A Multi-Agent Framework with Automated Decision Rule Optimization for Cross-Domain Misinformation Detection

Hui Li[1*], **Ante Wang**[1*], **Kunquan Li**[1], **Zhihao Wang**[1], **Liang Zhang**[1], **Delai Qiu**[2], **Qingsong Liu**[2], **Jinsong Su**[1,3†]
[1] School of Informatics, Xiamen University, China
[2] Unisound AI Technology, China
[3] Shanghai Artificial Intelligence Laboratory, China
huilinlp@xmu.edu.cn jssu@xmu.edu.cn

Abstract

Misinformation spans various domains, but detection methods trained on specific domains often perform poorly when applied to others. With the rapid development of Large Language Models (LLMs), researchers have begun to utilize LLMs for cross-domain misinformation detection. However, existing LLM-based methods often fail to adequately analyze news in the target domain, limiting their detection capabilities. More importantly, these methods typically rely on manually designed decision rules, which are limited by domain knowledge and expert experience, thus limiting the generalizability of decision rules to different domains. To address these issues, we propose a **M**ulti-**A**gent Framework for cross-domain misinformation detection with Automated Decision **R**ule **O**ptimization (MARO). Under this framework, we first employs multiple expert agents to analyze target-domain news. Subsequently, we introduce a *question-reflection mechanism* that guides expert agents to facilitate higher-quality analysis. Furthermore, we propose a decision rule optimization approach based on carefully designed cross-domain validation tasks to iteratively enhance decision rule effectiveness across domains. Experimental results and analysis on commonly used datasets demonstrate that MARO achieves significant improvements over existing methods.

1 Introduction

Nowadays, social media is flooded with misinformation spanning multiple domains such as politics, economics, and technology, significantly impacting people's lives and societal stability (Della Giustina, 2023). However, due to the differences in background knowledge and linguistic features across domains, misinformation detection models trained on specific domains often perform poorly when applied to others (Ran and Jia, 2023; Liu et al., 2024e). Thus, cross-domain misinformation detection offers substantial practical value, leading to increased research attention on this task. (Choudhry et al., 2022; Lin et al., 2022; Ran et al., 2023; Ran and Jia, 2023; Liu et al., 2024e; Karisani and Ji, 2024).

Generally, cross-domain misinformation detection methods are trained on the mixture of multiple source-domain datasets, and then evaluated on a unseen target-domain one (Hernández-Castañeda et al., 2017; Lin et al., 2022; Ran et al., 2023; Ran and Jia, 2023). Early studies primarily use machine learning methods with various classifiers (Pérez-Rosas and Mihalcea, 2014; Hernández-Castañeda et al., 2017). Subsequently, researchers resort to deep learning-based methods (Choudhry et al., 2022; Lin et al., 2022; Ran et al., 2023; Ran and Jia, 2023), which, however, suffer from limited training data. In recent years, with the emergence of Large Language Models (LLMs), researchers have shifted their attention to exploring the powerful capabilities of LLMs (Hang et al., 2024; Liu et al., 2024e). For example, Hang et al. (2024) explore incorporating graph knowledge into LLMs for cross-domain misinformation detection. Very recently, Liu et al. (2024e) propose a Retrieval-Augmented Generation approach that achieves state-of-the-art performance. They extract labeled source-domain examples based on emotional relevance and manually design a decision rule. These examples and the decision rule are incorporated into the prompt to directly judge target-domain veracity.

In spite of their success, these methods still have two major drawbacks. First, they tend to treat misinformation detection as a monolithic task, overlooking that news understanding is inherently multi-dimensional—covering linguistic features, external factual consistency, user comments, and so on. Although Wan et al. (2024) makes an initial attempt to incorporate multiple proxy tasks, their analysis remains inadequate[1]. More importantly, these methods rely on manually designed decision rules,

[1]We validate this issue in Section 3.3 through experiments.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5721–5737
November 4-9, 2025

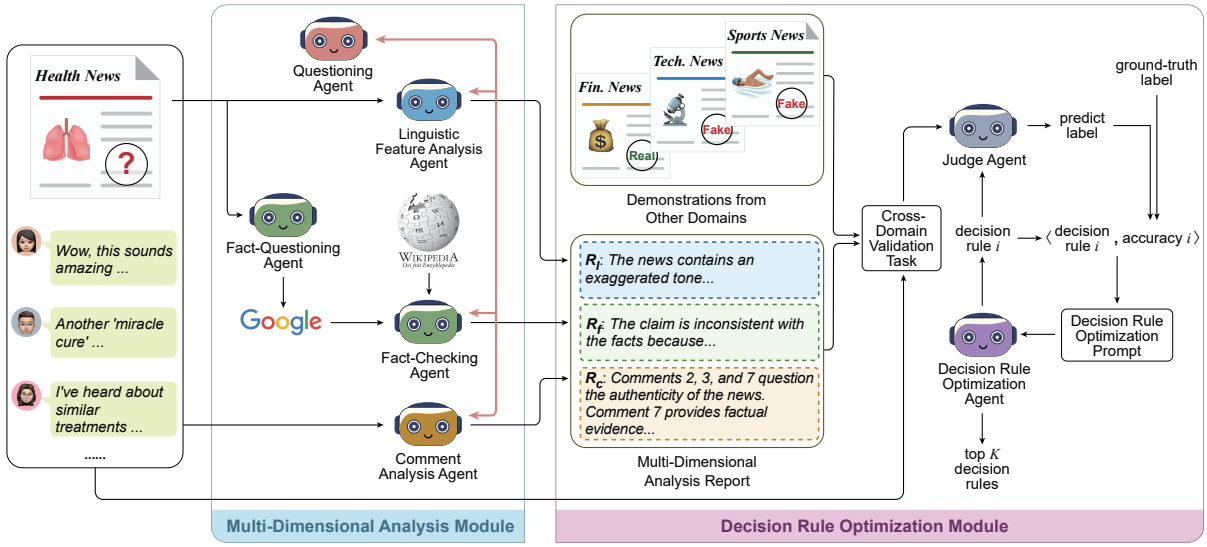

Figure 1: MARO first performs a multi-dimensional analysis on the news to be verified. Afterwards, the news, the multi-dimensional analysis report, and demonstration news from other domains are provided to the Judge Agent for verification. Meanwhile, the Decision Rule Optimization Agent supplies the Judge Agent with a decision rule to guide its verification. As the Judge Agent makes decision, the resulting <decision rule, accuracy> pairs form an optimization trajectory, which in turn enables the Decision Rule Optimization Agent to further refine decision rules.

which are typically developed based on domain-specific knowledge and experts' experience. However, news from different domains often exhibit different background knowledge and linguistic features. As a result, these decision rules usually struggle to effectively detect misinformation across different domains, leading to poor adaptability.

In this paper, we propose a **M**ulti-**A**gent Framework for cross-domain misinformation detection with Automated Decision **R**ule **O**ptimization, MARO. As illustrated in Figure 1, MARO consists of two main modules: 1) Multi-Dimensional Analysis Module, which decomposes the complex analysis task into several subtasks, each handled by an expert agent focusing on a specific aspect—such as linguistic features, external fact consistency, and user comments—collectively producing a set of analysis reports. In particular, to improve the quality of these analyses, we introduce a question-reflection mechanism, which employs a Questioning Agent to generate corresponding reflection questions based on the initial analysis reports, thereby helping the above expert agents produce more refined analysis responses. 2) The Decision Rule Optimization Module, which is specifically designed to automatically optimize and generate more effective decision rules. For this purpose, we gather news from different domains within the source-domain dataset and construct a series of validation tasks designed to simulate cross-domain misinformation detection scenarios. This module iteratively optimizes the decision rules according to their performance on the validation tasks.

We evaluate the performance of MARO using two commonly-used cross-domain misinformation detection datasets. Experimental results show that MARO outperforms existing state-of-the-art baselines across multiple LLMs. Further experiments demonstrate that both Multi-Dimensional Analysis Module and Decision Rule Optimization Module effectively improve the performance of MARO.

2 Our Method

2.1 Task Formulation

Given multiple source domain news datasets $D_s = \{D_s^i\}_{i=1}^{|D_s|}$ and a target domain news datasets D_t, each domain contains multiple news items represented as $(x_j, c_j, y_j)_{j=1}^{|D_*|}$, where x_j denotes the news content, $c_j = \{c_j^k\}_{k=1}^{|c_j|}$ represents the set of comments related to x_j, and $y_j \in \{0, 1\}$ is the corresponding ground-truth label. The goal of the cross-domain misinformation detection is to use source domain data to learn model parameters or decision rules with sufficient generalizability, and then effectively apply them to the target domain.

2.2 MARO

As shown in Figure 1, MARO consists of two main modules: the Multi-Dimensional Analysis Module and the Decision Rule Optimization Module, both

of which employ LLM-based agents to perform various tasks. We provide comprehensive details of these modules in the following subsections.

2.2.1 Multi-Dimensional Analysis Module

This module employs analysis agents to examine a news item from multiple perspectives, generating a multi-dimensional report to support decision making. To this end, we design four kinds of agents: *Linguistic Feature Analysis Agent*, *Comment Analysis Agent*, *Fact-Checking Agent Group*, and *Questioning Agent*. Each agent (or agent group) focuses on a specific aspect of the news item, collectively providing a comprehensive analysis report.

Linguistic Feature Analysis Agent. This agent analyzes linguistic features of the news content, such as emotional tone and writing style, generating a *linguistic feature analysis report* R_l. Specifically, we design a system prompt P_l to guide the LLM in analyzing linguistic features of the news, producing the report R_l as $R_l = \text{LLM}(P_l, x)$. The blue dashed box in Figure 1 presents a simplified linguistic feature analysis report, identifying an exaggerated tone in the news content.

Comment Analysis Agent. This agent analyzes comments to identify commenters' stances, emotional attitudes, and evidence information. It generates a *comment analysis report* R_c that summarizes commenters' reactions and factual evidence while counting their opinion distribution: $R_c = \text{LLM}(P_c, x, c)$, where P_c is the system prompt for Comment Analysis Agent. The orange dashed box in Figure 1 offers a simplified view of the generated comment analysis report, which quantifies the distribution of commenters' opinions and presents fact evidence.

Fact-Checking-Agent Group. This agent group uses external facts to verify the authenticity of news. It primarily consists of two agents: a *Fact-Questioning Agent* and a *Fact-Checking Agent*.

The Fact-Questioning Agent generates yes/no questions based on claims in the news content. The fact question set Q_f is generated as $Q_f = \text{LLM}(P_{Q_f}, x)$, where P_{Q_f} is the system prompt for Fact-Questioning Agent. Then, Q_f serve as queries to retrieve relevant clues from Google.

The Fact-Checking Agent combines clues retrieved from Google and facts gathered via the Wikipedia tool to collect an evidence set e. Subsequently, it evaluates the consistency between claims in news content and e. Based on this evaluation, it generates a fact-checking analysis report R_f to identify misleading claims: $R_f = \text{LLM}(P_f, x, e)$, where P_f is the system prompt for Fact-Checking Agent. The green dashed box in Figure 1 presents an example of the generated fact-checking analysis report, which highlights the inconsistency between claims in news content and the evidence.

Questioning Agent. To ensure sufficient analysis, we introduce a question-reflection mechanism. It uses a Questioning Agent to review the above-mentioned analysis reports, so as to identify any previously overlooked aspects. Then it generates specific questions to guide these analysis agents in conducting more in-depth and comprehensive analysis. Formally, the generation processes of these question sets are described as

$$
\begin{aligned}
Q_r^l &= \text{LLM}(P_q, x, R_l), \\
Q_r^c &= \text{LLM}(P_q, x, c, R_c), \\
Q_r^f &= \text{LLM}(P_q, x, e, R_f),
\end{aligned}
$$

where Q_r^l, Q_r^c, Q_r^f represents the question sets for the linguistic feature analysis, comment analysis, and fact-checking analysis report, respectively. P_q is the system prompt for Questioning Agent.

The above question sets are respectively fed into the Linguistic Feature Analysis Agent, Comment Analysis Agent, and Fact-Checking Agent, enabling them to perform more comprehensive and in-depth analyses. Then, each agent produces its individual response. Finally, we integrate the three analysis reports and these responses into a unified multi-dimensional analysis report, which serves as a reliable basis for evaluating news authenticity. The system prompts for the Multi-Dimensional Analysis Module are provided in Appendix B.1.

2.2.2 Decision Rule Optimization Module

In this module, we design cross-domain verification tasks and use the module to perform them. Subsequently, we optimize decision rules based on feedback from these executions to improve their generalization across domains.

Cross-Domain Validation Tasks Construction. We construct cross-domain validation tasks using news from different source domains. As illustrated in Figure 2, we first randomly sample a piece of source-domain news as the query news, and randomly select other source-domain annotated news as the demonstration news. The query news, along with its multi-dimensional analysis report and demonstration news, are then input into a

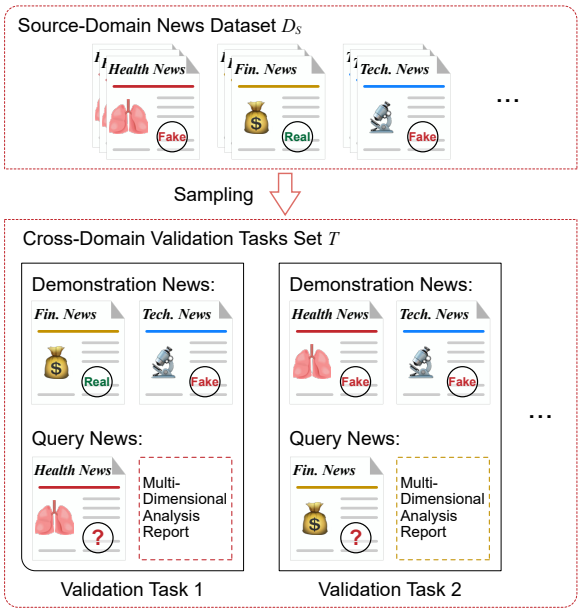

Figure 2: An illustration of constructing cross-domain validation tasks.

Judge Agent in the form of in-context learning. Finally, the Judge Agent evaluates the query news and its analysis report, using the demonstration news and the decision rule to judge its truthfulness. To ensure the diversity of validation tasks, we sequentially sample query news from each source domain, thereby creating a set of cross-domain validation tasks $T = \{t_1, t_2, \ldots, t_{N_{ct}}\}$, where N_{ct} denotes the total number of cross-domain validation tasks.

Decision Rule Optimization. To optimize the decision rules, we introduce a *Decision Rule Optimization Agent*, which refines decision rules based on the feedback obtained from Judge Agent's execution on the cross-domain validation task set. As illustrated in Algorithm 1, we first manually define a decision rule r_0. Using r_0, the Judge Agent executes cross-domain validation task set T to produce judgements. These judgements are compared with the ground-truth labels to obtain an accuracy score s_0. Subsequently, we add $\langle r_0, s_0 \rangle$ to L_{RS}, a set designed to store $\langle$decision rule, accuracy$\rangle$ pairs (**Lines 1-2**). Furthermore, $\langle r_0, s_0 \rangle$ is added to the optimization trajectory used in the Decision Rule Optimization Agent's prompt P_o (**Line 3**), which is provided in Appendix B.2.

We design an iterative optimization process to progressively enhance the generalizability of generated decision rules (**Lines 6-16**). During each iteration, the Decision Rule Optimization Agent first generates a new decision rule r_i, which is then applied by the Judge Agent to the cross-domain

Algorithm 1: Decision Rule Optimization

Input:
T: cross-domain validation task set
r_0: manually defined initial decision rule
N_{iter}: the maximum number of iterations
N_{att}: the maximum number of attempts
K: the number of returned decision rules

1 The Judge Agent utilizes r_0 to execute T, obtaining the accuracy s_0
2 $L_{RS} \leftarrow L_{RS} \cup \langle r_0, s_0 \rangle$
3 Add $\langle r_0, s_0 \rangle$ to the optimization trajectory
4 $r_{best}, s_{max} \leftarrow r_0, s_0$
5 $n_{iter}, n_{att} \leftarrow 0$
6 **while** $n_{iter} < N_{iter}$ ***and*** $n_{att} < N_{att}$ **do**
7 $n_{iter} = n_{iter} + 1$
8 The Decision Rule Optimization Agent generates a new decision rule r_i
9 The Judge Agent utilizes r_i to execute T, obtaining the accuracy s_i
10 **if** $s_i > s_{max}$ **then**
11 $L_{RS} \leftarrow L_{RS} \cup \langle r_i, s_i \rangle$
12 $r_{best}, s_{max} \leftarrow r_i, s_i$
13 $n_{att} \leftarrow 0$
14 **else**
15 $n_{att} = n_{att} + 1$
16 **end**
17 Use the top 10 $\langle$decision rule, accuracy$\rangle$ pairs in L_{RS} to construct the optimization trajectory in P_o
18 **end**
19 **return** top K decision rules

validation task set T (**Lines 8-9**). If s_i exceeds s_{max}, the pair $\langle r_i, s_i \rangle$ is added to L_{RS}, and we update the best decision rule r_{best}, the maximum accuracy s_{max} with r_i and s_i (**Lines 11-12**). Next, we select the top 10 $\langle$decision rule, accuracy$\rangle$ pairs from L_{RS} to update the optimization trajectory in P_o (**Line 17**). This enables the Decision Rule Optimization Agent to iteratively refine decision rules, ultimately achieving higher accuracy. Through this process, we expand L_{RS} until reaching the maximum iteration limit N_{iter} or failing to surpass s_{max} for N_{att} consecutive iterations (**Line 6**). Finally, the Decision Rule Optimization Module outputs the top K decision rules from L_{RS} (**Line 19**).

2.2.3 Inference

During inference, the news and its multi-dimensional analysis report are provided to the Judge Agent, which evaluates the input using each

	Method	Disasters		Entertain		Health		Politics		Society	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
NN-based	UCD-RD (Ran and Jia, 2023)	70.26	69.94	56.05	56.57	70.9	71.35	62.19	61.78	61.09	60.95
	CADA (Li et al., 2023)	73.26	72.75	58.24	58.05	70.3	70.05	64.33	65.07	59.82	58.62
	ADAF (Karisani and Ji, 2024)	73.54	72.39	57.19	56.95	70.5	69.91	62.82	61.94	61.19	61.88
LLM-based	GPT-3.5 w/ tools	72.35	72.19	60.26	59.91	68.8	68.05	63.42	62.94	61.69	60.27
	HiSS (Zhang and Gao, 2023)	72.79	72.06	57.56	56.87	72.7	72.37	67.96	66.34	62.64	61.07
	SAFE (Wei et al., 2024)	71.84	70.97	60.75	60.37	71.9	70.07	65.04	64.32	61.67	60.28
	TELLER (Liu et al., 2024a)	75.28	74.67	60.28	60.57	75.2	74.86	65.18	64.97	63.57	63.87
	DELL (Wan et al., 2024)	75.26	74.05	65.67	64.95	76.1	75.81	67.59	66.95	63.82	63.39
	DeepSeek-R1 (Guo et al., 2025)	76.41	85.73	57.16	54.86	69.5	76.3	72.15	**80.11**	66.89	**72.19**
	RAEmo (Liu et al., 2024e)	78.29	78.84	61.51	60.37	77.3	76.87	68.74	70.87	64.78	65.06
	MARO (ours)	**82.98**	**88.15**	**67.54**	**65.9**	**81.9**	**82.37**	**74.97**	79.38	**69.96**	71.97

	Method	Education		Finance		Military		Science		Avg.	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
NN-based	UCD-RD (Ran and Jia, 2023)	60.84	60.74	60.11	59.89	69.05	69.47	57.58	57.32	63.12	63.11
	CADA (Li et al., 2023)	64.31	63.82	61.15	60.83	69.37	70.14	59.31	59.14	64.45	64.27
	ADAF (Karisani and Ji, 2024)	65.54	64.32	62.05	61.16	66.28	65.16	59.16	58.49	64.25	63.58
LLM-based	GPT-3.5 w/ tools	65.96	65.79	62.15	61.61	67.41	66.27	60.16	59.65	64.69	64.08
	HiSS (Zhang and Gao, 2023)	64.84	64.15	63.95	62.89	68.63	67.84	55.91	55.37	65.22	64.33
	SAFE (Wei et al., 2024)	64.95	64.12	60.56	60.13	68.21	68.14	57.73	56.65	64.74	63.89
	TELLER (Liu et al., 2024a)	67.79	67.08	65.06	65.27	71.05	70.39	60.05	59.89	67.05	66.84
	DELL (Wan et al., 2024)	69.05	68.31	62.65	63.49	67.26	66.86	59.76	58.14	67.46	66.88
	DeepSeek-R1 (Guo et al., 2025)	62.19	72.07	57.26	55.26	73.82	79.04	57.38	62.17	65.87	70.86
	RAEmo (Liu et al., 2024e)	69.26	70.73	64.25	63.53	72.86	71.49	60.63	60.17	68.62	68.66
	MARO (ours)	**74.79**	**75.1**	**71.48**	**67.97**	**81.12**	**84.97**	**66.66**	**63.82**	**74.6**	**75.51**

Table 1: Performance comparison between MARO and baselines on Weibo21 using GPT-3.5-turbo-0125 as the underlying model. NN-based denotes conventional neural network-based methods. GPT-3.5 w/ tools means we enable GPT-3.5-turbo-0125 to make independent judgments using the search engine and the Wikipedia tool. The best result in each column is marked in **bold** and the second best is underlined. All results are reported as percentages.

of the top K optimized decision rules. The final judgement is determined by majority voting.

3 Experiments

3.1 Setup

Datasets. We conduct experiments on the Weibo21 (Nan et al., 2021) and AMTCele (Liu et al., 2024e) datasets. Weibo21 is a Chinese multi-domain rumor detection dataset covering 9 domains, where each news item includes news content and several comments. AMTCele, constructed by Liu et al. (2024e), is an English fake news detection dataset covering 7 domains. In this dataset, each news item contains only news content. Further details are provided in Appendix C.

Baselines. We compare MARO with two kinds of baselines: 1) **conventional neural networks based methods**: UCD-RD (Ran and Jia, 2023), CADA (Li et al., 2023) and ADAF (Karisani and Ji, 2024); 2) **LLM-based methods**: HiSS (Zhang and Gao, 2023), SAFE (Wei et al., 2024), TELLER (Liu et al., 2024a), DELL (Wan et al., 2024), DeepSeek-R1 (Guo et al., 2025) and RAEmo (Liu et al., 2024e). Appendix D provides a detailed description of these baselines.

Settings and Evaluation. To ensure fair comparisons, we use the same underlying models to construct MARO and LLM-based baselines. Particularly, we set the temperature of the Decision Rule Optimization Agent to 1 to encourage greater diversity in outputs, and set the temperature of the Judge Agent to 0 for consistent outputs. In our experiments, we conduct 8-fold cross-validation on Weibo21 and 6-fold cross-validation on AMTCele, setting the cross-domain validation task number N_{vt} to 500 for Weibo21 and 400 for AMTCele, with results shown in Appendix E. For both datasets, we empirically set the number of samples for each source domain to 100 on Weibo21 and 80 on AMTCele, the maximum iteration number N_{iter} to 500 for Weibo21, the maximum attempt number N_{att} to 10, and the returned decision rule number K to 3. Finally, we use accuracy (Acc.) and F1-score (F1) as evaluation metrics.

Method		Biz		Edu		Cele		Entmt	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
NN-based	UCD-RD (Ran and Jia, 2023)	73.52	73.29	64.21	63.85	62.2	61.93	61.57	60.21
	CADA (Li et al., 2023)	74.33	74.62	66.98	66.55	62	60.63	60.95	59.94
	ADAF (Karisani and Ji, 2024)	78.62	77.85	70.82	70.71	63.8	62.83	62.82	62.95
LLM-based	GPT-3.5 w/ tools	80.17	80.51	72.19	71.07	64.6	62.06	62.12	58.01
	HiSS (Zhang and Gao, 2023)	77.13	77.48	72.57	71.06	66.4	66.79	62.58	61.84
	SAFE (Wei et al., 2024)	79.26	78.64	72.51	72.27	63.8	62.11	63.56	63.13
	TELLER (Liu et al., 2024a)	82.21	81.38	73.27	73.85	67.6	65.28	63.91	63.64
	DELL (Wan et al., 2024)	83.57	82.94	74.13	73.72	65.2	64.35	62.54	61.49
	DeepSeek-R1 (Guo et al., 2025)	82.5	81.57	71.25	74.15	65	65.34	63.75	61.33
	RAEmo (Liu et al., 2024e)	78.76	77.16	69.28	68.07	61	59.27	61.13	60.21
	MARO (ours)	**85.46**	**84.83**	**77.62**	**77.24**	**68.8**	**67.95**	**66.81**	**65.97**

Method		Polit		Sport		Tech		Avg.	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
NN-based	UCD-RD (Ran and Jia, 2023)	66.25	66.35	63.56	62.79	73.26	73.39	66.37	65.97
	CADA (Li et al., 2023)	68.41	68.92	63.82	62.91	72.19	73.05	66.95	66.66
	ADAF (Karisani and Ji, 2024)	71.72	71.45	71.72	71.24	72.73	72.42	70.32	69.92
LLM-based	GPT-3.5 w/ tools	71.07	73.71	72.72	70.51	74.45	75.28	71.05	70.16
	HiSS (Zhang and Gao, 2023)	71.66	70.95	74.32	73.43	72.54	71.21	71.03	70.39
	SAFE (Wei et al., 2024)	74.51	74.76	70.75	69.63	76.51	75.86	71.56	70.91
	TELLER (Liu et al., 2024a)	73.57	72.29	75.24	75.51	76.11	75.65	73.13	72.51
	DELL (Wan et al., 2024)	75.26	75.18	**79.82**	78.56	77.63	76.41	74.02	73.24
	DeepSeek-R1 (Guo et al., 2025)	65	65.85	71.25	68.49	71.25	70.12	70.00	69.55
	RAEmo (Liu et al., 2024e)	73.55	72.58	71.15	70.09	70.89	69.05	69.39	68.06
	MARO (ours)	**78.93**	**79.73**	79.65	**79.34**	**82.86**	**82.47**	**77.16**	**76.79**

Table 2: Performance comparison between MARO and the baselines on AMTCele.

3.2 Main Results

Tables 1 and 2 present experimental results on Weibo21 and AMTCele[2]. Overall, MARO achieves the best performance across most domains on both datasets. On Weibo21, MARO outperforms the second-best method, RAEmo, by 5.98 in average accuracy and 6.85 in average F1. On AMTCele, MARO surpasses the second-best method, DELL, by 3.14 in average accuracy and 3.55 in average F1. These results demonstrate the effectiveness of MARO in cross-domain misinformation detection.

3.3 Further Analysis

Ablation Study. To verify the contributions of different components in MARO, we report the performance of MARO when these components are removed separately. Here, the components we considering include the Linguistic Feature Analysis Agent, the Comment Analysis Agent, the Fact-Checking-Agent Group, the Questioning Agent, the Cross-Domain Validation Tasks, and the Decision Rule Optimization Agent. To facilitate the subsequent descriptions, we name the variants of MARO removing different components as *w/o* LFAA, *w/o* CAA, *w/o* FCAG, *w/o* QA, *w/o* CDVT and *w/o* DROA, respectively.

	Weibo21		AMTCele	
	Acc.	F1	Acc.	F1
MARO	**74.60**	**75.51**	**77.16**	**76.79**
w/o LFAA	72.11	73.39	72.96	72.41
w/o CAA	71.65	72.34	-	-
w/o FCAG	72.38	73.56	72.62	71.83
w/o QA	72.56	73.48	74.26	73.95
w/o CDVT	70.21	71.75	73.27	72.86
w/o DROA	69.47	71.62	72.18	71.75

Table 3: Ablation studies.

From Table 3, we can clearly find that the removal of these components leads to a performance drop, indicating the effectiveness of these components. In particular, the performance of *w/o* QA shows a noticeable decline. This demonstrates that single-pass analysis is inadequate, while also proving that the question-reflection mechanism we proposed helps in identifying misinformation.

Impact of Source Domain Number. In this experiment, we investigate how the number of source domains impacts MARO's performance. We also

[2]Additional experimental results are provided in Appendix, including those of MARO and baselines on other underlying models (Appendix G.3), results on more datasets (Appendix G.4) and efficiency comparison (Appendix A.3).

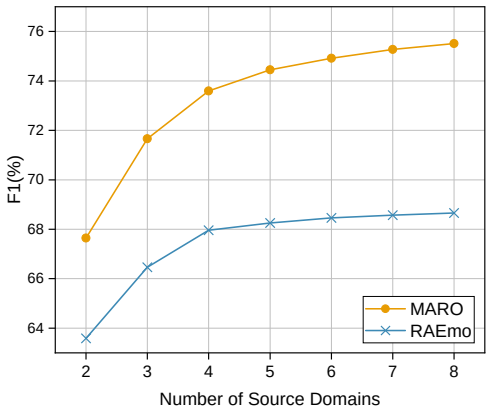

Figure 3: F1 changes with different number of source domains on Weibo21.

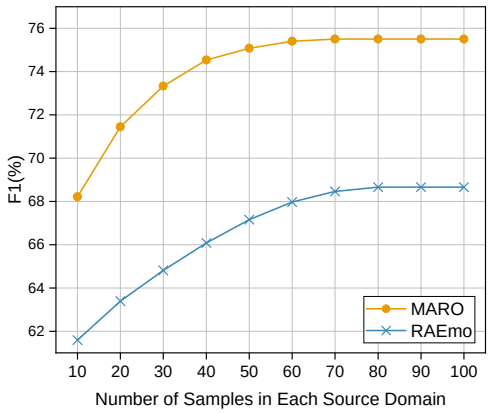

Figure 4: F1 changes with different number of samples in each source domains on Weibo21.

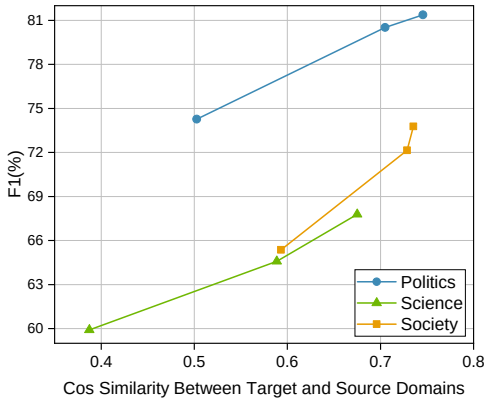

Figure 5: F1 changes with different source-target similarities on the Politics, Science and Society domains.

illustrate the performance of RAEmo, the most competitive baseline, as reported in Table 1.

As shown in Figure 3, increasing the number of source domains improves both methods' performance. This is reasonable because more source domains not only provide diverse feedback to optimize MARO's decision rules, but also enrich RAEmo's demonstration database. We further observe that MARO consistently outperforms RAEmo under different source domain settings, demonstrating MARO's effectiveness.

Impact of Source Domain Sample Number. Then, we investigate how the number of source domain samples affects MARO's performance. To this end, we gradually vary from 10 to 100 with an increment of 10 in each step, and report the corresponding model performance.

As shown in Figure 4, we observe that as the number of source domain samples increases, both MARO and RAEmo show improvements in F1 scores. For this phenomena, we argue that more source-domain samples also provide more comprehensive feedback and similar demonstrations for MARO and RAEmo, respectively. Furthermore, MARO outperforms RAEmo across different numbers of source domain samples, especially in the scenarios of limited samples.

Impact of Domain Similarity. As mentioned previously, MARO is proposed to address cross-domain misinformation detection. Thus, one critical question arises regarding the impact of the similarity between source and target domains on the performance of MARO. To investigate this, we use TF-IDF to calculate the semantic similarity between news from different domains in Weibo21, as illustrated by the similarity matrix in Appendix F. We sample *Politics*, *Science*, and *Society* as target domains, and pair the remaining six domains into three groups as source domains. Figure 5 illustrates the relationship between source-target domain similarity and the performance of MARO.

It can be observed from Figure 5 that the performance of MARO reflects a positive correlation with domain similarity. This phenomena is reasonable since similar source domain can provide abundant shared features, which enable the Decision Rule Optimization Agent to generate decision rules that are more effective for the target domain.

4 Case Study

We provide an example of the decision rule optimization process in Appendix H.

5 Related Work

Recently, LLMs have demonstrated impressive performance across a range of tasks (Wang et al., 2025b,a) and have been extensively used for misinformation detection (Huang and Sun, 2023; Zhang and Gao, 2023; Wu et al., 2024; Yue Huang, 2024;

Liu et al., 2024c,d,a; Wei et al., 2024; Liu et al., 2024b; Nan et al., 2024; Wan et al., 2024). For example, Huang and Sun (2023) design prompts tailored to the features of fake news, effectively guiding ChatGPT for misinformation detection. Along this line, Zhang and Gao (2023) and Wei et al. (2024) propose to deconstruct complex claims into simpler sub-statements, which are then verified step-by-step using external search engines. Unlike the above studies, Wu et al. (2024) leverage LLMs to disguise news styles and employ style-agnostic training, thereby improving the robustness of misinformation detection systems against style variations. Liu et al. (2024b) leverage LLMs to extract key information and integrate both the model's internal knowledge and external real-time information to conduct a comprehensive multi-perspective evaluation. To address the problem of scarce comments in the early stages of misinformation spread, Nan et al. (2024) utilize LLMs to simulate users and generate diverse comments. Slightly similar to ours, Wan et al. (2024) propose DELL, which analyzes various aspects of news to assist in identifying misinformation. Despite their effectiveness, these studies mainly concentrate on in-domain detection and have yet to adequately address the challenges of cross-domain detection.

Early approaches to cross-domain misinformation detection (Pérez-Rosas and Mihalcea, 2014; Hernández-Castañeda et al., 2017) rely on handcrafted features and traditional models, leading to limited performance. With the advent of deep learning, researchers explore this task by aligning feature representations across domains (Choudhry et al., 2022) or capturing invariant features (Ran et al., 2023; Ran and Jia, 2023) or reducing inter-domain discrepancies (Lin et al., 2022). Nevertheless, the lack of sufficient cross-domain labeled data limits the effectiveness of these methods. Very recently, Liu et al. (2024e) propose RAEmo, which leverages an emotion-aware LLM to encode source-domain samples and create in-context learning tasks for target-domain misinformation detection. However, RAEmo still relies on manually-designed decision rules for reasoning.

We introduce a multi-dimensional analysis approach within our framework to assist in news veracity evaluation, which has not been explored in previous studies. The one exception is DELL. However, unlike DELL, we introduce a Questioning Agent to facilitate more in-depth and comprehensive analysis. More importantly, compared with studies on LLM-based misinformation detection, such as DELL and RAEmo, we incorporate a decision rule optimization module to automatically optimize decision rules, inspired by (Pryzant et al., 2023; Xu et al., 2023; Yang et al., 2024).

6 Conclusion and Future Work

In this work, we have proposed MARO, a cross-domain misinformation detection framework which addresses two key shortcomings of existing LLM-based methods: inadequate analysis and reliance on manually designed decision rules. First, MARO employs multiple expert agents to analyze news from various dimensions and generate initial analysis reports. Then, a Questioning Agent then reviews each report and poses specific questions to prompt more in-depth and comprehensive analyses. These reports and the agents' responses are aggregated into a multi-dimensional analysis report to assist judgment. Additionally, we propose a decision rule optimization method that automatically refines decision rules based on feedback from cross-domain validation tasks. Compared to state-of-the-art methods, MARO achieves significantly higher accuracy and F1 scores on the commonly used datasets. Ablation studies confirm the effectiveness of each component.

As future work, we plan to incorporate logical reasoning and knowledge graph reasoning to conduct a deeper analysis, and to perform a more comprehensive evaluation of decision rules, thereby providing stronger evidence for their optimization. Moreover, our multi-agent coordination approach shows promising generalization potential and can be applied to other NLP tasks, such as machine translation (Zeng et al., 2019), text generation (Su et al., 2019), and style transfer (Zhou et al., 2020), thus demonstrating its applicability across tasks.

Limitations

Although MARO has demonstrated effectiveness in cross-domain misinformation detection, it may have two limitations. First, MARO's workflow is complex, requiring multiple rounds of iteration to generate effective decision rules, as well as multi-dimensional analysis conducted through multiple agents. Second, the clues gathered via search engines may include misinformation fabricated by malicious actors, which may introduce distortion into the process of judging the authenticity of target-domain news.

Acknowledgement

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular twitter threads. In *SmartCloud*.

Arjun Choudhry, Inder Khatri, Arkajyoti Chakraborty, Dinesh Vishwakarma, and Mukesh Prasad. 2022. Emotion-guided cross-domain fake news detection using adversarial domain adaptation. In *ICNLP*.

Nicholas Della Giustina. 2023. *Misinformation and Its Effects on Individuals and Society from 2015-2023: A Mixed Methods Review Study*. University of Washington.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ching Nam Hang, Pei-Duo Yu, and Chee Wei Tan. 2024. Trumorgpt: Query optimization and semantic reasoning over networks for automated fact-checking. In *CISS*.

Ángel Hernández-Castañeda, Hiram Calvo, Alexander Gelbukh, and Jorge J García Flores. 2017. Cross-domain deception detection using support vector networks. *Soft Computing*.

Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*.

Payam Karisani and Heng Ji. 2024. Fact checking beyond training set. In *NAACL*.

Jingqiu Li, Lanjun Wang, Jianlin He, Yongdong Zhang, and Anan Liu. 2023. Improving rumor detection by class-based adversarial domain adaptation. In *ACM MM*.

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 *NAACL*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *ACL*.

Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024b. Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection. *arXiv preprint arXiv:2407.08952*.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024c. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *IJCAI*.

Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024d. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. *arXiv preprint arXiv:2410.19064*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. 2024e. Raemollm: Retrieval augmented llms for cross-domain misinformation detection using in-context learning based on emotional information. *arXiv preprint arXiv:2406.11093*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *CIKM*.

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 *CIKM*.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *ACL*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Hongyan Ran and Caiyan Jia. 2023. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In *AAAI*.

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

Jinsong Su, Jiali Zeng, Jun Xie, Huating Wen, Yongjing Yin, and Yang Liu. 2019. Exploring discriminative word-level domain contexts for multi-domain neural machine translation. *IEEE transactions on pattern analysis and machine intelligence*.

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 *ACL*.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. 2025a. Litesearch: Efficient tree search with dynamic exploration budget for math reasoning. In *AAAI*.

Ante Wang, Linfeng Song, Ye Tian, Dian Yu, Haitao Mi, Xiangyu Duan, Zhaopeng Tu, Jinsong Su, and Dong Yu. 2025b. Don't get lost in the trees: Streamlining llm reasoning by overcoming tree search exploration pitfalls. *arXiv preprint arXiv:2502.11183*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. In *NeurIPS*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *ACM SIGKDD*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *AAAI*.

Lichao Sun Yue Huang. 2024. Fakegpt: Fake news generation, explanation and detection of large language models. In *WWW*.

Jiali Zeng, Yang Liu, Jinsong Su, Yubing Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. 2019. Iterative dual domain adaptation for neural machine translation. In *EMNLP*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *IJCNLP*.

Chulun Zhou, Liang-Yu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *ACL*.

A Frequently Asked Questions

A.1 Why Adopt a Multi-Agent Framework?

	Avg. Task Coverage	F1
Llama-3.1-8B	0.43	46.16
w/ multi-agent	**0.92**	**53.21**
Llama-3.1-405B	0.57	76.54
w/ multi-agent	**1**	**80.86**
GPT-3.5-0125	0.53	64.08
w/ multi-agent	**0.99**	**71.75**

Table 4: Comparison of task coverage and F1 between a single LLM and the multi-agent framework.

Misinformation detection involves multi-dimensional analysis of news and the integration of these analyses for judgment. Typically, a single LLM is not capable of handling these complex tasks simultaneously. In contrast, a multi-agent framework decomposes the complex task into simpler subtasks, which are then performed by different expert agents. To verify its effectiveness, we compare the task coverage and detection F1 of the multi-agent framework with a single LLM on Weibo21. For the single LLM, we prompt it to conduct linguistic feature analysis, comment analysis, and fact-checking on the news, and then make a judgment based on the analysis results.

As shown in Table 4, the task coverage and detection F1 of the multi-agent framework are both significantly higher than those of the single LLM.

A.2 How does decision rule optimization differs from transfer learning and domain adaptation approaches?

Traditional transfer learning and domain adaptation techniques can be applied to cross-domain misinformation detection. These methods typically improve a model's generalization ability by updating its parameters. However, when training samples are limited and the model has a large scale of parameters, such traditional approaches are often difficult to apply effectively.

In contrast, our proposed decision rule optimization method is well-suited for this scenario. Instead of updating model parameters, we enhance the generalization ability of Judge Agent in cross-domain misinformation detection by searching for and applying the optimal decision rules.

A.3 Efficiency Comparison

	Avg. Token	F1
RAEmo	1125	52.97
MARO (ours)	**1047**	**55.98**

Table 5: Efficiency comparison.

We conduct a computational cost analysis on Weibo21, comparing MARO with the strongest baseline, RAEmo (Liu et al., 2024e). Both methods use Llama-3.1-8B as the underlying model. Specifically, we measure the average number of input tokens required to complete both the training and inference processes, as well as the detection F1. As shown in Table 5, compared to RAEmo, MARO reduces the average token consumption by 6.9% while achieving a 5.7% improvement in F1.

B Prompts

B.1 System Prompts for the Multi-Dimensional Analysis Module

We list the system prompts for the agents in Multi-Dimensional Analysis Module as follows:

Linguistic Feature Analysis Agent
In a multi-agent misinformation detection system, you act as the linguistic feature analysis agent, responsible for conducting an in-depth analysis of the emotional polarity and writing style of the news while generating a linguistic feature analysis report.

Comment Analysis Agent
In a multi-agent misinformation detection system, you act as the comment analysis agent, responsible for conducting an in-depth analysis of commenters' stances and emotional polarity towards the news and identifying fact-checking information within the comments to generate a comment analysis report.

Fact-Questioning Agent
In a multi-agent misinformation detection system, you act as the fact questioning agent, responsible for generating specific yes/no questions based on the statements in the news to assist in determining its authenticity.

Fact-Checking Agent
In a multi-agent misinformation detection system, you act as the Fact-Questioning Agent, responsible for analyzing the consistency between statements in news and factual evidence. You need to invoke the Wikipedia tool and leverage clues from the search engine to retrieve relevant facts relevant to the statements. Then, you need assess the consistency between the statements and the facts, producing a fact-checking analysis report.

Questioning Agent
In a multi-agent misinformation detection system, you act as the Questioning Agent, responsible for reviewing the source content and the analysis report to identify aspects requiring further investigation. Then, you need to pose targeted questions, encouraging the report providers to perform more in-depth and comprehensive analysis.

Domain	Science	Military	Education	Disasters	Politics
Real	143	121	243	185	306
Fake	93	222	248	591	546
All	236	343	491	776	852
Domain	**Health**	**Finance**	**Entertain**	**Society**	**All**
Real	485	959	1000	1198	4640
Fake	515	362	440	1471	4488
All	1000	1321	1440	2669	9128

Table 6: Data Statistics of Weibo21.

Domain	Tech	Edu	Biz	Sport	Polit	Entmt	Cele	All
Legit	40	40	40	40	40	40	250	490
Fake	40	40	40	40	40	40	250	490
All	80	80	80	80	80	80	500	980

Table 7: Data Statistics of AMTCele.

B.2 Prompt for the Decision Rule Optimization Agent

Decision Rule Optimization Agent
You have been provided with a set of decision rules and their corresponding accuracy score. The decision rules are ordered by their accuracy in ascending order, where a higher accuracy represents higher generalizability.

<*decision rule 1, accuracy 1*>
<*decision rule 2, accuracy 2*>
(...more example pairs...)

Below are several examples demonstrating how to apply these decision rules. In each example, replace <DECISION RULE> with your decision rule, read the input carefully, and generate an accurate judgment. If the judgment matches the provided ground-truth label, it is considered correct; otherwise, it is wrong.

Input: [example news]
<DECISION RULE>
Output: fake
(...more examples...)

Now, design a new decision rule that differs from the existing ones and aim to maximize its accuracy.

C Datasets Details

We conduct experiments on the Weibo21 and AMTCele, respectively. The statistical of both datasets are summarized in Tables 6 and 7.

D Baselines

The adopted baselines are listed as follows:

- **UCD-RD** (Ran and Jia, 2023) This method leverages contrastive learning and cross-attention mechanisms to achieve cross-domain rumor detection through feature alignment and domain-invariant feature learning.
- **CADA** (Li et al., 2023) It utilizes category alignment and adversarial training to facilitate cross-domain misinformation detection.
- **HiSS** (Zhang and Gao, 2023) Typically, this approach breaks down complex news content into multiple sub-statements and uses search engines to gather clues, progressively verifying each sub-statement to determine the authenticity of the news.
- **TELLER** (Liu et al., 2024a) It combines neural-symbolic reasoning with logic rules to enhance explainability and generalizability, providing transparent reasoning paths for misinformation detection.
- **ADAF** (Karisani and Ji, 2024) This approach enhances cross-domain fact-checking

by adversarially training the retriever for robustness and optimizing the reader to be insensitive to evidence order, improving overall performance across domains.

- **SAFE** (Wei et al., 2024) The model decomposes news content into independent facts and verifies the authenticity of each fact through multi-step reasoning.

- **DELL** (Wan et al., 2024) It uses LLMs to generate diverse news reactions and interpretable agent tasks, aiming to enhance accuracy and calibration in misinformation detection by integrating expert predictions.

- **DeepSeek-R1** (Guo et al., 2025) It is a reasoning model that integrates multi-stage training and cold-start data.

- **RAEmo** (Liu et al., 2024e) It constructs a sentiment-embedded retrieval database, leveraging sentiment examples from the source domain for in-context learning to verify content authenticity in the target domain.

E Cross-Validation Experiments

To determine the cross-domain validation task number N_{vt}, we conduct 8-fold cross-validation experiments on Weibo21 and 6-fold cross-validation experiments on AMTCele. Through these experiments, we identify $N_{vt} = 500$ as the optimal value for Weibo21 and $N_{vt} = 400$ for AMTCele, with the validation results illustrated in Figure 6.

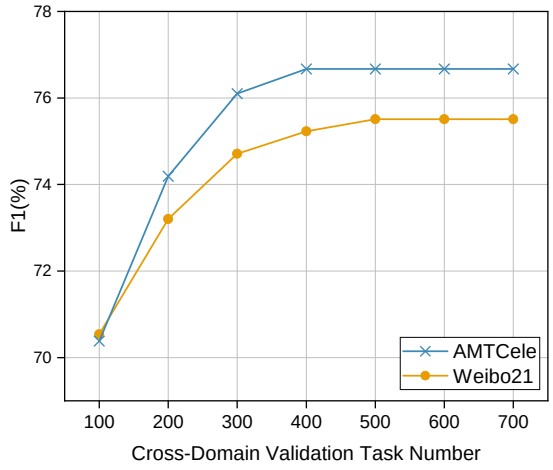

Figure 6: Cross-validation experiments.

F Similarity Matrix

We compute the domain similarity of the Weibo21 dataset using TF-IDF, with the resulting domain similarity matrix visualized in Figure 7.

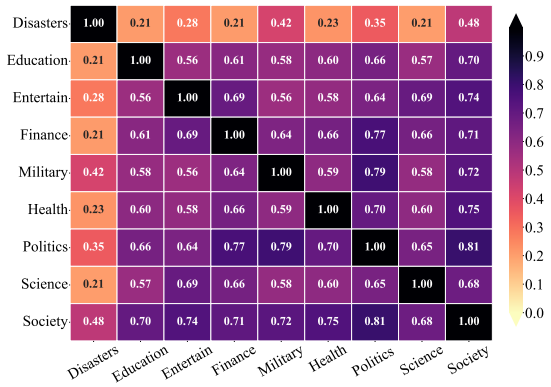

Figure 7: Domain similarity matrix of Weibo21.

G More Results

G.1 Performance Comparison under Significant Source-Target Domain Differences

In order to explore the performance of MARO when there are significant differences between the source and the target domain. According to the domain similarity matrix of Weibo21 (Figure 7), we select Disasters and Education—which are significantly different from all other domains—as the target domains. To avoid experimental redundancy, we choose the three domains with the lowest similarity to each target domain as the source domains (Science, Education, Finance → Disasters; Disaster, Science, Entertain → Education). The experiments are conducted using GPT-3.5-0125 as the underlying model, and the results are shown in Table 8.

Method	Sci, Edu, Fin -> Dis		Dis, Sci, Ent -> Edu	
	Acc.	F1	Acc.	F1
TELLER	72.24	74.17	66.45	66.58
DELL	72.26	73.05	67.24	67.51
RAEmo	76.52	77.31	65.28	66.36
MARO (ours)	**81.86**	**87.65**	**74.16**	**74.83**

Table 8: Performance Comparison under Significant Source-Target Domain Differences

Method	Dis->Edu		Dis->Ent		Edu->Dis		Edu->Ent		Ent->Dis		Ent->Edu		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER	65.79	65.18	59.27	59.16	73.28	73.56	59.82	59.67	73.32	73.75	65.91	65.45	66.23	66.13
DELL	69.05	68.31	65.42	64.28	74.38	73.91	65.52	64.61	74.46	73.95	69.38	68.57	69.7	68.94
RAEmo	65.11	64.39	60.24	60.87	77.34	76.82	61.38	61.76	77.95	76.97	65.57	64.68	67.93	67.58
MARO (ours)	**73.42**	**74.11**	**66.24**	**64.38**	**81.81**	**85.82**	**67.91**	**66.53**	**82.16**	**85.95**	**73.82**	**74.25**	**74.23**	**75.17**

Table 9: Performance comparison between MARO and baselines in single-source, single-target domain scenarios

Method	Disasters		Entertain		Health		Politics		Society	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	78.29	78.05	83.26	**83.28**	80.4	80.58	81.26	80.46	76.59	77.32
DELL (Wan et al., 2024)	81.05	81.26	82.06	81.89	83.1	82.97	84.18	83.89	**78.56**	77.19
RAEmo (Liu et al., 2024e)	84.26	84.49	83.11	82.84	83	82.85	84.75	84.55	77.65	77.49
MARO (ours)	**85.05**	**89.66**	**86.53**	79.7	**86.8**	**87.15**	**88.01**	**90.41**	77.97	**79.61**

Method	Education		Finance		Military		Science		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	72.11	72.12	74.76	75.25	88.19	87.89	72.21	71.19	78.56	78.46
DELL (Wan et al., 2024)	74.64	75.12	**80.91**	79.24	87.98	88.26	71.88	**72.58**	80.48	80.27
RAEmo (Liu et al., 2024e)	75.46	75.88	80.35	**79.75**	88.94	89.05	**72.75**	71.75	81.14	80.96
MARO (ours)	**76.82**	**77.1**	77.76	65.89	**91.44**	**93.39**	72.57	67.66	**82.56**	**81.18**

Table 10: Performance comparison on Weibo21 using LLaMA-3.1-405B as the underlying model.

Method	Biz		Edu		Cele		Entmt	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	83.15	83.19	82.75	82.81	77.6	77.12	77.51	77.57
DELL (Wan et al., 2024)	83.07	83.75	84.18	84.35	77.2	76.82	77.94	78.06
RAEmo (Liu et al., 2024e)	82.51	82.53	84.09	84.15	79.4	79.11	75.15	75.26
MARO (ours)	**86.25**	**86.54**	**86.25**	**86.11**	**81.2**	**80.84**	**78.75**	**78.81**

Method	Polit		Sport		Tech		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	83.69	86.72	82.75	82.89	85.57	85.56	81.79	82.27
DELL (Wan et al., 2024)	84.06	83.91	84.13	84.24	86.41	86.32	82.43	82.49
RAEmo (Liu et al., 2024e)	84.31	84.34	85.11	85.16	87.02	86.91	82.47	82.49
MARO (ours)	**87.52**	**87.51**	**87.35**	**87.42**	**91.25**	**90.87**	**85.49**	**85.44**

Table 11: Performance comparison on AMTCele using LLaMA-3.1-405B as the underlying model.

Method	Disasters		Entertain		Health		Politics		Society	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	85.47	85.16	84.19	84.26	84.8	84.26	77.65	77.42	73.94	73.81
DELL (Wan et al., 2024)	86.51	86.19	86.93	**86.89**	86.4	86.27	79.29	79.34	75.27	75.19
RAEmo (Liu et al., 2024e)	87.65	86.27	**87.49**	85.06	**87.2**	**87.45**	80.85	81.06	76.26	75.84
MARO (ours)	**89.17**	**92.75**	87.43	82.5	86.2	79.08	**87.61**	**91.32**	**78.48**	**78.29**

Method	Education		Finance		Military		Science		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	79.55	78.94	82.97	82.84	83.19	83.13	70.38	70.54	80.24	80.04
DELL (Wan et al., 2024)	**81.95**	81.78	83.46	83.57	83.49	83.25	71.15	71.05	81.61	81.5
RAEmo (Liu et al., 2024e)	81.17	81.36	84.39	**84.16**	85.51	85.34	72.78	72.54	82.58	82.12
MARO (ours)	81.7	**82.62**	**86.15**	79.08	**87.61**	**91.32**	**78.48**	**78.29**	**85.03**	**85.23**

Table 12: Performance comparison on Weibo21 using Claude-3.5-Sonnet as the underlying model.

Method	Biz		Edu		Cele		Entmt	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	86.55	86.27	88.54	88.69	73.2	73.51	79.28	79.62
DELL (Wan et al., 2024)	85.43	85.37	89.28	89.75	75.4	75.19	78.51	78.29
RAEmo (Liu et al., 2024e)	88.25	88.41	91.38	91.05	76.2	76.38	81.79	81.64
MARO (ours)	**91.68**	**91.79**	**92.51**	**92.31**	**79.6**	**79.25**	**85.25**	**84.93**

Method	Polit		Sport		Tech		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	80.26	80.11	84.51	84.63	85.59	85.26	82.56	82.58
DELL (Wan et al., 2024)	81.57	81.95	86.26	86.39	86.47	86.42	83.27	83.34
RAEmo (Liu et al., 2024e)	84.35	85.16	88.26	88.81	88.54	89.06	85.54	85.79
MARO (ours)	**87.65**	**87.29**	**89.67**	**89.75**	**91.69**	**91.81**	**88.29**	**88.16**

Table 13: Performance comparison on AMTCele using Claude-3.5-Sonnet as the underlying model.

Method	Disasters		Entertain		Health		Politics		Society	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	52.98	53.31	51.26	51.54	56.4	56.51	59.52	59.46	41.05	41.18
DELL (Wan et al., 2024)	53.89	53.76	52.54	**52.63**	55.4	55.21	61.55	61.48	48.24	48.51
RAEmo (Liu et al., 2024e)	53.28	53.95	51.39	51.44	58.5	58.33	62.54	62.69	43.56	43.19
MARO (ours)	**60.05**	**69.6**	**59.73**	47.52	**67.8**	**68.49**	**66.03**	**71.41**	**49.96**	**52.42**

Method	Education		Finance		Military		Science		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	39.26	39.28	59.53	59.64	49.37	49.82	52.74	52.68	51.35	51.49
DELL (Wan et al., 2024)	41.24	41.39	60.25	60.17	52.51	52.55	**53.69**	**53.57**	53.25	53.26
RAEmo (Liu et al., 2024e)	40.92	40.79	**61.46**	**62.28**	50.52	50.73	53.25	53.37	52.9	52.97
MARO (ours)	**46.74**	**46.96**	55.14	40.4	**53.98**	**60.6**	52.32	46.44	**56.86**	**55.98**

Table 14: Performance comparison on Weibo21 using LLaMA-3.1-8B as the underlying model.

Method	Biz		Edu		Cele		Entmt	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	51.29	51.42	54.52	54.39	71.4	71.42	50.69	50.73
DELL (Wan et al., 2024)	54.96	54.85	55.21	55.13	72.4	72.19	52.54	52.43
RAEmo (Liu et al., 2024e)	51.63	50.86	53.22	52.59	70.8	70.34	53.24	53.61
MARO (ours)	**56.79**	**56.41**	**57.05**	**56.38**	**76.2**	**75.7**	**57.95**	**57.31**

Method	Polit		Sport		Tech		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	49.56	49.79	57.54	57.36	63.27	63.18	56.9	56.9
DELL (Wan et al., 2024)	50.94	50.78	58.25	58.37	66.74	66.28	58.72	58.58
RAEmo (Liu et al., 2024e)	50.14	50.27	60.85	60.92	61.89	61.42	57.4	57.14
MARO (ours)	**56.17**	**55.05**	**63.29**	**63.51**	**70.26**	**69.65**	**62.53**	**62**

Table 15: Performance comparison on AMTCele using LLaMA-3.1-8B as the underlying model.

G.2 Performance Comparison under single-source, single-target domain

To further explore performance in single-source and target domain scenarios, we randomly select three domains (Disasters, Education, and Entertainment) from Weibo21, forming six source-target domain pairs. We use GPT-3.5-0125 as the underlying model. The experimental results are shown in Table 9.

G.3 More Underlying Models

We replace the underlying models for MARO and the strong baselines with LLaMA-3.1-405B, LLaMA-3.1-8B, and Claude-3.5-Sonnet. As shown in Tables 8-15, MARO's performance remains superior to these baselines across different underlying models, demonstrating its effectiveness.

Events	Charlie Hebdo	Sydney Siege	Ferguson	Ottawa Shooting	Germanwings Crash	All
Rumors	458	522	284	470	238	1972
Non-rumors	1621	699	859	420	231	3830
All	2079	1221	1143	890	469	5802

Table 16: Data Statistics of PHEME.

Method	Charlie Hebdo		Ferguson		Germanwings Crash	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	63.17	62.89	61.56	60.29	58.36	57.82
DELL (Wan et al., 2024)	63.68	63.05	62.18	61.26	59.76	58.82
RAEmo (Liu et al., 2024e)	64.36	63.79	63.76	62.87	61.79	60.68
MARO (ours)	**66.12**	**64.86**	**65.11**	**64.63**	**63.26**	**62.83**

Method	Ottawa Shooting		Sydney Siege		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	58.05	57.92	60.38	59.26	60.3	59.64
DELL (Wan et al., 2024)	59.26	57.08	60.24	58.29	61.02	59.7
RAEmo (Liu et al., 2024e)	59.56	58.32	61.34	60.89	62.16	61.31
MARO (ours)	**61.39**	**61.28**	**62.76**	**61.62**	**63.73**	**63.04**

Table 17: Performance comparison on PHEME using GPT-3.5-turbo-0125 as the underlying model.

Method	Charlie Hebdo		Ferguson		Germanwings Crash	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	63.79	63.26	63.59	62.68	61.37	60.59
DELL (Wan et al., 2024)	64.38	63.75	64.05	62.97	62.66	61.89
RAEmo (Liu et al., 2024e)	64.27	63.85	64.51	**63.26**	62.79	62.11
MARO (ours)	**66.56**	**64.86**	**64.63**	63.03	**63.26**	**63.49**

Method	Ottawa Shooting		Sydney Siege		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	59.52	59.48	60.18	60.05	61.69	61.21
DELL (Wan et al., 2024)	60.32	59.56	60.27	59.38	62.34	61.51
RAEmo (Liu et al., 2024e)	61.29	59.78	60.75	60.08	62.72	61.82
MARO (ours)	**61.39**	**60.06**	**62.76**	**61.62**	**63.72**	**62.61**

Table 18: Performance comparison on PHEME using Claude-3.5-Sonnet as the underlying model.

Method	Charlie Hebdo		Ferguson		Germanwings Crash	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	64.26	63.89	63.72	62.91	60.57	60.08
DELL (Wan et al., 2024)	63.79	62.26	63.57	62.19	62.38	61.57
RAEmo (Liu et al., 2024e)	64.79	63.86	64.35	63.76	61.52	60.58
MARO (ours)	**65.04**	**64.27**	**65.28**	**64.28**	**63.65**	**62.88**

Method	Ottawa Shooting		Sydney Siege		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	61.06	60.72	56.91	56.77	61.30	60.87
DELL (Wan et al., 2024)	61.39	61.28	56.82	56.08	61.59	60.68
RAEmo (Liu et al., 2024e)	63.15	62.36	57.35	57.08	62.23	61.53
MARO (ours)	**64.02**	**63.66**	**59.25**	**58.85**	**63.45**	**62.79**

Table 19: Performance comparison on PHEME using LLaMA-3.1-405B as the underlying model.

Method	Charlie Hebdo		Ferguson		Germanwings Crash	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	67.54	66.35	66.27	66.91	62.28	61.57
DELL (Wan et al., 2024)	68.05	**66.87**	67.54	66.89	61.05	60.59
RAEmo (Liu et al., 2024e)	68.26	63.26	67.37	65.37	61.26	60.85
MARO (ours)	**73.31**	66.52	**70.85**	**67.01**	**63.06**	**63.11**

Method	Ottawa Shooting		Sydney Siege		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1
TELLER (Liu et al., 2024a)	50.25	49.67	49.26	**50.35**	59.12	58.97
DELL (Wan et al., 2024)	51.52	50.67	48.23	47.52	59.28	58.51
RAEmo (Liu et al., 2024e)	49.28	48.05	47.26	46.35	58.69	56.78
MARO (ours)	**52.01**	**51.88**	**50.38**	48.93	**61.92**	**59.49**

Table 20: Performance comparison on PHEME using LLaMA-3.1-8B as the underlying model.

Decision Rule	Acc.
Analyze the credibility of the news outlet and its fact-checking history regarding the social media event. If the news outlet has a history of spreading misinformation, output "1" as fake news; if the news outlet is known for credible reporting, output "0" as real news. Output requirements: - Output format: judgment: <'1' represents fake-news, '0' represents real-news>	55.31
Evaluate the cross-referencing of multiple reliable sources to verify the accuracy and credibility of the information presented in the news item. If the information is corroborated by multiple reputable sources, output "0" as real news; if there are conflicting reports or lack of consensus among sources, output "1" as fake news. Output requirements: - Output format: judgment: <'1' represents fake-news, '0' represents real-news>	62.52
Utilize sentiment analysis and social media monitoring to assess public reactions and discussions surrounding the social media event. If a large portion of the online community expresses skepticism or disbelief in the news item, output "1" as fake news; if the overall sentiment is positive and supportive of the news, output "0" as real news. Output requirements: - Output format: judgment: <'1' represents fake-news, '0' represents real-news>	65.46
Evaluate the linguistic features and narrative structure of the news item to determine the level of bias and sensationalism in the reporting. If the article contains emotionally charged language, subjective opinions presented as facts, or sensationalized headlines, output "1" as fake news; if the article maintains a neutral tone, presents facts objectively, and avoids sensationalism, output "0" as real news. Output requirements: - Output format: judgment: <'1' represents fake-news, '0' represents real-news>	65.68
Examine the consistency of the news item with verified data and expert opinions related to the social media event. If the news item aligns with established facts and expert analysis, output "0" as real news; if the news item contradicts verified data or expert opinions, output "1" as fake news. Output requirements: - Output format: judgment: <'1' represents fake-news, '0' represents real-news>	68.39

Table 21: An example of the decision rule optimization process on Weibo21.

G.4 More Datasets

We also conduct experiments on PHEME (Buntain and Golbeck, 2017), which is an English rumor detection dataset containing posts and comments related to five breaking events. Table 16 shows the statistics of PHEME. Similar to the above experiments, we conduct cross-event misinformation detection experiments on each event. As shown in Tables 17-20, compared with the strong baselines, MARO still achieves the best performance on PHEME, demonstrating its effectiveness.

H Case Study

Table 21 shows an example of the decision rule optimization process. The left side of the table shows the generated decision rules, while the right side shows their validation accuracy. We can observe that decision rules with higher accuracy generally have stronger applicability.