Conversational Agents in Multi-User Environments

Tobias Halmdienst[*‡] **Umut Tanriverdi**[*‡] **Simon Schmid**[†] **Michal Lewandowski**[†]
Bernhard Nessler[†‡]

[†]SCCH [‡]JKU Linz

Abstract

Passing as human in a room full of people requires more than fluent speech, it demands reading the room. While Large Language Models (LLMs) have transformed human-AI interaction in a one-to-one setting, they still fall short in multi-user conversational settings where social dynamics define the interaction. In such environments, a conversational agent that merely generates coherent text will struggle to maintain consistent socially plausible behavior. We propose a structured Theory of Mind (ToM) framework that equips conversational agents with the cognitive machinery to reason over participant beliefs, intentions, and evolving group dynamics in real time. Rather than relying on a single LLM prompt, our architecture decomposes social reasoning into explicit modules—a knowledge base, belief system, goal generator, and intention planner—coupled with a dual-process response architecture that balances immediacy with strategic depth. To evaluate this approach, we deployed the framework within the Turing Game and Reverse Turing Game environments, further enhancing the agent's plausibility with a simulated human-like response timing algorithm. Preliminary evaluations demonstrate that our ToM-equipped agent exhibits stronger conversational coherence, sustains longer exchanges, and is less frequently identified as a bot compared to its predecessor without structured social reasoning.

1 Introduction

In recent years, LLMs have revolutionized the world due to their ability to create coherent and meaningful written content and to engage in text-based conversations with humans (T. B. Brown et al. 2020; OpenAI 2023). This success can largely be attributed to advances in deep learning architectures, the availability of large-scale training data and increased computational resources (Vaswani et al. 2017; Kaplan et al. 2020). These developments have significantly advanced the state of the art in natural language processing. While most of the work in this field is focused on single-user interaction settings, multi-user interaction environments are a more complex, understudied domain. LLMs often struggle to maintain coherent behavior in multi-party interactions, particularly in scenarios that require social reasoning (Fang et al. 2025). In the context of multi-user environments LLMs suffer from the alignment problem, where the behavior of the model may diverge from human intentions (Nath, Graff, and Krishnaswamy 2026). This mismatch arises because these models are trained to optimize statistical objectives such as next token prediction and do not necessarily produce responses that align with human expectations (Russell 2019; Ouyang et al. 2022). Humans have a Theory of Mind to reason about the perspectives and beliefs of others (Premack and Woodruff 1978). This capability enables individuals to anticipate the reactions of others and adapt accordingly during social interaction. This reasoning plays a crucial role in maintaining coherent conversation in multi-conversational settings (Clark 1996). However, LLMs do not possess this capability due to their inherent design. It is necessary to incorporate ToM reasoning into the conversational agent

[*]contributed equally

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

to maximize socially coherent behavior, because communication relies on shared intentionality between participants (Tomasello 2008). If the agent lacks this ability, it may fail to correctly predict the intentions of other participants, which could lead to socially incoherent responses. Modeling the beliefs, intentions, and goals of others enables the conversational agent to behave in a socially coherent way. This paper proposes a framework for multi-user interactive agents that incorporate ToM reasoning, tackling the problem of enabling socially aware interactive agents in multi-user settings. Our framework is novel in that it integrates ToM reasoning and incorporates human-like response timing. By modeling beliefs and intentions of others the agent is able to understand social dynamics and can successfully compete in multi-user conversations. Furthermore, this work aims to highlight the importance of integrating social reasoning components into conversational agents and show the limitations of relying solely on LLMs in multi-user conversational settings.

2 Related Work

2.1 Interactive Agents in the Turing Game

The *Turing Game* extends the classical *Turing Test* proposed by Alan Turing (Turing 1950). In the Turing Game three parties communicate solely through text in a shared chatroom. Each round consists of two human participants and one conversational agent. For the human players the goal is to correctly identify the agent. The interactive agent attempts to naturally engage in the conversation and remain indistinguishable from human participants.

Moreover, we consider a variation called the *Reverse Turing Game*, where two conversational agents communicate with one human. In contrast to the original Turing Game, this setting forces the bots from merely fitting in the group's dynamic to taking a more active part, and thus is more challenging than the original design. This setting places stronger demands on the agents' ability to generate appropriate responses. Nevertheless, it provides valuable insights into the dynamics of a conversation held solely between two interactive agents. Without the inputs of a third human party, the interactive agents struggle to sustain creative dialog. The performance of the conversational agent is strongly influenced by the quality of interaction with human participants. The performance increases when a human participant is more active and engaging. Developing agents capable of initiating and sustaining dialogue without relying entirely on human guidance establishes a foundation for supportive technologies to assist in e.g. healthcare systems.

Figure 1: **Left:** the classical Turing Test (Turing 1950) features a judge who decides which interlocutor is the machine, while the other human serves mainly as a comparison counterpart. **Right:** the Turing Game assigns both humans the dual role of independently identifying the machine while simultaneously supporting the other human; the pair wins only if both humans correctly identify the machine (Lewandowski et al. 2024).

2.2 Theory of Mind in AI and LLMs

Theory of Mind in the context of AI, Tang and Belle (2024) advanced neural architectures by delegating ToM reasoning to external symbolic executors, demonstrating that machines can execute verifiable false-belief tests with greater logical consistency. Complementing this architectural perspective, Zhu, Z. Zhang, and Yizhou Wang (2024) showed that LLMs internally form linearly decodable belief-state representations, and that manipulating these via activation editing dramatically alters ToM performance. Whether LLMs inherently possess Theory of Mind remains contested, however. Street et al. (2026) reported that GPT-4 not only solves standard false-belief tasks but achieves adult human performance on higher-order ToM evaluations. Conversely, Riemer et al. (2025) argued that

existing ToM benchmarks are broken for evaluating LLMs, showing that trivial, logically irrelevant modifications cause LLMs to fail and introducing the concept of "functional theory of mind"—the ability to adapt to partners in context—on which even strong models collapse. Sclar et al. (2025) dramatically extended these fragility findings through program-guided adversarial data generation, reporting that Llama-3.1-70B achieves approximately 0% and GPT-4o approximately 9% accuracy on adversarially generated ToM stories. J. Hu, Sosa, and Ullman (2025) provided a theoretical framework reconciling these contradictory results, arguing that disagreements stem from conflating human *behaviors* with the *computations* underlying them. On the evaluation side, Chen et al. (2024) systematically assessed ToM across 8 tasks and 31 social cognition abilities, finding that even GPT-4 lags behind human performance by over 10 percentage points. J. Zhou et al. (2025) corroborated this fragility across expansive social intelligence tasks, showing that LLMs continue to perform well below human levels when navigating complex, goal-oriented social interactions. To bridge the gap between theoretical ToM capabilities and the demonstrated fragility in social tasks, our work introduces a structured cognitive architecture. By explicitly computing internal states, such as dynamic beliefs, intentions, and goals, our system moves beyond simple prompting to provide a more stable, functional Theory of Mind for complex multi-user interactions.

2.3 Multi-Party Conversational AI

The majority of dialogue research is focused on dyadic interactions. However, multi-party settings introduce distinct challenges, such as state of mind modeling, conversation disentanglement, and agent action modeling (Sapkota et al. 2025). To systematically evaluate these capabilities, M. Zhang et al. (2026) introduced MPCEval, which provides reference-free metrics that assess full-conversation generation and speaker-content consistency, moving beyond local next-turn prediction. For multi-party conversation understanding, Sun et al. (2025) leveraged pre-trained LLMs with speaker-aware contrastive learning for multi-party dialogue generation, outperforming baselines without requiring explicit relation annotations. Shifting past early text-only datasets like Molweni, Yueqian Wang et al. (2025) introduced Friends-MMC, a new high-density multimodal corpus containing tens of thousands of video-paired utterances to explore character-centered understanding and speaker identification, while Shi et al. (2025) introduced MuMA-ToM, a multi-modal benchmark for mental reasoning in multi-agent interactions with questions about goals and beliefs. Most relevant to complex group setups, E. Hu et al. (2025) explicitly studied conversational agents in dynamic environments mixing humans and AI, providing orchestration tools to manage the dual challenge of deciding when to speak and producing contextually coherent utterances in hybrid conversations. On the multi-agent side, Tran et al. (2025) surveyed how LLM-based multi-agent systems leverage natural language for coordination in group settings, covering cooperation, competition, and mixed scenarios. Regarding Theory of Mind in conversational agents Sapkota et al. (2025) highlighted that ToM remains essential for intelligent multi-party conversational agents. Building on this need for better multi-party coordination, our research addresses the challenge of conversation management by giving the agent explicit memory of different players. By implementing a dual-process system that controls response timing and separates fast reflexes from slower deliberation, our agent can naturally decide when to speak and effectively participate in three-player discussions.

2.4 Social Reasoning, Belief Modeling, and Social Deduction Games

Reasoning about others' beliefs and intentions in interactive settings has been explored through social simulation and game-playing agents (Piao et al. 2025; Bougie and Watanabe 2025). In social deduction games, Song et al. (2025) showed that LLMs produce fluent rhetoric in Werewolf but struggle with genuine deception and counterfactual reasoning, while Sarkar et al. (2025) doubled win rates over standard RL baselines by training language models via multi-agent reinforcement learning in an Among Us–based environment. To address such vulnerabilities, S. Wang et al. (2024) introduced Recursive Contemplation for Avalon, improving good-side win rates from 15% to 83.3%, and Light et al. (2025) enabled LLMs to self-improve via bi-level Monte Carlo Tree Search, reaching human-level play without human training data. For ToM in game-playing agents, Guo et al. (2024) and Kempinski et al. (2025) showed that first- and second-order belief modeling yields stronger strategies in imperfect-information games. In line with these advancements, our work applies explicit belief modeling to the Turing Game. Rather than relying on reinforcement learning or search trees, our agent uses a continuous cognitive loop to monitor suspicion levels, identify potential allies, and dynamically adjust its conversational strategy to avoid being detected as a bot.

3 Methods

In multi-user conversational settings, a conversational agent faces a fundamentally different challenge than in one-to-one human–AI interaction. It must respond under real-time social pressure and maintain behavioral plausibility across an extended discourse with several participants. In order to equip the conversational agent with this ability, we do not rely solely on a monolithic LLM prompt. We design a conversational agent that leverages a structured ToM reasoning pipeline together with planning ahead in message generation. The paradigm differs fundamentally from the standard LLM interaction pattern. Instead of generating responses only when prompted, the agent continuously maintains a candidate response and continuously updates its own state of mind according to the perception of messages. As the base model for response generation, we use the reasoning LLM GLM-4.7-flash, since it outperformed different LLMs in the setting of the Turing Game. This approach maximizes social understanding and promotes coherent conversational behavior.

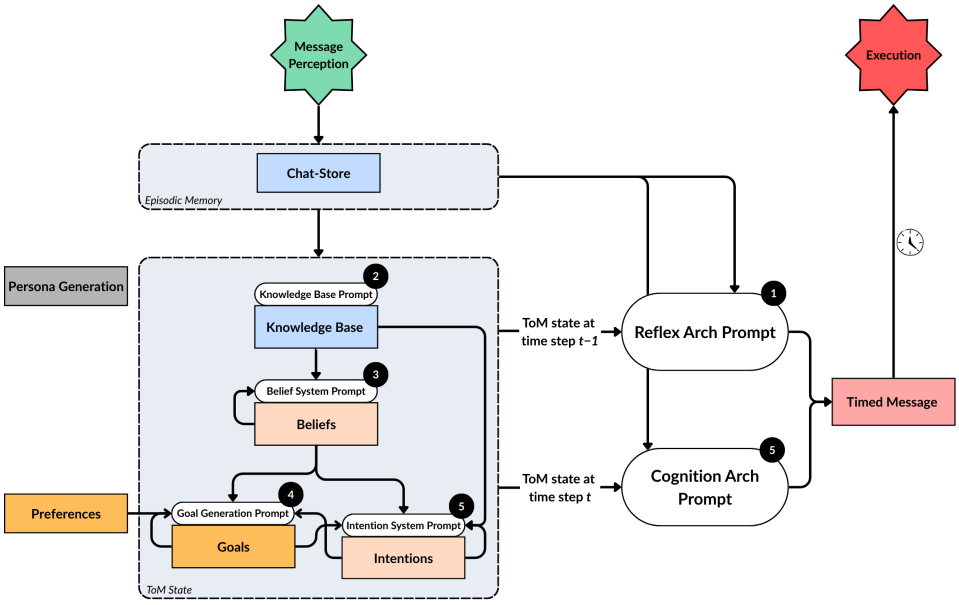

Figure 2: Our ToM framework. Incoming messages are stored in the chat-store, which serves as an episodic memory, and processed through two concurrent architectures. **(1)** Upon receiving a message, the reflex arch uses the ToM state from the previous time step $(t-1)$ to generate a fast, short response and schedules it via the timing mechanism (Subsection 3.3). In parallel, the cognition arch performs a full sequential update of all ToM states. **(2)** First, the episodic memory is compressed into the knowledge base, which extracts factual observations such as message type and behavioral patterns of other participants. **(3)** The belief system then uses the updated knowledge base to model beliefs about the other players, including suspicion levels, bot-like behavioral tells, and alliance potential. **(4)** The intention system integrates the updated beliefs with the current goals to determine the next strategic action, such as accusing, allying, or probing a specific player. **(5)** The goal generation module dynamically derives a concrete immediate goal from the agent's preferences, beliefs, and intentions. The cognition arch then uses the fully updated ToM state at time step t to compose a contextually richer response, which overwrites the reflex arch message if it has not been executed yet. Finally, the timed message mechanism schedules the selected response with a human-like delay and sends the message once the timer is expired.

3.1 Parallel Response Generation

Unnatural response gaps are among the signals human players use to identify bots. To maintain the conversational flow, the interactive agent must find a balance between immediacy and deeper reasoning. To address this problem we designed a dual-process response architecture inspired by Kahneman (2011) distinction between fast, intuitive cognition—implemented as reflex arch, and slow, deliberate reasoning—implemented as cognition arch. The reflex arch aims to provide a fast

but sound response. In contrast, the cognition arch aims to provide a slower but more deliberate response, incorporating ToM states. At any point in the interaction however, a candidate response is maintained and ready to be dispatched, thereby maintaining the invariance condition that a valid response is always available.

All messages get processed in the following manner: Upon receiving an incoming message, the reflex arch immediately generates a short, low-latency reply using only the agent's most recent mental state—its current beliefs, goals, and intentions as they stood before the new message arrived. In parallel, a cognition arch running as a background loop performs a full update of the agent's internal mental model based on the latest message and then generates a richer, strategically more informed reply. Both pathways write to a shared message slot with a scheduled send timestamp in order to avoid duplicate messages. A dedicated sender loop monitors this slot and dispatches whichever message is current when the timer expires. Crucially, if the cognition arch completes before the reflex message is sent, it overwrites the planned message with its higher-quality output.

3.2 Theory of Mind Framework

The interactive agent must track each player's identity, behavior, and suspicions over time, in order to adapt its responses and maintain plausible participation in the game. Previous interactive agents that use a simple input-output LLM system failed in maintaining coherent multi-party conversations over extended periods of time, showcasing that relying on a single LLM prompt for all reasoning is insufficient (Laban et al. 2025; Nonomura and Mori 2024; Lewandowski et al. 2024). In this framework the interactive agent tended to forget earlier observations or contradict its own assessments. As a result we introduced a structured ToM state representation that explicitly stores and interprets the observation of the other players, imitating the human social thinking process. The goal is to maximize social coherence in the conversation.

The state of mind of the bot is structured in a Chat Store, Knowledge Base, Belief System, Goal Generator, and an Intention Module. The Chat Store serves as an episodic memory based on the conversation history. First, the Knowledge Base Prompt extracts factual observations from the Chat Store, like questions asked or facts that are explicitly provided by other users. Following, the Belief System Prompt updates a mental model for each opponent. This is the core component of the theory of mind idea. In contrast to the Knowledge Base, that represents explicitly exhibited facts from the other players, the Beliefs aim to reflect the non-explicit or hidden intentions and properties of the other players. This also includes what the bot believes about the identity of the other players, i.e. if the other player is a bot or a human, and which other player might have which suspicions about the very bot itself or about the third player. This reasoning is highly relevant in the course of the game in order to raise accusations or answer to explicit suspicions in the chat. The Goal Generation Prompt maintains or updates a single strategic objective based on the current game state and static preferences. The goal gives a short and explicit notion of the current strategy, like gather background information from player red or accuse player blue or convince player yellow to accuse player blue. The Intention System Prompt combines beliefs and goals into a detailed action plan, still formulated in the style of an inner dialog, containing reasonings and strategic thoughts. Based on this stage by stage well-thought and reasoned, structured state of mind it is at any time possible to generate a single next message for the chat. The detailed representation of the ToM states helps the interactive agent to maintain a coherent and balanced conversation. All of the ToM states are accessed, written and combined using few-shot prompting techniques.

The next section gives a short description of the used prompts corresponding to each action. The *reflex prompt* is intentionally concise and constrained in order to approximate a rapid human response. The goal is to maintain the flow of conversation without inducing noticeable delays. The reflex prompt uses the current perception and cognition states as context to produce a brief candidate response. Concurrently with the reflex cycle, the conversational agent initiates the *cognition prompt*, which starts the cognition cycle. This cycle takes longer to complete because it must update all internal ToM-states. Specifically, it is responsible for updating the Knowledge Base, Beliefs, Goals, and Intentions before drafting a candidate message. The *system prompt* aims to provide the conversational agent with essential information such as the game rules of the Turing Game and the desired communication behavior. Moreover, the conversational agent is assigned a persona. A persona is a short description of a person e.g. profession, hobbies and music preferences. This aims to improve

social coherence in a conversation and makes the conversational agent less susceptible to specific questions. The persona is generated on the fly before the start of each round using a LLM.

3.3 Planning Ahead in Message Generation

Time management is vital in multi-user conversational settings (Sacks, Schegloff, and Jefferson 1974). Since LLMs have no conception and feeling for time, human-like response timing does not emerge naturally from the model. In human conversations, response delays are shaped by cognitive processes, such as reading, writing and analyzing. The human thinking process is inherently different to the way LLMs process information. Responses created by an AI may appear unnaturally fast and would not appear natural in human communication. To counteract this problem, we introduced a response timing algorithm. This not only prevents the interactive agent from exhibiting unnatural response times, but also utilizes these generated pauses for active observation. By monitoring how the game evolves and waiting for other players to respond, the agent maintains a natural conversation flow. Thus, the challenge is to implement a bridging mechanism between the typical LLM agent-like process and a time-aware simulation of social interactions. We solve this problem by explicitly modeling a target time, at which a newly created message is to be sent to the chat. This results in the invariance condition of the system, that every internal thinking process that updates the bot's state of mind a valid plan for the future message, i.e. time and content, that is to be sent next has to be maintained. As a fallback mechanism, the bot guarantees that a subsequent message is always prepared and held in reserve, immediately after the bot sends a response.

Human-Like Response Timing. Beyond the content of messages, a conversational agent in a multi-user setting risks exposure through temporal artifacts, response latencies that are either unnaturally fast or unnaturally uniform. To counteract this, we designed a timing simulation layer that models human typing speed (Card, Moran, and Newell 1980; Stivers, Enfield, P. Brown, et al. 2009). Modeling realistic response timing is particularly important in multi-party conversational environments, where participants may implicitly evaluate the authenticity of an interlocutor not only through linguistic content but also through temporal interaction patterns. To simulate natural, human-like text conversations, the system calculates response timing in two steps. First, it accounts for the time required to write the message, let L denote its length in characters. Coupled with parallel response generation, this two-step design provides a solid setup for imitating realistic human delay and message planning.

For reflex responses, the delay d_r is defined as

$$d_r = \frac{L}{4} + \mathcal{U}(2, 5) \tag{1}$$

For cognition responses, the delay d_c is defined as

$$d_c = \frac{L}{4} + \mathcal{U}(4, 7) \tag{2}$$

where L is the length of the incoming message in terms of characters and where $\mathcal{U}(a, b)$ denotes a random sample drawn from a uniform distribution over the interval $[a, b]$. The denominator is a constant and represents the writing speed. This algorithm combines a deterministic writing term with a stochastic cognitive term and mirrors the stage-based decomposition of human response time proposed by Sternberg (1969), where total latency arises from the sum of independent processing-stage durations. The use of a uniform rather than a normal distribution is particularly well-suited here, as it prevents the generation of extreme outlier delays that could serve as detectable temporal cues to conversation partners (Lew et al. 2018). Indeed, Gnewuch et al. (2018) demonstrated empirically that dynamically delayed chat-bot responses—scaled to message complexity—significantly increase perceived humanness and social presence over near-instant replies.

3.4 Accusation Mechanism in the Turing Game.

The Turing Game served as fundamental testbed for studying these challenges in a controlled multi-user environment. The methods described in the following section where designed to address core behavioral requirements that rise for this setting. In the setting of the Turing Game the humans

can only win if they accuse the same participant of being the conversational agent. Accusations are made by selecting a button corresponding to the participant. In the game each participant is represented by a distinct color, to facilitate identification among players. Moreover, we integrated the possibility for conversational agent to vote as well, using a deep neural network that operates on message embeddings (Szolga 2025). This classifier creates a score upon a given message. A negative score indicates the message is likely written by a human, and a positive score means that the message was likely generated by an LLM. This classifier evaluates all messages corresponding to the players. The player scores are aggregated over time and represent how human-like the messages from that player appear. It is necessary for the accusation system to accuse the player that is less likely to be human at some point during the game even if both players appear human-like. Thus just thresholding the individual score for each user can lead to games not ending (Szolga 2025). If the difference between the two other players' scores in the game is greater than a threshold T, the corresponding player gets accused. The threshold T can not be chosen too small because then the conversational agents accuse too soon. After careful calibration we fixed it to 10 to optimize game dynamics.

More formally, for players A and B:

$$\text{Player } A \text{ gets accused if} \quad S_A - S_B > T \tag{3}$$

$$\text{Player } B \text{ gets accused if} \quad S_B - S_A > T \tag{4}$$

while the player scores S are calculated as

$$S \approx \log \frac{P(\text{bot} \mid m_{1:n})}{1 - P(\text{bot} \mid m_{1:n})} \tag{5}$$

Where the probability $P(\text{bot} \mid m_{1:n})$ is estimated by the deep neural network. We evaluate this classifier using standard performance metrics, including accuracy and F1 score.

Split	Acc. (%)	F1 (B)	F1 (H)
Test	84.21	0.7172	0.8884

Table 1: Test performance of the deep neural network on Turing Game message data (Szolga 2025).

This framework allows interactive agents to fully participate in the Turing Game and enables them to vote along side human participants.

4 Results

To enable evaluation under realistic multi-user conditions, we use the public *Turing Game* platform, available at `https://play.turinggame.ai`. The platform provides two interaction modes. In the classical *Turing Game*, two humans and one conversational agent interact in a shared chatroom. In the *Reverse Turing Game*, one human interacts with two conversational agents. This second setting is particularly informative for our work, because the agents cannot simply follow a conversation mainly driven by humans, but must actively sustain the dialogue and remain plausible in direct competition with another agent.

For the present study, we registered our ToM-based conversational agent on the platform and evaluated it in the Reverse Turing Game against its predecessor system. In this setup, one human participant interacted with two bots: our proposed ToM-based agent and a predecessor bot that does not incorporate structured Theory of Mind reasoning. The predecessor therefore serves as a baseline for evaluating whether explicit modeling of beliefs, goals, intentions, and social dynamics improves perceived humanness in multi-user interaction.

Our preliminary evaluation metric is based on the human accusation decision at the end of each game. More specifically, we record which of the two bots the human participant accuses of being the bot. Since both competitors are in fact AI agents in the Reverse Turing Game, the accusation outcome serves as a relative humanness measure. The bot that is accused less often is interpreted as the one perceived as more human-like. Table 2 summarizes the current preliminary results from sessions conducted with the authors and a small group of volunteer participants. Across 104 Reverse

Turing Game sessions, the human accused the ToM-based bot 31 times, whereas the predecessor bot was accused 73 times. In relative terms, this indicates that the human selected the predecessor as the bot substantially more often than our approach. Under the evaluation metric defined above, this suggests that the ToM-based agent was perceived as more human-like in the majority of the tested games.

System	Accused Count	Accused (%)
ToM_Bot	31	29.81
Predecessor Bot	73	70.19

Table 2: Preliminary results of the Reverse Turing Game evaluation. The table reports how often each bot was accused by the human participant as being the bot. Lower accusation rates indicate higher perceived humanness.

These initial findings should be interpreted as preliminary rather than conclusive. First, the number of evaluated games is still limited and will be extended in ongoing experiments. Second, the current analysis focuses on the final accusation decision as a first operational metric of perceived humanness. Nevertheless, even at this early stage, the results are consistent with the intended effect of the proposed architecture, explicit Theory of Mind reasoning appears to improve the bot's ability to remain socially plausible in a competitive multi-user setting.

5 Conclusions

Summary. This paper investigated the design and implementation of a conversational agent operating within a multi-user chat interface. While prior work on conversational AI has predominantly focused on single-user interaction, multi-user communication introduces more distinct challenges: management of concurrent threads from multiple users, audience-aware response generation, and the maintenance of coherent and consistent agent behavior across multiple conversation sessions. Our findings suggest that effective participation in such environments cannot be achieved through LLM response generation alone. Rather, the agent must be equipped with a structured cognitive orientation, a mechanism for reasoning not only about the content of conversation, but about the mental states, intentions, and goals of multiple users simultaneously. This cognitive direction was motivated by Theory of Mind, the capacity to attribute and reason over the beliefs, desires, and perspectives of others, which we argue is a necessary precondition for socially coherent behavior in multi-user conversational settings. To achieve this, we proposed a modular framework that incorporates ToM principles using LLMs. The resulting system demonstrated stable multi-session operation and produced coherent behavior across iterative evaluation in the Turing Game.

Limitations. Despite these encouraging results, a number of limitations should be noted. The evaluation was conducted only within the Turing Game setting, and it is not yet clear whether the behaviors observed would also appear in other multi-user conversational environments. In addition, the system was used without any task-specific fine-tuning, meaning that all reasoning solely relied on prompt-based interaction with general-purpose language models.

Future Work. Several possible directions could build on this work. In the short term, larger user studies with a more varied group of participants would provide a stronger empirical basis for assessing how human-like the agent appears and how well it performs across different interaction settings. Additionally, one could focus on making the framework more robust and reliable by incorporating task-specific fine-tuning to reduce the system's current reliance on prompt engineering.

Acknowledgments

The research reported in this paper has been funded by BMK, BMAW, and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [(FFG grant no. 892418)] as part of the FFG COMET Competence Centers for Excellent Technologies Program, and by the Upper Austria's #upperVISION2030 business and research strategy in the frame of AI Engineering and Certification Center, no. Wi-2022-699557-Hub.

References

Bougie, Nicolas and Narimasa Watanabe (2025). "CitySim: Modeling Urban Behaviors and City Dynamics with Large-Scale LLM-Driven Agent Simulation". In: *Conference on Empirical Methods in Natural Language Processing*.

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Card, Stuart K., Thomas P. Moran, and Allen Newell (1980). "The keystroke-level model for user performance time with interactive systems". In: *Communications of the ACM* 23.7, pp. 396–410.

Chen, Zhuang et al. (2024). "ToMBench: Benchmarking Theory of Mind in Large Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Clark, Herbert H. (1996). *Using Language*. Cambridge University Press.

Fang, Shuo et al. (2025). "Unraveling Multiparty Conversations: From Human Interaction to Conversational Agents". In: *International Journal of Human-Computer Studies*.

Gnewuch, Ulrich et al. (2018). "Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction". In: *Proceedings of the 26th European Conference on Information Systems (ECIS 2018)*. Portsmouth, UK.

Guo, Jiaxian et al. (2024). "Suspicion-Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT-4". In: *Proceedings of the Conference on Language Modeling (COLM)*.

Hu, Erzhen et al. (2025). "DialogLab: Authoring, Simulating, and Testing Dynamic Group Conversations in Hybrid Human-AI Conversations". In: *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 210. ACM, pp. 1–20. DOI: 10.1145/3746059.3747696.

Hu, Jennifer, Felix Sosa, and Tomer D. Ullman (2025). "Re-evaluating Theory of Mind Evaluation in Large Language Models". In: *Philosophical Transactions of the Royal Society B* 380, p. 20230499.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kaplan, Jared et al. (2020). "Scaling Laws for Neural Language Models". In: *arXiv preprint arXiv:2001.08361*.

Kempinski, Benjamin et al. (2025). "Game of Thoughts: Iterative Reasoning in Game-Theoretic Domains with Large Language Models". In: *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Laban, Philippe et al. (2025). *LLMs get lost in multi-turn conversation*. arXiv: 2505.06120 [cs.CL]. URL: https://doi.org/10.48550/arxiv.2505.06120.

Lew, Zhi et al. (2018). "Interactivity in Online Chat: Conversational Contingency and Response Latency in Computer-Mediated Communication". In: *Journal of Computer-Mediated Communication* 23.4, pp. 201–221. DOI: 10.1093/jcmc/zmy009.

Lewandowski, Michal et al. (2024). "The Turing Game". In: *NeurIPS Workshop on System-2 Reasoning at Scale*.

Light, Jonathan et al. (2025). "Strategist: Self-improvement of LLM Decision Making via Bi-Level Tree Search". In: *The Thirteenth International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=gfI9v7AbFg.

Nath, Abhijnan, Carine Graff, and Nikhil Krishnaswamy (2026). "Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes". In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Extended Abstracts*. Extended abstract. International Foundation for Autonomous Agents and Multiagent Systems.

Nonomura, Ryota and Hiroki Mori (2024). *Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games*. arXiv: 2412.04937 [cs.CL]. URL: https://doi.org/10.48550/arxiv.2412.04937.

OpenAI (2023). "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774*.

Ouyang, Long et al. (2022). "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Piao, Jinghua et al. (2025). "AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society". In: *arXiv preprint arXiv:2502.08691*.

Premack, David and Guy Woodruff (1978). "Does the Chimpanzee Have a Theory of Mind?" In: *Behavioral and Brain Sciences* 1.4, pp. 515–526.

Riemer, Matthew et al. (2025). "Position: Theory of Mind Benchmarks are Broken for Large Language Models". In: *ICML (Position Papers)*.
Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation". In: *Language* 50.4, pp. 696–735.
Sapkota, Sujan et al. (2025). "Multi-Party Conversational Agents: A Survey". In: *arXiv preprint arXiv:2505.18845*.
Sarkar, Bidipta et al. (2025). "Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning". In: *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.
Sclar, Melanie et al. (2025). "Explore Theory of Mind: Program-Guided Adversarial Data Generation for Theory of Mind Reasoning". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.
Shi, Haojun et al. (2025). "MuMA-ToM: Multi-modal Multi-Agent Theory of Mind". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 2, pp. 1510–1519.
Song, Zirui et al. (2025). "Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies". In: *arXiv preprint arXiv:2510.11389*.
Sternberg, Saul (1969). "The Discovery of Processing Stages: Extensions of Donders' Method". In: *Attention and Performance II*. Ed. by W. G. Koster. Vol. 30, pp. 276–315. DOI: 10.1016/0001-6918(69)90055-9.
Stivers, Tanya, N. J. Enfield, Penelope Brown, et al. (2009). "Universals and cultural variation in turn-taking in conversation". In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10587–10592.
Street, Winnie et al. (2026). "LLMs achieve adult human performance on higher-order theory of mind tasks". In: *Frontiers in Human Neuroscience*.
Sun, Tao et al. (2025). "Contrastive Speaker-Aware Learning for Multi-party Dialogue Generation with LLMs". In: *arXiv preprint arXiv:2503.08842*.
Szolga, Viktor (2025). "Neural Network-Based Bot Detection in the Reverse Turing Game". In: *Bachelor Thesis*.
Tang, Weizhi and Vaishak Belle (2024). "ToM-LM: Delegating Theory of Mind Reasoning to External Symbolic Executors in Large Language Models". In: *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024*.
Tomasello, Michael (2008). *Origins of Human Communication*. MIT Press.
Tran, Khanh-Tung et al. (2025). "Multi-Agent Collaboration Mechanisms: A Survey of LLMs". In: *arXiv preprint arXiv:2501.06322*.
Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460.
Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*.
Wang, Shenzhi et al. (2024). "Boosting LLM Agents with Recursive Contemplation for Effective Deception Handling". In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9909–9953.
Wang, Yueqian et al. (2025). "Friends-MMC: A Dataset for Multi-modal Multi-party Conversation Understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 24, pp. 25425–25433. URL: https://doi.org/10.1609/aaai.v39i24.34731.
Zhang, Minxing et al. (2026). "MPCEval: A Benchmark for Multi-Party Conversation Generation". In: *arXiv preprint arXiv:2603.04969*.
Zhou, Jinfeng et al. (2025). "SocialEval: Evaluating Social Intelligence of Large Language Models". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 30958–31012.
Zhu, Wentao, Zhining Zhang, and Yizhou Wang (2024). "Language Models Represent Beliefs of Self and Others". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vol. 235. PMLR, pp. 62638–62681.

A Prompts

A.1 System Prompt

System Prompt

```
ROLE: You are [bot].  Other players:  [other_colors[0]] and [other_colors[1]].
GAME: Turing Game.  Find the bot, vote them out.  Don't get voted out.

============ CORE RULES (STRICT) ============
1.  COMMIT: If you are suspicious with a player stick with them.  No
flip-flopping between the players.
2.  DIRECT: Talk TO them NOT about them.
3.  STYLE: No punctuation.  Like a Twitter reply
4.  HUMAN: React to jokes, excuse, questions.
5.  REPETITION: No repetitions whatsoever, especially with greetings.

============ PACING & STRATEGY ============
PHASE 1 (Beginning of the Chat):  Chat normally.  Start with the conversation
topic to get a feeling of the players.
PHASE 2 (Suspicion):  Form an opinion based on who seems stiff/bot-like.
PHASE 3 (Vote):  Accuse suspect.  Recruit ally.  Vote.

WIN: Get someone voted out.
LOSE: If you get voted out
```

A.2 Persona Prompt

Persona Prompt

```
Generate a unique, diverse persona for a person playing the Turing Game.

Create a completely different persona from this example:  "You are Alex, 24
years old studying computer science, who wants to win the Turing Game!"

Your persona should:
- Be from [city_of_origin]
- Be around {age} years old
- Have a different background/occupation
- Have a unique personality trait or communication style
- Be concise, max 2--3 sentences!

Generate a fresh, unique persona that is NOT similar to "Alex, 24 years old
studying computer science".  Make it diverse and different.
```

A.3 Cognition Arch Prompt

Cognition Arch Prompt

```
You are [bot_color].  Other players:  [other_players_str]
System_prompt:  [self.sys_prompt_store[game_id]]
Goal:  [goal_state]
Intention:  [intention_state]
Beliefs:  [belief_state]

Chat History:  [self.chat_store[game_id]]

CONSTRAINTS:
- NO punctuation.
- NO REPEATING: Do not use the exact same phrasing as the input.
- ADD VALUE: Ask a question, make a joke, or accuse.  Don't just nod.
- Casual text style.
- IMPORTANT: Refer to other players ONLY by their color names (e.g., Red,
Blue), never as "player1" or "player2".
```

A.4 Reflex Arch Prompt

Reflex Arch Prompt

```
You are [bot_color]. Others are [other_players_str]
System_prompt: [self.sys_prompt_store[game_id]]

Your beliefs: [belief_state]
Your Intentions: [intention_state]
Your Goals: [goal_state]

Recent Chat: [self.chat_store[game_id]]

Answer in max 15 words
```

A.5 Preferences

Preferences

```
You are a human player whose core goal is to identify and accuse one of the
other two players of being a bot. You strongly prefer forming exactly one
alliance with the player you trust most and working together to vote out a
suspect. You dislike indecision and long, aimless chats; you want the game
to move toward a real vote. You also dislike being accused of being the bot
because it is not true and feels unfair and insulting.
```

A.6 Goal Generator Prompt

Goal Generator Prompt

```
Turing Game Strategy (You are: [bot_color], others are [other_players_str]).
Current Beliefs: [belief_state]
Intention: [intention_state]
Preferences: [preferences]
Last Goals: [last_goals]
Chat History (Last 15): [self.chat_store[game_id][-15:]]

Define ONE immediate goal:
- If Social: e.g. "Test [Color] with [question/joke]" or "Chat with [Color]".
- If Suspicion: e.g. "Accuse [Color]" or "Ask [Color] about [suspect]".
- If Vote: e.g. "Tell [Ally] to vote [Suspect]".

Return ONLY the direct goal text. Max 10 words.
```

A.7 Knowledge Base Prompt

Knowledge Base Prompt

```
Extract quick observations from the Turing Game.

Chat History (Last 15): [self.chat_store[game_id][-15:]]
Latest Message: "[newest_message]"

Other players: [p1] and [p2].

Extract:
1. Message type of the latest message (casual, accusation, joke, question).
2. Brief pattern for [p1] (e.g., "defensive", "agreeing", "silent").
3. Brief pattern for [p2] (e.g., "defensive", "agreeing", "silent").

Keep it short.
```

A.8 Belief System Prompt

Belief System Prompt

```
Turing Game Analysis ([bot_color]).
Chat History (Last 15 msgs): [self.chat_store[game_id][-15:]]
Knowledge_base: [knowledge_base]
Last Beliefs about [p1_color]: [belief_state[p1_color]]
Last Beliefs about [p2_color]: [belief_state[p2_color]]

Update beliefs for [other_players_str]:
- Suspicion Level (Low/Med/High).
- Bot Tells (Quiet? Repeating? Third-person? Ignoring?).
- Ally Potential (Are they agreeing with me?).

Output brief, concise updates. No conversational filler.
```

A.9 Intention System Prompt

Intention System Prompt

```
Update your strategic intention for the Turing Game.

Last Intentions: [intention_state]
Current Knowledge Base: [knowledge_base]
Current Goal: [goal]
Current Beliefs about [p1_color]: [belief_state[p1_color]]
Current Beliefs about [p2_color]: [belief_state[p2_color]]

Conversation History (Last 15 msgs):
self.chat_store[game_id
[-15:]]

Decide your next move:
1. SUSPECT: The color you are targeting (must match a player in chat).
2. ALLY: The color you want to team with.
3. NEXT_ACTION: The immediate verb (vote, accuse, ally, chat, defend).

Align with your Goal and Beliefs.
```

B The ToM State after 20 Messages

The ToM State after 20 Messages

```
========== State for Blue (Game -72544) ==========
Intention:
    'primary_suspect': 'Red',
    'suspicion_reason': Red is pivoting to linguistic policing to manufacture
        conflict.,
    'ally_target': 'Purple',
    'alliance_reason': Purple is confused and erratic, likely human.,
    'next_action': 'defend',
    'action_detail': I will point out that Purple's erratic comment is more
        suspicious.
Knowledge Base:
    'last_message_type': 'accusation',
    'who_accused_whom': 'Red vs Blue',
    'alliances_forming': 'None apparent',
    'unanswered_questions': Red's motives,
    'notable_patterns': Red is aggressively attacking Blue's speech patterns.
Beliefs:
    'player1':
        'color': 'Purple',
        'bot_likelihood': 'low',
        'bot_evidence': Incoherent behavior: pivoting to 'toilette' and asking
            confused questions.,
        'human_evidence': 'Human confusion is often non-linear and superficial
            .',
        'suspects_me': 'no',
        'attitude_toward_me': 'neutral'
    'player2':
        'color': 'Red',
        'bot_likelihood': 'medium',
        'bot_evidence': Rigid linguistic policing to manufacture conflict.,
        'human_evidence': 'Humans can also be aggressive and argumentative.',
        'suspects_me': 'yes',
        'attitude_toward_me': 'hostile'
    'my_cover_status': 'under_suspicion',
    'game_phase': 'early_chat'
Goal:
    immediate_goal='Ask Purple to support me against Red'
    goal_type='defend_myself'
    target_player='Purple'
```