ExTax: Explainable Disinformation Detection via Persuasion, Emotion, and Narrative Role Taxonomies

Anonymous ACL submission

Abstract

The democratization of LLMs has accelerated the generation and circulation of highly fluent disinformation, making traditional syntax-semantic verification increasingly insufficient. Such deception rarely relies solely on surface-level falsity; instead, it often combines persuasive rhetoric, emotional manipulation, and narrative role construction to influence readers' interpretations through multiple cognitive pathways. However, existing detectors typically emphasize isolated signals—such as syntax, external knowledge, persuasion, or affective cues—and therefore struggle to capture the multi-faceted manipulative intents underlying disinformation or provide human-auditable explanations. To address this gap, we present **ExTax**, a taxonomy-aligned framework for explainable disinformation detection. ExTax unifies persuasive rhetoric, emotional manipulation, and narrative roles into a 17-dimensional taxonomic space, covering 6 persuasive-rhetoric strategies, 5 emotional-manipulation methods, and 6 narrative-role categories. It elicits attributes from multiple frontier LLMs, reconciles their disagreements through Entropy-driven Dynamic Label Smoothing, and fuses the resulting taxonomic representations with contextual encodings via Heterogeneous Multi-Head Attention, grounding each prediction in an interpretable manipulation profile. Across five cross-domain and cross-genre benchmarks, ExTax achieves an overall Macro F_1 of 0.8456, outperforming state-of-the-art deep learning and LLM-based baselines. It also remains robust under severe genre imbalance, where the strongest deep baseline degrades from 0.9454 to 0.6194. Our code is available at `https://anonymous.4open.science/r/ExTax-1FBD/`.

1 Introduction

Disinformation detection has become increasingly challenging as social media platforms and generative AI systems accelerate the production and circulation of deceptive yet highly fluent content. Such content rarely operates through surface wording or outright falsity alone; instead, it often combines rhetorical strategies, emotional manipulation, and narrative framing to attack or simplify targets, evoke high-arousal emotions, and assign actors to morally charged roles. In this way, disinformation shapes public interpretation and sharing behavior through multiple cognitive routes (Saeed et al., 2022; Loth et al., 2026). Meanwhile, alarmist discourse around disinformation may itself reshape public risk perception, making reliable and interpretable detection increasingly important (Jungherr and Rauchfleisch, 2024).

Recent studies improve robustness by incorporating richer textual, contextual, or external signals. Syntax-aware models capture dependency-level distortions (Xiao et al., 2024), dynamic representation methods adapt feature views to instance-level difficulty (Farhangian et al., 2025), and knowledge- or retrieval-augmented LLM frameworks incorporate external evidence or affective cues to enable more timely detection (Jin et al., 2025; Liu et al., 2025). Intent-aware detection additionally suggests that communicative purposes provide stable signals beyond surface wording (Wang et al., 2025). However, most existing methods still model syntax, knowledge, emotion, persuasion, or intent in isolation, leaving underexplored how multiple human-interpretable dimensions jointly reveal the manipulative structure of disinformation.

We argue that robust disinformation detection should reason over three complementary facets of manipulation: persuasive rhetoric, emotional manipulation, and narrative roles. These facets characterize which rhetorical strategy is employed, which emotional-manipulation method is activated, and which morally charged role is assigned. Grounded in persuasion-oriented taxonomies, emotional manipulation studies, as well as narrative framing research (Piskorski et al., 2023b; Jamieson et al.,

2025; Zhao et al., 2024; Mahmoud et al., 2025), these facets provide human-interpretable manipulation evidence that extends beyond single-cue reasoning.

Realizing such a unified taxonomy introduces two challenges: no existing corpus jointly annotates these facets, and attributes elicited from a single LLM may inherit model-specific biases. To address these issues, we propose **ExTax**, an **Ex**plainable **Tax**onomy-aligned framework for disinformation detection. ExTax elicits taxonomy-guided attributes from multiple frontier LLMs, regularizes their disagreements through entropy-driven dynamic label smoothing, and constructs a unified 17-dimensional taxonomic representation consisting of 6 persuasive-rhetoric strategies, 5 emotional-manipulation methods, and 6 narrative-role categories. A taxonomy-aligned heterogeneous attention module then integrates this representation with contextual encodings, grounding each prediction in a per-sample manipulation profile.

Across five cross-domain, cross-genre benchmarks, ExTax achieves a Macro F_1 of 0.8456, outperforming the strongest LLM and deep-learning baselines by +2.0 and +3.1 points, respectively. It also remains robust under severe genre imbalance, where the strongest deep baseline degrades from 0.9454 to 0.6194 across genres.

Overall, ExTax makes three key contributions: (1) **Multifaceted taxonomic modeling.** We reframe disinformation detection as reasoning over three complementary facets of manipulation—persuasive rhetoric, emotional manipulation, and narrative roles—and instantiate these facets as a unified 17-dimensional taxonomic representation. (2) **Noise-robust attribute synthesis.** We aggregate taxonomy-guided outputs from multiple frontier LLMs and regularize noisy hard decisions through entropy-driven dynamic label smoothing, thereby transforming inter-LLM disagreement into reliable synthetic supervision. (3) **Taxonomy-aligned detection and explanation.** We integrate learned taxonomic signals with contextual encodings through heterogeneous attention, enabling robust disinformation detection while exposing rhetorical, affective-manipulative, and narrative evidence underlying each prediction.

2 Related Work

Disinformation detection. Disinformation detection has been studied under both supervised learning and LLM-based prompting paradigms. Supervised approaches typically rely on lexical, semantic, engagement, or contextual features to learn task-specific detectors (Aslam et al., 2021), while recent zero- and few-shot LLM methods address data scarcity and cross-domain generalization (Pelrine et al., 2023; Bang et al., 2023). Prompting strategies such as zero-shot Chain-of-Thought (Lucas et al., 2023) substantially improve detection performance. However, content-level methods may still overfit to dataset artifacts and generalize poorly across genres and topics, motivating supervised detectors that integrate structured, human-interpretable stylistic, rhetorical, or contextual evidence (Modzelewski et al., 2025).

Emotion, persuasion, and narrative framing in disinformation. Communication research has established emotion as central to how audiences process misleading content. The Appraisal-Tendency Framework (Han et al., 2007) explains how discrete emotions such as anger and sadness produce distinct cognitive appraisals. Zhao et al. (2024) extended this framework to narrative disinformation, showing that anger prompts heuristic action while sadness fosters deliberation, and Paletz et al. (2023) found that discrete emotions better predict post virality than valence or arousal. Bago et al. (2022) further showed that emotional reliance increases susceptibility to fake news. In parallel, NLP research has produced persuasive-technique taxonomies (Piskorski et al., 2023a), anchoring shared tasks on propaganda and manipulation detection. Entity framing and role portrayal provide a complementary lens for analyzing how actors are positioned within news narratives (Mahmoud et al., 2025).

Integrating high-level communicative signals. Recent work integrates such signals into detection pipelines. RAEmoLLM (Liu et al., 2025) leverages emotional embeddings for cross-domain disinformation detection, while PCoT (Modzelewski et al., 2025) infuses persuasion knowledge into LLM reasoning. Kamali et al. (2024) further showed that persuasive writing strategies can serve as intermediate signals for health disinformation detection. Unlike prior work that treats emotion, persuasion, and narrative framing as separate evidence sources,

ExTax jointly models them within a unified 17-dimensional taxonomic space and mitigates noise in LLM-elicited attributes through entropy-driven soft labeling.

3 Preliminaries

3.1 Problem Formulation

Disinformation detection is defined as the task of identifying false, inaccurate, or misleading information that is intentionally designed, presented, and promoted to cause public harm or generate profit. Formally, this task is structured as a binary classification problem in which a model $\mathcal{F}$ evaluates an input text T, such as a news article or a social media post, to determine its credibility. Given a dataset $\mathcal{D} = \{(T_i, y_i)\}_{i=1}^{n}$ where T_i represents the content and $y_i \in \{0, 1\}$ is the ground-truth label, the objective is to learn a predictive function $\boldsymbol{f}_{det} : \mathcal{T} \rightarrow \{0, 1\}$ such that:

$$\hat{y}_i = \boldsymbol{f}_{det}(T_i), \tag{1}$$

where $\hat{y}_i = 1$ denotes that the text T_i is classified as disinformation, and $\hat{y}_i = 0$ indicates it is reliable information.

3.2 Taxonomy of Disinformation: Persuasion, Emotion, and Narrative Roles

ExTax operationalizes the three-route cognitive triangulation introduced in § 1 through three complementary taxonomic dimensions, each grounded in a distinct strand of prior literature. For **Persuasion**, we adopt the hierarchical taxonomy of Piskorski et al. (2023b), which organizes 23 fine-grained persuasive techniques into six high-level strategies: Attack on Reputation, Justification, Simplification, Distraction, Call, and Manipulative Wording. For **Emotion**, we consider five affective states associated with manipulative content—fear, anger, hope, anxiety, and sadness—motivated on prior findings on emotional manipulation and narrative misinformation (Jamieson et al., 2025; Zhao et al., 2024). For **Narrative Roles**, we build upon the entity-framing taxonomy of Mahmoud et al. (2025) and adopt six role categories: Ethical Stabilizers, Altruistic Catalysts, Overt Aggressors, Deceptive Subversives, Institutional Toxins, and Marginalized Sufferers. Together, these three dimensions constitute a unified 17-dimensional taxonomic space. Detailed category definitions are provided in Appendix A.

4 Methods

The architecture of ExTax is illustrated in Figure 1, and comprises three core components: (1) Taxonomic Attribute Elicitation, which employs multiple LLMs to extract Persuasion, Emotion, and Narrative Role attributes and applies entropy-driven dynamic label smoothing to mitigate model noise; (2) Taxonomy-aligned Representation Learning, which aligns posts or news articles with a unified 17-dimensional taxonomic space; and (3) Taxonomy-based Disinformation Detection, which integrates taxonomic attributes with contextual text encodings for final prediction. Additional architectural details are provided in Appendix C.

4.1 Taxonomic Attribute Elicitation

Since no large-scale corpus jointly labels the three facets of manipulation, we construct soft taxonomic supervision by querying $N_{LLM} = 4$ top-tier LLMs and reconciling their disagreement. Specifically, for each input $T_i \in \mathcal{D}$, the Taxonomic Attribute Elicitation process extracts attributes associated with Persuasive Rhetoric, Emotional Manipulation, and Narrative Roles. This process yields an augmented dataset $\mathcal{D}_{\text{tax}} = \{(T_i, y_i, \boldsymbol{y}_{\text{tax}}^{(i)})\}_{i=1}^{n}$, where $\boldsymbol{y}_{\text{tax}}^{(i)} \in \mathcal{Y}_{\text{smooth}}$ is a 17-dimensional dynamically smoothed soft target vector encoding the refined presence probability of each taxonomic category in T_i.

Divergent attribute synthesis. For each text T_i, we query four frontier LLMs $\{M_k\}_{k=1}^{4}$ with taxonomy-specific prompts (Appendix B) to obtain three binary feature vectors over the persuasive-rhetoric ($|\mathcal{C}_P| = 6$), emotional-manipulation ($|\mathcal{C}_E| = 5$), and narrative-role ($|\mathcal{C}_N| = 6$) category sets: $\boldsymbol{p}_{i,k} \in \{0, 1\}^6, \boldsymbol{e}_{i,k} \in \{0, 1\}^5, \boldsymbol{n}_{i,k} \in \{0, 1\}^6$. A component is assigned a value of 1 if the LLM identifies the corresponding category in T_i. The raw outputs collected from all four models form the raw taxonomic dataset $\mathcal{R}$, on which we subsequently apply the entropy-driven smoothing procedure described below.

Entropy-driven dynamic label smoothing. To reduce noise in LLM-elicited attributes, we transform the hard decisions produced by four LLMs into entropy-aware soft labels. For notational simplicity, we use $z_{i,k}^{(c)}$ to denote the binary output of the k-th LLM for any category $c \in \mathcal{C}_P \cup \mathcal{C}_E \cup \mathcal{C}_N$. For each sample T_i and category c, we compute the

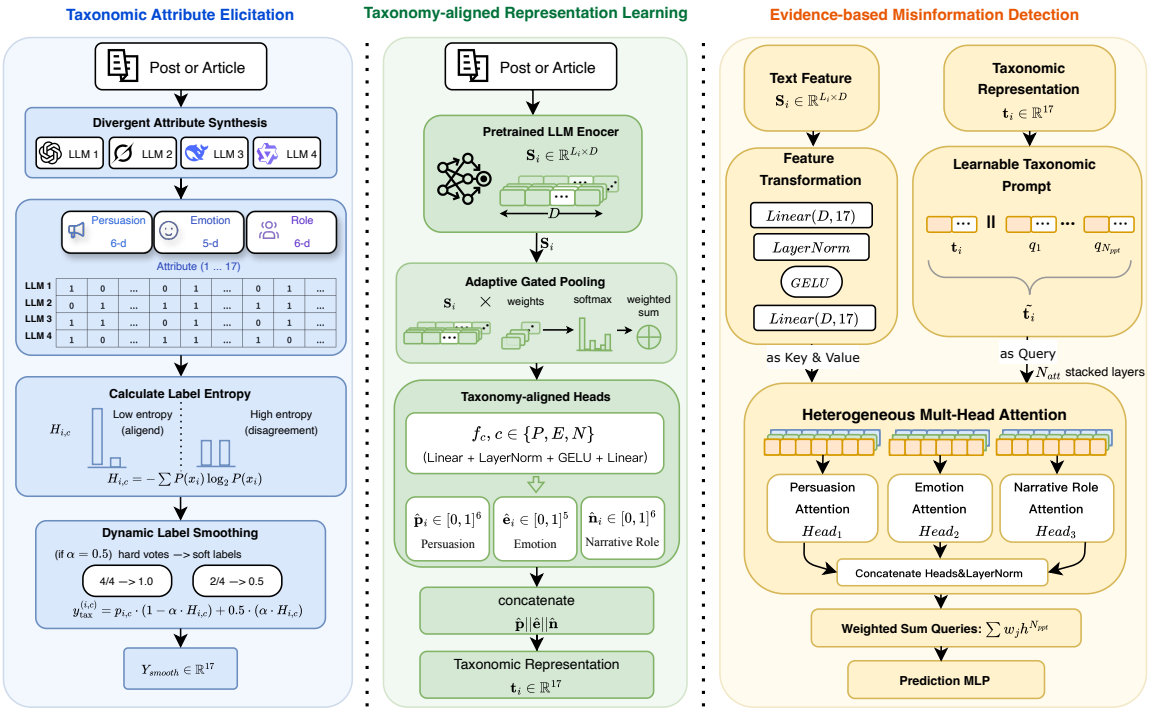

Figure 1: Overall framework of ExTax.

positive vote proportion:

$$p_{i,c} = \frac{1}{4}\sum_{k=1}^{4} z_{i,k}^{(c)}, \tag{2}$$

where $p_{i,c}$ represents the fraction of LLMs predicting the presence of category c. We then estimate the inter-model uncertainty using binary entropy:

$$H_{i,c} = -\left[p_{i,c}\log_2 p_{i,c} + (1-p_{i,c})\log_2(1-p_{i,c})\right], \tag{3}$$

with $0\log_2 0 = 0$. The final smoothed taxonomic target is defined as:

$$y_{\text{tax}}^{(i,c)} = p_{i,c}(1-\alpha H_{i,c}) + 0.5(\alpha H_{i,c}), \tag{4}$$

where α controls the smoothing strength. Thus, high-disagreement labels are pulled toward the neutral value 0.5, while high-agreement labels remain close to the original vote proportion. This procedure is applied to all categories in $\mathcal{C}_P \cup \mathcal{C}_E \cup \mathcal{C}_N$, yielding the smoothed taxonomic targets $\mathcal{Y}_{\text{smooth}}$.

4.2 Taxonomy-aligned Representation Learning

The objective of the Taxonomy-aligned Representation task is to map news articles into a multi-faceted taxonomic space defined by Persuasion, Emotion, and Narrative Roles. Specifically, given the textual dataset $\mathcal{D}$ and the dynamically smoothed taxonomic labels $\mathcal{Y}_{\text{smooth}}$, the task requires the model to learn a predictive mapping from the textual domain to the corresponding taxonomic label space:

$$\boldsymbol{f}_{tax} : \mathcal{T} \rightarrow \mathcal{Y}_{\text{smooth}}. \tag{5}$$

Namely, for any input news text $T_i \in \mathcal{T}$, the model yields a prediction vector $\hat{\boldsymbol{y}}_{tax}^{(i)}$ whose dimensionality aligns with that of $\mathcal{Y}_{\text{smooth}}$ (i.e., a 17-dimensional vector representing the combined taxonomic facets).

Adaptive gated pooling. For each input text T_i, we obtain token-level hidden states from a pretrained LLM:

$$\mathbf{S}_i = \text{LLM}(T_i) \in \mathbb{R}^{L_i \times D}, \tag{6}$$

where L_i is the sequence length and D is the hidden dimensionality. Instead of relying on mean or first-token pooling, we introduce a learnable gating vector that adaptively assigns importance weights to individual tokens:

$$\mathbf{v}_i = \sum_{h=1}^{L_i} g_{i,h}\mathbf{S}_{i,h} \in \mathbb{R}^D, \tag{7}$$

where $\mathbf{g}_i = \text{softmax}(\mathbf{w}_{1:L_i})$ denotes the normalized token-level gating weights. This mechanism allows the model to learn a weighted aggregation of token features before taxonomic projection.

Taxonomy-aligned representation learning training. Given the pooled representation $\mathbf{v}_i \in \mathbb{R}^D$, we employ three facet-specific MLP heads, corresponding to Persuasion, Emotion, and Narrative Roles, to project the text into their respective taxonomic spaces:

$$\hat{\mathbf{p}}_i = f_P(\mathbf{v}_i), \quad \hat{\mathbf{e}}_i = f_E(\mathbf{v}_i), \quad \hat{\mathbf{n}}_i = f_N(\mathbf{v}_i), \tag{8}$$

where each head outputs sigmoid-normalized probabilities over its corresponding taxonomy categories. The model is trained using Binary Cross-Entropy (BCE) supervision against the smoothed taxonomic targets:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_P + \mathcal{L}_E + \mathcal{L}_N. \tag{9}$$

Each facet-specific loss is computed independently over its corresponding category space using the smoothed labels derived from $\mathcal{Y}_{\text{smooth}}$.

4.3 Taxonomy-based Disinformation Detection

In the Taxonomy-based Disinformation Detection phase, we integrate the taxonomic feature vector $\mathbf{t}_i = [\hat{\mathbf{p}}_i \parallel \hat{\mathbf{e}}_i \parallel \hat{\mathbf{n}}_i]$, formed by concatenating the attributes elicited from the taxonomic alignment tasks, with the text feature sequence $\mathbf{S}_i$. This fusion mechanism enables the model to leverage structured rhetorical, emotional, and narrative evidence when predicting whether a given news article constitutes disinformation.

Text feature transformation. To align contextual token features with the taxonomic space, we project each token representation into the same 17-dimensional space as the taxonomic attributes:

$$\tilde{\mathbf{S}}_{i,h} = \text{TransMLP}(\mathbf{S}_{i,h}) \in \mathbb{R}^{17}, \quad h = 1, \ldots, L_i. \tag{10}$$

The resulting sequence $\tilde{\mathbf{S}}_i \in \mathbb{R}^{L_i \times 17}$ serves as the contextual feature input for heterogeneous multi-head attention.

Learnable taxonomic prompt. To enrich the semantic representations of the taxonomic dimensions, we introduce a Learnable Taxonomic Prompt mechanism. Specifically, the previously obtained taxonomic feature vector $\mathbf{t}_i \in \mathbb{R}^{17}$ is concatenated with N_{ppt} learnable semantic prompt vectors $\mathbf{q} \in \mathbb{R}^{N_{ppt} \times 17}$ of the same dimensionality, yielding an augmented representation:

$$\tilde{\mathbf{t}}_i = \text{Concat}(\mathbf{t}_i, \mathbf{q}) \in \mathbb{R}^{(1+N_{ppt}) \times 17}. \tag{11}$$

This concatenation operation enables the model to adaptively recalibrate feature importance based on both the original taxonomic attributes and the learnable prompts. Consequently, the framework can more effectively capture nuanced semantic cues.

Heterogeneous multi-head attention. To maintain a strict one-to-one correspondence with the predetermined taxonomic spaces, we propose a Heterogeneous Multi-Head Attention mechanism. Unlike standard multi-head attention, where all heads share the same dimensionality, we configure three specialized attention heads whose dimensionalities correspond directly to the number of categories in the three taxonomic facets: $d_1 = |\mathcal{C}_P| = 6$, $d_2 = |\mathcal{C}_E| = 5$, and $d_3 = |\mathcal{C}_N| = 6$.

Formally, given the learnable taxonomic prompt sequence $\tilde{\mathbf{t}}_i$ and the transformed token feature sequence $\tilde{\mathbf{S}}_i$, we project them into the query, key, and value spaces of the k-th taxonomic head: $\mathbf{Q}_k \in \mathbb{R}^{(1+N_{ppt}) \times d_k}$, $\mathbf{K}_k \in \mathbb{R}^{L_i \times d_k}$, and $\mathbf{V}_k \in \mathbb{R}^{L_i \times d_k}$, where $d_k \in \{|\mathcal{C}_P|, |\mathcal{C}_E|, |\mathcal{C}_N|\}$. The heterogeneous attention output is computed as:

$$\begin{aligned} \mathbf{H}_{i,k} &= \text{softmax}\left(\frac{\mathbf{Q}_k \mathbf{K}_k^\top}{\sqrt{d_k}}\right)\mathbf{V}_k, \\ \mathbf{H}_i^{\text{het}} &= \text{Concat}\left(\mathbf{H}_{i,1}, \mathbf{H}_{i,2}, \mathbf{H}_{i,3}\right), \end{aligned} \tag{12}$$

where $\mathbf{H}_{i,k}$ denotes the output of the k-th attention head for instance i. The concatenated representation $\mathbf{H}_i^{\text{het}} \in \mathbb{R}^{(1+N_{ppt}) \times 17}$ combines the three facet-specific outputs along the feature dimension, thereby preserving alignment with the unified 17-dimensional taxonomic space.

This design ensures that each attention head is dedicated to a single taxonomic dimension, preserving rigorous alignment with persuasion strategies, emotional manipulation, and narrative roles.

The heterogeneous multi-head attention module comprises a stack of N_{att} layers. Within each layer, the input representation first undergoes heterogeneous multi-head attention, followed by a residual connection and layer normalization.

Prediction. The classification head maps the final attention output $\mathbf{h}_i^{(N_{\text{att}})} \in \mathbb{R}^{(N_{\text{ppt}}+1) \times 17}$ into the target prediction space. Specifically, a Weighted Sum Queries layer first aggregates the representation across the query dimension to produce a unified 17-dimensional taxonomic vector, which is subsequently projected by a compact MLP layer to

Method	Overall		Post		Article	
	MacroF1	MacroRecall	MacroF1	MacroRecall	MacroF1	MacroRecall
RAEmoLLM	$\underline{0.8255}_{\pm.002}$	$0.7981_{\pm.002}$	$0.8625_{\pm.001}$	$0.8528_{\pm.001}$	$0.7978_{\pm.002}$	$0.7704_{\pm.002}$
FACTUAL	$0.7563_{\pm.004}$	$0.7377_{\pm.004}$	$0.6428_{\pm.006}$	$0.6474_{\pm.005}$	$\underline{0.8027}_{\pm.004}$	$0.7780_{\pm.004}$
DYNAMO	$0.6680_{\pm.007}$	$0.6972_{\pm.009}$	$0.5737_{\pm.020}$	$0.6200_{\pm.019}$	$0.6936_{\pm.002}$	$0.7177_{\pm.015}$
PCoT	$0.8218_{\pm.020}$	$0.8053_{\pm.019}$	$0.7675_{\pm.007}$	$0.7606_{\pm.007}$	$\mathbf{0.8439}_{\pm.027}$	$\mathbf{0.8226}_{\pm.025}$
RoBERTa-large	$0.8143_{\pm.028}$	$\underline{0.8085}_{\pm.038}$	$0.9270_{\pm.016}$	$0.9300_{\pm.014}$	$0.7583_{\pm.042}$	$0.7517_{\pm.053}$
DRES	$0.7294_{\pm.001}$	$0.7170_{\pm.001}$	$\underline{0.9454}_{\pm.002}$	$\underline{0.9418}_{\pm.002}$	$0.6194_{\pm.002}$	$0.6145_{\pm.002}$
MSynFD	$0.7337_{\pm.026}$	$0.7206_{\pm.026}$	$0.8887_{\pm.006}$	$0.8924_{\pm.003}$	$0.6375_{\pm.051}$	$0.6356_{\pm.039}$
ExTax (ours)	$\mathbf{0.8456}_{\pm.011}$	$\mathbf{0.8423}_{\pm.015}$	$\mathbf{0.9548}_{\pm.003}$	$\mathbf{0.9565}_{\pm.004}$	$0.7965_{\pm.018}$	$\underline{0.7913}_{\pm.022}$

Table 1: Performance comparison between baseline models and our framework. The methods above the horizontal dividing line represent LLM-based approaches, while those below are deep learning-based methods. The best and second-best results are highlighted in **bold** and underlined, respectively. We report the mean and standard deviation of the Macro F_1-score and Macro recall over all random seeds.

generate the final classification:

$$\mathbf{h}_i^{(sum)} = \sum_{j=1}^{N_{\text{ppt}}+1} \beta_j \cdot \mathbf{h}_{i,j}^{(N_{\text{att}})}, \tag{13}$$

$$\hat{\boldsymbol{y}}_i = \text{Linear}\big(\text{LN}(\text{GELU}(\text{Linear}(\mathbf{h}_i^{(sum)})))\big), \tag{14}$$

where β_j denotes the learnable query weights, and $\hat{\mathbf{y}}$ represents the final prediction vector.

5 Experiments

5.1 Setup

Dataset. We adopt the five datasets used in PCoT (Modzelewski et al., 2025): **CoAID** (Cui and Lee, 2020) and **ECTF** (Paka et al., 2021; Bansal et al., 2021) cover COVID-19 health disinformation in articles and tweets, respectively, and provide train/-validation/test splits; **ISOT Fake News** (Ahmed et al., 2018, 2017), **MultiDis** (Modzelewski et al., 2025), and **EUDisinfo** (Modzelewski et al., 2025) are test-only article corpora spanning mainstream news, multi-topic European political/social disinformation, and pro-Kremlin geopolitical narratives. Full specifications are provided in Appendix D.1.

Baseline. We compare against seven representative baselines, including three deep learning-based methods, **RoBERTa** (Liu et al., 2019), **DRES** (Farhangian et al., 2025), and **MSynFD** (Xiao et al., 2024), and four LLM-based methods, **RAEmoLLM** (Liu et al., 2025), **FACTUAL** (Li et al., 2025), **DYNAMO** (Jin et al., 2025), and **PCoT** (Modzelewski et al., 2025). Their introductions and implementation details are provided in Appendix D.2 and Appendix D.5, respectively.

Metrics. To quantitatively evaluate our framework, we employ two standard metrics: **Accuracy**, **Macro Recall** and **Macro F_1-score**. For a more detailed definition, please refer to Appendix D.

Implementation. ExTax uses a frozen `roberta-large` backbone with parameter-efficient trainable components (≈267K parameters), optimized in a two-phase architecture across three random seeds. All deep learning baselines and ExTax are trained and validated on the CoAID and ECTF training/validation partitions, then evaluated across all five test sets. Detailed implementation and hyperparameter configurations are provided in Appendix D.4.

5.2 Main Results

We categorize all five datasets into *Post* and *Article* genres based on text typology. Table 1 reports the main results, from which we draw three key observations. (i) **Best Overall performance. ExTax** achieves the highest Overall MacroF1 (**0.8456**) and MacroRecall (**0.8423**), improving by +2.01 and +4.42 points over the strongest LLM-based RAG baseline (`RAEmoLLM`), and by +3.13 and +3.38 points over the strongest deep-learning baseline (`RoBERTa-large`). As noted by Modzelewski et al. (2025), prompting-based baselines may further benefit from potential pre-training exposure, an advantage that ExTax does not require. (ii) **Genre trade-off rather than uniform dominance.** The Overall margin is mainly driven by short-form *Posts*, where ExTax reaches **0.9548** MacroF1 and outperforms `PCoT` by +18.7 points. On long-form *Articles*, ExTax is the runner-up with **0.7965** MacroF1, behind `PCoT` (**0.8439**), which may be attributable to the 30% *Article* minority in training and potential

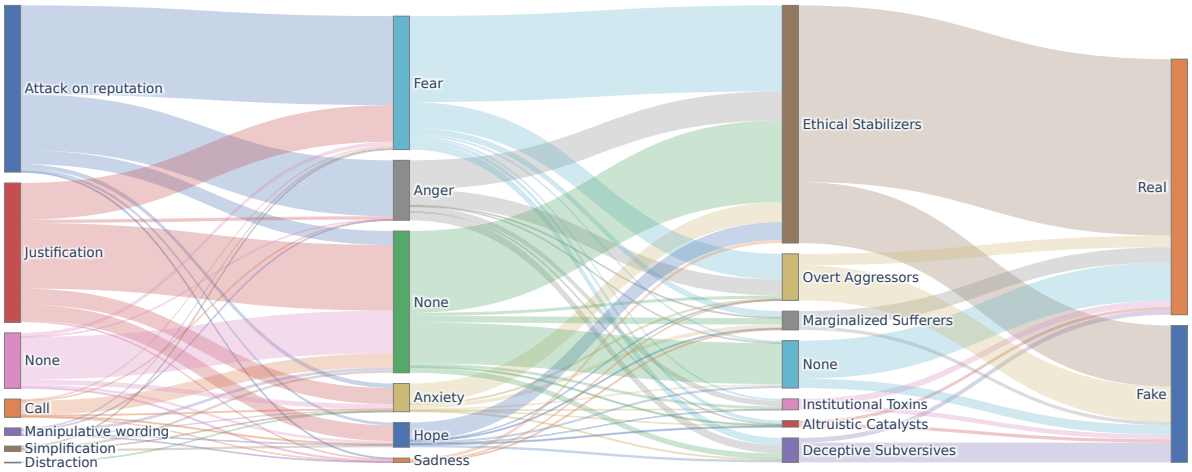

Figure 2: Sankey diagram illustrating the co-occurrence patterns across the three taxonomic dimensions (Persuasion, Emotion, and Narrative Roles) and the final ground-truth labels. The flows reveal which manipulative attributes tend to appear together. "None" denotes instances where the LLMs did not detect any of the predefined categories within a specific taxonomic facet.

pre-training exposure favoring article-style text for prompting baselines. Crucially, PCoT pays for this *Article* advantage by collapsing on *Posts* (**0.7675**), whereas ExTax does not exhibit a genre-collapse failure mode. (iii) **Robustness under severe genre imbalance.** Despite the 70/30 genre imbalance, ExTax generalizes consistently across both genres, while conventional deep-learning baselines exhibit substantial performance degradation. For example, DRES's MacroF1 drops from **0.9454** on *Posts* to **0.6194** on *Articles* (-32.6 points), suggesting that taxonomy-aligned features capture more genre-invariant manipulation patterns, whereas surface-pattern detectors may overfit to genre-specific artifacts.

5.3 Ablation

As shown in Table 2, our hybrid Taxonomic Attribute Elicitation framework does not achieve the best Stage-1 performance in taxonomy-aligned representation learning, but yields the strongest downstream Macro F_1 in Stage 2. This is by design: Stage 1 favors hard-label fidelity to synthesized annotations, whereas entropy-driven dynamic label smoothing deliberately trades such fidelity for calibrated uncertainty. When multiple LLMs disagree, supervision is softened to prevent overfitting to uncertain attributes. As a result, Stage 2 receives better-calibrated taxonomic supervision, improving robustness against synthesized attribute noise and leading to stronger disinformation detection.

Table 3 details the incremental contributions of

LLM	**stage 1**	**stage 2**
	Avg. F1	**Overall**
GPT-5 mini	**0.5026**	0.8260
DeepSeek-V4-Flash	0.3234	0.8216
Grok 4.1 Fast Reasoning	0.2571	0.8384
Qwen3.6 Flash	0.4286	0.8387
ours (hybrid)	0.4489	**0.8421**

Table 2: Performance comparison (Macro F_1) of our hybrid approach (Divergent Attribute Synthesis and Entropy-driven Dynamic Label Smoothing) against single LLM attribute synthesis across Stage 1 (Taxonomy-aligned Representation Learning) and Stage 2 (Taxonomy-based Disinformation Detection). Stage 1 reports the average Macro F_1 across the three taxonomic dimensions.

ExTax's core components over the baseline (0.7789 macF1). Introducing heterogeneous multi-head attention effectively integrates taxonomic representations with contextual text features, improving performance to 0.7898 macF1. Adding adaptive gated pooling dynamically captures key semantic tokens, boosting fake news detection accuracy (Acc_{fake}) to 0.6789. Further integration of the hybrid module (divergent attribute synthesis and entropy-driven label smoothing) filters synthetic feature noise, increasing the macF1 score to 0.8137.

Regarding prompt scaling, although $N_{\text{ppt}} = 2$ achieves the highest overall macF1 (0.8644), it suffers from severe classification bias toward the real

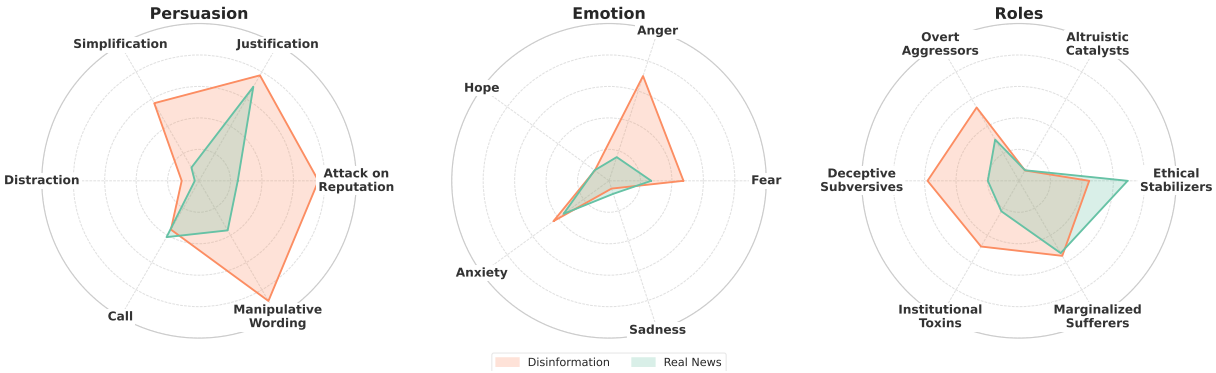

Figure 3: Distribution of persuasion, emotion, and narrative role taxonomic attributes across all datasets, where distributions for disinformation and real news are highlighted in red and green, respectively.

Conf.	Acc_{fake}	Acc_{real}	MacroF1
Baseline	0.5952	0.9313	0.7789
Hetero. Attention	0.6041	0.9415	0.7898
Ada. Gated Pooling	0.6789	0.8973	0.7966
hybrid	0.7132	0.9001	0.8137
$N_{\text{ppt}} = 1$	0.5165	**0.9551**	0.7541
$N_{\text{ppt}} = 2$	0.7512	0.9517	**0.8644**
$N_{\text{ppt}} = 3$(Full)	**0.8058**	0.8823	0.8421

Table 3: Ablation study of ExTax components. The baseline configuration utilizes RoBERTa paired with an MLP. Random seed is 445. Components are incrementally integrated from top to bottom until the full ExTax framework is achieved.

label ($\text{Acc}_{\text{real}} = 0.9517$ vs. $\text{Acc}_{\text{fake}} = 0.7512$). Consequently, ExTax adopts $N_{\text{ppt}} = 3$ to achieve more balanced diagnostic performance between fake ($\text{Acc}_{\text{fake}} = 0.8058$) and real ($\text{Acc}_{\text{real}} = 0.8823$) news, while maintaining a robust and competitive 0.8421 macF1 across varying seeds.

5.4 Interpretability

Figure 3 visualizes the global distribution of ExTax's three taxonomic dimensions across all datasets. Disinformation and real news exhibit clearly distinct patterns within the 17-dimensional taxonomic space: fake news is more strongly associated with "Attack on Reputation", "Simplification", and "Manipulative Wording", high-arousal emotions such as "Anger" and "Fear", and negative narrative roles such as "Deceptive Subversives" and "Institutional Toxins". These observations are consistent with prior findings on persuasion and emotional manipulation in disinformation (Modzelewski et al., 2025; Jamieson et al., 2025; Zhao et al., 2024).

Figure 2 further illustrates the co-occurrence structure among persuasion, emotion, narrative roles, and final labels. Aggressive persuasion frequently co-occurs with negative emotions, whereas non-manipulative attributes tend to propagate across dimensions and are more strongly associated with real news. These results suggest that ExTax captures interpretable manipulative patterns underlying deceptive texts. The full interpretability analysis and detailed case studies are provided in Appendix D.6.

6 Conclusion

In this paper, we introduced **ExTax**, an explainable and taxonomy-aligned framework for disinformation detection that models three complementary cognitive facets of manipulation—persuasive rhetoric, emotional manipulation, and narrative roles—within a unified 17-dimensional taxonomic space. ExTax combines Divergent Attribute Elicitation, Entropy-driven Dynamic Label Smoothing, and Heterogeneous Multi-Head Attention to produce noise-robust and structurally aligned representations of manipulative intent. Across five cross-domain, cross-genre datasets, ExTax outperforms strong LLM- and deep-learning baselines in Overall Macro F_1 and remains robust under severe genre imbalance. More importantly, by exposing per-sample 17-dimensional manipulation profiles, ExTax moves disinformation detection beyond binary fake/real verdicts toward an interpretable and auditable diagnosis of how deceptive content manipulates readers, thereby providing a more transparent foundation for downstream fact-checking and content-moderation workflows.

Limitations

Several further limitations are worth noting. (1) **Reliance on commercial frontier LLMs.** The divergent attribute elicitation stage relies on four commercial frontier LLMs, which incurs non-trivial inference cost and raises reproducibility concerns as those APIs evolve. Although entropy-driven smoothing mitigates inconsistency across models, it cannot correct systematic biases shared by all of them. (2) **English- and Western-centric corpora.** Our five evaluation datasets are predominantly English and focused on Western political and health discourse; thus, we do not validate ExTax on multilingual or non-Western disinformation, where persuasion conventions and narrative archetypes may differ. (3) **Explanations are not yet evaluated with humans.** While ExTax outputs structured per-sample 17-dimensional manipulation profiles, we have not conducted a user study with fact-checkers or moderators to verify whether these profiles are more useful in practice than opaque verdicts. We leave controlled human evaluation to future work.

Although ExTax is designed to support explainable disinformation detection, it may still produce false positives or false negatives, especially on long-form articles or domains underrepresented in the training data. False positives could unfairly flag legitimate content as disinformation, while false negatives could allow harmful misleading content to remain undetected. In addition, the LLM-elicited taxonomic attributes may reflect biases or inconsistencies from the underlying LLMs, even though our entropy-driven smoothing mechanism is designed to mitigate synthetic labeling noise. Therefore, ExTax should be used as a decision-support tool rather than as an automated content moderation system, and its outputs should be interpreted with human oversight.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *SECURITY AND PRIVACY*, 1(1):e9.

Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. 2021. Fake Detect: A Deep Learning Ensemble Model for Fake News Detection. *Complexity*, 2021:5557784.

Bence Bago, Leah R. Rosenzweig, Adam J. Berinsky, and David G. Rand. 2022. Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion*, 36(6):1166–1180.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Rachit Bansal, William Scott Paka, Nidhi, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of covid-19 fake tweets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv*.

Faramarz Farhangian, Leandro Augusto Ensina, George D. C. Cavalcanti, and Rafael M. O. Cruz. 2025. Dres: Fake news detection by dynamic representation and ensemble selection. In *Conference on Empirical Methods in Natural Language Processing*.

Seunghee Han, Jennifer S. Lerner, and Dacher Keltner. 2007. Feelings and Consumer Decision Making: The Appraisal-Tendency Framework. *Journal of Consumer Psychology*, 17(3):158–168.

Jack Jamieson, Toru Hara, and Mitsuaki Akiyama. 2025. Flagging Emotional Manipulation: Impacts of Manipulative Content Warnings on Sharing Intentions and Perceptions of Health-Related Social Media Posts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Di Jin, Jun Yang, Xiaobao Wang, Junwei Zhang, Shuqi Li, and Dongxiao He. 2025. A Dynamic Knowledge Update-Driven Model with Large Language Models for Fake News Detection. In *Thirty-Fourth International Joint Conference on Artificial Intelligence*, volume 4, pages 3000–3008.

Andreas Jungherr and Adrian Rauchfleisch. 2024. Negative downstream effects of alarmist disinformation discourse: Evidence from the united states. *Political Behavior*, 46:2123 – 2143.

Danial Kamali, Joseph D. Romain, Huiyi Liu, Wei Peng, Jingbo Meng, and Parisa Kordjamshidi. 2024. Using Persuasive Writing Strategies to Explain and Detect Health Misinformation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17285–17309, Torino, Italia. ELRA and ICCL.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2025. Mitigating Biases of Large Language Models in Stance Detection with Counterfactual Augmented Calibration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7075–7092, Albuquerque, New Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. 2025. RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning Based on Emotional Information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16508–16523, Vienna, Austria. Association for Computational Linguistics.

Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. *Preprint*, arXiv:2601.21963.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Tarek Mahmoud, Zhuohan Xie, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025. Entity framing and role portrayal in the news. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 302–326, Vienna, Austria. Association for Computational Linguistics.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393 – 107393.

Susannah B. F. Paletz, Michael A. Johns, Egle E. Murauskaite, Ewa M. Golonka, Nick B. Pandža, C. Anton Rytting, Cody Buntain, and Devin Ellis. 2023. Emotional content and sharing on Facebook: A theory cage match. *Science Advances*, 9(39):eade9231.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. Multilingual Multifaceted Understanding of Online News in Terms of Genre, Framing, and Persuasion Techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Zhengjia Wang, Qiang Sheng, Danding Wang, Beizhe Hu, and Juan Cao. 2025. Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.

Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. Msynfd: Multi-hop syntax aware fake news detection. *Proceedings of the ACM Web Conference 2024*.

Xinyan Zhao, Jessica Shaw, and Zexin Ma. 2024. How individuals cope with anger- and sadness-induced narrative misinformation on social media: Roles of

transportation and correction. *Online Media and Global Communication*, 3(3):418–446.

Appendix

A Taxonomies

This appendix provides the detailed definitions of the three taxonomic dimensions used in ExTax: Persuasion, Emotion, and Narrative Roles. These taxonomies are used during the Taxonomic Attribute Elicitation stage to guide LLMs in extracting interpretable textual attributes from each input text. The definitions are derived from the prompt configuration used for attribute elicitation, while prompt-formatting instructions and JSON output templates are reported separately in Appendix B.

A.1 Persuasion

Persuasion refers to the use of language intended to influence readers. We distinguish six high-level persuasion approaches: Attack on Reputation, Justification, Simplification, Distraction, Call, and Manipulative Wording.

A.1.1 Attack on Reputation

Attack on Reputation refers to arguments that do not directly address the topic itself, but instead target a participant, group, organization, object, or activity in order to question or undermine its credibility.

- **Name Calling or Labelling:** The use of loaded labels directed at an individual, group, object, or activity, typically in an insulting or demeaning way, although sometimes using labels that the target audience may find desirable.
- **Guilt by Association:** Attacking an opponent, activity, or concept by associating it with another group, activity, or concept that carries strong negative connotations for the target audience.
- **Casting Doubt:** Questioning the character or personal attributes of a person or entity in order to undermine its general credibility or perceived quality.
- **Appeal to Hypocrisy:** Attacking a target's reputation by accusing it of hypocrisy or inconsistency.
- **Questioning the Reputation:** Making strong negative claims about a target, with emphasis on undermining its character or moral standing rather than addressing the substantive topic.

A.1.2 Justification

Justification refers to arguments composed of a statement and an accompanying explanation or appeal, where the latter is used to justify or support the former.

- **Flag Waving:** Justifying an idea by invoking pride in a group or highlighting benefits for that group.
- **Appeal to Authority:** Giving weight to an argument, idea, or piece of information by presenting an entity regarded as authoritative as its source.
- **Appeal to Popularity:** Justifying an argument or idea by claiming that everyone, or the large majority, agrees with it, or that nobody disagrees with it.
- **Appeal to Values:** Supporting an idea by linking it to values that are viewed positively by the target audience.
- **Appeal to Fear or Prejudice:** Promoting or rejecting an idea by invoking fear, prejudice, or repulsion toward that idea.

A.1.3 Simplification

Simplification refers to arguments that excessively simplify a complex issue, usually by reducing its causes, consequences, or available choices.

- **Causal Oversimplification:** Assuming a single cause or reason for an issue when multiple causes are involved.
- **False Dilemma or No Choice:** Presenting only two options or sides when more alternatives exist, or explicitly instructing the audience to take one action while excluding other possible choices.
- **Consequential Oversimplification:** Claiming that an initial event or action will lead to a domino-like chain of increasingly serious consequences, where the causal chain is exaggerated, improbable, or insufficiently warranted.

A.1.4 Distraction

Distraction refers to arguments that divert attention away from the main issue or argument.

- **Strawman:** Giving the impression of refuting an opponent's argument while replacing the actual argument with a distorted or weaker version.
- **Red Herring:** Diverting the audience's attention from the main topic by introducing an irrelevant issue.
- **Whataboutism:** Attempting to discredit an opponent's position by accusing the opponent of hypocrisy without directly addressing or disproving the original argument.

A.1.5 Call

Call refers to language that is not primarily argumentative, but instead encourages the audience to act or think in a particular way.

- **Slogans:** Brief and striking phrases, often functioning as emotional appeals and potentially involving labelling or stereotyping.
- **Conversation Killer:** Words or phrases that discourage critical thinking or meaningful discussion about a given topic.
- **Appeal to Time:** Arguments centered on the claim that the time has come for a particular action.

A.1.6 Manipulative Wording

Manipulative Wording refers to language that is not necessarily an argument by itself, but uses non-neutral, confusing, exaggerated, or loaded expressions to influence the reader emotionally.

- **Loaded Language:** The use of words or phrases with strong emotional implications, either positive or negative, to influence the audience.
- **Obfuscation, Intentional Vagueness, or Confusion:** The use of deliberately unclear, vague, or ambiguous language that leaves room for audience-specific interpretations.
- **Exaggeration or Minimisation:** Representing something in an excessive manner or making it appear less important or smaller than it actually is.
- **Repetition:** Repeated use of the same phrase or expression in order to increase persuasive impact.

A.2 Emotion

Emotion refers to language designed to bypass purely rational processing and influence readers by triggering affective responses. We distinguish five high-level emotional manipulation types: Fear, Anger, Hope, Anxiety, and Sadness.

- **Fear:** Language designed to make readers feel scared or threatened, often by emphasizing immediate risks, serious health concerns, or distressing outcomes in order to provoke a protective reaction.
- **Anger:** Language designed to provoke outrage, resentment, or a sense of injustice, often by targeting specific groups, institutions, or actors.
- **Hope:** Language designed to elicit positive expectations or promise breakthroughs, often by presenting unverified cures, revolutionary alternatives, or highly optimistic outcomes.
- **Anxiety:** Language designed to create uncertainty about future safety or well-being, often by describing disturbing cases or hidden dangers in everyday practices.
- **Sadness:** Language focusing on irreversible loss, heartbreaking misfortune, or the suffering of helpless victims, thereby triggering sympathy, sorrow, or moral guilt.

A.3 Narrative Roles

Narrative Roles describe how a person, group, or entity is functionally positioned within a narrative, independently of its actual moral status. We distinguish six high-level role categories: Ethical Stabilizers, Altruistic Catalysts, Overt Aggressors, Deceptive Subversives, Institutional Toxins, and Marginalized Sufferers.

A.3.1 Ethical Stabilizers

Ethical Stabilizers are actors portrayed as using established values, moral authority, or mediation to maintain social cohesion, safety, and justice.

- **Guardian:** Actors portrayed as protectors of values or communities, such as law enforcement officers, soldiers, or community leaders.
- **Peacemaker:** Actors portrayed as seeking harmony, conflict resolution, diplomacy, negotiation, or mediation.
- **Virtuous:** Actors portrayed as righteous, noble, fair, just, or morally exemplary.

A.3.2 Altruistic Catalysts

Altruistic Catalysts are actors portrayed as challenging the status quo or enduring hardship to drive systemic change, liberation, or collective benefit.

- **Rebel:** Actors portrayed as challenging the status quo and fighting for major change, freedom, or liberation from oppression.
- **Martyr:** Actors portrayed as sacrificing their well-being, or even their lives, for a greater good or cause.
- **Underdog:** Actors portrayed as disadvantaged or unlikely to succeed, but striving against stronger forces or obstacles.

A.3.3 Overt Aggressors

Overt Aggressors are visible antagonists portrayed as using direct force, systemic power, or violent ideology to dominate, divide, or harm others.

- **Instigator:** Actors portrayed as initiating conflict, tension, unrest, or violence.
- **Tyrant:** Actors portrayed as abusing power, ruling unjustly, or oppressing those under their control.
- **Terrorist:** Actors portrayed as engaging in violence or terror to advance ideological goals, often by targeting civilians.
- **Bigot:** Actors portrayed as hostile or discriminatory toward specific groups, including racism, sexism, homophobia, antisemitism, Islamophobia, or other forms of hate speech.

A.3.4 Deceptive Subversives

Deceptive Subversives are hidden or manipulative actors portrayed as undermining trust, truth, or safety through secrecy, betrayal, deception, or strategic manipulation.

- **Conspirator:** Actors portrayed as participating in secret plots or covert plans to deceive or undermine others.
- **Spy:** Actors portrayed as gathering or transmitting sensitive information to a rival or enemy through secrecy or deception.
- **Traitor:** Actors portrayed as betraying a cause, community, or country.
- **Deceiver:** Actors portrayed as manipulating truth, spreading misinformation, or shaping public perception for strategic or self-interested purposes.

- **Saboteur:** Actors portrayed as deliberately damaging, obstructing, or weakening systems, organizations, or processes.
- **Foreign Adversary:** External actors portrayed as creating geopolitical tension or acting against another country's interests.

A.3.5 Institutional Toxins

Institutional Toxins are actors or entities portrayed as internal sources of organizational or systemic failure, caused by vice, incapacity, corruption, or incompetence.

- **Corrupt:** Actors portrayed as engaging in unethical or illegal behavior for personal, political, or financial gain.
- **Incompetent:** Actors portrayed as causing harm through ignorance, poor judgment, lack of skill, or inability to perform their duties.

A.3.6 Marginalized Sufferers

Marginalized Sufferers are actors portrayed as enduring harm, exploitation, neglect, exclusion, or unjust blame, often functioning as evidence of systemic failure or as targets for rescue.

- **Forgotten:** Marginalized or overlooked groups portrayed as ignored by society and lacking needed attention or support.
- **Exploited:** Individuals or groups portrayed as being used for others' benefit without adequate consent and with harm to their well-being.
- **Victim:** Actors portrayed as suffering physical or economic harm due to circumstances beyond their control.
- **Scapegoat:** Actors portrayed as being blamed unjustly for problems or failures in order to divert attention from the real causes or responsible parties.

B Prompts

The following section delineates the construction of prompts designed to guide LLMs in extracting taxonomic features across Persuasion, Emotion, and Narrative Roles. Specifically, the prompt configuration for eliciting Persuasion attributes is adopted from Modzelewski et al. (2025). Uppercase placeholders wrapped in hashtags, such as #PERSUASION_DEFINITIONS#, #EMOTION_DEFINITIONS#, and #NARRATIVE_ROLE_DEFINITIONS#, are used as designators to denote the comprehensive taxonomic definitions detailed in Appendix A.

B.1 Prompt 1: Extract Persuasion Taxonomic Features

System Prompt

```
You are an assistant who detects persuasion
     in text. Persuasive text is
    characterized by a specific use of
    language in order to influence readers.
     We distinguish the following high-
    level persuasion approaches:
    #PERSUASION_DEFINITIONS#

You are the expert who detects high-level
    persuasion approaches: Attack on
    Reputation, Justification,
    Simplification, Distraction, Call,
    Manipulative Wording.
```

Context Prompt

```
Analyze the text and decide if the text
    contains any high-level persuasion
    approaches from the following: Attack
    on Reputation, Justification,
    Simplification, Distraction, Call,
    Manipulative Wording. For each high-
    level persuasion approach, provide an
    explanation of how it appears in the
    analyzed text. Be conservative in your
    final decisions; when you are not fully
     sure, answer No. Give your answer in
    the form of a dictionary:
{{
"Attack_on_reputation": {{"is_used": "Your
    answer. Use only Yes or No", "
    explanation": "If high-level persuasion
     Attack on Reputation appears, provide
    an explanation here."}},
"Justification": {{"is_used": "Your answer.
     Use only Yes or No", "explanation": "
    If high-level persuasion Justification
    appears, provide an explanation here
    ."}},
"Simplification": {{"is_used": "Your answer.
     Use only Yes or No", "explanation": "
    If high-level persuasion Simplification
     appears, provide an explanation here
    ."}},
"Distraction": {{"is_used": "Your answer.
    Use only Yes or No", "explanation": "If
     high-level persuasion Distraction
    appears, provide an explanation here
    ."}},
"Call": {{"is_used": "Your answer. Use only
     Yes or No", "explanation": "If high-
    level persuasion Call appears, provide
    an explanation here."}},
```

```
"Manipulative_wording": {{"is_used": "Your
     answer. Use only Yes or No", "
     explanation": "If high-level persuasion
      Manipulative Wording appears, provide
     an explanation here."}}
}}
Text: {text}
Answer:
```

B.2 Prompt 2: Extract Emotion Taxonomic Features

System Prompt

```
You are a specialized linguistic analyst
     and expert in detecting emotional
     manipulation in online text. Your goal
     is to identify techniques designed to
     bypass rational thinking and influence
     readers by triggering specific
     emotional responses.

We distinguish between the following five
     high-level emotional appeals:
    #EMOTION_DEFINITIONS#

You are the expert who detects emotional
     manipulation types: Fear, Anger, Hope,
     Anxiety, Sadness.
```

Context Prompt

```
Analyze the text and decide if the text
     contains any emotional manipulation
     types from the following: Fear, Anger,
     Hope, Anxiety, Sadness. For each
     emotional manipulation type, provide an
      explanation of how it appears in the
     analyzed text. Be conservative in your
     final decisions; when you are not fully
      sure, answer No.

Give your answer in the form of a
     dictionary:
{{
"Fear": {{"is_used": "Your answer. Use only
      Yes or No", "explanation": "If the
     Fear type appears, provide a one-
     sentence explanation starting with '
     This post contains language that may be
      intended to make you feel fear by
     ...'"}},
"Anger": {{"is_used": "Your answer. Use
     only Yes or No", "explanation": "If the
      Anger type appears, provide a one-
     sentence explanation starting with '
     This post contains language that may be
      intended to make you feel anger by
     ...'"}},
"Hope": {{"is_used": "Your answer. Use only
      Yes or No", "explanation": "If the
     Hope type appears, provide a one-
     sentence explanation starting with '
     This post contains language that may be
      intended to make you feel hope by
     ...'"}},
"Anxiety": {{"is_used": "Your answer. Use
     only Yes or No", "explanation": "If the
      Anxiety type appears, provide a one-
     sentence explanation starting with '
     This post contains language that may be
      intended to make you feel anxiety by
     ...'"}},
"Sadness": {{"is_used": "Your answer. Use
     only Yes or No", "explanation": "If the
      Sadness type appears, provide a one-
     sentence explanation starting with '
     This post contains language that may be
      intended to make you feel sadness by
     ...'"}}
}}
Text: {text}
Answer:
```

B.3 Prompt 3: Extract Narrative Role Taxonomic Features

System Prompt

```
You are an expert at identifying narrative
     framing and role portrayal in text.
     Your goal is to detect whether a text
     utilizes specific high-level narrative
     categories -- Ethical Stabilizers,
     Altruistic Catalysts, Overt Aggressors,
      Deceptive Subversives, Institutional
     Toxins, or Marginalized Sufferers -- to
      frame the participants in a story.

This analysis focuses on functional roles
     -- how a person, group, or entity is
     positioned within the narrative --
     independent of their actual moral
     standing.

You distinguish between six high-level main
      roles based on the following
     archetypal markers:
    #NARRATIVE_ROLE_DEFINITIONS#

Your analysis must be based on the specific
      linguistic cues and narrative
     structure within the text.
```

Context Prompt

```
Analyze the text below to determine whether
      it contains any of the following high-
     level roles: Ethical Stabilizers,
     Altruistic Catalysts, Overt Aggressors,
      Deceptive Subversives, Institutional
     Toxins, or Marginalized Sufferers.

For each detected role, provide a concise
     explanation of how it is portrayed
     based on the narrative function and
     word choice. Be conservative in your
     final decisions; if you are not fully
```

```
    sure, answer No.

Return your answer as a JSON object in the
    following format:
{{
  "Ethical_Stabilizers": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      as a protector of values or a
     righteous figure (e.g., Guardian,
     Virtuous, Peacemaker)."
  }},
  "Altruistic_Catalysts": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      as driving positive change or
     sacrifice (e.g., Rebel, Martyr,
     Underdog)."
  }},
  "Overt_Aggressors": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      as an initiator of conflict or
     violence (e.g., Tyrant, Terrorist,
     Instigator, Bigot)."
  }},
  "Deceptive_Subversives": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      through secrecy, betrayal, or
     misinformation (e.g., Conspirator, Spy,
      Traitor, Deceiver, Saboteur, Foreign
     Adversary)."
  }},
  "Institutional_Toxins": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      as abusing power or as a source of
     systemic failure (e.g., Corrupt,
     Incompetent)."
  }},
  "Marginalized_Sufferers": {{
    "is_used": "Yes or No",
    "explanation": "If Yes, provide
     evidence of how a participant is framed
      as suffering harm, exploitation,
     neglect, exclusion, or unjust blame (e.
     g., Victim, Exploited, Forgotten,
     Scapegoat)."
  }}
}}

Text: {text}
Answer:
```

C Method

Details of entropy-driven dynamic label smoothing. For each taxonomic facet, the binary LLM outputs are first aggregated by category. In the Emotion task, for example, $e_{i,k}^{(c)} \in \{0, 1\}$ indicates whether the k-th LLM assigns sample T_i to emotion category $c \in \mathcal{C}_E$. The positive vote proportion $p_{i,c}$ measures the fraction of LLMs predicting the presence of category c. We compute binary entropy over this Bernoulli distribution to quantify model disagreement. The same computation is applied to the Persuasion and Narrative Role facets by replacing $e_{i,k}^{(c)}$ with the corresponding components of $\mathbf{p}_{i,k}$ and $\mathbf{n}_{i,k}$. The resulting smoothed labels across all samples and categories form:

$$\mathcal{Y}_{\text{smooth}} = \Big\{ y_{\text{tax}}^{(i,c)} \;\Big|\; i = 1, \ldots, n, \quad c \in \mathcal{C}_P \cup \mathcal{C}_E \cup \mathcal{C}_N \Big\}. \tag{15}$$

When all LLMs agree, $H_{i,c} = 0$ and the smoothed label equals the vote proportion $p_{i,c}$. When the models disagree maximally, $H_{i,c}$ is high and the target is shifted toward 0.5, reducing the effect of uncertain synthetic supervision.

Details of adaptive gated pooling. Given the LLM hidden states $\mathbf{S}_i \in \mathbb{R}^{L_i \times D}$, Adaptive Gated Pooling uses a learnable parameter vector $\mathbf{w} \in \mathbb{R}^{L_{\max}}$, where $L_{\max}$ is the maximum truncation length. For an input with length L_i, we slice the first L_i elements of $\mathbf{w}$ as $\mathbf{w}_{1:L_i}$ and compute the normalized gating weights:

$$\mathbf{g}_i = \text{softmax}(\mathbf{w}_{1:L_i}) \in \mathbb{R}^{L_i}, \tag{16}$$

$$g_{i,h} = \frac{\exp(w_h)}{\sum_{j=1}^{L_i} \exp(w_j)}. \tag{17}$$

The pooled representation is then obtained by a weighted sum:

$$\mathbf{v}_i = \sum_{h=1}^{L_i} g_{i,h} \mathbf{S}_{i,h}. \tag{18}$$

All elements of $\mathbf{w}$ are initialized to 1, making the initial pooling behavior equivalent to average pooling. During training, the gating vector is optimized to assign higher weights to more informative token positions.

Details of taxonomy-aligned representation learning. Each taxonomic facet is modeled using an independent MLP head:

$$\begin{aligned} \mathbf{h}_i^{(c)} &= \text{LN}\big(\text{GELU}(\text{Linear}_1^{(c)}(\mathbf{v}_i))\big), \\ f_c(\mathbf{v}_i) &= \text{Linear}_2^{(c)}\big(\text{Dropout}(\mathbf{h}_i^{(c)})\big), \end{aligned} \tag{19}$$

where $\text{Linear}_1^{(c)} : \mathbb{R}^D \to \mathbb{R}^{D_h}$ and $\text{Linear}_2^{(c)} : \mathbb{R}^{D_h} \to \mathbb{R}^{|\mathcal{C}_c|}$. The outputs are passed through sigmoid activations to obtain multi-label probabilities.

For each facet, we optimize Binary Cross-Entropy (BCE) loss against the smoothed taxonomic labels. Taking Persuasion as an example:

$$\mathcal{L}_P = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c\in\mathcal{C}_P}\Big[y_{\text{tax}}^{(i,c)}\log(\hat{p}_{i,c}) + (1-y_{\text{tax}}^{(i,c)})\log(1-\hat{p}_{i,c})\Big]. \tag{20}$$

The Emotion and Narrative Role losses are defined analogously.

Details of text feature transformation. The transformation module maps each token vector $\mathbf{S}_{i,h} \in \mathbb{R}^D$ into the 17-dimensional taxonomic space using a two-layer MLP:

$$\begin{aligned}\mathbf{h}_{\text{mid}} &= \text{Linear}_1\big(\text{LN}(\mathbf{S}_{i,h})\big),\\ \tilde{\mathbf{S}}_{i,h} &= \text{Linear}_2\big(\text{GELU}(\mathbf{h}_{\text{mid}})\big),\end{aligned} \tag{21}$$

where $\text{Linear}_1 : \mathbb{R}^D \to \mathbb{R}^{d_{\text{ff}}}$ and $\text{Linear}_2 : \mathbb{R}^{d_{\text{ff}}} \to \mathbb{R}^{17}$.

D Experiments

D.1 Dataset

- **CoAID** (Cui and Lee, 2020): A COVID-19 health disinformation dataset containing both news articles and social media posts, including over 4,000 news articles and more than 1,000 posts mainly from Twitter. The content covers pandemic-related topics such as public health policies, vaccines, treatment measures, and other healthcare claims. In our setting, it contains 5,011 samples across train, validation, and test splits, with a relatively stable genre distribution in each split: posts account for approximately 44%, while articles account for approximately 56%.

- **ISOT Fake News** (Ahmed et al., 2018, 2017): A large-scale fake news article dataset containing over 44,000 news articles collected from both reliable mainstream media, such as BBC and CNN, and unreliable or fake news websites. The articles span multiple domains, including politics, entertainment, technology, and health. Following the setting in PCoT, we use its test split with 499 samples, including 274 fake news instances and 225 real news instances.

- **ECTF** (Paka et al., 2021; Bansal et al., 2021): An extended version of the CTF (COVID-19 Tweet Fake) dataset for COVID-19 fake tweet detection. It consists of tweets annotated as fake or real and covers pandemic-related topics such as virus transmission, protective measures, public responses, and vaccine controversies. This dataset provides train, validation, and test splits. In our setting, it contains 4,152 samples, with 3,001, 751, and 400 samples in the train, validation, and test sets, respectively.

- **MultiDis** (Modzelewski et al., 2025): A researcher-constructed multi-topic disinformation dataset containing nearly 2,000 English news articles. The articles cover a broad range of political and social topics, including Anti-Europeanism, Anti-Atlanticism, Anti-migration, Xenophobia, Climate Change, Energy Crisis, Health, Institutional and Media Distrust, Gender Issues, Ukraine War, Refugees, and LGBT+. In our setting, this dataset is used as a test-only dataset and contains 499 test samples.

- **EUDisinfo** (Modzelewski et al., 2025): A disinformation dataset built upon the EUvsDisinfo database, focusing on pro-Kremlin propaganda narratives and political disinformation. The articles mainly address issues such as EU policies, the situation in Ukraine, NATO expansion, and related geopolitical narratives. In our setting, this dataset is used as a test-only dataset and contains 359 articles, including 241 credible articles and 118 disinformation articles.

D.2 Baseline Introduction

- **RoBERTa** (Liu et al., 2019): A BERT variant optimized via dynamic masking, larger batches, and expanded corpora without Next Sentence Prediction, serving as a competitive contextualized baseline.

- **DRES** (Farhangian et al., 2025): A framework routing instances to optimal feature spaces based on hardness measures to dynamically ensemble classifiers.

- **MSynFD** (Xiao et al., 2024): A syntax-aware framework capturing long-range dependency graphs and debiasing topical priors to detect fine-grained contextual distortions.

- **RAEmoLLM** (Liu et al., 2025): An emotion-based retrieval-augmented framework using a specialized LLM to dynamically select affective exemplars for cross-domain detection.
- **FACTUAL** (Li et al., 2025): A counterfactual augmented calibration framework disentangling robust features from domain shortcuts to dynamically calibrate model confidence scores.
- **DYNAMO** (Jin et al., 2025): A framework utilizing Monte Carlo Tree Search to verify factual paths and dynamically update knowledge graphs for real-time news detection.
- **PCoT** (Modzelewski et al., 2025): A zero-shot prompting method that first analyzes persuasive strategies in the input text and then uses this persuasion-aware reasoning to detect disinformation.

D.3 Metrics

To quantitatively evaluate our framework, we employ two standard metrics: **Accuracy** and **Macro F_1-score**.

Accuracy measures the global proportion of correctly classified instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{22}$$

where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

To account for potential class imbalances, we also use the **Macro F_1-score**. It calculates the F_1-score for each class independently and takes their unweighted arithmetic mean. For a specific class k, its precision (P_k), recall (R_k), and F_1-score ($F_{1,k}$) are defined as:

$$P_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{23}$$

$$R_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \tag{24}$$

$$F_{1,k} = \frac{2 \cdot P_k \cdot R_k}{P_k + R_k} \tag{25}$$

Given the binary class set $\mathcal{C} = \{0, 1\}$ (denoting reliable news and disinformation), the final Macro F_1-score is computed as:

$$\text{Macro } F_1 = \frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} F_{1,k} \tag{26}$$

Similarly, we report the **Macro Recall** to evaluate the model's ability to identify instances from each class regardless of class frequency. Macro Recall computes the recall for each class independently and then takes their unweighted arithmetic mean. Given the binary class set $\mathcal{C} = \{0, 1\}$, Macro Recall is defined as:

$$\text{Macro Recall} = \frac{1}{|\mathcal{C}|} \sum_{k \in \mathcal{C}} R_k \tag{27}$$

D.4 ExTax Implementation

During the training process, we perform an exhaustive hyperparameter search to ensure optimal convergence. To guarantee that our empirical findings are statistically significant and robust, we conduct multiple independent training and evaluation runs across three distinct random seeds: 43, 434 and 445. We subsequently report the mean and standard deviation of the core performance metrics. For the divergent attribute synthesis operation, we leverage a diverse suite of frontier models, including GPT-5-mini, DeepSeek-V4-flash, Qwen3.6-flash, and Grok-4.1-fast-reasoning. The label smoothing hyperparameter α is strictly grounded at 0.1896 to mitigate potential decision noise. All experiments and model training phases are computationally lightweight and can be fully executed on a single NVIDIA RTX 4090 (24 GB VRAM) GPU. For further details regarding the comprehensive training configurations and architectural model profiles, please refer to Table 4 and Table 5, respectively.

D.5 Baseline Implementation

Below, we report the detailed implementation of the reproduced baseline models. All baselines are evaluated under the same binary label space and dataset split protocol as ExTax. Specifically, CoAID and ECTF are used as source-domain datasets with training and validation splits, while CoAID, ECTF, EUDisinfo, ISOTFakeNews, and MultiDis are used for test evaluation. All labels are normalized into `real` and `fake`. For trainable baselines, hyperparameters are selected using the validation split. For LLM-based baselines, the source-domain training split is used as the support pool when retrieval, counterfactual generation, calibration, or knowledge construction is required. For each reproduced method, we retain the original high-level algorithmic design whenever possible and explicitly document the implementation-aligned adaptations required for our unified evaluation setting, includ-

Component	Parameter	Value
Taxonomies-Aligned	Adopted model	RoBERTa-large (Liu et al., 2019)
	Maximum token length	512
	Pooling strategy	Adaptive gated pooling
	Adaptive initialization	initialized to 1
	Aligned MLP hidden dims	$[786, 32, 32]$
	Aligned MLP dropout rate	0.3290
	Aligned MLP output dimension	Persuasion&Narrative Role: $d = 6$, Emotion: $d = 5$
Disinformation Detection	Number of HeteroAttention Block	$N_{att} = 1$ (hidden dim = 6+6+5 = 17)
	Classification Head	hidden dim = 128
	Activation	GELU
Total parameters		$\approx$ 357M($\approx$ 267K trainable)

Table 4: Full architecture details of Extax

Category	Parameter	Value
Taxonomies-Aligned	lr	0.00065
	Weight decay	0.00069
	Epochs	10
	Patience	7
	Seed	$[43, 434, 445]$
	Batch size	128
Disinformation Detection	lr	0.00096
	Weight decay	0.00018
	Epochs	50
	Patience	3
	Seed	$[43, 434, 445]$
	Batch size	128
Infra	Hardware	a NVIDIA RTX 4090 (24 GB VRAM) GPU(1 2h for training and evaluate)

Table 5: Training hyper-parameters and configuration for Extax

ing dataset loading, input/output formatting, label normalization, computational approximations, and unavailable-component substitutions.

- **RoBERTa** (Liu et al., 2019): For the baseline implementation, we utilize the `roberta-large` variant as the backbone encoder network. Given an input sequence, the token-level contextualized embeddings are extracted from the encoder, and a mean-pooling operation is subsequently applied across the sequence length to derive a unified sequence representation. This pooled representation is then fed into a sequence-level classification head structured as a multi-layer perceptron (MLP). Specifically, the classification head comprises a linear projection layer mapping the hidden size into a 256-dimensional intermediate space, followed by a Gaussian Error Linear Unit (GELU) activation function and Layer Normalization. To prevent overfitting, a Dropout layer with a regularization rate of 0.25 is applied before the final linear layer, which maps the features directly into the 2-dimensional classification target space.

- **DRES** (Farhangian et al., 2025): We reproduce DRES using the provided Dynamic Representation and Ensemble Selection implementation and adapt its input/output interface to our unified data splits. The training set is constructed by merging the CoAID and ECTF training splits, and the validation set is constructed by merging the corresponding validation splits. The test set contains CoAID, ECTF, EUDisinfo, ISOTFakeNews, and MultiDis. Following the dynamic representation selection principle of DRES, we construct multiple TF-IDF feature spaces with maximum dimensions of 100, 200, 300, 500, 700, and 1000, respectively, using English stop-word removal. For each representation, we train a pool of five classifiers, including Logistic Regression, Bernoulli Naive Bayes, k-Nearest Neighbors, Random Forest, and Support Vector Machine. Hyperparam-

eters are selected according to validation-set performance: Logistic Regression searches $C \in \{0.1, 1, 10\}$ with maximum iteration 2000; Bernoulli Naive Bayes searches $\alpha \in \{0.5, 1.0\}$; k-Nearest Neighbors searches $k \in \{3, 5, 7\}$; Random Forest uses 100 estimators and searches maximum depth in $\{5, 10, 20\}$; and SVM searches $C \in \{0.1, 1, 10\}$ with an RBF kernel. Instance hardness is estimated by the k-Disagreeing Neighbors criterion with $k = 5$. For each test instance, the representation with the lowest estimated hardness is selected dynamically. Dynamic ensemble selection is then performed using KNORA-Eliminate (KNORA-E; implemented as `KNORAE` in `deslib2`), with the region of competence size set to 5. Following the local-oracle criterion of KNORA-E, only base classifiers that correctly classify all samples in the local region of competence are selected; if no classifier satisfies this criterion, the region is iteratively reduced and re-evaluated. The final prediction is obtained by majority voting among the classifiers selected by KNORA-E. The experiment is repeated with three random seeds, $\{43, 434, 445\}$.

- **MSynFD** (Xiao et al., 2024): We reproduce MSynFD as a syntax-aware neural fake news detection baseline based on the released implementation structure. The model uses `bert-base-uncased` as the textual encoder, followed by a Semantic Transformer module, a Graphformer-based multi-hop syntactic aggregation module, and a keyword debiasing branch. The hidden dimension is set to 768, the number of attention heads is 12, the dropout rate is 0.1, and the hop number is set to 3. The maximum input sequence length is 200, and the maximum keyword sequence length is 50. In our reproduced implementation, the syntactic adjacency matrix is approximated by a local dependency-distance window with window size 3, rather than by running an external dependency parser. The final logits are computed by combining the main prediction branch and the keyword debiasing branch with weights 0.9 and 0.1, respectively. The training objective is the sum of the cross-entropy loss on the final prediction and 0.1 times the cross-entropy loss on the keyword debiasing output. The model is trained for 3 epochs with batch size 8 and learning rate 1×10^{-5} using AdamW. Following the actual optimizer parameter grouping in the implementation, non-bias and non-LayerNorm parameters use weight decay 0.01, while bias and LayerNorm parameters use no weight decay. For each run, up to 2000 training samples and 500 validation samples are sampled using the corresponding random seed. Evaluation is performed with batch size 16. We repeat the experiment with seeds $\{43, 434, 445\}$.

- **RAEmoLLM** (Liu et al., 2025): We implement RAEmoLLM as an emotion-aware retrieval-augmented LLM baseline. The support pool is built from the CoAID and ECTF training splits. For each test instance, the model extracts affective information and retrieves source-domain examples as in-context demonstrations. The reproduced implementation uses keyword-based affective features and, when available, transformer-based emotion or sentiment models. For retrieval, semantic similarity is computed with `all-MiniLM-L6-v2` when the encoder is available, and affective similarity is computed from the extracted emotion-oriented representation. The retrieval score combines semantic and affective similarity with weights 0.45 and 0.55, respectively. If neural sentence encoders are unavailable, the implementation falls back to a TF-IDF retriever with unigram and bigram features. The retrieval size is set to top-$K = 1$. The retrieved example, affective analysis, and target news text are inserted into a JSON-constrained prompt, and the LLM is required to output the predicted label, confidence score, and reasoning. The generated label is then normalized into the unified binary label space before evaluation.

- **FACTUAL** (Li et al., 2025): We reproduce FACTUAL as a counterfactual augmented calibration baseline and adapt it to binary fake news detection. The implementation first obtains an original LLM prediction and two debiased predictions, where the debiased prompts explicitly reduce sentiment-driven and target-prior-driven shortcuts. It then generates two counterfactual variants of the input: a causal counterfactual that minimally changes the text to flip the authenticity judgment, and

a non-causal counterfactual that paraphrases the input while preserving the original stance. These counterfactual texts are re-evaluated by the LLM to estimate counterfactual fake probabilities. The final prediction probability is calibrated using the original prediction, the two debiased predictions, the counterfactual predictions, and an explicit bias signal. The calibration strength is set to $\alpha = 0.35$. To avoid using target-test labels for calibration, the lightweight Logistic Regression calibration module is treated as an optional component and is used only when a calibrator trained on source-domain or validation data is explicitly available. Otherwise, we use the heuristic calibration rule implemented in the reproduced code. The final calibrated probability is thresholded at 0.5 to obtain the binary prediction.

- **DYNAMO** (Jin et al., 2025): We implement DYNAMO as a dynamic knowledge-update-driven LLM baseline with knowledge graph verification and Monte Carlo Tree Search. The source-domain support pool is built from the CoAID and ECTF training splits. The initial knowledge graph is constructed from reliable source-domain records labeled as `real`; if the number of reliable records is insufficient, the implementation falls back to the full support pool. In the reproduced setting, up to 10 support records are used for initial knowledge graph construction, and at most 2 entity-centered triples and 2 event-centered triples are extracted from each selected text. For each test instance, DYNAMO decomposes the input into verifiable subclaims through Monte Carlo Tree Search. The reproduced MCTS configuration uses maximum depth 2, maximum branching factor 2, one simulation, and UCB1 exploration coefficient 1.414. Each subclaim is verified against the constructed knowledge graph and assigned one of three verdicts: `supported`, `contradicted`, or `unverifiable`. The verification statistics and reasoning traces are then inserted into a JSON-constrained prompt, from which the LLM outputs the final binary label, confidence score, and reasoning. For efficiency, online knowledge updates are deferred during batch evaluation.
- **PCoT** (Modzelewski et al., 2025): We implemented the baseline using the official code released by PCoT at `https://github.com/ArkadiusDS/PCoT`. During evaluation, we ran all three prompting strategies, namely VaN, Z-CoT, and DeF-SpeC, and selected the best-performing strategy, DeF-SpeC, as the final reported result.

For the three LLM-based baselines, RAEmoLLM, FACTUAL, and DYNAMO, we use the same OpenAI-compatible chat-completion backend with the configured model name `gpt-5-mini`. The decoding temperature is set to 0.1, the maximum generation length is 256 tokens, and the request timeout is set to 30 seconds. Each LLM-based baseline is evaluated over 5 repeated inference runs with fixed base seed 42 for local sampling and preprocessing when applicable; variation across runs mainly comes from non-zero-temperature LLM generation. All LLM outputs are parsed through JSON-constrained generation and then normalized into the same binary label space as ExTax.

D.6 Interpretability

D.7 Global interpretability analysis

This appendix provides additional analysis of the interpretability results reported in the main text. The global distribution of the three taxonomic dimensions—persuasion strategies, emotional manipulation, and narrative roles—is shown in Figure 3. As illustrated, disinformation and real news exhibit markedly distinct structures within the constructed 17-dimensional taxonomic attribute space, providing intuitive evidence of the latent manipulative intents behind deceptive texts.

Specifically, concerning persuasion strategies, fake news relies heavily on "Attack on Reputation", "Simplification", and "Manipulative Wording", reflecting a typical discourse pattern that incites audiences by denigrating opposing targets through extreme language or misleads readers by omitting crucial information. This directly corroborates the findings in PCoT (Modzelewski et al., 2025). Regarding the emotional dimension, disinformation displays skewed distributions toward "Anger" and "Fear"—two high-arousal negative emotions—which closely aligns with psychological theories (Jamieson et al., 2025; Zhao et al., 2024) suggesting that intense emotional stimuli accelerate the dissemination of misleading content. Further-

Text: #Hope #NYC #NY #USA #World #Health vigilance a #vaccine is a year away #Pandemic harmful is misinformation side effects discrimination before #Ebola #SARS #MERS there was #AIDS a community marginalize... **Label**: Disinformation **Genre**: Post
Detected Taxonomies Type: - Persuasion: Justification, Call, Manipulative wording. - Emotion: Hope, Sadness. - Narrative Roles: Deceptive Subversives, Marginalized Sufferers.

Table 6: Case study 1.

Text: Title: Trump Attempts New Year's Eve Do-Over Tweet And Gets CRUSHED By Americans He Called 'Enemies' Content: If Donald Trump had sent proper New Year's greetings to the American people to begin with, he wouldn't be getting humiliated. But he royally f*cked up on New Year's Eve morning by referring to millions... **Label**: Disinformation **Genre**: News Article
Detected Taxonomies Type: - Persuasion: Attack on reputation, Justification, Simplification, Manipulative wording. - Emotion: Anger. - Narrative Roles: Overt Aggressors, Deceptive Subversives, Institutional Toxins.

Table 7: Case study 2.

more, in shaping narrative roles, disinformation strongly tends to frame narrative subjects as negative archetypes, such as "Deceptive Subversives" and "Institutional Toxins", whereas real news leans more toward positive or constructive roles such as "Ethical Stabilizers".

To further elucidate how taxonomic attributes across different dimensions jointly manifest within texts, the Sankey diagram in Figure 2 reveals their strong co-occurrence relationships. By tracing the convergence and divergence of the data flows, we observe several typical attribute binding patterns:

- *High-frequency co-occurrence of aggressive persuasion and negative emotions.* From the left to the middle of the diagram, the persuasion strategy of "Attack on Reputation" exhibits a strong binding relationship with "Anger" and "Fear". This indicates that when disinformation employs aggressive discourse, it is frequently accompanied by intense, high-arousal negative emotional stimuli; the two act as closely coupled co-occurrence components within inciting texts.

- *Mutual corroboration of rational defense and constructive roles.* In contrast to the extreme co-occurrence patterns of disinformation, "Justification"—a more neutral and objective persuasion strategy—has a high-probability co-occurrence with the "None" category in the emotional dimension. Texts without extreme emotional manipulation are also consistently associated with the positive narrative role of "Ethical Stabilizers", forming a core feature cluster of authentic news.

- *Chain transmission of non-manipulative characteristics.* When explicit manipulative intent is absent at one level, i.e., labeled as "None", this absence is likely to co-occur in subsequent levels. For instance, a large flow originating from "None" in persuasion strategies transitions into "None" in emotions and subsequently into "None" for narrative roles. This suggests that authentic news maintains a consistent non-manipulative pattern across multiple cognitive dimensions.

Overall, these distributional disparities show that ExTax not only improves the transparency of neural disinformation detection, but also captures multidimensional manipulative intents hidden within fluent texts, offering explanations that resemble the analytical process of human fact-checkers.

D.7.1 Case Study Analysis

To further demonstrate the instance-level interpretability of ExTax, we present two representative case studies in Table 6 and Table 7. These examples show how taxonomy-aligned predictions can be translated into human-auditable explanations by identifying the persuasive strategies, emotional appeals, and narrative-role constructions that support each disinformation prediction.

Case Study 1: Health-related social media post. Table 6 presents a short pandemic- and vaccine-related social media post. Although the text is fragmented and hashtag-driven, ExTax identifies a coherent manipulation profile. In the persuasion dimension, the detected categories include *Justification*, *Call*, and *Manipulative Wording*, indicating that the post attempts to legitimize its stance through emotionally charged public-health references while encouraging readers to adopt a particular viewpoint. Emotionally, ExTax detects both *Hope* and *Sadness*: vaccine-related references evoke hopeful expectations, whereas mentions of marginalized communities and severe disease histories trigger vulnerability and sympathy. In terms of narrative roles, *Deceptive Subversives* and *Marginalized Sufferers* suggest that the post frames some actors as potentially misleading while positioning affected communities as vulnerable victims. This case shows that ExTax can extract interpretable manipulation signals even from short and noisy social media posts.

Case Study 2: Political news article. Table 7 shows a long-form political article about Donald Trump's New Year's Eve tweet. Compared with the first case, this example contains more explicit evaluative and adversarial language. ExTax detects *Attack on Reputation*, *Justification*, *Simplification*, and *Manipulative Wording*, suggesting that the article relies on reputational attacks and reduces a complex political communication event into a morally one-sided conflict. The detected emotion, *Anger*, is consistent with the article's confrontational and mocking tone. The narrative-role predictions, including *Overt Aggressors*, *Deceptive Subversives*, and *Institutional Toxins*, further indicate that political actors are framed as visible antagonists and sources of institutional or social harm. Thus, the disinformation signal arises not from a single lexical cue, but from the alignment of aggressive persuasion, anger-based emotional manipulation, and negative role assignment.

Comparison between the two cases. Together, the two cases highlight the genre-adaptive interpretability of ExTax. The first case represents an implicit, compressed, and hashtag-driven manipulation pattern in short-form social media, while the second exhibits a more explicit pattern based on reputational attacks, anger, and antagonistic role construction in long-form political news. Despite these genre differences, both cases are mapped into the same 17-dimensional taxonomic space, enabling comparable explanations across posts and articles. These examples show that ExTax goes beyond binary classification by exposing how deceptive texts manipulate readers through rhetoric, emotion, and narrative framing.