Beyond Binary: Enhancing Misinformation Detection with Nuance-Controlled Event Context

Elijah Frederick Albertson, Retnani Latifah, Yi-Shin Chen
National Tsing Hua University, Hsinchu, Taiwan
yishin@gmail.com

Abstract

Misinformation rarely presents itself as entirely true or entirely false. Instead, it often embeds partial truths within misleading contexts, creating narratives that blur the boundary between fact and falsehood. Traditional binary fact-checking frameworks fail to capture this nuance, forcing complex claims into oversimplified categories. To address this gap, we introduce **MEGA**, a multidimensional graph framework designed to classify ambiguous claims, with a particular focus on those labelled "*Somewhat True*." MEGA integrates event evidence, spatio-temporal metadata, and a quantifiable nuance score. Its Event Candidate Extraction (ECE) module identifies supporting or contradicting evidence, while the Nuance Control Module (NCM) injects or removes nuance to assess its effect on classification. Experiments show that nuance is both detectable and learnable: adding nuance improves borderline discrimination, while stripping it leads the decisions toward false extremes and conceals partial truth. Our top model —nuance-injected without score weighting —improve accuracy and F1 score by 15 and 16 points over the claims-only baseline, and 6 and 9 points over the ECE-only variant. These results show that explicitly modeling nuance alongside context is crucial for classifying mixed-truth claims and advancing fact-checking beyond binary judgments.

Keywords: Misinformation detection, Linguistic nuance, Event-guided evidence

1 Introduction

The rapid growth of online media has fueled an overwhelming spread of misinformation (Sharma et al., 2019; Hu et al., 2025a). Because misleading narratives often interweave genuine facts with distortions, separating truth from fiction has become increasingly difficult. Traditional fact-checking pipelines, built on binary true/false labels (Wang et al., 2020a), are ill-suited for claims that fall into the borderline category—especially those tagged *Somewhat True*. Such claims typically contain accurate information that is exaggerated, stripped of context, or paired with omissions (Rashkin et al., 2017), making their classification inherently challenging.

This challenge connects to the notion of certainty, long studied in pragmatics and discourse through phenomena such as epistemic modality, evidentiality, doubt, and hedging (Rubin, 2007). These signals express how confidence is conveyed, and in computational terms can be characterised by polarity (support vs. contradiction) and intensity (strength of stance). Yet, recent work on causal epistemic consistency demonstrates that current language models struggle to remain stable when distinguishing such fine-grained cues (Cui et al., 2025). Motivated by these limitations, we manually analysed 150 *Somewhat True* claims and observed recurring linguistic patterns: hedging markers ("may," "could"), context-sensitive phrasing, and contrastive framing. These are not new facts, but structural signals—indicating that *Somewhat True* is not merely a midpoint between False and True, but a distinct category shaped by nuance.

Building on this observation, we design two key modules. A **Nuance Control Module (NCM)** manipulates hedging and ambiguity markers to probe how linguistic framing influences classification. An **Event Candidate Extraction (ECE)** module retrieves and summarises event-level snippets as external evidence, grounding claims in verifiable context. Together, these modules allow us to test whether nuanced linguistic cues help or hinder borderline judgments, and motivate our inclusion of score-aware evidence that weights semantic, temporal, spatial, and nuance features.

To integrate these signals, we propose the **Multidimensional Event-Guided Analysis Graph (MEGA)**, a graph-based framework that links claims to event evidence and metadata while encoding semantic, temporal, spatial, and nuanced relations. Experimental results show that injecting nuance improves performance in borderline cases: our best configuration, a nuance-injected model without score weighting, achieves a 15-point and 16-point improvement on accuracy and F1 scores over the claims-only baseline. Conversely, removing nuance pushes decisions toward extremes and obscures partial truths. These findings demonstrate that explicitly modelling nuance, alongside contextual evidence, is essential for reliable classification of mixed-truth claims.

The key contributions are:

- **Nuance Control Module (NCM)** — injects or removes hedging, conditional, and ambiguity markers to test framing effects.
- **MEGA** —a configurable graph that links claims to event evidence, metadata, and linguistic nuance features via semantic, temporal, spatial, and nuanced edges.
- **Event Candidate Extraction (ECE)** —automatically retrieves and summarises real-world events for each claim.
- **Score-Aware Graph Construction** — weights edges with temporal, spatial, semantic, and nuance scores to prioritise high-quality evidence.

2 Related Work

Research on misinformation has been extensively explored, with many studies adopting a binary classification approach. For example, Wang et al. (2020b) propose WeFEND, a reinforcement learning framework designed to filter noisy crowd-sourced reports, addressing the challenge of limited labeled data. While effective for binary fake news detection, WeFEND assumes all claims are either entirely true or entirely false, overlooking borderline or ambiguous cases. Earlier work on multi-class datasets has shown that mixture labels in between true and false are often predicted as hoaxes, mapping mostly to false (Torabi Asr and Taboada, 2018). Not accounting for this gray area can weaken detection, since some online users employ half-truths as propaganda to mislead readers (Hazra and Majumder, 2024). This stresses the importance of considering gray-area class labels. Using the PolitiFact dataset with six labels, the subquestion-based approach (Chen et al., 2022) improved multi-class veracity prediction, yet overall performance remained modest, highlighting the difficulty of distinguishing fine-grained cases such as half-true.

Beyond label design, model architecture also introduces limitations. ICP-BGCN (Hu et al., 2025b) combines tweet content and propagation structure into a graph but ignores external evidence, leaving it prone to echo-chamber bias. FrameTruth (Wang et al., 2024) extracts misleading narrative frames with an LLM, yet its text-only scope overlooks temporal, spatial, and source-level context. CAMOUFLAGE (Bethany et al., 2025) rewrites claims with hedges and ambiguity to evade detectors, but treats hedging solely as adversarial noise rather than an informative signal. More recently, Tang et al. (2025) introduced POLITIFACT-HIDDEN, a 15k-claim dataset annotated with omitted evidence and intent, and proposed TRACER, a framework that models omissions for half-truth detection. Integrated with existing verifiers, TRACER improved Half-True F1 by up to 16 points, underscoring the need to capture hidden context for trustworthy verification.

While several prior studies have explored half-truths, mixture labels, and omitted evidence (Chen et al., 2022; Tang et al., 2025), none have explicitly modelled linguistic nuance as the primary learnable signal for determining borderline claims. Existing approaches often collapse such borderline statements into either ”True” or ”False,” overlooking the linguistic and contextual subtleties that define partial truths. To the best of our knowledge, *MEGA* is the first framework to operationalize *Somewhat True* as an independent, learnable class, treating nuance not as noise but as a structural feature that bridges the gap between traditional binary classification and a more complex real-world claims.

In summary, prior work often relies on binary labels, internal propagation graphs, or text-only framing models, and sometimes treats linguistic nuance as noise. Our framework addresses this by modelling nuance with both a controllable module and a scoring mechanism, while incorporating event evidence and spatio-temporal metadata into the verification process.

3 Methodology

Our proposed framework, MEGA (Multidimensional Event-Guided Analysis), addresses the challenge of classifying borderline misinformation claims by combining real-world evidence, metadata, linguistic tone, and quality signals into a unified graph-based architecture. Our framework has four stages: (1) Event Candidate Extraction (ECE), (2) Nuance Control Module (NCM), (3) Evidence-Quality Assessment Score (EQAS), and (4) MEGA graph construction and classification.

3.1 Event Candidate Extraction(ECE)

The first step is to link each claim c_i (with metadata $m_i = (\text{date}, \text{platform})$) to external real-world evidence. We retrieve an event snippet e_i by generating structured queries using named entities extracted with spaCy (Honnibal et al., 2020), temporal expressions identified via rule-based patterns, and platform-specific keywords.These snippets were submitting to a SearXNG-powered search interface (SearXNG, 2021) for multi-engine lookups. Retrieved passages are embedded with Sentence-BERT (Reimers and Gurevych, 2019), clustered semantically, and summarised into a single factual event snippet e_i.

If search or clustering fails, we return a short "no reliable event context found" note, so downstream steps always receive a clear, interpretable output.

3.2 Nuance Control Module (NCM)

We change tone, not facts. This module manipulates the linguistic tone of event candidates before they are scored and selected, adjusting each event snippet e_i to convey varying levels of clarity, ambiguity, or caution. In this paper, *linguistic tone* refers to surface cues that influence how a statement is read—such as hedges and modality ("may", "could"), conditionality ("if", "unless"), attribution ("according to…"), and contrast markers ("however", "but"). The presence and strength of these cues are referred to as *nuance*.

We apply *linguistic reframing* to modify these nuances without adding or removing factual content. Specifically, we define two transformation mechanisms (Figure 1):

1. **Nuance injection** — introduces hedging/ambiguity (e.g., "reportedly", "suggests", "appears to").
2. **Nuance removal** — eliminates those markers to make the same content more assertive.

Formally, let e_i denote the event snippet retrieved by ECE for claim c_i. The NCM generates two rewrites: an *injected* version e_i^{inj} (adds hedging/ambiguity cues) and a *removed* version e_i^{rem} (strips them). Each experimental variant uses exactly one of these downstream; for brevity, we write

$$e_i^* \in \{\, e_i^{\text{inj}}, e_i^{\text{rem}} \,\}.$$

We generate e_i^{inj} and e_i^{rem} using Qwen2.5-14B-Instruct hosted locally via Ollama with fixed prompts and parameters to ensure consistency and reproducibility (Bai et al., 2023; Ollama, 2023). Only the event snippet is rewritten; the claim c_i remains unchanged. The resulting pair (c_i, e_i^*) is then used for Evidence-Quality Assessment Score (EQAS) and node-feature construction. This setup lets us directly measure how framing influences classification—especially for *Somewhat True* class.

3.3 Evidence-Quality Assessment Score (EQAS)

For each pair of claim and event snippet (c_i, e_i^*), we compute a four-dimensional score vector $S = \{s_T, s_S, s_M, s_N\}$:

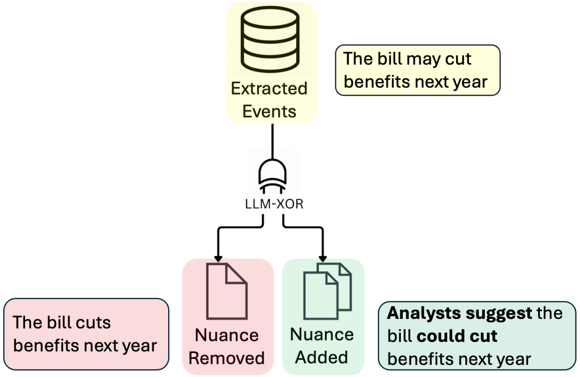
Extracted Events
The bill may cut benefits next year
LLM-XOR
The bill cuts benefits next year
Nuance Removed
Nuance Added
Analysts suggest the bill **could cut** benefits next year

Figure 1: Nuance Control Module (NCM). Given the extracted event snippet, NCM applies *one* rewrite: inject hedging/ambiguity or remove it, producing two alternative snippets used in our variants.

- **Temporal specificity (s_T)** —precision of temporal references in e_i^*, determined via rule-based parsing of explicit dates and scaled to $[0, 1]$.
- **Spatial specificity (s_S)** —granularity of location mentions in e_i^*, mapped by rule-based city/region/country resolution to $[0, 1]$.
- **Semantic similarity (s_M)** —cosine similarity between Sentence-BERT embeddings of c_i and e_i^* (Reimers and Gurevych, 2019).
- **Nuance score (s_N)** —strength of hedging or ambiguity cues in e_i^*, assigned by a locally hosted Qwen2.5-14B-Instruct using a short rubric; computed only when NCM is enabled (Bai et al., 2023).

For claim c_i, we denote $s_i = [\, s_{T,i},\, s_{S,i},\, s_{M,i},\, s_{N,i}\,]$, with $s_{N,i}$ omitted when NCM is disabled.

The score set S serves two purposes: (i) pruning edges via adaptive, type-specific thresholds, and (ii) augmenting node features during graph construction, which will be done in the next stage.

3.4 MEGA Graph Construction and Classification

Node Features. Each data point is $d_i = (c_i, m_i, y_i)$, where c_i is the claim text, $m_i = (\text{date}, \text{platform})$ is metadata, and $y_i \in \{0, 1, 2\}$ is the gold label (*Completely False*, *Somewhat True*, *True*). We encode: (1) c_i with SBERT $\to t_i$; (2) m_i into z_i using date buckets and platform one-hots; (3) e_i^* into EQAS $s_i = (s_{T,i}, s_{S,i}, s_{M,i}, s_{N,i})$. The node feature is:

$$x_i = [\, t_i \parallel z_i \parallel \text{enc}(e_i^*) \parallel s_i\,],$$

where enc$(\cdot)$ is the SBERT embedding of e_i^*. y_i is used only for training and evaluation purposes.

Graph and pruning. We construct a claim—evidence graph $\mathcal{G}$ over all claims, where each node v_i is assigned the feature vector x_i. Edges connect nodes whose claims and associated events are similar in semantic, temporal, or spatial terms, with the corresponding similarity scores stored as edge features.

Adaptive pruning. Using only the training split, we examine the distribution of each edge-score type (semantic, temporal, spatial) and select one cutoff per type (e.g., a chosen percentile). These cutoffs are then fixed and applied unchanged to validation and test splits to avoid leakage. An edge (i, j) is retained if it meets the semantic threshold, or if it satisfies both the temporal and spatial thresholds. We further keep only the top-k most similar neighbours (by semantic score) for each node to prevent any single node from dominating the graph. When the Nuance Control Module (NCM) is active, we increase the thresholds for edges whose endpoints have higher average nuance, $\bar{s}_N = \frac{s_{N,i}+s_{N,j}}{2}$, making the gate stricter when reframing is more ambiguous. This ensures that only well-supported links are preserved in high-nuance contexts.

Classifier. We employ a standard Graph Attention Network (GAT) without architectural modifications (Veličković et al., 2018). The combination of edge-aware construction and adaptive pruning biases the model toward stronger, contextually grounded relationships while reducing noise from weak or misleading connections.

3.5 Dataset and Labelling

We collect fact-checked claims from PolitiFact (2007—2024) (PolitiFact, 2024), including claim text, publish date, platform, and the original veracity label. PolitiFact uses six labels: *Pants on Fire*, *False*, *Mostly False*, *Half True*, *Mostly True*, and *True*.

Model Configuration	Scores	NCM
Claims only	No	No
Claims + metadata	No	No
ECE only	No	No
ECE + EQAS	s_T, s_S, s_M	No
Nuance injected (no EQAS)	s_N	Yes
Nuance removed (no EQAS)	s_N	Yes
sN-only	s_N	Yes
Full MEGA	All	Yes
Contrastive removal	s_N	Yes
ECE Core Isolation	No	No
Positional bias	No	No

Table 1: Feature and edge model configurations used in the experiments

For our experiments, we relabel to three classes to separate outright falsehoods, clear truths, and ambiguous cases:

- **Completely False** —merge *Pants on Fire + False*
- **Somewhat True** —merge *Half True + Mostly True*
- **True** —keep *True* as-is

We exclude *Mostly False* due to inconsistent annotation patterns and class imbalance in our corpus, which would introduce noise into the three-class distinction we aim to evaluate. The final dataset contains 26,500 labelled claims after cleaning (removing nulls, duplicates, extreme-length outliers, and formatting noise). For a balanced evaluation, we sample 6,000 claims (2,000 per class) with a fixed seed and use this same subset across all experiments.

4 Experiments

4.1 Experimental Setup

We conducted extensive experiments across multiple model configurations as shown in Table 1. All models use two GAT layers with a hidden size of 256 and 8 attention heads (Veličković et al., 2018), with each node linked to its top-7 semantic neighbours. Training uses cross-entropy loss, the AdamW optimiser with a learning rate of 5×10^{-4} (Loshchilov and Hutter, 2019), early stopping after 25 epochs without improvement, and a dropout rate of 0.30. We use sentence-BERT `all-mpnet-base-v2` to encode the text (Reimers and Gurevych, 2019).

Model	F1-Score by Class			Acc.
	T	SW True	CF	
Baseline Models				
Claims only	58	63	60	60
Claims + metadata	64	64	62	63
Real-World Context				
ECE only	72	65	70	69
Nuance Control Variants (no EQAS)				
Nuance injected	**77**	**74**	**73**	**75**
Nuance removed	74	70	72	72
Nuance-injected (EQAS) per dimension				
Nuance Score (s_N)	**78**	**80**	71	**77**
Contextual only	77	74	73	74
Temporal only	**78**	75	70	76
Spatial only	77	76	72	75
Spatial + Contextual	71	74	71	74
Spatial + Temporal	**78**	75	**74**	76
Contextual + Temporal	77	75	71	74
Full MEGA	76	74	73	74

Table 2: Performance metrics across models configurations. Abbreviations: T = True; SW True = Somewhat True; CF = Completely False; Acc. = Accuracy. The values are in percentage, applied for all the subsequent tables

The dataset is split into 70% training, 10% validation, and 20% test sets, stratified by class. We evaluate performance using Accuracy and per-class F1, and analyse confusion matrices to investigate misclassification boundaries, particularly for cases near decision edges (Fawcett, 2006). Unless otherwise stated, all tables report the same 20% test split with identical thresholds and prompts carried over from training.

4.2 Results and Discussion

Impact of External Evidence. Baseline models highlight the difficulty of claim classification without real-world context. The claims-only model reached just 60% accuracy, with weak performance across all labels (Table 1). Adding metadata such as platform and date improved accuracy by 3%, showing limited discriminative value on its own. A larger gain came from external evidence: incorporating ECE snippets raised accuracy to 69%. This supports the premise that linking claims to real-world events provides factual anchors through temporal and spatial cues. However, the model continued to struggle with *Somewhat True*, motivating the need for additional signals.

Nuance injection. The next significant shift occurs when the Nuance Control Module (NCM) introduces hedging and ambiguity into event snippets. Accuracy rises to 75%, with *Somewhat True* F1 improving by +9 points over ECE-only. Gains are also consistent for True and False classes. These improvements indicate that the model is not simply relaxing decision criteria but exploiting tone-related cues that clarify borderline distinctions. In particular, hedging and contrastive phrasing sharpen the boundary between *Somewhat True* and both extremes, showing that linguistic nuance functions as a meaningful signal rather than noise.

Nuance removal. When nuance is removed from the event snippet, the performance still improves compared to the base ECE configuration, with *Somewhat True* rising from 65% to 70% and overall accuracy from 69% to 72%. However, this configuration falls short of the injection gains, with *Somewhat True* reaching 74% and accuracy 75% under injection. This gap suggests that removing linguistic cues helps reduce some confusion but also strips away information that could aid the model in identifying fine-grained distinctions. Without these cues, the boundary between True and *Somewhat True* becomes less defined, and certain borderline cases may be pushed toward the wrong side of the decision threshold. The fact that removal still performs better than base ECE implies that not all nuance is helpful, and in some contexts, tone markers may distract the model from content-based reasoning.

Configuration	**T**	**SW True**	**CF**	**Acc.**
Nuance injection	**78**	**80**	71	**77**
Nuance removal	78	72	**72**	75
Contrastive removal	72	64	70	69

Table 3: Nuance Score s_N variants.

Score Config.	**T**	**SW True**	**CF**	**Acc.**
Contextual only	75	66	69	70
Temporal only	73	**77**	70	70
Spatial only	72	63	70	68
Spatial + Contextual	71	65	**71**	69
Spatial + Temporal	75	67	**71**	**71**
Contextual + Temporal	**76**	65	**71**	70
All combined	73	66	70	70

Table 4: EQAS applied to base ECE.

Nuance as Isolated Signal. To examine the effect of linguistic nuance in isolation, the Nuance Score s_N is used as a probe in two settings: using only s_N, and applying the same score to versions where nuanced phrasing has been removed. Using only s_N yields the highest overall accuracy at 77% and the strongest *Somewhat True* F1 at 80%, surpassing the Full MEGA configuration, which achieves 74% accuracy. When s_N is applied to the stripped versions, performance declines in proportion to the amount of nuance removed, indicating that s_N captures the influence of linguistic tone rather than memorising content. The comparative results are shown in Table 3. The values for nuance injection and removal differ from those in the previous table because this experiment measures the effect of nuance alone, without other cues. This indicates that s_N alone is a strong proxy for linguistic tone.

The Effect of Evidence-Quality Assessment Score (EQAS) Module. Applying EQAS on top of the base ECE produces only modest changes in performance (Table 4). Overall accuracy ranges from 68% to 71%, with the highest at 71% for the Spatial + Temporal configuration, a gain of two points over ECE-only at 69%. The Temporal-only setting pushes the *Somewhat True* F1 to 77% but does not raise overall accuracy beyond 70%. Other configurations mostly exchange small gains between classes without a consistent advantage. While these results show that EQAS adds useful signal, its contribution is secondary to the larger improvements achieved through nuance.

When nuance is reduced—either by removing all nuanced phrasing or only contrastive cues—EQAS still provides measurable but modest gains (Tables 5 and 6). Temporal and spatial scores occasionally lift accuracy by up to two points over the base setting, with Temporal-only and Spatial-only configurations performing best in their respective contexts. This shows that EQAS retains value even without nuanced language, but its effect is smaller and less consistent than when nuance is preserved (see Table 2).

Full MEGA Configuration. Full MEGA is the complete configuration of our framework, combining the ECE evidence snippet

Score Config.	T	SW True	CF	Acc.
Contextual only	72	73	72	72
Temporal only	**77**	**75**	**74**	**75**
Spatial only	74	72	72	73
Spatial + Contextual	73	74	72	73
Spatial + Temporal	77	**75**	73	**75**
Contextual + Temporal	70	73	70	71
All combined	74	74	71	74

Table 5: EQAS with all nuance removed.

Score Config.	T	SW True	CF	Acc.
Contextual only	72	66	70	69
Temporal only	74	66	**71**	70
Spatial only	**76**	66	70	**71**
Spatial + Contextual	73	66	70	70
Spatial + Temporal	73	66	69	69
Contextual + Temporal	75	**69**	70	**71**

Table 6: EQAS after contrastive removal.

e_i, an NCM rewrite e_i^*, and all EQAS scores $S = \{s_T, s_S, s_M, s_N\}$, which are encoded in the node features and also used as edge signals in the graph. As shown in Table 2, this configuration delivers strong and balanced performance across classes, although it is not the top performer for *Somewhat True*, where the nuance-injected model without EQAS achieves slightly higher F1 and accuracy. We evaluated both configurations on unseen claims, keeping all thresholds, hyper-parameters, and model settings fixed. Both maintain an F1 of 75% on *Somewhat True*, indicating that the nuance signal generalises beyond the training distribution. Full MEGA achieves the highest overall accuracy in this setting (77% vs. 76% for the nuance-injected variant) by combining temporal and spatial gating with semantic evidence, which slightly reduces off-class errors (Table 7).

We therefore regard Full MEGA as the comprehensive, stability-oriented configuration, while the nuance-injected variant without EQAS remains the most effective for borderline detection.

Model Variant	T	SW True	CF	Acc.
Nuance-injected ECE	77	**75**	75	76
Full MEGA	**78**	**75**	**76**	**77**

Table 7: Generalisation performance on unseen claims

Nuance Config.	T	SW True	CF	Acc.
Original (front-loaded)	77	74	73	75
Mid-loaded	75	72	73	73
Back-loaded	74	77	73	75

Table 8: Impact of shifting nuance position within a claim.

4.3 Diagnostics: Examining Model Behaviour

We conducted three controlled experiments to disentangle the contribution of linguistic nuance from other model cues: (1) *Positional Bias* —hedging cues (e.g., “may cause”) were moved to different positions in the sentence (front, middle, end) to test whether their location influences predictions. (2) *Contrastive Framing* —rhetorical pivots such as “however” and “although” were removed to evaluate reliance on explicit discourse contrast. (3) *Core Isolation* —each event was reduced to its factual core, removing all hedging, elaboration, and contextual detail, to assess how structural simplification affects classification.

Structural dependency via positional bias. The positional bias test examined whether the location of nuance changes the model’s decision-making. As shown in Table 8, shifting hedging cues had minimal effect, with only a 2% drop in accuracy for mid-position placement. This suggests the model’s detection of nuance is not tied to its syntactic location but rather to its lexical and semantic presence in the sentence. Performance stability across positions indicates that nuanced phrasing is treated as a content-level signal rather than a positional signal.

Contrastive removal (rhetorical pivots). The contrastive framing test evaluated the impact of removing explicit discourse markers that signal rhetorical shifts. Compared to the Full MEGA baseline, removing cues such as “however” and “although” reduced accuracy (Table 9), with the largest relative drop in *Somewhat True* performance. These pivots typically mark stance changes or qualifications, making them especially informative for detecting borderline or mixed-truth claims. Their removal reduces the model’ s ability to recognise such shifts, confirming that contrastive phrasing acts as a nuance-like signal in classification.

Model	F1-Score by Class			Acc.
	T	SW True	CF	
Nuance-Focused Baselines				
Nuance injection	**77**	74	**73**	**75**
Nuance Score (s_N)	**78**	**80**	71	**77**
Full MEGA	76	74	73	74
Structural Diagnostics				
Contrastive removal	76	68	72	72
ECE Core Isolation	**79**	**83**	**76**	**80**

Table 9: Comparison of nuance-focused models and structural diagnostic variants

Core isolation (higher-accuracy pitfall). Finally, we investigated the effect of stripping away all structural tone. The Core Isolation variant (which reduces events to bare factual statements without hedging or contextual detail) yielded the highest raw accuracy among non-EQAS settings (Table 9), but this created a problematic trade-off. As shown in the confusion matrices (Figures 2–3), predictions skewed toward extreme labels, particularly *Completely False*. Counts rose from 259 in the nuance-injected variant to 307 under Core Isolation, with ”True” → ”False” errors increasing from 12 to 18, and *Somewhat True* → ”False” from 46 to 54. Thus, accuracy gains came at the cost of misclassifying borderline cases, indicating sharper but less calibrated decision boundaries.

Interpreting the results. Event grounding (ECE) was necessary but not sufficient—linking claims to real-world events provided the first performance lift. The decisive change came from linguistic nuance: injecting hedging and conditional cues prevented the collapse of borderline cases into extremes, allowing the model to treat nuance as a distinct, learnable signal rather than noise. In contrast, Core Isolation simplified the problem rather than solving it, improving accuracy for the wrong reason by inflating binary decisions.

Nuance therefore acts as a dual-role structural signal. As text, it consistently stabilises *Somewhat True* predictions; as a graph feature, it retains influence via the nuance score, providing a direct input for model reasoning. These effects are position-independent, and contrastive phrasing behaves similarly to nuance, broadening the operational definition of nuanced language. EQAS complements this by anchoring decisions to temporal, spatial, and semantic context, but its impact is secondary when strong tone cues are present. Overall, the most robust configuration is ECE + Nuance Injection (no EQAS), which preserves calibration on borderline content while still generalising effectively to unseen claims.

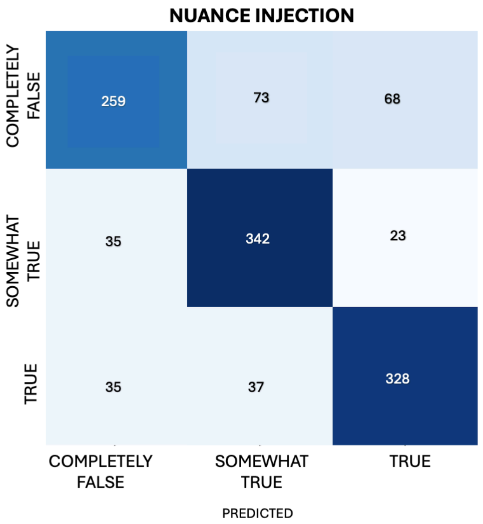

Figure 2: Confusion matrix for ECE with nuance injection.

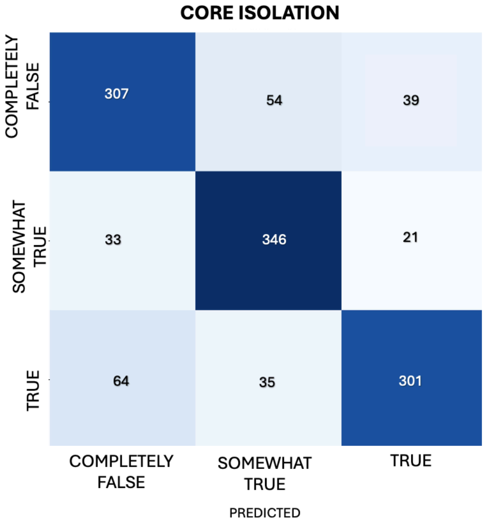

Figure 3: Confusion matrix for the ECE core isolation experiment.

5 Conclusion

Nuance stands out as the signal that defines our approach to misinformation detection. Real-world event grounding provides evidential anchoring, but it is the modelling of tone—hedging, conditionality, and contrast—that consistently enables accurate recognition of partial truths. This effect holds regardless of where cues appear, showing that their strength comes from presence, not position. Other signals, like temporal, spatial, and semantic scores, add stability but do not replace the interpretive weight of nuance. By embedding this signal into both the evidence and the graph, we show that subtle language patterns are not noise, but essential, learnable features for distinguishing misinformation with precision.

Limitations and Future Works

Our framework adopts a relatively simple architecture that combines Sentence-BERT embeddings with a Graph Attention Network, allowing us to isolate and highlight the effects of linguistic nuance. This design effectively captures the contribution of tone and event context; however, its simplicity may constrain the model's expressive capacity and ultimate performance ceiling. Consequently, the full potential of nuanced language understanding within state-of-the-art fact-verification architectures, which incorporate richer contextual modeling or explicit propagation dynamics, remains an open area for further exploration.

Recent fact-verification models use dense passage retrieval (Thorne et al., 2018), fine-tuned transformers trained on large-scale verification datasets (Schuster et al., 2019), or heterogeneous graphs that capture social propagation patterns (Hu et al., 2025b). Such architectures may already capture hedging and tonal variation through large-scale pre-training or by integrating evidence from multiple sources. However, it remains uncertain whether these implicit signals achieve the same interpretive precision as explicit nuance modeling. In other words, while advanced models may recognize linguistic uncertainty to some extent, they may not yet distinguish how specific tone markers influence veracity judgments.

Future work could therefore explore integrating the ECE and NCM modules into more advanced architectures would yield diminishing returns or, conversely, reveal complementary effects—and how stronger baselines might interact with nuance-aware modelling to either enhance or reduce their overall impact.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Ma, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Xu, Zhicu Yang, Zhenru Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Tianzhu Zhang, Bowen Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Mazal Bethany, Nishant Vishwamitra, Cho-Yu Jason Chiang, and Peyman Najafirad. 2025. Camouflage: Exploiting misinformation detection systems through llm-driven adversarial claim transformation. arXiv preprint arXiv:2505.01900.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shaobo Cui, Junyou Li, Luca Mouchel, Yiyang Feng, and Boi Faltings. 2025. Nuance matters: Probing epistemic consistency in causal reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23715–23723.

T. Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8506–8520, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jie Hu, Mei Yang, Bingbing Tang, and Jianjun Hu. 2025a. Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks. *Applied Sciences*, 15(7):3457.

Jie Hu, Mei Yang, Bingbing Tang, and Jianjun Hu. 2025b. Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks. *Applied Sciences*, 15(7):3457.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ollama. 2023. Run large language models locally. https://github.com/ollama/ollama.

PolitiFact. 2024. Politifact fact-check database. https://www.politifact.com/. Accessed 2025-07-10.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2931–2937.

N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel R Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

SearXNG. 2021. A privacy-respecting metasearch engine. https://github.com/searxng/searxng.

K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3):1–42.

Yixuan Tang, Jincheng Wang, and Anthony K. H. Tung. 2025. The missing parts: Augmenting fact verification with half-truth detection.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Fatemeh Torabi Asr and Maite Taboada. 2018. The data challenge in misinformation detection: Source reputation vs. content veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 10–15, Brussels, Belgium. Association for Computational Linguistics.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Guan Wang, Rebecca Frederick, Boshra Talebi Haghighi, B. L. William Wong, Verica Rupar, Weihua Li, and Quan Bai. 2024. Frametruth: A frame-based model utilizing large language models for misinformation detection. In *Proceedings of ACIIDS 2024 (LNAI 14795)*, pages 135–146.

Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, and J. Gao. 2020a. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6359–6369.

Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020b. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 516–523, New York, NY, USA.