Grounding Information Disorder in NLP: A Theoretical and Operational Framework

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This position paper proposes a theory grounded NLP framework for information disorder detection integrating three explicitly connected dimensions: epistemic status, intentionality, and contextual harm. Moving beyond binary fake news classification, we argue that reliable intervention requires structured differentiation between verification outcomes, manipulation indicators, and consequence assessment. We provide concrete annotation schemas with decision rules for ambiguous cases, formal aggregation operators with monotonicity and escalation guarantees, explicit conflict resolution strategies for inconsistent signals, and standardized risk profile templates that translate multidimensional outputs into actionable routing policies. Synthesizing work on harm taxonomies, uncertainty quantification, and automated fact checking pipelines, we introduce an integration layer that preserves interpretability while enabling policy aligned deployment. We further propose a reformed evaluation protocol incorporating conformal prediction for principled abstention, calibration analysis, disagreement modeling, harm weighted metrics, and human uplift assessment to measure real decision support utility rather than standalone classifier accuracy. We position this framework as a conceptual and operational roadmap for structured misinformation assessment, outlining phased validation pathways while acknowledging that empirical validation remains essential future work.

Keywords: information disorder, fake news detection, misinformation, NLP evaluation, harm taxonomy, uncertainty quantification

1. Introduction

The rapid expansion of digital media has transformed how information is produced, disseminated, and consumed globally. One of the most concerning manifestations has been the emergence of what is collectively known as fake news, a phenomenon attracting substantial research attention across computer science, communication studies, psychology, and political science. Initially conceptualized as a binary distinction between true and false information, fake news detection has historically focused on veracity classification using supervised machine learning. However, this binary framework is fundamentally insufficient to capture the full spectrum of misleading information circulating in contemporary media ecosystems (Wardle and Derakhshan, 2017).

The limitations of binary classification become apparent when considering the diversity of problematic information types. A satirical article misunderstood as genuine news, a genuine photograph shared with a misleading caption, an accurate statistic presented without crucial context, and a deliberately fabricated story designed to manipulate public opinion all represent fundamentally different phenomena demanding distinct analytical and interventional approaches. Yet conventional fake news detection systems typically collapse these distinctions into a single true/false dichotomy, losing critical information necessary for appropriate response.

This position paper shifts focus from binary classification towards a multidimensional, theory grounded approach to information disorder detection. We synthesize insights from existing work on harm taxonomies (Scheuerman et al., 2021; Sehat et al., 2024), uncertainty quantification in NLP (Xiao et al., 2022; Angelopoulos and Bates, 2021), integrated fact checking pipelines (Guo et al., 2022), and human centered AI research (Marusich et al., 2024; Prabhudesai et al., 2023) into a coherent framework with explicit integration mechanics. While no empirical validation is presented in this work, we ground all proposals in demonstrated capabilities from related literature and specify concrete validation pathways with measurable milestones, explicitly identifying the empirical questions that require answers before deployment.

Contributions. We position this as a conceptual roadmap with six contributions: (1) annotation schemas with decision rules for ambiguous cases; (2) formal aggregation operators specifying cross dimensional integration with conflict resolution; (3) adoption of FABLE harm taxonomy with concrete operationalizations and validation pathways; (4) integration of conformal prediction for principled abstention with coverage guarantees; (5) human uplift assessment framework measuring actual decision support utility; and (6) standardized risk profile templates for downstream stakeholder reporting.

2. Related Work

We organize related work around four themes that collectively motivate and inform the proposed

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 66–76
12 May 2026.

framework.

2.1. Harm Taxonomies and Content Moderation

A central challenge for information disorder research is moving beyond binary harm labels toward multidimensional severity assessment. Scheuerman et al. (2021) developed a comprehensive framework through grounded theory analysis with 52 participants, identifying four harm types (physical, emotional, relational, financial) and eight severity dimensions including perspective, intent, agency, and vulnerability. This work provides foundations for graduated response systems that differentiate content based on consequence severity rather than treating all harmful content equivalently. Building on this direction, the FABLE framework (Sehat et al., 2024) operationalizes harm specifically for fact checking prioritization through five dimensions: Fragmentation (social division potential), Actionability (harmful action likelihood), Believability (plausibility), Likelihood of spread (viral potential), and Exploitativeness (vulnerable group targeting). We adopt FABLE as our primary taxonomy due to its explicit operationalizability and alignment with downstream decision making.

Complementing these taxonomic efforts, Banko et al. (2020) synthesized platform guidelines into a unified taxonomy, highlighting definitional inconsistencies across platforms. Fortuna and Nunes (2018) surveyed hate speech taxonomies, documenting persistent challenges for consistent automated moderation. Wang et al. (2025) analyzed intent in algorithmic moderation, finding that current systems cannot reliably infer intent from text alone, a finding that motivates our proxy based approach for the intentionality dimension.

2.2. Uncertainty Quantification and Calibration

Reliable deployment of NLP systems in high stakes information disorder contexts requires principled handling of model uncertainty. Xiao et al. (2022) conducted a large scale empirical analysis showing that temperature scaling most effectively reduces expected calibration error (ECE) while preserving performance, and that larger pretrained language models calibrate better under domain shift. Desai and Durrett (2020) established calibration baselines for BERT family models, providing reference points for evaluating new systems. Xin et al. (2021) introduced error regularization that improves selective prediction, demonstrating accuracy coverage tradeoffs relevant to our abstention framework.

Conformal prediction offers a particularly promising direction by providing distribution free coverage guarantees without distributional assumptions. Angelopoulos and Bates (2021) provide an accessible introduction to conformal methods, while Quach et al. (2023) apply these methods to NLP classification, demonstrating that coverage guarantees hold even under distribution shift. These properties make conformal prediction well suited for misinformation contexts where content distributions change rapidly.

On the human side, research demonstrates that uncertainty information meaningfully affects decision quality. Marusich et al. (2024) showed that instance level uncertainty quantification improves human decision making, while Prabhudesai et al. (2023) found that users struggle to calibrate their reliance on AI without structured uncertainty communication. Devic et al. (2025) critique current UQ practices as insufficiently human centered, recommending evaluation of actual decision support utility rather than standalone calibration metrics.

2.3. Fact Checking Pipelines and Verification

Automated fact checking has matured into a modular pipeline encompassing claim detection, evidence retrieval, verdict prediction, and justification production, as comprehensively surveyed by Guo et al. (2022). Key resources include FEVER (Thorne et al., 2018) with 185K claims, LIAR (Wang, 2017) demonstrating six level graded assessment, and CheckThat! (Nakov et al., 2022) establishing multilingual benchmarks including Arabic. Our own involvement in the CheckThat! evaluation campaigns (Hasanain et al., 2024; Alam et al., 2021) has provided practical experience with the challenges of check worthiness estimation and multilingual claim detection at scale, informing the design requirements for the epistemic status dimension.

Zubiaga et al. (2018) survey rumor detection connecting to formal fact checking workflows. Nakov et al. (2021) emphasize that automated systems should support rather than replace human judgment, a principle central to our framework design. Stance detection research (Hardalov et al., 2022) provides computational foundations for estimating intent through analysis of how content positions itself relative to claims and entities. Related work on propaganda detection (Hasanain et al., 2023) and news media narrative analysis (Zaghouani et al., 2024) further demonstrates the practical complexity of identifying manipulation strategies in multilingual media ecosystems.

2.4. Multilingual, Multimodal, and Cultural Dimensions

Scaling information disorder detection beyond English requires addressing linguistic, cultural, and

resource diversity simultaneously. MuMiN (Nielsen and McConville, 2022) covers 41 languages with 13,000 claims. X FACT (Gupta and Srikumar, 2021) spans 25 languages. MM COVID (Li et al., 2020) offers multilingual multimodal data across six languages. Panchendrarajan and Zubiaga (2024) survey cross lingual claim detection approaches, while Panda and Levitan (2021) demonstrated mBERT for multilingual detection with variable performance across languages.

A critical consideration often underexplored is the cultural dimension of harm assessment. Recent work has shown that LLM persona simulation with cultural or demographic differences significantly affects the perception of harmful content. Piot et al. (2025) demonstrate that geographic persona assignment substantially influences LLM responses in hate speech detection, revealing systematic biases tied to country specific contexts. Plaza-del-Arco et al. (2024) show that LLMs exhibit culturally rooted biases in representing religion, with Eastern religions stereotyped and Abrahamic minority faiths stigmatized. These findings underscore that harm is not culturally universal and that information disorder frameworks must account for cultural variation in both annotation and deployment.

For multimodal detection, VERITE (Papadopoulos et al., 2024) provides robust benchmarks accounting for unimodal biases, while NewsCLIPpings (Luo et al., 2021) addresses out of context image detection. Work on multimodal propaganda detection in Arabic memes (Alam et al., 2024) illustrates the challenges of cross modal inconsistency where genuine images are paired with misleading text.

3. Information Disorder Framework

Following the influential taxonomy of Wardle and Derakhshan (2017), we adopt the tripartite distinction between misinformation (false content shared without intent to harm), disinformation (false content shared with intent to harm), and malinformation (genuine content shared with intent to harm). This theoretical foundation from communication studies motivates our three dimensional computational framework.

3.1. Conceptual Foundations

The information disorder framework addresses fundamental limitations of binary fake news classification. Traditional approaches treat all false claims as equivalent, ignoring crucial variations along three axes.

Epistemic Status concerns the degree to which a claim can be verified or falsified, the strength of available evidence, and the certainty with which verdicts can be assigned. A claim about yesterday's weather differs fundamentally from a claim about long term climate projections, even if both concern meteorology.

Intentionality concerns the motivations behind content creation and sharing. Identical false content may be shared by someone who genuinely believes it (misinformation), someone deliberately seeking to deceive (disinformation), or algorithms amplifying engaging content without regard to accuracy. These different origins suggest different interventions.

Contextual Harm concerns the potential consequences of content given its topic, timing, audience, and reach. A false claim about celebrity relationships differs fundamentally from a false claim about vaccine safety, even if both are equally false.

3.2. Relationship to Existing Approaches

We build on rather than dismiss existing progress. The LIAR dataset's six point veracity scale (Wang, 2017) demonstrates that graded epistemic assessment is both feasible and useful. CheckThat! shared tasks (Nakov et al., 2022) incorporate multi class and ranking formulations reflecting real world fact checking complexity. Stance detection research (Hardalov et al., 2022) provides computational foundations for estimating intent. Our contribution is not replacement of these existing components but systematic integration across dimensions with explicit aggregation mechanics and conflict resolution.

3.3. Framework Pipeline

The framework operates as a staged pipeline with three parallel assessment streams converging at an integration layer. **Stage 1 (Input Processing)** decomposes incoming content into atomic claims following Chen et al. (2022). **Stage 2 (Parallel Dimensional Assessment)** routes each claim through three parallel streams: the Epistemic Status stream performs evidence retrieval and entailment classification with conformal coverage; the Intentionality stream evaluates source history, linguistic manipulation markers, and coordination signals; the Contextual Harm stream applies FABLE dimensions to assess consequence severity. **Stage 3 (Integration)** applies the formal aggregation operator with conflict resolution rules to produce a composite risk score. **Stage 4 (Routing)** maps the integrated assessment to actionable decisions via the decision matrix, generating a structured risk profile for downstream stakeholders. Cases exceeding escalation thresholds or exhibiting cross dimensional conflicts are routed to human review.

4. Framework Architecture

Our framework comprises three dimensions with explicit integration mechanics. We provide concrete annotation schemas, formal aggregation operators, and standardized output templates.

4.1. Dimension 1: Epistemic Status

Epistemic Status concerns the degree of certainty and evidential support for a claim, aligning with traditional fact checking research.

Annotation Schema. We propose a five level scale aligned with existing graded resources:

Level 1 (Supported): Claim is supported by multiple independent, authoritative sources with high confidence.

Level 2 (Likely True): Claim is supported by credible evidence but with some uncertainty due to limited source diversity or domain complexity.

Level 3 (Uncertain): Evidence is insufficient, conflicting, or the claim involves legitimate scientific uncertainty or emerging situations.

Level 4 (Likely False): Claim contradicts credible evidence but with residual uncertainty.

Level 5 (Refuted): Claim is contradicted by multiple independent, authoritative sources with high confidence.

Decision Rules for Ambiguous Cases. When annotators encounter difficult cases: (a) if high quality sources genuinely conflict, assign Level 3 and document the conflict; (b) weight sources by domain expertise, independence, and track record; (c) for emerging situations where evidence is accumulating, assign Level 3 with temporal flag; (d) for claims mixing true and false components, annotate subclaims separately then aggregate; (e) document specific evidence and reasoning in justification field for all non obvious cases.

Operationalization. Implement through four steps: claim decomposition following Chen et al. (2022); evidence retrieval using FEVER style search, web APIs, or domain specific databases; NLI based entailment classification with aggregation across evidence; conformal prediction (Angelopoulos and Bates, 2021) for coverage guarantees rather than arbitrary confidence thresholds.

Output: (epistemic_level, conformal_set, confidence, evidence_ids[], justification_text).

4.2. Dimension 2: Intentionality

Intentionality concerns inferred motivation behind content creation. This dimension presents the greatest operationalization challenges.

Fundamental Limitations. Intent cannot be reliably inferred from content alone. As MacAvaney et al. (2019) note, intent is exceedingly difficult to capture algorithmically. Wang et al. (2025) document the gap between intent's prominence in platform policies and its absence from detection systems. We propose proxy based assessment estimating manipulation likelihood rather than claiming direct intent detection.

Annotation Schema. Three level proxy based assessment:

Low Manipulation Indicators: Source has established correction history, professional editorial standards, transparent authorship, no detected coordination signals.

Moderate Manipulation Indicators: Some concerning signals: elevated sensationalism markers OR mixed source credibility OR timing coincides with sensitive events OR minor coordination indicators.

High Manipulation Indicators: Multiple strong indicators: coordination with known problematic actors AND high sensationalism AND suspicious timing AND absent editorial standards AND network amplification patterns.

Proxy Validation Requirements. Source history proxies require temporal holdout validation: train on period T, evaluate on T+1 to detect strategic behavior modification. Linguistic manipulation indicators (Horne and Adali, 2017) require cross domain validation. Coordination signals require causal validation through counterfactual analysis.

Adversarial Robustness and Fairness Safeguards. For robustness: use behavioral features over easily gamed stylistic features; implement temporal monitoring; ensemble diverse proxy types. For fairness: separate false positive evaluation across source categories; exclude identity features; regular fairness audits; human review for high stakes decisions.

Output: (manipulation_level, proxy_scores[source, linguistic, coordination, timing], fairness_flags[], requires_review).

4.3. Dimension 3: Contextual Harm

We adopt FABLE (Sehat et al., 2024) as our primary harm taxonomy with concrete operationalizations.

Adopted Taxonomy. *Fragmentation* assesses social division potential via topic modeling against known divisive narratives and engagement pattern analysis. *Actionability* assesses harmful action prompting by detecting imperative constructions and claims with direct action implications. *Believability* assesses plausibility to target audiences via source presentation quality and alignment with audience prior beliefs. *Spread Likelihood* predicts viral potential using early engagement velocity and content virality features. *Exploitativeness* identifies vulnerable group targeting through linguistic markers and demographic signals.

Annotation Schema. Four tier classification:

Critical: Health or safety risk with high actionability and spread (e.g., dangerous medical advice during outbreak, incitement to imminent violence).

High: Significant harm potential in sensitive domain with moderate to high spread (e.g., election misinformation near voting).

Medium: Some harm potential but limited actionability or spread (e.g., misleading but not dangerous health claims).

Low: Minimal harm regardless of veracity (e.g., celebrity gossip, clearly satirical content).

Outcome Validation. Harm tiers should correlate with downstream impacts. Validate through: (a) expert panel consensus; (b) retrospective analysis linking content to documented harms; (c) prospective tracking of intervention effectiveness.

Output: (harm_tier, fable_scores[F, A, B, L, E], domain_category, expert_confidence).

4.4. Integration and Aggregation

The three dimensions interact in ways that determine appropriate downstream action. The integration layer is a structured aggregation procedure with explicit escalation guarantees and conflict handling rules.

Formal Aggregation Operator.

$$Risk = w_e \cdot f(E) + w_i \cdot f(I) + w_h \cdot f(H) + \lambda \cdot C(E, I, H)$$

where $f(\cdot)$ maps dimension levels to $[0, 1]$, weights w are context dependent, and $C(E, I, H)$ encodes predefined cross dimensional conflict rules. The conflict term $C(E, I, H)$ is not a learned latent interaction but a rule based adjustment that enforces escalation in predefined high risk configurations. For example, combinations such as high harm under epistemic uncertainty or weaponized true content trigger positive conflict adjustments regardless of linear score magnitude. This preserves interpretability and ensures that high consequence patterns cannot be suppressed by averaging effects.

The aggregation operator satisfies two desirable properties. First, *monotonicity*: increasing harm or manipulation cannot reduce the integrated risk score. Second, *escalation guarantees*: predefined high risk patterns always exceed minimum routing thresholds even if other dimensions are low.

Conflict Resolution Strategies.

Low Epistemic Certainty + High Harm: Escalate to human review regardless of intent signals.

High Epistemic (True) + High Manipulation + High Harm: Route to malinformation pathway requiring contextual assessment rather than factual correction.

High Epistemic (False) + Low Manipulation + Low Harm: Standard correction pathway.

Conflicting Intent Signals: When behavioral and linguistic proxies diverge, weight behavioral history features more heavily, as they are harder to game.

Uncertain Epistemic + Conflicting Evidence: Do not force resolution. Output uncertainty explicitly and route to domain expert review.

Decision Matrix.

Epistemic	Intent	Harm	Action (Risk Score)
Refuted	High	Critical	Immediate escalation ($\geq$0.9)
Refuted	High	High	Priority review ($\geq$0.8)
Refuted	Low	Critical	Expedited fact check ($\geq$0.75)
Refuted	Low	High	Standard fact check ($\geq$0.6)
Uncertain	High	Critical	Urgent investigation ($\geq$0.7)
Uncertain	Any	Medium	Monitor ($\geq$0.4)
Supported	High	Critical	Malinformation review ($\geq$0.5)
Supported	Any	Low	No action ($<$0.3)

Table 1: Decision matrix with indicative risk thresholds. Thresholds are deployment specific and may be tuned empirically.

The matrix operationalizes the aggregation logic as a routing policy rather than a purely numeric classifier. It ensures that qualitatively distinct configurations receive differentiated treatment.

Context Dependent Weighting. Aggregation weights are policy parameters reflecting institutional priorities. Fact checking organizations may set $w_e = 0.6, w_i = 0.1, w_h = 0.3$ to emphasize verification accuracy. Health misinformation contexts may set $w_e = 0.3, w_i = 0.2, w_h = 0.5$ to prioritize harm prevention. Election integrity monitoring may set $w_e = 0.3, w_i = 0.4, w_h = 0.3$ to emphasize coordination detection. Platform moderation may use near balanced weights $w_e = 0.33, w_i = 0.33, w_h = 0.34$. Weights are interpretable, externally auditable parameters rather than learned black box coefficients.

4.5. Illustrative End-to-End Example

To demonstrate operational feasibility, we provide a compact walkthrough of a hypothetical claim.

Example Claim. "Drinking high-dose vitamin C prevents COVID-19 infection."

Epistemic Status. Evidence retrieval identifies WHO and CDC guidance indicating no reliable evidence that vitamin C prevents COVID-19. Multiple authoritative sources contradict the claim. The system assigns *Level 5 (Refuted)* with a conformal prediction set $\{5\}$ at 90% coverage, indicating high epistemic confidence.

Intentionality (Proxies). Source history shows prior fact check flags and elevated sensationalism markers. Coordination signals are weak. Proxy aggregation yields *Moderate Manipulation Indicators*. Output: manipulation level = Moderate.

Contextual Harm. Actionability is high due to behavioral recommendation. Believability is moderate given medical framing. Spread likelihood is moderate. Overall harm tier: *High*, due to potential public health consequences.

Integration. Using health context weights ($w_e = 0.3$, $w_i = 0.2$, $w_h = 0.5$), the aggregation operator produces a risk score exceeding the High harm threshold. Routing decision: *Priority review and corrective labeling*.

This example illustrates three properties: epistemic refutation alone does not determine action; moderate intent proxies do not suppress escalation when harm is substantial; and the routing outcome is interpretable and traceable to dimensional inputs.

Malinformation Example. To illustrate the malinformation pathway, consider a genuine but selectively leaked internal corporate memo about product safety concerns, amplified by coordinated accounts during a competitor's product launch. *Epistemic Status:* Level 1 (Supported), the document is authentic. *Intentionality:* High Manipulation Indicators, coordination signals are strong and timing is suspicious. *Contextual Harm:* High, potential for financial harm and erosion of consumer trust through decontextualization. Under the integration logic, this triggers the malinformation pathway (true content being weaponized), routing to contextual assessment rather than factual correction. The framework correctly avoids labeling authentic content as false while still flagging the manipulative amplification pattern for review.

4.6. Risk Profile Template

For consistent, auditable downstream reporting, we propose a standardized risk profile template:

```
RISK PROFILE v1.0
Content ID: [identifier]
------------------------
EPISTEMIC: Level [1-5],
  Conformal Set [...],
  Evidence [...], Justification [...]
INTENT: Level [L/M/H],
  Scores [S,L,C,T],
  Fairness Flags [...]
HARM: Tier [C/H/M/L],
  FABLE [F,A,B,L,E], Domain [...]
------------------------
INTEGRATED: Risk [0-1],
  Action [...], Conflicts [...],
  Review Required [Y/N]
AUDIT: Status [...],
  Appeal History [...]
```

This template enables consistent reporting, auditability, appeal documentation, and systematic quality evaluation across deployment contexts.

5. Extensions

5.1. Multilingual and Cultural Considerations

Resources. MuMiN (Nielsen and McConville, 2022) covers 41 languages. CheckThat! (Nakov et al., 2022) provides resources for Arabic, Bulgarian, Dutch, Turkish, and English. X FACT (Gupta and Srikumar, 2021) spans 25 languages. MM COVID (Li et al., 2020) offers six languages with multimodal content.

Transfer Strategies. Zero shot cross lingual transfer uses multilingual encoders (mBERT, XLM R) trained on high resource languages; Panda and Levitan (2021) demonstrated feasibility though performance varies by language. Language specific fine tuning yields best performance where annotated data exists.

Cultural Adaptation. Harm assessment requires cultural calibration. We recommend region specific annotator pools with documented backgrounds, explicit documentation of cultural assumptions in guidelines, separate evaluation by language/region rather than aggregated global metrics, and qualitative analysis of cross cultural disagreement. This recommendation is reinforced by recent findings that LLM persona based approaches with geographic or demographic attributes significantly affect harm perception (Piot et al., 2025; Plaza-del-Arco et al., 2024), suggesting that cultural context must be integrated not only in human annotation but also in any LLM assisted assessment workflow, including persona simulation for hypothesis testing and guideline development.

Dimension Adaptations. For Epistemic Status, evidence retrieval must access language specific sources; Wikipedia coverage varies dramatically across languages. For Intentionality, credibility databases are primarily English centric; network coordination patterns may transfer better than linguistic markers. For Harm, topic sensitivity requires cultural adaptation.

5.2. Multimodal Considerations

Resources. MM COVID (Li et al., 2020) provides text plus images. NewsCLIPpings (Luo et al., 2021) addresses out of context images. VERITE (Papadopoulos et al., 2024) accounts for unimodal biases.

Dimension Adaptations. For Epistemic Status: image text consistency via CLIP/BLIP, reverse image search for provenance, metadata verification.

For Intentionality: deepfake and manipulation detection, metadata tampering indicators. For Harm: graphic content detection, emotional manipulation through imagery.

Integration Challenge. Multimodal misinformation often involves accurate text with misleading images or vice versa. The integration layer must handle cross modal inconsistencies, routing such cases to specialized review.

6. Evaluation Reform

We propose a reformed evaluation protocol addressing limitations of conventional metrics.

6.1. Limitations of Conventional Metrics

Standard precision, recall, and F1 fail to capture several critical quality dimensions. Regarding calibration, a system with 80% accuracy but overconfident predictions may be more dangerous than one with 75% accuracy and well calibrated confidence. Regarding selective prediction, systems recognizing their limitations and abstaining appropriately provide more value than those always producing predictions. Regarding harm weighted performance, standard metrics treat all errors equally, but false negatives on Critical content are far more consequential than on Low content. Regarding annotator disagreement, binary ground truth obscures legitimate disagreement reflecting genuine ambiguity rather than annotation error (Nie et al., 2020).

6.2. Proposed Evaluation Dimensions

We propose a multidimensional evaluation protocol aligned with the three dimensional framework.

Calibration. Beyond classification accuracy, systems must produce confidence estimates that align with empirical correctness. We measure Expected Calibration Error (ECE) and Brier scores following Xiao et al. (2022), reported separately for in domain and out of domain evaluation.

Conformal Prediction Quality. We adopt conformal prediction methods (Angelopoulos and Bates, 2021; Quach et al., 2023) providing distribution free coverage guarantees. We report empirical coverage, average prediction set size, and coverage conditional on harm tiers and demographic subgroups. Conditional coverage analysis is essential to detect systematic undercoverage on minority populations or rare claim types.

Harm Weighted Metrics. We define harm sensitive weights: Critical 4.0, High 2.0, Medium 1.0, Low 0.5. Weighted precision, recall, and F1 are computed by scaling errors according to harm tier. Reporting both weighted and unweighted metrics makes explicit the tradeoff between aggregate accuracy and consequence aware performance.

Generalization. We construct cross domain train and test splits spanning political misinformation, health claims, science controversies, and entertainment rumors. Temporal evaluation following Zhu et al. (2022) assesses entity bias by training on earlier time periods and testing on future data.

Disagreement Modeling. Binary majority vote labels obscure legitimate ambiguity. We report full annotator label distributions. Probabilistic aggregation methods such as Dawid Skene or MACE (Nie et al., 2020) jointly model annotator reliability and item difficulty. Disagreement entropy is reported per dimension to identify where guidelines require refinement versus where epistemic uncertainty is inherent. Additionally, for subjective dimensions such as harm assessment, we recommend incorporating subjectivity aware metrics such as cross replication reliability (xRR) (Wong et al., 2021), which benchmarks inter rater reliability against empirical baselines from replicated annotation rather than relying on fixed agreement thresholds. This is particularly relevant given the cultural sensitivity of harm judgments across different populations.

Human Uplift. Following Marusich et al. (2024) and critiques by Devic et al. (2025), we evaluate decision support utility rather than standalone classifier performance. Metrics include decision accuracy, time to decision, harm weighted error severity, appropriate reliance, and calibration of human confidence. Uplift is defined relative to human baseline performance.

6.3. Validation Pathway

Phase 1 (Months 1–3): Protocol Development. Develop guidelines with decision trees, worked examples, edge case specifications. Pilot 50–100 items using think aloud protocols. Deliverable: comprehensive annotation manual.

Phase 2 (Months 4–6): Pilot Study. Annotate 500–1000 items stratified across domains, 3+ annotators per item. Compute Krippendorff's alpha per dimension. Apply MACE for disagreement modeling. Incorporate subjectivity aware metrics including xRR (Wong et al., 2021) to benchmark annotation quality against replicated baselines rather than fixed thresholds. Deliverable: pilot dataset with reliability statistics.

Phase 3 (Months 7–9): Baseline Implementation. Implement dimensions using existing tools: FEVER pipeline for Epistemic Status, credibility APIs for Intent, toxicity classifiers for Harm. Implement conformal wrappers. Deliverable: benchmarks with coverage analysis.

Phase 4 (Months 10–12): Integration Evaluation. Compare integrated system against single

dimension baselines. User studies with fact checkers measuring decision quality, time, error severity. Deliverable: human uplift assessment and deployment recommendations.

7. Discussion

Relationship to Existing Systems. Our framework organizes and integrates existing capabilities rather than replacing them. FEVER style verification pipelines provide foundations for Epistemic Status. Credibility databases (NewsGuard) and coordination detection tools inform Intent Proxies. Content moderation classifiers contribute to Harm assessment.

Computational Considerations. Running three parallel analysis streams increases computational requirements. We recommend tiered deployment: fast classifiers for initial filtering, full dimensional analysis only for flagged content. Conformal prediction adds minimal overhead while providing principled uncertainty. For platform scale deployment, efficient batching and caching strategies become essential. We note that existing modular architectures in fact checking pipelines already employ sequential filtering strategies where claim detection precedes evidence retrieval, which in turn precedes verdict prediction. Our framework extends this principle by adding parallel harm and intent assessment streams that can be invoked selectively based on initial epistemic screening results.

Governance Requirements. Risk profiles require supporting infrastructure. Role definitions must specify reviewer qualifications by harm tier, with Critical tier cases requiring senior domain expertise. Appeal mechanisms must be available for content creators, with clear timelines and escalation paths. Audit procedures should assess systematic biases across source categories (mainstream vs. independent media, different geographic regions) and demographic proxies. Regular reporting on false positive rates by source type helps detect and correct systematic unfairness.

Relationship to Policy Frameworks. The framework's dimensional structure aligns with emerging regulatory approaches that distinguish between different types of harmful content and require proportionate responses. The European Digital Services Act, for instance, requires platforms to assess systemic risks including the dissemination of illegal content and the manipulation of services. Our risk profile template provides a structured mechanism for documenting assessment rationale in compliance with such transparency requirements. The explicit separation of epistemic status from harm assessment also supports contexts where factual accuracy and public interest considerations must be balanced, such as political speech during elections.

8. Conclusion

This position paper has introduced a theory grounded NLP framework for information disorder built on three explicitly integrated dimensions: Epistemic Status, Intentionality, and Contextual Harm. By moving beyond binary fake news classification, we articulate a multidimensional architecture that differentiates verification, manipulation signals, and consequence assessment, and specifies how these signals should be combined in practice.

Rather than proposing yet another classifier, we formalize cross dimensional aggregation through interpretable operators with monotonicity and escalation guarantees. We provide concrete annotation schemas with decision rules for ambiguous cases, explicit conflict resolution strategies, and a structured decision matrix translating model outputs into actionable routing policies. We operationalize contextual harm through the FABLE taxonomy with outcome validation pathways, incorporate conformal prediction for principled abstention under uncertainty, and define harm sensitive evaluation metrics aligned with real world risk, advancing a human uplift evaluation paradigm that measures decision support utility rather than standalone model accuracy.

The practical impact of this proposal depends on empirical validation, including reliable multidimensional annotation, calibrated automated estimation, and demonstrated improvement in expert decision making. We have outlined a phased validation pathway including pilot annotation studies, baseline implementation, and controlled user studies with practicing fact checkers. We offer this framework as a foundation for empirical validation and future system design.

9. Limitations

This paper proposes a conceptual framework without empirical validation, and several limitations require acknowledgment.

First, the proposed annotation schemas require pilot studies to establish inter annotator reliability and to characterize systematic disagreement. Without such studies, consistency and feasibility remain unverified.

Second, computational feasibility at platform scale remains undemonstrated. Running parallel dimensional analyses with uncertainty quantification may impose higher costs than single classifier pipelines, and efficient deployment strategies must be empirically evaluated.

Third, aggregation weights, risk thresholds, and harm tier boundaries require empirical tuning for

specific institutional contexts. The indicative values presented here are theoretically motivated but not validated in operational settings.

Fourth, the framework assumes access to metadata such as source history, coordination signals, and propagation patterns, and it requires cultural calibration for multilingual deployment. In addition, intent inference remains fundamentally limited: proxy based approaches provide correlational rather than causal evidence and must be interpreted cautiously. We view these limitations as a structured research agenda requiring phased empirical investigation rather than as defects in the conceptual design.

Ethical Considerations

Automated information disorder detection has profound implications for freedom of expression, privacy, and democratic governance. Our framework addresses these concerns through several design choices.

First, we emphasize triage based approaches that flag content for human review rather than automated removal. Second, explicit uncertainty communication acknowledges system limitations through conformal prediction and risk profile confidence levels. Third, structured outputs support explainability and appeal: the dimensional breakdown makes assessment reasoning transparent and contestable. Fourth, human in the loop integration ensures consequential decisions involve appropriate oversight.

Responsible deployment requires diverse annotator representation during guideline development, regular fairness audits for credibility and intent proxy components, appeal mechanisms for content creators, and ongoing monitoring for emergent biases. We recommend partnership with established fact checking organizations and civil society groups to ensure appropriate governance structures.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

10. Bibliographical References

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., and Zaghouani, W. (2021). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP 2021*, pp. 611–649.

Alam, F., Biswas, M. R., Shah, U., Zaghouani, W., and Mikros, G. (2024). Propaganda to hate: A multimodal analysis of Arabic memes with multi-agent LLMs. In *International Conference on Web Information Systems Engineering*, pp. 380–390. Springer.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137.

Sehat, C. M., Li, R., Nie, P., Prabhakar, T., and Zhang, A. X. (2024). Misinformation as a harm: Structured approaches for fact-checking prioritization. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Chen, J., Wang, X., Gao, S., Jiang, Y., and Shi, L. (2022). Generating literal and implied subquestions to fact check complex claims. In *Proceedings of EMNLP*, pp. 3495–3508.

Desai, S. and Durrett, G. (2020). Calibration of pre trained transformers. In *Proceedings of EMNLP*, pp. 295–302.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact checking. *TACL*, 10:178–206.

Gupta, A. and Srikumar, V. (2021). X-FACT: A new benchmark dataset for multilingual fact checking. In *Proceedings of ACL*, pp. 675–686.

Hardalov, M., Arora, A., Nakov, P., and Augenstein, I. (2022). A survey on stance detection for mis and disinformation identification. In *Findings of NAACL*, pp. 1259–1277.

Hasanain, M., Alam, F., Mubarak, H., Abdaljalil, S., Zaghouani, W., Nakov, P., Da San Martino, G., and Freihat, A. A. (2023). ArAIEval

shared task: Persuasion techniques and disinformation detection in Arabic text. *arXiv preprint arXiv:2311.03179*.

Hasanain, M., Suwaileh, R., Weering, S., Li, C., Caselli, T., Zaghouani, W., Barrón-Cedeño, A., Nakov, P., and Alam, F. (2024). Overview of the CLEF-2024 CheckThat! lab task 1 on checkworthiness estimation of multigenre content. In *Working Notes of CLEF 2024*, pp. 276–286.

Horne, B. D. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content. In *Proceedings of ICWSM*, 11(1):759–766.

Devic, S., Srinivasan, T., Thomason, J., Neiswanger, W., and Sharan, V. (2025). From calibration to collaboration: LLM uncertainty quantification should be more human-centered. *arXiv preprint arXiv:2506.07461*.

Li, Y., Jiang, B., Shu, K., and Liu, H. (2020). MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Luo, G., Darrell, T., and Rohrbach, A. (2021). NewsCLIPpings: Automatic generation of out of context multimodal media. In *Proceedings of EMNLP*, pp. 6801–6817.

MacAvaney, S., et al. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Marusich, L., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. (2024). Using AI uncertainty quantification to improve human decision making. In *Proceedings of ICML*, pp. 34949–34960.

Nakov, P., et al. (2021). Automated fact checking for assisting human fact checkers. In *Proceedings of IJCAI*, pp. 4551–4558.

Nakov, P., et al. (2022). Overview of the CLEF 2022 CheckThat! lab. In *CLEF 2022*, pp. 495–520.

Nie, Y., Zhou, X., and Bansal, M. (2020). What can we learn from collective human opinions on natural language inference data? In *Proceedings of EMNLP*, pp. 9131–9143.

Nielsen, D. S. and McConville, R. (2022). MuMiN: A large scale multilingual multimodal fact checked misinformation social network dataset. In *Proceedings of SIGIR*, pp. 3141–3153.

Panchendrarajan, R. and Zubiaga, A. (2024). Claim detection for automated fact checking: A survey. *Natural Language Processing Journal*, 7:100066.

Panda, S. and Levitan, S. I. (2021). Detecting multilingual COVID-19 misinformation via contextualized embeddings. In *Proceedings of NLP4IF Workshop*, pp. 125–129.

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., and Petrantonakis, P. C. (2024). VERITE: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multimedia Information Retrieval*, 13:4.

Piot, P., Martín-Rodilla, P., and Parapar, J. (2025). Personalisation or prejudice? Addressing geographic bias in hate speech detection using debias tuning in large language models. *arXiv preprint arXiv:2505.02252*.

Plaza-del-Arco, F. M., Cercas Curry, A., Paoli, S., Cercas Curry, A., and Hovy, D. (2024). Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of EMNLP 2024*, pp. 4346–4366.

Prabhudesai, S., et al. (2023). Understanding uncertainty: How lay decision makers perceive uncertainty in human AI decision making. In *Proceedings of IUI*, pp. 379–396.

Quach, V., Fisch, A., Schuster, T., et al. (2023). Conformal language modeling. In *Proceedings of ICLR*.

Scheuerman, M. K., Jiang, J. A., Fiesler, C., and Brubaker, J. R. (2021). A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):1–33.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: A large scale dataset for fact extraction and verification. In *Proceedings of NAACL HLT*, pp. 809–819.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pp. 422–426.

Wang, S., et al. (2025). The unappreciated role of intent in algorithmic moderation. *Harvard Kennedy School Misinformation Review*, 6(3).

Wardle, C. and Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework*. Council of Europe Report DGI(2017)09.

Wong, K., Paritosh, P., and Aroyo, L. (2021). Cross-replication reliability: An empirical approach to interpreting inter-rater reliability. In *Proceedings of ACL-IJCNLP*, pp. 7053–7065.

Xiao, Y., et al. (2022). Uncertainty quantification with pre trained language models: A large scale empirical analysis. In *Findings of EMNLP*, pp. 7273–7284.

Xin, J., Tang, R., Yu, Y., and Lin, J. (2021). The art of abstention: Selective prediction and error regularization for NLP. In *Proceedings of ACL*, pp. 1040–1051.

Zaghouani, W., Jarrar, M., Habash, N., Bouamor, H., Zitouni, I., Diab, M., El-Beltagy, S. R., and AbuOdeh, M. (2024). The FIGNEWS shared task on news media narratives. *arXiv preprint arXiv:2407.18147*.

Zhu, Y., et al. (2022). Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of SIGIR*, pp. 2120–2125.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.