GCML: Gradient Coherence Guided Meta-Learning for Cross-Domain Emerging Topic Rumor Detection

Zejiang He[1,2†], Jingyuan Huang[1,2†], Menglong Lu[1,3*], Zhen Huang[1,2*], Shanshan Liu[1,2], Zhiliang Tian[1,2], Dongsheng Li[1,2]

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

{hezejiang, jingyuanhuang, lumenglong, huangzhen, tianzhiliang, liushanshan17, dsli}@nudt.edu.cn

Abstract

With the emergence of new topics on social media as sources of rumor propagation, addressing the domain shift between the source and target domain and the target domain samples scarcity remains a crucial task in cross-domain rumor detection. Traditional deep learning-based methods and LLM-based methods are mostly focused on the in-domain condition, thus having poor performance in cross-domain setting. Existing domain adaptation rumor detection approaches ignore the data generalization differences and rely on a large amount of unlabeled target domain samples to achieve domain adaptation, resulting in less effective on emerging topic rumor detection. In this paper, we propose a **G**radient **C**oherence guided **M**eta-**L**earning approach (**GCML**) for emerging topics rumor detection. Firstly, we calculate the task generalization score of each source task (sampled from source domain) from a gradient coherence perspective, and selectively learn more "generalizable" tasks that are more beneficial in adapting to the target domain. Secondly, we leverage meta-learning to alleviate the target domain samples scarcity, which utilizes task generalization scores to re-weight meta-test gradients and adaptively updates learning rate. Extensive experimental results on real-world datasets show that our method substantially outperforms SOTA baselines.

1 Introduction

The rapid and wide spread of rumors on social media especially around emerging topics poses huge threats to public trust and social cohesion (He et al., 2023). Detecting rumors on emerging topics aims to automatically identify inaccurate and intentionally misleading news during the early stage of a topic's emergence. Deep learning-based methods have become mainstream in rumor detection.

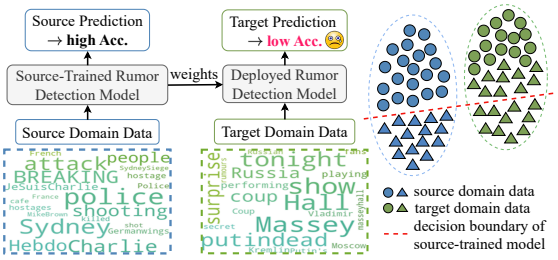


Figure 1: Domain Shift. Existing models trained on the source domain fail to detect rumors on the target domain (emerging topic domain).

Traditional deep learning-based methods mainly depend on news content (Yu et al., 2017; Ma et al., 2018), propagation structures (Bian et al., 2020; Matheven and Kumar, 2022), or user information (Huang et al., 2022; Gao et al., 2022) for rumor detection and have made significant progress. Recently, approaches based on large language models (LLMs) have also achieved desirable rumor detection results through prompt engineering (Chen et al., 2023a; Wan et al., 2024), in-context learning (Hu et al., 2024), or supervised fine-tuning (SFT) (Yang et al., 2024; Wan et al., 2024). However, these methods detect rumors under in-domain conditions, i.e., assuming that the training and test datasets are from the same data distribution. In practice, social platforms frequently release various news claims in diverse domains. Different domain presents their own characteristics (e.g., word usage, topic, writing style) as shown in Fig. 1. Thus, a model trained on the source domain will result in serious performance degradation when testing on newly emergent domain due to the domain shift.

To address the challenges of domain shift, some studies proposed crowdsourcing (Kou et al., 2022b; Chen et al., 2023b; Shang et al., 2024), which leverages domain experts or online resources to acquire domain knowledge. However, these methods require extensive human annotation costs. Another al-

† contributed equally to this work
* corresponding author

ternative approach is domain adaptation (Yue et al., 2022; Zeng et al., 2022; Yue et al., 2023), which transfers knowledge from the source domain to the target domain. These methods mitigate the domain shift by feature space alignment (Shu et al., 2022; Ran and Jia, 2023) or data space alignment (Shi et al., 2023; Chen et al., 2025). However, the above approaches still have two limitations: (1) They learn all source tasks[1] from the source domain equally without considering the generalization differences among tasks. Learning from source tasks with a large generalization gap to the target domain can lead to suboptimal performance. (2) Existing domain adaptation rumor detection methods, including semi-supervised (Li et al., 2021), and unsupervised methods (Mackey et al., 2021; Ran and Jia, 2023), mainly rely on a large amount of unlabeled target data to achieve domain adaptation. Since newly emergent domain topics are difficult to acquire sufficient data in time (Ran and Jia, 2023), these methods are less effective in few-shot learning scenarios where only few-shot target samples are available for emerging topic rumor detection.

In this paper, we propose a **G**radient **C**oherence guided **M**eta-**L**earning approach (**GCML**) for cross-domain emerging topic rumor detection, which improves domain generalization from both task-level selective learning and parameter-level adaptive update. For lacking generalization difference consideration, we introduce the Gradient Signal to Noise Ratio (GSNR) (Liu et al., 2020) to measure the generalization from a gradient coherence perspective. By comparing the model parameters' gradient coherency on the source task with that on few-shot target samples, we obtain the task generalization score of each source task for selectively learning more "generalizable" source tasks, which are more helpful in improving model generalization to the target domain. For addressing the target domain data scarcity, inspired by MAML (Finn et al., 2017), we propose a meta-learning based framework for few-shot domain adaptation. Our method aims to learn the generalization feature from the source data under the guidance of limited target samples. We first sample data batches from the source domain as source tasks to train the model. Then, we evaluate the updated model on the few-shot target samples to derive second-order meta-test gradients with respect to the original parameters. Finally, we fine-grained update the initial model parameters with re-weighted meta-test gradients based on task generalization scores, and adaptively update the learning rate of parameters with the high gradient coherency. Extensive domain adaptation experiments conducted on multiple datasets show that our method outperforms existing state-of-the-art (SOTA) baselines.

Our contributions are four-fold: (1) We propose GCML, a domain adaptation method for cross-domain emerging topic rumor detection, which 'learns-to-adapt' to target data distribution from both task-level and parameter-level. (2) We propose a GSNR-guided generalization calculation method to selectively learn more "generalizable" source tasks for addressing domain shift. (3) We propose a gradient coherence based meta-learning framework, which leverages task generalization scores to re-weight meta-test gradients and adaptively updates learning rate for few-shot domain adaptation. (4) Extensive experiments on real-world datasets show the effectiveness of GCML.

2 Related Work

Rumor Detection. Deep learning-based rumor detection methods can be categorized into three types: (1) content-based methods (Yu et al., 2017; Ma et al., 2018) employ deep learning models to capture the textual features of rumors for detecting rumors. (2) propagation structure-based methods capture temporal dynamics (Wu et al., 2020; Lu et al., 2022) and spatial structures (Bian et al., 2020; Sun et al., 2022a,b) in the news propagation process to identify rumors. (3) user-based methods (Huang et al., 2022; Gao et al., 2022) utilize user attributes and historical behavior for rumor detection.

Recently, some research has explored using LLMs as rumor detectors (Wang et al., 2024; Chen and Shu, 2024; Pelrine et al., 2023; Yang et al., 2024), and leveraging their commonsense reasoning to provide supplementary explanations for rumor detection. Cheung and Lam (2023) employ LoRA-tuning to train an LLaMA-based rumor detector. Hu et al. (2024) integrate LLMs to provide multi-perspective rationales for improved rumor detection. Tian et al. (2025) propose an SFT-based LLM rumor detection model with influence-guided sample selection and game-based multi-perspective analysis. However, these methods focus on improving in-domain performance, and their performance is poor in cross-domain rumor detection settings.

[1]As defined in § 3, batches of source data from the source domain are sampled as different "source tasks".

Domain Adaptation. Domain adaptation aims to train a model on a source dataset that can generalize well to a target dataset, even if the data distributions of source and target datasets differ (Singhal et al., 2023; Li et al., 2023). Such methods minimize the representation discrepancy (Kang et al., 2019; Na et al., 2021; Lu et al., 2023a; Singhal et al., 2023) between source and target domains to learn domain-invariant features. Domain adaptation has been applied to mitigate domain discrepancies in cross-domain rumor detection (Nan et al., 2022; Liu et al., 2025). Lin et al. (2022); Shu et al. (2022); Lu et al. (2023b) use domain-adversarial training to learn generalizable features for cross-domain rumor detection. Mosallanezhad et al. (2022) propose a reinforcement learning domain-aware feature extraction method for fine-grained domain adaptation. Ran and Jia (2023) adapt contrastive learning with cross-attention for unsupervised domain adaptation. However, these methods are not suitable for few-shot domain adaptation scenarios, and their equal learning from all source tasks limits their generalization performance.

Gradient-based Generalization. Generalization is considered the key to the performance of deep neural networks (Zhang et al., 2021; Chatterjee and Zielinski, 2022). Some studies (Chatterjee, 2020) indicate that stronger gradient coherence enhances a model's generalization capability. Liu et al. (2020) use the concept of gradient signal to noise ratio (GSNR) to establish a quantitative relationship between gradient coherence and model generalization. Subsequently, smoothing gradients to reduce sample gradient variance has been applied in tasks such as neural architecture search (Sun et al., 2023; Bai et al., 2025) and domain generalization (Michalkiewicz et al., 2023). Fort et al. (2019) introduce stiffness to measure the generalization, which focuses on how gradients in one sample affect loss changes in another. In this paper, we use gradient coherence to evaluate the task generalization for selective learning.

3 Problem Formulation

In this paper, we study the cross-domain emerging topic rumor detection, which is defined as a few-shot domain adaptation problem from a single-source domain[2] $\mathbb{D}_S$ to a target domain[3] $\mathbb{D}_T$.

[2]A high-resource domain with adequate annotated data.

[3]A low-resource domain that has a limited amount of data during the early stage of the emerging topic.

Definition 1. Source data (D_S): From source domain $\mathbb{D}_S$, we can obtain a source labeled training dataset $D_S = \{(x_i, y_i)\}_{i=1}^{N}$. The input x_i is a news claim, the corresponding label $y_i \in \{0, 1\}$ (i.e., false or true). During training, batches of source data are sampled as different "**source tasks**".

Definition 2. Few-shot target samples (D_T^l): we can access limited labeled data of the target domain during the early stage of the emerging topic. In our setting, we assume that only a K-shot subset D_T^l from $D_T \subseteq \mathbb{D}_T$ is provided for training, i.e., $D_T^l = \{(x_j, y_j)\}_{j=1}^{K}$, and the label $y_j \in \{0, 1\}$.

By utilizing source data D_S and limited few-shot target samples D_T^l, the objective of cross-domain emerging topic rumor detection is to train a rumor detection model $f(\theta)$ that optimizes the performance on the target test data D_T ($D_T \neq D_T^l$) from $\mathbb{D}_T$. Mathematically, the overall objective is minimizing the loss $\mathcal{L}$ of θ on D_T as Eq. (1):

$$\min_{\theta} \mathcal{L}(\theta, D_T) \tag{1}$$

4 Methods

4.1 Overview

Our method consists of two parts, as shown in Fig. 2. (1) The GSNR-guided generalization calculation module (§ 4.2) proposes task generalization score, which utilizes GSNR with few-shot target samples to measure the generalization of each source task for selectively learning more beneficial source tasks; (2) The gradient coherence based meta-learning module (§ 4.3) is a bi-level optimization framework consisting of Meta-Train, Meta-Test and Meta-Optimization. It leverages task generalization score (obtained from § 4.2) to re-weight meta-test gradients and adaptively update learning rates for bidirectional adaptation to target domain. The algorithm of our method is shown in App. A.

4.2 GSNR-Guided Generalization Calculation

To selectively learn source tasks that are more beneficial for enhancing the model's generalization to the target domain, we propose task generalization score that utilizes GSNR to measure the generalization of each source task, which is defined as the similarity of the model's gradient coherence on the source tasks to that on the few-shot target samples.

Existing domain adaptation methods learn equally from all source tasks without considering the distinctive generalization of each task, leading to suboptimal performance. Therefore, considering

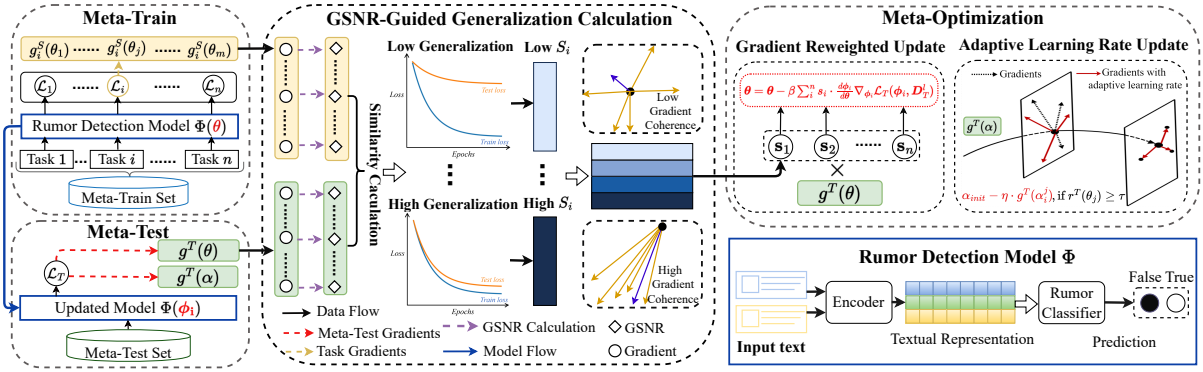


Figure 2: The overview of our framework. The process begins with the left top part, which trains the model on source tasks and derives task gradients. Then, we evaluate the updated model on the meta-test set (i.e., few-shot target samples) to derive meta-test gradients. Next, the middle part computes the generalization score of each source task. Finally, the right top part updates the initial model parameters with re-weighted meta-test gradients, accompanied by adaptive learning rate updates. The bottom-right part is the base rumor detection model.

this distinctive generalization is crucial for domain adaptation. Our GSNR-guided generalization calculation involves two steps as follows.

Task-Wise GSNR. We quantify the generalization of each source task using the model parameters' GSNR on these tasks. GSNR is the ratio of squared mean and variance of parameters' gradients on a particular data distribution, which represents gradient coherency and measures the generalization performance. Liu et al. (2020) prove that a higher GSNR exhibits a smaller generalization gap, i.e., stronger generalization.

Given a source dataset D_S, a rumor detection model parameterized by θ, we sample batches of source data from the source domain as different "source tasks". For source task i with corresponding data $\mathcal{B}_i$, we first calculate the task gradients $g_i^S(\theta_j)$ of the loss function $\mathcal{L}_S(\theta, \mathcal{B}_i)$ with respect to each parameter θ_j trained on $\mathcal{B}_i$ as Eq. (2):

$$g_i^S(\theta_j) = \frac{\partial \mathcal{L}_S(\theta, \mathcal{B}_i)}{\partial \theta_j} \tag{2}$$

for the j-th parameter θ_j, we can obtain its mean and variance of the data distribution $\mathcal{B}_i$ within the current source task i, enabling the calculation of the individual parameter θ_j's GSNR $r_i^S(\theta_j)$ as Eq. (3):

$$r_i^S(\theta_j) = \frac{\mathbb{E}^2_{(x,y)\sim\mathcal{B}_i}(g_i^S(x, y, \theta_j))}{\mathrm{Var}_{(x,y)\sim\mathcal{B}_i}(g_i^S(x, y, \theta_j))} \tag{3}$$

for model parameter $\{\theta_1, ..., \theta_j, ..., \theta_m\}$, where m denotes the number of parameters, we can obtain m GSNR: $R_i^S = [r_i^S(\theta_1), ..., r_i^S(\theta_j), ..., r_i^S(\theta_m)]$. Similarly, we can calculate the GSNR for each parameter θ_j on the few-shot target samples D_T^l based on the gradients $g^T(\theta_j)$, resulting in corresponding GSNR: $R^T = [r^T(\theta_1), ..., r^T(\theta_j), ..., r^T(\theta_m)]$.

Task Generalization Score Calculation. We obtain the generalization score of source task by comparing the model's gradient coherence (i.e., GSNR) on source task to that on few-shot target samples.

The GSNR of model parameters is positively correlated with the generalization. Thus, a source task with model parameters' GSNR similar to that of the few-shot target samples results in a small generalization gap, making it more beneficial for adapting to the target domain.

We calculate the similarity between the two sequences of GSNR R_i^S and R^T as the task generalization score in Eq. (4):

$$s_i = \mathrm{Cos}(R_i^S, R^T) \tag{4}$$

In each iteration, we sample n source tasks from source domain and compute the task generalization score for each source task with the few-shot target samples. These scores $[s_1, s_2, ..., s_n]$ are converted into probability weights via softmax function, i.e., $s = \mathrm{softmax}(s_1, s_2, ..., s_n)$, which will be utilized to re-weight the meta-test gradients in § 4.3 for fine-grained updating the initial model parameters.

4.3 Gradient Coherence based Meta-Learning

To mitigate the target domain data scarcity and improve the model's performance to the target domain with limited target samples, we propose a gradient coherence based meta-learning domain adaptation framework, which leverages task-level generalization scores and parameter-level adaptive learning rate for effective few-shot domain adaptation.

Meta-learning, or "learning-to-learn", enables quick adaptation to new tasks by learning general feature representations from multiple source tasks with the guidance of limited target samples. Therefore, we employ meta-learning to optimize the initial model parameters for optimal performance on unseen few-shot target tasks. Specifically, our meta-learning involves a series of loops over the "Meta-Train", "Meta-Test" and "Meta-Optimization".

Meta-Train on Source Tasks. Given the model f with initial parameters θ, source dataset D_S, and few-shot target samples D_T^l, we first sample n source tasks from D_S as the meta-train set in each iteration. Each source task's data is formalized as $\mathcal{B}_i = \{(x_1, y_1), (x_2, y_2), \dots, (x_{|\mathcal{B}_i|}, y_{|\mathcal{B}_i|})\}$, where $|\mathcal{B}_i|$ is the number of samples. The loss of model f on the current source task i is as Eq. (5):

$$\mathcal{L}_S(\theta, \mathcal{B}_i) = \frac{1}{|\mathcal{B}_i|} \sum_{x_j, y_j \in \mathcal{B}_i} \mathcal{L}(f(x_j; \theta), y_j) \quad (5)$$

Then, we perform a pseudo-update on the model parameters with Eq. (6). After several steps of gradient descent, the parameters locally converge. We donate the updated parameters as ϕ_i.

$$\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}_S(\theta, \mathcal{B}_i) \quad (6)$$

where α denotes the task learning rates and $\nabla_\theta \mathcal{L}_S(\theta, \mathcal{B}_i)$ represents the task gradient.

Meta-Test on Few-Shot Target Samples. After the pseudo-update in the meta-train phase, we calculate the meta-test loss of the model $f(\phi_i)$ on the K-shot target samples D_T^l as Eq. (7):

$$\mathcal{L}_T(\phi_i, D_T^l) = \frac{1}{|D_T^l|} \cdot \sum_{x_j, y_j \in D_T^l} \mathcal{L}(f(x_j; \phi_i), y_j) \quad (7)$$

We derive the gradient of the meta-test loss with respect to the initial parameters θ via the chain rule (Finn et al., 2017), which is second-order meta-test gradient as Eq. (8):

$$\frac{d\mathcal{L}_T(\phi_i, D_T^l)}{d\theta} = \frac{d\phi_i}{d\theta} \nabla_{\phi_i} \mathcal{L}_T(\phi_i, D_T^l) \quad (8)$$

Meta-Optimization by Gradient Coherence. In the meta-optimization, our objective is to learn optimal model parameters θ that minimize the meta-test loss on the few-shot target samples D_T^l. After obtaining the task gradient $\nabla_\theta \mathcal{L}_S(\theta, \mathcal{B}_i)$ of source task i and the meta-test gradients $\frac{d\phi_i}{d\theta} \nabla_{\phi_i} \mathcal{L}_T(\phi_i, D_T^l)$, we can calculate the task generalization score s_i for source task i as illustrated in § 4.2. Finally, we use s_i to re-weight corresponding meta-test gradient and update the initial model parameters as Eq. (9):

$$\theta = \theta - \beta \sum_i^n s_i \cdot \frac{d\phi_i}{d\theta} \nabla_{\phi_i} \mathcal{L}_T(\phi_i, D_T^l) \quad (9)$$

where β is learning rate during meta-optimization. This approach finds more "generalizable" source tasks and assigns larger weights to their correlated meta-test gradient. In this way, we can fully exploit the source domain knowledge and improve the model's performance on the target domain.

Adaptive Learning Rate Update. To further improve generalization and convergence, we propose a GSNR-based adaptive learning rate update strategy to optimize the meta-train. Previous studies (Antoniou et al., 2018) have theorized that using a static learning rate for all parameters not only reduces the generalization performance but also increases the hyperparameter tuning cost.

Therefore, we argue learning different learning rates for each iteration of the parameters whose GSNR in R^T exceeds a set threshold τ during the meta-train phase. Specifically, there will be m learning rates for corresponding parameters, i.e., $(\alpha_i^1, ..., \alpha_i^j, ..., \alpha_i^m)$ in the source task i's learning rates α_i of Eq. (6). We update each learning rate α_i^j through the adaptive learning steps as Eq. (10):

$$\alpha_i^j = \alpha_{init} - \eta \cdot \frac{d\phi_i}{d\alpha_i^j} \nabla_{\phi_i} \mathcal{L}(\phi_i, D_T^l), \text{if } r_j \geq \tau \quad (10)$$

where η represents learning rate for updating α.

5 Experiment

5.1 Experiment Setting

Datasets. We conduct experiments on multiple real-world source and target datasets. Following Yue et al. (2022, 2023), we use FEVER (Thorne et al., 2018), GettingReal (Risdal, 2016), GossipCop (Shu et al., 2020), LIAR (Wang, 2017) and PHEME (Buntain and Golbeck, 2017) as the source datasets. For target datasets, we adopt CoAID (Cui and Lee, 2020), Constraint (Patwa et al., 2021) and ANTiVax (Hayawi et al., 2022). The details of datasets are listed in the App. B.1.

Baselines. We compare our model with two types of baselines. (1) State-of-the-art domain adaptation

Source	Target	CoAID			Constraint			ANTiVax		
	Metric	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑
FEVER	CANMD	0.626	0.918	0.956	0.684	0.683	0.686	0.650	0.679	0.749
	ACLR	0.721	0.935	0.965	0.648	0.651	0.697	0.739	0.758	0.805
	ProtoNet	0.751	0.869	0.925	0.784	0.788	0.812	0.748	0.716	0.718
	MAML	0.780	**0.939**	**0.967**	0.812	0.808	0.797	0.826	0.808	0.823
	MetaAdapt	0.829	0.875	0.927	0.828	0.826	0.829	0.868	0.880	0.904
	CADM+	0.654	0.928	0.958	0.671	0.668	0.667	0.661	0.709	0.778
	GCML(Ours)	**0.856**	0.938	0.966	**0.834**	**0.833**	**0.838**	**0.911**	**0.908**	**0.928**
GettingReal	CANMD	0.669	0.935	0.965	0.744	0.742	0.737	0.582	0.632	0.729
	ACLR	0.693	0.928	0.961	0.683	0.689	0.736	0.660	0.695	0.751
	ProtoNet	0.720	0.639	0.757	0.672	0.664	0.608	0.736	0.756	0.804
	MAML	0.813	**0.937**	**0.965**	0.808	0.803	0.786	0.819	0.802	0.819
	MetaAdapt	0.830	0.928	0.960	0.819	0.819	0.823	0.886	0.882	0.902
	CADM+	0.689	0.925	0.955	0.731	0.727	0.717	0.604	0.654	0.751
	GCML(Ours)	**0.860**	0.934	0.963	**0.825**	**0.838**	**0.845**	**0.920**	**0.912**	**0.931**
GossipCop	CANMD	0.685	0.931	0.963	0.802	0.803	0.817	0.761	0.777	0.823
	ACLR	0.687	**0.933**	0.964	0.712	0.715	0.744	0.811	0.809	0.835
	ProtoNet	0.708	0.609	0.731	0.786	0.782	0.770	0.730	0.715	0.736
	MAML	0.816	0.926	0.959	0.813	0.809	0.801	0.826	0.810	0.826
	MetaAdapt	0.824	0.918	0.954	0.826	0.826	0.833	0.896	0.907	0.930
	CADM+	0.712	0.929	0.959	0.778	0.778	0.782	0.752	0.770	0.815
	GCML(Ours)	**0.855**	0.928	**0.965**	**0.841**	**0.837**	**0.846**	**0.904**	**0.915**	**0.933**
LIAR	CANMD	0.770	0.894	0.940	0.815	0.814	0.818	0.755	0.784	0.834
	ACLR	0.766	0.938	0.966	0.756	0.760	0.786	0.805	0.793	0.814
	ProtoNet	0.793	0.910	0.950	0.738	0.746	0.788	0.599	0.576	0.581
	MAML	0.813	**0.938**	**0.966**	0.813	0.809	0.800	0.824	0.807	0.824
	MetaAdapt	0.815	0.910	0.949	0.820	0.820	0.828	0.873	0.883	0.906
	CADM+	0.780	0.891	0.941	0.728	0.730	0.742	0.712	0.715	0.765
	GCML(Ours)	**0.848**	0.933	0.962	**0.830**	**0.834**	**0.844**	**0.903**	**0.909**	**0.930**
PHEME	CANMD	0.531	0.938	0.967	0.559	0.565	0.624	0.653	0.676	0.704
	ACLR	0.709	0.939	0.967	0.716	0.719	0.746	0.733	0.754	0.804
	ProtoNet	0.721	0.780	0.808	0.693	0.686	0.644	0.628	0.635	0.685
	MAML	0.800	**0.939**	**0.967**	0.816	0.812	0.802	0.819	0.805	0.823
	MetaAdapt	0.828	0.909	0.949	0.818	0.818	0.828	0.896	0.880	0.902
	CADM+	0.581	0.932	0.954	0.643	0.654	0.730	0.675	0.732	0.810
	GCML(Ours)	**0.861**	0.932	0.961	**0.819**	**0.821**	**0.838**	**0.911**	**0.905**	**0.926**

Table 1: Cross-domain rumor detection results, the best and second best results are in **bold** and underlined.

and few-shot learning rumor detection methods: ProtoNet (Snell et al., 2017), MAML (Finn et al., 2017), CANMD (Yue et al., 2022), ACLR (Lin et al., 2022), MetaAdapt (Yue et al., 2023), CADM+ (Zeng et al., 2024). (2) LLM-based methods: we select LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023) as rumor detectors to conduct zero-shot prompting, few-shot prompting, and supervised fine-tuning (SFT). We provide more baseline details in the App. B.2.

Metric. Following Yue et al. (2023), we adapt balance accuracy (BA), accuracy (Acc.), and F1 score (F1) to evaluate the performance.

Implementation Details. Similar to (Yue et al., 2022), we leverage Roberta (Liu et al., 2019) as the base rumor detection model. We follow the previous works (Kou et al., 2022a; Yue et al., 2022) and divide the dataset into training, validation, and test sets with the ratio of 7:2:1. We employ the 10-shot setting for few-shot domain adaptation. The batch size is set to 4. The learning rates for both the meta-train and meta-optimization are initialized to $1e-5$, and 3 source tasks are sampled in each meta-train iteration. More implementation details are provided in App. B.3.

5.2 Main Results

Cross-domain adaptation results. The results of all source-target combinations in cross-domain adaptation experiments are shown in Tab. 1. It can be observed that our method outperforms all baselines in all source-target adaptation scenarios in the BA metric. Specifically, on the CoAID dataset with imbalanced label classes[4], our method achieves the most significant improvement in BA compared to the baselines despite being slightly inferior in Acc. and F1. For instance, our method surpasses the second-best baseline MetaAdapt by an average of 3.1% on the CoAID target domain.

[4]Since the CoAID dataset contains over 90% positive labels, the BA metric is more reliable than Acc. and F1 for evaluating performance on this dataset.

Metric	CoAID			Constraint			ANTiVax		
	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑
Ours	0.856	0.933	0.963	0.829	0.831	0.842	0.910	0.908	0.929
w/o Task Generalization	0.816	0.912	0.952	0.809	0.806	0.810	0.865	0.857	0.882
w/o Adaptive LR	0.802	0.922	0.957	0.814	0.810	0.805	0.864	0.864	0.886
w/o Meta-Learning	0.794	0.929	0.959	0.805	0.797	0.775	0.846	0.846	0.873

Table 2: Ablation study. w/o Task Generalization: removing GSNR-guided generalization calculation module; w/o Adaptive LR: removing adaptive learning rate module; w/o Meta-Learning: removing meta-learning module.

The superiority of our method can be attributed to the following factors: first, the task generalization score helps to learn domain-invariant features that are more conducive to adapting to the target domain; second, we employ meta-learning to acquire relatively equitable features, alleviating the imbalanced target domain label distribution.

Comparison to LLMs. To further validate the effectiveness of our method in domain adaptation rumor detection, we compare our approach with LLM-based methods. Specifically, we use LLaMA and Alpaca to conduct zero-shot prompting, few-shot prompting, and SFT on the target domain dataset, with the results presented in Tab. 3.

The performance of most LLM-based methods with SFT significantly outperforms zero-shot and few-shot prompting methods. This indicates that LLM-based methods without SFT lack task-specific knowledge for rumor detection, while LLMs with SFT acquire this during fine-tuning. Our method enables smaller language model Roberta to outperform fine-tuned LLMs. For example, our method achieves a higher BA on the ANTiVax dataset by 16.8% compared to LLAMA-SFT and by 14.3% compared to Alpaca-SFT.

Setting	Dataset	BA ↑	Acc. ↑	F1 ↑
LLaMA-Zero-shot	CoAID	0.460	0.239	0.308
	Constraint	0.501	0.485	0.258
	ANTiVax	0.573	0.466	0.383
Alpaca-Zero-shot	CoAID	0.488	0.211	0.252
	Constraint	0.498	0.482	0.234
	ANTiVax	0.561	0.445	0.333
LLaMA-Few-shot	CoAID	0.500	0.906	0.951
	Constraint	0.500	0.523	0.687
	ANTiVax	0.500	0.664	0.798
Alpaca-Few-shot	CoAID	0.515	0.908	0.952
	Constraint	0.537	0.559	0.704
	ANTiVax	0.528	0.681	0.806
LLaMA-SFT	CoAID	0.749	0.874	0.928
	Constraint	0.724	0.721	0.718
	ANTiVax	0.742	0.756	0.811
Alpaca-SFT	CoAID	0.766	0.818	0.892
	Constraint	0.688	0.686	0.689
	ANTiVax	0.767	0.779	0.828
Ours	CoAID	**0.856**	**0.933**	**0.963**
	Constraint	**0.829**	**0.831**	**0.842**
	ANTiVax	**0.910**	**0.908**	**0.929**

Table 3: Comparison to large language models, the best and second best results are in **bold** and underlined.

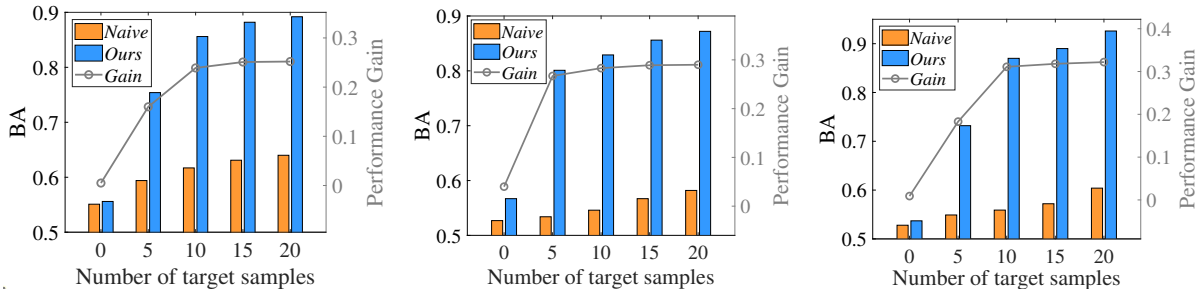
5.3 Ablation Study

To demonstrate the effectiveness of the proposed component, we conduct ablation study: *w/o Task Generalization*: the variant without GSNR-guided generalization calculation (§ 4.2); *w/o Adaptive LR*: the variant without adaptive learning rate update; *w/o Meta-Learning*: the variant without meta-learning(§ 4.3), i.e., learning with first-order approximation. The results of a single target dataset are averaged across the five source datasets.

As shown in Tab. 2, each component is essential for the effectiveness of our method. Specifically, the performance of variants *w/o Task Generalization* and *w/o Adaptive LR* drops by 3.5% and 3.8% in BA metric on average. This indicates that the GSNR-guided generalization calculation and adaptive learning rate update substantially contribute to the effectiveness of our method. Furthermore, the variant *w/o Meta-Learning* that replaces second-order Grads with first-order approximation results in the largest performance decrease, suggesting that meta-learning enables transferring knowledge from source domain to target domain effectively.

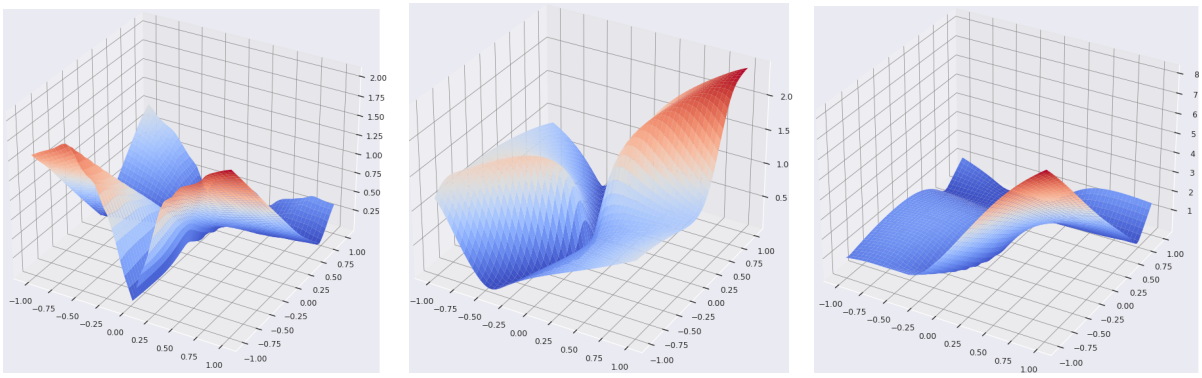
5.4 Analysis of Target Domain Sample Size

To validate the effectiveness of our method under different number of few-shot target samples, we compare our method with the naive-learning (Yue et al., 2023) method on three target datasets. Naive-learning indicates fine-tuning the model pre-trained on the source domain using few-shot target samples. Specifically, we set the number of available target domain samples to 0, 5, 10, 15, and 20, respectively, then compare BA between these two methods and calculate the performance gain. For the 0-shot setting, we train the model on source dataset and directly evaluate on target test data.



(a) CoAID dataset (b) Constraint dataset (c) ANTiVax dataset

Figure 3: Cross-domain emerging topic rumor detection performance with respect to the number of few-shot target samples, as the number of target samples increases from 0 to 20.



(a) Naive-Learning (b) MetaAdapt (c) Ours

Figure 4: The 3D loss surface comparison between Naive-Learning, MetaAdapt, and Ours. It can be seen that our method significantly smooths and flattens the loss surface, i.e., improving the model generalization.

As shown in Fig. 3, we observe: (1) our method outperforms naive-learning across all settings. As the number of target samples increases, the performance gap between our method and naive-learning widens. This highlights that our method enables more effectively utilizing the limited few-shot target samples; (2) the performance gain of our method increases rapidly during the initial stage (i.e., 0-shot to 5-shot) and then gradually stabilizes as the number of samples further increases. This indicates the stronger generalization ability of our method in scenarios with fewer target samples.

5.5 Analysis of Model Generalization

To have a visualization look at the model generalization, we present the loss landscapes of different methods. We follow Li et al. (2018) and show the 3D loss surface results in Fig. 4. Foret et al. (2021) prove that flatter and smoother regions of the loss surface are associated with better generalization performance. It can be observed that our method has flatter and smoother surfaces compared to others, which indicates that our method improves the generalization of models effectively. We also provide additional analysis of task generalization differences and the generalization results about the GSNR of model parameters over time in App. C.

6 Conclusion

In summary, we propose a gradient coherence-guided meta-learning approach for cross-domain emerging topic rumor detection, which improves domain generalization from both task-level selective learning and parameter-level adaptive updating. Firstly, we calculate the task generalization score of each source task by comparing the model's gradient coherence between the source task and few-shot target samples, which are used to selectively learn more "generalizable" source tasks. Secondly, we introduce a meta-learning based framework to learn the generalization feature from the source data and adaptively update parameters and learning rates under the guidance of limited target samples. Extensive experiments on real-world datasets show the superiority of our approach over baselines.

Limitations

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

Firstly, we only use text content for rumor detection but ignore multi-modal information like images. In social media, news typically contains text and images, where images can serve as supplementary information for rumor detection. However, it is worth mentioning that many existing studies also only focus on text. Our future research will consider incorporating multi-modal information to achieve comprehensive rumor detection.

Secondly, we do not consider the domain adaptation setting from multi-source domains to the target domain. In practice, much research also only explores domain adaptation with a single source domain. Future research will conduct domain adaptation rumor detection from the multi-source domains to the target domain.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62406332 and 62376284.

References

Antreas Antoniou, Harrison Edwards, and Amos Storkey. 2018. How to train your maml. *In International Conference on Learning Representations*.

Lichen Bai, Zixuan Xiong, Hai Lin, Guangwei Xu, Xiangjin Xie, Ruijie Guo, Zhanhui Kang, Hai-Tao Zheng, and Hong-Gee Kim. 2025. Frozen language models are gradient coherence rectifiers in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, pages 1817–1825.

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)*, pages 208–215. IEEE.

Satrajit Chatterjee. 2020. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. *International Conference on Learning Representations*.

Satrajit Chatterjee and Piotr Zielinski. 2022. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*.

Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

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

Songlin Chen, Xiaoliang Chen, Duoqian Miao, Hongyun Zhang, Xiaolin Qin, and Peng Lu. 2025. Ada-uda: A transferable transformer framework for rumor detection using adversarial domain alignment within unsupervised domain adaptation. *Expert Systems with Applications*, 261:125487.

Zhongwu Chen, Chengjin Xu, Fenglong Su, Zhen Huang, and Yong Dou. 2023b. Meta-learning based knowledge extrapolation for temporal knowledge graph. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2433–2443.

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*, pages 846–853. IEEE.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *ICLR 2021 - 9th International Conference on Learning Representations*.

Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. 2019. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*.

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*, pages 4687–4696.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

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, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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*.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022a. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022b. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *IJCAI*, pages 5087–5093.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Jinpeng Li, Yingce Xia, Xin Cheng, Dongyan Zhao, and Rui Yan. 2023. Learning disentangled representation via domain adaptation for dialogue summarization. In *Proceedings of the ACM Web Conference 2023*, pages 1693–1702.

Yichuan Li, Kyumin Lee, Nima Kordzadeh, Brenton Faber, Cameron Fiddes, Elaine Chen, and Kai Shu. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 668–676. IEEE.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556, Seattle, United States. Association for Computational Linguistics.

Jinlong Liu, Guo-Qing Jiang, Yunzhi Bai, Ting Chen, and Huayan Wang. 2020. Understanding why neural networks generalize well through gsnr of parameters. In *International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Shanshan Liu, Menglong Lu, Zhen Huang, Zejiang He, Liu Liu, Zhigang Sun, and Dongsheng Li. 2025. MONTROSE: LLM-driven Monte Carlo tree search self-refinement for cross-domain rumor detection. In *Findings of the Association for Computational Linguistics*, pages 21475–21487.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Menglong Lu, Zhen Huang, Binyang Li, Yunxiang Zhao, Zheng Qin, and DongSheng Li. 2022. Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 429–442.

Menglong Lu, Zhen Huang, Zhiliang Tian, Yunxiang Zhao, Xuanyu Fei, and Dongsheng Li. 2023a. Meta-tsallis-entropy minimization: A new self-training approach for domain adaptation on text classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5159–5169.

Menglong Lu, Zhen Huang, Yunxiang Zhao, Zhiliang Tian, Yang Liu, and Dongsheng Li. 2023b. DaMSTF: Domain adversarial learning enhanced meta self-training for domain adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1650–1668.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1989, Melbourne, Australia.

Tim K Mackey, Vidya Purushothaman, Michael Haupt, Matthew C Nali, and Jiawei Li. 2021. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on twitter. *The Lancet Digital Health*, 3(2):e72–e75.

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)*.

Mateusz Michalkiewicz, Masoud Faraki, Xiang Yu, Manmohan Chandraker, and Mahsa Baktashmotlagh. 2023. Domain generalization guided by gradient signal to noise ratio of parameters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6177–6188.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM web conference 2022*, pages 3632–3640.

Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. 2021. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1094–1103.

Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021*, pages 21–29.

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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429.

Hongyan Ran and Caiyan Jia. 2023. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13510–13518.

Megan Risdal. 2016. Getting real about fake news.

Lanyu Shang, Yang Zhang, Bozhang Chen, Ruohan Zong, Zhenrui Yue, Huimin Zeng, Na Wei, and Dong Wang. 2024. Mmadapt: A knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection. In *Proceedings of the ACM Web Conference 2024*, page 4653–4663, New York, NY, USA.

Yu Shi, Xi Zhang, Yuming Shang, and Ning Yu. 2023. Don't be misled by emotion! disentangle emotions and semantics for cross-language and cross-domain rumor detection. *IEEE Transactions on Big Data*, 10(3):249–259.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection*, pages 215–232. Springer.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. 2022a. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4611–4619.

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022b. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2789–2797. Association for Computing Machinery.

Zihao Sun, Yu Sun, Longxing Yang, Shun Lu, Jilin Mei, Wenxiao Zhao, and Yu Hu. 2023. Unleashing the power of gradient signal-to-noise ratio for zero-shot nas. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5763–5773.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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

Zhiliang Tian, Jingyuan Huang, Zejiang He, Zhen Huang, Menglong Lu, Linbo Qiao, Songzhu Mei, Yijie Wang, and Dongsheng Li. 2025. LLM-based rumor detection via influence guided sample selection and game-based perspective analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 28402–28414, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada.

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*.

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, and 1 others. 2017. A convolutional approach for misinformation identification. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3901–3907.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 2423–2433.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5223–5239.

Huimin Zeng, Zhenrui Yue, Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 159–162. IEEE.

Huimin Zeng, Zhenrui Yue, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Unsupervised domain adaptation via contrastive adversarial domain mixup: A case study on covid-19. *IEEE Transactions on Emerging Topics in Computing*, 12(4):1105–1116.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

A Algorithm of Our Method

The algorithm of our method is shown in Alg. 1.

Algorithm 1 GCML Algorithm

Require: Model parameters θ, source domain data D_S, K-shot target samples D_T^l, number of iterations $\mathcal{T}$, number of tasks n

for $t = 1 \rightarrow \mathcal{T}$ **do**
  for $i \in \{1, ..., n\}$ **do**
    Sample source task data $\mathcal{B}_i$ from D_S
    Meta-Train:
    Update parameter θ using $\mathcal{B}_i$ with Eq. (6)
    Derive the task gradient with Eq. (2)
    Meta-Test:
    Calculate meta-test loss and meta-test gradients using D_T^l as Eq. (7) and Eq. (8)
    GSNR Generalization Calculation:
    Compute each parameter's GSNR on source task $\mathcal{B}_i$ and D_T^l as Eq. (3)
    Calculate task generalization score s_i for source task i with Eq. (4)
  end for
  Meta-Optimization:
  Update original parameter θ with Eq. (9)
  Adaptive update learning rate with Eq. (10)
end for

B Experiment Setting Details

B.1 Datasets

We use FEVER (Thorne et al., 2018), GettingReal (Risdal, 2016), GossipCop (Shu et al., 2020), LIAR (Wang, 2017) and PHEME (Buntain and Golbeck, 2017) as the source datasets. For target datasets, we adopt CoAID (Cui and Lee, 2020), Constraint (Patwa et al., 2021) and ANTi-Vax (Hayawi et al., 2022). In the following, we provide the details of source and target datasets. The statistics of these datasets are shown in Tab. 4, where "Neg." and "Pos." indicate the proportions of rumor and non-rumor samples in the dataset, "Avg.Len" is the average token length of the text, and "Content Type" denotes the source type of the text. It is worth noting that CoAID is highly imbalanced, with over 90% positive samples.

- **FEVER** (Thorne et al., 2018) is a fact verification dataset which contains claims extracted and altered from Wikipedia.

- **GettingReal** (Risdal, 2016) is a fake news

Datasets	Neg.	Pos.	Avg.Len	Content Type
FEVER	29.6%	70.4%	9.4	Statement
GettingReal	8.8%	91.2%	738.9	News
GossipCop	24.2%	75.8%	712.9	News
LIAR	44.2%	55.8%	20.2	Statement
PHEME	34.0%	66.0%	21.5	Social Network
CoAID	9.7%	90.3%	54.0	News / Statement
Constraint	47.7%	52.3%	32.7	Social Network
ANTiVax	38.3%	61.7%	26.2	Social Network

Table 4: Statistics of the datasets.

dataset from Kaggle[5]. It contains text and metadata from online websites.

- **GossipCop** (Shu et al., 2020) is a part sub-dataset from the FakeNewsNet dataset, which includes news content, social context, and dynamic information from social media platforms.
- **LIAR** (Shu et al., 2020) is a public available dataset for fake news detection. LIAR comprises manually labeled short statements in various contexts collected from PolitiFact.com.
- **PHEME** (Buntain and Golbeck, 2017) contains a collection of Twitter rumors and non-rumors posted during breaking news. It also provides information about the structure of the conversation.
- **CoAID** (Cui and Lee, 2020) is a COVID-19 healthcare misinformation dataset, which includes fake news and related user engagements.
- **Constraint** (Patwa et al., 2021) is a manually annotated collection of social media posts and articles labeled as real or fake. It was released as the shared task at the CONSTRAINT workshop.
- **ANTiVax** (Hayawi et al., 2022) a Twitter dataset for COVID-19 vaccine misinformation detection. It is a collection of over 15,000 tweets related to COVID-19 vaccines, annotated using reliable sources and validated by medical experts.

B.2 Baselines

We compare our method with several state-of-the-art domain adaptation and few-shot learning rumor detection baselines and large language model based methods. The details are described below:

Domain Adaptation Detection Baselines.

- **ProtoNet** (Snell et al., 2017) uses prototypical networks to learn a metric space where classification is performed by computing distances to prototype representations of each class for few-shot classification. We use the same label space and base transformer model as the encoder for few-shot domain adaptation rumor detection.
- **MAML** (Finn et al., 2017) is a meta-learning algorithm that trains model parameters to enable rapid adaptation to new tasks with only a few gradient steps. MAML first updates model parameters on sampled tasks, then computes the meta-loss and derives second-order gradients with respect to the original parameters.
- **CANMD** (Yue et al., 2022) uses pseudo labeling and label correction to generate target samples and correct label shifts, and integrates a contrastive adaptation loss to learn domain-invariant features. We obtained the CANMD's results by incorporating the few-shot target samples into the training process.
- **ACLR** (Lin et al., 2022) presents an adversarial contrastive learning framework to improve cross-domain rumor detection performance with language alignment and supervised contrastive training. We replace the original graph convolution networks with our base transformer model for content-based rumor detection.
- **MetaAdapt** (Yue et al., 2023) is a meta-learning method for domain adaptive few-shot misinformation detection. It adapts models to target domains using limited data by rescaling meta-gradients based on gradient similarity.
- **CADM+** (Zeng et al., 2024) is an unsupervised domain adaptation framework, which uses adversarial domain mixup and contrastive learning to transfer knowledge from a source domain to a target domain for cross-domain rumor detection.

LLM-based Baselines. We select two representative LLMs: LLaMA2-7B (Touvron et al., 2023) and Alpaca2-7B (Taori et al., 2023) to conduct zero-shot prompting, few-shot prompting, and supervised fine-tuning on target domain for rumor detection. We adopt the following prompt template to perform zero-shot and few-shot prompting:

[5]https://www.kaggle.com/datasets/mrisdal/fake-news

Question: Given the following news, predict its veracity. If it is more likely to be real news, return 1; otherwise, return 0. Do not return any extra content. Please refrain from providing ambiguous assessments such as undetermined. The news is: [news text].
Input/output examples: [Optional for few-shot prompting].
Answer: [A predicted veracity label].

We design the following supervised fine-tuning (SFT) template:

Instruction: Next, I will give you a news claim; please determine whether the news is a rumor or not.
Input: News: [news text].
Output: The news label.

B.3 Implementation Details

We preprocess the data following Yue et al. (2022). Specifically, we convert special symbols (e.g., emojis) to English, remove special characters from the input text, and tokenize hashtags, mentions, and URLs. We implement our method and other baselines by applying PyTorch with CUDA 10.0 on Ubuntu 18.04.5 LTS servers with NVIDIA A100 GPU. For model optimization, we use AdamW with 0.01 weight decay. For the few-shot domain adaptation setting, we follow Yue et al. (2023) and select the first k samples (with 10-shot as default) from the original validation set and use the remaining samples for validation. The threshold τ is set based on the magnitude of GSNR values in R^T. The balanced accuracy (BA) in evaluation metric is defined as the average value of sensitivity and specificity as Eq. (11), which evaluates the adaptation performance in both classes equally, thus enabling a better and more reliable evaluation of domain adaptation performance under label imbalance.

$$\text{BA} = \frac{1}{2}(\text{TPR} + \text{TNR}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \quad (11)$$

where TPR denotes sensitivity, and TNR represents specificity. TP, TN indicate true positive and true negative, FP and FN represent false positive and false negative. For the results of domain adaptation baselines, we use the results from the original paper if provided. Otherwise, we reimplement the baseline methods following the original paper and its hyperparameter configuration. For LLM-based methods, supervised fine-tuning(SFT) is performed using low-rank adaptation (LoRA) (Hu et al., 2022), where the dimension $lora\_r$ of the LoRA low-rank matrix is 64, and the scaling factor $lora\_alpha$ of the LoRA low-rank matrix is 128. The batch size of LORA is 128, and the max token length is 1024. The regular expression used in LORA is [q_proj, v_proj, k_proj, o_proj, gate_proj, down_proj, up_proj]. In the meta-train phase, we pseudo-update the model three times on each source task to make it locally converge. We train the model for 500 iterations and validate it every 50 iterations. The best-performing model is then evaluated on the target test set.

C Additional Experiment Results

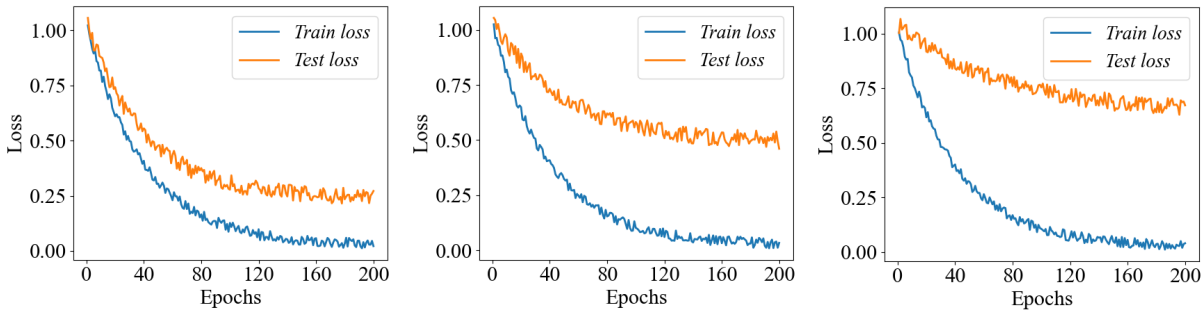
C.1 Analysis of Task Generalization

To validate that each source task contributes differently to adapt to the target domain (i.e., varying generalization gaps), we compare the one-step generalization ratios (OSGR) of the model training on different source tasks. The OSGR (Liu et al., 2020) is a metric that quantifies the ratio of the expected test loss decrease to the expected training loss decrease in one single training step. A higher OSGR, closer to 1, indicates better generalization performance, as it suggests that the model's improvements on the training set translate effectively to the test set. Specifically, we sample three source tasks from the source domain as the training set, with the target domain serving as the test set under the "PHEME → ANTiVax" scenario, and the model's OSGR results are shown in Fig. 5.

We observed: (1) the training loss decreases faster than the test loss, and $0 < OSGR(t) < 1$, resulting in a non-zero generalization gap between source task and target task at the end of training; (2) different source tasks have varying impacts on the model's OSGR. In Fig. 5 (a), the $OSGR(t)$ remains large in the entire training process, indicating a small generalization gap at the end of training and showing good generalization capability of the model. In contrast, Fig. 5 (c) shows a small $OSGR(t)$, resulting in a slower test loss decrease, corresponding to a larger generalization gap.

C.2 GSNR of Model Parameters over Time

To validate whether our approach enhances the overall GSNR of the model's parameters over time, we calculate an average GSNR of all parameters on few-shot target samples over the training period. A



(a) High $OSGR(t)$ (b) Middle $OSGR(t)$ (c) Low $OSGR(t)$

Figure 5: OSGR results under different source tasks. If OSGR(t) is large (closer to 1) throughout the training process, the generalization gap would be small, representing the good generalization ability of the model.

higher GSNR signifies better generalizability. The experiment is conducted under the "PHEME $\rightarrow$ COAID" domain adaptation scenario.

As shown in Fig. 6, our method demonstrates a higher GSNR compared to the baseline, which confirms that our approach enhances the model parameters' GSNR, i.e., improving the generalization capability on the target domain.

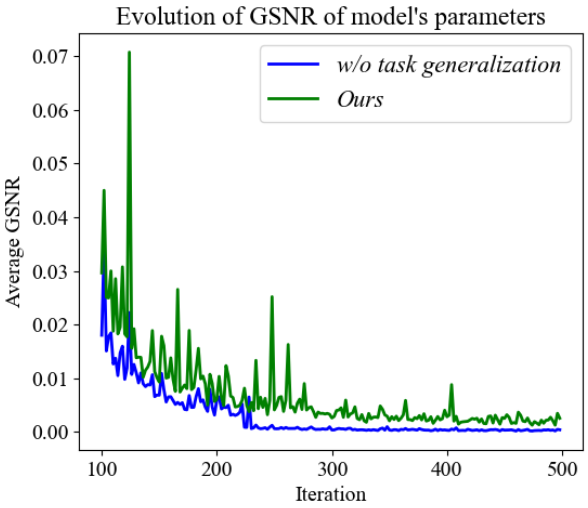


Figure 6: Our approach enhances the model's GSNR over time, compared to the baseline.