Dynamic Coalition Structure Detection in Natural Language-based Interactions

Abhishek N. Kulkarni*
University of Texas at Austin
Austin, USA
abhishek.kulkarni@austin.utexas.edu

Andy Liu*
Carnegie Mellon University
Pittsburgh, USA
andyliu@cs.cmu.edu

Jean-Raphaël Gaglione
University of Texas at Austin
Austin, USA
jr.gaglione@utexas.edu

Daniel Fried
Carnegie Mellon University
Pittsburgh, USA
dfried@cs.cmu.edu

Ufuk Topcu
University of Texas at Austin
Austin, USA
utopcu@utexas.edu

ABSTRACT

In strategic multi-agent sequential interactions, detecting dynamic coalition structures is crucial for understanding how self-interested agents coordinate to influence outcomes. However, natural-language-based interactions introduce unique challenges to coalition detection due to ambiguity over intents and difficulty in modeling players' subjective perspectives. We propose a new method that leverages recent advancements in large language models and game theory to predict dynamic multilateral coalition formation in Diplomacy, a strategic multi-agent game where agents negotiate coalitions using natural language. The method consists of two stages. The first stage extracts the set of agreements discussed by two agents in their private dialogue, by combining a parsing-based filtering function with a fine-tuned language model trained to predict player intents. In the second stage, we define a new metric using the concept of subjective rationalizability from hypergame theory to evaluate the expected value of an agreement for each player. We then compute this metric for each agreement identified in the first stage by assessing the strategic value of the agreement for both players and taking into account the subjective belief of one player that the second player would honor the agreement. We demonstrate that our method effectively detects potential coalition structures in online Diplomacy gameplay by assigning high values to agreements likely to be honored and low values to those likely to be violated. The proposed method provides foundational insights into coalition formation in multi-agent environments with language-based negotiation and offers key directions for future research on the analysis of complex natural language-based interactions between agents.

KEYWORDS

Coalition Structures; Game Theory; Multi-Agent Cooperation; Large Language Models

* Equal contribution.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

ACM Reference Format:
Abhishek N. Kulkarni*, Andy Liu*, Jean-Raphaël Gaglione, Daniel Fried, and Ufuk Topcu. 2025. Dynamic Coalition Structure Detection in Natural Language-based Interactions. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 8 pages.

1 INTRODUCTION

The process of coalition formation in multi-agent systems involves agents forming coalitions to work together towards aligned objectives by coordinating their actions [33, 38]. This process has been studied extensively for both static and dynamic cases in game theory and logic. Past game-theoretic approaches have focused on studying which coalitions are likely to form based on various types of equilibria [18] and evaluating the value of a coalition to an agent. Meanwhile, logic-based approaches have focused on evaluating whether a given coalition can enforce a temporal property, regardless of how the agents not in the coalition behave [1, 29]. However, neither of these approaches is suitable for studying coalition formation in natural language negotiation, where the ambiguity in language often leads players to interpret game states differently. This phenomenon is commonly observed in real-world scenarios such as human-robot teaming [8], computer games [23, 32].

In this work, we study the problem of predicting dynamic multilateral coalition structures in sequential multi-agent interactions where players coordinate their actions using natural language. A coalition structure [16] is a graph where nodes represent players and the edges represent agreements between players over coordinated actions. While traditional approaches define coalition structures to be a partition of players, we model a coalition as a multi-graph allowing multiple agreements between two players, in addition to allowing a player to form bilateral agreements with multiple players simultaneously. We model multi-agent interactions as reactive games, where the player can renegotiate their agreements in every round. Our aim is to predict these coalition structures from the perspective of an external observer, similar in spirit to an agency monitoring a computer network for anomalous behavior.

We use the board game Diplomacy [7] as a testbed for dynamic coalition structure detection over natural-language negotiations. Diplomacy is a seven-player board game that exemplifies key challenges in multi-agent systems research, combining semi-cooperative

strategic dynamics with natural language-based negotiation. Players aim to control a majority of 34 supply centers on a map of Europe by coordinating the movement of units. While Diplomacy is a zero-sum game, players must negotiate strategic coalitions to support their own plans or counteract the moves of other players.

Diplomacy highlights three key challenges central to studying coalition formation in games with natural-language negotiations: *decision-making under incomplete information*, *reasoning based on mental models of opponents*, and *multilateral negotiations*. Since the negotiations are pairwise and private, each player has incomplete information about the negotiations a second player has had with other players. As a result, players must anticipate others' actions without full knowledge of all agreements. This requires a player to construct a mental model of the other player's incentives to estimate the likelihood they would honor the agreement in addition to weighing their own incentives to honor the agreement. Lastly, since a player can simultaneously negotiate multiple agreements about the same unit with different players, a player must ultimately select a subset of agreements to honor based on the strategic advantage they offer to the player and the likelihood of them being honored by the player with whom the agreement is made.

We define a novel method that addresses these challenges in dynamic coalition structure detection over natural language negotiations, as visualized in Figure 1. Our approach consists of two stages: **agreement detection** and **strategic reasoning**. To detect agreements negotiated via natural language in Diplomacy, we leverage large language models to parse dialogue, as well as fine-tuned "intent" models from CICERO [11], a Diplomacy-playing agent. By comparing the distribution over all moves involving parsed territories for a given unit before and after a phase of dialogue, we can learn whether a coalition was formed in the dialogue for that phase. Given a set of potential agreements identified, we then predict the set of honored agreements, which defines a coalition structure, using a deep reinforcement-learning based method. Due to the large action space and incomplete-information nature of Diplomacy, traditional enumerative game-theoretic approaches are intractable. To address these challenges, we extend the approach in [4] to compute the strategic value of an agreement for both players. We then sample from the intent model to determine the likelihood that both players will uphold the coalition, allowing us to measure the rationalizability of a coalition structure.

The three main contributions of this work are:

(1) **Approach.** We introduce a novel method that integrates large language models and game theory to predict dynamic multilateral coalition formation in multi-agent systems where agents negotiate coalitions using natural language.
(2) **Agreement detection.** We develop a procedure that combines pretrained language models with game dynamics to extract agreements from dialogues between agents, enabling the detection of coalition structures in real-time interactions.
(3) **Strategic reasoning.** We propose a new metric based on subjective rationalizability from hypergame theory, which evaluates the likelihood that agents will adhere to agreements by accounting for their subjective views of the game and strategic uncertainty.

We validate our method on a dataset of online Diplomacy gameplay experiments. We find that our hybrid agreement detection outperforms existing baselines and that our rationalizability metric effectively distinguishes between when players will honor coalition agreements and when they will not. These findings highlight the value of integrating natural language techniques with game-theoretic analysis. They extend existing game-theoretic dynamic coalition prediction approaches to handle natural language negotiations, bridging toward more realistic real-world applications.

1.1 Related Work

Game theory. The study of coalition formation in game theory focuses on identifying and characterizing stable coalitions by estimating the value of possible coalitions to an agent. Solution concepts such as the core [2], the kernel [37], the nucleolus [26], and the Shapley value [3] have been introduced to analyze stability in transferable utility games, where side payments are allowed, and non-transferable utility games, where they are prohibited. However, these approaches do not account for sequential interactions where agents strategically join coalitions to achieve their goals.

The study of dynamic coalition formation in game theory has mainly focused on understanding the effects of externalities, where the formation of one coalition impacts the gains of other co-existing coalitions. [31] presented a computational study of coalitional games with externalities, arguing that such externalities are common in real-world settings. The study of such externalities was extended by work including [24, 39, 40]. While these approaches were able to better capture the impacts of externalities on other coalitions, they assume that complete information available to all agents, which is not applicable to games such as Diplomacy.

Logic. Strategic decision-making within coalitions has been studied within the logic community. Coalition Logic [29] and Alternating-time Temporal Logic [1] formalize reasoning about the existence of joint strategies for agents in coalitions to achieve their goals regardless of how the non-coalitional agents act. However, these logics only study static coalitions.

There is an ongoing effort to extend coalition logic to handle dynamic settings. In [41], authors introduce coordinated coalitions that represent a predefined sequence of coalitions for model checking. [17] presents a complex framework that enriches Concurrent Game Models (CGM) by incorporating negotiations, where promises—represented as epistemic logic formulas—are embedded into states and existence of strategies that ensure goal satisfaction are verified through model checking. These methods requires full access to the game model, which is impractical for large-scale games like Diplomacy. Moreover, they model negotiations as deterministic statements, failing to capture the inherent ambiguity of natural language.

Negotiation in Natural Language. While significant past work has studied negotiation and coordination as a natural language task, the analysis of coalition formation in games involving natural language negotiations remains relatively understudied. [21] collected a dataset of human negotiation dialogues in a semi-cooperative negotiation task, then trained natural language agents to perform the same task using the dataset. More recently, large language model-based agents have been used to achieve stronger performance on a

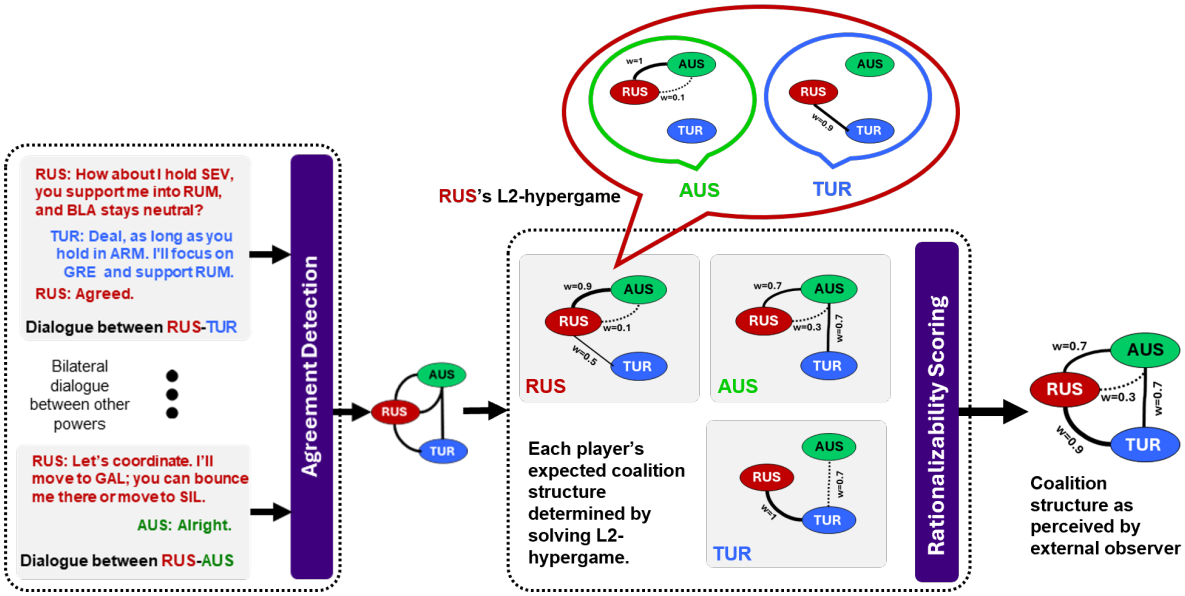

Figure 1: Proposed two-stage approach for learning coalition structures from natural language interactions in Diplomacy games. Stage 1 extracts agreements from pairwise dialogues to form an unweighted coalition structure. Stage 2 applies hypergame theory to assess the rationalizability of agreements for each player separately, which are then integrated into a weighted coalition structure representing the likelihood that an external observer believes agreements will be honored.

range of social influence tasks [9], including negotiation in a self-play environment over both zero-sum [13] and non-zero-sum [22] games. [14] demonstrates that incorporating more explicit search and belief tracking into language models can improve their negotiation performance over a wide range of environments. [25] specifically seeks to model political coalition formation with language model-based agents, arguing that previous language model-based approaches to negotiation do not fully capture the full complexity and iterative nature of human negotiations. They contribute a multilingual dataset of European political party manifestos, as well as coalitions that they formed with other parties. While we similarly seek to model the multi-issue, iterative dynamics of natural-language coalition formation, we additionally analyze this problem from a game-theoretic perspective that accounts for agents' models of each other.

Diplomacy has also attracted attention from the natural language community as a testbed for the analysis of coordination dynamics in a strategic multi-agent environment. [28] studies the formation and termination of long-term alliances from a linguistic perspective, finding linguistic cues that presage acts of betrayal. [30] models deception over long-term relationships in Diplomacy, finding that a model that uses both game dynamics and dialogue cues can predict player deception at a near-human level. [42] analyzes games between CICERO and human Diplomacy players, noting that despite CICERO's strong strategic capabilities, it is still less persuasive compared to human players. Finally, [27] devises a novel positive-sum variant of Diplomacy, finding that language model-based agents are capable of attaining high joint welfare in this setting.

2 PROBLEM FORMULATION

We use Diplomacy as a testbed to study dynamic multilateral coalition formation. Diplomacy is a deterministic game where players negotiate before concurrently submitting actions for their units, such as hold, move, or support. The game state transitions based on these actions, and the negotiation phase repeats. In this paper, we only consider the movement phases of the Diplomacy game. We note the high complexity of the game: each unit has an average of 26 valid orders, with up to 34 units on the board, making enumerative approaches intractable [4].

Game model. Diplomacy can be modeled as a concurrent multiplayer game [1] with rewards, $G = (N, S, A, T, s_0, R)$, where $N = \{P_i \mid i = 1, 2, \ldots n\}$ is a set of players, S is a set of states, A is a set of actions, $T : S \times A \to S$ is a deterministic transition function, $s_0 \in S$ is an initial state, and $R : S \times A \to \mathbb{R}$ is a reward function. A game play in G is determined in two phases: Given a state $s \in S$, the players in N *privately* negotiate non-binding bilateral agreements with each other.

Definition 1 (Agreement)**.** Given a state $s \in S$, an agreement between two players P_i, P_j is a tuple (u_1, u_2, a_1, a_2), where u_1 is a unit controlled by P_i, u_2 is a unit controlled P_j, and a_1, a_2 are legal actions for units u_1, u_2 in state s.

The content of these negotiations and the agreements are only known to the players involved in the negotiation, unless one of these players explicitly shares this information with other players. At the conclusion of negotiation phase, all players choose an action, assigning an order to each unit controlled by them. Together, these actions determine the joint action $a = (a_1, a_2, \ldots, a_n)$, which in turn uniquely determines the next state $s' = T(s, a)$. We denote

by d_t the set of all natural language messages exchanged between any pair of players in round t. Therefore, a game can be denoted as the sequence of state-dialogue-action pairs, $\rho = s_0 d_0 a_0 s_1 d_1 \dots s_n$. A game in Diplomacy is of finite duration since a player will either win the game, or the game will be declared a draw. For a more detailed description, see [36].

A policy for a player P_i is a map $\pi_i : S \to \mathcal{D}(A_i)$, where $\mathcal{D}(A_i)$ is a set of probability distributions over actions A_i of player P_i. A policy profile is a collection of policies of all players, $\pi = (\pi_1, ..., \pi_n)$.

Coalition structure. A coalition is a set of honored agreements. The coalition structure, given a game state, is represented as an undirected multigraph with players as nodes and parallel edges indicating agreements between them. Since agreements are inferred from potentially ambiguous natural language dialogue, we assign weights to the edges. Intuitively, these weights represent the likelihood of each agreement being honored.

Definition 2. A coalition structure is a graph $C = (N, E, \text{Agmt}, \text{wt})$, where N is the set of players, Agmt is a set of agreements, $E \subseteq N \times N \times \text{Agmt}$ is the set of edges, and $\text{wt} : E \to \mathbb{R}$ is a function that assigns a real-valued weight to each edge.

Note that in Diplomacy, the coalition structure is not static; we denote the coalition structure in round t as C_t. This background motivates the problem of **coalition structure prediction**.

Problem 1. Given a concurrent multiplayer game G, a round $t \geq 0$, and the play $\rho = s_0 d_0 a_0 \dots s_t d_t$ until round t, predict the coalition structure C_t.

3 BACKGROUND: HYPERGAME THEORY

Diplomacy is characterized by both *incomplete information* and *unawareness*. In Diplomacy, players make decisions without full knowledge, as they may be unaware of message exchanges between other players or the content of those messages. As a result, a player's rationality must be assessed based on their subjective view of the game, shaped by their knowledge (c.f. [19, 20]).

Hypergame is a game-theoretic model designed for games with incomplete information and unawareness [5, 35]. In a hypergame, each player has a subjective view of their interaction, shaped by their knowledge of the game and others' perspectives. This structure allows players to independently form subjective views and make decisions based on their own *subjective game*, effectively capturing player unawareness within the model.

Formally, hypergames are defined inductively based on players' levels of perception. A level-0 (L0) hypergame represents a game with complete, symmetric information, where both players have the same perception of the game, identical to the true game. In a level-1 (L1) hypergame, at least one player misperceives the game, but neither player is aware of this discrepancy. Each player believes their perceptual game is the true game and plays accordingly, with these perceptual games being level-0 hypergames. In a level-2 (L2) hypergame, one player becomes aware of the misperception and can reason about the other player's perceptual game. This concept can be extended to higher hypergame levels; however, in this paper, we restrict ourselves to L2-hypergames, which are most directly relevant to Diplomacy.

Subjective rationalizability [34] is a solution concept for hypergames that evaluates the rationality of players' actions based on their subjective views of the game, considering their knowledge and beliefs about other players' perspectives and actions.

Definition 3 (Subjective Rationalizability)**.** Let $H^2 = \langle H_1^1, H_2^1 \rangle$ denote a L2-hypergame, where $H_i^1 = (G_1^i, G_2^i)$ is player P_i's L1-hypergame and G_1^i is the subjective game of P_1 as perceived by P_i. Then, a strategy $\pi_i^{*,2}$ is said to be subjectively rationalizable for player P_2 if and only if it satisfies the following condition for all π_i:

$$u_i^2(\pi_i^{*,2}, \pi_j^{*,2}, x) \geq u_i^2(\pi_i, \pi_j^{*,2}, x),$$

where $(i, j) \in \{(1, 2), (2, 1)\}$ and x is a distribution over Φ representing P_2's hypothesis over some aspect of P_1's game. In this case, the utility is calculated based on the expectation, that is, $u_i^2(\pi_i, \pi_j, x) = \sum_{\varphi \in \Phi} x(\varphi) u_i^2(\pi_i, \pi_j, \varphi)$. The strategy $\pi_1^{*,1}$ is subjectively rationalizable for P_1 if and only if it satisfies the following condition for all π_1,

$$u_1^1(\pi_1^{*,1}, \pi_2^{*,2}, \varphi_1) \geq u_1^1(\pi_1, \pi_2^{*,2}, \varphi_1),$$

where $\pi_2^{*,2}$ is subjectively rationalizable for P_2.

Def. 3 enables evaluating when a player's strategy is rational within their own subjective view of the game. For P_2, a strategy is subjectively rationalizable if, given its information about P_1's game (H_2^1), P_2 cannot improve their utility by choosing a different strategy. Specifically, P_2's utility from its chosen strategy, given the other player's strategy and their own beliefs (represented by a distribution x), must be at least as high as the utility from any other strategy they might choose. Subjective rationalizability is understood similarly for P_1.

4 COALITION STRUCTURE PREDICTION METHODOLOGY

We introduce a two-stage approach as shown in Fig. (1) that integrates recent developments in LLMs with subjective rationalizability in hypergames to solve the problem of dynamic coalition structure prediction, as defined in Section 2.

The first stage identify the set of candidate agreements Agmt_t being discussed given a play ρ_t. We assume that all agreements are discussed in natural language and no side channels exist for forming agreements. Letting C_t be the coalition structure at round t, the set Agmt_t determines the set of edges of C_t.

The second stage assigns weights to the edges in Agmt_t, referred to as the *rationalizability score*. For an agreement $\alpha \in \text{Agmt}_t$, this score represents the likelihood that an external observer, with access to the full state, action, and dialogue history, believes that α will be honored by both players. To compute the rationalizability score $\text{wt}_i(\alpha)$, we estimate the likelihood of a player P_i honoring the agreement using a L2-hypergame constructed by filtering the messages to include only those exchanged between the two players involved in α.

Formally, the rationalizability score of an agreement α for a player P_i is computed by evaluating the strategic value (utility) of α for P_i in its hypergame H_i^1. Formally, it is given by

$$\text{wt}_i(\alpha) = V_i(\alpha) * V_j^i(\alpha),$$

where $V_i(\alpha)$ is the game-theoretic value of agreement for P_i and $V_j^i(\alpha)$ is P_i's belief about the likelihood of P_j honoring the agreement. Based on Def. 3, this weight reflects how subjectively rationalizable an agreement is for P_i, with higher weights indicating greater rationalizability.

Given the rationalizability scores $\mathrm{wt}_i(\alpha)$ and $\mathrm{wt}_j(\alpha)$ of P_i and P_j, respectively, the rationalizability score of the agreement α for an external observer is given by

$$\mathrm{wt}(\alpha) = \mathrm{wt}_i(\alpha) * \mathrm{wt}_j(\alpha).$$

Note that the proposed two stage approach separates the language-based reasoning from game-theoretic one. The strategic values V_i and V_j^i are derived from game-theoretic solution concepts and do not rely on dialogue. Whereas, the set of agreements Agmt_t is inferred directly from the dialogue.

In the remainder of this section, we first outline how agreements are identified from dialogue, followed by the computation of their rationalizability score.

4.1 Agreement Detection

To detect agreements in Diplomacy gameplay from game transcripts, we combine (1) a filtering stage where mentioned locations that a coalition can be formed over are extracted by a language model, and (2) an intent extraction stage where specialized Diplomacy models are used to extract player intents for classification.

Figure 2 outlines our method for agreement detection. First, for each state-players tuple (S, P_1, P_2), we first prompt GPT-4o[1], a strong language model, with the dialogue between the two players at state S and information about the Diplomacy board. We then use the model to extract all locations that were explicitly mentioned in negotiation between P_1 and P_2. In addition to information about whether two countries are sufficiently close to form a coalition, we will use this in a later filtering step.

We then leverage the intent models used in CICERO [11]; these are 2.7-billion parameter language models that predict player actions from dialogue. Specifically, they are trained using behavioral cloning over a subset of "truthful" player dialogues collected from WebDiplomacy. Notably, this intent model only takes the conversation between P_1 and P_2 for a given phase, excluding any dialogue either player had with other players, to restrict the model to direct coordination between the two players. By computing a distribution of move likelihoods over all possible moves for a unit before and after player dialogue, we can estimate whether a coalition was formed over the unit in question. We extract for each $(S, P_1, P_2, u \in u_1 \cup u_2)$ a most likely action a^* for the unit in this state. We also compute the probability of a^* before and after dialogue, as well as the entropy of the distribution of moves $P(a|S, P_1, P_2)$ before and after dialogue. After filtering out all units where a coalition is not possible, or where none of the territories involved in the move are mentioned in the dialogue, we then train a logistic regression classifier on these features to predict whether a coalition was formed over the unit in question.

This method allows us to leverage the advantages of both using a larger, general language model and a smaller, Diplomacy-specialized language model. While using the intent model allows us to capture more implicit coalition agreements that may not be identified with an explicit parser, it may also raise many false negatives due to noise in how the distribution changes as a result of unrelated dialogue. Adding a filtering step allows us to identify cases where the distribution shifts due to identifiable discussion of the provinces in question, as identified by the larger model. Indeed, in Section 5.2, we show that this hybrid method outperforms methods that only rely on large language model annotation or learning from intent distributions.

[1]https://platform.openai.com/docs/models/gpt-4o

4.2 Strategic Value of Agreements

Determining the strategic value $V_i(\alpha)$ of an agreement $\alpha = (u_1, u_2, a_1, a_2) \in \mathrm{Agmt}$ for a player P_i is a challenging task in large games like Diplomacy. It requires P_i to determine the rational actions for all units controlled by P_i as well as the other players conditioned on the unit u_1 being assigned action a_1 and the unit u_2 being assigned action a_2. Traditional game-theoretic approaches [15] enumerate all possible actions and evaluate them under a solution concept to determine the action that yields highest value from a given state. These approaches are inapplicable to games like Diplomacy due to the large size of players' action spaces.

Instead, we employ a deep reinforcement learning approach that first learns a probability distribution

$$\Pr(a \mid s_0, \ldots, s_t, a_0, \ldots, a_{t-1}, \alpha) \tag{1}$$

over joint actions of all players conditioned on P_i and P_j honoring a given agreement in addition to the state and action histories. Intuitively, every joint action in the support of the distribution in Eq. (1) constitutes a Nash equilibrium in which P_i and P_j honor the agreement α.

Learning joint action distribution. We leverage order sampling models trained as part of the CICERO agent [11], which use Double oracle reinforcement learning for action exploration (DORA) [4] to learn the distribution in Eq. (1). DORA simultaneously learns a state-value function and an joint action probability distribution using neural networks trained by bootstrapping on an approximate Nash equilibrium for the stage game each turn.

DORA is a Nash Q-Learning based approach to approximate Nash equilibrium in games with large state and action spaces. It accommodates the large action spaces of Diplomacy by training a neural network $\pi(s; \theta_\pi)$ to predict joint action probability distribution with parameters θ_π that approximates the distribution of actions under the equilibrium policy at state s. The candidate actions to explore are determined by sampling a large number of actions from $\pi(s; \theta_\pi)$ for each player and selecting actions with highest likelihood. The Nash equilibrium is then estimated using regret minimization [12] in the matrix sub-game that includes only the sampled actions, assuming that the values of successor states are given by a learned network $\mathbf{V}(s; \theta_v)$, using the following update equation:

$$\mathbf{V}(s) \leftarrow (1 - \beta)\mathbf{V}(s) + \beta(r(s) + \gamma\sigma(a)\mathbf{V}(T(s, a))).$$

We refer interested readers to [4] for more details on the implementation. While we rely on CICERO-trained models for this work, versions of all of the specialized Diplomacy models used can be trained in novel game settings where human data is available [4].

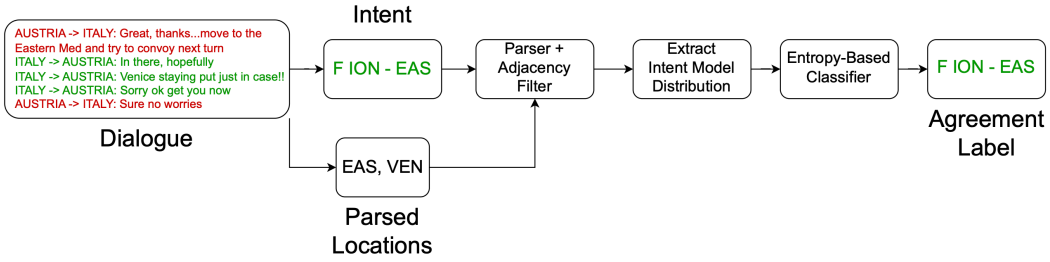

Figure 2: An overview of our agreement detection framework. In this case, we are analyzing whether Italy and Austria have come to an agreement over Italy's unit F ION, and determine that an agreement has been reached for Italy to move this unit to the Eastern Mediterranean Sea (EAS).

Value of agreement. The order sampling model and the value model provide a way to determine not only the joint action probabilities conditioned on an agreement, but also the value of the resulting state. Hence, we determine the value of an agreement by sampling from these distributions and computing the expected value of next state reached by the player by honoring α,

$$V_i(\alpha) = \sum \Pr(a \mid \vec{s}_t, \vec{a}_{t-1}, \alpha)\mathbf{V}(s'), \tag{2}$$

where $s' = T(s, a)$ is the new state reached when joint action a is performed in state s.

4.3 Perceived Value of Agreement to Opponent

While Eq. (2) determines the value of an agreement for P_i, it does not allow P_i to estimate its value for P_j due to incomplete information about P_j's negotiations with others. Instead, P_i must infer P_j's intent from their mutual dialogue and by estimating the value of various actions in the current state for P_j.

We interpret a players' intent as a probability distribution over the actions they assign to their units in the next round. To estimate the likelihood that P_j will honor an agreement from P_i's perspective, we approximate the intent distribution discussed in Sec. 4.1. This enables us to extract action probabilities from the dialogue to inform the agreement value computation.

Given the distribution in Eq. (1), we compute the likelihood of P_j respecting α using the following equation. We denote the support of a probability distribution $\mathbf{d}$ by $\text{Supp}(\mathbf{d})$. Given a joint action $a \in A$, let $\mathbf{1}_j(a, \alpha) \mapsto \{\top, \bot\}$ denote whether the action a assigns the same action with P_j's unit as that assigned under α.

$$V_j^i(\alpha) = \sum_{a \in \text{Supp}(\mathbf{d})} \frac{\beta \Pr(a \mid \vec{s}_t, \vec{d}_t, \vec{a}_{t-1})}{\beta \Pr(a \mid \vec{s}_t, \vec{d}_t, \vec{a}_{t-1}) + (1-\beta)\Pr(a' \mid \vec{s}_t, \vec{d}_t, \vec{a}_{t-1})},$$

where $\beta = \mathbf{1}_j(a, \alpha)$ and $a' \neq a$ is a valid action assigned to unit u_2 by P_j.

Intuitively, $V_j^i(\alpha)$ measures the relative value that P_j achieves by selecting an action that honors α when compared with selecting an action that does not honor α, as players are more likely to honor agreements that are more strategically advantageous.

5 EXPERIMENTS

We evaluate the two stages of our proposed method separately. First, we outline the dataset employed for evaluation, and then present the results for each of the two stages.

5.1 Dataset

We source previous Diplomacy games from WebDiplomacy[2], a multiplayer online implementation of Diplomacy. We consider a dataset of 140 full-press games played over the standard Diplomacy map. In order to calibrate our agreement detection classifier, we manually annotate five games from this dataset, randomly sampling from games with at least 250 total messages sent. This gives us a total of 16962 (S, P_1, P_2, u) tuples over 1603 combinations of a state S and players P_1, P_2. This dataset is highly imbalanced, with 444 (2.6%) of all (S, P_1, P_2, u) tuples having a coalition formed.

After we validate our usage of the agreement detection classifier, we then use it to label the remainder of the games with detected agreements. This resulting dataset consists of 415001 total (S, p_1, p_2, u) tuples. Of these tuples, 11008 have agreements detected by our automatic method, 8344 of which are upheld (i.e. the player played the agreed-upon move).

5.2 Validating Agreement Detection Method

We use the manually-annotated data sample described in Section 5.1 to test our agreement detection method, with an 80-20 train-test split. While this dataset is strongly imbalanced, we mitigate the impact of the dataset imbalance by only training on instances that pass our language model-based filter, which reduces our classifier training data to 1768 tuples, and by tuning a classification threshold to optimize F1-score on our training dataset. We benchmark three methods on this dataset:

- **GPT-4o**, which prompts a strong language model to directly identify units over which an agreement has been reached,
- **Classifier**, which trains a classifier on intent model distributions before and after dialogue over all (S, P_1, P_2, u) tuples in the dataset, and

[2]https://webdiplomacy.net/

Method	F1 Score	Precision	Recall
GPT-4o	0.34	0.26	0.47
Classifier	0.44	0.43	0.45
Hybrid	**0.55**	**0.63**	**0.48**

Table 1: Classification metrics over the test dataset for our three methods. Hybrid methods outperform both purely language model-based and intent model distribution-based approaches at detecting whether an agreement has been reached over a specific unit.

		Hand-Labelled			Hybrid		
Honored?	Metric	Value	R-Score@1	R-Score@5	Value	R-Score@1	R-Score@5
Yes	MRR (↑)	0.2842	0.9444	**0.9722**	0.3602	0.7416	**0.8294**
No	MRR (↓)	0.2583	**0.0**	0.125	0.3682	**0.2628**	0.3354
Yes	Brier (↓)	0.0802	**0.0422**	**0.0422**	0.0739	**0.0311**	**0.0311**
No	Brier (↑)	0.7303	**0.7494**	**0.7494**	0.5706	**0.6145**	**0.6145**

Table 2: Evaluation metrics for honored and violated agreements based on hand-labeled and hybrid datasets. The ranking induced by rationalizability score (RScore) on the set of agreements assigns lower ranks to honored agreements and higher ranks to violated agreements when compared to the ranking induced by Nash approximation-based predictions.

- **Hybrid**, our approach, which first filters using GPT-4o-parsed locations and player adjacency before training a classifier on the filtered data.

The results of our evaluation are in Table 1. Our hybrid method outperforms both training a classifier on unfiltered intent data and prompting a strong language model on identifying units over which agreements have been reached. Extraction of coalition agreements from Diplomacy dialogue is a challenging task, due to the length of many dialogues in Diplomacy as well as the implicitness and fluctuating nature of negotiation over a multi-party dialogue. Fine-tuning more generally capable language models following the intent model formula in CICERO, in combination with more sophisticated parsers such as the one trained in [42], could yield even stronger performance improvements, which we leave to future work in this direction.

5.3 Evaluating Rationalizability Score

The rationalizability score establishes a ranking of potential agreements for a unit within a specific game state. To evaluate the effectiveness of the score in predicting coalition structures, we analyze the rankings induced by the score on honored agreements in comparison to those of violated agreements.

We utilize both hand-labeled data and data labeled through the hybrid approach for the evaluation. We consider a total of 7434 agreements labeled using the hybrid approach for evaluation. For each agreement identified in the agreement detection stage, we generate a set of alternative agreements by sampling different orders for the units involved in the agreement. The results of this evaluation for honored and violated agreements are presented in Table 2. Given that the output of our model is a ranked list based on the rationalizability score, we employ two information retrieval metrics: mean reciprocal rank (MRR) [10] and Brier score [6]. The MRR is calculated using both the top-1 and top-5 ranked elements.

The ranking generated by the rationalizability score effectively differentiates between honored and violated agreements. Our findings indicate that honored agreements typically receive lower ranks, while violated agreements tend to rank higher. This is observed through both the MRR and Brier scores. When calculating the Brier score, we normalize the rationalizability scores, such that a score close to 1 reflects that honored agreements usually have low ranks and violated agreements have higher ranks. Notably, MRR scores that are close to 1 in the top-1 case suggest that honored agreements are frequently assigned a rank of 0, suggesting that this metric can very precisely recognize upheld coalitions.

We also compare our rationalizability score to a more conventional coalition formation prediction method, approximate Nash equilibrium (as estimated by the CICERO value model). We find that even when such approximate equilibrium-based methods are adapted for games with large state and action spaces, they remain inadequate for predicting coalition formation in such dynamic environments. Our R-Score yields a significantly higher MRR and a lower Brier score than the value model score in all cases. This suggests that our rationalizability framework is significantly better at distinguishing between coalitions that are upheld and coalitions that are not upheld than Nash approximation-based predictions.

6 CONCLUSION

The detection of dynamic coalition structures is a key problem in understanding sequential interactions in strategic multi-agent environments. While many such environments use language, the study of coalition structure detection over natural language-based coordination is relatively understudied. This is compounded in settings like the board game Diplomacy, where players make decisions with incomplete information using dialogue-informed mental models of their opponents' future actions, and where relationships between players can shift drastically between turns as information is revealed.

Drawing from hypergame theory and the concept of subjective rationalizability, we propose a general method to dynamically predict coalition structures over sequential multi-agent interactions. In our method, we first extract detected agreements using the combination of a large language model-based parser and a specialized language model to predict player intents before and after the negotiation phase. We then compute the value of the agreement using a deep reinforcement-learning based value function, which we use in combination with player intents to compute the likelihood that each player will honor the agreement.

We validate the success of our method over sampled interactions between human Diplomacy players, using components of Meta's CICERO agent to compute player intents and action values. When compared to approximate Nash Equilibrium-based methods, our rationalizability score is significantly better at predicting the coalition structure at a given timestep. Our method can also generalize to other multi-agent, dialogue-based games, as long as sufficient human data exists upon which similar, game-specific models can be trained.

Extending coalition structure detection to natural language-based negotiation environments such as Diplomacy presents unique challenges in a setting where agents have incomplete information, negotiations are both multi-issue and multi-party, and where agents must reason over mental models of their opponents. However, for artificial agents to handle such complex environments properly, they must be capable of understanding the coalition dynamics of the environment at a given state. Our method and experiments serve as an important first step in this direction. We hope that future work will be able to extend our framework to new settings, including those with more complex negotiations and less existing domain-specific models, paving the way for agents that can reason over such information in deployment settings.

ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112490410, the Army Research Lab under Agreement ARO W911NF-23-1-0317 and the Office of Naval Research under Agreement N00014-24-1-2097. We thank WebDiplomacy for supporting this research by providing access to online gameplay data.

REFERENCES

[1] Rajeev Alur, Thomas A Henzinger, and Orna Kupferman. 2002. Alternating-time temporal logic. *Journal of the ACM (JACM)* 49, 5 (2002), 672–713.
[2] Tone Arnold and Ulrich Schwalbe. 2002. Dynamic coalition formation and the core. *Journal of economic behavior & organization* 49, 3 (2002), 363–380.
[3] Robert J Aumann and Roger B Myerson. 2003. *Endogenous formation of links between players and of coalitions: An application of the Shapley value.* Springer.
[4] Anton Bakhtin, David Wu, Adam Lerer, and Noam Brown. 2021. No-press diplomacy from scratch. *Advances in Neural Information Processing Systems* 34 (2021), 18063–18074.
[5] Peter G Bennett. 1980. Hypergames: developing a model of conflict. *Futures* 12, 6 (1980), 489–507.
[6] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.
[7] Allan Calhamer. 1974. The invention of diplomacy. *Games & Puzzles* 21 (1974).
[8] Tathagata Chakraborti, Kartik Talamadupula, Yu Zhang, and Subbarao Kambhampati. 2016. A formal framework for studying interaction in human-robot societies. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.*
[9] Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.* 750–766.
[10] Nick Craswell. 2009. Mean reciprocal rank. *Encyclopedia of database systems* (2009), 1703–1703.
[11] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. https://doi.org/10.1126/science.ade9097 arXiv:https://www.science.org/doi/pdf/10.1126/science.ade9097
[12] Dean P Foster and Rakesh Vohra. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* 29, 1-2 (1999), 7–35.
[13] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142* (2023).
[14] Kanishk Gandhi, Dorsa Sadigh, and Noah Goodman. [n.d.]. Strategic Reasoning with Language Models. In *NeurIPS 2023 Foundation Models for Decision Making Workshop.*
[15] Robert Gibbons et al. 1992. A primer in game theory. (1992).
[16] Joseph Greenberg. 1994. Coalition structures. *Handbook of game theory with economic applications* 2 (1994), 1305–1337.
[17] Dimitar P Guelev. 2023. Of Temporary Coalitions in Terms of Concurrent Game Models, Announcements, and Temporal Projection. In *International Workshop on Logic, Rationality and Interaction.* Springer, 126–134.
[18] Jana Hajduková. 2006. Coalition formation games: A survey. *International Game Theory Review* 8, 04 (2006), 613–641.
[19] Abhishek N Kulkarni and Jie Fu. 2021. Synthesis of deceptive strategies in reachability games with action misperception. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence.* 217–223.
[20] Abhishek N Kulkarni, Huan Luo, Nandi O Leslie, Charles A Kamhoua, and Jie Fu. 2020. Deceptive labeling: hypergames on graphs for stealthy deception. *IEEE Control Systems Letters* 5, 3 (2020), 977–982.
[21] Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2443–2453.
[22] Austen Liao, Nicholas Tomlin, and Dan Klein. 2024. Efficacy of Language Model Self-Play in Non-Zero-Sum Games. *arXiv preprint arXiv:2406.18872* (2024).
[23] Parisa Mazrooei, Christopher Archibald, and Michael Bowling. 2013. Automating collusion detection in sequential games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27. 675–682.
[24] Tomasz Michalak, Dorota Marciniak, Marcin Szamotulski, Talal Rahwan, Michael Wooldridge, Peter McBurney, and Nicholas Jennings. 2010. A logic-based representation for coalitional games with externalities. (2010).
[25] Farhad Moghimifar, Yuan-Fang Li, Robert Thomson, and Gholamreza Haffari. 2024. Modelling Political Coalition Negotiations Using LLM-based Agents. *arXiv preprint arXiv:2402.11712* (2024).
[26] Maria Montero. 2006. Noncooperative foundations of the nucleolus in majority games. *Games and Economic Behavior* 54, 2 (2006), 380–397.
[27] Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. [n.d.]. Welfare Diplomacy: Benchmarking Language Model Cooperation. In *Socially Responsible Language Modelling Research.*
[28] Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong and Michael Strube (Eds.). Association for Computational Linguistics, Beijing, China, 1650–1659. https://doi.org/10.3115/v1/P15-1159
[29] Marc Pauly. 2002. A modal logic for coalitional power in games. *Journal of logic and computation* 12, 1 (2002), 149–166.
[30] Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL.*
[31] Talal Rahwan, Tomasz Michalak, Nicholas Jennings, Michael Wooldridge, and Peter McBurney. 2009. Coalition structure generation in multi-agent systems with positive and negative externalities. (2009).
[32] Manuel J García Rodríguez, Vicente Rodríguez-Montequín, Pablo Ballesteros-Pérez, Peter ED Love, and Regis Signor. 2022. Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction* 133 (2022), 104047.
[33] Tuomas Sandholm, Kate Larson, Martin Andersson, Onn Shehory, and Fernando Tohmé. 1999. Coalition structure generation with worst case guarantees. *Artificial intelligence* 111, 1-2 (1999), 209–238.
[34] Yasuo Sasaki. 2014. Subjective rationalizability in hypergames. (2014).
[35] Yasuo Sasaki and Kyoichi Kijima. 2012. Hypergames and Bayesian games: a theoretical comparison of the models of games with incomplete information. *Journal of Systems Science and Complexity* 25, 4 (2012), 720–735.
[36] Richard Sharp. 1978. The Game of Diplomacy. (1978).
[37] Onn Shehory and Sarit Kraus. 1996. A kernel-oriented model for coalition-formation in general environments: Implementation and results. In *AAAI/IAAI, Vol. 1.* 134–140.
[38] Onn Shehory and Sarit Kraus. 1998. Methods for task allocation via agent coalition formation. *Artificial intelligence* 101, 1-2 (1998), 165–200.
[39] Oskar Skibski, Szymon Matejczyk, Tomasz P Michalak, Michael J Wooldridge, and Makoto Yokoo. 2016. k-Coalitional Cooperative Games.. In *AAMAS.* 177–185.
[40] Youcef Sklab, Samir Aknine, Onn Shehory, and Abdelkamel Tari. 2020. Coalition formation with dynamically changing externalities. *Engineering Applications of Artificial Intelligence* 91 (2020), 103577.
[41] Raza Umar and Wessam Mesbah. 2016. Coordinated coalition formation in throughput-efficient cognitive radio networks. *Wireless Communications and Mobile Computing* 16, 8 (2016), 912–928.
[42] Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon M Stewart, Jonathan K Kummerfeld, Denis Peskoff, and Jordan Lee Boyd-Graber. 2024. More Victories, Less Cooperation: Assessing Cicero's Diplomacy Play. *arXiv preprint arXiv:2406.04643* (2024).