LLM-based Rumor Detection via Influence Guided Sample Selection and Game-based Perspective Analysis

Zhiliang Tian[1,2†], Jingyuan Huang[1,2†], Zejiang He[1,2], Zhen Huang[1,2*], Menglong Lu[1,3], Linbo Qiao[1,2], Songzhu Mei[1,2], Yijie Wang[1,2*], Dongsheng Li[1,2]

[1]College of Computer Science and Technology, National University of Defense Technology, China
[2]National Key Laboratory of Parallel and Distributed Computing
[3]Key Laboratory of Advanced Microprocessor Chips and Systems

{tianzhiliang, jingyuanhuang, hezejiang, huangzhen, lumenglong}@nudt.edu.cn

Abstract

Rumor detection on social media has become an emerging topic. Traditional deep learning-based methods model rumors based on content, propagation structure, or user behavior, but these approaches are constrained by limited modeling capacity and insufficient training corpora. Recent studies have explored using LLMs for rumor detection through supervised fine-tuning (SFT), but face two issues: 1) unreliable samples sometimes mislead the model learning; 2) the model only learns the most salient input-output mapping and skips in-depth analyses of the rumored content for convenience. To address these issues, we propose an SFT-based LLM rumor detection model with **I**nfluence guided **S**ample selection and **G**ame-based multi-perspective **A**nalysis (**ISGA**). Specifically, we first introduce the Influence Score (IS) to assess the impact of samples on model predictions and select samples for SFT. We also approximate IS via Taylor expansion to reduce computational complexity. Next, we use LLMs to generate in-depth analyses of news content from multiple perspectives and model their collaborative process for prediction as a cooperative game. Then we utilize the Shapley value to quantify the contribution of each perspective for selecting informative perspective analyses. Experiments show that ISGA excels existing SOTA on three datasets.

1 Introduction

The wide and fast spread of rumors online has posed real-world threats in critical domains like politics (Hartley and Vu, 2020), and economy (CHEQ, 2023). Rumor detection aims to automatically detect inaccurate and intentionally misleading news items, and deep learning-based rumor detection has become the mainstream of this task.

Traditional deep learning-based rumor detection methods can be divided into three categories. The first is the content-based method, which typically employs the deep-learning model to model the textual content of rumors (Yu et al., 2017; Ma et al., 2018a). This method is easy to implement but suffers from unreliable content. The second is the propagation structure-based method, which captures the presence of temporal shifts (Ma et al., 2018b; Wu et al., 2020) and spatial structures (Bian et al., 2020; Matheven and Kumar, 2022) in the news propagation process to avoid being misled by unreliable contents. However, the propagation structure is usually simple and disorganized, thus sometimes failing to carry sufficient information for rumor detection. The third is the user-based method, which models the users who participated in rumor propagation since users with similar attributes (Huang et al., 2022) and historical behavior (Gao et al., 2022) tend to perform similarly in the rumor propagation process. Given the prevalence of malicious users, such as social bots, in the rumor propagation process, the inaccurate modeling of user relationships limits the effectiveness. Though the above methods achieved desirable performance, their ability to model the required features (i.e. content, propagation structure, or user behavior) is also limited by the model capability and amount of training data.

To increase the model capability, some researchers use large language models (LLMs) (Touvron et al., 2023a) for rumor detections. LLMs with many parameters, trained on extensive corpora, achieved remarkable capabilities (Wei et al., 2023). Some work (Chen et al., 2023; Wan et al., 2024) prompts LLMs to detect the rumors with the task-specific instruction and in-context learning (ICL) examples. However, on the one hand, given the prompt length limitations, only a few ICL examples can be provided. The limited quality and quantity of ICL examples and their patterns weaken the ability to detect rumors. On the other hand, it is hard to acquire task-specific knowledge without

[†] contributed equally to this work
[*] corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28402–28414

supervised fine-tuning (SFT) in the given scenario.

Hence, some works (Yang et al., 2024; Wan et al., 2024) apply SFT to LLMs on specific scenarios to detect the rumors. LLMs acquire task-specific knowledge via SFT and thus are quite sensitive to the quality of SFT samples. The sample quality in rumor detections has high variance since the samples are usually collected from social media, where the text is colloquial and informal. Therefore, SFT for LLMs in rumor detection faces two issues: (1) The unreliable training samples sometimes mislead the model learning, thus ensuring the quality of SFT samples is crucial. (2) The current SFT guides the LLMs in learning input-output mapping to accomplish the tasks. It results in the fine-tuned LLMs capturing only the most salient and shallow characteristics from the limited SFT samples, learning that end-to-end mapping tends to skip the in-depth analyses of the rumored content and ignore the specific reasons behind making the final decision. Lacking the ability to analyze the in-depth reasons results in failing to correctly detect the cases unseen during the SFT.

In this paper, we propose an LLM-based rumor detection method via SFT, which selects samples via the influence of samples and conducts multi-perspective analyses via a multiplayer game. Our model aims to select reliable samples for SFT and exploit in-depth analyses to avoid SFT capturing only the salient and shallow characteristics in detection. Firstly, we introduce the Influence Score (IS) to assess the impact of training samples on model predictions and select high-quality samples for SFT. As directly calculating IS is complex and time-consuming, we propose applying Taylor expansion with a chain rule to approximate it. Secondly, we use LLMs to generate in-depth analyses of news content from multiple perspectives. Considering conflict and interference between different perspective analyses, we model the collaborative process of these perspective analyses for predicting the news label as a cooperative game and each perspective as a "player". We further introduce the PVI as the value function in the cooperative game and utilize the Shapley value to quantify the contribution of these perspectives to the game's outcome. Experiments on three public datasets show that our model excels existing baselines in rumor detection tasks and achieves SOTA results.

Our contributions are four-fold: (1) We propose ISGA, a novel LLM-based rumor detection model that first considers enhancing the LLMs with SFT for rumor detection from both sample-level and analysis-level. (2) We propose an influence-based sample selection method to deal with the negative impact of unreliable samples on SFT. (3) We propose a shapely value-based multi-perspective analysis selection approach to relieve the challenge that the SFT-based LLM lacks in-depth analysis of rumored content. (4) Experiments on three datasets show that our model achieves SOTA performance.

2 Related Work

2.1 LLMs in Rumor Detection

Recently, Large Language Models (LLMs) have demonstrated remarkable performance in various text classification and reasoning tasks (Chang et al., 2024; Achiam et al., 2023; Ouyang et al., 2022; Touvron et al., 2023b). Existing studies have proven that LLMs possess significant potential in rumor detection (Hu et al., 2024; Wang et al., 2024; Luo et al., 2024). Some works explore LLMs as rumor detectors (Chen and Shu, 2023; Pelrine et al., 2023), and use their commonsense reasoning to provide supplementary explanations for rumor detection. (Cheung and Lam, 2023) uses LoRA tuning to train a LLama-based detector with external retrieved knowledge. However, these methods focus on detecting rumors but fail to perform in-depth analysis of rumored content.

On a related front, recent investigations have found that LLMs can act as high-quality data generators (Su et al., 2023; Luo et al., 2024; Lucas et al., 2023a; Pan et al., 2023; Huang et al., 2023; Wang et al., 2024; Hu et al., 2024). (Luo et al., 2024) utilizes LLMs to generate and inject malicious messages into social media message propagation trees. (Lucas et al., 2023b) uses LLMs to generate authentic and deceptive content to create a comprehensive dataset for rumor detection. (Huang et al., 2023) generates training data by using an LLM to replace a salient sentence in authentic news articles with a plausible but fake piece of information. (Wang et al., 2024) proposes a defense-based explainable rumor detection framework leveraging LLMs to generate justifications. In this work, we use LLM as a rumor detector and enhance its ability to learn task-specific knowledge via SFT. Considering the contextual understanding and text generation capabilities of LLMs, we also regard them as analysis generators and conduct in-depth multiple-perspective analyses of news content.

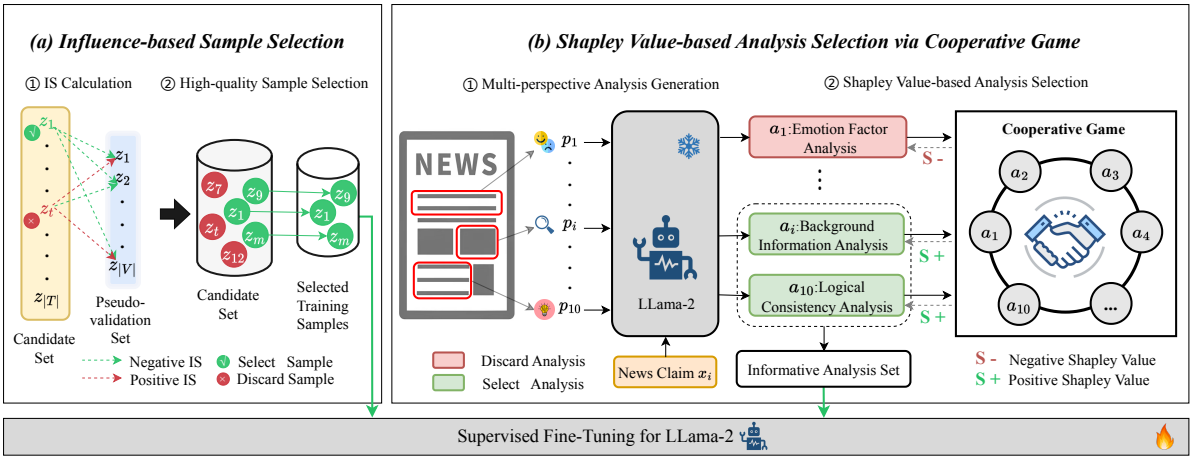


Figure 1: The overview of our framework. The process begins with the left part, which evaluates samples using IS and selects samples. Next, the right part generates multi-perspective analyses and models them as players in a cooperative game. It then quantifies the contribution of perspective analysis using Shapley values and selects informative analysis. Finally, the selected samples and analyses are utilized to conduct SFT on the LLM.

2.2 Influence-based Data Selection

The application of influence functions for data selection in natural language processing has been widely studied. (Koh and Liang, 2017) estimate the marginal effect of a training example on the loss of a validation example by influence functions, which assess the impact of data points on models. (Pruthi et al., 2020) proposes a gradient-based method, TracIn, to estimate influence by tracking the change in loss on a validation example during training, using the dot product of gradients at each training step. (Akyürek et al., 2022) shows gradient info can trace knowledge back to training data, but practical uses are still being explored. (Lam et al., 2022) found vanilla influence functions insufficient for optimal retrieval performance when identifying erroneous training data with synthetic noise. (Schioppa et al., 2022) and (Han et al., 2020) proposed methods for scaling influence functions and explaining the predictions of NLP models. Recently, (Xia et al., 2024) selects influential data for task-specific LLM fine-tuning by adapting influence formulations to work with the Adam optimizer and variable-length instruction data. (Pan et al., 2024) utilizes the influence function to quantify the impact of individual training examples on the model during training and selects supervised fine-tuning (SFT) data by analyzing these impacts. In our work, we select representative and reliable SFT training samples by introducing the influence score to measure the sample-level impact on the validation samples and performing a Taylor expansion to simplify calculations.

3 Methods

3.1 Overview

Our model consists of three modules, as shown in Fig. 1. (1) The influence-based sample selection module (§ 3.2) utilizes the influence score (IS) to quantify the impact of the individual SFT sample on model prediction to select representative and reliable training samples. (2) The estimating IS via Taylor expansion module (§ 3.3) approximates the IS via Taylor expansion with a chain rule to reduce the computational complexity. (3) The Shapley value-based analysis selection via cooperative game module (§3.4) conducts the multi-perspective analysis of the news as a cooperative game and utilizes the Shapley value as a criterion to select informative analysis. After obtaining the training samples and analysis, we use them to conduct SFT on LLM for rumor detection, and the SFT details are listed in the App. A.

3.2 Influence-based Sample Selection

To better select representative and reliable samples for supervised fine-tuning (SFT), we use the influence score (IS) to measure the impact of the samples on the model prediction and select the samples for SFT based on IS.

Existing rumor detection overlooks the quality of SFT training samples, while unreliable training samples sometimes mislead the model learning. Thus ensuring the quality of SFT training samples is crucial. Selecting high-quality samples for SFT involves two steps as § 3.2.1 and § 3.2.2.

3.2.1 Influence Score (IS) Calculation

We propose an influence score (IS) to help select representative and reliable samples, which quantifies the impact of the training sample on the model prediction. For training samples $\{z_1, z_2, ..., z_n\}$, as shown in Eq. 2, the optimization objective of the rumor detection model is to find an optimal model weight θ^* that minimizes the loss $R(\theta)$ on the training set:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(z_i, \theta) \tag{1}$$

$$\theta^* := \arg\min_{\theta} R(\theta) \tag{2}$$

where θ denotes the model weight and $\mathcal{L}(z_i, \theta)$ is the cross-entropy loss. When we perturb a training sample z_m (i.e. give a factor ϵ to perturb the loss $\mathcal{L}(z_m, \theta)$), the optimized model weight θ^* after perturbing changes accordingly as shown in Eq. 4:

$$R(\theta_\epsilon) = R(\theta) + \epsilon\mathcal{L}(z_m, \theta) \tag{3}$$

$$\theta^*_\epsilon = \arg\min_{\theta}[R(\theta) + \epsilon\mathcal{L}(z_m, \theta)] \tag{4}$$

The impact of training sample z_m for model prediction is evaluated on the validation set. That is, the training sample z_m should aid the model in accurately predicting the label of the validation sample z_v. The influence score $\mathcal{I}_{\theta^*}(z_m, z_v)$ is impact of disturbing the training samples z_m's loss to validation sample z_v, i.e. the change of loss value $\mathcal{L}(z_v, \theta^*_\epsilon)$ compare to $\mathcal{L}(z_v, \theta^*)$ as Eq. 5:

$$\mathcal{I}_{\theta^*}(z_m, z_v) = \mathcal{L}(z_v, \theta^*_\epsilon) - \mathcal{L}(z_v, \theta^*) \tag{5}$$

3.2.2 High-quality Sample Selection

We select samples based on IS obtained in § 3.2.1. Specifically, we randomly extract a subset of samples from the training set as pseudo-validation set $\mathcal{D}_{pval}$, and the rest is the candidate set $\mathcal{D}_{can}$. For each sample z_m in $\mathcal{D}_{can}$, if the IS of the candidate sample z_m for all samples in the pseudo-verification set is negative (the loss of the pseudo-validation sample z_s decreases after disturbing $\mathcal{L}(z_m, \theta)$), we select z_m as a training sample for SFT as illustrated in Eq. 6:

$$\forall z_s \in \mathcal{D}_{pval}, \mathcal{I}_{\theta^*_\epsilon}(z_m, z_s) < 0 \tag{6}$$

where z_s is the sample in $\mathcal{D}_{pval}$, and $\mathcal{I}_{\theta^*_\epsilon}(z_m, z_s)$ is the IS of sample z_m on sample z_s.

3.3 Estimating IS via Taylor Expansion

As the calculation of influence score (IS) is complex (with $O(n^2)$ complexity) [1], we propose to approximate the IS (Eq. 5) via Taylor expansion. In this way, we only need to calculate the gradient n times, reducing the complexity from $O(n^2)$ to $O(n)$. Specifically, we assume the factor ϵ converges to zero in the limit and introduce the marginal change variable Δ_ϵ which is used to approximate the derivative of the optimized weight θ^*_ϵ for ϵ. Then we use the Taylor expansion to approximate Δ_ϵ. Finally, we use the chain rule to obtain IS. Detailed steps are as follows.

Step 1: Define marginal change variable Δ_ϵ. The marginal change variable Δ_ϵ quantifies the marginal change of the model weights while giving a factor ϵ to perturb the loss $\mathcal{L}(z_m, \theta)$. We introduce Δ_ϵ since the derivative of the optimized model weight θ^*_ϵ to the factor ϵ is hard to compute directly. Therefore, we use a differential method to approximate calculate IS.

We assume that the factor ϵ converges to zero in the limit and define the marginal change $\Delta_\epsilon = \theta^*_\epsilon - \theta^*$ that is the difference of the model weights between the model trained with and without disturbing. Based on the differential method, Δ_ϵ divided by the ϵ can approximate the derivatives of θ^*_ϵ for ϵ when ϵ is zero as demonstrated in Eq. 7:

$$\frac{d\theta^*_\epsilon}{d\epsilon} = \frac{\theta^*_\epsilon - \theta^*}{\epsilon - 0} \tag{7}$$

where $\frac{d\theta^*_\epsilon}{d\epsilon}$ is used to obtain IS in step 3.

Step 2: Approximate marginal change variable Δ_ϵ via Taylor expansion. We use Taylor expansion to approximate Δ_ϵ. The optimized model weight θ^*_ϵ obtained by minimizing the loss $R(\theta_\epsilon)$ under the factor ϵ, thus the first derivative of the loss $R(\theta_\epsilon)$ on θ^*_ϵ is zero as shown in Eq. 8:

$$0 = \bigtriangledown R(\theta^*_\epsilon) + \epsilon \bigtriangledown \mathcal{L}(z_m, \theta^*_\epsilon). \tag{8}$$

Then, we approximate Δ_ϵ via Taylor expansion based on Eq. 8: we first derive a first-order Taylor expansion of Eq. 8 at θ^*, which provides an

[1] For each sample z_m, we have to train the model twice on the entire candidate sample set with and without sample z_m to compute IS. For n candidate samples, we need to train the model for $2n$ times and calculate the gradient for $2n^2$ times. For example, calculating IS on Pheme datasets takes over 40,000 hours.

approximation for θ^*_ϵ and contains Δ_ϵ as Eq. 9 [2]:

$$\begin{aligned} 0 \approx & [\bigtriangledown R(\theta^*) + \epsilon \bigtriangledown \mathcal{L}(z_m, \theta^*)] \\ & + [\bigtriangledown^2 R(\theta^*) + \epsilon \bigtriangledown^2 \mathcal{L}(z_m, \theta^*)]\Delta\epsilon + o(\Delta_\epsilon) \end{aligned} \tag{9}$$

After discarding the higher-order term $o(\Delta_\epsilon)$ in Eq. 9, we obtain Δ_ϵ after a shift operation in Eq. 10,

$$\begin{aligned} \Delta_\epsilon \approx & -[\bigtriangledown^2 R(\theta^*) + \epsilon \bigtriangledown^2 \mathcal{L}(z_m, \theta^*)]^{-1} \\ & \times [\bigtriangledown R(\theta^*) + \epsilon \bigtriangledown \mathcal{L}(z_m, \theta^*)] \end{aligned} \tag{10}$$

Since θ^* is the optimized model weight obtained by minimizing the loss $R(\theta)$, we obtain $\bigtriangledown R(\theta^*) = 0$. Then we substitute this into Eq.10 and remove the higher order infinitesimal terms, i.e. retaining only the $o(\epsilon)$ term, then obtain Δ_ϵ as follows.

$$\Delta_\epsilon \approx - \bigtriangledown^2 R(\theta^*)^{-1} \bigtriangledown \mathcal{L}(z_m, \theta^*)\epsilon \tag{11}$$

Step 3: Get IS $\mathcal{I}_{\theta^*}(z_m, z_v)$ via chain rule. Recall that $\mathcal{I}_{\theta^*}(z_m, z_v)$ is IS of samples z_m on a validation sample z_v, which is the change of loss $\mathcal{L}(z_v, \theta^*_\epsilon)$ compare to $\mathcal{L}(z_v, \theta^*)$ as defined in §3.2.1 (Eq. 5). We evaluate the impact of disturbing training sample z_m for model prediction in the validation sets and use the chain rule to calculate IS $\mathcal{I}_{\theta^*}(z_m, z_v)$. Considering directly calculating the derivative of the loss $\mathcal{L}(z_v, \theta^*_\epsilon)$ for the factor ϵ is complex.[3] To reduce complexity, we apply the chain rule, which transforms complex derivative calculations into the product of two simple derivative calculations. Specifically, we introduce the intermediate variable θ^*_ϵ and define the derivative $\frac{d\mathcal{L}(z_v,\theta^*_\epsilon)}{d\epsilon}$ as the the product of $\frac{d\mathcal{L}(z_v,\theta^*_\epsilon)}{d\theta^*_\epsilon}$ and $\frac{d\theta^*_\epsilon}{d\epsilon}$. The former is computed during the back-propagation stage, while the latter is derived in Step 2 via Taylor expansion approximation as Eq. 12.

$$\begin{aligned} \mathcal{I}_{\theta^*}(z_m, z_v) &= \frac{d\mathcal{L}(z_v, \theta^*_\epsilon)}{d\epsilon}|_{\epsilon=0} \\ &= \frac{d\mathcal{L}(z_v, \theta^*_\epsilon)}{d\theta^*_\epsilon} \frac{d\theta^*_\epsilon}{d\epsilon}|_{\epsilon=0} \\ &= \bigtriangledown_{\theta|\theta^*} \mathcal{L}(z_v) \frac{d\theta^*_\epsilon}{d\epsilon}|_{\epsilon=0} \\ &= - \bigtriangledown_{\theta|\theta^*} \mathcal{L}(z_v) \mathbf{H}^{-1}_{\theta^*} \bigtriangledown_{\theta|\theta^*} \mathcal{L}(z_m) \end{aligned} \tag{12}$$

where $\mathbf{H}^{-1}_{\theta^*} = \bigtriangledown^2 \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(z_i, \theta^*)$ represents the Hessian matrix, which is the second-order partial derivative matrix indicating the impact of all training samples on the model. The gradient $\bigtriangledown_{\theta|\theta^*} \mathcal{L}(z_m)$ and $\bigtriangledown_{\theta|\theta^*} \mathcal{L}(z_v)$ indicate the impact of the training sample z_m and the validation sample z_v on the model respectively.

[2] When ϵ in Eq. 7 converges to zero in the limit, the optimized weight θ^*_ϵ converges to θ^*.

[3] The calculation needs to obtain the optimized model weight θ^* and θ^*_ϵ before and after the disturbing to compute the marginal loss, then obtain the division of the marginal loss and the factor ϵ as IS.

3.4 Shapley Value-based Analysis Selection via Cooperative Game

To obtain the most informative analysis for rumor detection, we first prompt LLMs to generate news analysis from multiple perspectives. Subsequently, we regard the multi-perspective analysis of the news as a cooperative game and employ Shapley values to quantify the contribution of individual perspective analysis to selecting informative analyses for rumor detection.

Rather than existing methods (Wan et al., 2024) that assume the equal contribution of all perspectives for detecting rumors, our model considers the different perspectives generate contradictory analyses, and not all perspective analyses benefit the prediction as some even introduce noise. Hence, we construct a cooperative game to capture potential relationships between multiple perspective analyses and quantify the contribution of each perspective analysis to select informative perspective analyses. Obtaining the informative analyses involves two steps as § 3.4.1 and § 3.4.2.

3.4.1 Multi-Perspective Analysis Generation

To comprehensively analyze the news content, the module emulates the conventional fact-checking procedure (Tsang, 2023) to prompt LLMs for generating news analysis from multiple perspectives.

We emulate the standard fact-checking procedure to select 10 perspective analyses. For each sample, we obtain multiple perspectives analyses set $A = \{a_1, a_2, ..., a_{10}\}$ by prompting LLMs with different prompt templates, the complement detail in App. B. Then we pass set A to the Shapley value-based analysis selection module (§ 3.4.2) for selecting the informative perspective analyses.

3.4.2 Shapley Value-based Analysis Selection

We utilize the Shapley value to quantify the contribution of perspective analysis to rumor detection and select informative perspective analyses. The Shapley value is a concept derived from cooperative game theory. It measures each player's contribution to the coalition's total value. In the rumor detection task, the Shapley value evaluates the contribution of each perspective analysis to predict the news label, which helps select the informative perspective analyses.

Inspired by cooperative game theory (Elkind and Rothe, 2016), we regard the collaborative process of multiple perspective analyses for predicting the news label as a cooperative game and each per-

spective analysis (i.e., emotion factor, background information, logical consistency, etc.) as a "player", while the Shapley value measures the contribution of each perspective analysis in this process. Specifically, we first calculate the PVI, then regard PVI as the value function to obtain Shapley value, and finally select informative analyses. The detailed process is as follows:

Step 1: Calculate PVI value. PVI quantifies the input's contribution to the model's prediction by comparing the difference of the model's prediction entropy with and without the input. Specifically, for a given input x with its label y, where the input $x = \{t, A\}$ contains both news content t and the perspective analyses set A, g and g' are the models after fine-tuning the rumor detection model with and without the input respectively. The PVI$(x \to y)$ is the difference in the log-probability assigned to the ground truth by these models as Eq. 13:

$$\text{PVI}(x \to y) = -\log_2 g[\emptyset]y + \log_2 g'[x](y) \quad (13)$$

Step 2: Calculate the Shapley value. Given a sample x_i, we define a cooperative game with a player set $A_i = \{a_i^1, a_i^2 \ldots, a_i^j, \ldots, a_i^m\}$ (i.e., a set of perspective analysis as mentioned in § 3.4.1), and a value function f measures the total value obtained by the coalition (a set of perspective analysis) in the cooperative game. The value function $f(A_i)$ is the PVI for predicting the label y_i given the news content t_i and the perspective analyses set A_i, i.e. $f(A_i) = \text{PVI}([t_i : A_i] \to y_i)$. The Shapley value of perspective analysis a_i^j represents its average marginal contribution across all possible combinations of perspective analysis. Specifically, we iterate all subsets $A_i' \subseteq A_i$ and compute the marginal contribution of a_i^j when it is added to A_i'. For each subset A_i', the number of all perspective analysis arrangements in A_i' is $|A_i'|!$ and the number of remaining perspective arrangements after excluding $|A_i'|!$ and a_i^j is $(|A_i| - |A_i'| - 1)!$, thereby the probability $\mathcal{P}(a_i^j, A_i')$ of a_i^j joins A_i' as shown in Eq. 14:

$$\mathcal{P}(a_i^j, A_i') = \frac{|A_i'|!(|A_i| - |A_i'| - 1)!}{|A_i|!} \quad (14)$$

we consider $\mathcal{P}(a_i^j, A_i')$ as the weight and the Shapley value $\phi_{ij}(f)$ is the weighted sum of the marginal contributions as Eq. 15:

$$\phi_{ij}(f) = \sum_{A_i' \subseteq A_i, a_j \notin A_i'} p(a_i^j, A_i') \times [f(A_i' \cup \{a_i^j\}) - f(A_i')] \quad (15)$$

where $|A_i|$ is the number of perspective analysis in set A_i, and the $f(A_i' \cup \{a_i^j\}) - f(A_i')$ is the marginal contribution of a_i^j joins A_i'.

Step 3: Select the informative perspective analysis. We select the informative analysis perspectives based on the Shapley value calculated in Step 2. Specifically, for each perspective analysis a^j, we obtain the mean value of the Shapley value for a^j in the validation set $\mathcal{D}_{val}$. If this mean value is positive, i.e. $\frac{1}{m}\sum_{v=1}^{m} \phi_{vj}(f) > 0$, indicting a^j is useful for rumor detection, we select a^j as the informative perspective analyses for SFT.

4 Experiment

4.1 Experiment Setting

Datasets. We conduct experiments on three public datasets, including human-written Pheme (Buntain and Golbeck, 2017), Weibo21 (Nan et al., 2021), and the LLM-generated LLM-mis (Chen and Shu, 2022). Following the existing work (Yuan et al., 2019), we select 20% of samples as the validation set, and the rest of the samples are split into the training set and testing set with a ratio of 3:1. The detail of Datasets in App. C.1

Baselines. We compare our model with three types of state-of-the-art baselines: (1) **LLM-based *w/o SFT***: Zero-shot, Few-shot, Zero-shot CoT, and Few-shot CoT. (2) **SLMs-based**: RvNN (Ma et al., 2018b), DEFEND (Shu et al., 2019), GLAN(Yuan et al., 2019), QSAN (Tian et al., 2020), EBGCN (Wei et al., 2021), and WSDNS (Wu et al., 2023); (3) **LLM-based**: DELL (Wan et al., 2024) and GenFEND (Nan et al., 2024). We provide more details about baselines in the App. C.2.

Implementation Details. We leverage LLama2-7B (Touvron et al., 2023a) as the base LLMs for supervised fine-tuning and prompt learning. It should be noted that instead of performing full parameter fine-tuning, we use Lora (Hu et al., 2021) to fine-tune LLama2. The batch size is 128, the learning rate is 1e-4, and the maximum length is 512; for the configuration of Lora, where the dimension $lora\_r$ of the LoRA low-rank matrix is 64, the scaling factor $lora\_alpha$ of the LoRA low-rank matrix is 128. See more details of experimental implementations in App. C.3

Metric. For evaluation metrics, we follow the existing works (Wan et al., 2024; Nan et al., 2024) and adopt macro F1 score (macF1) and accuracy (Acc.) to evaluate the performance.

Category	Method	Pheme		Weibo21		LLM-mis	
		macF1	Acc.	macF1	Acc.	macF1	Acc.
LLMs-based *w/o SFT*	Zero-shot	0.459	0.460	0.580	0.570	0.597	0.600
	Few-shot	0.490	0.500	0.610	0.622	0.565	0.567
	Zero-shot CoT	0.499	0.470	0.635	0.644	0.566	0.570
	Few-shot CoT	0.510	0.508	0.650	0.665	0.620	0.610
SLMs-based Methods	QSAN	0.668	0.751	0.675	0.710	-	-
	EBGCN	0.800	0.850	0.815	0.832	-	-
	GLAN	0.785	0.832	0.814	0.824	-	-
	RvNN	0.791	0.790	0.918	0.921	0.888	0.892
	ENDEF	0.775	0.762	0.770	0.772	0.860	0.862
	DEFEND	0.727	0.751	0.800	0.800	0.823	0.84
	WSDMS	0.810	0.805	0.912	0.923	0.862	0.873
LLMs-based Methods	DELL	0.820	0.852	0.858	0.862	0.897	0.906
	GenFEND	0.832	0.837	0.810	0.818	0.879	0.883
	ISGA(Ours)	**0.939**	**0.944**	**0.980**	**0.980**	**0.932**	**0.940**

Table 1: Rumor detection experiment results. "-" indicates that the LLM-mis datasets do not have a propagation structure, so the method based on the propagation structure cannot be used directly. "**Bold**" indicates optimal results, and "underline" indicates sub-optimal results. Our improvements are significant under the t-test with $p < 0.0025$ (See details in App. D.1).

4.2 Main Result

Tab. 1 shows the results of ISGA and the compared methods on three datasets. The results show that ISGA surpasses all baselines across three datasets. Noticeably, we achieved an accuracy improvement of over 10% compared with the second-best baseline on the Pheme dataset, and the accuracy exceeded 98% on the Weibo21 dataset. We also conduct an additional experiment about ISGA on different series and scales LLMs in App. D.2.

Specifically, the performance of LLMs-based and SLMs-based methods significantly outperforms the LLMs-based *w/o SFT* on all datasets. This indicates the LLMs-based *w/o SFT* lacks rumor detection task-specific knowledge, while the LLMs-based and SLMs-based methods learn this during fine-tuning. Furthermore, our model outperforms all baselines and achieves the most improvement on the Pheme. This superiority is attributed to two factors: first, the influence-based sample selection module utilizes sample selection to enhance the LLMs' ability to acquire task-specific knowledge for rumor detection via SFT; second, the Shapley value-based analysis selection module designs the multiple perspective analysis collaborative process to make the model capture in-depth analysis of news content for prediction.

4.3 Ablation Study

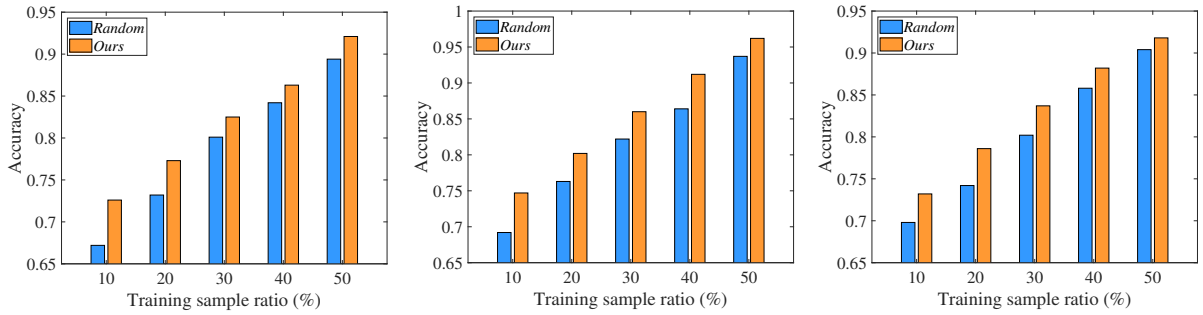
To demonstrate the effectiveness of the proposed component, we conduct ablation study: **w/o SSM**: the variant without Influence-based Sample Selection module (§ 3.2); **w/o MPAM**: the variant without Multi-Perspective Analysis Generation module (§ 3.4.1); **w/o ASM**: the variant without Shapley Value-based Analytical Selection module (§ 3.4.2).

As shown in Tab. 2, each component is essential for the effectiveness of our method. Specifically, the performance of both variants **w/o SSM** and **w/o ASM** drops significantly. This indicates that the quality of the samples and the generated analyses substantially impact the effectiveness of SFT on LLMs. Furthermore, the variant **w/o MPAM** shows the largest decrease in performance, suggesting that incorporating multi-perspective analyses into SFT enables LLMs to conduct a more in-depth and comprehensive analysis of news content.

4.4 Analysis experiment of sample selection

To validate the effectiveness of the Influence-based Sample Selection module in improving SFT, we compared our proposed sample selection method with the random sampling method. Specifically, we set the ratio of the training samples in the dataset to 10%, 20%, 30%, 40%, and 50%[4], and compare

[4]Since we split 60% of the dataset as the training set, the training sample ratio cannot exceed 60%.



(a) PHEME dataset (b) Weibo dataset (c) LLM-mis dataset

Figure 2: IS experiment results under different training sample ratios. "Random" indicates that selecting the training samples from the candidate set randomly.

Method	Pheme	Weibo	LLM-mis
Ours	**0.944**	**0.980**	**0.940**
w/o SSM	0.932	0.965	0.930
w/o MPAM	0.902	0.943	0.910
w/o ASM	0.920	0.962	0.923

Table 2: Ablation study (Acc.). w/o SSM: remove Influence-based Sample Selection module; w/o MPAM: remove Multi-Perspective Analysis Generation module; w/o ASM: remove Shapley Value-based Analytical Selection module.

accuracy (Acc.) between these two methods.

As shown in Fig. 2, the results indicate that our method consistently outperforms the random method across all scenarios. As the number of training samples increases, the performance gap between our method and the random method narrows. This highlights the stronger generalization ability of our method in scenarios with fewer samples. Additionally, it underscores the importance of sample quality for SFT, especially when the number of samples is limited.

4.5 Analysis experiment of analysis selection

To confirm that our Shapley Value-based Analytical Selection via Cooperative Game module provides more informative analyses, we conducted an experiment comparing accuracy (Acc.) between our proposed multi-perspective analysis selection method and other analysis selection variants. Specifically, we designed the following variants: **All**: the variant selects all perspective analysis; **One**: the variant selects one perspective analysis randomly; **Two**: the variant selects two perspective analysis randomly.

As shown in Tab. 3, Variant **All** significantly outperforms variants **One** and **Two**, indicating that incorporating more perspective analysis provides more information for rumor detection. However, our method outperforms Variant **All**, suggesting that our analytical selection strategy can effectively choose perspectives with high information content by leveraging the Shapley value.

Method	Pheme	Weibo	LLM-mis
All	0.920	0.962	0.923
One	0.912	0.942	0.906
Two	0.908	0.939	0.902
Ours	**0.944**	**0.980**	**0.940**

Table 3: The analysis experiment of the perspective analysis quantity (Acc.). "All" means selecting all perspective analysis, "one" means selecting one perspective analysis randomly, and "two" means selecting two perspective analyses randomly.

5 Conclusion

In summary, we propose an LLM-based rumor detection method via SFT, which selects samples based on the influence of samples and conducts multi-perspective analyses via a multiplayer game. Firstly, we employ the IS to evaluate the impact of training samples on model predictions for selecting reliable samples, where we utilize the Taylor expansion with a chain rule to approximate IS for reducing computation cost. Secondly, we use LLMs to generate an in-depth multi-perspective analysis of new content, model each perspective analysis as a player in a cooperative game, and measure each perspective analysis's contribution to rumor detection using the Shapley value to obtain informative perspective analyses for SFT. Finally, we conduct experiments on three real-world datasets to

show the superiority of our approach over existing methods in rumor detection.

6 Limitation

Although our method produces promising results on three datasets, it has certain limitations. We will continue to investigate these concerns in the future.

Firstly, We only used text content for rumor detection, ignoring multimodal information like images. In real world social media, news often combines text and images, with images providing additional clues for rumor detection. Thus, relying solely on text may be insufficient for comprehensive detection. However, it is worth noting that in the field of rumor detection, many existing studies also focus only on text. Our future research will consider incorporating multimodal information to further enhance the practicality of rumor detection.

Secondly, this work lacked explanations for the model's prediction results. In practice, many rumor detection scenarios require explanations to enhance the credibility of the model's predictions. However, most studies in this field only provide results without explanations. Future research will introduce interpretability mechanisms to enhance the credibility of the model's predictions.

Acknowledgments

This work is supported by the following foundations: the National Natural Science Foundation of China under Grant No.62376284, 62406332, 62202487, and 62306330, the National Science Foundation for Distinguished Young Scholars under Grant No.62025208, and the Young Elite Scientist Sponsorship Program by CAST under Grant No.YESS20230367.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing factual knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Canyu Chen and Kai Shu. 2022. Can llm-generated misinformation be detected? In *In The Twelfth International Conference on Learning Representations*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Mengyang Chen, Lingwei Wei, Han Cao, Wei Zhou, and Songlin Hu. 2023. Can large language models understand content and propagation for misinformation detection: An empirical study. *CoRR*.

CHEQ. 2023. The economic cost of bad actors on the internet. *https://info.cheq.ai/hubfs/Research/THE ECONOMIC COST Fake News fnal.pdf. Accessed: 2023-08-13*.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Edith Elkind and Jörg Rothe. 2016. Cooperative game theory. *Economics and computation: an introduction to algorithmic game theory, computational social choice, and fair division*, pages 135–193.

Li Gao, Lingyun Song, Jie Liu, Bolin Chen, and Xuequn Shang. 2022. Topology imbalance and relation inauthenticity aware hierarchical graph attention networks for fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*.

Kris Hartley and Minh Khuong Vu. 2020. Fighting fake news in the covid-19 era: policy insights from an equilibrium model. *Policy sciences*, pages 735–758.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Zhen Huang, Zhilong Lv, Xiaoyun Han, Binyang Li, Menglong Lu, and Dongsheng Li. 2022. Social bot-aware graph neural network for early rumor detection. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Tsz Kin Lam, Eva Hasler, and Felix Hieber. 2022. Analyzing the use of influence functions for instance-specific data filtering in neural machine translation. *arXiv preprint arXiv:2210.13281*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023a. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. *arXiv preprint arXiv:2310.15515*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023b. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Yifeng Luo, Yupeng Li, Dacheng Wen, and Liang Lan. 2024. Message injection attack on rumor detection under the black-box evasion setting using large language model. In *Proceedings of the ACM on Web Conference 2024*, pages 4512–4522.

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

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018b. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Anand Matheven and Burra Venkata Durga Kumar. 2022. Fake news detection using deep learning and natural language processing. In *2022 9th International Conference on Soft Computing Machine Intelligence (ISCMI)*.

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*.

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*. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15395–15406, Bangkok, Thailand. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2022. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

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*.

Tian Tian, Yudong Liu, Xiaoyu Yang, Yuefei Lyu, Xi Zhang, and Binxing Fang. 2020. Qsan: A quantum-probability based signed attention network for explainable false information detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Hugo Touvron, Louis Martin, and Kevin R. 2023a. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Stephanie Jean Tsang. 2023. Hkbu fact check. *https://factcheck.hkbu.edu.hk/home/en/fact-check/our-process/ [Accessed: (Accessed:December 5, 2023)]*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2023. Emergent abilities of large language models. *ArXiv*.

Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2023. 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*.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision tree-based co-attention networks for explainable claim verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A convolutional approach for misinformation identification. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3901–3907.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE International Conference on Data Mining (ICDM)*.

A Rumor Detection Module

In this module, we use the news text x_i and the set of informative perspective analyses to determine whether the news is a rumor or not. We have designed the following supervised fine-tuning (SFT) template:

- **Instruction**: Next, I will give you a news story and an analysis of the veracity of the news; please combine the analysis of the news to ultimately determine whether the news is a rumor or not.
- **Input**: News: {news text}; analysis:{analysis text}.
- **Output**: The news label.

B The Detail of Multi-perspective Analysis Generation

We design the following prompt set $P = \{p1, p2, ..., p_{10}\}$, as illustrated in Tab. 4, where 10 denote the number of perspective analysis. For a given news x_i and prompt p_j, the prompt template T as:

You are a professional fact-checker, and you have been tasked with verifying the authenticity of news to determine whether it is a rumor. Given a news: $\{x_i\}$, and your verification content will be: $\{p_j\}$.

Num.	Prompt Template
1	Does the news contain sufficient background information?
2	Is the background information in the news accurate and objective?
3	Is there any content deliberately omitted that distorts the meaning?
4	Is there any improper intent in the news (political motives, commercial purposes, etc.)?
5	Is the news report based on facts, or does it mainly rely on speculation or opinions?
6	Are there any logical fallacies or misleading arguments in the news report?
7	Does the news exhibit bias?
8	Are there any grammatical or spelling errors in the news report that might indicate a lack of professional editing?
9	Does the news report use inflammatory language or engage in personal attacks?
10	Is the news purely an expression of emotion without involving judgments of truthfulness?

Table 4: Prompt Learning Template

Dataset	Train		Validation		Test		Totol
	Fake	Real	Fake	Real	Fake	Real	
Weibo21	2,883	2,179	540	702	539	724	7,567
LLM-mis	410	289	118	82	53	42	1,000
Pheme	1,153	2,298	384	766	384	766	5,802

Table 5: Datasets statistic.

C Experiment Setting Details

C.1 Dataset Details

We evaluate ISGA and baselines on three public datasets on the rumor detection task, and the dataset statistics are shown in Tab. 5.

- **Pheme** (Buntain and Golbeck, 2017) is a dataset of potential rumors on Twitter and journalistic assessments of their accuracies.
- **LLM-mis** (Chen and Shu, 2022) is an LLM-generated misinformation dataset with different LLM generators and generation approaches.
- **Weibo21** (Nan et al., 2021) is a comprehensive dataset that encompasses news from nine domains, namely science, military, education, disaster, politics, health, finance, entertainment, and society. Each domain contains news content, publication timestamps, corresponding images, and comments. In total, Weibo21 comprises 4,488 instances of false news and 4,640 instances of true news.

C.2 Baseline Details

We compare our method with several methods, the details are described as follows:

- Zero-Shot asks LLMs to conduct detection.
- Few-shot first provides LLMs with some pairs of news instances and labels and then asks LLMs to conduct detection.
- Zero-shot CoT uniquely leverages LLMs' self-formulated rationales by integrating a standard instruction with the simple phrase, "Let's think step by step known as Chain of Thoughts (CoT).
- Few-shot Cot is based on the Zero-shot CoT with extra pairs of news instances and labels.
- RvNN (Ma et al., 2018b) proposes two recursive neural models based on bottom-up and top-down tree-structured neural networks for rumor representation learning and classification.
- DEFEND (Shu et al., 2019) conducts explainable detection by the attention weights; we set the maximum sentence length and maximum comment length as 96, the maximum sentence count as 64, and the maximum comment count as 10 to reproduce so that the approach is applicable to our tasks and datasets.
- GLAN (Yuan et al., 2019) is designed for rumor detection on social media by jointly encoding local semantic relations between tweets and retweets and global structural information of the heterogeneous propagation graph to improve classification accuracy and early detection performance.
- QSAN (Tian et al., 2020) is a quantum-based signed attention network that uses a quantum-inspired textual representation to model textual semantics and captures post-commentary

Dataset	pheme		Weibo		LLM-mis	
	macF1	Acc.	macF1	Acc.	macF1	Acc.
Bartlett's Test	2.6128×10^{-5}	0.000114843	2.78354×10^{-7}	9.96033×10^{-5}	0.002388062	0.000599668

Table 6: The p values of t-test on our method. The p values are all smaller than 0.025.

Method	Pheme		Weibo21		LLM-mis	
	macF1	Acc.	macF1	Acc.	macF1	Acc.
LLaMA2-13B	<u>0.942</u>	**0.948**	<u>0.982</u>	<u>0.984</u>	<u>0.936</u>	**0.945**
LLaMA2-7B	0.939	<u>0.944</u>	0.980	0.980	0.932	0.940
LLaMA3-8B	**0.945**	0.940	0.972	0.972	**0.937**	<u>0.942</u>
Qwen2.5-7B	0.932	0.935	**0.985**	**0.984**	0.928	0.935

Table 7: Different LLMs' results on our method, "**Bold**" indicates optimal results, and "<u>underline</u>" indicates sub-optimal results

relations through a new symbolic attention mechanism

- EBGCN (Wei et al., 2021) is a novel model for rumor detection on social media that addresses the uncertainty in propagation structures by using a Bayesian approach to adaptively adjust edge weights and an edge-wise consistency training framework to optimize the model, achieving improved performance in both rumor detection and early rumor detection tasks.

- WSDMS (Wu et al., 2023) needs bag-level labels for training but possesses the capability to infer both sentence-level misinformation and article-level veracity, facilitated by pertinent social media conversations meticulously contextualized with news sentences.

- DELL (Wan et al., 2024) is a model that leverages large language models (LLMs) to generate diverse user reactions, create explainable proxy tasks, and ensemble expert predictions to enhance the detection of misinformation in news articles.

- GenFEND (Nan et al., 2024) enhances fake news detection by generating diverse user comments through large language models (LLMs) and analyzing these comments from multiple demographic subpopulations to provide comprehensive feedback.

C.3 Implementation Details

In our implementation of the Influence-based Sample Selection Module in §3.2, we use the gradients of the model's Multilayer Perceptron (MLP) parameters in $\{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$-th layers to speed up calculate the influence score.

D Additional Experiment Results

D.1 T-test

To substantiate the efficacy of our proposed model, we conducted a series of experiments, each involving the replacement of the seed across ten distinct runs of fake news detection on various datasets. For each iteration, we calculate the t-test confidence interval P to statistically assess the comparative metrics of our model against the established baseline in each metric. We consider $P < 0.025$ as a significant enhancement and $P > 0.025$ as an insignificant enhancement. As presented in Tab. 6, the empirical findings reveal that, across all three datasets, our model has achieved a significant enhancement in most metrics when compared to the baseline.

D.2 Experiment on Different LLMs

We have selected LLaMA2-13B, LLaMA3-8B, and Qwen2.5-7B as the base models for experiments. As shown in Tab.7, our method performs well across different series and scales of LLMs, showing its effectiveness and generalizability.