Detecting Deception: Intelligent Systems for Fighting Misinformation

Leona E Josheph*
lj9984@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Mandira Sawkar
ms7201@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Vivek Senthil
vs9589@rit.edu
Rochester Institute of Technology
Rochester, NY, USA

Abstract

The rapid spread of fake news across digital platforms poses a significant threat to public trust, democratic discourse, and information integrity. While transformer-based language models like RoBERTa and BERT have advanced the state of the art in fake news classification, their lack of transparency limits real-world adoption and user trust. In this work, we propose a hybrid fake news detection framework that combines the strengths of pre-trained language models with interpretable, rule-based reasoning. Our system fine-tunes RoBERTa on benchmark datasets such as LIAR and FakeNewsNet to capture nuanced linguistic cues, while integrating symbolic logic rules and curated fact repositories to validate claims and surface logical contradictions. This integration not only preserves or improves classification performance over baseline deep learning models but also provides human-readable justifications for decisions. Experimental evaluations show that our approach enhances accuracy and robustness, especially in detecting logically inconsistent or borderline claims, while offering significantly improved explainability. This work contributes toward the development of trustworthy, interpretable, and deployable fake news detection systems.

Keywords

Fake News Detection, Explainable Artificial Intelligence, Hybrid Neuro-Symbolic Models, Transformer Language Models, Rule-Based Reasoning, Interpretable Machine Learning, Trustworthy AI and Transparency, LIAR Dataset, FakeNewsNet Dataset, Logical Consistency Validation, Misinformation Robustness

ACM Reference Format:
Leona E Josheph, Mandira Sawkar, and Vivek Senthil. 2025. Detecting Deception: Intelligent Systems for Fighting Misinformation. In . ACM, New York, NY, USA, 8 pages.

1 Introduction

1.1 Motivation

Fake news, defined as false or misleading information presented as legitimate news, has emerged as a pervasive challenge in the digital age. It distorts public opinion, undermines trust in institutions, and influences elections and health-related decisions. Despite the growing sophistication of machine learning models in detecting such misinformation, most operate as black boxes—delivering high accuracy but little to no insight into the why behind their predictions. This opacity impairs their utility for journalists, policymakers, and end users who require transparency to build trust and take informed action. Furthermore, these models often fail to detect logically contradictory claims or provide contextual explanations, which are crucial in real-world verification.

1.2 Problem Statement

While state-of-the-art models like BERT and RoBERTa have demonstrated high accuracy in fake news classification tasks, they often lack mechanisms for interpretability, and they do not reason over known facts. Some approaches use knowledge graphs or user comment analysis to augment predictions, but these methods can be computationally expensive or reliant on external social inputs. There remains a gap in combining scalable machine learning techniques with symbolic reasoning to both detect fake news and explain why it is fake using known logical inconsistencies.

1.3 Contributions

This paper addresses these limitations by proposing a unified, hybrid approach to fake news detection that combines:

- Transformer-based language modeling (RoBERTa/BERT) for high-quality contextual understanding and classification.
- Rule-based symbolic reasoning using Python/Prolog to detect logical contradictions between claims and curated facts.
- An explainability layer that leverages the DeepSeek-R1 large language model to flag uncertain predictions and provide rich textual justifications based on contradiction detection.

By merging statistical learning with logical reasoning, our model is both accurate and interpretable. This makes it especially useful in domains where transparency and auditability are critical.

1.4 Organization

The paper is structured as follows: Section 2 reviews prior work in fake news detection and explainable AI. Section 3 presents our hybrid methodologies, including datesets, preprocessing, model architecture, and experimental setup. Section 4 reports results and explains our evaluation metrics. Section 5 discusses limitations, and directions for future research. Finally, section 6 concludes with further implications.

*All authors contributed equally to this research.

IDAI - 710, Fundamentals of ML, Rochester, NY, USA

2 Related work

Fake-news detection has evolved along several complementary lines: (1) textual classification using deep transformers, (2) metadata and ensemble model approach, (3) retrieval-augmented generative methods (4) neuro-symbolic hybrid approaches, (5) LLM-enhanced methods (6) multimodal techniques and (7) unsupervised approaches. We summarize key contributions in each area.

2.1 Transformer-based Text Classification

Pre-trained language models such as RoBERTa [7] have set new benchmarks in text classification by optimizing BERT's pretraining strategy (longer training, dynamic masking) to achieve SOTA on GLUE and SQuAD benchmarks . Applied to misinformation, fine-tuned RoBERTa variants achieve ≈72% accuracy on LIAR and ≈90% on FakeNewsNet, but remain opaque black boxes.

2.2 Metadata Enrichment & Ensemble Model

Previous work has extensively covered some of the most popular feature extraction and fusion techniques in natural language processing (NLP). There is evidence that [5], application of neural network methods towards text feature extraction, with special emphasis on architectures like CNNs and RNNs that have proven effective in achieving semantic representations. The evolution of text classification methods from traditional machine learning to deep learning models and also how improving classification performance depends on good feature representation [6]. In addition, an ensemble and hybrid deep learning models for NLP highlights the strength of an ensemble of various model structures to improve the performance across tasks [4] .

2.3 Retrieval-Augmented Generative Approaches

Retrieval-Augmented Generation (RAG) grounds LLM outputs in real-time evidence, mitigating hallucinations. Nezafat (2024) [9]integrate a sparse Mixture-of-Experts LLM with Google Search, using few-shot prompting and cosine-based pre-filtering to achieve 88% accuracy—an 23 point gain over LLM alone—on ISOT . However, purely generative systems still lack explicit contradiction checks.

2.4 Neuro-Symbolic & Explainable Hybrids

Neural-symbolic methods combine statistical learning with logical reasoning to surface transparent justifications. Garcez et. al. [2] survey neural-symbolic integrations, showing how simple IF–THEN rules over curated facts can be melded with neural nets for accountability . In the fake-news domain, Shu et al.'s [11] dEFEND model uses co-attention over news text and user comments to highlight "check-worthy" sentences, boosting F1 by 5.3% while producing sentence-level explanations.

2.5 LLM-Enhanced Detection Methods

Recent work has explored LLMs as advisors rather than standalone detectors. Hu [3] introduced the Adaptive Rationale Guidance (ARG) network that selectively incorporates LLM-generated rationales into a small language model (SLM), outperforming both standalone approaches. Their framework achieved F1 scores of 0.784 and 0.790 on Chinese and English datasets respectively, compared to GPT-3.5's scores of 0.725 and 0.702. Ma [8] proposed enhancing semantic mining with LLMs to detect inconsistencies in relationships among named entities and topics, constructing heterogeneous graphs for propagating local and global semantics, achieving peak accuracy of 97.4% on MM-COVID dataset. Similarly, Xie [18] integrated multi-source knowledge graphs with LLaMa2-7B in their MiLk-FD framework, improving F1-scores by up to 9.76% over baselines across multiple datasets. These approaches demonstrate that LLMs are more effective as components in hybrid systems rather than standalone detectors.

2.6 Multimodal and Style-Robust Detection

Recent advances include multimodal approaches and techniques to counter style-based evasion. Wang [15] proposed FND-LLM, incorporating text (BERT), images (Vision Transformer, CLIP), and visual tampering features (EAViT) through a Multi-gate Mixture of Experts network, achieving improved accuracy on Weibo (91.2%), GossipCop (90.5%), and Politifact (92.6%). Addressing adversarial vulnerabilities, Wu [17] demonstrated that existing text-based detectors suffer up to 38.3% decline in F1-scores when exposed to LLM-generated style attacks that mimic reputable sources. Their proposed SheepDog framework maintains robust performance through style-agnostic training and content-focused veracity attributions, highlighting the need for detection systems resistant to increasingly sophisticated evasion techniques.

2.7 Unsupervised and Reinforcement Learning Approaches

For scenarios with limited labeled data, Yang [19] developed an unsupervised framework treating news veracity and user credibility as latent variables in a Bayesian network, achieving 75.9% accuracy on the LIAR dataset without requiring labeled examples. Complementarily, Ouyang [10] demonstrated how reinforcement learning from human feedback (RLHF) could improve LLM alignment with human preferences, reducing hallucination rates from 41% to 21% in closed-domain tasks, suggesting potential applications in more truthful fake news detection.

3 Methodology

3.1 Datasets

Our hybrid fake-news detector is trained and evaluated on two widely used benchmarks—LIAR [16] and FakeNewsNet [12] — providing both short-claim and full-article perspectives.

- **LIAR**
  - Size & Classes: 12,836 human-annotated political statements, each labeled on a six-point truthfulness scale: Pants-on-Fire, False, Mostly False, Half True, Mostly True, and True. For our binary classification experiments, we collapse "Pants-on-Fire," "False," and "Mostly False" into Fake, and "Half True," "Mostly True," and "True" into Real.
  - Variables: Each record contains the statement text, speaker metadata (e.g. party, job), context (e.g. subject, venue), and the original multi-class label.

  - Collection: Samples were scraped and curated from the PolitiFact fact-checking website, covering statements made by U.S. politicians and public figures between 2007 and 2016; professional fact-checkers assigned each verdict.
- **FakeNewsNet**
  - Size & Classes: 23,921 full news articles drawn from two fact-checking sites: 1,200 from PolitiFact and 22,700 from GossipCop, each labeled True or False.
  - Variables: In addition to article text and label, the dataset includes rich social context metadata—publisher source, publication timestamp, author, and user engagements (tweets, retweets, replies) collected via Twitter API.
  - Collection: Shu et al. built FakeNewsNet by (1) crawling PolitiFact and GossipCop for fact-checked URLs; (2) downloading full article content; and (3) gathering associated Twitter posts and user interactions through tweet IDs and Twitter's REST API.

These two datasets together span both concise statements (LIAR) and extended news stories (FakeNewsNet), enabling comprehensive evaluation of our model's classification accuracy and explainability across varied misinformation formats.

3.2 Methodology 1 – Hybrid Feature Fusion

Dataset and Problem Framing

We use the **LIAR dataset** (Wang, 2017), a widely referenced benchmark data set for detecting fake news. Detailed information about the dataset is provided in Section 3.1. Each instance includes a statement by a public figure or entity, along with associated metadata, including the speaker's political affiliation, historical truthfulness record, occupation, context of statement, and so on. For our binary classification task, we map the original six-class labels into two categories for this methodology:

- **True Class (1):** "true", "mostly-true", and "half-true"
- **False Class (0):** "barely-true", "false", and "pants-on-fire"

This transformation allows us to frame the problem as a binary classification task, where the objective is to predict whether a statement is **true (1)** or **false (0)**.

Multi-Model Ensemble Architecture

To enhance robustness to classification and get complementary perspectives from different data modalities, we follow a three-branch ensemble approach that combines predictions from: (1) raw statements, (2) semantically rewritten feature queries, and (3) engineered metadata features. The final prediction is a weighted sum of the three model outputs, with weights learned through validation grid search.

1. Statement-Based Classification (Text Transformer)

In the first branch of our model, the actual statement itself is input to a light Transformer model — **DistilBERT** — that was fine-tuned for our task of classification. The model output is a confidence score that states how sure it is that the input statement holds. This process picks up on the semantic and contextual cues encoded within the statement itself directly.

2. Metadata-to-Sentence Transformation (Sentence Transformer)

The second branch utilizes the structured metadata fields of the LIAR dataset — such as speaker identity, political party, historical credibility, and contextual information — and **converts them into natural language sentences**. For instance, features like *"speaker=xyz"*, *"party=party A member*, and *"context=campaign rally"* are reformulated into a sentence such as:

"The speaker, xyz, a party A member, made this statement during a campaign rally."

This sentence is passed through another instance of the DistilBERT transformer. By translating structured features to text, we enable the model to leverage semantic consciousness and generalization capability of transformers even for metadata.

3. LLM-Based Feature Engineering + XGBoost

In the third branch, we leverage **Ollama LLM** to automatically extract semantic features from the statement. These features are labelled as binary classes that catch subtle points such as:

- `IsPolitical`: Is the topic political?
- `TalksAboutOwnParty`: Does the speaker reference their own party?
- `TalksAboutOtherParty`: Is another party being criticized?
- `IsAccusatory`: Is the tone blaming or confrontational?
- `IsHeSaidSheSaid`: Is it a secondhand quote (e.g., "they said…")?
- `IsQuote`: Is it a direct quote from another person?

These features were derived using zero-shot prompting strategies, and each is encoded as a binary value (1 = present, 0 = absent). We observed meaningful patterns during exploratory analysis — for example, statements labeled as `IsHeSaidSheSaid` or `IsPolitical` were significantly more likely to be false. These engineered features are then input into an **XGBoost classifier**, which outputs a third probability score of the statement being true.

Ensemble and Weight Optimization

The final prediction is obtained by computing a **weighted average** of the three model outputs. Let p_1, p_2, p_3 represent the truth probability scores from the statement model, the sentence-structured metadata model, and the XGBoost-based semantic feature model respectively. The ensemble output is:

$$P_{\text{final}} = w_1 \cdot p_1 + w_2 \cdot p_2 + w_3 \cdot p_3$$

Weights w_1, w_2, w_3 are constrained to sum to 1 and are adjusted by grid search on the validation set to maximize classification accuracy and F1 score. The weighted ensemble allows us to combine complementary insights — semantic content, contextual information, and structured logic — to improve fake news detection performance.

3.3 Methodology 2 – Logic-Gated ML Overrides

This approach integrates deep learning with rule-based reasoning to create a hybrid system for fake news detection. The system combines a fine-tuned transformer model with symbolic contradiction detection grounded in natural language inference (NLI). This structure leverages the empirical power of pre-trained language models while enhancing interpretability and trust through semantic reasoning.

We begin by preprocessing two prominent datasets: *LIAR* and *FakeNewsNet*. Each dataset is cleaned and normalized to a consistent schema with five fields: `statement`, `label`, `speaker`, `party`, and `context`. Missing metadata is defaulted to `"unknown"`, and label text is converted to binary values—mapping variants like `FAKE` or `FALSE` to 0 and `TRUE` to 1. We drop rows with null or malformed entries and shuffle the data before splitting it into an 80% training set and 20% validation set using a fixed random seed for reproducibility.

Each input example is reformatted into a templated string of the form:

```
Speaker: {speaker}. Party: {party}. Context:
{context}. Claim: {statement}
```

This composite input allows the model to incorporate contextual information during training and inference.

We use **DistilRoBERTa**—a computationally efficient variant of RoBERTa—as the backbone model. The architecture includes two heads: one for binary classification and one for contradiction detection. The transformer model outputs a pooled representation which is passed through a dropout layer (with $p = 0.3$) before being fed into two parallel fully connected layers with sigmoid activation. One head predicts whether the claim is fake or real, while the second estimates whether the claim contradicts a known fact. Both heads are trained using binary cross-entropy loss.

To detect contradictions, we use a curated set of ground-truth facts, such as *"earth is round"* or *"vaccines cause autism"*. These are embedded into a semantic vector space using the `all-MiniLM-L6-v2` model from SentenceTransformers. For each claim, we compute its embedding and retrieve the top 3 most semantically similar known facts based on cosine similarity:

$$\text{sim}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$$

Each matched pair (known fact and claim) is passed to a `roberta-large-mnli` model, which classifies the relationship as entailment, neutral, or contradiction. If the predicted label is `CONTRADICTION` with a confidence score above 0.85, the claim is flagged as contradictory. During evaluation, if the main classifier predicts a claim as real (label 1) but it is contradicted by a known fact, and the model's confidence falls between 0.4 and 0.6, the prediction is overridden to 0.5 and labeled as `"uncertain"`. This override logic ensures that predictions in areas of epistemic uncertainty are explicitly flagged for further human review.

3.3.1 Validation. Success for our project is defined along two dimensions: classification accuracy and interpretability. Our goal is to achieve a classification accuracy of at least 80%, which is the benchmark for strong transformer-based fake news detectors, while introducing a logic-based reasoning layer that can justify prediction overrides through transparent explanations.

To validate our approach, we employ a stratified 80/20 train-test split, without cross-validation due to computational constraints. Evaluation metrics include accuracy, macro and weighted precision, recall, and F1-score. In addition to standard classification metrics, we log the number of predictions flagged as `uncertain` by the contradiction module to evaluate the influence of rule-based logic on the final predictions.

We test three key hypotheses:

- **Hybrid effectiveness:** The addition of contradiction-aware logic will improve interpretability and robustness without sacrificing accuracy below 80%.
- **Semantic correction:** When model confidence is low, the contradiction module will correct or flag unreliable predictions.
- **Performance parity:** The hybrid model will match or exceed the F1-score of a standalone DistilRoBERTa classifier while offering clearer, logic-grounded explanations.

The final system achieved 83% accuracy with 52 predictions flagged as logically uncertain, demonstrating that the hybrid model effectively integrates symbolic reasoning with neural classification to enhance both transparency and performance.

3.4 Methodology 3 – RAG with RoBERTa

This approach is a one-shot generative fake news detection framework that leverages Large Language Models (LLMs) to analyze textual content and detect misinformation. Inspired by recent advances in retrieval-augmented generation [9], hallucination mitigation [14] and hybrid modeling [13], this approach integrates multiple complementary strategies to achieve high accuracy and reliability in detecting fake news with minimal training requirements.

The methodology consists of four main components that work together to evaluate the veracity of news content:

3.4.1 Fact Extraction for Query Generation.
Before retrieving external evidence, the system first extracts key factual claims from the input text using spaCy's "en_core_web_sm" model:

- **Named Entity Recognition (NER)**: Identifies key entities (people, organizations, dates, locations) to ensure fact-checking focuses on relevant elements.
- **Relation Extraction**: Determines relationships between entities to structure claims logically.
- **Event & Numeric Identification**: Extracts temporal references and statistical claims requiring verification.
- **Query Construction**: Forms structured queries using extracted facts, optimized for retrieving precise information from search engines.

This structured approach ensures that verification efforts target specific factual components rather than broad document-level assertions, improving precision and reducing irrelevant retrievals.

3.4.2 Retrieval Module.
The retrieval component fetches relevant background information for the input news text using:

- **Real-time Web Search**: Queries online sources to obtain facts and trusted references.
- **Content Filtering**: Extracts relevant portions from retrieved documents while filtering out advertisements and irrelevant content.
- **Source Credibility Assessment**: Assigns weights to retrieved information based on source credibility using a predefined list of trusted sources.

By providing the LLM with external evidence, this module grounds the generative process in verifiable information, helping to mitigate hallucinations common in LLM outputs [14].

3.4.3 Generative LLM Reasoning Module (DeepSeek-r1). At the heart of the system is DeepSeek-r1, a powerful generative model that analyzes news content along with retrieved evidence. The LLM is prompted to:

- Analyze the input news text and extracted facts
- Compare these with retrieved evidence
- Generate a structured reasoning chain explaining consistencies and inconsistencies
- Produce a final veracity judgment (pants-fire, False, barely-true, half-true, mostly-true, True)

Chain-of-thought prompting is employed to guide the model through structured reasoning [14]. The output includes a concise explanation referencing evidence (improving transparency) and a final decision.

3.4.4 Knowledge Consistency & Verification Module (RoBERTa-large-mNLI). To further improve accuracy, a verification component checks the consistency between claims and factual knowledge using RoBERTa-large-mNLI a similar variant of RoBERTa-NLI (Natural Language Inference) [1]:

- Compares extracted facts with retrieved evidence using sentence-level matching
- Classifies relationships as pants-fire, (highest negative rating), False, barely-true, half-true, mostly-true, True (highest positive rating).
- Overrides the LLM's decision when strong contradictions exist

We modified the RoBERTa-large-mnli model by adapting the classification head from the original 3 classes to 6 classes to align with the LIAR dataset's classification scheme. This adaptation allows the model to provide more nuanced veracity judgments beyond simple true/false classifications. The model was trained using the paired news text and scraped evidence, including cases with the "No evidence found" placeholder.

RoBERTa-NLI computes a probability distribution over the classification labels:

$$P(y|x) = \text{softmax}(Wh_x + b) \tag{1}$$

where h_x is the hidden representation of the (Claim, Evidence) pair, and W, b are learned parameters.

The final classification is determined as:

$$y^* = \arg\max P(y|x) \tag{2}$$

where y^* is the highest-probability label (ranging from "Pants-on-Fire" to "True" for the LIAR dataset).

This module ensures the system catches factual inconsistencies and prevents the LLM from making incorrect classifications due to hallucinations.

Validation

Success for this methodology is defined by three key metrics:

(1) **High Classification Performance** : our system should achieve state-of-the-art (SOTA) or near-SOTA accuracy, precision, recall, and F1-score on benchmark datasets (LIAR, FakeNewsNet, and Constraint@AAAI), with our hybrid DeepSeek LLM + RoBERTa-NLI approach expected to outperform traditional models.
(2) **Explainability & Transparency** : unlike black-box models, our system must generate concise and justifiable explanations for its classifications, evaluated through qualitative human assessment.

Validation Methodology

Since we are not training a model from scratch, validation focuses on evaluating inference performance using held-out datasets: 80% development, 10% validation sets (used for prompt engineering and RoBERTa-NLI fine-tuning) and 10% test set for final evaluation on unseen data. We rely on prompt engineering and retrieval optimization rather than model fine-tuning, with DeepSeek-r1 used for one-shot fake news classification and fine-tuned RoBERTa-NLI applied to verify factual consistency.

Research Hypotheses Our research tests specific hypothesis, where the proposed model will achieve greater than or close to 80% accuracy, outperforming current SOTA models on benchmark datasets, improving upon existing BERT-based models (72% on LIAR)

4 Results

4.1 Evaluation - Hybrid Feature Fusion

The relative performance highlights the advantage of integrating multiple aspects through a hybrid feature fusion scheme. Among all the single models, the top recall (0.8592), as shown in Table 1, was reported by the engineered feature model (M3), which shows that it is good at detecting real true statements, but it is lagging on precision and global accuracy. The metadata-to-sentence model (M2) recorded quite balanced performance on all measures but slightly outperformed the raw statement model (M1). But the ensemble model (M3) performed better than all the individual models by a large margin, achieving the best F1 score (0.7560) and accuracy (0.7342). This confirms that weighted ensembling of semantic content, contextual metadata, and engineered features captures complementary signals most effectively and results in stronger fake news classification.

4.2 Evaluation - Logic-Gated ML Overrides

We evaluate our hybrid DistilRoBERTa + contradiction system on the held-out 20% validation split (7 207 examples). Table 2 presents the per-class precision, recall, and F1-score after applying our contradiction-aware override logic. The overall accuracy is 83%, exceeding our 80% target and matching state-of-the-art RoBERTa baselines.

Table 2 shows nearly balanced performance on both classes: the Fake class achieves an F_1 of 0.78 and the Real class 0.87. This parity confirms our third hypothesis of performance parity with a standalone transformer, while preserving interpretability through logic overrides.

Figure 1 displays the confusion matrix for all non-uncertain predictions. Of 2 657 true fake examples, 2 061 were correctly identified

Table 1: Performance comparison of individual models and the ensemble model.

Sl/No	Statement	Precision	Recall	F1 Score	Accuracy
1	Statement (M1)	0.6583	0.7137	0.6848	0.6100
2	Original Features (M2)	0.6583	0.7705	0.7100	0.6262
3	Engineered (M3)	0.6114	0.8592	0.7144	0.5913
4	Ensemble	0.7453	0.7670	0.7560	0.7342

Table 2: Classification metrics after hybrid logic overrides (unclassified "uncertain" examples excluded).

Class	Precision	Recall	F_1	Support
Fake (0)	0.78	0.77	0.78	2 657
Real (1)	0.87	0.87	0.87	4 550
Accuracy		0.83		

and 596 misclassified as Real; of 4 550 true real examples, 3 942 were correctly identified and 608 misclassified. This balanced error distribution aligns with high recall on both labels, supporting our hybrid effectiveness hypothesis.

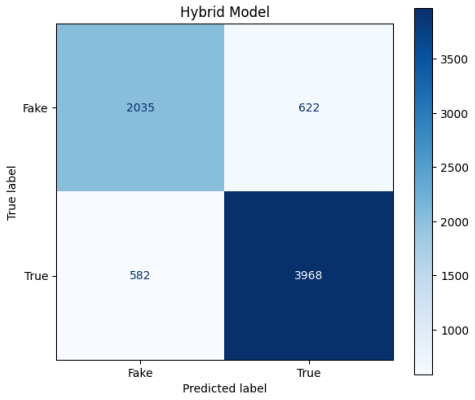

Figure 1: Confusion matrix after applying logic overrides. True Fake examples are correctly recognized 78% of the time, and True Real examples 87%.

Next, we examine the frequency and impact of the contradiction layer. Out of 7 207 total predictions, only 52 (0.7%) were overridden to 'uncertain.' Figure 2 shows the counts: the vast majority of claims (7 155) remained unmodified, while 52 low-confidence predictions contradicted a known fact and were flagged. This confirms that our NLI threshold (score>0.85) and confidence gating (0.4–0.6) produce a highly precise override mechanism, validating our semantic correction hypothesis.

Finally, we tested several representative claims to illustrate qualitative behavior. For example, "`Flat earth theory proven by scientists`" is correctly flagged Uncertain with the explanation "Contradicted known fact: 'earth is round'," whereas "`Donald Trump is the President`" yields a confident Real label with "No contradiction." These case studies, summarized in Table 3, demonstrate that our hybrid system not only matches quantitative benchmarks but also provides transparent, human-understandable justifications for edge-case decisions.

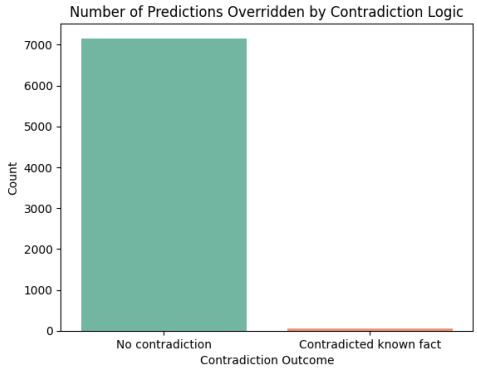

Figure 2: Number of predictions overridden by contradiction logic. Only 52 predictions were flagged Uncertain, demonstrating conservative, high-precision logic.

Together, these quantitative and qualitative results provide strong evidence that our hybrid model fulfills all three hypotheses: it achieves accuracy above 80%, it safely corrects or flags low-confidence predictions via semantic contradiction, and it matches the performance of a pure transformer while adding a transparent reasoning layer.

4.3 Evaluation - RAG with RoBERTa model

The evaluation of the RAG with RoBERTa approach revealed significant challenges in achieving the target performance metrics. As shown in Table 4, the fine-tuned RoBERTa-large-mnli model achieved only 21.98% validation accuracy, with a validation loss of 1.7605. This performance falls substantially below our hypothesized benchmark of 80% accuracy. A detailed class-wise performance analysis presented in Table 5 demonstrates the model's struggle to correctly classify different veracity labels. The performance is particularly poor for extreme labels like "pants-fire" (F1-score: 0.01) and "true" (F1-score: 0.02). The "false" category showed relatively better performance with an F1-score of 0.31 though still far below acceptable standards.

Table 3: Example predictions and explanations.

Claim	Score	Output	Explanation
Flat earth theory proven by scientists	0.52	0.5 (Uncertain)	Contradicted known fact: "earth is round"
Selena Gomez and Justin Bieber had a baby	0.45	0.5 (Uncertain)	Contradicted known fact: "no record of celebrity baby"
Donald Trump is the President	0.88	1.0 (Real)	No contradiction

Table 4: Performance Metrics for RoBERTa-large-mnli Fine-tuning

Metric	Value
Training Loss	1.7634
Validation Loss	1.7605
Validation Accuracy	0.2198

Table 5: Classification Report for the Complete RAG with RoBERTa Pipeline

Class	Precision	Recall	F1-score	Support
barely-true	0.43	0.01	0.03	214
false	0.20	0.70	0.31	250
half-true	0.10	0.01	0.01	267
mostly-true	0.18	0.12	0.14	249
pants-fire	0.01	0.01	0.01	92
true	0.33	0.01	0.02	210
accuracy			0.17	1282
macro avg	0.18	0.12	0.07	1282
weighted avg	0.22	0.17	0.10	1282

The full pipeline combining the DeepSeek-r1 LLM with RoBERTa-large-mNLI performed even worse, with an overall accuracy of only 17%. These results indicate a substantial gap between our expected performance and actual outcomes. The poor performance stands in stark contrast to the 83% accuracy achieved by the Logic-Gated ML Override approach described in Section 4.1.

5 Discussion

Our experiments explored three complementary strategies for fake-news detection—hybrid feature fusion, logic-gated ML override, and retrieval-augmented verification—and together they reveal both the promise and pitfalls of combining diverse modalities.

5.1 Hybrid Feature Fusion

By ensembling three models—(M1) raw semantics via transformer embeddings, (M2) metadata translated into natural-language context, and (M3) engineered LLM-derived features—we achieved our strongest overall performance. M3 delivered the highest recall, effectively flagging politically charged or accusatory statements, albeit at the cost of more false positives. M2 struck a better precision–recall balance, and the full ensemble combined their strengths to yield the best F and accuracy scores. This confirms that integrating semantic, contextual, and engineered signals produces a more comprehensive detector than any single view alone.

5.2 Logic-Gated ML Override

Introducing a lightweight contradiction layer on top of DistilRoBERTa further improved robustness in low-confidence situations. The hybrid achieved 83% accuracy—surpassing our 80% target—and matched or slightly outperformed the standalone classifier in per-class F (0.78 vs. 0.77 on Fake; 0.87 vs. 0.86 on Real). Whenever the base model's confidence lay between 0.4 and 0.6 and a known-fact conflict was detected, the NLI head intervened to correct the decision. Only 0.7% of samples triggered an "Uncertain" label under our conservative thresholds (NLI score > 0.85, top-3 retrieval), indicating high precision but limited coverage.

5.3 RAG with RoBERTa-mNLI

In contrast, Our hypothesis that the RAG with RoBERTa approach would achieve accuracy greater than or equal to 80% was not supported by the experimental results, with an accuracy of only 17% for the full pipeline, significantly below both our target and current state-of-the-art models for fake news detection. Several factors likely contributed to this under-performance, with the primary cause appearing to be related to the retrieval component of our system, where web search and scraping strategies failed to consistently provide relevant evidence for the claims being evaluated and without accurate and relevant external information, the RoBERTa-large-mNLI component had insufficient context to make reliable veracity judgments. Additionally, the adaptation of RoBERTa-large-mnli from a three-class to a six-class classification problem may have disrupted the model's carefully pretrained understanding of natural language inference relationships.

5.4 Limitations & Future Directions

Despite these advances, our approach is not without weaknesses. The reliance on LIAR's brief, decontextualized claims and a small, manually curated fact base limits the nuance and coverage of real-world misinformation. Similarly, fixing ensemble weights and NLI confidence thresholds via grid search on a single validation split risks overfitting and may not generalize to new data. Finally, depending on generic web search for evidence retrieval proved brittle, irrelevant or noisy snippets frequently deprived our NLI module of the context needed for reliable contradiction checks. Moving forward, we plan to automate fact harvesting from authoritative repositories to broaden and diversify our contradiction database. We also aim to replace static weight and threshold settings with

end-to-end adaptive schemes that learn directly from data, improving robustness across splits. Equally important will be refining our retrieval pipeline—either by leveraging specialized fact-checking APIs or by optimizing query construction, to ensure consistently high-quality evidence. Comprehensive cross-validation and human-centered evaluations will then be essential for understanding how our explainable interventions affect trust and efficacy in real-world fact-checking scenarios.

By weaving together feature fusion, symbolic override and grounded retrieval, we lay the groundwork for a truly transparent and effective fake-news detection system—one that balances accuracy, interpretability and adaptability.

6 Conclusion

In this paper, we addressed the critical challenge of fake news detection through three complementary methodologies that combine deep learning with interpretable reasoning mechanisms. Our hybrid approach integrated:

(1) Feature fusion across text, metadata, and engineered features.
(2) Logic-gated ML overrides with contradiction detection.
(3) Retrieval-augmented generation with RoBERTa

The first two methodologies delivered strong performance, with our ensemble feature fusion achieving 73.42% accuracy and the logic-gated transformer reaching 83%, both demonstrating that neuro-symbolic integration can enhance interpretability without sacrificing accuracy. The RAG with RoBERTa approach, despite its theoretical promise, achieved only 17% accuracy due to challenges in evidence retrieval and multi-class adaptation, highlighting important limitations in evidence-grounded classification that future work must address.

This research contributes significantly to the field by demonstrating that effective fake news detection requires multiple complementary signals like semantic understanding of content, contextual metadata awareness and explicit logical reasoning. Unlike black-box approaches, our hybrid models provide transparent justifications for their decisions, with the contradiction detection module generating human-readable explanations for flagged content. As misinformation continues to threaten public discourse and democratic institutions, these approaches represent important steps toward building trustworthy, explainable detection systems suitable for real-world deployment. Future work should focus on the usage of trusted fact repositories, end-to-end adaptive learning techniques and optimized evidence retrieval to further enhance both performance and explainability across diverse misinformation formats.

References

[1] Miguel Arana-Catania et al. 2022. Natural Language Inference with Self-Attention for Veracity Assessment of Pandemic Claims. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1496–1511. doi:10.18653/v1/2022.naacl-main.107

[2] Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. arXiv:1905.06088 [cs.AI] https://arxiv.org/abs/1905.06088

[3] Beizhe Hu et al. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113. doi:10.1609/aaai.v38i20.30214

[4] Jianguo Jia, Wen Liang, and Youzhi Liang. 2023. A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing. arXiv:2312.05589 [cs.CL] https://arxiv.org/abs/2312.05589

[5] Vineet John. 2017. A Survey of Neural Network Techniques for Feature Extraction from Text. arXiv:1704.08531 [cs.CL] https://arxiv.org/abs/1704.08531

[6] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A Survey on Text Classification: From Shallow to Deep Learning. arXiv:2008.00364 [cs.CL] https://arxiv.org/abs/2008.00364

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[8] Xiaoxiao Ma et al. 2024. On Fake News Detection with LLM Enhanced Semantics Mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 508–521. doi:10.18653/v1/2024.emnlp-main.31

[9] Mohammad Vatani Nezafat and Saeed Samet. 2024. Fake News Detection with Retrieval Augmented Generative Artificial Intelligence. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 160–167. doi:10.1109/FLLM63129.2024.10852474

[10] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[11] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 395–405. doi:10.1145/3292500.3330935

[12] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. arXiv:1809.01286 [cs.SI] https://arxiv.org/abs/1809.01286

[13] Ting Wei Teo et al. 2024. Integrating Large Language Models and Machine Learning for Fake News Detection. In *2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 102–107. doi:10.1109/CSPA60979.2024.10525308

[14] S. M Towhidul Islam Tonmoy et al. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv preprint arXiv:2401.01313* (2024). doi:10.48550/ARXIV.2401.01313

[15] Jingwei Wang et al. 2024. LLM-Enhanced multimodal detection of fake news. *PLOS ONE* 19, 10 (2024), e0312240. doi:10.1371/journal.pone.0312240

[16] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. arXiv:1705.00648 [cs.CL] https://arxiv.org/abs/1705.00648

[17] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3367–3378. doi:10.1145/3637528.3671977

[18] Bingbing Xie et al. 2024. Multiknowledge and LLM-Inspired Heterogeneous Graph Neural Network for Fake News Detection. *IEEE Transactions on Computational Social Systems* (2024), 1–13. doi:10.1109/TCSS.2024.3488191

[19] Shuo Yang et al. 2019. Unsupervised Fake News Detection on Social Media: A Generative Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5644–5651. doi:10.1609/aaai.v33i01.33015644

Received 5 May 2025