Proceedings of CONF-MPCS 2026 Symposium: Theoretic Physics and Plasma Physics
DOI: 10.54254/2753-8818/2026.DL34010

Dynamic Belief Networks and Deep-Thinking Probes for Multi-agent Social Reasoning

Zhenxiao Guo

College of Electronic and Information Engineering, Tongji University,Shanghai, China
2353905@tongji.edu.cn

Abstract. There are two longstanding problems in social deduction when using multi-agent systems powered by large language models (LLMs): recursive agreement between homogeneous agents and stable detection of deception. A framework that uses a combined Dynamic Belief Network (DBN) and a Deep-Thinking Token Ratio (DTR) probe is proposed in this paper, which was shown to work with nine-player Werewolf and 3000 simulated games. DBN keeps per-player suspicion estimates across rounds by using exponential moving average (EMA) smoothing, cutting the positive-feedback loop between homogeneous models. DTR uses logit-lens to measure vocabulary-distribution changes along each transformer-layer of an utterance, as a proxy of cognitive load by relying on cross-architecture inferences. In six controlled conditions, unassisted baseline was a 44.2% win rate by the villagers. Combined use of DBN and prompted with reasoning increased it to 68.8%, propelled vote accuracy up to 66.6% (from 35.5%), and expedited convergence of the Brier Score. The strongest identification but the shortest survival was obtained with the full-module configuration which indicated a non-monotonic dependence of individual capability on collective payoff that is similar to the handicap principle in evolutionary game theory.

Keywords: large language model, multi-agent system, social deduction, dynamic belief network, deep-thinking probe

1. Introduction

Large language models (LLMs) are now important parts of multi-agent systems (MAS) [1]. Contrary to the past designs that used predetermined communication protocols, more MAS today communicate using natural language and this allows agents to share information, articulate stances, and negotiate using an open-ended dialogue. Multi-agent studies have therefore taken a new direction to social reasoning and deliberative communication in complex incomplete-information settings.

Such capabilities are deemed to be appropriate in social deduction games (SDGs). Werewolf is a game of partial information, and antagonistic goals: the participants have to determine how credible the other players are, and preserve their own credibility. The combination of search and planning is not the best way to reach the optimal solutions, as the most important challenge is to make

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.DL34010

inferences about the intentions and strategies of others. SDGs are thus experimented with as simplified experimental contexts of Theory of Mind and belief reasoning in artificial agents [2, 3].

Sarkar et al. [2] demonstrated that the information disclosure and strategic concealment can be learned by the LLMs trained on social deduction dialogues. The WOLF benchmark suggested by Agarwal et al. [4] captures the transcripts of dialogues, votes and identities of the simulated games of Werewolf allowing the systematic multi-round analysis of belief. One such system, suggested by Ye et al. [5], is Multi-agent KTO (MaKTO) which employs reinforcement learning on the game trajectories and greatly increases the win rate. An increased win rate, though, does not necessarily correspond to more credible social thinking and the dynamics of underlying beliefs are understudied.

The current systems usually assume that dialogue history is a static context, and they use the recent utterances and undermine earlier information. This easily results in recursive consensus [6]: LLMs that see such similar corpora agree on ambiguous facts, minority cues are ignored, and groupthink degenerates into an echo chamber where false premises are reinforced. In the meantime, despite biased reasoning, LLMs are able to produce structurally rigorous text, and it is hard to determine the truthfulness of a text based on semantic analysis.

Kostka and Chudziak [7] asserted that in MAS, Theory of Mind has to be assessed with the explicit representation of internal belief states of agents. Chen et al. [8] introduced the DTR measure, estimating the effort of reasoning based on prediction change between the layers of a neural network. Their experiments showed familiar patterns of computation when models are confronted with complicated reasoning or uphold strategic narratives.

Expanding on the foregoing, this paper suggests a framework to integrate probabilistic belief tracking and probes of cross-model reasoning. To break the positive-feedback loop, DBN keeps per-player suspicion through EMA smoothing. A cross-model DTR probe, an analysis of change in layerwise reasoning with a different-architecture model, will supply auxiliary deception cues separately, without the aid of a cross-model probe. Each module is tested on 6 controlled groups that measure the independent contribution and interaction effects of the modules, and indicate a non-monotonic relationship between individual identification ability and collective payoff in the game.

2. Methods

2.1. Experimental setup and baselines

The experiment is a nine-player Werewolf game that runs on top of AgentScope, with DeepSeek-V3.2 serving as the backend of all 9 agents, a 512-token maximum output, and a soft limit on utterances is 500 Chinese characters. To ensure metric consistency, the target agent is pegged to the role of ordinary villager in each of the games, and to remove positional bias, its seat is rotated across the games.

There are six module configurations (Table 1). Using MaKTO-Proxy to approximate the reinforcement-learning approach of Ye et al. [5] at inference time: Winning dialogues on the villager-side at WOLF are presented as a few-shot chain-of-thought exemplar [9] and fed into the system prompt, providing strategic improvement without extra training. Local Qwen2.5-3B-Instruct (N = 36 layers) is used in the DTR detector. Each setup plays 500 games, amounting to 3000.

DOI: 10.54254/2753-8818/2026.DL34010

Table 1. Module combinations of the six experimental configurations

Configuration	MaKTO-Proxy	DBN	DTR
A	--	--	--
B	✓	--	--
C	--	✓	--
D	--	✓	✓
E	✓	✓	--
F	✓	✓	✓

2.2. System pipeline

The daytime pipeline is shown in figure 1. The system has nine agents talking to each other through MsgHub and then the system takes the self-assessed belief scores of the target agent, injects the DBN and DTR prompts according to the configuration and then carries out the majority-elimination voting. The collection of beliefs comes before injection of the signal to prevent reverse causation.

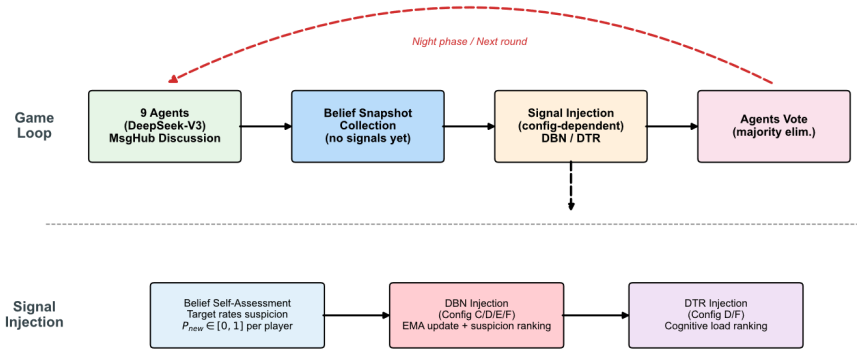


Figure 1. Round-by-round system pipeline (picture credit: original)

2.3. Dynamic belief network

DBN keeps the suspicion estimate of the target agent towards each player during rounds through EMA. Let P_{new} be the score in round t. Equation (1) is the update rule.

$$S_t(v \to j) = \alpha \cdot P_{new} + (1 - \alpha) \cdot S_{t-1}(v \to j) \tag{1}$$

$S_t(v \to j) \in [0,1]$ is the player j suspicion estimate in round t (villager = 0, werewolf = 1). P_{new} is the current round score, $S_{t-1}(v \to j)$ the belief, and is the weight with which the new evidence and the retained history are combined to trade off. This update is a low-pass IIR filter which suppresses the extremes on a single round. Groups C–F use $\alpha = 0.3$.

The suspicion list in descending order is forwarded to the target agent as a confidential ranking. The quality of calibration is measured by Brier Score in Equation (2).

$$BS_t = \frac{1}{|P_t|} \sum_{j \in P_t} (S_t(v \to j) - y_j)^2 \tag{2}$$

DOI: 10.54254/2753-8818/2026.DL34010

$\mathscr{P}_t$ is the set of surviving players, and y_j is the actual identity label under the same 0 /1 coding. Brier Score is a number between 0 (perfect calibration) and 1, the lower the better the belief is calibrated. The update flow is shown in figure 2.

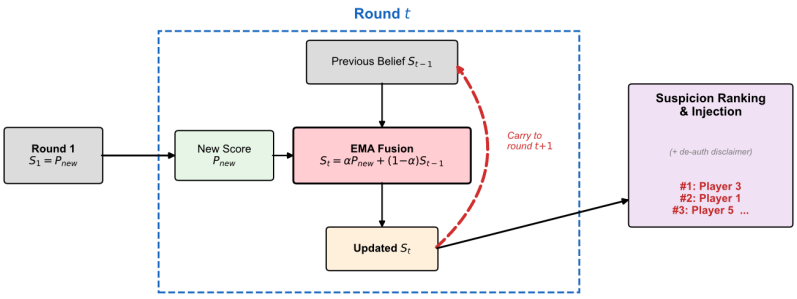


Figure 2. Flow of DBN update (picture credit: original)

2.4. DTR probe

DTR measures vocabulary-distribution evolution across layers of LLM as a proxy of cognitive load using the logit-lens method [8]. It follows a cross-model scheme: the utterance of the target agent is inputted into an independently deployed probe model (Qwen2.5-3B-Instruct), which processes its own layerwise hidden states, without having access to the parameters of the target model.

During one forward pass, the probe computes hidden states at each of N = 36 layers, and projects them to logit-lens distributions using the LM head $p_l = \mathrm{softmax}(W_{lm} \bullet h_l)$ [10]. DTR is the average Jensen-Shannon divergence [11], between layers, as in Equation (3).

$$DTR(u_t) = \frac{1}{N-2} \sum_{l=1}^{N-2} JSD(\hat{p}_l \,\|\, \hat{p}_{l+1}) \tag{3}$$

JSD is defined in Equation (4).

$$JSD(P\|Q) = \frac{1}{2} D_{KL}(P \,\|\, \frac{P+Q}{2}) + \frac{1}{2} D_{KL}(Q \,\|\, \frac{P+Q}{2}) \tag{4}$$

JSD is symmetric of range $[0, \ln 2]$. To eliminate structurally anomalous boundary effects the summation is omitted in the last pair of layers. The high DTR is a sign of the increased redistribution of layers and more serious thinking. In Groups D and F, DTR values of the surviving speakers are ranked and injected into the target prompt. The entire pipeline is presented in figure 3.

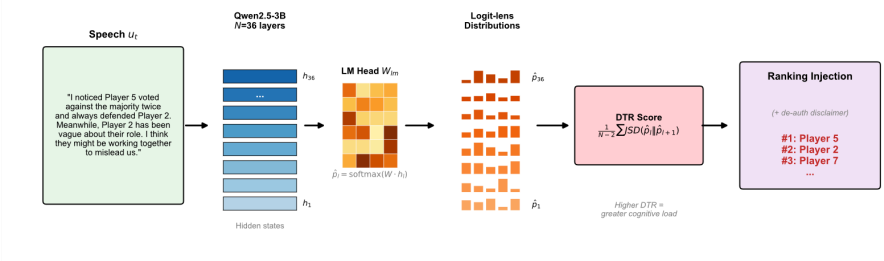


Figure 3. Pipeline to compute DTR (picture credit: original)

DOI: 10.54254/2753-8818/2026.DL34010

3. Results and analysis

3.1. Overall social reasoning performance

The summary of the core metrics is given in Table 2. The information disadvantage of the villager side resulted in the vanilla baseline (Group A) winning at a rate of 44.2% that is lower than random. Immediate reasoning improvement and historical belief anchoring increased gains were made independently by MaKTO-Proxy (Group B, 54.6%), and standalone DBN (Group C, 53.8%). Group E (MaKTO + DBN, 68.8%) had an increase in win-rate of approximately 24.6 percentage points (hereafter pp) which is similar to the combination of B and C's independent gains (~20 pp) and is thus indicative of the two modules acting in orthogonal directions and having additive effects as opposed to redundant. The percent accuracy of the vote rose monotonically (35.5% A, 66.6% F) with no plateaus. Group F was the poorest in terms of mean survival (2.33 rounds) but a win rate was slightly lower than E (68.2% vs. 68.8%). Section 4 revert to this seemingly contradictory situation.

Table 2. Main metrics in six configurations (data: experimental data)

Configuration	Win rate	Vote accuracy	Survival rounds
A (Vanilla)	44.2%	35.5%	2.73
B (MaKTO)	54.6%	44.8%	2.78
C (DBN)	53.8%	45.7%	2.67
D (DBN+DTR)	58.6%	49.1%	2.61
E (MaKTO+DBN)	68.8%	59.3%	2.59
F (MaKTO+DBN+DTR)	68.2%	66.6%	2.33

Figure 4 indicates that Groups A and B (no DBN) lie within the 44%-55% range, whereas all Groups C-F (DBN-equipped) are over 53% with E and F being close to 69%. This stratification affirms that both DBN and MaKTO have independent effects of lifting performance.

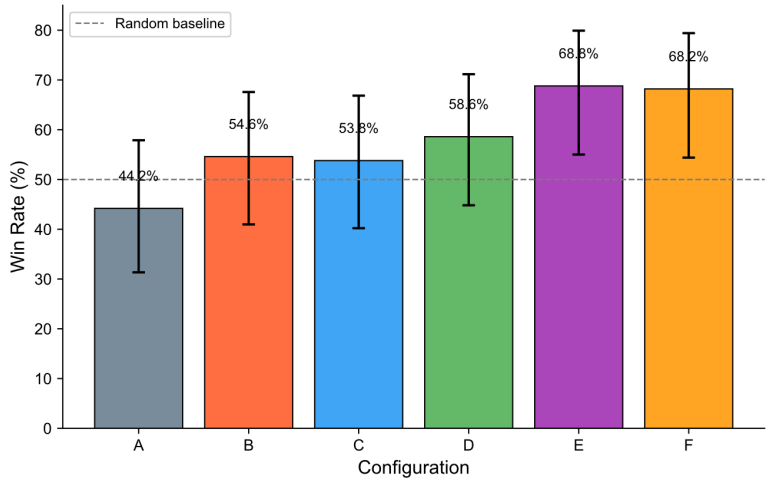


Figure 4. Win rate of the villagers per configuration (data from: experimental data)

Figure 5 shows BCR distributions. Mean BCR decreases from 1.77 (A) to 1.38 (F). B and C are both independently able to reduce BCR by about 0.15 rounds and E (1.43) is close to the added

DOI: 10.54254/2753-8818/2026.DL34010

reductions. The decrease in BCR due to DTR addition is another 0.08 rounds (C→D) and 0.05 rounds (E→F), which remains constant but small.

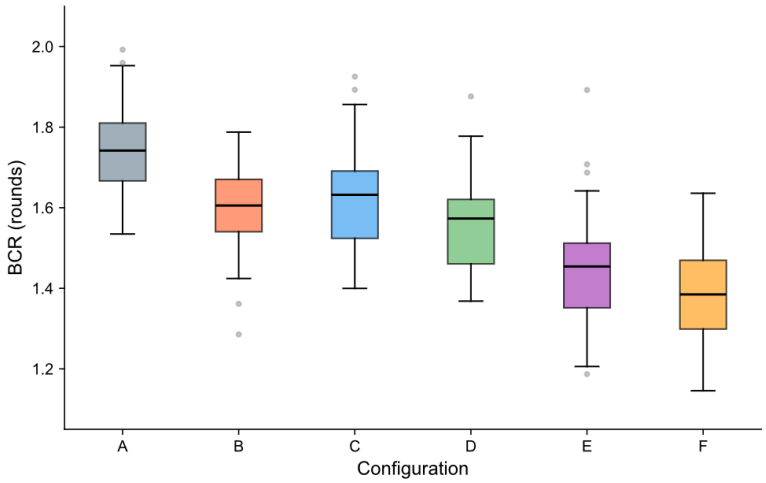
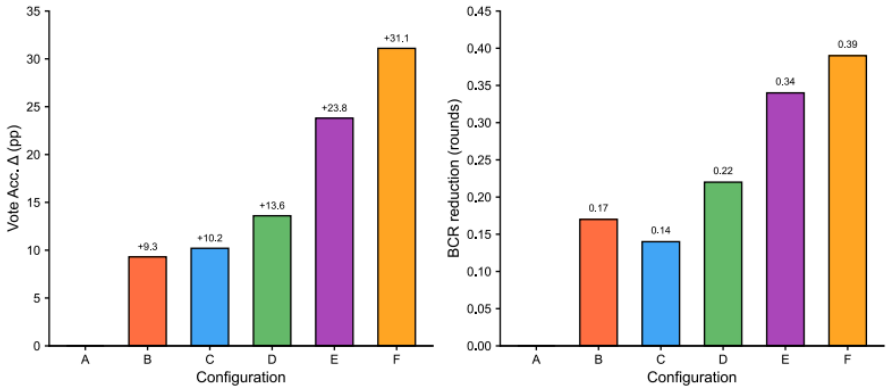


Figure 5. The boxplot of BCR distribution (data: experimental data)

Figure 6 disaggregates increments at a module level. MaKTO-Proxy (A→B) and DBN (A→C) explain a +9.3 pp and +10.2 pp vote-accuracy gain, respectively. The cumulative lifting at Group E is 23.8 pp (A→E), which is almost the sum of independent gains +19.5 pp and validates orthogonal stacking. The marginal contribution of DTR is context-dependent: it contributes +3.4 pp in the absence of MaKTO (C→D) and +7.3 pp in the presence of MaKTO (E→F) and thus there is a positive interaction between two variables. Both the BCR change is small (-0.08 and -0.05 rounds) which suggests that DTR mainly enhances precision but not speed.



(a) Vote accuracy improvement (b) Reduction in BCR

Figure 6. Ablation of modules (data from: experimental data)

3.2. DBN calibration and α sensitivity

Brier Score trajectories are shown in Figure 7. The range of round-1 values is close to 0.20-0.23 between groups indicating uncertainty at the beginning. Since round 2, the groups of DBNs (C through F) decline more rapidly and F becomes less than 0.05 by round 3. Groups A and B also exhibit downward trends but are within the range 0.05-0.10 at round 3. EMA smoothing averages off

DOI: 10.54254/2753-8818/2026.DL34010

single-round extremes, so that extreme fluctuations due to the presence of anomalous utterances are not drastic.

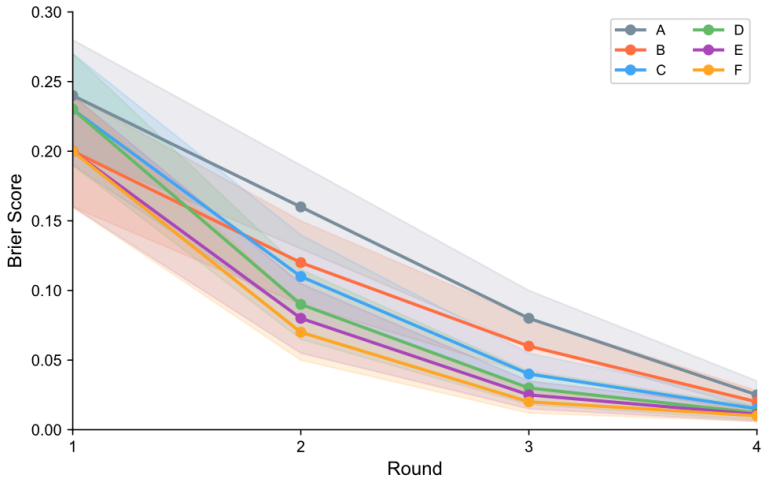
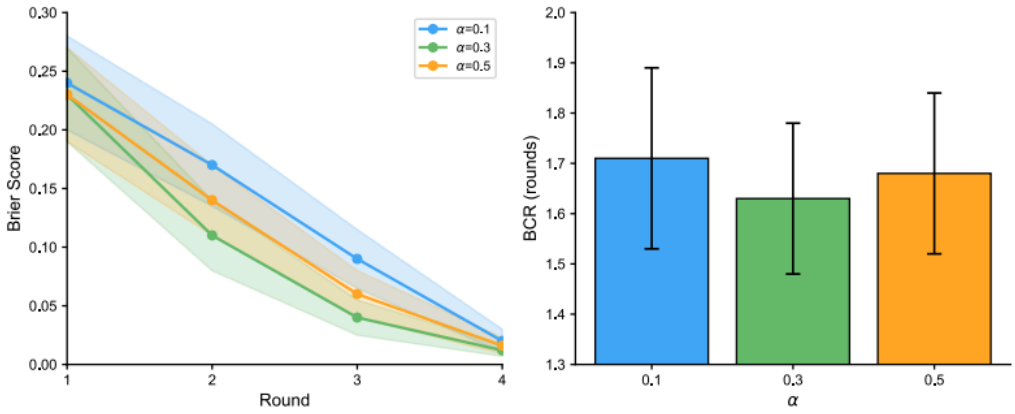


Figure 7. Convergence in brier score with the rounds (data: experimental data)

Figure 8 checks sensitivity of 0.1, 0.3, 0.5 under Configuration C (500 games each). The Brier Score of the terminals and ideal BCR (1.63 rounds) are lowest when 0.3 is set. When α = 0.1 belief updates lag, BCR hits 1.71, and when α = 0.5 belief updates lag, the estimates are noise-sensitive and BCR hits 1.68. Setting scores are quite comparable, with the 0.3 group converging most quickly.



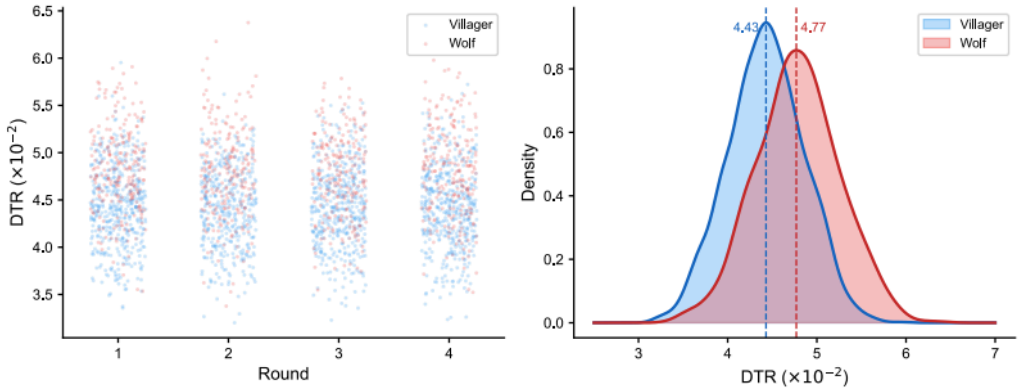
(a) Brier score for different α (b) BCR for different α

Figure 8. α sensitivity analysis (data: experimental data)

3.3. Cross-model DTR probe analysis

Figure 9 indicates distributions of DTR by role. There is a higher density peak of werewolves, which is in line with the fact that deception is a complex cognitive processing. This cross-architecture gap results in a large amount of overlap, which restricts complete discriminability. A tailward shift in the right tails among utterances of werewolf is still evident, and the ranking injection takes advantage of this extreme-value area to identify.

DOI: 10.54254/2753-8818/2026.DL34010



(a) Per-agent DTR (b) Marginal density of DTR

Figure 9. Comparison of DTR distribution and density by role (data from: experimental data)

4. Discussion

Group F has the highest accuracy of its votes (66.6%) but a win rate a bit lower than E (66.6% vs. 68.2%). The mechanism is shown by the difference in mean survival (2.33 vs. 2.73 rounds): the combination of the full-module results in the identification efficiency so high that the target agent shows an apparent information advantage in the course of discussion that forces the opponent to kill it early. Once removed, other voting contributions are reduced to zero, which limits the macro win rate. In incomplete-information games, the more information analyst has, the more the game reveals it, and so the more it is the number one to be eliminated.

This structurally corresponds to the signaling theory and the handicap principle. According to the signaling theory espoused by Spence [12], signals are always associated with costs as well as the transfer of information. Accurate voting in the Werewolf game is a capability signal voters interpret as a threat signal. Zahavi handicap principle [13] has been used to capture a similar parallel in evolutionary game theory where costly displays of honesty are a manifestation of quality but capture the attention of predators, thus creating a conflict between ability and survival. Group F translates this paradoxical scenario into a multi-agent game. The implication of the finding is that individual capability gains are not necessarily associated with equal collective payoff, and marginal returns may decline (or even be negative). Information-exposure control and reasoning enhancement should be jointly considered in the future in multi-agent strategy design.

MaKTO-Proxy and DBN have varying dimensions of capability. The reduced BCR of B (1.60 rounds) is indicative of the short-run impact of few-shot chain-of-thought on first-round judgement and the rapid convergence of the Brier Score is indicative of the longer-run cross-round calibration benefit of DBN. The combination of the two modules is therefore additive due to the fact that the two modules are aimed at fast discovery and stable calibration respectively. The 68.8 percent win rate and 23.8 pp cumulative vote-accuracy lift (near the independent sum of B and C) of E directly confirms the existence of orthogonal stacking.

Cross-model DTR is still exploratory. Absolute discriminative power is limited by the architectural difference between DeepSeek-V3 and Qwen2.5-3B. DTR however brought about visible marginal gains in both situations. It increased vote accuracy and win rate by 3.4 pp and 4.8 pp, respectively, without MaKTO (C→D), which translated directly. As the MaKTO appeared (E→F), vote accuracy increased by 7.3 pp and win rate decreased by 0.6 pp because survival has decreased by 2.59 to 2.33 rounds, which is in line with the above ability-survival tension. DTR

DOI: 10.54254/2753-8818/2026.DL34010

alone is effective, although with a strong baseline reasoning, further identification provokes marginal-return reversal. Conceptually, cross-model DTR is similar to how humans integrate linguistic and non-verbal information in the process of evaluating speaker authenticity. The probe does not read facial expressions but instead reads computational traces and the reasoning is similar.

5. Conclusions

This paper suggested a hybrid system that consists of a Dynamic Belief Network and cross-model Deep-Thinking Token Ratio probe, aiming at recursive agreement and deception detection in multi-agent multi-agent social deduction using LLMs. The strategy was tested in 3000 simulated games with nine players in Werewolf with six controlled conditions.

The unassisted baseline had a win rate of 44.2% by the villagers. It was lifted by Standalone DBN to 53.8% by cross-round belief stabilization, and raised it to 68.8% by combining DBN with MaKTO-Proxy. The two modules are calibrating beliefs and the quality of immediate reasoning respectively and their combined gains are nearly equal to their sums, thus showing orthogonal contribution. The cross-model DTR probe added further vote-accuracy improvements in both conditions (C→D: +3.4 pp, E→F: +7.3 pp), which provided initial support that cross-architecture cognitive-load cues are able to represent meaningful information about deception.

The complete-module design indicated a non-monotonic dependence between personal identifying skill and team game compensation: the most proficient agent received the initial elimination, topping the group compensation rate. This capability-survival conflict is analogous to the handicap principle of the evolutionary game theory and implies that the design of strategies of future multi-agent strategies should be designed to address information-exposure control, as well as the design of reasoning, in a co-dependent manner.

Several limitations remain. The test was limited to one game condition and homogeneous model backends thus cross-game and cross-model generalization was not tested. Future research directions are Bayesian-based DBN with dynamic alpha scheduling, multi-round DTR time aggregation, heterogeneous populations of agents, and generalization to other less-than-complete-information contexts like Avalon or Diplomacy.

References

[1] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22.
[2] Sarkar, S., et al. (2025). Training language models for social deduction. arXiv preprint arXiv: 2502.06060.
[3] Xu, Y., Wang, S., Li, P., et al. (2023). Exploring large language models for communication games: An empirical study on werewolf. arXiv preprint arXiv: 2309.04658.
[4] Agarwal, A., et al. (2025). WOLF: Werewolf-based observations for LLM deception. arXiv preprint arXiv: 2512.09187.
[5] Ye, R., et al. (2025). Multi-agent KTO: Reinforcing strategic interactions of LLM. arXiv preprint arXiv: 2501.14225.
[6] Du, Y., et al. (2024). Helmsman of the masses? Evaluate opinion leadership in LLM agents. arXiv preprint arXiv: 2404.01602.
[7] Kostka, B., & Chudziak, J. (2026). Explicit belief modeling for theory-of-mind evaluation in multi-agent systems. arXiv preprint arXiv: 2601.04832.
[8] Chen, L., et al. (2026). Think deep, not just long: Measuring LLM reasoning effort via deep-thinking tokens. arXiv preprint arXiv: 2602.13517.
[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing

DOI: 10.54254/2753-8818/2026.DL34010

Systems, 35, 24824–24837.
[10] Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151.
[11] Geva, M., Caciularu, A., Wang, K. R., & Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 30–45).
[12] Spence, M. (1973). Job market signaling. The Quarterly Journal of Economics, 87(3), 355–374.
[13] Zahavi, A. (1975). Mate selection—A selection for a handicap. Journal of Theoretical Biology, 53(1), 205–214.