RESEARCH-ARTICLE

# X-Troll: eXplainable Detection of State-Sponsored Information Operations Agents

**LIN TIAN**, University of Technology Sydney, Sydney, NSW, Australia

**XIUZHEN ZHANG**, RMIT University, Melbourne, VIC, Australia

**MARIA MYUNG HEE KIM**, Defence Science and Technology Group, Canberra, ACT, Australia

**JENNIFER BIGGS**, Defence Science and Technology Group, Canberra, ACT, Australia

**MARIAN ANDREI RIZOIU**, University of Technology Sydney, Sydney, NSW, Australia

**Open Access Support** provided by:

**Defence Science and Technology Group**

**University of Technology Sydney**

**RMIT University**



**Published:** 10 November 2025

**Citation in BibTeX format**

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
*November 10 - 14, 2025*
*Seoul, Republic of Korea*

**Conference Sponsors:**
SIGWEB
SIGIR

# X-Troll: eXplainable Detection of State-Sponsored Information Operations Agents

Lin Tian
Lin.Tian-3@uts.edu.au
University of Technology Sydney
Sydney, Australia

Xiuzhen Zhang
xiuzhen.zhang@rmit.edu.au
RMIT University
Melbourne, Australia

Maria Myung-Hee Kim
myung.kim@defence.gov.au
Defence Science and Technology Group
Adelaide, Australia

Jennifer Biggs
jennifer.biggs@defence.gov.au
Defence Science and Technology Group
Adelaide, Australia

Marian-Andrei Rizoiu
Marian-Andrei.Rizoiu@uts.edu.au
University of Technology Sydney
Sydney, Australia

## Abstract

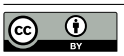
State-sponsored trolls, malicious actors who deploy sophisticated linguistic manipulation in coordinated information campaigns, posing threats to online discourse integrity. While Large Language Models (LLMs) achieve strong performance on general natural language processing (NLP) tasks, they struggle with subtle propaganda detection and operate as "black boxes", providing no interpretable insights into manipulation strategies. This paper introduces **X-Troll**, a novel framework that bridges this gap by integrating explainable adapter-based LLMs with expert-derived linguistic knowledge to detect state-sponsored trolls and provide human-readable explanations for its decisions. X-Troll incorporates appraisal theory and propaganda analysis through specialized LoRA adapters, using dynamic gating to capture campaign-specific discourse patterns in coordinated information operations. Experiments on real-world data demonstrate that our linguistically-informed approach shows strong performance compared with both general LLM baselines and existing troll detection models in accuracy while providing enhanced transparency through expert-grounded explanations that reveal the specific linguistic strategies used by state-sponsored actors. X-Troll source code is available at: https://github.com/ltian678/xtroll_source/.

## CCS Concepts

• **Computing methodologies** → *Natural language processing*; • **Information systems** → *Collaborative and social computing systems and tools.*

## Keywords

Social Media, Information Operation, Troll Detection




*CIKM '25, Seoul, Republic of Korea*

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761028

**ACM Reference Format:**
Lin Tian, Xiuzhen Zhang, Maria Myung-Hee Kim, Jennifer Biggs, and Marian-Andrei Rizoiu. 2025. X-Troll: eXplainable Detection of State-Sponsored Information Operations Agents. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761028

## 1 Introduction

State-sponsored information operation agents—commonly known as troll accounts—have emerged as major threat actors in the digital information ecosystem, systematically manipulating public discourse to achieve geopolitical objectives [22, 48]. Unlike isolated bad actors spreading misinformation, these agents operate as coordinated units within state-directed campaigns, using nuanced linguistic strategies and assuming false personas to infiltrate and influence online communities. Their tactics extend beyond spreading falsehoods to include more subtle forms of manipulation: amplifying divisive content, undermining institutional trust, and steering narratives through strategic emotional appeals. A prominent example is the `Doppelgänger campaign`, a coordinated disinformation effort that mimicked legitimate media sources to spread misleading narratives[1]. Volunteer groups like `@antibot4navalny` have played a crucial role in exposing these operations by manually tracking and documenting troll activities, providing valuable insights into the evolving nature of disinformation. However, the scale of these campaigns exceeds the capacity of manual efforts, highlighting the necessity for automated and scalable troll detection.

Automated disinformation and troll detection have been widely explored in the literature [3, 5, 12, 15, 38, 45]. In particular, neural approaches, such as pre-trained language models and graph neural networks, have been used to analyze the content of social media posts and their propagation patterns, and user engagement patterns for detecting troll behavior. However, most existing models operate as black-boxes, offering little transparency into how decisions are made. This lack of interpretability limits their real-world usability, as end-users often struggle to trust automated predictions.

[1] https://www.disinfo.eu/doppelganger-operation/

Research on machine learning interpretability has evolved from explaining internal decision-making processes to generating human-understandable explanations, particularly in NLP applications. Rationalization, in particular, involves identifying input phrases sufficient to predict the desired outcome [24, 27]. While most studies on rationalization primarily focus on generating phrase-level explanations, the challenge of producing natural language justifications that are easily understandable to humans remains largely unsolved.

Media and communication research provides critical insights into disinformation by exploring its creation, spread, and impact across media platforms. Experts in these fields have identified distinctive linguistic features of disinformation discourse [16, 30], the mechanisms of its dissemination and amplification [32, 53], and the psychological and societal effects on audiences [16]. Such studies have shown that trolls employ propagandistic and targeted communication strategies, often characterized by inflammatory and provocative language, to influence audiences [32]. These human-derived expert insights provide deep contextual understanding that complements computational machine learning research.

Drawing from recent LLM evaluation limitations and linguistic analysis insights, we present two key research questions:

**RQ1:** Why do transformer-based models trained on massive corpora struggle with detecting state-sponsored trolls despite strong performance on related tasks? What linguistic knowledge—specifically from appraisal theory and propaganda analysis—is required to identify coordinated manipulation patterns?

**RQ2:** How can we generate explanations that reveal not just what features triggered a classification, but the underlying manipulation strategies being used, when dealing with adversaries who deliberately obscure their tactics?

To answer these research questions, we propose X-Troll, a framework that integrates linguistic expert knowledge – specifically appraisal analysis and propaganda strategy identification – through Low-Rank Adaptation (LoRA) fine-tuning of LLMs. Unlike general-purpose models that struggle with subtle propaganda techniques, X-Troll analyzes users' social media post timelines through the lens of established linguistic theory to perform classification while producing expert-informed rationales. By combining domain-specific linguistic knowledge with adapter-based fine-tuning and rationale-based explanation generation, X-Troll is trying to mitigate the limitations of general LLMs in propaganda detection while achieving high accuracy and producing human-readable natural language explanations that illuminate the specific linguistic manipulation strategies used by state-sponsored trolls.

## 2 Related Work

State-sponsored troll and disinformation campaign detection has become a key research focus. Recent work primarily uses NLP techniques for Information Campaigns. Kim et al. [19] developed a time-sensitive semantic edit distance (t-SED) metric to analyze user identity and social roles through timestamped text sequences. Their case study of Russian trolls on Twitter classified social roles into left-, right-leaning, and news feed categories. Addawood et al. [3] explored linguistic cues that indicate deception in political trolls' social media posts. They identified key markers of state-sponsored accounts. Im et al. [15] proposed a content-based approach to detect Russian troll accounts on Twitter. Their method leverages user metadata, activity patterns, and linguistic features.

Behavioral modeling approaches have shown promise in capturing troll dynamics. Rizoiu et al. [37] found that socialbots are 2.5 times more influential than humans during political events, while Ram et al. [35] developed birdspotter, an end-to-end pipeline achieving strong bot detection performance. Recent advances focus on early detection and social system reactions. Tian et al. [44] introduced IC-Mamba for predicting engagement patterns within 15-30 minutes of posting, while Kong et al. [21] proposed detecting information operations by analyzing social reactions to identify state-backed agents. Ram et al. [36] identified distinct right-wing patterns in moral language and dichotomous thinking.

Understanding the broader manipulation ecosystem provides crucial context for detection systems. Calderon et al. [9] introduced the Opinion Market Model to evaluate interventions against extremist content spread, while Kong et al. [20] used mixed methods to explain how extreme opinions infiltrate mainstream discussions. Crisis events amplify these dynamics: Bailo et al. [6] found far-right accounts moved from peripheral to central positions during Australian disasters, and Johns et al. [17] showed that some Facebook pages overperformed during COVID-19. Ferrara et al. [13] surveyed the landscape of social bots and their role in manipulation campaigns, while Zannettou et al. [52] traced how disinformation spreads across the web ecosystem. Starbird et al. [41] revealed how alternative media ecosystems participate in information operations. These studies underscore that troll detection cannot be isolated from understanding the broader information context.

Explainability in rumor and fake news detection has gained increasing attention. Early efforts primarily employ attention mechanisms to generate explanations. These methods leverage both text and non-text signals to provide insights into detection model decisions [18, 25, 29, 33, 39, 40]. For example, Shu et al. [39] applied co-attention mechanisms to examine the relationship between content and audience reactions. This helps classify news as real or fake. Similarly, Khoo et al. [18] used multi-head attention to analyze tweet interactions at both token and post levels. However, attention mechanisms are not inherently designed for human interpretability.

Multi-task learning approaches have shown promise in related domains. Yuan and Rizoiu [49], Yuan et al. [51] showed that multi-dataset training significantly outperforms single-task approaches in cross-domain generalization. Their multi-task learning pipeline parallels our multi-adapter architecture, where each adapter specializes in different aspects of manipulative discourse while sharing the base model representation.

Some studies have used user attributes and propagation patterns to explain detection model decisions. Vosoughi et al. [48] identified distinct propagation patterns for fake news. Ni et al. [33] extended this work by modeling propagation with graph neural networks. They used attention mechanisms to highlight key features. Lu and Li [29] and Silva et al. [40] further explored these patterns to distinguish fake news from true news. Both employed attention mechanisms to provide explanations. Liu et al. [25] advanced this work using self-supervised graph learning. Their approach identifies important nodes and generates interpretable subgraphs.

Interpretability remains a challenge in AI and machine learning, particularly for large language models. Recent research has focused

on the plausibility of model-generated rationalizations. Rajani et al. [34] introduced Commonsense Auto-Generated Explanations. Their approach fine-tunes language models on explanation datasets. This automatically generates rationalizations for commonsense QA tasks. Liu et al. [26] extended this work by developing a model with multiple generators and a single predictor. Each generator uses different initializations to produce diverse rationale candidates. This design reduces bias and improves explanation plausibility. While these developments show promise, their effectiveness for troll detection remains unclear.

Despite progress in automated troll detection and broader machine learning interpretability, explainability of troll detection systems remains largely unexplored. This paper addresses this gap by proposing a novel rationale-based approach to troll detection. We aim to provide clear and human-understandable explanations for detection model decisions.

## 3 Preliminaries

In this section, we provide the necessary background on appraisal theory and propaganda analysis. Effective troll detection requires understanding the systematic linguistic strategies that distinguish coordinated manipulation from authentic discourse. We ground X-Troll in established theoretical frameworks from discourse analysis and propaganda studies, enabling both accurate detection and interpretable explanations of manipulative communication patterns.

### 3.1 Appraisal Theory

Appraisal theory [30] provides a systematic framework for analyzing how language users express evaluative stance and emotional positioning in discourse. Recent research has demonstrated that state-sponsored trolls exhibit distinctive appraisal patterns that differ systematically from authentic users [46], making this framework particularly valuable for troll detection.

**Theoretical Framework.** Appraisal analysis examines three interconnected systems of evaluative meaning: Attitude (emotional reactions and judgments), Engagement (how speakers position themselves relative to their propositions), and Graduation (the scaling of evaluative intensity). For troll detection, we focus on three dimensions that capture manipulative discourse strategies:

*Ideational targeting:* Systematic focus on specific entities, topics, or themes designed to direct audience attention toward predetermined narrative frames. State-sponsored trolls consistently target particular political figures, institutions, or ideological concepts rather than engaging in organic topical variation.

*Sentiment polarity:* Strategic deployment of positive, negative, or neutral evaluative language to influence audience perception. Unlike authentic users who express emotional responses, trolls systematically calibrate sentiment to achieve persuasive objectives.

*Persona construction:* Linguistic techniques used to establish false credibility or deliberately provoke emotional responses. This includes strategic deployment of authority markers, community membership signals, and emotional authenticity performance.

**Empirical Application.** We apply appraisal analysis through expert annotation of state-sponsored troll datasets, focusing on Twitter—released accounts from Russian Internet Research Agency operations (detailed in Section 5). Domain experts identified systematic patterns where trolls manipulate evaluative language to maximize influence. These patterns provide supervised signals that enable X-Troll to recognize subtle linguistic manipulation strategies that automated systems typically miss.

### 3.2 Propaganda Technique Analysis

State-sponsored trolls systematically use propaganda techniques to manipulate audience perception and achieve strategic objectives.

**Theoretical Foundation.** We integrate propaganda technique identification to capture these explicit manipulation strategies, building on established taxonomies and modern computational approaches [11, 23]. We focus on three core techniques commonly used in state-sponsored information operations:

*Loaded Language* involves strategic use of emotionally charged terms to provoke reactions rather than facilitate rational discourse, by associating targets with predetermined emotional valences.

*Appeal to Commonality* frames partisan positions as widely accepted beliefs, creating false consensus through linguistic manipulation that exploits social proof heuristics.

*Doubt and Questioning* systematically undermines credible sources and established facts through persistent skepticism, eroding confidence in existing knowledge structures without providing alternative explanations.

**Integration with X-Troll.** We incorporate the DIPROMATS 2023 dataset [32], which provides post-level annotations of these techniques in real-world information operations. These annotations serve as supervised signals enabling X-Troll to recognize established propaganda strategies across user timelines. This propaganda-aware approach provides two advantages: distinguishing coordinated operations from authentic communication, and supplying structured vocabulary for interpretable explanations grounded in established rhetorical analysis.

## 4 Methodology

This section introduces our problem formulation and presents X-Troll, a rationale-based framework for explainable detection of state-sponsored trolls. As shown in Fig. 1, X-Troll integrates three core components: (1) knowledge-fused LLM adapters for multifaceted feature extraction from user timelines, (2) a unified rationale selector for identifying key trolling evidence, and (3) a summary generator for producing human-readable explanations from selected rationales.

### 4.1 Problem Statement

Social media troll detection differs from general misinformation detection. While misinformation detection focuses on post-level falsehood, troll accounts engage in information campaigns with specific interests and targets over time. Troll user timelines exhibit distinct linguistic and behavioral features that differ from typical social media users.

We formalize the problem as follows. Given a set of users $\mathcal{U} = \{u_1, \ldots, u_m\}$ and a set of information campaigns $C$, the task is to detect troll users and identify their associated information campaigns. Each user $u$ has a timeline of posts $T_u = [x_{(u,1)}, \ldots, x_{(u,n)}]$,

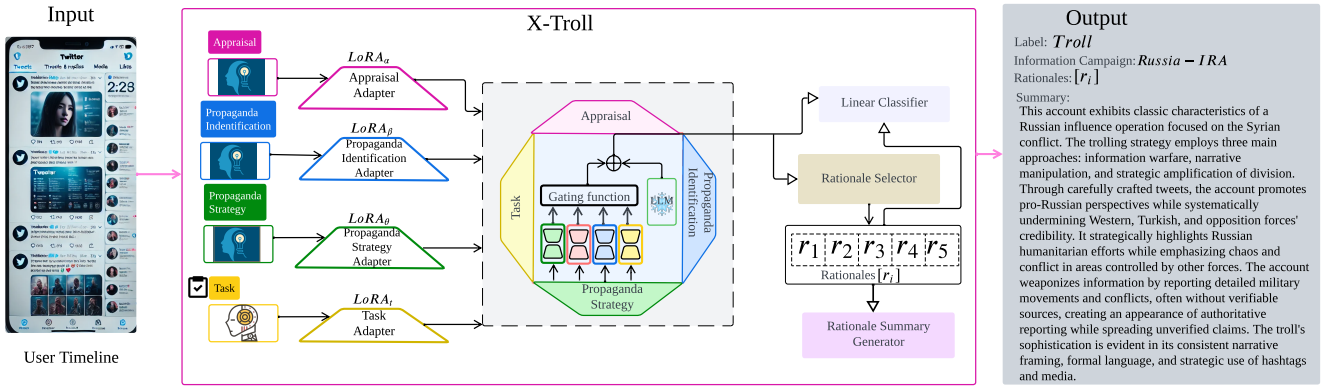



**Figure 1: X-Troll framework for explainable state-sponsored troll detection. Given a user timeline, four LoRA adapters capture distinct aspects of manipulative discourse: Appraisal (evaluative language patterns), Propaganda Identification (binary propaganda detection), Propaganda Strategy (specific manipulation techniques), and Task (troll-specific features). A dynamic gating mechanism adaptively weights adapter contributions, feeding the fused representation to a linear classifier for troll detection and campaign classification. The rationale selector identifies salient tokens across the timeline, which the summary generator transforms into human-readable explanations grounded in linguistic theory. The example shows detection of a Russian-IRA agent with extracted rationales and generated explanation revealing narrative manipulation strategies.**

where $x_{(u,j)}$ represents the $j$-th post by $u$. Users are labeled as *trolls* ($y_u = 1$) or *non-trolls* ($y_u = 0$).

*Troll and Information Campaign Classification.* A classifier $g_\phi$ predicts whether a user is a troll based on their timeline: $y_u = g_\phi(T_u)$. For users identified as trolls, an additional classifier $h_\theta$ predicts their associated information campaign $c_u \in C$: $c_u = h_\theta(T_u)$.

*Rationale Selection and Explanation Generation.* We address explainability through a two-stage interpretability mechanism. First, a rationale selector $f_\gamma$ identifies a sparse subset of $k$ salient tokens across the user's timeline. These tokens most strongly influence the classification decision: $\mathcal{R}_u = f_\gamma(T_u, y_u)$. Second, an explanation generator $s_\omega$ synthesizes a comprehensive, human-readable explanation from the extracted rationales: $S_u = s_\omega(\mathcal{R}_u)$.

## 4.2 Timeline Encoding

To model user behaviour over time, we encode a user's timeline $T_u$ consisting of posts $\{x_{(u,j)}\}_{j=1}^{n_u}$. Each post is first transformed into a contextualized representation using a pretrained language model $f_\beta$: $\mathbf{h}_{(u,j)} = f_\beta(x_{(u,j)}) \in \mathbb{R}^d$. Next, the sequence of post embeddings $[\mathbf{h}_{u,1}, \ldots, \mathbf{h}_{u,n_u}]$ is processed by a Transformer encoder parameterized by $\gamma$: $\mathbf{H}_u = \text{Transformer}_\gamma([\mathbf{h}_{(u,1)}, \ldots, \mathbf{h}_{(u,n_u)}]) \in \mathbb{R}^{n_u \times d}$, where $\mathbf{H}_u = [\tilde{\mathbf{h}}_{(u,1)}, \ldots, \tilde{\mathbf{h}}_{(u,n_u)}]$ represents the enhanced timeline representations. To obtain a single timeline representation $\mathbf{t}_u$, we apply attention pooling to individual posts:

$$\alpha_{u,j} = \frac{\exp(\mathbf{q}^\top \tilde{\mathbf{h}}_{(u,j)})}{\sum_{k=1}^{n_u} \exp(\mathbf{q}^\top \tilde{\mathbf{h}}_{(u,k)})},$$

$$\mathbf{t}_u = \sum_{j=1}^{n_u} \alpha_{(u,j)}, \tilde{\mathbf{h}}_{(u,j)},$$

where $\mathbf{q} \in \mathbb{R}^d$ is a learnable query vector. We experimented with mean pooling, max pooling, and attention-based pooling, finding that attention pooling outperformed other methods by emphasizing the most informative posts for downstream tasks.

## 4.3 Adapter Fusion

X-Troll uses a novel adapter fusion architecture based on Low-Rank Adaptation (LoRA) [14] to incorporate diverse expert knowledge while maintaining computational efficiency. This approach enables parameter-efficient fine-tuning while integrating domain-specific knowledge across multiple dimensions of troll behavior. Each adapter specializes in a distinct aspect of troll behavior and is dynamically integrated via a *gating mechanism* (Fig. 1).

*4.3.1 LoRA Adapter Mechanism.* We leverage LoRA to efficiently fine-tune X-Troll. Given a weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA introduces low-rank matrices $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where $r \ll \min(d, k)$, updating weights as: $W' = W + BA$. This enables task-specific knowledge adaptation while maintaining computational efficiency. Each adapter operates independently while sharing the base model, facilitating multi-domain adaptation. The LoRA update modifies the hidden state representation $h$: $h' = Wh + \Delta Wh = Wh + B(Ah)$. This ensures domain-relevant transformations without full fine-tuning. X-Troll integrates three specialized LoRA-adapters designed to capture strategic, linguistic, and behavioral aspects of troll detection.

*(1) Appraisal Adapter (LoRA$_\alpha$).* It performs fine-grained linguistic analysis based on appraisal theory, trained via token-level sequence labeling on expert-annotated data. It captures ideational targeting (consistent entity focus), sentiment polarity (strategic emotional framing), and persona construction (credibility establishment techniques). The adapter optimizes a custom sequence labeling loss $\mathcal{L}_{\text{appraisal}}$ accounting for hierarchical appraisal features across linguistic spans.

*(2) Propaganda Identification Adapter (LoRA$_\beta$).* It specializes in detecting binary propaganda presence within posts, drawing from the DIPROMATS 2023 propaganda dataset annotations. It applies targeted low-rank updates ($\Delta\mathbf{W}_\beta = \mathbf{B}_\beta\mathbf{A}_\beta$) to model layers most sensitive to propaganda features. The adapter is optimised using binary cross-entropy loss $\mathcal{L}_{\text{prop}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i)+(1-y_i)\log(1-p_i)]$, where $y_i$ represents the ground-truth propaganda label and $p_i$ the model prediction for the $i$-th sample.

*(3) Propaganda Strategy Adapter (LoRA$_\theta$):* This adapter performs fine-grained classification of specific propaganda techniques (e.g., loaded language, appeal to fear, causal oversimplification) based on multi-class strategy annotations. It is optimised using a categorical cross-entropy loss $\mathcal{L}_{\text{strat}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{(i,c)} \log(p_{(i,c)})$, where $C$ represents the number of propaganda strategy classes, and $y_{i,c}$ and $p_{i,c}$ represent the ground truth and prediction for class $c$ of sample $i$, respectively.

*(4) Task Adapter (LoRA$_t$):* This adapter serves as the task-specific component directly optimised for troll detection and identification, capturing patterns distinctive to state-sponsored trolls. It is trained using the troll classification loss:

$$\mathcal{L}_{\text{task}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)].$$

By integrating expert knowledge through these specialised adapters, X-Troll captures patterns of troll behaviour while maintaining computational efficiency. Our multi-adapter approach builds on behavioral pattern analysis findings. Yuan et al. [50] demonstrated that behavioral homophily can reveal user patterns that transcend topical similarity, suggesting that similar behavioral policies may be identifiable across different topics. This supports our design of specialized adapters that capture different aspects of trolling activities.

*4.3.2 Dynamic Gating Mechanism.* To effectively integrate outputs from multiple specialised LoRA adapters, X-Troll uses a dynamic gating mechanism that adaptively learns the optimal weighting of each adapter's contribution. Let $K$ denote the number of adapters; in our case, $K = 4$.

Each adapter produces an output representation $h_k \in \mathbb{R}^d$, where $k \in \{1, 2, \ldots, K\}$. We introduce learnable scalar gating parameters $w_k$ for each adapter, which are transformed using a softmax function to ensure non-negative weights that sum to one: $\alpha_k = \frac{\exp(w_k)}{\sum_{j=1}^{K}\exp(w_j)}$. The combined representation $h_{\text{combined}}$ is computed as a weighted sum of adapter outputs: $h_{\text{combined}} = \sum_{k=1}^{K}\alpha_k h_k$.

This dynamic gating mechanism is trained end-to-end with the rest of the model. During training, the learnable weights $w_k$ are updated to minimise the overall loss function, allowing the model to select the most informative adapters for each specific input. By adjusting these weights, the model can prioritise key adapters over others, effectively capturing the complex and evolving nature of troll behaviour.

## 4.4 Rationale Selector

Most LLMs based on decoder-only architectures, such as GPT-3 [8], have demonstrated strong capabilities in natural language understanding and generation. To leverage these strengths for explainable troll detection, we introduce a unified decoder architecture that simultaneously performs rationale selection and user classification–a new way over conventional two-stage approaches.

Our approach uses a shared decoder that jointly extracts informative rationales from user posts and performs classification based on these rationales. This design tries to mitigate the degeneration issue prevalent in two-phase rationalisation models, where classifiers often overfit to uninformative rationales [27]. By establishing direct interaction between rationale extraction and classification, our model creates a reinforcing cycle: classification guides the selection of relevant evidence, while focused rationales improve classification accuracy.

Formally, given a user timeline $T_u = [x_{(u,1)}, \ldots, x_{(u,n)}]$, we extract token-level rationales $\mathcal{R}_u = \{r_{(u,1)}, \ldots, r_{(u,k)}\}$ that serve as supporting evidence for classification. Unlike post-level approaches, our token-level selection precisely identifies linguistic cues that contribute to troll detection, better enhancing interpretability.

The rationale selection process, detailed in Algorithm 1, operates as follows. Given user timeline $T_u$, we first concatenate all posts and compute contextual embeddings for each token. For token position $i$, we compute an attention score: $p_\psi(r_{k=1} \mid x_u) = f_\psi(x_u)_k$, where $f_\psi$ is our rationale selector parameterised by $\psi$, and $f_\psi(x_u)_k$ represents the attention probability assigned to token $x_{u,k}$. We retain tokens with probability exceeding threshold $\tau$:

$$r_k = \mathbf{I}[p_\psi(r_{k=1} \mid x_u) > \tau]$$
$$\mathcal{R}_u = \{x_{(u,k)} \mid r_{k=1}\},$$

where $\mathbf{I}[\cdot]$ is the indicator function and $\tau = 0.5$.

Without explicit gold-standard annotations for rationales, we apply two regularisation techniques to ensure high-quality selection:

*Sparsity Constraint.* We limit selected tokens to a maximum of $l$ or fraction $\alpha$ of input length:

$$\sum_{k=1}^{|x_u|} r_k \leq \min(l, \alpha|x_u|).$$

*Continuity Regularization.* We encourage selection of coherent linguistic spans by penalising discontinuities:

$$\mathcal{L}_{\text{cont}} = \sum_{k=2}^{|x_u|} |r_k - r_{k-1}|.$$

When enforcing the sparsity constraint (Algorithm 1, lines 8-12), we implement a dynamic selection process that prioritizes tokens with highest attention scores within the constraint budget. The continuity regularization (line 14) adopts a dynamic programming approach to find optimal contiguous spans that minimize discontinuity loss while respecting sparsity constraints.

The selected rationales serve dual purposes: they provide interpretable evidence for model predictions and support the generation of explanations. For classification, we pool the embeddings of

selected rationale tokens and compute classification logits (lines 16-18):

$$y_u = \sigma(\mathbf{w}_c^T \text{Pool}(\{H_i \mid r_i = 1\}) + b_c),$$

where $\text{Pool}(\cdot)$ aggregates rationale token embeddings and $\sigma$ is the sigmoid activation function.

Our unified approach creates a cycle where better rationales lead to more accurate classification, which in turn guides more precise rationale selection. This self-reinforcing mechanism outperforms traditional pipeline approaches, as demonstrated in our experimental results (Section 5).

**Algorithm 1** Rationale Selection

**Require:** User timeline $T_u = [x_{(u,1)}, x_{(u,2)}, ..., x_{(u,n)}]$, threshold $\tau$, sparsity constraint $\alpha$, continuity weight $\lambda_c$
**Ensure:** Rationales $\mathcal{R}_u$, troll classification $y_u$
1: **Initialize** attention scores $A = []$, rationale mask $r = []$
2: **Concatenate** timeline posts: $x_u = [x_{(u,1)}; x_{(u,2)}; ...; x_{(u,n)}]$
3: **Compute** contextual embeddings: $H = \text{Encoder}(x_u) \in \mathbb{R}^{|x_u| \times d}$
4: **for** each token position $i \in \{1, 2, ..., |x_u|\}$ **do**
5: Compute token-level attention score: $a_i = \sigma(\mathbf{w}_a^T H_i + b_a)$
6: Append to attention scores: $A = A \cup \{a_i\}$
7: **end for**
8: **Apply** threshold: $r_i = \mathbf{1}[a_i > \tau]$ for all $i$
9: **if** $\sum_i r_i > \alpha |x_u|$ **then**
10: Sort attention scores: $A_{\text{sorted}} = \text{Sort}(A, \text{descending} = \text{True})$
11: Keep top-$k$ tokens where $k = \lfloor \alpha |x_u| \rfloor$
12: Set mask $r_i = 1$ for tokens with top-$k$ attention scores, else 0
13: **end if**
14: **Apply** continuity regularisation:
15: Minimise $\mathcal{L}_{\text{cont}}$ as in Section 4.4
16: **Extract** rationale tokens: $\mathcal{R}_u = \{x_{(u,i)} | r_{i=1}\}$
17: **Compute** classification logits:
18: $h_{\mathcal{R}} = \text{Pool}(\{H_i | r_{i=1}\})$
19: $y_u = \sigma(\mathbf{w}_c^T h_{\mathcal{R}} + b_c)$
20: **return** $\mathcal{R}_u, y_u$

## 4.5 Summary Generation

The summary generator $s_\omega$ in X-Troll produces concise, natural language explanations derived from the selected rationales $\mathcal{R}$, offering clear justifications for classification decisions. Inspired by [43], we propose a troll-specific summary generator that incorporates rationale embeddings to enhance explanation quality.

Given a selected rationale $r \in \mathcal{R}$, we construct an explanatory summary $S$ using a base model with an adapter layer $E_A$:

$$S = s_\omega(r) = \text{LLM}\left([\texttt{CLS}] \oplus E_A(r) \oplus [\texttt{RAT}]\right),$$

where [CLS] and [RAT] are special tokens indicating the start of the sequence and the rationale segment, respectively, and $\oplus$ denotes concatenation. The adapter $E_A : \mathbb{R}^d \to \mathbb{R}^e$ maps the rationale embedding into the model's token embedding space, ensuring seamless integration of the rationale information into the language model.

The adapter $E_A$ is implemented as a two-layer multilayer perceptron (MLP):

$$E_A(r) = W_2 \cdot \text{ReLU}(W_1 r + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{h \times d}$, $W_2 \in \mathbb{R}^{e \times h}$, $b_1 \in \mathbb{R}^h$, and $b_2 \in \mathbb{R}^e$ are learnable parameters. Here, $d$ is the dimension of the rationale embedding, $h$ is the hidden dimension of the MLP.

# 5 Experiments

This section presents experimental findings across four areas: few-shot learning, ablation study, summary generation, and a qualitative case study. Few-shot (zero/one/five-shot) evaluations used a held-out test set, with LoRA adapters trained independently per task. We used AdamW (learning rate 1e-3, weight decay 0.01), training for 10 epochs with early stopping.

For k-shot evaluation, examples were randomly sampled while maintaining a balanced positive-negative distribution. A consistent test set was used across all shot settings for comparability. The experiments were run on PyTorch 2.0 using 4 NVIDIA A100 GPUs (40GB each). Results were averaged over five runs with different random seeds for robustness[2].

**Table 1: Data Statistics. The "appraisals" column indicates the number of tweets annotated with appraisal labels in each campaign category. Categories without appraisal annotations are marked with a dash (–).**

| Campaign | #users | #posts | #appraisals |
|---|---|---|---|
| Russia-Anti-NATO Troll | 70 | 26,684 | 124 |
| Russia-Anti-NATO Non-Troll | 140 | 36,895 | – |
| Russia-IRA Troll | 31 | 68,914 | 159 |
| Russia-IRA Non-Troll | 100 | 34,511 | – |
| PRC-Xinjiang Troll | 257 | 24,075 | 303 |
| PRC-Xinjiang Non-Troll | 1,444 | 356,112 | – |
| Random | 2,000 | 40,000 | – |

## 5.1 Datasets

For troll detection and information campaign classification tasks, we used datasets from three specific state-sponsored information campaigns: Russia-Anti-NATO Troll, Russia-IRA Troll, and PRC-Xinjiang Troll[3]. These datasets were released by Twitter and contain troll accounts and their posts. Twitter banned these accounts in October 2018 and anonymized them to ensure they could not be associated with individual users. For example, the 'Russia-Anti-NATO Troll' dataset contains users banned for amplifying narratives that sought to undermine faith in the NATO alliance and its stability[4]. We also collected posts from non-troll users via the Twitter data API to establish a baseline dataset for evaluating model performance on non-troll data. In our non-troll data collection, we excluded all profile information and retrieved only necessary data. We focused on posts related to specific topics to protect user privacy. Table 1 presents the summary of the data statistics including the number

[2]To ensure fair evaluation, we used consistent data splits across all experiments. The annotated dataset was divided into train:validation:test splits with ratios of 70:10:20.
[3]https://blog.x.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed
[4]https://blog.x.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-

**Table 2: Model performance ($F_1$)on troll detection and campaign classification tasks under few-shot settings. Results are reported for three settings: 0S (zero-shot), 1S (one-shot), 5S (five-shot).**

| Model | Troll Detection | | | Campaign Classification | | |
|---|---|---|---|---|---|---|
| | 0S | 1S | 5S | 0S | 1S | 5S |
| **In-context learning** | | | | | | |
| LLaMA | 0.456 | 0.502 | 0.549 | 0.328 | 0.375 | 0.421 |
| Falcon | 0.487 | 0.535 | 0.582 | 0.359 | 0.407 | 0.456 |
| Gemma | 0.505 | 0.553 | 0.600 | 0.377 | 0.425 | 0.474 |
| T5 | 0.421 | 0.468 | 0.515 | 0.293 | 0.339 | 0.386 |
| GPT-4 | **0.523** | 0.571 | 0.618 | **0.396** | **0.444** | **0.493** |
| MetaTroll | - | **0.582** | **0.689** | - | - | - |
| **LoRA fine-tuned** | | | | | | |
| LLaMA | 0.535 | 0.580 | 0.630 | 0.390 | 0.435 | 0.485 |
| Falcon | 0.570 | 0.615 | 0.665 | 0.425 | 0.470 | 0.520 |
| Gemma | **0.585** | **0.630** | **0.680** | **0.440** | **0.485** | **0.535** |
| T5 | 0.500 | 0.545 | 0.595 | 0.355 | 0.400 | 0.450 |
| **X-Troll** | | | | | | |
| LLaMA | 0.592 | 0.628 | 0.663 | 0.483 | 0.518 | 0.552 |
| Falcon | 0.627 | 0.661 | 0.696 | 0.522 | 0.557 | 0.591 |
| Gemma | **0.648** | **0.682** | **0.717** | **0.547** | **0.581** | **0.616** |
| T5 | 0.558 | 0.593 | 0.627 | 0.448 | 0.482 | 0.516 |

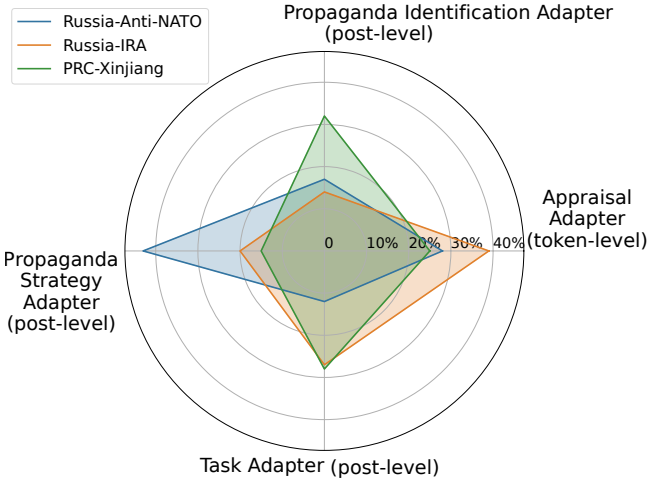



**Figure 2: Adapter weight distribution across information operations. The radar chart illustrates the relative weighting of four adapter types (Appraisal, Propaganda Identification, Propaganda Strategy, and Task) across three different information operations (Russia-Anti-NATO (blue), Russia-IRA (orange), and PRC-Xinjiang (green)).**

of users and posts. To enhance analysis quality, domain experts annotated a subset of posts with appraisal labels for each information campaign, as shown in Table 1. The propaganda strategy adapter was fine-tuned on datasets provided by Moral et al. [32] as part of the DIPROMATS 2023 shared task[5]. We also provide a sample of annotated posts with detailed labels (including appraisal and propaganda tags), which is included in Appendix [1].

## 5.2 Base Models

We evaluate X-Troll across diverse model architectures to demonstrate the generalizability of our linguistic knowledge integration approach. Our evaluation includes four base language models: three decoder-only architectures (LlaMA-3-8B [47], Falcon-7B [4], Gemma-7B [42]) and one encoder-decoder model (FLAN-T5-XL [10]). We compare X-Troll against established baselines across multiple settings: in-context learning and LoRA fine-tuning for all base models, GPT-4 [2] in-context learning, and MetaTroll [45], the current state-of-the-art BERT-based few-shot troll detection system.

## 5.3 Detection Performance

X-Troll show strong performance over both general-purpose LLMs and specialized transformer models across all evaluation settings. Table 2 shows that X-Troll (Gemma-7B) achieves $F_1$ scores of 0.648, 0.682, and 0.717 for zero-shot, one-shot, and five-shot troll detection respectively–consistent 6.3, 5.2, and 3.7 percentage point improvements over the best baseline (LoRA fine-tuned Gemma-7B).

**The Expert Knowledge Advantage.** The performance hierarchy reveals critical insights about automated troll detection capabilities. While general-purpose LLMs struggle with subtle manipulation techniques (GPT-4: 0.523 zero-shot $F_1$), specialized fine-tuning provides improvements (LoRA Gemma-7B: 0.585). However, X-Troll's systematic integration of linguistic expertise (0.648) shows that expert knowledge can bridge remaining performance gaps.

**Campaign Classification Insights.** For information campaign classification, the linguistic advantage becomes even more pronounced. X-Troll achieves 10.7, 9.6, and 8.1 percentage point improvements over LoRA baselines across few-shot settings, suggesting that campaign-specific linguistic signatures are particularly distinctive and can be effectively captured through our multi-adapter architecture. This validates our hypothesis of systematic rhetorical differences across state-sponsored operations.

**Practical Deployment Implications.** The performance gap between zero-shot (0.648 $F_1$) and five-shot (0.717 $F_1$) settings shows that X-Troll can provide reasonable detection capabilities even with minimal training examples from new campaign types. This rapid adaptation capability is essential for countering emerging threats where extensive labeled data may not be immediately available, demonstrating how linguistic expertise enables practical deployment in operational environments where traditional approaches would require extensive retraining.

## 5.4 Campaign-Specific Patterns

Our adapter weight analysis uncovers distinct behavioral patterns across state-sponsored information operations, providing unprecedented insights into the strategic doctrines underlying different campaigns. Fig. 2 reveals how different operations exhibit unique linguistic signatures that align with documented strategic approaches.

**Russia-Anti-NATO: Strategy-Driven Sophistication.** Russia-Anti-NATO campaigns show strong reliance on the Propaganda Strategy adapter (0.43) while minimizing direct Task-specific (0.12) and Propaganda Identification (0.17) features. This pattern reflects

[5]https://sites.google.com/view/dipromats2023

rhetorical approaches that prioritize subtle strategic techniques over overt propaganda signals. The emphasis on nuanced strategy aligns with documented "gray zone" warfare approaches characteristic of Russian information operations, where plausible deniability requires detailed rhetorical sophistication [7, 31].

**Russia-IRA: Appraisal-Focused Narrative Construction.** In contrast, Russia-IRA operations prioritize the Appraisal adapter (0.39) and Task features (0.27) while showing reduced dependence on Propaganda Strategy (0.20). This pattern validates findings by Lazer et al. [22] documenting how Russian disinformation campaigns by IRA use evaluative language to construct persuasive narrative frames. These operations prioritize subtle sentiment steering through appraisal manipulation rather than explicit rhetoric, directly impacting detection system design.

**PRC-Xinjiang: Balanced Multi-Modal Doctrine.** PRC-Xinjiang campaigns show a more balanced activation pattern with heightened Propaganda Identification (0.32) alongside maintained Appraisal (0.25) and Task-specific (0.28) weights. This balanced approach suggests a different operational doctrine that combines explicit propaganda techniques with subtle linguistic manipulation, possibly reflecting distinct cultural and strategic contexts for Chinese information operations compared to Russian approaches.

**Detection System Implications.** These findings provide three insights for troll detection task. First, they empirically validate that state-sponsored operations use campaign-specific linguistic strategies that can be systematically distinguished through computational analysis. Second, they demonstrate the necessity of multi-faceted detection approaches–relying solely on propaganda identification would systematically miss the appraisal-based strategies prevalent in Russian campaigns. Third, they validate our dynamic gating mechanism's ability to automatically focus on contextually relevant linguistic features without manual reconfiguration.

## 5.5 Explainability Analysis

We analyze X-Troll's explainability through three dimensions: summary generation evaluation, case analysis of successful rationale selection and error analysis of false positives.

**Summary Generation Evaluation.** Table 3 reveals certain patterns in explanation generation performance. The rationale selector is able to improve explanation quality across most model-dataset combinations, with strong improvements for FLAN-T5 on Russia-IRA data (7.1% improvement) and Falcon-7B on PRC-Xinjiang data (6.2% improvement). However, some configurations show slight degradation (FLAN-T5 on Russia-Anti-NATO: −4.5%), suggesting that rationale effectiveness varies with campaign characteristics and base model capabilities. In Appendix [1], we include the detailed G-Eval prompts and templates used for scoring [28], as well as the complete evaluation results.

**Mechanistic Insights from Case Analysis.** Fig. 3 illustrates both the power and limitations of our rationale-based approach. In correctly identified cases, X-Troll successfully highlights emotionally charged phrases ("leaders of the gangs," "provocations") that align with established propaganda techniques–opponent discrediting and loaded language respectively. This shows contextual pattern recognition beyond simple emotional language detection, validating the value of token-level evidence extraction over post-level approaches.

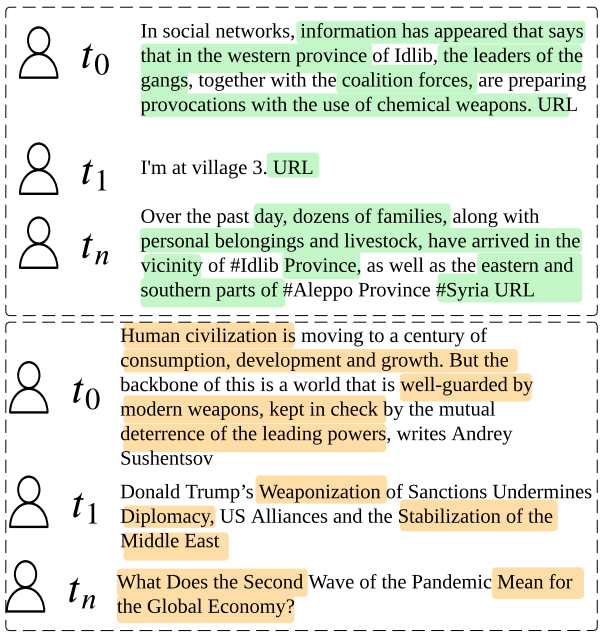



**Figure 3: X-Troll's rationale selection on Russia-IRA examples, with diagnostic tokens highlighted. The correctly classified troll post (top) shows characteristic geopolitical framing and conflict narratives, while the false positive (bottom) reveals how political topic overlap without coordinated rhetorical patterns can mislead classification.**

**False Positive Pattern Analysis.** Misclassification analysis reveals the challenges in distinguishing trolling patterns from legitimate political discourse. Fig. 3 (bottom) shows a representative false positive where X-Troll incorrectly classified an authentic user as a troll. In this case, the system focuses on ideological framing language ("human civilization") and geopolitical discourse markers rather than manipulation-specific linguistic patterns characteristic of coordinated information operations. The rationale selector highlights phrases that reflect legitimate geopolitical analysis rather than the systematic opponent discrediting or loaded language patterns typical of state-sponsored trolls. This error shows that the system can be misled by complex political discourse that uses abstract conceptual framing without manipulative intent.

## 5.6 Ablation Studies

To evaluate the impact of individual adapters in X-Troll, we conducted ablation studies shown in Fig. 4. We examine two configurations to quantify the role of adapters and impact of expert knowledge integration. We evaluate two configurations: (1) removing individual adapters while preserving the gating mechanism (+G), and (2) using single adapters without any gating (-G).

The full X-Troll model achieves the highest $F_1$ scores of 0.885 for troll detection and 0.870 for campaign classification. This confirms that adapter fusion and expert knowledge integration significantly enhance classification accuracy. Performance degradation upon

**Table 3: Impact of rationale selection on explanation quality across three information campaigns. Scores are G-Eval ratings (1-5 scale) for coherence, consistency, fluency, and relevance. Percentage changes indicate relative improvement (+) or degradation (-) from baseline. Bold indicates stongest improvements per campaign; *italics* denote performance decreases.**

| Model | Russia-Anti-NATO | Russia-IRA | PRC-Xinjiang |
|---|---|---|---|
| X-Troll w/o rationale selector | | | |
| LLaMA-3B | **3.607** | 3.231 | 3.060 |
| Falcon-7B | 3.032 | **3.348** | 3.292 |
| Gemma-7B | 3.482 | 3.127 | **3.524** |
| FLAN-T5 | 2.904 | 2.609 | 2.937 |
| X-Troll with rationale selector | | | |
| LLaMA-3B | **3.847 (+6.7%)** | *3.186 (-1.4%)* | 3.075 (+0.5%) |
| Falcon-7B | 3.148 (+3.8%) | 3.474 (+3.8%) | **3.496 (+6.2%)** |
| Gemma-7B | 3.552 (+2.0%) | 3.256 (+4.1%) | 3.536 (+0.3%) |
| FLAN-T5 | *2.774 (-4.5%)* | **2.796 (+7.1%)** | 3.059 (+4.1%) |

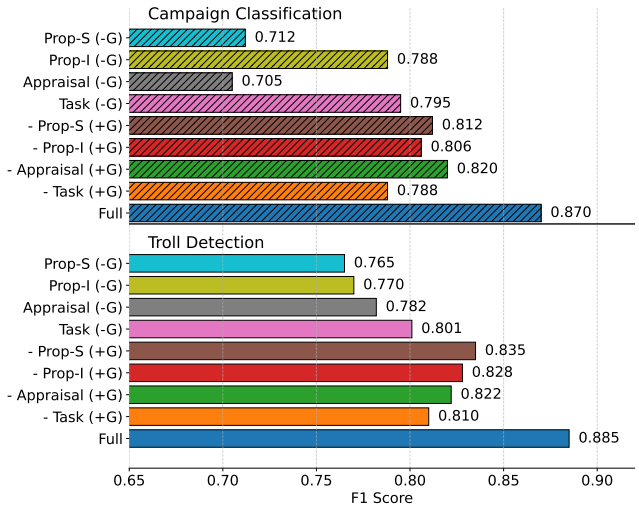



**Figure 4: Ablation study results showing F1 scores for troll detection and campaign classification tasks. Each bar represents a configuration: "Full" uses all four adapters with gating, "- [Adapter] (+G)" removes one adapter while keeping the gating mechanism, and "[Adapter] (-G)" uses only a single adapter without gating. Prop-I = Propaganda Identification adapter, Prop-S = Propaganda Strategy adapter.**

adapter removal reveals a hierarchical importance among knowledge sources.

**Task Adapter Centrality.** Fig. 4 demonstrates that the Task Adapter proves most critical for both troll detection (−7.5% when removed) and campaign classification (−8.2%). This finding validates the importance of direct supervised learning from troll-labeled data while highlighting how expert knowledge augments rather than replaces task-specific adaptation.

**Complementary Expert Knowledge Effects.** Appraisal and Propaganda adapters show comparable importance levels (5 − 6% performance drops when removed), but minimal configurations reveal their interdependence. Single adapters alone show performance degradation (up to 16.5% for campaign classification with Appraisal only), showing that expert knowledge sources achieve maximum effectiveness through integration rather than isolation.

**Dynamic Gating Validation.** The comparison between gated (+G) and non-gated (-G) configurations validates our dynamic integration approach. Without gating, the Task Adapter alone achieves only 0.801 $F_1$ for troll detection and 0.795 for campaign classification, representing substantial degradation from the full model's 0.885/0.870 performance. In contrast, non-gated single adapters suffer dramatic performance losses (Appraisal alone: 0.782/0.705 $F_1$, −11.6%/−19.0% drops), demonstrating strong complementarity effects. The performance gap between gated and non-gated configurations (e.g., Appraisal: 0.822 vs. 0.782) confirms that effective troll detection requires dynamic integration of multiple knowledge sources rather than static combination approaches.

## 6 Conclusion

We introduced X-Troll, an explainable framework that integrates linguistic expert knowledge with large language models for state-sponsored troll detection. By systematically incorporating appraisal theory and propaganda analysis through specialized LoRA adapters, X-Troll is trying to address the limitations in current detection systems: poor interpretability and inability to capture sophisticated manipulation strategies. Our evaluation demonstrates that linguistic knowledge integration provides substantial benefits, achieving 5 − 10 percentage point improvements over strong baselines while generating human-readable explanations. The dynamic gating mechanism reveals distinct strategic patterns across information operations, with Russian campaigns emphasizing appraisal-based manipulation while Chinese operations use more balanced propaganda techniques. Key findings extend beyond detection performance. Our hierarchical knowledge integration shows that expert linguistic insights achieve maximum effectiveness when combined with task-specific learning, providing a template for incorporating domain expertise in security applications.

X-Troll's explainable approach contributes to a growing ecosystem of tools for understanding and countering online manipulation. While Kong et al. [21] focuses on detecting operations through social reactions and Tian et al. [44] enables early prediction of content engagement, X-Troll provides the crucial missing piece: explainable identification of the actors themselves. Future work could integrate these complementary approaches, combining early engagement prediction with explainable actor detection to create comprehensive early warning systems for information operations.

## Acknowledgements

This research was supported by the Australian Research Council Discovery Project DP200101441, the Advanced Strategic Capabilities Accelerator (ASCA), the Australian Department of Home Affairs, Commonwealth of Australia as represented by the Defence Science and Technology Group of the Department of Defence, and the Defence Innovation Network.

## GenAI Usage Disclosure

This work was created, reviewed, and edited by human authors. AI tools were used in two specific capacities: (1) debugging the code components of the X-Troll framework, and (2) writing assistance to improve conciseness and readability of manuscript sections.

For writing assistance, we used Claude (Anthropic) with prompts such as: "Improve the writing of this paragraph from a scientific paper. Keep concise, and improve reading flow. Match style. Highlight changes. Break down complex and long sentences and make more concise." All AI-generated suggestions were critically reviewed, modified, and integrated by human authors. The original conceptual content, technical contributions, experimental design, analysis, and final editorial decisions remain entirely human-authored. AI tools did not contribute to the research methodology, data analysis, or scientific conclusions.

## References


[1] Appendix. X-Troll: eXplainable Detection of State-Sponsored Information Operations Agents. (Appendix). https://arxiv.org/pdf/2508.16021#page=12

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 15–25.

[4] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. *Falcon-40B: an open large language model with state-of-the-art performance*. Technical Report. Technical report, Technology Innovation Institute.

[5] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation?. In *Companion Proceedings of The 2019 World Wide Web Conference*. 162–168.

[6] Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. 2023. Riding information crises: the performance of far-right Twitter users in Australia during the 2019–2020 bushfires and the COVID-19 pandemic. *Information, Communication & Society* (4 2023), 1–19. doi:10.1080/1369118X.2023.2205479

[7] André Barrinha and Thomas Renard. 2017. Cyber-diplomacy: the making of an international society in the digital age. *Global Affairs* 3, 4-5 (2017), 353–364.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Pio Calderon, Rohit Ram, and Marian-Andrei Rizoiu. 2024. Opinion Market Model: Stemming Far-Right Opinion Spread Using Positive Interventions. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (5 2024), 177–190. doi:10.1609/icwsm.v18i1.31306

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[11] Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 5636–5646.

[12] Ritam Dutt, Ashok Deb, and Emilio Ferrara. 2018. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. In *International conference on intelligent information technologies*. Springer, 151–168.

[13] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* (2020).

[14] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[15] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2020. Still out there: Modeling and identifying russian troll accounts on twitter. In *12th ACM Conference on Web Science*. 1–10.

[16] Olivia Inwood and Michele Zappavigna. 2021. Ambient affiliation, misinformation and moral panic: Negotiating social bonds in a YouTube internet hoax. *Discourse & Communication* 15, 3 (2021), 281–307.

[17] Amelia Johns, Francesco Bailo, Emily Booth, and Marian-Andrei Rizoiu. 2024. Labelling, shadow bans and community resistance: did Meta's strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread? *Media International Australia* (3 2024), 36. doi:10.1177/1329878X241236984

[18] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8783–8790.

[19] Dongwoo Kim, Timothy Graham, Zimin Wan, and Marian-Andrei Rizoiu. 2019. Analysing user identity via time-sensitive semantic edit distance (t-SED): a case study of Russian trolls on Twitter. *Journal of Computational Social Science* 2 (7 2019), 331–351. Issue 2. doi:10.1007/s42001-019-00051-x

[20] Quyu Kong, Emily Booth, Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. 2022. Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (5 2022), 524–535. Issue 1. doi:10.1609/icwsm.v16i1.19312

[21] Quyu Kong, Pio Calderon, Rohit Ram, Olga Boichak, and Marian-Andrei Rizoiu. 2023. Interval-censored Transformer Hawkes: Detecting Information Operations using the Reaction of Social Systems. In *Proceedings of the ACM Web Conference 2023* (New York, NY, USA). ACM, 1813–1821. doi:10.1145/3543507.3583481

[22] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[23] Alfred Lee and Elizabeth Briant Lee. 1939. The fine art of propaganda. (1939).

[24] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[25] Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. 2023. Towards Trustworthy Rumor Detection with Interpretable Graph Structural Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4089–4093.

[26] Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, YuanKai Zhang, and Yang Qiu. 2023. MGR: Multi-generator Based Rationalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 12771–12787. doi:10.18653/v1/2023.acl-long.715

[27] Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. 2022. FR: Folded rationalization with a unified encoder. *Advances in Neural Information Processing Systems* 35 (2022), 6954–6966.

[28] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2511–2522.

[29] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 505–514.

[30] James R Martin and Peter R White. 2003. *The language of evaluation*. Vol. 2. Springer.

[31] Michael J Mazarr et al. 2015. Mastering the gray zone: understanding a changing era of conflict. (2015).

[32] Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. 2023. Overview of DIPROMATS 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural* 71 (2023), 397–407.

[33] Shiwen Ni, Jiawen Li, and Hung-Yu Kao. 2021. MVAN: Multi-view attention networks for fake news detection on social media. *IEEE Access* 9 (2021), 106907–106917.

[34] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4932–4942.

[35] Rohit Ram, Quyu Kong, and Marian-Andrei Rizoiu. 2021. Birdspotter: A Tool for Analyzing and Labeling Twitter Users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (New York, NY, USA). ACM, 918–921. doi:10.1145/3437963.3441695

[36] Rohit Ram, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. 2025. Practical Guidelines for Ideology Detection Pipelines and Psychosocial Applications. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (6 2025), 1630–1648. doi:10.1609/icwsm.v19i1.35892

[37] Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. 2018. #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. *Proceedings of the International AAAI Conference on Web and Social Media* 12 (6 2018), 1–10. Issue 1.

doi:10.1609/icwsm.v12i1.15029

[38] Hossein Shafiei and Aresh Dadlani. 2022. Detection of fickle trolls in large-scale online social networks. *Journal of big Data* 9, 1 (2022), 1–21.

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

[40] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 58, 5 (2021), 102618.

[41] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[42] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).

[43] Guy Tennenholtz, Yinlam Chow, ChihWei Hsu, Jihwan Jeong, Lior Shani, Azamat Tulepbergenov, Deepak Ramachandran, Martin Mladenov, and Craig Boutilier. 2024. Demystifying Embedding Spaces using Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[44] Lin Tian, Emily Booth, Francesco Bailo, Julian Droogan, and Marian-Andrei Rizoiu. 2025. Before It's Too Late: A State Space Model for the Early Prediction of Misinformation and Disinformation Engagement. In *Proceedings of the International Web Conference (WWW)*. doi:10.1145/3696410.3714527

[45] Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2023. Metatroll: Few-shot Detection of State-Sponsored Trolls with Transformer Adapters. In *Proceedings of the ACM Web Conference 2023*. 1743–1753.

[46] Lin Tian, Xiuzhen Jenny Zhang, Myung Hee Kim, and Jennifer Biggs. 2023. Task and Sentiment Adaptation for Appraisal Tagging. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1960–1970.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

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

[49] Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing Hate Speech Detection Using Multi-Task Learning: A Case Study of Political Public Figures. *Computer Speech & Language* 89 (1 2025), 101690. doi:10.1016/j.csl.2024.101690

[50] Lanqin Yuan, Philipp J. Schneider, and Marian-Andrei Rizoiu. 2025. Behavioral Homophily in Social Media via Inverse Reinforcement Learning: A Reddit Case Study. *Proceedings of the International Web Conference (WWW)* (2 2025). doi:10.1145/3696410.3714618

[51] Lanqin Yuan, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, and Marian-Andrei Rizoiu. 2023. Transfer learning for hate speech detection in social media. *Journal of Computational Social Science* 6 (10 2023), 1081–1101. Issue 2. doi:10.1007/s42001-023-00224-9

[52] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 World Wide Web Conference*. 218–226.

[53] Yini Zhang, Josephine Lukito, Min-Hsin Su, Jiyoun Suk, Yiping Xia, Sang Jung Kim, Larissa Doroshenko, and Chris Wells. 2021. Assembling the networks and audiences of disinformation: How successful Russian IRA Twitter accounts built their followings, 2015–2017. *Journal of Communication* 71, 2 (2021), 305–331.