ASIDE: Adaptive Self-Learning for Misinformation Detection via Dual Supervision

Anonymous ACL submission

Abstract

Detecting misinformation on high-volume social media platforms such as X and Facebook is challenging, exacerbated by the cost and inconsistency of human annotation. To address this, we propose a novel adaptive self-learning framework, ASIDE, that leverages active learning with limited labels and a dual teacher-student approach. Our framework uniquely employs a teacher text model for content and a hyper-teacher to capture conversational structure. During active learning, ASIDE dynamically transitions its data acquisition strategy based on its confidence, prioritizing uncertain but influential samples using a hybrid method and an uncertainty threshold. Simultaneously, it employs a dynamic sampling strategy to adapt to performance changes. Extensive experiments demonstrate that ASIDE significantly outperforms state-of-the-art methods, including active learning, graph-based, and unsupervised approaches, in misinformation detection.

1 Introduction

Social media platforms like Facebook, Weibo, and X have become major sources of information, but their lack of fact-checking has enabled the rapid spread of misinformation (Yan et al., 2019). This has led to serious consequences, including over 800 deaths from a false COVID-19 cure (Islam et al., 2020). According to Olan et al. (2024), distinguishing truth from falsehood might become increasingly elusive and challenging, raising concerns about manipulation and the erosion of informed discourse. Although misinformation detection has been investigated ubiquitously, current methods struggle to keep pace with false content and often rely on expensive manual labeling. Graph-based models and active learning offer potential (Farinneya et al., 2021; Cui et al., 2025; Ma et al., 2017; Liu et al., 2022; Hao et al., 2024), but still depend heavily on human annotation. To address this, we propose ASIDE, a novel framework that combines few-label supervision with active learning in a dual teacher–student setup. ASIDE adaptively selects samples using uncertainty, diversity, and influence, while minimizing annotation through teacher-driven pseudo-labeling. Our framework introduces several key contributions. First, to the best of our knowledge, a dual teacher-student setup comprising textual and hypergraph modules on the teacher and student sides has not been previously introduced. This novel architecture allows the textual teacher (e.g., GPT-2) and the hypergraph teacher to guide the corresponding dual student components collaboratively. This fusion enables a holistic learning process that would capture both semantic content and structural patterns within conversation threads. Second, the acquisition strategy dynamically shifts from uncertainty-based sampling to influence-based selection, allowing the model to adapt and prioritize high-impact samples as learning progresses. Third, after a brief warm-up with minimal human-labeled data, the system transitions to fully self-supervised learning via teacher-generated pseudo labels, removing the need for ongoing annotation. Finally, unlike fixed-batch active learning strategies, our method dynamically determines how many samples to acquire in each iteration based on model learning progress. This ensures the proposed framework remains responsive to performance changes during the learning process.

2 Related Work

2.1 Misinformation Detection Revolution

Misinformation detection has undergone a revolutionary stage, beginning with early text-based approaches such as sentiment analysis, bag-of-words, and user profiling (Castillo et al., 2011; Ma et al., 2015; Tolosi et al., 2016; Enayet and El-Beltagy, 2017; Kumar et al., 2022). While effective to some extent, these methods have struggled to capture the temporal and interactional dynamics of online discourse. Neural architectures addressed these gaps: RNNs modeled temporal sequences (Ma et al., 2017), CNNs extracted local textual features (Yu et al., 2020), and attention mechanisms improved semantic focus (Veyseh et al., 2019). To further advance detection, graph-based models were introduced to represent the structural and relational flow of misinformation (Zhou et al., 2020). Hypergraph approaches like CBT-HGCN capture higher-order tweet relationships (Li et al., 2023), while graph entropy techniques model propagation over time (Hao et al., 2024), and retweet-based graphs identify influential users (Liu et al., 2022).

2.2 Semi-Supervised Learning with Clustering and Active Learning

Semi-supervised learning offers a promising direction for misinformation detection, especially when labeled data is limited. Few-shot methods combine metric learning with pre-trained language models (Ran et al., 2023), and memory-augmented meta-learning has been shown to improve generalization (Mishra et al., 2017). Graph-based approaches like Positive-Unlabeled (PU) learning with GATs further reduce label dependency (Marques et al., 2023). Clustering is widely used to organize unlabeled data for public opinion analysis (Park et al., 2022), accident studies (Xu et al., 2022), and topic detection (Martínez et al., 2022), using algorithms like K-means (Zhang et al., 2023), hierarchical clustering (Murtagh and Contreras, 2011), DBSCAN, and GMMs (Shafi et al., 2024). It also improves few- and zero-shot learning (Alsuhaibani et al., 2024; Gretz et al., 2023; Freitas, 2024). Active learning helps prioritize informative samples (Hino, 2020; Barnabò et al., 2023; Zhang et al., 2015), with recent work like DISAL (Wan et al., 2024) focusing on influence-based selection. Our method uses cosine similarity instead. Combined strategies (Klein et al., 2021; Xiao et al., 2023) show promise, though multi-round AL for LLMs remains an open challenge (Margatina et al., 2023).

3 Our Novel ASIDE Model

Each component of ASIDE is s carefully integrated to address distinct challenges associated with misinformation detection in situations where the annotated data is scarce. The text encoder captures tweet semantics, while the hypergraph encoder models structural relationships in conversations. A teacher–student framework guides pseudo-labeling during active learning, with EMA ensuring stable teacher pseudo-labeling. The acquisition strategy begins with hybrid sampling during early uncertainty, then switches to DISAL to select high-influence samples via gradient-based scoring, enhancing performance as the model gains confidence.

3.1 Notations and Problem Description

We are given a dataset $\mathcal{D} = \{(T_i, y_i)\}_{i=1}^{N}$, where N is the total number of conversation threads. Each T_i represents a conversation thread on social media initiated by a root tweet (source of the conversation thread) r_i, and $y_i \in \{0, 1\}$ denotes the class label, with 1 representing a *rumor* and 0 a *non-rumor*. Each thread can be defined as:

$$T_i = \{r_i, c_i^1, c_i^2, \ldots, c_i^n\},$$

where each c_i^j denotes a reaction tweet replying to the root tweet (source of the conversation thread) r_i. The dataset is split into a small labeled subset $\mathcal{D}_l$ and a larger unlabeled subset $\mathcal{D}_u$, such that $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, reflecting a low-resource scenario.

Each thread T_i is structurally represented as a subgraph $G_i = (V_i, E_i)$, where V_i is the set of tweets and $E_i \subseteq V_i \times V_i$ encodes relationships (reactions to the source tweet) such as replies and retweets. We further model each G_i as a hypergraph within the collection $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$, incorporating both semantic and structural information. Each node (tweet) is associated with an initial embedding $h_i^{(0)}$ derived from its textual content, while the subgraph is characterized by an incidence matrix $\mathbf{H}_i$ and node features.

Given conversation thread T_i, the goal is to predict whether the T_i expresses a *rumor* or *non-rumor*, denoted by $y_i \in \{0, 1\}$. To address this task, we introduce a dual-teacher dual-student active learning framework that minimizes labeling cost by adaptively selecting the most informative unlabeled samples in each iteration (***dynamic sample selection instead of a fixed number decided via Dynamic Sampling Size Strategy***). The sample selection process follows one of two strategies depending on the model's confidence: a *hybrid strategy* that combines entropy-based uncertainty and representativeness via clustering, or a *DISAL (Dual*

Influence Self-Adaptive Learning) strategy, which leverages cosine similarity of gradient projections and KL divergence to identify high-influence samples.

This framework promotes effective learning by allowing the student model to learn from the most informative samples encountered during the active learning phase. Rather than relying on oracle assistance, the dual teacher will assign pseudo-labels to these samples, enabling the student model to be self-supervised. Detailed implementation and experimental protocols, including data split, sampling heuristics, and training schedules, are described in Section 4.

4 ASIDE Description

Figure 1 illustrates the proposed framework, which integrates several complementary strategies. We used a dual-teacher architecture comprising a text-based teacher and a hypergraph-based teacher. textbf A) Dual Teacher Initialization, a warm start phase fine-tunes two complementary teacher models, a text-based encoder and a hypergraph-based encoder on a small set of labeled data (see figure 1). This dual-teacher architecture captures both semantic and structural properties of conversation threads using the post content (FT-LLM) and hyperedge incidence matrix (HyperGCN-T). **B) Acquisition Strategies**: Two active learning strategies are incorporated to guide sample selection: the *Hybrid* strategy combines entropy-based uncertainty sampling with cluster-based representativeness when the teacher's confidence is low. Alternatively, the *DISAL* (Dual Influence Self-Adaptive Learning) strategy is used when the teacher's confidence exceeds a threshold. DISAL identifies high-influence samples based on gradient projection using cosine similarity and KL divergence between current and previous student predictions. **C) Active Learning Loop and EMA.** At each iteration, the student model is retrained using all pseudo-labeled samples accumulated thus far. The most informative unlabeled samples are dynamically selected based on uncertainty, influence, and labeling progress. The dual-teacher model assigns pseudo-labels without human supervision. To ensure stable learning, the teacher parameters are updated via an Exponential Moving Average (EMA) based on the student's learning trajectory. **(D) Final Inference.** After several iterations of active learning, the trained dual-student model is evaluated on unseen conversation threads. It combines both textual and structural features to classify each conversation as a rumor or a non-rumor. Unlike previous active learning approaches that rely solely on uncertainty or require manual labeling at every iteration, ASIDE introduces a novel dual-teacher mechanism to provide robust self-supervision, incorporates dynamic switching from uncertainty state (Hybrid) to influence state (DISAL) for influence-aware sample acquisition, and dynamically adjusts the sampling budget based on model progress all while eliminating the need for oracle involvement after the first iteration. To the best of our knowledge, this is the first work to combine text and hypergraph encoders in a dual-teacher dual-student setup under a fully automated active learning paradigm for misinformation detection. While our framework leverages some existing models, its novelty lies in the specific integration of their complementary strengths to address a research gap that none of the individual models can effectively tackle alone. By strategically combining these models, we create a synergistic effect, resulting in a novel approach with emergent capabilities and improved performance beyond the sum of its parts. This unique architecture and the resulting advancements constitute a significant contribution to the field.

ASIDE is built upon a dual teacher-student architecture, where both the teacher and the student models comprise two collaborative components: a text encoder and a hypergraph-based encoder. The teacher is initialized through a warm-start procedure using a small set of labeled examples (few labels), while the student is iteratively retrained within a dynamic active learning loop. The following steps elaborate core components and processes of the proposed framework:

Step 1: Teacher Warm Start. The teacher model is initially warm-started using a small set of labeled data $\mathcal{D}_l = \{(G_i, y_i)\}_{i=1}^{N_l}$, where each G_i represents a conversation thread encoded as a hypergraph substructure, and $y_i \in \{0, 1\}$ is the corresponding ground truth label (rumor or non-rumor). Each subgraph G_i consists of a sequence of tweets with both textual features and a hypergraph incidence matrix capturing relational structure. The dual-teacher model comprises a text-based encoder f_{text} and a hypergraph-based encoder f_{hyper}, and produces a final prediction by averaging their output logits:

$$\hat{y}_i = \frac{1}{2}\left(f_{\text{text}}(T_i) + f_{\text{hyper}}(H_i, \mathcal{I}_i)\right),$$

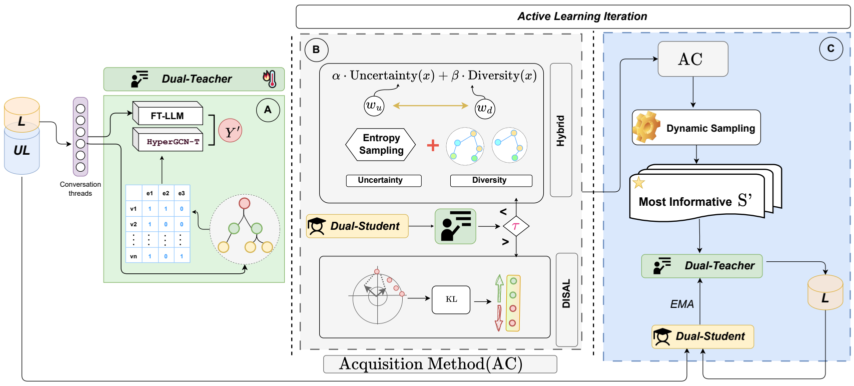


Figure 1: **ASIDE**: Adaptive Self-Learning for Misinformation Detection via Dual Supervision

where T_i is the conversation text, H_i denotes the initial node embeddings, and $\mathcal{I}_i$ is the hypergraph incidence matrix for subgraph G_i. This warm-start phase is essential and will allow the teacher model to capture both semantic and structural aspects of misinformation early in the learning process, serving as a strong and fundamental basis for generating high-quality pseudo-labels in subsequent active learning iterations. **Step 2: Iterative Student Training.** After the initial iteration, the framework departs from traditional oracle-based labeling and fully commits to a teacher-guided paradigm. Specifically, from the second active learning iteration onward, the teacher model, which is initially warmed up using a small subset of labeled data under the assumption that the student model is uncertain in the early stages of learning, generates ***pseudo labels*** for all newly acquired samples. These pseudo-labels serve as the sole source of supervision for the student model. Consequently, no additional Oracle annotations are required beyond the first iteration, which enhances the scalability and practicality of the framework in low-resource scenarios. The student model comprises two branches: a text encoder and a hypergraph encoder. During each iteration, the student is retrained using all examples labeled by the teacher up to that point. The loss function is a balanced combination of text and hypergraph classification losses:

$$\mathcal{L}_{\text{student}} = \frac{1}{2} \left(\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{hyper}} \right) \tag{1}$$

This joint formulation encourages consistent learning across both modalities while depending entirely on pseudo-labeled supervision.

Step 3: EMA Teacher Update. To progressively align the teacher with the improving student, we employ an exponential moving average (EMA) strategy over the student's weights. The teacher parameters θ_T are updated via:

$$\theta_T \leftarrow \alpha \cdot \theta_T + (1 - \alpha) \cdot \theta_S, \tag{2}$$

where α is the EMA coefficient (typically 0.99), and θ_S denotes the student's parameters.

Step 4: Sample Selection via DISAL or Hybrid. We place significant emphasis on the need to dynamically switch between uncertainty-based and influence-driven acquisition strategies. In other words, the framework is designed to be responsive to the evolving learning state of the model by adaptively selecting the most suitable acquisition method at each iteration.

At each iteration, the selection strategy is chosen dynamically:

- If the teacher's accuracy exceeds the uncertainty threshold (e.g., 0.90), we adopt **DISAL**.
- Otherwise, we use the **Hybrid** method based on uncertainty and representativeness.

(a) DISAL Acquisition. Each unlabeled sample is scored using a combination of gradient alignment and KL divergence. First, the gradient of the sample is projected against the accumulated gradient of labeled data (**labeled by the teacher dual**). Cosine similarity is used as the projection mode:

$$\text{Proj}(g_i, G) = \cos(g_i, G) = \frac{g_i^\top G}{\|g_i\| \cdot \|G\|}, \tag{3}$$

where g_i is the gradient of the current sample and G is the average gradient from labeled samples.

In addition, the KL divergence between the student's current prediction and its previous snapshot measures the influence of this sample:

$$\mathrm{KL}(p^{\text{new}} \| p^{\text{old}}) = \sum_k p_k^{\text{new}} \log \left(\frac{p_k^{\text{new}}}{p_k^{\text{old}}} \right). \quad (4)$$

The total DISAL score combines both:

$$s_i = \alpha \cdot \mathrm{Proj}(g_i, G) + \beta \cdot \mathrm{KL}(p^{\text{new}} \| p^{\text{old}}), \quad (5)$$

where α and β are balancing coefficients (default: 0.5).

(b) Hybrid Acquisition. When the teacher is not sufficiently confident, we revert to the hybrid strategy, combining student uncertainty and cluster-based representativeness:

$$s_i = \lambda \cdot u_i + (1 - \lambda) \cdot (1 - r_i), \quad (6)$$

where u_i is the normalized entropy of prediction, and r_i is the representativeness distance to its cluster center, with $\lambda = 0.6$ by default.

Step 5: Dynamic Sampling Size Strategy. Rather than querying a fixed number of samples at each iteration, our framework employs a dynamic strategy to determine the number S of newly selected data points. Specifically, we compute:

$$\mathbf{S} = \min(S_{\max}, \max(S_{\min}, \rho)), \quad (7)$$

where the score ρ is defined as:

$$\boldsymbol{\rho} = \alpha \cdot \bar{U} + \beta \cdot (1 - \mathcal{A}) + \gamma \cdot \left(1 - \frac{N_{\text{labeled}}}{N_{\text{total}}}\right), \quad (8)$$

Here, $\bar{U}$ denotes the average uncertainty or influence score depending on the acquisition method (Hybrid or DISAL), $\mathcal{A}$ is the student model's current validation accuracy, and the final term $\left(1 - \frac{N_{\text{labeled}}}{N_{\text{total}}}\right)$ reflects the proportion of remaining unlabeled data in the pool. Importantly, **our framework does not rely on an oracle**; instead, the dual-teacher model generates *pseudo labels* for the selected samples. Thus, the *labeling budget* in this context refers to the number of samples for which the teacher model is queried to provide pseudo-supervision. This setup enables scalability and makes the framework particularly well-suited for low-resource scenarios where manual annotation is expensive or infeasible.

5 Experiments and Results

In the following section, we evaluate the proposed ***ASIDE*** model against several state-of-the-art methods to assess its effectiveness in both non-rumor and rumor detection tasks. Subsequently, we test the proposed method with the baseline models, followed by a comparison against GNN structures and standard uncertainty methods.

5.1 Experimental Settings

Three datasets, including Twitter15, Twitter 16 (Ma et al., 2017), and Pheme (Zubiaga et al., 2016), were intensively used during the experimental setup of this work.

Dataset	Non-rumor	Rumor
Twitter 15	372	1086
Twitter 16	205	613
PHEME	1860	1860

Table 1: Dataset statistics showing the number of rumor and non-rumor conversation threads.

Teacher and student configurations including model types, hidden dimensions, number of layers, and learning rates are detailed in Section A.1. This section also specifies dataset-specific split ratios for training, validation, and testing, as well as active learning parameters, including dynamic sample bounds and the maximum number of iterations.

5.2 Comparative Models

ASIDE was evaluated against a diverse set of comparative models under consistent low-resource settings. These included large-scale LLMs (Llama-2-13B, GPT-2, GPT-Neo-1.3B, GPT-J-1.3B, BLOOM-1.7B, OLMo-7B), instruction-tuned variants (TinyLlama-1.1B-Chat, birkhauser/causal), and compact transformer models (DeBERTa-v3-large, Qwen1.5-0.5B, ALBERT-base). All models were fine-tuned using the same learning rate of 3.0×10^{-4}, as shown in Table 6.

To structure the comparison, we group existing rumor detection models into three main categories: (1) graph-based models (Bian et al., 2020; Ma et al., 2018), which exploit conversational structures to improve prediction; (2) active learning methods (Farinneya et al., 2021; Alalawi et al., 2025), which minimize labeling cost by querying high-utility samples; and (3) few-label and clustering-based approaches, such as IdoFew (Alsuhaibani et al., 2024) and $\mathrm{BERT}_{\mathrm{IT:Cluster}}$ (Shnarch et al., 2022), which enhance performance with minimal supervision by grouping similar examples early in training.

Our proposed framework integrates strengths from all three categories: it combines dual encoders for modeling text and structure, integrates adaptive active learning strategies (Hybrid and DISAL), and enhances few-label learning via pseudo-labeling, clustering, and optional paraphrasing. For fairness, all models were trained under the same labeling budget and learning rate (Table 8), with an equal number of labeled examples used during training and evaluation.

5.3 Discussion

Our proposed framework, ASIDE, demonstrates superior performance compared to existing state-of-the-art approaches. Table 2 indicaties the performance of ASIDE across Twitter15, Twitter16, and PHEME, where it consistently surpasses leading frameworks including active learning methods (FLAL, ATL), graph-based models (BiGCN (Bian et al., 2020), TDRvNN (Ma et al., 2018), ClaHi-GAT (Lin et al., 2021), RAGCL (Cui et al., 2025)), and unsupervised clustering-driven approaches (IdoFew (Alsuhaibani et al., 2024), $\text{BERT}_{IT\text{:Cluster}}$ (Shnarch et al., 2022)). ASIDE delivers a clear improvement in accuracy by at least 4.5% on Twitter15, 4.6% on Twitter16, and 2.2% on PHEME showing strong and consistent gains across datasets. Unlike earlier graph-based models like BiGCN and TDRvNN that rely on fixed tree structures, ASIDE takes a more flexible approach by using hypergraphs to capture the complex, multi-way nature of conversation threads. Furthermore, compared to transformer-based models like $\text{BERT}_{\text{Cluster}}$ and RAGCL, which utilize pretrained embeddings, ASIDE achieves significantly better performance across nearly all metrics, particularly excelling in Recall and F1-score for the rumor class.. These results underscore ASIDE's effectiveness in low-resource settings and its strong generalization to complex, real-world misinformation.

5.4 Uncertainty standard methods and Baselines comparison

Across all three benchmark datasets (Twitter15, Twitter16, and PHEME), ASIDE consistently outperforms existing baselines for misinformation detection (Tables 3, 4, 6). Compared to traditional uncertainty-based active learning strategies (e.g., Entropy, Least Confidence), it improves accuracy by up to 4.6% on Twitter16 and 3.9% on PHEME,

Table 2: Comparison of ASIDE with state-of-the-art methods on Twitter15, Twitter16, and PHEME datasets. NR = Non-Rumor, R = Rumor.

Method	ACC	NR-P	NR-R	NR-F1	R-P	R-R	R-F1
Twitter15							
BiGCN (Bian et al., 2020)	88.16	84.15	62.83	70.49	89.17	96.10	92.39
TDRvNN (Ma et al., 2018)	86.53	69.19	72.30	70.41	92.44	91.10	89.85
ClaHi-GAT (Lin et al., 2021)	74.90	50.20	13.40	21.60	76.10	96.40	85.10
IdoFew (Alsuhaibani et al., 2024)	74.90	56.20	13.40	21.60	76.10	96.40	85.10
$\text{BERT}_{\text{IT:Cluster}}$ (Shnarch et al., 2022)	74.40	51.50	19.50	27.70	76.80	93.80	84.50
FLAL (Alalawi et al., 2025)	80.20	81.10	19.00	27.70	74.00	30.10	48.50
RAGCL (Cui et al., 2025)	84.60	–	–	–	–	–	80.60
ASIDE	**93.10**	**90.40**	**80.18**	**85.00**	**93.91**	**97.29**	**95.57**
Twitter16							
BiGCN (Bian et al., 2020)	87.30	87.12	52.17	63.34	87.15	98.03	92.14
TDRvNN (Ma et al., 2018)	84.83	58.04	65.57	76.11	93.78	90.25	87.03
ClaHi-GAT (Lin et al., 2021)	62.90	22.00	16.20	18.60	72.70	70.50	71.90
IdoFew (Alsuhaibani et al., 2024)	62.90	22.00	16.20	18.60	72.70	79.50	75.90
$\text{BERT}_{\text{IT:Cluster}}$ (Shnarch et al., 2022)	53.70	32.20	68.40	43.80	81.10	48.60	60.80
FLAL (Alalawi et al., 2025)	82.70	83.00	93.00	88.60	76.50	37.50	48.80
RAGCL (Cui et al., 2025)	89.10	–	–	80.20	–	–	94.50
ASIDE	**93.60**	**83.30**	**88.90**	**80.20**	**96.80**	**95.00**	**95.90**
PHEME							
BiGCN (Bian et al., 2020)	80.37	79.14	84.18	81.40	83.09	77.78	80.21
TDRvNN (Ma et al., 2018)	70.43	67.43	81.17	73.42	77.78	59.09	65.75
ClaHi-GAT (Lin et al., 2021)	73.80	80.70	62.20	70.70	68.50	84.80	74.50
IdoFew (Alsuhaibani et al., 2024)	73.80	80.70	62.90	70.70	69.50	84.90	76.40
$\text{BERT}_{\text{IT:Cluster}}$ (Shnarch et al., 2022)	75.00	83.10	45.40	73.20	71.40	89.60	78.30
ATL (Farinneya et al., 2021)	–	–	–	**78.90**	–	–	**78.90**
FLAL (Alalawi et al., 2025)	79.40	83.30	76.70	80.10	81.70	71.90	76.70
RAGCL (Cui et al., 2025)	76.80	–	–	–	–	–	–
ASIDE	**83.80**	**86.80**	**80.10**	**83.30**	**81.20**	**87.60**	**84.30**

Table 3: Comparison of ASIDE with standard uncertainty sampling methods (Entropy, Least Confidence, Margin) across Twitter15, Twitter16, and PHEME datasets. NR = Non-Rumor, R = Rumor.

Dataset	Method	ACC	NR			R		
			Prec.	Rec.	F1	Prec.	Rec.	F1
Twitter15	Entropy	89.88	86.70	75.36	80.62	90.90	95.51	93.15
	Least Confidence	91.09	100.00	68.12	81.03	89.00	100.00	94.18
	Margin	87.04	87.76	62.32	72.81	86.89	96.63	91.49
	ASIDE	**93.10**	**90.40**	**80.18**	**85.00**	**93.91**	**97.29**	**95.57**
Twitter16	Entropy	83.81	93.39	44.93	60.78	82.24	98.88	89.80
	Least Confidence	87.85	82.35	65.22	75.00	87.76	96.63	91.98
	Margin	87.45	95.24	57.97	72.07	85.86	98.88	91.91
	ASIDE	**93.66**	**83.00**	**88.90**	**86.02**	**96.82**	**95.89**	**96.35**
PHEME	Entropy	81.29	41.61	67.88	51.54	78.90	83.15	80.92
	Least Confidence	82.80	83.59	81.43	82.45	82.05	84.15	83.07
	Margin	81.94	81.05	83.16	82.09	82.86	80.78	81.88
	ASIDE	**83.80**	**86.80**	**80.10**	**83.30**	**81.20**	**87.60**	**84.30**

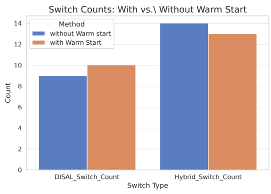
Table 4: Comparison of GCN, GAT, SAGE, and ASIDE across Twitter15, Twitter16, and PHEME datasets. NR = Non-Rumor, R = Rumor.

Dataset	Method	ACC	NR			R		
			Prec.	Rec.	F1	Prec.	Rec.	F1
Twitter15	GCN	73.31	0.00	8.20	0.00	75.30	100.00	85.91
	GAT	77.24	61.67	34.43	44.21	76.86	100.00	86.92
	SAGE	78.45	88.89	72.07	79.60	81.22	93.01	86.71
	ASIDE	**93.10**	**90.40**	**80.18**	**85.00**	**93.91**	**97.29**	**95.57**
Twitter16	GCN	73.31	50.00	6.56	11.94	75.12	97.39	85.47
	GAT	79.75	57.14	24.59	34.42	76.25	98.39	85.92
	SAGE	78.45	68.18	88.89	80.21	75.56	96.24	87.10
	ASIDE	**93.66**	**83.33**	**88.89**	**86.02**	**96.82**	**95.89**	**96.35**
PHEME	GCN	69.43	93.02	42.71	57.99	63.11	93.54	76.01
	GAT	74.84	80.10	61.04	71.88	71.27	84.29	77.24
	SAGE	68.28	92.28	38.97	54.82	96.82	61.96	73.56
	ASIDE	**83.80**	**86.80**	**80.10**	**83.30**	**81.20**	**87.60**	**84.30**

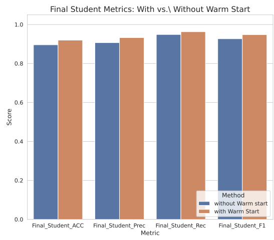
Method	Accuracy	Precision	Recall	F1-Score	DISAL / Hybrid
ASIDE w/o Warm Start	**0.896**	**0.907**	0.949	**0.927**	9 / **14**
ASIDE w/ Warm Start	0.920	0.933	**0.963**	0.948	**10** / 13

Table 5: ASIDE performance comparison with and without Warm Start (WS). Metrics include Accuracy, Precision, Recall, F1-Score, and switch counts.

Figure 2: Summary plots comparing with and without Warm Start (WS) for the proposed framework ASIDE.



(a) Switch Counts: With vs. Without Warm Start (WS)



(b) Final Student Metrics Comparison (WS)

Table 6: Comparison of base LLMs with ASIDE across three datasets. ASIDE consistently outperforms baselines in accuracy and F1-score.

Dataset	Base Model	Accuracy	Precision	Recall	F1
Twitter15	meta-llama/Llama-2-13b	71.20	71.78	69.18	70.25
	microsoft/deberta-v3-large	75.00	75.00	100.00	85.71
	Qwen/Qwen1.5-0.5B	70.29	67.02	72.73	69.78
	GPT-2	70.46	72.04	64.98	67.90
	birkha-user/causal	70.00	70.00	72.73	71.29
	TinyLlama-1.1B-Chat	73.20	73.90	68.83	70.78
	OPT-1.3B	70.30	70.97	68.13	69.32
	GPT-Neo-1.3B	70.60	73.20	66.88	69.89
	BLOOM-1.7B	70.69	72.28	68.13	70.14
	OLMo-7B	75.00	75.00	100.00	85.71
	ALBERT-base	70.03	71.08	67.10	69.03
	ASIDE	**93.10**	**93.91**	**97.29**	**95.58**
Twitter16	meta-llama/Llama-2-13b	72.61	73.48	70.29	71.75
	microsoft/deberta-v3-large	70.29	67.02	72.73	69.78
	Qwen/Qwen1.5-0.5B	71.71	69.80	73.91	71.80
	GPT-2	72.36	71.63	71.43	71.53
	birkha-user/causal	69.33	70.05	67.19	68.59
	TinyLlama-1.1B-Chat	72.61	73.48	70.29	71.75
	OPT-1.3B	70.29	67.02	72.73	69.78
	GPT-Neo-1.3B	70.60	73.20	66.88	69.89
	BLOOM-1.7B	70.69	72.28	68.13	70.14
	OLMo-7B	75.00	75.00	100.00	85.71
	ALBERT-base	70.03	71.08	67.10	69.03
	ASIDE	**93.66**	**96.82**	**95.89**	**96.35**
PHEME	meta-llama/Llama-2-13b	74.20	75.25	72.73	73.96
	microsoft/deberta-v3-large	69.78	0.00	0.00	0.00
	Qwen/Qwen1.5-0.5B	70.68	67.07	74.70	70.63
	GPT-2	70.47	73.81	71.43	72.60
	birkha-user/causal	70.00	70.59	72.73	71.64
	TinyLlama-1.1B-Chat	72.61	73.48	70.29	71.75
	OPT-1.3B	70.30	70.97	68.13	69.32
	GPT-Neo-1.3B	73.81	76.15	71.11	73.55
	BLOOM-1.7B	73.20	73.91	71.43	72.65
	OLMo-7B	75.00	75.00	100.00	85.71
	ALBERT-base	70.63	70.73	72.73	71.72
	ASIDE	**83.80**	**81.28**	**87.63**	**84.34**

Method	Accuracy	Precision	Recall	F1-Score	DISAL / Hybrid
ASIDE w/o EMA	0.877	0.873	**0.964**	0.916	0 / **23**
ASIDE w/ EMA	**0.920**	**0.933**	0.963	**0.948**	**10** / 13

Table 7: ASIDE performance comparison with and without Exponential Moving Average (EMA). Metrics include Accuracy, Precision, Recall, F1-Score, and switch counts.

highlighting the strength of combining hypergraph modeling, few-shot learning, and dual-teacher supervision. ASIDE also surpasses leading GNNs (GCN, GAT, SAGE) and LLMs (LLaMA, DeBERTa, GPT-2, OPT), achieving up to 20% higher F1 in some cases. Its strong performance on both rumor and non-rumor classes, particularly in recall and F1, underscores its ability to handle class imbalance and generalize effectively with limited supervision—setting a new benchmark in misinformation detection.

Overall, the integration of a hypergraph structure, a dual teacher-student paradigm, and a custom active learning strategy with few-label supervision has empowered the proposed ASIDE model to outperform state-of-the-art methods across multiple domains, including graph-based learning, active learning, and unsupervised approaches. This synergy enables the model to effectively capture complex conversational structures, leverage minimal annotations, and generalize well across diverse misinformation detection scenarios.

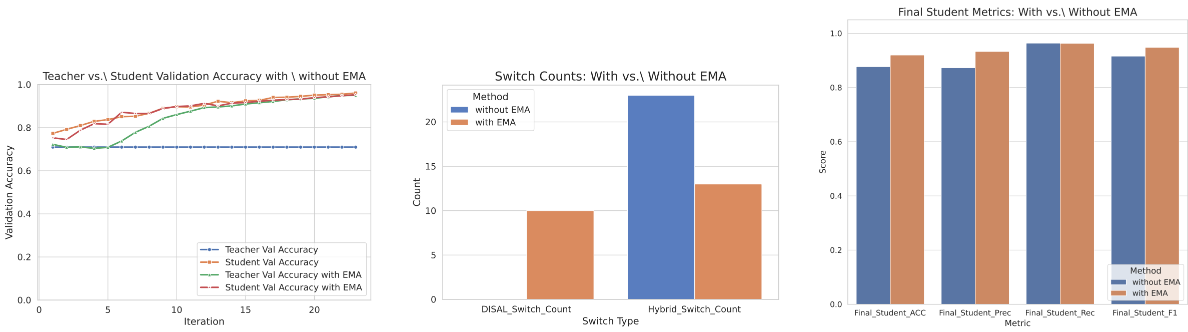
6 Ablation Study

To investigate the contributions of individual components, experiments have been conducted on the Twitter15 dataset to evaluate the impact of different components in our proposed framework. The dataset was partitioned with 5% used as few labeled data, simulating a low-resource setting, 70% as unlabeled data for active learning, and the remaining 25% is reserved for final testing.

6.1 Impact of EMA and Acquisition Switching

As part of our ablation study, we assess the individual contributions of Exponential Moving Average (EMA) and dynamic switching between acquisition strategies (i.e., hybrid and DISAL) to the overall performance of our proposed framework. Demonstrating its robustness and effectiveness under low-resource constraints. To further investigate the role of EMA and acquisition switching, Figure 7 presents a detailed comparison of three key perspectives. Figure 7. (a) shows the evolution of teacher and student accuracy over active learning iterations with and without EMA. EMA leads to a smoother and more stable learning curve, especially for the teacher model, allowing it to provide

Figure 3: Summary plots comparing with and without EMA for the proposed framework.



(a) Teacher vs. Student Accuracy (b) Switch Counts: EMA vs. No EMA (c) Final Student Metrics

more reliable pseudo-labels throughout the process. Figure 7(b) visualizes the number of times the framework dynamically switches between hybrid and DISAL acquisition strategies. In addition, the presence of a dynamic switching mechanism would ensure that the framework adapts its querying strategy based on model confidence and performance, thereby avoiding over-reliance on a single sampling approach. Figure 7. (c) highlights the final student metrics concerning accuracy, precision, recall, and F1 score. Moreover, it appears that the utilization of EMA during the active learning process leads the mode with EMA consistently and remarkably would outperform their counterparts without EMA, affirming its effectiveness in stabilizing learning and improving generalization. Thus, it can be confirmed that leveraging both EMA and dynamically switching between acquisition methods in correlation with model performance during the training phase has boosted the proposed methodology's overall performance.

6.2 Using zero-shot with teacher and student dual

To evaluate the effect of Warm Start, we compare ASIDE with and without teacher initialization using a brief supervised phase. As shown in Table 5, Warm Start leads to higher accuracy (0.920 vs. 0.896), precision (0.933 vs. 0.907), and F1-score (0.948 vs. 0.927), while recall remains comparable. This highlights the benefit of initializing the teacher with a few labeled examples for stable pseudo-labeling. The table also shows that Warm Start results in more DISAL acquisitions (10 vs. 9) and slightly fewer hybrid switches (13 vs. 14), indicating a more confident transition toward gradient-based sampling. These findings confirm that Warm Start improves both model performance and acquisition dynamics in the active learning process.

7 Conclusion

We have developed a novel ASIDE model that leverages a few labels, active learning, teacher-student paradigms, and a hyper-graph to effectively detect misinformation while simultaneously reducing the need for annotators' involvement in the model learning and training phase via combining hybrid and influence-driven acquisition strategies to satisfy the labeling requirements. Instead of relying on an oracle to aid the model in figuring out the most informative and elusive data samples, ASIDE leverages teacher-dual pseudo-labeling participation to guide the student dual during the training process. We have considered the quality of the generated pseudo labels by applying the EMA method to invoke the teacher dual to match the aggressive learning progress of the student dual during the active learning phase and not fall behind, which eventually would hinder the learning progress of the proposed framework. ASIDE has outperformed the state-of-the-art methods.

8 Limitations

While ASIDE has demonstrated strong performance on various English-language datasets, its effectiveness in other language contexts remains untested. In addition, temporal data should not be underestimated, as it may offer further advantages in effectively modeling the propagation of misinformation on social media platforms. Also, one of the limitations is a lack of interpretability. Therefore, integrating an explainable layer would help make the model's decisions more transparent and accessible to human users.

8.1 References

References

Abdulrahman Alalawi, Abdullah Alsuhaibani, Usman Naseem, Basem Suleiman, Shoaib Jameel, and Imran Razzak. 2025. Few labels with active learning: From weak to strong labels for misinformation detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2592–2596.

Abdullah Alsuhaibani, Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2024. Idofew: Intermediate training using dual-clustering in language models for few labels text classification. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 18–27.

Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Chaoqun Cui, Siyuan Li, Kunkun Ma, and Caiyan Jia. 2025. Enhancing rumor detection methods with propagation structure infused language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7165–7179.

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 470–474.

Parsa Farinneya, Mohammad Mahdi Abdollah Pour, Sardar Hamidian, and Mona Diab. 2021. Active learning for rumor identification on social media. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4556–4565.

Lucas José Gonçalves Freitas. 2024. Text clustering applied to unbalanced data in legal contexts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 639–642.

Shai Gretz, Alon Halfon, Ilya Shnayderman, Orith Toledo-Ronen, Artem Spector, Lena Dankin, Yannis Katsis, Ofir Arviv, Yoav Katz, Noam Slonim, and 1 others. 2023. Zero-shot topical text classification with llms-an experimental study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9647–9676.

Xiaorong Hao, Bo Liu, Xinyan Yang, Xiangguo Sun, Qing Meng, and Jiuxin Cao. 2024. Multi-stage dynamic disinformation detection with graph entropy guidance. *World Wide Web*, 27(2):8.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, AH Mostofa Kamal, and SM Hasan. 2020. M, kabir, a. *Yeasmin, D., Islam, M. A., Amin Chowdhury, KI, Anwar, KS, Chughtai, A., Seale, H*, pages 1621–1629.

Jan Klein, Sandjai Bhulai, Mark Hoogendoorn, and Rob Van der Mei. 2021. Plusmine: Dynamic active learning with semi-supervised learning for automatic classification. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 146–153.

Akshi Kumar, MPS Bhatia, and Saurabh Raj Sangwan. 2022. Rumour detection using deep learning and filter-wrapper feature selection in benchmark twitter dataset. *Multimedia Tools and Applications*, 81(24):34615–34632.

Fangfang Li, Zhi Liu, Junwen Duan, Xingliang Mao, Heyuan Shi, and Shichao Zhang. 2023. Exploiting conversation-branch-tweet hypergraph structure to detect misinformation on social media. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–20.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on twitter with claim-guided hierarchical graph attention networks. *arXiv preprint arXiv:2110.04522*.

Bo Liu, Xiangguo Sun, Qing Meng, Xinyan Yang, Yang Lee, Jiuxin Cao, Junzhou Luo, and Roy Ka-Wei Lee. 2022. Nowhere to hide: Online rumor detection based on retweeting graph neural networks. *IEEE transactions on neural networks and learning systems*.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. *arXiv preprint arXiv:2305.14264*.

Henrique O Marques, Lorne Swersky, Jörg Sander, Ricardo JGB Campello, and Arthur Zimek. 2023. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. *Data Mining and Knowledge Discovery*, 37(4):1473–1517.

José Manuel Guaita Martínez, Patricia Carracedo, Dolores Gorgues Comas, and Carlos H Siemens. 2022. An analysis of the blockchain and covid-19 research landscape using a bibliometric study. *Sustainable Technology and Entrepreneurship*, 1(1):100006.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Fionn Murtagh and Pedro Contreras. 2011. Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

June Young Park, Evan Mistur, Donghwan Kim, Yunjeong Mo, and Richard Hoefer. 2022. Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of covid-19 policy in transportation hubs. *Sustainable Cities and Society*, 76:103524.

Hongyan Ran, Caiyan Jia, and Jian Yu. 2023. A metric-learning method for few-shot cross-event rumor detection. *Neurocomputing*, 533:72–85.

Imran Shafi, Mahnoor Chaudhry, Elizabeth Caro Montero, Eduardo Silva Alvarado, Isabe De La Torre Diez, Md Abdus Samad, Imran Ashraf, and 1 others. 2024. A review of approaches for rapid data clustering: Challenges, opportunities and future directions. *IEEE Access*.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. *arXiv preprint arXiv:2203.10581*.

Laura Tolosi, Andrey Tagarev, and Georgi Georgiev. 2016. An analysis of event-agnostic features for rumour classification in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 151–158.

Amir Pouran Ben Veyseh, My T Thai, Thien Huu Nguyen, and Dejing Dou. 2019. Rumor detection in social networks via deep contextual modeling. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 113–120.

Tianjiao Wan, Kele Xu, Long Lan, Zijian Gao, Dawei Feng, Bo Ding, and Huaimin Wang. 2024. Tracing training progress: Dynamic influence based selection for active learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9417–9425.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. Freeal: Towards human-free active learning in the era of large language models. *arXiv preprint arXiv:2311.15614*.

Hui Xu, Yi Liu, Chi-Min Shu, Mingqi Bai, Mailidan Motalifu, Zhongxu He, Shuncheng Wu, Penggang Zhou, and Bing Li. 2022. Cause analysis of hot work accidents based on text mining and deep learning. *Journal of loss prevention in the process industries*, 76:104747.

Ruidong Yan, Yi Li, Weili Wu, Deying Li, and Yongcai Wang. 2019. Rumor blocking through online link deletion on social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–26.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations.

Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller. 2015. Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 275–278.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. *arXiv preprint arXiv:2305.14871*.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*.

A Appendix

A.1 Hyperparameter Settings

To ensure reproducibility and clarity regarding the diverse settings across our dual-model architecture and dataset-specific variations, we have tabulated all the parameter values in Table 8, which provides a comprehensive summary of all hyperparameter values and experimental configurations used across our proposed framework. All hyperparameters were selected based on the conducted experiments. For both the text-based and hypergraph-based encoders, we set the hidden dimension to 768 and the number of layers to 2 for consistency and efficiency. The teacher model was warm-started for 25 epochs using a learning rate of 3×10^{-4}, with a learning rate scheduler controlled by `step_size` and `gamma` values. Similarly, the student model was retrained for 7 epochs per iteration using the same base learning rate and a slightly more aggressive decay ($\gamma = 0.9$). We applied dataset-specific train/validation/test splits for Twitter15, Twitter16, and PHEME, taking care to balance class distributions and label scarcity in each case. The teacher-student interaction included EMA updates ($\alpha = 0.90$) for both Twitter 15 and 16, whereas ($\alpha = 0.88$) for Pheme due to the variation between datasets' size. The switching threshold was applied to transition between acquisition strategies (Hybrid and DISAL) when the teacher's accuracy exceeded 0.90. DISAL parameters ($\alpha = 0.5$, $\beta = 0.5$) were tuned to balance gradient projection and KL divergence contributions. For dynamic sample selection, the number of queried samples per iteration ranged from 10 to 50, depending on the unlabeled pool size and learning progress. All experiments were run with cosine similarity as the projection mode to ensure stability in influence estimation.

A.2 ASIDE Pseudo-Labels quality

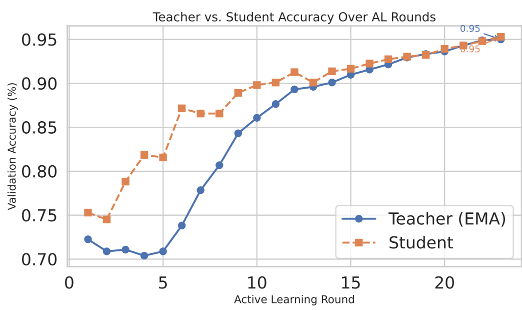
ASIDE maintains pseudo-label quality from the start by relying on a teacher model stabilized through Exponential Moving Average (EMA). After a brief warm start using a few ground-truth labels, which are being deployed to avoid noisy pseudo labels that might degrade the model's performance, the EMA teacher's performance enhances steadily, reaching 95 % accuracy during validation phase and 92.1 % in testing phase and 0.8902 macro F1, as shown in Table 9.

Table 8: Summary of hyperparameters and settings used across datasets and modules.

Parameter	Value
Teacher Settings	
Text teacher model	gpt2
Hypergraph teacher model	bert-base-uncased
Teacher hidden dim	768
Teacher HGC layers	2
Warm-up teacher epochs	25
Teacher learning rate	3×10^{-4}
Teacher LR step size	2
Teacher LR gamma	0.5
Student Settings	
Text student model	gpt2
Hypergraph student model	bert-base-uncased
Student hidden dim	768
Student HGC layers	2
Epochs per iteration	7
Student learning rate	3×10^{-4}
Student LR step size	1
Student LR gamma	0.9
Dataset Split Ratios	
Train ratio	Tw15 = 0.10, Tw16 = 0.10, PHEME = 0.15
Validation ratio	Tw15 = 0.60, Tw16 = 0.65, PHEME = 0.60
Test ratio	Tw15 = 0.30, Tw16 = 0.25, PHEME = 0.25
Framework Behavior	
Dynamic consistency enabled	True
Dynamic α	0.1
DISAL α	0.5
DISAL β	0.5
Gradient label mode	teacher
Teacher accuracy threshold	0.9
EMA enabled	True
EMA α	Tw15 & 16 = 0.90, PHEME = 0.88
Projection mode	cosine
Active Learning	
Max iterations	Tw15 & 16 = 23, PHEME = 70
Dynamic min samples	10
Dynamic max samples	50

Table 9: Ablation results and teacher-student trend on Twitter15. EMA enhances both score and learning stability.

Metric	With EMA	Without EMA	Zero-shot
Accuracy	92.1	87.7	89.6
Macro F1	89.0	84.1	87.2
Rumor F1	94.8	91.7	92.8
Non-Rumor F1	83.2	76.4	81.6
Recall (Macro)	87.7	81.7	86.0
Support	366 samples		



Teacher vs. Student validation accuracy over rounds.