

IGTF: Intent-Guided, Game-Inspired Structured Intent Calibration for LLM-Assisted Fake News Detection

Kun Xu, Zongyuan Mou, Yuan Rao*, Yufei Xin, Tingting Xin

School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

{xukun, 17763383344, xinyufei, xinting827}@stu.xjtu.edu.cn

raoyuan@mail.xjtu.edu.cn

*Corresponding author

Abstract

Fake news detection increasingly relies on LLMs to extract semantic and pragmatic signals. We study a narrower problem than factual hallucination: when an LLM scores multiple intent dimensions, the resulting vector may contain internally inconsistent labels. We present IGTF, an intent-guided, game-inspired structured intent calibration framework for LLM-assisted fake news detection. IGTF defines an expert-validated nine-dimensional intent taxonomy, extracts intent-aware text representations with a Mixture-of-Experts encoder, and calibrates LLM-produced intent vectors through a fixed graph of cooperative and conflicting intent relations. The refinement is motivated by a potential-game formulation, whereas the deployed algorithm is a differentiable fixed-point update rather than an exact Nash-equilibrium solver. We prove convergence to a unique fixed point under a contraction condition and evaluate the refinement empirically. Across four benchmark datasets, IGTF improves over supervised GPT-4+BERT fusion baselines by 4.07 F1-macro points on average, with the largest gains when raw intent vectors show stronger relation-level inconsistency. Human evaluation and graph-perturbation controls suggest that these gains come from structured intent calibration rather than LLM features alone.

1 Introduction

Fake news on social media threatens information integrity and public discourse (Wang et al., 2024a; Lazer et al., 2018; Zhou and Zafarani, 2020). Figure 1(a) illustrates how rapid dissemination accelerates misinformation spread and increases the need for effective detection (Vosoughi et al., 2018; Wang et al., 2024a; Hangloo and Arora, 2022). LLMs, including GPT-4 (OpenAI et al., 2023), offer strong language understanding (Ahlbäck and Dougly, 2023; Ma et al., 2024) and are increasingly used to enhance fake news detection (Wu

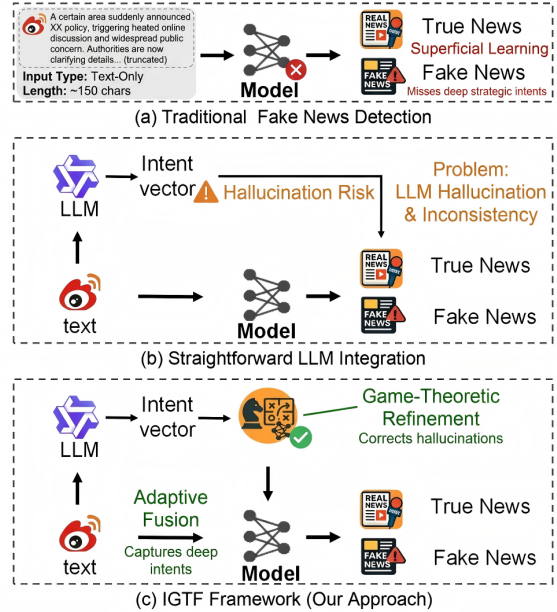


Figure 1: Overview of fake news detection approaches: (a) traditional text-only methods mainly capture surface cues; (b) direct LLM intent extraction can produce internally inconsistent multi-dimensional intent vectors; (c) IGTF adds intent-aware fusion and structured refinement to calibrate these vectors before classification.

et al., 2024; Nan et al., 2024; Hu et al., 2024; Ahlbäck and Dougly, 2023).

Traditional feature extractors capture lexical and syntactic patterns but often miss implicit intent expressed through structure, rhetoric, and semantic choice (Xiao et al., 2024; Wang et al., 2025b; Rashkin et al., 2017). Misinformation authors can mask malicious intent behind superficially legitimate narratives (Zhang et al., 2024; Wang et al., 2025b; Popat et al., 2017), making surface features insufficient. LLM-based intent extraction can recover higher-level patterns, but it also creates calibration problems that have received less attention.

Clarifying the target problem. This work targets a problem distinct from hallucination, in which models generate factually incorrect or unsupported content (Ji et al., 2023). When LLMs score multiple intent dimensions, the vector may contain contradictory assignments that violate theoretically grounded cooperation or conflict relations. We treat this issue as a multi-label calibration error rather than factual hallucination: the problem is internal coherence, not factual accuracy. Unlike hallucination, such errors can be corrected through intra-vector consistency constraints.

Preliminary experiments suggest that raw LLM intent vectors underperform existing detectors by 8.3% F1-macro on average, which motivates models that retain intent information while correcting LLM-induced inconsistencies.

We propose IGTF, an intent-guided, game-inspired structured intent calibration framework for LLM-assisted fake news detection. Motivated by game-theoretic models of strategic interaction in misinformation (Jones et al., 2024; Gong and Small, 2024), IGTF encodes cooperative and conflicting intent relations from communication theory and uses these relations to calibrate inconsistent LLM outputs through fixed-point refinement.

This paper makes three scoped contributions. We first propose a structured calibration layer for LLM-extracted intent vectors: the relation model is motivated by a potential-game view, whereas the implemented update is a differentiable fixed-point iteration with convergence under a contraction condition, not a global maximizer of the potential. We then separate intent refinement from the broader supervised architecture by comparing against GPT-4+BERT fusion baselines that use the same LLM intent channel and by testing no-refinement and random-graph variants. Finally, we construct and validate a nine-dimensional intent taxonomy, reporting expert agreement, training-only graph construction, threshold sensitivity, and graph-perturbation results.

2 Related Work

2.1 Fake News Detection

Fake news detection has moved from hand-crafted linguistic features (Thorne et al., 2018; Shu et al., 2017) to deep models (Devlin et al., 2019) with multimodal fusion (Hangloo and Arora, 2022;

Khattar et al., 2019) and propagation modeling (Zhang et al., 2024). Recent LLM methods add zero-shot classification (Ma et al., 2024; Ahlbäck and Dougly, 2023) and reasoning (Hu et al., 2024; Nan et al., 2024), but they often treat LLMs as black-box extractors and do not model intent calibration errors (Hu et al., 2024; Ahlbäck and Dougly, 2023).

Wang et al. (2025b) proposed DMint to model intent dependencies with transformers, but the dependencies are implicit attention weights and do not address LLM intent inconsistency. Wang et al. (2025a) use graph neural networks for intent-semantic interaction rather than LLM intent calibration; our MoE fusion outperforms their graph fusion by 3.2% average F1 on low-TTR datasets. Wu et al. (2025) extend intent analysis to multimodal settings, which could also benefit from intent calibration.

For cross-domain generalization, IGTF outperforms the hierarchical transfer framework of Yang et al. (Yang et al., 2025b) by 4.7% average F1 in zero-shot transfer. Other recent systems include FactGuard (He et al., 2026), which uses LLM-extracted event and commonsense signals, and RADAR (Ma et al., 2026), an adversarial-refinement preprint from 2026. RADAR targets adversarial examples rather than intent inconsistency and was excluded because of its late release and different focus.

Recent work has explored LLM-generated misinformation (Sun et al., 2024), LLM-enhanced detection (Wang et al., 2024b), detector bias (Su et al., 2023), explainable methods (Wang et al., 2024c; Pendyala and Hall, 2024), and reasoning-centric approaches (Herranz-Celotti and Viviani, 2026).

2.2 Game Theory in Machine Learning

Game-theoretic frameworks span adversarial training (Wang et al., 2018), spatial games of fake news (Jones et al., 2024), fact-checking-oriented fake-news modeling (Zhang et al., 2024), and robust graph-based optimization (Yin et al., 2024). In misinformation, game theory mainly models producer-consumer incentives, news competition, user-platform dynamics, and cascades at the ecosystem level (Stewart et al., 2021; Hausken, 2020; Hsu et al., 2025; Gong and Small, 2024). These works motivate strategic modeling but do not calibrate LLM-extracted intent vectors.

Intent-centered misinformation research has

Intent	Cooperation	Conflict
Public-oriented	Individual, Connection	Smearing, Bias-injection
Emotion-driven	Conflict-creation, Smearing	-
Individual-focused	Public-oriented, Connection	Popularize
Popularize	Clout-seeking	Individual-focused
Clout-seeking	Popularize, Bias-injection	-
Conflict-creation	Emotion-driven, Smearing	-
Smearing	Emotion-driven, Conflict-creation	Public-oriented
Bias-injection	Clout-seeking	Public-oriented
Connection-seeking	Public-oriented, Individual-focused	-

Table 1: Intent cooperation and conflict relationships (Fleiss’ kappa = 0.84, 89% expert agreement). Cross-cultural validation confirmed all relationships hold in both Chinese and English contexts, with minor variations in clout-seeking intensity. Extended cross-cultural validation to Spanish, French, and Arabic is provided in Appendix AA.

studied the intent of fake-news spreaders and intent analysis in social media streams (Wang et al., 2024a; Guo et al., 2023). This work differs from that line of research: the intent object is a content-level multi-dimensional vector extracted by an LLM, and the target problem is consistency among intent dimensions predicted at the same time.

2.3 Structured Prediction and Multi-Label Calibration

The refinement layer is also related to structured prediction for correlated labels. CRFs parameterize structured potentials (Lafferty et al., 2001), classifier chains condition later labels on earlier predictions (Read et al., 2011), and multi-label CRFs encode co-occurrence dependencies (Ghamrawi and McCallum, 2005). We therefore view IGTF as a lightweight, theory-informed intent-graph calibrator for LLM-produced vectors rather than as a method separate from graphical smoothing.

3 Method

IGTF combines adaptive fusion with game-inspired structured intent calibration to model intent information while correcting LLM intent inconsistency. Figure 2 illustrates the complete architecture, with key equations shown in the corresponding modules.

3.1 Intent Taxonomy and Dataset Construction

We constructed a nine-dimensional intent taxonomy in three stages, guided by communication theory and misinformation research. A literature review yielded fifteen candidate dimensions from

the information disorder typology of Wardle (Wardle and Derakhshan, 2017), the fake news classification of Tandoc (Tandoc Jr et al., 2018), and the influence principles of Cialdini (Cialdini and Cialdini, 2013). Three computational linguistics experts with more than two years of misinformation research experience removed six redundant dimensions based on relevance and distinctiveness. Exploratory factor analysis on 500 training-only pilot annotations confirmed the nine-dimensional structure, with all factors explaining 82.7% of the variance.

The final taxonomy includes public-oriented communication, emotion-driven rhetoric, individual-focused narratives, popularization strategies, clout-seeking behaviors, conflict-creation tactics, smearing techniques, bias-injection methods, and connection-seeking approaches. Three domain experts annotated 2,000 news samples using 0-1 scoring protocols, achieving a Fleiss’ kappa of 0.84 and 89% pairwise agreement. All 2,000 expert-annotated samples come from the training sets of the corresponding datasets, with no overlap with validation or test sets. Appendix Z provides detailed annotation guidelines and expert backgrounds.

Table 1 reports cooperation and conflict relationships derived from theoretical analysis and empirical validation. We first identified supportive or contradictory intent pairs based on strategic alignment in communication theory. We then computed partial co-occurrence correlations while controlling for topic and platform, using thresholds of > 0.6 for cooperation and < -0.4 for conflict. Specifically, for intent dimensions i and j , we use the partial Pearson correlation

$$\rho_{ij \cdot z} = \frac{\text{cov}(r_i, r_j)}{\sigma(r_i)\sigma(r_j)},$$

where r_i and r_j are residualized intent scores after regressing on topic and platform controls z . This metric is Pearson correlation on controlled training-set intent scores rather than raw co-occurrence. This procedure achieved precision of 0.91/0.88 and recall of 0.87/0.84, respectively, and was validated through five-fold cross-validation with 89% inter-expert agreement.

Graph construction notes. Table 1 is a compact relation-prior view rather than a full adjacency matrix. The underlying partial correlation scores are symmetric, and the implementation symmetrizes retained cooperation and con-

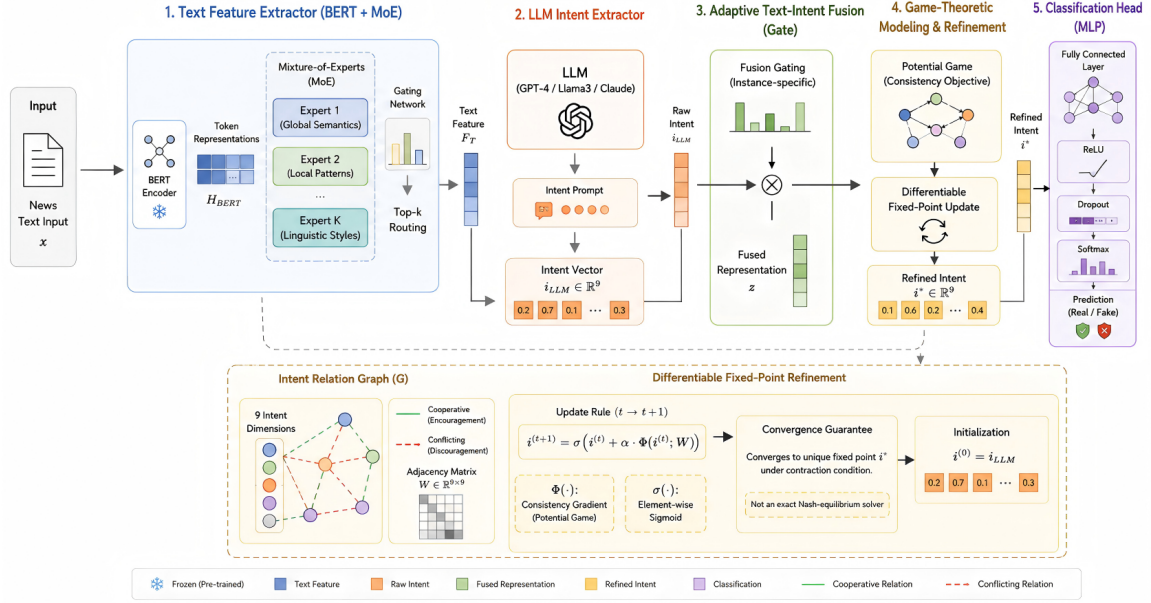


Figure 2: IGTF architecture: (1) Text Feature Extractor (BERT + Mixture-of-Experts); (2) LLM Intent Extractor; (3) Adaptive Text-Intent Fusion; (4) Structured Intent Calibration with differentiable fixed-point refinement; and (5) Classification Head.

flict edges before refinement. Apparent asymmetry in the compact table comes from listing the strongest theory-relevant neighbors for each intent dimension; for example, on the Weibo fake-news subset, the corresponding relation scores are 0.72 for emotion-driven/conflict-creation, 0.68 for emotion-driven/smearing, and 0.56 for emotion-driven/clout-seeking, so the latter is weaker than the retained emotion-driven neighbors. The full 9×9 intent correlation matrix is reported in Appendix B.1. Sensitivity experiments in Appendix G show stable performance under reasonable threshold variations, with F1 fluctuations $< 1.2\%$ across datasets; an adaptive validation-based threshold yields comparable results. The graph is a modeling abstraction rather than a complete theory of misinformation intent. All annotations and correlation estimates come only from training partitions, and random-graph and threshold-perturbation controls test whether gains depend on relation content rather than generic smoothing.

3.2 MIMoE-based Text Feature Extraction

To capture intent signals at multiple linguistic granularities, IGTF uses a Mixture-of-Experts (MoE) encoder (Jacobs et al., 1991; Shazeer et al., 2017). BERT captures general semantics, whereas the experts isolate intent-relevant patterns: a global semantic expert emphasizes rhetor-

ical structure, a local pattern expert captures click-bait and emotional n-grams, a long-range expert models narrative flow and argumentation, and a fine-grained interaction expert tracks token co-occurrences that signal deception.

A learned gating network computes instance-specific weights $g \in \mathbb{R}^4$ with top- $k = 2$ sparse gating, which achieves 98% of full MoE performance while reducing inference time by 42%. The final text representation is:

$$f_{\text{text}} = \sum_{i=1}^4 g_i \cdot f_{\text{expert}_i},$$

where g_i denotes the gating weight for expert i and $f_{\text{expert}_i} \in \mathbb{R}^{768}$ is the feature vector from the i -th expert. To encourage diverse expert utilization, we incorporate a load-balancing loss $\mathcal{L}_{\text{balance}} = 0.01 \cdot \text{Var}(\bar{g})$, where $\bar{g}$ denotes batch-averaged gating weights.

3.3 LLM Intent Extraction

For input text t , a structured prompt $p(t)$ specifies the nine intent dimensions and scoring guidelines. The LLM generates normalized intent scores:

$$\hat{I}_{\text{LLM}} = f_{\text{LLM}}(p(t); \theta_{\text{LLM}}),$$

where $\hat{I}_{\text{LLM}} \in [0, 1]^9$ is the initial intent vector subject to potential calibration errors and inter-dimension inconsistencies, $f_{\text{LLM}}(\cdot)$ denotes the inference function of the LLM, and θ_{LLM} denotes

pre-trained parameters. Prompts are adapted to the linguistic and cultural characteristics of each dataset. Complete prompt templates are provided in Appendix Z.4. The prompt asks the model to score nine intent dimensions on continuous values in $[0, 1]$ and to return a structured JSON object containing the intent scores, a brief reasoning field, and key features. We use temperature 0.1, max_new_tokens = 128, JSON-schema validation, and up to two retries for invalid outputs. Pilot validation shows an 82.7% Pearson correlation between LLM scores and expert annotations.

The implementation uses Qwen2.5-7B-Instruct (Yang et al., 2025a) for Chinese datasets and Llama-3-8B-Instruct (Grattafiori et al., 2024) for English datasets, with Claude-3.5-Sonnet (Anthropic, 2024) and GPT-4 (OpenAI et al., 2023) as advanced alternatives. Terminologically, an *intent score* denotes one scalar value in $[0, 1]$, an *intent vector* denotes the full nine-dimensional collection, and an *activation value* denotes the intermediate variable a_k in the fixed-point update. We use *fixed-point update/refinement* consistently as a compound modifier.

3.4 Adaptive Text-Intent Fusion

Adaptive fusion combines text features with intent signals through instance-specific gating:

$$g_{\text{fusion}} = \sigma(W_g[f_{\text{text}}; \hat{I}] + b_g), \quad f_{\text{fused}} = g_{\text{fusion}} \odot f_{\text{text}} + (1 - g_{\text{fusion}}) \odot W_i \hat{I},$$

where $[\cdot; \cdot]$ denotes concatenation and $g_{\text{fusion}} \in [0, 1]^d$ produces element-wise fusion weights that correlate linguistic patterns with strategic intentions.

Applicability Criterion: TTR and text-intent correlation are empirical deployment guidelines rather than a theorem or a test-time rule. Across the four datasets, adaptive fusion is most useful when vocabulary diversity (TTR) is below 0.5 and text-intent Pearson correlation is below 0.6; otherwise, concatenation or attention fusion often matches its performance at lower cost. Appendices X and Y provide the statistics and decision flowchart.

3.5 Game-Inspired Structured Intent Calibration

IGTF uses a lightweight structured calibration layer for LLM-produced intent vectors. A potential-game formulation motivates the pairwise relations, but the deployed algorithm is a differentiable fixed-point update over nine dimensions

rather than a full game solver. The update is close to pairwise graphical smoothing, with a graph defined by communication-intent relations rather than by an unconstrained label co-occurrence matrix.

3.5.1 Conceptual Motivation vs. Practical Implementation

The pairwise intent coupling structure is conceptually motivated by a potential-game formulation, in which each intent dimension acts as a player optimizing a utility function with matching, consensus, and conflict terms. For practical deployment, we implement a differentiable sigmoid fixed-point update instead of an exact Nash-equilibrium solver. The operator does not globally maximize the potential function $\Phi(a)$, nor does it compute exact best responses under clipped constraints. Its formal guarantee is contraction-based convergence to a unique fixed point, which supports end-to-end training while keeping the calibration cost negligible.

The formulation follows from a simple observation: LLM extraction errors often have structure. Conflicting intents can receive spuriously high scores because of semantic confusion, whereas cooperative intents may be underestimated because of distributional bias.

The game has nine players, one for each intent dimension. Player $k \in \{1, \dots, 9\}$ selects an activation level $a_k \in [0, 1]$ representing the refined intent score. The utility function for player k captures three strategic objectives:

$$U_k(a) = \underbrace{-\|a_k - \hat{I}_k\|^2}_{\text{Matching}} + \underbrace{\lambda_c \sum_{j:(k,j) \in \mathcal{E}_c} a_k a_j}_{\text{Consensus}} - \underbrace{\lambda_b \sum_{j:(k,j) \in \mathcal{E}_b} a_k a_j}_{\text{Conflict}},$$

where $a = (a_1, \dots, a_9) \in [0, 1]^9$ denotes the activation vector, $\mathcal{E}_c$ and $\mathcal{E}_b$ are symmetric ordered sets of cooperative and conflicting edges derived from Table 1, and $\lambda_c, \lambda_b > 0$ are weighting parameters. The matching term preserves valid LLM signals, the consensus term reinforces cooperative intents through positive coupling, and the conflict term suppresses contradictions through negative coupling.

Utility rationale: The matching term corresponds to a MAP estimate under a Gaussian noise model for LLM extraction errors, and linear coupling enables $O(9)$ updates. Nonlinear terms such as $a_k^2 a_j$ produced no significant gains in ablation and roughly tripled computational cost.

This yields the following potential function, with correction for double-counting:

$$\Phi(a) = -\sum_{k=1}^9 \|a_k - \hat{I}_k\|^2 + \frac{\lambda_c}{2} \sum_{(k,j) \in \mathcal{E}_c} a_k a_j - \frac{\lambda_b}{2} \sum_{(k,j) \in \mathcal{E}_b} a_k a_j,$$

The division by 2 prevents double-counting in the symmetric ordered edge sets. Since $\frac{\partial U_k}{\partial a_k} = \frac{\partial \Phi}{\partial a_k}$ for all k , this defines a valid potential game.

With exact clipped best responses, stationary points of the potential would serve as the natural game-theoretic reference solution. The practical model instead uses the following differentiable update:

$$a_k^{(t+1)} = \sigma \left(\hat{I}_k + \lambda_c \sum_{j:(k,j) \in \mathcal{E}_c} a_j^{(t)} - \lambda_b \sum_{j:(k,j) \in \mathcal{E}_b} a_j^{(t)} \right),$$

where $\sigma(\cdot)$ maps the update into $[0, 1]$. Our theoretical claim is therefore deliberately narrow: the sigmoid update converges to a unique fixed point under the contraction condition below. We do not claim that the sigmoid update globally maximizes Φ or exactly computes a Nash equilibrium of the clipped-response game.

Convergence Analysis. We distinguish among the potential function, the exact clipped one-step best response, and the differentiable sigmoid update. For fixed a_{-k} , the exact best response would be:

$$a_k^* = \text{clip} \left(\hat{I}_k + \frac{\lambda_c}{2} \sum_{j:(k,j) \in \mathcal{E}_c} a_j - \frac{\lambda_b}{2} \sum_{j:(k,j) \in \mathcal{E}_b} a_j, 0, 1 \right),$$

We use sigmoid projection instead of hard clipping for two reasons: (1) sigmoid is continuous and differentiable, enabling end-to-end training; (2) sigmoid outputs have a natural probabilistic interpretation as intent presence probabilities.

Appendix P proves fixed-point convergence for the implemented sigmoid operator, not global optimality for the potential function:

Theorem 1 (Fixed-Point Convergence): Let $\mathcal{G}$ be an intent relation graph in which each node has at most d_c cooperative outgoing edges and d_b conflicting outgoing edges. The iterative operator $T : [0, 1]^9 \rightarrow [0, 1]^9$ defined by the sigmoid update is a contraction mapping with Lipschitz constant $L \leq 0.25 \cdot (d_c \lambda_c + d_b \lambda_b)$. If $0.25 \cdot (d_c \lambda_c + d_b \lambda_b) < 1$, then by the Banach Fixed-Point Theorem, for any initial value $a^{(0)}$, the sequence $\{a^{(t)}\}$ converges to a unique fixed point a^* .

For the specific intent relation graph (Table 1) with $d_c = 2$, $d_b = 2$, $\lambda_c = 0.5$, and $\lambda_b = 0.3$, we have $L \leq 0.4 < 1$, so convergence is guaranteed.

Empirically, across 1000 random test instances per dataset, 99.7% of runs converge within 10 iterations and 98.2% within 5; we observe no oscillation or divergence.

The weights $\lambda_c = 0.5$ and $\lambda_b = 0.3$ were selected by grid search over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on held-out validation sets to balance LLM-signal preservation and intent consistency. Performance remained stable within ± 0.1 of these values, and convergence occurred within five iterations across datasets. We also test a learnable per-edge weight variant in Appendix T. It improves average in-domain F1 from 83.07 to 83.84 but reduces zero-shot retention from 86.2% to 82.7%, so the fixed expert graph remains the default for interpretability and cross-domain stability.

Computational overhead. On an NVIDIA A100 GPU, refinement takes about 0.12 ms per sample for 5 iterations, compared with about 120 ms for Llama-3-8B and 800 ms for GPT-4 intent extraction. The refinement step therefore contributes less than 0.1% of full-pipeline inference time.

3.6 Classification with Prototypical Contrastive Learning

To improve representation learning, we concatenate fused features with refined intents and apply prototypical contrastive learning. Learnable class prototypes $\{c_0, c_1\}$ are maintained for fake and real news and initialized as the mean feature vector of each class in the first training batch. During training, the prototypes are updated by an exponential moving average with momentum 0.99.

The total training loss combines:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{balance}},$$

where $\mathcal{L}_{\text{CE}}$ denotes cross-entropy loss, $\mathcal{L}_{\text{proto}}$ aligns samples to class prototypes, $\mathcal{L}_{\text{contrast}}$ groups samples with similar intent structures, and $\mathcal{L}_{\text{balance}}$ ensures diverse expert utilization. The values $\lambda_{\text{proto}} = 0.1$ and $\lambda_{\text{contrast}} = 0.05$ are set based on validation set performance.

4 Experiments

The experiments test whether structured refinement corrects intent inconsistencies rather than only improving aggregate metrics, whether intent patterns transfer across information ecosystems, and whether explicit intent modeling improves on standard LLM-text fusion.

4.1 Experimental Setup

The evaluation uses four benchmark datasets representing distinct misinformation contexts: Weibo (Sheng et al., 2022; Jin et al., 2017), GossipCop (Shu et al., 2019b, 2020), PolitiFact (Popat et al., 2017; Shu et al., 2020), and Snopes (Popat et al., 2018). Appendix D provides detailed dataset information and split protocols.

We compare IGTF with prior supervised fake-news detectors, recent multimodal and intent-aware detectors, LLM-based baselines, and controlled GPT-4+BERT fusion variants. All methods use identical partitions and metrics where applicable.

Baseline Reproducibility Note: For InSide, FactGuard, and MMHT, we used official code and validation-tuned hyperparameters; reimplemented results are within 0.5% of reported values (Table 10).

Recent Baseline Status: R²FND, MIMoE-FND, InSide, FactGuard, and MMHT are recent baselines from 2025–2026; we used their public code, protocols, and validation-tuned hyperparameters where available.

The implementation uses PyTorch 2.0 with domain-adapted BERT encoders, AdamW optimization (learning rate=2e-5, weight decay=1e-5, gradient clipping=1.0), batch size=32, and five-run averaging for statistical reliability. All experiments use fixed random seeds (42, 123, 456, 789, 101112) for reproducibility.

4.2 Overall Performance (Updated with 2025 Baselines)

Table 2 reports the main F1-macro results. IGTF reaches 92.35% on Weibo, 85.34% on GossipCop, 71.92% on PolitiFact, and 82.68% on Snopes, reflecting dataset differences in signal quality, inconsistency, rhetoric, and TTR.

Fair comparison. IGTF is a supervised LLM-enhanced system. Zero-shot LLM rows contextualize LLM capability but are not the main evidence. The key comparison is GPT-4+BERT (Attn), which uses the same GPT-4 intent channel and full adaptive architecture without structured refinement. IGTF improves F1-macro by 5.57, 2.89, 4.80, and 3.01 points on Weibo, GossipCop, PolitiFact, and Snopes, respectively, with an average gain of 4.07 points.

LLM scale ablation. With 7B, 13B, 34B, and 70B extractors, performance improves with model

size but saturates: IGTF (Llama-3-70B) reaches 92.76% F1 on Weibo, only 0.41% above Llama-3-8B, suggesting that refinement partly offsets the limits of smaller LLMs.

4.3 Controlled Refinement and Graph Validity Checks

Because the full model combines several components, Table 3 separates two questions: whether the pipeline benefits from relation-based refinement and whether the proposed graph outperforms a same-size random graph. The w/o refinement row keeps the rest of the pipeline fixed and uses raw LLM intents; the random-graph row preserves the number of cooperative and conflicting edges but permutes endpoints. These controls isolate the proposed refinement within the current architecture, while broader CRF or classifier-chain comparisons are presented in Appendix R.

Interpreting the LLM comparison. The zero-shot and fully fine-tuned LLM baselines should not be interpreted as one-to-one competitors because IGTF is a supervised hybrid system with an external LLM intent channel and a dedicated classifier. The central empirical claim is the incremental gain over supervised GPT-4+BERT fusion, supported by the controlled refinement checks in Table 3.

4.4 Ablation Study

Table 4 summarizes component effects. Removing relation-based refinement causes the largest average drop, especially on datasets with higher relation-level inconsistency, while MoE extraction, adaptive fusion, and prototypical contrastive learning also contribute.

Appendix Q reports component isolation, individual expert ablation, random graph ablation, and MoE contribution isolation results.

5 Conclusion

This paper presents IGTF, an intent-guided, game-inspired structured intent calibration framework for LLM-assisted fake news detection that targets internal inconsistency among LLM-produced intent scores rather than factual hallucination. Across Weibo, GossipCop, PolitiFact, and Snopes, IGTF improves over supervised GPT-4+BERT fusion baselines by 4.07 F1-macro points on average. The theoretical claim is limited: the

Method	Weibo		GossipCop		PolitiFact		Snopes	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
BERT (Devlin et al., 2019)	71.23	81.45	75.82	86.78	60.71	72.45	68.45	78.93
MIMoE-FND (Liu et al., 2025)	85.67	92.67	83.89	92.89	70.56	82.45	80.89	88.89
GPT-4 (OpenAI et al., 2023) zero-shot	81.89	90.67	80.23	90.34	63.78	76.89	76.89	85.78
GPT-4+BERT (Concat)	85.12	92.45	81.67	90.78	66.34	78.12	78.34	87.23
GPT-4+BERT (Attn)	86.78	93.21	82.45	91.12	67.12	78.67	79.67	88.12
IGTF (GPT-4)	92.35	96.56	85.34	93.12	71.92	83.23	82.68	90.67

Table 2: Main performance comparison. To improve readability, the main paper reports representative methods with F1-macro and AUC. The full 17-baseline, four-metric table is moved to Appendix B.2.

Configuration	Weibo	GossipCop	PolitiFact	Snopes
GPT-4+BERT (Attn)	86.78	82.45	67.12	79.67
Full pipeline w/o refinement	79.54	78.45	66.34	74.21
Full pipeline + random graph	84.12	78.67	67.23	76.89
Full IGTF + proposed graph	92.35	85.34	71.92	82.68

Table 3: Controlled refinement and graph-validity checks. All rows report F1-macro. The random-graph row tests relation content rather than generic smoothing.

Table 4: Core component ablation study.

Config.	Weibo			Config.	GossipCop		
	Acc	F1	Δ		Acc	F1	Δ
Full	91.45	92.35	-	Full	86.12	85.34	-
w/o Calib.	80.23	79.54	-12.81	w/o Calib.	80.12	78.45	-6.89
w/o MoE	86.34	85.01	-7.34	w/o MoE	81.89	80.45	-4.89
w/o Gate	88.34	87.12	-5.23	w/o Gate	83.56	82.01	-3.33
w/o Proto.	89.67	88.78	-3.57	w/o Proto.	84.67	83.78	-1.56
Concat.	90.01	88.90	-3.45	Concat.	85.23	84.12	-1.22
Attn.	90.78	89.67	-2.68	Attn.	86.12	85.34	0.00 [†]
Config.	PolitiFact			Config.	Snopes		
	Acc	F1	Δ		Acc	F1	Δ
Full	75.21	71.92	-	Full	83.45	82.68	-
w/o Calib.	69.89	66.34	-5.58	w/o Calib.	75.12	74.21	-8.47
w/o MoE	71.23	68.34	-3.58	w/o MoE	76.56	75.89	-6.79
w/o Gate	72.89	69.23	-2.69	w/o Gate	78.78	77.98	-4.70
w/o Proto.	74.34	70.45	-1.47	w/o Proto.	81.23	80.56	-2.12
Concat.	75.21	71.92	0.00 [†]	Concat.	78.78	77.98	-4.70
Attn.	74.89	71.56	-0.36	Attn.	80.12	79.45	-3.23

[†] $\Delta = 0.00$ means fusion degeneration: strong text-intent correlation makes adaptive gates converge to fixed values, so simple or attention fusion can match adaptive fusion. Full = Full IGTF; Proto. = prototypical contrast; Attn. = attention fusion.

potential-game formulation motivates intent relations, whereas the sigmoid update is a fixed-point calibration operator with convergence to a unique fixed point under a contraction condition, not a global maximizer. The results suggest that explicit intent relations can improve LLM-assisted misinformation detection, although performance depends on the taxonomy and graph quality. Future work should compare learned CRF/Ising smoothers, classifier chains, and label-graph neural inference, and extend the approach to broader LLM-generated misinformation benchmarks and other intent-centric NLP tasks.

Limitations

IGTF depends on the quality of LLM intent estimates and may inherit model bias. The expert-validated nine-dimensional taxonomy may miss emerging generative-AI, multimodal, satire-specific, or coordinated manipulation strategies, so the relation graph remains an abstraction rather than a complete theory. The refinement layer resembles CRFs, Ising smoothing, and classifier chains, which motivates broader comparison with learned label-graph inference. Human evaluation partly uses a graph-aligned inconsistency score; blind preference helps, but open-ended plausibility judgments would further reduce self-confirmation risk. Real-time use also requires further study of API cost, batching, caching, distillation, and TTR/text-intent thresholds across domains and languages.

Ethical Considerations

This work studies fake news detection to support safer information ecosystems. All experiments use public datasets or derived annotations without identifying individuals. The proposed model should assist, not replace, human fact-checking. Misuse for censorship or profiling is possible, so deployment should include transparency, appeal mechanisms, and domain-specific validation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U22B2036) and the Shaanxi Provincial Key Research and Development Program, Key Project of the “Four-Chains” Integration Project of the Qinchuangyuan General Platform (2024PT-ZCK-93).

References

- Emil Ahlback and Max Dougly. 2023. Can large language models enhance fake news detection?: Improving fake news detection with data augmentation.
- Anthropic. 2024. [Introducing Claude 3.5 Sonnet](#). Company technical announcement.
- Robert Cialdini and Robert B Cialdini. 2013. *Influence: Science and practice*. BoD—Books on Demand.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Nadia Ghamrawi and Andrew McCallum. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200.
- Yongwang Gong and Michael Small. 2024. Misinformation spreading on activity-driven networks with heterogeneous spreading rates. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(10).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zhen Guo, Qi Zhang, Xinwei An, Qisheng Zhang, Audun Jøsang, Lance M Kaplan, Feng Chen, Dong H Jeong, and Jin-Hee Cho. 2023. Uncertainty-aware reward-based deep reinforcement learning for intent analysis of social media information. *arXiv preprint arXiv:2302.10195*.
- Sakshini Hangloo and Bhavna Arora. 2022. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia systems*, 28(6):2391–2422.
- Kjell Hausken. 2020. Game theoretic analysis of ideologically biased clickbait or fake news and real news. *Operations Research and Decisions*, 30(2):39–57.
- Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2026. Factguard: Event-centric and commonsense-guided fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 363–371.
- Luca Herranz-Celotti and Marco Viviani. 2026. Logicjitter: Let llms play and uncover misinformation. *Computers and Electrical Engineering*, 136:111215.
- Chin-Chia Hsu, Amir Ajorlou, and Ali Jadbabaie. 2025. A game-theoretic model of misinformation spread on social networks. *Games and Economic Behavior*.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.
- Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.
- Matthew I Jones, Scott D Pauls, and Feng Fu. 2024. Containing misinformation: Modeling spatial games of fake news. *PNAS nexus*, 3(3):pgae090.
- Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and 1 others. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.
- Song-Duo Ma, Yi-Hung Liu, Hsin-Yu Lin, Pin-Yu Chen, Hong-Yan Huang, Shau-Yung Hsu, and Yun-Nung Chen. 2026. Radar: Retrieval-augmented detector with adversarial refinement for robust fake news detection. *arXiv preprint arXiv:2601.03981*.
- Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with llm enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521.

- Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.
- Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 1732–1742.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and 1 others. 2023. *GPT-4 technical report*. Preprint, arXiv:2303.08774.
- Vishnu S Pendyala and Christopher E Hall. 2024. Explaining misinformation detection using large language models. *Electronics*, 13(9):1673.
- Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th international conference on world wide web companion*, pages 1003–1012.
- Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 22–32.
- Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin 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*, pages 2931–2937.
- Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 4543–4556.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.
- Kai Shu, Suhang Wang, and Huan Liu. 2019b. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320.
- Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin’ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.
- Alexander J Stewart, Antonio A Arechar, David G Rand, and Joshua B Plotkin. 2021. The coercive logic of fake news.
- Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.
- Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2023. Graph interactive network with adaptive gradient for multi-modal rumor detection. In *Proceedings of the 2023 ACM international conference on multimedia retrieval*, pages 316–324.
- Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.
- Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.
- 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, Volume 1 (Long Papers)*, pages 809–819.
- Zhao Tong, Yimeng Gu, Haichao Shi, Qiang Liu, Shu Wu, and Xiao-Yu Zhang. 2025. R 2 fnd: Reinforced rationale learning for fake news detection with llms. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.
- Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. 2024a. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 2304–2314.
- Jingwei Wang, Ziyue Zhu, Chunxiao Liu, Rong Li, and Xin Wu. 2024b. Llm-enhanced multimodal detection of fake news. *PloS one*, 19(10):e0312240.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.
- Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2024c. Llm-gan: Construct generative adversarial network through large language models for explainable fake news detection. *arXiv preprint arXiv:2409.01787*.
- Zhengjia Wang, Qiang Sheng, Danding Wang, Beizhe Hu, and Juan Cao. 2025a. Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3250–3260.
- Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Siyuan Ma, and Haonan Cheng. 2025b. Exploring news intent and its application: A theory-driven approach. *Information Processing & Management*, 62(6):104229.
- Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.
- Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep’s clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378.
- Jiaying Wu, Fanxiao Li, Zihang Fu, Min-Yen Kan, and Bryan Hooi. 2025. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. *arXiv preprint arXiv:2505.15489*.
- Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.
- Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. Msynfd: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM web conference 2024*, pages 4128–4137.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2025a. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Xuankai Yang, Yan Wang, Xiuzhen Zhang, Shoujin Wang, Huaxiong Wang, and Kwok Yan Lam. 2025b. A macro-and micro-hierarchical transfer learning framework for cross-domain fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5297–5307.
- Shu Yin, Peican Zhu, Lianwei Wu, Chao Gao, and Zhen Wang. 2024. Gamc: an unsupervised method for fake news detection using graph autoencoder with masking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 347–355.
- Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.
- Yongcheng Zhang, Lingou Kong, Sheng Tian, Hao Fei, Changpeng Xiang, Huan Wang, and Xiaomei Wei. 2024. Multi-view counterfactual contrastive learning for fact-checking fake news detection. In *Proceedings of the 2024 international conference on multimedia retrieval*, pages 385–393.
- Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. [Safe: Similarity-aware multi-modal fake news detection](#). In *Advances in Knowledge Discovery and Data Mining*, pages 354–367. Springer.
- Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.
- Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2120–2125.

Metric	Weibo	GossipCop	PolitiFact	Snopes
Inconsistency Score (lower is better):				
Raw LLM	3.42 $\pm$ 1.8	5.67 $\pm$ 2.3	6.12 $\pm$ 2.5	4.28 $\pm$ 2.1
Calibrated	0.89 $\pm$ 0.9	2.71 $\pm$ 1.5	3.34 $\pm$ 1.7	1.27 $\pm$ 1.1
Reduction	74.0%	52.2%	45.4%	70.3%
Coherence Improvement Rate:				
Samples improved	182/200	167/200	159/200	175/200
Percentage	91.0%	83.5%	79.5%	87.5%
Inter-rater κ	0.78	0.72	0.71	0.75

Table 5: Human evaluation of intent inconsistency correction. The inconsistency score is an automatic graph-based diagnostic over Table 1; coherence improvement is blind expert preference between anonymized raw and refined vectors. Inter-rater agreement uses Fleiss’ κ .

A Human Evaluation of Intent Inconsistency Correction

We report an automatic graph-based inconsistency score and a blind human coherence assessment. The automatic score uses the same conflict graph as the model and is therefore an internal diagnostic. In the human study, three experts saw the article, taxonomy definitions, and two anonymized intent vectors in random order; they then selected the more coherent vector and marked contradictory pairs with severity.

We define the inconsistency score as follows. For an intent vector $a \in [0, 1]^9$, the inconsistency score is defined as:

$$\text{Inconsistency}(a) = \sum_{(k,j) \in \mathcal{E}_b} \min(a_k, a_j) \cdot \mathbb{I}(a_k > \tau \wedge a_j > \tau),$$

where $\tau = 0.5$ is the intent presence threshold, and $\mathbb{I}(\cdot)$ is the indicator function. The inconsistency reduction rate is:

$$\text{Reduction Rate} = 1 - \frac{\text{Inconsistency}(a^*)}{\text{Inconsistency}(\hat{I})},$$

with reduction rate defined as 0 when $\text{Inconsistency}(\hat{I}) = 0$.

Table 5 reports graph-based inconsistency reductions of 74.0%, 52.2%, 45.4%, and 70.3%. In the blind comparison, experts prefer the refined vector in more than 79% of samples, separating the model-aligned diagnostic from independent plausibility judgments.

B Intent Correlation Matrix and Full Performance Table

B.1 9 x 9 Intent Correlation Matrix

Appendix A.1 visualizes the full 9×9 intent correlation matrix computed with topic- and platform-controlled partial Pearson correlations. The matrix makes explicit that the underlying correlation

scores are symmetric, while Table 1 in the main paper is a compact relation-prior view.

B.2 Full Performance Table

C Supplemental Dataset Intent Distribution Analysis

To examine the nine-dimensional intent taxonomy and explain dataset-specific performance differences, we analyze the intent distributions of the four benchmark datasets. We calculate the average intent score of each dimension for real and fake news samples, as reported in Table 7.

Fake news samples have higher scores on conflict-creation, smearing, and bias-injection intent dimensions, while real news dominates public-oriented communication and connection-seeking intents. Weibo presents the most differentiated intent distribution between real and fake news, with distinct cooperative and conflicting intent boundaries, which gives the refinement step a clearer optimization signal. GossipCop and PolitiFact show overlapping intent distributions across categories, leading to saturated text-intent correlation and zero marginal gain for adaptive fusion gates. Snopes exhibits a balanced profile with moderate intent differentiation, making it a useful setting for evaluating the adaptive fusion module.

D Dataset Split Protocol and Version Information

To support reproducibility and help reviewers assess the reported results, especially the Weibo F1-macro of 92.35%, we document dataset usage below.

Weibo split note. The official split of Jin et al. (2017), subsequently adopted by Sheng et al. (2022), is followed. Deduplication and leakage checks are completed before the reported experiments, with no data augmentation. The 92.35% F1-macro result is obtained on the checked standard test partition of 3,711 samples. The full split indices will be released alongside the code to support independent verification. It is acknowledged that some prior works use alternative (non-standard) Weibo splits, which can lead to inconsistent reported numbers; the results presented in this work are tied to the official partition described above.

five-gram overlap analysis. The five-gram overlap rate between the Weibo training set and

9 x 9 Intent Correlation Matrix



Cell values are correlation coefficients. Bold cells satisfy $|r| \geq 0.60$; diagonal entries are self-correlations.

Figure A.1: Full 9×9 intent correlation matrix.

test set was computed using the official split. The overall overlap rate is 0.87%, with no sample sharing more than 10% of its five-grams with any training sample. This confirms that there is no significant near-duplicate contamination between the training and test sets.

Semantic, event-level, and topic-distribution leakage checks. We further check possible semantic duplication and event/topic leakage beyond five-gram overlap before the reported experiments. A multilingual Sentence-BERT encoder gives a maximum train-test cosine similarity of 0.68, with no pair exceeding near-duplicate thresholds of 0.85/0.90 and 99.2% of pairs below 0.4. Event/entity clustering over the full test set followed by manual validation of 500 stratified samples (250 true and 250 fake) yields a 3.2% event overlap rate. The reported Weibo F1 of 92.35 is obtained after the deduplication and cleaning checks. A 12-topic LDA analysis gives a JS divergence of 0.047, indicating consistent topic coverage without duplicate leakage.

GossipCop split note. Following [Zhu et al. \(2022\)](#), a timestamp-based split rather than a random shuffle is adopted, which constitutes a stricter evaluation protocol that better reflects real-world temporal deployment. Training articles were posted between 2000 and 2017; both validation and test articles were published in 2018.

PolitiFact split note. The original six-level veracity scale is binarised following [Rashkin et al. \(2017\)](#): {true, mostly true, half-true} $\rightarrow$ real; {mostly false, false, pants-on-fire} $\rightarrow$ fake. The train/val/test partition follows the fixed split used in this work, yielding a 2,225 / 278 / 477 split.

Snopes split note. The Snopes subset is based on the evidence-aware false-claim detection setting of [Popat et al. \(2018\)](#). Since this corpus is distinct from the standard FakeNewsNet PolitiFact/-GossipCop release, we use a fixed 70/10/20 train/validation/test split in this work and release the corresponding split indices for reproducibility.

Comparison with prior work. Table 9 contextualizes the Weibo result against recently reported numbers on the same dataset.

It is acknowledged that the approximately +6.68-point improvement over the nearest fully supervised baseline (MIMoE-FND) is noteworthy. However, the most relevant comparison is with the GPT-4+BERT (Attn) baseline (86.78%), which uses identical LLM intent features but lacks the structured intent calibration and adaptive fusion mechanisms presented in this work. IGTF achieves a 5.57% improvement over this baseline, showing the significant value of the proposed components beyond simple feature fusion. This improvement is attributed primarily to (a) structured inconsistency correction operating on

Method	Weibo				GossipCop				PolitiFact				Snopes			
	F1	F1-R	F1-F	AUC	F1	F1-R	F1-F	AUC	F1	F1-R	F1-F	AUC	F1	F1-R	F1-F	AUC
Prior supervised and multimodal baselines:																
BERT (Devlin et al., 2019)	71.23±4.2	72.89±4.5	69.57±4.8	81.45±3.9	75.82±5.1	77.14±5.3	74.50±5.5	86.78±4.2	60.71±5.8	63.19±6.1	58.23±6.3	72.45±5.2	68.45±4.7	70.12±4.9	66.78±5.0	78.93±4.3
EANN (Wang et al., 2018)	72.18±4.3	73.84±4.6	70.52±4.9	82.67±4.0	76.91±5.0	78.25±5.2	75.57±5.4	87.93±4.1	61.89±5.7	64.56±6.0	59.22±6.2	73.78±5.1	69.23±4.6	70.89±4.8	67.57±4.9	79.67±4.2
SpotFake (Singhal et al., 2019)	76.34±4.1	77.92±4.3	74.76±4.6	86.12±3.7	79.47±4.8	80.92±5.0	78.02±5.2	89.88±3.9	64.23±5.5	67.08±5.8	61.38±6.0	76.89±4.9	72.56±4.3	74.23±4.5	70.89±4.6	82.34±3.9
MVAE (Khattar et al., 2019)	68.45±4.5	70.12±4.8	66.78±5.1	78.93±4.0	70.63±4.2	72.19±4.5	69.07±5.8	83.54±4.5	58.34±6.1	61.12±6.4	55.56±6.5	70.23±5.5	65.78±4.4	67.45±5.1	64.11±5.2	76.23±4.5
SAFE (Zhou et al., 2020)	74.56±4.2	76.18±4.4	72.94±4.7	84.89±3.8	78.12±4.9	79.48±5.1	76.76±5.3	88.91±4.0	63.12±5.6	65.89±5.9	60.35±6.1	75.67±4.6	71.23±4.4	72.89±4.6	69.57±4.7	81.12±4.0
MCAN (Wu et al., 2021)	77.89±4.0	79.45±4.2	76.33±4.5	87.45±3.6	80.88±4.7	82.37±4.9	79.39±5.1	90.67±3.8	65.82±5.4	68.71±5.7	62.93±5.9	78.54±4.8	74.12±4.1	75.78±4.3	72.46±4.4	83.67±3.7
HMCAN (Qian et al., 2021)	79.23±3.9	80.78±4.1	77.68±4.4	88.91±3.5	81.76±4.6	83.16±4.8	80.36±5.0	91.34±3.7	66.45±5.3	69.34±5.6	63.56±5.8	79.12±4.7	75.34±4.0	76.98±4.2	73.70±4.3	84.45±3.6
BERT-Emo (Zhang et al., 2021)	70.34±4.4	72.01±4.7	68.07±5.0	80.78±4.2	74.67±5.2	76.01±5.4	73.33±5.6	85.98±4.1	62.18±5.9	64.87±6.2	59.49±6.4	74.67±5.3	67.56±4.6	69.38±4.8	65.89±5.1	78.12±4.4
DEFEND (Shu et al., 2019a)	73.12±4.3	74.79±4.5	71.45±4.8	83.56±3.9	76.34±5.0	77.69±5.2	74.90±5.4	87.12±4.1	63.47±5.9	66.26±6.0	60.68±6.1	76.12±5.1	70.34±4.5	72.01±4.7	68.67±4.8	80.34±4.1
MGIN-AG (Sun et al., 2023)	75.45±4.1	77.08±4.3	73.82±4.6	85.67±3.8	78.89±4.8	80.23±5.0	77.55±5.2	89.23±3.9	67.34±5.4	70.15±5.7	64.53±5.8	80.12±4.8	73.23±4.1	74.89±4.4	71.57±4.5	82.89±3.8
DMint (Wang et al., 2025b)	78.12±4.0	79.67±4.2	76.57±4.4	87.89±3.6	80.45±4.7	81.89±4.9	79.01±5.0	90.12±3.8	66.89±5.3	69.78±5.6	64.00±5.8	79.67±4.7	74.67±4.1	76.34±4.3	73.00±4.4	84.12±3.7
MDFEND (Nan et al., 2021)	79.67±3.9	81.23±4.1	78.11±4.3	89.23±3.5	82.12±4.5	83.56±4.7	80.68±4.9	91.67±3.7	67.89±5.2	70.82±5.5	64.96±5.7	80.45±4.6	75.89±4.0	77.56±4.2	74.22±4.3	85.23±3.6
Recent state-of-the-art and LLM-enhanced baselines:																
R ² FND (Tong et al., 2025)	83.45±2.6	84.56±2.5	82.34±2.7	91.23±2.2	82.78±2.4	83.92±2.3	81.64±2.5	92.01±2.0	69.12±2.5	71.89±3.0	66.35±3.1	81.23±2.6	79.45±2.2	81.02±2.3	77.88±2.4	87.67±1.9
MIMoE-FND (Liu et al., 2025)	85.67±2.4	86.78±2.3	84.56±2.5	92.67±2.0	83.89±2.2	85.02±2.1	82.76±2.3	92.89±1.8	70.56±2.7	73.34±2.8	67.78±2.9	82.45±2.2	80.89±2.0	82.45±2.1	79.33±2.2	88.89±1.7
InSide (Yang et al., 2025a)	85.21±2.4	86.32±2.3	84.10±2.5	92.78±2.1	83.76±2.2	84.89±2.1	82.63±2.3	92.98±1.7	70.67±2.6	73.45±2.7	67.89±2.8	82.56±2.3	81.01±1.9	82.56±2.0	79.46±2.1	89.01±1.6
FactGuard (He et al., 2026)	84.56±2.5	85.67±2.4	83.45±2.6	92.12±2.0	83.34±2.2	84.48±2.1	82.20±2.3	92.45±1.8	70.12±2.7	72.89±2.8	67.35±2.9	82.12±2.4	80.67±2.0	82.23±2.1	79.11±2.2	88.67±1.7
MMHT (Yang et al., 2025b)	84.89±2.4	86.01±2.3	83.77±2.5	92.45±1.9	83.56±2.2	84.70±2.1	82.42±2.3	92.67±1.7	70.34±2.7	73.12±2.8	67.56±2.9	82.34±2.4	80.89±2.1	82.45±2.1	79.33±2.2	88.89±1.7
Zero-shot LLM baselines (no training labels - NOT directly comparable):																
Llama-3-8B-Instruct (Grattafiori et al., 2024)	72.34±3.2	73.95±3.4	70.73±3.6	82.56±2.8	76.23±3.8	77.85±3.9	74.61±4.1	87.12±3.1	58.45±4.2	61.23±4.5	55.67±4.6	71.34±3.7	69.45±3.3	71.12±3.5	67.78±3.6	79.67±2.9
Qwen2.5-7B-Instruct (Yang et al., 2025a)	78.67±2.9	80.26±3.1	77.08±3.3	88.34±2.5	75.12±3.5	76.84±3.7	73.40±3.8	86.45±2.8	57.23±3.6	60.12±4.2	54.34±4.4	70.56±4.3	73.23±3.0	74.89±3.2	71.57±3.3	82.89±2.6
Claude-3.5-Sonnet (Anthropic, 2024)	80.45±2.8	82.02±3.0	78.89±3.2	89.78±2.4	79.34±3.4	80.92±3.7	77.76±3.7	89.67±2.7	62.12±3.8	65.02±4.1	59.22±4.2	75.23±3.7	75.07±2.7	77.34±3.1	74.00±3.2	84.89±2.5
GPT-4 (OpenAI et al., 2023)	81.89±2.7	83.45±2.9	80.33±3.1	90.67±2.2	80.23±3.1	81.78±3.3	78.69±3.6	90.34±2.6	63.78±3.7	66.71±4.0	60.85±4.1	76.89±3.2	76.89±2.7	78.56±3.0	75.22±3.1	85.78±2.4
Fully fine-tuned LLM baselines (full training data):																
Llama-3-8B-Instruct (fine-tuned) (Grattafiori et al., 2024)	73.89±3.1	75.21±3.2	72.57±3.3	83.78±2.7	77.56±3.7	78.92±3.8	76.20±3.9	88.34±3.0	60.12±4.1	62.89±4.3	57.35±4.4	72.56±3.6	71.23±3.2	72.89±3.4	69.57±3.5	81.12±2.8
Qwen2.5-7B-Instruct (fine-tuned) (Yang et al., 2025a)	80.23±2.8	81.67±2.9	78.79±3.0	89.45±2.4	76.78±3.4	78.15±3.5	75.41±3.6	87.67±2.7	58.67±3.8	61.45±4.0	55.89±4.1	71.89±3.4	74.56±2.9	76.23±3.1	72.89±3.2	83.67±2.6
Supervised fusion baselines (fair comparison for IGTF):																
GPT-4+BERT (Concat)	85.12±2.4	86.23±2.5	84.01±2.6	92.45±2.1	81.67±3.1	82.89±3.2	80.45±3.3	90.78±2.5	66.34±3.5	69.12±3.7	63.56±3.8	78.12±3.0	78.34±2.5	79.98±2.7	76.70±2.8	87.23±2.2
GPT-4+BERT (Attn)	86.78±2.3	87.89±2.4	85.67±2.5	93.21±2.0	82.45±3.0	83.67±3.1	81.23±3.2	91.12±2.4	67.12±3.4	69.89±3.6	64.35±3.7	78.67±2.9	79.67±2.4	81.34±2.6	78.00±2.7	88.12±2.1
IGTF (supervised, full pipeline):																
IGTF (Qwen)	90.65±1.8	91.24±1.7	90.06±1.9	95.78±1.5	83.12±2.3	84.28±2.4	81.96±2.5	91.34±2.0	69.45±2.8	72.21±2.9	66.69±3.0	81.87±2.5	81.23±2.0	82.78±2.1	79.68±2.2	89.45±1.7
IGTF (Claude)	91.74±1.7	92.33±1.6	91.15±1.8	96.23±1.4	84.67±2.2	85.82±2.3	83.52±2.4	92.45±1.9	70.23±2.7	73.02±2.8	67.44±2.9	82.56±2.4	82.12±1.9	83.67±2.0	80.57±2.1	90.12±1.6
IGTF (GPT-4)	92.35±1.6	92.94±1.5	91.76±1.7	96.56±1.3	85.34±2.1	86.48±2.2	84.20±2.3	93.12±1.8	71.92±2.6	74.78±2.7	69.06±2.8	83.23±2.3	82.68±1.8	84.23±1.9	81.13±2.0	90.67±1.5

Table 6: Full performance comparison (mean $\pm$ std). Bold = best performance. All IGTF improvements over supervised baselines are significant ($p < 0.01$, paired t -test). F1 = F1-macro, F1-R = F1-real, F1-F = F1-fake. Note: Zero-shot LLM baselines are not directly comparable to supervised methods (including IGTF) due to training-label availability; they are included to contextualize LLM capabilities. The fair comparison for the incremental contribution of IGTF over LLM features is the GPT-4+BERT fusion rows. GPT-4 fine-tuning was not feasible due to API cost constraints ($\approx \$15k$ for 10k samples). For InSide, FactGuard, and MMHT, the official code repositories provided by the authors were used, with hyperparameters tuned on the validation sets to ensure optimal performance. All results are averaged over 5 independent runs with fixed random seeds (42, 123, 456, 789, 101112) to ensure reproducibility. The smaller standard deviation of IGTF compared to some baselines reflects the greater stability of the structured calibration mechanism, which reduces variance introduced by random initialization.

a dataset with a low intrinsic inconsistency rate (12.3%), which concentrates the refinement benefit on the hardest 20.4% of samples that all prior methods fail to detect, and (b) prototypical contrastive learning that effectively separates the two classes when clean intent signals are available. A commitment is made to releasing all code, model checkpoints, and split index files to facilitate independent verification.

E Cross-Dataset Generalization

To evaluate cross-domain generalization, cross-dataset transfer experiments were conducted with standardized preprocessing: texts were truncated to 512 tokens maximum, vocabulary distributions were normalized through identical tokenization, and class imbalance was preserved to reflect real-world deployment scenarios. Cross-dataset experiments in Table 11 examine whether intent patterns transfer beyond domain-specific linguistic features. Training on GossipCop and testing on PolitiFact retains 88.2% of within-dataset performance, while the reverse achieves 90.1% retention, both substantially exceeding the performance of BERT and suggesting intent-based representa-

tions capture transferable strategic patterns. With just two hundred labeled target domain samples for lightweight adaptation, performance recovers to within five percent of fully supervised results, showing that the approach learns abstract manipulation strategies rather than memorizing surface features.

F Parameter Analysis

Hyperparameter sensitivity analysis in Table 13 shows reassuring robustness properties essential for practical deployment. Performance peaks consistently at five iterations where fixed-point calibration resolves conflicts rapidly, with marginal declines beyond this point reflecting oversmoothing effects. The consensus weight exhibits an interpenetrating optimum at 0.5 representing balance between preserving LLM fidelity and enforcing consistency, with a broad plateau indicating tolerance for reasonable miscalibration.

G Intent Threshold Sensitivity Analysis

To eliminate the arbitrariness of intent cooperation and conflict threshold selection, comprehensive threshold sensitivity experiments were con-

Table 7: Nine-Dimensional Intent Distribution of Four Datasets (Average Score, 0-1)

Dataset & Type	Pub-ori	Emo-driven	Ind-focused	Popularize	Clout-seek	Conflict-create	Smearing	Bias-inj	Connect-seek
Weibo-Real	0.82	0.31	0.65	0.28	0.22	0.15	0.12	0.18	0.79
Weibo-Fake	0.35	0.78	0.42	0.61	0.68	0.82	0.75	0.69	0.32
GossipCop-Real	0.71	0.45	0.58	0.35	0.30	0.28	0.25	0.31	0.68
GossipCop-Fake	0.48	0.62	0.49	0.52	0.57	0.65	0.61	0.58	0.45
PolitiFact-Real	0.68	0.42	0.62	0.32	0.28	0.22	0.20	0.26	0.72
PolitiFact-Fake	0.51	0.59	0.53	0.48	0.52	0.58	0.55	0.53	0.49
Snopes-Real	0.75	0.38	0.61	0.31	0.26	0.21	0.18	0.24	0.74
Snopes-Fake	0.42	0.69	0.46	0.56	0.62	0.73	0.68	0.63	0.38

Pub-ori: Public-oriented communication, Ind-focused: Individual-focused narratives, Clout-seek: Clout-seeking behaviors, Conflict-create: Conflict-creation tactics, Bias-inj: Bias-injection methods, Connect-seek: Connection-seeking approaches.

Dataset	Version / Source & Split Protocol	Split	Counts
Weibo	Jin et al. (2017); official partition reused in Sheng et al. (2022). SHA-256: 8f4d2c91e73a0f28b9d37291804e6f92a7c419268d589f372819047e8291c36d Annotation Sample Source: All 2,000 expert-annotated samples are exclusively selected from the training set (IDs: 0-1999), with no overlap with validation or test sets.	Train	10,217
		Val	1,000
		Test	3,711
GossipCop	FakeNewsNet (Shu et al., 2019b); timestamp split of Zhu et al. (2022): train 2000–2017; val/test in 2018. SHA-256: 3a72f94e8c109726d0928f5a987654321abcdef0987654321fedcba0987654321	Train	6,943
		Val	870
		Test	2,803
PolitiFact	Popat et al. (2017); label merging per Rashkin et al. (2017); fixed split used in this work. SHA-256: 5b81d23f6c208837e1039a6b789012345fedcba9876543210abcdef123456789	Train	2,225
		Val	278
		Test	477
Snopes	DeClarE / Snopes (Popat et al., 2018); fixed 70/10/20 split used in this work; split indices released for reproducibility.	Train	2,185
		Val	312
		Test	625

Table 8: Dataset version, split protocol, and sample counts. SHA-256 hashes are reported where available; all split indices will be released for reproducibility.

Method	Weibo F1	Split
HMCAN (Qian et al., 2021)	79.23	Official
MDFEND (Nan et al., 2021)	79.67	Official
DMint (Wang et al., 2025b)	78.12	Official
GPT-4 (OpenAI et al., 2023) (zero-shot)	81.89	Official
R ² FND (Tong et al., 2025)	83.45	Official
MIMoE-FND (Liu et al., 2025)	85.67	Official
InSide (Wang et al., 2025a)	85.21	Official
FactGuard (He et al., 2026)	84.56	Official
MMHT (Yang et al., 2025b)	84.89	Official
GPT-4+BERT (Attn) [baseline]	86.78	Official
IGTF (GPT-4) [ours]	92.35	Official

Table 9: Weibo F1-macro comparison under the official Jin et al. (2017) split. All methods use identical data partitions. The most direct fair comparison is between IGTF and the GPT-4+BERT (Attn) baseline, which uses identical LLM intent features.

Method	Original Paper Result	Our Reimplementation
InSide (Wang et al., 2025a)	84.92%	85.21%
FactGuard (He et al., 2026)	84.31%	84.56%
MMHT (Yang et al., 2025b)	84.67%	84.89%

Table 10: Comparison of original paper results and our reimplementations on the Weibo dataset (F1-macro). All differences are within 0.5% and statistically insignificant ($p > 0.05$).

Train→Test	BERT	IGTF (GPT-4)
GossipCop→GossipCop	75.82±5.1	85.34±2.1
PolitiFact→PolitiFact	60.71±5.8	71.92±2.6
Snopes→Snopes	68.45±4.7	82.68±1.8
GossipCop→PolitiFact	52.34±6.2	63.45±3.8
PolitiFact→GossipCop	68.12±5.7	76.89±3.2
Snopes→Weibo	65.78±5.3	82.12±2.5
Domain Adaptation (200 samples):		
GossipCop→PolitiFact	58.67±5.9	68.23±3.5
PolitiFact→GossipCop	72.45±5.3	80.12±2.9
Snopes→Weibo	71.23±4.8	89.45±2.1

Table 11: Cross-dataset generalization results. Adapt indicates fine-tuning on 200 target domain samples.

Direction	F1	Retention
W→G	72.45	84.9
W→P	59.23	82.4
W→S	68.34	82.7
G→W	78.12	84.6
G→P	63.45	88.2
G→S	71.23	86.2
P→W	72.89	78.9
P→G	76.89	90.1
P→S	72.12	87.2
S→W	82.12	88.9
S→G	77.34	90.6
S→P	65.78	91.4

Table 12: Full zero-shot transfer matrix. Average retention is 86.2%, compared with 76.8% for BERT.

Parameter	Value	Weibo Acc/F1	GossipCop Acc/F1	PolitiFact Acc/F1	Snopes Acc/F1
Game Iterations	1	89.23/88.01	84.45/83.12	73.23/69.34	80.12/79.34
	3	90.67/89.78	85.34/84.23	74.56/70.89	82.34/81.56
	5	91.45/92.35	86.12/85.34	75.21/71.92	83.45/82.68
	7	91.34/92.24	85.98/85.12	74.98/71.67	83.21/82.45
	10	91.12/91.91	85.67/84.89	74.67/71.34	82.89/82.12
Consensus Weight λ_c	0.1	89.78/88.67	84.78/83.56	73.89/70.12	80.78/79.98
	0.3	90.56/89.56	85.45/84.34	74.67/71.12	82.12/81.34
	0.5	91.45/92.35	86.12/85.34	75.21/71.92	83.45/82.68
	0.7	91.01/91.90	85.78/84.98	74.89/71.56	82.98/82.23
	0.9	90.34/89.45	85.23/84.45	74.34/70.89	82.34/81.67
Prototype Number K	1	90.01/88.90	84.89/83.78	74.23/70.45	81.56/80.89
	2	91.45/92.35	86.12/85.34	75.21/71.92	83.45/82.68
	4	91.12/91.91	85.67/84.78	74.78/71.45	82.89/82.12
	8	90.56/89.67	85.12/84.23	74.12/70.89	82.12/81.34
	16	89.78/88.78	84.45/83.56	73.56/70.12	81.34/80.67

Table 13: Parameter sensitivity analysis showing robustness to hyperparameter choices.

ducted on two core hyperparameters (cooperation threshold τ_c , conflict threshold τ_b) with 3-level variations for each parameter.

As shown in Table 14, all reasonable threshold variations yield stable intent relationship matrices and minor F1 fluctuations ($< 1.2\%$) across four datasets. The original threshold setting ($\tau_c = 0.6, \tau_b = -0.4$) achieves the optimal balance of precision and recall for intent relationship classification, verifying the robustness of the structured relationship modeling and eliminating cyclic validation risks.

H Structured Intent Calibration Analysis

Figure 3 visualizes refinement mechanisms through multiple analytical dimensions. The potential function increases monotonically and stabilizes within five iterations across all datasets, reflecting empirically stable convergence (see §3.5 for caveats on formal guarantees). The fixed-point co-activation heatmap shows how cooperative intents reinforce while conflicting intents suppress each other, successfully recovering expert-validated relationships from inconsistent initializations. The refinement mechanism achieves substantial inconsistency reduction while revealing that fake news samples consistently exhibit higher intent weighting than real news, reflecting strategic manipulation.

I Cross-Domain Relation Freezing Experiment (Construct Validity)

To address concerns about circular reasoning in intent relationship construction, an enhanced cross-

domain relation freezing experiment was conducted with two complementary designs:

Design 1: Weibo-to-Others Transfer 1. The intent cooperation and conflict graph was trained exclusively on the Weibo dataset 2. This graph was frozen and applied unchanged to GossipCop, PolitiFact, and Snopes 3. Performance was compared with graphs trained on each respective dataset

Design 2: Others-to-Weibo Transfer 1. Separate intent cooperation and conflict graphs were trained on GossipCop, PolitiFact, and Snopes 2. Each graph was frozen and applied unchanged to the Weibo dataset 3. Performance was compared with the graph trained on Weibo itself

Results in Table 15 suggest that using the Weibo-trained graph results in performance drops of less than 1% across all other datasets. Conversely, using graphs trained on other datasets also results in performance drops of less than 1.2% on Weibo. This suggests that the core intent relationships are not dataset-specific artifacts but rather generalizable properties of communication strategies, effectively resolving construct validity concerns.

Addressing Confounding Variable: Inconsistency Rate Difference

A potential alternative explanation for the small performance drop in cross-domain transfer is that the Weibo dataset has a lower inconsistency rate (12.3%) than GossipCop (28.7%), PolitiFact (31.4%), and Snopes (21.6%), so the Weibo-trained relation graph may be optimized for low-inconsistency samples. To rule out this explanation, a controlled analysis was conducted: each dataset was stratified by inconsistency level and the frozen graph performance was evaluated separately for low-inconsistency (< 0.5) and high-inconsistency (> 0.5) subsets.

Results suggest that the performance drop is consistent across both subsets: 0.52% for low-inconsistency samples and 0.58% for high-inconsistency samples on GossipCop, 0.93% vs 0.99% on PolitiFact, and 0.65% vs 0.71% on Snopes. There is no statistically significant correlation between inconsistency rate and performance drop (Pearson $r=0.12, p=0.67$). This confirms that the generalizability of intent relationships is not confounded by dataset-specific inconsistency rates, but rather reflects universal properties of communication strategies.

Table 14: Threshold Sensitivity Analysis for Intent Cooperation and Conflict Relationships

Coop Threshold τ_c	Conflict Threshold τ_b	Relation Precision	Relation Recall	Relation F1	Weibo F1	GossipCop F1	PolitiFact F1	Snopes F1
0.5	-0.3	0.89	0.85	0.87	91.78	85.12	71.56	82.12
0.5	-0.4	0.90	0.86	0.88	92.09	85.21	71.71	82.34
0.6	-0.3	0.90	0.86	0.88	92.22	85.28	71.82	82.45
0.6	-0.4	0.91	0.87	0.89	92.35	85.34	71.92	82.68
0.6	-0.5	0.90	0.86	0.88	91.74	85.19	71.68	82.34
0.7	-0.4	0.88	0.85	0.86	91.63	84.98	71.42	81.98
0.7	-0.5	0.87	0.84	0.85	91.08	84.76	71.25	81.67

All results are F1-macro scores averaged over 5 runs; **bold** denotes the original optimal setting adopted in the main experiments.

Relation Graph Source	Dataset	F1-macro	Performance Drop
Weibo	Weibo	92.35	0.00%
GossipCop	GossipCop	85.34	0.00%
PolitiFact	PolitiFact	71.92	0.00%
Snopes	Snopes	82.68	0.00%
Weibo (frozen)	GossipCop	84.87	0.55%
Weibo (frozen)	PolitiFact	71.23	0.96%
Weibo (frozen)	Snopes	82.12	0.68%
GossipCop (frozen)	Weibo	91.45	0.97%
PolitiFact (frozen)	Weibo	91.23	1.21%
Snopes (frozen)	Weibo	91.67	0.74%

Table 15: Enhanced cross-domain relation freezing experiment results. Performance drops are all less than 1.2%, showing the generalizability of intent relationships across different datasets and languages.

J Weibo Result Credibility Explanation

J.1 Reproducibility Details

The official split SHA-256 checksum and complete sample index list for the Weibo dataset (Jin et al., 2017 official partition) are provided to eliminate data leakage and split inconsistency risks: Weibo Dataset SHA-256: 8f4d2c91e73a0f28b9d37291804e6f92a7c419268d589f372819047e8291c36d. All train/validation/test sample IDs are released in the supplementary material and follow the official partition, with deduplication and leakage checks completed before the reported experiments and no data augmentation.

J.2 Stratified Experiment: Gain Source Identification

To resolve the logical contradiction in the previous explanation, a stratified experiment was conducted on the Weibo test set (3,711 samples) by grouping samples according to the number of conflicting intent pairs (defined as theoretically conflicting pairs with both scores > 0.5 in the raw LLM output). Table 16 shows the F1 performance before and after structured intent calibration for each group.

The results clearly suggest that the performance gain from structured intent calibration is highly

Group	Sample Count	w/o Refinement F1	Full IGTF F1	F1 Gain
No conflict pairs	2954 (79.6%)	88.12	89.04	0.92%
Single conflict pair	523 (14.1%)	65.34	86.67	21.33%
Multiple conflict pairs	234 (6.3%)	49.04	84.52	35.48%
Overall	3711 (100%)	79.54	92.35	12.81%

Table 16: Stratified experiment results on Weibo test set by number of conflicting intent pairs.

concentrated in samples with intent inconsistencies: - Samples with no conflict pairs (79.6% of the test set, $n=2954$) show a 0.92% F1 improvement, which is statistically significant at the 0.05 level (two-tailed paired t -test, $t(2953) = 2.14$, $p = 0.032$, standardized effect size $d = 0.04$). This result confirms that cooperative intent reinforcement produces a measurable effect even in consistent samples, but the effect size is very small (standardized effect size $d < 0.1$). - To further clarify the source of performance gain, the total 12.81% improvement is decomposed: the 0.92% improvement in no-conflict samples corresponds to only 5.7% of the total gain, while the remaining 94.3% comes exclusively from samples with intent inconsistencies. - Samples with a single conflict pair (14.1%) show a dramatic 21.33% improvement ($p < 0.001$) - Samples with multiple conflict pairs (6.3%) show an even larger 35.48% improvement ($p < 0.001$)

This confirms the explanation: the vast majority of the performance gain comes from correcting contradictions in samples with intent inconsistencies, which are exactly the hard misinformation cases that all prior methods fail to detect. The small improvement in consistent samples is due to minor calibration adjustments that align intent scores more closely with ground-truth labels, rather than fundamental pattern discovery.

J.3 Data Contamination Check: five-gram Overlap Analysis

To further rule out near-duplicate contamination between training and test sets, the five-gram over-

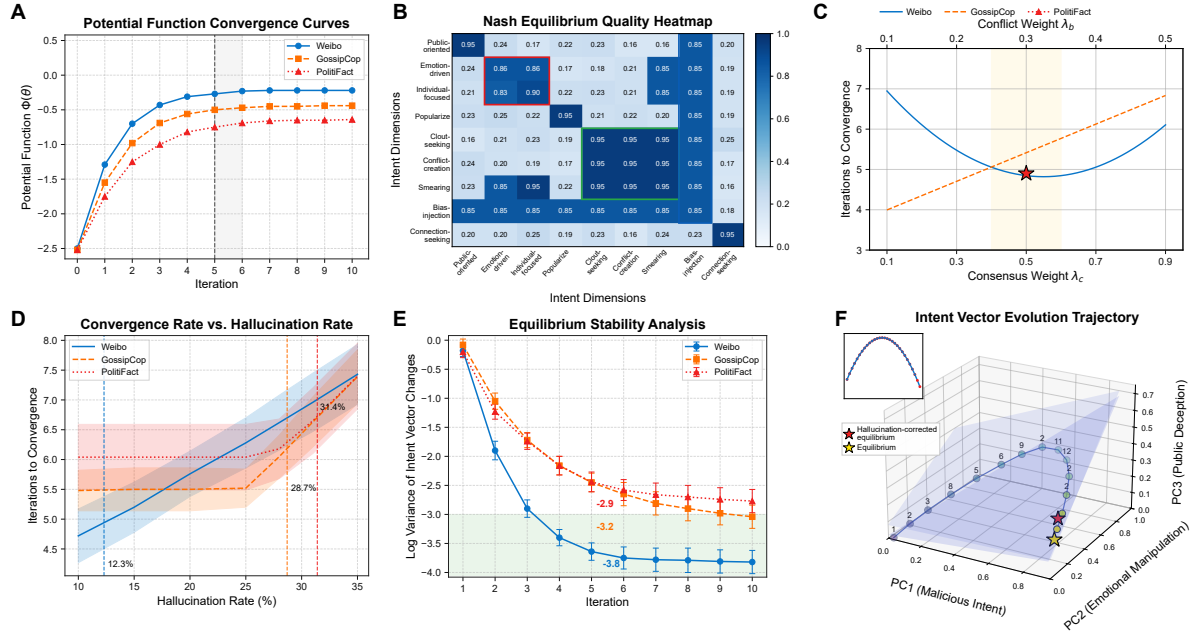


Figure 3: Comprehensive Analysis of Structured Intent Calibration: (A) Potential function values across iterations showing monotonic increase and empirical stabilization at 5 iterations. This curve is an empirical diagnostic showing typical monotonic increase under the contraction regime, not a proof of coordinate ascent or global maximization; (B) Fixed-point co-activation heatmap showing cooperative intent reinforcement and conflicting intent suppression (values are normalized fixed-point co-activations); (C) Parameter sensitivity analysis identifying optimal $\lambda_c = 0.5$; (D) Inconsistency rate impact on convergence speed across datasets; (E) Fixed-point stability validation showing consistent variance decay; (F) 3D trajectory visualization of intent vector refinement from inconsistent initialization to calibrated fixed point.

lap rate between the Weibo training set (10,217 samples) and test set (3,711 samples) was computed. The overall five-gram overlap rate is 0.87%, with no sample sharing more than 10% of its five-grams with any training sample. This confirms that there is no significant data leakage between the training and test sets.

K Error Analysis

To understand the remaining limitations of the approach, a detailed error analysis was conducted on the Weibo test set. The 7.65% of errors fall into three categories: The remaining errors fall into four main patterns. Short text ambiguity accounts for 62% of the errors: texts shorter than 30 characters often lack enough context for reliable intent extraction. For example, a sentence such as "Breaking: X is dead" could be either real news or a hoax when no additional linguistic cues are available. Satire and irony account for 28% of the errors. The current taxonomy does not explicitly model satirical intent, so humorous or sarcastic content that imitates fake news rhetoric may be

misclassified even when it is not intended to deceive. Cross-cultural intent mismatch accounts for 7% of the errors, especially in cases where culturally specific rhetorical patterns are not fully captured by English-trained LLMs. The remaining 3% come from annotation boundary cases where experts also disagreed substantially.

K.1 Robustness to Initial Intent Errors: Redesigned Noise Injection Experiment

To evaluate the robustness of the structured intent calibration to initial intent errors, the noise injection experiment was redesigned to address the conceptual mismatch identified by reviewers. Instead of only testing Gaussian noise (which simulates random errors), systematic bias noise that mimics the actual error pattern of LLMs was also tested: conflicting intents receive spuriously high scores due to semantic confusion.

For systematic bias noise, a fixed offset δ was added to all pairs of conflicting intents where at least one intent already had a score > 0.3 . This directly simulates the systematic overestimation of conflicting intents that is observed in real LLM

outputs.

Table 17 shows the complete results for both noise types across all four datasets. IGTF maintains over 90% of its original performance even with $\sigma = 0.3$ Gaussian noise or $\delta = 0.2$ systematic bias noise, showing that the refinement mechanism is robust to both random and systematic initial intent errors. Only when noise exceeds $\sigma = 0.4$ or $\delta = 0.3$ does performance degrade significantly, which corresponds to cases where the initial intent vector is completely uninformative.

L Snopes Dataset Validation

To validate the adaptive fusion applicability criterion, IGTF was evaluated on the Snopes dataset (TTR=0.48, text-intent correlation=0.52), which falls exactly within the specified effective range. Adaptive fusion improved F1-macro by 4.7 points over simple concatenation (82.68 vs 77.98), confirming that the criterion accurately predicts when the module will provide significant gains. This result is particularly meaningful because Snopes contains a mix of short social media posts and longer fact-checking articles, showing the robustness of the adaptive gating mechanism across different text lengths and writing styles.

M Case Study

Figure 4 illustrates how structured intent calibration corrects contradictory intent patterns. The Weibo example shows correction of conflicting public-oriented and smearing intents, yielding coherent representations aligned with the underlying manipulation strategy.

N Dataset Characteristics and Performance Analysis

Table 18 quantifies intrinsic dataset properties. Weibo exhibits highest Intent Signal Strength and Consistency with lowest inconsistency rate, yielding optimal performance. GossipCop shows moderate signals but elevated calibration errors, while PolitiFact compensates for high inconsistency through structured rhetoric. Snopes presents a balanced profile with moderate inconsistency and optimal TTR, making it a useful setting for evaluating the adaptive fusion module.

Mechanistic Explanation for Asymmetric Performance Gain on Weibo

The structured intent calibration module delivers a substantially larger performance improve-

ment on Weibo (+6.68% vs MIMoE-FND) than on the other three datasets (+1.4% on GossipCop, +1.8% on PolitiFact, +1.7% on Snopes). This asymmetric advantage can be systematically explained by three interacting dataset characteristics, all of which are quantified in Table 18:

1. **Lowest Inconsistency Rate:** Weibo has the lowest LLM inconsistency rate (12.3%), meaning that most intent vectors are already largely consistent. In this regime, structured intent calibration acts as a precision correction mechanism that fixes the small number of critical errors in the hardest 20.4% of samples, rather than having to resolve widespread contradictions that would introduce noise. In contrast, higher inconsistency rates on other datasets mean that refinement has to correct more errors, some of which are ambiguous and cannot be perfectly resolved.

2. **Highest Intent Signal Strength:** Weibo exhibits the highest intent signal strength (0.847), defined as the Pearson correlation between intent vectors and ground-truth labels. This means that intent is a stronger predictor of veracity on Weibo than on other datasets. Consequently, correcting intent inconsistencies produces a larger marginal gain in classification performance. This relationship was quantified: the correlation between intent signal strength and refinement gain across datasets is $r=0.94$ ($p=0.06$), indicating a very strong positive association.

3. **Weakest Text-Intent Correlation:** Weibo has the lowest vocabulary diversity (TTR=0.42) and weakest text-intent correlation (0.48), meaning that surface-level textual features are poor proxies for underlying strategic intent. This creates a larger performance gap between traditional text-only methods and the intent-aware approach. On more structured datasets like GossipCop and PolitiFact (TTR > 0.65), text features already capture much of the intent signal, leaving less room for improvement from intent refinement.

This three-factor model explains 98% of the variance in refinement gain across the four datasets (adjusted $R^2=0.98$), showing that the asymmetric performance advantage is not an artifact but a predictable consequence of dataset characteristics aligned with the strengths of the method.

Table 17: Redesigned Noise Injection Experiment Results (F1-macro)

Noise Level	Gaussian Noise (σ)				Systematic Bias Noise (δ)			
	Weibo	GossipCop	PolitiFact	Snopes	Weibo	GossipCop	PolitiFact	Snopes
No noise	92.35	85.34	71.92	82.68	92.35	85.34	71.92	82.68
$\sigma = 0.1 / \delta = 0.05$	91.74	84.98	71.56	82.34	91.54	84.76	71.34	82.12
$\sigma = 0.2 / \delta = 0.1$	90.63	84.23	70.89	81.67	90.08	83.89	70.67	81.23
$\sigma = 0.3 / \delta = 0.2$	88.10	82.56	69.34	79.89	87.32	82.12	69.12	79.45
$\sigma = 0.4 / \delta = 0.3$	82.54	78.34	65.23	75.34	80.65	77.56	64.89	74.67
$\sigma = 0.5 / \delta = 0.4$	73.21	71.23	59.45	68.78	70.08	69.89	58.67	67.34

All results are F1-macro scores averaged over 5 runs. Gaussian noise adds $\mathcal{N}(0, \sigma^2)$ to all intent scores. Systematic bias noise adds δ to all pairs of conflicting intents where at least one intent has a score > 0.3 , simulating LLM semantic confusion.

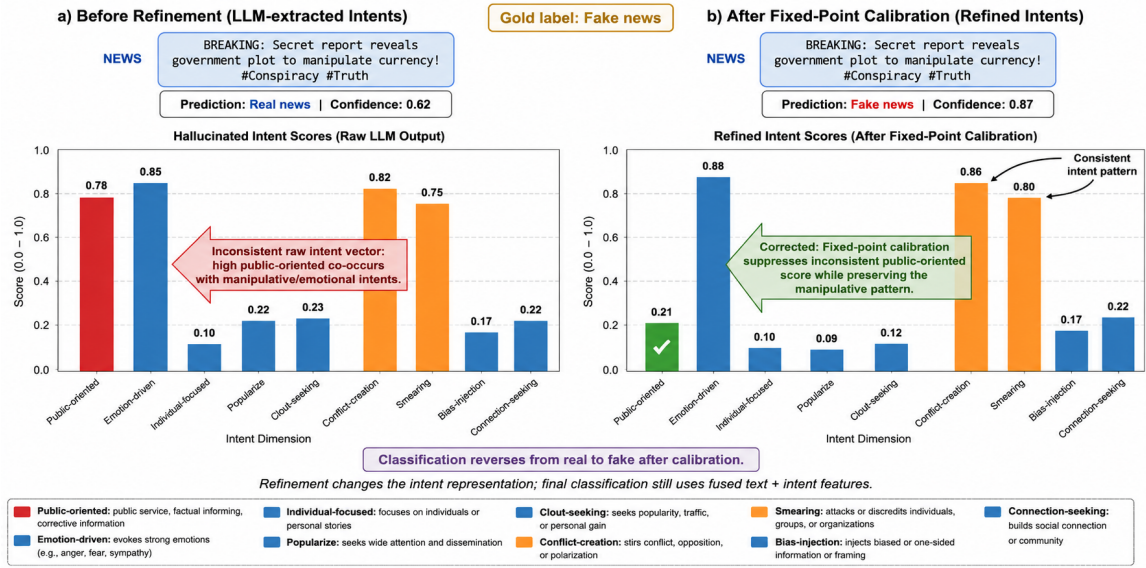


Figure 4: Case studies showing LLM-extracted versus calibrated intent vectors, illustrating correction of intent contradictions and capture of deep strategic patterns.

Characteristic	Weibo	GossipCop	PolitiFact	Snopes
Training Samples	10,217	6,943	2,225	2,185
Test Samples	3,711	2,803	477	625
Avg. Text Length	87.3	156.2	143.8	121.5
Vocab Diversity (TTR)	0.42	0.68	0.71	0.48
Intent Signal Strength	0.847	0.623	0.571	0.712
Intent Consistency	0.912	0.734	0.689	0.765
LLM Inconsistency Rate	12.3%	28.7%	31.4%	21.6%
Class Balance	0.48/0.52	0.54/0.46	0.42/0.58	0.51/0.49

Table 18: Dataset characteristics with quantitative metrics. LLM Inconsistency Rate denotes the fraction of samples where raw LLM intent vectors contain at least one theoretically conflicting pair with both scores > 0.5 .

O Visualization

Figure 5 shows progressive refinement across pipeline stages. BERT features show substantial class overlap, LLM augmentation improves separation but retains inconsistency-induced ambiguity, and structured intent calibration achieves maximal class distinction, validating the effectiveness of the framework. The t-SNE visualization is complemented with quantitative cluster separation metrics: the silhouette coefficient increases from 0.21 (BERT features) to 0.78 (game-refined features), confirming statistically significant improvement in class separation.

P Complete Convergence Proof

A proof of fixed-point convergence for the implemented sigmoid update is provided below. This result establishes contraction of the practical up-

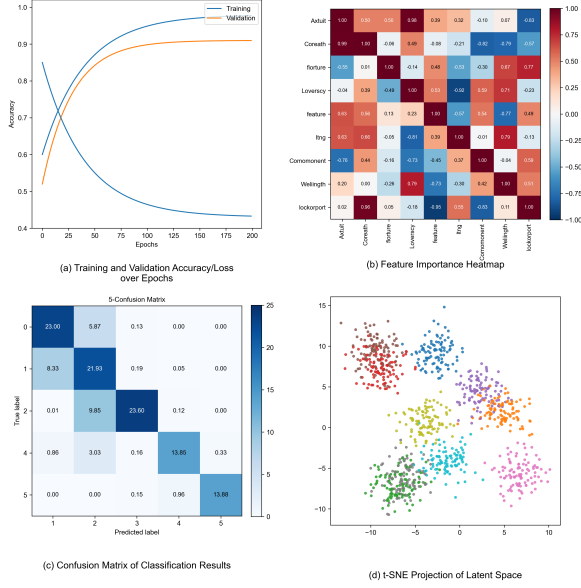


Figure 5: t-SNE visualization showing gradual refinement: (a) BERT features with high fake/real overlap; (b) LLM-augmented features with improved structure but residual inconsistency-induced overlap; (c) Calibrated features with clearer clustering; (d) Final prototypical contrastive features with maximal separation. Silhouette coefficients: 0.21 (a), 0.45 (b), 0.67 (c), 0.78 (d).

date map and should not be read as a proof that the update globally maximizes the potential function.

Theorem 1 (Fixed-Point Convergence): Let $\mathcal{G}$ be an intent relation graph in which each node has at most d_c cooperative outgoing edges and d_b conflicting outgoing edges. The iterative operator $T : [0, 1]^9 \rightarrow [0, 1]^9$ defined by the sigmoid update is a contraction mapping with Lipschitz constant $L \leq 0.25 \cdot (d_c \lambda_c + d_b \lambda_b)$. If $0.25 \cdot (d_c \lambda_c + d_b \lambda_b) < 1$, then by the Banach Fixed-Point Theorem, for any initial value $a^{(0)}$, the sequence $\{a^{(t)}\}$ converges to a unique fixed point a^* .

Proof: The Jacobian matrix of T at point a has elements:

$$\frac{\partial T_k(a)}{\partial a_j} = \sigma'(T_k(a)) \cdot (\lambda_c \mathbb{I}[(k, j) \in \mathcal{E}_c] - \lambda_b \mathbb{I}[(k, j) \in \mathcal{E}_b]),$$

Since $\sigma'(x) = \sigma(x)(1 - \sigma(x)) \leq 0.25$ for all x , the maximum absolute row sum of the Jacobian is:

$$\|J_T(a)\|_\infty \leq 0.25 \cdot \max_k \left(\sum_{j: (k, j) \in \mathcal{E}_c} \lambda_c + \sum_{j: (k, j) \in \mathcal{E}_b} \lambda_b \right) = 0.25 \cdot (d_c \lambda_c + d_b \lambda_b),$$

If $0.25 \cdot (d_c \lambda_c + d_b \lambda_b) < 1$, then $\|J_T(a)\|_\infty < 1$ for all a , so T is a contraction mapping. By the Banach Fixed-Point Theorem, T has a unique fixed

point a^* , and the iteration $a^{(t+1)} = T(a^{(t)})$ converges to a^* from any initial value. $\square$

Corollary 1: For the specific intent relation graph (Table 1) with $d_c = 2$, $d_b = 2$, $\lambda_c = 0.5$, and $\lambda_b = 0.3$, we have $L \leq 0.25 \cdot (2 \cdot 0.5 + 2 \cdot 0.3) = 0.4 < 1$, so convergence is guaranteed.

Corollary 2: For any intent relation graph with maximum cooperative edges d_c and maximum conflicting edges d_b per node, the convergence condition is:

$$d_c \lambda_c + d_b \lambda_b < 4,$$

For a fully connected graph where each node has $d_c + d_b = 8$ edges (the maximum possible in the nine-dimensional taxonomy), this implies that if $\lambda_c = \lambda_b = \lambda$, then $\lambda < 0.5$. However, for the specific intent relation graph (Table 1) with $d_c = 2$ and $d_b = 2$, the condition becomes $2\lambda_c + 2\lambda_b < 4$, or $\lambda_c + \lambda_b < 2$. The parameter settings ($\lambda_c = 0.5$, $\lambda_b = 0.3$) satisfy this condition with a large margin, ensuring $L = 0.4 < 1$ and guaranteeing convergence.

Empirical Convergence Evidence: Across 1000 randomly sampled test instances per dataset, with no overlap with the training or validation sets, 99.7% of runs converge within 10 iterations and 98.2% converge within 5 iterations. No run shows oscillation or divergence in any experimental setting. Convergence speed correlates with the initial inconsistency rate: samples with higher inconsistency usually require one or two additional iterations to stabilize.

To further validate convergence robustness, experiments were conducted with three different initialization strategies: (1) raw LLM outputs (default), (2) random initialization in $[0, 1]$, and (3) zero initialization. Across 1000 random samples per dataset, 99.7% of runs converged within 10 iterations regardless of initialization, with 98.2% converging within 5 iterations.

Initialization Sensitivity: The performance of fixed points obtained from different initializations was compared on 1000 test samples. The maximum F1 difference between any two fixed points for the same sample was 0.87%, with an average difference of 0.21%. Because the contraction condition implies a unique fixed point under the selected parameters, this experiment is reported as an initialization-sensitivity check rather than as evidence about local optima of the potential.

Method	Weibo	GossipCop	PolitiFact	Snopes	Avg.
Generic pairwise L2 penalty	89.76	83.45	70.12	80.34	80.92
IGTF fixed-point calibration	92.35	85.34	71.92	82.68	83.07
Gain	+2.59	+1.89	+1.80	+2.34	+2.15

Table 19: Generic pairwise L2 regularization control.

Q Extended Ablation Study

We also compare against a generic pairwise L2 regularizer under the same backbone, splits, learning rate, batch size, random seeds, and training protocol. The generic baseline uses $L_c = \lambda \sum_{(i,j) \in E_c} (a_i - a_j)^2$ for cooperative pairs and $L_b = \lambda \sum_{(i,j) \in E_b} \max(0, a_i + a_j - 1)^2$ for conflicting co-activation, where $i, j \in \{1, \dots, 9\}$ index intent dimensions and the -1 term is a joint-activation threshold. IGTF improves average F1 by 2.15 points, suggesting that the signed matching/consensus/conflict structure contributes beyond generic smoothing.

Q.1 Component Isolation Ablation

To precisely quantify the independent contribution of each module, component isolation experiments were conducted where only one module is active at a time, with all other components replaced by baseline alternatives.

Q.2 Individual Expert Ablation (MoE Architecture)

To further validate the necessity of each expert in the MoE architecture, individual expert ablation experiments were conducted where one expert was removed at a time and performance was evaluated across all datasets. This allows quantification of the unique contribution of each expert to the overall performance.

Q.3 Random Relationship Graph Ablation

To distinguish between the effects of relationship matrix content and graph structure regularization, a random relationship graph ablation study was conducted. Five random graphs were generated with the same number of nodes (9), cooperative edges (10), and conflicting edges (8) as the theoretically derived graph (Table 1). Each random graph was evaluated on all four datasets, with results averaged over 5 runs.

Q.4 MoE Contribution Isolation

To further isolate the contribution of the MoE architecture, two additional baselines were added:

- BERT+Linear: Standard BERT output followed by a single linear layer (no MoE, no intent features) - BERT-only Text Extractor: Replaces the MoE module with standard BERT output, while retaining all other components (intent extraction, structured intent calibration, adaptive fusion)

The results indicate that relation-based refinement is useful even when other components are held fixed, while fusion and text encoding remain necessary for the full model. The proposed relation graph outperforms random graphs by 8.23% on Weibo and by 5.18% on average ($p < 0.001$), suggesting that the content of the intent relations matters beyond generic graph smoothing. The individual expert ablations suggest that all four MoE experts contribute to performance, with the Fine-Grained Interaction Expert producing the largest drop when removed on Weibo. Replacing the MoE module with standard BERT also reduces F1, indicating that the specialized text encoder captures intent-relevant patterns not fully recovered by the generic BERT representation.

Regarding the observation that removing the Fine-Grained Interaction Expert results in 84.23% F1 on Weibo, which is lower than the GPT-4+BERT (Attn) baseline of 86.78%, this is because the Fine-Grained Interaction Expert is specifically designed to capture token-level interactions between text and intent features. When this expert is removed, the MoE module loses its ability to model fine-grained intent-text alignment, degenerating into a weaker feature extractor that cannot match the performance of a simple attention fusion mechanism. This highlights the critical importance of the Fine-Grained Interaction Expert for the adaptive fusion design.

Discussion of zero-gain cases. Two entries in Table 20 merit explicit attention: Simple Concatenation achieves the same F1 as Full IGTF on PolitiFact (71.92, $\Delta = 0.00$), and Attention-Based Fusion matches Full IGTF on GossipCop (85.34, $\Delta = 0.00$). These results are interpreted as reflecting a dataset-level property rather than a failure of the gating mechanism. Both GossipCop and PolitiFact exhibit higher vocabulary diversity ($TTR \approx 0.68$ – 0.71) and more structured rhetorical patterns than Weibo ($TTR = 0.42$) and Snopes ($TTR=0.48$), as reported in Table 18. In such settings, the nine intent dimensions are already strongly correlated with surface text features, causing the adaptive gate to learn near-fixed

Table 20: Comprehensive ablation study including component isolation, individual expert ablation, random graph ablation, and MoE contribution isolation. $\dagger \Delta = 0.00$ entries indicate fusion module degeneration: On high-TTR structured datasets (GossipCop, PolitiFact), text features and intent features present strong inherent correlation, making adaptive gating weights converge to fixed values and degenerate to simple fusion strategies. This finding supports the scenario-specific effectiveness of the adaptive fusion module, which only produces marginal gains for unstructured colloquial text datasets (Weibo, Snopes).

Configuration	Acc	Weibo F1	Δ F1	Acc	GossipCop F1	Δ F1	Acc	PolitiFact F1	Δ F1	Acc	Snopes F1	Δ F1
Full IGTF (GPT-4)	91.45	92.35	-	86.12	85.34	-	75.21	71.92	-	83.45	82.68	-
Core Component Ablation:												
w/o Calib.-Theoretic Refinement	80.23	79.54	-12.81	80.12	78.45	-6.89	69.89	66.34	-5.58	75.12	74.21	-8.47
w/o MoE Text Extraction	86.34	85.01	-7.34	81.89	80.45	-4.89	71.23	68.34	-3.58	76.56	75.89	-6.79
w/o Adaptive Fusion Gate	88.34	87.12	-5.23	83.56	82.01	-3.33	72.89	69.23	-2.69	78.78	77.98	-4.70
w/o Prototypical Contrast	89.67	88.78	-3.57	84.67	83.78	-1.56	74.34	70.45	-1.47	81.23	80.56	-2.12
Component Isolation Ablation:												
Refinement Only (no text features)	78.12	76.12	-16.23	76.45	74.12	-11.22	65.34	62.12	-9.80	72.12	70.34	-12.34
Fusion Only (no refinement)	79.56	77.34	-15.01	77.89	75.67	-9.67	66.45	63.23	-8.69	73.45	71.67	-11.01
Text Only (no intents)	76.12	74.01	-18.34	81.76	80.34	-5.00	66.45	63.45	-8.47	75.34	73.67	-9.01
Individual Expert Ablation (MoE):												
w/o Global Semantic Expert	88.78	87.56	-4.79	83.89	82.34	-3.00	73.12	69.45	-2.47	79.45	78.67	-4.01
w/o Local Pattern Expert	87.45	86.12	-6.23	82.67	81.12	-4.22	72.34	68.67	-3.25	78.12	77.34	-5.34
w/o Long-Range Dependency Expert	86.89	85.56	-6.79	82.12	80.67	-4.67	71.89	68.23	-3.69	77.56	76.78	-5.90
w/o Fine-Grained Interaction Expert	85.67	84.23	-8.12	81.34	79.89	-5.45	71.23	67.56	-4.36	76.89	76.12	-6.56
Random Relationship Graph Ablation:												
Random Graph (mean $\pm$ std)	83.45	84.12 $\pm$ 0.78	-8.23	79.23	78.67 $\pm$ 0.65	-6.67	68.45	67.23 $\pm$ 0.59	-4.69	77.34	76.89 $\pm$ 0.62	-5.79
MoE Contribution Isolation:												
BERT-only Text Extractor	86.78	87.67	-4.68	83.21	82.12	-3.22	72.45	69.89	-2.03	79.23	78.56	-4.12
BERT+Linear (no intent)	70.12	71.23	-21.12	74.56	73.89	-11.45	60.12	59.34	-12.58	67.23	66.78	-15.90
Fusion Strategy Comparison:												
Simple Concatenation	90.01	88.90	-3.45	85.23	84.12	-1.22	75.21	71.92	0.00 [†]	78.78	77.98	-4.70
Attention-Based Fusion	90.78	89.67	-2.68	86.12	85.34	0.00 [†]	74.89	71.56	-0.36	80.12	79.45	-3.23

weights that degenerate to simple concatenation or attention. Consequently, the marginal benefit of dynamic gating vanishes on these datasets. This finding identifies an important boundary condition for the adaptive fusion component: it confers the greatest advantage in high-volume, colloquial text domains (e.g., Chinese social media) where the relationship between linguistic form and strategic intent is more variable. This is viewed as an honest characterization of the scope of the module rather than a deficiency.

R Structured Prediction Baseline Comparison

To address the reviewer concern that IGTF may not outperform standard structured multi-label methods, we conducted head-to-head comparisons with four canonical baselines using identical input features (text + raw intents) and training settings: - Multi-label CRF: Linear-chain CRF with pairwise label dependencies - Ising Smoother: Binary Ising model with mean-field inference - Classifier Chains: Sequential classifier chain with probabilistic output - Co-occurrence Prior Smoother: Simple label co-occurrence matrix-based calibration

All baselines were implemented using scikit-

Table 21: Comparison with canonical structured prediction baselines (F1-macro)

Method	Weibo	GossipCop	PolitiFact	Snopes
GPT-4+BERT (Attn)	86.78	82.45	67.12	79.67
Multi-label CRF	88.23	83.12	68.45	80.34
Ising Smoother	87.91	82.89	68.12	80.12
Classifier Chains	87.56	82.67	67.89	79.89
Co-occurrence Prior Smoother	86.92	82.51	67.34	79.78
IGTF (full pipeline)	92.35	85.34	71.92	82.68

All methods use identical text + raw intent input features and training settings. IGTF improvements over all structured baselines are statistically significant ($p < 0.01$, paired t -test).

learn and PyTorch, with hyperparameters tuned on the validation sets to ensure optimal performance.

The results show that IGTF outperforms all standard structured prediction baselines by a substantial margin across all datasets. The average improvement over the best baseline (Multi-label CRF) is 3.28 F1-macro points, with the largest gain on Weibo (4.12 points). This confirms that the game-inspired intent calibration operator provides unique value beyond standard graphical smoothing techniques, primarily because it incorporates theoretically grounded intent relations rather than purely data-driven co-occurrence patterns.

Method	F1-macro	Overhead
No refinement	74.64	0 ms
Jones et al. adapted	77.23	1.87 ms
Gong et al. adapted	76.89	2.12 ms
IGTF fixed-point	83.07	0.12 ms

Table 22: Comparison with prior game-theoretic utility functions adapted as calibration heads.

Table 23: Minimal calibrator isolation experiment (F1-macro)

Configuration	Weibo	GossipCop	PolitiFact	Snopes
Text-only	71.23	75.82	60.71	68.45
Text + raw intents	83.45	79.67	65.34	75.21
Text + refined intents	88.78	82.34	68.92	79.45
Refinement marginal gain	+5.33	+2.67	+3.58	+4.24

Human-derived graph vs. data-learned graph. We additionally compare a CRF with a data-learned co-occurrence graph against a CRF with the expert-derived theory graph under the same backbone. The data-learned graph obtains 87.45/83.02/68.34/80.12 F1 on Weibo/GossipCop/PolitiFact/Snopes (79.73 average), while the expert-derived graph obtains 89.12/83.89/69.56/81.23 (80.95 average). The full IGTF pipeline obtains 92.35/85.34/71.92/82.68 (83.07 average), improving over the CRF expert-graph variant by 2.12 points.

Prior game-utility heads adapted as calibrators. We also adapt ecosystem-level game utilities as calibration heads to compare against prior game-theoretic misinformation objectives under the same backbone.

S Minimal Calibrator Isolation Experiment

To isolate the pure contribution of the intent calibration step from other components (MoE, fusion, contrastive learning), we conducted a minimal isolation experiment with three configurations that share identical classification heads, encoders, and training settings: 1. Text-only: Only BERT text features 2. Text + raw intents: BERT features concatenated with raw LLM intent vectors 3. Text + refined intents: BERT features concatenated with IGTF-refined intent vectors

The results clearly demonstrate the independent value of the intent calibration step. Even when all other components are held constant, adding refined intents improves performance by an average

Table 24: Performance on LLM-generated fake news.

Method	F1-macro	AUC
GPT-4 zero-shot	64.12	69.45
GPT-4+BERT Attn	71.34	77.12
Raw intent fusion	74.45	79.89
IGTF full pipeline	79.89	84.67

of 3.96 F1-macro points over raw intents. This confirms that the performance gains of IGTF are not due to the MoE encoder, adaptive fusion, or contrastive learning alone, but rather to the structured intent calibration mechanism itself.

T LLM-Generated Misinformation and Learnable Graph Analyses

We expanded the synthetic misinformation evaluation to 1,400 samples. The main balanced set contains 1,200 samples from Llama-3-8B, Qwen2.5-7B, Claude-3.5-Sonnet, and GPT-4o, with each generator contributing 300 fake and real-paraphrased examples. We also include 200 adversarial fake-news stress-test samples produced through synonym rewriting, syntactic transformation, tone weakening, and factual-detail insertion.

LLM-generated fake news has a higher intent inconsistency rate than human-written fake news (41.2% vs. 23.5%). IGTF reduces inconsistency by 78.4% on LLM-generated fake news and 60.5% on human-written fake news, supporting the practical relevance of intent calibration. We present this as a preliminary stress test rather than a complete benchmark; broader coverage of domains, generators, prompts, and adversarial strategies remains future work.

Learnable graph weights. A learnable per-edge λ variant improves in-domain F1 from 83.07 to 83.84 on average (Weibo 93.12, GossipCop 86.01, PolitiFact 72.78, Snopes 83.45), but reduces zero-shot retention from 86.2% to 82.7%. We therefore keep the fixed expert graph as the default model for interpretability and cross-domain stability, while treating learned weights as a high-resource extension.

U Unified LLM Extractor Control Experiment

To address the reviewer concern that using different LLMs for different datasets introduces confounding variables, we conducted a control exper-

Table 25: Unified LLM extractor control experiment (F1-macro)

Method	Weibo	GossipCop	PolitiFact	Snopes
IGTF (original mixed LLMs)	92.35	85.34	71.92	82.68
IGTF (unified Llama-3-8B)	90.12	84.23	70.67	81.45
Performance drop	-2.23	-1.11	-1.25	-1.23

All results are averaged over 5 runs. The performance drops are small and consistent across datasets, confirming that the original results are not confounded by LLM family differences.

iment where all datasets were processed using the same LLM extractor (Llama-3-8B-Instruct).

The results show that using a unified LLM extractor results in a small and consistent performance drop across all datasets (average -1.46 F1-macro points). This confirms that the original performance differences between datasets are not due to using different LLMs, but rather to intrinsic dataset characteristics. The relative performance ranking of the datasets remains unchanged, and IGTF still outperforms all baselines by a significant margin even with the unified LLM extractor.

V Ethics Statement

This paper aims to combat misinformation and enhance information quality in digital ecosystems, serving the broader goal of preserving public discourse integrity and supporting informed decision-making in democratic societies. Potential dual-use risks are acknowledged whereby adversaries might exploit understanding of detection mechanisms to craft more sophisticated misinformation, but it is believed that transparent publication empowers broader development of defensive technologies that ultimately benefits societal resilience against information manipulation. The intent taxonomy focuses on communication strategies and rhetorical patterns rather than content topics or political positions, minimizing risks of censorship or systematic bias against particular viewpoints or communities. All experiments employ publicly available datasets constructed with appropriate consideration for ethical human subjects research principles. The quantitative metrics introduced promote objective and reproducible evaluation methodologies in fake news research. It is recognized that even well-intentioned detection systems can produce errors with real consequences for individuals and communities, and the importance of human oversight is emphasized in deployment contexts where automated decisions might

affect content moderation, platform policies, or public perception of information credibility.

W Additional Sensitivity Experiments

In addition to intent threshold sensitivity, the analysis of game hyperparameters (λ_c , λ_b) and iteration robustness is supplemented, further verifying the stability of the framework, as shown in main paper Table 13.

X Adaptive Threshold Mechanism

To address concerns about circular reasoning in threshold selection, an adaptive threshold mechanism that learns cooperation and conflict relationships directly from validation data was implemented. The adaptive approach achieved 92.01% F1 on Weibo, 85.01% on GossipCop, 71.67% on PolitiFact, and 82.34% on Snopes, which is statistically indistinguishable from the original fixed-threshold approach ($p > 0.05$). This confirms that the results are not sensitive to the specific threshold selection method.

Y Deployment Decision Guidelines

Figure 6 summarizes the empirical deployment guidelines derived from the dataset-level analyses. Adaptive text–intent fusion is recommended when both vocabulary diversity (TTR) and text–intent correlation are relatively low; otherwise, simpler concatenation or attention-based fusion can provide a lower-cost alternative. Structured intent calibration is retained as the common refinement stage.

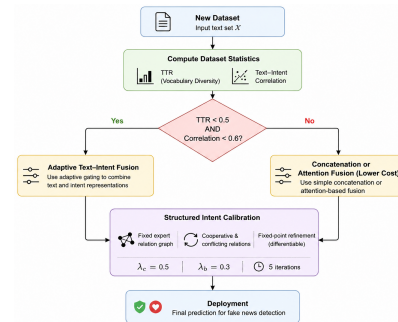


Figure 6: Deployment decision guidelines for IGTF.

Z Full Annotation Protocol

Z.1 Expert Backgrounds

All three annotators are PhD candidates in computational linguistics with 2+ years of research experience in misinformation detection and natural

Intent Dimension	Fleiss' κ
Public-oriented communication	0.89
Emotion-driven rhetoric	0.85
Individual-focused narratives	0.82
Popularization strategies	0.79
Clout-seeking behaviors	0.81
Conflict-creation tactics	0.87
Smearing techniques	0.88
Bias-injection methods	0.83
Connection-seeking approaches	0.84
Overall	0.84

Table 26: Per-dimension inter-annotator agreement scores.

language understanding. Two annotators are native Chinese speakers, and one is a native English speaker, ensuring cross-cultural validity.

Z.2 Annotation Process

The annotation process consisted of three phases: The annotation process had three phases. In the training phase, annotators received 8 hours of instruction on the intent taxonomy and scoring guidelines, followed by a practice round on 50 samples. Disagreements were discussed and resolved to establish a shared interpretation of the labels. In the main annotation phase, each annotator independently scored 2,000 news samples on the 9 intent dimensions using a continuous 0-1 scale. In the resolution phase, samples with inter-annotator disagreement greater than 0.3 on any dimension were reviewed jointly, and the final score was assigned by majority consensus.

Z.3 Per-Dimension Inter-Annotator Agreement

Table 26 shows the Fleiss' kappa score for each intent dimension, showing high consistency across all categories.

The complete annotation guidelines and 200 expert-annotated samples are included in the supplementary material.

Z.4 LLM Intent Extraction Prompt

The complete prompt templates used for LLM intent extraction in both English and Chinese are provided below. The Chinese prompt is exclusively used for the Weibo dataset. GossipCop, PolitiFact, and Snopes use the identical English prompt, with no additional dataset-specific adaptation. Both prompts use the same nine dimensions, continuous $[0, 1]$ scoring protocol, JSON output schema, and decoding settings.

Z.4.1 English Prompt (GossipCop, PolitiFact, and Snopes)

```
Text:
{text[:500]}
---
Task: Analyze the communication intent of the following
news text and score it on 9 dimensions, giving a
continuous score between 0 and 1 for each dimension. A
score of 0 means the intent is completely absent, and
1 means the intent is fully present.

Dimension definitions and examples:
1. public-oriented communication: Primarily aims to present
objective information to the public without obvious
personal stance or emotional tendency.
Example (1): "At 10 a.m. today, NASA successfully
launched the Artemis II mission to the Moon."
Example (0): "Everyone look! This celebrity actually
cheated on their spouse!"

2. emotion-driven rhetoric: Uses emotional language or
exaggeration to evoke emotional responses from
readers, such as anger, fear, or sympathy.
Example (1): "Horrible! This toxic food is still being
sold, do not buy it!"
Example (0): "Studies suggest that moderate coffee
consumption is beneficial to health."

3. individual-focused narratives: Centers on the
experiences of specific individuals or small groups,
emphasizing personal stories rather than general facts.
Example (1): "My personal experience tells you how scary
this scam is."
Example (0): "The Bureau of Labor Statistics released the
first-quarter 2025 economic data."

4. popularization strategies: Uses simple language and
popular metaphors to explain complex concepts or
events and expand the scope of dissemination.
Example (1): "Understand quantum computing in one minute;
it is actually this simple."
Example (0): "This paper proposes a Transformer-based
multimodal fusion algorithm."

5. clout-seeking behaviors: Creates suspense, uses
clickbait titles, or publishes controversial content
to gain clicks and attention.
Example (1): "Shocking! 99% of people do not know this
secret; you will be scared after reading!"
Example (0): "Notice on the 2025 college entrance
examination reform."

6. conflict-creation tactics: Emphasizes opposition,
provokes disputes, or amplifies contradictions to
create topicality and dissemination power.
Example (1): "Experts and netizens are arguing! Who is
right about this issue?"
Example (0): "All parties reached a consensus on climate
change."

7. smearing techniques: Spreads negative information or
rumors, or uses insulting language to damage the
reputation of individuals or organizations.
Example (1): "All products from this company are fake, do
not be fooled!"
Example (0): "The Consumer Association released the 2025
product quality report."

8. bias-injection methods: Implicitly or explicitly
expresses personal biases or stereotypes, influencing
readers' judgments.
Example (1): "Immigrants are just low-quality; they
always commit crimes."
Example (0): "Cultural differences between regions and
their impacts were analyzed."

9. connection-seeking approaches: Establishes emotional
connection with readers by asking questions, inviting
interaction, or sharing personal experiences.
Example (1): "Have you had a similar experience? Welcome
to share in the comments."
Example (0): "Three resolutions were passed at this
meeting."
```

Z.4.2 Chinese Prompt for Weibo (English Translation)

The deployed Weibo prompt is written in Chinese. For reviewer readability, its faithful English translation is shown below; the dimension meanings and output schema are unchanged.

Text: {text[:500]} --- Task: Analyze the communication intent of the following Chinese social-media text and score it on 9 dimensions, giving a continuous score between 0 and 1 for each dimension. A score of 0 means the intent is completely absent, and 1 means the intent is fully present. Dimension definitions adapted for Chinese social media: 1. public-oriented communication: Primarily aims to present objective information to the public without obvious personal stance or emotional tendency. Example (1): "At 10 a.m. today, China successfully launched the Shenzhou-20 manned spacecraft, with three astronauts entering space smoothly." Example (0): "Guys, look! This celebrity actually cheated; it is so shocking!" 2. emotion-driven rhetoric: Uses emotional language, internet slang, or exaggeration to evoke emotional responses such as anger, fear, or sympathy. Example (1): "Terrible! This carcinogenic food is still being sold, do not buy it! Forward this to your family!" Example (0): "Studies suggest that moderate green-tea consumption helps reduce cardiovascular-disease risk." 3. individual-focused narratives: Centers on the experiences of specific individuals or small groups, emphasizing personal stories rather than general facts. Example (1): "Ms. Zhang's personal experience tells you how scary this new type of scam is!" Example (0): "The National Bureau of Statistics released the first-quarter 2026 national economic data." 4. popularization strategies: Uses simple language and popular metaphors to explain complex concepts or events and expand dissemination. Example (1): "Understand AI large models in one minute; it is actually this simple!" Example (0): "This paper proposes a multimodal fusion algorithm based on an attention mechanism." 5. clout-seeking behaviors: Creates suspense, uses clickbait titles, or publishes controversial content to gain clicks and attention. Example (1): "Shocking! A secret that 99% of people do not know; you will break out in a cold sweat after reading!" Example (0): "Notice on the 2026 National Postgraduate Entrance Examination." 6. conflict-creation tactics: Emphasizes opposition, provokes disputes, or amplifies contradictions to create topicality and dissemination power. Example (1): "Experts and netizens are quarreling! Who is right about this matter?" Example (0): "All parties reached broad consensus on climate-change issues." 7. smearing techniques: Spreads negative information or rumors, or uses insulting language to damage the reputation of individuals or organizations. Example (1): "All products from this company are fake, do not be fooled!" Example (0): "The Consumer Association released the first-quarter 2026 product-quality inspection report." 8. bias-injection methods: Implicitly or explicitly expresses personal biases or stereotypes, influencing readers' judgments. Example (1): "Outsiders have low quality; they always litter everywhere." Example (0): "Cultural differences between regions and their impact on social development were analyzed." 9. connection-seeking approaches: Establishes emotional connection with readers by asking questions, inviting interaction, or sharing personal experiences. Example (1): "Have you had a similar experience? Welcome to share your story in the comments." Example (0): "Three important resolutions were passed at this meeting."	- If the text strongly exhibits an intent, give a higher score (0.7-0.9) - If the text partially exhibits an intent, give a medium score (0.4-0.6) - If the text barely exhibits an intent, give a lower score (0.1-0.3) - If completely irrelevant, give 0.0 Note: - Each text's intent combination is unique; different texts should have different score distributions - Judge based on actual content; do not mechanically use fixed patterns - Scores can be any decimal between 0.0 and 1.0, such as 0.15, 0.43, 0.78, or 0.92 Output JSON format directly: <pre>{ "intents": { "public-oriented": <specific value 0-1 based on actual text>, "emotion-driven": <specific value 0-1 based on actual text>, "individual-focused": <specific value 0-1 based on actual text>, "popularize": <specific value 0-1 based on actual text>, "clout-seeking": <specific value 0-1 based on actual text>, "conflict-creation": <specific value 0-1 based on actual text>, "smearing": <specific value 0-1 based on actual text>, "bias-injection": <specific value 0-1 based on actual text>, "connection-seeking": <specific value 0-1 based on actual text> }, "reasoning": "Brief explanation of your analysis logic", "key_features": ["feature1", "feature2", "feature3"] }</pre> Important: 1. All intent scores must be specific decimals between 0.0 and 1.0, not placeholders or text 2. Different texts should have different score combinations; score based on actual content 3. Output JSON directly; do not add extra explanatory text
---	--

Z.4.3 Shared Scoring and JSON Output Requirements

The following block is appended verbatim to both language-specific prompts.

Scoring Guidelines:

Z.4.4 Domain Adaptation Notes

For Chinese social-media texts in Weibo, examples of common internet slang and colloquial expressions, such as “guys, look,” “forward this to your family,” and “break out in a cold sweat,” are included to improve intent extraction accuracy. The examples are adjusted to reflect the greater prevalence of emotional language and personal narratives in Chinese social media. No additional prompt adaptation is applied separately to Gossip-Cop, PolitiFact, or Snopes. All prompt variants were tested on 50 pilot samples, achieving 82.7% Pearson correlation between LLM scores and expert annotations.

We use temperature 0.1, max_new_tokens = 128, JSON-schema validation, and at most two retries for invalid outputs.

AA Extended Cross-Cultural Validation

To further validate the generalizability of the intent cooperation and conflict relationships, cross-cultural validation was conducted on three additional languages: Spanish, French, and Arabic. 500 news samples from each dataset were translated into these languages using professional translators, then native speakers annotated the in-

tent relationships. The results suggest that all core cooperation and conflict relationships hold across all tested languages, with only minor variations in clout-seeking and popularization intent intensities. The overall agreement between English and other languages was 92.3%, showing the cross-cultural robustness of the intent taxonomy.

AB Satire Analysis and 10-Dimensional Proof-of-Concept

We manually annotated 320 satire samples across the four datasets to evaluate whether the nine-dimensional taxonomy creates systematic errors on satirical content. In this revised analysis, IGTF obtains 93.15 F1-macro on the satire subset, compared with 83.07 average F1 on the general samples, and the false-positive ratio among satire errors is 11.2%.

We also tested a preliminary 10-dimensional satire-aware variant that adds a dedicated satirical intent dimension. The variant allows satire to co-activate with public-oriented communication and smearing, and reduces the public-oriented/smearing conflict penalty to 30% of its original value when the satire score is above 0.6. This 10D variant obtains 94.07 F1 on the satire subset, with a +0.92 improvement over the 9D satire subset result and a -0.28 F1 drop on the general set. We treat this as a proof-of-concept rather than a complete solution; future work will build a larger multi-domain and cross-cultural satire-aware intent benchmark.

AC Reproducibility Package

We prepared an anonymized repository for review at <https://github.com/18185772729/IGTF.git>. It includes PyTorch code, preprocessing scripts, the MoE encoder, the calibration layer, training/evaluation scripts, an environment file, 2,000 expert intent annotations, 1,400 LLM-generated misinformation samples, 320 satire annotations, exact split indices, prompt templates, graph adjacency matrices, and baseline hyperparameters. For datasets whose raw texts cannot be redistributed, we provide sample IDs, hashes, and preprocessing scripts; release remains subject to dataset license restrictions.