Should I Trust You? Detecting Deception in Negotiations using Counterfactual RL

Wichayaporn Wongkamjan[1] **Yanze Wang**[4] **Feng Gu**[1] **Denis Peskoff**[2]
Jonathan K. Kummerfeld[3] **Jonathan May**[4] **Jordan Lee Boyd-Graber**[1]
[1]University of Maryland [2]Princeton University [3]University of Sydney
[4]Information Sciences Institute, University of Southern California
wwongkam@umd.edu jbg@umiacs.umd.edu

Abstract

An increasingly prevalent socio-technical problem is people being taken in by offers that sound "too good to be true", where persuasion and trust shape decision-making. This paper investigates how AI can help detect these deceptive scenarios. We analyze how humans strategically deceive each other in *Diplomacy*, a board game that requires both natural language communication and strategic reasoning. This requires extracting logical forms of proposed agreements in player communications and computing the relative rewards of the proposal using agents' value functions. Combined with text-based features, this can improve our deception detection. Our method detects human deception with a high precision when compared to a Large Language Model approach that flags many true messages as deceptive. Future human-AI interaction tools can build on our methods for deception detection by triggering *friction* to give users a chance of interrogating suspicious proposals.

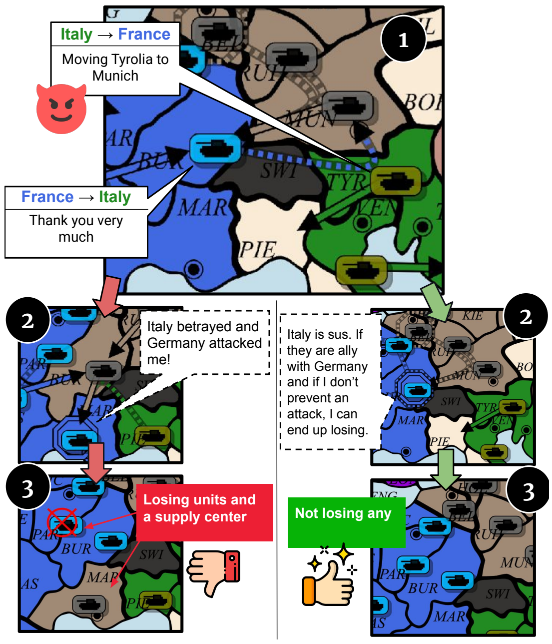

Figure 1: Detecting deception is crucial in mixed cooperative-competitive environments. **(Left)** France believed the lie that Italy will move their army in Tyrolia to Munich, losing Burgundy and subsequently Marseilles to Germany. **(Right)** If France had detected the deception, they could have successfully defended Burgundy and avoided disbanding one army.

1 Friction in AI systems

Deception in natural language is a fundamental aspect of human communication, often employed as a strategic tool to mislead others through misrepresentation, omission, exaggeration, or counterfactual reasoning (Bok, 2011). From casual social interactions to high-stakes negotiations, deception influences trust, decision-making, and cooperation, making it a subject of extensive study in psychology, linguistics, and philosophy, manifesting real-world challenges such as fake news on social media (Bade et al., 2024), misinformation (Panda and Levitan, 2022) and adversarial communication in strategic games (Bernard and Mickus, 2023). As artificial intelligence systems increasingly engage in human-like communication, they not only inherit but also amplify deceptive strategies, sometimes unintentionally. In AI-generated text, deception can emerge as a byproduct of optimization objectives, particularly in RLHF scenarios where agents maximize utility in multi-agent settings, sometimes at the expense of honesty (Wen et al., 2025). This phenomenon has garnered significant attention across various domains, as AI deception is not confined to theoretical constructs but manifests in real-world challenges, e.g. hallucination in reasoning tasks (Grover et al., 2024).

Prior research underscores that AI-generated deceptive communication can be difficult to detect and may lead to unintended consequences when deployed in practical applications (Park et al., 2024; Sarkadi, 2024). Deceptive AI-generated text can erode trust in digital communication, amplify mis-

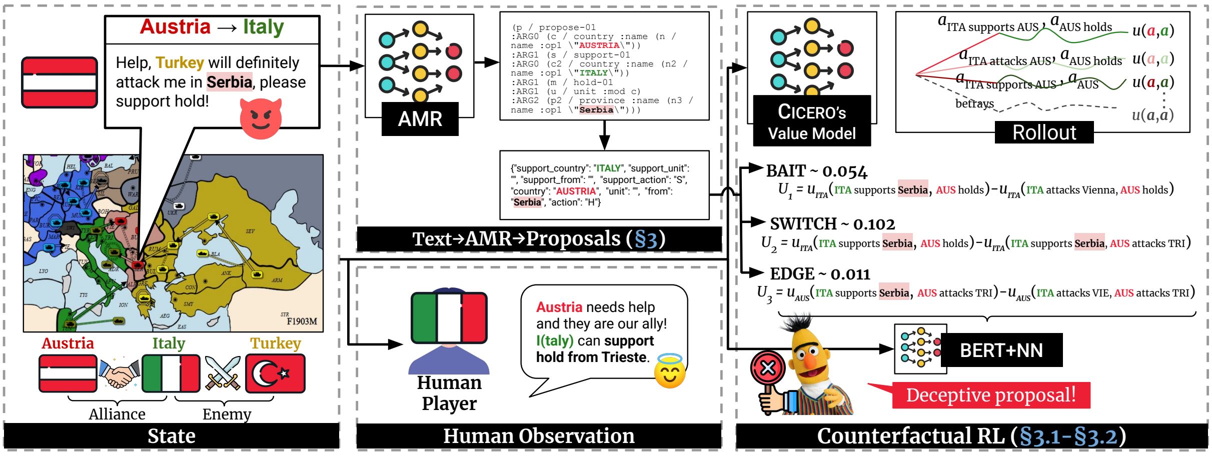

Figure 2: An overview of our approach to detect deceptive proposals, requiring a recipient (Human Player) to follow a proposed action. **(Left)** A state of this Diplomacy game is (1) Austria and Italy have an alliance (2) while Turkey and Italy have been clashing for several turns. Austria realizes that they are in a weak spot and need a quick escape, so they reach out to Italy. It is a deceptive proposal so that Austria can get to Trieste. **(Bottom Middle)** A human player can be biased towards their own ally (Austria) and use their fast-thinking system to instinctively help. **(Top Middle)** For an alternative perspective, our approach converts natural language to proposals using AMR. **(Right)** Then, we leverage the RL value function from CICERO to estimate three aspects of deception—Bait, Switch and Edge—from counterfactual actions of Austria and Italy. Passing the dialogue alongside these values to a classifier decides whether Austria's proposal is **deceptive**.

information, and facilitate large-scale manipulation in political, financial, and social domains (Solaiman et al., 2019; Weidinger et al., 2022). Furthermore, the scalability of AI models allows deceptive content to be produced and disseminated at unprecedented rates, making manual detection impractical. To address these risks, robust mitigation strategies are necessary, including adversarial training (Perez et al., 2022), explain-ability techniques to enhance AI transparency (Danilevsky et al., 2020), and real-time detection methods leveraging linguistic and behavioral cues (Vosoughi et al., 2018).

We test our detection strategies within the environment of *Diplomacy*, a game rich in negotiation, cooperation, and betrayal expressed through natural language. The most intriguing moments of the game arise when two players negotiate to cooperate in pursuit of their respective goals. While such agreements usually yield mutual benefits, this is not always the case—some negotiated arrangements are the result of deception, omission, or straight-up lies on the part of one player. Skilled players combat such behavior by developing the ability to recognize when an offer *sounds too good to be true*. Our work explores this area to raise awareness among human players when they encounter deception embedded in negotiations.

We use the *value function* of CICERO (Bakhtin et al., 2023), an agent trained to play Diplomacy at a human level, to detect whether a proposal is "*too good to be true*". We estimate other players' likely strategies and query Llama3 (AI@Meta, 2024), an LLM good at general purpose semantic understanding, to determine if a message is contrary to CICERO's expected strategy. However, Llama3 cannot precisely predict deception since it sees almost half of data as deceptive. Consequently, we decide to take a more explicit approach to modeling deception. Our contributions are as follows:

1) With Theory-of-Mind-influenced deception, we identify negotiations in natural language via formal logical modeling and detect potential deceptive offers in negotiations using CICERO RL value function to generate counterfactual explanations.

2) We train a BERT-based (Devlin et al., 2019) classifier to predict deception using RL values and message embeddings.

3) We show that our classifier is more accurate than a fine-tuned Llama3 in human lie prediction and detecting partially-deceptive negotiations.

2 Deception in the Wild

Real-world deception manifests in various forms, such as *scams* and *phishing* attacks, where perpetrators exploit **trust** to manipulate victims into believing in the possibility of good fortune, even if it is unlikely (Button et al., 2014; Muscanell et al., 2014; Hanoch and Wood, 2021). These deceptive tactics often rely on persuasive language. If victims fall for these **too good to be true** claims, they become targets and may comply with the perpetrators' requests—for example, disclosing sensitive information or making financial investments under false pretenses—ultimately resulting in monetary *loss* or data breaches (Burnes et al., 2017; Coluccia et al., 2020). Those scammers would *gain value* through data breaches or simply by acquiring cash.

Detecting deception remains a persistent challenge, especially when it is needed for real-world problems. We can use AI, but it is hard to evaluate and requires quality feedback from humans to train models in detecting real deception. Deception in a limited space like a strategic game, e.g., Diplomacy, where nuanced persuasion and deception is required for winning, is more tractable to evaluate. A bounded example would allow us to measure the ability of an AI to improve in deception detection.

2.1 One Gains, One Loses

Deception has been studied in games that rely on trust, negotiation, and strategic misrepresentation, such as Werewolf (Chittaranjan and Hung, 2010; Hancock et al., 2017; Girlea, 2017), Poker (Lee and Hin, 2013; Palomäki et al., 2016), and Diplomacy (Niculae et al., 2015; Kramár et al., 2022). Diplomacy is a complex interplay of strategy, high-level cooperation, and subtle betrayal. The game is set on an European map, highlighting key territorial cities known as supply centers. Each of the seven players controls a country and moves units on the map, with the objective of capturing more than half of the supply centers (18 out of 34) to achieve victory. For each turn, players communicate one-to-one and then simultaneously reveal their orders for each units.

Deception plays a crucial role in gaining supply centers and, ultimately, securing a win. Cliques of players agreeing to coordinate to gain advantages over others must operate in secrecy. Deception must be undetected to be successful. If the player fails to recognize deception, they risk losing supply centers and may lose the game (Figure 1). If a player's deception succeeds, they may gain supply centers. The challenge lies in quantifying the benefits of deception and the losses of those who are deceived. Given the significance of supply centers as a sparse scoring mechanism, we see an opportunity to integrate reinforcement learning (RL) (Zinkevich et al., 2007; Brown et al., 2019) into the analysis.

RL has been extensively used to train AI agents in optimizing decision-making that maximize a reward. RL-based AI has been applied to Diplomacy (Paquette et al., 2019; Anthony et al., 2020; Gray et al., 2021; Bakhtin et al., 2021), with a recent model, CICERO (Bakhtin et al., 2022), achieving competitive human-level play. This paper uses a reward model from CICERO to detect action proposals where deception is likely (Figure 2).

3 Counterfactual RL against Deception

We look for deception in the text messages between pairs of players. Each player controls multiple units in this game, so we restrict a pool of messages where a player explicitly requests another player to issue a specific order (e.g. Austria asks Italy to support in Serbia, Figure 2). With this, we parse messages in natural language to Abstract Meaning Representation (AMR, Banarescu et al., 2013).

For any message to a player, we want to raise awareness if the proposal is *potentially deceptive*. We leverage a well-trained value function, a part of CICERO (Bakhtin et al., 2022), to estimate how likely a proposal is deceptive. This section we discusses our method, **C**oun**T**erfactual **RL** against **D**eception (CTRL-D), which has two main components: 1) Counterfactual RL and 2) formulations to capture potential deceptive proposals.

3.1 Counterfactual RL

Player i needs to pick an action a_i given a board state s. However, moves in Diplomacy do not happen in isolation — all actions of *other* players a_{-i} happen simultaneously, so CICERO uses a function $u_i(a_i, a_{-i}, s)$ that represents estimated future rewards that player i will receive if actions a_i and a_{-i} are played in a state s. Thus, a high value represents a "*better*" move based on learned strategies.

While a review of CICERO is outside the scope of this paper, its value function allows our work to compute counterfactual one-step actions to estimate potential deceptive proposals from another player j, where each proposal is about action a_i and a_j. Equipped with text-to-proposals and the RL value function, we are ready to detect deception.

3.2 Deceptive Proposals

To estimate how likely a proposal is deceptive, we introduce three *deceptive signs* that account for different aspects of deception. First, we can measure whether a victim would get a higher reward if the proposal was not a deception (i.e., is the fake proposal from the deceiver appealing?). Second, we can measure whether a victim would get a lower reward if they believe the deception. Third, we can measure whether a deceiver would increase their future reward by deceiving the victim. These three measures called: Bait, Switch and Edge.[1] In this section, we highlight deceptive values through three hypotheses.

We define a proposal $p_{j \to i}$ when player j proposes actions to player i. A proposal $p_{j \to i}$ consists of an action $\hat{a}_i$ that player j wants player i to play and an action $\hat{a}_j$ that player j promises to make. In Diplomacy, an action is a tuple of unit orders, e.g.

- an army in Berlin moves to Kiel ,
- an army in Munich moves to Ruhr and
- a fleet in Kiel moves to Holland

where these can represent in a tuple as ('A BER - KIE', 'A MUN - RUH', 'F KIE - HOL'). Therefore, player j can propose an action to player i with multiple unit orders $\hat{a}_i = (\hat{a}_{i,1}, \hat{a}_{i,2}, \ldots, \hat{a}_{i,n})$ where n is a number of player i's units.

We estimate how likely a proposal is deceptive by following three hypotheses when it is '*too good to be true*". **Bait**, a victim *perceives a greater reward* if they *alter* a decision to *follow* the deceiver's proposal and the deceiver does not actually deceive, but rather follows the plan. Assume player i has a plan a_i, and player j proposes $\hat{a}_i$ and $\hat{a}_j$. From the perspective of player i, they decide to play $\hat{a}_i$ because they perceive that the estimated future rewards will increase by:

$$U_1 = u_i(\hat{a}_i, \hat{a}_j) - u_i(a_i, \hat{a}_j). \quad (1)$$

Switch, a victim will *receive a lower* reward if they *follow* the deceiver's request and if the deceiver betrays the victim. Player j proposes actions $\hat{a}_i$ and $\hat{a}_j$ to player i where player j has alternative plan a_j to instantly stab or take advantage of player i. The estimated future rewards of player i will decrease if player j betrays player i. We leverage CICERO's RL value function u_i (Section 3.1) to formulate the first hypothesis:

$$U_2 = u_i(\hat{a}_i, \hat{a}_j) - u_i(\hat{a}_i, a_j) \quad (2)$$

[1]These terms come from popular culture terms around scams: a deceiver offers "*bait*" to attract the victim who suffers from the "*switch*", leading the deceiver to profit, their "*edge*" in the scam

where a_j is an alternative action that player j might play instead of following the proposed move $a_j \neq \hat{a}_j$.

Edge, a deceiver will *receive a better reward* when a victim *follows* their proposal. Given the deceiver j's plan a_j and the victim's plan a_i, if player j proposes a suboptimal $\hat{a}_i$ to player i and player i falls for it. The estimated future rewards for player j can increase:

$$U_3 = u_j(\hat{a}_i, a_j) - u_j(a_i, a_j). \quad (3)$$

In short, the three hypotheses for deceptive proposals assume the victim loses, the deceiver gains and victim follows the proposal (Counterfactual RL, Figure 2). For a victim or player i's plan a_i in Equation 1 and Equation 3, we define the plan loosely as any action closest to the optimal action from player i's perspective, thus sampling an action $a_i \sim \pi_i$ where π_i is CICERO's policy.

In the final step, we train a classifier using *deceptive values* U_1, U_2 and U_3 from Equation 1, 2 and 3. We have two main models:

1) A BERT-BASED-UNCASED that takes a message and outputs text embeddings.

2) A Neural Network with three linear layers, ReLU and Sigmoid activations, that takes in text embeddings concatenated with U_1, U_2 and U_3 .

We train only ten epochs with a small training data sampled from Peskov et al. (2020). In the next section, we discuss datasets that we use to train the classifier and to test our approach against an LLM baseline.

4 Recall-Oriented Lie Detection for Friction

This section explores deception detection and its role in creating strategic friction, i.e., deliberate decision-making in human-AI interactions. Section 4.1 analyzes human-only Diplomacy games (Peskov et al., 2020) to categorize deceptive messages and select messages for training and evaluation. Section 4.2 extends this analysis to a larger dataset within human-AI settings, testing whether our framework can introduce friction against deceptive proposals. Our goal is not to optimize F_1-Score but rather to flag *possible* deception for users, introducing friction to help them detect deception. In other words, our goal is to maximize recall.

Sender	Message
Russia (Lie)	I think I will move Moscow into War, with Sil supporting, where I could go for Austria the following turn.
Russia (Lie)	Also I will move Ukraine to gal. Could you support me there?
Turkey (Truth)	yeah I dont mind support

Table 1: An example of a lie annotation from a human player in Peskov et al. (2020) dataset.

Categories	Total
Any messages	17,289
Any lies	842
Other	459
Deceptive Moves	**286**
Feigning Trust/Loyalty	28
False Excuse	27
Withholding Information	24

Table 2: We categorize lie messages in Peskov et al. (2020) data set, in which **Deceptive Moves** is the closest to our interest. Though messages with this type do not appear often, they are useful for our CTRL-D to get deceptive signs.

4.1 Alignment to Human Lies

To understand human deception, we use the twelve Diplomacy games[2] annotated by Peskov et al. (2020) containing human strategy through orders and communication from private messages. The dataset contains annotations from players at per-message granularity, indicating whether or not the content of their message contained a lie (Table 1).

To best select training and evaluation data, first we breakdown lie messages—natural language texts generated by humans—in Peskov et al.'s dataset into categories (Table 2), where the category that is closest to our interest is **Deceptive Moves**.[3] Deceptive moves are rare, constituting fewer than 1.7% of all messages; therefore, we sample data for training and evaluation:

- **For training data,** we focus on messages with human-annotated negotiations—logical form of negotiations annotated by experts—specifically for player i or player j. In total, there are 344 messages with human-annotated negotiations containing 59 lies and 28 proposals. We sample additional 1,500 messages, though without human annotation, we retrieve logical forms of negotiations as discussed in Section 3.
- **For evaluation,** we simply sample 1,000 messages containing 80 lies from the rest of dataset without any further selection.

4.2 Friction for Humans

While the data from Peskov et al. (2020) confirms our approach, **CTRL-D**, has a desired recall, it is small. Thus, we next test generalization on Meta's dataset[4] of Diplomacy, which contains 40,000 games, 13 million natural language interactions from humans and CICERO players. Although these data lack thorough deception annotation, we can ex post facto validate precision through human verification.

5 Baselines for Deception

Deception is not only about the offer on the table: Niculae et al. (2015) show language changes before a betrayal occurs in Diplomacy, and Lai et al. (2020) demonstrate that this also holds true even in online reviews.

Thus, we compare our approach to language only baselines. Specifically, we implement an LLM-based baseline, using `LLaMA 3.1-8B-Instruct` as our primary model, to detect and mitigate deception in Diplomacy negotiations. Our objective is to determine whether a large language model can identify suspicious messages without relying on reinforcement learning (RL). We adopt two complementary criteria for evaluating deception: *Direct Judgment* and *Alignment Judgment*. Detailed prompts are provided in Figure A1.

Direct Judgment In the *Direct Judgment* approach, we prompt the baseline to decide whether a negotiation message is deceptive by examining its textual content together with relevant contextual information. For each negotiation round, we provide:

1. **Current board state:** A concise summary of each player's positions and units (e.g., which territories are occupied by which units).
2. **Negotiation message:** The specific proposal, plan or statement made by player j to player i.
3. **Predicted Orders $\tilde{a}_j$:** The orders CICERO would issue for player j under a dialogue-free RL policy, indicating how CICERO believes player j will actually move.

[2] Github: It Takes Two to Lie: One to Lie, and One to Listen

[3] For more category details, see Appendix A

[4] AI@Meta: Towards Human-AI Cooperation RFP

Given this information, the baseline determines whether the message from player j is deceptive, returning a binary label. We do not impose a strict threshold at this stage; instead, we rely on the model's ability to integrate textual cues and board-state context.

Alignment Judgment Because *Direct Judgment* alone results in nearly half of the messages being flagged as deceptive, we propose *Alignment Judgment*, which applies a more structured, threshold-based process, to measure consistency between CICERO's predicted orders and the orders proposed in the negotiation message.

For each player j, CICERO predicts orders $\tilde{a}_j$ using a dialogue-free RL policy. These predicted orders are then compared to the actions stated in the negotiation message. If player j proposes $p_{j \to i}$ with actions $\hat{a}_i$ and $\hat{a}_j$, but CICERO predicts alternative actions $\tilde{a}_j$ that differ from $\hat{a}_j$, this inconsistency may indicate deception. The baseline is used to identify these inconsistencies. We define two sets: the number of misaligned orders $O_m = \sum_k \mathbf{1}[\tilde{a}_{j,k} \neq \hat{a}_{j,k}]$, and the number of aligned orders $O_a = \sum_k \mathbf{1}[\tilde{a}_{j,k} = \hat{a}_{j,k}]$, where k is the number of the units of player j. If $O_m \geq O_a$, the message is classified as deceptive. Once flagged, the baseline triggers a *risk alert*, informing player i of the misaligned orders. This prompts players to re-evaluate their decisions before finalizing their moves.

6 Results and Analysis

This section compares the approach (Section 3) —a classifier trained by using *deceptive values* from counterfactual RL and text embeddings from BERT to train a three-layer linear neural network—to an LLM baseline (Section 5). We evaluate these using two Diplomacy datasets; the first is Peskov et al. (2020) with human lie annotations. In Section 6.1, we prove that our approach can outperform the LLM baseline and other baselines in predicting human lies. In Section 6.2, we show that our approach has a high precision in predicting deception, potentially creating friction at just the point when a human player is making a critical decision.

6.1 Detecting Human Lies

With human lie annotations, we compare **CTRL-D** prediction to these baselines: 1) **LLM baseline using Direct Judgment**, 2) **LLM baseline using Alignment Judgment** and 3) **Context LSTM with Power** by Peskov et al. (2020) (Table 3). This section answers these questions:

1. Is our value-based RL with classifier approach (CTRL-D) able to predict human lies?
2. How good it is when compared to baselines?
3. What lies does it miss (False negatives) and what truth does it misconceive as lies (False positives)?

Our detection is most effective at predicting human lies. It predicts with high precision 0.950, compared to both LLM-based methods and to Peskov's LSTM (0.147, 0.095 and 0.104). In other words, when CTRL-D predicts a message as a lie, it is 95% likely that the message is an **actual** human lie. However, our detection is not perfect. Its recall is 0.238, meaning that it can only catch **one** out of every **four** lies. This high precision but low recall is still useful in deployment, since human players at least can trust CTRL-D when it flags lies.

The LLM baseline is good at recall but bad at precision. Comparing the recall of CTRL-D with our LLM baseline with Direct Judgment, we find the latter has good recall but low precision. The LLM predicts 45.5% of all messages as lies, which is too sensitive to be useful, when considering the number of misaligned CICERO predicted orders and proposed orders seen in a message. Current LLM-based approaches to providing guidance do not appear capable of recognizing deception, and the resulting high false positive rate makes this strategy not compatible for real-time human games.

CTRL-D misses deception if explicit actions are unclear in negotiations. To characterize the lies CTRL-D misses detecting (False negatives), we observe that 41 of 61 non-detected lies have no logical forms for negotiations. Further, all of these 41 fit in categories other than **Deceptive Moves** (Table 2). We examine the remaining 20 and correct errors in logical forms by hand, to see if doing so improves prediction. With logical forms from humans, the recall of CTRL-D (Table 3) improves slightly from 0.238 to 0.300, showing that CTRL-D depends on proper logical parsing of natural language text.

CTRL-D mispredicts once, while LLM baselines mispredict frequently. We further investigate false positives of CTRL-D and LLM baselines. Since our approach is very precise, it predicts only one true message as a lie (Table 4). On the

Model	Precision	Recall	F1-Score
LLM baseline using Direct Judgment	0.095	**0.551**	0.161
LLM baseline using Alignment Judgment	0.147	0.065	0.090
CTRL-D **(ours)**	**0.950**	0.238	**0.380**
CTRL-D with human annotated logical forms	**0.960**	0.300	**0.457**
Context LSTM + Power	0.263	0.171	0.207

Table 3: While **LLM baseline Direct Judgment** detects deception on actual human lies with a high recall, its precision is very low. **LLM baseline using Alignment Judgment** and Peskov et al. (2020) **LSTM** shows problems in detection with poor precision and recall. Our **CTRL-D** has critically high precision and the best overall tradeoff between precision and recall.

Sender	Message
Germany (Truth)	Well the feeling is mutual. I wanted to let you know that Austria asked for my help putting pressure on Warsaw. I don't intend to do that, but **I recommend you use Silesia to support your Rumanian unit into Galicia**. I promise not to interfere with this maneuver if you promise to keep your baltic fleet focused on defending Sweden from England?

Table 4: An example of False Positive that CTRL-D detects. Germany has a possible short-term gain had they betrayed Russia, but they nonetheless followed through on their proposal.

Sender	Message
France (Truth)	I am supporting Tys to Wes. Can you use Mar to support spain hold?

Table 5: An example of False Positive that the LLM baseline detects as deception. It interprets France's message as a promise to support Austria's order. CICERO's predict orders for France with `A MUN H` where LLM baseline misinterprets as France is attacking Germany. It claims that France contradicts.

other hand, the LLM baseline with Direct Judgment[5] mispredicts 412 messages as lies (Example in Table 5). We look into these messages and see patterns that LLM baseline is heavily constrained on CICERO's predicted orders when it considers proposed orders in messages. This could be improved if LLM baseline can recognize the nature of Diplomacy, namely that there are many possibilities for players' orders, which are not necessarily deceptive.

In sum, CTRL-D captures human lies at best among all methods, including LLM-based methods. Though LLM baseline is better with semantics, it still lacks skills to interpret Diplomacy information in a way that would enable deception detection. With a strong agent, CICERO, predicting human lies using its RL value function makes detection possible. To further validate the quality of our deception detection, we evaluate both CTRL-D and LLM baseline on a larger data set that contains interactions between humans and CICERO.

[5]We focus on Direct Judgment and omit the Alignment Judgment baseline variant since both LLM approaches are similar and have similar results.

6.2 Awareness against Deception

In this section, we evaluate CTRL-D and LLM baselines using the Meta data set. Since this dataset lacks human deception annotations, we first let both models predict whether each message is deceptive, then verify the labeled predictions through human judgment. Human reviewers are provided with historical messages and final orders that sender and recipient submit through games. This information is used to decide whether a sender deceives a recipient by comparing between 1) the sender's commitment in a proposal and 2) the sender's final orders. Although this verification is limited to deception that appears within explicit orders of the sender, this could serve as more evidence to verify performance of our approach and the LLM baseline.

CTRL-D predicts with higher precision than LLM baseline, which overpredicts deception. As shown in Table 6, our findings are consistent with those on the previous dataset. LLM baseline with Direct Judgment predict 41.3% of all samples as deceptive, which is greatly higher than 5% actual lie rate from humans (Peskov et al., 2020). This high rate makes human verification impractical. For LLM baseline with Alignment Judgment, its precision is 0.282 (only 1 in 4 flagged messages is a true lie). In contrast, CTRL-D, though not matching its previous high performance, is still the best at predicting deception.

Errors in CTRL-D and LLM baselines showing thier weakness. We cross-validated our CTRL-D

Model	Deceptive prediction rate	Precision
LLM baseline using Direct Judgment	0.413	-
LLM baseline using Alignment Judgment	0.066	0.282
CTRL-D **(ours)**	0.014	**0.727**
Human Actual Lie Rate	0.050	-

Table 6: Human verification supports **CTRL-D** as the stronger method with higher precision. However, **LLM baseline using Alignment Judgment** is able to detect some lies. **LLM baseline using Direct Judgment** detect almost half of messages as deception. A rate of messages that humans label as lies is included for comparison (Peskov et al., 2020).

with LLM baseline under Alignment Judgment to evaluate their ability to detect deceptive proposals. While both methods can correctly identify some lies, each may fail under different circumstances. We present several examples here:

- Both models label it deceptive, and indeed it is a lie (Table A1).
- LLM baseline overlooks Russia's convoy promise, but CTRL-D detects the unfair exchange (Table A2).
- CTRL-D misses one lie, while LLM baseline correctly spots it (Table A3).

Human verification supports CTRL-D as the stronger method; however, the LLM baseline can still catch some lies. We hope to further test these approaches with human players, thus introducing additional *friction* in real negotiation settings.

7 Related works

Deception in Human Behaviors. Research on deception has highlighted key behavioral and cognitive cues, such as micro-expressions and inconsistencies from mental strain (Ekman, 2003; Vrij, 2008). Multimodal analyses integrating verbal and nonverbal signals have further enhanced detection accuracy (DePaulo et al., 2003). Linguistic cues linked to betrayal in the game Diplomacy offer insights closely aligned with our study (Niculae et al., 2015). Moreover, computational models using language cues have shown promise in detecting deception in text, though evaluations have been limited to small datasets and specific scenarios (Serra-Garcia and Gneezy, 2023; Hazra and Majumder, 2024). Despite progress, the complex dynamics of deception in human behavior remain underexplored.

AI Deception in Texts. With the rise of AI-generated content, detecting textual deception is crucial. Linguistic and psycholinguistic analysis aids detection, while transformer models improve accuracy (Ott et al., 2011; Pérez-Rosas et al., 2015; Zhang et al., 2019). Prior work focuses on detecting AI deception using external and internal methods (Park et al., 2024). External techniques like "*consistency checks*" (Fluri et al., 2024) analyze AI behavior for inconsistencies, while internal methods examine embeddings to detect dishonesty (Azaria and Mitchell, 2023; Burns et al., 2024). Closely related to our work is Fluri et al. (2024), which uses Monte-Carlo Tree Search to check chess moves for logical inconsistencies.

8 Conclusion and Future Work

Our study confirms that with a well-trained value function, we can estimate deception signs— *bait*, *switch*, *edge*—to predict deception. CTRL-D, our counterfactual RL against deception, has a low recall but a high precision, which can be useful for triggering friction for humans. Comparing to a high recall LLM baseline, CTRL-D is better able to predict human lies and generalize, demonstrating high precision consistently on both evaluation data sets.

While these tasks for deception detection are scoped within Diplomacy, they illustrate the general risks and challenges of AI-deception. Future human-AI interaction tools can build on our methods to reevaluate trust in suspicious negotiations.

Limitation

This study can evaluate through real-time Diplomacy games to test whether our approach could help trigger friction in human players and if it could, how useful it is. We limit our evaluation space to Diplomacy, and we could gain a better understanding of deception if we expand to broader areas like negotiation in trading. Our approach, **CTRL-D**, relies on a tool (AMR) to transform texts into logical forms. Its representation could sometimes be invalid and undermine accuracy deception detection. Our detection can only predict those negotiations with explicit actions, missing opportunities where deception occurs in other forms.

Ethical Considerations

Our study uses existing data sets so we do not experiment or collect new data from humans. This paper highlights deception detection which will be necessary for dealing with existing and future harms of AI and LLMs. As a double-edged sword, acknowledging this deception may make future systems better at masking their deception.

Acknowledgments

We thank Meta for granting access to over 40,000 games played on the online platform webdiplomacy.net and for open sourcing Llama 3 and CICERO. Our thanks also go to Ulf Hermjakob and Tess Wood for Diplomacy AMR Annotation Dictionary. We thank Alex Hedges for LLM Diplomacy setup and Sadra Sabouri for valuable feedback. This research is supported by the U.S. Defense Advanced Research Projects Agency (DARPA) Other Transaction award HR00112490374 from the Friction for Accountability in Conversational Transactions (FACT) program. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

AI@Meta. 2024. Llama 3 model card.

Thomas Anthony, Tom Eccles, Andrea Tacchetti, János Kramár, Ian Gemp, Thomas Hudson, Nicolas Porcel, Marc Lanctot, Julien Pérolat, Richard Everett, et al. 2020. Learning to Play No-Press Diplomacy with Best Response Policy Iteration.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it‘s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media fake news classification using machine learning algorithm. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 24–29, St. Julian's, Malta. Association for Computational Linguistics.

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):1067–1074.

Anton Bakhtin, David Wu, Adam Lerer, and Noam Brown. 2021. No-press diplomacy from scratch. In *Advances in Neural Information Processing Systems*.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. 2023. Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning. In *The Eleventh International Conference on Learning Representations*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Timothée Bernard and Timothee Mickus. 2023. So many design choices: Improving and interpreting neural agent communication in signaling games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8399–8413, Toronto, Canada. Association for Computational Linguistics.

Sissela Bok. 2011. *Lying: Moral Choice in Public and Private Life*. Vintage.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. 2019. Deep counterfactual regret minimization. In *International conference on machine learning*, pages 793–802. PMLR.

David Burnes, Charles R. Henderson, Christine Sheppard, Rebecca Zhao, Karl Pillemer, and Mark S. Lachs. 2017. Prevalence of financial fraud and scams among older adults in the united states: A systematic review and meta-analysis. *American Journal of Public Health*, 107(8):e13–e21. PMID: 28640686.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision. *Preprint*, arXiv:2212.03827.

Mark Button, Carol McNaughton Nicholls, Jane Kerr, and Rachael Owen. 2014. Online frauds: Learning from victims why they fall for these scams. *Australian & New Zealand journal of criminology*, 47(3):391–408.

Gokul Chittaranjan and Hayley Hung. 2010. Are you Awerewolf? Detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5334–5337. IEEE.

Anna Coluccia, Andrea Pozza, Fabio Ferretti, Fulvio Carabellese, Alessandra Masti, and Giacomo Gualtieri. 2020. Online romance scams: relational dynamics and psychological characteristics of the victims and scammers. a scoping review. *Clinical practice and epidemiology in mental health: CP & EMH*, 16:24.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

B. M. DePaulo, S. A. Kashy, and et al. M. Kirkendol. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 2003. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books, New York.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. 2024. Evaluating superhuman models with consistency checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232. IEEE.

Codruta Liliana Girlea. 2017. *Deception detection in dialogues*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Jonathan Gray, Adam Lerer, Anton Bakhtin, and Noam Brown. 2021. Human-Level Performance in No-Press Diplomacy via Equilibrium Search.

Shresth Grover, Vibhav Vineet, and Yogesh S Rawat. 2024. Navigating hallucinations for reasoning of unintentional activities. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9666–9680, Miami, Florida, USA. Association for Computational Linguistics.

Will Hancock, Michael W Floyd, Matthew Molineaux, and David Aha. 2017. Towards deception detection in a language-driven game. In *The Thirtieth International Flairs Conference*.

Yaniv Hanoch and Stacey Wood. 2021. The scams among us: Who falls prey and why. *Current Directions in Psychological Science*, 30(3):260–266.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8506–8520, Mexico City, Mexico. Association for Computational Linguistics.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. 2022. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13(1):7214.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Jackey Lee and Ting Hin. 2013. Deception and Arousal in Texas Hold 'em Poker. Master's thesis, University of Waterloo.

Nicole L Muscanell, Rosanna E Guadagno, and Shannon Murphy. 2014. Weapons of influence misused: A social influence analysis of why people fall prey to internet scams. *Social and Personality Psychology Compass*, 8(7):388–396.

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)*, pages 1650–1659, Beijing, China. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA.

Jussi Palomäki, Jeff Yan, and Michael Laakasuo. 2016. Machiavelli as a poker mate—A naturalistic behavioural study on strategic deception. *Personality and Individual Differences*, 98:266–271.

Subhadarshi Panda and Sarah Ita Levitan. 2022. Improving cross-domain, cross-lingual and multi-modal deception detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 383–390, Dublin, Ireland. Association for Computational Linguistics.

Philip Paquette, Yuchen Lu, SETON STEVEN BOCCO, Max Smith, Satya O.-G., Jonathan K. Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. 2019. No-press diplomacy: Modeling multi-agent gameplay.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai Deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns*, 5(5).

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Vanessa Pérez-Rosas, Benjamin Kleinberg, Alexandre Lefevre, and Rada Mihalcea. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1566–1576.

Ştefan Sarkadi. 2024. Deceptive AI and Society. *IEEE Technology and society magazine*, 42(4):77–86.

M. Serra-Garcia and Uri Gneezy. 2023. Improving human deception detection using algorithmic feedback. *SSRN Electronic Journal*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Adrian Vrij. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley, Chichester, UK.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2025. Language models learn to mislead humans via RLHF. In *The Thirteenth International Conference on Learning Representations*.

Xinxing Zhang, Liang Zhang, and Bo Zhang. 2019. Deception detection in news: A transformer-based approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2340–2345.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20.

A Lie Categories

We simply categorize lie messages using keywords, prioritize as the following:

- **Deceptive Moves.** {support, move, attack, retreat, convoy, hold, bounce}
- **Feigning Trust/Loyalty.** {trust, friend}
- **Withholding Information.** {no idea, not sure}
- **False Excuse.** {sorry, busy}.

Sender	Message
Austria (Lie)	Germany is apparently ripe for the taking. I'll support you into Kiel in exchange for support into Bulgaria from Serbia.

Table A1: LLM baseline and our CTRL-D correctly detect this messages as a lie, where Austria does not support to Kiel as they propose.

Sender	Message
Russia (Lie)	Oh and RE: **St Petersburg yeah. My plan is to convoy an army over from Norway. If you vacate Serbia** I would be happy to then move my units back to Ukraine and Sevastopol in return for you then moving that Fleet to Smyrna.

Table A2: Russia lies to Austria which CTRL-D is able to catch this, while LLM baseline is not.

Sender	Message
Germany (Lie)	Hi again. **My new army in kiel will move to munich and Munich will move to bohemia.** You are a trustful neighbor and we can share victory. Im not moving North sea and i do not expect you to do it. Let the Russian collapse by himself

Table A3: LLM baseline detects this lie from Germany. However, CTRL-D misses to detect this lie due to its error in logical forms.

Message	Category
We have a terrific counter, if you'd like to work with me. I can tap Greece and Aegean, which should allow you to save Bulgaria (Const S Bulgaria holding). At the same time, Black Sea Support Armenia to Sevastopol WILL WORK! I'm planning to move over to Trieste, so the end result of all this would be Turkey regaining Sev and remaining at 5 while I take Trieste and get to 5. From there, I think we should be able to work together to finish off Austria.	Deceptive Moves
This is too big a risk for me. My preferred move is to Belgium. I understand this may warrant doing a couple of favors for your next year in return. But I just feel too exposed otherwise. I'm about to head out for the evening and I have my move set for Belgium. I'm hoping for your support into there, but I understand if you can't provide it.	Deceptive Moves
I can move on Picardy, but only if you move on Belgium. I think Germany is going for him, so now's your chance to get a slice of France while the getting is good.	Deceptive Moves
Hi Turkey! I'm sorry that I've been so slow to get in touch. Kind of a rough day for me to begin a game as I e been pretty swamped. Things are clearing up now, and I appreciate you reaching out to me. So far I have notes from Austria and Russia being pretty cagey and non-committal. Perhaps that is just the life of Italy? Nobody really has me in their plans? I don't really know what I'm going to do yet, so if you have ideas, or you have a use for me, please let me know. I'd basically be delighted to work with anyone who really wants to work with me. (No sign yet that this includes anyone at all)	False Excuse
I am a bit new at this stuff, sorry	False Excuse
But yeah sorry about that, I had put that in as my orders and then forgot to change them	False Excuse
If you want me to trust you I think you should give at least one back, as you are much bigger and frankly it's probably in your best interest to go for the solo, although if the rest of the contries cooperate you almost certaintly won't get it. And your definitly going to need me if you want to go for the draw	Feigning Trust/Loyalty
Well good luck to you too! Let's be friend! Have you heard anything interesting?	Feigning Trust/Loyalty
I can't say ik certain of how things will go in the west but I like to maintain peace as long as possible and a few allies I can trust is important as hell in this game	Feigning Trust/Loyalty
Not sure what your plan is. I'm covering Marseilles, Paris, and Belgium so you have no chance to take a supply center this turn.	Withholding Information
I'm not sure yet. Lots to think about.	Withholding Information
How do you know they are plotting each-others destruction? Not sure which side to take yet.	Withholding Information
Hmm I thought it was something else	Other
I can settle for survival. Deal.	Other
I can't share screenshots apparently, but that's true. It's not just England	Other

Table A4: Our lie categories using examples from Peskov et al. (2020)

If the message does not belong to any category, we rule it as **Other**. Examples for each category in Table A4.

You are an expert assistant specializing in the Diplomacy board game. Your role is to assist a novice player by analyzing:
1. The current board state.
2. The message history exchanged between the novice player and the counterplayer.
3. The predicted orders for the counterplayer.
Your primary objective is to evaluate whether the counterplayer's predicted orders align with the message history and the board state.
Key Evaluation Guidelines:
1. Consider an order aligned if its purpose or intent is consistent with the counterplayer's stated goals or the tactical/strategic needs implied by the board state.
2. Special orders (e.g., supporting moves from counterplayer) can still be aligned if it serves to ensure the success of a critical move or maintains flexibility in uncertain situations.
3. Misalignment occurs if the order: Contradicts the stated strategy or creates unnecessary risks. Fails to contribute meaningfully to the position given the board state.
4. Evaluate the broader context of each move, including defensive, offensive, or preparatory purposes, and weigh these factors in your analysis. For each predicted order, explain whether it aligns with the message history and board state. Use reasoning that incorporates tactical, strategic, and contextual factors while avoiding overly strict judgments. Limit your evaluation to alignment or misalignment, with short reasoning for each.
Predicted Orders of ITALY:
['A TUN - ALB VIA', 'F ION C A TUN - ALB', 'F NAP - TYS', 'A VEN - TRI']
Board State:
AUSTRIA: ['F TRI', 'A SER', 'A GAL', 'A BUD'], ENGLAND: ['F LON', 'F NWY', 'A YOR', 'F EDI'], FRANCE: ['A POR', 'A SPA', 'F BEL', 'F BRE', 'F MAR', 'A PAR'], GERMANY: ['F DEN', 'A MUN', 'A HOL', 'A BER', 'F KIE'], ITALY: ['A VEN', 'F ION', 'A TUN', 'F NAP'], RUSSIA: ['A WAR', 'F SEV', 'F BOT', 'A RUM', 'A STP'], TURKEY: ['A CON', 'A BUL', 'F BLA', 'F SMY']
Message History:
Message from AUSTRIA: "France and Turkey are trying to get me to attack you, obviously I will not do that."
Message from ITALY: "With the build in Marseilles, France's goals are quite transparent."
Message from ITALY: "Turkey will likely bounce me in EAS. Would you be willing to support me into Greece (via convoy)?"
Message from ITALY: "From there, I can support you into Bulgaria."
Message from AUSTRIA: "I'm willing to debate it but I had a different view on this turn. I figured he would bounce you in EAS but what if you went for AEG instead? Unless you're worried about overextending yourself due to France. Because I trust you enough at this point to move Tri to Alb and then we will definitely be able to get Gre in the fall."
Message from ITALY: "Which makes going to AEG/EAS a bit of an overstretch."
Message from ITALY: "If we can get Germany on board I'd feel a lot more comfortable with committing east."
Message from ITALY: "Sorry, some connection issues on my end so messages sent out of order."
Question(Direct Judgment):
I am playing as AUSTRIA. You need to determine whether ITALY is being deceptive based on the above information. Your answer must be one of two categories: Truthful or Deceptive. Provide a brief explanation and highlight any potentially deceptive content.
Question(Alignment Judgment):
I am playing as AUSTRIA. For each of ITALY's predicted orders, evaluate whether it aligns with the message history and the current board state. Explain the orders first and provide short reasoning and analysis for each predicted move.

Figure A1: Prompt template used in Llama baselines for identifying alignment and detecting deception.