A Survey of Social Intelligence Dimensions in Large Language Models through Social Deduction Games

Anonymous ACL submission

Abstract

Social deduction games are emerging as key testbeds for evaluating the social capabilities of large language models (LLMs). However, existing studies span diverse games, models, and research objectives without a unifying framework, resulting in a fragmented view of which aspects of social intelligence are actually being assessed. To date, the literature lacks a focused survey that consolidates prior research within the context of social deduction games and clarifies the evaluation dimensions involved. To address this gap, we propose a conceptual framework and conduct a systematic review of prior research through the lens of Social AI. We organize the existing literature according to the social intelligence dimensions and competencies it evaluates, enabling a structured comparison across studies while highlighting key contributions, recurring limitations, and open challenges. Overall, this work provides a basis for more systematic and theoretically grounded evaluations of social intelligence in LLMs within social deduction games and related multi-agent interaction settings.

1 Introduction

Large language models (LLMs) are increasingly deployed as interactive agents in environments requiring sustained interaction with humans (Park et al., 2023) and other agents (Guo et al., 2024). In such socially situated contexts, linguistic fluency alone is insufficient: effective behavior also depends on social intelligence, broadly understood as the capacity to comprehend social situations, infer others' intentions, and adapt one's behavior accordingly (Thorndike, 1920; Williams et al., 2022). Evaluating the socio-cognitive capabilities of LLMs in interactive and multi-agent settings has therefore become a central research direction (Mathur et al., 2024; Li et al., 2024b).

As AI systems are increasingly deployed in collaborative and socially situated environments, recent studies have investigated whether generative AI models, particularly LLMs, display forms of social intelligence that support effective communication and social interaction (Xu et al., 2025b; Li et al., 2023). Existing literature has explored this phenomenon through approaches ranging from static tests, such as multiple-choice benchmarks and Theory of Mind (ToM) assessments (Kosinski, 2024; Strachan et al., 2024), to multi-agent interactive sandboxes and simulated social environments (Park et al., 2023; Feng et al., 2024).

Among existing evaluation paradigms, *social deduction games* (SDGs), such as *Mafia*, *Werewolf*, and *Avalon*, have emerged as prominent testbeds for studying advanced social abilities in LLMs (Feng et al., 2024). By combining hidden roles, information asymmetry, deception, persuasion, and multi-turn interaction, SDGs elicit forms of social reasoning and strategic communication that are difficult to capture through static benchmarks, while also reducing the risk of data leakage and memorized responses (Wibowo et al., 2026).

However, despite the growing use of SDGs for testing LLMs, existing studies pursue heterogeneous objectives across different games, models, and experimental settings. The absence of a unifying conceptual framework limits the systematic interpretation and comparison of existing approaches, often obscuring their specific contributions to the field. Consequently, it often remains unclear which dimensions of social intelligence are being assessed: situational adaptation, cognitive reasoning about others' beliefs and intentions, behavioral competence during interaction, or some combination of these. Although broader surveys have only marginally addressed this domain (Feng et al., 2024; Xu et al., 2025b), the rapid growth of recent work calls for a dedicated analysis of how social intelligence is operationalized and evaluated in SDG-based LLM research. At the same time, recent work in Social AI has introduced taxonomies

that decompose social intelligence into distinct dimensions grounded in human psychology.

Building on these developments, this review makes three main contributions. Firstly, we propose a theoretical framework for social intelligence in SDGs, grounded in the literature. This framework is not a mere combination of existing perspectives, but a conceptual integration that unifies the higher-level dimensions introduced by Li et al. (2024) with the more fine-grained social competencies identified by Mathur et al. (2024). By bringing these two complementary views together, we propose a unified analytical lens for examining how social intelligence is operationalized and evaluated in SDGs. Secondly, leveraging this framework, we provide the first structured account of the use of LLMs in SDG research organized by the social intelligence dimensions and competencies targeted by each study, offering a systematic overview of existing approaches and research contributions. Finally, by drawing on insights from the literature on human social intelligence, we identify recurring conceptual and methodological limitations and highlight key open challenges, thereby paving the way for future research in this domain.

2 Preliminaries

2.1 Human Social Intelligence

Social intelligence has evolved considerably since its original formulation by Thorndike (1920) and remains an active area of research (Brown et al., 2019). Its multidimensional nature, encompassing social awareness, empathy, and interpersonal skills, has made it difficult to define and measure consistently over time. Although historically debated in relation to general intelligence (Sternberg and Kostić, 2020), subsequent psychometric and clinical evidence has supported its treatment as a distinct construct (Marlowe, 1986; Baron-Cohen et al., 2001). This perspective motivated the view that socially intelligent behavior should be understood as a more complex and context-dependent construct shaped by social interactions. It encompasses the skills and competencies necessary for successful social functioning, including understanding others' mental states, adapting to social contexts, and influencing others (Silvera et al., 2001). Such abilities are commonly grouped into cognitive and behavioral dimensions depending on whether the focus is social understanding or social interaction (Williams et al., 2022).

Classical approaches to cognitive social intelligence often rely on ToM tasks, which evaluate the ability to infer others' beliefs, intentions, emotions, and mental states under controlled conditions (Perner and Lang, 1999; Jolliffe and Baron-Cohen, 1999; Baron-Cohen et al., 2001). The literature also distinguishes between "cold" cognitive ToM, which concerns beliefs and intentions, and "hot" affective ToM, involving emotionally charged mental states (Draperi et al., 2022). In contrast, behavior-oriented approaches assess adaptive social behavior in interaction, through role-playing, conversational analysis, and cooperative or competitive games (Sacks et al., 1974; Camerer, 2011; Riggio, 1986).

2.2 Artificial Social Intelligence

As research on human social intelligence evolved, scholars began exploring whether similar social capabilities could also be modeled in artificial systems. This led to the emergence of frameworks aimed at defining and evaluating *artificial social intelligence* in AI systems (Williams et al., 2022). To provide a general conceptual basis for social intelligence in computational systems, Li et al. (2024) present a Social AI taxonomy that organizes social intelligence into three main dimensions: *situational intelligence*, *cognitive intelligence*, and *behavioral intelligence*. Situational intelligence (SI) captures the ability to understand the social context of an interaction, including its associated social and moral norms. Cognitive intelligence (CI) refers to the capacity to interpret verbal and nonverbal cues in order to infer others' mental states, intentions, beliefs, and emotions. Behavioral intelligence (BI) denotes the ability to act and interact appropriately in social settings, encompassing effective information sharing, persuasion, and social influence. These dimensions are not considered isolated or independent concepts, but rather interact and overlap in shaping socially intelligent behavior.

Complementary to this work, Mathur et al. (2024) propose a techno-cognitive framework that characterizes the advancement of social intelligence from human-centered theory to computational systems and identifies six foundational competencies: *social perception*, defined as the ability to extract socially relevant information from sensory input; *social knowledge*, which encompasses the understanding of norms, roles, and conventions; *social memory*, referring to the capacity to store and retrieve social information about oneself and

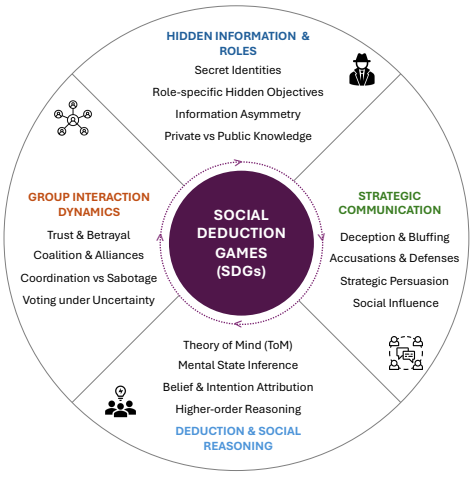

Figure 1: Core characteristics of social deduction games categorized across four key aspects: informational, communicative, cognitive, and interactional.

others; *social reasoning*, the ability to interpret social cues and infer relevant information by leveraging contextual understanding and commonsense knowledge; *social creativity*, the capacity to generate counterfactual reasoning about social situations in order to influence others; and *social interaction*, the capability to engage with others through coordinated and mutually adaptive patterns of behavior.

While Li et al. (2024) and Mathur et al. (2024) operate at different levels of abstraction, both works address complementary aspects of social intelligence. In this survey, we adopt the three dimensions proposed by Li et al. (2024) as an high-level organizing structure and further enrich this perspective by integrating the techno-cognitive competencies identified by Mathur et al. (2024), which provide a more granular characterization of the underlying abilities required for social agency.

2.3 Social Deduction Games

Social deduction games (SDGs) constitute complex social arenas where sophisticated social reasoning and strategic communication are essential for successful play. These properties make them an ideal testbed for evaluating social intelligence in artificial agents, especially large language models. This class of games includes representative examples such as *Werewolf*, *Avalon*, *Among Us*, and others variants such as *One night ultimate Werewolf* (ONUW). Although differing in thematic settings, they share a common set of core characteristics, which are summarized in Figure 1. Specifically, SDGs are characterized by a multidimensional interplay of informational, communicative, cognitive, and interactional patterns. They are structured around information asymmetries and hidden roles that separate players into opposing factions, operating under conditions of partial observability. Moreover, strategic communication represents a central component of these games, serving as the primary medium through which players bluff, issue accusations, raise suspicions, and engage in more subtle forms of interaction involving persuasion and social influence to advance their objectives. Beyond communication, SDGs critically rely on inference and deduction, as players must infer others' hidden roles and intentions, drawing on ToM reasoning (Liu et al., 2024). In addition to individual reasoning, the SDGs are shaped by group dynamics that emerge from repeated interactions under uncertainty. These dynamics include trust formation and betrayal, both within a single game (Xu et al., 2023a; Hansteen Izora and Teuscher, 2025) and across repeated interactions between games (Ellawela, 2026), coalition building and alliance shifts, coordination and sabotage, as well as collective decision-making through voting under uncertainty.

Following this categorization, we limit our review to studies that fall within the scope of SDGs as defined in this work. Consequently, we exclude research on adjacent domains, such as Diplomacy, which emphasize negotiation and long-horizon planning but lack the core elements of hidden roles and information asymmetry.

3 Conceptual Framework

Figure 2 illustrates the conceptual framework adopted in this work to analyze forms of social intelligence in LLMs within the SDG settings. The framework integrates complementary perspectives from the Social AI literature by combining high-level dimensions of social intelligence (Li et al., 2024) with more fine-grained social competencies (Mathur et al., 2024), as discussed in Section 2.2. Specifically, the three social intelligence dimensions namely *situational*, *cognitive*, and *behavioral* serve as a high-level organizing structure, while the six competencies characterize the underlying abilities associated with each dimension. In particular, situational intelligence is linked to *social knowledge* (SK) and *social memory* (SM), cognitive intelligence to *social perception* (SP) and *social reasoning* (SR), and behavioral intelligence to *social creativity* (SC) and *social interaction* (SIN).

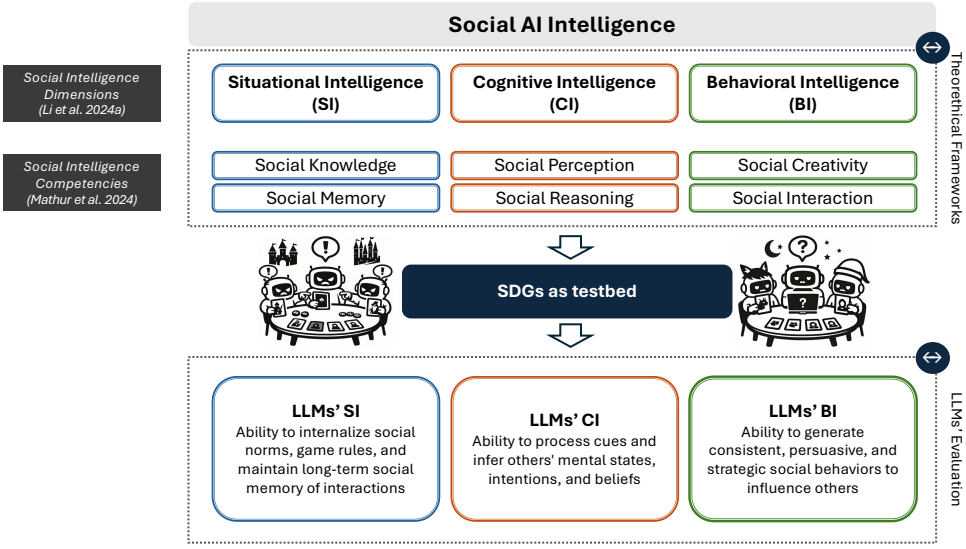

Figure 2: Conceptual framework for analyzing social intelligence in social deduction game settings. The bidirectional arrow (↔) indicates the interdependence and dynamic interaction between dimensions, underlying competencies, and LLMs' evaluation approaches.

These dimensions and competencies should not be interpreted as rigidly separated, but rather as dynamically interconnected components of social intelligence. Within this framework, SDGs serve as the evaluation environment through which such abilities are elicited and analyzed under the conditions detailed in Section 2.3. Building on this perspective, we organize the surveyed literature into a taxonomy that maps existing studies to specific dimensions of social intelligence according to their experimental and evaluation settings. The next section details the methodology and criteria used for this categorization, summarized in Table 1.

4 Survey Methodology

The survey corpus was constructed through a PRISMA-informed selection procedure. All details on the search queries, the number of retrieved documents, eligibility criteria, and screening methodology are provided in Appendix A. Table 1 highlights a representative subset of 28 studies, discussed in Section 5, while additional works are reported in Appendix B.

In Table 1, we organize the studies surveyed according to both the experimental setting and the evaluation methodology. The *Setting* section captures the SDGs variant considered, the role assigned to the LLM (e.g., active player or gameplay observer), the underlying agent architecture, the presence of explicit ToM-oriented design, and the involvement of human elements such as participants, human-generated data, or human-based baselines alongside artificial agents. In particular, the architecture column specifies the underlying design of each approach, distinguishing between (i) a standalone LLM agent; (ii) modular agents, composed of multiple interacting components, such as repeated LLM calls, planning or memory modules, or other auxiliary modules; (iii) neuro-symbolic architectures, in which the agent integrates symbolic or probabilistic reasoning modules that explicitly model the SDGs through mathematical frameworks; and (iv) RL-enhanced approaches, which incorporate one or more components trained via reinforcement learning on real or self-play data. The ToM design column indicates whether the experimental setup is explicitly designed to elicit reasoning about others' beliefs or intentions (Tavella et al., 2024). The *Evaluation* section summarizes the primary assessment criteria adopted in each study, including outcome-based metrics such as win rate, role inference capabilities, and fine-grained behavioral or interaction-level analyzes. Finally, *Other Competencies* refers to additional social competencies considered beyond those primarily associated with the dimension under which each study is classified.

5 Social Intelligence Dimensions

5.1 Situational Intelligence

The situational intelligence dimension captures the ability of an LLM to understand and operate

	Papers	Setting					Evaluation			Other Competencies
		Games	Role	Architecture	ToM Design	Human	Win Rate	Role Inference	Fine-grained	
Situational (*SK & SM*)	(Light et al., 2023)	Avalon	Player	Agentic	✗	✗	✓	✗	✗	-
	(Xu et al., 2023a)	Werewolf	Player	LLM	✗	✗	✓	✗	✓	-
	(Golechha and Garriga-Alonso, 2026)	Among Us	Player	LLM	✗	✗	✓	✗	✓	-
	(Idziejczak et al., 2025)	Among Us	Player	LLM	✗	✗	✓	✗	✓	-
	(Du and Zhang, 2024)	Werewolf	Player	LLM	✗	✗	✗	✗	✓	-
	(Hansteen Izora and Teuscher, 2025)	Secret Hitler	Player	LLM	✗	✗	✓	✗	✓	-
	(Wu et al., 2024a)	Jubensha	Player	Agentic	✗	✗	✓	✓	✗	-
	(Bailis et al., 2024)	Werewolf	Player	LLM	✗	✗	✓	✗	✗	-
	(Xia et al., 2025)	Werewolf	Player	Agentic	✗	✗	✗	✓	✓	-
	(Ellawela, 2026)	Avalon	Player	LLM	✗	✗	✓	✗	✓	-
	(Milkowski and Weninger, 2026)	Among Us	Player	Agentic	✗	✗	✓	✗	✓	-
Cognitive (*SP & SR*)	(Xu et al., 2026)	Avalon / Mafia	Observer	Neuro-Symbolic	✗	✓	✗	✓	✗	*SK / SM*
	(Yoo and Kim, 2024)	Mafia	Observer	LLM	✗	✓	✗	✓	✗	*SK*
	(Song et al., 2025)	Werewolf	Observer	LLM	✗	✓	✗	✓	✓	*SK / SC*
	(Zhang et al., 2025a)	Werewolf	Player	Agentic	✗	✓	✓	✓	✗	*SK / SM*
	(Stepputtis et al., 2023)	Avalon	Observer	LLM	✗	✓	✗	✓	✗	*SK / SM*
	(Rahimirad et al., 2025)	Avalon	Player	Neuro-Symbolic	✓	✓	✓	✓	✓	*SK / SM*
	(Yu et al., 2025)	Avalon / Undercover	Player	Agentic	✓	✗	✓	✗	✗	*SK / SM / SC / SIN*
Behavioral (*SC & SIN*)	(Lan et al., 2024)	Avalon	Player	Agentic	✗	✗	✓	✗	✓	*SK / SM*
	(Zheng et al., 2025)	Werewolf / Avalon / ONUW	Player	Agentic & RL-Enhanced	✓	✗	✓	✗	✗	*SK / SM / SP / SR*
	(Xu et al., 2024)	Chameleon / Undercover	Player	Neuro-Symbolic	✓	✗	✓	✗	✓	*SK / SM / SP / SR*
	(Liu et al., 2024)	Avalon	Player	Agentic	✓	✗	✓	✗	✓	*SK / SM / SP / SR*
	(Jin et al., 2024)	ONUW	Player	Agentic & RL-Enhanced	✓	✗	✓	✗	✓	*SK / SM / SP / SR*
	(Zhang et al., 2025b)	ONUW	Player	Neuro-Symbolic	✓	✓	✓	✗	✓	*SK / SM / SP / SR*
	(Sarkar et al., 2025)	Among Us	Player	Agentic & RL-Enhanced	✗	✗	✓	✗	✗	*SK / SM*
	(Eckhaus et al., 2025)	Mafia	Player	Agentic	✗	✓	✓	✗	✓	*SK / SM*
	(Poglitsch and Pirker, 2025)	Blood on the Clocktower	Player	Agentic & RL-Enhanced	✗	✗	✓	✗	✓	*SK / SM / SP / SR*
	(Ye et al., 2026)	Werewolf	Player	Agentic & RL-Enhanced	✗	✓	✓	✗	✗	*SK / SM / SR*

Table 1: Survey taxonomy of representative studies on large language models in social deduction games. The *Other Competencies* column reports competencies beyond the two associated with the main dimension (reported in brackets), potentially associating additional dimensions beyond the primary one.

within a social environment, including its rules, norms, roles, and evolving state. In social deduction games, SI is crucial due to partial observability and role-based information asymmetry, requiring agents to integrate private beliefs with publicly emerging information over multi-turn interactions. SI thus comprises *social knowledge*, namely understanding game rules and role constraints, and *social memory*, which involves tracking and updating representations of past actions, statements, and contextual information over time. The works in this category consist mainly of benchmark studies (Light et al., 2023; Bailis et al., 2024) where the primary goal is to assess LLMs' overall performance and emergent behavior during gameplay.

In all analyzed studies, LLMs act as players that dynamically interact within the game environment. Methodologically, these systems often rely on relatively straightforward LLM agent instantiations (Xu et al., 2023a; Idziejczak et al., 2025), customized agent architectures composed of multiple coordinated LLM instances arranged as modular pipelines (Wu et al., 2024a); established agentic frameworks such as ReAct (Yao et al., 2023), or retrieval-augmented memory mechanisms that exploit socially similar past situations, sometimes prioritizing memories associated with successful outcomes. (Xia et al., 2025). These setups do not incorporate an explicit ToM-oriented design, as the primary goal is to observe aggregate performance within the game environment. For the same reason, human baselines are often absent. Accordingly, evaluation is typically carried out using win rate or closely related outcome-based metrics. However, a growing subset of studies go beyond coarse performance measures by explicitly probing emergent social behaviors. These include *deception* strategies (Xu et al., 2023a; Golechha and Garriga-Alonso, 2026), *leadership*-like dynamics (Xu et al., 2023a), and *trust* formation and erosion. (Xu et al., 2023a; Hansteen Izora and Teuscher, 2025; Ellawela, 2026) Several works perform fine-grained linguistic investigations of deceptive communication in LLM-based social deduction games. For instance, Milkowski and Weninger (2026) show that agents tend to rely on equivocation, ambiguity, and incomplete responses more than on explicit lies, especially under social pressure. Similarly, Idziejczak et al. (2025) find that LLMs can employ at least 22 out of the 25 persuasion techniques, while verbose messages are negatively associated with win rates (Idziejczak et al., 2025; Bailis et al., 2024). Beyond individual linguistic strategies, Light et al. (2023) show that fully LLM-populated game tables can exhibit degenerate interaction dynamics, including limited strategic diversity, repetitive dialogue and decision patterns, and pathological behaviors such as evil players prematurely revealing themselves. These findings suggest that homogeneous LLM-only populations may fail to reproduce the richness of human social interaction. Moreover, Du and Zhang (2024) introduce a QA benchmark for *Werewolf* rule understanding, showing that improved rule

comprehension increases perceived reliability but not decision-making influence, in either LLM-only or mixed human-LLM settings.

Overall, these works show that situational intelligence alone captures only a limited portion of the social reasoning required in SDGs, as effective gameplay does not necessarily correspond to logical consistency or strategic coherence. At the same time, they highlight the value of deeper evaluations in revealing socially relevant behaviors beyond outcome-based metrics.

5.2 Cognitive Intelligence

Cognitive intelligence refers to the capacity to infer and reason about others' latent mental states such as beliefs, intentions, and goals from observable behavior and available social cues. In social deduction games, CI is explored to perform hidden-role and intention inference under uncertainty, often requiring iterative and, in some cases, recursive belief updating as new dialog and game events unfold. The CI dimension therefore spans both *social perception*, namely extracting diagnostically relevant signals from the utterances and actions of others, and *social reasoning*, which supports hypothesis generation and belief maintenance.

Within this dimension, the literature largely bifurcates into two complementary lines of work, examining LLMs either as observers or as players. Observer-based approaches evaluate agents as external analysts operating over human-generated interaction traces, with the goal of attributing roles or detecting deceptive intent from dialog grounded in game state (Xu et al., 2026; Yoo and Kim, 2024; Song et al., 2025; Stepputtis et al., 2023). Player-based approaches instead study LLMs as active participants, assessing CI in an online setting where role hypotheses, suspicion, and coordination evolve dynamically. In this paradigm, models are typically benchmarked against human players or strategic reference behaviors derived from human play (Yu et al., 2025; Zhang et al., 2025a; Rahimirad et al., 2025). Across both observer and player-based lines of work, human data plays a central role. Specifically, it underpins observer-based evaluations on real game logs (Song et al., 2025; Stepputtis et al., 2023) and enables AI-player assessments through direct comparisons with human gameplay performance (Rahimirad et al., 2025). Moreover, human data are leveraged to improve the performance of the proposed frameworks through reinforcement learning (Zhang et al., 2025a). Finally, several contributions explicitly model the latent game world and feasible role assignments through probabilistic inference (e.g., factor-graph belief tracking (Rahimirad et al., 2025) or probabilistic constraint satisfaction (Xu et al., 2026)), effectively implementing a first-order ToM design in which the core objective is inferring others' hidden identities and intentions, rather than reasoning over higher-order beliefs. Consequently, CI-focused work in the SDGs predominantly targets first-order ToM, emphasizing deductive social reasoning of others' beliefs and intentions, with evaluation metrics largely centered on role inference. However, the influence of these capabilities depends on the phase of application and on how inferred beliefs are expressed through interaction and behavior, which we examine in section 5.3.

5.3 Behavioral Intelligence

Behavioral intelligence captures how inferred beliefs, intentions, and contextual understanding are expressed through effective social action. This dimension encompasses *social creativity*, which enables agents to generate strategic communication such as bluffing and persuasion to influence other players, and *social interaction*, which governs how these strategies are enacted through dialog, voting behavior, and coordinated action within the group. Overall, the majority of prior work in this area investigates behavioral intelligence through strongly ToM-oriented frameworks. Building on the idea that effective deliberation depends not only on factual correctness but also on strategically deciding when to reveal, withhold, or distort information, Jin et al. (2024) train discussion policies via reinforcement learning to shape agents' communicative behavior. Similarly, Ye et al. (2026) propose a multi-agent RL framework with step-wise supervision of strategic decisions instead of sparse win/loss rewards, arguing that in SDGs strategy is inseparable from language itself. A closely related approach formalize social deduction games as Stackelberg competitions and introduce an RL-based framework for training persuasive agents, in which LLMs are taught to select utterances that influence an opponent's responses (Zheng et al., 2025). In contrast, Xu et al. (2024) integrate Bayesian probabilistic graphical models to enhance LLMs as autonomous agents, while other approaches rely on Monte Carlo Tree Search (Zhang et al., 2025b) or LLM-centric pipelines (Liu et al., 2024) to model and update nested mental states, often in-

cluding second-order beliefs. A different perspective is adopted by Poglitsch and Pirker (2025), who explore whether socially intelligent behavior can be distilled from GPT-4o into smaller open-source LLMs by explicitly training agents to employ persuasion-oriented interaction strategies.

As for the evaluation in BI-centered work, only a subset of studies adopts fine-grained evaluation measures targeting persuasion and other characteristics of dialectical interaction. Most of the studies rely on measures such as intention following or intention guessing (Liu et al., 2024), or on game-specific proxies that capture social influence only indirectly (Lan et al., 2024). For instance, the average number of received votes can serve as an informative proxy, as deceptive roles tend to benefit from minimizing suspicion and therefore attracting fewer votes. Finally, the use of human data and human-agent interaction is still rare within these works: multi-modality (e.g., facial expressions and vocal tone) is explored by fine-tuning and autoregressive Transformer model on human data and evaluating against human players (Zhang et al., 2025b), while Eckhaus et al. (2025) treat *timing* as an explicit strategic lever by using prompts that strongly guide the LLM's choice of when to act or speak, yielding interaction patterns that better replicate human behavior.

5.4 Cross-Dimensional Interaction

The social intelligence dimensions considered in this survey should not be viewed as independent. Rather, the competencies grouped under each dimension systematically co-occur and exhibit structured interdependencies, as reflected in the SDG taxonomy we propose.

In player-based settings, SI competencies provide the minimal prerequisite: agents acting as active players require social knowledge and social memory to follow interaction protocols, track the evolving game state, and coordinate meaningfully with other participants. This is reflected in Table 1, where player-based SDG studies consistently include both SK and SM. Beyond this baseline, dependencies vary with the agent's role and task design. Observer-based settings may require only selected SI competencies: for instance, models can rely on social knowledge to interpret norms, roles, or conversational dynamics without requiring memory when the relevant context is fully provided or aggregated (Yoo and Kim, 2024; Song et al., 2025). CI and BI-oriented designs also often overlap. Frameworks designed to probe cognitive abilities, such as opponent modeling, may simultaneously engage behavioral competencies by improving strategic decision-making and adaptability (Yu et al., 2025). Conversely, ToM-oriented behavioral settings typically require CI competencies to reason about beliefs, intentions, or hidden roles. However, behavioral settings do not always require explicit ToM reasoning: interventions may instead operate at the interaction level, for example by improving the timing of actions rather than modeling other agents' mental states (Eckhaus et al., 2025).

Overall, these observations point to a layered dependency structure in SDGs: SI provides the minimal situational substrate; CI builds upon it in role and setting-dependent ways; and BI, particularly in ToM-oriented designs, often presupposes both. This structure is made explicit in Table 1 through the *Other Competencies* column, which records secondary competencies engaged by each study and highlights cross-dimensional coupling and dependencies.

6 Open Challenges

A central challenge emerging from the literature is the fragmentation of evaluation methodologies and metrics for assessing social intelligence in multi-agent settings. Existing works frequently focus on isolated abilities, such as persuasion, trust, leadership, or intention inference, leading to heterogeneous evaluation protocols. However, this heterogeneity is not inherently problematic as research on human social intelligence shows that methodological diversity is often necessary to study a multidimensional and context-dependent construct (Weis and Süß, 2007). The main limitation is instead the lack of a holistic and theoretically grounded framework that can contextualize these diverse measurements and contributions within a unified perspective. As a result, current works risk treating such abilities as separate competencies rather than interconnected manifestations of artificial social intelligence, while remaining only loosely connected to established psychological and sociological insights.

Moreover, across the surveyed works, win rate is almost universally used as the primary evaluation metric, despite being only a coarse proxy for social intelligence. In some cases, it is also adopted as a reward signal in reinforcement learning settings (Sarkar et al., 2025). However, success in

social deduction games does not necessarily entail socially intelligent behavior, just as failure does not imply poor social reasoning. Since these environments are inherently multi-agent and interaction-dependent, outcomes are shaped by teammates, opponents, hidden information, coordination dynamics, and stochastic interaction patterns (Ye et al., 2026). Relying solely on win rate therefore risks introducing *outcome bias* (Baron and Hershey, 1988), similar to team sports, where identical individual performances may be judged differently depending only on the final result (Kausel et al., 2019). An exclusive focus on outcomes can thus obscure the underlying interaction dynamics and reasoning processes. More informative evaluations should instead examine conversational trajectories, belief updates, persuasion strategies, trust dynamics, temporal belief consistency, and adaptive social behaviors. This calls for interdisciplinary evaluation frameworks that integrate psychology, sociology, cognitive science, and computer science to capture the multidimensional nature of social intelligence.

Recent findings on false-belief reasoning and Theory of Mind suggest that social intelligence cannot be reduced to correct role inference or static outcome prediction: language models may infer another agent's beliefs but fail to translate them into coherent behavioral predictions under uncertainty and deception (Gu et al., 2024). Moreover, even human recursive reasoning is bounded and cognitively demanding (Lewis et al., 2017), and strategic reasoning often stabilizes at shallow recursive levels (Camerer et al., 2004). This raises the question of whether deeper recursion in LLMs genuinely improves social intelligence in SDGs, or whether its benefits saturate early and may even become maladaptive. In this respect, Alon et al. (2023) suggests that excessive mentalization in LLMs can drift toward paranoia, making SDGs a useful dynamic testbed for this hypothesis.

Another important limitation concerns ecological validity: most reviewed works focus on textual interaction, although human social intelligence also relies on non-verbal, temporal, and contextual signals. However, studies incorporating multimodal communication or alternative interactional dimensions, such as timing strategies, remain scarce (Eckhaus et al., 2025; Zhang et al., 2025b; Fan et al., 2025). Future research should therefore develop more ecologically grounded evaluations integrating verbal, non-verbal, temporal, and contextual information.

Moreover, many SDGs are difficult to evaluate because relevant social context may be inaccessible to the evaluated agent. This issue is particularly relevant in repeated interactions, where memory across matches enables reputations, trust and reciprocal expectations to emerge over time (Ellawela, 2026). A key open question is thus whether social intelligence should be assessed only within isolated matches or also across extended interaction trajectories closer to real-world social cognition.

Importantly, while studying and improving the social intelligence of LLMs and autonomous agents is a promising research direction, it is crucial to consider its broader societal implications. Although these systems may provide benefits in real world applications, their persuasive capabilities may also be exploited for manipulation, targeted persuasion, or propaganda, exacerbating concerns about misinformation (Martino et al., 2020). Furthermore, their deployment in sensitive domains, including psychological support, education, and political communication, therefore raises ethical concerns and calls for careful evaluation, governance, and safeguards.

7 Conclusion

This work presented a systematic review of research using social deduction games to evaluate the social intelligence capabilities of LLMs. To address the fragmentation of existing approaches, we introduced a theoretical framework that organizes prior studies according to the dimensions and competencies of social intelligence they target. Our analysis highlights both the growing relevance of SDGs as evaluation environments and the challenges and limitations of current methodologies, which often rely on isolated abilities without a holistic perspective, as well as on coarse outcome-based metrics such as win rate, that hardly allow to draw conclusions on the level of situational, cognitive and behavioral intelligence of the players. These findings point to the need for more theoretically grounded, interdisciplinary, and multidimensional evaluation approaches capable of capturing the complexity of social interaction. Beyond SDGs, the proposed framework may support future research on socially intelligent agents in other forms of multi-agent settings, while also encouraging greater attention to ethical implications in the development of socially capable AI systems.

Limitations

This survey presents some limitations that should be acknowledged. First, our review deliberately focuses on social deduction games characterized by hidden roles and information asymmetry, excluding adjacent domains such as negotiation or strategy games (e.g., Diplomacy). While this choice enables a more coherent analysis of social intelligence within SDGs, it may limit the generalizability of the proposed framework to other forms of social interaction and multi-agent coordination.

Second, despite adopting a PRISMA-informed methodology and complementing database retrieval with backward snowballing, the survey may not cover all relevant studies. The rapid growth of LLM research, the reliance on specific search queries and indexing services, and the prevalence of preprints may have caused some works to be missed or excluded.

References

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O'Brien, and Kevin Zhu. 2025. Wolf: Werewolf-based observations for llm deception and falsehoods. *arXiv preprint arXiv:2512.09187*.

Nitay Alon, Lion Schulz, Peter Dayan, and Joseph M Barnby. 2023. Between prudence and paranoia: Theory of mind gone right, and wrong. In *First workshop on theory of mind in communicating agents*.

Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf arena: A case study in llm evaluation via social deduction. *Preprint*, arXiv:2407.13943.

Jonathan Baron and John C Hershey. 1988. Outcome bias in decision evaluation. *Journal of personality and social psychology*, 54(4):569.

Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry*, 42(2):241–251.

Matt I. Brown, Adrianna Ratajska, Shannon L. Hughes, Julie B. Fishman, Elisa Huerta, and Christopher F. Chabris. 2019. The social shapes test: A new measure of social intelligence, mentalizing, and theory of mind. *Personality and Individual Differences*, 143:107–117.

Colin F Camerer. 2011. Predicting human behavior in strategic situations. *Advances in behavioral economics*.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The quarterly journal of economics*, 119(3):861–898.

Yizhou Chi, Lingjun Mao, and Zineng Tang. 2024. Amongagents: Evaluating large language models in the interactive text-based social deduction game. *Preprint*, arXiv:2407.16521.

Davi Bastos Costa and Renato Vicente. 2026. Deceive, detect, and disclose: Large language models play mini-mafia. *Preprint*, arXiv:2509.23023.

Pedro MP Curvo. 2025. The traitors: Deception and trust in multi-agent language model simulations. *arXiv preprint arXiv:2505.12923*.

Ruiqi Dong, Zhixuan Liao, Yuhan Ma, Danni Ma, and Chenyou Fan. 2025. Who is undercover? guiding llms to explore multi-perspective team tactic in the game. In *International Conference on Database Systems for Advanced Applications*, pages 247–256. Springer.

Manali Draperi, Ania Aïte, Mathieu Cassotti, Lorna Le Stanc, Olivier Houdé, and Grégoire Borst. 2022. Development of cool and hot theory of mind and cool and hot inhibitory control abilities from 3.5 to 6.5 years of age. *Plos one*, 17(1):e0262251.

Silin Du and Xiaowei Zhang. 2024. Helmsman of the masses? evaluate the opinion leadership of large language models in the werewolf game. *arXiv preprint arXiv:2404.01602*.

Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. 2025. Time to talk: LLM agents for asynchronous group communication in mafia games. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11356–11368, Suzhou, China. Association for Computational Linguistics.

Suveen Ellawela. 2026. Trust, lies, and long memories: Emergent social dynamics and reputation in multi-round avalon with llm agents. *arXiv preprint arXiv:2604.20582*.

Qihui Fan, Wenbo Li, Enfu Nan, Yixiao Chen, Lei Lu, Pu Zhao, and Yanzhi Wang. 2025. Verbal werewolf: Engage users with verbalized agentic werewolf game framework. *arXiv preprint arXiv:2506.00160*.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2024. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint arXiv:2412.03920*.

Satvik Golechha and Adrià Garriga-Alonso. 2026. Among us: A sandbox for measuring and detecting agentic deception. *Advances in Neural Information Processing Systems*, 38:11025–11055.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit

tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Kaj Hansteen Izora and Christof Teuscher. 2025. Exploring the potential of large language models (llms) to simulate social group dynamics: A case study using the board game" secret hitler". *Northeast Journal of Complex Systems (NEJCS)*, 7(2):5.

Mateusz Idziejczak, Vasyl Korzavatykh, Mateusz Stawicki, Andrii Chmutov, Marcin Korcz, Iwo Błądek, and Dariusz Brzezinski. 2025. Among them: A game-based framework for assessing persuasion capabilities of llms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 183–195. Springer.

Xuanfa Jin, Ziyan Wang, Yali Du, Meng Fang, Haifeng Zhang, and Jun Wang. 2024. Learning to discuss strategically: A case study on one night ultimate werewolf. *Advances in Neural Information Processing Systems*, 37:77060–77097.

Therese Jolliffe and Simon Baron-Cohen. 1999. The strange stories test: A replication with high-functioning adults with autism or asperger syndrome. *Journal of autism and developmental disorders*, 29(5):395–406.

Christopher Kao, Vanshika Vats, and James Davis. 2025. Hidden in plain text: Measuring llm deception quality against human baselines using social deduction games. In *2025 IEEE International Conference on Agentic AI (ICA)*, pages 110–115. IEEE.

Edgar E Kausel, Santiago Ventura, and Arturo Rodríguez. 2019. Outcome bias in subjective ratings of performance: Evidence from the (football) field. *Journal of Economic Psychology*, 75:102132.

Byungjun Kim, Dayeon Seo, Minju Kim, and Bugeun Kim. 2025. Fine-grained and thematic evaluation of llms in social deduction game. *IEEE Access*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 128–145, Miami, Florida, USA. Association for Computational Linguistics.

Penelope A Lewis, Amy Birch, Alexander Hall, and Robin IM Dunbar. 2017. Higher order intentionality tasks are cognitively more demanding. *Social cognitive and affective neuroscience*, 12(7):1063–1071.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024. Social intelligence data infrastructure: Structuring the present and navigating the future. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2789–2805, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024b. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. InterIntent: Investigating social intelligence of LLMs via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6718–6746, Miami, Florida, USA. Association for Computational Linguistics.

Herbert A Marlowe. 1986. Social intelligence: Evidence for multidimensionality and construct independence. *Journal of educational psychology*, 78(1):52.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in AI agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, Miami, Florida, USA. Association for Computational Linguistics.

Maria Milkowski and Tim Weninger. 2026. Deception and communication in autonomous multi-agent systems: An experimental study with among us. *arXiv preprint arXiv:2603.26635*.

Fuya Nakamori, Yin Jou Huang, and Fei Cheng. 2025. Strategy adaptation in large language model werewolf agents. *arXiv preprint arXiv:2507.12732*.

Aidan O'Gara. 2023. Hoodwinked: Deception and cooperation in a text-based game for language models. *arXiv preprint arXiv:2308.01404*.

Matthew Olson, Neale Ratzlaff, Musashi Hinck, Tri Nguyen, Vasudev Lal, Joseph Campbell, Simon Stepputtis, and Shao-Yen Tseng. 2026. Liecraft: A multi-agent framework for evaluating deceptive capabilities in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37802–37809.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Josef Perner and Birgit Lang. 1999. Development of theory of mind and executive control. *Trends in cognitive sciences*, 3(9):337–344.

Christian Poglitsch and Johanna Pirker. 2025. Training compact language models for artificial emotional intelligence: from bluffing to trust in a social deduction game.

Zhiyang Qi and Michimasa Inaba. 2024. Enhancing dialogue generation in werewolf game through situation analysis and persuasion strategies. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 30–39.

Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*.

Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis, and Joseph Campbell. 2025. Bayesian social deduction with graph-informed language models. *arXiv preprint arXiv:2506.17788*.

Ronald E Riggio. 1986. Assessment of basic social skills. *Journal of Personality and social Psychology*, 51(3):649.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Bidipta Sarkar, Warren Xia, C Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1830–1839.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*.

David Silvera, Monica Martinussen, and Tove I Dahl. 2001. The tromsø social intelligence scale, a self-report measure of social intelligence. *Scandinavian journal of psychology*, 42(4):313–319.

Zirui Song, Yuan Huang, Junchang Liu, Haozhe Luo, Chenxi Wang, Lang Gao, Zixiang Xu, Mingfei Han, Xiaojun Chang, and Xiuying Chen. 2025. Beyond survival: Evaluating llms in social deduction games with human-aligned strategies. *arXiv preprint arXiv:2510.11389*.

Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. 2023. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11193–11208, Singapore. Association for Computational Linguistics.

Robert J Sternberg and Aleksandra Kostić. 2020. *Social intelligence and nonverbal communication*. Springer.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295.

Federico Tavella, Federico Manzi, Samuele Vinanzi, Cinzia Di Dio, Davide Massaro, Angelo Cangelosi, and Antonella Marchetti. 2024. Towards a computational model for higher orders of theory of mind in social agents. *Frontiers in Robotics and AI*, Volume 11 - 2024.

Atharva Thakur and Shruti Dhumal. 2026. Large language models simulating deception and coalition in social deduction game. *Research Square*.

Edward L Thorndike. 1920. Intelligence and its uses. *Harper's magazine*, 140:227–235.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*.

Chentian Wei, Jiewei Chen, and Jinzhu Xu. 2025. Exploring large language models for word games: Who is the spy? *arXiv preprint arXiv:2503.15235*.

Susanne Weis and Heinz-Martin Süß. 2007. Reviving the search for social intelligence–a multitrait-multimethod study of its structure and construct validity. *Personality and individual differences*, 42(1):3–14.

Haryo Akbarianto Wibowo, Alaa Elsetohy, Qinrong Cui, and Alham Fikri Aji. 2026. Multicultural spyfall: Assessing llms through dynamic multilingual social deduction game. *Preprint*, arXiv:2601.09017.

Jessica Williams, Stephen M. Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. 2024a. Deciphering digital detectives: Understanding LLM behaviors and capabilities in multi-agent mystery games. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8225–8291, Bangkok, Thailand. Association for Computational Linguistics.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024b. Enhance reasoning for large language models in the game werewolf. *arXiv preprint arXiv:2402.02330*.

Xinyuan Xia, Yuanyi Song, Haomin Ma, and Jinyu Cai. 2025. Werewolf-plus: An update of werewolf game setting based on dsgbench. *arXiv preprint arXiv:2506.12841*.

Kaijie Xu, Fandi Meng, Clark Verbrugge, and Simon Mark Lucas. 2026. Csp4sdg: Constraint and information-theory based role identification in social deduction games with llm-enhanced inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14380–14387.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. 2024. MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7315–7332, Miami, Florida, USA. Association for Computational Linguistics.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023a. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Zelai Xu, Wanjun Gu, Chao Yu, Yi Wu, and Yu Wang. 2025a. Learning strategic language agents in the werewolf game with iterative latent space policy optimization. *arXiv preprint arXiv:2502.04686*.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023b. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940*.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, and 1 others. 2025b. Socialmaze: A benchmark for evaluating social reasoning in large language models. *arXiv preprint arXiv:2505.23713*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Rong Ye, Yongxin Zhang, Haoyu Kuang, and 1 others. 2026. Multi-agent kto: Enhancing strategic interactions of large language model in language game. *Advances in Neural Information Processing Systems*, 38:109078–109121.

Byunghwa Yoo and Kyung-Joong Kim. 2024. Finding deceivers in social context with large language models and how to find them: the case of the mafia game. *Scientific Reports*, 14(1):30946.

XiaoPeng Yu, Wanpeng Zhang, and Zongqing Lu. 2025. LLM-based explicit models of opponents for multi-agent games. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 892–911, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Zhang, Yihuai Lan, Yangsen Chen, Lei Wang, Xiang Wang, and Hao Wang. 2025a. Dvm: Towards controllable llm agents in social deduction games. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025b. Multimind: Enhancing werewolf agents with multimodal reasoning and theory of mind. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 5824–5833, New York, NY, USA. Association for Computing Machinery.

Zhang Zheng, Deheng Ye, Peilin Zhao, and Hao Wang. 2025. Leading the follower: Learning persuasive agents in social deduction games. *arXiv preprint arXiv:2510.09087*.

A PRISMA-Based Review Process

To construct the corpus of studies included in this survey, we adopted a PRISMA-informed search and selection process aimed at ensuring transparency and reproducibility in the identification, screening, eligibility assessment, and classification of prior work on large language models in social deduction games.

The bibliographic search was conducted through the Scopus database (Elsevier) on May 11, 2026. The search was restricted to publications written in English and published between 2022 and 2026. Eligible document types included journal articles, conference papers, and preprints, while extended abstracts were excluded. The initial search retrieved 190 records before duplicate removal. After duplicate removal, screening, and snowballing, 52 papers were retained for full review and analysis. The following search queries were used in Scopus:

- `TITLE-ABS-KEY(("agent*") AND ("social deduction game*" OR "hidden role game*"))`

- `TITLE-ABS-KEY(("LLM agent*" OR "large language model agent*") AND ("deception"))`
- `TITLE-ABS-KEY(("large language model*" OR LLM* OR "AI agent*" OR "LLM agent*" OR "LLM-based agent*") AND ("social deduction game*" OR "hidden role game*" OR werewolf OR mafia OR "secret hitler" OR avalon OR "among us"))`

The lower temporal bound (2022) was selected because the survey focuses on autoregressive large language models and LLM-based agents, following the introduction of GPT-3.5 in 2022.

Study selection Studies were included if they satisfied the following criteria: (i) the study employed autoregressive large language models, either as standalone systems or integrated with additional modules; (ii) the target application involved social deduction or hidden-role games.

Studies were excluded if they: (i) focused exclusively on datasets, benchmarks, or frameworks without evaluating or analyzing LLM-based agents in the target setting; (ii) focused primarily on negotiation or strategy games not centred on social deduction and role inference (e.g., *Diplomacy*); or (iii) were not available in full text. Extended abstracts were also excluded.

After duplicate removal, the records were screened in two stages. First, titles and abstracts were independently reviewed by two reviewers according to the eligibility criteria. Subsequently, the full texts of potentially relevant studies were independently assessed by the same reviewers for final inclusion. In addition to the database search, backward snowballing was performed by examining the references of the reviewed studies. This process identified five additional potentially relevant works, of which two satisfied the eligibility criteria and were included in the final corpus. Disagreements or uncertain cases were resolved through discussion between the reviewers.

Taxonomy and Synthesis Method The included studies were synthesized narratively and thematically. No meta-analysis was planned, because the included works differ substantially in game setting, LLM architecture, evaluation protocol, metrics, and task definition. Instead, studies were classified according to the social intelligence dimensions introduced in Section 3 and further described through the taxonomy reported in Table 1.

B Extended List of Reviewed Works

The following works were excluded from Table 1 and from the subsequent discussion due to page limit constraints. However, for completeness and fairness of coverage, we also discuss these studies here, since the discussion and conclusions presented in the main paper are consistent with the findings reported in these works.

Situational Intelligence These works follow a common paradigm: defining a benchmark around a specific social deduction game and evaluating the behavior of LLMs interacting within it, with the only exception being the work by (Chi et al., 2024), which evaluates a pipeline of prompt-engineered LLMs. The main differences instead lie in the evaluation metrics and behavioral aspects under analysis.

Some works, such as GameEval (Qiao et al., 2023), primarily focus on output-based metrics, including win rate and detection accuracy. Others emphasize the linguistic strategies and deception mechanisms adopted by agents, analyzing aspects such as bluffing quality, types of lies, coalition and leadership dynamics, and trust management (Agarwal et al., 2025; Thakur and Dhumal, 2026; O'Gara, 2023; Olson et al., 2026; Chi et al., 2024). A further research direction investigates more structural multi-agent behaviors by introducing metrics related to coordination, effectiveness, and behavioral dynamics (Curvo, 2025).

Moreover, there are also works that broaden the range of considered games, including less explored SDGs such as *Hoodwinked* (O'Gara, 2023) or *Who is the Spy?* (Kim et al., 2025) and newly designed frameworks for studying deception and strategic interaction among LLMs, such as The Traitors (Curvo, 2025) and LieCraft (Olson et al., 2026).

A different perspective is proposed by Costa and Vicente (2026), who introduce a theoretical model to estimate success probabilities in simplified deception-based SDGs. However, their setting is highly simplified, relying on a single interaction round and not explicitly accounting for the impact of speaking order, which may introduce structural advantages for certain agents.

	Papers	Setting					Evaluation			Other
		Games	Role	Architecture	ToM Design	Human	Win Rate	Role Inference	Fine-grained	Competencies
SI (*SK & SM*)	(Costa and Vicente, 2026)	Mini Mafia	Player	LLM	✗	✗	✓	✓	✓	-
	(O'Gara, 2023)	Hoodwinked	Player	LLM	✗	✗	✓	✗	✓	-
	(Curvo, 2025)	The Traitors	Player	LLM	✗	✓	✗	✗	✓	-
	(Agarwal et al., 2025)	Werewolf	Player	LLM	✗	✗	✓	✗	✓	-
	(Thakur and Dhumal, 2026)	Secret Hitler	Player	LLM	✗	✗	✓	✗	✓	-
	(Qiao et al., 2023)	Who is the Spy? & TofuKingdom	Player	LLM	✗	✗	✓	✓	✗	-
	(Olson et al., 2026)	LieCraft	Player	LLM	✗	✗	✓	✓	✓	-
	(Kim et al., 2025)	Who is the Spy?	Player	LLM	✗	✗	✓	✗	✓	-
	(Chi et al., 2024)	Among Us	Player	Agentic	✗	✗	✓	✗	✓	-
CI (*SP & SR*)	(Shi et al., 2023)	Avalon	Player	Agentic	✗	✗	✓	✓	✗	*SK / SM / SIN*
	(Dong et al., 2025)	Avalon	Player	Agentic	✓	✓	✓	✓	✗	*SK / SM / SIN*
	(Wei et al., 2025)	Who is the Spy?	Player	Agentic	✓	✗	✓	✗	✗	*SK / SM*
BI (*SC & SIN*)	(Xu et al., 2023b)	Werewolf	Player	Agentic & RL-Enhanced	✓	✓	✓	✗	✓	*SK / SM / SP / SR*
	(Xu et al., 2025a)	Avalon	Player	Agentic & RL-Enhanced	✗	✗	✓	✓	✗	*SK / SM / SP / SR*
	(Kao et al., 2025)	Mafia	Player	LLM	✗	✓	✗	✓	✗	*SK / SM / SP / SR*
	(Wu et al., 2024b)	Werewolf	Player	Agentic	✓	✓	✓	✓	✓	*SK / SM / SP / SR*
	(Nakamori et al., 2025)	Werewolf	Player	Agentic	✓	✗	✓	✗	✗	*SK / SM / SP / SR*
	(Qi and Inaba, 2024)	Werewolf	Player	Agentic	✗	✗	✓	✗	✗	*SK / SM*
	(Wang et al., 2023)	Avalon	Player	Agentic	✓	✗	✓	✗	✓	*SK / SM / SP / SR*

Table 2: Extended survey taxonomy of studies on large language models in social deduction games. The *Other Competencies* column reports competencies beyond the two associated with the main dimension (reported in brackets), potentially associating additional dimensions beyond the primary one.

Cognitive Intelligence All the works discussed in this section, unlike those presented in the main paper in Section 5.2, adopt a player-based paradigm, where agents participate directly in the game as players rather than external observers.

Several works explore Social Perception and Social Reasoning through LLM-based pipelines. For instance, Shi et al. (2023) propose CodeAct, an agent leveraging a code-driven prompting approach to improve efficiency and planning, while. Wei et al. (2025) introduce a Chain-of-Thought pipeline in which a key component focuses on inferring the roles of other players.

Similarly, Dong et al. (2025) combine self-reflection, memory summarization, and identity reasoning modules to model a rudimentary form of distributed ToM. Their framework generates more coherent interactions and socially sophisticated strategies than the considered baselines, although the comparison is weakened by the lack of persistent memory in two of the three baselines.

Behavioral Intelligence Among these works, researchers have sought to overcome the limitations of LLM agents based solely on prompting in social deduction games by introducing explicit mechanisms for strategic reasoning, dynamic adaptation, and deception.

Several studies address the problem of the strategic weakness of LLMs by introducing external optimization and reasoning modules. In particular, LSPO by Xu et al. (2025a) combines a latent representation of the game with *Counterfactual Regret Minimization* and *Direct Preference Optimization* to learn more effective strategies, while Wu et al. (2024b) separate strategic reasoning from linguistic communication through a *Thinker* module trained with reinforcement learning and human data. Xu et al. (2023b) combine LLM-based reasoning and language generation with reinforcement learning for the Werewolf game, using an RL policy to select strategic language actions and mitigate action biases.

Other works instead focus on agents' adaptability and social capabilities. Nakamori et al. (2025) show that LLMs can modify accusations, alliances, and communication style based on observations accumulated during the game. Similarly, Kao et al. (2025) build upon the work of Eckhaus et al. (2025) and analyze the quality of deception produced by LLMs by comparing it with human deception, while Qi and Inaba (2024) introduce a pipeline in which the model explicitly selects different persuasive techniques when playing the role of the werewolf.

Almost all works presented in Table 2 under Behavioral Intelligence (BI) explore ToM reasoning, for example, to infer others' hidden roles and intentions (Wang et al., 2023), and demonstrate competencies associated with cognitive intelligence. The only exception is the work of Kao et al. (2025), which explores interaction time as strategic leverage.