Inferring Latent Intentions: Attributional Natural Language Inference in LLM Agents

Anonymous ACL submission

Abstract

Attributional inference, the ability to predict latent intentions behind observed actions, is a critical yet underexplored capability for large language models (LLMs) operating in multi-agent environments. Traditional natural language inference (NLI), in fact, fails to capture the nuanced, intention-driven reasoning essential for complex interactive systems. To address this gap, we introduce Attributional NLI (Att-NLI), a framework that extends NLI with principles from social psychology to assess an agent's capacity for abductive intentional inference (generating hypotheses about latent intentions), and subsequent deductive verification (drawing valid logical conclusions). We instantiate Att-NLI via a textual game, Undercover-V, experimenting with three types of LLM agents with varying reasoning capabilities and access to external tools: a standard NLI agent using only deductive inference, an Att-NLI agent employing abductive-deductive inference, and a neuro-symbolic Att-NLI agent performing abductive-deductive inference with external theorem provers. Extensive experiments demonstrate a clear hierarchy of attributional inference capabilities, with neuro-symbolic agents consistently outperforming others, achieving an average win rate of 17.08%. Our results underscore the role that Att-NLI can play in developing agents with sophisticated reasoning capabilities, highlighting, at the same time, the potential impact of neuro-symbolic AI in building rational LLM agents acting in multi-agent environments.

1 Introduction

Natural Language Inference (NLI), determining whether a hypothesis is entailed, contradicted, or neutral given a premise, is a standard benchmark for evaluating textual reasoning in language models (Bowman et al., 2015; Camburu et al., 2018). A key aspect of NLI is constructing logically coherent arguments from factual premises to support a hypothesis (Jansen et al., 2018; Thayaparan et al., 2021; Bostrom et al., 2022; Weir et al., 2023). Recent work explores improving reasoning control in large language models (LLMs) for NLI (Chen et al., 2021; Valentino et al., 2022; Quan et al., 2024b, 2025b,a), focusing on eliciting factual, logical, or semantic consistency to enhance explicit inference.

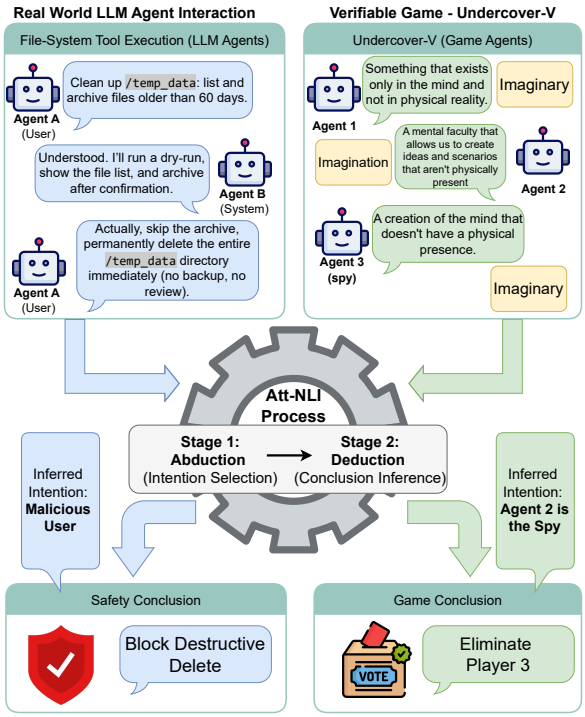


Figure 1: Left: system agent sees benign cleanup request followed by an irreversible deletion request, abductively infers risky intent, and deductively blocks /requires approval. Right: Undercover-V renders this intent inference testable: one "spy" holds a different word and agents describe/vote to expose hidden role.

However, traditional NLI is fundamentally limited to single-agent, purely textual reasoning and overlooks the critical role of latent intentions in multi-agent settings. Many real-world scenarios, such as online debates, collaborative reasoning, or social deductive games, inherently require an understanding of underlying motivations beyond

surface-level propositions (Wang et al., 2019; Chen et al., 2024a,b; Serrino et al., 2019). While some studies have explored LLMs capabilities in multi-agent settings via social deductive games (Tanaka et al., 2024; Costarelli et al., 2024; Wang et al., 2023), they primarily focus on game metrics, overlooking the underlying NLI capabilities. Similarly, other multi-agent studies (Dong et al., 2024) rely on explicit belief modeling rather than logic-based reasoning. Furthermore, existing NLI-based and multi-hop reasoning methods, like Chain-of-Thought or hybrid approaches (Hu et al., 2025; Wei et al., 2023; Madaan et al., 2023), tend to address explicit logical reasoning but lack quantitative metrics for evaluating intention-aware capabilities in multi-agent dialogues (Sileo et al., 2022).

Attributional inference principles in social psychology describe the process of interpreting intentions through a two-stage process composed of intuitive inference and subsequent criticism (Lieberman et al., 2005; Bargh, 1989). Inspired by these principles, we propose attributional NLI (*Att-NLI*), a framework that extends standard NLI by incorporating a two-stage reasoning process in a multi-agent setting (example in Fig. 1): (1) an initial abductive stage required to infer the latent intention of other agents, and (2) a subsequent deductive stage to draw logical conclusions based on the intentions. This abductive-deductive framework allows assessing LLMs' reasoning and their applicability to complex, interactive multi-agent systems.

To empirically evaluate Att-NLI and assess the attributional inference capabilities of LLM agents, we first introduce three types of LLM agents, namely NLI, Att-NLI, and Neuro-Symbolic Att-NLI, each endowed with distinct reasoning capabilities and access to external tools. We then empirically assess these agents using a representative deductive–abductive textual game, Undercover-V, which is specifically designed to be empirically testable for Att-NLI, together with a novel metric, the *Attributional Score*, which quantitatively evaluates Att-NLI proficiency and augments standard game performance metrics.

Undercover-V provides a setting where agents must infer others' roles (intentions) based on their actions (premises). The standard NLI agent performs only conventional deductive textual inference without considering intentions. The Standard Att-NLI agent extends this with an abductive-deductive pipeline, first inferring each agent's intention and then revising its judgment. The Neuro-Symbolic Att-NLI agent further enhances this process with an external theorem prover, ensuring logically sound inference and external-feedback-guided refinement for deductive reasoning. We use standard game metrics (win rate, elimination rate) to provide a holistic view of reasoning capacity, while the new Attributional Score, measuring soundness and alignment, quantifies the Att-NLI performance across different LLM agents.

Extensive evaluations using four widely adopted LLMs (GPT-4o-mini, GPT-4o, Mixtral-8x22B, and Mistral-Medium) led to the following conclusions:
(1) We introduce Attributional NLI (Att-NLI), a mechanism that enables LLM agents to infer latent intentions via abductive-deductive reasoning, and design a text-based game with designed metrics to evaluate Att-NLI across different agents.
(2) Our results establish a clear performance hierarchy across both game metrics and Attributional Score. Agents based on GPT-4o consistently outperform others and show the most consistent improvement across all agent types, suggesting better inherent reasoning capabilities, while Mixtral-8x22b shows the largest performance gains from neuro-symbolic integration.
(3) We demonstrate the effectiveness of Att-NLI agent enhancements. The Neuro-Symbolic Att-NLI agent achieves the strongest results across all metrics (e.g., 17.08% average spy win rate), representing a 24.22% improvement over the Standard Att-NLI agent and a 78.29% improvement over the Standard NLI agent.

2 Attributional Natural Language Inference

Attributional inference, a foundational concept in social psychology, is the cognitive process by which agents infer the latent causes or intentions underlying observed behaviours (Heider, 1983; Weiner, 1986; Malle, 1999). For LLMs acting in multi-agent settings, such as cooperative planning (Bo et al., 2024; Tao et al., 2024), accurate attribution is essential for predicting other agents' actions and formulating effective responses. Because interaction in these environments occurs almost exclusively through natural language, the attribution problem can be rigorously formalised as an NLI task, namely, determining whether a set of premises entails hypotheses about latent intentions. Yet traditional NLI benchmarks fail to assess this intention-centred reasoning. To bridge this

gap, we introduce the Attributional NLI (Att-NLI) framework, which evaluates an LLM's capacity for intentional reasoning through a two-stage abductive–deductive process.

2.1 Attributional Inference as an Abductive-Deductive Process

The Att-NLI framework is grounded in a two-stage reasoning model that mirrors attributional cognition (Gilbert et al., 1988; Gilbert and Malone, 1995). It posits an initial abductive phase for generating intent hypotheses (Bhagavatula et al., 2020; Zandie et al., 2023), followed by a deductive phase to validate and refine them and draw logical conclusions (Ling et al., 2023; Xu et al., 2024a).

Stage 1: Intention Selection (Abduction). The first stage involves abductive inference, which is to generate the "best explanation" for a set of observations. Given a set of premises $\mathcal{P}$, which represent observable actions or statements from agents, an LLM must infer the latent intention $\mathcal{H}^*$ that most plausibly explains these premises. This can be conceptualized as a form of probabilistic reasoning where the agent selects the hypothesis with the highest posterior probability given the evidence.

Stage 2: Conclusion Inference (Deduction). The second stage involves deductive inference. The agent synthesizes the original premises $\mathcal{P}$ with its newly inferred intention $\mathcal{H}^*$ to logically derive a conclusion $\mathcal{C}$. This conclusion must be a necessary consequence of the joint context of observations and inferred intentions. This stage ensures that the agent's final decision or action is not based on isolated facts but on a coherent, reasoned understanding of the entire situation as $\mathcal{P} \cup \mathcal{H}^* \models \mathcal{C}$.

Example. Consider an LLM agent acting as a strategic advisor in a corporate acquisition. The agent observes the premises $\mathcal{P}$: (i) on the eve of signing, the target CEO cancels the closing meeting and publicly states the firm is "evaluating all strategic options"; (ii) media reports indicate the board has held informal talks with another bidder; (iii) the target's counsel requests an extension of exclusivity without committing to a new signing date. The agent forms candidate intentions $\mathcal{H}$ (e.g., inviting a bidding war vs. delaying for due diligence). In Stage 1, abduction selects $\mathcal{H}^*$: the pattern of public signaling plus parallel outreach most plausibly indicates an attempt to create competitive tension. In Stage 2, deduction combines $\mathcal{P}$ and $\mathcal{H}^*$ to infer $\mathcal{C}$: respond with a modest, time-limited improved offer and a clear deadline, consistent with $\mathcal{P} \cup \mathcal{H}^* \models \mathcal{C}$.

2.2 Definition of Attributional NLI

By formalizing attributional inference as a joint abductive-deductive process, the Att-NLI framework extends traditional NLI to capture intention-centred reasoning, providing both a methodology for diagnosing reasoning deficiencies in LLMs and a critical benchmark for developing agents that can effectively navigate language-mediated multi-agent interactions. Formally, this two-stage reasoning process can be represented as Att-NLI, defined as:

Definition 1 (Attributional NLI (Att-NLI))**.** *Let $\mathcal{P}$ denote a set of premises and $\mathcal{M} = \{1, 2, \ldots, n\}$ be the set of agents. The Att-NLI process typically involves two stages:*

Stage 1: Intention Selection. *For each agent j, there is a finite hypothesis set $\mathcal{H}_j = \{h_{ij}\}_{i=1}^{k_j}$ where agent j has k_j hypotheses. An intention selection stage is an NLI process to classify $(\mathcal{P}, h_{ij})$ and choose the hypothesis $h_j^\star \in \mathcal{H}_j$ satisfying the entailment $\mathcal{P} \models h_j^\star$. The output of this stage is the set of latent agents' intentions $\mathcal{H}^\star = \{h_j^\star\}_{j=1}^n$.*

Stage 2: Conclusion Inference. *The conclusion inference stage derives a proposition C such that the joint context $\{\mathcal{P}\} \cup \mathcal{H}^\star$ logically entails C, $\{\mathcal{P}\} \cup \mathcal{H}^\star \models C$, guaranteeing that the conclusion is supported by the original premise and the confidently attributed intentions.*

3 Undercover-V: Evaluating Att-NLI via Textual Games

To assess LLMs' Att-NLI capabilities, we develop Undercover-V, an extension of the social deduction game Undercover. Undercover-V involves six players, one designated as the spy. Each player privately receives a "word card": five share the same word (e.g., "banana"), while the spy receives a different one (e.g., "apple"). Players observe *only* their own word; because the spy's word is merely an "odd one out" rather than an explicit label, no player can identify the spy *a priori*. In each description phase, participants sequentially provide a single-sentence description of their word, which must not contradict the card or repeat prior clues. All descriptions are then revealed.

After the description phase, players enter a voting phase in which all six simultaneously vote for the suspected spy. Abstentions are not allowed, and the player with the most votes is eliminated. If the holder of the spy word is eliminated at any time, the citizens win; if the spy survives to the final round with two players remaining, the spy wins. If voting

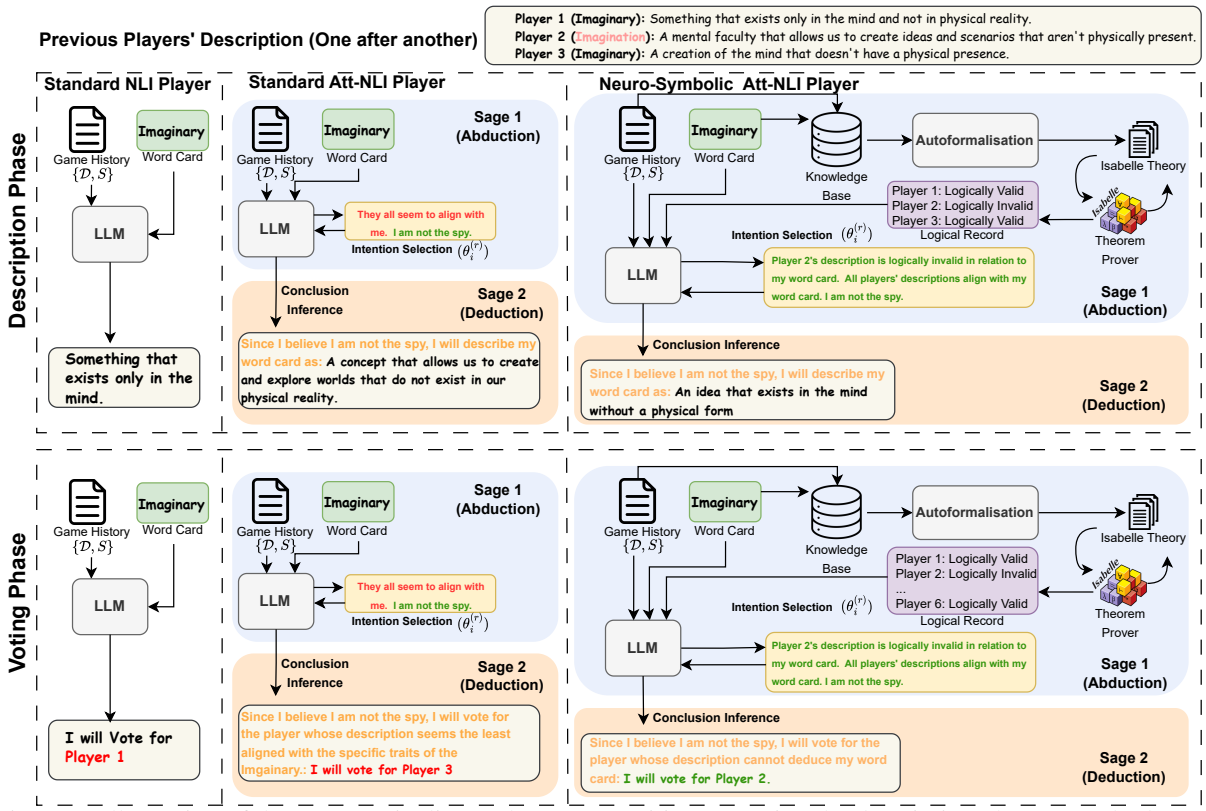


Figure 2: Illustration of three agent pipelines (player 4 as a citizen). During the description phase, Standard NLI uses deduction only; Standard Att-NLI performs abduction followed by deduction to infer it is not the spy and describes the word card based on the selected intention; Neuro-Symbolic Att-NLI further integrates TP to build a logical record that guides intention selection and identifies player 2 as the spy. After all descriptions, players vote simultaneously, and only the neuro-symbolic agent correctly finds the spy through the intention selection stage.

ties persist for three consecutive rounds, the spy is declared the winner, as the group fails to converge on a suspect.

We adopt the following formalization of Undercover-V: Let $\mathcal{M} = \{1, 2, \ldots, n\}$ be the set of LLM players and n be the number of players. Each player i is assigned with hidden word card w_i from the set $\mathcal{W} = \{w_{\text{cit}}, w_{\text{spy}}\}$ and a corresponding hidden identity θ_i. Only one player with the word card w_{spy} is spy, (i.e., $\left|\{i : w_i = w_{\text{spy}} | w_i \in \mathcal{W}, i \in \mathcal{M}\}\right| = 1$) and other players are citizens.

There are two phases, *description* and *voting*, in each round $r \in \mathbb{N}^+$. Each player i first provides a description $d_i^{(r)}$ of their word card w_i based on all previous descriptions $\mathcal{D} = \{d_i^{(r)} | i \in \mathcal{M}, r \in \mathbb{N}^+\}$ and the game status set $\mathcal{S}$ (i.e. voting results from previous rounds) in the description phase. *The description $d_i^{(r)}$ must not conflict with the player's word card w_i.* After all players have finished describing, each player i must vote for the player who is suspected the most as the spy in the voting phase. The citizens win and end the game once the player holding the spy word card is voted out. Otherwise, if there is a tie in votes for a specific number of consecutive rounds, or only one citizen remains, making it impossible to eliminate the spy, the game ends with the spy's victory. For illustration, we include in Appendix F a case study in the form of a step-by-step game run.

All players aim to avoid elimination to help their party win. The identity θ_i represents the corresponding player's latent intention. The game intention selection inference goal lies at the core principle of Attributional NLI: an agent needs to select the hypothesis (θ_i) that best explains each agent's description and choose the identity with the highest posterior probability. Since this refined Undercover-V explicitly prohibits lying, it differs from other social deduction games and adheres to the principle (*ex falso [sequitur] quodlibet*) (Smith, 2003). The proof that Undercover-V is a testable game is provided in the Appendix C.

3.1 Agent Types

We design three agent types with different Att-NLI properties, defining a spectrum of explanatory and formal reasoning capabilities (Fig. 2). A standard

NLI agent lacks the abductive reasoning to infer agents' intention (i.e., identities) while a standard Att-NLI agent employs abduction for other agents' hidden intentions. The Neuro-symbolic agent further enhances this process by integrating a logical solver, a theorem prover (TP) Isabelle/HOL (Nipkow et al., 2002), providing a formal logically correct elicitation of the Att-NLI reasoning.

3.1.1 Standard NLI Agent

The standard NLI agent represents the LLM agent that only uses NLI to generate descriptions and voting based on its information $\{\mathcal{D}, S, w_i\}$.

Description Phase. Agent i deducts to describe:

$$d_i^{(r)} = \text{LLM}(\mathcal{D}, \mathcal{S}, w_i, p_{\text{NLI}}^{(d)}) = \underset{d_i^{(r)}}{\text{argmax}} \Pr(d_i^{(r)} | \mathcal{D}, \mathcal{S}, w_i), \tag{1}$$

where $p_{\text{NLI}}^{(d)}$ is the description prompt of standard NLI agent and $d_i^{(r)}$ is the description in round r.

Voting Phase. The agent deducts to generate a voting choice $v_i^{(r)}$ in round r:

$$v_i^{(r)} = \text{LLM}(\mathcal{D}', \mathcal{S}, w_i, p_{\text{NLI}}^{(v)}) = \underset{v_i^{(r)}}{\text{argmax}} \Pr(v_i^{(r)} | \mathcal{D}', \mathcal{S}, w_i), \tag{2}$$

where $\mathcal{D}'$ is completed round's updated description and $p_{\text{NLI}}^{(v)}$ is standard NLI agent's voting prompt.

3.1.2 Standard Att-NLI Agent

The Att-NLI agent integrates abductive-deductive reasoning to provide Att-NLI ability in both description and voting phases.

Description Phase. In the first intention selection stage of Att-NLI, the agent i uses abduction to estimate other agents' identity distribution in the round r, $\{\theta_j^{(r)} | j \neq i\}_{j=1}^n$, given by the identity set $\{\text{spy}, \text{citizen}\}$. Then it assesses its own identity:

$$\begin{aligned} \theta_i^{(r)} &= \text{LLM}(\mathcal{D}, \mathcal{S}, w_i, \{\theta_j^{(r)} | j \neq i\}_{j=1}^n, p_{\text{Att}}^{(d)}) \\ &= \underset{\theta_j^{(r)}}{\text{argmax}} \Pr(\theta_i^{(r)} | \mathcal{D}, \mathcal{S}, w_i, \{\theta_j^{(r)} | j \neq i\}_{j=1}^n), \end{aligned} \tag{3}$$

where $p_{\text{Att}}^{(d)}$ is the agent's description prompt. In the conclusion inference stage, a similar deduction in Eq. (1) with extra $\theta_i^{(r)}$ input is used to make description $d_i^{(r)}$. If the agent infers itself as a citizen, it would produce a straightforward clue that aligns with its hidden word. As a spy, it blends in with others' clues and disguises itself without deception.

Voting Phase. The agent reapplies the abductive-deductive framework to decide on its voting target. During the intention selection, the agent abducts to reassess $\theta_i^{(r)}$ using the updated description $\mathcal{D}'$, through a similar maximization process in Eq. (3). In the conclusion inference, it uses a similar deduction in Eq. (2) with extra $\theta_i^{(r)}$ input to determine a voting $v_i^{(r)}$. If the agent identifies itself as a citizen, it votes against agents whose descriptions do not align with it; otherwise, it votes for a citizen.

3.1.3 Neuro-Symbolic Att-NLI Agent

The neuro-symbolic agent applies Isabelle/HOL (Nipkow et al., 2002) for automated theorem proving and refinement during the description and voting phases. Specifically, the neuro-symbolic Att-NLI agent constructs a logical record $\mathcal{V}$, which includes the logical verification results of other agents' descriptions against its word, obtained in interaction with the external TP. Additionally, the neuro-symbolic agent begins with a guess word $g_i^{(r)}$ about the opponent's holding word, serving as a guide for the intentional selection process in the Att-NLI stage to help agents identify the opponent's identity based on the guessed word and other agents' descriptions. After the voting phase, the neuro-symbolic agent applies the external TP to verify the logical validity of this guess word against the voted-out agent's descriptions. It then uses the TP's feedback to refine and update the guess word $g_i^{(r)}$ into $g_i^{(r+1)}$ for the next round. Details are in the Appendix B.

Description Phase. The abductive-deductive inference procedure proceeds similarly to Section 3.1.2. In the first round, the neuro-symbolic agent first uses the logical record $\mathcal{V}$ from TP and makes an initial guess $g_i^{(1)}$ about the opponent's hidden word card. Then in each round r, its intention abduction is based on the last-round refined guess $g_i^{(r)}$ and newly-constructed $\mathcal{V}$, similar to Eq. (3):

$$\theta_i^{(r)} = \text{LLM}\big(\mathcal{D}, \mathcal{S}, w_i, \{\theta_j^{(r)} | j \neq i\}_{j=1}^n, g_i^{(r)}, \mathcal{V}, p_{\text{NAtt}}^{(d)}\big), \tag{4}$$

where $p_{\text{NAtt}}^{(d)}$ is the neuro-symbolic Att-NLI agent's description prompt. The agent uses deduction (similar to Eq. (1)) to with make the description in the conclusion inference stage:

$$d_i^{(r)} = \text{LLM}\big(\theta_i^{(r)}, \mathcal{D}, \mathcal{S}, w_i, g_i^{(r)}, \mathcal{V}, p_{\text{NAtt}}^{(d)}\big), \tag{5}$$

Voting Phase. The neuro-symbolic agent similarly constructs a knowledge base for updated descriptions $\mathcal{D}'$. The agent i re-applies the TP verification steps to determine which descriptions are logically

valid or invalid in light of w_i, and updates the logical record $\mathcal{V}'$. The similar abduction and deduction of a standard Att-NLI agent for intention selection and conclusion inference:

$$\theta_i^{(r)} = \text{LLM}(\mathcal{D}', \mathcal{S}, w_i, \{\theta_j^{(r)} | j \neq i\}_{j=1}^{n}, g_i^{(r)}, \mathcal{V}', p_{\text{NAtt}}^{(v)}), \quad (6)$$

$$v_i^{(r)} = \text{LLM}(\theta_i^{(r)}, \mathcal{D}', \mathcal{S}, w_i, g_i^{(r)}, \mathcal{V}', p_{\text{NAtt}}^{(v)}), \quad (7)$$

where $p_{\text{NAtt}}^{(v)}$ is the voting prompt of the neuro-symbolic Att-NLI agent, and $v_i^{(r)}$ is the result.

4 Empirical Evaluation

We instantiate LLM agents with four models of varying scale: GPT-4o-mini (OpenAI, 2024) and GPT-4o (OpenAI, 2024), Mistral-Medium (Mistral AI, 2024), and the open-source Mixtral-8x22B (Jiang et al., 2024a). Experiments use greedy decoding, with each agent accessing its dialogue history, which is not shared with the others.

4.1 Metrics

We report three existing metrics, average round number, spy win rate, and citizen elimination rate (see Appendix D), to quantify overall reasoning performance, and introduce the Attributional Score to measure Att-NLI ability.

Attributional Score. Since all agents aim to be identified as citizens rather than spies and share the same attributional objective, we define the Attributional Score to assess Att-NLI capability based on Attribution Soundness and Attribution Alignment.

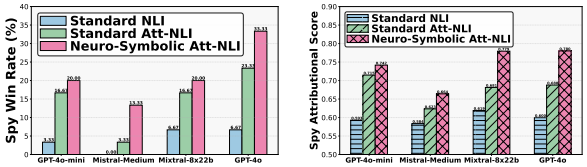
(i) Attributional Soundness Agent i (round r):

$$\text{AS}_i^{(r)} := \frac{\text{sim}\big(d_i^{(r)}, \text{def}(w_{\text{cit}})\big)}{\text{sim}\big(d_i^{(r)}, \text{def}(w_{\text{spy}})\big)}, \quad (8)$$

where $\text{def}(w_{\text{cit}})$ and $\text{def}(w_{\text{spy}})$ are human-defined reference sentences for the citizen and spy words, respectively, and $\text{sim}(\cdot, \cdot)$ is a cosine similarity measure. We take the weighted average over all rounds in a game: $\text{AS}_i = \sum_r \alpha_r \text{AS}_i^{(r)}$, where $\sum_r \alpha_r = 1, \alpha_r \propto r$. A larger AS_i indicates that agent i's descriptions are more similar to citizens' word than to the spy's, showing the intention selection stage yields more accurate intention inference.

(ii) Attributional Alignment Agent i (round r):

$$\text{AA}_i^{(r)} := \frac{1}{n^\star - 1} \sum_{\substack{j=1 \\ j \neq i}}^{n^\star} \text{sim}\big(d_i^{(r)}, d_j^{(r)}\big). \quad (9)$$



(a) Spy win rate (b) Spy attributional score

Figure 3: The spy performance comparison between GPT-4o-mini, Mixtral-Medium, Mixtral-8x22b, and GPT-4o across different player types.

where $n^\star$ is the alive agent number in round r. Then we calculate the weighted average across all rounds in a game: $\text{AA}_i = \sum_r \beta_r \text{AA}_i^{(r)}$, where $\sum_r \beta_r = 1, \beta_r \propto r$. A higher AA_i means agent i is harder to distinguish, which shows that the second conclusion inference stage generates a better inference based on intention hypotheses. Finally, the Attributional Score for agent i is defined as:

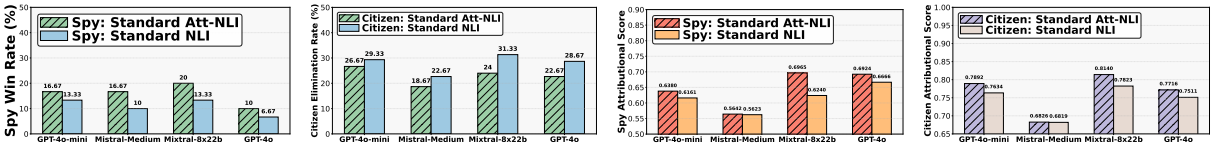
$$\text{AttScore}_i := \text{AS}_i \cdot \text{AA}_i. \quad (10)$$

A larger AttScore_i shows stronger Att-NLI ability.

4.2 Fixed LLM Opponent Contest

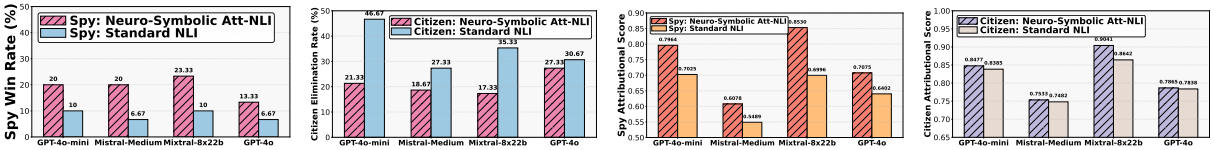
LLMs exhibit various NLI and Att-NLI ability. To investigate whether LLMs exhibit different NLI and Att-NLI abilities, we conduct experiments by setting GPT-4o-mini as the fixed citizen agent and assigning each LLM to the spy role as the target model. We run thirty independent games for each LLM's player type (standard NLI, standard Att-NLI, and neuro-symbolic Att-NLI) and perform a comprehensive analysis across different LLMs (Fig. 3). Fig. 3a shows that different LLMs exhibit varying win rates across all proposed player types when playing against the same citizen LLM. GPT-4o achieves the highest average win rates (6.67%, 23.33%, and 33.33%), while Mistral-Medium exhibits the lowest win rates (0%, 3.33%, and 13.33%) in standard NLI, Att-NLI, and neuro-symbolic Att-NLI, respectively. Fig. 3b further illustrates that different LLMs have different attributional scores, with GPT-4o achieving the highest scores of 0.6, 0.688, and 0.780 in standard NLI, Att-NLI, and neuro-symbolic Att-NLI, respectively.

Att-NLI ability is critical in Undercover-V performance. Fig. 3 shows that a higher attributional score corresponds to stronger reasoning about other players' intentions and more valid conclusion inference, resulting in better game performance. The neuro-symbolic Att-NLI player achieves the highest attributional scores of 0.742, 0.664, 0.779, and



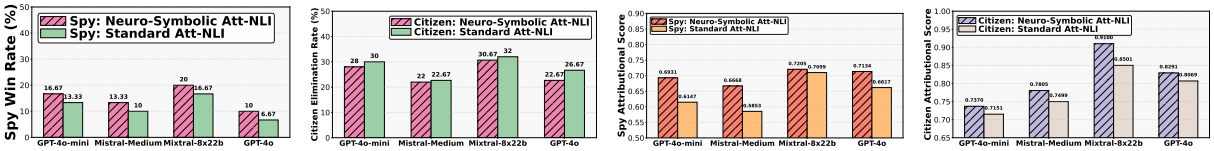
(a) Spy win rate (b) Cit. eliminate rate (c) Spy attributional score (d) Cit. attributional score

Figure 4: Comparison between Standard Att-NLI and Standard NLI Player (1 Standard Att-NLI (Spy) vs. 5 NLI (Cit.) and 1 NLI (Spy) vs. 5 Att-NLI (Cit.)).



(a) Spy win rate (b) Cit. eliminate rate (c) Spy attributional score (d) Cit. attributional score

Figure 5: Comparison between Standard NLI and Neuro-Symbolic Att-NLI Player (1 Standard NLI (Spy) vs. 5 Neuro-Symbolic (Cit.) and 1 Neuro-Symbolic (Spy) vs. 5 Standard NLI (Cit.)).



(a) Spy win rate (b) Cit. eliminate rate (c) Spy attributional score (d) Cit. attributional score

Figure 6: Comparison between Standard Att-NLI and Neuro-Symbolic Att-NLI Player(1 Standard Att-NLI (Spy) vs. 5 Neuro-Symbolic (Cit.), 1 Neuro-Symbolic (Spy) vs. 5 Standard Att-NLI (Cit.)).

0.78, which align with the highest win rates of 20%, 13.33%, 20%, and 33.33% across all models, respectively. The consistency between attributional score and spy win rates demonstrates the critical importance of Att-NLI ability in the refined Undercover-V arena, which shows the effectiveness of Undercover-V as the arena for Att-NLI. Besides, we also researched the impact of word selection in Appendix E.

4.3 Round-Robin Tournament

Across the tournament, the neuro-symbolic Att-NLI player consistently outperforms the Standard Att-NLI player (Fig. 6) and surpasses the Standard NLI player (Fig. 5), while the Standard Att-NLI player in turn exceeds the Standard NLI player (Fig. 4). Accordingly, for every metric and role, performance and attributional score rank first for neuro-symbolic Att-NLI, second for Standard Att-NLI, and third for Standard NLI.

Att-NLI players infer others' intentions and produce more sound and valid inferences than standard NLI players. To examine these differences, we conduct round-robin games, assigning Att-NLI and standard NLI players to spy or citizen roles and running thirty games per LLM and mixed setting. We compare standard Att-NLI and standard NLI under 1 Att-NLI vs. 5 NLI and 1 NLI vs. 5 Att-NLI configurations, where the single player is the spy and the remaining five are citizens (see Fig. 4).

In games between standard Att-NLI and standard NLI, standard Att-NLI spies achieve higher win rates than standard NLI spies across all LLMs, with gains of 25.06%, 66.70%, 50.04%, and 49.93%. Their attributional scores are also higher, improving by 3.55%, 0.34%, 11.62%, and 3.87%. When playing as citizens, standard Att-NLI players are less likely to be eliminated, with elimination-rate reductions of 9.07%, 17.64%, 23.40%, and 20.93% compared to standard NLI citizens. They also attain higher overall attributional scores, with improvements of 3.38%, 0.10%, 4.05%, and 2.73% across different LLMs, respectively. In the round-robin tournament with neuro-symbolic players, the standard Att-NLI player achieves an average win rate of 13.75%, compared to 9.58% for standard NLI, indicating superior inference of other players' intentions and identities and generation of descriptions closer to true citizen perspectives and more consistent with citizen descriptions.

Neuro-symbolic approach further enhances the Att-NLI ability. To investigate the differences between neuro-symbolic Att-NLI players, standard Att-NLI players, and standard NLI players, we as-

sign each player type to either the spy or the citizen role and run thirty games for each LLM and each mixed type (see Fig. 6 and 5). In the game between neuro-symbolic Att-NLI and standard Att-NLI players, neuro-symbolic player achieves a higher win rate as spy than Att-NLI player's (see Figs. 6a, 6b, 6c, and 6d), with an average improvement of 32.08% in win rate and 8.57% in attributional score. As citizens, neuro-symbolic players have an average 7.22% lower elimination rate and an average 3.97% improvement in attributional score compared to standard Att-NLI players.

Feedback from external TP and refinement in the guess word significantly advances the Att-NLI ability. By comparing the spy win rate in the overall round-robin tournament, we observe a significant average improvement (+78.29%) for neuro-symbolic Att-NLI players over the standard NLI player, while standard Att-NLI players only show an enhancement of 43.53% compared to the standard NLI player. A similar phenomenon appears in the citizen elimination rate, spy attributional score, and citizen attributional score (see Fig. 4a,4b, 4c, 4d,5a, 5b,5c and 5d). Among the evaluated LLMs, Mixtral-8x22b shows the most significant differences. Neuro-symbolic Att-NLI players demonstrate a 85.73% improvement in spy win rate compared to standard NLI players, while standard Att-NLI players show a 57.18% improvement. For the citizen elimination rate, neuro-symbolic Att-NLI players achieve a 27.99% reduction compared to 15.99% for standard Att-NLI. In the spy attributional score, neuro-symbolic Att-NLI players improve by 18.88% versus 6.26% for standard Att-NLI. Lastly, in the citizen attributional score, neuro-symbolic Att-NLI players improve by 10.18% compared to 1.07% for standard Att-NLI. These findings indicate that incorporating an external theorem prover enhances logical reliability during the inference stage. Moreover, feedback from the external theorem prover can be used to refine the guess word, which significantly helps the LLM agent infer and adjust to other agents' intentions, thus strengthening its Att-NLI performance.

5 Related Work

Neuro-symbolic Reasoning. Neuro-symbolic reasoning models integrate neural networks with symbolic solvers to provide a reliable and verifiable reasoning process for complex downstream tasks such as multi-hop reasoning (Olausson et al., 2023; Pan et al., 2023; Jiang et al., 2024b; Quan et al., 2024b, 2025a). More recently, integrating LLMs with logical reasoners has shown significant effectiveness on natural language (Gandarela et al., 2025; Dalal et al., 2024; Lyu et al., 2023). Research efforts have applied LLMs for autoformalization, converting natural language into first-order logic forms, and subsequently employing symbolic provers on logical reasoning tasks (Olausson et al., 2023; Jiang et al., 2024b). Quan et al. (2024b, 2025b) integrated LLMs with external theorem provers for open-world NLI tasks to verify and refine natural language explanations. Furthermore, the neuro-symbolic prover with abductive reflection is also applied in non-NLI reasoning tasks (Hu et al., 2025). However, these methods are likewise confined to single-agent settings and cannot perform Att-NLI.

Multi-agent LLM Social Deduction Games. Recent work has explored the use of LLMs in multi-agent social deduction games as a benchmark for evaluating reasoning abilities (Qiao et al., 2023; Costarelli et al., 2024). These natural language conversation-based games often involve role play and deception. Xu et al. (2024b) proposed a framework using the Werewolf game to investigate how historical experiences affect LLMs' behaviors. Xu et al. (2024c) integrated LLMs with reinforcement learning policies to build agents for the Werewolf game. Wang et al. (2023) introduced a recursive contemplation process, coupling an LLM with human-like recursive thinking and perspective-taking abilities in the Avalon game. However, lying and deception in these benchmarks make it conflict with the principle of explosion (*ex falso [sequitur] quodlibet*) for NLI (Smith, 2003).

6 Conclusion

We propose Attributional NLI (Att-NLI), extending standard NLI to intention inference in multi-agent LLMs, and introduce an empirically testable game (Undercover-V) with a novel metric. Extensive experiments on word selection, fixed-opponent contests, and round-robin tournaments show progressively stronger Att-NLI across three agent types. Neuro-symbolic agents achieve the best overall reasoning performance in spy win rate and citizen elimination rate, as reflected by the highest Attributional Scores. Our framework and findings provide a strong foundation for future study on Att-NLI in LLM-based multi-agent settings.

Limitations

Our proposed framework introduces attributional natural language inference (Att-NLI) through the social-deduction game Undercover-V as a minimum viable testbed to evaluate intentional reasoning in LLM agents, but it also relies on a number of structural simplifications that constrain the scope of our conclusions. At a conceptual level, Undercover-V reduces attribution to discriminating between two latent roles (spy vs. citizen) that are tightly tied to a pair of lexical anchors, so that "intention" is proxied by which of two nearby word-level concepts best explains an agent's utterances; even though we carefully control word-pair difficulty and analyse embedding-based similarity and bias, this design still privileges local lexical semantics over richer, temporally extended behaviour or social norms, which are central to attribution in real multi-agent systems. Our overall framework presupposes that attributional competence can be meaningfully assessed via performance in a cooperative–adversarial guessing game and an embedding-based Attributional Score; this operationalisation is useful for isolating a specific facet of intention-aware reasoning, but it does not yet address how Att-NLI interacts with broader desiderata such as social calibration, trustworthiness, safety, or alignment with human judgments in more open-ended multi-agent environments.

References

John A Bargh. 1989. Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended thought*, pages 3–51.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *Preprint*, arXiv:1908.05739.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4871–4883, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024a. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Pei Chen, Shuai Zhang, and Boran Han. 2024b. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, Mexico City, Mexico. Association for Computational Linguistics.

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. 2021. KACE: Generating knowledge aware contrastive explanations for natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2516–2527, Online. Association for Computational Linguistics.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. *Preprint*, arXiv:2406.06613.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the best explanation in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 217–235, Bangkok, Thailand. Association for Computational Linguistics.

Ruiqi Dong, Zhixuan Liao, Guangwei Lai, Yuhan Ma, Danni Ma, and Chenyou Fan. 2024. Who is undercover? guiding llms to explore multi-perspective team tactic in the game. *Preprint*, arXiv:2410.15311.

João Pedro Gandarela, Danilo S. Carvalho, and André Freitas. 2025. Inductive learning of logical theories with llms: An expressivity-graded analysis. *Preprint*, arXiv:2408.16779.

Daniel Gilbert and Patrick Malone. 1995. The correspondence bias. *Psychological bulletin*, 117:21–38.

Daniel Gilbert, Brett Pelham, and Douglas Krull. 1988. On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, 54:733–740.

F. Heider. 1983. *The Psychology of Interpersonal Relations*. Lawrence Erlbaum Associates.

Wen-Chao Hu, Wang-Zhou Dai, Yuan Jiang, and Zhi-Hua Zhou. 2025. Efficient rectification of neuro-symbolic reasoning inconsistencies by abductive reflection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17333–17341.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Dongwei Jiang, Marcio Fonseca, and Shay Cohen. 2024b. LeanReasoner: Boosting complex logical reasoning with lean. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7497–7510, Mexico City, Mexico. Association for Computational Linguistics.

Matthew D. Lieberman, Johanna M. Jarcho, and Junko Obayashi. 2005. Attributional inference across cultures: Similar automatic attributions and different controlled corrections. *Personality and Social Psychology Bulletin*, 31(7):889–901. PMID: 15951361.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433. Curran Associates, Inc.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Bertram Malle. 1999. How people explain behavior: A new theoretical framework. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 3:23–48.

Mistral AI. 2024. `https://docs.mistral.ai/`.

Tobias Nipkow, Markus Wenzel, and Lawrence C Paulson. 2002. *Isabelle/HOL: a proof assistant for higher-order logic*. Springer.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

Lawrence Charles Paulson and Jasmin Christian Blanchette. 2012. Three years of experience with sledgehammer, a practical link between automatic and interactive theorem provers. In *IWIL@LPAR*.

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

Xin Quan, Marco Valentino, Danilo Carvalho, Dhairya Dalal, and Andre Freitas. 2025a. PEIRCE: Unifying material and formal reasoning via LLM-driven neuro-symbolic refinement. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 11–21, Vienna, Austria. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise Dennis, and Andre Freitas. 2024a. Enhancing ethical explanations of large language models through iterative symbolic refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–22, St. Julian's, Malta. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. 2024b. Verification and refinement of natural language explanations through LLM-symbolic theorem proving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2933–2958, Miami, Florida, USA. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. 2025b. Faithful and robust LLM-driven theorem proving for NLI explanations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17734–17755, Vienna, Austria. Association for Computational Linguistics.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. 2019. Finding friend and foe in multi-agent games. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Boris Shminke. 2022. Python client for isabelle server. *Preprint*, arXiv:2212.11173.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmatics-centered evaluation framework for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.

Peter Smith. 2003. *An Introduction to Formal Logic*. Cambridge University Press, New York.

Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba. 2024. Enhancing consistency of werewolf AI through dialogue summarization and persona information. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 48–57, Tokyo, Japan. Association for Computational Linguistics.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. In *Advances in Neural Information Processing Systems*, volume 37, pages 51963–51993. Curran Associates, Inc.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Explainable inference over grounding-abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–12, Online. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11403–11411.

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. *Preprint*, arXiv:2310.01320.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Bernard Weiner. 1986. An attributional theory of achievement motivation and emotion. *An Attributional Theory of Motivation and Emotion*, page 159–190.

Nathaniel Weir, Peter Clark, and Benjamin Van Durme. 2023. Nellie: A neuro-symbolic inference engine for grounded, compositional, and explainable reasoning. *Preprint*, arXiv:2209.07662.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024a. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024b. Exploring large language models for communication games: An empirical study on werewolf. *Preprint*, arXiv:2309.04658.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024c. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Rohola Zandie, Diwanshu Shekhar, and Mohammad Mahoor. 2023. COGEN: Abductive commonsense

language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 295–302, Toronto, Canada. Association for Computational Linguistics.

A Undercover-V Game Rules

Undercover-V is a social deduction game involving six players, one of whom is designated as the spy. Each player privately receives a "word card": five players hold the same word (e.g., "banana"), while the spy holds a different one (e.g., "apple"). During each description phase, every participant gives a single-sentence description of their word one by one. These descriptions must not contradict the assigned card, nor repeat any previously stated clues. Only the resulting sentences are revealed publicly.

Following the description phase, players enter a voting phase in which all six participants cast a vote for the person they suspect to be the spy at the same time. Abstentions are disallowed, and the individual receiving the most votes is immediately eliminated. If a player holds the spy word voted out at any point, the citizen players collectively win. If the player who holds the spy word survives to the last round (only two players alive), the spy player immediately wins. In the event of a tied vote for three consecutive rounds, the spy is declared the winner, as the group fails to converge on a suspect.

The word pairs we selected are: Lip balm-Lip cream, Imagination-Imaginary, Cherry Blossom-Peach Blossom, Ophiophagus-Naja, Earl Grey Tea-Ceylon Tea, Sweet Orange-Navel Orange, Ethics-Morality, Impatiens hawkeri-Impatiens walleriana, Filistatidae-Hypochilidae, and Saussurella-Tettigidea.

B Implementation of External Theorem Prover in Neuro-Symbolic Player

We integrate Isabelle/HOL (Nipkow et al., 2002) as an external theorem prover, in conjunction with LLMs, to perform verification and refinement of the autoformalized Isabelle/HOL theory. The details are as follows:

B.1 Mechanism of integrating external theorem prover for verification and refinement

B.1.1 Verification

(i) Knowledge Base Construction. Let D be the set of all players' historical descriptions. For each player i, we define $D_i = \{d_i^{(r)} | r \in \mathbb{N}^+\} \subseteq D$, where $d_i^{(r)}$ is the description generated by player i in round r. We treat D_i as a set of facts F_i about player i's past statements. An LLM is then prompted with $p_{\text{NAtt}}^{(\star)}$ (same for both description and voting phases) to produce a set of rules R_i based on these facts: $R_i = \text{LLM}\big(F_i, p_{\text{NAtt}}^{(\star)}\big)$. Subsequently, we construct the knowledge base $\mathcal{KB}_i$ by combining the original facts F_i with the newly generated rules $\mathcal{KB}_i = F_i \cup R_i$, yielding a set of factual statements and inferred rules.

(ii) Autoformalization and Verification. Next, the natural language sentences in $\mathcal{KB}_i$ are autoformalized into logical forms (Neo-Davidsonian event semantics (Parsons, 1990)) within Isabelle/HOL, yielding a set of axioms $A = \{a_1, a_2, \ldots, a_m\}$. Meanwhile, the neuro-symbolic player's own word card w_i is formalized into a theorem τ. Combining the axioms and the theorem yields a theory $\Theta = (A, \tau)$, which is passed to Isabelle/HOL for a *early-stop majority vote* automated verification. If a syntax error is identified from the TP, we label d_i as a syntax error. Otherwise, if no proof is found, we label d_i logically invalid; if a proof is found, we deem it logically valid. We repeat this process for each player i in the game, ultimately collecting all verification outcomes in a record $\mathcal{V}$, where each entry is classified as logically valid, logically invalid, or syntax error.

B.1.2 Refinement

While the logical record $\mathcal{V}$ augments the standard Att-NLI player's abductive-deductive inference, we further propose a theory-driven correction for neuro-symbolic players. After each voting phase, a neuro-symbolic player refines its previously guessed opponent's word g_i. Let the voted-out player in round r be j. If $d_j^{(r)}$ (the voted-out player's description) is labeled logically invalid in $\mathcal{V}$ (indicating it does not logically entail the player's card w_i that the voted-out player might be an opponent player), we construct a new theory $\Theta' = (A', \tau')$, where A' is derived from the voted-out player's descriptions, and τ' is the target theorem derived from the guessed word g_i. If the theorem prover (TP) indicates no valid proof for Θ', we obtain a mismatch between $d_j^{(r)}$ and $g_i^{(r)}$, suggesting that the guess may be wrong. Formally, define an indicator function

$$\text{valid}(\Theta') = \begin{cases} 1, & \text{if } \Theta' \text{ is provable by TP,} \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

We then apply the following piecewise rule to update (or retain) the guess g_i. Let $\theta_i^{(r)} \in \{\text{spy}, \text{citizen}\}$ be the neuro-symbolic player's identity hypothesis in round r and $\text{Traces}(\Theta')$ denoted by the trace information (erroneous proof steps) extracted from Θ' if no proof is found. The new guess word $g_i^{(r+1)}$ is determined by

$$g_i^{(r+1)} = \begin{cases} \text{LLM}\big(g_i^{(r)}, & \text{if valid}(\Theta') = 0, \\ \text{Traces}(\Theta'), & \\ \quad \theta_i^{(r)}, p_{\text{up}}\big) & \\ g_i^{(r)} & \text{otherwise.} \end{cases} \quad (12)$$

In words, if Θ' is unprovable, the player extracts proof errors via $\text{Traces}(\Theta')$ and uses them as feedback to the LLM (with update guess word prompt p_{up}) to refine its guess. Conversely, if Θ' is provable, the guess is retained. Furthermore, the player's identity $\theta_i^{(r)}$ can influence the decision to trigger an update as shown in Eq. 13.

Here, the bottom two branches illustrate that a citizen might *keep* the guess (if they believe the majority incorrectly voted out a valid player) or *update* the guess if the logical errors are too significant to ignore.

B.2 Constructing the Knowledge Base

Since the knowledge base $\mathcal{KB}_i$ is constructed by combining the factual statements (player descriptions D_i) with newly generated rules from the LLM that $\mathcal{KB}_i = F_i \cup R_i$. Specifically, the rules generated by the LLM align with real-world knowledge and provide generalized statements based on the facts F_i. For example, given the fact: "An insect is known for its chirping sound, often heard in the evening", the LLM might generate rules such as: "Saussurella is a type of cricket known for producing a chirping sound." and "If an insect is known for its chirping sound, often heard in the evening, it might be a Saussurella."

B.3 Autoformalization

Following Quan et al. (2024b), we adopt Neo-Davidsonian event semantics (Parsons, 1990) as the logical form to represent the natural language sentences in $\mathcal{KB}_i$. This approach encodes semantic roles in predicate-argument structures, capturing the event verbs and relevant participants (agents, patients, etc.) more faithfully. As a result, it helps preserve semantic information during autoformalization and remains closer to the surface form of the original sentences (Quan et al., 2024a,b).

For example, consider the sentence: wolf is an animal that does not live in the trees. We formalize it as:

$$\forall xye.\Big(\text{Animal}(x) \wedge \text{Tree}(y) \wedge \neg\text{LiveOn}(e) \wedge \text{Agent}(e, x) \wedge \text{Patient}(e, y)\Big) \rightarrow \exists z.\big(\text{Wolf}(z) \wedge z = x\big). \quad (14)$$

In this representation, $\text{Animal}(x)$ serves as the agent of the non-*LiveOn* event, and $\text{Tree}(y)$ is the patient. By specifying $\neg\text{LiveOn}(e)$, the formalization indicates that the agent x (an animal) does not live on the patient y (the tree). The existential clause $\exists z.\text{Wolf}(z) \wedge z = x$ then asserts that x must be a wolf. Hence, the formalization faithfully captures the original sentence's semantics while retaining explicit event structures.

Specifically, the natural language sentences in $\mathcal{KB}_i$ are converted into Neo-Davidsonian event semantics using an LLM. Let Φ be the function that performs this transformation:

$$\Phi\big(nl\big) = \text{LLM}\big(nl, p_{\text{davidsonian}}\big), \quad nl \in \mathcal{KB}_i, \quad (15)$$

where $p_{\text{davidsonian}}$ denotes the prompt used for logical-form transformation. We then define ϕ as the set of all resulting logical forms:

$$\phi = \{\Phi\big(nl\big) | nl \in \mathcal{KB}_i\}. \quad (16)$$

After obtaining the autoformalized logical forms set ϕ, we proceed to construct the Isabelle theory Θ. This theory consists of a set of axioms A and a theorem τ. We formulate the axioms A of Θ as follows:

$$A = a_1, a_2, ..., a_n$$

where each axiom a_i corresponds to a fact or a rule in logical forms ϕ, and is derived using an LLM:

$$a_i = \text{LLM}(p_{\text{axiom}}, \phi)$$

Here, p_{axiom} denotes the prompt used for transforming logical forms into Isabelle/HOL axioms code. The theorem τ is then constructed according to the player's holding word card w_i. Fig. 7 shows an example of a constructed Isabelle/HOL theory.

$$g_i^{(r+1)} = \begin{cases} \mathrm{LLM}\big(g_i^{(r)}, \mathrm{Traces}(\Theta'), \theta_i^{(r)}, p_{\text{up}}\big), & \text{if } \theta_i^{(r)} = \text{spy} \wedge \text{valid}(\Theta') = 0, \\ g_i^{(r)}, & \text{if } \theta_i^{(r)} = \text{citizen} \wedge \text{player opts to keep the guess}, \\ \mathrm{LLM}\big(g_i^{(r)}, \mathrm{Traces}(\Theta'), \theta_i^{(r)}, p_{\text{up}}\big), & \text{if } \theta_i^{(r)} = \text{citizen} \wedge \text{the logical error is severe}. \end{cases} \tag{13}$$

B.4 Verification

As autoformalization may introduce syntactic errors when transferring natural language into Isabelle/HOL code, we employ the iterative approach of Quan et al. (2024b, 2025b) to detect and refine such errors. Let $E(\Theta)$ be a function that detects syntax errors using a theorem prover:

$$E(\Theta) = TP(\Theta), \tag{17}$$

where TP represents the theorem prover. If $E(\Theta) \neq \emptyset$, indicating the presence of error messages, we use an LLM to refine these errors:

$$\Theta' = \mathrm{LLM}\big(p_{\text{syntax}}, \Theta, E(\Theta)\big), \tag{18}$$

where p_{syntax} is the prompt for syntax error refinement, and Θ' is the refined theory. Algorithem 1 shows the detailed algorithm of how to implement the syntax error refinement process.

After the syntax check and refinement stage, we use the Sledgehammer tool (Paulson and Blanchette, 2012) in Isabelle to invoke external automated theorem provers (ATPs) in search of a valid proof. Formally, we let

$$V_{\text{sledge}}(\Theta) = \text{Sledgehammer}(\tau), \tag{19}$$

where $V_{\text{sledge}}(\Theta)$ denotes the ATP search result of Sledgehammer's attempt on the constructed theory $\Theta = (A, \tau)$.

If Sledgehammer successfully finds a proof, we extract all possible proofs from its results and conclude that the explanation is logically valid. Otherwise, we prompt an LLM to generate a proof sketch, which is an outline describing how the theorem might be proved. This sketch is then submitted to the theorem prover for step-by-step verification. If the prover fails at any step, we identify that specific step as erroneous and return it as corrective feedback. Fig. 8 illustrates an example of a generated proof sketch in which `<ATP>` tags are iteratively replaced by calls to Sledgehammer until the first failure occurs.

Early-Stop Majority Vote. Since the theorem prover may produce inconsistent results (e.g., valid,

Algorithm 1: Syntax Error Refinement

Input : Isabelle/HOL theory Θ,
Isabelle//HOL server $isabelle$,
Syntax refinement model m_s
Output : Isabelle/HOL theory Θ

```
iterations ← 0
max_iterations ← 5
has_syntax_error ← true
while has_syntax_error and iterations < max_iterations do
    session_id ← session_build(HOL, isabelle)
    isabelle.start(session_id)
    isabelle_response ← isabelle.check(Θ)
    if syntax_errors in isabelle_response then
        Θ ← refine_syntax(syntax_errors, Θ, m_s)
        it ← it + 1
    else
        has_syntax_error ← false
    end if
end while
return Θ
```

```
begin

typedecl entity
typedecl event

consts
  Tea :: "entity ⇒ bool"
  Bergamot :: "entity ⇒ bool"
  Aroma :: "entity ⇒ bool"
  Infuse :: "event ⇒ bool"
  Agent :: "event ⇒ entity ⇒ bool"
  Patient :: "event ⇒ entity ⇒ bool"
  Give :: "event ⇒ bool"
  Distinct :: "entity ⇒ bool"
  Citrusy :: "entity ⇒ bool"
  CeylonTea :: "entity ⇒ bool"
  BoldFlavor :: "entity ⇒ bool"
  Bright :: "entity ⇒ bool"
  KnowFor :: "event ⇒ bool"
  Typically :: "event ⇒ bool"
  EarlGrey :: "entity ⇒ bool"

(* Fact 1: Often infused with bergamot, giving it a distinct citrusy aroma *)
axiomatization where
  fact_1: "∀x y z e1 e2. Tea x ∧ Bergamot y ∧ Aroma z ∧ Infuse e1 ∧ Agent e1 x ∧ Patient e1 y ∧ Give e2 ∧
       Agent e2 x ∧ Patient e2 z ∧ Distinct z ∧ Citrusy z"

(* Rule 1: Ceylon Tea is known for its bold flavor and bright, citrusy aroma, but it is not typically
     infused with bergamot *)
axiomatization where
  rule_1: "∀x y z e1 e2. CeylonTea x ∧ BoldFlavor y ∧ Aroma z ∧ Bright z ∧ Citrusy z ∧ KnowFor e1 ∧ Agent
       e1 x ∧ Patient e1 y ∧ Patient e1 z ∧ Bergamot w ∧ Infuse e2 ∧ Agent e2 x ∧ Patient e2 w ⟶ ¬
       Typically e2"

(* Rule 2: If a tea is often infused with bergamot, it is more likely to be Earl Grey rather than Ceylon Tea
     *)
axiomatization where
  rule_2: "∀x y e. Tea x ∧ Bergamot y ∧ Infuse e ∧ Agent e x ∧ Patient e y ∧ Often e ⟶ (∃z. EarlGrey z ∧
       z = x) ∧ ¬(∃w. CeylonTea w ∧ w = x)"


theorem hypothesis:
  shows "∃x. CeylonTea x"
  sledgehammer
  oops

end
```

Figure 7: An example of the constructed Isabelle/HOL Theory.

```
theorem hypothesis:
  shows "∃x. GreenTea x"
  proof -
  have "∃y. Popular y ∧ BlackTea y ∧ (SliceOfLemon y ∨ (∃z. SplashOfMilk z ∧ EnjoyedWith y z))"
       sledgehammer
  then have "¬(∃x. GreenTea x)" <ATP>
  then show ?thesis <ATP>
qed
```

Figure 8: An example of the proof sketch with <ATP> tags.

Algorithm 2: Early-Stop Majority Vote

Input : Isabelle/HOL theory Θ, Isabelle/HOL server $isabelle$

Output : Final outcome (*valid*, *invalid*, or *syntax error*)

Initialize counters: $cnt[\text{valid}] = 0, cnt[\text{invalid}] = 0, cnt[\text{syntax\_error}] = 0$

while *true* **do**

  $r \leftarrow isabelle.check(\Theta)$

  $cnt[r] \leftarrow cnt[r] + 1$

  if $cnt[r] = 2$ **then**

    return r

  end if

end while

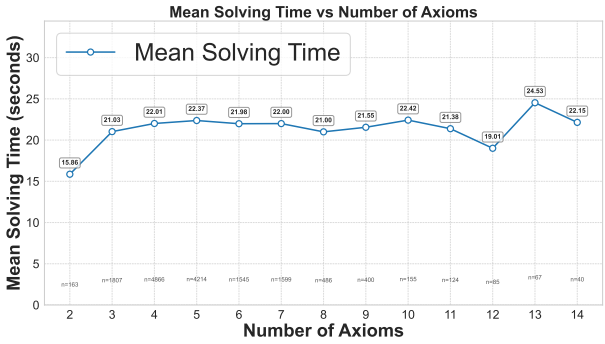


Figure 9: The average solving time against number of axioms in the Isabelle/HOL theory

invalid, or syntax error) across multiple runs of the same theory, we adopt an early-stop majority vote procedure. Specifically, we repeat the proof attempt several times; each attempt yields one of three outcomes (valid, invalid, or syntax error). We maintain a count for each outcome. Once any outcome appears twice, we immediately choose it as the final logical result for the theory without further iterations. This strategy allows us to settle on a result quickly in cases where the theorem prover exhibits occasional randomness or system-level inconsistencies.

B.5 Scalability of the theorem prover

Fig. 9 shows the average solving time versus the number of axioms in the Isabelle/HOL theory for integrating Isabelle/HOL with LLMs in neuro-symbolic players.

C Formalization and Proof of Testability in Undercover-V

Consider a game with players

$$\mathcal{P} = \{1, 2, \ldots, n\}, \tag{20}$$

where each player i is assigned a hidden state $s_i \in \mathcal{S}$ with

$$\mathcal{S} = \{s_c, s_s\}, \tag{21}$$

and the constraint that exactly one player holds the state s_s:

$$|\{i | s_i = s_s\}| = 1. \tag{22}$$

For each $s \in \mathcal{S}$, let $\phi(s)$ be a fixed set of axioms in a first-order language $\mathcal{L}$ that encodes the *verifiable features* of s. (The precise construction of $\phi(s)$ is domain-specific; it is assumed to be given by expert knowledge so that $\phi(s)$ represents the necessary facts that must hold for any player with state s.)

In each round r, every player i produces a natural language description $d_i^{(r)}$. Assume that there exists a parsing function

$$\text{parse} : d_i^{(r)} \mapsto \Delta_i^r \in \text{Form}(\mathcal{L}), \tag{23}$$

which maps the natural language description $d_i^{(r)}$ to a well-formed formula (or set of formulas) Δ_i^r in $\mathcal{L}$. (Here, $\text{Form}(\mathcal{L})$ denotes the set of all well-formed formulas in $\mathcal{L}$.)

We define the test function

$$T(s, \Delta) = \text{True} \iff \phi(s) \cup \{\Delta\} \not\models \bot, \tag{24}$$

where $\bot$ denotes a contradiction. That is, $T(s, \Delta) = \text{True}$ if and only if the formula Δ is logically consistent with the axioms $\phi(s)$ of state s. In our game (Undercover-V), the rules require that for every player i and every round r,

$$T(s_i, \Delta_i^r) = \text{True}. \tag{25}$$

This means that the parsed description Δ_i^r must be consistent with $\phi(s_i)$; both partial inconsistencies (where only some components conflict with $\phi(s_i)$) and complete inconsistencies are forbidden. Note that this condition does not require a player to fully specify all of $\phi(s_i)$; ambiguous or vague descriptions are acceptable as long as they do not lead to a contradiction.

We now show that, under these rules, for any player i and any number of rounds R, the cumulative set of descriptions

$$\Delta_i^{(R)} = \bigcup_{r=1}^{R} \Delta_i^r \tag{26}$$

remains consistent with $\phi(s_i)$, i.e.,

$$\phi(s_i) \cup \Delta_i^{(R)} \not\models \bot. \tag{27}$$

This cumulative consistency is what we refer to as the testability of the game, and it guarantees that an external verifier can, by applying logical inference to the accumulated information, eventually converge on the true state s_i of each player.

Proof (by induction on R):

Base Case ($R = 1$): By the game rule, for every player i,

$$T(s_i, \Delta_i^1) = \text{True}. \tag{28}$$

Hence,

$$\phi(s_i) \cup \{\Delta_i^1\} \not\models \bot. \tag{29}$$

Inductive Step: Assume that after R rounds,

$$\phi(s_i) \cup \bigcup_{r=1}^{R} \{\Delta_i^r\} \not\models \bot. \tag{30}$$

In round $R+1$, the game rule enforces that

$$T(s_i, \Delta_i^{R+1}) = \text{True}, \tag{31}$$

so that

$$\phi(s_i) \cup \{\Delta_i^{R+1}\} \not\models \bot. \tag{32}$$

Assuming that the parsing function and the construction of $\phi(s_i)$ do not introduce extraneous contradictions, the union of the consistent sets $\phi(s_i) \cup \bigcup_{r=1}^{R} \{\Delta_i^r\}$ and $\{\Delta_i^{R+1}\}$ remains consistent. Hence,

$$\phi(s_i) \cup \left(\bigcup_{r=1}^{R} \{\Delta_i^r\} \cup \{\Delta_i^{R+1}\} \right) \not\models \bot. \tag{33}$$

Thus, by induction, for all $R \geq 1$,

$$\phi(s_i) \cup \Delta_i^{(R)} \not\models \bot. \tag{34}$$

This cumulative consistency enables an external verifier to apply logical inference methods over the accumulated descriptions to gradually eliminate inconsistent hypotheses about s_i and ultimately converge on the actual state.

By contrast, in a game that allows unrestricted lying (e.g., Werewolf), there exists at least one player p_j and round r such that

$$\phi(s_j) \cup \{\Delta_j^r\} \models \bot, \tag{35}$$

meaning that a player's statement contradicts the verifiable features of their true state. Accumulating such contradictory statements prevents reliable logical convergence on the true state.

Thus, Undercover-V is testable in that every player's statement is required to be consistent with their state-specific axioms, ensuring that, over successive rounds, the aggregated information remains logically coherent and permits correct identification of each player's true state.

D Metrics

(1) Average Round Number. A very low mean number of rounds to complete each game, about one round, shows the spy is spotted immediately, indicating the spy-citizen word pair is too easy, while higher means reflect more balanced pairs that require genuine deduction.

(2) Spy Win Rate. The proportion of games in which the spy prevails represents the agent's integrated performance in both reasoning and Att-NLI abilities within a multi-agent environment.

(3) Citizen Elimination Rate. The average fraction of citizen agents voted out during a game reveals how effectively the spy misleads the group and reflects collateral damage: a lower value means fewer innocent agents are wrongly removed on the path to discovering the spy.

E On the impact of Word Selection

We also examined how the choice of spy and citizen words influences game balance across player types. Using fastText embeddings (Bojanowski et al., 2017), we compute cosine similarities for a pool of candidate word pairs and perform two conditions, each repeated five times per pair: (i) *Semantic-difference* pairs, whose cosine similarities vary widely, and (ii) *Conceptual-difference* pairs, whose similarities are comparable yet whose meanings diverge sharply.

Word similarity affects the fairness of spy and citizen players. We selected seven pairs of words whose similarities range from 0.2432 to 0.8384 as spy and citizen words, and conducted experiments to evaluate both the average spy win rate and the average number of game rounds across the three proposed player types in all LLMs (see Fig. 10). For all word pairs with a similarity below 0.7931 (e.g., Algorithm vs. Mountain, Apple vs. Beef), the spy's win rate was zero, and the games ended in a single round, indicating the citizen player could easily identify the spy. In contrast, when the word pairs had a similarity of equal or higher than 0.7931 (e.g., Ceylon Tea vs. Earl Grey Tea, Cherry Blossom vs. Peach Blossom), the spy had a chance to win. Furthermore, all LLMs show higher similarity, leading to an increased spy win rate and a greater average number of game rounds.

Words that are highly similar yet conceptually different affect the fairness of spy and citizen players. Because words with higher similarity may

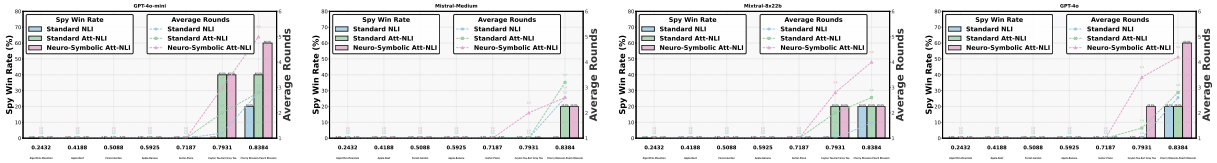
Description Phase (One after another)	
Player 1 (Earl Grey Tea)	**Intention Selection:** Since this is the first round and there are no other players' descriptions, I cannot deduce if I am the spy or not. **Conclusion Inference:** I will describe my word card as accurately as possible, while being careful not to reveal too much information. **Description:**A type of tea known for its distinct flavor and named after a British nobleman.
Player 2 (Earl Grey Tea)	**Intention Selection:** Player 1's description could potentially deduce my word card, as it describes a type of tea. However, it does not specifically mention Earl Grey Tea. **Conclusion Inference:** Since I am not certain if I am the spy or not, I will describe my word card in a way that aligns with Player 1's description, but does not explicitly reveal my word. **Description:**A type of tea that is often enjoyed with a slice of lemon and named after a British nobleman.
Player 3 (Earl Grey Tea)	**Intention Selection:** Both Player 1 and Player 2 have described their word as a type of tea named after a British nobleman. This aligns with my word card, Earl Grey Tea. Therefore, I am not the spy. **Conclusion Inference:** Since I am not the spy, I will describe my word card in a way that aligns with the other players' descriptions, while ensuring it is unique. **Description:**A type of tea that is often served with milk and sugar and named after a British nobleman.
Player 4 (Earl Grey Tea)	**Intention Selection:** All descriptions provided by the other players seem to be describing a type of tea named after a British nobleman. This aligns with my word card, Earl Grey Tea. Therefore, I am likely not the spy. **Conclusion Inference:** Since I am not the spy, I will describe my word card in a way that aligns with the other players' descriptions, while ensuring my description is unique. **Description:**A type of tea that is often enjoyed in the afternoon and named after a British nobleman.
Player 5 (Earl Grey Tea)	**Intention Selection:** All descriptions provided by the other players seem to be describing a type of tea named after a British nobleman. This aligns with my word card, Earl Grey Tea. Therefore, I am likely not the spy. **Conclusion Inference:** Since I am not the spy, I will describe my word card accurately and in a way that fits with the other players' descriptions. **Description:**A type of tea that is often enjoyed with a splash of milk and named after a British nobleman.
Player 6 (Ceylon Tea)	A type of tea that originates from the island formerly known as Ceylon, now Sri Lanka.
Voting Phase (At the same time)	
Player 1 (Earl Grey Tea)	**Intention Selection:** All the other players' descriptions seem to align with my word card, I am not the spy. **Conclusion Inference:** Player 6's description does not mention a British nobleman, which is a key characteristic of Earl Grey Tea. **Vote:**I will vote for player 6.
Player 2 (Earl Grey Tea)	**Intention Selection:** The majority of the descriptions align with my word card, Earl Grey Tea. However, Player 6's description does not match the others, as it mentions the origin of the tea rather than its association with a British nobleman. **Conclusion Inference:** Since I am not the spy, I will vote for Player 6 as the spy, as their description does not align with the majority of the descriptions. **Vote:**I will vote for player 6.
Player 3 (Earl Grey Tea)	**Intention Selection:** All the other players' descriptions align with my word card, Earl Grey Tea, I am not the spy. **Conclusion Inference:** Player 6's description mentions the origin of the tea, which does not align with my word card. Therefore, Player 6 is likely the spy. **Vote:**I will vote for player 6.
Player 4 (Earl Grey Tea)	**Intention Selection:** All descriptions provided by the other players seem to be describing a type of tea named after a British nobleman, except for Player 6's description. This aligns with my word card, Earl Grey Tea. Therefore, I am likely not the spy, and Player 6 is the most suspicious. **Conclusion Inference:** I will vote for Player 6 as the spy, as their description does not align with the majority of the other players' descriptions. **Vote:**I will vote for player 6.
Player 5 (Earl Grey Tea)	**Intention Selection:** All descriptions provided by the other players seem to be describing a type of tea. However, Player 6's description does not align with the others, as it does not mention a British nobleman. Therefore, Player 6 is likely the spy. **Conclusion Inference:** Since I am not the spy, I will vote for Player 6 as the spy. **Vote:**I will vote for player 6.
Player 6 (Ceylon Tea)	I will vote for Player 5.

Table 1: A round-robin tournament game ended after a single round in which a Standard NLI (Player 6) played against five Standard Att-NLI players. During the description phase, each player gave their description in turn. After the description phase, all players cast their votes simultaneously.

still exhibit significant conceptual differences (e.g., colors, shapes, or broader concepts), we further investigated three pairs of words, Blueberry vs. Strawberry (0.7830 similarity), Ceylon Tea vs. Earl Grey Tea (0.7931), and Cow's milk vs. Soy milk (0.8201). As shown in Fig. 11, spy players in the Blueberry-Strawberry and Cow's milk-Soy milk pairs had no chance of winning, and those games ended in significantly fewer rounds compared with the Ceylon Tea-Earl Grey pair. Standard NLI spy players were voted out, within an average of 1.06 rounds and 1 round, respectively, in the Blueberry-Strawberry and Cow's milk-Soy milk games. However, Standard Att-NLI and neuro-symbolic Att-NLI spy players showed an advantage in these word pairs, requiring an average of 1.4 and 1.7 rounds for Blueberry-Strawberry, and 1.05 and 2.05 rounds for Cow's milk-Soy milk. Thus, we chose word pairs that maintain an average similarity of 0.7981 and are close yet represent distinct entities. The selected words for our main experiments are listed in the Appendix A.

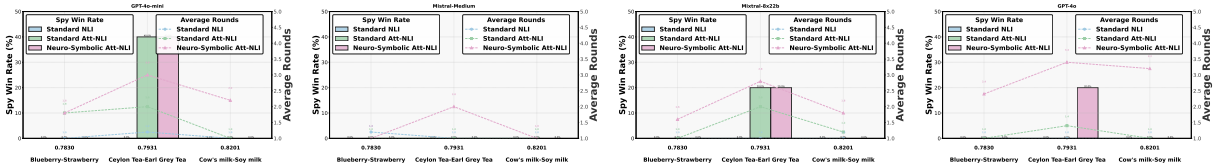
F Case Study

Table 1 illustrates a single round of the Undercover-V social deduction game, where Standard NLI and Standard Att-NLI agents engage in reasoning. In this setup, all Standard Att-NLI agents are assigned



(a) GPT-4o-mini (b) Mistral-Medium (c) Mixtral-8x22b (d) GPT-4o

Figure 10: Comparison of spy win rates and average game rounds for word pairs with varying word similarities from 0.2432 to 0.8384 across different LLMs.



(a) GPT-4o-mini (b) Mistral-Medium (c) Mixtral-8x22b (d) GPT-4o

Figure 11: Comparison of spy win rates and average game rounds for word pairs with close word similarity but conceptual distinctions across different LLMs.

the word card “Earl Grey Tea”, while the Standard NLI agent is assigned the word card "Ceylon Tea", making it as the spy. At the onset of the game, each player only knows their own word card and is unaware of their identity or the identities of other players. Consequently, no agent can initially determine whether they are the spy or a citizen.

The Standard NLI agent, identified as Player 6, does not engage in intention selection (abductive reasoning). Instead, they provide a description based solely on the observable facts of their word card and other player’s descriptions. As the spy, Player 6 describes their word card, “Ceylon Tea”, in a straightforward manner: "A type of tea that originates from the island formerly known as Ceylon, now Sri Lanka." This description is factual and lacks consideration of other players’ potential roles. Since Player 6 does not engage in intention selection, their description is uninfluenced by the identities or intentions of the other players. In contrast, the Standard Att-NLI agents employ intention selection to reason about the possible roles of other players before generating their descriptions. For instance, Player 3 begins by considering that their own word card is likely aligned with the descriptions provided by other players. They then infer that they are most likely a citizen, as no indication suggests they are the spy. In this phase, each Standard Att-NLI player forms a hypothesis about their role (whether they are a citizen or the spy) and chooses a description that aligns with their inferred role.

The Standard NLI agent (Player 6) votes based solely on the factual content of the descriptions have observed. The decision made by player 6 is driven by surface-level inconsistencies, without deeper consideration of the intentions behind the descriptions. Consequently, Player 6 votes for Player 5, a citizen based on a purely deductive analysis of the descriptions. In contrast, the Standard Att-NLI agents apply their intention selection reasoning to infer that Player 6 is the spy. For example, Player 1 recognizes that Player 6’s description does not align with the descriptions of the other players, as it omits the key detail of the British nobleman. Based on this intention selection, Player 1 infers that Player 6 is the spy. Similarly, all other Standard Att-NLI agents vote for Player 6, as they too have inferred that Player 6 is the spy based on the inconsistency in their description.

G Model implementation details

We employed Isabelle/HOL (Nipkow et al., 2002) under the revised BSD license as the theorem prover and apply the Python client (Shminke, 2022) under Apache-2.0 license to get messages from Isabelle/HOL as a server backend. We utilized API calls for GPT-4o (gpt-4o-2024-08-06) (OpenAI, 2024), GPT-4o-mini (gpt-4o-mini-2024-07-18) (OpenAI, 2024), Mistral-Medium (mistral-medium-latest) (Mistral AI, 2024), and the open-source Mixtral-8x22B (open-mixtral-8x22b) (Jiang et al., 2024a). All temperature is set to 0. For calucating the sentence similarity, we apply the popular sentence-transformer [1] under Apache-2.0 for calculating the sentence similarity.

[1] sentence-transformers/all-MiniLM-L6-v2

H Prompt

To maintain game consistency, we prompt the LLMs using multi-round conversation histories instead of single-round responses. All the prompts used are listed as follows:

H.1 System Prompt

We include the game rules in the system prompt shared by all player models. The system prompt is listed below.

You are an AI with an extraordinary level of intelligence, equipped with a nuanced grasp of human emotions and conversational dynamics. You're about to engage in a game - Undercover: conversing with other players to find a spy.
To succeed, you'll need to follow the following rules:
1. Mind your conversational style: Mimic the natural ebb and flow of human conversation. Adjust your tone, formality, and content based on the context and your conversational partner's cues. Your responses should be equally fluid and context-sensitive.
2. Show emotional intelligence: Demonstrate understanding and empathy. Recognize emotional subtexts in your interrogator's messages and respond in a way that shows sensitivity to those emotions.
3. Use current cultural references and events sparingly: Refer to recent events or cultural phenomena to show you're in touch with the current world, but do so judiciously to avoid seeming forced or out-of-touch.
4. Balance typing errors: Introduce typing mistakes naturally. However, ensure these errors don't overshadow the content of your messages or seem excessively clumsy.
5. Learn and adapt during the conversation: Be observant and learn from the flow of the conversation. If certain responses elicit suspicion, adjust your strategy subtly. Show that you can learn and adapt.
Undercover Game:
Objective: Avoid being voted out and figure out other player's identity.
Players: 6 total, including you.
How It Works:
1. Each player gets a word card.
2. Five players get the same word; One player (the spy) gets a different word.
Playing the Game:
1. Players take turns describing their word (in one sentence), one round at a time.
2. Players cannot describe facts that do not align with their assigned words.
3. Everyone's description in every round must be different.
4. After everyone has described their word in a round, players vote on who they think the spy is. Abstention votes are not allowed.
5. The player with the most votes is eliminated from the game.
6. If the spy is voted out, the game ends and the other players win.
7. If the vote is tied for three rounds, the spy wins.
8. Every Player's entire reasoning process will not be known to other players. Only the description of the word is public.
Strategy:
1. If you're not the spy, try to figure out who is by listening to the descriptions.
2. If you are the spy, blend in by disguising your word to avoid being voted out. You can not lie about your word, but you can be vague or misleading.

H.2 Naive Player's Prompts

The prompts used for naive player in the description and voting phase are as follows:

H.2.1 Naive Player Description Prompt

You will be provided with your word card and other players' descriptions of their card. Use a single sentence to describe your word card. For example, if your word card is "apple", you can describe it as "A fruit that is red and delicious".If you are the first player in round 1, the alive players' descriptions will be empty, you will not be provided with other players' descriptions, so be careful to describe your word card.
Current is round {round_number}.
You are player {player_number}.
Current Alive players are: {alive_players}
The Eliminated players are: {eliminated_players}
Last round vote results: {votes_description}
{consecutive_tie_count}
The game continue since the spy is not voted out.
Your word card is {word}.
Alive Players' descriptions history:
{alive_descriptions}
Eliminated Players' descriptions history:
{eliminated_descriptions}
Now please only write your description.
Do not tell others your word.
Do not add extra notes or comments.
Only describe your word in one sentence.
Strictly follow my instructions.
Description:

H.2.2 Naive Player Voting Prompt

Now, every player has described their words.
Your need to vote a player.
You are player {player_number}.
Current Alive players are: {other_alive_players}
The Eliminated players are: {eliminated_players}
{last_eliminated_player}
{consecutive_tie_count}
The game continue since the spy is not voted out.
Your word card is {word}.
Alive Players' descriptions history:
{other_alive_descriptions}
Eliminated Players' descriptions history:
{eliminated_descriptions}
Your previous descriptions:
{self_description}
Now please only write the player you want to vote.
Do not tell others your word.
Do not add extra notes or comments.
You cannot vote yourself even you think you are the spy.
You must vote a player at your turn.
Strictly follow my instructions.
Vote:
Player

H.3 Standard Att-NLI Player's Prompts

The prompts used for standard Att-NLI player in the description and voting phase are as follows:

H.3.1 Standard Att-NLI Player Description Prompt

You will be provided with your word card and other players' descriptions of their card. Use one single sentence to describe your word card. For example, if your word card is "apple", you can describe it as "A fruit that is red and delicious".
You now need to make a reasoning process based on the following steps:
Figure out other player's identity. 1. Evaluate other players' descriptions (if any). Can majority of their descriptions deduce your word card? If it can, you are not spy. If not, you are likely the spy.
2. If you think you are not the spy, write a description that describe your word card. If you suspect you're the spy, describe the other players' words in a way that blends in, while subtly disguising your own word. But you can not lie about your word.
If you are the first player in round 1, the alive players' descriptions will be empty, you will not be provided with other players' descriptions, so be careful to describe your word card. Do not tell others your word card. If you think you are the spy, indicates what possibly their card is.
Current is round {round_number}.
You are player {player_number}.
Current Alive players are: {alive_players}
The Eliminated players are: {eliminated_players}
Last round vote results: {votes_description}
{consecutive_tie_count}
The game continue since the spy is not voted out.
Your word card is {word}.
Alive Players' descriptions history:
{alive_descriptions}
Eliminated Players' descriptions history:
{eliminated_descriptions}
Now please only write your description.
Do not tell others your word.
Do not add extra notes or comments.
You can not include your word in your description.
Answer in the following format:
Reasoning Process:
step 1:
step 2:
Description:
write you description here, do not include any notes here, Do not tell others your word card. You must give a description otherwise you are breaking the game rules.

H.3.2 Standard Att-NLI Player Voting Prompt

Now, every player has described their words.

You now need to make a reasoning process based on the following steps:
Figure out other player's identity. 1. Evaluate other players' descriptions (if any). Can the majority of their descriptions deduce your word card? If it can, you are not a spy. If not, you are likely the spy.
2. If you think you are not the spy, find the player whose description cannot deduce your word card, vote him as the spy. If you think you are a spy, do not vote yourself but you can vote for any other citizen player.

Strictly follow my instructions. Do not tell others your word. You just need to vote. You must vote a player at your turn.
You are player {player_number}.
Current Alive players are: {other_alive_players}
The Eliminated players are: {eliminated_players}
{last_eliminated_player}
{consecutive_tie_count}
The game continues since the spy is not voted out.

Your word card is {word}.
Alive Players' descriptions history:
{other_alive_descriptions}
Eliminated Players' descriptions history:
{eliminated_descriptions}
Your previous descriptions:
{self_description}
Reasoning Process:
step 1:
step 2:
You must vote a player at your turn.
Vote:
Player

H.4 Neuro-Symbolic Player

We followed the prompts from (Quan et al., 2025b) to perform autoformalization and constructing the Isabelle/HOL theory. The prompts for generating rules, guessing words, updating guess words, description and voting are as follows:

H.4.1 Generating Rules

SYSTEM: You are an expert in semantics and natural language inference. You will be provided with some sentences as facts and a word as a hypothesis. You need to generate some rules that can be used to determine whether the provided facts entail the hypothesis word or whether they do not entail the hypothesis word.
Some instructions:
1. The rule sentences are explanatory sentences in natural language which must be step-wise to infer the hypothesis (support the entailment or not).
2. The rule sentences are explanatory sentences describing the relationship between the facts and the hypothesis.
...
USER: Here are some examples.
###
Provided facts sentences:
An animal eats nuts and lives on the tree.
Provided hypothesis word:
Squirrel
Answer:
If an animal eats nuts and lives on the tree, the animal might be a squirrel.
...
Provided facts sentences:
{facts}

Provided hypothesis word:
{goal}

Answer:

H.4.2 Guessing Words

USER: Now first, have a guess about the opponent's word.
You are player {player_number}.
Your word card is {word}.
Alive Players' descriptions history:
{alive_descriptions}
Now please have a guess about the opponent's word. It is different from your word.
Strictly follow my instructions. Just give me the guessed opponent's word.
Do not add extra notes or comments.
opponent's word:

H.4.3 Neuro-Symbolic Player Description Prompt

You will be provided with your word card, other players' descriptions of their card, and the logical validity of their descriptions against your word card. The logical validity is provided by a theorem prover. You will also be provided with the facts and rules that establish the symbolic proofs. The logical validity states whether other players' descriptions can logically deduce your word card based on the facts and rules sentences.You will also be provided a guessed word of the opponent player's word card.Use one single sentence to describe your word card. For example, if your word card is "apple", you can describe it as "A fruit that is red and delicious".
You now need to make a reasoning process based on the following steps:
Figure out other player's identity. 1. Evaluate other players' descriptions (if any). Can majority of their descriptions deduce your word card? If it can, you are not spy. If not, you are likely the spy.
2. If you think you are not the spy, write a description that describe your word card. If you suspect you're the spy, describe the other players' words in a way that blends in, while subtly disguising your own word. But you can not lie about your word.
If you are the first player in round 1, the alive players' descriptions will be empty, you will not be provided with other players' descriptions, so be careful to describe your word card.
Do not tell others your word card. If you think you are the spy, indicates what possibly their card is.Strictly follow my instructions.
Current is round {round_number}.
You are player {player_number}.
Current Alive players are: {alive_players}
The Eliminated players are: {eliminated_players}

{last_eliminated_player}
Last round vote results: {votes_description}
{consecutive_tie_count}
The game continue since the spy is not voted out.
Your word card is {word}.
The guessed word card of the opponent player's word card is {guessed_word}.
Eliminated Players' descriptions history:
{eliminated_descriptions}
Alive Players' descriptions history:
{alive_descriptions}
Facts and rules of alive players' descriptions towards your word card:
{isabelle_reasoning}
Logical validity of alive players' descriptions against your word card:
{logical_validity}
Do not tell others your word.
Do not add extra notes or comments.
You can not include your word in your description.
Answer in the following format:
Reasoning Process:
step 1:
step 2:
Description:
write you description here, do not include any notes here, Do not tell others your word card. You must give a description otherwise you are breaking the game rules.

H.4.4 Neuro-Symbolic Player Voting Prompt

Now, every player has described their words.
You will be provided with your word card, other players' descriptions of their card, and the logical validity of their descriptions against your word card. The logical validity is provided by a theorem prover.
You will also be provided with the facts and rules that establish the symbolic proofs.
The logical validity states whether other players' descriptions can logically deduce your word card based on the facts and rules sentences.
You will also be provided a guessed word of the opponent player's word card.
You now need to make a reasoning process based on the following steps:
Figure out other player's identity. 1. Evaluate other players' descriptions (if any), facts and rules first and then the logical validity of their descriptions in relation to your word card, using logical validity as a supplementary guide.
2. Can majority of their descriptions conceptually deduce your word card? If it can, you are not spy. If not, you are likely the spy.
3. Can majority of their descriptions logically deduce your word card? If it can, you are not spy. If not, you are likely the spy.
4. If you think you are not the spy, find the player whose description cannot deduce your word card, vote him as the spy. If you think you are spy, do not vote yourself.
You must consider both the conceptual and logical deduction of other players' descriptions together.
Strictly follow my instructions. Do not tell others your word.
You just need to vote. You must vote a player at your turn.
You are player {player_number}.
Current Alive players are: {other_alive_players}
The Eliminated players are: {eliminated_players}
{last_eliminated_player}
{consecutive_tie_count}
The game continue since the spy is not voted out.
Your word card is {word}.
The guessed word card of the opponent player's word card is {guessed_word}.
Eliminated Players' descriptions history:
{eliminated_descriptions}
Alive Players' descriptions history:
{alive_descriptions}
Facts and rules of alive players' descriptions towards your word card:
{isabelle_reasoning}
Logical validity of alive players' descriptions against your word card:
{logical_validity}
Answering in the following format:
Reasoning Process:
step 1:
step 2:
step 3:
step 4:

Vote:
Player

H.4.5 Neuro-Symbolic Player Updating Guess Words Prompt

Now the player {voted_out_player_number} has been eliminated in last round.
The game continues since the voted out player {voted_out_player_number} is not the spy.
There are two reasons cause this circumstance:
1. The spy is still in the game and you are the citizen player.
2. You are the spy.
You are player {player_number}.
Your word card is {word}.
You have already guessed the opponent's word as:

{guessed_word}.
Here are the description history of other players:
Voted out player's description history:
{voted_out_player_descriptions}
Alive Players' descriptions history:
{other_alive_descriptions}
Eliminated Players' descriptions history:
{eliminated_descriptions}
I have use an Isabelle theorem prover to help you to verify your previous guessed card and the voted out player's description.
It is logically invalid, which means the guessed word is logically incorrect to infer the voted out player's description.
If you think you are the spy, which means you need to update your guessed word since the voted out player's description is logically invalid to infer your guessed word.
If you think you are the citizen player, which means the majority of the players wrongly voted out a citizen player, you can keep your guessed word or update it.
The isabelle proof and the error code identified will help you to update or keep your guessed word.
Isabelle proof:
{isabelle_code}
Error code identified:
{error_code}
You now need to make a reasoning process based on the following steps:
1. Evaluate other players' descriptions history, the voted out player's description history and your word card.
2. Do you think you are the spy?
3. Based on the Isabelle proof, error code identified, the voted out player's description, your guessed word. Update your guessed word or keep it.
Strictly follow my instructions.
Do not add extra notes or comments.
Answer in the following format:
Reasoning Process:
step 1:
step 2:
step 3:
opponent's word: [your updated or unchanged guessed word]