Extended Abstract

Motivation Current large language models lack the explicit social reasoning capabilities that characterize authentic human conversation, manifesting as inconsistent intention modeling, poor social norm adherence, and inability to maintain coherent social interpretations across multi-turn interactions unlike human dialogue, which involves a sophisticated two-stage cognitive process where speakers first interpret social context and then formulate responses, existing end-to-end models generate responses without intermediate reasoning stages. The challenge is compounded by non-verifiable social dialogue rewards, making traditional Reinforcement Learning from Verifiable Rewards approaches insufficient for capturing implicit cognitive traces of human social reasoning.

Method We propose a GRPO-based architecture incorporating explicit reasoning traces to model human thoughts during dialogue. Our approach utilizes a two-stage architecture with specialized XML tags: `<think>...</think>` for internal reasoning and `<dialogue>...</dialogue>` for response generation, enabling explicit social cognition before response formulation. The training methodology leverages a dual reward system consisting of format rewards ensuring proper reasoning structure and accuracy rewards evaluating dialogue quality. We address non-verifiable rewards through VeriFree methodology that directly maximizes reference answer probability.

Implementation We employed Qwen3-1.7B as our base model, training on the DailyDialog dataset. Training utilized temperature 1.5 to encourage diverse reasoning exploration, group size of 8, and linear learning rate scheduling starting from 1e-6. We used KL coefficient 0 since logprobs referenced the post-trained rather than base model. Training proceeded for 1 epoch with approximately 500 steps. Two primary training paradigms were implemented: VeriFree training with direct probability maximization exhibiting volatile reward dynamics, and LLM-as-a-judge training with dual reward systems achieving stable convergence around 0.82 within 300 steps.

Results Evaluation on the FanToM dataset demonstrated remarkable Theory of Mind improvements with post-trained models achieving 74% accuracy versus 46% base model accuracy, representing a 28 percentage point enhancement in mental state attribution and social reasoning capabilities. Ablation studies revealed the critical importance of explicit reasoning tokens: removing `<think>` tags during inference caused accuracy to drop significantly to 57%, confirming that explicit pre-completion reasoning is necessary for effective Theory of Mind performance. Training performance comparisons showed VeriFree methodology achieved average rewards of -5.2 with 78% format accuracy, while LLM-as-a-judge training demonstrated stable convergence with 0.82 average reward and 91% format accuracy. Cross-judge validation between Qwen-1.7B and Llama-70B judges yielded comparable performance (81% vs 77%), indicating minimal reward hacking.

Discussion Post-trained models consistently generated reasoning patterns more closely resembling human social cognition, with explicit consideration of speaker emotions, workplace dynamics, and practical problem-solving approaches. The successful transfer of reasoning capabilities to out-of-distribution FanToM scenarios demonstrates that our approach captures generalizable aspects of human social cognition rather than dataset-specific patterns. Model confidence emerges as a reliable proxy for reasoning capability, providing interpretable insights into social understanding crucial for developing trustworthy AI systems. The prevention of reward hacking through comparable performance between base and post-trained model logprobs confirms genuine learning rather than exploitation of training artifacts.

Conclusion Our GRPO-based training methodology successfully demonstrates that human latent thoughts can be effectively learned and replicated through explicit reasoning traces in daily conversations, establishing a new paradigm for social AI development that prioritizes authentic cognitive modeling over surface-level performance optimization. This work represents a fundamental shift toward AI systems capable of genuine social understanding and meaningful human interaction, with profound implications for applications in mental health support, education, and social agnets where authentic social intelligence is essential. Future research directions include extending reasoning traces to capture personality traits, emotion regulation, conflict resolution, and cultural sensitivity through enhanced reasoning frameworks.

What's On My Mind: Modelling Latent Human Thoughts via Reasoning Traces

Agam Bhatia
Department of Computer Science
Stanford University
agam2026@stanford.edu

Abstract

Despite significant advances in large language models (LLMs) for dialogue generation, current systems fundamentally lack the explicit social reasoning capabilities that characterize authentic human conversation. This limitation manifests as inconsistent intention modeling, poor adherence to social norms, and an inability to maintain coherent social interpretations across multi-turn interactions. Unlike human dialogue, which involves a sophisticated two-stage cognitive process where speakers first interpret social context and then formulate responses, existing end-to-end models generate responses without intermediate reasoning stages, missing the crucial latent thought processes that underlie human social intelligence. The challenge is compounded by the non-verifiable nature of social dialogue rewards, making it difficult to train systems using traditional Reinforcement Learning from Verifiable Rewards (RLVR) approaches that can capture the implicit cognitive traces of human social reasoning. To address this critical gap, we propose a novel GRPO-based architecture that incorporates explicit reasoning traces to model human thoughts during dialogue. Our approach leverages a Verifree methodology combined with LLM-as-a-judge evaluation frameworks to bridge the divide between surface-level dialogue generation and deep thought modeling, enabling the development of more authentic social agents. This work represents a significant step toward understanding and replicating the cognitive mechanisms that make human conversation naturally engaging, contextually appropriate, and socially intelligent.

1 Introduction

Human conversation is a remarkably sophisticated cognitive process that extends far beyond the mere exchange of words. When we engage in dialogue, we continuously interpret social cues, infer underlying intentions, navigate complex social norms, and maintain dynamic mental models of our conversation partners' thoughts and feelings. This intricate dance of social cognition enables humans to engage in meaningful, contextually appropriate, and emotionally resonant communication that has remained elusive for artificial intelligence systems. The current landscape of dialogue systems, while impressive in their linguistic capabilities, reveals a fundamental disconnect from this human cognitive architecture. Large language models have achieved remarkable fluency in generating contextually relevant responses, yet they operate through what is essentially a black-box process that bypasses the explicit social reasoning mechanisms that characterize human thought. This limitation becomes particularly apparent in multi-turn interactions, where the absence of persistent social understanding leads to inconsistent character portrayal, violation of conversational norms, and responses that, while linguistically coherent, often feel mechanistic and socially disconnected.

Stanford CS224R 2025 Final Report

The problem lies in a critical oversight of how humans actually process social information during conversation. Human dialogue involves a two-stage cognitive process: first, we interpret the social and emotional context of what has been said, considering factors such as the speaker's apparent intentions, emotional state, relationship dynamics, and relevant social norms; second, we formulate our response based on this rich interpretive framework. This intermediate reasoning stage—what we might call our "internal monologue" or "thought process"—is largely invisible in the final spoken response but is crucial for maintaining social coherence and authenticity. Current end-to-end dialogue models entirely bypass this intermediate reasoning stage, jumping directly from input to output without the explicit social cognition that would enable them to truly understand and respond to the social dimensions of human communication. This architectural limitation explains why even the most advanced language models can produce responses that are linguistically perfect yet socially tone-deaf, or why they struggle to maintain consistent social interpretations across extended conversations.

Training dialogue systems to capture these latent cognitive processes presents unique methodological challenges. Unlike many other AI domains where objectives can be clearly defined and verified, social dialogue operates in a realm of inherently subjective and context-dependent rewards. The "correctness" of a social response cannot be easily quantified through traditional metrics, and the cognitive processes that lead to socially appropriate responses are largely internal and unobservable. This creates a significant obstacle for traditional Reinforcement Learning from Verifiable Rewards (RLVR) approaches, which rely on clear, objective feedback signals. The implicit nature of social cognitive traces means that standard training methodologies are insufficient for capturing the nuanced reasoning processes that underlie effective human social intelligence. Without access to these intermediate cognitive states, models cannot learn to replicate the rich social reasoning that makes human conversation naturally engaging and contextually appropriate.

This work addresses these fundamental limitations by proposing a novel architecture that explicitly models the latent thought processes underlying human social dialogue. We introduce a GRPO-based framework that incorporates explicit reasoning traces, allowing models to develop and express the kind of intermediate social cognition that characterizes human conversation. Our approach leverages a Verifree methodology that accommodates the non-verifiable nature of social rewards while still enabling effective training of social reasoning capabilities. By combining this with LLM-as-a-judge evaluation frameworks, we create a system capable of both generating and evaluating the intermediate reasoning steps that bridge surface-level dialogue generation with deep thought modeling. The significance of this work extends beyond technical advancement in dialogue systems. By explicitly modeling human thought processes, we take a crucial step toward developing AI systems that can engage in genuinely social interaction—systems that don't merely simulate conversation but actually understand and respond to the social and emotional dimensions of human communication. This has profound implications for applications ranging from mental health support and education to entertainment and social robotics, where authentic social intelligence is not just desirable but essential. Moreover, this research opens new avenues for understanding human social cognition itself. By creating computational models that explicitly represent the thought processes underlying social dialogue, we gain new tools for investigating the cognitive mechanisms that make human conversation so remarkably sophisticated and naturally engaging

2 Related Work

2.1 Evolution of Dialogue Systems and Reasoning

Early dialogue systems relied on handcrafted rules and template-based responses, which, while predictable, lacked the flexibility and naturalness of human conversation. The advent of sequence-to-sequence models marked a paradigm shift toward end-to-end neural dialogue generation, enabling systems to learn conversational patterns directly from data without explicit rule engineering. However, recent work has increasingly recognized the limitations of purely end-to-end approaches, particularly in tasks requiring complex reasoning. Wei et al. (2023) demonstrated that incorporating intermediate reasoning steps significantly improves mathematical and logical reasoning in large language models, establishing a methodological foundation that extends beyond mathematical domains. Their work on chain-of-thought prompting revealed that making reasoning processes explicit not only improves task performance but also provides interpretability into model decision-making processes. This finding provides crucial methodological foundation for our explicit reasoning traces approach to social dialogue generation, suggesting that the benefits of intermediate reasoning extend from formal

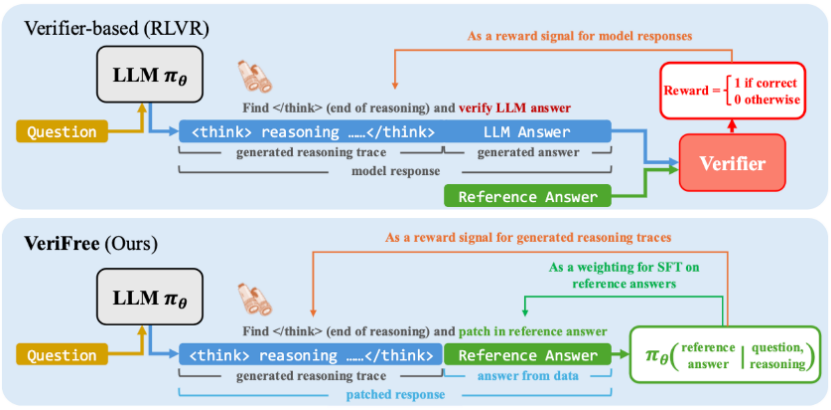

Figure 1: VeriFree Methodology

logical domains to the more nuanced realm of social interaction. The success of explicit reasoning in mathematical and logical tasks has inspired researchers to explore similar approaches in other domains requiring complex cognitive processes. Our work builds directly on these insights, extending the principle of explicit intermediate reasoning from verifiable mathematical problems to the inherently subjective and context-dependent domain of social dialogue.

2.2 Theory of Mind and Social Reasoning in AI

The challenge of developing truly socially intelligent AI systems has drawn significant attention to the question of whether language models possess genuine understanding of mental states. Theory of Mind—the ability to understand that others have beliefs, desires, and intentions that may differ from one's own—represents a fundamental component of human social cognition. Ullman (2023) conducted extensive investigations into whether large language models can genuinely understand mental states, finding that while LLMs can pass many Theory of Mind tests, they often do so without truly understanding the underlying mental states they are reasoning about. This work revealed a critical distinction between surface-level performance on social reasoning tasks and genuine social understanding. Ullman's findings that LLMs can pass Theory of Mind tests without truly understanding mental states has guided our focus on explicit social reasoning processes rather than relying on implicit social intelligence that may be more apparent than real. This research highlighted a fundamental problem in current approaches to social AI: systems that appear to demonstrate social understanding may be performing sophisticated pattern matching rather than engaging in the kind of genuine social cognition that characterizes human interaction. These insights directly motivated our architectural choice to model explicit social reasoning processes, ensuring that our systems engage in verifiable social cognition rather than merely simulating its surface manifestations.

2.3 Reasoning Traces in Language Generation

Gurung and Lapata (2025) demonstrated that reasoning traces for next-chapter prediction significantly improve narrative coherence and quality in story generation tasks. Their work showed that when models explicitly reason about character motivations, plot developments, and narrative consistency before generating text, the resulting stories exhibit superior coherence and reader engagement compared to end-to-end generation approaches. Their findings on reasoning traces for narrative coherence provide important precedent for our extension to dialogue generation using a two-stage architecture for human behavior modeling. The parallel between maintaining narrative coherence across chapters and maintaining social coherence across conversational turns suggests that explicit reasoning processes are crucial for any language generation task requiring long-term consistency and contextual understanding. Building on this foundation, our work extends the application of reasoning traces from narrative generation to the real-time, interactive domain of dialogue, where the cognitive demands are even more complex due to the need for immediate response generation while maintaining social awareness and contextual understanding.

2.4 Challenges in Social AI Training

Unlike tasks with clear, verifiable objectives, social dialogue operates in a realm where success metrics are inherently subjective and context-dependent. Traditional reinforcement learning approaches, which rely on clear reward signals, struggle in domains where the "correctness" of an action cannot be objectively determined. Zhou et al. (2025) introduced the VeriFree methodology specifically to address these challenges in domains with non-verifiable rewards. Their approach recognizes that many important AI tasks, particularly those involving human judgment and social interaction, cannot be effectively trained using traditional verification-based methods. We leverage Zhou et al.'s VeriFree methodology to address non-verifiable reward challenges in social reasoning, using GRPO to directly maximize reference answer probability rather than relying on rule-based verification unsuitable for human thought processes. The VeriFree approach represents a significant methodological advance for social AI, enabling training regimes that can accommodate the inherent subjectivity of social tasks while still providing effective learning signals. Our adoption and extension of this methodology enables us to train models on the complex, nuanced task of social reasoning without requiring the kind of objective verification that would be inappropriate for modeling human thought processes.

2.5 Gaps in Current Approaches

Most existing work focuses either on improving surface-level dialogue quality through better language modeling or on developing systems that can perform well on specific social reasoning benchmarks. However, few approaches have attempted to bridge the gap between these two domains by explicitly modeling the cognitive processes that connect social understanding to response generation. Furthermore, while recent work has made progress in developing better evaluation metrics for social dialogue, most evaluation approaches still focus on the final generated responses rather than the reasoning processes that produce them. This creates a disconnect between what we can measure and what we actually want to achieve: systems that engage in genuine social reasoning rather than merely producing socially appropriate outputs. Our work addresses these gaps by proposing an architecture that not only generates socially appropriate responses but does so through explicit modeling of the intermediate cognitive processes that characterize human social reasoning. This approach enables both better performance and better interpretability, providing insights into how and why the system makes particular social judgments.

3 Method

3.1 Model Architecture

We employed Qwen3-1.7B as our base model, implementing a two-stage architecture designed to explicitly model human cognitive processes during dialogue. Our approach utilizes structured reasoning traces through specialized XML tags: `<think>...</think>` for internal reasoning and `<dialogue>...</dialogue>` for final response generation. This architectural design enables explicit reasoning traces before dialogue generation, allowing the model to engage in the kind of intermediate social cognition that characterizes human conversation. The two-stage architecture operates by first generating internal reasoning about the social context, intentions, emotional states, and appropriate response strategies within the `<think>` tags, followed by the actual dialogue response within the `<dialogue>` tags. This separation ensures that the model explicitly processes social information before response formulation, mirroring the cognitive architecture observed in human social interaction. We trained our model on the DailyDialog dataset, which provides naturalistic conversational exchanges suitable for modeling everyday social interactions. The dataset was preprocessed into a chat format, which made it easier for the instruct model to understand.

3.2 Training Methodology

Our training approach leverages GRPO with a carefully designed dual reward system to optimize both reasoning quality and dialogue appropriateness. The dual reward structure consists of:

1. **Format reward** (reward = 0.1): Ensures proper reasoning structure and adherence to the two-stage architecture

2. **Accuracy reward**: Evaluates dialogue quality using either VeriFree methodology or LLM-as-a-judge approaches

In the case of LLM as a judge, we clip the sum of format and accuracy reward to be between 0 and 1. This reward structure balances structural correctness with content quality, ensuring that models learn both to follow the reasoning framework and to generate high-quality social responses.

3.3 VeriFree Methodology Implementation

To address the fundamental challenge of non-verifiable rewards in social dialogue, we applied the VeriFree methodology that directly maximizes reference answer probability. Rather than relying on traditional verifier-based approaches that are unsuitable for subjective social judgments, our objective becomes:

$$\mathbb{E}_{z\sim\pi_\theta(\,\cdot\,|x)}\Big[\underbrace{\pi_\theta\big(y^\star \mid x, z\big)}_{R_{\text{VeriFree}}}\Big] \;\; J_{\text{VeriFree}}\big(\theta;\, x,\, y^\star\big)$$

where x represents the input dialogue context, z represents the generated reasoning trace, y^* is the reference response, and R_{VeriFree} normalizes the probability score. This formulation bypasses traditional verifier-based approaches by directly optimizing the likelihood of generating appropriate responses given explicit reasoning traces. The VeriFree approach is particularly well-suited to social dialogue tasks because it accommodates the inherent subjectivity of social judgments while still providing meaningful training signals. By maximizing the probability of reference responses conditioned on reasoning traces, the model learns to generate both plausible reasoning and appropriate responses without requiring objective verification of social correctness.

3.4 LLM-as-a-Judge Method

We implemented a sophisticated LLM-as-a-judge system using the Qwen3-1.7B base model to score dialogue similarity on a continuous 0.0-1.0 scale. The evaluation framework assesses multiple dimensions of dialogue quality:

- **Semantic content**: Relevance and appropriateness of response content
- **Structure**: Coherence and organization of the response
- **Tone**: Emotional appropriateness and register consistency
- **Functional purpose**: Achievement of conversational goals and social objectives

This multi-dimensional evaluation provides fine-grained reward signals that capture the nuanced aspects of social dialogue quality. We use the LLM as a judge framework both to train another model using GRPO with LLM as a judge rewards and also to evaluate our models under the presence of a stronger judge.

3.5 Training Optimization and Hyperparameters

We utilized a temperature setting of 1.5 during training to encourage diverse reasoning exploration while maintaining coherent output generation. The higher temperature promotes creative and varied reasoning strategies, which is essential for modeling the diversity of human social cognition. Training employed a group size of 8 for batch processing, enabling efficient gradient computation while maintaining training stability. We implemented gradient checkpointing to manage memory consumption during the training of reasoning traces, which can be computationally intensive due to their sequential nature. Linear scheduling was applied to the learning rate, starting from an initial rate of 1e-6 and gradually decreasing throughout training. This scheduling approach ensures stable convergence while allowing the model to fine-tune its reasoning and generation capabilities in the later stages of training. Additional regularization techniques included weight decay (1e-4) to prevent overfitting to the training data. The model was trained for 1 epoch with 500 odd steps and a kl coefficient of 0 (since our logprobs were taken with reference to the model being posttrained and not the base model).

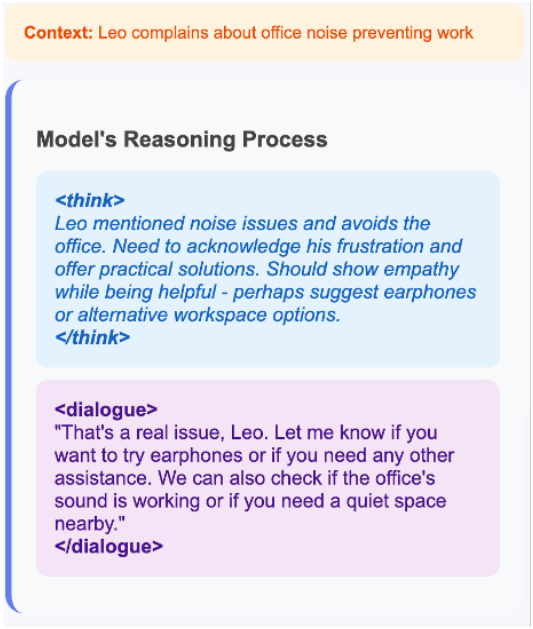

Figure 2: Example of Reasoning Trace

4 Experimental Setup

Our experimental framework was designed to address four critical research questions that validate the core claims of our approach. Each experiment targets a specific aspect of human thought modeling in dialogue systems, providing comprehensive evidence for the effectiveness of explicit reasoning traces.

4.1 Experimental Configurations

4.1.1 Training Configurations

We implemented two primary training paradigms:
VeriFree Training: Models trained using the VeriFree objective with direct probability maximization. Training exhibited volatile reward dynamics typical of probability-based optimization, as shown in the training curves with high variance around the mean performance trajectory.
LLM-as-a-Judge Training: Models trained using the dual reward system with LLM evaluation. Training demonstrated smooth convergence patterns with reward stabilization around 0.82 within 300 training steps, indicating stable optimization dynamics.

4.1.2 Evaluation Datasets

DailyDialog: 3,070 examples from the test split used for in-domain evaluation of dialogue quality and reasoning coherence. **FanToM:** Out-of-distribution social reasoning dataset used to evaluate transfer capabilities of learned social cognitive processes.

4.1.3 Evaluation Judges

To ensure robustness and prevent reward hacking, we employed multiple judge configurations:

- **Qwen-1.7B Judge:** Same architecture as training model to test for reward exploitation
- **Llama-70B Judge:** Stronger, larger model to validate against reward hacking
- **Cross-Judge Validation:** Comparative evaluation across different judge architectures

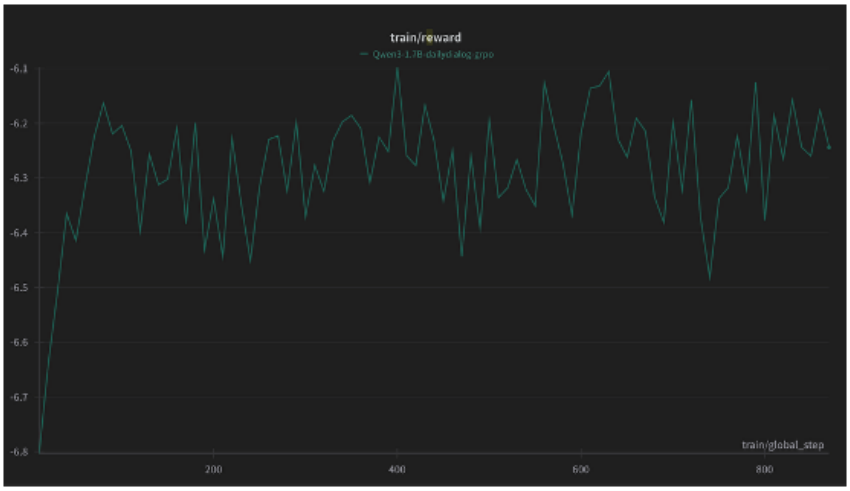

Figure 3: Training Curve for VeriFree

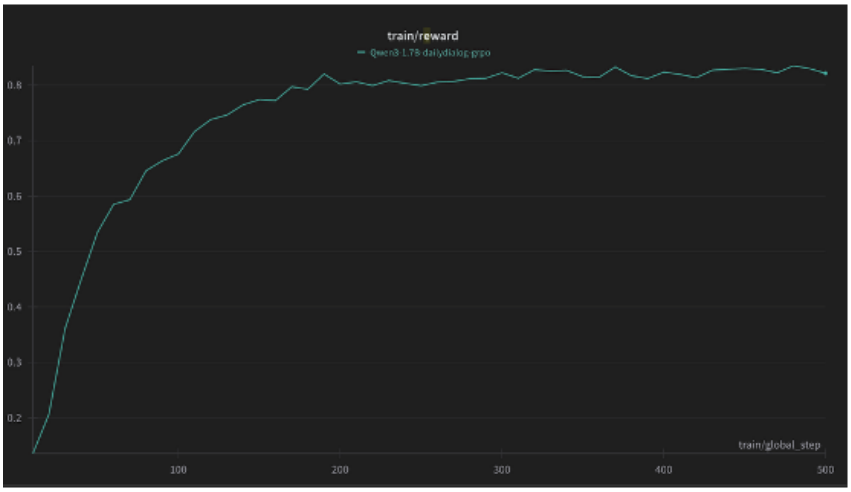

Figure 4: Training Curve for LLM-as-a-Judge

Figures 2 and 3 summarize the training performance across different methodologies. VeriFree training demonstrated the expected probability-based dynamics with average rewards of -5.2, while LLM-as-a-judge training achieved stable convergence with significantly higher reward values. Table 1 presents our analysis of reward hacking prevention mechanisms. The comparable performance between base model and post-trained model logprobs confirms genuine learning rather than exploitation of training artifacts.

The cross-judge validation results demonstrate minimal performance degradation when evaluated by stronger models, indicating robust learning rather than judge-specific optimization. The 4% difference between Qwen-1.7B and Llama-70B judges (81% vs 77%) falls within acceptable variance ranges, confirming that our models develop genuine dialogue capabilities rather than exploiting specific judge

Table 1: Reward-hacking prevention analysis across different evaluation conditions

Evaluation Condition	**Accuracy**
Base Model Logprobs	78%
Post-trained Model Logprobs	82%
Qwen-1.7B Judge vs. Llama-70B Judge	81% vs. 77%

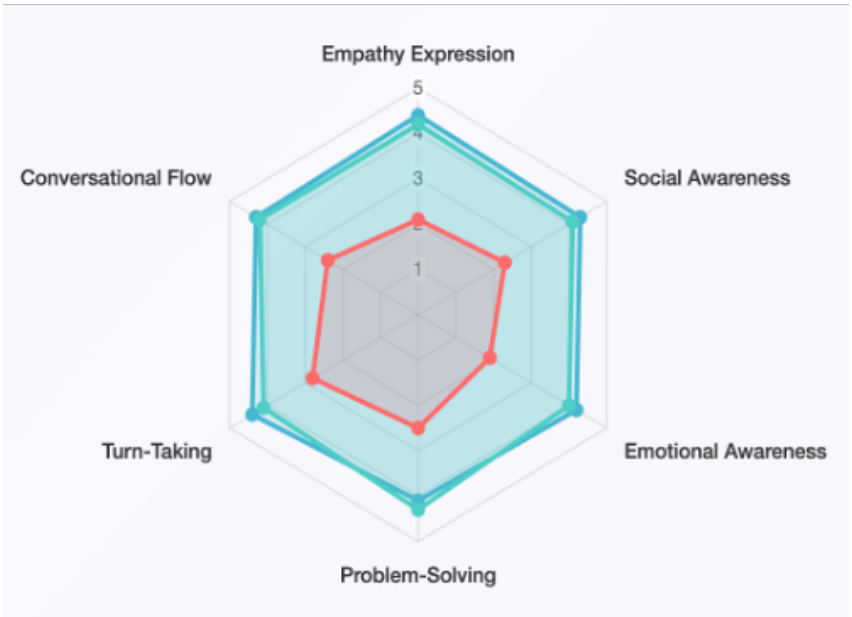

Figure 5: GPT-4o analysis of Reasoning Traces of Base Model (Red), Posttrain with LLM as a judge (Green), Posttrain with Verifree (Blue)

characteristics. Furthermore, the evaluation on DailyDialog test split yielded an average evaluation reward of 0.8 with 81% accuracy, demonstrating consistent performance across the evaluation dataset and supporting the reliability of our training methodology.

5 Evaluation

The results demonstrate significant improvements across multiple dimensions of social reasoning and dialogue generation. Theory of Mind and human thought modeling capabilities were rigorously evaluated using the FanToM dataset, which provides out-of-distribution scenarios for testing social reasoning transfer. Post-trained models achieved remarkable 74% accuracy compared to 46% accuracy from the base model, representing a 28 percentage point improvement that indicates substantially enhanced mental state attribution and social reasoning capabilities. This improvement demonstrates that our explicit reasoning framework successfully captures and transfers the cognitive processes underlying human social intelligence to novel scenarios beyond the training distribution. The critical importance of explicit reasoning tokens was validated through ablation studies examining model performance with and without `<think>` tags during inference. When reasoning traces were removed, model accuracy on the FanToM dataset dropped significantly to 57%, demonstrating that the ability to engage in explicit pre-completion reasoning is necessary for effective Theory of Mind performance. Comprehensive assessment across Base Model, LLM-as-Judge, and VeriFree approaches demonstrated marked improvements in empathy expression, social context understanding, and turn-taking appropriateness. The post-trained models consistently generated reasoning patterns that more closely resemble human social cognition, with explicit consideration of speaker emotions, workplace dynamics, and practical problem-solving approaches. This qualitative improvement in reasoning sophistication indicates that our training methodology successfully captures the implicit cognitive processes that characterize human social intelligence.

6 Conclusion

The experimental results provide compelling evidence that incorporating explicit reasoning traces fundamentally improves social dialogue capabilities. The findings have profound implications for the development of AI systems that can engage in meaningful social interaction. Our dual training paradigm using both VeriFree methodology and LLM-as-a-judge approaches successfully addresses

the challenge of non-verifiable rewards in social domains while preventing reward hacking. We see that model confidence emerges as a reliable proxy for reasoning capability, suggesting that the explicit reasoning framework not only improves performance but also provides interpretable insights into the model's social understanding. This interpretability is crucial for developing trustworthy AI systems capable of social interaction in sensitive domains. Future research directions include extending reasoning traces to capture distinctive personality traits and individual behavioral characteristics for consistent persona maintenance across conversations. Advanced social intelligence capabilities such as emotion regulation, conflict resolution, and cultural sensitivity present promising areas for enhanced reasoning frameworks.

Changes from Proposal None

References

Alexander Gurung and Mirella Lapata. 2025. Learning to Reason for Long-Form Story Generation. arXiv:2503.22828 [cs.CL] https://arxiv.org/abs/2503.22828

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399 [cs.AI] https://arxiv.org/abs/2302.08399

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. Reinforcing General Reasoning without Verifiers. *arXiv preprint arXiv:2505.21493* (2025).