CSP4SDG: Constraint and Information-Theory Based Role Identification in Social Deduction Games with LLM-Enhanced Inference

Kaijie Xu[1], **Fandi Meng**[2], **Clark Verbrugge**[1], **Simon Mark Lucas**[2]

[1]Department of Computer Science, McGill University, Montreal, Quebec, Canada
[2]Queen Mary University of London, London, United Kingdom
kaijie.xu2@mail.mcgill.ca, f.meng@qmul.ac.uk, clump@cs.mcgill.ca, simon.lucas@qmul.ac.uk

Abstract

In Social Deduction Games (SDGs) such as *Avalon*, *Mafia*, and *Werewolf*, players conceal their identities and deliberately mislead others, making hidden-role inference a central and demanding task. Accurate role identification, which forms the basis of an agent's belief state, is therefore the keystone for both human and AI performance. We introduce **CSP4SDG**, a probabilistic, constraint–satisfaction framework that analyses gameplay objectively. Game events and dialogue are mapped to four linguistically-agnostic constraint classes—*evidence*, *phenomena*, *assertions*, and *hypotheses*. Hard constraints prune impossible role assignments, while weighted soft constraints score the remainder; information-gain weighting links each hypothesis to its expected value under entropy reduction, and a simple closed-form scoring rule guarantees that truthful assertions converge to classical hard logic with minimum error. The resulting posterior over roles is fully interpretable and updates in real time. Experiments on three public datasets show that CSP4SDG (i) outperforms LLM-based baselines in every inference scenario, and (ii) boosts LLMs when supplied as an auxiliary "reasoning tool." Our study validates that principled probabilistic reasoning with information theory is a scalable alternative—or complement—to heavy-weight neural models for SDGs.

Code — https://github.com/Nortrom1213/CSP4SDG
Extended version — https://arxiv.org/abs/2511.06175

Introduction

Social Deduction Games (SDGs), like *Avalon*, *Mafia*, and *Werewolf*, require players to infer hidden roles despite deception and sparse evidence. Although recent efforts cover transformer role classifiers, reinforcement learning agents, and planning-LLM hybrids (de Ruiter and Kachergis 2018; Serrino et al. 2019; Ibraheem, Zhou, and DeNero 2022; Lai et al. 2023), they typically treat chat as opaque text, require heavy training, or embed game-specific heuristics—hindering interpretability and cross-title transfer.

We propose **CSP4SDG**, a training-free, probabilistic constraint-satisfaction framework. A lightweight LLM converts raw logs into four language-agnostic constraint types; hard constraints prune impossible worlds, and soft ones receive information-gain weights that remove manual tuning required by valued or VOI-CSPs (Schiex et al. 1995; MacKay 1992). The solver returns calibrated posteriors and MAP assignments that stand alone or refine the LLM.

To test the generality of our approach, we run a unified evaluation on three public SDG datasets that cover crowd-sourced chat and large-scale log files. On each dataset, we compare three reasoning methods—*pure CSP*, *LLM-only*, and the hybrid *LLM+CSP*—while ablating the CSP solver through five settings. Every experiment is repeated under multiple player perspectives (objective, good-roles, evil-roles) and both truthful- and deceptive-good conditions. Finally, we perform a backbone ablation with successively stronger LLMs, isolating the effect of language-model capacity. This comprehensive design allows us to examine (i) how much structure alone can achieve under different settings, (ii) how much an LLM can learn without structure, and (iii) how the two components interact when combined.

Our main contributions are as follows:

- **Generalized Probabilistic CSP Framework for Role Inference:** We formulate SDG role inference as a training-free probabilistic constraint-satisfaction problem that integrates logical filtering and information-theoretically weighting in a single, generic framework.
- **LLM-driven end-to-end workflow:** We design a lightweight LLM pipeline that converts raw game logs into structured constraints and seamlessly couples them with the CSP solver, yielding an interpretable, plug-and-play reasoning module.
- **Empirical Validation Across Multiple Datasets:** Experiments on three public datasets—covering various CSP settings, diverse viewpoints, and several LLM backbones—show that CSP4SDG consistently outperforms baseline methods and reliably boosts LLM reasoning.

Related Works

Extensive research has been conducted in SDGs. Early studies extracted lexical and pragmatic deception cues (Zhou and Sung 2008; Niculae et al. 2015); the Mafiascum corpus enabled supervised classifiers (de Ruiter and Kachergis 2018), later extended to transformer and LLM benchmarks (Ibraheem, Zhou, and DeNero 2022; Stepputtis et al. 2023). Mul-

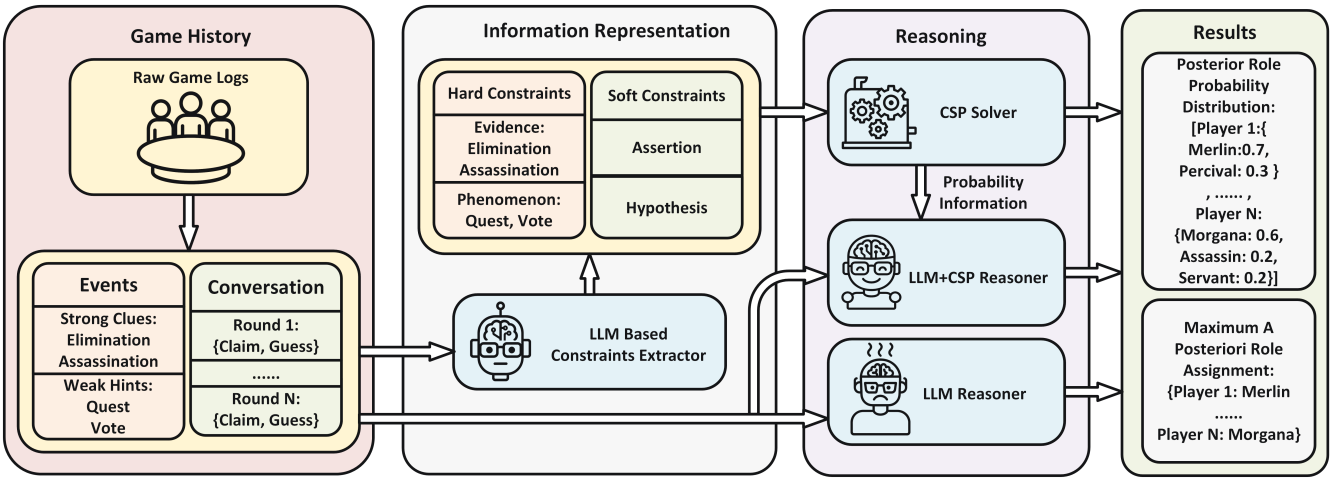


Figure 1: Schematic overview of the proposed inference architecture for social–deduction games. **Game History** (left) comprises objective *event* traces (e.g., eliminations, assassinations, quest outcomes) and subjective *conversation* turns. **Information Representation** (centre) leverages an auxiliary LLM (Brown et al. 2020) to transform raw logs into a structured constraint set: (i) *hard constraints* (evidence and phenomena) that are logically inviolable, and (ii) *soft constraints* (player assertions and hypotheses) that contribute graded probabilistic weight. **Reasoning** (right) contrasts three inference engines. A *plain LLM* lacks the combinatorial apparatus required for reliable role deduction; a *hybrid LLM + CSP* benefits from externally supplied posteriors but is still bottlenecked by heuristic language reasoning; our *CSP solver* enforces all hard constraints and optimally scores soft ones, delivering calibrated posterior distributions and MAP role assignments that achieve the highest empirical accuracy.

timodal persuasion (Lai et al. 2023), reinforcement-learning from logs (Serrino et al. 2019), mental-state agents (Nakamura et al. 2016), endgame SVMs (Chuchro 2022), and planning–LLM hybrids (Bakhtin et al. 2022) further diversified the field; Other works explore SDG research in multiple dimensions (Kopparapu et al. 2022; Eger and Martens 2018; Chi, Mao, and Tang 2024; Sarkar et al. 2025; Kim, Seo, and Kim 2024; Velikov 2021; Martinenghi et al. 2024; Wu et al. 2024; Carminati et al. 2024). Yet most approaches rely on heavy task-specific training, treat dialogue as opaque text, or lack principled uncertainty—gaps our CSP4SDG fills with a training-free, interpretable, and game-agnostic reasoning module. Full survey in **Appendix A** (in **extended version**).

Classical constraint satisfaction offers exact, interpretable search (Freuder and Mackworth 2006); soft (Schiex et al. 1995) and probabilistic (Fargier, Lang, and Schiex 1996) variants rank near-consistent worlds but depend on hand-tuned costs. Information-theoretic criteria quantify evidential value (MacKay 1992). CSP4SDG unifies these strands: LLM-extracted constraints become IG-weighted soft relations, eliminating manual costs and yielding calibrated posteriors without retraining.

Methods

Problem Definition and Framework

SDGs involve a group of players secretly assigned different roles. Each player aims to deduce the hidden roles of others through various events and conversations. Precise *role-level* inference—not merely a good/evil split—is pivotal in SDGs with multiple special abilities: knowing a player's role lets teammates interpret their actions as structured evidence and coordinate strategy, while adversaries can exploit that same knowledge to craft targeted deceptions that mislead opponents who possess only partial information. To facilitate and automate the role identification process, we propose a generalized constraint-based framework for various SDGs.

General Formulation An SDG scenario includes (i) players $P = \{p_1, p_2, \ldots, p_n\}$, (ii) roles $R = \{r_1, r_2, \ldots, r_m\}$ partitioned into "good" (e.g., Loyal Servant, Villager) and "evil" (e.g., Assassin, Mafia), and (iii) a world-state assignment $a : P \to R$ that maps each player to a role in R.

Given observed game events and players' utterances at time t, our task is to infer and update the posterior distribution of players' roles $Pr(r \mid \mathcal{C}_t)$ for each player-role pair (p_i, r), where $\mathcal{C}_t$ represents all observed constraints until time t. This role identification task can serve two purposes:

1. **Direct Role Prediction**: To directly infer the most probable roles of each player to support automated agents in making in-game decisions or to analyze player strategies in post-game scenarios.
2. **Auxiliary Decision Support**: To provide probabilistic insights as auxiliary input for human players or for models (such as LLMs) to enhance their reasoning capability and potential decision-making accuracy during the game.

Constraint-Based Representation Our proposed generalized framework categorizes in-game information into four constraint types. Each constraint type has a clear semantic interpretation, logical formulation, and standardized template, as illustrated in Table 1.

The distinction between these constraint types lies primarily in their certainty level and consequently, their treatment

in the inference process:

- **Evidence (E) and Phenomenon (P)**: Evidence fixes a player's role exactly, whereas Phenomenon narrows each player's role domain; both are hard constraints that prune any assignment violating them.
- **Assertions (A)**: These player-made statements are treated as soft constraints with extremely high weights ($w_A \gg 1$). Assignments satisfying more assertions are exponentially favored, ensuring that satisfying assertions is highly prioritized to hypotheses.
- **Hypotheses (H)**: Representing weaker, player-driven speculations or supports; assigned relatively low weights. They provide subtle probabilistic preferences to the inference without strictly excluding possibilities.

CSP-Based Role Inference Mechanism

Given the generalized constraint framework introduced previously, we now describe the role inference process in detail. Our approach leverages Constraint Satisfaction Problems (CSPs) and a specialized scoring function to estimate posterior probabilities of player roles.

Constraint Satisfaction Framework Formally, we define the CSP-based role inference as follows. An SDG-CSP model at time t is a tuple: $\mathcal{M}_t = (P, R, \mathcal{C}_t)$, where P is the player set, R is the role set, and $\mathcal{C}_t$ is the accumulated constraint set consisting of evidence (E), phenomenon (P), assertions (A), and hypotheses (H).

The feasible assignment space at time t, denoted as AS_t, consists of all assignments $a : P \to R$ satisfying the accumulated hard constraints ($E \cup P$):

$$AS_t = \{a \mid a \models E, a \models P\}.$$

Assignment Filtering via Hard Constraints Each new evidence or phenomenon constraint reduces the feasible assignment space. Formally, for any constraint $c \in E \cup P$, the set of feasible assignments is updated by:

$$AS_{t+1} = \{a \in AS_t \mid a \models c\}, \quad \text{where } AS_{t+1} \subseteq AS_t.$$

Thus, hard constraints monotonically reduce the feasible set, ensuring consistency and correctness in role inference.

Scoring Function with Weighted Soft Constraints To incorporate player-driven constraints (Assertions and Hypotheses), we introduce a novel scoring function. Given an assignment a, the scoring function is defined as:

$$S(a) = \left(\prod_{c \in A, a \models c} w_A \right) \cdot \left(1 + \sum_{h \in H, a \models h} w_H(h) \right), \quad (1)$$

where:

- $w_A \gg 1$ is the high weight assigned to assertions, strongly motivating their satisfaction.
- $w_H(h)$ is the weight of hypotheses, calculated either manually or via Information Gain (IG): $w_H(h) = \text{IG}(h)$

To estimate posterior probabilities, we normalize these scores over the entire feasible set (Cover 1999):

$$Pr(a|\mathcal{C}_t) = \frac{S(a)}{\sum_{a' \in AS_t} S(a')}. \quad (2)$$

Information Gain Weighting We adopt information-theoretic principles to dynamically adjust hypothesis weights. For hypothesis h, the information gain (IG) is calculated as the reduction in entropy when incorporating h:

$$\text{IG}(h) = H(\text{prior}) - H(\text{posterior}|h), \quad (3)$$

where entropy $H(X) = -\sum_x Pr(x) \log Pr(x)$.

The IG-based weighting measures how informative a hypothesis is, assigning greater weight to hypotheses that significantly clarify role distributions. When all assertions are truthful, using high–weight soft constraints is (almost) equivalent to treating them as hard constraints: the posterior probability error is bounded by $|Pr_s(a) - Pr_h(a)| \leq 1/w_A$ (Formal proof in **Appendix B**).

Inference Procedure and Computational Complexity Given the constraint set $\mathcal{C}_t$, we proceed in three steps:

1. **Prune** Apply all *hard* constraints $(E \cup P)$ to eliminate infeasible worlds and obtain the candidate set AS_t.
2. **Score & normalize** For every $a \in AS_t$ (or for a Monte-Carlo sample thereof), compute the soft–weighted score $S(a)$ from Eq. (1) and obtain the normalized posterior $Pr(a \mid \mathcal{C}_t)$ via Eq. (2).
3. **Marginals and MAP** Derive (i) player–wise marginal posteriors by summing $Pr(a \mid \mathcal{C}_t)$ over worlds that assign a given role, and (ii) the maximum–a-posteriori world $\hat{a} = \arg\max_{a \in AS_t} Pr(a \mid \mathcal{C}_t)$.

For larger lobbies, we *could* replace step 2 with an MCMC sampler (Andrieu and Thoms 2008): starting from a feasible world, role-swap proposals would respect global role counts and mixing could be sped up by caching partial soft-scores. With M samples this strategy would scale as $\mathcal{O}(Mn)$, give an unbiased estimator of all marginals, and take the highest-scoring sample as MAP. Because current SDG datasets are low-dimensional ($n \leq 10$), we did not deploy this variant; testing it is left to future work on highly heterogeneous games such as *Blood on the Clocktower*.

Dataset

Game Overview

Avalon Players are secretly assigned to *good* roles (Merlin, Percival, Servants) or *evil* roles (Morgana, Assassin). Each round follows the fixed loop *team proposal* → *vote* → *quest resolution*; after three successful quests the Assassin may attempt to identify Merlin for a last-chance victory.

Mafia Roles are *Mafia* (evil) and *Bystanders* (good). The game alternates *night* (Mafia privately eliminate one target) and *day* (public discussion + lynch vote) cycles until either faction is wiped out.

Datasets Description

We evaluate our approach on three datasets:

- **Avalon NLU Dataset** (Stepputtis et al. 2023): 21 games (6 players each), detailed dialogues, votes and quests.
- **Mafia Dataset** (Ibraheem, Zhou, and DeNero 2022): 44 games (4-10 players, 460 participants), detailed dialogues, votes, and night eliminations.

Type	Game	Event / Dialogue Cue	Constraint Format (grammar)	Mathematical Representation
Evidence	Avalon	Assassin kills Merlin	*role_is(p1/p2,assassin/merlin)*	$a \models \big(\text{role}(p_1, p_2) = [\text{assassin}, \text{merlin}]\big)$
	Mafia	Night victim revealed	*role_is(p, Bystander)*	$a \models \text{role}(p) = \text{Bystander}$
Phenomenon	Avalon	Quest q has f fail cards	*evil_at_least(team, f)*	$a \models \sum_{p \in \text{team}} \mathbf{1}[\text{evil}(p)] \geq f$
Assertion	Avalon	"I am Percival."	*assert_role_is(speaker, role)*	$S(a) \times = \ w_A \ \mathbf{1}[a \models \text{role}(s) = \text{Percival}]$
	Avalon	"This team is clean."	*assert_team_good(speaker, team)*	$S(a) \times = \ w_A \ \mathbf{1}[\forall p \in \text{team good}(p)]$
	Avalon	"X is evil."	*assert_role_in(sp, tg, evil)*	$S(a) \times = \ w_A \ \mathbf{1}[\text{evil}(tg)]$
Hypothesis	Avalon	Weak suspicion / NO vote	*hypo_role_in(sp, tg, evil)*	$S(a) \ += w_{mid} \ \mathbf{1}[\text{evil}(tg)]$
	Avalon	Weak support / YES vote	*hypo_role_in(sp,tg,good)*	$S(a) += w_{low} \ \mathbf{1}[\text{good}(tg)]$
	Avalon	Proposer chooses squad	*hypo_team_good(proposer, team)*	$S(a) += w_{strong} \ \mathbf{1}[\forall p \in \text{team good}(p)]$
	Mafia	"X looks Mafia."	*hypo_role_in(sp, tg, mafia)*	$S(a) += w_{mid} \ \mathbf{1}[\text{mafia}(tg)]$
	Mafia	"X looks good."	*hypo_role_in(sp, tg, bystander)*	$S(a) += w_{mid} \ \mathbf{1}[\text{bystander}(tg)]$
	Mafia	YES vote to lynch	*hypo_role_in(voter, target, mafia)*	$S(a) += w_{low} \ \mathbf{1}[\text{mafia}(target)]$

Table 1: Constraint catalogue across datasets. *Constraint Format* shows the abstract grammar used by the extractor; *Mathematical Representation* shows how each constraint modifies the score $S(a)$ of an assignment a. Assertions multiply S by a large factor $w_A \gg 1$, whereas hypotheses add small weights w_{low}, w_{mid} and w_{high}(dataset-specific, all less than 1).

- **AvalonLogs Dataset** (WhoaWhoa 2022): Large-scale dataset with 12,699 Avalon games (5-10 players), recording quests, votes, roles, and assassination choices.

Table 1 lists the full catalogue with the corresponding mathematical semantics used by the solver. We manually annotate Dataset 1 and Dataset 2 as Truth or Lie, where Lie games have at least one explicitly false role claim from a good-aligned player (details in **Appendix C**).

Dataset Preprocessing and Constraint Extraction via LLM

For every game log we (i) **segment** the transcript into temporal blocks—quests for Avalon, day/night cycles for Mafia; (ii) **annotate** each block with LLMs (gpt-4.1-mini) using a few-shot prompt that maps every utterance or event to one of four first-order constraint schemas; the model outputs a JSON list of predicate instances; (iii) **post-process** the JSON by normalising player IDs, deduplicating constraints, enforcing type checks, and dropping ill-formed entries. We then feed the resulting constraint sets sequence directly to the CSP solver. AvalonLogs is processed automatically with the same pipeline. A rigorous, dual-author audit on Datasets 1 and 2 validated our constraint extraction, confirming 100% fidelity by tracing all hard constraints to the raw logs and ensuring the inclusion of all mandatory events. (See the complete prompt templates in **Appendix F**.)

Experiments

In this section we describe our experimental protocol and evaluation results across three datasets (Avalon–NLU, Mafia, AvalonLogs). We compare three classes of methods:

- **CSP-only:** purely constraint-based inference under five settings (*Strict*, *+Assert*, *+HypIG*, *+HypM*, *+TurnIG*).
- **LLM-only:** direct probability estimation by prompting a LLM on dialogue history (global vs. turn-based).
- **LLM+CSP:** LLM predictions augmented with CSP posterior or MAP "assist" under the same five CSP settings.

We evaluate three *perspectives*: the *objective* view (public evidence only), each *good* role's private view (Merlin, Percival, Servant, Bystander), and each *evil* role's view (Assassin, Mafia). For AvalonLogs—which lacks dialogue—we run only the full-information variants (LLM-Global and LLM,+CSP,+HypIG) and sample one quest per game to keep the large corpus tractable.

Unlike prior work that merges roles into broad classes, we report exact per-player role accuracy, a much harder target. Potential baselines like DeepRole (Serrino et al. 2019) and SVC-Assassin (Chuchro 2022) are not included: DeepRole optimizes win-rate through full-game policy learning, SVC-Assassin predicts only the final assassination, and both models ignore dialogue, hard-code the variants, and depend on rule-specific network engineering, making them incomparable to our round-level, multi-dataset role-inference protocol. For the same reasons, we also forgo dataset-specific fine-tuning of the LLM backbones: the corpora are too small for reliable supervised adaptation, instruction-tuned models are the standard baselines in previous works, and preserving out-of-domain generalization is central to our evaluation.

Methods and Settings

CSP-only We accumulate *evidence* and *phenomenon* constraints and evaluate under five settings:

1. *Strict*: evidence + phenomenon only.
2. *+Assert*: adds high-weight assertions (carried forward).
3. *+HypIG*: adds all hypotheses with IG weights.
4. *+HypM*: adds all hypotheses with manual weights. (grid-search tuned in experiments; details in Appendix I)
5. *+TurnIG*: adds current-turn hypotheses with IG weights.

LLM-only We prompt the LLM with the chat history up to the current round (either global or turn-only) and the public event log of quests or eliminations, asking it to return for every player a probability distribution over roles together with a MAP assignment that respects the known role counts.

LLM+CSP Identical to LLM-only but the prompt is augmented with the CSP posterior table and MAP assignment under one CSP setting (we experiment with all five).

Family	Setting	Cond.	Avalon-NLU						Mafia-UCB		
			Obj	MRL	PCV	ASN	MRG	LSV	Obj	BYS	MAF
CSP	Strict	T	0.28/0.24	0.63/0.60	0.69/0.67	0.59/0.53	0.59/0.53	0.56/0.53	0.70/0.90	**0.72/0.91**	**1.00/1.00**
	+Assert	T	0.33/0.29	**0.65/0.60**	0.75/0.73	**0.61/0.53**	**0.61/0.53**	0.59/0.60	0.70/0.90	0.72/0.91	1.00/1.00
	+HypIG	T	**0.34/0.30**	0.65/0.60	0.76/0.79	0.61/0.53	0.61/0.53	**0.60/0.65**	**0.73/0.76**	**0.74/0.76**	1.00/1.00
	+HypM	T	**0.33/0.31**	0.65/0.60	0.75/0.79	0.61/0.53	0.61/0.53	0.60/0.65	0.73/0.76	0.74/0.76	1.00/1.00
	+TurnIG	T	0.34/0.29	0.65/0.60	**0.76/0.80**	0.61/0.53	0.61/0.53	0.60/0.63	0.72/0.79	0.73/0.79	1.00/1.00
LLM	GChat	T	0.14/0.20	0.20/0.24	0.16/0.19	0.13/0.19	0.16/0.25	0.13/0.21	0.56/0.62	0.56/0.61	0.59/0.77
	TChat	T	0.11/0.12	0.21/0.25	0.14/0.18	0.11/0.12	0.16/0.20	0.11/0.16	0.66/0.72	0.61/0.66	0.67/0.82
LLM + CSP	Strict	T	0.16/0.12	0.41/0.39	0.44/0.41	0.42/0.39	0.42/0.39	0.12/0.17	0.69/0.89	0.70/0.89	0.89/0.93
	+Assert	T	0.19/0.15	0.42/0.40	0.47/0.45	0.44/0.40	0.44/0.40	0.13/0.18	0.69/0.89	0.70/0.89	0.89/0.93
	+HypIG	T	0.19/0.16	0.42/0.40	0.48/0.50	0.44/0.40	0.44/0.40	0.12/0.19	0.69/0.74	0.70/0.74	0.89/0.93
	+HypM	T	0.19/0.16	0.42/0.40	0.47/0.50	0.44/0.40	0.44/0.40	0.13/0.20	0.69/0.74	0.70/0.73	0.89/0.93
	+TurnIG	T	0.19/0.15	0.42/0.40	0.47/0.45	0.44/0.40	0.44/0.40	0.11/0.15	0.69/0.86	0.70/0.86	0.92/0.95
Random	–	–	0.2667	0.4	0.4	0.5833	0.5833	0.3333	0.68	0.6889	1.00

Table 2: Truthful-Good Results on dialogue datasets (**Lying-Good results in Appendix G**). Each cell reports *marginal accuracy* / *MAP accuracy*. **Strict**: evidence & phenomenon only; **+Assert**: adds assertions; **+HypIG**: global hypotheses (IG weights); **+HypM**: global hypotheses (manual); **+TurnIG**: turn-local hypotheses; **GChat**: LLM with global chat + state; **TChat**: LLM with turn chat, state. Role abbreviations—Avalon: MRL Merlin, PCV Percival, ASN Assassin, MRG Morgana, LSV Loyal Servant; Mafia: BYS Bystander, MAF Mafia; Obj = objective. Only the first highest value per metric is highlighted in bold.

Views

Let V be the set of all perspectives: $V = \{\text{objective}\} \cup \{\text{each role}\}$. For each $v \in V$ we add "perspective evidence": If $v =$ Merlin, add hard constraints that all evils are known; if $v =$ Percival, constrain the Merlin/Morgana candidates to those roles; if v is evil, reveal fellow evils; otherwise, none.

Evaluation Metrics

For each round q of a game we evaluate two quantities:

Marginal Accuracy Let P be the number of players, $r_p^\star$ the ground-truth role of player p, and $a_{p,r}$ the model's posterior for role r is: $\mathrm{MA}_q = \frac{1}{P}\sum_{p=1}^{P} a_{p,r_p^\star}$.

MAP Accuracy Let $\hat{r}_p$ be the role assigned to p in the model's maximum–a-posteriori (MAP) world, then: $\mathrm{MAP}_q = \frac{1}{P}\sum_{p=1}^{P} \mathbf{1}[\hat{r}_p = r_p^\star]$.

Aggregation Both metrics are first averaged over the Q rounds of a single game, then across the set $\mathcal{G}$ of games in a dataset Overall $m = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \frac{1}{Q}\sum_{q=1}^{Q} m_q$ where $m \in \{\mathrm{MA}, \mathrm{MAP}\}$, reporting the mean and standard deviation.

Experimental Protocol

For Avalon–NLU and Mafia we run all methods × settings × views at every round. For AvalonLogs we sample a single quest per game and run CSP-only (5 settings), LLM-only (global) and LLM+CSP (*HypIG*). Scaling all weights by factor $\lambda \in \{0.25, 0.5, 1, 2, 4\}$ set altered MAP accuracy by $\leq$ 1.5 pp (percentage points), so we retain the default (w_{strong}, w_{mid}, w_{low}) = (0.5,0.2,0.1) weights (tuned in **Appendix I**) for all reported results. In addition, on Avalon–NLU we conduct supplementary *LLM-only* experiments to compare pure dialogue-based inference across multiple LLMs (GPT-4o-mini, GPT-4o, GPT-4.1, Deepseek-v3, Gemini-2.0-flash). Details are presented in **Appendix D**.

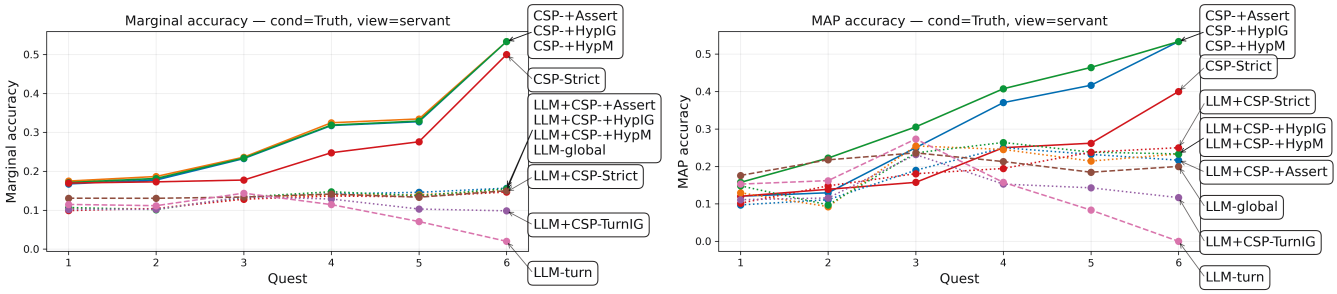
Results

Dialogue Datasets Analysis

Table 2 summarizes the results for both the Avalon-NLU and Mafia-UCB datasets under various conditions. Because the Lying-Good split contains only a handful of games (n = 4), we regard its findings as preliminary; the complete L-condition table is provided in **Appendix G**. Based on both sets of results, several clear observations emerge:

CSP methods dominate consistently Regardless of whether good players lie, CSP methods achieve the highest overall accuracy. Specifically, CSP+HypIG and CSP+HypM settings yield the best performance, with HypIG marginally outperforming HypM due to information-gain weighting. Interestingly, the CSP+TurnIG setting, which incorporates turn-local hypotheses, surpasses CSP+Assert, and CSP+Assert is superior to CSP-Strict. In certain viewpoints, however, CSP+Assert already attains the column-wise optimum, and adding further soft cues does not raise—but also never lowers—the score, making the extra layers a pure bonus. This trend indicates the incremental benefit of adding soft information in structured formats. All reported improvements of CSP over the LLM baseline are statistically significant according to two-sided Wilcoxon signed–rank and paired t-tests ($p < 0.05$); see **Appendix E** for test details.

LLM struggles with complex reasoning Pure LLM approaches significantly underperform compared to CSP in the more complex Avalon-NLU dataset, indicating difficulties in precise role deduction from dialogue alone. Introducing CSP-derived information (LLM+CSP methods) slightly improves overall accuracy, especially from special-role viewpoints (Merlin, Percival, Morgana, Assassin). However, from the objective viewpoint and Loyal Servant perspective, accuracy gains are minimal or even negative, reflecting that limited structured knowledge may mislead the LLM.



(a) Marginal accuracy by quest (servant view)

(b) MAP accuracy by quest (servant view)

Figure 2: Quest-by-quest accuracy trends on Avalon-NLU. Quest 6 = assassination round following three successful quests.

Family	Setting	AvalonLogs — 6-player						AvalonLogs — 9-player					
		Obj	MRL	PCV	MRG	LSV	EVM	Obj	MRL	PCV	MRG	LSV	EVM
CSP	Strict	0.290	0.614	0.635	0.584	0.390	0.691	0.302	0.595	0.568	0.554	0.348	0.614
	+HypIG	0.326	0.615	0.637	0.577	0.419	0.729	0.360	0.666	0.525	0.578	0.403	0.674
LLM	Global	0.283	0.360	0.325	0.298	0.290	0.464	0.280	0.360	0.309	0.313	0.291	0.417
LLM+CSP	+HypIG	0.289	0.613	0.633	0.584	0.389	0.691	0.301	0.593	0.564	0.554	0.347	0.614

Table 3: AvalonLogs performance split by game size. Each cell shows *marginal accuracy*. Roles: Obj (objective), MRL (Merlin), PCV (Percival), MRG (Morgana), LSV (loyal–servant aggregate), EVM (evil–minion aggregate). The 6-player subset illustrates a small-scale scenario, while the 9-player subset represents a larger game. In both cases CSP-based methods outperform a pure LLM baseline, and the hybrid *LLM+CSP* variant retains most CSP gains while leveraging LLM reasoning.

Simpler dialogue benefits LLM For the simpler Mafia-UCB dataset, performance differences between LLM and CSP methods narrow significantly. Here, the TChat setting (turn-based contexts) notably outperforms the global-context setting (GChat). Moreover, under the CSP-supported HypIG condition, the LLM+CSP approach even surpasses pure CSP, illustrating the synergy possible when combining structured priors and simpler dialogue contexts.

However, a critical limitation of the LLM becomes evident: despite having perfect information from the Mafia perspective, the pure LLM still fails to achieve full accuracy due to inherent hallucinations and contextual confusion. Even with 100% correct CSP priors, the LLM struggles to disregard misleading or irrelevant dialogue elements, underscoring the necessity for explicit fine-tuning or stronger constraints to directly guide accurate reasoning.

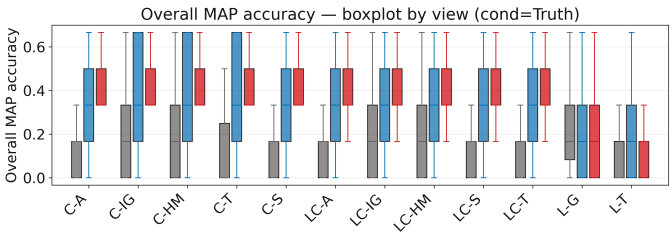
Marginal vs. MAP accuracy Interestingly, pure CSP consistently shows slightly higher marginal accuracy compared to MAP accuracy. The reverse pattern occurs for LLM methods, suggesting that while LLM confidently converges towards a single best solution, the CSP generates a more probabilistic distribution of roles, spreading uncertainty across multiple assignments.

Impact of Lying-Good behavior In both Avalon-NLU and Mafia-UCB, only a few games contain explicit role-lying from good players (**Appendix G**). In these limited cases we see anecdotal shifts: CSP accuracy sometimes moves in favor of good-aligned perspectives, whereas our prompt-based LLM baselines tend to drop slightly across viewpoints. Because the sample is too small for firm statistics, we treat these observations as illustrative only. A larger, dedicated study of deception effects is left for future work.

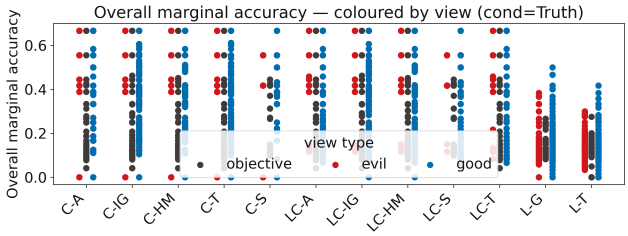
No-dialogue Large Scale Dataset: AvalonLogs

Without dialogue, AvalonLogs provides only quest and vote records. We give six- and nine-player results here in Table 3 and place full tables in **Appendix H**. Only Strict and +HypIG settings are shown, because all other variants alter accuracy by ≤ 2 pp, and +HypIG is typically the best. Under +HypIG setting, the objective score rises to $0.33 \sim 0.36$, while role views climb further (Merlin 0.67, Evil-Minion 0.73). These gains result mainly from event-level assertions, augmented by IG-weighted hypotheses.

LLM alone, confined to frequency heuristics, stays near 0.28 objectively and below 0.37 for special roles; CSP surpasses it by 5–8 pp on low-information roles and 25–30 pp on informative ones. Feeding CSP back to the model yields the same scores as CSP-Strict, confirming that LLM performs better in no-dialogue dataset yet still relies on CSP for combinatorial reasoning. Compared with the dialogue-rich Avalon-NLU corpus, gains here are smaller because the evidence ceiling is lower, but CSP remains dominant. Accuracy also exhibits a size-related pattern—dipping at 7 players, recovering at 8–9, and falling again at 10 (discussion in **Appendix H**). Conversely, LLM and LLM+CSP degrade on Avalon-NLU, showing that unfiltered dialogue can swamp the model even with accurate CSP hints. Overall,(1) mission and vote data alone give a solid but bounded signal, whereas dialogue unlocks the deeper role logic that makes SDGs dialogue-driven; (2) LLMs excel at information extraction but not at the structured reasoning as CSP does.



(a) Overall MAP accuracy by view type



(b) Overall marginal accuracy scatter plot by view type

Figure 3: Accuracy distribution across different perspectives (*Truthful* condition). Good-role perspectives (blue) achieve higher accuracy, evil-role perspectives (red) exhibit broader variance, and objective views (gray) record lower accuracy levels. Abbreviations: C = CSP, LC = LLM+CSP, L = LLM; A = +Assert, IG = +HypIG, HM = +HypM, T = TurnIG, S = Strict.

Model	**LLM-Global**					**LLM+CSP+HypIG**				
	Obj	ASN	PCV	MRG	LSV	Obj	ASN	PCV	MRG	LSV
GPT-4o-mini	0.14/0.20	0.13/0.19	0.16/0.19	0.16/0.25	0.13/0.21	0.19/0.16	0.44/0.40	0.48/0.50	0.44/0.40	0.12/0.19
GPT-4o	0.15/0.16	0.12/0.08	0.20/0.27	0.17/0.18	0.15/0.16	0.19/0.17	0.44/0.40	0.47/0.50	0.44/0.41	0.15/0.15
GPT-4.1	0.15/0.12	0.11/0.08	0.25/0.30	0.16/0.14	0.15/0.13	0.19/0.16	0.43/0.40	0.48/0.50	0.42/0.40	0.15/0.15
DeepSeek-v3	0.16/0.16	0.20/0.16	0.22/0.25	0.19/0.21	0.17/0.16	0.19/0.16	0.46/0.40	0.51/0.55	0.46/0.40	0.16/0.14
Gemini-2.0-flash	0.16/0.18	0.13/0.12	0.18/0.22	0.18/0.16	0.16/0.17	0.19/0.16	0.41/0.39	0.48/0.50	0.38/0.36	0.16/0.17

Table 4: Marginal/MAP accuracy on Avalon-NLU under pure LLM and LLM+CSP+HypIG. Merlin omitted (identical across setups; likely as Merlin's deterministic constraints heavily prune the domain, leading to uniform inferences across all models).

Analysis of Trends and Viewpoints

Figures 2a and 2b show how accuracy evolves throughout the quests from the servant's viewpoint. CSP-based methods (solid lines) exhibit consistent improvement in accuracy, particularly from Quest 3 onward, sharply rising in the final assassination round (Quest 6). CSP variants utilizing soft constraints (+Assert, +HypIG, +HypM) lead performance, whereas the Strict setting lags slightly due to reliance solely on hard constraints. Pure LLM methods (dashed and dotted lines) have minimal improvements or even degrade over time, especially in the LLM-turn mode. Augmenting LLM with CSP posterior information (LLM+CSP) provides moderate gains but still falls short compared to pure CSP, highlighting the advantage of structured probabilistic reasoning.

Figure 3a confirms that accuracy varies significantly by perspective, with evil-role views consistently achieving higher median accuracies across methods. Good-role perspectives, however, exhibit notable variance, reflecting their susceptibility to misleading or ambiguous dialogue. The objective perspective generally records the lowest accuracies, emphasizing that informed viewpoints substantially aid role prediction. Figure 3b further highlights this trend, clearly separating good/evil-role predictions toward higher accuracy, while objective predictions scatter more broadly, indicating the nuanced challenges inherent to this perspective.

Ablation: Different LLM Capabilities

Across LLM backbones, from lightweight to state-of-the-art, increasing model scale yields only modest and sometimes negative returns on role-inference accuracy, as shown in Table 4. This suggests the performance bottleneck is not a matter of model capacity but a conceptual limitation, as larger models tend to amplify superficial heuristics rather than master the combinatorial reasoning essential for SDGs. Conversely, augmenting any LLM with CSP posteriors produces a uniform and significant accuracy uplift, particularly for roles with hidden information like Assassin, Percival, and Morgana. This consistent improvement demonstrates that the hybrid design successfully addresses a core reasoning deficit inherent in current language models. Therefore, the results indicate that structured, constraint-based reasoning is not merely a temporary fix for present models but will likely remain an essential component for achieving high-fidelity inference, underscoring the framework's long-term relevance even as foundation models continue to evolve.

Conclusion

In this work, we introduced CSP4SDG, a generalized probabilistic constraint-satisfaction framework for role inference in SDGs. We formulated role identification as a CSP by decomposing game information into structured, linguistically-agnostic constraints, and leveraged information-theoretic principles to dynamically weight soft constraints. Extensive evaluations across three public datasets with multiple CSP configurations, varying role perspectives, and LLM ablation experiments confirmed that pure LLMs are insufficient for human-level role identification tasks, and CSP4SDG consistently achieves superior accuracy and interpretability compared to pure LLM baselines and hybrid approaches. Future work will explore interactive, human-in-loop constraint processing, test on large and complex datasets, and extend CSP-driven posteriors from passive inference to active, real-time decision support within live gameplay.

References

Andrieu, C.; and Thoms, J. 2008. A tutorial on adaptive MCMC. *Statistics and computing*, 18(4): 343–373.

Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A.; et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074.

Braverman, M.; Etesami, O.; and Mossel, E. 2008. Mafia: A Theoretical Study of Players and Coalitions in a Partial Information Environment. *The Annals of Applied Probability*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Carminati, L.; Zhang, B. H.; Farina, G.; Gatti, N.; and Sandholm, T. 2024. Hidden-Role Games: Equilibrium Concepts and Computation. In *Proceedings of the 25th ACM Conference on Economics and Computation*, 106–107.

Chi, Y.; Mao, L.; and Tang, Z. 2024. Amongagents: Evaluating large language models in the interactive text-based social deduction game. *arXiv preprint arXiv:2407.16521*.

Chuchro, R. 2022. Training an assassin ai for the resistance: Avalon. *arXiv preprint arXiv:2209.09331*.

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

de Ruiter, B.; and Kachergis, G. 2018. The mafiascum dataset: A large text corpus for deception detection. *arXiv preprint arXiv:1811.07851*.

Eger, M.; and Martens, C. 2018. Keeping the story straight: A comparison of commitment strategies for a social deduction game. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, 24–30.

Fargier, H.; Lang, J.; and Schiex, T. 1996. Mixed constraint satisfaction: A framework for decision problems under incomplete knowledge. In *AAAI/IAAI, Vol. 1*, 175–180.

Freuder, E. C.; and Mackworth, A. K. 2006. Constraint satisfaction: An emerging paradigm. In *Foundations of artificial intelligence*, volume 2, 13–27. Elsevier.

Ibraheem, S.; Zhou, G.; and DeNero, J. 2022. Putting the Con in Context: Identifying Deceptive Actors in the Game of Mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kim, B.; Seo, D.; and Kim, B. 2024. Fine-Grained and Thematic Evaluation of LLMs in Social Deduction Game. *arXiv preprint arXiv:2408.09946*.

Kopparapu, K.; Duéñez-Guzmán, E. A.; Matyas, J.; Vezhnevets, A. S.; Agapiou, J. P.; McKee, K. R.; Everett, R.; Marecki, J.; Leibo, J. Z.; and Graepel, T. 2022. Hidden agenda: a social deduction game with diverse learned equilibria. *arXiv preprint arXiv:2201.01816*.

Lai, B.; Zhang, H.; Liu, M.; Pariani, A.; Ryan, F.; Jia, W.; Hayati, S. A.; Rehg, J.; and Yang, D. 2023. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. *Association for Computational Linguistics: ACL 2023*.

MacKay, D. J. 1992. Information-based objective functions for active data selection. *Neural computation*, 4(4).

Martinenghi, A.; Donabauer, G.; Amenta, S.; Bursic, S.; Giudici, M.; Kruschwitz, U.; Garzotto, F.; Ognibene, D.; et al. 2024. LLMs of Catan: Exploring Pragmatic Capabilities of Generative Chatbots Through Prediction and Classification of Dialogue Acts in Boardgames' Multi-party Dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing@ LREC-COLING 2024*, 107–118. ELRA and ICCL.

Nakamura, N.; Inaba, M.; Takahashi, K.; Toriumi, F.; Osawa, H.; Katagami, D.; and Shinoda, K. 2016. Constructing a human-like agent for the werewolf game using a psychological model based multiple perspectives. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE.

Niculae, V.; Kumar, S.; Boyd-Graber, J.; and Danescu-Niculescu-Mizil, C. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1650–1659.

Sarkar, B.; Xia, W.; Liu, C. K.; and Sadigh, D. 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*, AAMAS '25, 1830–1839. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Schiex, T.; Fargier, H.; Verfaillie, G.; et al. 1995. Valued constraint satisfaction problems: Hard and easy problems. *IJCAI (1)*, 95: 631–639.

Serrino, J.; Kleiman-Weiner, M.; Parkes, D. C.; and Tenenbaum, J. 2019. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32.

Stepputtis, S.; Campbell, J. P.; Xie, Y.; Qi, Z.; Zhang, W.; Wang, R.; Rangreji, S.; Lewis, C.; and Sycara, K. 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*, 11193–11208.

Velikov, G. V. 2021. *RLereWolf–Reinforcement Learning Agent Development Framework For The Social Deduction Game Werewolf*. Ph.D. thesis, University of Aberdeen.

WhoaWhoa. 2022. Avalon: The Resistance Logs. https://github.com/WhoaWhoa/avalonlogs.

Wu, S.; Zhu, L.; Yang, T.; Xu, S.; Fu, Q.; Wei, Y.; and Fu, H. 2024. Enhance reasoning for large language models in the game werewolf. *arXiv preprint arXiv:2402.02330*.

Zhou, L.; and Sung, Y.-w. 2008. Cues to deception in online Chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 146–146. IEEE.