Cognitive Threat Detection for SOC Operations: Automating Manipulation Tactic Analysis in Election Security

Keerthana Madhavan
School of Computer Science
University of Guelph
Guelph, ON, Canada
kmadhava@uoguelph.ca

Luiza Antonie
School of Computer Science; CARE-AI
University of Guelph
Guelph, ON, Canada
lantoine@uoguelph.ca

Stacey D. Scott
School of Computer Science; CARE-AI
University of Guelph
Guelph, ON, Canada
stacey.scott@uoguelph.ca

***Abstract*—Election security Security Operations Centers (SOCs) face an expanding mandate: beyond traditional network defense, they must now detect cognitive threats, content that manipulates audiences through psychological tactics rather than explicit falsehoods. Existing tools provide binary labels without explaining how manipulation occurs, limiting triage and response. We present E-MANTRA, a Large Language Model (LLM)-based framework that integrates agentic Artificial Intelligence (AI) into SOC workflows by identifying six manipulation tactics (emotional manipulation, conspiracy framing, discrediting, trolling, impersonation, polarization) with explainable classifications. Evaluated on 900 election-related samples, E-MANTRA attains 54.2% triage accuracy and an estimated 57% workload reduction under confidence-based decision-making. Results confirm exploitable model specialization: Llama-3 70B excels at conspiracy detection (F1=0.71), GPT-3.5 at emotional manipulation (F1=0.66), Mistral-Small at discrediting (F1=0.63). Category-aware routing improves accuracy by 2.4 percentage points over the best single model at $0.005 per classification. We provide a practitioner-oriented deployment checklist, cost models, and Security Information and Event Management (SIEM)/Security Orchestration, Automation, and Response (SOAR) integration guidelines to support operational adoption in election security SOCs.**

***Index Terms*—SOC operations, cognitive security, election infrastructure, manipulation detection, LLM deployment, threat intelligence**

I. Introduction

Election security has shifted from a primarily technical problem to a cognitive one. During recent election cycles, adversaries have weaponized social media to manipulate voters through psychological tactics rather than explicit falsehoods, exploiting fear, anger, and institutional distrust to erode confidence in democratic processes. In 2020, for example, the "Sharpiegate" incident saw false narratives about ballot-handling procedures circulate widely despite official assurances, demonstrating how emotionally charged but factually thin claims can still mobilize protests and intimidate election workers [1]. National guidance now explicitly treats election infrastructure as a high-value cyber target, urging election authorities to adopt dedicated security capabilities, structured monitoring, and formal incident response processes [2], [3].

This guidance increasingly recognizes that election operations must be run like security-critical environments. Federal and sector bodies provide playbooks for incident response, communications, and social-engineering defense that assume some form of continuous monitoring and coordinated response, even for smaller jurisdictions [4]–[6]. Private organizations that administer high-stakes elections, such as member votes and corporate governance ballots, are likewise being urged to adopt more formalized election-security practices even though they lack government-grade resources [7]. While dedicated "election SOCs" remain an emerging concept rather than an established standard, analogous capabilities are increasingly embedded within state Fusion Centers, the Elections Infrastructure Information Sharing and Analysis Center (EI-ISAC), and ad-hoc election security task forces [8]. These election-security teams face a dual mandate: protect technical infrastructure (registration systems, result reporting, networks) while also detecting and countering cognitive threats that can suppress turnout or delegitimize outcomes.

This mandate collides with a crisis of scale. Consider a representative scenario: at 10,000 flagged posts per day, a plausible mid-scale load for monitoring election-related hashtags across a single state, and a conservative estimate of five minutes of manual review per post, full human coverage would require more than 800 analyst-hours daily, far beyond realistic staffing for most election-security operations. Empirical studies report that alert overload drives SOC analyst burnout and turnover [9], [10], degrading institutional expertise precisely when election security requires continuity and experience. Industry analyses further indicate that nearly 70% of SOC analyst time is consumed by triage [11], leaving limited capacity for proactive threat hunting, campaign-level analysis, or process improvement.

Workshop on Security Operation Center Operations and Construction (WOSOC) 2026
23 February 2026, San Diego, CA, USA
ISBN 978-1-970672-03-9
https://dx.doi.org/10.14722/wosoc.2026.23010
www.ndss-symposium.org

Current detection approaches exacerbate this problem. Most misinformation and fake-news systems treat content as a binary classification task (true vs. false, harmful vs. benign) without explaining how manipulation operates or what response it demands. Surveys and systems work on misinformation, propaganda, and fake-news detection confirm that even when models reach respectable accuracy, their outputs remain coarse labels that offer little tactical insight for operations teams [12]–[14]. For election-security SOCs, this binary framing is operationally blind: a voter-suppression campaign using fear appeals and a coordinated impersonation campaign spoofing election officials both receive similar "suspicious content" labels, despite requiring fundamentally different countermeasures.

Large language models offer a potential path beyond this binary paradigm. Unlike traditional Natural Language Processing (NLP) classifiers relying on bag-of-words or keyword features, LLMs process language through self-attention mechanisms that model contextual relationships, enabling detection of manipulation tactics requiring interpretation beyond surface patterns [15]. Recent work demonstrates LLM capabilities for rhetorical tactic classification, deception detection, and propaganda analysis (reviewed in Section II), suggesting potential for SOC integration. However, deploying LLMs in security-critical election environments raises practical questions: Can they reliably detect manipulation tactics? Do different model architectures exhibit useful specializations? And can LLM-based systems integrate into SOC workflows at acceptable cost?

To address these questions, we introduce E-MANTRA (**E**lection-based **M**anipulation **A**nnotation for **N**arrative **T**actic **R**ecognition and **A**nalysis), an LLM-driven framework that reconceptualizes misinformation detection as tactical intelligence generation. E-MANTRA classifies social media content into six manipulation tactics grounded in cognitive psychology [16]: emotional manipulation, conspiracy framing, discrediting, trolling, impersonation, and polarization. Each classification includes an explanation of the manipulation mechanism, enabling analysts to select appropriate countermeasures.

Our central research question is: *Can LLM-based manipulation tactic detection reduce SOC analyst workload while providing actionable intelligence for election security operations?* We decompose this into three focused inquiries: (RQ1) To what extent can confidence-based triage automate high-confidence classifications while preserving accuracy for security-critical decisions? (RQ2) Do different LLM architectures exhibit complementary strengths for detecting specific manipulation tactics, and can routing posts to specialist models improve overall performance? (RQ3) What confidence thresholds, cost structures, and integration workflows enable practical deployment in election SOCs?

Evaluating six LLMs on 900 election-related public samples, we find that model specialization is real and exploitable. No single model dominates all categories: Llama-3 70B excels at conspiracy detection (F1=0.71), GPT-3.5 at emotional manipulation (F1=0.66) and impersonation (F1=0.51), and Mistral-Small at discrediting (F1=0.63). Category-aware routing to specialist models improves accuracy by 2.4 percentage points (˜4.6% relative) over the best single-model baseline. A three-tier triage architecture (automated, AI-assisted, manual) routes 57% of posts to full automation at high confidence, yielding an estimated 57% reduction in analyst review volume. At $0.005 per classification (6 runs for confidence estimation), a mid-scale SOC processing 10,000 posts daily incurs $1,500 monthly—while potentially saving hundreds of analyst-hours.

Our contributions are fourfold:

- We address an *underserved SOC domain*, election security operations, by presenting E-MANTRA as a tactical intelligence framework that provides explainable, tactic-level classifications and moves beyond binary detection toward actionable SOC intelligence.
- We document *model specialization patterns* across six large language models, demonstrating that heterogeneous deployment outperforms single-model approaches.
- We propose a *confidence-based triage architecture* with empirically validated thresholds that balance the benefits of automation against the risks of error.
- We provide a *practitioner-oriented deployment checklist* (Section VII) that includes cost models, SIEM and SOAR integration guidelines, and risk mitigation strategies designed for immediate operational adoption.

II. Background and Related Work

A. The DEPICT Framework for Manipulation Tactics

Effective cognitive threat detection requires a principled taxonomy of manipulation techniques. We adopt van der Linden's DEPICT framework [17]—**D**iscrediting, **E**motional manipulation, **P**olarization, **I**mpersonation, **C**onspiracy, and **T**rolling—which identifies six tactics grounded in cognitive psychology research. Each tactic provides distinct detection signatures for automated classification:

Discrediting erodes trust through systematic credibility attacks on institutions or individuals, often using ad hominem arguments without substantive evidence.

Emotional manipulation weaponizes fear, anger, or outrage to bypass critical evaluation. Emotionally charged content spreads significantly faster than neutral information [18], making early detection critical for limiting amplification.

Polarization exploits group identity to transform discourse into tribal conflict, mobilizing existing supporters rather than persuading undecided audiences.

Impersonation fabricates authority through fake credentials or personas, often requiring coordination across platforms to maintain consistency.

Conspiracy theories construct alternative narratives that are internally consistent yet unfalsifiable, typically requiring sustained effort to develop and distribute.

Trolling disrupts discourse through provocative content optimized for engagement, exhausting defenders and polluting information spaces.

Unlike binary misinformation labels, these tactics inform response selection: emotional manipulation may warrant inoculation messaging, while impersonation requires identity verification and platform coordination.

B. Limitations of Current Detection Approaches

Existing misinformation detection systems primarily pursue binary classification (true/false, harmful/benign). Zhou and Zafarani's survey [14] documents approaches achieving 70–85% accuracy through linguistic analysis and network modeling, yet these systems provide no insight into *how* content manipulates or *what* response it warrants, leaving analysts without actionable intelligence for differentiated response.

Propaganda detection research offers finer granularity. Da San Martino et al. [19] achieve 50–55% F1 on 18-category propaganda classification, and SemEval-2020 [13] advanced multi-label detection. However, these approaches operate on curated news datasets rather than real-time social media, and none address confidence calibration for security-critical decisions or integration with SOC triage workflows.

Commercial SOAR platforms (Cortex XSOAR, Splunk Phantom) automate up to 80% of Level-1 SOC tasks [20], but operate on structured indicators (IPs, hashes, domains)—not semantic understanding of manipulation. The automation capability exists; the tactical intelligence to drive it does not.

C. LLMs for Cognitive Security

Large language models offer semantic understanding beyond pattern matching [21], [22]. Unlike keyword-based systems, LLMs can interpret narrative framing, rhetorical strategy, and psychological intent—detecting emotional manipulation in neutral-appearing language or identifying conspiracy framing in factual presentations [23], [24]. Recent studies confirm that LLMs achieve parity with fine-tuned models on propaganda detection [25], while chain-of-thought prompting enables step-by-step reasoning for nuanced classification tasks [26].

Beyond general capabilities, recent work demonstrates LLM potential for security applications specifically. Krašovec et al. [27] apply LLMs to threat intelligence processing; Casino et al. [28] explore cognitive security applications. However, neither addresses the operational requirements of election SOCs: confidence calibration for security-critical decisions, workflow integration, or economic viability at scale.

E-MANTRA addresses these gaps by providing (1) tactic-level classification rather than binary labels, (2) confidence-based triage for workload reduction, and (3) deployment guidance including cost models and integration workflows. Our evaluation reveals systematic model specialization, where different architectures excel at different tactics, enabling heterogeneous deployment that improves both accuracy and potential adversarial robustness.

III. Threat Model

Election security SOCs face adversaries ranging from state-sponsored disinformation units [29] to domestic operatives and commercial opportunists engaged in organized social media manipulation. These actors deploy cognitive manipulation tactics targeting human decision-making rather than technical vulnerabilities, aiming to suppress turnout, undermine electoral legitimacy, or amplify divisions [30]. Research characterizes these operations as collaborative, participatory efforts that exploit platform affordances and audience vulnerabilities [31].

Adversary Capabilities. We assume adversaries: (1) understand platform algorithms and moderation policies, designing content that maximizes reach while evading filters; (2) can generate high-volume coordinated campaigns across multiple accounts; (3) are aware that detection systems may be deployed and may adapt tactics accordingly; and (4) can evolve approaches in response to countermeasures.

Manipulation Taxonomy. We operationalize the six DEPICT tactics (Section II) as classification targets. This taxonomy was selected for its grounding in cognitive psychology and its operational relevance: each tactic suggests distinct countermeasure strategies.

Heterogeneous Defense. Adversaries aware of detection systems may develop evasions targeting specific model weaknesses, such as exploiting Llama-3 70B's limitations in emotional manipulation detection. Category-aware routing to specialist LLMs complicates adaptation: evasions optimized for one architecture may fail against alternatives with different training data and reasoning strategies. Systematic adversarial robustness evaluation remains future work (Section VIII).

IV. Methodology and Experimental Design

A. Dataset Construction

We constructed a 900-sample benchmark dataset through stratified sampling from three established misinformation corpora: FakeNewsNet [32] (300 news articles with fact-checker labels), Twitter Elections Study [33] (300 tweets from November 2020–January 2021), and LIAR [34] (300 political statements with PolitiFact ratings). Critically, these source datasets provide only binary veracity labels (true/false) or credibility ratings—none include manipulation tactic annotations. This label gap necessitated the GPT-4-based annotation process described in Section IV-B to generate the six-category DEPICT tactic labels required for our evaluation. Preprocessing applied Personally Identifiable Information (PII) anonymization (412 instances removed), deduplication (587 entries eliminated), and quality filtering (minimum 5 tokens, substantive content only), ensuring privacy compliance while preserving manipulation indicators. The final dataset exhibits realistic class imbalance: Emotional manipulation (28%), Polarization (22%), Discrediting (18%), Conspiracy (15%), Trolling (12%), and Impersonation (5%).

B. Ground Truth Annotation (Phase 1)

We established ground truth labels using Azure OpenAI GPT-4 (version 2024-04-09) through a systematic comparison of three prompting paradigms, building upon the prompt engineering validation framework established in our preliminary study [35]. **Zero-Shot** (task instructions only, 2,015

tokens/instance, $0.06 cost) achieved 90.8% majority agreement but exhibited severe category bias (39.4% Emotion, 32.3% Discrediting). **Few-Shot** (12–18 examples, 9,830 tokens/instance, $0.30 cost) suffered catastrophic failure with 53% collapse to "Inconclusive" and 337 parsing errors across 900 samples, a specific failure mode diagnosed in our prior error analysis [36]. **In-Context Learning** (ICL: comprehensive guidance with definitions, examples, and reasoning templates; 9,111 tokens/instance, $0.27 cost) achieved optimal performance: 85.8% majority agreement, zero parsing failures, balanced category distribution, and BERTScore semantic consistency of 0.89. All costs reported in this section are in USD. Based on this validation, we selected ICL for final ground truth annotation, achieving Cohen's $\kappa = 0.81$ inter-run agreement across the six independent GPT-4 classifications per sample.

Epistemic Scope of Annotation. The manipulation-tactic labels produced in Phase 1 are synthetic, model-generated references used to provide a consistent benchmark for comparative evaluation rather than human-validated detection truth. However, human analysts were involved in reviewing all "Inconclusive" outputs, validating the seven-category error taxonomy, and assigning error types for robustness analysis. Thus, human verification in our study applies to model failure characterization and ambiguity resolution rather than validating tactic labels at scale. This mirrors evaluation practices in explanation-quality research (e.g., FinGrAct [37]), where human review grounds failure modes without requiring full human annotation of underlying class labels.

This human review does not replace the synthetic manipulation-tactic labels; instead, it ensures that parsing errors, failure types, and ambiguous outputs are grounded in analyst judgment.

C. Model Evaluation (Phase 2)

Building on Phase 1's selection of In-Context Learning (85.8% reliability, κ=0.81), Phase 2 evaluates six diverse LLMs to establish baseline detection accuracies, identify model-specific strengths, and assess deployment implications.

Model Selection. We selected models representing architectural diversity and practical deployment constraints: GPT-3.5-Turbo (commercial API baseline), Llama-3 70B and Mistral-Small (open-weight alternatives for data sovereignty), DeepSeek-R1 (reasoning-optimized), and Phi-4/Phi-4-Mini (compact models for resource-constrained deployment). This portfolio spans 3.8B–175B parameters, testing whether model scale correlates with tactic detection capability.

All models were accessed via Azure AI Foundry [38] with standardized parameters (temperature=1.0, top-p=1.0, max_tokens=800) and 6 independent classifications per sample through isolated API threads. Confidence scores derive from multi-sampling consistency [26]: the proportion of runs agreeing with the majority label (e.g., 5/6 agreement = 0.83 confidence), providing empirically grounded uncertainty estimates robust to model miscalibration.

Data Sovereignty: DeepSeek-R1 evaluations used locally-hosted weights via Azure AI Foundry without external API dependencies. For operational deployment, we recommend on-premise hosting or substitution with fine-tuned Llama-3/Mistral variants if geopolitical constraints apply.

Evaluation Metrics. We selected metrics addressing both classification performance and operational reliability: (1) *per-category F1 scores* to handle class imbalance [39] (Impersonation: 5% vs. Emotion: 28%). F1 is the harmonic mean of precision (what fraction of detected items are correct) and recall (what fraction of actual items are detected), ranging from 0 to 1, where higher values indicate better performance; it is preferred over raw accuracy when categories have unequal sizes. (2) *Cohen's* κ for inter-run agreement beyond chance [40]; (3) *Cramér's V* for effect size independent of sample size; and (4) *BERTScore* [41] for semantic consistency of explanations across runs. Confidence thresholds (0.80, 0.40) were empirically derived through precision-recall analysis on a held-out validation set, targeting 90%+ precision for Tier 1 automation.

V. System Architecture and Results

A. LLM-Based Detection Pipeline

The E-MANTRA framework integrates three core components for operational SOC deployment (Figure 1). First, the **Preprocessing** stage ingests post streams from social media monitoring systems, performing PII scrubbing and content normalization to ensure privacy compliance while preparing data for analysis. Second, the **Tactic Recognition Engine** employs In-Context Learning (ICL) with a curated prompt bank to guide LLM classification across N=6 independent iterations, generating manipulation tactic predictions for each post. Third, the **Consistency Oracle** aggregates these N=6 classifications through majority vote while computing semantic consistency via BERTScore to derive confidence scores robust to model miscalibration.

The framework then routes posts through confidence-based thresholds (τ) to three operational tiers (Table III): **Tier 1 (Auto-Queued with Deferred Audit)** handles high-confidence cases (Conf ≥ 0.80) for rapid triage while remaining subject to retrospective human audit and campaign-level validation (57% volume); **Tier 2 (AI-Assisted)** presents medium-confidence predictions (0.40–0.79) with explanations for analyst validation (24% of volume); and **Tier 3 (Manual)** routes low-confidence cases (Conf < 0.40) to full analyst review (19% of volume). This architecture enables seamless integration with existing SOC SIEM/SOAR platforms while maintaining human oversight for ambiguous cases.

B. Model Specialization and Performance

We evaluated six diverse LLMs to assess specialization patterns (Table I):

Key finding: No single model achieves top-3 performance across all categories. Llama-3 70B dominates Conspiracy detection (F1=0.71), GPT-3.5 excels at Emotion (F1=0.66), Impersonation (F1=0.51), and Polarization (F1=0.54), Mistral-Small leads Discrediting (F1=0.63), while DeepSeek-R1 shows strongest Trolling detection (F1=0.48). This systematic

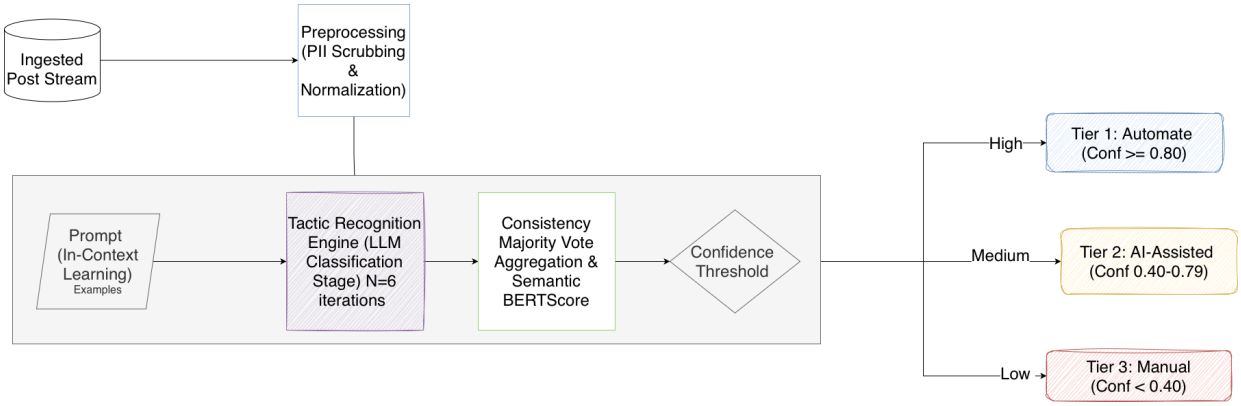

Fig. 1: E-MANTRA Framework Architecture. Ingested posts undergo preprocessing (PII scrubbing), then multi-iteration tactic classification through the LLM engine guided by ICL examples. The Consistency Oracle aggregates N=6 runs via majority vote and BERTScore semantic validation, producing confidence scores that route posts to three operational tiers: Tier 1 (auto-queued with deferred audit, Conf $\geq$ 0.80, 57% volume), Tier 2 (AI-assisted, 0.40–0.79, 24%), and Tier 3 (manual, Conf $<$ 0.40, 19%).

TABLE I: Comprehensive Model Evaluation Across Six LLMs

Model	Overall	Consp.	Emot.	Disc.	Troll.	Impers.	Polar.
GPT-3.5	**51.8%**	0.46	**0.66**	0.46	0.35	**0.51**	**0.54**
Mistral-Small	51.7%	0.34	0.63	**0.63**	0.24	0.11	0.41
DeepSeek-R1	47.5%	0.64	0.64	0.50	**0.48**	0.41	0.46
Llama-3 70B	38.6%	**0.71**	0.57	0.46	0.28	0.16	0.25
Phi-4	40.6%	0.56	0.61	0.34	0.10	0.23	0.38
Phi-4-Mini	28.2%	0.26	0.55	0.15	0.09	0.18	0.27
Category-Aware Routing (Best model per category)							
Hybrid	**54.2%**	**0.71**	**0.66**	**0.63**	**0.48**	**0.51**	**0.54**

Notes: Abbreviations: Consp. (Conspiracy), Emot. (Emotional manipulation), Disc. (Discrediting), Troll. (Trolling), Impers. (Impersonation), Polar. (Polarization). Hybrid routing achieves 54.2% accuracy (+2.4 pp or ˜4.6% relative vs best single model). F1 scores are point estimates on 900 samples.

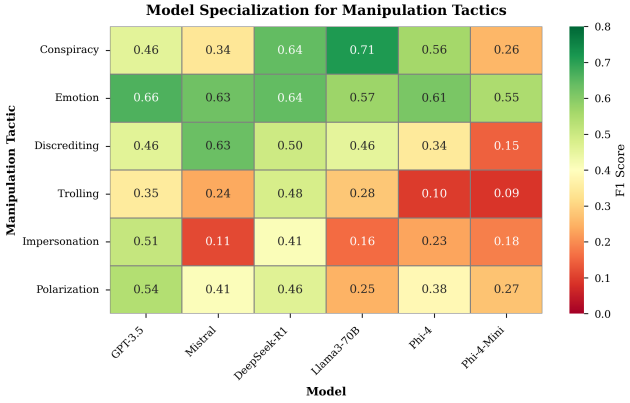

Fig. 2: Model Specialization Heatmap: F1 scores across manipulation tactics reveal distinct model strengths. Bold values indicate best performer per category.

specialization enables heterogeneous deployment with +2.4 percentage points (˜4.6% relative) accuracy improvement. For context, these F1 scores (0.48–0.71) are comparable to prior work on fine-grained propaganda detection, which reports 0.50–0.55 F1 across 18 categories [42]; our confidence-based triage ensures that lower-scoring categories receive human review rather than fully automated decisions.

Implications for SOC Deployment. The observed model specialization highlights the importance of aligning SOC priorities with model strengths. For instance, SOCs prioritizing conspiracy detection may benefit from deploying Llama-3 70B, while those focusing on emotional manipulation could leverage GPT-3.5. Hybrid routing strategies enable SOCs to optimize performance by dynamically assigning tasks to models best suited for specific manipulation tactics.

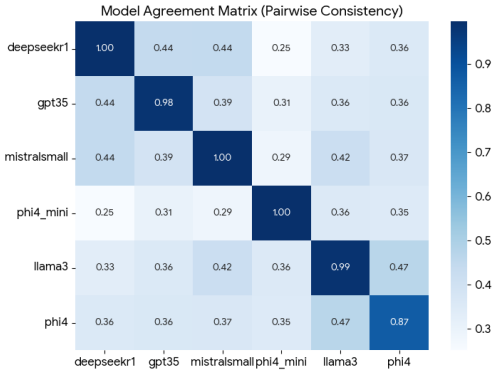

Fig. 3: Pairwise Agreement Matrix: Low mean agreement ($\mu = 0.38$) validates ensemble approach—models detect orthogonal manipulation features.

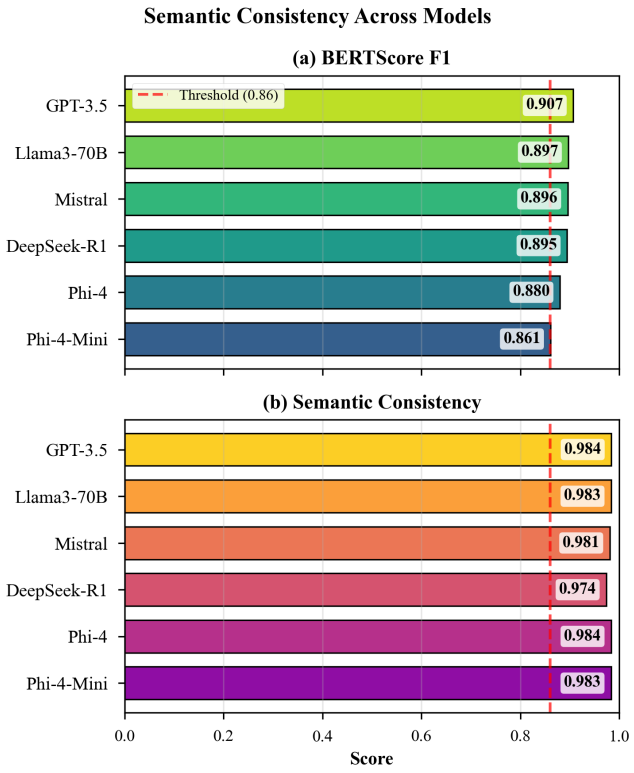

Fig. 4: BERTScore Semantic Consistency: All models achieve F1 $\geq$ 0.86, indicating stable reasoning across independent runs.

C. Architectural Specialization and Ensemble Validity

To explain the mechanics of this improvement, we observe that LLMs are not 'generalists' in cognitive security; rather, they exhibit distinct behavioral profiles that necessitate heterogeneous deployment. Analysis of 5,400 independent classifications (900 samples $\times$ 6 models) reveals significant architectural specialization (Figure 2). GPT-3.5 demonstrates high sensitivity to Impersonation (F1=0.51) and Polarization (F1=0.54), whereas Mistral-Small specializes in Discrediting (F1=0.63), and Llama-3 70B excels at Conspiracy detection (F1=0.71).

This divergence is further quantified in Figure 3, which presents the pairwise agreement matrix between models. The mean pairwise agreement remains low ($\mu = 0.38$), with no two models exceeding 0.47 agreement (excluding self-correlation). This low consistency validates the hypothesis that heterogeneous LLMs detect distinct linguistic and semantic features of cognitive manipulation. If models were highly correlated, an ensemble approach would yield diminishing returns; instead, the orthogonality of these detection vectors supports the use of a category-aware routing architecture.

Comparison against ground truth labels (N=900) confirms that this heterogeneous routing improves detection balance. While single models like Llama-3 70B suffer in specific categories such as Impersonation (F1=0.16), the hybrid architecture maintains a floor of F1 $\geq$ 0.48 across all categories by routing each category to its best-performing model. Furthermore, the ensemble majority vote successfully normalizes individual model biases, recovering a class distribution that closely mirrors the ground truth dataset composition. These results confirm that the specialist capability of individual models can be effectively harvested to create a robust, general-purpose detection system.

D. Validation Results

Overall system performance demonstrates substantial improvement over baseline approaches. Single-model deployments achieve 51.8% accuracy (GPT-3.5), while category-aware routing improves performance to 54.2%. These results achieve statistical significance ($p < 0.001$, Cramér's V = 0.23–0.47) and represent a ˜3.2$\times$ improvement over a random 16.7% baseline.

A critical consideration for SOC deployment is whether LLM explanations are trustworthy enough to support security-critical analyst decisions. Figure 4 shows all six models achieve high semantic consistency (BERTScore F1 $>$ 0.86) with low variance (std=0.016–0.026), indicating stable reasoning across independent classifications. This high consistency indicates linguistic stability in generated explanations across repeated runs, though it does not guarantee factual correctness or causal faithfulness to the underlying manipulation process.

Error analysis reveals that technical failures (such as JSON parsing errors or timeout issues) are addressable through prompt engineering refinements. Semantic failures primarily involve context insufficiency (15%), multi-tactic overlap where posts employ multiple simultaneous manipulation strategies (12%), and subtle framing that requires deep contextual understanding (10%).

E. Case Study: Explainable Tactic Detection

To demonstrate E-MANTRA's operational utility, consider the Sharpiegate incident from Section 1. A post claiming *"They're using Sharpies to invalidate your ballot in Arizona! Don't let them steal your vote!"* would be classified as **Emotional Manipulation** (expected confidence $\geq$0.83, based on similar fear-appeal patterns in our corpus), with explanation: *"Fear-inducing language ('invalidate', 'steal your vote') transforms mundane procedural details into existential threats, characteristic of voter suppression tactics."* This tactic signature, combining emotional appeals with procedural misinformation, enables targeted SOC response: rapid election official PSAs rather than content takedowns.

F. Operational Deployment Architecture

Cost Model: Our confidence estimation requires 6 independent classification runs per post (Section 4.2). Table IV summarizes per-model costs and latencies. For GPT-3.5-Turbo at 2025 pricing ($0.50/1M input, $1.50/1M output) with 1,200 input tokens and 150 output tokens, single-run cost is $0.000825, yielding $0.005 per post ($\times$6 runs). Production deployments may optimize to 3–4 runs for 40–50% cost reduction. Small SOCs (1k posts/day) incur $150 monthly, medium SOCs (10k/day) $1,500 monthly, large SOCs (100k/day) $15,000

TABLE II: E-MANTRA Classifications from Benchmark Dataset (N=900)

Sample Text	Tactic	Conf.	System Explanation
"...convicted liar lawyer...media scum reporters..." (Twitter Elections Study [**?**], #13)	Discrediting	1.00	Character attacks without evidence using derogatory language; ad hominem targeting individuals and media institutions to undermine credibility
"Democrats suffered crushing loss...massive fraud..." (Twitter Elections Study [**?**], #135)	Conspiracy	1.00	Suggests secret plot to rig election; connects real event (election loss) with implausible fraud claim, a hallmark of conspiracy narratives

Notes: Confidence scores (0–1.00) based on 6-run majority agreement. Sample IDs enable reproducibility. Classifications and explanations generated by GPT-4 (Phase 1 ground truth annotation).

TABLE III: Confidence-Based Triage Architecture

Tier	Confidence	Volume	Action
Tier 1 (Auto)	≥ 0.80	57%	Deferred audit
Tier 2 (Assisted)	0.40–0.79	24%	AI + analyst
Tier 3 (Manual)	< 0.40	19%	Full review

TABLE IV: Model Cost and Latency Comparison (6 runs/post)

Model	Cost/Post	Latency
GPT-3.5-Turbo	\$0.005	14.3s
Mistral-Small	\$0.005	**9.8s**
DeepSeek-R1	\$0.007	56.2s
Llama-3 70B	\$0.005	19.4s
Operational Estimates (10k posts/day):		
Monthly cost		\$1,500
Workload reduction		57% (est.)
Analyst hours saved/day		475 (est.)

Notes: Costs assume 2025 API pricing. Analyst time based on 5 min/post manual review; actual savings depend on organizational context.

monthly. For 10k posts/day with 5 min/review baseline, the estimated workload reduction saves 475 analyst hours daily, representing potential time savings rather than direct monetary conversion, as actual value depends on organizational context and analyst redeployment effectiveness.

G. Model Selection for Deployment

Budget-constrained SOCs benefit from single-model deployment using GPT-3.5 or Mistral-Small (10,000+ accuracy points per dollar). Mistral-Small offers 31% faster processing (9.8s vs 14.3s) for high-throughput scenarios (see Table IV).

Sophisticated SOCs can exploit heterogeneous deployment through category-aware routing: posts are directed to specialist models based on predicted category—Llama-3 70B for conspiracy (F1=0.71), GPT-3.5 for emotional manipulation (F1=0.66), impersonation (F1=0.51), and polarization (F1=0.54), Mistral-Small for discrediting (F1=0.63), and DeepSeek-R1 for trolling (F1=0.48). This hybrid approach achieves 54.2% accuracy (+2.4 pp or ˜4.6% relative improvement) and delivers superior F1 performance (Avg F1=0.59 vs 0.50 best single-model), representing an 18% relative F1 improvement at the same \$0.005 per-classification cost.

VI. Threat Intelligence Applications

The tactic-level classifications produced by E-MANTRA suggest several potential threat intelligence applications. While validating these applications requires operational deployment with labeled threat actor data, we include this discussion because SOC practitioners consistently identify threat actor attribution and campaign detection as key requirements for cognitive security tools [**?**]. We outline hypothesized use cases to illustrate how tactic-level outputs could address these operational needs.

A. Threat Actor Profiling

Manipulation tactic patterns may provide operational intelligence for attributing content to adversary classes. The model specialization patterns in Table I suggest distinct adversary signatures: sophisticated state-sponsored actors may favor conspiracy-impersonation combinations (exploiting Llama-3 70B's strength), while domestic operatives tend toward emotional-polarization overlap (where GPT-3.5 excels). The low pairwise model agreement ($\mu = 0.38$, Figure 3) implies that different adversary types may be detectable by which models agree or disagree on classification. Full attribution requires external intelligence sources beyond the scope of this dataset, but tactic-level outputs provide a foundation for pattern-based profiling.

B. Campaign Detection

Beyond individual post classification, temporal correlation of tactic distributions may enable detection of coordinated multi-vector campaigns. We hypothesize three campaign signatures based on tactic co-occurrence patterns:

Institutional trust erosion: Synchronized surges in discrediting (targeting election officials) combined with conspiracy narratives (questioning vote integrity). Response: Official statements and platform escalation.

Voter suppression: Coordinated emotional fear appeals, impersonation of election authorities, and trolling to create procedural confusion. Response: Rapid PSAs clarifying voting logistics.

Polarization amplification: Simultaneous us-versus-them framing targeting both political bases. Response: Cross-partisan counter-messaging.

Validating these signatures requires labeled campaign data from operational deployments, which we identify as priority future work.

VII. Integration Guidelines and Recommendations

A. Deployment Checklist

Pre-deployment validation requires: (1) API access and cost validation through 1,000-sample pilot testing, (2) analyst calibration to establish confidence thresholds aligned with risk tolerance, (3) monitoring dashboards tracking tactic distributions and override rates, and (4) analyst training on explanation interpretation.

Operational monitoring maintains performance through key indicators: analyst override rates (<15% for Tier 1), tactic distribution shifts for emerging campaign detection, workflow analytics validating time savings, and weekly calibration sessions maintaining human-AI trust.

B. Risk Mitigation

Three primary risks require mitigation. Adversarial adaptation, the most significant threat, is addressed through category-aware routing providing diversity defense, as evasions optimized for one model architecture fail to transfer to specialized alternatives. API dependency is mitigated through open-source alternatives (Llama-3 70B) enabling self-hosting, multi-vendor redundancy, and contractual Service Level Agreements (SLAs)[1]. Bias concerns are addressed through mandatory Tier 2–3 human oversight, periodic audits stratified by political orientation, and diverse analyst teams.

VIII. Limitations and Future Work

A. Current Limitations

Deployment Maturity: E-MANTRA represents proof-of-concept validation rather than production-tested system. Field validation with operational election security SOCs is required to verify analyst acceptance, SIEM/SOAR integration compatibility, election-day surge performance, and adversarial robustness.

Performance Ceiling: 54.2% accuracy suits confidence-based triage with human oversight but not full automation. We report overall accuracy and per-category F1 scores without tier-specific precision/recall breakdowns; operational deployment requires validating Tier 1 ($\geq$0.80 confidence) achieves target 90%+ precision.

Statistical Precision: Point estimates for F1 scores lack confidence intervals due to space constraints. Given class imbalance, rare category scores (e.g., Trolling F1=0.09–0.48) likely have wider intervals (±0.12–0.15) than common categories (±0.06–0.08).

Synthetic Ground Truth: While our error taxonomy and failure analyses are human-validated, the underlying manipulation-tactic class assignments remain synthetic and have not been validated against large-scale human ground truth. Consequently, our results should be interpreted as benchmark comparisons rather than absolute detection accuracy.

Political Neutrality: While annotation guidelines focus on manipulation tactics rather than political orientation, we have not conducted empirical bias audits. Operational deployment requires: (1) stratified bias analysis by political orientation, (2) parity metrics, and (3) diverse analyst review teams.

Additional Constraints: Dataset from 2016–2021 may not represent current tactics (AI-generated content, deepfakes). English-only processing limits international applicability. Economic analysis assumes stable API pricing.

[1]A Service Level Agreement is a formal contract with an API or cloud provider that specifies guaranteed levels of service, typically including availability, performance, and support.

B. Future Enhancements

Future work should prioritize operational field testing with election security SOCs to validate analyst workflows and measure actual workload reduction. Systematic red-team evaluation is needed to quantify adversarial robustness and validate heterogeneous deployment resilience. The framework should extend to multi-label classification (12% of content exhibits multi-tactic overlap), evaluation on 2024+ election data, multi-modal image/video analysis, and temporal correlation for real-time coordinated campaign detection.

IX. Conclusion

E-MANTRA addresses critical SOC operational challenges for election security by automating cognitive threat analysis with tactic-level outputs. Our confidence-based workflow achieves an estimated 57% reduction in analyst review volume through three-tier routing that preserves human oversight for ambiguous cases. At $0.005 per classification, mid-scale deployment processing 10,000 posts per day costs $1,500 monthly while enabling 24/7 processing. This cost structure makes LLM-based triage economically viable for election security operations of varying scale.

Heterogeneous routing leveraging specialized models improves accuracy to 54.2%, a 2.4 percentage point gain over the best single-model baseline. This improvement stems from systematic model specialization: no single LLM dominates all manipulation categories, but category-aware routing harvests each model's strengths. This architectural diversity may also provide resilience against adversarial evasion, as attacks optimized for one model's weaknesses may fail against specialist alternatives.

For election security practitioners, the key takeaway is that LLM-based cognitive threat triage is operationally practical today. The technology is economically viable, integrates with existing SIEM/SOAR infrastructure, and delivers explainable outputs that support analyst judgment rather than replacing it. visWe provide a deployment checklist, cost models, and integration guidelines to support adoption. The E-MANTRA benchmark subset, prompting templates, and deployment artifacts are available upon request to the corresponding author.

E-MANTRA's contributions extend beyond election security SOCs. Its framework for cognitive threat detection can be adapted to other domains, such as corporate governance, public health misinformation, and crisis communication. By providing explainable, tactic-level intelligence, E-MANTRA offers a versatile solution for addressing cognitive security challenges across diverse operational contexts.

Acknowledgments

This work builds upon operational insights from election security practitioners documented in the Cybersecurity and Infrastructure Security Agency's (CISA) Election Infrastructure Security guidance [2] and the Canadian Centre for Cyber Security's (CCCS) guidance for elections authorities [3].

REFERENCES

[1] G. Sands, B. Ortega, A. Fantz, and K. Scannell, "Sharpies in az: False rumors went viral, sparking outrage and a lawsuit — cnn politics," 11 2020. [Online]. Available: https://www.cnn.com/2020/11/05/politics/arizona-sharpie-ballots-maricopa-trump

[2] CISA, "Election security — cybersecurity and infrastructure security agency cisa." [Online]. Available: https://www.cisa.gov/topics/election-security

[3] CCCS, "Cyber security guidance for elections authorities," Canadian Centre for Cyber Security, Tech. Rep., 8 2020. [Online]. Available: https://publications.gc.ca/collections/collection_2021/cstc-csec/D97-4-10-020-2020-eng.pdf

[4] "Election security spotlight – social engineering." [Online]. Available: https://www.cisecurity.org/insights/spotlight/cybersecurity-spotlight-social-engineering

[5] H. Kennedy, "Election cyber incident communications plan template for state and local officials defending digital democracy defending digital democracy project the cybersecurity campaign playbook defending digital democracy," Harvard Kennedy School Belfer Center for Science and International Affairs, Tech. Rep., 2 2018. [Online]. Available: https://www.belfercenter.org/sites/default/files/2024-08/CommunicationsTemplate.pdf

[6] U. S. E. A. Commission, "Election security preparedness — u.s. election assistance commission," 10 2024. [Online]. Available: https://www.eac.gov/election-officials/election-security-preparedness

[7] surveyandballotsystems, "The importance of election security — sbs." [Online]. Available: https://www.surveyandballotsystems.com/blog/security/national-election-security-for-private-organizations/

[8] B. Freed, "The election infrastructure-isac wants to 'reintroduce' itself — statescoop." [Online]. Available: https://statescoop.com/election-infrastructure-isac-reintroduce/

[9] S. Tariq, M. B. Chhetri, S. Nepal, and C. Paris, "Alert fatigue in security operations centres: Research challenges and opportunities," *ACM Comput. Surv.*, vol. 57, 4 2025. [Online]. Available: https://doi.org/10.1145/3723158

[10] T. Micro, "70

[11] M. Brozek, "Forrester study: The 2020 state of security operations," 9 2020. [Online]. Available: https://www.paloaltonetworks.com/blog/2020/09/state-of-security-operations/

[12] Y. Zhou and L. Shen, "Processing of misinformation as motivational and cognitive biases," *Frontiers in Psychology*, vol. 15, 2024. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1430953

[13] G. D. S. Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "Semeval-2020 task 11: Detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. International Committee for Computational Linguistics, 12 2020, pp. 1377–1414. [Online]. Available: https://aclanthology.org/2020.semeval-1.186/

[14] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, 9 2020. [Online]. Available: https://doi.org/10.1145/3395046

[15] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data," *Knowl. Inf. Syst.*, vol. 67, pp. 3267–3292, 1 2025. [Online]. Available: https://doi.org/10.1007/s10115-024-02321-1

[16] J. Roozenbeek, S. van der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky, "Psychological inoculation improves resilience against misinformation on social media," *Science Advances*, vol. 8, p. eabo6254, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.abo6254

[17] J. Roozenbeek, C. S. Traberg, and S. V. D. Linden, "Technique-based inoculation against real-world misinformation," *Royal Society Open Science*, vol. 9, 2022.

[18] S. Vosoughi, D. K. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146 – 1151, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4549072

[19] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 11 2019, pp. 5636–5646. [Online]. Available: https://aclanthology.org/D19-1565/

[20] C. Lawson and P. Shoard, "Market guide for security orchestration, automation and response solutions," 6 2023. [Online]. Available: https://www.gartner.com/en/documents/4470299

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[23] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

[24] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," *ArXiv*, vol. abs/2301.04246, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255595557

[25] K. Sprenkamp, D. G. Jones, and L. Zavolokina, "Large language models for propaganda detection," *ArXiv*, vol. abs/2310.06422, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830582

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2022.

[27] A. Krasovec, G. Steri, G. Karopoulos, and M. Trapani, "Large language models for cyber threat intelligence: Extracting mitre with llms," *Lecture notes in Computer Science (Including Subseries Lecture Notes on Artificial Intelligence in Bioinformatics)*, vol. 15995, pp. 80–89, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-032-00633-2_5

[28] F. Casino, "Unveiling the multifaceted concept of cognitive security: Trends, perspectives, and future challenges," *Technology in Society*, vol. 83, p. 102956, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X25001460

[29] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. C. Matney, R. Fox, J. Albright, and B. Johnson, "The tactics & tropes of the internet research agency." US Congress, 10 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:158638305

[30] Y. Benkler, R. Farris, and H. Roberts, "Network propaganda," *Network Propaganda*, p. 472, 11 2018.

[31] K. Starbird, A. Arif, and T. Wilson, "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, 11 2019. [Online]. Available: https://doi.org/10.1145/3359229

[32] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, pp. 171–188, 2020, pMID: 32491943. [Online]. Available: https://doi.org/10.1089/big.2020.0062

[33] Z. Sanderson, M. Brown, R. Bonneau, J. Nagler, and J. Tucker, "Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform," *Harvard Kennedy School Misinformation Review*, vol. 2, 12 2021.

[34] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," pp. 422–426. [Online]. Available: https://doi.org/10.18653/v1/P17-2067

[35] K. Madhavan, L. Antonie, and S. D. Scott, "Manipulation vector identification: A security framework for detecting and classifying election misinformation attacks," 7 2025.

[36] K. Madhavan, L. Antonie, and S. Scott, "Flare: An error analysis framework for diagnosing llm classification failures," in *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, P. Przybyła, M. Shardlow, C. Colombatto, and N. Inie, Eds. INCOMA Ltd., Shoumen, Bulgaria, 9 2025, pp. 40–44. [Online]. Available: https://aclanthology.org/2025.ommm-1.4/

[37] I. Eldifrawi, S. Wang, and A. Trabelsi, "Fingract: A framework for fine-grrained evaluation of actionability in explainable automatic fact-checking," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Association for Computational Linguistics, 11 2025, pp. 9882–9901. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.525/

[38] "Foundry models — microsoft azure." [Online]. Available: https://azure.microsoft.com/en-us/products/ai-foundry/models/

[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, pp. 427–437, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457309000259

[40] M. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, vol. 22, pp. 276–282, 12 2012.

[41] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[42] F. Pierri, L. Luceri, N. Jindal, and E. Ferrara, "Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine," *ACM International Conference Proceeding Series*, vol. 10, pp. 65–74, 4 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3578503.3583597