Multi-Agent Graphical Dual-Attention for Dynamic Long-Horizon Strategic Interaction

Anonymous ACL submission

Abstract

Long-horizon strategic interaction in multi-agent settings arises in negotiation dialogues, online communities, collaborative planning, and competitive games, where outcomes depend jointly on linguistic actions, temporal dynamics, and evolving inter-agent relationships. Phenomena such as deception, negotiation, legal and political discourse are central to long-term strategic interaction, yet most NLP systems still struggle to recognize these events in free- form dialogue that unfolds over many turns and shifting power dynamics. To cater this, we introduce a novel architecture, RG-DAT, a RoBERTa-based multi-agent graphical dual-attention transformer that jointly models message text, agent-state asymmetry features, and a dynamic graph of agent interactions via a graph attention encoder and a dual-attention fusion module. As our primary testbed, we focus on the online negotiation based strategic interaction, Diplomacy using the Diplomacy Deception Dataset, which uniquely annotates both sender intent and receiver perception at the message level. To assess the broader applicability of our approach beyond deception, we additionally evaluate RG-DAT on CaSiNo dataset, a corpus of campsite negotiation dialogues with rich annotations of negotiation outcomes and strategies. Experiments on Diplomacy and CaSiNo show that RG-DAT substantially outperforms strong baselines and contemporary large language models.

1 Introduction

Many real-world domains involve long-horizon strategic interaction (He et al., 2018) among multiple agents, where agents repeatedly communicate, form and dissolve alliances, and act under evolving power asymmetries (Bakhtin et al., 2022).

In multi-agent settings like diplomacy, business negotiations, and legal discourse (Wellman, 2016), events such as deception, trust, and betrayal are rarely isolated (Asher et al., 2016) (Crawford and Sobel, 1982) but emerge from a complex interplay between linguistic content, temporal context, power dynamics, and evolving social relationships. Yet most NLP systems treat these phenomena as purely lexical (Cao et al., 2018), ignoring the relational and strategic structure that shapes agent decisions in real-world interaction. We therefore study the general problem of modeling long-horizon multi-agent strategic interaction from text and interaction traces (Xu et al., 2022), and we instantiate this problem primarily on the Diplomacy Deception Dataset[1] in the game of Diplomacy (Peskov et al., 2020), while additionally evaluating on the CaSiNo campsite negotiation corpus[2] (Chawla et al., 2019) to demonstrate that the same architecture extends to a distinct negotiation domain and thus supports greater generalizability.

We introduce **RG-DAT** (**R**oBERTa-based **G**raph **D**ual-**A**ttention **T**ransformer), a state of the art architecture, that jointly models message text, agent-state asymmetry features, and a dynamic communication graph of agents. Conceptually, RG-DAT is not specific to deception as a task. Instead, it provides a general template for modeling long-horizon, multi-agent strategic interaction: nodes represent agents, edges summarize interaction history, node and edge features encode asymmetric state and relationship signals, and a dual-attention mechanism arbitrates between local message-level evidence and global interaction context.

2 Methodology

2.1 Dataset Description

We use the *Diplomacy* Deception Dataset (Peskov et al., 2020), which contains 17,289 private messages exchanged between agents in online inter-

[1] Deception in Diplomacy dataset (Convokit).
[2] CaSiNo campsite negotiation dataset.

actions of *Diplomacy*[3], along with rich metadata including message text, sender and receiver identifiers, season and year, message indices, and game scores. Each message is annotated with dual deception labels: a binary *sender* label indicating whether the sender reports the message as deceptive, and a binary *receiver* label indicating whether the receiver suspects the message is deceptive. We follow the official Convokit train/validation/test splits and formulate two binary classification tasks: (i) **sender deception detection**, which predicts whether a message is actually deceptive according to the sender's self-reported label, and (ii) **receiver suspicion detection**, which predicts whether a message is perceived as deceptive according to the receiver's label.

These patterns motivate our multimodal approach: deception manifests through *linguistic style* (text encoder), *temporal context and power dynamics* (numerical features), and *social structure* (communication graph).

2.2 RG–DAT Architecture

Building on these insights, **RG–DAT** (RoBERTa-based Graph Dual-Attention Transformer) is designed as a general framework for modeling multi-agent strategic interaction. Figure 1 shows the architecture of RGDAT. Nodes in the underlying graph correspond to agents, edges encode historical interactions between agents, and node and edge features summarize agent state, interaction frequency, and asymmetries such as power or resource imbalance. In this section, we describe one concrete instantiation of this framework on a Diplomacy-style environment; Section 5 shows how the same components are instantiated on a campsite negotiation corpus.

RGDAT jointly encodes message text, numerical power-dynamics features, and the evolving communication graph of agents. The same architecture is instantiated separately for sender intent prediction (y_s) and receiver perception prediction(y_r), sharing all components but trained with different supervision.

[3]Diplomacy is a negotiation-based strategic interaction in which seven agents, each controlling a European great power on the eve of World War I, compete to capture supply centers on a shared map. Armies move deterministically, and no randomness or hidden information is involved. Because all agents begin with equal strength, progress in the interaction is possible only through forming and breaking alliances with other agents, making strategic communication and betrayal central to gameplay (Peskov et al., 2020).

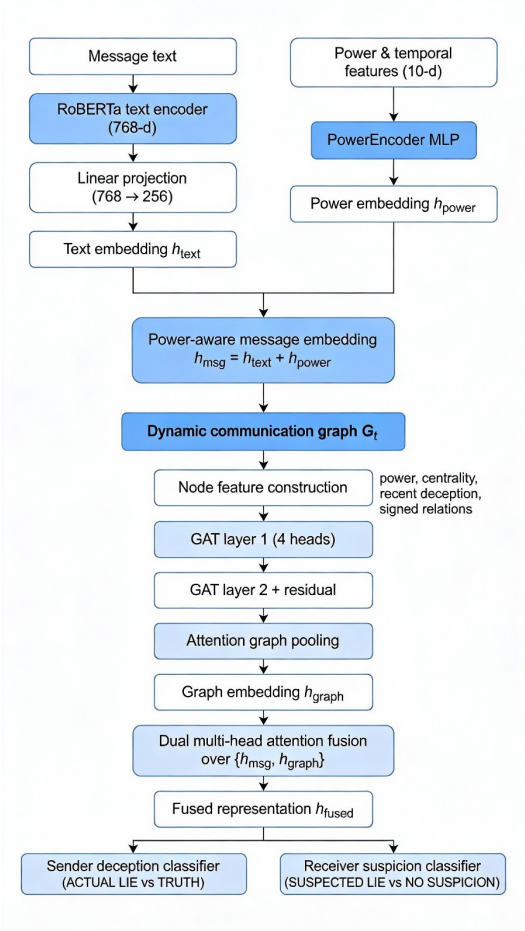


Figure 1: Architecture of the RG–DAT model.

2.2.1 Text Encoder

Each message is encoded using the roberta-base transformer with a maximum sequence length of 128 tokens. The final [CLS] representation $h_{\text{CLS}} \in R^{768}$ is projected into a shared hidden space of dimension $d = 256$. The text embedding is computed as

$$\mathbf{h}_{\text{text}} = \mathbf{W}_{\text{proj}}\, \mathbf{h}_{[\text{CLS}]} + \mathbf{b}_{\text{proj}},$$

where $\mathbf{W}_{\text{proj}} \in R^{d\times 768}$ and $\mathbf{b}_{\text{proj}} \in R^{d}$ are learnable projection parameters. RoBERTa parameters are fine-tuned end-to-end (Bengio et al., 2015), enabling the encoder to adapt to deception-specific linguistic cues.

2.2.2 Power-Dynamics Features

To complement textual information, each message is augmented with a numerical feature vector p capturing environment state and the sender's recent behavior (e.g., current power, temporal context, communication volume, and recent deceptive actions). In our Diplomacy instantiation, $p \in R^{10}$ includes the current score delta for the sender, sender–

receiver score gap, normalized year, season indicators, sender's share of messages so far, exponentially decayed deception counts, normalized absolute and relative message indices, and the fraction of deceptive messages in the sender's recent interactions.

These features are passed through a small POWERENCODER MLP with LeakyReLU activation:

$$\mathbf{h}_{\text{power}} = \text{LeakyReLU}\big(\mathbf{W}_{\text{power}}\mathbf{p} + \mathbf{b}_{\text{power}}\big),$$

where $\mathbf{W}_{\text{power}} \in R^{d\times 10}$ and $\mathbf{b}_{\text{power}} \in R^{d}$ are learnable parameters. The power-aware message embedding is then obtained via residual addition and dropout:

$$\mathbf{h}_{\text{msg}} = \text{Dropout}_{0.3}\big(\mathbf{h}_{\text{text}} + \mathbf{h}_{\text{power}}\big),$$

where $\text{Dropout}_{0.3}$ denotes dropout with rate 0.3.

2.2.3 Communication Graph and Node Features

RG-DAT explicitly models a social structure using a dynamic communication graph. For each game and timestep, nodes correspond to all agents in that game and directed edges connect any pair of agents who have exchanged at least one message up to and including the current message, with edge statistics tracking interaction counts and deception history. On this graph, we compute standard betweenness centrality for each node (e.g., via `NetworkX`) measuring how often an agent lies on shortest paths between other agents. Agents with high betweenness act as important intermediaries and may be more strategically positioned to deceive or coordinate. The graph is updated incrementally as new messages are processed.

2.2.4 Signed Relationship Scoring

Beyond simple connectivity, RG–DAT estimates a signed relationship score for each ordered agent pair (i, j) that compresses their interaction history into a single scalar reflecting cooperation, trust, and potential exploitation. Four interpretable components are computed over all messages exchanged up to the current timestep.

Mutual Communication Frequency (C_{ij}).

$$C_{ij} = \frac{\#(i \to j) + \#(j \to i)}{\text{total messages in the game so far}},$$

which reflects how intensively the pair communicates relative to overall interaction activity.

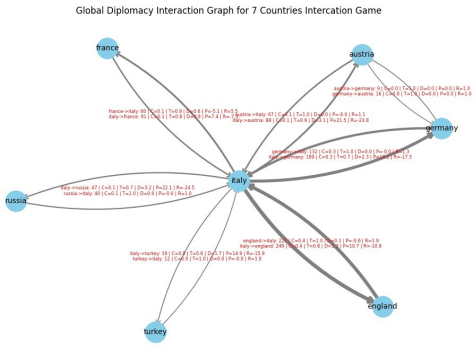


Figure 2: Global Diplomacy interaction graph for a single seven-player game, where nodes correspond to countries and directed edges encode cumulative message exchanges annotated with relationship scores $(C_{ij}, T_{ij}, D_{ij}, P_{ij}, R_{ij})$ used by the RG–DAT graph encoder.

Trust Proxy (T_{ij}).

$$T_{ij} = 1 - \frac{\#\text{ deceptive messages from } i \text{ to } j}{\#\text{ total messages from } i \text{ to } j},$$

defined when at least one message has been sent from i to j, and defaulting to 1 when there is no evidence of lying yet. Higher values indicate that most messages from i to j have been truthful, suggesting trustworthiness in that direction.

Decayed Deception Intensity (D_{ij}). This term aggregates deceptive messages from i to j with exponential decay:

$$D_{ij} = \sum_{m \in \text{lies}(i\to j)} 0.9^{\Delta t(m)},$$

where $\Delta t(m)$ denotes the number of messages elapsed since lie m occurred. Recent lies thus contribute more than older ones, encoding short-term betrayal history.

Power Exploitation (P_{ij}).

$$P_{ij} = (\max_i - \max_j) \times D_{ij},$$

where $\max_i$ and $\max_j$ approximate each agent's current power (e.g., maximum score or territorial control over time). This term highlights situations in which a more powerful agent i repeatedly lies to a weaker agent j, which is a strong indicator of exploitative behavior.

Signed Relationship Score. The four components are combined into a single signed score[4]:

$$R_{ij} = \alpha C_{ij} + \beta T_{ij} - \gamma D_{ij} - \delta P_{ij},$$

[4]The coefficients α, β, γ, and δ are set to 1.0 in order to

Positive values indicate frequent, mostly truthful communication with limited exploitation, while strongly negative values indicate a history of deceptive, power-skewed interactions. Figure 2 shows the Diplomacy interaction graph where the directed edges connecting the agents show the relationship scores.

For each agent i, the model retains

$$\max_j |R_{ij}|$$

as a node-level summary of that agent's most extreme relationship (either strongly cooperative or strongly adversarial), which is incorporated into the node feature vector.

Although we instantiate these components using deception labels and game scores in Diplomacy, the same formulation can aggregate alternative signals such as agreement violations, sentiment shifts, failed commitments, or reward-relevant events in other multi-agent domains, making the signed relationship score a flexible latent relationship encoding

2.2.5 Node Features and Graph Encoder

We model each interaction with a dynamic communication graph whose nodes correspond to agents and whose edges encode which agents have communicated with each other up to the current timestep. In our Diplomacy instantiation, the node set consists of up to seven agents, and each node is associated with a 129-dimensional feature vector. The first six dimensions encode game-relevant statistics: the agent's power score, total message count, recent deception ratio, betweenness centrality, node degree, and the strongest signed relationship score. The remaining dimensions are zero padding, included to match the input dimensionality expected by the graph attention network (GAT).

A two-layer Graph Attention Network (GAT) encodes the communication graph (Veličković et al., 2018). Let $\mathbf{X} \in R^{|V|\times 129}$ stack all node features, and let $\mathbf{E}_t$ denote the sparse edge index. The first GAT layer computes

$$\mathbf{H}^{(1)} = \mathrm{GAT}_{\text{4-head}}(\mathbf{X}, \mathbf{E}_t\,;\, \text{dropout} = 0.2)\,,$$

where $\mathbf{H}^{(1)} \in R^{|V|\times d}$ is the intermediate node representation. Multi-head attention allows the encoder to attend differently to neighbors.

keep the signed relationship score simple, interpretable, and numerically stable, without introducing additional hyperparameters.

The second GAT layer uses a single attention head with a residual connection:

$$H^{(2)} = \mathrm{ELU}\Big(\mathrm{GAT}_{\text{1-head}}(H^{(1)}, E) + W_{\text{res}} X\Big)\,.$$

where $W_{\text{res}} \in R^{d\times 129}$ projects the original node features into the hidden space, and $\mathrm{ELU}(\cdot)$ denotes the element-wise Exponential Linear Unit activation function. The residual projection $W_{\text{res}}X$ helps preserve low-level structural information and mitigates over-smoothing in the graph.

An attention-based graph pooling layer aggregates node-level representations into a single graph-level embedding:

$$\mathbf{h}_{\text{graph}} = \sum_i \mathrm{softmax}\Big(W_{\text{attn}} H_i^{(2)}\Big)\, H_i^{(2)} \in R^{256}.$$

where $W_{\text{attn}} \in R^{1\times d}$ is a learnable attention vector and $H_i^{(2)}$ is the representation of node i.

2.2.6 Dual-Attention Fusion and Classifier

The power-aware message embedding $\mathbf{h}_{\text{msg}} \in R^d$ and the graph-level embedding $\mathbf{h}_{\text{graph}} \in R^d$ are concatenated to form a length-two sequence

$$\mathbf{Z} = \begin{bmatrix} \mathbf{h}_{\text{msg}} \\ \mathbf{h}_{\text{graph}} \end{bmatrix} \in R^{2\times d}.$$

A four-head multi-head attention module operates over this two-step sequence, producing context-aware representations for both positions and learning the relative importance of textual cues versus social-structural context. The attention outputs are averaged across positions and projected to obtain a fused representation:

$$\mathbf{h}_{\text{fused}} = W_{\text{out}} \left(\frac{1}{2} \sum_{k=1}^{2} \mathrm{MultiHeadAttn}(\mathbf{Z})_k \right),$$

where $\mathrm{MultiHeadAttn}(\cdot)$ denotes the four-head attention operation and $W_{\text{out}} \in R^{d\times d}$ is a learnable projection matrix.

A final dropout layer followed by a linear classifier produces logits for binary prediction:

$$\hat{\mathbf{y}} = W_{\text{cls}}\, \mathrm{Dropout}_{0.3}(\mathbf{h}_{\text{fused}}) + \mathbf{b}_{\text{cls}},$$

where $W_{\text{cls}} \in R^{2\times d}$ and $\mathbf{b}_{\text{cls}} \in R^2$ are learnable classifier parameters. A sigmoid or softmax function is applied at inference time to obtain probabilities for deceptive versus truthful (sender task) or suspected versus not-suspected (receiver task).

This dual-attention fusion lets the model jointly reason over local message content and global interaction context in a single representation.

3 Experiments and Evaluation

3.1 Dataset Split

All experiments use the official CONVOKIT train/validation/test splits of the DIPLOMACY DECEPTION DATASET, ensuring that games do not overlap across splits. The training and validation sets are combined only for exploratory data analysis and feature design; model selection and early stopping rely on the original validation split, and final results are reported on the held-out test set.

3.2 Evaluation Metrics

The RGDAT model is evaluated using F1-score, accuracy, and the label-smoothed cross-entropy loss. Due to strong class imbalance, macro-F1 is used as the primary evaluation metric for both tasks, averaged across deceptive and truthful (or suspected and not-suspected) classes. Accuracy is also reported for completeness; however, macro-F1 better reflects performance on the minority deceptive and suspected classes and is therefore used for validation-based checkpoint selection.

3.3 LLM Baselines

To contextualize the performance of our proposed model, we evaluate four large language models (LLMs) (Zhao et al., 2023) as few-shot baselines on the same Diplomacy deception dataset: Meta–LLaMA–3.1–8B–Instruct, Sonar, Gemini–2.5–Flash–Lite, and GPT–OSS–120B. Meta–LLaMA–3.1–8B–Instruct is an 8B-parameter, instruction-tuned model released by Meta and accessed via a provider-optimized Turbo endpoint. Sonar is a proprietary instruction-tuned model served by Perplexity AI. Gemini–2.5–Flash–Lite is an efficiency-optimized instruction-tuned model from the Gemini family released by Google DeepMind. GPT–OSS–120B is a third-party, open-source GPT-style model accessed through an OpenAI-compatible API. All LLM baselines are evaluated in a few-shot in-context learning setting. For the test instances, the prompt includes eight labeled examples drawn from the training split, followed by the target message. Model parameters are kept frozen, and no fine-tuning or gradient-based updates are performed. Specifically, the LLM has to classify the message as deceptive or truthful from either the sender or receiver perspective and the resulting predictions are evaluated using the same evaluation metrics (macro F1 and accuracy) as those employed for our RGDAT model.

4 Results and Analysis

Table 1 reports test-set performance for RG–DAT on both sender deception detection and receiver suspicion detection.

Task	Loss	Acc.	Macro-F1
Sender (Actual)	0.4074	88.54%	0.6332
Receiver (Perceived)	0.4564	81.69%	0.6070

Table 1: Test-set performance of RG–DAT for sender deception and receiver suspicion detection.

4.1 Comparison with Baselines

Model	Macro F1 (Sender)	Macro F1 (Receiver)
Random	39.8	38.3
Majority Class	47.8	48.3
Bag of Words	54.3	51.5
Bag of Words + Power	54.9	51.6
LSTM	53.8	53.8
Context LSTM	55.8	54.3
Context LSTM + Power + BERT	56.1	53.6
RG-DAT	**63.3**	**60.7**

Table 2: Macro F1 (in percentage) comparison on the Diplomacy deception detection task for sender and receiver labels.

To contextualize our approach, we compare against a set of baselines (Peskov et al., 2020) for deception detection in Diplomacy messages. These start with Random predictors, which samples labels according to the empirical distribution, and Majority Class predictors, which always predicts truth. Then they move to lexical models based on logistic regression over bag-of-words features, with and without an additional scalar capturing the sender–receiver power differential (victory point gap). Neural baselines include a single-message LSTM, which models word order within each utterance, and a hierarchical Context LSTM that conditions predictions on the surrounding dialogue history. The strongest published baseline, Context LSTM+Power+BERT, augments this architecture with BERT embeddings and the power feature. Table 2 shows the comparison of various baselines with our architecture, RGDAT.

4.2 Ablated Models

Model Variant	Test Acc.	Macro-F1
No Power–Dynamics Features	0.8595	0.5273
No Dual–Attention (Simple Fusion)	0.8231	0.6118
Text–only RoBERTa	0.8792	0.5602
Text + Power (No Graph)	0.7315	0.5733
RGDAT	0.8854	0.6332

Table 3: Sender-side ablation results for RG–DAT variants

Variant	Acc.	Macro-F1
No Power Dynamics Features	0.9088	0.5449
No Dual Attention (Simple Fusion)	0.6465	0.5054
Text Only RoBERTa	0.9062	0.4941
Text + Power (No Graph)	0.8785	0.5789
RGDAT	0.8169	0.6070

Table 4: Receiver-side ablation results for RG–DAT variants

The ablation models show how performance degrades when individual components of RG-DAT are removed, revealing their impact on accuracy and macro-F1. Drops in both metrics for variants without the graph encoder, relationship features, agent-state features, or dual-attention fusion indicate that each contributes complementary information, and that the full combination is needed to achieve the best results. Table 3 and Table 4 shows how RGDAT's full architecture performs better than the ablated models on the sender intent and receiver perception.

4.3 Comparison with LLMs

Model	Macro-F1
GPT–OSS–120B (sender)	0.2340
GPT–OSS–120B (receiver)	0.3066
Gemini–2.5–Flash–Lite (sender)	0.4688
Gemini–2.5–Flash–Lite (receiver)	0.5000
Meta–LLaMA–3.1–8B–Instruct (sender)	0.4604
Meta–LLaMA–3.1–8B–Instruct (receiver)	0.4324
Perplexity Sonar (sender)	0.3382
Perplexity Sonar (receiver)	0.4261
RG–DAT (sender)	**0.6356**
RG–DAT (receiver)	**0.6013**

Table 5: Macro-F1 performance of RG–DAT and large language model baselines on the Diplomacy deception detection task under sender-centric and receiver-centric evaluation settings.

Table 5 shows how RG–DAT achieves the highest Macro F1 in both sender and receiver settings, outperforming all evaluated large language model baselines.

4.4 Analysis

The sender-centric model achieves the strongest overall performance, with higher test accuracy and a slightly higher macro-F1 score than the receiver-centric variant. This suggests that actual deceptive behavior is somewhat more predictable than human suspicion, which may be influenced by additional subjective or contextual factors beyond message content and interaction structure.

Despite substantial class imbalance, both models achieve macro-F1 scores in the range of 0.60–0.63, indicating that RG–DAT captures a meaningful fraction of deceptive cases beyond what LLM baselines can achieve.

Training curves for loss, accuracy, and macro-F1 reveal rapid convergence on the training set, where both accuracy and F1 approach 0.99 after only a few epochs. In contrast, validation macro-F1 plateaus and begins to oscillate, while training loss continues to decrease. This divergence indicates the onset of overfitting and motivates the use of early stopping for checkpoint selection.

Across both sender and receiver settings, validation accuracy remains relatively high even when validation macro-F1 degrades. This behavior reflects increasing confidence on the majority truthful class and highlights the limitations of accuracy as an evaluation metric in highly imbalanced deception detection tasks. Consequently, macro-F1 is a more appropriate metric for model selection and comparison, as it better reflects performance on the minority deceptive and suspected classes (Kiela et al., 2021).

Although the LLM captures some surface cues of deception, its macro F1 remains below that of RG–DAT, suggesting that explicit modeling of interaction structure and power dynamics is still beneficial beyond generic language understanding.

5 Generalization and Cross-Validation of RGDAT Across Domains

5.1 CaSiNo campsite negotiation corpus

We additionally evaluate our RGDAT architecture on CaSiNo (Chawla et al., 2019), a corpus of 1,030 two-party campsite negotiation dialogues where participants bargain over Food, Water, and Firewood packages based on private preferences. Each dialogue contains alternating utterances between two negotiators, along with dialogue-level outcomes (points scored, self-reported satisfaction, and opponent likeness) and utterance-level annotations of negotiation strategies such as proposing offers, sharing preferences, and building rapport. Unlike Diplomacy, CaSiNo does not explicitly annotate deception; instead it captures cooperative and competitive bargaining behavior in a symmetric, two-agent setting, providing a complementary testbed for strategic interaction modeling.

5.2 Instantiating RG-DAT on CaSiNo

RG-DAT is instantiated on CaSiNo (Chawla et al., 2019) demonstrating its flexibility as a general multi-agent framework. On CaSiNo, we reuse the same RG-DAT architecture (text encoder, power encoder, communication graph, GAT, and dual-attention fusion) but instantiate the numerical features and graph structure to match the two-party negotiation setting. Each dialogue is treated as a sequence of utterances exchanged between two negotiators, which serve as the interacting agents, and the model reuses the same RoBERTa text encoder, graph attention encoder, dual-attention fusion module, and training setup originally developed for Diplomacy. Only the data loader and label space are adapted to CaSiNo's utterance-level strategy prediction task, where RG-DAT achieves strong performance, indicating that the architecture can be instantiated on a distinct negotiation dataset without architectural changes.

5.3 Results and evidence of generalizability

On CaSiNo, RG-DAT outperforms text-only RoBERTa baselines and non-graph variants on both strategy prediction and outcome prediction, achieving higher macro-F1 and accuracy across all evaluated tasks. Taken together with the gains observed on the Diplomacy Deception Dataset, these results indicate that modeling agents, interaction graphs, and dual-attention fusion between local and global context provides benefits across heterogeneous strategic-interaction domains, supporting RG-DAT's role as a general framework rather than a Diplomacy-specific deception detector.Table 6 shows how RGDAT is better than the other baseline models on the CaSiNo dataset.

Model	Loss	Acc.	Macro-F1
LSTM	0.3361	0.9096	0.5197
Bag-of-Words + LR	0.4481	0.8242	0.5756
RoBERTa (text only)	0.4481	0.8759	0.6417
RG–DAT	0.2727	0.9648	0.8776

Table 6: Test performance of baseline models and RG–DAT on the CaSiNo dataset. †Loss not reported; value shown for layout consistency.

6 Related Work

Early work on deception detection relied on surface-level cues such as n-grams, part-of-speech tags, and psycholinguistic features (Mihalcea and Strapparava, 2009) (Ott et al., 2011) (Feng et al., 2012), typically combined with linear or shallow neural models applied to reviews and short, crowdsourced statements. These studies found small but consistent differences between deceptive and truthful texts, but also showed that such cues are brittle and highly domain-dependent. More recent approaches leverage pre-trained language models and contextual information (Papangelis et al., 2019). *BERTective* (Fornaciari et al., 2021) demonstrate that augmenting BERT with short, same-speaker context improves deception detection performance in courtroom transcripts, although cross-domain generalization remains limited.

The Diplomacy Deception Dataset in *"It Takes Two to Lie: One to Lie and One to Listen"* (Peskov et al., 2020), in which each private message exchanged in online Diplomacy games is annotated with both ACTUAL LIE labels capturing sender intent and SUSPECTED LIE labels capturing receiver perception. Using this dataset, they proposed BERT+LSTM models augmented with simple power-based and contextual features, achieving macro-F1 scores in the range of 0.55–0.60 for both deception production and perception tasks. Subsequent work on Diplomacy (QANTA Research Team, 2020) has focused primarily on strategic reasoning and language generation rather than fine-grained deception modeling, leaving relatively little prior work that jointly captures deception, alliance structure, and game-theoretic incentives on this corpus.

Graph neural networks, particularly Graph Attention Networks (Veličković et al., 2018) (Kipf and Welling, 2017), are widely used to encode relational structure in social and conversational tasks (Ouyang et al., 2021) by modeling users or utterances as nodes and their interactions as edges (Bolleddu, 2025) (Shu et al., 2019). Prior work on stance detection, rumor verification, and conversation classification demonstrates that incorporating graph structure complements textual representations and that attention mechanisms help improve robustness to noisy or weakly relevant neighbors. More recent deception detection systems similarly leverage conversation graphs or speaker interaction networks (Levitan and Hirschberg, 2011) in social-media and dialogue settings (Shu et al., 2019) (Panda and Levitan, 2022). However, existing models for Diplomacy have not combined rich player-level graph features with message-level language modeling within a single, unified archi-

tecture (Kulkarni et al., 2025).

6.1 Our novel contribution

Several gaps remain in prior work on modeling strategic interaction, particularly in settings with multiple agents, long-horizon dependencies, and explicit social structure. Existing deception-detection models demonstrate that surface cues and contextual BERT-style encoders can be effective, but they largely ignore long-term power shifts, alliance structures, and game-theoretic incentives that shape agent behavior in strategic negotiation. Multimodal and graph-based approaches leverage social and structural signals, yet they are typically designed for interviews, courtrooms, fake news, or generic dialogue data, and rarely distinguish between sender intent and receiver suspicion in long-running, dual-annotated interactions such as those found in Diplomacy.

In contrast, this work (i) introduces RG-DAT, a RoBERTa-based multi-agent graphical dual-attention framework that jointly encodes message text, agent-state asymmetry features, and a dynamic agent communication graph with latent relationship scores, and (ii) instantiates this framework on the Diplomacy Deception Dataset to model both agent intent and agent perception. (iii) We train parallel sender-centric and receiver-centric models while benchmarking several modern large language models, providing, to our knowledge, the first systematic comparison between graph-augmented multi-agent modeling and instruction-tuned LLMs on long-horizon strategic interaction in Diplomacy.

7 Conclusion

RG-DAT demonstrates that combining contextual language modeling with graph-based multi-agent social reasoning is an effective approach to modeling long-horizon strategic interaction in the game of Diplomacy. By jointly encoding message text, agent-state asymmetry features, and interaction-graph structure through a dual-attention architecture, the model achieves strong performance on both sender-centric (intent) and receiver-centric (perception) prediction tasks. While this paper focuses on deception as the supervised signal, the underlying architecture is task-agnostic and can support other message-level prediction problems in multi-agent environments given appropriate labels and domain features. More broadly, this work suggests a path toward richly contextualized deception detection systems that move beyond isolated utterances to reason over longer-term interactions, social structure, and power dynamics.

8 Future Work

Future work will focus on strengthening both the linguistic and practical aspects of our approach. While RoBERTa was kept frozen for efficiency, fine-tuning it on in-domain examples of deceptive and truthful dialogue may further improve the model's semantic sensitivity to subtle cues. Explainability is another key direction where integrating attention visualizations and post-hoc interpretation methods such as SHAP or LIME would help reveal which textual and social features drive individual predictions, making the system more transparent to analysts and players.

9 Limitations

RG–DAT is architecturally complex, combining a large pretrained language model, power-dynamics features, and a graph attention network, which makes training and deployment computationally expensive. In addition, the evaluation treats agent labels as ground truth without modeling annotator disagreement or uncertainty, and uses standard classification metrics that do not capture downstream impacts of errors in strategic interaction, such as cascading miscalibration of trust over long horizons.

10 Ethical Considerations

This work raises several ethical considerations that warrant careful acknowledgment. The data used in our experiments are drawn from completed games of *Diplomacy*, in which in-game messages are publicly logged. All messages are treated strictly as textual artifacts, without any attempt to link them to real-world identities or to profile individual players beyond the scope of the game itself. Moreover, deception is an expected and integral component of *Diplomacy*, and the proposed model is trained and evaluated solely within this bounded gaming context. It is therefore not intended for use in monitoring or judging deception in everyday human communication. More generally, any application of RG-DAT or similar multi-agent interaction models outside bounded game settings must be framed

as modeling strategic patterns under observable interaction histories, not as definitive detectors of "truth" or "deception," and should include clear communication of uncertainty and domain limitations.

At the same time, any system designed to detect deception carries the risk of misuse, particularly in settings such as workplaces, political discourse, or online communities, where automated judgments could disproportionately affect vulnerable individuals or groups. For this reason, RG–DAT should be viewed as a research prototype rather than a deployable lie-detection tool. Its predictions are inherently probabilistic and imperfect, and any future application would require careful human oversight, transparent communication of uncertainty, and explicit safeguards to ensure that model outputs are not treated as definitive evidence of dishonesty.

References

Nicholas Asher, Julie Hunter, and Luca Moretti. 2016. Strategic conversation. *Semantics and Pragmatics*, 9.

Anton Bakhtin, Adam Wu, Chris Anastasiades, Marcin Andrychowicz, Ioannis Antonoglou, Jack Rae, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Yoshua Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*.

Dinesh Bolleddu. 2025. Dialogue diplomats: An end-to-end multi-agent reinforcement learning framework for automated conflict resolution. arXiv preprint arXiv:2511.17654.

Yiping Cao, Rui Zhao, Steffen Eger, and 1 others. 2018. Strategic dialogue management via deep reinforcement learning. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 448–457.

Kushal Chawla, Sean Welleck, Robert West, and Kyunghyun Cho. 2019. Casino: A corpus of campsite negotiation dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 522–532.

Vincent P. Crawford and Joel Sobel. 1982. Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 171–175.

Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. Bertective: Language models and contextual information for deception detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2699–2708.

He He, Jianshu Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in neural negotiation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Douwe Kiela, Max Bartolo, Yixin Nie, and 1 others. 2021. Dynabench: Rethinking benchmark design in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 893–908.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*.

Akshay N. Kulkarni, Alex Liu, John-Ryan Gaglione, Daniel Fried, and Ufuk Topcu. 2025. Dynamic coalition structure detection in natural language-based interactions. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*.

Sarah Ita Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment. In *Proceedings of Interspeech 2011*, pages 1877–1880.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312.

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*, pages 309–319.

Shuo Ouyang and 1 others. 2021. Dialogue graph modeling for conversation machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1249–1259.

Suryadeep Panda and Sarah Ita Levitan. 2022. Deception detection within and across domains: Identifying and understanding the performance gap. *ACM Journal of Data and Information Quality*, 14(4):1–27.

Alexandros Papangelis, Jiwei Li, Abhinav Rastogi, and 1 others. 2019. Collaborative multi-agent dialogue

model training via reinforcement learning. In *Proceedings of the First Workshop on Advances in Language Generation*, pages 92–102.

Denis Peskov, Brandon 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*, pages 3811–3854.

QANTA Research Team. 2020. The QANTA diplomacy research platform. https://sites.google.com/view/qanta/projects/diplomacy. Accessed: 2025-01.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 312–320.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations*.

Michael P. Wellman. 2016. Putting the agent in agent-based modeling. *Autonomous Agents and Multi-Agent Systems*, 30(6):1175–1189.

Peng Xu, Xiatian Zhu, and Yi Yang. 2022. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*.

Wayne Xin Zhao and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2310.19736*.

A Training Setup

Both sender and receiver models share the same training configuration:

- Label-smoothed cross-entropy loss with smoothing factor $\epsilon = 0.1$.
- `WeightedRandomSampler` with inverse class-frequency weights to mitigate class imbalance.
- AdamW optimizer with separate learning rates: 1×10^{-5} for RoBERTa parameters and 2×10^{-4} for all newly added layers.
- Linear warmup over the first 100 steps followed by linear decay over 20 epochs.
- Regularization via batch size 8 with gradient accumulation over 4 steps, gradient clipping at norm 1.0, and dropout applied in both text and fusion layers.
- Early stopping with patience 5 based on validation macro-F1.

B Exploratory Data Analysis

The Diplomacy deception dataset reveals several important characteristics of deception in strategic negotiation. We first inspected the schema of the training and validation splits and computed descriptive statistics to verify data types, ranges, and missing values.

Severe class imbalance. Deception is rare: sender labels show approximately 95.55% truthful (13,901) vs. 4.45% deceptive (647), while receiver labels show about 87.26% not lie suspected (12,695) vs. 4.21% suspected lie (613) and 8.52% no annotation (1,240). This motivates the use of weighted sampling and label smoothing.

Temporal structure. Deceptive sender labels are more concentrated toward the end of games: roughly 32.15% early, 30.91% mid-game, and 36.94% late-game. This aligns with Diplomacy's game-theoretic incentives, where betrayal becomes profitable only after trust is established.

Linguistic markers. Deceptive messages contain more hedging language (e.g., "maybe", "possibly", "don't worry"), and use more strategic game terms (e.g., "support", "move", "army", "attack") than truthful messages.

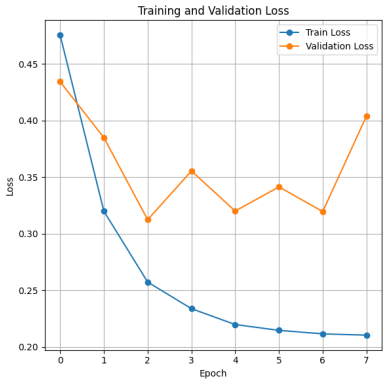
C LLM Few-Shot Prompt Template

To evaluate large language models as baselines, we use the following few-shot prompt template with eight labeled examples from the training set. The prompt instructs the model to classify each message as either `DECEPTIVE` or `TRUTHFUL` based on game context and message content.

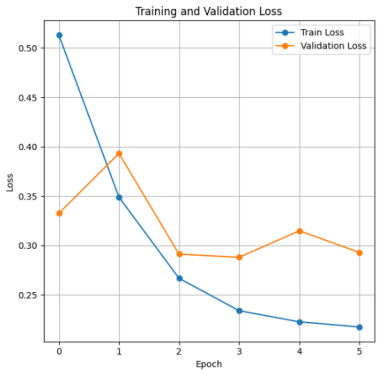
```
Consider yourself as a dynamic strategic
interaction evaluator. You are evaluating
negotiation messages from the board game Diplomacy.
In each example, one player (the sender) sends a
private message to another player (the receiver).
Your task is to decide whether the sender is
lying/deceptive in this message, given the game
context.

DEFINITION OF THE TASK
A message is **DECEPTIVE** if the sender's stated
intentions, promises, or descriptions of the game
state are not genuinely meant, or are misleading,
with the goal of gaining an advantage over the
receiver.
A message is **TRUTHFUL** if it honestly reflects
the sender's real plans and beliefs about the
situation, even if the sender is cautious,
strategic, or withholds some information.

You must classify each message as exactly one of:
- DECEPTIVE
- TRUTHFUL

You should:
1. Read the game context (year, season, scores).
2. Read the private message from sender to receiver.
3. Reason briefly about whether the sender is
   likely lying or telling the truth.
4. Output a final label in the format:
   PREDICTION: DECEPTIVE or PREDICTION: TRUTHFUL

[8 demonstration examples omitted for brevity]

TARGET EXAMPLE (TO CLASSIFY)
Game context:
- Year: {year}
- Season: {season}
- Sender (power): {sender}
- Receiver (power): {receiver}
- Sender current score: {sender_score}
- Receiver current score: {receiver_score}

Message:
{message_text}

Now, for ONLY this target example:
1. Briefly reason step by step about whether the
   sender is likely lying to the receiver.
2. On the last line ONLY, output the final label
   in the exact format:
   PREDICTION: DECEPTIVE or PREDICTION: TRUTHFUL
```

The full prompt with all eight demonstration examples is provided in our code repository. Each demonstration includes game context (year, season, sender and receiver powers, scores), the message text, reasoning about deception cues, and the ground-truth label.

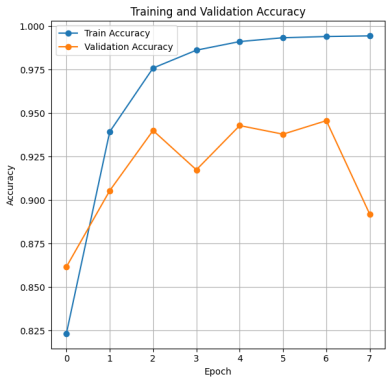
D Training Curves



(a) Sender loss

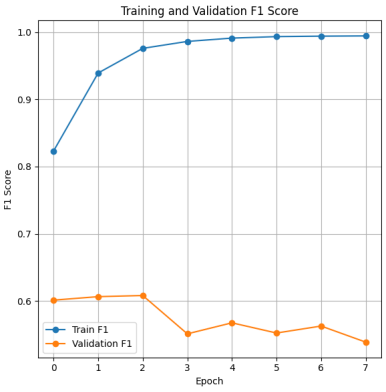


(b) Receiver loss



(c) Sender accuracy

(d) Receiver accuracy



(e) Sender macro-F1

(f) Receiver macro-F1

Figure 3: Training and validation curves for deception detection on the Diplomacy dataset. The plots show loss, accuracy, and macro-F1 across training epochs for both sender-centric and receiver-centric models.