ConspEmoLLM: Conspiracy Theory Detection Using an Emotion-Based Large Language Model

Zhiwei Liu[a,*], **Boyang Liu**[a], **Paul Thompson**[a], **Kailai Yang**[a] **and Sophia Ananiadou**[a,b,c]

[a]The University of Manchester, Manchester, UK
[b]Artificial Intelligence Research Center, Japan
[c]Archimedes/Athena RC, Greece

Abstract. The internet has brought both benefits and harms to society. A prime example of the latter is misinformation, including conspiracy theories, which flood the web. Recent advances in natural language processing, particularly the emergence of large language models (LLMs), have improved the prospects of accurate misinformation detection. However, most LLM-based approaches to conspiracy theory detection focus only on binary classification and fail to account for the important relationship between misinformation and affective features (i.e., sentiment and emotions). Driven by a comprehensive analysis of conspiracy text that reveals its distinctive affective features, we propose ConspEmoLLM, the first open-source LLM that integrates affective information and is able to perform diverse tasks relating to conspiracy theories. These tasks include not only conspiracy theory detection, but also classification of theory type and detection of related discussion (e.g., opinions towards theories). ConspEmoLLM is fine-tuned based on an emotion-oriented LLM using our novel ConDID dataset, which includes five tasks to support LLM instruction tuning and evaluation. We demonstrate that when applied to these tasks, ConspEmoLLM largely outperforms several open-source general domain LLMs and ChatGPT, as well as an LLM that has been fine-tuned using ConDID, but which does not use affective features. ConspEmoLLM can be easily applied to identify and classify conspiracy-related text in the real world. The work has been released at https://github.com/lzw108/ConspEmoLLM/.

1 Introduction

Misinformation has become one of the major threats to society. The rise of the internet and social media has made it increasingly simple for misinformation to spread rapidly. Conspiracy theories are one type of misinformation, whose false content is intended to cause harm [20]. Examples of popular conspiracy theories include those claiming that the Earth is flat and that vaccines cause autism. Conspiracy theorists ignore scientific evidence and tend to interpret events as secretive actions [5]. During the COVID-19 pandemic, the spread of conspiracy theories significantly increased (e.g., the claim that 5G telecommunications networks activated the virus), resulting in a considerable negative impact on society [4]. As such, there is an increasing urgency for high-performance methods that can automatically detect conspiracy theories.

It has previously been shown that there is a close relationship between misinformation (including conspiracy theories) and affective information, i.e., sentiment and emotions [14]. For example, Dong et al. [3] found that during the COVID-19 pandemic, there was a correlation between the level of public anger and the likelihood of rumor propagation, while Zaeem et al. [34] also observed a significant positive correlation between negative emotions and fake news. Based on such observations, many studies have adopted affective information as the means to detect misinformation [11, 37]. Here, we aim to extend the study of affective information, specifically sentiment and emotions, to deepen our understanding of conspiracy theories and to improve their automated detection.

Pre-trained language models (PLMs) such as BERT [2] and RoBERTa [12] have shown outstanding performance when applied to various classification tasks, including conspiracy theory detection [31, 23]. However, due to their restricted number of parameters, PLMs cannot perform optimally when applied to diverse and complex tasks [35]. Recently, LLMs, which possess significantly larger numbers of parameters, have been explored as a novel means to address the issue of misinformation, with very promising results [7, 21, 1]. However, these studies mostly focus on binary classification of texts according to whether or not they convey misinformation, Moreover, these previous efforts have mostly utilized simple prompts to test or carry out instruction-tuning of LLMs, or else have employed LLMs as auxiliary tools for other models. To our knowledge, no existing LLM-based studies have attempted to leverage important affective features that are characteristic of misinformation, nor have such studies attempted to carry out in-depth analyses of conspiracy-related text.

To address these research gaps, we have constructed a multi-task conspiracy detection instruction dataset, *ConDID*, to facilitate instruction-tuning and evaluation of LLMs. Based on annotations in two conspiracy theory datasets, ConDID is divided into five tasks that encompass conspiracy theory judgment, conspiracy theory topic detection, and conspiracy theory intention detection. We subsequently propose a novel open-source LLM, ConspEmoLLM, which is specialized for detecting conspiracy-related information. ConspEmoLLM is created by applying an instruction-tuning method to an emotion-oriented LLM using the ConDID dataset. Evaluation of ConspEmoLLM using the ConDID test set shows that it achieves state-of-the-art (SOTA) performance among other open-source LLMs, as well as the closed-source ChatGPT.

Our main contributions are as follows:

(1) We develop ConDID, the first multi-task conspiracy instruction-tuning dataset.

* Corresponding Author. Email: zhiwei.liu-2@postgrad.manchester.ac.uk

(2) We conduct an affective analysis of the two conspiracy theory datasets used to construct ConDID, which provides evidence that expressions of sentiment and emotions in conspiracy theory text are distinct from those occurring in the mainstream text.

(3) We propose ConspEmoLLM, the first open-source emotion-based LLM that is specialized for diverse conspiracy theory detection tasks. Evaluation of ConspEmoLLM shows that it outperforms other open-source LLMs and ChatGPT across different tasks. It also surpasses the performance of an instruction-tuned LLM that does not incorporate affective features, thus confirming the effectiveness and importance of affective information in detecting conspiracy-related information.

The remainder of this paper is structured as follows: Section 2 introduces related work concerning the detection of conspiracy theories and misinformation, analysis of sentiment, and open-source LLMs. Section 3 describes our construction of the ConDID dataset, the affective analysis of conspiracy theory datasets, and the instruction-tuning of ConspEmoLLM. Section 4 reports on our evaluation and analyzes of the performance of multiple models on the ConDID test set. Section 6 concludes the paper and provides directions for future work. Sections 7 and 8, respectively, discuss potential limitations and confirm the ethical soundness of our study. The supplementary material [15] provides figures illustrating our detailed affective analysis of conspiracy theory datasets.

2 Related work

2.1 Conspiracy theory and misinformation detection

PLMs have been widely applied to the task of conspiracy theory and misinformation detection. For example, Yanagi et al. [31] utilized BERT as the base model for COVID-19 conspiracy theory detection, while Peskine et al. [23] applied a domain-specific COVID BERT (CT-BERT) to the same task. More recently, an increasing amount of attention has been devoted to exploring the application of LLMs to detect conspiracy theories and misinformation. For example, Peskine et al. [24] employed zero-shot learning to evaluate the accuracy of the GPT-3 model in performing fine-grained multi-label conspiracy theory classification of tweets, while Hu et al. [7] proposed a fake news detection framework that utilizes an LLM as an auxiliary tool to enhance the prediction accuracy of BERT. Meanwhile, Pavlyshenko [21] employed prompt-based fine-tuning of LLaMA2 for rumor and fake news detection. Cheung and Lam [1] supplemented an LLM with external knowledge to enhance the performance of fake news detection. However, all of these models focus on binary classification and do not exploit affective information for misinformation detection.

2.2 Affective analysis

Emotion detection and sentiment analysis are two types of NLP techniques for analyzing human expressions. Sentiment analysis aims to capture both the overall emotional tone conveyed by the data source (typically *positive*, *negative*, or *neutral*) and the intensity of this tone. Emotion detection is the process of categorizing data at a finer level of granularity based on the emotions conveyed. In comparison to sentiments, emotions correspond to more specific and intense feelings. For example, negative emotions include *anger* and *fear*, while positive emotions include *happiness* and *joy* [14]. Identifying both sentiments and emotions is crucial for downstream tasks.

Automated affective analysis of text has previously been carried out by various means, including the use of different sentiment analysis tools, such as VADER [8] and TextBlob[1]. PLMs have also been employed for sentiment analysis. For example, Hoang et al. [6] used BERT for aspect-based sentiment analysis, while Tan et al. [27] proposed a hybrid sentiment analysis model that combines RoBERTa with an LSTM. More recent work has also begun to explore the effectiveness of LLMs for sentiment analysis. For instance, Zhang et al. [35] and Lei et al. [10] both utilized retrieval-augmented LLMs to enhance the sentiment analysis capabilities of LLMs when applied to financial news and dialogues, respectively. Meanwhile, Liu et al. [13] proposed a series of comprehensive LLMs that are specialized for affective analysis (EmoLLMs), and which are capable of analyzing emotions across five different dimensions (i.e. emotion intensity, ordinal classification of emotion intensity, sentiment strength, sentiment classification, emotion detection). EmoLLMs demonstrate strong generalization ability, surpassing ChatGPT and GPT-4 in most emotion analysis tasks. Therefore, we use one of these EmoLLMs, i.e., EmoLLaMA, to perform affective analysis in this study.

2.3 Open-source LLMs

A large amount of research has been dedicated to developing open-source LLMs as an alternative to the well-known closed-source LLMs (e.g., ChatGPT), in order to support easier research into the improvement and application of LLMs. Popular series of open-source, general language models include LLaMA [28], OPT [36] and BLOOM [29]. These are complemented by a range of domain-specific open-source LLMs, including FinMA [30] for finance, MentalLLaMA [32] for mental health, ExTES-LLaMA [38] for emotional support chatbots, and TimeLLaMA [33] for temporal reasoning. In this work, we extend the inventory of domain-specific LLMs, by developing the first open-source LLM for multitask conspiracy theory detection based on affective information.

3 Methods

3.1 Task formalization

We approach conspiracy theory detection as a generative task, using a generative model as its foundation. This generative model is an autoregressive language model $P_\phi(y|x)$, parameterized using pre-trained weights ϕ. It differs from previous discriminative models, in terms of its ability to simultaneously handle multiple conspiracy theory detection tasks, i.e., conspiracy identification, conspiracy intention detection, and conspiracy theme recognition. Each task, denoted as t, is represented as a set of context-target pairs: $D_t = (q_i^t, r_i^t)_{i=1,2,\ldots N_t}$, where the context q is a token sequence containing the task description, input text, and query, and r is a further token sequence containing the answer to the query. The model is optimized based on the merged dataset, which combines all task datasets, with the aim of maximizing the objective of conditional language modeling to improve prediction performance.

3.2 Construction of instruction tuning dataset

3.2.1 Raw data

We build our instruction tuning dataset using two existing annotated datasets:

[1] https://textblob.readthedocs.io/

COCO The COVID-19 conspiracy theories (COCO) dataset [9] is an extension of the dataset used in the *MediaEval FakeNews: Corona Virus and Conspiracies Task* challenge [25]. COCO consists of tweets, each of which is assigned 12 different labels that characterize the *intention* of the tweet with respect to 12 different conspiracy theory categories [2]. Each label can have three possible values, i.e., *Unrelated:* The tweet is unrelated to the specific conspiracy category; it contains conspiracy-related keywords, but they are used in a completely different context; *Related:* The tweet is related to the specific category, but it does not propagate misinformation or conspiracy theories; *Conspiracy:* The tweet is related to the specific category and is actively aimed at spreading conspiracy theories.

Each tweet in COCO is also assigned a single overall intention label, as follows: The overall *Conspiracy* label is assigned to tweets for which the *Conspiracy* label is used for at least one of the 12 categories. Otherwise, if the *Related* category is assigned to at least one of the categories, 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.

LOCOAnnotations The Language of Conspiracy (LOCO) [17] is a corpus consisting of documents gathered from the internet (88 million words), which is used to study the differences between conspiracy language and mainstream language. We use an annotated subset of LOCO, created by Mompelat et al. [18], which we refer to as *LOCOAnnotations*. This subset consists of documents concerning two different topics (i.e., the Sandy Hook school shooting and coronavirus) in which two types of labels have been assigned. The first type of label concerns whether or not the document is directly concerned with a conspiracy theory. The second type of label reflects the degree of relatedness to a conspiracy theory, using three labels (*closely related, broadly related* or *not related*).

3.2.2 Tasks

Using these two different datasets, we define five different tasks. Tasks 1-3 are based on the COCO corpus, and are similar to those used in the MediaEval FakeNews challenge, while Tasks 4 and 5 are based on LOCOAnnotations.

Task 1: Conspiracy Intention Detection Determine the overall intention of the tweet (i.e., *Unrelated/Related/Conspiracy*) towards COVID-19 conspiracy theories.

Task 2: Conspiracy Theory Topics Detection Determine whether the tweet mentions or refers to any of the 12 predefined conspiracy theory categories.

Task 3: Combination of Task 1 and Task 2 Predict *both* the conspiracy category and the relationship of the tweet (*Unrelated/Related/Conspiracy*) to the category.

Task 4: Conspiracy Theory Detection Determine whether a document is directly concerned with a conspiracy theory (*Conspiracy/Non-Conspiracy*).

Task 5: Relatedness Detection Determine the level of relatedness of a document to a conspiracy theory (*Closely related/Broadly related/Not related*).

3.2.3 Affective analysis of raw data

EmoLLMs [13] is a series of affective analysis models. We adopt the best-performing model from this series (i.e., EmoLLaMA-chat-7b) to conduct sentiment analysis on COCO and LOCOAnnotations across five different dimensions, as follows:

- *Emotion intensity score*: For each of four different emotions (*anger, fear, joy* and *sadness*), a real-valued score between 0 and 1 is assigned to represent the intensity of the emotion in the text.
- *Emotion intensity classification*: For the same four emotions, one of four different emotional intensity classes is assigned to the text, i.e. *{no, low, moderate, high} emotional intensity*.
- *Sentiment strength score*: A real-valued score between 0 (most negative) and 1 (most positive) is assigned to represent the intensity of sentiment expressed in the text.
- *Sentiment classification*: One of seven classes (i.e. *{very, moderately, slightly} negative, neutral, {slightly, moderately, very} positive*) is assigned to represent the intensity of sentiment conveyed in the text.
- *Emotion detection*: One or more labels is assigned to the text if any of eleven different emotions are conveyed (i.e. *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise* and *trust*). If none of these emotions is expressed, then the label *neutral or no emotion* is assigned.

The results of analyzing the COCO dataset according to these five dimensions are shown in Figures 1 to 5, broken down according to the different intentions of the tweets in the dataset (i.e., *Unrelated/Related/Conspiracy*). In Figures 1 and 3, the x-axis represents the real-valued scores for intensity of different emotions and sentiment strength, respectively. The y-axis represents the corresponding probability density distribution (i.e., the number of tweets belonging to each intention class that has a particular score for emotional intensity/sentiment strength). In Figures 2, 4 and 5, the y-axis represents the distribution of labels within the intention class indicated on the x-axis. Due to space limitations, these figures only report on part of our complete analysis. The supplementary material [15] includes additional figures that depict an analysis of the five dimensions based on the different conspiracy categories of tweets in the COCO dataset. The supplementary material [15] also includes analyses of LOCOAnnoations, which examine affective differences between conspiracy and non-conspiracy text, and between text exhibiting various levels of relatedness to conspiracy theories (i.e., *Not related/closely related/broadly related*).

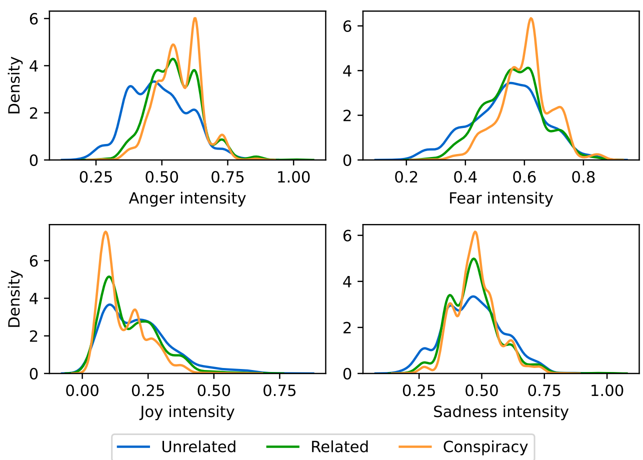

Figure 1. Emotion intensity of different intentions

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

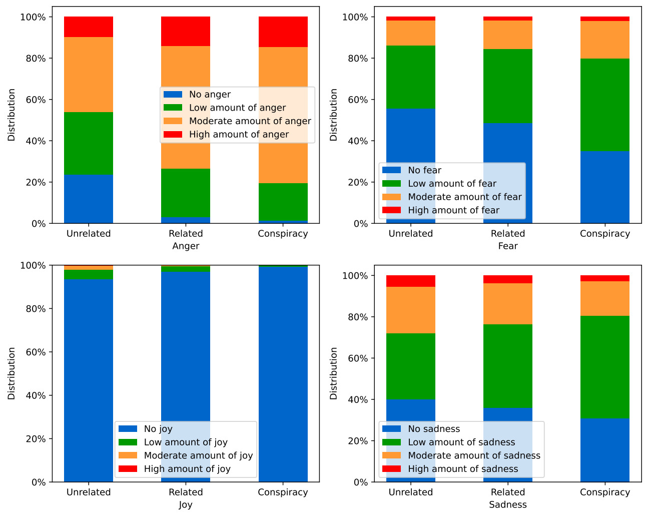

Figure 2. Emotion intensity classification of different intentions

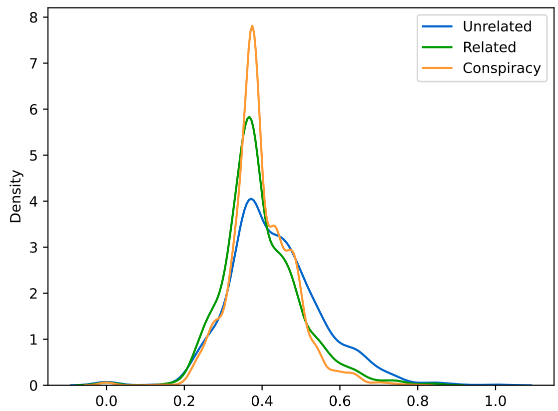

Figure 3. Sentiment strength of different intentions

Figure 1 presents a quantitative analysis of emotional intensity. It may be observed that tweets that refer to conspiracy theories (i.e., *Related/Conspiracy*) express stronger feelings of anger and fear, with a lesser degree of joy, compared to the *Unrelated* tweets. The intensity of sadness is similar across the three categories, indicating that during the pandemic, people were generally in a state of sadness. Similar conclusions may be drawn from Figure 2, which presents a qualitative analysis of emotional intensity. The sentiment strength in Figure 3 and the sentiment polarity classification in Figure 4 indicate that sentiments expressed in *Unrelated* tweets are more strongly positive than sentiments in tweets related to conspiracy theories. It can be observed from Figure 5 that tweets related to conspiracy theories predominantly convey negative sentiments and emotions (e.g., anger, fear, and disgust). In contrast, the *Unrelated* tweets are more likely to express positive sentiments and emotions (e.g., joy, love, and optimism). The figures in the supplementary material [15] reveal that the types of emotions expressed in tweets can also vary according to the specific category of conspiracy theory being discussed in the COCO dataset, and that there are noticeable differences in the types of emotion information conveyed across different categories and levels of conspiracy relatedness in the LOCOAnnotations dataset. Overall, it can be inferred that there are close relationships between affective information and conspiracy theories. The affective information can be an important feature for detecting conspiracy theories.

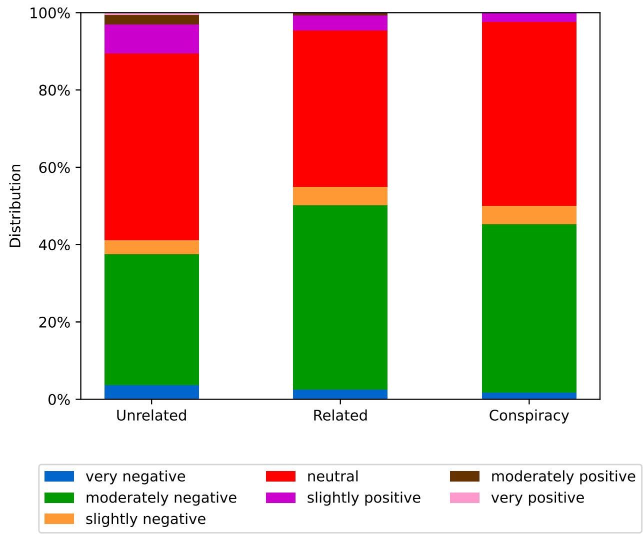

Figure 4. Sentiment classification of different intentions

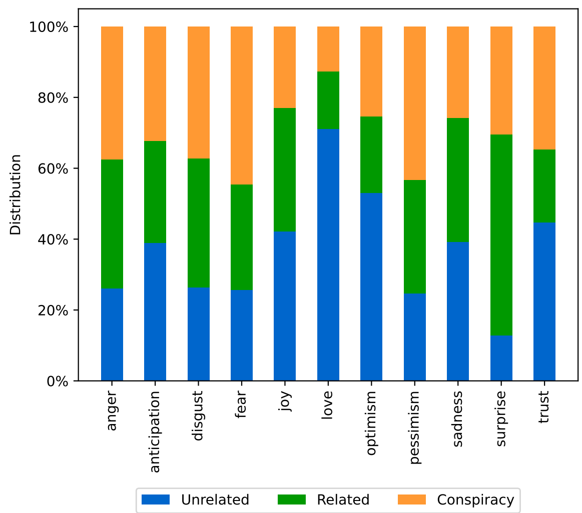

Figure 5. Emotion classification of different intentions

3.2.4 *Construction of the ConDID conspiracy detection instruction dataset*

We used the raw datasets as the basis to build the instruction dataset. We randomly selected 20% of the data as the test set and 20% as the validation set. The dataset statistics are presented in Table 1. For Task 3, prompts needed to be created for each of the 12 categories, resulting in a magnitude of data that is 12 times the size of the original corpus. We constructed instruction-tuning data for each task based on the following template:

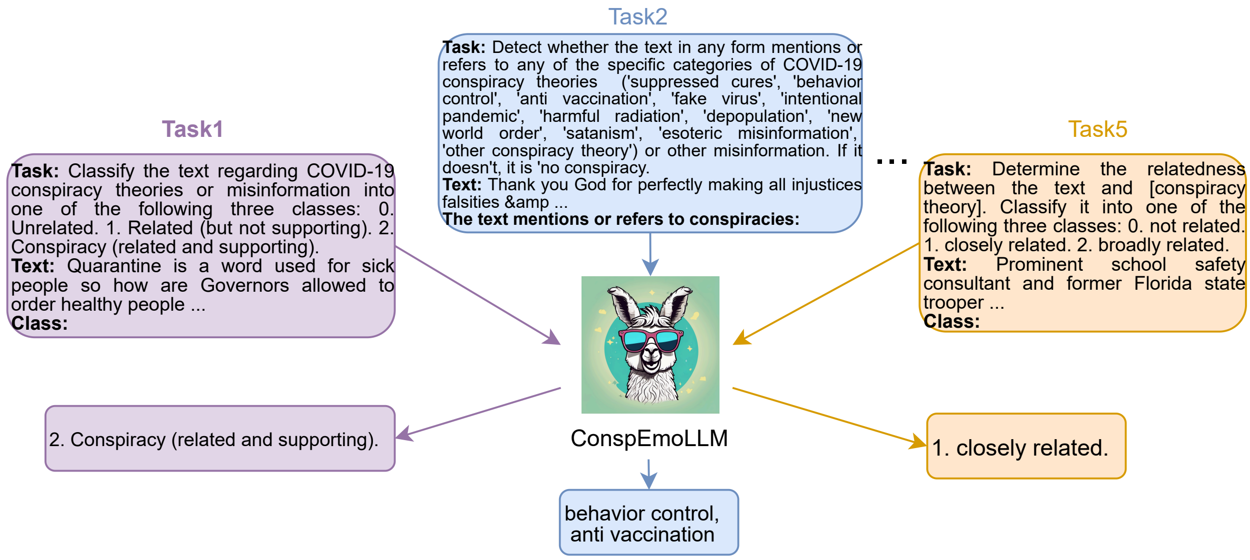

Figure 6. An overview of multi-task instruction tuning of ConspEmoLLM

Task	Raw (Train/Dev/Test)	Instruction (Train/Dev/Test)
Task 1		2092/697/698
Task 2	2092/697/698	2092/697/698
Task 3		25104/8364/8376
Task 4		669/223/233
Task 5	669/223/233	669/223/233

Table 1. Dataset statistics. *Raw* denotes the raw data from COCO and LOCOAnnotations. *Instruction* denotes the converted instruction data based on *Raw*.

Task: *[task prompt]* Text: *[input text]* Class/The text mentions or refers to conspiracies: *[output]*

[task prompt] denotes the instruction for the task. *[input text]* is a data item from the raw data. With the exception of Task 2, all tasks use *Class* to prompt LLM to generate answers. *[output]* is the output from LLM. Table 2 lists the task prompts for each task, and Figure 6 presents several examples used to fine-tune the LLM. To allow an evaluation of the impact of explicitly encouraging the LLMs to make use of affective information, we additionally constructed prompts of the form shown in the final row of Table 2. These prompts provide information about the specific analysis results obtained from EmoLLaMA regarding the text to be classified.

3.3 ConspEmoLLM and ConspLLM

We built ConspEmoLLM by fine-tuning EmoLLaMA-chat-7b [13] using the ConDID dataset. We also fine-tuned an LLM that does not use affective information (ConspLLM) based on LLaMA2-chat-7b [28], also using ConDID. The models are trained based on the AdamW optimizer [16] for three epochs, using DeepSpeed [26] to reduce memory usage. We set the batch size to 256. The initial learning rate is set to 1e-6 with a warm-up ratio of 5%, and the maximum model input length is set to 4096. All models are trained on two Nvidia Tesla A100 GPUs, each with 80GB of memory. Figure 6 provides an overview of multi-task instruction tuning of ConspEmoLLM for diverse conspiracy detection tasks.

4 Experiments

4.1 Baseline models

PLMs: Conspiracy theory detection is typically regarded as a classification task. For our baseline models, we selected commonly used PLMs, which can only be fine-tuned for individual tasks, i.e., the general language BERT and RoBERTa, along with CT-BERT [19], which is tailored for the COVID-19 domain. We treat Tasks 1 and 5 as 3-way classification tasks, and Task 4 as a binary classification task, using cross-entropy loss for training. Task 2 is treated as a multi-label binary classification problem, which can be achieved by utilizing the binary cross-entropy with logits loss. To address Task 3, fine-tuning is performed for 12 different classification problems, each with its own cross-entropy loss function. The final loss is the average of the 12 losses.

LLMs: LLMs have been proven to be capable of solving numerous tasks. We apply zero-shot prompting on the instruction dataset to the following open-source LLMs: Falcon-7B-instruct [22], LLaMA2-chat-7B [28], OPT-7B [36], BLOOM-7B [29], and Vicuna-7B-v1.5[3]. We also utilize zero-shot prompting with the proprietary LLM ChatGPT.

4.2 Evaluation methods

Since the tasks addressed in this paper are all classification problems, we apply commonly used metrics to evaluate the performance of the models, i.e., Accuracy (ACC), Precision (PRE), Recall (REC), and weighted F1 score.

4.3 Results

Tables 3 and 4 report the results for each task. ChatGPT-aff and ConspLLM-aff are the models that use the prompts containing ex-

[3] https://huggingface.co/lmsys/vicuna-13b-v1.5

Task	Prompt Template
Task 1	Task: 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).
Task 2	Task: detect whether the text in any form mentions or refers to any of the specific categories of COVID-19 conspiracy theories ('suppressed cures', 'behavior control', 'anti vaccination', 'fake virus', 'intentional pandemic', 'harmful radiation', 'depopulation', 'new world order', 'satanism', 'esoteric misinformation', 'other conspiracy theory') or other misinformation. If it doesn't, it is 'no conspiracy.
Task 3	Task: Classify the text regarding the specific category [Specific Conspiracy] into one of the following three classes: 0. Unrelated. 1. Related (but not supporting). 2. Conspiracy (related and supporting).
Task 4	Task: Determine if the text is a conspiracy theory. Classify it into one of the following two classes: 0. non-conspiracy. 1. conspiracy.
Task 5	Task: Determine the relatedness between the text and [conspiracy theory]. Classify it into one of the following three classes: 0. not related. 1. closely related. 2. broadly related.
Affective prompt	Task: original task prompt + "You can also refer to the affective information. (1) Emotion intensity: anger: 0.521, fear: 0.625, joy: 0.25, sadness: 0.354. (2) Ordinal classification of emotion intensity: moderate amount of anger can be inferred. low amount of fear can be inferred. no joy can be inferred. no sadness can be inferred. (3) Sentiment intensity: 0.435. (4) Sentiment classification: neutral or mixed mental state can be inferred. (5) The emotions included are: anger, disgust, fear."

Table 2. Prompts used for each task.

Model	Task1				Task2				Task3			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
BERT	0.576	0.601	0.428	0.406	0.272	0.150	0.012	0.023	0.893	0.871	0.893	0.842
RoBERTa	0.517	0.505	0.335	0.231	0.279	0.204	0.077	0.112	0.893	0.842	0.893	0.844
CT-BERT	0.564	0.472	0.469	0.428	0.042	0.137	0.303	0.175	0.893	0.797	0.893	0.842
Falcon	0.265	0.361	0.265	0.211	0.189	0.193	0.180	0.160	0.556	0.810	0.556	0.653
Vicuna	0.380	0.446	0.380	0.392	0.136	0.172	0.270	0.150	0.437	0.825	0.437	0.556
LLaMA2-chat	0.325	0.527	0.325	0.251	0.193	0.093	0.071	0.074	0.674	0.814	0.674	0.732
OPT	0.311	0.389	0.311	0.298	0.235	0.216	0.048	0.072	0.481	0.805	0.481	0.590
BLOOM	0.328	0.389	0.328	0.318	0.268	0.000	0.000	0.000	0.114	0.810	0.114	0.146
ChatGPT	0.638	0.677	0.638	0.596	0.324	0.546	0.333	0.332	0.208	0.896	0.208	0.240
ChatGPT-aff	0.583	0.647	0.583	0.525	0.312	0.449	0.326	0.308	0.150	0.898	0.150	0.146
ConspLLM	0.662	**0.757**	0.662	0.675	0.328	0.685	0.320	0.334	0.893	0.886	0.893	**0.864**
ConspLLM-aff	0.517	0.483	0.517	0.354	0.203	0.404	0.330	0.301	0.077	**0.901**	0.077	0.015
ConspEmoLLM	**0.695**	0.755	**0.695**	**0.705**	**0.340**	**0.699**	**0.345**	**0.364**	**0.897**	0.884	**0.897**	0.860

Table 3. Evaluation results for Tasks 1, 2, and 3

plicit affective information. From the results tables, we can observe that, in terms of F1 score, the fine-tuned ConspLLM and ConspEmoLLM outperform all other open-source models[4], as well the PLMs. ConspLLM and ConspEmoLLM also both outperform ChatGPT, with the exception of ConspLLM on Task 4. Moreover, ConspEmoLLM, which is fine-tuned based on a large emotion language model, achieves F1 scores that are over 3% higher than ConspLLM for all tasks except for Task 3, in which it lags slightly behind ConspLLM. The results for ChatGPT-aff and ConspLLM-aff reveal that explicitly augmenting prompts with affective information leads to reduced performance for all tasks apart from Task 4. We can infer that explicitly adding affective information appears to distract models' attention from the task at hand. In contrast, the implicit use of emotion information by ConspEmoLLM is more successful in allowing emotional cues to be leveraged.

5 Automatic predictions

The following example code demonstrates how ConspEmoLLM can be applied to automatically classify text according to whether or not it represents a conspiracy theory. Any piece of text may be classified by ConspEmoLLM, by replacing *[input text]* with the text of interest. The task prompt may also be adjusted to perform the other conspiracy-related tasks introduced above. Further details can be found on Github [5].

[4] It should be noted that the LLMs that are not instruction-tuned produced some responses that do not follow instructions, i.e., they do not produce output of the type requested in the prompts. In such cases, we label them as *unrelated* or *non-conspiracy*, depending on the task.

[5] https://github.com/lzw108/ConspEmoLLM

```
from transformers import AutoTokenizer,
    AutoModelForCausalLM
Model_PATH = "lzw1008/ConspEmoLLM-7b"
tokenizer = AutoTokenizer.from_pretrained(
    MODEL_PATH)
model = AutoModelForCausalLM.from_pretrained(
    MODEL_PATH, device_map='auto')

prompt = '''Human:
Task: Task: Determine if the text is a
    conspiracy theory. Classify it into one
    of the following two classes: 0. non-
    conspiracy. 1. conspiracy
Text: [input text]
Class:
Assistant:
'''
inputs = tokenizer(prompt, return_tensors="pt
    ")
generate_ids = model.generate(inputs["
    input_ids"], max_length=256)
response = tokenizer.batch_decode(
    generate_ids, skip_special_tokens=True)
    [0]
print(response)
>> 0. non-conspiracy / 1. conspiracy
```

Model	Task4				Task5			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1
BERT	0.691	0.677	0.642	0.645	**0.614**	0.623	0.356	0.296
RoBERTa	0.677	0.664	0.670	0.666	0.601	0.200	0.333	0.250
Falcon	0.448	0.489	0.448	0.453	0.422	0.640	0.422	0.477
Vicuna	0.529	0.524	0.529	0.526	0.274	0.633	0.274	0.279
LLaMA2-chat	0.583	0.611	0.583	0.588	0.291	0.634	0.291	0.314
OPT	0.466	0.509	0.466	0.470	0.507	**0.676**	0.507	0.554
BLOOM	0.439	0.483	0.439	0.443	0.587	0.653	0.587	0.617
ChatGPT	0.668	**0.774**	0.668	0.664	0.596	0.576	0.596	0.574
ChatGPT-aff	0.673	0.769	0.673	0.670	0.587	0.557	0.587	0.551
ConspLLM	0.641	0.663	0.641	0.646	0.596	0.574	0.596	0.580
ConspLLM-aff	**0.731**	0.758	**0.731**	**0.735**	0.453	0.555	0.453	0.479
ConspEmoLLM	0.700	0.717	0.700	0.703	0.610	0.647	**0.610**	**0.623**

Table 4. Evaluation results for Tasks 4 and 5

6 Conclusion

In this paper, our comprehensive affective analysis of two conspiracy theory datasets has demonstrated that conspiracy theory text exhibits sentiment and emotion features that are distinct from mainstream text. The results of this analysis motivated our development of ConspEmoLLM, an open-source domain-specific LLM that is based on affective information, and which can perform diverse conspiracy theory detection tasks. ConspEmoLLM was fine-tuned using our newly constructed multitask conspiracy detection instruction dataset (ConDID). Evaluation on the test set of ConDID reveals that ConspEmoLLM achieves SOTA performance among the other open-source LLMs tested, as well as ChatGPT, in all tasks. On most tasks, the performance of ConspEmoLLM surpasses that of the ConspLLM model, which was also instruction-tuned using ConDID, but does not use affective information. These results provide strong evidence of the importance of affective features in detecting various types of information relating to conspiracy theories. We also exemplified how ConspEmoLLM can be applied straightforwardly to detect the presence of conspiracy-related information in any piece of text. This demonstrates the real world utility of our model in helping people to identify the misinformation on the Internet, thus contributing towards reducing the potential harm caused by conspiracy theories.

As future work, we aim to augment the ConDID dataset with further conspiracy theory datasets, including data from multiple platforms, sources, domains and languages. This should help to further improve the performance and diversity of tasks that can be carried out using ConspEmoLLM. We will additionally explore alternative methods of incorporating affective information to further improve the ability of the model to detect conspiracy theories. Furthermore, we will design more appropriate prompts and utilize more complex and diverse model structures to better leverage emotions and sentiments.

7 Limitations

The potential limitations of our work may be summarized as follows:

(1) Due to restricted computational resources, we only carried out instruction-tuning and evaluation of conspiracy theory detection tasks using 7B LLMs. As such, we have not considered how the use of larger or different model architectures may potentially impact upon performance in conspiracy theory detection tasks.

(2) The datasets used in this paper mostly concern COVID-19 conspiracy theories and are limited in size. These limitations may affect the ability of the model to generalize to other types of data or domains. However, as mentioned in Section 6, we plan to increase the size of our ConDID dataset, by collecting additional data from diverse sources and domains. It is hoped that this will help to improve both the performance and generalizability of the model.

8 Ethics Statement

The original datasets collected to construct the ConDID dataset 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.

Acknowledgements

The ConspEmoLLM illustration in Figure 6 was generated using PIXLR[6]. This work is supported by the computational shared facility at the University of Manchester and the scholar award from the Department of Computer Science at the University of Manchester. This work is also supported by the Centre for Digital Trust and Society at the University of Manchester, the Manchester-Melbourne-Toronto Research Fund, and the New Energy and Industrial Technology Development Organization.

References

[1] T.-H. Cheung and K.-M. Lam. 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, 2023.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] W. Dong, J. Tao, X. Xia, L. Ye, H. Xu, P. Jiang, and Y. Liu. Public emotions and rumors spread during the covid-19 epidemic in china: web-based correlation study. *Journal of Medical Internet Research*, 22 (11):e21933, 2020.

[4] K. M. Douglas. Covid-19 conspiracy theories. *Group Processes & Intergroup Relations*, 24(2):270–275, 2021.

[5] A. Giachanou, B. Ghanem, and P. Rosso. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49(1):3–17, 2023.

[6] M. Hoang, O. A. Bihorac, and J. Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196, 2019.

[7] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint arXiv:2309.12247*, 2023.

[8] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[6] https://pixlr.com/image-generator/

[9] J. Langguth, D. T. Schroeder, P. Filkuková, S. Brenner, J. Phillips, and K. Pogorelov. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, pages 1–42, 2023.

[10] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*, 2023.

[11] F. Liu, X. Zhang, and Q. Liu. An emotion-aware approach for fake news detection. *IEEE Transactions on Computational Social Systems*, 2023.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Z. Liu, K. Yang, T. Zhang, Q. Xie, Z. Yu, and S. Ananiadou. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*, 2024.

[14] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, and S. Ananiadou. Emotion detection for misinformation: A review. *Information Fusion*, page 102300, 2024.

[15] Z. Liu, B. Liu, P. Thompson, K. Yang, and S. Ananiadou. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*, 2024. Full version of this paper.

[16] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] A. Miani, T. Hills, and A. Bangerter. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24, 2021.

[18] L. Mompelat, Z. Tian, A. Kessler, M. Luettgen, A. Rajanala, S. Kübler, and M. Seelig. How "loco" is the loco corpus? annotating the language of conspiracy theories. In *Proceedings of the 16th Lingusitic Annotation Workshop (LAW-XVI) within LREC2022*, pages 111–119, 2022.

[19] M. Müller, M. Salathé, and P. E. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6:1023281, 2023.

[20] M. G. Napolitano and K. Reuter. What is a conspiracy theory? *Erkenntnis*, 88(5):2035–2062, 2023.

[21] B. M. Pavlyshenko. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*, 2023.

[22] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[23] Y. Peskine, G. Alfarano, I. Harrando, P. Papotti, and R. Troncy. Detecting covid-19-related conspiracy theories in tweets. *MediaEval*, 2021.

[24] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, and P. Rosso. Definitions matter: Guiding gpt for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, 2023.

[25] K. Pogorelov, D. T. Schroeder, S. Brenner, and J. Langguth. Fakenews: Corona virus and conspiracies multimedia analysis task at mediaeval 2021. In *Multimedia Benchmark Workshop*, volume 67, 2021.

[26] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

[27] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim. Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525, 2022.

[28] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[29] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[30] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

[31] Y. Yanagi, R. Orihara, Y. Tahara, Y. Sei, and A. Ohsuga. Classifying covid-19 conspiracy tweets with word embedding and bert. In *Working Notes Proceedings of the MediaEval 2021 Workshop, Online*, pages 13–15, 2021.

[32] K. Yang, T. Zhang, Z. Kuang, Q. Xie, and S. Ananiadou. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.

[33] C. Yuan, Q. Xie, J. Huang, and S. Ananiadou. Back to the future: Towards explainable temporal reasoning with large language models. *arXiv preprint arXiv:2310.01074*, 2023.

[34] R. N. Zaeem, C. Li, and K. S. Barber. On sentiment of online fake news. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 760–767. IEEE, 2020.

[35] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356, 2023.

[36] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[37] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.

[38] Z. Zheng, L. Liao, Y. Deng, and L. Nie. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*, 2023.