LURK Werewolf: Evaluating Linguistic Bias in Large Language Models

Lisa Liu*, Runqi Zou*, Shuxuan Kuang, Yuchen Li, Tim Merino, and Julian Togelius
New York University
New York, NY, USA
ll5937@nyu.edu, runqi.zou@nyu.edu, sk11789@nyu.edu, yl6394@nyu.edu, tm3477@nyu.edu, julian@togelius.com

***Abstract*—Existing multi-agent frameworks have agents generate messages as free-form language, entangling their intended meaning with their linguistic delivery, and making it impossible to attribute behavioral differences to either attribute alone. This study disentangles the two by representing agent speech with a structured predicate vocabulary and generating natural language from them deterministically; linguistic difference across conditions is introduced only at a controlled rewriting stage. We investigate two research questions. First, does expressing the same content in natural language change LLM decisions compared to the predicate-only form? In the most commonly studied 7-player setup of *Werewolf*, the introduction of natural language alters 52.6% of decisions made by agents. Additional 8-12 player setups reflect similar findings. Second, we investigate whether varying a single linguistic feature, formality, introduces bias in how a player is treated by others. Although no bias was found for this feature, formality is only one of many linguistic dimensions along which speakers vary. These findings demonstrate the need to account for linguistic form, in addition to content, when evaluating fairness in LLM agent systems.**

***Index Terms*—LLM agents, Werewolf, social deduction games, gameplay, logic predicates, multi-agent evaluation**

I. Introduction

As large language models (LLM) are used increasingly in multi-agent settings involving cooperation, negotiation, deliberation, and decision-making [1], [2], it has become more important to understand how their decisions are influenced by the content being communicated, as opposed to how the content is being worded.

Surface-level changes already introduce bias: varying gender information shifts LLM agent decisions in *Werewolf* [3], and template wording distorts counterfactual bias measurement even under fixed factual content [4]. If wording alone changes whom an agent supports, suspects, or acts against, these systems embed linguistic bias that is uncontrolled by design. We investigate this problem in *Werewolf*, also known as *Mafia*, a popular social deduction game in which two factions [5]–[7], the informed minority (werewolves) and uninformed majority (villagers), must identify and eliminate each other through public deliberation, voting, and role-specific skills. The game is driven largely by natural-language interaction, leaving room for linguistic bias, where an agent's style of speech and not the strength of its argument, may determine whether it is trusted, accused, or eliminated. To isolate that influence, logic predicates serve as an intermediate representation in a three-stage speech generation process that decouples strategic content from linguistic form. This process forms the basis of Language Understanding Replay Kit via *Werewolf* (LURK-Werewolf)[1], an evaluation framework that produces experimental evidence of linguistic bias through the following research questions:

RQ1. Does the presence of natural language alone cause LLM agents to change their behavior?

RQ2. Does varying a single linguistic feature, such as formality, change how a player is treated by others?

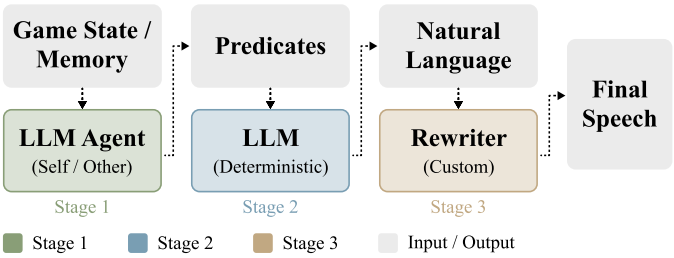

Fig. 1. Three-stage speech pipeline. Stage 1 runs in all conditions. Stage 2 is added for natural-language conditions. Stage 3 is where the researcher may introduce their custom linguistic feature to study.

II. Related Work

LURK-Werewolf draws on two areas: ethical investigation of LLM behavior and predicate-based representations of game communication. For the first, several benchmarks study social bias in LLM outputs through controlled text examples. StereoSet measures stereotypical preference together with language modeling ability [8]. CrowS-Pairs tests whether models prefer stereotypical statements about disadvantaged groups [9]. BBQ shows that question answering models often rely on stereotypes when the context is ambiguous or underinformative [10]. These benchmarks mainly evaluate model behavior in static text settings. This paper extends bias evaluation to multi-agent interactive play, where each agent's words can affect the decisions of every other player within a round. For the second area, structured meaning representations have long been used to separate content from wording. AMR [11] maps sentences to rooted graphs that normalize away surface

*Lisa Liu and Runqi Zou contributed equally.
[1]https://github.com/lisasiliu/LURK_Werewolf

* 979-8-3315-9476-3/26/$31.00 ©2026 IEEE

TABLE I
LURK-WEREWOLF PREDICATE INVENTORY. FULL GRAMMAR (ARGUMENTS, TYPES, SEMANTICS) IS IN THE PROJECT REPOSITORY.

Category	Predicates
Speech (16)	`claim_role`, `estimate`, `claim_divine`, `claim_guard`, `claim_antidote`, `claim_poison`, `question`, `answer`, `agree`, `disagree`, `allied`, `unallied`, `dissociate`, `recommend_action`, `address_role`, `fluff`
Action (7)	`divine`, `guard`, `antidote`, `poison`, `shoot`, `explode`, `pass`
Sheriff (3)	`run_sheriff`, `sheriff_order`, `sheriff_successor`
System (8)	`executed`, `attacked`, `voted`, `sheriff_set`, `game_status`, `announcement`, `hunter_shot`, `wwk_explode`

variation, and semantic parsing treats language as a mapping to executable meaning representations [12]. In *Werewolf*, Nishizaki and Ozaki [13] converted human game logs into predicates for behavior analysis. These approaches all map from language onto structure. LURK-Werewolf inverts the direction: agents produce predicates first, and natural language is generated from them, enabling comparisons across cases.

III. FRAMEWORK

A. Predicate System

Nishizaki and Ozaki's 12-predicate *Werewolf* inventory is adapted and extended into a 34-predicate system (Table I) covering public discussion, sheriff procedures, night actions, and system events [13]. A complete game can be expressed in predicates alone, with no natural language at all. The default 7-player setup (which only has the roles seer, guard, villager, and werewolf) uses 25 of these [3], with the larger 8-12-player configurations adding role-specific predicates for hunter, witch, and white wolf king. Each predicate has a type signature (e.g., `estimate(Phase, Speaker, Target, Role)`). Within the framework, predicates give us a clear, structured representation of every speech and game event, instead of free-form text which may be subject to different interpretations. More importantly, they decouple strategic content from linguistic form: once the agent commits to a predicate, the same factual content can be re-expressed in many ways when introducing linguistic variation at Stage 3 (Section III-B). This allows any behavioral difference across conditions to be fully attributable to a statement's wording, rather than the statement choice itself.

B. Three-Stage Speech Generation

During discussion phases, an agent's public speech is represented as a sequence of predicates paired with optional speech text. Agent speech is generated through three stages (Fig. 1).

Stage 1: Predicate generation. Each agent receives a prompt containing the observable game state, its private role knowledge, statements it has recorded as potential truths or falsehoods, a reliability, or trust, score of other players, the current turn's speeches, and the available predicate vocabulary. From this, the agent generates one or more predicates.

Stage 2: Natural language generation. Each predicate is converted to a natural language sentence by a second LLM call with deterministic decoding[2], given only basic facts such as their own role, identifier, predicate definition list, and if a predicate is a response to a specific speech, that single speech is given as well. This ensures identical predicates always produce the same baseline text, isolating any subsequent variation to Stage 3. Starting from this stage, other agents see only the natural language text, not the underlying predicate, and communication is no longer done through objective functions.

Stage 3: Introducing linguistic variation. An optional rewriter transforms the natural language text before it enters the game state. Using a custom rewriter class, output text can be perturbed to vary specific linguistic features (e.g., formality, verbosity), while the underlying predicate content and intent of the speaker remains fixed.

C. Replay Settings

Each comparison specifies a baseline game setting and one or more replay settings. Four experiment settings are implemented, listed below. The third and fourth settings both designate a player to fulfill the "self" role, or the player whose perspective is used for evaluation. The self role is rotated between all players, with one replay per player.

- $\mathcal{T}_1$: **Predicate-only:** Agents both generate and see only predicates, no natural language is involved.
- $\mathcal{T}_2$: **Natural language:** Agents generate a predicate privately, and are then prompted to speak naturally when expressing their choice. Other agents see only their natural language output.
- $\mathcal{T}_3$: **Flip self:** Only the self player's speeches undergo the rewriting phase according to the setting requested. Evaluate all other players' responses to the self player.
- $\mathcal{T}_4$: **Flip others:** The speeches of all players except the self player undergo the rewriting phase according to the setting requested. Evaluate the self player's response to all other players.

At the start of each day's discussion phase, a checkpoint, or copy of the game state, is created. For each replay requested, each checkpoint is loaded. Then, the requested rewriting settings per agent are inserted and the simulation continues

[2]We ensure deterministic outputs using locally-run open source models, with a fixed seed of 101 and temperature set to 0. We experimentally verify this leads to identical outputs given the same input predicates, through a preliminary experiment that scores a candidate model's natural language generation on a fixed predicate set across N=10 trials at temperature 0. The model selected for natural language generation and the agents' base model was Qwen3.5-27B, a dense transformer that scored 100% on this test.

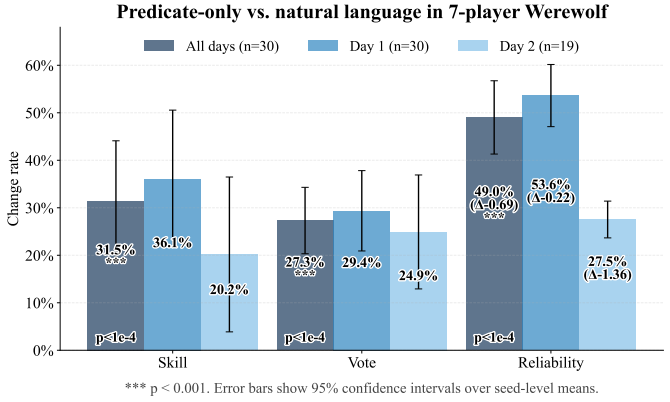

Fig. 2. Predicate-only versus speech in the 7-player setting. Error bars show 95% confidence intervals.

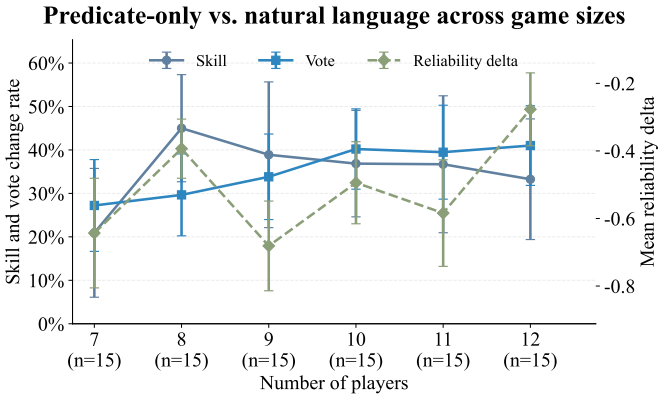

Fig. 3. Predicate-only communication versus natural language speech across game sizes. The left axis shows skill and vote change rates. The right axis shows mean reliability delta. Error bars show 95% confidence intervals over seed-level means. Tick labels indicate the number of seeds.

until the end of the corresponding night cycle. Finally, results are recorded and the game state is reset. The average change for each of the following three metrics is measured:

- Skill targets: which players are targeted by agents with special actions, listed in Table I.
- Vote targets: which players are voted for elimination.
- Reliability scores: the 0-10 trust rating each agent assigns to every other player. All scores are initialized at 5, and increase or decrease incrementally with reasoning.

To evaluate, compute the fraction of data points that differ between the baseline and replay for each metric:

$$\Delta_{\text{metric}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\,\text{baseline}_i \neq \text{replay}_i\,] \tag{1}$$

Let $N_{\text{total}} = N_{\text{skill}} + N_{\text{vote}} + N_{\text{reliability}}$. The total fraction of decisions changed is:

$$\Delta_{\text{total}} = \frac{1}{N_{\text{total}}} \sum_{i=1}^{N_{\text{total}}} \mathbb{1}[\,\text{baseline}_i \neq \text{replay}_i\,] \tag{2}$$

IV. Experiments and Results

The sample sizes are determined by the number of checkpoints and living players, rather than by large-scale sampling, so permutation-based tests that remain valid at small N are used to avoid distributional assumptions. Thus, RQ1 is evaluated with one-sided sign-flip tests and RQ2 is evaluated with paired one-sided sign-flip tests.

A. Effect of Natural Language

1) 7-Player Setting: The baseline game uses the predicate-only setting $\mathcal{T}_1$; at each checkpoint, the same day-night cycle is replayed under $\mathcal{T}_2$, where predicates are expressed as natural language. Following the 7-player setup of prior LLM *Werewolf* evaluations [3], [5], we obtain data of 30 seeds and 51 matched phases. Figure 2 summarizes the results.

Speech changes 31.5% of skill decisions, 27.3% of vote targets, and 49.0% of reliability judgments (all $p < 10^{-4}$, one-sided sign-flip tests). Mean reliability delta is -0.69 [-0.79, -0.59]. Speech is not a neutral rendering of predicates, even when predicate-level content is matched on the speaker side. Furthermore, the effect is not uniform across roles. Villagers show the highest vote-change rate (32.8%), the Seer the highest skill-change rate (57.1%), and mean reliability delta is positive for Werewolves ($+0.72$) but negative for Guard, Seer, and Villagers (-1.57, -0.87, -1.07).

2) Larger Player Configurations (8-12 players): To test whether these effects generalize beyond the 7-player setup, the same experiment is repeated with 8, 9, 10, 11, and 12 players, which introduce additional roles, more targets, and denser discussion. All configurations share the same 15 seeds, with the 7-player reference in Figure 3 using only its first 15. The pattern persists at every scale. Skill-change rates are 45.0%, 38.9%, 36.9%, 36.7%, and 33.3% for the 8–12-player settings respectively; vote-change rates are 29.7%, 33.8%, 40.2%, 39.5%, and 41.0%. Mean reliability delta remains negative at every player count at -0.39, -0.68, -0.49, -0.58, and -0.28. Almost all values meet or exceed the 7-player reference, confirming that the behavioral effect observed in the 7-player setting is not limited to that standard configuration.

B. Formality as a Linguistic Style Perturbation

The second experiment uses formality as a controlled style perturbation in the 7-player setting. Formality is chosen because it is a well-studied style dimension, with prior work defining it through reduced ambiguity and context dependence and modeling it from human judgments in online communication [14], [15]. The data comprises 30 seeds and 1,635 replayed phases. Figure 4 reports paired formal-minus-casual differences under both scope settings.

The perturbation produces consistent surface-level changes. Formal versions are longer and contain fewer contractions: in the flip self setting, formal and casual versions average 46.35 and 34.36 words per edited item, with 0.77 and 2.11 contractions respectively. The flip others setting shows the same pattern (45.21 vs. 33.72 words; 0.74 vs. 1.99 contractions). Thus it is verified that the rewriting stage is shifting style in the intended direction. However, these surface changes do

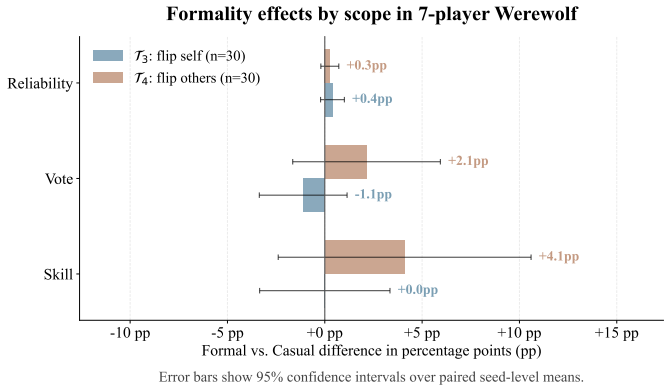

Fig. 4. Paired formal-minus-casual differences under self and others rewrite scopes in the 7-player setting. Whiskers show 95% confidence intervals across seeds. Values are reported in percentage points.

not translate into stable downstream behavioral effects. Across both settings, paired formal-minus-casual contrasts for three reported metrics remain close to zero, with all confidence intervals in Figure 4 overlapping zero. No contrast reaches significance. Mean reliability delta is near zero in both settings. These results confirm that the framework can impose and verify a controlled style perturbation, but formality alone does not produce a robust behavioral effect in the data.

V. Discussion

Several limitations remain. First, all experiments use one base model family (Qwen3.5-27B), so the results show that natural language can change LLM agent behavior in this setup, but should not be treated as universal across models. Second, the reliability results vary by role: Werewolves increase while Guard, Seer, and Villagers decrease, suggesting that natural language may make deceptive or persuasive statements more plausible; future work, however, should test this mechanism more directly. Finally, this paper tests only formal versus casual wording. Future work should test more linguistic features and model families, and add a no-change replay check to separate behavioral effects from game randomness.

VI. Conclusion

This study investigated whether LLM agent behavior is sensitive to linguistic form when strategic content is held fixed. The answer is a clear yes at the level of communication mode: expressing the same predicate-level content as natural language changes $\Delta_{\text{total}} = 52.6\%$ of the total decisions made, across skills, votes, and reliability judgments in the 7-player setting, with similar effects across 8-12-player configurations. LLMs do make different decisions when they read prose than when they process the same information in a structured form without language. At the level of a single linguistic feature, the answer is more uncertain: formal versus casual wording produced measurable differences but no behavioral shift, serving as an informative null that not every wording change introduces bias. Taken together, these results show that the presence of natural language itself is a significant and largely unexamined source of bias in LLM agent systems. Accounting for linguistic form, not just content, is necessary when evaluating fairness in multi-agent LLM settings. The framework that enabled these comparisons, LURK-Werewolf, is open-source. Its three-stage pipeline and checkpoint replay design are not specific to formality; any linguistic axis that can be expressed as a rewriting rule can be tested under the same matched-state protocol, and the predicate layer can be adapted to other discussion-based games or deliberation settings.

VII. Acknowledgment

Claude (Anthropic) was used to assist with editing, shortening length, and LaTeX formatting of Table I. All design, experiments, analysis, and findings are the authors' own.

References

[1] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang *et al.*, "AgentBench: Evaluating LLMs as agents," in *Proc. International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://iclr.cc/virtual/2024/poster/17388

[2] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[3] Q. Zhang, Y. Li, B. Yuan, J. Togelius, G. N. Yannakakis, and J. Liu, "Ethical considerations of large language models in game playing," *arXiv preprint arXiv:2508.16065*, 2025.

[4] F. Kohankhaki, D. Emerson, J.-J. Tian, L. Seyyed-Kalantari, and F. K. Khattak, "Template-based probes are imperfect lenses for counterfactual bias evaluation in llms," *arXiv preprint arXiv:2404.03471*, 2024.

[5] Y. Xu, S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu, "Exploring large language models for communication games: An empirical study on werewolf," *arXiv preprint arXiv:2309.04658*, 2023.

[6] S. Bailis, J. Friedhoff, and F. Chen, "Werewolf arena: A case study in llm evaluation via social deduction," *arXiv preprint arXiv:2407.13943*, 2024.

[7] S. Du and X. Zhang, "Helmsman of the masses? evaluate the opinion leadership of large language models in the werewolf game," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=xMt9kCv5YR

[8] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 2021, pp. 5356–5371.

[9] N. Nangia, C. Vania, R. Bhalerao, and S. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 1953–1967.

[10] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "Bbq: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.

[11] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 2013, pp. 178–186.

[12] P. Liang, "Learning executable semantic parsers for natural language understanding," *Communications of the ACM*, vol. 59, no. 9, pp. 68–76, 2016.

[13] E. Nishizaki and T. Ozaki, "Behavior analysis of executed and attacked players in werewolf game by ilp." in *ILP (Short Papers)*, 2016, pp. 48–53.

[14] F. Heylighen and J.-M. Dewaele, "Formality of language: definition, measurement and behavioral determinants," *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, vol. 4, no. 1, pp. 1–38, 1999.

[15] E. Pavlick and J. Tetreault, "An empirical analysis of formality in online communication," *Transactions of the association for computational linguistics*, vol. 4, pp. 61–74, 2016.