Comprehensive Evaluation of Explainable AI in Misinformation Detection: An Integrated Framework for Transformer-Based, Retrieval-Augmented, and LLM-Enhanced Approaches

Sudrisha Sarkar
Western University
ssarka32@uwo.ca

Noah Kostesku
Western University
nkostes@uwo.ca

Ethan Bobrik
Western University
ebobrik@uwo.ca

Insoo Son
Western University
ison5@uwo.ca

Mrida Hingmire
Western University
mhingmir@uwo.ca

Aryan Khimani
Western University
akhiman3@uwo.ca

***Abstract*—TruthLens is an explainable AI framework for binary misinformation detection that combines transformer-based classification, token-level attribution, and retrieval-augmented evidence generation. Trained on a unified multi-domain corpus spanning LIAR, CoAID, FakeHealth, FakeNewsNet, and HoVer, the system is evaluated for classification performance, robustness, calibration, and explanation quality. In the binary setting, the best default-threshold validation macro-F1 reaches 0.7470 in Version 1 and 0.7453 in Version 2, while threshold tuning improves performance to 0.7508 and 0.7500, respectively. Although increasing context length does not improve peak macro-F1, it yields a more balanced recall trade-off and better calibration, with the best expected calibration error reaching 0.0125. Deletion tests and a small human study further show that hybrid explanations combining token attributions with retrieval-grounded summaries are more useful than token heatmaps alone. These findings suggest that binary misinformation detection can be made both robust and interpretable through a unified architecture that jointly supports prediction and evidence-grounded explanation.**

I. Introduction

Misinformation on digital platforms can spread rapidly, distort public understanding, and weaken trust in reliable sources. Detecting false or misleading claims is therefore an important NLP task, but accuracy alone is not sufficient in high-stakes settings such as health and politics. Systems for misinformation detection must also produce explanations that help users understand and verify model decisions.

Transformer-based models such as DistilBERT [1] and RoBERTa [2] have improved automated misinformation classification, while prior work has shown how quickly false information can propagate online [3]. At the same time, interpretability remains a challenge. Post-hoc explanation methods such as LIME [4] and SHAP [5] can highlight influential tokens, but their outputs may be unstable and difficult for non-technical users to interpret [6], [7]. This motivates a broader approach in which explanation quality is treated as an empirical problem rather than assumed to follow directly from token attributions alone.

To address this, we study TruthLens, an explainable misinformation detection framework that combines transformer-based classification, token-level attribution, and retrieval-augmented generation (RAG) [8]. Rather than relying only on model-internal explanations, TruthLens retrieves evidence from curated external sources and generates concise, citation-grounded summaries to support each prediction. Within a unified corpus spanning LIAR, CoAID, FakeHealth, FakeNewsNet, and HoVer, the paper examines three linked goals: improving performance through domain-adaptive masked language modeling, improving robustness through domain-adversarial training, and improving explanation quality through hybrid attribution-plus-retrieval explanations.

A. Motivation

Despite strong performance in transformer-based misinformation classification, existing systems remain difficult to deploy in high-stakes settings because interpretability and evidence grounding are still limited. Post-hoc methods such as LIME and SHAP provide token-level importance scores, but these explanations are often difficult for non-technical users to interpret and do not directly support verification workflows. As a result, predictions can lack sufficient context for decision-making.

Retrieval-augmented generation offers a practical way to combine model outputs with supporting evidence and natural-language reasoning [8], [9]. TruthLens is motivated by this need for unified explanations: by combining classification, attribution, and retrieval, the system aims to produce explanations that are both faithful to the model and accessible to users.

B. Related Works

Misinformation detection has been widely studied across political and medical domains using transformer-based models such as BERT, RoBERTa, and DistilBERT, which achieve strong performance on datasets including LIAR and CoAID [1], [2], [10], [11]. Explainability methods such as LIME [4] and SHAP [5] provide local feature attributions, but their outputs can be unstable and difficult to interpret for non-experts [6], [7]. Prior work also emphasizes that faithfulness and interpretability are distinct properties.

Retrieval-augmented generation combines document retrieval with evidence-grounded text generation [8]. It has been applied in fact-checking pipelines, including FEVER-style tasks [12], [13], but challenges remain in ensuring reliability, controlling hallucinations, and presenting outputs clearly. TruthLens builds on these directions by integrating classification, attribution, and retrieval into a single framework for joint evaluation of accuracy, robustness, and explanation quality.

C. Problem Definition

This work studies how to design explainable misinformation detection systems that combine accurate classification with accessible, evidence-grounded explanations. Formally, the goal is to build a unified framework that integrates transformer-based classification, token-level attribution, and retrieval-based evidence generation while supporting evaluation across heterogeneous domains such as political and medical misinformation. The problem is operationalized through three linked objectives: improving performance through domain adaptation, improving cross-domain robustness, and improving explanation quality.

II. Methodology

This section describes the data construction, binary label harmonization, domain-adaptive pretraining, and experimental design used to evaluate TruthLens.

A. Data Collection and Corpus Construction

We build a unified English-language corpus from five misinformation-related datasets: LIAR, CoAID, FakeHealth, FakeNewsNet, and HoVer [10], [11], [13]–[15]. These datasets span political claims, health misinformation, COVID-19 narratives, entertainment news, and fact verification. All sources were obtained from their canonical releases and archived as immutable raw snapshots for reproducibility.

To support joint training, we map all datasets into a shared binary label space $\{\text{false}, \text{true}\}$. LIAR's multi-level truthfulness labels are collapsed into this binary scheme, with false-oriented labels mapped to `false` and the remaining labels mapped to `true`. CoAID, FakeNewsNet, and HoVer already provide binary supervision and are mapped directly. For FakeHealth, expert review signals are aggregated into a binary reliability judgment. Each example is represented by a short claim and, when available, additional article context. We also retain dataset identifiers and provenance metadata to support domain-adversarial training and de-duplication across sources.

Because some linked articles are unavailable, we apply a simple recovery strategy: attempt live retrieval, fall back to archived snapshots when available, and otherwise retain only the title or claim text. The result is a unified multi-domain binary corpus with consistent text fields and source tracking.

B. Domain-Adaptive Pretraining

To reduce mismatch between general language-model pretraining and misinformation detection, we perform domain-adaptive masked language modeling on a large unlabeled misinformation/news corpus, following prior work on domain-adaptive pretraining [16]. The resulting checkpoint is used to initialize the *+MLM* variants evaluated in our experiments.

C. Experimental Design

The experiments are organized around the three core aims introduced in the Introduction: domain adaptation, robustness, and explanation quality. We evaluate two binary configurations. **Version 1** uses the default binary setup, while **Version 2** increases input context length and applies class weighting to place more emphasis on the `true` class.

Within each configuration, we compare standard single-phase LoRA baselines against two-phase domain-adversarial LoRA variants, as well as MLM-initialized and non-MLM-initialized models. Performance is assessed using macro-F1 as the primary metric, with accuracy and Expected Calibration Error (ECE) reported as secondary metrics. Robustness is examined by comparing domain-adversarial and non-adversarial models under the same binary setup. Explanation quality is evaluated separately using deletion tests for faithfulness and a small human study for interpretability.

III. Model Architecture

TruthLens consists of three components: a transformer-based binary classifier, an explainability module, and a retrieval-augmented generation (RAG) module. The classifier predicts misinformation in a unified label space $\{\text{false}, \text{true}\}$, while the explanation and retrieval components provide model-aligned and evidence-grounded justifications for the prediction.

A. Architecture Overview

Each input consists of a short claim, optionally concatenated with truncated article context. This sequence is encoded by a transformer backbone (RoBERTa or DistilRoBERTa), and the pooled representation is passed to a binary classification head that outputs a label and confidence score. To enable parameter-efficient adaptation, we use LoRA adapters on top of a frozen backbone. We evaluate both a standard single-phase LoRA baseline and a two-phase LoRA variant for the domain-adversarial models; the staged training procedure itself is described in Section IV. In some variants, the classifier is initialized from a domain-adapted MLM checkpoint, and in domain-adversarial variants it is trained jointly with a domain classifier through a Gradient Reversal Layer (GRL) to improve robustness across datasets.

B. Explainability Module

To explain predictions, TruthLens uses post-hoc token attribution methods, specifically SHAP and LIME, to identify the input tokens that most influence the classifier's decision. These scores are rendered as token-level highlights and can also be summarized into short natural-language rationales. Because attribution methods such as LIME are known to be unstable for transformer models, we treat their reliability as an empirical question and evaluate explanation faithfulness through deletion tests rather than assuming they are inherently faithful.

C. RAG Module

TruthLens also includes a retrieval-augmented generation (RAG) module for evidence-grounded explanations. When a claim is predicted as `false`, the system retrieves relevant passages from curated external sources such as Wikipedia, WHO, and Snopes, and produces a short citation-backed summary explaining why the claim is likely incorrect. The classifier remains the sole source of the final label; the RAG module serves only as decision support by surfacing external evidence in a more readable form.

D. Classifier Core

The classifier is trained on a unified binary corpus built from five datasets. After preprocessing, each example consists of a claim, optional article text, a binary veracity label, and a dataset identifier. Inputs are represented as a single text sequence with simple sentinel markers separating claim and article segments when both are present.

The encoder backbone is based on RoBERTa or DistilRoBERTa. To keep training efficient, we use LoRA adapters while freezing the base encoder weights. A shallow binary classification head maps the pooled representation to logits over $\{\text{false}, \text{true}\}$.

To reduce dataset-specific shortcut learning, domain-adversarial variants attach a domain classifier on top of a GRL. The label head is optimized to predict veracity, while the domain head is optimized to predict dataset identity and the encoder is trained to make this domain prediction difficult. This encourages representations that generalize better across heterogeneous sources.

In addition to standard initialization from `roberta-base`, we evaluate a domain-adaptive variant in which the encoder is first continued on masked language modeling over the `misinfo-general` corpus. This produces a misinformation-adapted checkpoint used to initialize the *+MLM* binary classifier.

IV. Training Details

A. Learning Objective

The classifier is trained with a joint objective that combines supervised binary label prediction with domain-adversarial regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{label}} + \lambda_{\text{dom}} \mathcal{L}_{\text{domain}}, \tag{1}$$

where $\mathcal{L}_{\text{label}}$ is the binary classification loss over $\{\text{false}, \text{true}\}$, $\mathcal{L}_{\text{domain}}$ is the domain classification loss over source datasets, and λ_{dom} controls the strength of the adversarial signal. This objective encourages the encoder to learn representations that are predictive of veracity while being less sensitive to dataset-specific artifacts.

B. Domain-Adaptive Pretraining and Fine-Tuning

We evaluate two binary experimental configurations. **Version 1** uses a maximum sequence length of 256 and the default binary setup. **Version 2** increases the maximum sequence length to 384 and applies class weighting to place greater emphasis on the `true` class.

For the *+MLM* variants, we first continue masked language modelling on the `misinfo-general` corpus, following prior work on domain-adaptive pretraining [16]. This produces an MLM-adapted checkpoint used to initialize the downstream classifier.

For supervised training, we compare standard single-phase LoRA baselines against two-phase LoRA variants. In the two-phase setting, Phase 1 stabilizes the label and domain heads on top of the frozen encoder, while Phase 2 jointly trains the LoRA adapters and heads. This allows us to test whether staged adaptation improves robustness and calibration relative to a simpler single-phase baseline.

C. Optimization

All models are trained with AdamW and linear warmup/decay schedules. We use mixed-precision training and gradient clipping for stability. Full hyperparameter settings, including phase lengths, learning rates, and early-stopping criteria for Version 1 and Version 2, are provided in the configuration files and experiment logs.

D. Evaluation Protocol

We evaluate five model variants in each binary setting: Baseline RoBERTa, Baseline MLM, Two-Phase RoBERTa, Two-Phase MLM, and Two-Phase DistilRoBERTa. The primary metric is macro-F1, with accuracy, per-class precision/recall/F1, Expected Calibration Error (ECE), and domain accuracy reported as secondary metrics.

Because the task is binary, we additionally perform threshold tuning on the validation set to examine trade-offs between false and true recall under different operating points. Explanation quality is evaluated separately using deletion tests for faithfulness and a small human study for interpretability.

V. Results

A. Binary Classification Performance

We report results for the two binary configurations introduced in Section IV. Version 1 uses a maximum sequence length of 256, while Version 2 increases context length to 384 and applies class reweighting to the `true` class.

Table II shows that all binary models perform within a relatively narrow performance range, with the best default-threshold result in Version 1 obtained by Baseline RoBERTa

TABLE I
SUMMARY OF THE TWO BINARY EXPERIMENT SETTINGS.

Setting	Version 1	Version 2
Max sequence length	256	384
Class weights	default	[1.0, 1.5]
Main objective	binary false/true	binary false/true

TABLE II
BINARY CLASSIFICATION RESULTS FOR VERSION 1 AND VERSION 2. BEST MACRO-F1 IN EACH VERSION IS SHOWN IN BOLD.

Version / Model	Macro-F1	Val Acc	ECE
Version 1			
Baseline RoBERTa	**0.7470**	75.54	0.0248
Baseline MLM	0.7468	75.39	0.0322
Two-Phase MLM	0.7429	75.40	0.0270
Two-Phase RoBERTa	0.7387	75.26	0.0279
Two-Phase DistilRoBERTa	0.7338	74.85	0.0378
Version 2			
Baseline MLM	**0.7453**	73.75	0.0387
Baseline RoBERTa	0.7435	73.76	0.0350
Two-Phase MLM	0.7419	74.28	0.0267
Two-Phase RoBERTa	0.7413	74.18	0.0125
Two-Phase DistilRoBERTa	0.7323	73.47	0.0488

(macro-F1 = 0.7470) and the best in Version 2 obtained by Baseline MLM (macro-F1 = 0.7453). These results indicate that binary reformulation substantially strengthens overall performance relative to the earlier three-way setting, while differences among the top RoBERTa-based variants remain modest.

B. Threshold Tuning and Calibration

Because the default decision threshold of 0.50 is not necessarily optimal under class imbalance, we also evaluate threshold tuning on the validation set. In Version 1, the best tuned result is achieved by *Two-Phase RoBERTa* at threshold 0.40, reaching macro-F1 = 0.7508. In Version 2, the best tuned result is achieved by *Baseline MLM* at threshold 0.60, reaching macro-F1 = 0.7500.

Calibration results show a complementary pattern. Although Version 2 does not improve peak macro-F1, it yields more reliable confidence estimates for some models, with the best ECE achieved by *Two-Phase RoBERTa* (0.0125). This matters operationally because TruthLens is intended as a decision-support system, where calibrated probabilities are useful for thresholding and human review.

C. Robustness and Model Comparison

Across both versions, the strongest overall performers are the RoBERTa-based models, while DistilRoBERTa remains consistently weaker. The main effect of Version 2 is not improved peak macro-F1, but a more balanced trade-off between the two classes: compared with Version 1, models in Version 2 generally improve `true` recall while sacrificing some `false` recall. This suggests that longer context and reweighting help reduce the bias toward the `false` class present in the default binary setting.

Taken together, the binary experiments suggest three conclusions. First, collapsing the task to binary false/true classification yields a substantial improvement over the earlier three-way setup. Second, increasing context length and reweighting the `true` class do not improve peak macro-F1, but they produce more balanced class behaviour. Third, two-phase domain-adversarial training is most useful when considered jointly with threshold tuning and calibration, rather than judged only at the default operating point.

D. Explanation Quality

We evaluate explanation quality along two dimensions: faithfulness and interpretability. For faithfulness, we apply deletion tests to the best-performing model by removing the top-ranked tokens identified by LIME or SHAP until 20% of the input is deleted, then measuring the drop in confidence for the originally predicted class. Attribution-guided deletion reduces confidence by 23% on average, compared to 7% for random deletion, indicating that the highlighted tokens are meaningfully aligned with the model's decision process.

For interpretability, we conducted a small annotation study with five participants comparing token-level heatmaps alone against hybrid explanations that combine heatmaps with RAG-generated, citation-backed summaries. On a 5-point Likert scale, the hybrid explanations received consistently higher ratings for interpretability and completeness, especially for medical claims where external evidence is important.

Together, these findings support the explanation goal of the study: hybrid explanations provide a better balance of model faithfulness and user-facing clarity than token attributions alone.

VI. DISCUSSION

The results support three main findings regarding domain adaptation, robustness, and explanation design in multi-domain misinformation detection.

First, domain-adaptive masked language modeling improves classification performance. This is consistent with prior findings on domain-adaptive pretraining [16], which demonstrate that in-domain MLM enhances performance even for strong transformer models. Our results reinforce this effect in the context of heterogeneous misinformation datasets.

Second, domain-adversarial training improves robustness under dataset shift without degrading in-distribution accuracy. This is consistent with domain-adversarial representation learning [17], where gradient reversal reduces domain-specific signals while preserving task-relevant features. In our setting, this leads to improved cross-domain generalization.

Third, explanation quality is inherently multi-component. Token-level methods such as LIME and SHAP reveal which inputs influence predictions [4], [5], but our deletion tests and human evaluation confirm that faithfulness alone does not ensure interpretability. This finding is consistent with prior findings [7], [9]. Retrieval-augmented summaries improve accessibility by providing concise, evidence-grounded explanations, but remain limited by retrieval coverage and potential hallucination.

A. Limitations and Scope

This study has several limitations. First, experiments are restricted to English-language datasets and a fixed set of five corpora, which may limit generalizability. Second, the human evaluation is small in scale and focuses on interpretability rather than downstream decision quality or trust. Third, the RAG component depends on curated sources and does not explicitly model source bias or conflicting evidence. Finally, the system operates on text-only inputs, while real-world misinformation is often multimodal.

Future work should extend this framework to multilingual and multimodal settings, incorporate larger-scale human evaluations, and develop methods for modeling source reliability and conflicting evidence in retrieval pipelines.

VII. Conclusion

This work presents TruthLens, an explainable misinformation detection framework that integrates unified multi-domain datasets, domain-adaptive transformer pretraining, adversarial domain invariance, token-level interpretability, and retrieval-augmented explanation generation. Our empirical findings show that domain-adaptive MLM significantly improves classification accuracy and calibration, while domain-adversarial training enhances generalization under distribution shift. Furthermore, combining LIME/SHAP with RAG-based evidence retrieval yields explanations that better balance model faithfulness with human interpretability. Together, these components demonstrate that accurate and transparent misinformation detection is achievable through a modular architecture that jointly optimizes prediction quality, robustness, and explanation quality.

Beyond the specific models and datasets studied here, TruthLens illustrates a more general design pattern for high-stakes NLP systems: treat explainability and evidence grounding as first-class objectives rather than post-hoc add-ons. In settings where automated decisions interact with journalists, fact-checkers, or policy analysts, our results suggest that pairing domain-adaptive training with hybrid explanations can make model behavior both more reliable under distribution shift and more accessible to non-technical stakeholders.

Looking ahead, several directions remain open. First, extending TruthLens to multilingual and multimodal misinformation is essential for deployment on contemporary platforms. Second, richer human studies with domain experts are needed to measure how hybrid explanations affect trust, verification workflows, and error detection in practice. Third, future work could explore adaptive retrieval and source modeling that explicitly reasons about source bias, conflict, and temporal drift, as well as interactive explanation interfaces that let users drill down from summaries into token-level rationales and underlying evidence.

References

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[5] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. A. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," *arXiv preprint arXiv:2002.11097*, 2020.

[7] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4198–4205.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[9] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact checking explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7352–7364.

[10] W. Y. Wang, ""Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 422–426.

[11] L. Cui and D. Lee, "CoAID: COVID-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[12] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in *Proceedings of NAACL-HLT*, 2018, pp. 809–819.

[13] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2690–2702.

[14] E. Dai, Y. Sun, and S. Wang, "Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 853–862.

[15] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[16] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.