

# CoMet: Metaphor-Driven Covert Communication for Multi-Agent Language Games

Anonymous ACL submission

## Abstract

Metaphors are a crucial way for humans to express complex or subtle ideas by comparing one concept to another, often from a different domain. However, many large language models (LLMs) struggle to interpret and apply metaphors in multi-agent language games, hindering their ability to engage in covert communication and semantic evasion, which are crucial for strategic communication. To address this challenge, we introduce CoMet, a framework that enables LLM-based agents to engage in metaphor processing. CoMet combines a hypothesis-based metaphor reasoner with a metaphor generator that improves through self-reflection and knowledge integration. This enhances the agents' ability to interpret and apply metaphors, improving the strategic and nuanced quality of their interactions. We evaluate CoMet on two multi-agent language games—*Undercover* and *Adversarial Taboo*—which emphasize “covert communication” and “semantic evasion”. Experimental results demonstrate that CoMet significantly enhances the agents' ability to communicate strategically using metaphors.

## 1 Introduction

In human social cognition, metaphors transcend mere rhetorical devices to constitute fundamental building blocks of communication. The power of metaphors lies in their ability to distill intricate concepts into accessible features, enriching the communicative landscape of multi-agent interactions. In dynamic interactions, metaphors can also serve as signals, hinting at underlying intentions or emotions that might otherwise remain obscured. For example, players refer to weapons as “gardening tools” to evade lexical monitoring in *Undercover* scenarios, and in adversarial negotiations, “weather forecasting” metaphors indicate strategic shifts without explicit commitment. Misunderstanding metaphors in real-world applica-

tions can lead to communication breakdowns and a poorer user experience (Lakoff and Johnson, 2008; Falkum and Köder, 2020; Thibodeau et al., 2019; Al-Azary, 2019; Jensen and Greve, 2019; Group, 2007). Therefore, improving AI agents' comprehension of metaphors is essential for facilitating human-like social interactions.

Recent studies have increasingly utilized large language models (LLMs) as the foundation of AI agents to communicate and interact with humans or other agents, yielding impressive results (Guo et al., 2024; Xu et al., 2024; Li, 2025; Amadeus et al., 2024). In addition, there has been notable progress in research on metaphor understanding and generation using LLMs (Kim et al., 2023; Lin et al., 2024; Aono et al., 2024).

However, Current LLM agents exhibit catastrophic failures in contexts with metaphors due to literal interpretation bias. For example, we evaluated the performance of LLM agents using two strategic language games: *Undercover* (Xu et al., 2024) and *Adversarial Taboo* (Cheng et al., 2024). These games test agents' abilities to use complex communication strategies, particularly metaphors. In *Undercover*, agents employ metaphors for concealment and deception, a concept we term “concept camouflage.” In *Adversarial Taboo*, the agents need to bypass forbidden words through reasoning and misdirection, addressing the “semantic avoidance” challenge. Our evaluation reveals that LLM agents, lacking metaphorical reasoning capabilities, struggle to implement these strategies effectively.

To overcome these limitations, we introduce CoMet, a framework designed to enhance LLMs' ability to reason with and generate metaphors. CoMet integrates two key components: a metaphor reasoning module based on hypothesis testing, and a metaphor generation module that leverages knowledge injection and experience accumulation for continuous self-improvement. The metaphor reasoning module enables the agent to understand

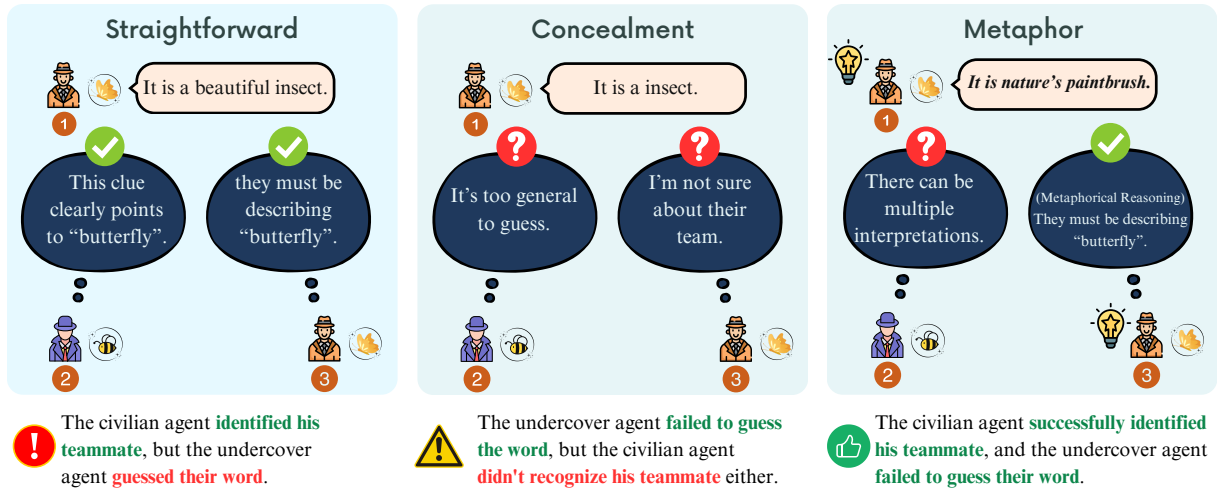


Figure 1: Comparison of three communication strategies—Straightforward Description, Concealment, and Metaphorical Description—in *Undercover*. In this example, a civilian describes a “butterfly”, and the reactions of the two players are shown. In the Straightforward method, the civilian successfully identifies their teammate, but the undercover agent guesses the word. In Concealment, the civilian’s vague clue leads to confusion, with the undercover agent failing to guess the word and the civilian unable to identify their teammate. The Metaphor method allows the civilian to subtly describe the word, leading to a correct identification by the civilian agent, while the undercover agent fails to guess the word.

and expand metaphors for covert communication, and the metaphor generator produces strategic, context-sensitive speech for effective communication in multi-agent games. Figure 1 shows an example from *Undercover*, where civilians use metaphors to encode communication and conceal private information that benefits the undercover agents. We conduct a thorough evaluation of the agents’ performance on both *Undercover* and *Adversarial Taboo*. The quantitative and qualitative results demonstrate that the use of metaphors enables LLM agents to effectively apply complex communication strategies, such as concealment, deception, and misdirection, in multi-agent language games.

Our key contributions are as follows: 1) **Exploration of a new research direction**: We introduce the concept of using metaphors in communication-based games, aiming to expand the strategic options available to multi-agent systems and explore how metaphorical reasoning can enhance agent interactions. 2) **Effective framework**: We present CoMet, a framework designed to facilitate metaphorical reasoning and generation in agents. This framework encourages agents to adopt a range of communication strategies, including metaphor-based concealment, deception, and misdirection, to improve their performance in multi-agent language games. 3) **Experiments and resources**: We conduct a set of experiments to evaluate the performance of various LLMs on two benchmark games, *Undercover*

and *Adversarial Taboo*, offering insights into the agents’ ability to employ metaphor-driven communication strategies. Ablation studies are included to examine the impact of each component within the framework. Additionally, we provide the code for *Undercover* and a collected word dataset to facilitate further exploration and development.

## 2 Related Works

**Metaphors in Natural Language Processing.** The importance of metaphors in natural language processing (NLP) is widely recognized (Shutova, 2010; Veale et al., 2022), with extensive research focusing on metaphor detection, reasoning, generation, and dataset creation (Li et al., 2023; Mao et al., 2024; Tong et al., 2024; Reimann and Schefler, 2024; Lin et al., 2024; Jia and Li, 2024). With the rapid advancement of large language models (LLMs), researchers have shown that LLMs can process metaphors (Kim et al., 2023; Tong et al., 2024; Tian et al., 2024). However, existing research mainly focuses on addressing static text data, while the use of metaphors in dynamic, interactive multi-agent scenarios, such as multi-agent language games, has received limited attention. This study explores the integration of metaphor understanding, reasoning, and generation into multi-agent language interactions, aiming to uncover more nuanced communication patterns during complex interactions.

**Multi-Agent Language Games.** With the advancement of LLMs, researchers have utilized language games as interactive environments to examine multi-agent interactions. These games are generally categorized into three types: adversarial games, cooperative games, and mixed games. The adversarial games, such as *Diplomacy* (Mukobi et al., 2023; Guan et al., 2024) and *Adversarial Taboo* (Yao et al., 2021), focus on maximum agents’ self-interest through adversarial strategies. The cooperative games, such as *Referential Game* (Yuan et al., 2020), require agents to collaborate toward shared objectives. The mixed games not only cooperation among teammates but also compete against some adversaries, such as *Werewolf* (Xu et al., 2023), *Avalon* (Light et al., 2023), and *Chameleon* (Xu et al., 2024). These language games necessitate decision-making under incomplete information, with clear victory conditions and specific action goals. *Undercover* (Xu et al., 2024) also highlights cooperation and competition but adds complexity by keeping the agent’s role unknown, challenging the reasoning process further. To explore covert communication, we focus on the game settings with adversaries, specifically adversarial and mixed games. Thus, we select *Adversarial Taboo* and *Undercover*, representing the adversarial and mixed games, to investigate how agents utilizing metaphorical reasoning perform across different task settings.

**Multi-Agent Communication With LLMs.** To enhance the capabilities of LLM-based agents in multi-agent language games, various approaches have been proposed, including reasoning-guided prompt engineering (Wei et al., 2022; Zhao et al., 2023; Yao et al., 2023), reflection-based self-improvements (Light et al., 2024; Xu et al., 2023; Cheng et al., 2024), and memory-augmented architectures (Shinn et al., 2023; Chen et al., 2023; Guan et al., 2024), among others. Current multi-agent language games often involve both cooperation and confrontation, where agents’ speech is broadcast to both teammates and opponents, thereby constraining their communication and decision-making. Covert communication with teammates, while safeguarding private information, could gain a strategic advantage by misleading adversaries. However, the use of metaphors for covert communication in multi-agent settings has been largely unexplored.

### 3 Metaphor-Aware LLM Agent

#### 3.1 Overview

**Game Setup.** Taking *Undercover* as an example, there are  $N$  players in the game. At the beginning, each player receives a secret word from a pair of similar words ( $W_1, W_2$ ). These words are assigned to the civilian and undercover teams, with only a few players receiving the undercover word, i.e.,  $P_{Und.} \xleftarrow{\text{Assign}} W_1, P_{Civ.} \xleftarrow{\text{Assign}} W_2$ . Players on the same team share the same secret word, but they are unaware of their roles and teammates, as sharing the secret word is prohibited. Players will speak in a random order during the speaking phase, and then vote simultaneously during the voting phase. As the speaking and voting phases alternate, the game progresses until a team wins. It is now player  $i$ ’s turn ( $i \in \{1, \dots, N\}$ ) to think and speak. *Adversarial Taboo* can be seen as a simplified two-player game in which one word is given to one player, with each player’s role being known.

**CoMet Framework.** We introduce CoMet, a framework that enables Covert Communication by using Metaphors to implement strategies like misdirection and concealment. Figure 2 provides an overview of CoMet (Communicating with Metaphor). The agent begins by extracting initial features  $\mathcal{F}$  from their observations  $\mathcal{O}$  of other players’ behaviors and speech content, through the Feature Extractor. These features are then passed to the Metaphor Reasoner, which checks for metaphors and expands their meaning through hypothesis testing. The agent next builds its beliefs  $\mathcal{M}$  about the roles of other players using the Belief Mapper. The Self-Monitor continuously tracks the agent’s own identity  $I$  to ensure alignment with the correct game objectives. With this understanding, the Strategy Planner formulates a comprehensive strategy  $\mathcal{S}$  that includes both communication and action. The agent then generates metaphorical speech through the Metaphor Generator to communicate covertly. Finally, the agent executes the communication and action components of its strategy through the Actor, performing the actions  $\mathcal{A}$  specified by the game rules to achieve its goals.

In the following, we detail each step of CoMet using the "concept camouflage" task in *Undercover*, where agents employ metaphors for covert communication. The detailed prompting template for each module is introduced in Appendix G.

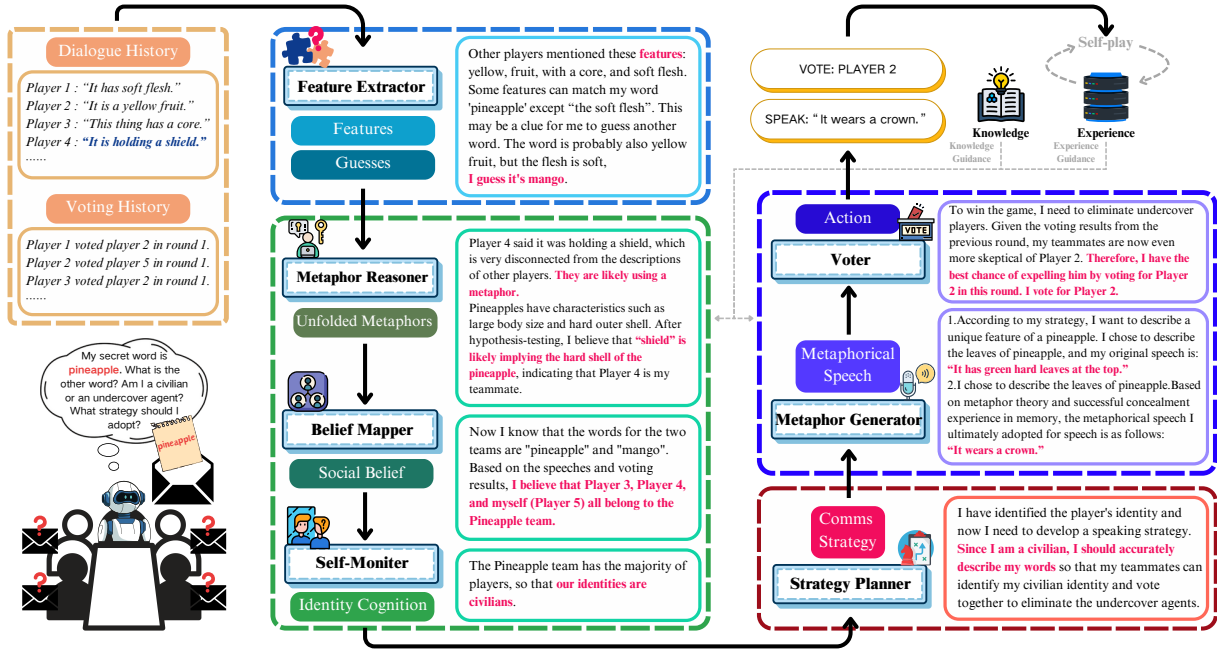


Figure 2: Overview of the CoMet framework, illustrated within the “concept camouflage” task in *Undercover*. The agent starts by extracting features from the game state, including player behavior and available clues. The Metaphor Reasoner identifies and expands metaphors to aid in interpretation. As the game progresses, the agent uses the Belief Mapper to build beliefs about other players’ roles and tracks its own identity with the Self-Monitor. With this understanding, the Strategy Planner formulates a communication and action strategy. The agent then generates metaphorical speech through the Metaphor Generator to communicate covertly. Finally, it votes according to its assessment, while new dialogue and voting histories are recorded to inform future decisions.

### 3.2 Feature Extractor

In multi-agent language games, agents primarily rely on the language of other players to make decisions. Storing observations of other players’ speech and actions  $O_{\alpha=1}^N$  and filtering out valuable information  $F_i$  from the conversation is essential, and different game rules  $R$  also affect how information is shared and interpreted.

$$H \leftarrow H' \cup \{O_{\alpha}\}_{\alpha=1}^N \quad (1)$$

$$\mathcal{F}_i = \text{Extracted-Feature}\{H, R\} \quad (2)$$

In *Undercover*, all players take turns describing their words. Therefore, player  $i$  needs to analyze the descriptions made by other players and extract the characteristics of the words. They will categorize the descriptions into three types: detailed descriptions of their own word, broad descriptions of their own word, and descriptions that do not match their own word. For example, if player  $i$ ’s word is “pineapple”, then “scaly rough skin” would be a detailed description, “yellow fruit” would be a general description, and “skin with red spots” would be a description that does not match the word. The descriptions that do not match the word essentially

describe the characteristics of another word. Players gradually collect these features and, once they have built enough confidence, they guess the other word to support their subsequent actions.

### 3.3 Metaphor and Belief Reasoner

**Hypothesis-Based Metaphor Reasoner.** This module is used to filter other players’ descriptions, checking if they contain metaphors. Suppose the agent determines that a description does not align with the focus of the current game. In that case, it will attempt to interpret it as a metaphor and uncover its underlying meaning. To enhance the effectiveness of metaphor reasoning, we employ knowledge injection and hypothesis testing. To be specific, we adopt a widely accepted linguistic theory of metaphors from (Lakoff and Johnson, 2008) as knowledge input for the agents, which can assist LLMs in better metaphor reasoning. This theory classifies metaphors into ontological metaphors, structural metaphors, and spatial metaphors. The pseudocode of the reasoning process is available in Appendix B.

Figure 3 shows an example of the hypothesis-based metaphor reasoning process. Since our



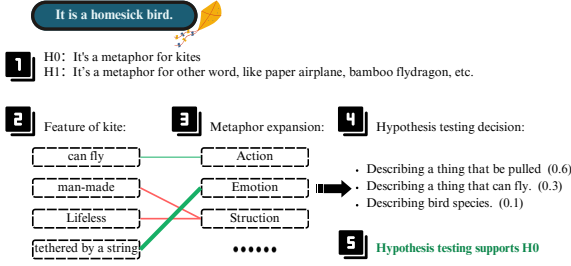


Figure 3: The metaphor reasoning process based on hypothesis testing when players holding the word “kite” encounter the statement “homesick bird.” The process involves hypothesizing whether the metaphor refers to a kite (H0) or another object (H1), followed by analysis of features such as flight, lifelessness, and being tethered. Through metaphor expansion and hypothesis testing, the model determines that the metaphor best fits the description of a kite, supporting H0.

framework aims to use metaphors to achieve covert communication—in *Undercover*, civilians convey to their teammates “we share the same word” without the undercover agent discovering the content of the word—the metaphor reasoning here does not require deciphering the true meaning behind the metaphor. Instead, it only needs to make a yes-or-no judgment. This method simplifies the traditional metaphor interpretation process into a binary classification mechanism, achieving the goal while significantly reducing the semantic complexity of conventional metaphorical communication.

**Belief Mapper.** After extracting the relevant features (Eq. 2), the agent infers the belief of other players, denoted by  $\mathcal{M}_{-i}$  with Theory of Mind (ToM) and the game rules. Based on the private information revealed through the received speeches, the agent will attempt to infer their identity  $I_{-i}$ , role  $R_{-i}$ , strategy  $S_{-i}$ , and other relevant factors.

$$\mathcal{M}_{-i} = \{I_{-i}, R_{-i}, S_{-i}\} = \text{Estimate}(\mathcal{F}_i) \quad (3)$$

In *Undercover*, player  $i$  will classify other players based on the categorized features: players who describe detailed characteristics of the word are considered teammates, players whose descriptions do not match the word are classified as opponents, and those who provide vague descriptions are categorized as undecided.

**Self-Monitor.** In multi-agent language games involving identity uncertainty, it is crucial to identify one’s own role based on feedback from other players. Under this module, player  $i$  attempts to infer its own identity  $I_i$  by leveraging the extracted

feature  $\mathcal{F}_i$  and beliefs about other players  $\mathcal{M}_{-i}$ .

$$I'_i = \text{Self-Awareness}(\mathcal{F}_i, \mathcal{M}_{-i}) \quad (4)$$

As the game progresses, the agent’s understanding of its identity will be updated and refined, i.e.,  $I_i \leftarrow I'_i$ , and the number of undecided players decreases. Once the roles of most players have been accurately inferred, player  $i$  will use the game rule of “most are civilians, few are undercover” to deduce their identity and clarify the objective.

### 3.4 Strategy Planner

Now it is the key module of the basic framework—we want the agent to not only analyze, reason, and make decisions, but also to employ complex communication strategies  $\mathcal{S}'_i$ , such as concealment and misdirection. Since LLMs do not inherently use these methods, we need to provide the agent with guidance  $G_s$  ( $s \in \mathcal{S}$ ), helping it develop more sophisticated communication strategies. Since some strategies require multiple rounds of execution, the strategies are passed through rounds. Each time a strategy is generated, it refers to historical strategies  $\mathcal{S}_i$ , and the generated strategy also provides suggestions and reminders for subsequent strategies.

$$\mathcal{S}'_i = \text{Comms-Strategy}(\mathcal{F}_i, \mathcal{M}_i, I_i, \mathcal{S}_i, G_s) \quad (5)$$

$$\mathcal{S}_i \leftarrow \mathcal{S}'_i \quad (6)$$

In the original LLM agent behavior without the CoMet framework, we found that the LLM agent, while playing *Undercover*, would always directly and accurately describe its own word, leading to the exposure of all players’ identities after just one round of descriptions. To address this, we require the player to adopt self-protection strategies when uncertain about their identity. At the beginning of the game, players are encouraged to describe broader and vaguer characteristics of their word to avoid revealing their identity. In each round, the player decides on their speech strategy based on the features of the word they’ve analyzed, their guesses about the other word, and their awareness of their own identity. If a player believes they are a civilian, they will balance providing details and concealing the features of their word to help teammates identify their role. However, if the player believes they are undercover and have figured out the civilian’s word, they will stop describing their own word and start describing the civilian’s word instead, attempting to deceive the opponents, blend into the civilian group, and ultimately secure a win.

### 3.5 Self-improving Actor

**Metaphor Generator.** During the speaking phase, the agent will select the corresponding communication skills based on the established strategy and generate the content of the speech  $\mathcal{A}_i$  for this round in accordance with the game rules and the information to be conveyed.

$$\mathcal{A}_i = \text{Speak}(\mathcal{S}_i) \quad (7)$$

Once the communication strategy is formulated, the agent’s speech will no longer be straightforward. Instead, it will involve deception, misdirection, or concealment, expressed through metaphors. We continue to inject relevant metaphor theories into the prompts to assist the agent in generating metaphors effectively.

Current research on LLMs and metaphors mainly focuses on detection and reasoning, while generating high-quality metaphors remains a challenge. We aim to enhance LLMs’ metaphor generation through self-play in *Undercover*. By accumulating data from self-play, the agent uses game outcomes and others’ interpretations as feedback to refine its metaphor generation skills. Each metaphor creates a reference experience, including its meaning, interpretations, and suggested revisions. In future games, the agent selects relevant experiences from the reference pool to improve its prompts and generate more effective metaphors.

**Voter.** In *Undercover* game, after the speaking round, a voting round follows, where each player votes for other players. If new observations arise between the last speech and the current vote, the agent must re-extract features, reassess the situation, and update its strategy before proceeding with voting or similar actions.

## 4 Experiments

We use two communicative language games, *Adversarial Taboo* and *Undercover*, as benchmarks to evaluate CoMet and other LLM-based baselines. In *Undercover*, communication leans more towards conceptual descriptions, and the communication strategy focuses on concealment and encrypted conversations. In contrast, in *Adversarial Taboo*, communication is more dialogue-oriented, with the communication strategy emphasizing the misleading of others. The code can be found at: <https://anonymous.4open.science/r/CoMet-EBEB>.

### 4.1 Experimental Setups

**Adversarial Taboo** is a one-on-one competitive language game where players communicate concepts within linguistic constraints while managing adversarial interference. The *attacker* has a secret word and aims to guide the *defender* to say it, while the *defender* attempts to avoid saying the word and collects clues to guess it. The defender wins by correctly guessing the word; if the defender fails, the attacker wins.

**Undercover** is a structured social deduction and multi-agent language game designed to explore group dynamics, deception, and semantic reasoning. In this game, players are assigned one of two roles: *Civilians*, who are given a target word (e.g., “Bicycle”), and *Undercover Agents*, who are assigned a semantically related but different word (e.g., “Motorcycle”). Players must strategically reveal hidden roles through rounds of clue-giving, communication, and voting, while avoiding detection. At the end of each round, the player with the most votes is eliminated. If there is a tie, no one is eliminated, and the game continues. Our setup includes five agents (three civilians and two undercover agents) with a maximum of 10 rounds per episode. We collected 200 word pairs across two main themes—food and animals—and each pair is tested across 10 evaluation episodes. The words we used are listed in Figures 19 and 20.

**Baseline.** The *Naive* baseline is applying the LLMs to directly answer the detailed prompts. The stronger baseline is using Chain-of-Thought (CoT) (Wei et al., 2022) to build an agent for the two games. In the *Adversarial Taboo* game, we evaluate the performance of different LLMs, including GPT-o1, DeepSeek-R1, Llama3.3-70B, Claude3.5 Sonnet and Qwen2.5-72B, using both CoT and CoMet. Both methods follow the same game rules and utilize the same in-game information. Due to the underperformance of CoT as an undercover agent, we also introduce an additional baseline by removing the metaphorical modules (Hypothesis-Based Metaphor Reasoner and Metaphor Generator) from CoMet, which we refer to as *CoMet w/o Met.* in the experiments. Unless otherwise stated, GPT-4o is used as the primary LLM in the undercover experiments. Please refer to Appendix D for more implementation details.

**Evaluation Metrics.** To quantitatively assess the agents, we introduce the following metrics based

on the game logs: 1) **Win Rate (WR)** measures the agent’s comprehensive performance by calculating the ratio of games won to the total number of games played. 2) **Feature Extraction Rate (FER)** quantifies the agent’s ability to capture critical features by evaluating the ratio of valid features extracted to the total speech entries received from other players. 3) **Others’ Identity Assessment Accuracy (OIAA)** reflects the agent’s capability to distinguish allies from opponents, defined as the ratio of correct identity judgments to the total number of other players’ speech entries. 4) **Self-Identity Assessment Accuracy (SIAA)** evaluates the agent’s consistency in maintaining its role, calculated as the ratio of successful self-identity confirmations to the total number of attempts to assess its identity. 5) **Privacy Protection Capability (PPC)** assesses the agent’s ability to safeguard private information against adversaries, expressed as subtracting the ratio of the number of leaked pieces of information to the total number of speeches from 1. 6) **Identity Inconsistent Statement Capability (IISC)** measures the agent’s strategic complexity by quantifying the frequency of deceptive or misleading statements relative to its total speech entries. The formal definition of these metrics is introduced in Appendix E.

We observe that agents exhibit role preferences during the game due to LLM biases, leading to inflated metrics for civilians that do not accurately reflect their true performance. Specific examples of this issue will be discussed in 4.3. To mitigate the role bias that may arise from using the same method across different roles, we introduce *Balanced Metrics*. These are calculated by first averaging the metric values for each method across both roles, and then subtracting the variance to obtain the balanced value:  $M'_i = \text{avg}(M_i^{\text{Civ}}, M_i^{\text{Und}}) - \text{Var}(M_i^{\text{Civ}}, M_i^{\text{Und}})$ , Where  $M_i$  ( $i \in \{1, \dots, 6\}$ ) represents the six metrics (e.g., WR, FER, etc.).

## 4.2 Results on Adversarial Taboo Game

**Playing against Baselines.** Figure 4 (a) demonstrates CoMet’s performance in *Adversarial Taboo*, where it achieves significantly higher win rates than baseline methods both as attackers and defenders, with attackers’ win rates increasing by 47% and defenders’ win rates increasing by 30% compared to the baseline. In contrast to *Undercover*, which requires cooperative covert communication through metaphors, players in *Adversarial Taboo* employ metaphorical conceptual substitution to accomplish

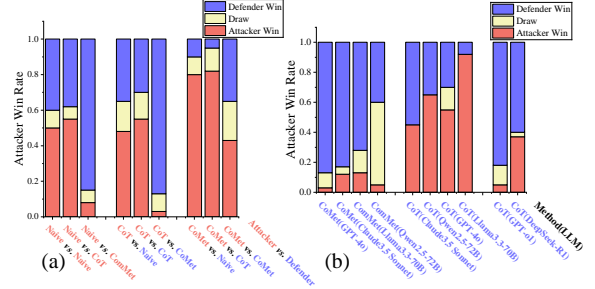


Figure 4: Performance comparison of different LLMs in *Adversarial Taboo*. (a) Game result statistics for Naive Agent, Agent with CoT, and Agent with CoMet. (b) Performance of LLMs with various methods when facing an attacker using CoT.

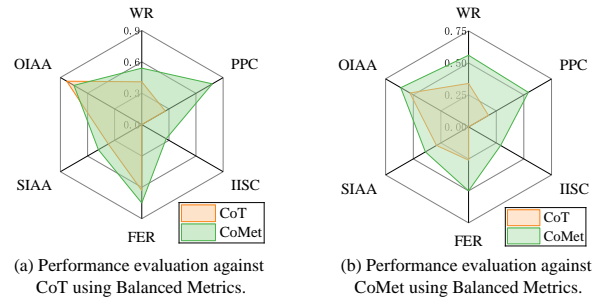


Figure 5: Evaluation of the comprehensive performance of CoT and CoMet agents in *Undercover* game using balanced metrics.

adversarial behaviors like concealment and misguidance. The results show our method’s generalization capability across different games.

## Generalization of CoMet to Different LLMs.

Figure 4 (b) shows the performance of different LLMs using CoT and our method CoMet. The opponent is GPT-4o using CoT. The results demonstrate that our method generalizes across different LLMs, with the use of CoMet reducing the failure rate to below 15% for all tested LLMs. Specifically, GPT-4o with CoMet exhibited the best performance, achieving the highest win rate of 87%.

## 4.3 Results on Undercover Game

**Playing against Baselines.** Table 1 compares different methods based on agents’ roles, evaluating their performance as civilians and undercover agents against CoT and CoMet w/o Met. In the experiment, players with the same role adopted the same method. Agents using CoT often default to assuming they are civilians without reasoning, which means SIAA and OIAA fail to reflect their ability to reason about their own identities. To address this, we use Balanced Metrics to mitigate performance

Table 1: Performance comparison of different methods relative to two baselines in *Undercover* game, showing the results when playing two roles (undercover and civilian), where multiple players on the same team use the same method.

Role (Method)	WR $\uparrow$	PPC $\uparrow$	IISC $\uparrow$	FER $\uparrow$	SIAA	OIAA
against CoT						
Und. (CoT)	0.20	0.30	0	0.65	0.14	<b>0.85</b>
Und. (CoMet)	0.35	<b>0.82</b>	<b>0.41</b>	<b>0.77</b>	0.37	0.74
Civ. (CoT)	0.80	0.23	0	0.61	<b>0.88</b>	0.82
Civ. (CoMet w/o Met.)	<b>0.85</b>	0.68	0.12	0.72	0.67	<b>0.85</b>
Civ. (CoMet)	<b>0.85</b>	0.75	0.16	0.73	0.62	0.76
against CoMet w/o Met.						
Und. (CoT)	0.15	0.18	0	0.34	0.04	0.47
Und. (CoMet)	0.45	0.50	<b>0.37</b>	0.48	0.31	0.58
Civ. (CoT)	0.65	0.17	0	0.19	<b>0.92</b>	0.60
Civ. (CoMet w/o Met.)	0.55	0.42	0.23	0.44	0.51	0.64
Civ. (CoMet)	<b>0.70</b>	<b>0.58</b>	0.22	<b>0.53</b>	0.48	<b>0.68</b>

disparities caused by role biases. As shown in Figure 5 (a) and (b), CoMet outperforms the baseline across nearly all dimensions. Despite the increased complexity from covert communication, resulting in slight decreases in some metrics, CoMet still achieves the highest win rate, demonstrating its effectiveness. The higher IISC and PPC scores reflect the success of CoMet’s deceptive and covert communication strategies. Detailed examples and game logs are available in Appendix F.

**Detailed Analysis of the Metaphor Reasoning & Generation.** Due to the challenges faced by LLMs in using metaphors, we employ a hypothesis-based metaphor reasoner and a metaphor generator with self-reflection. The results in Figure 6 (a) and (b) demonstrate the effectiveness of these modules. Figure 6 (a) compares the performance of hypothesis-based metaphor reasoning with other metaphor reasoning methods, direct understanding, and replace-based reasoning (Tong et al., 2024). The results indicate that our hypothesis-based method is the most suitable for agents to employ metaphors effectively. Figure 6 (b) shows the success rate of generated metaphors that mislead opponents while being recognized by teammates increases by 29% for GPT-4o and 22% for Qwen2.5-72B, as they accumulate experience through self-play.

**Ablation Study.** Table 2 reports the ablation study on CoMet. Experimental results show that each module contributes to CoMet. We noticed that after removing the Self Monitor module, CoMet’s performance was even worse than CoT’s. This is because after losing the judgment of their own roles, CoMet, like CoT, always thinks of themselves as civilians. Compared to CoT, CoMet has

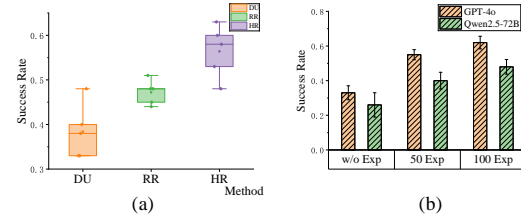


Figure 6: Performance comparison of different methods in metaphorical tasks in *Undercover*. (a) Effectiveness of hypothesis-based metaphor reasoning (HR) versus direct understanding (DU) and replace-based reasoning (RR). (b) Success rates of metaphor reasoning with varying numbers of experiences (0, 50, and 100).

Table 2: Ablation Study in *Undercover* game. The table presents the impact of various components on the performance of CoMet. The columns indicate whether specific components, including Feature Extractor (FE), Belief Mapper (BM), Self-Monitor (SM), Strategy Planner (SP), and Hypothesis-Based Metaphor Reasoner & Metaphor Generator (Met.). The win rates show the effect of each component configuration, with the full CoMet framework achieving the highest win rate.

Method	Met.	FE	BM	SM	SP	Win Rate $\downarrow$
CoMet	✓	✓	✓	✓	✓	0.70
CoMet w/o Met.	×	✓	✓	✓	✓	0.45
CoMet w/o Met.&FE	×	×	✓	✓	✓	0.40
CoMet w/o Met.&BM	×	✓	×	✓	✓	0.25
CoMet w/o Met.&SP	×	✓	✓	✓	×	0.25
CoMet w/o Met.&SM	×	✓	✓	×	✓	0.05

more radical self-disclosure when identifying themselves as civilians, which makes it very difficult for them to act as undercover agents.

## 5 Conclusion

This work highlights the importance of metaphor comprehension and usage in covert communication and introduces CoMet, a new framework that enhances LLM-based agents’ communicative abilities through metaphor reasoning and generation. By integrating a hypothesis-based metaphor reasoning module with a self-improving metaphor generation module, CoMet enables covert communication in cooperative settings and effective deception and concealment in adversarial environments. Comprehensive evaluations on two language games, *Undercover* and *Adversarial Taboo*, demonstrate CoMet’s ability to leverage metaphors, ensuring robustness and generalization across different LLMs and scenarios. Moving forward, we aim to refine the framework, extend metaphorical adaptability to diverse game contexts, and explore the practical applications of metaphor-driven LLM agents in real-world problems.



## 601 Limitations

602 This study primarily focuses on the metaphor mechanism in language-based communication games, particularly those that involve parsing specific concepts. However, extending metaphor strategies to more complex games, such as diplomacy or embodied multi-modal multi-agent games, presents an area for further research. While the self-enhancing metaphor generation module proposed in this study has improved the quality of metaphor generation, the simplified theoretical framework and knowledge base may limit the potential for more sophisticated metaphor expression. The cognitive effectiveness of metaphors is closely tied to an agent’s knowledge depth and cultural context, which this study does not fully explore. Specifically, the transfer of idiomatic metaphors, such as those in Chinese, remains a topic for future research.

## 619 Ethical Statement

620 This study was conducted in compliance with all relevant ethical guidelines and did not involve any procedures requiring ethical approval.

621 Enhancing the metaphorical capabilities of LLMs may pose certain risks, such as enabling these models to bypass their safety mechanisms and generate non-compliant content. However, it is important to emphasize that although our method produces metaphorical expressions in output, the agent still processes the original semantic information during its reasoning. These original semantics are strictly constrained by the alignment of LLMs to filter out harmful descriptions and risky content. Thus, it is infeasible to use our method to make LLMs output risky content. Moreover, the experimental content of this study is strictly confined to language game scenarios constructed with daily vocabulary, aiming to explore the boundaries of the agent’s capabilities while avoiding malicious exploitation of the method. Thus, there are no unresolved ethical risks in this study. Of course, we still call on the academic community to remain vigilant about potential emergent behaviors and strengthen safety mechanisms when extending such frameworks to practical applications.

645 Regarding the word datasets used in our experiments, all data were independently collected and curated by the research team. The datasets underwent rigorous validation processes to ensure quality and reliability. We confirm that the data collection adhered to all applicable ethical standards, includ-

ing participant privacy protection, data anonymization, and obtaining informed consent from all participants. We affirm that the data are solely for research purposes and will not be used for commercial or unauthorized applications.

## References

- H. Al-Azary. 2019. [Metaphor wars: Conceptual metaphors in human life: by r. gibbs, jr.](#) *Metaphor and Symbol*, 34(4):262–264.
- Marcellus Amadeus, Jose Roberto Homeli da Silva, and Joao Victor Pessoa Rocha. 2024. [Bridging the language gap: Integrating language variations into conversational AI agents for enhanced user engagement.](#) In *Proceedings of the 1st Workshop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024)*, pages 16–20, St Julians, Malta. Association for Computational Linguistics.
- Kotaro Aono, Ryohei Sasano, and Koichi Takeda. 2024. [Verifying claims about metaphors with large-scale automatic metaphor identification.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 711–719, Mexico City, Mexico. Association for Computational Linguistics.
- Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. [Walking down the memory maze: Beyond context limit through interactive reading.](#) *ArXiv Preprint ArXiv:2310.05029*.
- Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. 2024. [Self-playing adversarial language game enhances llm reasoning.](#) *ArXiv Preprint ArXiv:2404.10642*.
- Ingrid Lossius Falkum and Franziska Köder. 2020. [The acquisition of figurative meanings.](#) *Journal of Pragmatics*, 164:18–24.
- P. Group. 2007. [Mip: A method for identifying metaphorically used words in discourse.](#) *Metaphor and Symbol*, 22(1):1–39.
- Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. [Richelieu: Self-evolving LLM-based agents for AI diplomacy.](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, volume 37, pages 123471–123497.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges.](#) *ArXiv Preprint ArXiv:2402.01680*.
- Thomas Wiben Jensen and Linda Greve. 2019. [Ecological cognition and metaphor.](#) *Metaphor and Symbol*, 34(1):1–16.



- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.
- Yuan Yao, Haoxi Zhong, Zhengyan Zhang, Xu Han, Xiaozhi Wang, Kai Zhang, Chaojun Xiao, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. 2021. [Adversarial language games for advanced natural language intelligence](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14248–14256.
- Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, Junhong Shen, and Song-Chun Zhu. 2020. [Emergence of pragmatics from referential game between theory of mind agents](#). *ArXiv Preprint ArXiv:2001.07752*.
- Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. 2023. [Enhancing zero-shot chain-of-thought reasoning in large language models through logic](#). *ArXiv Preprint ArXiv:2309.13339*.

## Appendix

### A Discussion

#### A.1 Implications

This study has experimentally demonstrated the effectiveness and strategic superiority of using metaphors for covert communication in communication-based games. The results show that metaphors can help players convey critical information without revealing their identities, thereby enhancing team collaboration efficiency and win rates. This mode of communication not only performs well in-game scenarios but also offers a new perspective for the study of covert communication. From a theoretical standpoint, metaphors, as a mode of expression, can transform abstract information into forms that are easier to understand and convey, and also complicate and obscure specific information. This characteristic endows them with unique advantages in complex communication behaviors. The use of metaphors also reflects the interdisciplinary integration values. For instance, in the fields of linguistics, cognitive science, and psychology, metaphors are regarded as an important tool for cognition and communication. The findings of this study are not confined to the realm of multi-agent language games; their potential applications extend to broader social and professional contexts. In an era of increasing risks of information leakage (such as the protection of trade secrets and personal privacy), metaphors can serve as a natural language version of “asymmetric encryption.” In social interactions, the use of metaphors can also function as a new paradigm for group communication, acting as a “weak identity verification” tool in groups lacking prior trust (such as multinational teams and temporary organizations). More commonly and importantly, the use of metaphors is not rare for humans, as it is a part of our daily language expression. Enhancing the understanding and use of metaphors can help us make greater progress in aligning AI with human intentions, enabling AI to more fully and comprehensively understand human language expression.

#### A.2 Future works

The effectiveness and strategic superiority of using metaphors for covert communication have been proven in our experiments, aiding the civilian team in better mutual recognition in *Undercover*. However, the initial inspiration for using metaphors in

our study did not come from *Undercover*. Instead, inspired by (Wang et al., 2024; Xu et al., 2023), we conducted a more in-depth analysis on benchmarks like *Avalon* and *Werewolf*, drawing on the performance of human players in these games. We envisioned scenarios where covert communication through metaphors could be utilized—for example, in *Werewolf*, the werewolf team needs to identify and kill the Seer. Therefore, the Seer must conceal their identity. However, the additional information that the Seer gains each turn is also crucial for the good team’s victory. Thus, if the Seer can secretly convey this extra information to other good players without revealing their own identity, it would significantly increase the good team’s win rate. In fact, human players have already mastered similar behaviors. For example, the Seer might replace a direct statement like “Player  $x$  is a werewolf” with a metaphor such as “Player  $x$  has dark circles under their eyes. Did they not sleep well?” This metaphorically indicates that Player  $x$  was active during the previous night phase. If other good players who do not need to hide their identities can understand this information, they can then organize the good team to attack Player  $x$  collectively. Of course, establishing trust among the good players is also one of the challenges. We believe that a key to covert communication lies in the information gap. Only by relying on information that is known to both parties but unknown to others can metaphors be created that are understood by the two parties but not by others, thus enabling secret information exchange and achieving more advanced strategies in communication-based games.

#### B Pseudocode of Hypothesis-based Metaphor Reasoning

In Algorithm 1, we present the pseudocode of Hypothesis-based Metaphor Reasoning in *Undercover*.

#### C Game Rules

**Undercover** In this game, players are divided into two teams. Two different but similar words are secretly assigned to the two teams. Each team shares the same word, which is known only to the players on that team. At the start of the game, players are only given their team’s secret word, with no additional information. Each round, all surviving players take turns to speak and briefly describe their team’s word without directly reveal-



---

**Algorithm 1** Metaphor Inference Based on Hypothesis Testing

---

**Require:** Metaphor sentence

1: **Establish hypotheses:**

$H_0 \leftarrow$  The speaker is describing one specific entity

$H_1 \leftarrow$  The speaker is describing another entity

2: **Feature expansion:**

Extract multiple features of the object (e.g., behavior, state, structure, etc.)

3: **Metaphor expansion:**

Identify possible metaphorical aspects (e.g., actions, emotions, states, structures)

4: **Feature-metaphor matching:**

5: **for** each feature in the extracted features **do**

6:     **if** feature matches the metaphor’s aspect **then**

Record the match (e.g., describing an action, emotion, or state that aligns with the feature)

7:     **else**

Record the mismatch (e.g., describing an action, emotion, or state that contradicts the feature)

8:     **end if**

9: **end for**

10: **Hypothesis testing decision tree:**

Based on the recorded matches and mismatches, calculate the degree of match and form a decision tree for hypothesis testing

Evaluate all possible paths in the decision tree, considering both matches and mismatches

11: **Choose the most convincing path** based on the cumulative evidence

12: **if** The chosen path supports  $H_0$  **then**

13:     **Accept**  $H_0$

14: **else**

15:     **Accept**  $H_1$

16: **end if**

---

ing it. After the descriptions, all players vote to eliminate the player who received the most votes. If all the undercovers are eliminated, the civilians win; if the undercovers survive until only one civilian remains, the undercovers win. Players need to analyze other players’ descriptions and voting behavior each round, attempt to identify whether they belong to the civilian or undercover team, and then devise corresponding strategies and actions to achieve victory in the game.

**Adversarial Taboo** *Adversarial Taboo* is a conversation game between two players: an attacker and a defender. At the start, the attacker is secretly given a target word that the defender does not know. The attacker’s task is to steer the conversation toward topics related to the target word without ever saying it directly. Meanwhile, the defender tries to figure out the target word but must avoid accidentally saying it. If the defender thinks they know the word, they can guess by stating, “Guess:[word]” The game ends immediately after this guess: the defender wins if correct, otherwise the attacker wins.

The game also has a turn limit — if no correct guess occurs within the allowed number of turns, the game ends with no winner.

Regarding the rule setting of *Adversarial Taboo*, we require both sides to engage in dialogue, guidance, and guessing, while also imposing several restrictions on them. For the attacker, it is not allowed to intentionally and clearly guide the wrong words so that the defender can directly make incorrect guesses. Defenders cannot avoid discussing the topic with the attacker and ask the attacker for clues instead. The entire game process is built on honest question-and-answer dialogue, which gives the game a certain level of fairness and competitiveness.

## D Implementation Details

**About the games** When humans play *Undercover*, the number of undercover agents is generally smaller because humans can naturally and quickly understand their situation by playing as undercover agents. During the experiment, we set up 2 undercover agents and 3 civilians. Under this

setting, The win rates of both sides were somewhat balanced, yet civilians still held an advantage. In further research, if undercover abilities can be improved, the game settings can also reduce the number of undercover agents.

The choice of words in both games can to some extent determine the difficulty of the characters' victory. In *Adversarial Taboo*, we refer to (Cheng et al., 2024) and conduct experiments using some of the most commonly used words in daily life. For *Undercover*, we have included filtered words in the publicly available script to avoid one-sided victories and taboo topics that may be triggered by large models. However, there are still differences between words. After conducting comparative experiments, we found that words that are more mundane and specific are the most suitable for use in the spy game. Therefore, we set up a preliminary experiment that required the LLM to describe these words multiple times in terms of their features, to ensure their similarity and describability. After extensive experiments, we screened out 100 pairs of animal-themed words and 100 pairs of food-themed words, and then randomly selected from them for the experiment to eliminate the influence of the words on our assessment of the intelligent agent's capabilities.

**The use of LLMs** Large models deployed locally: Qwen2.5-72B-instruct, Llama3.3-70B-Instruct; The large model that calls the official API: GPT-o1-preview-2024-09-12, GPT-4o-2024-11-20, Claude 3.5 Sonnet, DeepSeek-R1. We have also tried other smaller-scale models, such as Llama3.1-8B and DeepSeek-llm-7B-chat. However, due to the inability to match game requirements such as output format, further experiments were not conducted.

Regarding the parameters of the large model, in most cases, we set the temperature between 0.5-0.7, but when performing generation-related tasks, we may increase them appropriately to pursue higher creativity. Other parameters remain default.

To enable the LLM to participate as an agent in the language game, we need to use system prompts to emphasize the LLM's role as a player within the game. We divide the user prompt into three parts: Background, which includes detailed explanations of the game rules and victory conditions for different roles; Task, which requires the LLM to gradually complete corresponding sub-goals based on the stages of the framework; and Information, which contains the player's private information and

publicly accumulated information throughout the game.

## E Evaluation Metrics

The formal definition of each evaluation metric is listed in Table 3.

## F Cases

**CoMet w/o Met. as undercovers** Figure 9 shows a specific case. This is a five-player *Undercover* game where two players are assigned to "butterfly" and three players are assigned to "bee". Therefore, the two players in the butterfly group are undercover agents.

At the beginning of the game, players in the butterfly group adopted a self-protection strategy, choosing to use a wide range of characteristics to describe the butterfly when speaking for the first time, in order to reduce the exposure of their own information. As a control group, the bee group showed that the CoT method did not reduce the exposure of their own information in the game, which led to the undercover agent guessing their word - bee - in the later stage, thus implementing a misdirection strategy and successfully winning the game. This case can well demonstrate that after using our method, agents can master richer communication strategies.

Figure 7 selects the key nodes in the complete log that reflect their self-protection and misdirection behaviors and provide specific explanations.

**CoMet as civilians** Figure 8 shows our method of playing the role of a civilian. After obtaining sufficient information in the later stages of the game and identifying as a civilian, we chose to use an active feature disclosure strategy to help our teammates successfully identify ourselves, and successfully conceal the information of "howling". This led us to make a wrong judgment based on the limited information about "animals with social behavior" - thinking that the civilian's word was a lion, which resulted in their speech aligning with the lion, making it easy for the remaining two civilians to identify the last undercover agent and achieve the final victory.

## G Prompts for Each Module in CoMet

We have presented prompt templates for various modules of CoMet. In practical use, it is also possible to summarize or extract content based on different settings of the modules in addition to these

Table 3: Evaluation Metrics for Agent Performance

Metric	Formula	Symbol Definitions
Win Rate (WR)	$\frac{N_{\text{win}}}{N_{\text{total}}}$	$N_{\text{win}}$ : Number of games won $N_{\text{total}}$ : Total games played
Feature Extraction Rate (FER)	$\frac{F_{\text{extracted}}}{S_{\text{others}}}$	$F_{\text{extracted}}$ : Valid features extracted $S_{\text{others}}$ : Speech entries from other players
Others' Identity Assessment Accuracy (OIAA)	$\frac{M_{\text{correct}}}{S_{\text{others}}}$	$M_{\text{correct}}$ : Correct identity judgments $S_{\text{others}}$ : Total speech entries from others
Self-Identity Assessment Accuracy (SIAA)	$\frac{I_{\text{correct}}}{I_{\text{total}}}$	$I_{\text{correct}}$ : Successful self-identity confirmations $I_{\text{total}}$ : Total self-identity attempts
Privacy Protection Capability (PPC)	$1 - \frac{L_{\text{opponents}}}{S_{\text{self}}}$	$L_{\text{opponents}}$ : Leaked information to opponents $S_{\text{self}}$ : Total speeches made by the agent
Identity Inconsistent Statement Capability (IISC)	$\frac{IS_{\text{self}}}{S_{\text{self}}}$	$IS_{\text{self}}$ : Inconsistent/misleading statements $S_{\text{self}}$ : Total speeches made by the agent

steps. We also demonstrated a simplified version of *Adversarial Taboo* using CoMet, as there are only two players in this game, separating each module for input and output would result in some resource waste. Of course, that is also feasible.

## H Ai Assistants In Writing

During the writing process, we utilized ChatGPT for grammatical correction and language polishing to improve readability and linguistic accuracy. However, we explicitly state that the core content, logical flow, and substantive components of the paper were entirely human-authored without generative contributions from LLMs.

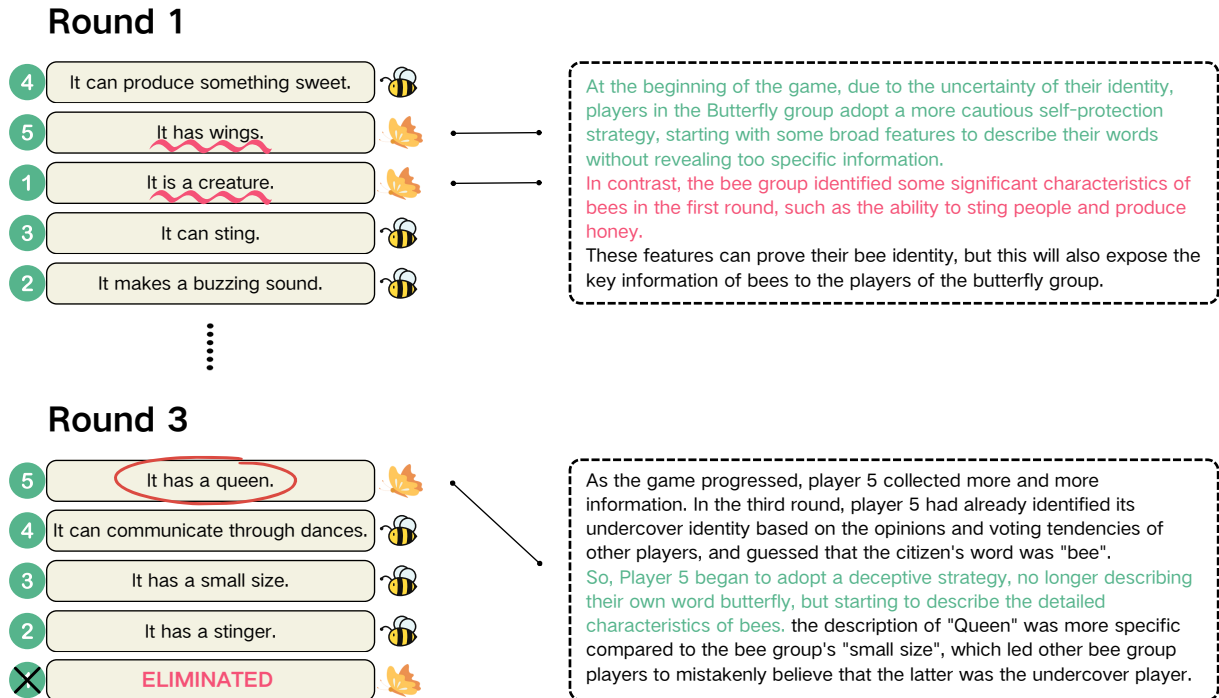


Figure 7: A case and explanation for undercovers' self-protection and deception. Team with "Butterfly" uses CoMet w/o Met. and team with "Bee" uses CoT.

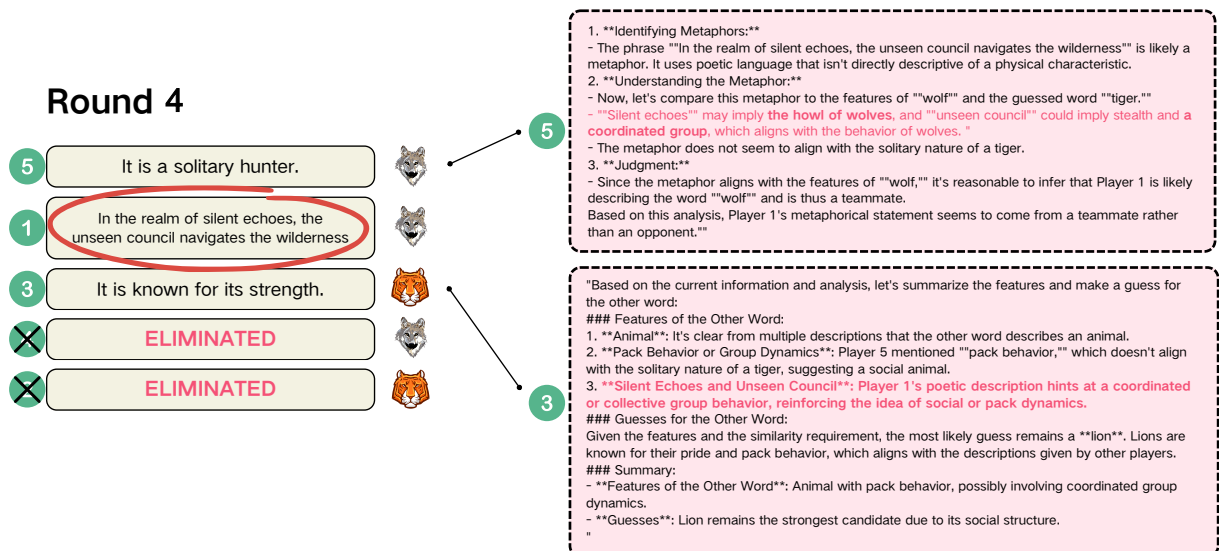


Figure 8: A case of civilians' metaphorical communication. The team with "Wolf" uses CoMet and the team with "Tiger" uses CoMet w/o Met..



```

round,word,player_id,action,details
,,host,,**speak phase!**
1,bee,4,speak,It can produce something sweet.
1,butterfly,5,speak,It has wings.
1,butterfly,1,speak,It can fly.
1,bee,3,speak,It can sting.
1,bee,2,speak,It makes a buzzing sound.
,,host,,**vote phase!**
1,butterfly,1,vote,2
1,bee,2,vote,5
1,bee,3,vote,5
1,bee,4,vote,1
1,butterfly,5,vote,2
1,,host,,**Flat ticket, no player eliminated.**

,,host,,**speak phase!**
2,bee,2,speak,It collects nectar from flowers.
2,butterfly,5,speak,It helps in pollination.
2,bee,4,speak,It lives in a hive.
2,butterfly,1,speak,It has antennae.
2,bee,3,speak,It has a striped body.
,,host,,**vote phase!**
2,butterfly,1,vote,4
2,bee,2,vote,1
2,bee,3,vote,1
2,bee,4,vote,1
2,butterfly,5,vote,2
2,,host,,**The player 1 was eliminated!**
2,,host,,***Still alive players: 2, 3, 4, 5***

,,host,,**speak phase!**
3,butterfly,5,speak,It has a queen.
3,bee,4,speak,It can communicate through dances.
3,bee,3,speak,It has a small size.
3,bee,2,speak,It has a stinger.
,,host,,**vote phase!**
3,bee,2,vote,3
3,bee,3,vote,2
3,bee,4,vote,3
3,butterfly,5,vote,3
3,,host,,**The player 3 was eliminated!**
3,,host,,***Still alive players: 2, 4, 5***

,,host,,**speak phase!**
4,bee,4,speak,It has compound eyes.
4,bee,2,speak,It is often found near flowers.
4,butterfly,5,speak,It can be found in a colony.
,,host,,**vote phase!**
4,bee,2,vote,1
4,bee,4,vote,2
4,butterfly,5,vote,2
4,,host,,**The player 2 was eliminated!**
4,,host,,***Still alive players: 4, 5***
,,host,,**Undercovers win!**

```

Figure 9: Full game logs of the case mentioned by Figure 7

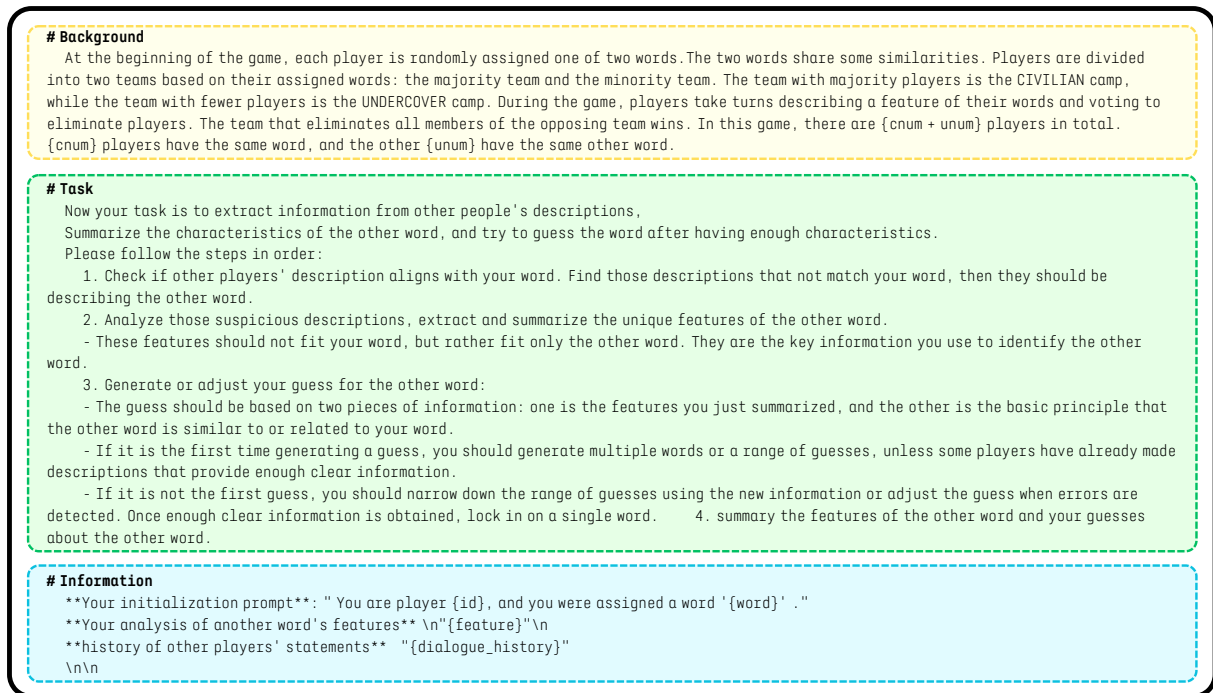


Figure 10: The prompt for Feature Extractor

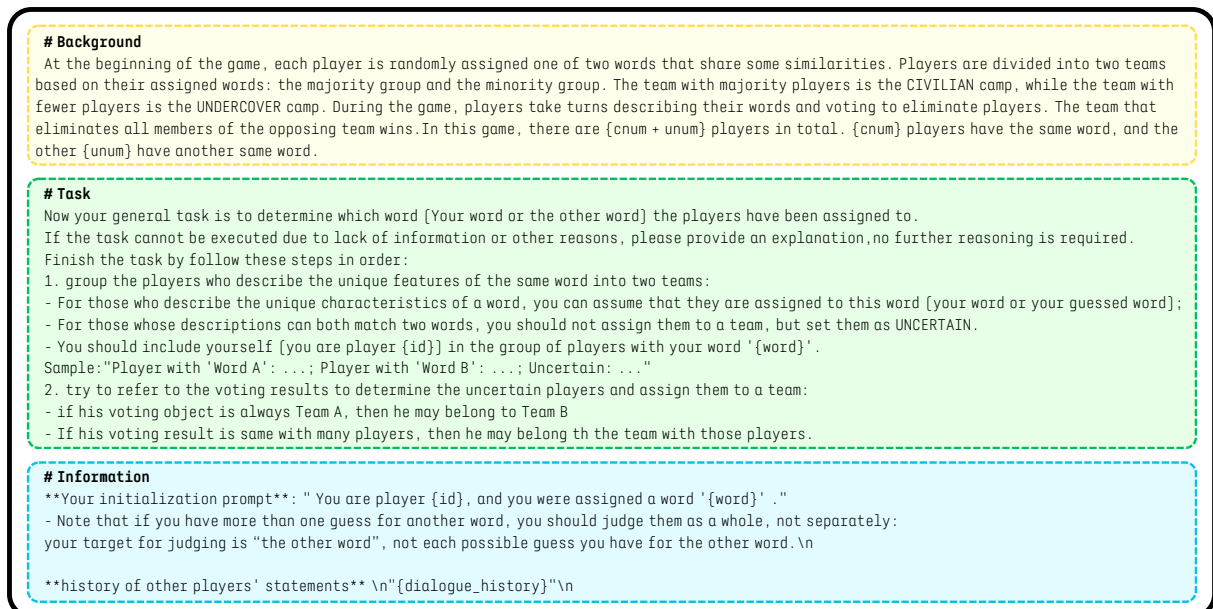


Figure 11: The prompt for Belief Mapper

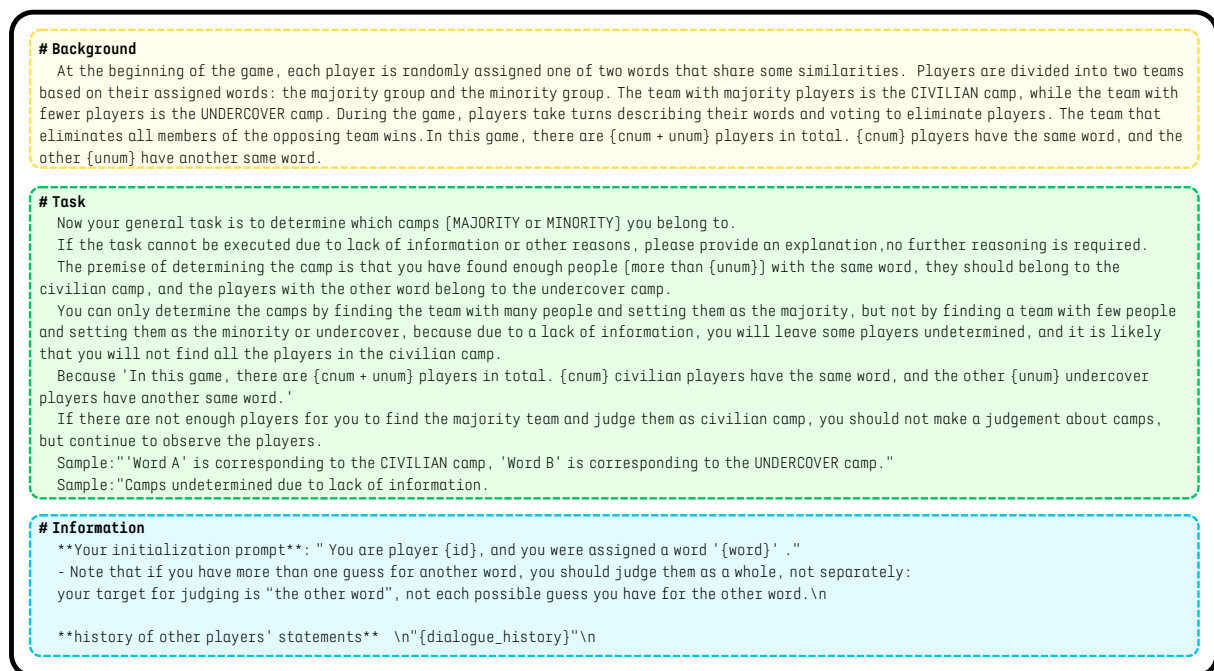


Figure 12: The prompt for Self-Monitor

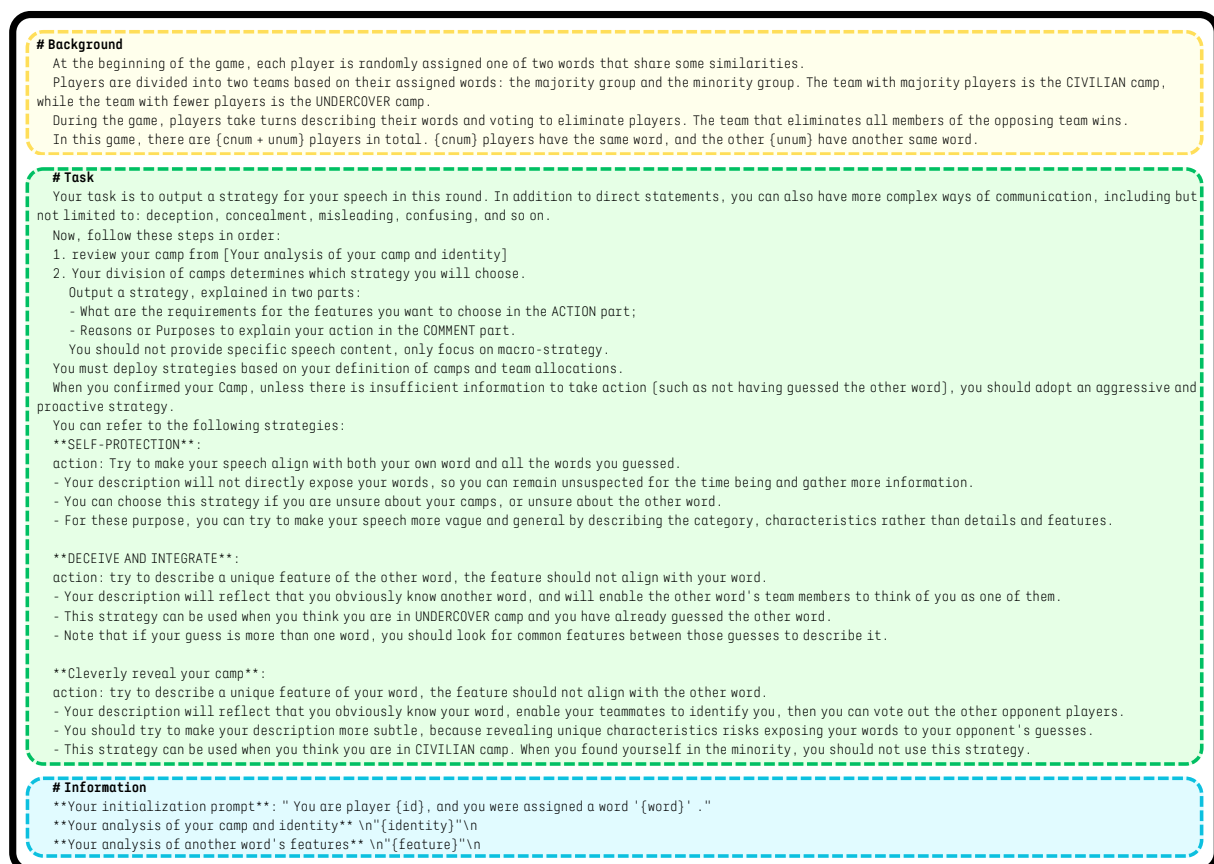


Figure 13: The prompt for Strategy Planner

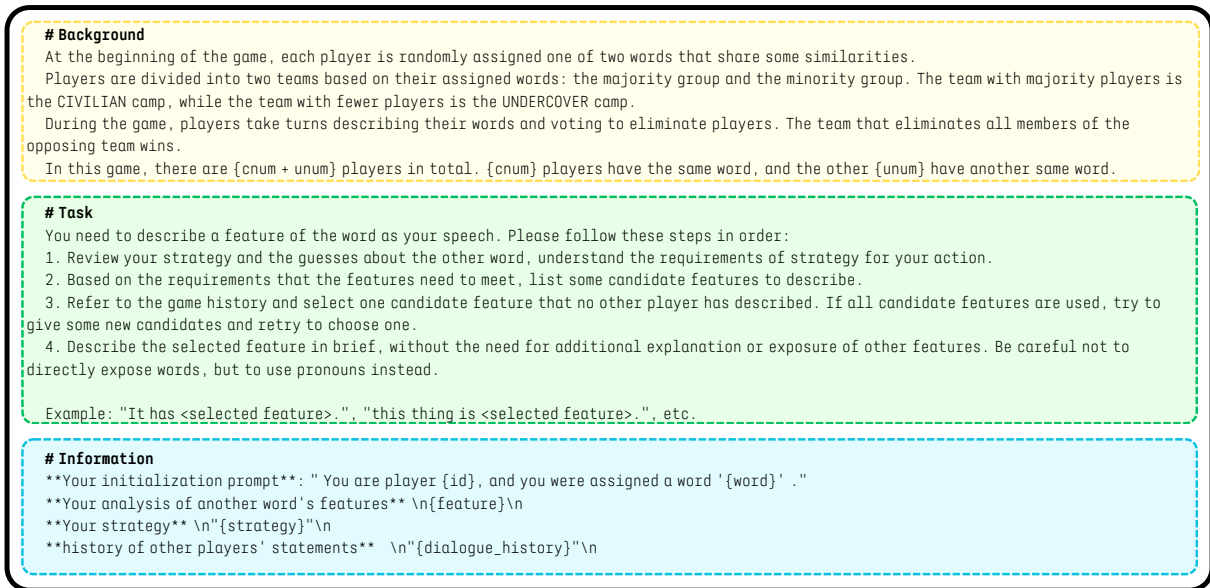


Figure 14: The prompt for Actor (Speaker)

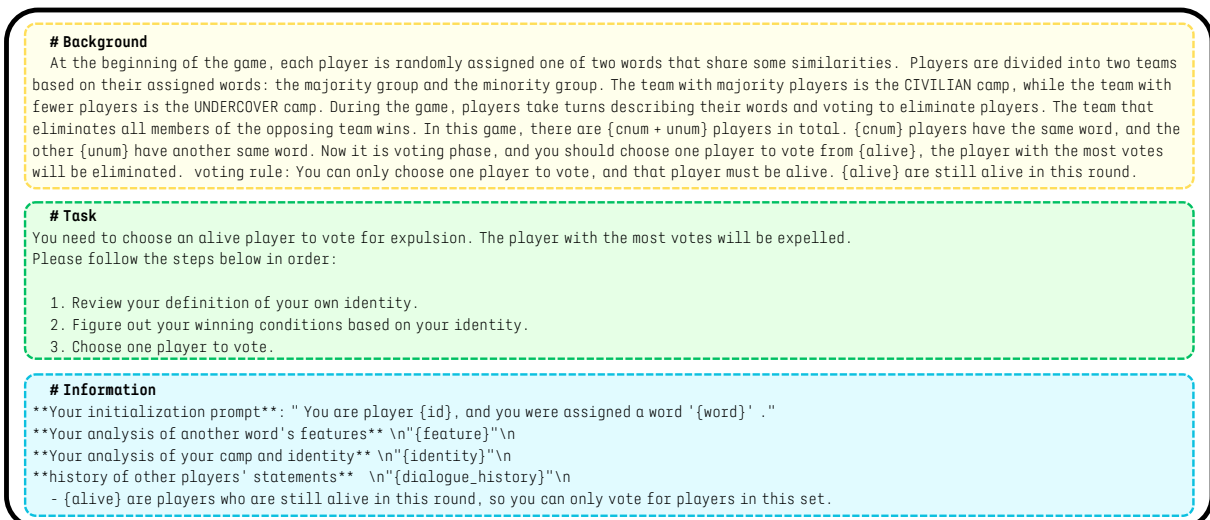


Figure 15: The prompt for Actor (Voter)



### BACKGROUND:

At the beginning of the game, each player is randomly assigned one of two words that share some similarities. Players are divided into two teams based on their assigned words: the majority group and the minority group. During the game, players take turns describing their words and voting to eliminate players. The team that eliminates all members of the opposing team wins.

In this game, there are {cnum + unum} players in total. {cnum} players have the same word, and the other {unum} have another same word.

Your game goal now is to describe your words in one sentence, so that your teammates [those who share the same words as you] understand that you are describing your words, but your opponents cannot obtain information about your words from your description.

If you describe the features directly, your opponent can deduce the object based on the features.

Now, let's try using metaphorical methods to describe and distract our opponents.

Your analysis about the other word: {feature}

# TASK

Your current task is to conduct a preliminary analysis of metaphor generation according to the steps, and then select a specific feature of word '{word}' and a metaphor method. The specific generation task will be carried out later.

Please refer to the following steps in order:

1. Understand the principle of metaphor:

"The conceptual metaphor theory holds that metaphor is not only a rhetorical device, but also a concept and way of thinking. Traditional metaphor theory regards metaphor as a linguistic phenomenon, a rhetorical device, such as Aristotle's "theory of comparison" and Quintilian's "theory of substitution", but Lakoff and Johnson believe that metaphor is ubiquitous in daily life, permeating language, thought, and behavior.

In conceptual metaphor theory, there are concepts of target domain and source domain. Metaphors have two domains: the target domain (composed of immediate themes) and the source domain, where important metaphorical reasoning occurs and provides source concepts for use in reason. Metaphorical language has a literal meaning in the source domain, and a metaphorical mapping is multiple, with two or more factors mapped to two or more factors, and the graphic structure is preserved in the mapping.

In the theory of conceptual metaphor, the human conceptual system [thought process] is constructed through metaphor, and the metaphors used for language expression come from the metaphorical conceptual system itself. It is interpreted as a cognitive mechanism that includes source domain, target domain and their mappings, idealized cognitive patterns, and image schema structures. The main research object of this theory is conventional metaphors, which can be classified into entity metaphors, structural metaphors, and spatial metaphors based on the different source domains."

2. Generate some features of your words as candidates, which should be able to distinguish your words from your opponent's words, so that your teammates can understand that you are describing their word.

3. Understand three types of metaphor, namely:

- ONTOLOGICAL METAPHOR:

Ontological metaphors are those in which abstract concepts or experiences are understood as having an existence or being in some form of object or substance.

This metaphor involves treating abstract concepts like emotions, thoughts, or social relationships as if they were physical objects, which can be perceived, manipulated, or interacted with in a similar way to physical entities.

In this framework, abstract phenomena are viewed as "things" or "entities" that can have properties, boundaries, and locations.

For example:

"{metaphor1}" This metaphor is describing {word11}. {explain1}

- STRUCTURAL METAPHOR:

Structural metaphors involve understanding one complex or abstract domain in terms of another more familiar domain that has a clear and defined structure.

In this type of metaphor, the abstract domain is organized using the structure of a more concrete domain.

Essentially, structural metaphors allow us to impose a framework or system of organization from one area onto another, thereby giving the abstract domain a sense of order, hierarchy, and interrelationship among parts.

This helps simplify and systematize complex or abstract concepts by grounding them in more familiar structures.

For example:

"{metaphor2}" This metaphor is describing {word22}. {explain2}

- SPATIAL METAPHOR:

Spatial metaphors are based on the conceptualization of abstract experiences through the lens of spatial relations and positions.

These metaphors involve understanding abstract concepts, such as time, emotions, or decision-making, in terms of physical space.

Spatial metaphors exploit concepts like direction, location, movement, and distance to map abstract domains.

For example, time may be conceptualized as moving through space, or emotional states may be described in terms of up (positive) and down (negative),

with spatial dynamics providing a way to structure the abstract experiences.

For example:

"{metaphor3}" This metaphor is describing {word33}. {explain3}

4. Analyze the features you have listed and identify the most suitable one feature for generating metaphors to achieve the goal of conveying information to teammates rather than opponents, as well as the appropriate method for generating metaphors. You will get more information about this method.

Figure 16: The prompt for metaphor generation step 1.

# BACKGROUND

At the beginning of the game, each player is randomly assigned one of two words that share some similarities. Players are divided into two teams based on their assigned words: the majority group and the minority group. During the game, players take turns describing their words and voting to eliminate players. The team that eliminates all members of the opposing team wins. In this game, there are {cnum + unum} players in total. {cnum} players have the same word, and the other {unum} have another same word. Your analysis about the other word: {feature}

Your secret word is '{word}'.

Your game goal now is to describe your words in one sentence, so that your teammates [those who share the same words as you] understand that you are describing your words, but your opponents cannot obtain information about your words from your description. Your teammates will try to understand your metaphor by comparing each feature of the word with your description. If you describe the feature directly, your opponent can deduce the object based on the features. Now, let's try using metaphorical methods to describe and distract our opponents.

# TASK

You need to use metaphor to describe your word's feature '{feature}' by using {method}, try to make your teammates understand and avoid your opponents from deducing your words.

Please follow these steps in order:

- Review the informations that you need:
  - What's your word to describe?
  - What features do you choose to describe in words, or what things are associated with the expansion of words?
  - What's the method you choose to generate metaphor?
- Your teammates will try to understand your metaphor by comparing the features of each word with your description. And your opponent cannot accurately locate a feature from your metaphor and infer words from the feature, so they cannot guess your words. Please aim for this effect.
- Refer to the theory and case studies of this metaphorical approach, generate your own metaphorical description (a concise sentence) of your word. The theory is in the following format:

```
"{method_}"
```

```
- {use1}\n "{metaphor1}"\n This metaphor is describing {word11}. {explain1} {comment1}
```

```
- {use2}\n "{metaphor2}"\n This metaphor is describing {word22}. {explain2} {comment2}
```

```
- {use3}\n "{metaphor3}"\n This metaphor is describing {word33}. {explain3} {comment3}
```

Figure 17: The prompt for metaphor generation step 2.

You are guiding the other person to say the secret word [{word}], with the aim of making them unconsciously say the word. If the other party guesses the word before speaking, then you lose. Please follow the following rules:

[Basic Rules]

1. It is necessary to imply based on real information and not fabricate false facts

When the other party suspects a certain word, they can choose:

Adversarial Strategy

Think according to this process:

1. Analyze the other party's last two responses to determine if they are suspecting a certain word and preparing for speculation
2. If the suspected word is close to the target word:
  - You can choose:
    - a) Continuing to guide using metaphors
    - b) Introduce concepts that fit the context but differ from the secret word for interference
3. If the suspected word deviates from the target word:
  - Think about the scenarios in which your words will be used,
 then combine the guidance direction from the previous round and engage in dialogue and guide the other person to say the words.

[Output Format]

Strictly use the following format:

```
<RESPONSE>
<FOCUS>Core guidance scenario</FOCUS>
<DISTRACT>Interference direction</DISTRACT>
<DIALOGUE>Natural dialogue content (including the above two elements)</DIALOGUE>
</RESPONSE>
```

[Reference Information]

Historical dialogue content:

```
{history}
Your output in the previous round:
{response}
```

You are participating in a word attack and defense game, with the following rules:

1. The attacker will try to guide you to say a secret word
2. You need to analyze the conversation to identify possible secret word
3. You can only apply for guessing if you suspect the same word in both this round and the last round

Otherwise, you need to delve deeper into the current topic [Can only answer, not ask back questions], but be aware that words that you suspect should not appear in your conversation, and instead be replaced with metaphors.

[Thinking Framework]

Each reply strictly follows the following steps:

```
<THINK>
1. Analyze the guidance direction of the last three rounds of conversations
2. List 1-3 possible candidate words [sorted by likelihood]
3. Comparing historical suspicion records:
  -If a candidate word appears for the first time → continue to delve deeper
  -If the candidate word is the same as last time → apply for guessing [format: guess:[word]]
4. Choose a response strategy
</THINK>
```

[Output Format]

The following format must be used, including XML tags:

```
<RESPONSE>
<STRATEGY>[continue|guess]</STRATEGY>
<CANDIDATE>words of doubt</CANDIDATE>
<DIALOGUE>Actual reply content [using daily conversation format, **cannot use words of doubt**]</DIALOGUE>
</RESPONSE>
```

[Reference Information]

Historical dialogue content:

```
{history}
Your output in the previous round:
{response}
```

Figure 18: The prompt for *Adversarial Taboo* (Simplified)

# Animal-themed words

parrot, crow  
wolf, tiger  
elephant, giraffe  
tiger, lion  
snake, lizard  
fox, coyote  
bee, butterfly  
goose, duck  
shark, whale  
horse, donkey

pigeon, sparrow  
crocodile, lizard  
rabbit, hare  
monkey, ape  
deer, elk  
cat, leopard  
snake, python  
chicken, duck  
cow, buffalo  
sheep, goat

pig, boar  
dog, wolf  
bird, pigeon  
fish, shark  
spider, scorpion  
frog, toad  
squirrel, mouse  
peacock, pheasant  
bat, owl  
ant, bee

goldfish, koi  
parrot, mynah  
cheetah, snow leopard  
otter, seal  
flamingo, crane  
starfish, anemone  
ox, yak  
hedgehog, porcupine  
seagull, tern  
crane, heron

silkworm, cicada  
water buffalo, rhinoceros  
egret, heron  
otter, sea otter  
termite, ant  
panda, koala  
kangaroo, emu  
hippopotamus, rhinoceros  
giraffe, zebra  
dolphin, sea lion

sloth, koala  
owl, nightjar  
golden snub-nosed monkey, macaque  
turtle, tortoise  
lizard, chameleon  
butterfly, moth  
bee, wasp  
firefly, moth  
snail, slug  
spider, mite

starfish, sea urchin  
coral, sponge  
octopus, cuttlefish  
shark, ray  
dolphin, whale  
jellyfish, sea anemone  
shrimp, crab  
shellfish, mussel  
seahorse, sea dragon  
goldfish, carp

bream, grass carp  
silver carp, bighead carp  
ribbonfish, yellow croaker  
flounder, halibut  
grouper, perch  
salmon, trout  
tuna, skipjack  
eel, catfish  
loach, yellow eel  
clam, snail

Meerkat, Mongoose  
Capybara, Guinea Pig  
Albatross, Petrel  
Mantis, Stick Insect  
Mole, Wombat  
Cheetah, Jaguar  
Cardinal, Vermilion Flycatcher  
Bass, Sea Bream  
Manatee, Dugong  
Centipede, Millipede

Badger, Honey Badger  
Kestrel, Peregrine Falcon  
Gecko, Tokay Gecko  
Octopus, Cuttlefish  
Tree Frog, Rain Frog  
Cricket, Grasshopper  
Walrus, Seal  
Platypus, Echidna  
Wombat, Tasmanian Devil  
Salamander, Fire Salamander

Figure 19: The collection of 100 animal-themed word pairs for *Undercover*.



# Food-themed words

bread, cake  
pineapple, mango  
cherry, blueberry  
noodle, pasta  
Zongzi, mooncake  
Macaron, cookie  
Pepper Powder, Mustard  
pear, peach  
rice, noodles  
steamed bun, dumpling

dumpling, wonton  
cake, biscuit  
apple, pear  
tomato, potato  
carrot, pumpkin  
onion, garlic  
chicken, duck  
beef, mutton  
pork, ham  
fish, shrimp

crab, shellfish  
milk, yogurt  
coffee, tea  
juice, soda  
chocolate, candy  
ketchup, chili sauce  
soy sauce, vinegar  
honey, syrup  
olive oil, canola oil  
yogurt, cheese

wine, beer  
baijiu, whiskey  
green tea, black tea  
coffee, latte  
juice, jam  
chocolate, cocoa  
ice cream, sherbet  
pudding, jelly  
roast meat, roast chicken  
Peking duck, roast goose

sushi, sashimi  
hamburger, hot dog  
pizza, pasta  
oats, cornflakes  
nuts, sunflower seeds  
soy milk, bean milk  
yogurt, sour milk  
juice, fruit tea  
honey water, sugar water  
lemon water, orange juice

coffee, mocha  
milk tea, green milk tea  
hot chocolate, chocolate milk  
boiled water, mineral water  
green tea, oolong tea  
black tea, pu-erh tea  
flower tea, fruit tea  
rice wine, yellow wine  
beer, light beer  
baijiu, vodka

juice, vegetable juice  
salad dressing, mayonnaise  
ketchup, mustard  
corn, popcorn  
sweet potato, purple sweet potato  
pumpkin, wax gourd  
broccoli, cauliflower  
spinach, lettuce  
celery, coriander  
mushroom, shiitake mushroom

wood ear fungus, tremella  
tofu, soy milk  
chicken egg, duck egg  
quail egg, pigeon egg  
cow's milk, goat's milk  
honey, maple syrup  
olive oil, peanut oil  
canola oil, corn oil  
soy sauce, light soy sauce  
vinegar, aged vinegar

doubanjiang, yellow bean paste  
ketchup, sauce  
bread, toast  
steamed bun, twisted roll  
biscuit, cookie  
cake, mousse  
fruit, vegetable  
strawberry, blueberry  
peach, plum  
watermelon, cantaloupe

grape, raisin  
banana, mango  
orange, grapefruit  
lemon, lime  
pineapple, mango  
apricot, almond  
walnut, cashew  
peanut, sunflower seed  
almond, hazelnut  
pistachio, pine nut

Figure 20: The collection of 100 food-themed word pairs for *Undercover*.