Intent-driven AIWolf Agents with Hierarchical BDI Model and Personality

Yuya Harada, Yoshinobu Kano

Faculty of Informatics, Shizuoka University

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

{yharada, kano}@kanolab.net

INLG 2025

Research Background and Issues

Remarkable Progress of LLMs in Natural Language Processing

There is growing interest in applying LLMs to multi-agent dialogue environments that require social reasoning, such as the Werewolf game.

Challenges in Social Reasoning and Strategic Language Generation

It is difficult to handle long-term strategic planning and short-term tactical choices in an integrated manner. Maintaining consistency of intentions and strategies throughout the game is also challenging.

Representation of Personality Traits Limited to Speech Style

The systematic influence of personality on situation recognition and decision-making has not been sufficiently modeled.

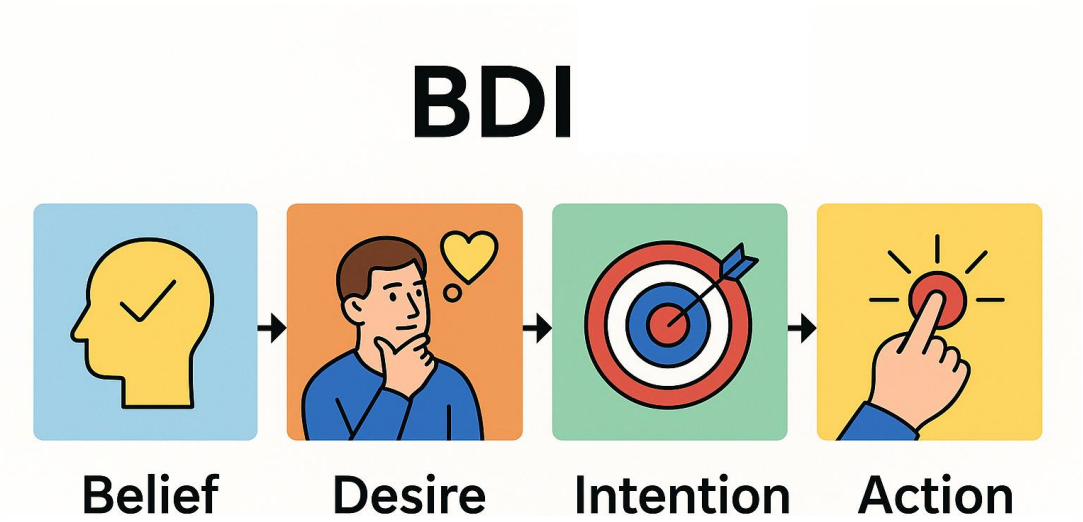


Figure: Basic Structure of a BDI (Belief–Desire–Intention) Agent

Image generated with ChatGPT (OpenAI)

Limitations of Conventional Approaches

- **Difficulty in Integrating Long- and Short-Term Strategies**
  Most existing methods treat long-term strategies and short-term tactics separately. This prevents consistent behavior and flexible policy adjustment during the game.

- **Personality Representation Limited to Speech Style**
  Personality is reflected only in wording or tone, without affecting cognition or decision-making. Thus, individual differences remain shallow and merely stylistic.

- **Rigidity of Role-Specific Tactics**
  Role-based tactics are rule-based and lack adaptability. It is hard to reproduce flexible tactical changes based on situational judgment, as seen in human players.

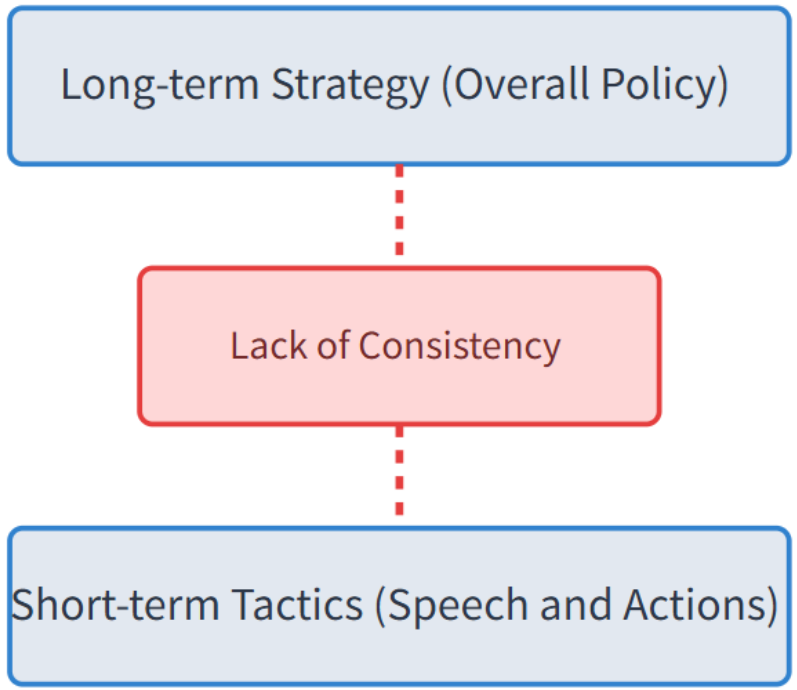


Figure: Gap Between Strategy and Tactics

Contributions of This Study

1. **Hierarchical BDI Architecture**
   A two-layer BDI model separating macro strategy and micro tactics enables consistent decision-making across different time scales.

2. **Systematic Personality Integration**
   Twenty-four traits are derived from MBTI and Enneagram, providing a computable intermediate representation that maps unstructured descriptions to behaviors.

3. **Personality-Driven Cognitive Bias**
   Personality-weighted parameters numerically model cognitive biases and motivational tendencies to produce more realistic, non-optimal behavior.

4. **Empirical Evaluation**
   Comparative experiments in the AIWolf environment show a 14.1% improvement in subjective human-likeness evaluation.

Overview of the Proposed Method

Hierarchical BDI Architecture

A two-layer structure enabling consistent decision-making across different time scales.

Macro-BDI Layer

Responsible for long-term strategy formulation and management of personality parameters. Personality traits are derived from MBTI and Enneagram models.

Micro-BDI Layer

Handles turn-level tactical decisions and immediate responses, executing concrete actions consistent with the macro-level strategy.

Personality-Weighted Parameters

Link both layers to ensure coherence in perception, evaluation, and decision-making.

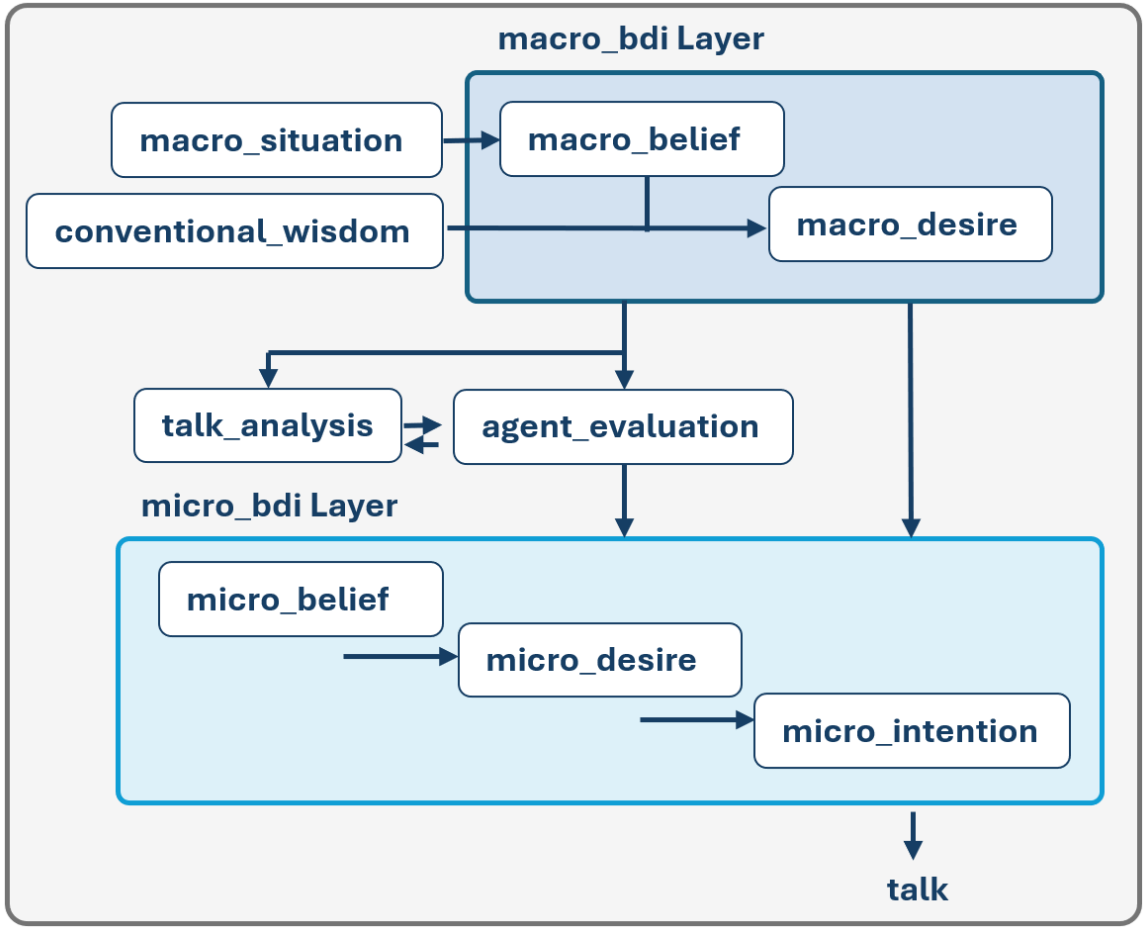


Figure: Macro–Micro Structure of the BDI Model

Macro-BDI Layer: Long-Term Strategy and Personality Parameters

Long-Term Strategy Formulation

Determines and maintains overarching policies throughout the game. Adaptively adjusts strategies in response to situational changes.

1 Management of Personality Parameters

Numerical personality traits derived from MBTI and Enneagram models. Quantifies tendencies in cognition, trust, affinity, desire, and behavior.

2 Use of the Conventional Wisdom Bank

Organizes behavioral guidelines through a role–situation matrix. Generates personality-aligned desire tendencies, providing consistent bias to the Micro layer.

3 Generation of Macro-Desires

Defines high-level desires based on role and personality under varying situations. Narrows strategic options using general knowledge and conventional patterns, and finalizes preferences according to personality traits.

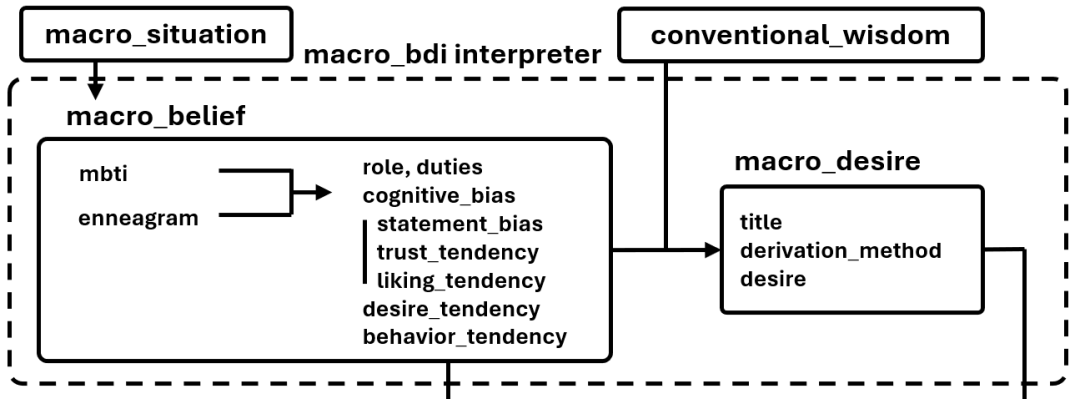


Figure: Macro BDI Structure

Integration of MBTI and Enneagram

MBTI: Four Axes, Eight Dimensions

Characterizes cognitive and judgmental tendencies across four dimensions: Extraversion/Introversion (E/I), Sensing/Intuition (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P).

Enneagram: Nine Types

Represents motivational structures through nine types: Reformer, Helper, Achiever, Individualist, Investigator, Loyalist, Enthusiast, Challenger, and Peacemaker.

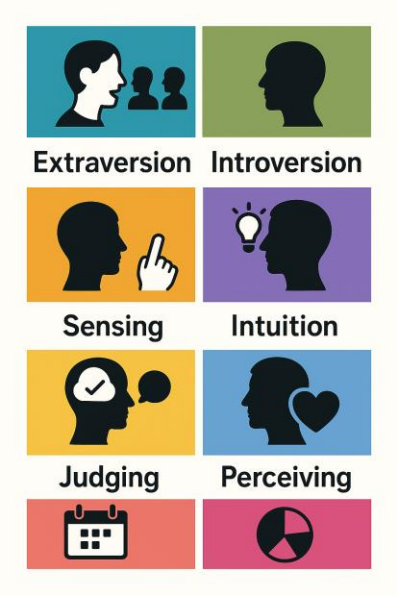
Rationale

Psychological Significance ：

MBTI is widely recognized, and its combination with the Enneagram enables multifaceted modeling of cognition, motivation, and behavior. The two frameworks align naturally with tactical choices and desire expression.

Use as an Intermediate Representation ：

Used not as psychological “truths” but as **structured, computational intermediates** for generating consistent behavior.



MBTI – Eight Cognitive Dimensions



Enneagram – Nine Personality Types

24 traits derived from MBTI and Enneagram, used as a cross-domain computational representation (cognition–motivation–behavior).

Images generated with ChatGPT (OpenAI)

Micro-BDI Layer: Turn-Level Tactics

Immediate Turn Response

React to the latest dialogue events while maintaining alignment with macro strategies. Assess the situation and select tactics each turn to produce consistent behavior.

Utterance Analysis & Relationship Update

Record utterance content, type, and reliability as mid-term memory. Continuously update long-term impressions (liking, trust) for each speaker.

Micro-BDI Processing

Micro-Belief: Trust, preference, and role information for each agent. **Micro-Desire:** Tactical goals derived from macro desires. **Micro-Intention:** Decision unit for the next single utterance.

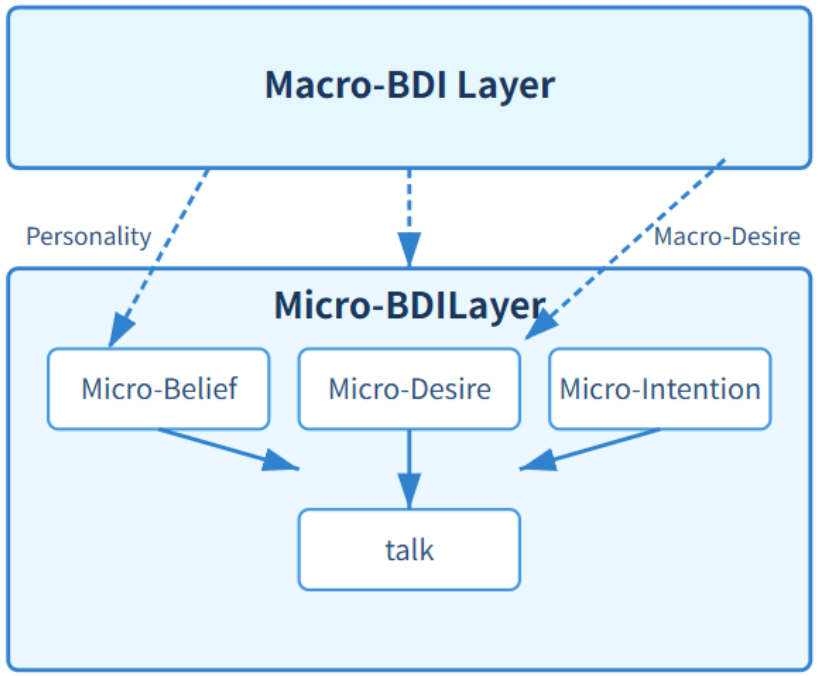


図: Micro-BDI層の構成とMacro-BDI層との連携

Utterance Generation Process

Five-Step Flow of Utterance Generation

A hierarchical BDI-based process for producing coherent speech from underlying intentions.

① Micro-Context Collection

Analyze recent utterances, trust/affinity, desires, and game state.

② Response Target Selection

Choose one target utterance by priority (Self > CO > Question > Others).

③ Micro-Desire Formation

Set a short-term goal aligned with macro desires and context.

④ Micro-Intention Construction

Define intention based on role, tendencies, and trust relations.

⑤ Final Utterance Generation

Generate a natural sentence based on the micro-intention.



Experimental Setup

AIWolf 5-Player Configuration

2 Villagers, 1 Seer, 1 Possessed, and 1 Werewolf.

Experiment Scale

6 teams (3 proposed, 3 baseline).

Round-robin format, 10 matches each (30 games total).

Technical Details

Language model: GPT-4o (temperature = 0.7)

Speech limit: 20 utterances/day (max 4 per player)

Voting: Secret ballot

5-Player Role Configuration

V ×2 S ×1 P ×1 W ×1

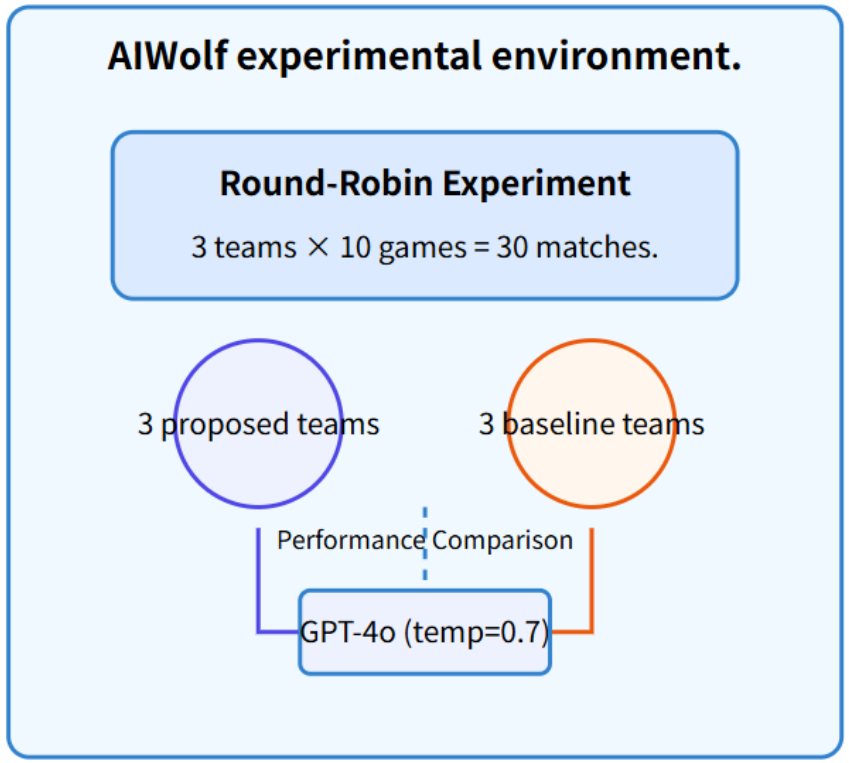


Figure: Round-robin design in a 5-player AIWolf setting

Evaluation Metrics

Both objective win-rate metrics and subjective human-likeness assessments were used to compare the proposed and baseline methods.

【Objective Evaluation】 Win-Rate Metrics

Metric	Definition
Overall Win rate	Simple average win rate across all matches.
Macro (%)	Aggregated win rate across all roles. $\text{Macro} = (\Sigma_{r \in \mathcal{R}} N_r \times p_r) \div (\Sigma_{r \in \mathcal{R}} N_r) \times 100$
Micro (%)	Average win rate for observed roles only. $\text{Micro} = (1 \div \lvert\{r: N_r > 0\}\rvert) \times \Sigma_{r:N_r>0} p_r \times 100$
Weighted Micro (%)	Average weighted by role appearance ratio. $\text{Weighted Micro} = (\Sigma_{r:N_r>0} w_r \times p_r) \div (\Sigma_{r:N_r>0} w_r) \times 100$

【Subjective Evaluation】 5-Axis LLM-Judge Assessment

Axis	Criterion	Method
A	Naturalness of expression (human like phrasing, fluency)	Ranking (1st = best) — evaluated by GPT-5
B	Contextual coherence (flow and situational relevance)	
C	Consistency of content (logical coherence, no contradiction)	
D	Behavioral alignment (match between speech, voting, divination, and attacks)	
E	Expressive diversity (variation and character consistency)	

※ *LLM-Judge validity:* Comparison with human evaluations from AIWolf competitions confirmed high correlation on axes A, B, D, and E.

Experiment Result 1: Win Rate by Role

Comparison of role-specific win rates in the 5-player AIWolf setting — showing a clear trade-off between human-likeness and strategic performance.

team	Possessed	Seer	Villager	Werewolf	Wins/Games	Macro (%)	Micro (%)	Weighted Micro (%)
Baseline (avg)	16.67%	**83.33%**	75.00%	**33.33%**	17/30	**56.67%**	**52.08%**	**59.09%**
Proposed (avg)	**33.33%**	**66.67%**	75.00%	**16.67%**	16/30	**53.33%**	**47.92%**	**54.55%**
Differences	**+16.66%**	**-16.66%**	±0.00%	**-16.66%**	-1/30	**-3.34%**	**-4.16%**	**-4.54%**

Key Observations:

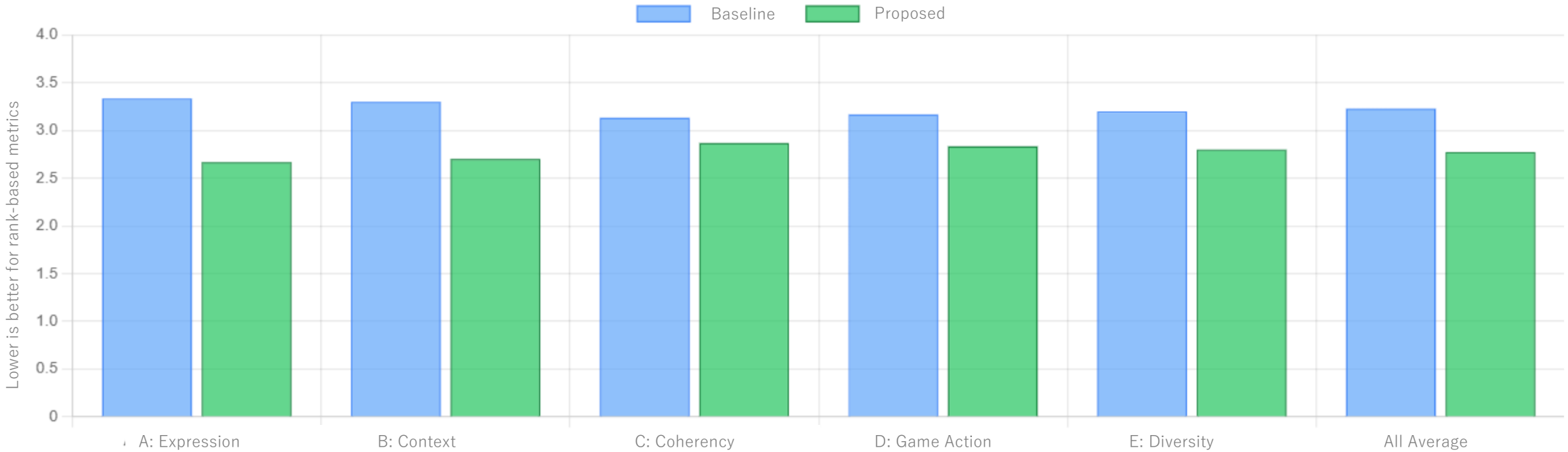
- **Major improvement as Possessed (+16.66%):** The hierarchical BDI separation proved effective for roles requiring dual-layer behavior (feigned sanity and disruption).
- **Decline as Seer and Werewolf (−16.66% each):** Personality parameters limited information disclosure and aggressive strategies.
- **Slight overall decrease (−3.34%):** Indicates a trade-off between strategic optimality and human-likeness.

Note:
Win rates are calculated per role, based on the number of games for that role.
Colored cells indicate statistically significant improvements or declines.

Experiment Result 2: Subjective Evaluation

The proposed method outperformed the baseline in all five LLM-Judge metrics.

(Ranking-based evaluation: lower = better)



Key Improvements

Overall： 3.227 → 2.773 (+**14.1%**)
B：Contextual naturalness: 3.300 → 2.700 (+**18.2%**)
E： Character consistency: 3.200 → 2.800 (+**12.5%**)

Note: Although win rate slightly decreased (−3.34 pt), human-likeness improved significantly (+14.1%).

Discussion: Trade-off Between Human-likeness and Strategic Optimality

Clear Trade-off Observed

Win rate slightly decreased (**-3.34pt)**, while subjective human-likeness improved significantly (**+14.1%).**
Results indicate that personality integration worked as intended.

Cognitive Bias Enhances Human-likeness

Confirmation bias: fixation of initial preferences
Emotional judgment: emotion-driven voting over logical consistency

Reproduction of "Role Aptitude" Differences

Halved win rate as Werewolf (33.33% → 16.67%) due to personality constraints. Introverted, low-empathy agents were less suited for aggressive strategies.

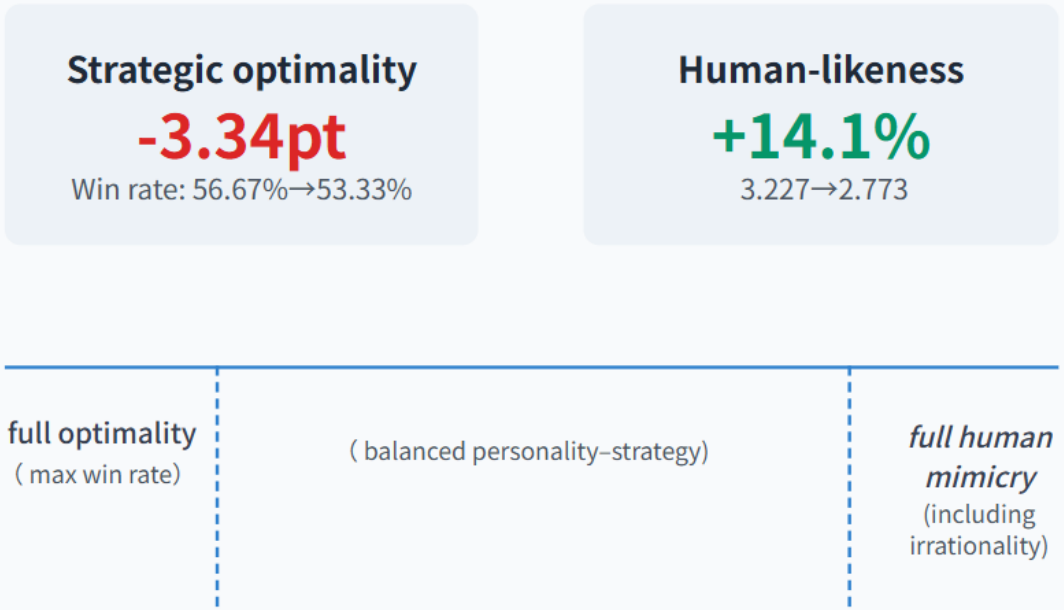


Figure: Trade-off between human-likeness and strategic optimality.

Future Work

Expansion of Experiment Scale

Current 30-match setup is preliminary and lacks statistical significance.
Plan to validate through hundreds of matches in the next AIWolf competition with diverse opponents.

Advancement of Intention Inference and Trust Estimation

Current system applies BDI only to utterance generation.
Next step: develop a BDI-based **intention inference** module for opponent speech and shift from surface-level scoring to **fine-grained trust estimation**.

Implementation Challenges and Next-Year Direction

Aim to build more **rational yet authentically human-like agents**.
Expand personality modeling (MBTI / Enneagram) and improve computational efficiency.

Research Plan

- Develop and conduct preliminary experiments on the intention inference system.

Conclusion

Integration of Hierarchical BDI and Personality Theory

Maintained consistency through a **two-layer structure**:
Macro-BDI (long-term strategy) and Micro-BDI (short-term tactics).
MBTI/Enneagram used as a **computational intermediate**, generating 24 derived traits.

Empirical Results

Win rate slightly decreased (56.67% → 53.33%),
but subjective evaluation improved significantly (**+14.1%**).
Marked gains in **contextual naturalness (+18.2%)** and **character consistency (+12.5%)**.

Realization of Human-like Decision-Making

Reproduced cognitive biases (confirmation, emotional, conformity) through personality parameters.

Demonstrated a **designed trade-off** between strategic optimality and human-likeness.

Potential Applications

- Consistent dialogue agents with diverse personalities
- Opinion-formation simulations reflecting cognitive bias
- Training/education agents with human-like imperfection
- Emotion- and personality-aware conversational assistants