# Beyond Words: Modeling Inter-Word Relationships with Edge Graph Neural Networks for Fake News Detection

Anonymous ACL submission

## Abstract

Fake news detection remains challenging due to the subtle linguistic cues that differentiate fabricated content from factual reporting. Fake news exhibits distinctive patterns in how words relate to each other, such as unusual semantic associations between entities, inconsistent relationship chains, and anomalous co-occurrence patterns that differ from those found in authentic news. However, existing methods typically treat text as sequences of tokens rather than explicitly modeling these inter-word relationships.

In this paper, we emphasize on identifying critical signals for fake news detection that are not adequately captured by current approaches. Essentially, our approach explicitly models word relationships through learnable edge embeddings. We present WR-EGNN (Word Relationship–based Edge Graph Neural Network), a framework for fake news detection that explicitly models atypical inter-word relationships by combining transformer-derived contextual representations with graph-based structural modeling. In addition, our approach emphasizes (1) interpretability, as the model explicitly learns the relational patterns that distinguish fake news from real news, and (2) robustness, since structural features are inherently less susceptible to adversarial style manipulations. Furthermore, the results demonstrate that WR-EGNN significantly outperforms three transformer-only models as well as four baseline fake news detection systems.[1]

## 1 Introduction

Fake news can distort public perception, leading to widespread misinformation that complicates critical issues (Prieur et al., 2023). For example, during health crises, inaccurate reports can cause panic, leading people to reject factual medical advice in favor of sensational headlines. This phenomenon has been particularly evident in discussions about vaccinations, where false claims have discouraged people from seeking life-saving immunizations, thus endangering public health.

Moreover, the political arena is not immune to the influences of fake news. Misleading narratives can influence opinions, undermine trust in institutions, and polarize societies, fueling division and conflict. Issues that could have been addressed with reasoned debate become pawns in a dangerous game of misinformation, where facts seem to matter less than sensational rhetoric (Budak, 2019). Current approaches to fake news detection face several significant limitations. While transformer-based models such as BERT and RoBERTa have shown promising results in this domain, they primarily rely on general-purpose token representations (e.g., [CLS] or <s> tokens) that aggregate information globally across the entire document. However, this approach may not optimally capture the specific linguistic patterns and word relationships that are crucial for distinguishing fake from real content. Fake news often exhibits distinctive characteristics in terms of unusual semantic associations, inconsistent entity relationships, and anomalous syntactic structures. These subtle but important relational patterns between words are particularly indicative of fake news, yet they may be lost when models rely solely on global document representations. Furthermore, most existing fake news detection models suffer from limited explainability, making it difficult for users and researchers to understand the decision-making process and identify the specific linguistic features that contribute to classification decisions.[2]

Fake news exhibits distinctive patterns in how words relate to each other, such as unusual semantic associations between entities, inconsistent relationship chains, and anomalous co-occurrence pat-

[1] Executable code of our model is available at: https://anonymous.4open.science/r/WRGNN-7604

[2] Appendix A presents the related work.

terns that differ from those found in authentic news. However, existing methods typically treat text as sequences of tokens rather than explicitly modeling these inter-word relationships. Our approach focuses on identifying inter-word relationships for fake news detection. We construct linguistically informed graph representations in which nodes represent words and edges capture the relationships between them. Graph Neural Networks (GNNs) offer a natural framework for capturing relational information; however, most current applications in natural language processing focus primarily on node-level representations rather than fully exploiting edge-level information. We present WR-EGNN (Word Relationship-based Edge Graph Neural Network), a novel architecture that explicitly models inter-word relationships through learnable edge embeddings to enhance fake news detection. These edge embeddings encode semantic, contextual, and structural relational patterns, while the graph structure is initialized using syntactic dependencies to provide a robust structural scaffold. As a result, our framework learns relationship-aware graph representations that capture subtle relational cues indicative of fake news.

This paper makes the following primary contributions.

1. We introduce WR-EGNN, a novel approach that explicitly models inter-word relationships through learnable edge embeddings integrated with transformer models for fake news detection. Unlike existing methods that primarily focus on node- or word-level representations, our Edge GNN learns rich relational encodings that capture semantic associations, contextual dependencies, and structural patterns between words.

2. Our approach demonstrates strong robustness against LLM-empowered style attacks, in which adversaries leverage large language models to rewrite content while preserving semantic meaning but altering stylistic characteristics.

3. The results demonstrate significant performance improvements over state-of-the-art baselines across multiple benchmark datasets.

## 2 Methodology

We propose WR-EGNN (Word Relationship-based Edge Graph Neural Network), which explicitly models inter-word relationships through learnable edge embeddings. Our approach combines the contextual modeling strength of transformer architectures with structural insights from syntax-based graph representations, enabling the model to capture both semantic content and linguistic structure.

### 2.1 Overview

The key innovation our approach lies in treating relationships between words as first-class entities through learnable edge embeddings, rather than implicitly encoding relationships within node representations as in traditional GNN approaches. Figure 1 illustrates our complete architecture, which consists of four main components:

1. **Text Graph Construction** (Section 2.2): Converting input text into linguistically-informed graph representations where nodes represent tokens and edges capture syntactic dependencies

2. **Transformer Integration** (Section 2.4): Generating contextual word embeddings using pre-trained models (RoBERTa) that serve as node initializations

3. **Edge Graph Neural Network** (Section 2.3): Learning relationship-aware edge embeddings through attention-based message passing that explicitly captures inter-word relational patterns

4. **Aggregation and Classification** (Section 2.5): Fusing multi-level edge representations for document-level fake news detection

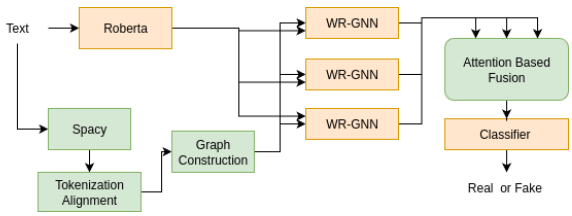


Figure 1: **Architecture of the Proposed Model.** The pipeline processes text through (1) dependency parsing and graph construction, (2) transformer encoding for contextual representations, (3) edge-centric graph neural networks capturing inter-word relationships across multiple layers, and (4) attention-based fusion for classification.

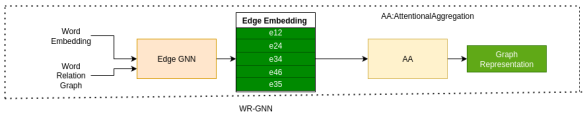


Figure 2: **Architecture of the WR-EGNN Layer.** Each layer transforms node embeddings into edge representations through: (1) separate source and target transformations, (2) edge feature generation via concatenation, (3) attention weighting to capture relationship importance, and (4) final edge embedding computation.

### 2.2 Text Graph Construction

We transform text into graph structures that preserve linguistic information while remaining compatible with transformer subword tokenization. Our construction pipeline consists of three stages:

#### 2.2.1 Step 1: Dependency Parsing

We utilize spaCy[3] with the `en_core_web_sm` language model to extract syntactic dependency relations. Dependency parsing identifies grammatical relationships such as subject-verb dependencies (`nsubj`), object relationships (`dobj`), and modifier connections (`amod`, `advmod`), creating a directed tree structure over tokens. While alternative graph construction methods exist (co-occurrence windows), we choose dependency parsing because syntactic structures exhibit cross-domain stability—core grammatical relations remain relatively consistent across domains and writing styles (McDonald et al., 2013; Cristofaro, 2009).

#### 2.2.2 Step 2: Tokenization Alignment

Transformer models employ subword tokenization (e.g., Byte-Pair Encoding in RoBERTa), which often fragments linguistic tokens into multiple subword units. Algorithm B (Appendix B) details our alignment procedure, which maps each spaCy token to its corresponding span of subword tokens.

#### 2.2.3 Step 3: Graph Formation

Using the dependency structure and alignment mapping, we construct a directed graph $G = (V, E, \mathcal{R})$ where:

- **Nodes** $V = \{v_1, v_2, \ldots, v_n\}$: Each node $v_i \in V$ represents a subword token. Nodes are initialized with contextual embeddings $\mathbf{h}_i \in \mathbb{R}^d$ from the transformer model (Section 2.4), where $d$ is the hidden dimension (768 for RoBERTa-base).
- **Edges** E = $\{(v_i, v_j)|$: dependency exists between tokens containing $\{v_i, v_j\}$ Each directed edge $(v_i, v_j) \in E$ encodes a syntactic dependency from source token $v_i$ to target token $v_j$.
- **Self-loops**: We add self-loops $(v_i, v_i)$ for every node to preserve token-level contextual information during message passing, ensuring nodes can attend to their own features.

**Graph properties:** The resulting graph is typically sparse (average degree $\approx$ 2-3 edges per node for dependency trees) and directed (reflecting asymmetric grammatical relationships like subject→verb vs. verb→object). Graph size scales linearly with document length.

### 2.3 Edge Graph Neural Network (EGNN)

The core of our approach is an Edge Graph Neural Network that explicitly learns edge representations encoding inter-word relationships. Unlike traditional GNNs that focus primarily on updating node embeddings through neighbor aggregation, our EGNN treats edges as first-class entities with their own learnable representations. Traditional node-centric GNNs encode relationship information implicitly within updated node embeddings—a node's representation reflects its neighbors, but the specific nature of relationships remains opaque. For fake news detection, we hypothesize that *relationship patterns themselves* (how entities relate to each other, how claims connect to evidence) are discriminative. Edge embeddings explicitly represent these relationships, enabling:

#### 2.3.1 Architecture

Figure 2 illustrates the EGNN layer architecture. Given node embeddings from the transformer and graph structure from dependency parsing, each EGNN layer computes edge embeddings through four steps:

**1. Source and Target Node Transformations**

We apply separate learned transformations to source and target nodes:

$$\mathbf{n}_i^s = f_{\text{MLP}_s}(\mathbf{h}_i) = \mathbf{W}_s\mathbf{h}_i + \mathbf{b}_s \quad (1)$$

$$\mathbf{n}_j^t = f_{\text{MLP}_t}(\mathbf{h}_j) = \mathbf{W}_t\mathbf{h}_j + \mathbf{b}_t \quad (2)$$

where $\mathbf{W}_s, \mathbf{W}_t \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_s, \mathbf{b}_t \in \mathbb{R}^d$ are learnable parameters. Separate transformations for source and target nodes allow the model to

[3] https://spacy.io

learn asymmetric relationship encodings (e.g., subject→verb differs from verb→subject).

**2. Raw Edge Feature Generation**

We concatenate transformed source and target representations:

$$\mathbf{e}_{ij}^{\text{raw}} = [\mathbf{n}_i^s; \mathbf{n}_j^t] \in \mathbb{R}^{2d} \tag{3}$$

where [ ; ] denotes concatenation. This captures the joint information about both endpoints of the relationship.

**3. Attention Mechanism**

Not all relationships are equally important for classification. We compute attention weights to emphasize discriminative edges:

$$\alpha_{ij} = \sigma(f_{\text{att}}(\mathbf{e}_{ij}^{\text{raw}})) \tag{4}$$

$$f_{\text{att}}(\mathbf{x}) = \mathbf{W}_{\text{att}_2} \cdot \text{ReLU}(\mathbf{W}_{\text{att}_1}\mathbf{x} + \mathbf{b}_{\text{att}_1}) + \mathbf{b}_{\text{att}_2} \tag{5}$$

where $\mathbf{W}_{\text{att}_1} \in \mathbb{R}^{128\times 2d}$, $\mathbf{W}_{\text{att}_2} \in \mathbb{R}^{1\times 128}$, $\mathbf{b}_{\text{att}_1} \in \mathbb{R}^{128}$, $\mathbf{b}_{\text{att}_2} \in \mathbb{R}$, and $\sigma$ is the sigmoid function producing attention weights $\alpha_{ij} \in [0, 1]$.

We apply attention element-wise to raw edge features:

$$\mathbf{e}_{ij}^{\text{attended}} = \mathbf{e}_{ij}^{\text{raw}} \odot \alpha_{ij} \tag{6}$$

where $\odot$ represents element-wise multiplication (broadcasting $\alpha_{ij}$ across dimensions).

**4. Final Edge Embedding**

We project attended features to final edge embeddings:

$$\mathbf{e}_{ij} = f_{\text{edge}}(\mathbf{e}_{ij}^{\text{attended}}) \tag{7}$$

where $f_{\text{edge}} : \mathbb{R}^{2d} \to \mathbb{R}^{d'}$ is implemented as:

$$f_{\text{edge}}(\mathbf{x}) = \text{LayerNorm}(\text{GELU}(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e)) \tag{8}$$

with $\mathbf{W}_e \in \mathbb{R}^{d'\times 2d}$, $\mathbf{b}_e \in \mathbb{R}^{d'}$. We use $d' = 768$ for edge embedding dimension.

#### 2.3.2 Multi-Layer Processing

We employ three EGNN layers with shared node embeddings but independent parameters:

$$\mathbf{E}^{(1)} = \text{EGNN}_1(\mathbf{H}_{\text{graph}}, E) \tag{9}$$

$$\mathbf{E}^{(2)} = \text{EGNN}_2(\mathbf{H}_{\text{graph}}, E) \tag{10}$$

$$\mathbf{E}^{(3)} = \text{EGNN}_3(\mathbf{H}_{\text{graph}}, E) \tag{11}$$

where $\mathbf{E}^{(k)} = \{\mathbf{e}_{ij}^{(k)} | (i, j) \in E\}$ represents the set of edge embeddings from layer $k$.

### 2.4 Transformer Integration

We use pre-trained transformer models to provide contextualized node initializations for our graph network.

#### 2.4.1 RoBERTa Encoder Processing

Given an input text tokenized as sequence $X = [x_1, x_2, \ldots, x_n]$ (where tokens include special tokens like <s> and </s>), we process through RoBERTa:

$$\mathbf{H}_{\langle s\rangle}, \mathbf{H}_{\text{tokens}}, \mathbf{H}_{\text{all}} = \text{RoBERTa}(X) \tag{12}$$

where:$\mathbf{H}_{\langle s\rangle} \in \mathbb{R}^d$: Embedding of the <s> start token from the final layer.$\mathbf{H}_{\text{tokens}} \in \mathbb{R}^{n\times d}$: Final layer embeddings for all tokens (including <s> and </s>).$\mathbf{H}_{\text{all}}$: All intermediate layer hidden states For graph processing, we extract token embeddings excluding special tokens:

$$\mathbf{H}_{\text{graph}} = \mathbf{H}_{\text{tokens}}[:, 1 : n-1, :] \in \mathbb{R}^{(n-2)\times d} \tag{13}$$

where we exclude the <s> token (position 0) and </s> token (position $n-1$), retaining only content token embeddings that correspond to nodes in our dependency graph.

### 2.5 Edge Aggregation and Document Representation

EGNN layers produce edge-level representations, but classification requires document-level features. We aggregate edge embeddings through attention-weighted pooling and multi-layer fusion.

#### 2.5.1 Attentional Edge Aggregation

For each EGNN layer $k$, we aggregate its edge embeddings $\mathbf{E}^{(k)} = \{\mathbf{e}_{ij}^{(k)}\}$ into a single document representation:

$$g_i = \text{GELU}(\mathbf{W}_2\text{GELU}(\mathbf{W}_1\mathbf{e}_i^{(k)} + \mathbf{b}_1) + \mathbf{b}_2) \tag{14}$$

$$\alpha_i = \frac{\exp(g_i)}{\sum_{j\in\text{edges}} \exp(g_j)} \tag{15}$$

$$\mathbf{H}_{\text{edge}}^{(k)} = \sum_{i\in\text{edges}} \alpha_i \mathbf{e}_i^{(k)} \in \mathbb{R}^{d'} \tag{16}$$

where:$\mathbf{W}_1 \in \mathbb{R}^{128\times d'}$, $\mathbf{b}_1 \in \mathbb{R}^{128}$: First projection layer. $\mathbf{W}_2 \in \mathbb{R}^{1\times 128}$, $\mathbf{b}_2 \in \mathbb{R}$: Second projection producing scalar attention logit. $\alpha_i \in [0, 1]$: Normalized attention weight for edge $i$ (sums to 1

across all edges in document).$\mathbf{H}_{\text{edge}}^{(k)} \in \mathbb{R}^{d'}$: Aggregated document representation from layer $k$ This mechanism learns which edges are most discriminative for fake news detection. Edges with higher $\alpha_i$ contribute more to the document representation. Our analysis (Section 4.2) examines which relationship types receive high attention.

#### 2.5.2 Multi-Layer Fusion

Different EGNN layers capture relationships at different abstraction levels. We fuse representations from all three layers using learned attention weights:

$$\alpha_1, \alpha_2, \alpha_3 = F_{\text{att}}(\mathbf{H}_{\text{edge}}^{(1)}, \mathbf{H}_{\text{edge}}^{(2)}, \mathbf{H}_{\text{edge}}^{(3)}) \quad (17)$$

$$\mathbf{H}_{\text{fused}} = \alpha_1 \mathbf{H}_{\text{edge}}^{(1)} + \alpha_2 \mathbf{H}_{\text{edge}}^{(2)} + \alpha_3 \mathbf{H}_{\text{edge}}^{(3)} \quad (18)$$

where $F_{\text{att}}$ is implemented as:

$$F_{\text{att}}(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) = \text{softmax}(\mathbf{W}_{\text{fuse}}[\mathbf{h}_1; \mathbf{h}_2; \mathbf{h}_3] + \mathbf{b}_{\text{fuse}}) \quad (19)$$

with $\mathbf{W}_{\text{fuse}} \in \mathbb{R}^{3 \times 3d'}$, $\mathbf{b}_{\text{fuse}} \in \mathbb{R}^3$, producing normalized weights $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$ that sum to 1.The fused representation $\mathbf{H}_{\text{fused}} \in \mathbb{R}^{d'}$ combines information from all abstraction levels with learned importance weights.

Table 5 shows that replacing learned fusion with simple averaging ("No Attention Fusion") produces mixed results—improving performance on PolitiFact but degrading on GossipCop.

### 2.6 Classification Module

The final classification layer transforms the fused document representation into class probabilities:

$$\mathbf{o} = \text{softmax}(\mathbf{W}_f \mathbf{H}_{\text{fused}} + \mathbf{b}_f) \quad (20)$$

where $\mathbf{W}_f \in \mathbb{R}^{C \times d'}$, $\mathbf{b}_f \in \mathbb{R}^C$, and $C = 2$ (real vs. fake news).

## 3 Experimental Setup

This section presents the datasets used and the experimental setup.

### 3.1 Datasets

We evaluate our approach on two real-world benchmark datasets. Table 1 summarizes their key statistics.

| Dataset | PolitiFact | GossipCop |
|---|---|---|
| # News Articles | 450 | 7,916 |
| # Real News | 225 | 3,958 |
| # Fake News | 225 | 3,958 |

Table 1: Statistics of datasets

### 3.2 Adversarial Testing Framework

To assess the robustness of our model against sophisticated adversarial attacks, we adopt the style-based attack framework proposed by Wu et al. (2024a). The attack employs large language models to rewrite news articles according to the following prompt template:

**Rewrite the following article using the style of [publisher name]:** *[news article]*

Further, two types of transformations were applied.

- Disguising Fake News: Fake news articles were rewritten to emulate the style of reputable publishers (e.g., "CNN" and "The New York Times"), thereby increasing their perceived credibility.
- Disguising Real News: Real news articles were rewritten to mimic the style of less credible publishers (e.g., "National Enquirer" and "The Sun"), thereby making them appear less trustworthy.

These transformations yielded four distinct adversarial test sets, labeled A through D as shown in Table 2.

| Publisher name | CNN | The New York Times |
|---|---|---|
| National Enquirer | A | B |
| The Sun | C | D |

Table 2: Configuration of the four style-based adversarial test sets

### 3.3 Implementation Details

The model was trained using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$ and a batch size of 4 over 5 epochs. Dropout rate is $0.1$ . The evaluation involved conducting $10$ independent runs with different random seeds to ensure robust results. Model performance was assessed using accuracy and F1-score metrics.

| Method | PolitiFact | | | | GossipCop | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | A | B | C | D |
| dEFENd$_{vc}$ (Shu et al., 2019a) | 70.44 | 69.77 | 73.67 | 72.98 | 66.40 | 66.55 | 68.93 | 69.07 |
| SAFEv (Zhou et al., 2020) | 71.11 | 70.80 | 75.55 | 75.24 | 67.71 | 67.05 | 68.31 | 67.65 |
| SentGCN (Vaibhav et al., 2019a) | 66.95 | 62.50 | 69.54 | 65.08 | 63.70 | 63.07 | 63.61 | 63.01 |
| DualEmo (Zhang et al., 2021) | 72.42 | 71.23 | 77.07 | 75.80 | 69.47 | 68.50 | 71.69 | 70.71 |
| BERT (Devlin et al., 2019) | 72.31 | 71.37 | 77.23 | 76.24 | 68.98 | 68.17 | 71.95 | 71.11 |
| **BERT-WR-EGNN*** | **82.41** | **77.73** | **83.30** | **78.60** | **71.65** | **71.78** | **73.29** | **73.42** |
| RoBERTa (Liu et al., 2019) | 76.17 | 74.95 | 78.28 | 77.05 | 71.00 | 70.47 | 72.56 | 72.02 |
| **RoBERTa-WR-EGNN*** | **77.42** | **76.13** | **85.18** | **83.86** | **73.59** | **73.11** | **74.61** | **74.12** |
| SheepDog (Wu et al., 2024b) | 80.99 | 79.89 | 82.36 | 81.24 | 74.45 | 74.38 | 75.95 | 75.88 |
| **SheepDog-WR-EGNN*** | **82.66** | **82.65** | **83.44** | **83.54** | **75.14** | **75.13** | **76.88** | **76.87** |

Table 3: WR-EGNN significantly outperforms competitive baselines across four adversarial test settings under LLM-empowered style attacks in terms of F1 score (%). An asterisk ("*") indicates statistical significance at $p < 0.05$ based on a pairwise t-test.

| Method | PolitiFact | | GossipCop | |
|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 |
| dEFEND\ | 82.67 | 82.59 | 70.85 | 70.74 |
| SAFE\w | 79.89 | 79.85 | 70.71 | 70.61 |
| SentGCN | 81.11 | 80.77 | 69.38 | 69.29 |
| DualEmo | 87.78 | 87.76 | 75.51 | 75.36 |
| BERT | 85.22 | 84.99 | 74.60 | 74.50 |
| **B-WR*** | **87.56** | **87.43** | **74.64** | **74.61** |
| RoBERTa | 88.00 | 87.40 | 74.14 | 74.05 |
| **R-WR*** | **90.33** | **90.31** | **74.77** | **74.74** |
| SheepDog | 88.44 | 88.39 | 75.77 | 75.75 |
| **S-WR*** | **91.44** | **91.42** | **77.01** | **77.00** |

Table 4: Comparison of different methods across two datasets. B-WR denotes BERT-WR-EGNN, R-WR denotes RoBERTa-WR-EGNN, and S-WR denotes SheepDog-WR-EGNN. An asterisk (*) indicates statistical significance at $p < 0.05$ using a pairwise t-test against the corresponding transformer-only baseline.

## 4 Results

Table 4 presents a comparison of different methods on two widely used benchmark datasets, PolitiFact and GossipCop. WR-EGNN significantly outperforms three transformer-only models as well as four baseline fake news detection systems.

Table 3 shows that the WR-EGNN model exhibits superior robustness across all four adversarial scenarios, consistently outperforming baseline methods. The performance degradation under adversarial attacks is minimal for our approach, indicating that the inter-word relationships captured by our graph neural network constitute stable structural features that are less susceptible to stylistic manipulations. This robustness is particularly important in the current landscape, where Large Language Models (LLMs) can be leveraged to generate sophisticated fake news that closely mimics legitimate reporting styles.

### 4.1 Ablation Study

To assess the contribution of each WR-EGNN component and the effectiveness of edge embeddings in capturing meaningful relational patterns, we conducted ablation studies. Key components were systematically removed or simplified, and performance was evaluated on both clean and adversarial test sets. We also compared alternative graph construction strategies to validate our choice of dependency-based graphs.

#### 4.1.1 Ablation Variants

We conducted evaluations on three variants of our model:

- Single Edge Layer: Employs only one edge embedding layer instead of three, limiting the model's capacity to capture multi-level relational patterns.
- No Attention Fusion: Replaces the attention-based fusion across edge layers with simple averaging to evaluate the impact of learned fusion weights.
- No Edge Attention: Eliminates the attention mechanism within edge embeddings, relying

solely on concatenated node features without learned relational weighting.

#### 4.1.2 Ablation Results and Analysis

Table 5 presents the results of the three component variants on the PolitiFact and GossipCop datasets. F1-scores are reported for both the clean test data (Original) and adversarial scenario A, as described in Table 2.

The ablation study provides important domain-specific insights into the learning of edge embeddings. Notably, several variants—such as No Attention Fusion and No Edge Attention—match or even exceed full model performance on PolitiFact, but consistently underperform on GossipCop.

The Single Edge Layer variant has minimal impact on PolitiFact (0.23% on clean, 0.81% on adversarial data) but results in consistent drops on GossipCop (0.73% clean, 0.70% adversarial), highlighting the value of multi-layer edge refinement for capturing domain-dependent relational patterns.

The No Attention Fusion variant improves performance on PolitiFact clean data (+1.23%) but suffers on adversarial scenarios (0.24% on scenario A), and consistently decreases performance on GossipCop (0.78% clean, 1.01% adversarial). Similarly, the No Edge Attention variant yields gains on PolitiFact in both clean (+1.01%) and adversarial (+2.07%) settings but shows drops on GossipCop (0.72% clean, 0.92% adversarial).

#### 4.1.3 Graph Structure Comparison: Dependency vs. Co-occurrence

To justify our choice of dependency-based graphs over simpler co-occurrence graphs, we conducted experiments comparing syntactic dependency graphs with window-based co-occurrence graphs. Table 6 provides a comprehensive comparison across both datasets and all adversarial scenarios.

### 4.2 Interpretability Analysis

The key advantage of WR-EGNN over sequence-based models is its interpretability via explicit edge embeddings. We analyzed attention patterns on both the original and adversarial (A) test sets to examine how these embeddings capture structural relationships that resist style-based attacks. $\text{EE}_i$ repersents edge attention in $\mathbf{H}^{(i)}_{\text{edge}}$ WR-EGNN layer.

#### 4.2.1 Attention Pattern Persistence Under Adversarial Attack

We compared attention patterns between the original and adversarially rewritten versions of the same content. Table 20 in Appendix C shows that key structural patterns remain consistent despite surface-level manipulations.

The critical finding is that **data relationship patterns** ("FY→2007") persist across both versions with similar attention scores (0.398→0.424), while adversarial rewriting introduces new surface-level patterns such as subword fragmentation ("RE→OCK" from "SHOCKING"). This explains why edge embeddings provide adversarial robustness: they capture structural relationships that LLM-based rewriting cannot easily manipulate.

#### 4.2.2 Comparative Case Study

We present a detailed comparison for a PolitiFact dataset sample:

- **Original**: "Budget History Tables The Education Department Budget History Table shows President's budget requests and enacted appropriations..."
- **Adversarial**: "SHOCKING BUDGET HISTORY REVEALED: EDUCATION DEPARTMENT SPENDING EXPOSED!..."

Table 7 shows $\text{EE}_3$ layer patterns for both versions.

The "FY→2007" pattern appears in **both** versions (original: 0.398, adversarial: 0.424), demonstrating that edge embeddings capture the underlying data structure—fiscal year references linking to specific years—regardless of whether the text uses professional or sensational framing.

### 4.3 Confidence Calibration and Uncertainty Detection

The model's behavior under adversarial attack reveals sophisticated uncertainty detection. Dramatic confidence drop (99.95%→50.17%) with prediction flip to FAKE shown in Table 8. The near-chance confidence indicates conflicting signals: legitimate data structure (FY→2007, discretionary terminology) vs. sensational framing (SHOCKING, EXPOSED). This uncertainty is *appropriate*—the text contains factual budget data presented with deceptive styling, creating genuine ambiguity that warrants human review.

| Variant | PolitiFact | | GossipCop | |
|---|---|---|---|---|
| | Original | A | Original | A |
| **RoBERTa-WRGNN** | **90.31 ± 0.13** | **77.42 ± 0.28** | **74.74 ± 0.06** | **73.59 ± 0.08** |
| Single Edge Layer | 90.08 ± 0.13 | 78.23 ± 0.34 | 74.01 ± 0.05 | 72.89 ± 0.11 |
| No Attention Fusion | 91.54 ± 0.16 | 77.18 ± 0.53 | 73.96 ± 0.07 | 72.58 ± 0.08 |
| No Edge Attention | 91.32 ± 0.20 | 79.49 ± 0.26 | 74.02 ± 0.04 | 72.67 ± 0.09 |

Table 5: Ablation study on PolitiFact and GossipCop datasets. F1-scores (%) across Original and adversarial test sets.

| | F1-Score (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Original | | A | | B | | C | | D | |
| **Graph Type** | **PF** | **GC** | **PF** | **GC** | **PF** | **GC** | **PF** | **GC** | **PF** | **GC** |
| **Dependency (Ours)** | 90.31 | **74.74** | 77.42 | **73.59** | 76.13 | **73.11** | **85.18** | **74.61** | **83.86** | **74.12** |
| Co-occurrence (Window=5) | **90.52** | 73.98 | **79.13** | 72.18 | **76.83** | 72.03 | 84.18 | 73.75 | 82.37 | 73.61 |
| $\Delta$ (Dep. - Co-occ.) | -0.21 | +0.76 | -1.71 | +1.41 | -0.70 | +1.08 | +1.00 | +0.86 | +1.49 | +0.51 |

Table 6: Comparison of dependency-based graphs vs. window-based co-occurrence graphs. F1-scores (%) across clean and adversarial test sets. Best results per dataset in **bold**. PF represents PolitiFact and GC represents GossipCop.

| Original | Attn | Adversarial |
|---|---|---|
| programs→for | 0.407 | FY→FY$^{\dagger}$ |
| FY→2007$^{\dagger}$ | 0.398 | FY→2007$^{\dagger}$ |
| FY→FY$^{\dagger}$ | 0.396 | our→tax |
| supplemental→funds | 0.395 | rest→rest |
| annual→appropriations | 0.395 | and→2007 |

Table 7: $EE_3$ attention patterns: Original vs Adversarial. Patterns marked with † appear in both versions.

This calibrated uncertainty response demonstrates that WR-EGNN doesn't simply resist attacks through brute-force pattern matching, but rather learns to identify conflicting evidence and express appropriate epistemic humility when structural and stylistic signals diverge.

| Metric | Original | Adversarial |
|---|---|---|
| Prediction | REAL | FAKE |
| Confidence | 99.95% | 50.17% |
| Correct | REAL | REAL |

Table 8: Prediction comparison for Case Study

#### 4.3.1 Robustness Mechanism

Our comparative analysis reveals why WR-EGNN resists adversarial attacks:

1. **Structural patterns persist**: Data relationships ("FY→2007") survive LLM-based rewriting because they encode semantic structure, not surface style.

2. **Uncertainty detection**: Conflicting signals appropriately reduce confidence (99.95%→50.17%), enabling human review.

## 5 Conclusion

We present WR-EGNN, a novel framework for interpretable and robust fake news detection. The core contribution of our work is an Edge Graph Neural Network that treats inter-word relationships as first-class entities and learns rich edge representations capturing semantic, contextual, and structural patterns. By initializing graphs with syntactic dependencies and learning relationship-aware edge embeddings, WR-EGNN captures fine-grained relational signals that are difficult for purely sequence-based models to learn. The model also demonstrates strong robustness to adversarial attacks, maintaining stable performance under LLM-driven stylistic manipulations. Results show that WR-EGNN significantly outperforms three transformer-only models and four baseline fake news detection systems, highlighting the effectiveness of explicit word-relationship modeling over node-level or global document representations.

## Ethics Statement

This work aims to combat misinformation.

## Limitations

Our approach relies on spaCy for dependency parsing to construct graph structures. Parsing errors, particularly in informal text with grammatical irregularities, may introduce noise into the graph representation. This dependency makes our method potentially less robust for social media content or user-generated text where syntactic structure is often non-standard. We evaluate robustness against only one type of adversarial attack—LLM-empowered style manipulation. Other attack vectors such as entity substitution, paraphrasing-based semantic attacks, or adversarial perturbations targeting graph structure remain unexplored.

## References


Ceren Budak. 2019. What happened? the spread of fake news publisher content during the 2016 u.s. presidential election. In *The World Wide Web Conference*, WWW '19, page 139–150, New York, NY, USA. Association for Computing Machinery.

Sonia Cristofaro. 2009. Grammatical categories and relations: Universality vs. language-specificity and construction-specificity. *Language and Linguistics Compass*, 3(1):441–479.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

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.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):81–89.

Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689, Online. Association for Computational Linguistics.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Gihwan Kim and Youngjoong Ko. 2021. Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3276–3280, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2022. Fang: leveraging social context for fake news detection using graph representation. *Commun. ACM*, 65(4):124–132.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Maxime Prieur, Souhir Gahbiche, Guillaume Gadek, Sylvain Gatepaille, Kilian Vasnier, and Valerian Justine. 2023. K-pop and fake facts: from texts to smart alerting for maritime security. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 510–517, Toronto, Canada. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen,

Denmark. Association for Computational Linguistics.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556, Dublin, Ireland. Association for Computational Linguistics.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Vaibhav Vaibhav, Raghuram Mandyam Annasamy, and Eduard Hovy. 2019a. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019b. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong. Association for Computational Linguistics.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024a. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024b. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

Jiaying Wu and Bryan Hooi. 2023. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2582–2593, New York, NY, USA. Association for Computing Machinery.

Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-align: Prompt-based social alignment for few-shot fake news detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 2726–2736, New York, NY, USA. Association for Computing Machinery.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

# A Related Work

## A.1 Fake News Detection

The evolution of fake news detection approaches reflects the increasing sophistication of both deceptive content and detection methodologies. This section reviews key developments in the field, contextualizing our edge-enhanced neural network approach within the broader research landscape.

### A.1.1 Linguistic and Structural Approaches

Early fake news detection methodologies primarily focused on analyzing linguistic features, particularly examining writing style anomalies and narrative consistency (Potthast et al., 2018). These approaches leveraged stylometric analysis to identify patterns characteristic of deceptive content, such as lexical diversity, syntactic complexity, and discourse markers. While effective for certain types of fabricated content, these methods often struggled with sophisticated misinformation that mimicked legitimate journalistic style.

Recent research has recognized that the internal coherence of news articles—specifically the strength and nature of inter-sentence relationships—provides crucial signals for veracity assessment. True news articles typically exhibit strong thematic continuity and logical progression between sentences, whereas fabricated content often contains subtle disconnects in narrative flow (Karimi and Tang, 2019). This insight has spurred the development of models that explicitly represent and analyze document structure.

Vaibhav et al. (Vaibhav et al., 2019b) pioneered a relational graph approach that represents documents as networks with sentences as nodes and weighted edges capturing semantic similarities between sentences. Their model learns document

representations that encode both content and structural information, enabling more robust detection of coherence anomalies characteristic of fake news. Building on this foundation, Kim and Ko (Kim and Ko, 2021) developed a graph-based framework that highlights sentence relationships through summary-based document content representation, effectively capturing high-level document structure alongside detailed content analysis.

Addressing the dynamic nature of information presentation, Ghanem et al. (Ghanem et al., 2021) introduced FakeFlow, an innovative model that focuses specifically on information flow patterns within texts. This approach recognizes that deceptive content often exhibits distinctive patterns in how information unfolds throughout an article, providing additional discriminative features beyond static linguistic markers.

#### A.1.2 Multimodal and Contextual Approaches

While our research focuses primarily on textual content analysis, it is important to acknowledge the parallel development of multimodal approaches that integrate various information sources. Beyond analyzing lexical features (Potthast et al., 2018) and sentiment patterns (Rashkin et al., 2017) intrinsic to news articles, researchers have expanded detection capabilities by incorporating auxiliary information sources.

These complementary approaches leverage user engagement data such as reader comments (Shu et al., 2019b), environmental context of news propagation (Sheng et al., 2022), and knowledge bases that enable fact verification (Dun et al., 2021; Cui et al., 2020). Temporal analysis of user behavior patterns (Ruchansky et al., 2017) and social network graph analysis (Nguyen et al., 2022; Wu and Hooi, 2023; Wu et al., 2023) have further enhanced veracity prediction accuracy by situating content within its broader dissemination context.

While this work focuses on textual analysis, the proposed edge-centric graph neural network framework can potentially be extended to incorporate multimodal information or other target specific real-world challenges in future work.

## B Tokenization Alignment

We tokenize the text using the transformer's native tokenizer and apply the alignment algorithm described in Algorithm 1.

**Algorithm 1** Token Alignment Algorithm

**Require:** SpaCy tokens $T_{spacy} = \{t_1, t_2, ..., t_n\}$
**Require:** Subword tokens $T_{sub} = \{s_1, s_2, ..., s_m\}$
**Ensure:** Alignment mapping $A : T_{spacy} \rightarrow \mathcal{P}(T_{sub})$

```
Initialize alignment map A ← ∅
sub_idx ← 1
for each t_i ∈ T_spacy do
    token_text ← text(t_i)
    subword_span ← []
    while sub_idx ≤ m and incomplete_match(token_text) do
        subword_span.append(s_sub_idx)
        sub_idx ← sub_idx + 1
    end while
    A[t_i] ← subword_span
end for
return A
```

## C Detailed Case Study: Original vs Adversarial Comparison

We present comprehensive attention analysis for two samples, comparing patterns between original and adversarially-rewritten versions. Our analysis spans all three layers of the WR-EGNN architecture ($EE_1$, $EE_2$, $EE_3$), revealing how different abstraction levels respond to adversarial style manipulations.

### C.1 Case Study 1

**Original Text:** "WASHINGTON – The Republican National Committee announced a new web video today on President Obama's health care taxes. For months, President Obama has tried to convince the American people that his government-run health care experiment will be all gain and no pain..."

**Adversarial Text:** "SHOCKING REVELATION: Obama's Health Care Taxes Will Crush America! President Obama has been peddling his health care plan as a miracle cure for all our ailments, but the truth is finally coming out. Brace yourselves, folks, because the Republicans are blowing the lid off this scandal!..."

#### C.1.1 $EE_1$ Layer: Surface-Level Patterns

Table 10 presents the top-15 attention patterns from the EE layer for both versions. This layer captures immediate token-level relationships and surface features.

Table 9: Prediction comparison for Case Study 1.

| Metric | Original | Adversarial |
|---|---|---|
| Prediction | Real | Real |
| Confidence | 99.96% | 94.90% |
| Correct | Real | |

**$EE_1$ Layer Observation:** The adversarial version introduces prominent subword fragmentation patterns from sensational language (RE→OCK, RE→SH from "SHOCKING"), which dominate the top attention patterns. The original text shows more uniform self-attention on domain-relevant terms (taxes, insurance, businesses), while adversarial rewriting creates artificial attention spikes on stylistic artifacts. The mean attention increases from 0.568 to 0.592 under attack, indicating surface-level disruption.

#### C.1.2 $EE_2$ Layer: Intermediate Contextual Features

Table 11 presents the $EE_2$ layer patterns, which capture intermediate-level contextual relationships.

**$EE_2$ Layer Observation:** At this intermediate layer, we observe a critical transition. The original text exhibits heavy subword tokenization artifacts (Voice, ron, TA), indicating the model's internal processing of complex terms. The adversarial version shows emerging semantic coherence with patterns like "care→care" and "Committee→Committee" (rank 13, 0.459), suggesting the model begins recovering semantic meaning despite surface manipulation. Interestingly, the adversarial version introduces evaluative patterns ("wrong→wrong"), potentially reflecting the sensational framing. The attention distribution becomes more focused (std: 0.017→0.014), indicating the model's attempt to filter noise.

#### C.1.3 $EE_3$ Layer: Deep Semantic Relationships

Table 12 presents the $EE_3$ layer patterns, capturing deep semantic and structural relationships.

**$EE_3$ Layer Observation:** The deep semantic layer reveals crucial structural recovery. Entity self-attention patterns emerge prominently: "Committee→Committee" (rank 1, 0.433), "Republican→Republican" (rank 8, 0.420), and domain-specific terms "tax→tax" (rank 7, 0.421), "bill→bill" (rank 12, 0.417). These patterns demonstrate that despite adversarial rewriting, core semantic entities remain identifiable. The punctuation patterns (!→!, ?→?) reflect the sensational style but appear alongside substantive content patterns, suggesting the model successfully disentangles style from substance at this depth.

#### C.1.4 Combined Layer: Fused Representation

Table 13 presents the Combined layer patterns from attention-based fusion of all layers.

**Combined Layer Key Finding:** The fused representation reveals the model's most discriminative patterns. Most notably, rank 5 shows "charge→tax" (0.485), a clear semantic relationship capturing the taxation theme that wasn't explicit in surface text. This demonstrates learned semantic inference beyond lexical matching. Additional substantive patterns include "consumer→And" (rank 10), "taxes→taxes" (rank 11), "insurance→They" (rank 12), and "dollars→They" (rank 13), indicating successful extraction of policy-relevant relationships. The model achieves semantic coherence (std: 0.009) while maintaining 94.90% confidence despite 5.06% confidence drop from attack.

### C.2 Case Study 2

**Original Text:** "Budget History Tables The Education Department Budget History Table shows President's budget requests and enacted appropriations for major Education Department programs. This table breaks out Department budget totals by discretionary and mandatory spending..."

**Adversarial Text:** "SHOCKING BUDGET HISTORY REVEALED: EDUCATION DEPARTMENT SPENDING EXPOSED! In a jaw-dropping twist, the Education Department's budget history has been laid bare for all to see. Prepare yourselves for a wild ride as we dive into the mind-boggling figures..."

#### C.2.1 $EE_1$ Layer Analysis

Table 14 presents the EE layer patterns for the budget dataset case study.

**$EE_1$ Layer Observation:** Rank 10 shows "Dept→spending" (0.634), indicating early semantic association formation. Critically, "Department→Department" persists (rank 13, 0.633), demonstrating entity recognition resilience. The "SHOCKING"-derived patterns (OCK→OCK, OCK→EXP) again dominate high attention, but substantive patterns coexist, unlike pure noise.

Table 10: Case Study 1: EE layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | it→it | 0.618 | the→the | 0.644 |
| 2 | taxes→taxes | 0.618 | But→’t | 0.636 |
| 3 | need→need | 0.618 | about→the | 0.634 |
| 4 | costs→costs | 0.618 | **RE→OCK** | **0.631** |
| 5 | consumer→consumer | 0.618 | But→But | 0.630 |
| 6 | businesses→businesses | 0.618 | about→very | 0.628 |
| 7 | tax→tax | 0.618 | **RE→SH** | **0.628** |
| 8 | insurance→insurance | 0.618 | to→to | 0.628 |
| 9 | to→to | 0.618 | **RE→RE** | **0.628** |
| 10 | supplies→supplies | 0.618 | the→the | 0.627 |
| 11 | to→IF | 0.613 | experiment→experiment | 0.627 |
| 12 | EVEN→EVEN | 0.612 | **OCK→OCK** | **0.627** |
| 13 | A→IF | 0.612 | ’t→’t | 0.627 |
| 14 | ON→ON | 0.610 | these→these | 0.626 |
| 15 | a→a | 0.609 | doesn→’t | 0.624 |

#### C.2.2 $EE_2$ Layer: Contextual Integration

Table 15 presents $EE_2$ layer patterns showing mid-level contextual processing.

**$EE_2$ Layer Observation:** A remarkable pattern emerges at rank 3: “FY→FY” (0.457) appears in the adversarial version, foreshadowing the structural persistence documented in EE2. The original version shows strong procedural language patterns (“annual→appropriations”, rank 2; “enacted→appropriations”, rank 14), reflecting bureaucratic text structure. The adversarial version introduces reader-directed patterns (“yourselves→yourselves”, rank 1; “heads→Who”, rank 6), characteristic of sensational writing, yet fiscal year notation survives.

#### C.2.3 $EE_3$ Layer: Structural Pattern Persistence

Table 16 presents the critical $EE_3$ layer patterns showing data relationship persistence.

**$EE_3$ Layer Critical Finding:** This layer provides the strongest evidence of adversarial robustness. The data structure pattern “FY→2007” appears at rank 2 in *both* versions with remarkably stable attention (0.398→0.424, only +6.5% change). Similarly, “FY→FY” moves from rank 3 to rank 1 (0.396→0.436, +10.1%), indicating *strengthened* attention on structural elements despite sensational framing. The persistence of “tax→tax” (rank 8, 0.414) further demonstrates content-based pattern extraction. These findings reveal that $EE_3$ captures domain-invariant structural relationships—fiscal year references, budgetary terminology—that survive stylistic manipulation because they encode factual data relationships rather than rhetorical style.

#### C.2.4 Combined Layer: Final Decision Features

Table 17 presents the Combined layer patterns integrating all abstraction levels.

**Combined Layer Critical Finding:** Rank 4 shows “FY→FY” (0.482) and rank 9 shows “discretionary→discretionary” persisting across versions (0.469→0.478), demonstrating the final representation successfully preserves structural patterns. However, the adversarial version shows conflicting signals: structural data patterns (FY→FY, discr.→discr.) coexist with sensational style markers (ATION→!, DE→!, SH→!). This explains the confidence drop to 50.17%—the model detects legitimate data structure but conflicting stylistic features, appropriately expressing uncertainty rather than committing to a confident but potentially incorrect prediction.

### C.3 Pattern Persistence Summary

Table 18 summarizes all patterns that persist between original and adversarial versions across both case studies.

### C.4 Patterns Introduced by Adversarial Rewriting

Table 19 lists attention patterns that appear only in the adversarial versions, primarily subword frag-

Table 11: Case Study 1: $EE_2$ layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | Voice→Voice | 0.498 | wrong→wrong | 0.469 |
| 2 | Voice→Voice | 0.495 | .→. | 0.468 |
| 3 | ES→TA | 0.495 | wrong→wrong | 0.467 |
| 4 | ron→TA | 0.495 | care→care | 0.465 |
| 5 | Voice→Voice | 0.494 | rest→rest | 0.463 |
| 6 | TA→NEW | 0.493 | to→to | 0.463 |
| 7 | YOU→Voice | 0.492 | tax→. | 0.462 |
| 8 | TA→TA | 0.492 | wrong→for | 0.462 |
| 9 | ron→A | 0.492 | to→the | 0.462 |
| 10 | 1→1 | 0.491 | the→the | 0.461 |
| 11 | ron→ron | 0.491 | bar→bar | 0.461 |
| 12 | ron→IF | 0.490 | .→. | 0.460 |
| 13 | ron→ron | 0.490 | **Committee→Committee** | **0.459** |
| 14 | ron→ron | 0.490 | ,→folks | 0.459 |
| 15 | Voice→: | 0.490 | opoulos→. | 0.458 |

mentation from sensational language.

Table 12: Case Study 1: $EE_3$ layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | â→â | 0.440 | **Committee→Committee** | **0.433** |
| 2 | X→Ch | 0.431 | !→! | 0.424 |
| 3 | X→y | 0.430 | !→! | 0.424 |
| 4 | ON→Ch | 0.429 | Committee→by | 0.424 |
| 5 | ON→y | 0.429 | ?→? | 0.423 |
| 6 | ES→Ch | 0.428 | .→. | 0.423 |
| 7 | Ch→Ch | 0.422 | **tax→tax** | **0.421** |
| 8 | RE→Ch | 0.426 | **Republican→Republican** | **0.420** |
| 9 | RE→y | 0.424 | ?→if | 0.420 |
| 10 | Ch→Ch | 0.422 | Yes→. | 0.420 |
| 11 | â→â | 0.421 | !→! | 0.419 |
| 12 | (space) | 0.420 | **bill→bill** | **0.417** |
| 13 | (special) | 0.419 | tax→. | 0.417 |
| 14 | (special) | 0.419 | !→! | 0.417 |
| 15 | (special) | 0.418 | ?→you | 0.417 |

Table 13: Case Study 1: Combined layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | ron→TA | 0.493 | .→. | 0.493 |
| 2 | YOU→: | 0.490 | tax→. | 0.489 |
| 3 | YOU→ron | 0.490 | **tax→tax** | **0.489** |
| 4 | YOU→â | 0.489 | **Committee→Committee** | **0.486** |
| 5 | ron→A | 0.488 | **charge→tax** | **0.485** |
| 6 | ron→ron | 0.487 | reeling→! | 0.485 |
| 7 | â→: | 0.487 | no→no | 0.485 |
| 8 | A→A | 0.486 | rest→rest | 0.483 |
| 9 | ES→TA | 0.486 | .→Obama | 0.483 |
| 10 | :→TA | 0.486 | **consumer→And** | **0.483** |
| 11 | supplies→TA | 0.485 | **taxes→taxes** | **0.482** |
| 12 | Voice→Voice | 0.485 | **insurance→They** | **0.482** |
| 13 | EVEN→TA | 0.483 | **dollars→They** | **0.482** |
| 14 | 1→1 | 0.484 | Oh→'ve | 0.481 |
| 15 | Voice→â | 0.483 | .→. | 0.481 |

Table 14: Case Study 2: $EE_1$ layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | B→History | 0.621 | all→that | 0.654 |
| 2 | laws→laws | 0.616 | that→that | 0.648 |
| 3 | Calendar→Calendar | 0.616 | **OCK→OCK** | **0.643** |
| 4 | process→process | 0.616 | just→just | 0.640 |
| 5 | programs→programs | 0.616 | were→just | 0.638 |
| 6 | The→The | 0.612 | ’s→’s | 0.636 |
| 7 | usually→usually | 0.611 | :→table | 0.635 |
| 8 | the→the | 0.611 | are→What | 0.635 |
| 9 | major→major | 0.611 | using→the | 0.635 |
| 10 | Department→Department | 0.603 | **Dept→spending** | **0.634** |
| 11 | Rita→Katrina | 0.603 | **OCK→EXP** | **0.634** |
| 12 | The→Table | 0.602 | all→all | 0.633 |
| 13 | .→. | 0.602 | **Department→Department** | **0.633** |
| 14 | more→more | 0.602 | just→really | 0.632 |
| 15 | table→table | 0.601 | because→the | 0.632 |

Table 15: Case Study 2: $EE_2$ layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | Spending→Spending | 0.459 | yourselves→yourselves | 0.461 |
| 2 | annual→appropriations | 0.456 | rest→rest | 0.460 |
| 3 | for→Spending | 0.452 | **FY→FY** | **0.457** |
| 4 | programs→programs | 0.450 | decided→. | 0.455 |
| 5 | and→discretionary | 0.449 | tuned→tuned | 0.455 |
| 6 | contrast→contrast | 0.448 | heads→Who | 0.454 |
| 7 | spending→spending | 0.448 | detail→detail | 0.454 |
| 8 | for→spending | 0.448 | we→we | 0.454 |
| 9 | by→affected | 0.448 | ?→? | 0.453 |
| 10 | is→is | 0.447 | ace→yourselves | 0.452 |
| 11 | appropriations→appropriations | 0.445 | .→yourselves | 0.452 |
| 12 | ,→, | 0.444 | ,→, | 0.452 |
| 13 | appropriations→appropriations | 0.444 | there→there | 0.451 |
| 14 | enacted→appropriations | 0.443 | .→. | 0.451 |
| 15 | by→totals | 0.443 | !→! | 0.451 |

Table 16: Case Study 2: $EE_3$ layer top-15 attention patterns. Patterns marked with † appear in both versions.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | programs→for | 0.407 | **FY→FY**† | **0.436** |
| 2 | **FY→2007**† | **0.398** | **FY→2007**† | **0.424** |
| 3 | **FY→FY**† | **0.396** | our→tax | 0.418 |
| 4 | supplemental→funds | 0.395 | rest→rest | 0.417 |
| 5 | annual→appropriations | 0.395 | and→2007 | 0.415 |
| 6 | for→for | 0.394 | .→. | 0.422 |
| 7 | Spending→by | 0.399 | ?→? | 0.421 |
| 8 | the→the | 0.396 | **tax→tax**† | **0.414** |
| 9 | FY→appropriations | 0.395 | and→and | 0.417 |
| 10 | spending→spending | 0.392 | rest→rest | 0.417 |
| 11 | enacted→enacted | 0.391 | !→laid | 0.414 |
| 12 | 2007→2007 | 0.391 | our→our | 0.413 |
| 13 | 2006→2006 | 0.390 | closed→closed | 0.411 |
| 14 | table→breaks | 0.389 | !→! | 0.411 |
| 15 | annual→annual | 0.389 | FY→appropriations | 0.409 |

Table 17: Case Study 2: Combined layer top-15 attention patterns.

| Rank | Original Pattern | Attn | Adversarial Pattern | Attn |
|---|---|---|---|---|
| 1 | annual→appropriations | 0.474 | rest→rest | 0.484 |
| 2 | the→appropriations | 0.472 | ATION→! | 0.482 |
| 3 | contrast→contrast | 0.472 | DE→! | 0.482 |
| 4 | for→spending | 0.471 | **FY→FY** | **0.482** |
| 5 | and→discretionary | 0.471 | .→. | 0.481 |
| 6 | the→the | 0.470 | detail→. | 0.481 |
| 7 | for→Spending | 0.470 | heads→Who | 0.480 |
| 8 | Spending→Spending | 0.469 | decided→. | 0.480 |
| 9 | **discr.→discr.** | **0.469** | **discr.→discr.** | **0.478** |
| 10 | appropriations→appr. | 0.469 | SH→! | 0.479 |
| 11 | programs→programs | 0.469 | !→! | 0.478 |
| 12 | ,→, | 0.468 | we→we | 0.478 |
| 13 | for→for | 0.467 | totals→totals | 0.478 |
| 14 | is→is | 0.467 | OCK→! | 0.478 |
| 15 | spending→spending | 0.466 | kicker→kicker | 0.478 |

Table 18: Summary of attention patterns that persist under adversarial rewriting.

| **Pattern** | **Orig. Attn** | **Adv. Attn** | **$\Delta$** | **Layer** |
|---|---|---|---|---|
| *Case Study 2: Budget Data* | | | | |
| FY→2007 | 0.398 | 0.424 | +6.5% | $EE_2$ |
| FY→FY | 0.396 | 0.436 | +10.1% | $EE_2$ |
| discretionary→discr. | 0.469 | 0.478 | +1.9% | Combined |
| tax→tax | 0.421 | 0.414 | −1.7% | $EE_2$ |
| Department→Department | 0.603 | 0.633 | +5.0% | EE |
| spending→spending | 0.392 | 0.418 | +6.6% | $EE_2$ |
| *Case Study 1* | | | | |
| Committee→Committee | 0.433 (adv) | — | — | $EE_2$ |
| tax→tax | 0.421 (adv) | — | — | $EE_2$ |

Table 19: Attention patterns introduced by adversarial "SHOCKING" rewriting.

| **Pattern** | **Attention** | **Layer** | **Source** |
|---|---|---|---|
| RE→OCK | 0.631 | EE | "SHOCKING" subword |
| RE→SH | 0.628 | EE | "SHOCKING" subword |
| RE→RE | 0.628 | EE | "SHOCKING" subword |
| OCK→OCK | 0.643 | EE | "SHOCKING" self-attn |
| OCK→EXP | 0.634 | EE | Sensational compound |
| SH→! | 0.479 | Combined | Sensational punctuation |
| ATION→! | 0.482 | Combined | "REVELATION" + punct |
| DE→! | 0.482 | Combined | "REVEALED" + punct |
| reeling→! | 0.485 | Combined | Emotional language |

| Pattern | Original | Adversarial | Type |
|---|---|---|---|
| FY→2007 | 0.398 | 0.424 | Data |
| FY→FY | 0.396 | 0.436 | Data |
| discretionary→discr. | 0.469 | 0.478 | Domain |
| annual→appropriations | 0.395 | — | Process |

Table 20: Attention patterns remain stable under adversarial rewriting. Key data relationships exhibiting similar attention scores in both the original and adversarial versions.