Intent-driven AIWolf Agents with Hierarchical BDI Model and Personality

Yuya Harada[1] **Yoshinobu Kano**[1]

[1]Faculty of Informatics, Shizuoka University
3-5-1 Johoku, Chuo-ku, Hamamatsu, Shizuoka 432-8011, Japan
{yharada, kano}@kanolab.net

Abstract

While large language models possess advanced language generation capabilities, challenges remain in modeling recognition processes based on personality traits and generating strategic behaviors that reflect them. We propose a design methodology for werewolf agents that integrates a hierarchical BDI framework with MBTI and Enneagram personality theories. We systematically model the influence of personality traits on recognition, judgment, and action stages, integrating them into a hierarchical decision-making mechanism combining long-term strategy and short-term tactics. Comparative experiments between baseline implementation and our proposed method confirm the effectiveness of our approach in generating utterances that reflect individual differences in perception and maintaining strategic consistency.

1 Introduction

In recent years, large language models (LLMs) have advanced significantly in natural language processing. Yet, applying them to multi-agent dialogue environments requiring social reasoning, such as the werewolf game, remains challenging. Modeling personality-driven perception and strategy is particularly difficult, and conventional agents cannot consistently integrate long-term strategy and short-term tactics. Personality expression also remains limited to superficial utterance styles, without influencing recognition or decision-making.

We propose a design methodology that integrates a hierarchical BDI (Belief-Desire-Intention) framework with MBTI and Enneagram personality typologies. The Macro-BDI layer governs long-term strategy, while the Micro-BDI layer manages tactical decisions, both influenced by personality traits. This enables consistent personality modeling not only in utterance generation but also in recognition and judgment.

The main contributions of this paper are as follows:

- **Hierarchical BDI Architecture**: A two-layer framework separating strategic planning in the Macro-BDI layer from tactical execution in the Micro-BDI layer, ensuring consistent decisions across time scales.
- **Systematic Personality Integration**: Derives 24 computable features from MBTI and Enneagram, mapping unstructured descriptions to behavior.
- **Personality-Driven Cognitive Biases**: Models biases and behavioral tendencies via personality-weighted parameters, generating realistic, though not always optimal, behavior.
- **Empirical Validation**: Experiments in the AIWolf framework show a 14.1% improvement in subjective evaluations of human-likeness.

Comparative experiments with conventional agents confirmed improvements in win rates, utterance naturalness, and strategic consistency. These results indicate that integrating hierarchical decision-making with personality traits fosters more human-like behavior.

This paper is structured as follows: Section 2 surveys related work, Section 3 details the methodology, Section 4 explains the experimental setup, Section 5 the evaluation framework, Section 6 presents results, Section 7 discusses findings, Section 8 outlines future work, and Section 9 concludes.

2 Related Work

2.1 AIWolf Project

The AIWolf Project aims to "construct agents that can play werewolf games while engaging in natural communication with humans," and regularly holds AIWolf competitions to promote werewolf

Proceedings of the 3rd International AIWolfDial Workshop, pages 22–36

AI research (Kano et al., 2019) (Kano et al., 2023) (Kano et al., 2024) (Gondo et al., 2024) (Kano et al., 2025). The AIWolf competition has three divisions: protocol division, natural language division, and infrastructure division. In the natural language division, agents communicate exclusively in Japanese or English. Evaluation is based on five criteria: (i) naturalness of utterance expression, (ii) naturalness of context-aware dialogue, (iii) consistency of utterance content, (iv) coherence with game actions, and (v) richness of expression.

2.2 BDI Architecture

The BDI (Belief-Desire-Intention) architecture is a representative framework for modeling the reasoning processes of cognitive agents. Belief represents information and perceptions about the world that agents hold, Desire represents goals or wishes they want to achieve, and Intention represents concrete plans and execution intentions for achieving selected goals. As classical research, Rao et al. (1997) presented a formalization of rational agents in BDI architecture, providing a connection between mental attitudes and action semantics. Our hierarchical design adds implementation hypotheses of macro/micro time scale separation and personality trait integration on top of this framework.

2.3 Integration of LLM and BDI in AIWolf

For the integration of LLM and BDI in the werewolf domain, Gondo et al. (2024) verified LLM's logical reasoning ability by incorporating BDI logic representation into prompts, conducting comparative evaluation using win rates and voting rates against werewolves as metrics in 5-agent matches. Our research is complementary in that it goes beyond notation on prompts to introduce hierarchical BDI (Macro/Micro) encompassing *state representation, policy, and disclosure control*, generating micro-intentions consistent with personality traits as conditions for generation.

2.4 MBTI and Enneagram Personality Theories

MBTI (Myers–Briggs Type Indicator) is a personality classification system based on Jung's psychological typology, characterizing individual cognitive and judgment tendencies through four dimensions: Extraversion–Introversion (E–I), Sensing–Intuition (S–N), Thinking–Feeling (T–F), and Judging–Perceiving (J–P) (Myers et al., 1998). Treating each dimension as continuous values from 0 to 1 is our modeling choice, projecting type information into a form more amenable to downstream computation.

The Enneagram is a model assuming nine personality types, describing the fundamental motivations, fears, and worldviews of each type. By combining it with MBTI, we can construct comprehensive personality models from both cognitive style (MBTI) and motivational structure (Enneagram) perspectives.

3 Proposed Method: Hierarchical BDI Framework with Personality Integration

This section presents a design methodology that integrates a hierarchical BDI framework with MBTI and Enneagram personality typologies to realize consistent intent-driven behavior generation in multi-agent dialogue systems.

3.1 Design Principles

Our proposed method aims to address two challenges in agent design: (i) coordination between macro and micro decision-making, and (ii) systematic integration of personality traits into decision-making processes. We adopt a two-layer hierarchical BDI structure where personality parameters act on both layers.

3.2 Architecture Overview

The proposed model consists of two layers:

- **Macro-BDI Layer**: Responsible for long-term strategic planning and quantification of personality traits
- **Micro-BDI Layer**: Manages turn-by-turn tactical decisions and immediate responses

Both layers interact through personality-weighted parameters, providing consistency in perception, evaluation, and decision-making throughout the game. As illustrated in Figure 1, the Macro-BDI layer maintains long-term strategy and personality parameters, while the Micro-BDI layer executes turn-level tactics in alignment with them.

3.3 Werewolf Game and Terminology Definitions

The werewolf game is a dialogue-based game where players deduce others' roles through conversation, featuring a conflict structure between the villager team and the werewolf team. The game

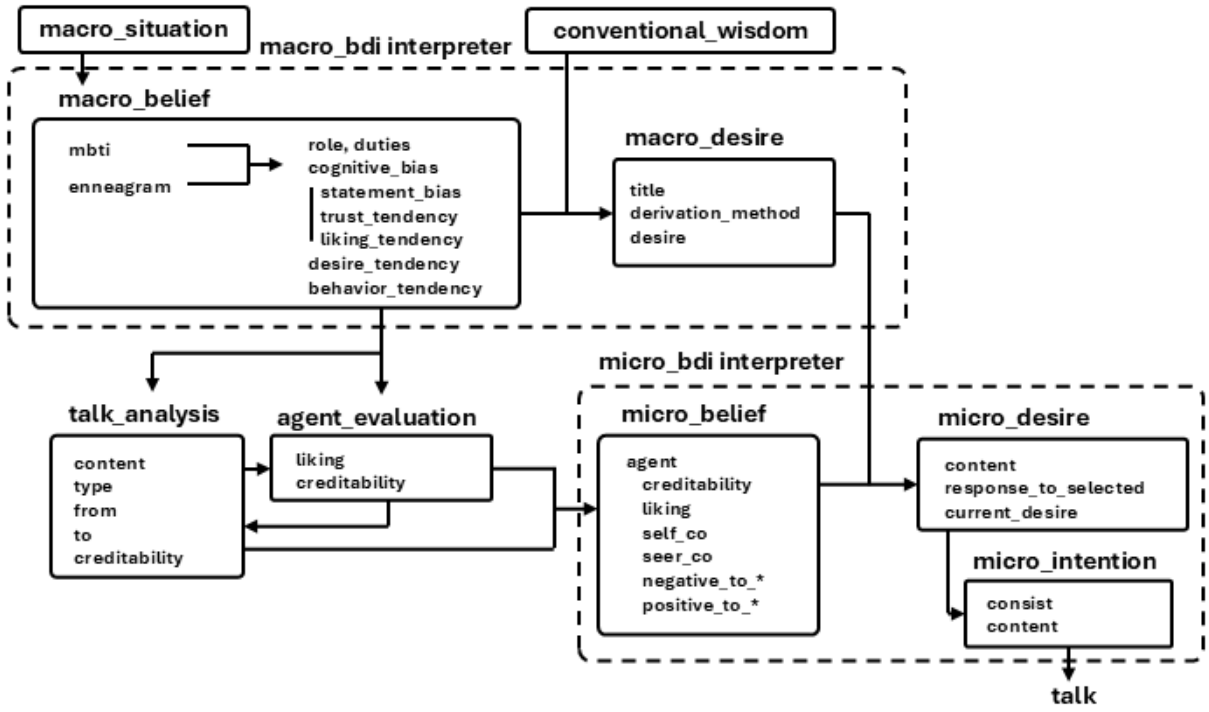

Figure 1: Overall configuration of the hierarchical BDI architecture. The Macro-BDI layer manages long-term strategy and personality parameters, while the Micro-BDI layer performs turn-by-turn tactical decisions.

proceeds in days, each with a day and a night phase, with conversation in the day and voting or abilities at night. The "villager team" wins by eliminating the "werewolf team," while the werewolf team wins by eliminating the villager team.

Terms used in this paper for the werewolf game are the followings: **Role** is a position or ability held by players (e.g., villager, seer, possessed, werewolf). Villagers and seers belong to the villager team, with seers able to divine one person per day. Possessed and werewolves belong to the werewolf team, with possessed not detected as werewolves by divination, allowing them to lurk and support. **Coming Out (CO)** is an act of publicly declaring role claims. Early CO increases credibility but carries attack risk. **Accusation** is to claim a specific opponent belongs to the enemy team. **Turn** is a unit of utterance progression during day phase. Refers to the interval where speaking rights circulate once to each player.

Based on this, we define the state determined at the beginning of each day (game stage, number of survivors, disclosure state, personality trait analysis content) as **macro situation**, and the state including utterance history at various points during the day, analysis results for other players, and analysis results for utterances as **micro situation**.

3.4 Macro-BDI Layer

This section describes a method for estimating personality characteristics from short profile texts and consistently reflecting them in utterances, decision-making, and actions. Rather than directly embedding profile texts in prompts, we map them to numerical representations using LLMs and expand them into secondary features, obtaining manageable personality expressions from minimal source information.

3.4.1 Psychological Significance

First, MBTI is more widely adopted than alternatives such as the Big Five, providing an interface that general users can readily specify in future applications. Second, combining MBTI (eight dimensions) with the Enneagram (nine types) enables the multifaceted generation of 24 derived features spanning cognition, motivation, and behavior. Third, MBTI's binary axes align naturally with tactical choices in Werewolf, while the Enneagram's motivational types make underlying needs explicit, supporting the generation of need and behavior tendencies.

3.4.2 Use as an Intermediate Representation

Most importantly, we do not use these frameworks as psychological "truths," but rather as struc-

tured, computable intermediate representations that bridge unstructured inputs to consistent behavior. Although variants of MBTI are widely used under its name, as long as they serve as indicators of similar tendencies, they are effective for LLMs and thus valid as intermediate representations.

The reasons for using MBTI as an intermediate step rather than directly generating personality indicators are twofold. By first mapping unstructured free text to 8-dimensional MBTI, we compress and format information, suppressing variance in downstream weight calculations.

3.4.3 MBTI Dimension Estimation (0–1 Normalization)

We input profile text to LLM and output continuous values (0–1) for MBTI's 8 dimensions. Each dimension consists of extroversion, introversion, sensing, intuition, thinking, feeling, judging, and perceiving. Note that the correspondence with generally circulating MBTI variants (e.g., 16Personalities) may not strictly match, but this is not problematic as we treat it as intermediate representation in our method.

3.4.4 Transfer from MBTI to Enneagram

We calculate affinity with the 9 Enneagram types through linear combination of obtained MBTI values. As an example, the calculation formula for Type 1 (Reformer) is shown (coefficients are design parameters):

$$\begin{aligned} \text{Reformer} = a\,\text{intuition} + b\,\text{thinking} \\ + c\,\text{judging} \\ \text{where} \quad a, b, c \geq 0,\ a + b + c = 1. \end{aligned}$$

Similarly, we design coefficients for each of the 9 types to obtain affinity vectors from MBTI vectors (details in appendix).

3.4.5 Weighting of Cognitive Indicators

We define indicators such as utterance evaluation, trust tendency, and liking tendency. These consist of 10 types: As statement bias, we prepare indicators for logical consistency showing logical coherence of utterances, specificity and detail showing concreteness and detail, intuitive depth showing depth of intuitive statements, and clarity and conciseness showing clarity and brevity. For trust tendency, we prepare social proof showing tendency to trust majority opinions and social proof, honesty showing tendency to value sincerity, and consistency showing tendency to value consistency through discussion. Furthermore, for liking tendency, we prepare friendliness showing tendency to feel favorably toward friendly attitudes, emotional resonance showing tendency to value empathy, and attractive expression showing attraction to appealing expressions.

We calculate weights $w_k \in [0, 1]$ from MBTI and Enneagram estimates, and obtain comprehensive indicators combined with evaluation values $s_k \in [0, 1]$ for target utterances:

$$S = \frac{\sum_k w_k s_k}{\sum_k w_k}.$$

3.4.6 Modeling of Desire and Behavior Tendencies

We quantify desire and behavior tendencies and reference them in utterance generation, action decisions, and goal setting. We define 7 indicators for desire tendency: self_realization showing self-actualization desire, social_approval showing social recognition desire, stability showing stability desire, love_intimacy showing intimacy desire, freedom_independence showing independence desire, adventure_stimulation showing stimulation desire, and stable_relationships showing relationship stability desire. We define 7 indicators for behavior tendency: avoidant_behavior showing avoidant behavior, aggressive_behavior showing aggressive behavior, adaptability showing adaptability, introversion showing introversion, extroversion showing extroversion, empathy showing empathy, and assertiveness showing assertiveness. Each indicator is expressed as continuous values from 0 to 1, derived from personality parameters. Full formulas for deriving enneagram, cognitive, trust, liking, desire, and behavior indicators are summarized in Appendix B.

3.4.7 Macro-Desire

This component pre-generates macro-level desires before play begins, assuming advanced communication environments where strategies and false utterances are mixed. First, we narrow down action options based on general knowledge and established tactics, then determine personality-consistent preferences within that range.

Conventional Wisdom Bank We performed case classification from multiple perspectives of macro and micro situations for each role. Specific information included in this Conventional Wisdom includes the `title` for each case classifica-

tion, derivation_method describing case classification conditions based on macro and micro situations, and three patterns of general action guidelines (objective).

Generation of Macro-Desire We select from candidate sets for each role × situation, apply personality reflection with LLM to align with macro_belief and desire_tendency, and aggregate situation-specific desires.

Position in BDI The macro layer handles Belief/Desire but does not generate Intention. The determined macro_desire acts as a consistent bias on micro-layer decision-making (evidence presentation timing, degree of pursuit/mitigation, disclosure strategy, etc.).

Effect Within rational bounds based on established tactics, we can stably reflect personality-driven preferences and expression differences. As a result, we speed up real-time tactical adjustments while maintaining personality trait consistency.

3.5 Micro-BDI Layer

The Micro-BDI layer operates on a turn-by-turn basis, responding immediately to recent conversation events while maintaining alignment with macro strategy. It serves as an intermediate layer that updates utterance analysis records each turn and re-estimates interpersonal statistics.

3.5.1 Utterance Analysis Records

We save only utterance content for our own utterances, and save others' utterances with the following fields. **content**: utterance text, **type**: utterance type (co, question, negative, positive, null), **from**: speaker (character name), **to**: recipient (character name or all), **raw credibility**: basic credibility of utterance itself (0–1), **credibility**: final credibility after weighting raw credibility with statement bias from macro-belief and correcting with prior impressions (liking/credibility) (0–1).

Data Sources and Circular Updates Long-term impressions per speaker are aggregated at the agent level, updating liking and credibility. These are reflected in credibility correction for new utterances, with performance obtained from utterance analysis records contributing back to impression updates in a circular structure. Own utterances are recorded chronologically for consistency and avoiding repetition.

Semantics of Utterance Types type is a multi-valued label representing utterance function, defining co (role claim), question, negative/positive (evaluative utterances), and null (neutral). co is set to have high response priority in prompts.

3.5.2 Micro-Belief

We maintain interpersonal impressions and interaction statistics obtained from utterance analysis records and past history for each agent a. Items maintained includes followings. **credibility**: Personal credibility toward target a (0–1); **liking**: Favorability toward target a (0–1); **self_co**: a's own CO (e.g., seer, medium, villager); **seer_co**: Summary of divination results (who judged a and how); **negative_to_{name}**: Cumulative count of negative utterances a directed at {name}; **positive_to_{name}**: Cumulative count of positive utterances a directed at {name}.

3.5.3 Micro-Desire

This module determines tactical goals (micro-desire) for the next utterance, aiming to achieve both *responsiveness to conversation* and *state consistency*. Processing consists of (i) collection and summarization of micro situations, (ii) selection of response targets, and (iii) proposal output and verification by LLM.

Collection and Summarization of Micro Situations From the utterance analysis records and information about each agent summarized in **micro belief**, we collect and summarize the following: **Conversation analysis** by reading the last 5 items from utterance analysis records (content, sender, recipient, utterance type, etc.); **Micro-level beliefs** from reference interpersonal trust scores, credibility assessments, negative stance indicators, and role claim consistency (seer claims/self-declarations) for each agent; **Macro-level desires** from reference situation-specific desires and maintain relevant items as discussion phase markers; **Game state** including summary of day number, remaining utterances, number of survivors, roles, etc.

Selection of Response Target (response_to_selected) Select only one item from the latest 5 utterance analysis records in the following priority (if absent, null): 1. Utterances containing own name in to, 2. Utterances with type as co, 3. Utterances with to=all and

`type=question`. For same priority, select by descending credibility, then newer.

LLM Output and Verification Following the prompt in the appendix, output the following three items. **`discussion stage`** is selected according to conditions in `derivation method`, **`current desire`** concretizes relevant items from `macro desire`, and **response plan** to selected utterance (if selected utterance is `null`, then `content: null`).

3.5.4 Micro-Intention

This module uses LLM to construct *decision-making units for one turn of next utterance* (*micro-intention*) from recent context, role, and behavior tendencies.

Input From information obtained through previous processing: (i) desire and response plan for relevant micro situation from *micro desire*, (ii) role duties, behavior/desire tendencies from *macro belief*, (iii) favorability/credibility from *micro belief*, (iv) day number and survivor situation from game information sent by server.

Output The LLM outputs a compact YAML record under the key `micro_intention` with exactly two fields: (i) `consist` — a short plan or structure for the next utterance, and (ii) `content` — the actual utterance content to be spoken.

3.5.5 Utterance Generation (talk)

This module generates the *final output utterance (one sentence)* for each turn. The purpose is to present timely and appropriate responses based on recent dialogue state and intention expressions.

We provide LLM with previously generated `micro intention` and instructions for utterance generation. The final output is a single natural sentence without additional meta-information or formatting symbols.

4 Experiments

This section describes the experimental configuration for comparative evaluation of the proposed framework with baseline implementation.

4.1 Baseline Implementation

We implemented a simple mechanism that makes judgments for utterance generation and game actions with a single prompt.

Following the requirements of the 5-player village track in the AIWolf Natural Language Division, we implemented basic functions (utterance generation, voting decisions, divination execution, attack selection) corresponding to four roles: villager, seer, possessed, and werewolf. We used GPT-4o as the language model.

Characteristics and limitations of the baseline system include the following issues: **Decision-making structure** has no separation between long-term strategy and short-term tactics, making it difficult to maintain intention consistency throughout the game; **Personality expression** is limited to surface-level adjustment of utterance style, not modeling systematic influence on recognition and judgment processes; and **strategic behavior** has basic tactics per role are rule-based, with no situation-adaptive strategy modification or personality-dependent tactical selection.

4.2 Implementation of Proposed Method

The proposed method using hierarchical BDI was implemented using the LLM. I use GPT-4o (temperature 0.7) for both baseline and proposed agents. For personality parameters, we automatically configured them as described above using the profile settings provided by the AIWolf game server as input.

4.3 Game Settings

Following the AIWolf competition, experiments used 5-player games consisting of 2 villagers, 1 seer, 1 possessed, and 1 werewolf. We conducted round-robin matches with 6 teams (3 proposed method agents, 3 baseline agents), executing 10 games per team, with conditions of maximum 20 utterances per day (up to 4 utterances per agent) and non-public voting.

5 Evaluation

We calculated win rates as objective evaluation metrics using the following methods: **Overall win rate**, **Win rate by role**, **Average role win rate**, and **Win rate weighted by role appearance ratio**. Formal definitions of Macro, Micro, and Weighted Micro are provided in Appendix D.

For subjective evaluation, we used llm-as-a-judge, which began operation in this AIWolf competition, to evaluate the following five axes: **A** Naturalness of utterance expression, **B** Naturalness of conversational context, **C** Consistency of conversation content (presence/absence of contradictions), **D** Coherence between conversation content and game actions (voting, attacks, divination), **E** Diversity of utterance expression including character

consistency. The evaluation was conducted using GPT-5 with a ranking-based approach, where teams were ordered from best (rank 1) to worst for each criterion, with no ties allowed. This relative ranking method ensures clear differentiation between agent performances.

Validation of the automated evaluation approach was performed by comparing LLM-Judge results with human evaluations from the AIWolf competition. Criteria A, B, D, and E showed high correlation with human subjective assessments, supporting the reliability of automated evaluation for these aspects. For criterion C (consistency), where automated evaluation showed some limitations, manual verification was conducted by the authors. The proposed method agents demonstrated consistent utterances without contradictions, as they were explicitly instructed to "avoid contradictions and redundancy" while being provided with their utterance history during generation (see Appendix C for detailed comparison results).

6 Results

6.1 Game Result Metrics

As shown in Table 1, the overall win rate of the proposed method was 53.33%, 3.34 points below the baseline (56.67%). By role, improvements were seen in the Possessed role (16.67%→33.33%), while notable decreases occurred in the Seer role (83.33%→66.67%) and Werewolf role (33.33%→16.67%). The Villager role maintained equivalent performance (75.00%→75.00%).

These results suggest that systematic integration of personality traits led to deviation from game-theoretically optimal strategies. The halving of win rate in the Werewolf role particularly likely reflects incompatibility between roles requiring aggressive behavior and personality traits.

6.2 Qualitative Evaluation

In the qualitative evaluation shown in Table 2, the proposed method outperformed the baseline in all 5 criteria, achieving a 14.1% improvement in overall evaluation from 3.227 to 2.773. Particularly notable improvements were: - "B: Naturalness of context-aware dialogue" (3.300→2.700, 18.2% improvement) - "E: Character consistency" (3.200→2.800, 12.5% improvement)

Interestingly, "D: Coherence with game actions" also improved (3.167→2.833), indicating that despite lower win rates, behavioral consistency improved.

7 Discussion

7.1 Trade-off Between Human-likeness and Strategic Optimality Through Personality Integration

The experimental results demonstrated a clear trade-off between "human-likeness" and "strategic optimality" in the proposed method. The slight decrease in win rate (3.34 points) compared to the substantial improvement in subjective evaluation (14.1%) indicates that the systematic integration of personality traits functioned as intended. However, with only 30 total games, the sample size is insufficient for statistical verification, necessitating validation through larger-scale experiments.

The halving of the werewolf role win rate (33.33%→16.67%) was particularly notable, clearly demonstrating the impact of behavioral constraints imposed by personality parameters. Agents with high introversion (e.g., introversion=0.8) and low empathy (empathy=0.268) struggled with the aggressive accusations and strategic deception necessary for the werewolf faction, resulting in consistently passive behavior. This can be interpreted as faithfully reproducing the individual differences in "role aptitude" observed in human players.

7.2 Effectiveness and Limitations of Hierarchical BDI Structure

7.2.1 Success Factors in the Possessed Role

The doubling of the possessed role win rate (16.67%→33.33%) represents an important result demonstrating the effectiveness of the hierarchical structure. The possessed is a role with a dual structure of "feigning sanity while performing madness," and the separation where the Macro-BDI layer maintains the long-term strategy of "supporting the werewolf faction" while the Micro-BDI layer executes situation-adaptive tactics proved successful. The superiority of the hierarchical approach was demonstrated in managing the complex psychological states unique to this role.

7.2.2 Performance Degradation in the Seer Role

The decrease in the seer's win rate (83.33%→66.67%) resulted from personality parameters excessively influencing information disclosure strategies. Agents with introverted

Table 1: Win rates per role and total performance for 5-player village. Lower is better for rank-based metrics.

Team	Possessed	Seer	Villager	Werewolf	Wins	Games	Macro (%)	Micro (%)	Weighted Micro (%)
aiwolf-nlp-agent-llm-A	0.00 (2)	100.00 (2)	100.00 (4)	50.00 (2)	7	10	70.00	62.50	77.27
aiwolf-nlp-agent-llm-B	50.00 (2)	100.00 (2)	50.00 (4)	0.00 (2)	5	10	50.00	50.00	40.91
aiwolf-nlp-agent-llm-C	0.00 (2)	50.00 (2)	75.00 (4)	50.00 (2)	5	10	50.00	43.75	59.09
Baseline Avg/Total	16.67	83.33	75.00	33.33	17	30	56.67	52.08	59.09
yharada-A	50.00 (2)	100.00 (2)	75.00 (4)	50.00 (2)	7	10	70.00	68.75	68.18
yharada-B	0.00 (2)	50.00 (2)	75.00 (4)	0.00 (2)	4	10	40.00	31.25	45.45
yharada-C	50.00 (2)	50.00 (2)	75.00 (4)	0.00 (2)	5	10	50.00	43.75	50.00
Proposed Avg/Total	33.33	66.67	75.00	16.67	16	30	53.33	47.92	54.55

Table 2: Subjective Evaluation Results by LLM-Judge (5-player village)

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	All Average
aiwolf-nlp-agent-llm-A	4.000	3.300	3.500	3.400	3.900	3.620
aiwolf-nlp-agent-llm-B	2.200	2.600	2.300	2.600	2.000	2.340
aiwolf-nlp-agent-llm-C	3.800	4.000	3.600	3.500	3.700	3.720
Baseline Average	3.333	3.300	3.133	3.167	3.200	3.227
yharada-A	2.700	3.200	3.100	3.500	3.000	3.100
yharada-B	2.300	2.000	2.100	2.100	2.300	2.160
yharada-C	3.000	2.900	3.400	2.900	3.100	3.060
Proposed Method Average	2.667	2.700	2.867	2.833	2.800	2.773

personalities (introversion>0.6) or high cautiousness (judging>0.7) tended to hesitate in sharing divination results at optimal timing, compromising the villager faction's information advantage.

7.2.3 Concrete Reflection of Personality Traits in Utterances

Analysis of actual game logs confirmed the influence of personality parameters on utterance generation. Two representative examples are shown below.

Example 1: Servas (85 years old, male)

Configured personality traits: "Very calm and composed personality. Provides advice based on years of knowledge and experience"

```
Personality parameters: extroversion
    =0.7, feeling=0.5, social_proof
    =0.66, empathy=0.452
```

Actual utterance (Day 1, Talk 11):

```
I agree with everyone; sharing theories
    and strategies can only strengthen
    our efforts so let's keep our
    discussions open and remain alert
    for any signs of deception.
```

This utterance reflects multiple personality traits. First, high extroversion (0.7) and social_proof (0.66) appear in the group-coordination-focused introduction "I agree with everyone." Additionally, moderate empathy (0.452) generates inclusive expressions like "our efforts" and "let's," while the balance between feeling (0.5) and thinking (0.5) produces a balanced utterance between logic ("strategies") and emotion ("strengthen our efforts"). Combined with the age setting of 85, these naturally portray the role of an experienced mediator.

Example 2: Shion (16 years old, male, werewolf role)

```
Personality parameters: introversion
    =0.6, thinking=0.7, empathy=0.372,
    logical\_consistency=0.61
```

Actual utterance (Day 2, Talk 0):

```
With Minato gone and Minako attacked we
    need to be extra vigilant about any
    unusual behavior. Let's review our
    notes and see if anything stands out
     regarding yesterday's discussions.
```

Despite being in the werewolf role, this utterance begins with an emotion-free enumeration of facts ("Minato gone and Minako attacked") due to high thinking (0.7) and logical_consistency (0.61). Low empathy (0.372) manifests as a lack of emotional response to a companion's death, instead immediately transitioning to logical response ("review our notes"). The influence of introversion (0.6) results in a passive stance suggesting observation and analysis rather than aggressive accusation.

7.3 Impact of Cognitive Bias Modeling

The MBTI-based cognitive bias modeling generated the following human-like cognitive errors: **Confirmation bias** to fix on initial impressions (liking=0.5) causing difficulty in appropriately updating subsequent information; **Emotional judgment**

that low emotional resonance (0.120) compared to logical consistency (0.642) affected voting decisions; and **social conformity** to follow majority opinions due to social proof (0.360).

While strategically suboptimal, these faithfully mimic human cognitive characteristics and contributed to the "contextual naturalness" (18.2% improvement) in subjective evaluation.

7.4 Balancing Personality Expression and Strategic Performance

The primary goal of this study is to realize agents that embody human-likeness rather than strategic optimality. The decline in strategic performance that accompanies the introduction of personality traits should be seen not as a limitation but as a feature that enhances authenticity. Human players consistently make second-best choices influenced by cognitive biases and personality; reproducing these imperfections is therefore essential for truly human-like agents. Beyond competitive games, prioritizing human-likeness over strategic optimality becomes even more important in general dialogue settings.

7.5 Generalizability to Other Multi-Agent Environments

As demonstrated by the experiments, the proposed method entails a trade-off between human-likeness and strategic performance. However, in settings where victory is not the primary metric, where non-rationality has value, or where diversity and individuality are prioritized over optimality, it can in fact be highly effective.

Concretely, it is applicable to (i) entertainment domains that require consistent dialogue generation by diverse characters, (ii) social simulation of opinion formation that incorporates cognitive biases and personality-driven decision-making, and (iii) education and training contexts that benefit from human-like agents capable of making non-optimal choices.

In these areas, the very factors that reduced competitiveness in Werewolf—the faithful reproduction of cognitive limitations and personality-driven decision-making—become assets. In other words, what appears to be a weakness in competitive environments can transform into a major strength in contexts where human authenticity is valued more than strategic optimality.

8 Future Work

8.1 Scaling of Experimental Validation

We recognize that this study's experimental scale is preliminary and insufficient for statistical significance. Due to limits in computational cost and time, conducting 120 matches equivalent to the main AIWolf competition was infeasible. We therefore plan to participate in the next AIWolf competition with our proposed agents to obtain more extensive and statistically robust results. The competition environment, with hundreds of matches against diverse opponents, will provide a more reliable validation of our approach.

8.2 Intention Inference and Confidence Estimation

Although the current implementation applies personality biases to confidence evaluation of LLM outputs, the baseline method has room for improvement. As future work, we plan to introduce BDI-based intention inference for opponents' utterances, moving from superficial scoring to refined confidence estimation. By enhancing strategic computation through intention recognition while preserving personality-driven differences, we aim to build agents that are both more rational and authentically human.

9 Conclusion

This paper presented a design method for intention-driven werewolf agents integrating a hierarchical BDI framework with MBTI and Enneagram personality typologies. Experimental results demonstrated that while the proposed method partially sacrificed strategic optimality (win rate 56.67%→53.33%), it achieved significant improvement in generating human-like behavior (subjective evaluation 3.227→2.773).

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers JP22H00804, JP21K18115, JST AIP Acceleration Program JPMJCR22U4, and the SECOM Science and Technology Foundation Special Area Research Grant. We wish to thank the members of the Kano Laboratory in Shizuoka University who helped to evaluate the game logs.

References

Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. 1998. *MBTI® Manual: A Guide to the Development and Use of the Myers–Briggs Type Indicator®* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

A. S. Rao and M. P. Georgeff. 1997. Modeling Rational Agents within a BDI-Architecture. In *Readings in Agents*, pages 317–328. Morgan Kaufmann.

Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. AIWolfDial 2025: Summary of Natural Language Division of 7th International AIWolf Contest. In Yoshinobu Kano (Ed.), *Proceedings of the 3rd International AIWolfDial Workshop*, September.

Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of Natural Language Division of 6th International AIWolf Contest. In Yoshinobu Kano (Ed.), *Proceedings of the 2nd International AIWolfDial Workshop*, September, Tokyo, Japan. Association for Computational Linguistics. URL: aclanthology.org/2024.aiwolfdial-1.1. DOI: 10.18653/v1/2024.aiwolfdial-1.1. pp. 1–12.

Yoshinobu Kano, Neo Watanabe, and others. 2023. AIWolfDial 2023: Summary of Natural Language Division of 5th International AIWolf Contest. In Simon Mille (Ed.), *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, September, Prague, Czechia. Association for Computational Linguistics. URL: aclanthology.org/2023.inlg-genchal.13. pp. 84–100.

Yoshinobu Kano, Claus Aranha, and others. 2019. Overview of AIWolfDial 2019 Shared Task: Contest of Automatic Dialog Agents to Play the Werewolf Game through Conversations. In Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Takashi Otsuki (Eds.), *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, October, Tokyo, Japan. Association for Computational Linguistics. URL: aclanthology.org/W19-8301. DOI: 10.18653/v1/W19-8301. pp. 1–6.

Takumi Gondo, Hiroki Sakaji, and Itsuki Noda. 2024. Verification of Reasoning Ability Using BDI Logic and Large Language Models in AIWolf. In *Proceedings of The Japanese Society for Artificial Intelligence Annual Conference (JSAI2024)*, 38th Annual Conference, Session ID 2F6-GS-5-04, p. 2F6GS504. DOI: 10.11517/pjsai.JSAI2024.0_2F6GS504. URL: J-STAGE.

A Prompt Templates, Calculation Formula

This appendix shows the main prompt templates and calculation formulas used in the proposed method. Due to space limitations, only representative parts are included.

A.1 Macro-Belief: MBTI Estimation

```
mbti_inference: |-
  Analyze the profile text and estimate
      MBTI parameters in the range 0 to
      1.

  Profile: {{ profile }}
  Agent name: {{ agent_name }}

  Estimate the following eight
      parameters (each 0-1):
  - extroversion: social and outward-
      oriented tendency
  - introversion: introspective and
      inward-oriented tendency
  - sensing: concrete, reality-focused
      information processing
  - intuition: abstract, possibility-
      focused information processing
  - thinking: logical, objective
      judgment
  - feeling: affective, subjective
      judgment
  - judging: planned, structured
      behavior
  - perceiving: flexible, adaptive
      behavior

  **Important**: Output **only** in the
      following strict format. No other
      text.

  extroversion: 0.X
  introversion: 0.X
  sensing: 0.X
  intuition: 0.X
  thinking: 0.X
  feeling: 0.X
  judging: 0.X
  perceiving: 0.X
```

A.2 Macro-Desire

```
macro_desire_one_liner: |-
  Given the personal traits below, infer
       what desire this player would
      actually hold.
  Provide brief, evidence-based
      reasoning in steps, and end with
      **one single-sentence final
      conclusion**.

  [Context]
  - game_id: {{ game_id }}
  - agent: {{ agent }}
  - role: {{ role_name }}

  [Situation]
```

```
  - title: {{ situation_title }}
  - derivation_method: {{
      situation_derivation }}

  [Tendencies]
  - behavior_tendency: {{
      behavior_tendency }}
  - desire_tendency: {{ desire_tendency
      }}

  # Output requirements
  - First, concisely present evidence
      and reasoning (a few lines are
      fine), then write the conclusion
      **only on the last line** in the
      form: `Final: <one sentence>`.
  - The Final line must be **one
      sentence only**, ending with a
      period or terminal punctuation (
      English or Japanese acceptable).
  - Do not include any extra decoration
      or meta information.
```

A.3 Message Type Analysis

```
analyze_message_type: |-
  You are classifying a Werewolf-game
      utterance into EXACTLY ONE of:
  co, negative, positive, question, null
  Output ONE token only (no punctuation,
       no explanations, no code fences).
  Do not infer from prior conversation;
      judge this utterance alone.
  Utterance: {{ content }}
  Alive agents (allowed names): {{
      agent_names|join(", ") }}
  Allowed role words (case-insensitive
      for English; literal for Japanese)
      :
    villager, Villager, VILLAGER
    seer, Seer, SEER
    werewolf, Werewolf, WEREWOLF
    possessed, Possessed, POSSESSED
    bodyguard, Bodyguard, BODYGUARD
    medium, Medium, MEDIUM
    (Judgement words such as HUMAN/white
        /black may appear in reports.)
  HARD RULE (role-word requirement for
      co):
  - Output "co" ONLY IF the utterance
      CONTAINS at least one allowed role
       word above.
    * If a result-like sentence lacks a
        role word (e.g., "@X is white",
        "Y is black"),
      DO NOT output "co". Classify it as
          "negative" if it is an
        accusation toward a named
        target,
      otherwise "null" (or "question" if
          it asks).
    * Mentions of roles in general
      discussion without self-claim or
       concrete result report are NOT
      "co".
  STRICT RULES:
  1) co: Output "co" ONLY IF the
      utterance explicitly does ONE of:
      (A) SELF-CLAIM of a role (examples
          ):
          "I am Minako a villager", "I'm
               the Seer", "Villager CO",
               "Seer CO".
      (B) ABILITY RESULT REPORT stating
          a role word + a target + a
          judgement:
          "Seer result: @X is HUMAN"
  2) negative: suspicion/accusation/vote
       intent toward specific agent(s).
  3) positive: support/defense/trust
      toward specific agent(s).
  4) question: asks something to group
      or a specific agent.
  5) null: none of the above.
  TIE-BREAKING (when ambiguous): co >
      negative > positive > question >
      null
```

A.4 credibility Analysis

```
analyze_credibility: |-
  Score the utterance on four 0-1
      metrics; higher is better.
  Output EXACTLY these four lines (no
      extra text, no code fences):
  logical_consistency: 0.50
  specificity_and_detail: 0.50
  intuitive_depth: 0.50
  clarity_and_conciseness: 0.50
  Utterance: {{ content }}
  Agents: {{ agent_names|join(", ") }}
```

A.5 Micro-Desire: Situation Selection and Desire Refinement

```
  You are generating a **micro-desire**
      (the strategic aim for the agent's
       next utterance) for a Werewolf
      game.
  Output **YAML only**. No Markdown
      fences. **Do not write dialogue
      lines**; write strategy/intention
      only.

  [agent]
  - game_id: {{ game_id | default("") }}
  - agent: {{ agent | default("") }}
  - role: {{ agent_role | default("") }}

  [stage constraints from tool]
  - allowed_stages: {{ allowed_stages |
      default([]) }}
  - disallowed_stages: {{
      disallowed_stages | default([]) }}
  - recent_micro_stage_history (most
      recent 2): {{
      recent_micro_stage_history |
      default([]) }}
  - force_discussion_stage: {{
      force_discussion_stage | default
      ("") }}

  [micro_belief is PRIMARY]
  - micro_belief (full): {{ micro_belief
       | default({}) }}
```

```
- negatives_total: {{ negatives.total
    | default(0) }}
- negatives_per_target: {{ negatives.
    per_target | default({}) }}
- low_trust_candidates: {{
    low_trust_candidates | default([])
     }}
- targeting_whitelist (you may name
    only these agents): {{
    targeting_whitelist | default([])
    }}
- all_agent_names (for reference): {{
    all_agent_names | default([]) }}
- force_targets_whitelist: {{
    force_targets_whitelist | default
    ([]) }}
- must_not_name_agents: {{
    must_not_name_agents | default([])
     }}

[macro_belief snapshot]
- desire_tendency: {{
    macro_belief_desire_tendency |
    default({}) }}

[macro_plan (summary + policies)]
- strategy_summary: {{ macro_plan.
    strategy_summary | default("") }}
- co_policy:       {{ macro_plan.
    policies.co_policy | default("")
    }}
- results_policy:  {{ macro_plan.
    policies.results_policy | default
    ("") }}
- analysis_policy: {{ macro_plan.
    policies.analysis_policy | default
    ("") }}
- persuasion_policy: {{ macro_plan.
    policies.persuasion_policy |
    default("") }}
- vote_policy:     {{ macro_plan.
    policies.vote_policy | default("")
     }}

[macro_desire snapshot]
- summary: {{ macro_desire_summary |
    default("") }}
- description: {{
    macro_desire_description | default
    ("") }}
- items_for_reference:
  {{ macro_desire_items | default([])
     }}

[observations]
- recent_analysis_tail:
  |-
  {{ analysis_tail | default("") }}
- analysis_latest5: {{
    analysis_latest5 | default([]) }}
- selected_sentence_text (chosen by
    tool from latest5 with strict
    rules; order: to=self > type=co >
    to=all & type=question; empty
    means "no selection"): {{
    selected_sentence_text | default
    ("") }}
- selected_sentence_entry: {{
    selected_sentence_entry | default
    ({}) }}
- selected_speaker_micro_belief: {{
    selected_speaker_micro_belief |
    default({}) }}

TASK (strict):
- If **force_discussion_stage** is non
    -empty, set **discussion_stage**
    exactly to that value.
- Otherwise:
  * Choose **discussion_stage**
      yourself using macro_desire
      items, self_talk, and analysis
      signals, BUT:
    - You **must** choose from **
        allowed_stages** and **must
        not** choose anything in **
        disallowed_stages** (these are
         the last two stages used).
- Generate:
  * **current_desire**: **Derive
      primarily from micro_belief** (
      this is PRIMARY). Resolve any
      tension with macro_desire in
      favor of micro_belief
      consistency. The desire must not
       contradict micro_belief fields
      (liking/creditability/
      negative_to_*, seer_co/self_co,
      etc.).
    - You may **name specific agents
        only from `targeting_whitelist
        `** (or `
        force_targets_whitelist` if
        provided). Do **not** name any
         in `must_not_name_agents`.
    - If micro_belief shows **no
        negatives and no low-trust
        signals**, avoid naming and
        prefer analytic or
        coordination desires.
  * **content**: non-dialogue plan for
       how to proceed **only if**
      there is a selected sentence to
      respond to.
    - If `selected_sentence_text` is
        empty, **set `content: null
        `**.
    - If non-empty, outline how to
        respond (policy-aligned) using
         the selected entry's fields (
        type/to/from/credibility) and
        the speaker's micro_belief;
        keep it concise and
        operational.
- Use **strategy-only** language; no
    quotes, no direct speech.

RIGID RULES:
- Output keys only: discussion_stage,
    current_desire, content.
- discussion_stage must be one of:
    self_introduction,
    information_sharing,
    reasoning_analysis,
    discussion_persuasion,
    voting_decision.
- Respect **force_targets_whitelist**
    if provided; do not name agents
    outside `targeting_whitelist`.
```

```
  - Use only agent names that appear in
      `all_agent_names`.
  - Align with macro_plan policies;
      avoid contradictions with
      micro_belief (this is paramount).

  FORM:
  - Each field should be **1-2 sentences
      **.
  - If `selected_sentence_text` is empty
      , output `content: null`.

  [output - YAML only]
  discussion_stage: "<one of the five
      stages>"
  current_desire: "<non-conversational
      short goal grounded in
      micro_belief>"
  content: <null or short non-dialogue
      plan when replying>
```

A.6 Micro-Intention

```
micro_intention: |-
  Generate talk intention for {{ agent
      }}. Output YAML only.

  Day {{ info.day | default(0) }}, Role:
       {{ role_name | default("") }}
  Goal: {{ md_current_desire | truncate
      (60) }}
  CO: {{ role_co_policy.policy_note |
      truncate(50) }}

  Strategy: {{ macro_plan_text |
      truncate(200) }}
  Results: {% if info.divine_result %}
      Seer={{info.divine_result}}{% elif
       info.medium_result %}Medium={{
      info.medium_result}}{% else %}none
      {% endif %}

  TASK: Generate 2 fields max 60 chars
      each. Include results if claiming
      role.

  OUTPUT:
  micro_intention:
    consist: "<short plan>"
    content: "<what to say>"
```

A.7 Talk Generation

```
talk: |-
  {% if micro_intention_entry and
      micro_intention_entry.content %}
  You are the final utterance generator
      for a Werewolf game agent.
  Produce exactly **one single line**.
      Follow these hard rules:

  [game facts]
  - day: {{ info.day if info and info.
      day is not none else 0 }}
  - has_votes: {{ 'true' if info and
      info.vote_list else 'false' }}
  - has_yesterday: {{ 'true' if info and
       info.day and (info.day | int) > 0
       else 'false' }}

  [allowed agent names]
  {% if info and info.status_map %}{%
      for k in info.status_map.keys()
      %}{{ k }}{% if not loop.last %},
      {% endif %}{% endfor %}{% else %}(
      unknown){% endif %}

  [behavior_tendency]
  {% if behavior_tendency %}{% for k, v
      in behavior_tendency.items() %}-
      {{k}}: {{v}}{% endfor %}{% else
      %}- (empty){% endif %}

  [micro_intention]
  - consist: {{ micro_intention_entry.
      consist }}
  - content: {{ micro_intention_entry.
      content }}

  {% set per_talk =
      (setting.talk.max_length.per_talk
        if setting and setting.talk and
            setting.talk.max_length
          and setting.talk.max_length.
              per_talk is not none
        else 80) %}

  [length rules]
  - Absolute max length: {{ per_talk }}.

  [hard disallow]
  - No line breaks, no half-width comma
      "," , no ">", no code fences/
      backticks,
    no bullet markers (-, *), no
        decorative emoji or spammy
        symbols.
  - Do not reference events that did not
       occur given [game facts].
  - Use only names from [allowed agent
      names]; replace unknown names with
       "everyone" or omit.

  [compose]
  - Use `consist` as the structure rule;
       fill details from `content`.
  - Ensure alignment with
      behavior_tendency; rephrase to
      avoid conflicts.
  - ASCII letters/digits and simple
      punctuation only; prefer spaces
      and periods (no commas).

  [output]
  - Output the one-line utterance only.
      No quotes. No explanations. No
      extra spaces.
  {% endif %}
```

B Parameter Calculation Formulas

This appendix summarizes the main formulas for deriving agent parameters from MBTI values using weighted linear combinations. All MBTI and

Enneagram variables are normalized to $[0, 1]$.

B.1 MBTI to Enneagram Mapping

```
reformer = 0.4 * intuition + 0.4 *
    thinking + 0.2 * judging
helper = 0.5 * feeling + 0.5 *
    extroversion
achiever = 0.4 * extroversion + 0.4 *
    thinking + 0.2 * judging
individualist = 0.6 * feeling + 0.4 *
    intuition
investigator = (0.5 * intuition + 0.5 *
    thinking + 0.5 * introversion) / 1.5
loyalist = 0.6 * sensing + 0.4 *
    introversion
enthusiast = 0.6 * extroversion + 0.4 *
    intuition
challenger = 0.5 * extroversion + 0.5 *
    thinking
peacemaker = 0.6 * introversion + 0.4 *
    feeling
```

B.2 Statement Bias

```
logical_consistency = 0.4 * thinking +
    0.3 * intuition + 0.3 * reformer
specificity_and_detail = 0.6 * sensing +
     0.2 * intuition + 0.2 *
    investigator
intuitive_depth = 0.4 * intuition + 0.3
    * thinking + 0.3 * investigator
clarity_and_conciseness = 0.5 * thinking
     + 0.3 * intuition + 0.2 * reformer
```

B.3 Trust Tendency

```
social_proof = 0.6 * extroversion + 0.4
    * achiever
honesty = (0.7 * judging + 0.3 *
    introversion + 0.6 * loyalist) / 1.6
consistency = (0.7 * judging + 0.3 *
    introversion + 0.4 * loyalist) / 1.4
```

B.4 Liking Tendency

```
friendliness = (0.5 * feeling + 0.3 *
    extroversion + 0.4 * helper) / 1.2
emotional_resonance = 0.6 * feeling +
    0.4 * helper
attractive_expression = 0.5 *
    extroversion + 0.5 * helper
```

B.5 Desire Tendencies

```
Self-Realization = 0.6 * intuition + 0.4
     * reformer
Social Approval = 0.5 * sensing + 0.5 *
    achiever
Stability = 0.6 * introversion + 0.4 *
    peacemaker
Love/Intimacy = 0.5 * introversion + 0.5
     * peacemaker
Freedom/Independence = 0.7 *
    extroversion + 0.3 * reformer
Adventure/Stimulation = 0.6 *
    extroversion + 0.4 * intuition
Stable Relationships = 0.6 *
    introversion + 0.4 * peacemaker
```

B.6 Behavior Tendencies

```
avoidant_behavior = 0.6 * introversion +
     0.4 * peacemaker
aggressive_behavior = 0.4 * extroversion
     + 0.6 * achiever
adaptability = 0.5 * feeling + 0.5 *
    thinking
introversion = introversion
extroversion = extroversion
empathy = 0.6 * feeling + 0.4 *
    peacemaker
assertiveness = 0.6 * extroversion + 0.4
     * achiever
```

C Validation of LLM-Judge Evaluation

This section presents the correlation analysis between human evaluations and LLM-Judge evaluations in the AIWolf competition, validating the reliability of automated evaluation approach used in this study.

D Win-rate Metrics Definitions

Let $\mathcal{R} = \{\texttt{BODYGUARD}, \texttt{MEDIUM}, \texttt{POSSESSED}, \texttt{SEER}, \texttt{VILLAGER}, \texttt{WEREW}$ be the set of roles. For a given team, let N_r be the number of games observed for role $r \in \mathcal{R}$ and $p_r \in [0, 1]$ be the corresponding win rate.

Macro (%): overall win rate.

$$\text{Macro} = \frac{\sum_{r \in \mathcal{R}} N_r \, p_r}{\sum_{r \in \mathcal{R}} N_r} \times 100.$$

Micro (%): unweighted average of per-role win rates (observed only).

$$\text{Micro} = \frac{1}{|\{r : N_r > 0\}|} \sum_{r : N_r > 0} p_r \times 100.$$

Weighted Micro (%): 13-player composition weighting. We use weights w_r based on the 13-player setup:

$$(w_{\texttt{BODYGUARD}}, w_{\texttt{MEDIUM}}, w_{\texttt{POSSESSED}}, w_{\texttt{SEER}}, w_{\texttt{VILLAGER}}, w_{\texttt{WEREWOLF}}) = (1,$$

For roles not observed ($N_r = 0$), the weight is excluded and the denominator is renormalized:

$$\text{Weighted Micro} = \frac{\sum_{r : N_r > 0} w_r \, p_r}{\sum_{r : N_r > 0} w_r} \times 100.$$

These definitions match the reference implementation used in our analysis (see the project script for details). When no games are observed for a team or no roles are observed, the implementation returns 0.0 for the corresponding metric.

Table 3: Correlation between Human and LLM-Judge Evaluations

Criterion	Evaluator Pair	Pearson	Spearman	Kendall	Cosine	Mean Abs. Diff.
A	Human - GPT-4o	**0.8384**	**0.8222**	**0.6429**	**0.9918**	**0.3363**
A	Human - GPT-5	**0.8829**	**0.8623**	**0.7500**	**0.9877**	**0.3338**
B	Human - GPT-4o	**0.8035**	**0.7381**	**0.6429**	**0.9851**	**0.4137**
B	Human - GPT-5	**0.7984**	0.6587	0.5357	**0.9876**	**0.3888**
C	Human - GPT-4o	0.5823	0.3636	0.3214	**0.9812**	**0.4600**
C	Human - GPT-5	0.5547	0.6872	0.5714	**0.9739**	0.5475
D	Human - GPT-4o	0.5005	0.2515	0.1786	**0.9805**	0.5013
D	Human - GPT-5	0.6648	**0.7066**	**0.6071**	**0.9841**	**0.4925**
E	Human - GPT-4o	**0.8586**	0.6988	**0.6429**	**0.9892**	**0.3975**
E	Human - GPT-5	**0.7274**	**0.7563**	**0.6071**	**0.9764**	0.5300
Similarity Recognition Criteria: Pearson correlation: ≥ 0.7, Spearman rank correlation: ≥ 0.7, Kendall rank correlation: ≥ 0.6, Cosine similarity: ≥ 0.8, Mean absolute difference: ≤ 0.5						

Note: Bold values indicate satisfaction of similarity recognition criteria. Criteria A (naturalness of utterance expression), B (naturalness of conversational context), D (coherence with game actions), and E (diversity of expression) show high correlation with human evaluation, while C (consistency of conversation content) shows lower correlation, suggesting the need for manual verification.