RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning based on Emotional Information

Anonymous ACL submission

Abstract

Misinformation is prevalent in various fields such as education, politics, health, etc., causing significant harm to society. However, current methods for cross-domain misinformation detection rely on time and resources consuming fine-tuning and complex model structures. With the outstanding performance of LLMs, many studies have employed them for misinformation detection. Unfortunately, they focus on in-domain tasks and do not incorporate significant sentiment and emotion features (which we jointly call *affect*). In this paper, we propose RAEmoLLM, the first retrieval augmented (RAG) LLMs framework to address cross-domain misinformation detection using in-context learning based on affective information. It accomplishes this by applying an emotion-aware LLM to construct a retrieval database of affective embeddings. This database is used by our retrieval module to obtain source-domain samples, which are subsequently used for the inference module's in-context few-shot learning to detect target domain misinformation. We evaluate our framework on three misinformation benchmarks. Results show that RAEmoLLM achieves significant improvements compared to the zero-shot method on three datasets, with the highest increases of 20.69%, 23.94%, and 39.11% respectively. This work will be released on Github[1].

1 Introduction

The internet is flooded with misinformation (Scheufele and Krause, 2019), which has a significant impact on people's lives and societal stability (Della Giustina, 2023). Misinformation is pervasive across various domains such as education, health, technology, and especially on the internet, which requires people to invest significant time and effort in discerning the truth (Pérez-Rosas et al., 2018). However, models trained in specific known domains are often fragile and prone to making incorrect predictions when presented with samples from new domains (Saikh et al., 2020). As a result, detecting cross-domain misinformation has become an urgent global issue and poses greater challenges and difficulties.

Although some studies address cross-domain misinformation detection (Comito et al., 2023; Tang et al., 2023; Shi et al., 2023), they require time-consuming fine-tuning, and apply only traditional machine learning methods or complex deep learning methods. Recently, LLMs have achieved impressive results in various tasks through zero-shot, few-shot (Li, 2023), or instruction tuning (Zhang et al., 2023a). Many researchers have applied LLMs to identify misinformation (Li et al., 2023; Hu et al., 2024; Cheung and Lam, 2023). However, these methods perform only in-domain misinformation detection. Moreover, emotions and sentiments (which we jointly call *affect*) are important characteristics of human expression and communication (Hakak et al., 2017). When authors publish misinformation, they often consciously choose specific emotions to capture the attention and resonance of readers to encourage rapid spread (Keen, 2006; Liu et al., 2024d). Unfortunately, there are few LLMs that utilize affective information to detect misinformation, and the only ConspEmoLLM (Liu et al., 2024b) are developed based on an emotional LLM, which does not make full use of affective information, has no cross-domain ability, and also needs time-consuming fine-tuning.

In-context learning (ICL) needs only task instructions and few-shot examples (input-label pairs), eliminating fine-tuning on specific task labels (Dong et al., 2022b). A few studies have used ICL to address cross-domain problems (Long et al., 2023; Wu et al., 2024). To the best of our knowledge, there is currently no application of ICL for cross-domain misinformation detection based on affective information retrieval.

[1]The code and data for review can be found in Software

To address these issues, we propose the first retrieval augmented (RAG) LLMs framework based on emotional information (RAEmoLLM), to address cross-domain misinformation detection using in-context learning based on affective information. RAEmoLLM contains three modules: (1) In the *index construction* module, we apply EmoLLaMA-chat-7B (Liu et al., 2024c) to encode all domain corpora, obtaining implicit embeddings to construct the retrieval database as well as explicit affective labels. We also conduct a comprehensive affective analysis to demonstrate the effectiveness of affective information for discriminating between true and misinformation. (2) The *retrieval* module recommends the top K affect-related examples (text-label pairs) from the source domain corpus according to the target domain content, obtained from the retrieval database. (3) These examples are utilized as the few-shot demonstrations in the *inference* module, which is driven by a prompt template to guide the LLM to verify the target content for misinformation. The template helps combine implicit and explicit affective information.

In this work, we make three main contributions:

- We conduct affective analysis on different kinds of misinformation datasets and construct the retrieval database according to the implicit affective information for misinformation datasets.
- We propose RAEmoLLM, the first framework for cross-domain misinformation detection using ICL based on affective information, which does not require fine-tuning. Experimental results show that RAEmoLLM outperforms the zero-shot methods and the few-shot methods without using affective information.
- We evaluate RAEmoLLM on a variety of misinformation benchmarks, including fake news, rumors, and conspiracy theory datasets. Results show that RAEmoLLM achieves significant improvements compared to the zero-shot method on three datasets, with the highest increases of 20.69%, 23.94%, and 39.11% respectively, which illustrate the effectiveness of RAEmoLLM framework.

2 Related Work

2.1 Misinformation detection

Cross-domain misinformation detection: Comito et al. (2023) propose a deep learning-based architecture able to mitigate this problem by yielding high-level cross-domain features. Tang et al. (2023) design the News Optimal Transport to learn transferable features across domains by aligning the source and target news using Optimal Transport (OT) techniques. Shi et al. (2023) develop a rough-fuzzy graph learning framework that uses representations of cross-domain sample uncertainty structural information, and captures shared general features across domains. But these methods all require complex structures and fine-tuning strategies.

Based on affective information: Emotion and sentiment are important features for misinformation detection (Liu et al., 2024d). Zhang et al. (2023b) combine the use of semantic and sentiment information, along with propagation information for rumor detection. Dong et al. (2022a) design a sentiment-aware hyper-graph attention network for fake news detection. Liu et al. (2024b) develop a conspiracy theory detection LLM by fine-tuning EmoLLaMA (Liu et al., 2024c). Choudhry et al. (2022) utilize emotional information for fake news detection based on an adversarial learning structure. Unfortunately, these works either have complex structural designs or fine-tuned models, which require significant time and computational resources. The RAEmoLLM in this article applies the ICL method based on affective information, which has a simple structure and does not involve fine-tuning.

2.2 In-context learning

Liu et al. (2024a) develop in-context curriculum learning, a simple but helpful demonstration ordering method for ICL that gradually increases the complexity of prompt demonstrations. Xu and Zhang (2024) propose in-context reflection to strategically select demonstrations that reduce the discrepancy between the LLM's outputs and the actual input-output mappings. Long et al. (2023) propose a retrieval-enhanced language model to address cross-domain problems, in which they train language models by learning both target domain distribution and the discriminative task signal simultaneously with the augmented cross-domain in-context examples. Inspired by this work, we propose the RAEmoLLM.

3 Methodology

This section introduces our method of cross-domain misinformation detection, using the *index construction* module, *retrieval* module and *infer-*

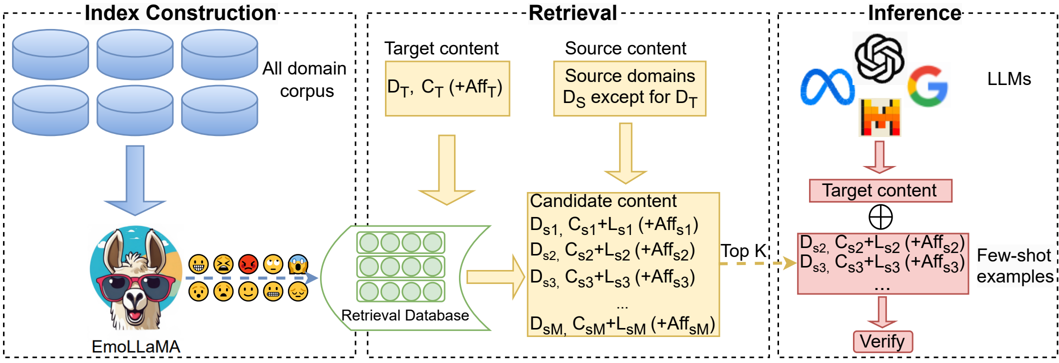

Figure 1: The architecture of RAEmoLLM. D: Domain. T: Target domain. S: Source domain. C: Corpus. L: Label. Aff: Affective information.

ence module. The overall architecture of RAEmoLLM is shown in Figure 1. In the *index construction* module we collect domain datasets, and employ an emotional LLM to obtain embeddings as well as affective labels to conduct a comprehensive affective analysis on them to detect the affective differences between real and false information. The implicit embeddings are adopted to construct the retrieval database. This database is used by the *retrieval* module to obtain source-domain examples. These results are the few-shot examples used by the *inference* module's in-context learning to detect target domain misinformation.

3.1 Index Construction Module

3.1.1 Datasets

We collect FakeNewsAMT (Pérez-Rosas et al., 2018), Celebrity (Pérez-Rosas et al., 2018), PHEME (Kochkina et al., 2018), and COCO (Langguth et al., 2023) datasets. The statistics of these datasets are presented in Table 1. We combine FakeNewsAMT and Celebrity as AMTCele. For AMTCele and PHEME, each domain will take turns being the target domain data. For COCO, we select 3 of 12 domains as target domains, and others as source domains (The details of datasets can be found in Appendix A.1).

3.1.2 Affective Analysis

We firstly conduct a comprehensive affective analysis after collecting datasets. EmoLLaMA-chat-7B, which has the best overall performance among the EmoLLMs (Liu et al., 2024c), is used for affective analysis. EmoLLaMA-chat-7B can be used to extract the following affective dimensions (which we jointly call affect):

(1) *Emotion intensity (EIreg):* For each of four different emotions (anger, fear, joy and sadness), assign a score between 0 and 1 to represent the intensity of emotion of the text;

(2) *Emotion intensity classification (EIoc):* The text can be classified into one of four classes of the intensity of emotion (anger, fear, joy, sadness), i.e. no/low/moderate/high emotional intensity;

(3) *Sentiment (valence) strength (Vreg):* Assign a real-valued score between 0 (most negative) and 1 (most positive) to represent the sentiment intensity of the text.

(4) *Sentiment (valence) classification (Voc):* The text can be categorized into one of seven ordinal classes (i.e. *{very, moderately, slightly} negative, neutral, {slightly, moderately, very} positive*);

(5) *Emotion detection (Ec):* The text can be classified as 'neutral or no emotion' or as one, or more, of eleven given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust).

Obtain implicit and explicit affective information: Following the guidelines of EmoLLMs (Liu et al., 2024c), we add prompts provided by EmoLLMs for each data point in order to obtain vectors from the last hidden layer (i.e., 4096d) for each affective dimension, as well as final labels using EmoLLaMA-chat-7B. We subsequently determine the distribution of affective information in different categories in each dataset.

Explicit affective analysis: Table 2 shows regression information (i.e., EIreg and Vreg) of final labels. The t-value and p-value calculated between *legit/non-rumours/related* and *fake/rumours/conspiracy* demonstrate that there are statistically significant affective differences between the different categories. Figure 3 to Fig-

Table 1: Statistic of datasets

AMTCele			PHEME			COCO		
Domain	Legit	Fake	Events	Rumours	Non-rumours	Topics	Related	Conspiracy
Technology	40	40	Charlie Hebdo	458	1621	Fake Virus		
Education	40	40	Sydney siege	522	699	Harmful Radiation	248	612
Business	40	40	Ferguson	284	859	Depopulation		
Sports	40	40	Ottawa shooting	470	420	Other 9 domains	540	1181
Politics	40	40	Germanwings-crash	238	231	Total	788	1793
Entertainment	40	40	Putin missing	126	112			
Celebrities	250	250	Prince Toronto	229	4			
Total	490	490	Gurlitt	61	77			
			Ebola Essien	14	0			
			Total	2402	4023			

Table 2: Statistics values of EIreg and Vreg on different datasets. The t-test is conducted between *legit/non-rumours/related* and *fake/rumours/conspiracy*.

Datasets	Affective	sub-emotion	legit/non-rumours/related		fake/rumours/conspiracy		t-test	
			mean	var	mean	var	t	p
AMTCele	EIreg	Anger	0.3584	0.0064	0.4055	0.0060	-9.3294	6.91E-20
		Fear	0.3587	0.0137	0.4047	0.0124	-6.2861	4.90E-10
		Joy	0.3392	0.0180	0.2897	0.0142	6.1054	1.48E-09
		Sadness	0.3341	0.0109	0.3697	0.0106	-5.3726	9.70E-08
	Vreg		0.5471	0.0204	0.4940	0.0170	6.0656	1.88E-09
PHEME	EIreg	Anger	0.4547	0.0102	0.4233	0.0075	12.7093	1.44E-36
		Fear	0.5337	0.0170	0.5632	0.0198	-8.5027	2.28E-17
		Joy	0.2134	0.0121	0.1817	0.0133	11.0177	5.58E-28
		Sadness	0.5215	0.0152	0.5177	0.0182	1.1442	0.2526
	Vreg		0.4331	0.0143	0.3842	0.0139	15.9786	2.18E-56
COCO	EIreg	Anger	0.5475	0.0088	0.5641	0.0068	-4.5211	6.43E-06
		Fear	0.5623	0.0097	0.6034	0.0077	-10.5568	1.56E-25
		Joy	0.1800	0.0111	0.1514	0.0075	7.2230	6.66E-13
		Sadness	0.4701	0.0098	0.4773	0.0073	-1.8808	0.0601
	Vreg		0.3961	0.0095	0.3973	0.0066	-0.3325	0.7395

Table 3: Statistics values of cosine similarity between embeddings of different affective information. Top K denotes retrieval top K examples. In addition to Vreg, the results of other affective information are all based on top 4. "A-B" represents the calculation of cosine similarity between each data point in A and each data point in B. Each element (i, j) in the resulting calculation represents the cosine similarity between the i-th vector in the A group embeddings and the j-th vector in the B group embeddings. The top 4 refers to selecting the four highest values from each row. The t-value and p-value represent the t-test results for the "A-B" results of the two lines above.

Datasets	Values	Vreg					Voc	Ec	EIreg				EIoc			
		top 4	top 8	top 16	top 32	top 64			anger	fear	joy	sadness	anger	fear	joy	sadness
AMT	fake-legit	0.791	0.771	0.753	0.736	0.718	0.852	0.812	0.801	0.801	0.801	0.801	0.840	0.840	0.840	0.840
	fake-fake	0.848	0.810	0.783	0.761	0.741	0.894	0.862	0.855	0.855	0.855	0.855	0.885	0.885	0.885	0.885
	t	-22.516	-14.875	-10.951	-8.976	-8.037	-20.550	-22.617	-22.434	-22.433	-22.462	-22.461	-22.260	-22.246	-22.267	-22.244
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	legit-fake	0.787	0.765	0.747	0.729	0.711	0.848	0.807	0.797	0.797	0.797	0.797	0.836	0.836	0.836	0.836
	legit-legit	0.841	0.798	0.768	0.743	0.721	0.886	0.856	0.848	0.848	0.848	0.848	0.877	0.877	0.877	0.877
	t	-21.568	-12.845	-8.052	-5.263	-3.452	-17.138	-21.024	-21.399	-21.387	-21.407	-21.396	-19.364	-19.328	-19.335	-19.315
	p	0.000	0.000	0.001	0.008	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PHEME	nonr-rum	0.930	0.927	0.924	0.921	0.917	0.982	0.952	0.940	0.940	0.940	0.939	0.972	0.972	0.972	0.972
	nonr-nonr	0.957	0.946	0.938	0.932	0.927	0.989	0.971	0.963	0.963	0.963	0.963	0.983	0.983	0.983	0.983
	t	-75.127	-49.017	-35.035	-27.844	-24.327	-69.237	-78.344	-77.082	-77.231	-76.869	-78.103	-71.392	-71.732	-71.005	-72.538
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	rum-nonr	0.935	0.932	0.929	0.925	0.921	0.984	0.957	0.945	0.944	0.945	0.944	0.974	0.974	0.974	0.974
	rum-rum	0.961	0.950	0.942	0.935	0.928	0.990	0.974	0.966	0.966	0.967	0.966	0.984	0.984	0.984	0.984
	t	-58.813	-38.823	-27.206	-19.693	-14.156	-54.654	-58.600	-59.494	-59.637	-59.377	-60.266	-55.874	-56.306	-56.033	-56.759
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
COOC	rela-consp	0.873	0.870	0.866	0.861	0.856	0.955	0.905	0.885	0.885	0.886	0.885	0.936	0.936	0.937	0.936
	rela-rela	0.907	0.887	0.875	0.865	0.857	0.967	0.931	0.916	0.916	0.916	0.916	0.953	0.953	0.954	0.954
	t	-44.603	-23.007	-11.581	-5.437	-2.012	-37.288	-43.522	-44.744	-44.772	-44.253	-44.800	-38.201	-38.337	-37.684	-38.281
	p	0.000	0.093	0.428	0.457	0.312	0.004	0.000	0.000	0.000	0.001	0.000	0.001	0.001	0.002	0.002
	consp-rela	0.863	0.858	0.852	0.846	0.838	0.950	0.897	0.876	0.876	0.877	0.876	0.929	0.929	0.930	0.929
	consp-consp	0.911	0.891	0.878	0.868	0.859	0.968	0.933	0.919	0.919	0.920	0.920	0.954	0.954	0.955	0.954
	t	-74.176	-47.239	-33.132	-25.606	-21.079	-54.114	-69.563	-73.828	-73.876	-73.190	-73.709	-60.255	-60.393	-59.577	-60.204
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

ure 8 in Appendix B confirm that other classifications using affective information are also related to misinformation. However, we can also observe some special cases that cannot effectively distinguish real and false information (e.g. EIreg-sadness in PHEME and COCO, Vreg in COCO). Liu et al. (2024b) also conducted some experiments that demonstrated that simply utilizing explicit affective information does not enhance the model's capability. Therefore, we introduce how to utilize implicit affective information.

Implicit affective analysis: Table 3 shows statistics of different affective embeddings. We perform t-tests on the top-K cosine similarity within categories and the cosine similarity between categories. For example, "fake-legit" denotes computing the cosine similarity between each data point in the "fake" category and each data point in the "legit" category. We then selected the top-K similarity values and performed t-test on them. The t-value and p-value of the top-4 similarity values between "fake-legit" and "fake-fake" are -22.516 and 0, which demonstrates that the top 4 similar data retrieved based on cosine similarity within the "fake" category are highly likely to belong to the same "fake" category. We can see from the results that all affective information leads to the same conclusion in the top-4 scenarios[2]. We also visualize the data distribution reduced to 3 dimensions using PCA in Figures 9 and 10. It can be observed that different categories are clearly separated in the latent space. All the above demonstrates the close relationship between affective information and misinformation.

3.1.3 Retrieval Database Construction

We obtained the implicit affective embeddings in the last step. Due to the high dimensionality of the original vectors, there is a possibility of encountering the curse of dimensionality (Marimont and Shapiro, 1979), which can increase retrieval time. To address this, we employ PCA to reduce the dimensions to 3d, 8d, 16d, 128d, and 512d, respectively (Figure 12 in Appendix D shows the time-F1 trade-off). As a result, we construct a retrieval database comprising vectors of six kinds of dimensions representing affective information.

3.2 Retrieval Module

After constructing the retrieval database, we first obtain the target domain corpus embedding $E_T = [e_{t1}, e_{t2}, ..., e_{tN}]$ and source domain corpus embedding $E_S = [e_{s1}, e_{s2}, ..., e_{sM}]$ through the embedding retrieval database R based on the target domain corpus $D_T = [c_{t1}, c_{t2}, ..., c_{tN}]$ and source domain corpus $D_S = [c_{s1}, c_{s2}, ..., c_{sM}]$. ($N$ and M are the numbers of target domain texts and source domain texts respectively.) Subsequently, we traverse the target domain embeddings E_T and calculate the similarity values with embeddings E_S from source domains using the cosine method. Finally, we select the top k examples based on the similarity values that will be the few-shot examples for LLM inference. Algorithm 1 shows the retrieval process.

Algorithm 1 Retrieval process

Require: Target domain corpus D_T, source domain corpus D_S, retrieval database R.
Ensure: Target domain corpus with top K retrieval examples D_{retri}.

```
E_T ← R(D_T)
E_S ← R(D_S)
for e_t in E_T do
    for e_s in E_S do
        score == cosine(e_t, e_s)
        if score > threshold(0.2) then
            Sco ← score
        end if
    end for
    D_retri ← select top k examples in R(D_S)
    according to Sco
end for
```

3.3 Inference Module

> **Template 1**
> **Task:** *[task prompt]*
> **Target text:** *[input text]*
> **Here are a few examples:** *[examples]*
> **According to the above information, the label of target text:** *[output]*

We apply template 1 to construct the instruction datasets for inference once get top examples for each target domain data. *[task prompt]* denotes the instruction for the task (The different *[task prompts]* for each datasets can be found in Appendix A.2). *[input text]* is a data item from the target domain data. *[examples]* are the retrieval examples (text-label) from source domain data and the *[output]* is the output from LLM.

We also apply template 2 to add explicit affective information. *[affective information]* contains five dimensions described in Section 3.1.2. The format

[2]It should be noted that in Vreg, as the value of K increases, the second p-value in AMTCele and the first p-value in COCO dataset also gradually increase, which may affect the results. Therefore, we choose K to be 4. The analysis of different values of K can be found in Section E.

of *[examples]* is "*Text: [text]. [Affective info]: [value]. The label of text: [label]*". Table 6 shows one complete example.

Template 2
Task: *[task prompt]*
Target text: *[input text]* + *[affective info]*
Here are a few examples retrieved by *[affective info]*: *[examples]*
According to the above information, the label of target text: *[output]*

4 Experiments

4.1 Base Models

We apply ChatGPT (gpt-3.5-turbo-0125), GPT-4o[3], LLaMA3-8b-Instruct[4], Gemma-instruct-(2b, 7b) (Team et al., 2024), Mistral-7b-Instruct (Jiang et al., 2023) and Vicuna-(7b, 13b, 33b) (Chiang et al., 2023) as base models to test our methods. We also select BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) as fine-tuning baselines. Specifically, one domain is selected as the target domain, source domains are used as the training dataset to fine-tune. All open-sourced models use one Nvidia Tesla A100 GPU (80G) for inference.

4.2 Evaluation Metric

Misinformation detection is typically regarded as a classification task, therefore we employ a variety of metrics—Accuracy, Precision, Recall, and F1 for evaluation (Su et al., 2020) (All metrics use the weighted variant).

4.3 Results

4.3.1 Results of different LLMs on the data retrieved based on Vreg

We firstly select the instruction data based on Vreg to test the effectiveness of the RAEmoLLM framework on different LLMs. The results in Table 4 represent the best-performing results of retrieval using different dimensions of each model. The result is the overall performance, which means that in AMTCele and PHEME, every domain is considered as the target domain test set, and the overall result is the performance of the combination of each domain test set. For Gemma series and Vicuna series, we only show the best-performing one in the table. In the content of this section, we will be discussing results exclusively based on the F1 score.

[3]https://openai.com/
[4]https://llama.meta.com/llama3/

Comparison with zero-shot method and random sample examples method without using affective information (random few-shot): From Table 4, we can observe that the RAEmoLLM framework increases the LLMs with zero-shot method largely in most cases. The largest increase in AMTCele is gemma2b (+20.69%), in PHEME is llama3-8b (+23.94%), and in COCO is Vicuna7b (+39.11%). The random few-shot method also performs lower than the RAEmoLLM in most cases. Moreover, in some cases, this method decreases the performance of the model compared to the zero-shot method. A special case is that in the AMTCele dataset, ChatGPT's and GPT-4o's performance is reduced by adding few-shot examples. One possible reason is that the AMTCele dataset is collected from fact-checking websites, and ChatGPT's and GPT-4o's training set includes these data and can effectively utilize this information.

Comparison with fine-tuning method: We can observe that most LLMs with RAEmoLLM framework outperform fine-tuned RoBERTa and BERT on AMTCele and COCO datasets, but it slightly underperforms fine-tuned models in the PHEME dataset. One possible reason is that in cross-domain misinformation detection tasks, the fine-tuning method may perform better for simple short-text discrimination problems (e.g. PHEME). However, the fine-tuning methods may not be suitable for long texts (e.g. AMTCele) or complex tasks (e.g. intent recognition in COCO).

Comparison between approaches utilizing different affective information: We can see LLMs with explicit affective information (i.e. LLMs-Vreg-expl) exceed LLM without explicit affective information (i.e. LLMs-Vreg) in many cases. However, in the case of the llama3-8B and a few exceptional scenarios, it is the opposite. And applying only explicit Vreg information (i.e. random-addexpl) has little effect in most cases. This demonstrates that utilizing explicit affective information places higher demands on the model, requiring it to be able to handle not only the main text but also pay attention to affective cues.

Table 7 and Table 8 in Appendix C present the performance of mistral7b on each domain on AMTCele and PHEME separately. It can be observed that mistral7b with RAEmoLLM framework overtakes mistral with zero-shot in most domains except for prince, gurlitt, and ebola domains in PHEME, which are significant imbalanced data.

Table 4 shows that mistral7b has the best per-

Table 4: Overall results on three datasets. 'zs' denotes the zero-shot method. 'random' denotes randomly sample four examples without using affective information. 'random-addexpl' denotes adding explicit Vreg information for the randomly sample examples. 'Vreg' denotes retrieving four examples based on Vreg information using Template 1. 'Vreg-addexpl' denotes adding explicit Vreg information using Template 2.

	AMTCele				PHEME				COCO			
Model	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
BERT	0.5414	0.5453	0.5414	0.5322	0.7214	0.7203	0.7214	0.7208	0.7288	0.7510	0.7288	0.6356
RoBERTa	0.5678	0.7228	0.5678	0.4730	0.7199	0.7213	0.7199	0.7204	0.7328	0.7851	0.7328	0.6388
mistral7b-zs	0.7194	0.7417	0.7194	0.7127	0.5910	0.6460	0.5910	0.5956	0.4453	0.7081	0.4453	0.5391
mistral7b-random	0.7071	0.7762	0.7071	0.6876	0.6193	0.6326	0.6193	0.6238	0.6663	0.7429	0.6663	0.7025
mistral7b-random-addexpl	0.6357	0.7051	0.6357	0.6021	0.5788	0.6137	0.5788	0.5856	0.6360	0.7254	0.6360	0.6765
mistral7b-Vreg	0.7510	0.7801	0.7510	0.7444	0.6761	0.6825	0.6761	0.6786	0.7395	0.7966	0.7395	0.7667
mistral7b-Vreg-addexpl	**0.7786**	**0.7881**	**0.7786**	**0.7767**	**0.6942**	**0.6928**	**0.6942**	**0.6935**	**0.7488**	**0.7978**	**0.7488**	**0.7725**
gemma2b-zs	0.4592	0.4469	0.4592	0.4259	0.3988	0.4988	0.3988	0.3742	0.3314	0.4578	0.3314	0.3845
gemma2b-random	0.4980	0.4997	0.4980	0.4649	0.4268	0.5797	0.4268	0.3573	0.4477	0.6336	0.4477	0.4816
gemma2b-random-addexpl	0.4918	0.4918	0.4918	0.4916	0.5914	0.5777	0.5914	0.5820	0.6221	0.6164	0.6221	0.5587
gemma2b-Vreg	**0.6367**	0.6429	**0.6367**	**0.6328**	0.4643	0.5528	0.4643	0.4503	0.5291	**0.7333**	0.5291	0.5810
gemma2b-Vreg-addexpl	0.6184	**0.6533**	0.6184	0.5953	**0.6039**	**0.5915**	**0.6039**	**0.5953**	**0.6767**	0.6932	**0.6767**	**0.5990**
llama3-8b-zs	0.5143	0.5256	0.5143	0.4538	0.3757	0.6409	0.3757	0.2085	0.7198	0.7661	0.7198	0.6427
llama3-8b-random	0.6235	0.6479	0.6235	0.6073	0.3788	0.5440	0.3788	0.2242	0.6686	0.6865	0.6686	0.6270
llama3-8b-random-addexpl	0.5684	0.5685	0.5684	0.5682	0.3745	0.6616	0.3745	0.2051	0.6802	**0.7718**	0.6802	**0.6982**
llama3-8b-Vreg	**0.6408**	**0.6824**	**0.6408**	**0.6191**	**0.4682**	**0.5778**	**0.4682**	**0.4479**	**0.7151**	0.7653	**0.7151**	0.6683
llama3-8b-Vreg-addexpl	0.5786	0.5788	0.5786	0.5783	0.3855	0.6823	0.3855	0.2317	0.4488	0.8132	0.4488	0.5551
ChatGPT-zs	**0.7541**	**0.7649**	**0.7541**	**0.7515**	0.5835	0.6483	0.5835	0.5864	0.7709	0.7938	0.7709	0.7245
ChatGPT-random	0.7082	0.7459	0.7082	0.6965	0.6112	0.6494	0.6112	0.6173	0.7651	0.7808	0.7651	0.7152
ChatGPT-random-addexpl	0.7112	0.7321	0.7112	0.7046	0.5986	0.6604	0.5986	0.6022	0.7558	0.7660	0.7558	0.7002
ChatGPT-Vreg	0.6939	0.7578	0.6939	0.6736	**0.6539**	0.6622	**0.6539**	**0.6570**	0.8012	0.8091	0.8012	0.7743
ChatGPT-Vreg-addexpl	0.6939	0.7326	0.6939	0.6806	0.6224	**0.6795**	0.6224	0.6266	**0.8047**	**0.8114**	**0.8047**	**0.7814**
GPT4o-zs	**0.8929**	**0.9003**	**0.8929**	**0.8924**	0.6218	0.6633	0.6218	0.6275	0.8209	0.8482	0.8209	0.7951
GPT4o-random	0.8796	0.8928	0.8796	0.8786	0.6511	0.6925	0.6511	0.6564	0.8547	0.8704	0.8547	0.8394
GPT4o-random-addexpl	0.8612	0.8819	0.8612	0.8593	0.6134	0.6641	0.6134	0.6184	0.8628	0.8690	0.8628	**0.8523**
GPT4o-Vreg	0.8765	0.8894	0.8765	0.8755	0.6718	0.7076	0.6718	0.6770	0.8593	0.8696	0.8593	0.8467
GPT4o-Vreg-addexpl	0.8878	0.8984	0.8878	0.8870	**0.6962**	**0.7310**	**0.6962**	**0.7011**	**0.8640**	**0.8745**	**0.8640**	0.8521
Vicuna-7b-zs	0.5306	0.5361	0.5306	0.5333	0.4931	0.5950	0.4931	0.4791	0.3337	0.5995	0.3337	0.3382
Vicuna-7b-random	0.5735	0.5789	0.5735	0.5760	0.4143	0.5828	0.4143	0.3317	0.6709	0.6159	0.6709	0.6033
Vicuna-7b-random-addexpl	0.5653	0.6030	0.5653	0.5349	0.5348	0.5764	0.5348	0.5458	0.6651	0.5754	0.6651	0.5766
Vicuna-7b-Vreg	0.6031	0.6217	0.6031	0.6017	0.4585	**0.6267**	0.4585	0.4070	**0.7547**	**0.7560**	**0.7547**	**0.7293**
Vicuna-7b-Vreg-addexpl	**0.6082**	**0.6346**	**0.6082**	**0.6088**	**0.5866**	0.6034	**0.5866**	**0.5947**	0.6826	0.7483	0.6826	0.6961

formance among open-sourced LLMs. We choose mistral7b to conduct the following experiments.

4.3.2 Results on the data retrieved based on different affective information

Figure 2 presents the best results of retrieval with each affective dimension embedding (Figure 11 shows all the F1 scores of mistral7b with four few-shot examples using different affective dimension information). It is evident that using *affect-addexpl* method can improve compared to solely relying on implicit affective information retrieval (i.e. Vreg, EIreg). However, when using affective classification information, adding explicit affective information may confuse the model (e.g. Voc, EIoc in AMTCele and PHEME). In the COCO dataset, all the *affect-addexpl* method outperforms *affect* except for EIreg-anger. Regarding the *affect-addexpl* method, in AMTcele, we can see the results retrieval based on Vreg are best, followed by EIreg-joy and EIreg-sadness. And the final three rankings are retrieved based on Voc, EIoc-anger, and Ec. It seems that affective intensity and strength are more suitable for cross-domain fake news detection tasks. In PHEME, retrieval based on Ec exhibits the highest performance, with the Vreg and EIreg series closely trailing behind. While the last few are the EIoc series, which may suggest that a coarse-grained emotional intensity classification is not suitable for rumor detection. However, it is the opposite in the conspiracy theory dataset. In COCO, the performance of retrieval based on the EIoc series is better than that based on the EIreg series.

4.3.3 Special cases analysis

Misinformation and true information often convey different affective information (as shown in Table 2 and Table 3). For example, fake news and conspiracy theories tend to evoke more negative sentiments and emotions (e.g. anger or fear) and less joy. However, these results are based on statistics derived from the entire dataset. The special cases need to be analyzed. We investigate some special

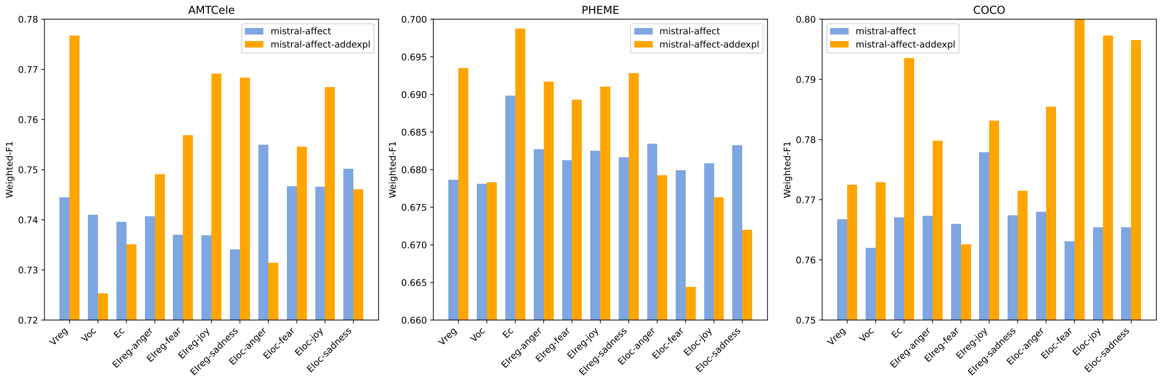

Figure 2: Results of mistral7b based on different affective information on three datasets. 'affect' denotes retrieving four examples based on one affective information using Template 1. 'affect-addexpl' denotes adding explicit affective information using Template 2.

Table 5: Special cases retrieval based on EIoc. 'num' denotes number. 'non-rum' denotes non-rumours. 'consp' denotes conspiracy. The '0', '2', and '3' in the EIoc column represent 'no', 'moderate', and 'high' emotional intensity.

Datasets	EIoc	num	F1	mean num of retrieval	
				legit/non-rum/related	fake/rumour/consp
AMTCele	fake anger=0	218	0.7718	1.0780	2.9220
	legit anger=2/3	29	0.9455	2.2414	1.7586
	fake joy=2/3	14	0.6000	1.5000	2.5000
	legit joy=0	304	0.8824	2.1217	1.8783
PHEME	non-rum fear=2/3	446	0.6959	2.4776	1.5224
	rumour fear=0	1039	0.3727	2.4658	1.5342
	non-rum joy=0	3795	0.8966	2.9057	1.0943
	rumour joy=2/3	25	0.2759	3.7600	0.2400
COCO	related fear=2/3	47	0.5313	2.0426	1.9574
	consp fear=0	171	0.8787	0.9708	3.0292
	realted joy ==0	246	0.7385	2.2927	1.7073

cases retrieved based on EIoc. The results are listed in Table 5.

For AMTCele, we investigate cases where fake news lacks anger or exhibits higher levels of joy, as well as cases where legit news displays higher levels of anger or lacks joy. We can see that the examples retrieved are mostly of the same category as the target, and their results have not been greatly influenced. For PHEME and COCO, we calculate statistics on cases of rumour and conspiracy without fear or exhibiting higher levels of joy (we do not report conspiracy with higher joy due to its low occurrence), as well as cases where non-rumour and related display higher levels of fear or without joy. We can see that the results for rumours in PHEME and related in COCO are relatively poor. The most likely reason is due to the imbalance of categories in the original data, and these special cases are in the minority. This has resulted in the retrieval of more data from the larger category in original datasets, causing the model to learn less useful information and ultimately affecting the final results.

5 Conclusion and Future Work

In this paper, we propose RAEmoLLM, the first RAG framework to address cross-domain misinformation detection using in-context learning based on affective information. We introduce the three modules of RAEmoLLM. We also conduct a comprehensive affective analysis for three public misinformation datasets. We evaluate the performance of RAEmoLLM on the three misinformation benchmarks based on various LLMs. The results show that RAEmoLLM can significantly improve LLMs compared to the zero-shot method and randomly few-shot method, which illustrates the effectiveness of RAEmoLLM. We also analyze the performance of retrieval based on different affective information, and some special cases, which provide a foundation for further improvements in the future.

In the future, we will explore the application of multimodal affective information in the task of detecting misinformation. We will also evaluate the application of the RAEmoLLM framework in other fields (e.g. mental health and finance). In addition to affective information, there are many other influencing factors in misinformation, such as stance and topic. We will combine sentiments and emotions with other features to construct a more robust retrieval database. Furthermore, the retrieval process can be slowed down by a large amount of data. In the future, we will also explore more efficient retrieval methods.

6 Limitations

Due to restricted computational resources, we only carried out inference of 2B, 7B, 8B, 13B, and 33B open-sourced LLMs. As such, we have not considered how the use of larger or different model architectures may potentially impact upon performance in cross-domain misinformation detection tasks.

Though achieving outstanding performance, RAEmoLLM still bears limitations. Firstly, for domain data with imbalanced distribution, RAEmoLLM performs worse compared to zero-shot methods (e.g. prince, gurlitt, and ebola domains in PHEME). The special cases analysis in Section 4.3.3 also illustrates in the imbalanced datasets, the retrieval in RAEmoLLM will be influenced for some special cases. Therefore, further exploration is needed to address such issues. Secondly, in the PHEME dataset, RAEmoLLM performs worse than fine-tuning methods without emotional information. This indicates that for simple tasks with shorter texts, the model still struggles to effectively balance textual features and emotional information.

7 Ethics Statement

The datasets we use in this paper are sourced from public social media platforms and websites. We strictly adhere to privacy agreements and ethical principles to protect user privacy and to ensure the proper application of anonymity in all texts.

References

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Arjun Choudhry, Inder Khatri, Arkajyoti Chakraborty, Dinesh Vishwakarma, and Mukesh Prasad. 2022. Emotion-guided cross-domain fake news detection using adversarial domain adaptation. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 75–79.

Carmela Comito, Francesco Sergio Pisani, Erica Coppolillo, Angelica Liguori, Massimo Guarascio, and Giuseppe Manco. 2023. Towards self-supervised cross-domain fake news detection.

Nicholas Della Giustina. 2023. Misinformation and its effects on individuals and society from 2015-2023: A mixed methods review study.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diwen Dong, Fuqiang Lin, Guowei Li, and Bo Liu. 2022a. Sentiment-aware fake news detection on social media with hypergraph attention networks. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2174–2180. IEEE.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022b. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 international conference on computer, communications and electronics (COMPTELIX)*, pages 397–402. IEEE.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Suzanne Keen. 2006. A theory of narrative empathy. *Narrative*, 14(3):207–236.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, pages 1–42.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*.

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*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024a. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, Raghav Jain, and Sophia Ananiadou. 2024b. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024c. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*.

Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024d. Emotion detection for misinformation: A review. *Information Fusion*, page 102300.

Quanyu Long, Wenya Wang, and Sinno Jialin Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. *arXiv preprint arXiv:2311.11551*.

Rosalind B Marimont and Marvin B Shapiro. 1979. Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, 24(1):59–70.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A deep learning approach for automatic detection of fake news. *arXiv preprint arXiv:2005.04938*.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Jiao Shi, Xin Zhao, Nan Zhang, Yu Lei, and Lingtong Min. 2023. Rough-fuzzy graph learning domain adaptation for fake news detection. *IEEE Transactions on Computational Social Systems*.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13.

Wei Tang, Zuyao Ma, Haifeng Sun, and Jingyu Wang. 2023. Learning sparse alignments via optimal transport for cross-domain fake news detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yi Wu, Ziqiang Li, Chaoyue Wang, Heliang Zheng, Shanshan Zhao, Bin Li, and Dacheng Tao. 2024. Domain re-modulation for few-shot generative domain adaptation. *Advances in Neural Information Processing Systems*, 36.

Shangqing Xu and Chao Zhang. 2024. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023a. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xuewen Zhang, Yaxiong Pan, Xiao Gu, and Gang Liang. 2023b. Sentiment analysis-based social network rumor detection model with bi-directional graph convolutional networks. In *International Conference on Computer Application and Information Security (ICCAIS 2022)*, volume 12609, pages 463–469. SPIE.

A Datasets

A.1 Preprocessing of raw datasets

All datasets we used are publicly available. FakeNewsAMT is a cross-domain dataset, including six domains. The legitimate news in FakeNewsAMT was obtained from various mainstream news websites. The authors adopted crowdsourcing via Amazon Mechanical Turk (AMT) to generate fake versions of legitimate news items. The Celebrity dataset was derived from online magazines. We combine FakeNewsAMT and Celebrity as AMTCele. PHEME contains a collection of Twitter rumors and non-rumors posted during nine breaking events news. COCO dataset consists of 12 conspiracy theory categories[5]. Each tweet in COCO is assigned an overall intention label, as follows: *Conspiracy* is assigned to tweets for which the tweet is related to at least one of the 12 categories and is actively spreading conspiracy theories. Otherwise, if the tweet is related to the specific category, but it does not propagate misinformation or conspiracy theories, then the overall label of *Related* is used. The overall label of *Unrelated* is only used for tweets that are unrelated to all 12 conspiracy categories. We remove the *Unrelated* text since the aim of the cross-domain test. For COCO dataset, we select the *Fake Virus*, *Harmful Radiation*, and *Depopulation* topics as the test set, and the other topics as the retrieval dataset.

A.2 Task Prompt and Instruction Dataset Example

For AMTCele, we utilize *"Determine whether the target text is 0. Fake or 1. Legit."* For PHEME, we employ *"Determine if the target text is 0. non-rumours or 1. rumours."* For COCO, we apply *"Classify the text regarding COVID-19 conspiracy theories or misinformation into one of the following three classes: 0. Unrelated. 1. Related (but not supporting). 2. Conspiracy (related and supporting)."* Here we keep the *0. Unrelated* category to test the robustness of the LLM by increasing the complexity of the task.

B Affective analysis

We show the statistics values and distribution of labels and embeddings in this Appendix. In Figures 3 to Figure 8, the y-axis represents the distribution of labels within the intention class indicated on the x-axis. The affective analysis on COCO has been done by ConspEmoLLM (Liu et al., 2024b).

[5]Suppressed Cures, Behavior Control, Anti Vaccination, Fake Virus, Intentional Pandemic, Harmful Radiation, Depopulation, New World Order, Esoteric Misinformation, Satanism, Other Conspiracy Theory, Other Misinformation.

Table 6: An example in PHEME instruction dataset.

Task: Determine if the target text is 0. non-rumours or 1. rumours.
Target text: UPDATE: Reports of 1 more shooter being SHOT. This is in addition to one shot and killed earlier in Parliament Hill #OttawaShooting. Sentiment intensity: 0.234.
Here are a few examples retrieved through sentiment intensity:
Text: UPDATE: Reports gunman says four devices are located around Sydney. Security response underway. Police calling for calm. #9News. Sentiment intensity: 0.429. The label of this text: 1. rumours.
Text: JUST IN: Police confirm to ABC there is a second hostage situation underway in eastern Paris. Sentiment intensity: 0.328. The label of this text: 1. rumours.
Text: UPDATE: There are reports police have discovered the identity of the lone gunman, with the #SydneySiege in its sixth hour. #9News Sentiment intensity: 0.435. The label of this text: 1. rumours.
Text: JUST IN: A separate shooting and hostage situation at a supermarket in eastern Paris has been reported ... developing. Sentiment intensity: 0.236. The label of this text: 1. rumours.
According to the above information, the label of target text:

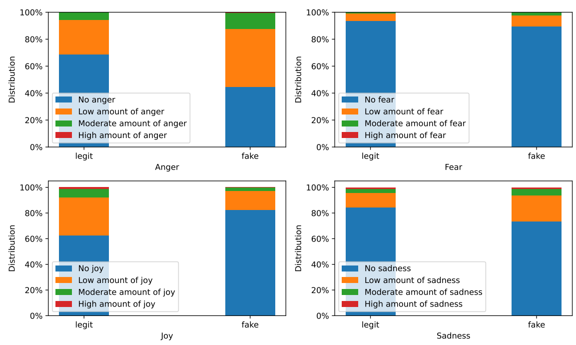

Figure 3: Emotion intensity classification on AMTCele

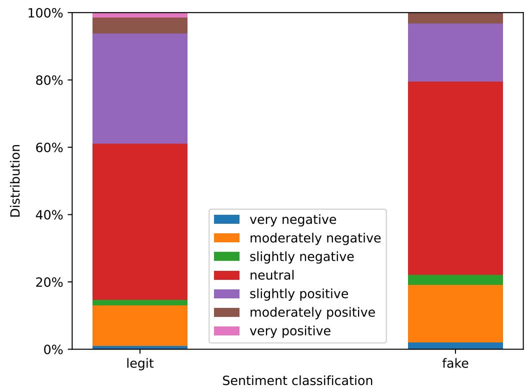

Figure 4: Sentiment classification on AMTCele

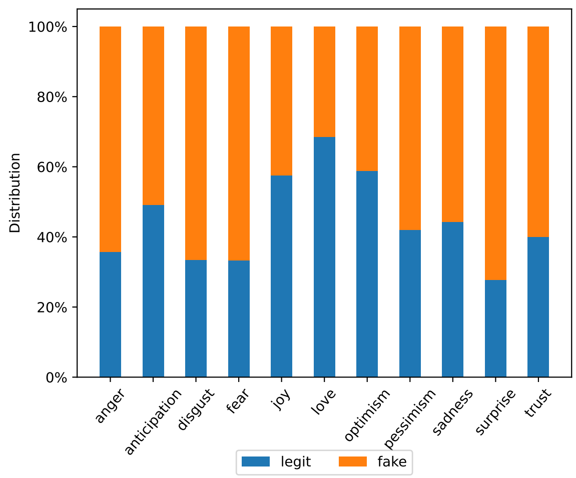

Figure 5: Emotion classification on AMTCele

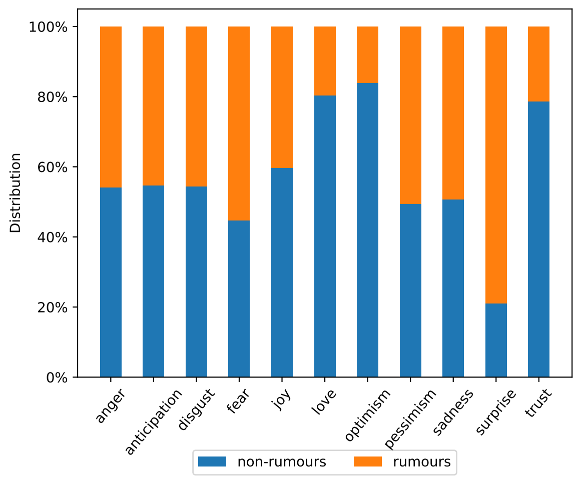

Figure 8: Emotion classification on PHEME

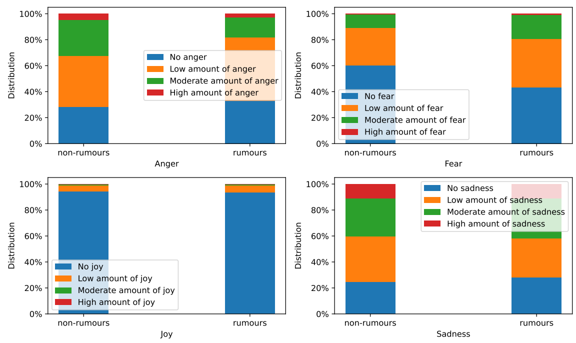

Figure 6: Emotion intensity classification on PHEME

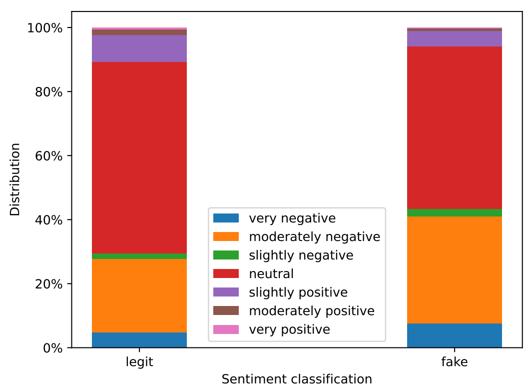

Figure 7: Sentiment classification on PHEME

C The results from different domains in the AMTCele dataset and the results from different topics in the PHEME dataset. (Table 7 and Table 8)

D Effectiveness analysis

Figure 11 shows different performances of mistrail7b retrieval examples using different affective dimension embeddings. From Figure 11, we can see in AMTCele, the F1 score of retrieval based on 16d embeddings is similar to original 4096d embedding retrieval, and in PHEME and COCO, the performance of retrieval based on larger dimensions is significantly better than lower dimensions retrieval. Figure 12 presents the Time-F1 score trade-off. We can see the time cost of the original 4096d retrieval is significantly higher than the time cost of other dimension retrievals. The closer to the upper left corner (i.e., less time, higher F1 score), the better. Overall, it seems that the LLM with 16d affective information retrieval has the highest trade-off efficiency.

E Results of different numbers of retrieval examples

Table 9 presents the F1 score of retrieval of different numbers of examples based on Vreg (we only tested 16 examples in the AMTCele dataset due to its long text). From the table, it can be observed that increasing the retrieval examples does not consistently improve the model's performance, and it may even lead to a decline in its performance (e.g. Vreg-addexpl in COCO). One possible reason is that when the model has multiple examples as references, it needs to consider a large amount of information comprehensively, which depends on the model's capability. Another reason we can infer from Table 3. For the three datasets, the p-values in retrieval top 4 examples are all zero. However, as the number of retrieval examples increases, the second p-values in AMTCele and the first p-value in COCO dataset also gradually increase. This indicates that the retrieved content may come from

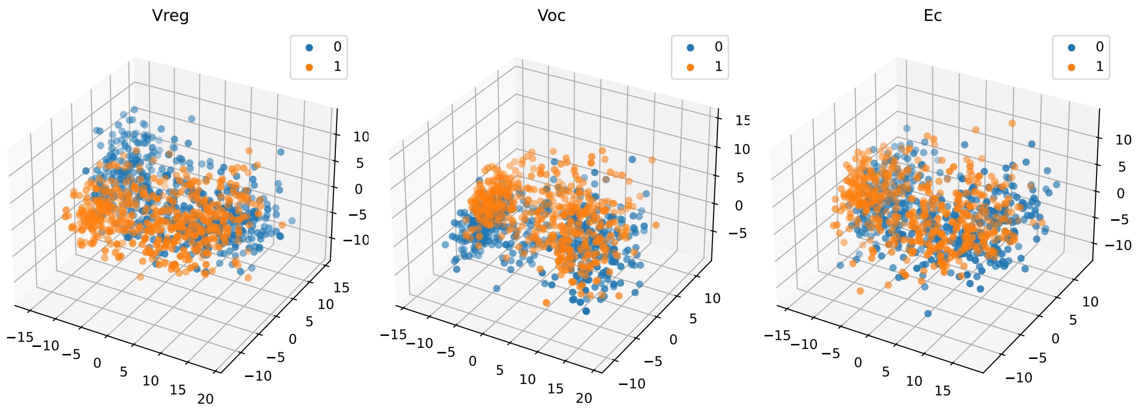

Figure 9: 3D visualization of affective embeddings on AMTCele. 0: Fake. 1: Legit

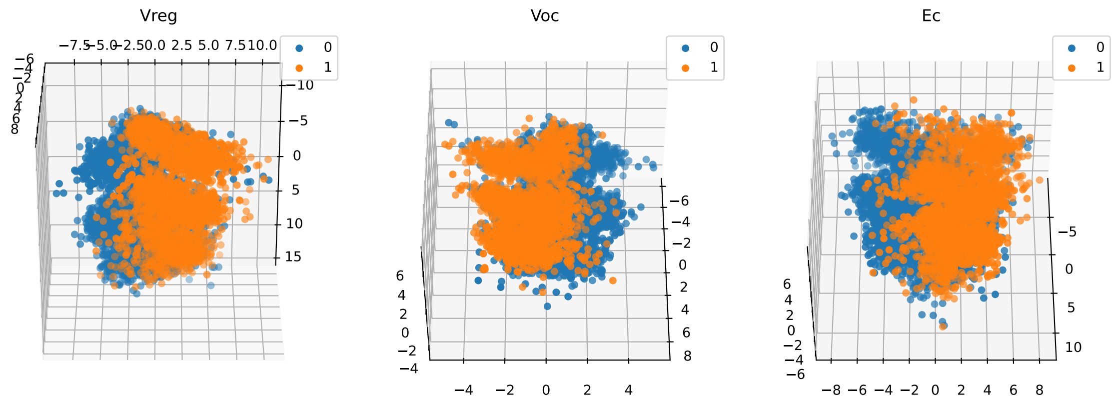

Figure 10: 3D visualization of affective embeddings on PHEME. 0: Non-rumours. 1: Rumours

Table 7: The results from different domains in the AMTCele dataset

	biz		edu		entmt		polit		sports		tech		celebrity	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	0.5975	0.5930	0.5725	0.5436	0.5800	0.5610	0.5450	0.5180	0.5525	0.5293	0.5650	0.5409	0.5152	0.5039
mistral7b-zs	0.7375	0.7324	0.8250	0.8222	0.7000	0.6952	0.5750	0.5248	0.8000	0.7980	0.6625	0.6260	0.7200	0.7174
mistral7b-random	0.7250	0.7067	0.8000	0.7968	0.8250	0.8249	0.5250	0.3866	0.6750	0.6366	0.5500	0.4508	0.6260	0.5778
mistral7b-Vreg	0.7750	0.7656	0.8250	0.8222	0.8250	0.8222	0.6375	0.5983	0.8125	0.8089	0.7000	0.6755	0.7400	0.7357
mistral7b-Vreg-addexpl	0.8000	0.7968	0.8750	0.8743	0.8750	0.8743	0.6750	0.6577	0.8375	0.8373	0.8625	0.8607	0.7380	0.7366

Table 8: The results from different domains in the PHEME dataset

	sydneysiege		ottawashooting		charliehebdo		ferguson		germanwings		prince		putinmissing		gurlitt		ebola	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	0.7463	0.7418	0.7497	0.7490	0.7971	0.8113	0.7053	0.7147	0.7275	0.7260	0.1296	0.1985	0.5866	0.5297	0.5391	0.4949	0.5714	0.7220
mistral7b-zs	0.6495	0.6512	0.6416	0.6418	0.6147	0.6476	0.4094	0.4213	0.6716	0.6659	0.7339	0.8315	0.5546	0.4807	0.4565	0.4518	0.4286	0.6000
mistral7b-random	0.6822	0.6833	0.6348	0.6289	0.6950	0.7193	0.5223	0.5542	0.6162	0.5970	0.3219	0.4660	0.5420	0.4597	0.5290	0.4278	0.4286	0.6000
mistral7b-Vreg	0.7215	0.7189	0.6640	0.6575	0.7321	0.7508	0.5844	0.6126	0.7271	0.7266	0.5365	0.6807	0.6008	0.5716	0.4783	0.4357	0.5000	0.6667
mistral7b-Vreg-addexpl	0.7486	0.7449	0.6708	0.6630	0.7456	0.7635	0.6623	0.6723	0.7015	0.7009	0.4421	0.5964	0.6008	0.5983	0.4420	0.4372	0.4286	0.6000

another category or unrelated examples, thereby affecting the model's judgment ability. Therefore, when employing retrieval augmentation techniques, it is not just about blindly increasing the number of examples, but rather selectively choosing the most useful examples.

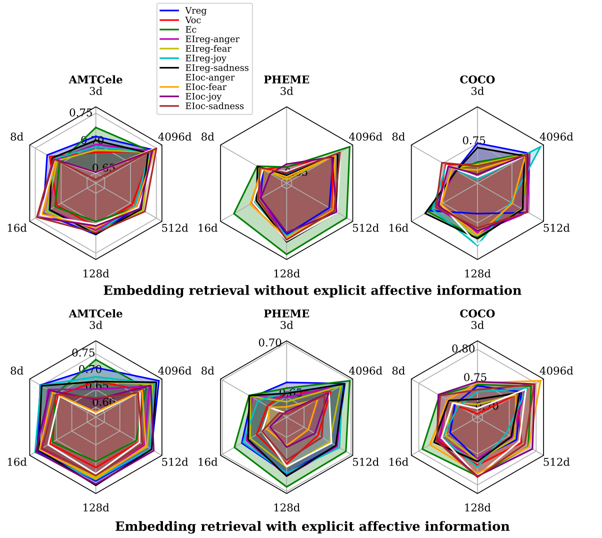

Figure 11: F1 score of mistral7b based on different affective dimension information retrieval on three datasets

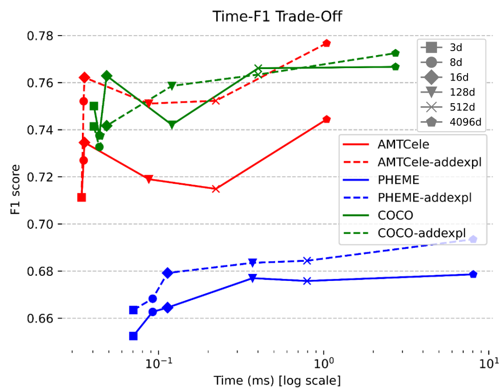

Figure 12: Time-F1 Trade-Off. The x-axis represents the average time (ms) consumed for retrieving the top 4 examples for each data point. The y-axis represents F1 score of mistral7b-Vreg on three datasets. The solid line represents the results obtained only using Vreg retrieval, while the dashed line (i.e., data-addexpl) represents the results with the addition of explicit Vreg.

Table 9: F1 score of mistral7b with retrieval of different numbers of examples based on Vreg. 'random' denotes randomly sampling four examples. 'Vreg' denotes retrieval of four examples based on Vreg. 'Vreg-addexpl' denotes adding explicit Vreg.

Datasets	methods	4	8	16	32	64
	Random	0.6876	0.7110	0.6174	-	-
AMTCele	Vreg	0.7444	0.7385	0.7322	-	-
	Vreg-addexpl	0.7767	0.7663	0.7680	-	-
	Random	0.6238	0.6218	0.6244	0.6389	0.6333
PHEME	Vreg	0.6786	0.6854	0.6821	0.6941	0.6997
	Vreg-addexpl	0.6935	0.6957	0.6965	0.6957	0.6920
	Random	0.7025	0.7472	0.7274	0.7524	0.7130
COCO	Vreg	0.7667	0.7610	0.7739	0.7838	0.7534
	Vreg-addexpl	0.7725	0.7128	0.7190	0.6933	0.6963