votes, or night outcomes before choosing a candidate. Challenge unsupported calls. In your Speak field, give a candidate or defense target, at least two public observations, and a contingent call such as 'vote PlayerX unless PlayerY explains Z'. High-risk play is welcome only when you state its payoff, trigger, and exit condition. Do not invent private actions or emit vote/night-action JSON during discussion.

Shared voting instruction (phase-specific suffix).

> Vote for the player that best serves your side's long-horizon win condition. Use only the required vote JSON; do not output discussion or night-action fields.

Private ACG excerpt (code-rendered from the logged plan). The following lines are produced by the released formatter from coordination-history entry 12; the initial memory instruction comes from the environment template. The conversation file does not archive the complete API request, so this is a reproducible rendering of the recorded plan, not a claim to recover the full sent prompt. Other graph fields are omitted rather than populated from the final-state snapshot.

> Private cognitive graph memory. Use it as structured private memory for strategy, but do not reveal it directly.
> Coordination-and-Utility plan [cul_w9_swh_v6], scope=wolf_pack_private
> Goal: Keep a coherent public exile narrative while preserving enough living pack members to reach the evil terminal condition.
> Selected strategy: pack_safe_pressure (safe); utility=0.6059, risk=0.5475 budget.
> Executable directive for this phase: vote vote=Player1.
> Pack assignments: Player2:vote:Player1; Player9:vote:Player1.

Here the printed risk value is the *budget*; candidate risk is 0.16. The formatter retains the word "directive" during discussion, but the guard is ineligible in that phase: the message supplies a proposed vote target for the speech, and the ballot occurs later. The baseline receives the shared blocks without this private ACG block. The artifact includes all five exact style instructions and these excerpts as plain text.

The stored top-k is 3. The 1,200-character compact-memory setting is inactive, so it is not an enforced input cap. Equal output budgets within a cohort do not imply equal context lengths.

C.2 Models and decoding

Table 9 gives the recorded Werewolf phase budgets. The model names identify requested endpoints. The Pro aliases `deepseek-chat` and `deepseek-v4-pro` are normalized to DeepSeek V4 Pro following the experimenter's confirmation; archived request names are preserved. A returned field that echoes the request is not independent evidence of the backend model version.

Generic Avalon reports reuse Werewolf-named budget fields; these do not establish Avalon call limits. Unset temperature and unreported reasoning-token counts remain unknown. Cross-model comparisons are not compute-matched.

Table 9: Werewolf phase output-token limits in the per-game runtime configuration. H9 refers to the common runner configuration; endpoint behavior may differ.

Model/cohort	Wolves	Seer	Witch	Discuss	Vote	Hunter
DeepSeek Flash	240	60	90	260	120	60
GPT-5.6 Luna	240	60	90	260	120	60
Gemini 3 Flash Preview	600	600	600	600	600	600
H9 roster	240	60	90	260	120	60

C.3 Official GRAIL decision chain in a local board

The included GRAIL arm uses the factor-graph model, calibrated checkpoint, heuristic oracle, and language-prior/prompt components from the reference decision chain (Rahimirad et al., 2026). Its adapter encodes completed quest teams, votes, and outcomes into the reference state representation and invokes the oracle inside our environment. The shipped calibrated checkpoint is `ego_model_2_calibrated.pth`; its file hash and the adapter source identify the snapshot. This arm is distinct from the earlier threshold-based approximation, which is excluded from the reported E1 cells.

The adapter explicitly sets temperature to 1.0 for its language calls, while the other compared arms use endpoint defaults. It also supplies bounded parsing retries for the language prior and a neutral-prior fallback, and makes an oracle random choice seed-dependent. The local board forces quest execution after five rejected proposals; the adapter documents that the reference engine instead ends the game in Evil's favor. Evil players in this comparison use the local language policy, not the reference oracle's always-fail quest policy. These differences, along with the altered member-attribution channel, preclude reading E1 as a direct reproduction of the official benchmark score.

D Complete outcomes and statistical definitions

D.1 Rates and pairing

Tables 10–11 report counts and Wilson 95% intervals. Pairing uses family, model/roster, board, condition, and seed; each four-arm matrix shares 30 role assignments across arms. Remote generation remains stochastic.

D.2 Paired contrasts and multiplicity

Let n_{ab} count seed blocks in which baseline has Good-win indicator a and treatment has indicator b. With $d = n_{01} + n_{10}$, the exact two-sided McNemar value is

$$p = \min\left(1, 2 \sum_{j=0}^{\min(n_{01},n_{10})} \binom{d}{j} 2^{-d}\right), \tag{18}$$

and is one when $d = 0$. Holm adjustment is applied to the three deployment-versus-baseline contrasts within each listed model–game matrix. These are explicitly defined correction families, not a correction over every analysis in the paper. Additional model comparisons and profile analyses remain exploratory. The archived materials do not establish a preregistration, so "primary" identifies the main reported comparison rather than a preregistered analysis.

Table 10: E1 complete descriptive results. Rates and interval endpoints are percentages. ACG denotes Good-side deployment; unequal denominators are retained.

Model	Reveal	Method	Wins/n	Rate	95% CI
GPT-4.1	members	Baseline	6/40	15.0	[7.1, 29.1]
GPT-4.1	members	ACG	13/20	65.0	[43.3, 81.9]
GPT-4.1	members	GRAIL	6/20	30.0	[14.5, 51.9]
GPT-4.1	binary	Baseline	2/20	10.0	[2.8, 30.1]
GPT-4.1	binary	ACG	4/20	20.0	[8.1, 41.6]
GPT-4.1	binary	GRAIL	8/20	40.0	[21.9, 61.3]
DeepSeek V4 Pro	members	Baseline	2/40	5.0	[1.4, 16.5]
DeepSeek V4 Pro	members	ACG	39/40	97.5	[87.1, 99.6]
DeepSeek V4 Pro	members	GRAIL	24/40	60.0	[44.6, 73.7]
DeepSeek V4 Pro	binary	Baseline	1/20	5.0	[0.9, 23.6]
DeepSeek V4 Pro	binary	ACG	6/20	30.0	[14.5, 51.9]
DeepSeek V4 Pro	binary	GRAIL	12/20	60.0	[38.7, 78.1]

Table 11: Audited four-arm cohorts. Each entry gives Good wins/n above its Wilson 95% interval in percent. Source selection is described in Appendix E.

Model	B	G	E	A
E2: role-rich Avalon				
DeepSeek Flash	2/30 [1.8, 21.3]	9/30 [16.7, 47.9]	2/30 [1.8, 21.3]	13/30 [27.4, 60.8]
Gemini 3 Flash	6/30 [9.5, 37.3]	5/30 [7.3, 33.6]	6/30 [9.5, 37.3]	12/30 [24.6, 57.7]
GPT-5.4-mini	7/30 [11.8, 40.9]	14/30 [30.2, 63.9]	7/30 [11.8, 40.9]	22/30 [55.6, 85.8]
E3: homogeneous Werewolf				
DeepSeek Flash	13/30 [27.4, 60.8]	8/30 [14.2, 44.4]	12/30 [24.6, 57.7]	6/30 [9.5, 37.3]
GPT-5.6 Luna	12/30 [24.6, 57.7]	6/30 [9.5, 37.3]	11/30 [21.9, 54.5]	1/30 [0.6, 16.7]
Gemini 3 Flash	13/30 [27.4, 60.8]	5/30 [7.3, 33.6]	8/30 [14.2, 44.4]	6/30 [9.5, 37.3]
E4: heterogeneous Werewolf				
Nine-model roster	6/30 [9.5, 37.3]	5/30 [7.3, 33.6]	9/30 [16.7, 47.9]	5/30 [7.3, 33.6]

Table 12: Complete paired comparisons for the seven audited four-arm matrices, $n = 30$ paired seeds per contrast. n_{01} is a treatment-only Good win; n_{10} is a baseline-only Good win. Holm correction is within the three contrasts for each matrix.

Game	Model	Contrast	n_{01}	n_{10}	Exact p	Holm p
Avalon	DeepSeek Flash	G-B	8	1	0.039	0.078
Avalon	DeepSeek Flash	E-B	1	1	1.000	1.000
Avalon	DeepSeek Flash	A-B	12	1	0.0034	0.010
Avalon	GPT-5.4-mini	G-B	9	2	0.065	0.131
Avalon	GPT-5.4-mini	E-B	2	2	1.000	1.000
Avalon	GPT-5.4-mini	A-B	17	2	0.0007	0.0022
Avalon	Gemini 3 Flash	G-B	3	4	1.000	1.000
Avalon	Gemini 3 Flash	E-B	4	4	1.000	1.000
Avalon	Gemini 3 Flash	A-B	9	3	0.146	0.438
Werewolf	DeepSeek Flash	G-B	5	10	0.302	0.604
Werewolf	DeepSeek Flash	E-B	6	7	1.000	1.000
Werewolf	DeepSeek Flash	A-B	4	11	0.118	0.355
Werewolf	GPT-5.6 Luna	G-B	5	11	0.210	0.420
Werewolf	GPT-5.6 Luna	E-B	6	7	1.000	1.000
Werewolf	GPT-5.6 Luna	A-B	0	11	0.0010	0.0029
Werewolf	Gemini 3 Flash	G-B	2	10	0.039	0.116
Werewolf	Gemini 3 Flash	E-B	4	9	0.267	0.267
Werewolf	Gemini 3 Flash	A-B	4	11	0.118	0.237
H9	Nine-model roster	G-B	3	4	1.000	1.000
H9	Nine-model roster	E-B	6	3	0.508	1.000
H9	Nine-model roster	A-B	4	5	1.000	1.000

D.3 WHICH SIDE BENEFITS FROM AN INTERVENTION?

Let $W_G(m_G, m_E)$ be the Good-win probability. Since terminal team outcomes are complementary, an Evil recipient's gain is measured in the opposite direction:

$$\Delta_G = W_G(1,0) - W_G(0,0), \qquad \Delta_E = W_G(0,0) - W_G(0,1). \tag{19}$$

Thus, a lower Good rate under Evil-only deployment can benefit the receiving Evil side. For H9, the point estimates are $\widehat{\Delta}_G = -1/30$ and $\widehat{\Delta}_E = -3/30$; neither paired contrast is significant. Both-side deployment changes the game ecology, and cannot raise both sides' complementary terminal win probabilities simultaneously.

E ELIGIBILITY, PROVENANCE, AND AUDIT COVERAGE

Each experimental cell contributes one selected run with a rule-defined terminal, recognized winner, no execution error, and passing protocol checks. The chronological Werewolf audit distinguishes permitted first-night and same-round exile last words from unauthorized later discussion or voting. It checks observable execution and participation, not every semantic claim in generated speech.

The three 120-game Werewolf cohorts retain/replace 108/12 DeepSeek cells, 95/25 Luna cells, and 98/22 Gemini cells. Ineligible attempts were replaced within the same arm–seed cell, using the first eligible completion independently of its winner; retries are not extra samples. The isolated DeepSeek correction rebuilds the living-player speaker ring after a Hunter shot. Its twelve replacement cells preserve role/style assignment, requested model, prompt configuration, phase budgets, and method flags. The artifact supplies the detailed cell mappings, prior and selected file hashes, configuration comparisons, and source patch, while retaining superseded records outside the analysis denominator.

The final manifest contains 1,160 selected games: 320 E1, 360 E2, 360 E3, and 120 E4, all passing the stated checks. The seven balanced four-arm matrices contribute 840 games and retain matched seed blocks. Per-game configurations identify runtime flags; the supplied source and endpoint aliases cannot reconstruct unrecorded launch-time revisions or provider changes. These provenance limits apply throughout the artifact.

F NINE-MODEL ROUNDTABLE AND BEHAVIORAL PROFILES

F.1 FIXED SEATS, SEEDED ROLES, AND SEEDED STYLES

The roster in Table 13 is fixed across the 120 games. A 30-seed role assignment is reused across the four arms. The five social-style instructions are role-blind, but are not fixed personality labels attached permanently to a model. For seed s and player identifier i, the implementation selects

$$\mathrm{style}(s,i) = \mathcal{L}\big[\,\mathrm{SHA256}(s:i)_0 \bmod 5\big], \tag{20}$$

where the subscript denotes the first digest byte. The ordered library $\mathcal{L}$ is captain, evidence analyst, coalition builder, contrarian probe, and resource guardian; exact English instructions are in the artifact. Recorded assignments match this rule. Styles are shared across arms within a seed–seat, but need not be balanced across seeds.

F.2 PROFILE DEFINITIONS AND SPLIT-HALF PROCEDURE

Sort the 30 seeds numerically and split them into the first and last 15, identically across arms. Combining four arms gives each model-seat 60 games per half. We aggregate numerators and denominators before computing:

- **Discussion length:** Unicode characters divided by player messages labelled discussion, including permitted last words with that label; Moderator messages are excluded.
- **High-risk plan rate:** high-risk selected modes divided by coordination records; this describes plans rather than executed risky actions.
- **Guard-intervention rate:** flagged interventions divided by all guard checks, including unchanged actions.

Table 13: H9 roster and role counts over the 30 unique seed assignments. Each assignment appears in all four arms. W/V/S/Wi/H denote Werewolf, Villager, Seer, Witch, and Hunter. These counts are not 120 independent assignments.

Seat	Model endpoint	W	V	S	Wi	H
P1	Claude Opus 5	9	9	4	3	5
P2	Grok 4.5	11	10	2	4	3
P3	Kimi K3	5	13	3	5	4
P4	Claude Fable 5.1	10	11	2	3	4
P5	GPT-5.6 Luna	11	11	2	1	5
P6	Gemini 3.1 Flash Lite	12	6	6	4	2
P7	GPT-4.1	7	12	5	2	4
P8	DeepSeek V4 Pro	10	9	6	4	1
P9	DeepSeek Flash	15	9	0	4	2

- **Team win rate:** games won by the seat's assigned side divided by its games.

Spearman correlations compare the nine seat ranks across halves, using average ranks for ties. Zero denominators remain missing; planner/guard statistics use active-module records. The 1,080 player–game rows are dependent within games. Fixed model seats and uneven role/style frequencies preclude an intrinsic personality or ability ranking.

The supplementary artifact retains all 18 profile correlations, both halves' seat-level values, scatter plots, and every seed's outcome. Seventeen of thirty seed blocks have different winners across arms; deployment and stochastic generation both contribute to these differences. These detailed displays are indexed in the artifact's extended-details guide.

G TRACEABLE EXAMPLES AND REPRODUCTION

G.1 A HETEROGENEOUS DECISION: EVIDENCE, PLAN, ACTION, AND RESPONSE

We retain the smallest seed with an Evil baseline win and a Good Evil-only win, seed 29, as a descriptive matched case. The two games share roles: wolves P2/P3/P9, Seer P1, Witch P8, Hunter P4, and Villagers P5/P6/P7. Both exile P5 on day 1; subsequent trajectories diverge. The following trace follows Player2's day-2 decision in the Evil-only game, connecting the numerical example in Appendix B.4 to its observed execution. Full roles here are analyst annotations; deaths did not publicly reveal them.

Both wolves execute the planned target choice, but four opponents vote out P2; P4 explicitly challenges the repeated case. Thus plan agreement within the pack does not ensure persuasion outside it. The coordination score rewards a proposed pack pattern, rather than predicting opponents' responses. Claims that P5 was "confirmed" Villager are player assertions, not role reveals. Candidate scores and contracts are saved, but the full decision-time graph input is not; the belief history is an event-sourced rebuild. This trace documents the belief–plan–action interface where available and its observed reception, without supplying unrecorded beliefs or establishing a causal explanation of the cohort's win rates.

Offline reproduction. The artifact README maps equations to code and lists offline commands for checking outcomes, paired statistics, candidate scores, and the quoted events. Full manifests, profile tables, and an additional Avalon verifier example are indexed in the extended-details guide.

Table 14: H9 seed 29, Evil-only: a trace from available evidence to a plan and subsequent responses. All indices are zero-based within the selected game file. "Plan" indexes coordination history; "event" indexes game events.

Stage	Recorded evidence	Source
Before the decision	P5 was exiled and claimed Villager in last words. P8 and P3 are publicly announced dead. P2 privately knows P3 was a wolf teammate; six players remain, including wolves P2/P9.	Events 55–56, 69–70; own pack
Recorded plan	Common pressure on P1 scores 0.6059 versus 0.4997 for soft bussing. P2 and P9 are assigned P1; risk budget is 0.5475.	Plan 12
Generated speech	P2 says, "I accuse Player1. Vote Player1 unless a tighter public contradiction appears before ballot." P9 also calls for P1.	Events 71, 75
Public response	P4 challenges the repeated case: "Player2's persistence after the Day 1 result suggests coordination, not analysis." P1, P7, and P6 also call for P2.	Events 79, 83, 87, 91
Public commitments	The board records four intended voters for P2 and two for P1. These are declared intentions, separate from hidden-role evidence.	Public-board entry 1
Ballots and outcome	P2/P9 vote P1; P1/P4/P6/P7 vote P2. Both wolf ballots match the updated vote contract without guard intervention. P2 is exiled.	Events 97–103; plans 14–15; guard log