Deception Detection in Diplomacy

Satwik Garg[1], Shreyansh Srivastav[1], Ashutosh Dwivedi[1]

[1]Indraprastha Institute of Information Technology Delhi, India

{satwik22461, shreyansh22485, ashutosh22116}@iiitd.ac.in

Abstract

Detecting deception in online communication remains a critical challenge, particularly in complex strategic interactions. This paper addresses deception detection in the "Diplomacy" strategy game, utilizing the dataset by Peskov et al. (2020) that uniquely includes labels for both sender intent and receiver perception of deception. We propose GRAPHITE, a novel hybrid architecture designed to capture both message semantics and evolving player relationships. GRAPHITE represents player interactions as a fully-connected graph where nodes represent players and edges capture relationship strengths. Our architecture integrates a BERT encoder with a relationship-aware LSTM that models temporal dynamics between player pairs, processing multiple input modalities including message text, player scores, and game context. Despite significant class imbalance (only 4-5% deceptive messages), GRAPHITE achieves 81.25% accuracy on the sender task and 88.73% on the receiver task, with macro F1 scores of 0.578 and 0.524 respectively, and Lie F1 scores of 0.264 (sender) and 0.109 (receiver). Our results demonstrate that modeling dynamic strategic relationships significantly enhances deception detection capabilities, offering insights applicable beyond games to real-world scenarios where relationship context is critical for deception detection.

1 Introduction

In today's increasingly digital world, human communication predominantly takes place via text-based online platforms where trust and deception play pivotal roles. From online negotiations and dating applications to phishing scams and social media misinformation, the detection of deceptive intent has emerged as both a challenging computational problem and a pressing social need. Despite considerable research in deception detection, existing approaches are often hampered by limitations such as reliance on synthetic data, short-term interactions, or role-constrained scenarios that do not adequately capture the complexity of genuine human behavior.

The strategic board game *Diplomacy* offers a uniquely realistic context for studying deception. In *Diplomacy*, players must cultivate long-term alliances, engage in intricate strategic planning, and occasionally resort to betrayal as a tactical maneuver. Distinctively, each message within the game is annotated from two complementary perspectives: the sender's intended veracity and the receiver's perceived truthfulness. This dual annotation provides a richer, more nuanced framework for analyzing deceptive behavior.

For instance, consider a message such as:

> *"You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long-term thing! Thank you."*

In this example, while the sender categorizes the message as deceptive, the receiver interprets it as sincere—thereby highlighting the intricate and often subtle discrepancies between deceptive intent and perceived authenticity.

Motivated by these observations, our paper aims to develop robust, context-aware models that transcend superficial linguistic cues. Specifically, we explore how conversational context, power dynamics, and inherent linguistic patterns jointly contribute to both the intention and perception of deception in extended, strategic interactions. Ultimately, the insights gleaned from this work are poised to advance applications in cybersecurity, digital diplomacy, and social media moderation, where accurately gauging trustworthiness is crucial.

This study leverages the distinctive *Diplomacy* deception dataset to bridge the gap between syn-

thetic deception detection paradigms and the complexities of real-world strategic communication.

2 Related Work

Traditional deception detection research has largely centered on isolated utterances within narrow domains like product reviews (Ott et al., 2012) or court testimony (Louwerse et al., 2010), limiting their applicability to real-world, dynamic interactions. Games such as *Werewolf* and *Box of Lies* offer controlled environments but often rely on fixed roles, reducing ecological validity.

In contrast, the board game *Diplomacy* fosters open-ended, strategic communication, making it an ideal testbed for studying deception within evolving social contexts. Peskov et al. (2020) introduced a labeled corpus of *Diplomacy* messages annotated by both senders and receivers, highlighting the complex interplay between intent and perception in deceptive exchanges. Their work demonstrated that incorporating linguistic, contextual, and relational signals boosts deception detection, with neural models approaching human performance.

Building on this, Wongkamjan et al. (2025) proposed CTRL-D—a framework integrating AMR parsing, counterfactual reasoning, and RL-based modeling (via CICERO) to identify deceptive proposals through plausibility estimation. Their results showed that combining structured reasoning with BERT-based embeddings significantly outperforms LLM-only baselines, particularly in identifying nuanced signals like *Bait*, *Switch*, and *Edge*.

Our work extends this line of research by systematically evaluating classical, graph-based, and LLM approaches across both actual and perceived deception, using the rich, multi-agent setting of *Diplomacy* as a foundation.

3 Dataset

We utilize a dataset derived from 12 full games of *Diplomacy*, comprising player-to-player negotiation messages annotated for both intended and perceived deception. Each game consists of seven players: *England, France, Germany, Italy, Austria, Russia*, and *Turkey*. The dataset is divided into:

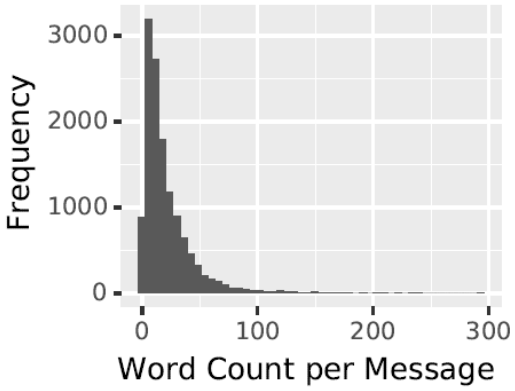
- **Training set:** 189 conversations, 13132 messages
- **Validation set:** 21 conversations 1416 messages
- **Test set :**42 conversations 2741 messages

The dataset includes the following fields:

- **speakers**: The sender of each message, represented as a string (e.g., `"russia"`, `"turkey"`).
- **receivers**: The intended recipient of the message, also a string from the same set of player names.
- **messages**: The raw message text exchanged between players. Messages vary in length, from single words to multi-paragraph communications.
- **sender_labels**: A boolean or string indicator (`true`/`false`) reflecting whether the sender marked the message as truthful or deceptive. This is used for computing ACTUAL_LIE.
- **receiver_labels**: A string indicating the receiver's perception of the message as truthful (`true`), deceptive (`false`), or unannotated (`"NOANNOTATION"`). This is used to derive SUSPECTED_LIE.
- **game_score**: The number of supply centers controlled by the sender at the time of the message, ranging from 0 to 18 (string format).
- **score_delta**: The difference between the sender's and receiver's supply center counts at message time, ranging from -18 to $+18$ (string format).
- **absolute_message_index**: The position of the message in the entire game (integer).
- **relative_message_index**: The index of the message within its specific dialogue thread (integer).
- **seasons**: The current season (`Spring`, `Fall`, or `Winter`) in which the message was sent.
- **years**: The year of gameplay, ranging from 1901 to 1918.
- **game_id**: An integer (1–12) indicating which of the 12 Diplomacy games the message is drawn from.

This dataset provides a rich multimodal context for deception modeling by incorporating strategic gameplay metadata (e.g., supply centers, time, and

seasonal progression) alongside both subjective and objective labels of deceptive intent and perception.



Category	Value
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79

4 Methodologies

4.1 Dataset and Preprocessing

We utilize the *Diplomacy* dataset provided by Peskov et al. (2020), which consists of annotated messages extracted from the strategic negotiation game *Diplomacy*. The dataset is uniquely valuable as it contains dual annotations: one for the sender's intent (i.e., whether the sender intended to deceive) and another for the receiver's perception (i.e., whether the receiver believed the message to be deceptive).

Each game involves seven players representing distinct European powers: England, France, Germany, Italy, Austria, Russia, and Turkey. During gameplay, messages are exchanged bilaterally as players negotiate, form alliances, and potentially betray one another. Our preprocessing pipeline transforms the raw JSON data into structured instances incorporating the following components:

- **Message Text:** The full text of the negotiation message.
- **Player Identities:** Identifiers for both the sender and receiver.
- **Deception Labels:** Binary labels for sender intent and receiver perception.
- **Game State Features:**
  - *Player Scores:* Supply centers controlled by each player.
  - *Temporal Indicators:* A game progression timestamp.
  - *Seasonal Information:* Categorical variables indicating Spring, Fall, or Winter.
  - *Game Year:* Normalized values spanning 1901–1909.

For missing player scores, we implement a cascading imputation strategy that first leverages the last available score for that player, defaulting to a value of 3.0 in the absence of prior data.

4.2 GRAPHITE Architecture

We introduce **GRAPHITE** (Graph-based Relational Analysis for Player Honesty In Tactical Exchanges), a novel neural architecture that reframes deception detection as a multi-modal learning problem, integrating textual analysis with dynamic relationship modeling.

4.2.1 Graph-Based Modeling with Relationship Dynamics

Each game of Diplomacy involves seven players: England, France, Germany, Italy, Austria, Russia, and Turkey. We model interactions among players using a fully connected undirected graph with $N = 7$ nodes, corresponding to the players. This results in $\binom{7}{2} = 21$ edges, each representing a pairwise relationship.

We employ a LSTM to compute dynamic *relationship scores* for each of the 21 player pairs. At each message timestamp, the input to the LSTM consists of:

- $S_i \in R^7$: the vector of in-game scores for all players at that point,
- $T_i = \frac{\text{abs\_index}}{\text{total\_messages}}$: the normalized timestamp within the game,
- TSY_i: a normalized encoding of the season and year of the message.

Simultaneously, the current message is encoded using BERT, yielding an embedding $E_i \in R^{768}$ for the message content.

For a given message from sender i to receiver j, we extract the relationship score $R_{ij} \in R$ from

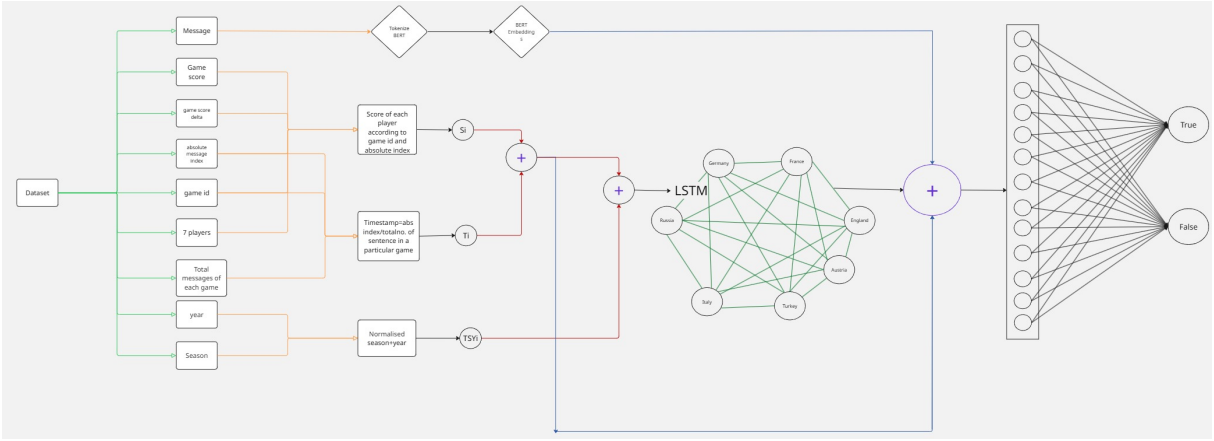


Figure 1: Overview of the GRAPHITE architecture. The model fuses linguistic cues from the message text with relational dynamics between players, using a graph-based relationship encoder to contextualize deception in gameplay.

the LSTM. This value, along with the message embedding E_i, score vector S_i, and timestamp T_i, are concatenated to form a joint representation:

$$x_i = [E_i; R_{ij}; S_i; T_i] \in R^{778}$$

This fused vector is passed through a fully connected layer to produce the final deception prediction.

4.3 Training Details

We train separate models for the two deception tasks:

- Sender Intent Detection (Actual Deception)
- Receiver Perception Detection (Perceived Deception)

4.3.1 Implementation Details

We use the `bert-base-uncased` model as our encoder with a maximum sequence length of 128 tokens. The model is trained for 15 epochs using a batch size of 16 and a learning rate of 2×10^{-5}. To address the significant class imbalance (only 4-05% of messages labeled as deceptive), we apply class weights of [1.0, 30.0] for sender and [1.0, 50.0] for receiver tasks Optimization is performed using the AdamW optimizer with a linear learning rate scheduler and no warmup steps. We have used RTX 4050 with 6GB RAM. The code and implementation details are available at our GitHub repository: link.

4.3.2 Relationship Loss

In deception detection within strategic communication, the relationship between players significantly influences the interpretation and intent of a message. To explicitly model this aspect, we incorporate a specialized *relationship loss* that biases the model toward learning from the most relevant interaction: the one between the sender and the receiver of each message.

Let $\mathbf{R} \in R^{21}$ denote the vector of predicted relationship strengths between all possible pairs of the seven players in a game, and let $\mathbf{M} \in \{0, 1\}^{21}$ be a binary *edge mask* that zeros out all entries except the one corresponding to the current sender-receiver pair.

We define the relationship loss as:

$$\mathcal{L}_{\text{relation}} = \left((\mathbf{R} \odot \mathbf{M})^2 \right)_{\text{mean}} \quad (1)$$

Here, $\odot$ denotes element-wise multiplication. This formulation penalizes high relationship scores for irrelevant edges, encouraging the model to focus on the specific sender-receiver interaction pertinent to the current message.

This loss is added to the overall training objective alongside the primary classification loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{relation}} \quad (2)$$

where $\mathcal{L}_{\text{cls}}$ is the weighted binary cross-entropy loss for deception prediction, and λ is a tunable hyperparameter that controls the influence of the relationship loss. In our experiments, we set $\lambda = 1.0$ after validation-based tuning.

This additional loss term significantly improves the model's ability to distinguish between genuine and strategic communication by grounding the deception prediction in evolving inter-player dynamics.

4.4 Evaluation Metrics

Our models are evaluated using a suite of complementary metrics:

- **Accuracy:** The overall proportion of correct predictions.
- **Macro F1 Score:** The harmonic mean of precision and recall, averaged across classes to account for class imbalance.
- **Lie F1 Score:** The F1 score computed specifically for the deceptive (positive) class, which is the primary focus of our analysis.

We utilize early stopping based on validation macro F1 and lie F1 scores, preserving the best-performing models for final evaluation on the test set.

4.5 Baseline Models

For comparative evaluation, we implement several baseline methods.

4.5.1 LLM-Based Baseline

To benchmark our GRAPHITE model against strong contemporary language models, we implement a zero-shot baseline using LLaMA 3.1-8B-Instruct, a state-of-the-art open-source large language model. This baseline evaluates the capability of powerful generative models to detect deception without explicit training on the task. .

Prompt Structure. Each prompt includes key contextual information:
Context: Game of Diplomacy;
Sender: `[sender_name]`.
Receiver: `[receiver_name]`.
Previous Message in this dialog:
"`[previous_message_text]`".
Current Message in this dialog: "`[current_message]`".

System Prompt. We configure the LLaMA 3.1 model with the following system-level instruction to ensure consistent and constrained outputs:

```
You are an analyst of the
game Diplomacy. Analyze
messages based on the
user's specific request.
Respond ONLY with the
single word 'True' or
'False'. Do not add
explanations.
```

Sender Task. For detecting actual deception, we append the following instruction:

```
Task: Predict the
sender's label for the
Current Message.
'True' means the sender
seems strategically
genuine/cooperative at
this moment.
'False' means
the sender seems
deceptive/manipulative
at this moment.
```

Receiver Task. For detecting perceived deception, we instead use:

```
Task: Predict the
receiver's likely
perception of the Current
Message.
'True' means the receiver
likely perceives the
message as strategically
genuine/cooperative from
the sender.
'False' means the
receiver likely
perceives the message as
deceptive/manipulative
from the sender.
```

Output Constraints. To constrain the model's response space, we prepend a system prompt instructing it to return only a single token—either "True" or "False". We also set the decoding temperature to 0.1 to minimize stochasticity in generation. The model's output is parsed using regular expression pattern matching to ensure robust and deterministic binary classification.

4.5.2 Prior Model Baselines

We further compare our approach against established baselines from Peskov et al. (2020):

- **BERT+Context:** Integrates text and game context features but excludes relationship modeling.
- **LSTM:** Implements sequential processing of messages.
- **Bag of Words:** Utilizes simple lexical features.

- **Random and Majority Class:** Serve as statistical baselines.

These comparisons underscore the contributions of our graph-based relationship modeling approach in advancing the state-of-the-art in deception detection.

5 Results

We evaluate three types of models on the deception classification tasks: (1) NLP models and human, (2) a graph-based neural model incorporating relationship dynamics, and (3) a local large language model (LLM) evaluated via prompt-based interaction.

5.1 Baseline Performance: Human and Traditional Models

We begin by comparing human performance with a variety of traditional machine learning models, including Bag-of-Words, LSTM, BERT-enhanced models, and variants incorporating power dynamics. Figure 3 summarizes the performance on both the **Actual Lie** and **Suspected Lie** tasks, using Macro F1 and Lie F1 scores.

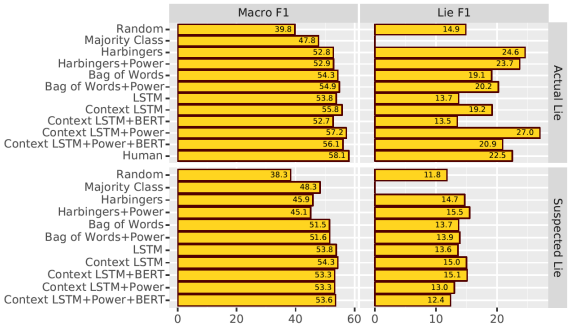


Figure 2: Comparison of models and humans on deception detection. Top: Actual Lie task. Bottom: Suspected Lie task. Left: Macro F1. Right: Lie F1.

Key observations:

- **Actual Lie Task**: Humans achieve the highest Macro F1 (58.1), narrowly outperforming all automated baselines. However, the best Lie F1 is achieved by the *Context LSTM + Power* model (27.0), indicating its relative effectiveness at identifying actual deceptive messages. Human Lie F1 is slightly lower at 22.5, followed closely by Harbingers (24.6).
- **Power and Contextual Features Help**: Adding power features or contextual modeling consistently improves performance. For instance, Bag of Words + Power (54.9) and Context LSTM + Power (57.2) outperform their respective baselines. The same pattern is observed in Lie F1 improvements.
- **Suspected Lie Task**: All models struggle more in this setting, with lower scores overall. The highest Macro F1 (51.6) is again achieved by Bag of Words + Power, while the best Lie F1 (15.5) comes from Harbingers + Power. These results underscore the greater difficulty of detecting perceived deception as opposed to ground-truth lies.
- **Random and Majority Class Baselines**: These models perform the worst, confirming that deception classification requires more than naive prediction. For example, Random achieves just 39.8 Macro F1 and 14.9 Lie F1 on the Actual Lie task.

Overall, while traditional models can approach or even exceed human-level performance in some metrics, especially when enriched with contextual or relational signals, human intuition remains strong—especially in general macro-level accuracy.

5.2 GRAPHITE Model Results

To address relational and contextual nuances, we employ a GRAPHITE model. Table 4 presents the results on both sender and receiver tasks.

Model	Macro F1	Lie F1
Human	0.581	0.225
Harbingers	0.528	0.246
Harbingers+Power	0.529	0.237
LSTM	0.538	0.137
Context LSTM	0.558	0.192
Context LSTM+ BERT	0.527	0.135
Context LSTM + Power	0.572	0.27
Context LSTM+Power+BERT	0.561	0.209
Bag of words	0.543	0.191
Bag of words+Power	0.549	0.202

Table 1: Baseline(sender) on test set.

Model	Macro F1	Lie F1	Accuracy
Context LSTM + Power	0.533	0.13	
Sender (LLM)	0.4529	0.1397	0.6359
Receiver (LLM)	0.4230	0.1151	0.5774
Sender (GRAPHITE)	**0.5781**	**0.2636**	**0.8125**
Receiver (GRAPHITE)	**0.5242**	**0.1086**	**0.8873**

Table 2: GRAPHITE model results on test set.

While the macro-level performance is competitive with traditional models, the Lie F1

Task	Accuracy	Macro F1	Lie F1
Sender (LLM)	0.6359	0.4529	0.1397
Receiver (LLM)	0.5774	0.4230	0.1151

Table 3: Performance of local LLM model (prompt-based).

Model	Macro F1	Lie F1
Harbingers	0.459	0.147
Harbingers+Power	0.451	0.155
LSTM	0.538	0.136
Context LSTM	0.543	0.15
Context LSTM+ BERT	0.533	0.151
Context LSTM + Power	0.533	0.13
Context LSTM+Power+BERT	0.536	0.124
Bag of words	0.515	0.137
Bag of words+Power	0.516	0.139

Table 4: Baseline(receiver) on test set.

scores—especially on the receiver task—are quite low. This highlights the challenge of detecting lies from the perspective of the receiver, who lacks ground-truth knowledge and relies solely on perception.

5.3 Local LLM-Based Evaluation

Finally, we assess a local large language model (LLaMA 3 via Ollama) under a zero-shot, prompt-based setup. Table 3 shows its classification performance.

The LLM underperforms compared to structured models. While accuracy is decent, Lie F1 scores are low due to overpredicting the majority class. This highlights the difficulty of zero-shot deception classification and the need for domain-specific fine-tuning.

6 Experimental Setup

Our experiment utilizes the Diplomacy dataset from (Peskov et al., 2020), comprising 13,132 training, 1,416 validation, and 2,741 test instances. GRAPHITE integrates a BERT encoder (bert-base-uncased) with a relationship-focused LSTM model that processes 10-dimensional input features (player scores, season, year, and timestamp) through a 64-dimensional hidden layer to produce 21-dimensional relationship vectors. These relationship representations are combined with BERT embeddings (768 dimensions) and additional game state features to form a 777-dimensional feature vector for classification. We train separate models for sender deception and receiver perception, using AdamW optimization (learning rate 2×10^{-5}) with batch size 16 and apply class weights ([1.0, 30.0] for sender, [1.0, 50.0] for receiver) to address the significant class imbalance. Early stopping based on validation Macro F1 and Lie F1 metrics prevents overfitting. For comparison, we implement a zero-shot LLaMA 3.1-8B-Instruct baseline using structured prompts and also benchmark against previous models from Peskov et al., including BERT+Context and simpler baselines. Performance evaluation focuses on accuracy, Macro F1, and Lie F1 scores, with particular emphasis on Lie F1 due to the rarity of deceptive messages in the dataset (only 4-5%).

7 Observation

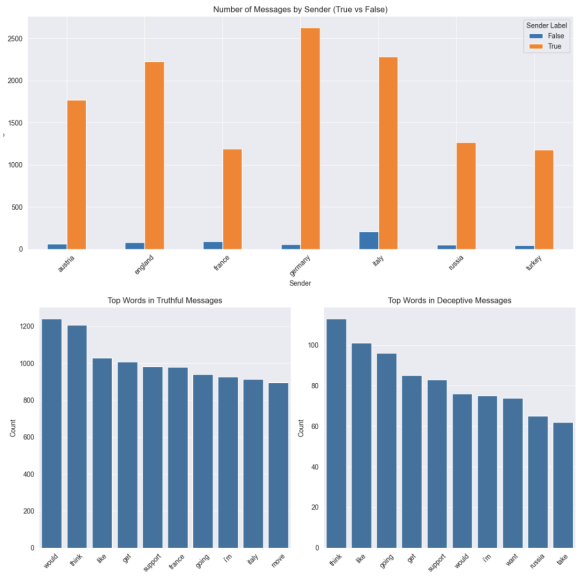


Figure 3: Observation

The most interesting observation is that Italy is the country which has told the most number of lies yet it's name comes the most in the table of top truthful words . Meanwhile, Russia's name comes the most in the table of top words in Deceptive messages.

8 Conclusion and Future Work

In this work, we address the challenging problem of deception detection in strategic, multi-party communication by leveraging the naturalistic dialogue found in the board game *Diplomacy*. Unlike prior work that often relies on artificially constructed lies or role-constrained deception scenarios, our dataset captures organically occurring deception with annotations from both the sender and the receiver, enabling a richer understanding of both intentional dishonesty and perceived manipulation.

We evaluate a wide range of approaches—from traditional machine learning classifiers and human baselines to graph-based neural models and prompt-based large language models (LLMs). Our results demonstrate several important trends. First, traditional models like Random Forests and Bag-of-Words variants perform competitively on average metrics due to class imbalance but are less effective at detecting rare, high-stakes deception events. Second, graph-based models that explicitly model inter-player relationships and game-state context can capture more nuanced forms of strategic lying, particularly when combined with contextual embeddings. Third, zero-shot prompting with LLMs is feasible but underperforms compared to task-specific models, highlighting the need for fine-tuning or structured reasoning components in such setups.

Most notably, human participants consistently set a strong benchmark, especially for macro-level accuracy and subtle deception. However, some models outperform humans in detecting explicit lies (Lie F1), suggesting that certain deception cues are detectable through pattern recognition even when they elude human interpretation. This points to a future where hybrid systems—combining human intuition with machine learning's pattern-sensitivity—could outperform either in isolation.

Beyond the realm of games, this research has broader implications. Strategic deception is not confined to fictional settings; it manifests in real-world domains such as cybersecurity, political discourse, business negotiations, and digital communication. From phishing emails and disinformation campaigns to negotiation tactics in diplomacy and commerce, the ability to detect deception has significant societal and technological value. By developing robust models that can detect lies in a realistic, interactive, and temporally-evolving context, we lay the foundation for more general-purpose tools capable of enhancing trust and transparency in digital spaces.

Our contributions are : (1) we introduce a graph-based neural architecture tailored for modeling relational dynamics in multi-agent strategic dialogue; and (2) we assess the limits and potential of LLMs in prompt-based deception classification. Together, these contributions offer a comprehensive evaluation framework for deception detection and highlight avenues for advancing future work in this domain.

Future Work. There are several promising directions for future research:

- **Dialogue Modeling**: Incorporating full conversation history through transformers or dialogue-aware graph structures could better capture evolving intent and suspicion.
- **Cross-perspective Learning**: Designing architectures that jointly model sender intent and receiver perception could improve interpretability and robustness.
- **Generalization to Other Domains**: Extending deception detection techniques to other high-stakes settings—such as negotiations, cybersecurity, or online moderation—would validate model transferability and robustness.

We hope this work lays the foundation for future systems that can model, interpret, and eventually assist in identifying complex forms of human deception across diverse communication settings.

References

Max Louwerse and 1 others. 2010. Deception detection in court testimonies. *Journal of Language and Social Psychology*, 29(3):319–341.

Myle Ott and 1 others. 2012. Deceptive opinion spam. In *Proceedings of the ACM SIGKDD Conference*, pages 201–209.

Dmitry Peskov and 1 others. 2020. Deception in diplomacy: A study of human-human communication. In *Proceedings of the 2020 International Conference on Computational Linguistics*, pages 120–130.

Thanapon Wongkamjan and 1 others. 2025. Ctrl-d: A framework for deception detection in diplomacy. In *Proceedings of the 2025 Conference on Computational Deception*, pages 77–88.