Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

From Past Outbreaks to Future Threats: Detecting Medical Conspiracy Theories With LLMs and Limited Labels

Ipek Baris Schlicht[1,2] | Damir Korenčić[3] | Berta Chulvi[4] | Lucie Flek[5,6] | Paolo Rosso[1,7]

[1]Universitat Politècnica de València, Valencia, Spain | [2]DW Innovation, Bonn, Germany | [3]Division of Computing and Data Science, Ruđer Bošković Institute, Zagreb, Croatia | [4]Universitat de València, Valencia, Spain | [5]Bonn-Aachen International Center for IT, University of Bonn, Bonn, Germany | [6]Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany | [7]ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain

Correspondence: Ipek Baris Schlicht (ibarsch@doctor.upv.es)

Received: 19 April 2025 | **Revised:** 21 June 2026 | **Accepted:** 22 June 2026

Keywords: BERT-like models | conspiracy theory detection | large language models | medical domain | transfer learning | transformers

ABSTRACT

Online dissemination of conspiracy theories (CTs) during epidemics poses significant risks to public health. This paper addresses the problem of detecting CTs in social media posts with an emphasis on the resource-constrained scenarios characterized by the scarcity of labelled datasets and expert annotations, and the lack of computational budget for large-scale LLM inference. To address these challenges, we investigate resource-efficient methods for CT detection across multiple epidemics. We construct a novel dataset of CT-labelled social media posts covering four major epidemics from the past decade: Ebola, Zika, COVID-19 and Monkeypox. We conduct extensive experiments addressing four research questions: (1) the performance of BERT-like models on individual epidemics, (2) the ability to transfer knowledge from past epidemics to new ones, (3) the efficacy of zero-shot classification using Large Language Models (LLMs) and (4) the feasibility of training BERT-like models on LLM-labelled datasets. Our findings indicate that BERT-like models exhibit highly variable performance across epidemics. Transfer learning from prior epidemics can be effective and their performance can be improved with the number of prior datasets. Zero-shot LLM classifiers, including ensemble methods, achieve performance that matches or surpasses that of fine-tuned BERT-like models. Finally, we demonstrate that BERT-like models trained on LLM-labelled datasets achieve results close to the models trained on expert-annotated data, offering a practical alternative when expert labelling is infeasible. While automated methods can be useful for data analysis, we caution against automatization of content filtering due to the inherent difficulty of CT detection and the potential biases of language models.

1 | Introduction

Conspiracy theories (CTs) are a type of misinformation that attempts to explain the reasons for societal events as the outcome of malicious actors with hidden agendas (Douglas and Sutton 2023). Medical CTs focus on health-related topics such as causes and consequences of an epidemic, vaccination and public health policies. The spread of medical CTs can have serious consequences, such as increasing vaccine hesitancy and undermining of public health efforts (Ullah et al. 2021). To make matters worse, societal crises such as pandemics increase the tendency of people to believe in CTs (Van Prooijen and Douglas 2017).

Tools for automatic detection of CTs based on machine learning could help mitigate the negative effects of medical CTs on public health. However, the existing research has focused exclusively on medical CTs related to the COVID-19 pandemic (Langguth et al. 2023) or multi-topic CT datasets (Miani et al. 2022; Phillips et al. 2022). In contrast, we compose and analyse a novel multi-epidemic dataset of social media posts annotated with CT labels. Epidemic-related CTs represent

an important case of CTs since, in addition to their ability to cause health-related harm (Kou et al. 2017; Ullah et al. 2021), they have a potential to become prominent in public discourse. Namely, epidemics can create an environment for an 'infodemic', a rapid spread of health-related misinformation (Chong et al. 2020). Additionally, scientific and public information related to epidemics may remain uncertain for longer periods than for other health-related topics (Kou et al. 2017), which is beneficial to unsupported CT claims.

We also address the need to perform CT detection under two distinct resource constraints. The first is a lack of human annotations. When attempting to detect CTs about an epidemic, labelled data for supervised models is costly, since it requires work of trained human annotators. Existing datasets with CT-labelled texts about the same disease might help, but in the use-case of an emerging new epidemic they are not available. In this scenario, one can attempt to use datasets from previous epidemics. Regardless of whether the epidemic is emerging or not, one can attempt to use zero-shot methods to avoid labelling and Large Language Models (LLMs) have shown potential for zero-shot text classification (Ziems et al. 2024).

The second constraint is a budget-limit for large scale of LLM inference. Accessing LLMs via commercial API interfaces can be expensive when processing large datasets, and owning an LLM infrastructure requires non-trivial investment. In this setting, LLMs may still be useful as annotators for creating training data, while smaller BERT-like models can be used for large-scale inference.

Our experiments examine all of the resource-related concerns on datasets from multiple epidemics. We frame the CT detection problem as a binary classification of texts into *CT* and *not-CT* classes, and approach the classification problem using BERT-like models (Devlin et al. 2019) and LLMs (Zhao et al. 2023). We construct a dataset of epidemic-related posts from X (previously Twitter) that covers four recent epidemics: Ebola, Zika, COVID-19 and Monkeypox. We focus on social media posts, as they are a prominent tool for spreading misinformation and CTs (Vosoughi et al. 2018). We consider the standard use-case of training and testing the supervised BERT-like models on the posts from the same epidemic. Next we address the use-case of an emerging epidemic by training BERT-like models on CT-labelled post from previous epidemics. We proceed with extensive experiments on zero-shot LLMs for CT detection. Finally, we address the scenario where large-scale use of LLMs is not feasible by using LLMs as annotators that create training data for BERT-like models. Figure 1 shows the workflows of the experiments with cross-epidemic transfer, LLM-based classification and the use of LLMs as annotators.

Consequently, our paper focuses on four main research questions (*RQs*):

- *RQ1*. How well do BERT-like models classify conspiracy posts about different epidemics?
- *RQ2*. How well do BERT-like models transfer knowledge about medical CTs from previous to emerging epidemics?

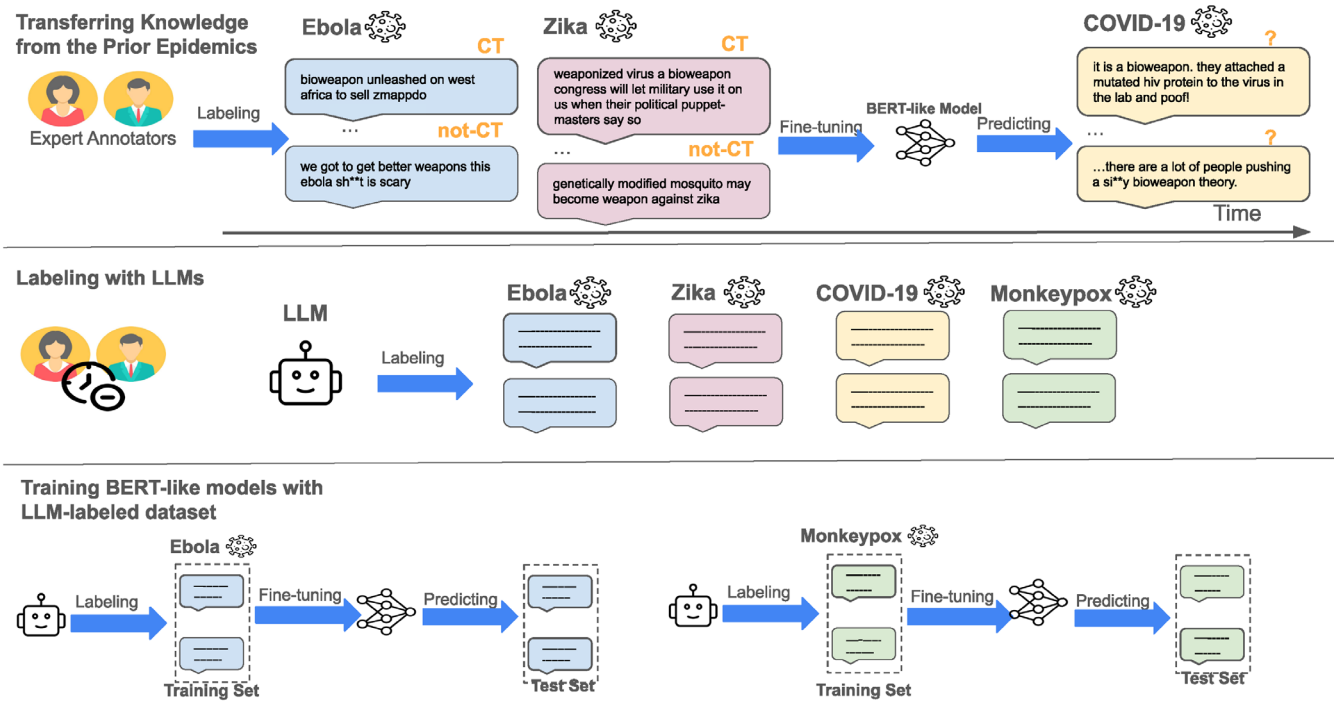

FIGURE 1 | Overview of the experiments that we examined for detecting CTs with limited resources. *Transferring knowledge from the prior epidemics*: During an ongoing epidemic (e.g., COVID-19), only available knowledge comes from past epidemics (e.g., Zika and Ebola) labelled by expert annotators (e.g., fact-checkers, researchers, journalists). BERT-like models are fine-tuned on the past epidemic datasets for predicting CTs about the ongoing epidemic. *Labelling with LLMs*: The expert annotators are unavailable, LLMs are utilized for labelling datasets. *Training BERT-like models with LLM-labelled dataset*: There are limited resources for inference with LLMs and experts are unavailable, so the LLM-labelled datasets are used for training the BERT-like models.

- *RQ3*. How well do LLMs detect medical conspiracies using zero-shot prompts?
- *RQ4*. How well do BERT-like models identify medical CTs when they are trained on data annotated by LLMs?

The basis of our experiment is a novel dataset covering four prominent epidemics from the last 10 years, and it contains 10,258 posts labelled as CT or not-CT. To construct the dataset, we started from existing unlabelled corpora containing posts related to three epidemics: Ebola (Tamine et al. 2016), Zika (Pruss et al. 2019) and Monkeypox (Thakur 2022). To add COVID-19 data, we re-used the already annotated COCO dataset (Langguth et al. 2023). In the next step, we sub-sampled the Ebola, Zika and Monkeypox datasets in order to increase the number of CT posts, and annotated the data in a manner consistent with the COCO corpus. The resulting dataset is, to the best of our knowledge, the only multi-epidemic dataset with CT labels.

In the first experiments BERT-like models were trained and evaluated, for each epidemic, using train-test splits of texts from the same epidemics (*RQ1*). The results show that the models' performance varies widely for different epidemics, ranging from $F1$ of approx. 58 for Ebola to $F1$ of approx. 81 for COVID-19.

The second experiment, depicted in the top of Figure 1, examines transferring knowledge, in the form of labelled data, from the prior epidemics by using BERT-like models (*RQ2*). We use COVID-19 and Monkeypox data for testing the models, which are fine-tuned on data from Ebola and Zika. This was done to faithfully simulate the 'new epidemic' scenario, since the models we used (BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019) and XLNet (Yang et al. 2019)) are pretrained on data from a period that overlaps with the period of Ebola and Zika, but not with the period of COVID-19 and Monkeypox. The experiments demonstrate that training on previous epidemics can lead (in the case of Monkeypox) to performance close to that of in-domain training, and that having access to a larger number of previous datasets increases performance.

The third experiment, depicted in middle of Figure 1, examines how well LLMs can predict human labels (*RQ3*). For this purpose, we experimented with several popular LLMs including GPT-4o (OpenAI et al. 2024) and Llama3-70b (Grattafiori et al. 2024). We used zero-shot prompting to elicit the predictions from the LLMs, with the key part of a prompt being a definition of a conspiracy theory. One set of prompts was based on a simple, general and short definition, while the other set uses a longer expert definition that was used for data annotation. In addition, we experimented with ensembles of zero-shot LLM classifiers. The experiments show that LLMs can perform competitively, matching or even slightly surpassing the fine-tuned models. While the performance of individual LLMs is unstable across datasets, the ensemble of LLMs consistently leads to good performance.

Lastly, the fourth experiment, depicted at the bottom of Figure 1, examines how well BERT-like models can learn from data labelled by zero-shot LLM classifiers (*RQ4*). In essence, this experiment performs knowledge distillation (He et al. 2021), by attempting to distil task-related LLM knowledge into BERT-like models, which are much cheaper to run, especially on large datasets. For all datasets except Zika, distilled BERT-like models perform slightly below or comparably to in-domain models, which demonstrates the feasibility of the approach for the problem of CT detection.

It must be underlined that CT detection is in general a hard problem—the best models reach $F1$ scores of approximately 65, 70 and 72 for Ebola, Zika and Monkeypox, respectively. When this is combined with the potential for different type of biases embedded in language models (Gallegos et al. 2024), one must be cautious when applying automatic CT detection models. Specifically, we do not advocate the use of CT detection models without human supervision for the use-case of content filtering (human-centred AI). However, even imperfect models might prove useful for data analysis.

The main contributions of this paper are:

1. Construction of a novel dataset of social media posts annotated for conspiracy, which spans four prominent epidemics of the last decade. The dataset enables the benchmarking of language models across distinct diseases in order to evaluate their robustness and generalizability, as well as the benchmarking of models' potential for transfer learning across epidemics. To the best of our knowledge, this is the first dataset of this kind.
2. A comprehensive set of experiments that investigates cross-epidemic transfer learning with BERT-like models, and the potential of LLMs for both zero-shot CT detection and for data annotation. The experiments show, among other findings, that ensembles of LLMs produce conspiracy labels that align well with expert annotations, and that BERT-like models trained on LLM-labelled datasets can perform comparably to the models trained on expert-labelled datasets.
3. Analyses of data and of the experimental results related to CTs across multiple epidemics. These include the analysis of conspiracy-related terms, the evidence that the success of cross-epidemic transfer is determined by the similarity of the CT narratives of different epidemics and the error analysis that indicates that models tend to misclassify texts containing political criticism and satirical and ironic language. Lastly, our source code, the IDs of the tweets and the trained models are publicly available.[1] The full texts of the available tweets can be obtained via the official X API.[2] Additionally, the scripts and the pipeline in our source code can be adapted to collect other similar social data.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 describes the multi-epidemic dataset. We give the details of its construction, labelling and share exploratory analysis on the final dataset. In Section 4, we present the details of BERT-like and LLMs that we used for the experiments and other implementation details, followed by the results and analysis in Section 5. Finally, we discuss broader implications and limitations in Sections 6 and 7 and conclude with future directions in Section 8.

2 | Related Work

2.1 | Health Misinformation Detection

Several studies have focused on the automatic detection of health misinformation, particularly in the context of COVID-19 misinformation on social media. Before the emergence of LLMs, researchers employed a range of approaches, from traditional machine learning methods to BERT-like models. Early methods relied on classical machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM) and Random Forests, often using linguistic, affective features or medical information (Mukherjee et al. 2014; Kinsora et al. 2017; Ghenai and Mejova 2018; Di Sotto and Viviani 2022). Some approaches such as (Sharifpoor et al. 2025; Du et al. 2021) utilize in Convolutional Neural Networks.

With the introduction of transformer architectures, BERT-like models became the dominant approach. Models such as BERT and RoBERTa demonstrated significant improvements in misinformation detection. These models have been widely used in health misinformation detection (Medina Serrano et al. 2020; Alam et al. 2021; Dharawat et al. 2022). For a comprehensive overview of pre-LLM approaches to health misinformation detection, including a taxonomy of tasks, methods and datasets, readers can refer to the survey by Schlicht et al. (2024).

LLMs can reduce the need for human annotations and can be more generalize more effectively across the domains for misinformation detection (Chen and Shu 2024). These opportunities have attracted the researchers to explore LLMs such as GPT-family models, for instance GPT-3, GPT-4 for detection of political misinformation (Buchholz 2023; Pelrine et al. 2023). In the medical domain, Thapa et al. (2024) examined several LLMs for identifying misinformation about COVID-19 and Monkeypox. GPT-4 performed better than other models across datasets and different experiment settings.

2.2 | Conspiracy Theory Detection

Medical CTs are category of health misinformation (Schlicht et al. 2024). CT detection in texts has become a significant area of research, particularly with the rise of misinformation and disinformation and the widespread reach of social media. Related studies have focused on building datasets for detecting CTs across various platforms and domains, as well as evaluating the performance of BERT-like models and LLMs in identification of conspiratorial texts.

As for datasets, for instance, Phillips et al. (2022) developed a Twitter dataset covering Global Warming, Epstein-Maxwell and COVID-19 topics. Similarly, Langguth et al. (2023) analysed conspiracy narratives about COVID-19 on Twitter. Other works have constructed LOCO which consists of news articles from mainstream and conspiracy websites (Miani et al. 2022) and performed textual analysis on CTs from conspiracy subreddits on Reddit (Samory and Mitra 2018; Klein et al. 2019). More recent datasets have been constructed from groups on Telegram (Pustet et al. 2024; Korenčić, Chulvi, Casals, et al. 2024) and video sharing platforms such as Tiktok (Corso et al. 2025) and YouTube (Medina Serrano et al. 2020). Although most datasets focus on CTs about politics or COVID-19 as medical domain, we constructed our dataset from multiple epidemics to investigate the cross-epidemic transfer of CTs in medical domain.

When constructing CT datasets, large quantity of content on social media platforms and low prevalence of conspiratorial posts (Langguth et al. 2023) create a need for data filtering and re-balancing methods. Such methods, commonly relying on keyword-based filtering (Langguth et al. 2023) and re-ranking (Korenčić, Chulvi, Casals, et al. 2024), introduce the risk of selection bias (Ousidhoum et al. 2020). Dataset creators commonly acknowledge the potential for selection bias (Klein et al. 2019; Miani et al. 2022; Phillips et al. 2022; Langguth et al. 2023), and sometimes quantify biases in the final dataset against available metadata such as temporal (Miani et al. 2022; Corso et al. 2025) or geographic data (Langguth et al. 2023). Since our data processing pipeline also contains filtering, we examine the differences between the filtered and unfiltered datasets by means of visual analysis of semantic embeddings, and we also analyse the effect of the pipeline on the recall of CTs.

Early studies of textual CT detection have employed lexical and linguistic analyses for comparing conspiratorial texts with non-conspiratorial texts by using Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker 2010) and Empath (Fast et al. 2016) lexicons. Samory et al. (Samory and Mitra 2018) analysed conspiracies from r/conspiracy subreddit in Reddit. They proposed agent-action-target triples to cluster narrative motifs that highlight commonalities across different CTs. BERT-like models have emerged as a popular method to identify CTs in texts. Several studies (Phillips et al. 2022; Corso et al. 2025) demonstrated that RoBERTa (Liu et al. 2019) has a superior performance in identifying CTs. George et al. (2024) extended RoBERTa by incorporating moral foundation scores and emotions to enhance conspiracy detection, fine-tuning the model on the multi-domain CT dataset (Phillips et al. 2022). Fort et al. (2023) investigated transferability of CTs from one domain to another domain (e.g., from vaccine to climate change) on LOCO. The researchers experimented with SVM and BART by using a random split of training/validation/test sets from LOCO. Additionally, they applied a word bleaching method that replaces domain-specific words with their corresponding POS tags to enhance the generalizability of the models. In our transferability experiments, we focus on social media texts and transferability among epidemic-related datasets. Unlike the random split for training models, we evaluate models in a more realistic scenario by using data from newer epidemics as test sets, which are not included in training datasets that BERT-like model pretrained with. Older epidemic datasets are utilized for fine-tuning and validation to ensure that the models have no prior information about the test sets.

2.3 | Large Language Models as Annotators

LLMs (Zhao et al. 2023) have been used in Computational Social Sciences (CSS) both as classification and annotation tools. A large study (Ziems et al. 2024) on 20 CSS classification

tasks found that fine-tuned BERT-like models often outperform zero-shot LLMs in CSS classification tasks. A study (Bucher and Martini 2024) performed on 4 classification datasets found that BERT-like models always outperform zero-shot LLMs, often by a large margin. A study of LLMs on 6 CSS classification tasks found that BERT-like models tend to outperform (Mu et al. 2024) LLMs configured with a variety of zero- and few-shot prompts.

Regarding conspiracy-oriented classification, an experiment on detection of conspiratorial German Telegram texts (Pustet et al. 2024) showed that zero-shot LLMs can perform well, with performance comparable to that of fine-tuned models. Similarly, a study (Corso et al. 2025) on detecting conspiratorial TikTok video transcripts also demonstrated the competitiveness of zero-shot LLMs. However, several experiments (Pogorelov et al. 2021; Peskine et al. 2023) examined a more challenging task of multi-label classification of fine-grained conspiracy theories and found that LLMs are significantly outperformed by the fine-tuned models.

Previously listed studies show that it is hard to generalize about performance of zero-shot LLMs on CSS-related classification tasks (Ziems et al. 2024; Bucher and Martini 2024; Mu et al. 2024; Pustet et al. 2024; Corso et al. 2025). Namely, even in the studies that found LLMs to be mostly or always inferior, LLMs can either outperform the fine-tuned models in a number of benchmarks, or closely match their performance. Therefore, it is accurate to say that the effectiveness of a zero-shot LLM is highly dependent on both the dataset and the classification problem. Our experiments investigate in-context learning with LLMs on an unexamined and important problem of CT detection during an emerging epidemics, and provide related data points and application guidelines.

3 | Epidemic Datasets

We selected four epidemics for our experiments: Ebola (Jacob et al. 2020), Zika (Chang et al. 2016), COVID-19 (Ciotti et al. 2020) and Monkeypox (Beer and Rao 2019). These diseases had widespread occurrences across multiple countries and garnered significant attention on online social communities. Given the popularity of Twitter as a platform for social data analysis, disease-related datasets that we found were constructed from Twitter. The datasets of Ebola, Zika and Monkeypox contain solely tweet ids. Therefore, we fetched the full content of the tweets by applying Hydrator.[3]

We present the details of the datasets as follows:

- *Ebola [2014–2016]*: The 2.7 million samples were collected using the keyword 'ebola' (Tamine et al. 2016). We obtained the full content of 58.83% of the samples.
- *Zika [2015–2016]*: The dataset (Pruss et al. 2019) was constructed by searching for keywords related to Zika, such as 'zika' and 'ZIKV'. It is a multilingual dataset and consists of approximately 6.9 million English tweets. We obtained 63.62% of the samples.
- *COVID-19 [2019–present]*: We used COCO (Langguth et al. 2023) for our experiments, since it is a diverse dataset about COVID-19 conspiracies. The dataset was annotated as multi-class conspiracy detection. Since we focus on binary classification, we unified the labels into two classes based on the annotation task described in Section 3.2.
- *Monkeypox [2022–2023]*: The dataset (Thakur 2022) was constructed by searching tweets containing 'monkeypox' or 'monkey pox'. We were able to fetch 494k tweets from the dataset.

3.1 | Dataset Sampling

Although COCO is already labelled, the collections for Ebola, Zika and Monkeypox are unlabelled and large. Manual annotation of these collections is a costly and time-consuming task. Typical methods for constructing misinformation datasets are searching for related keywords and using claims from authoritative sources as queries (Schlicht et al. 2024). However, relying on keyword searches can be challenging, especially when the keywords are disease-specific and require laborious interdisciplinary work. Additionally, building a conspiracy dataset poses a challenge due to the infrequent occurrence of CT samples (Langguth et al. 2023). Several strategies were proposed to mitigate these challenges. One is to employ the credibility websites such as Media Bias Fact Check (MBFC),[4] a website for assessing publishers, as a labelling resource (Stefanov et al. 2020; Sakketou et al. 2022). Yang et al. (2021) propose a framework for identifying claims automatically which involves several steps, including embedding tweets using sentence transformers (SBERT) (Reimers and Gurevych 2019), community-based clustering and summarization. Rahman et al. (2021) used a set of machine learning methods and active learning to select tweets for annotation in hate speech detection.

We use an approach that combines elements from Sakketou et al. (2022) and Yang et al. (2021) in order to obtain a diverse collection of conspiracy examples. In our approach we tackle the issue by curating claims from multiple sources and utilizing them as queries to fetch similar tweets using a semantic search engine.[5] For each unlabelled epidemic dataset, we constructed search queries composed of conspiratorial texts from different sources and searched the indexed collections from the dataset for tweets that are semantically similar to these queries by the search engine. Subsequently, we randomly sampled the retrieved tweets for annotating the dataset.

3.1.1 | Indexing

To construct search engine indices for each unlabelled collection, we initially cleaned the unlabelled collections from the unrelated samples. The cleaning step involves of removing duplicates, eliminating samples written in languages other than English from the collection, selecting the posts written by a user who have a minimum of 5 unique posts about a given disease assuming that these users are actively engaged in the discussions. After this step, we split each cleaned collection into two categories: samples citing external links and those that do not. To create a collection similar to COCO, we follow the approach

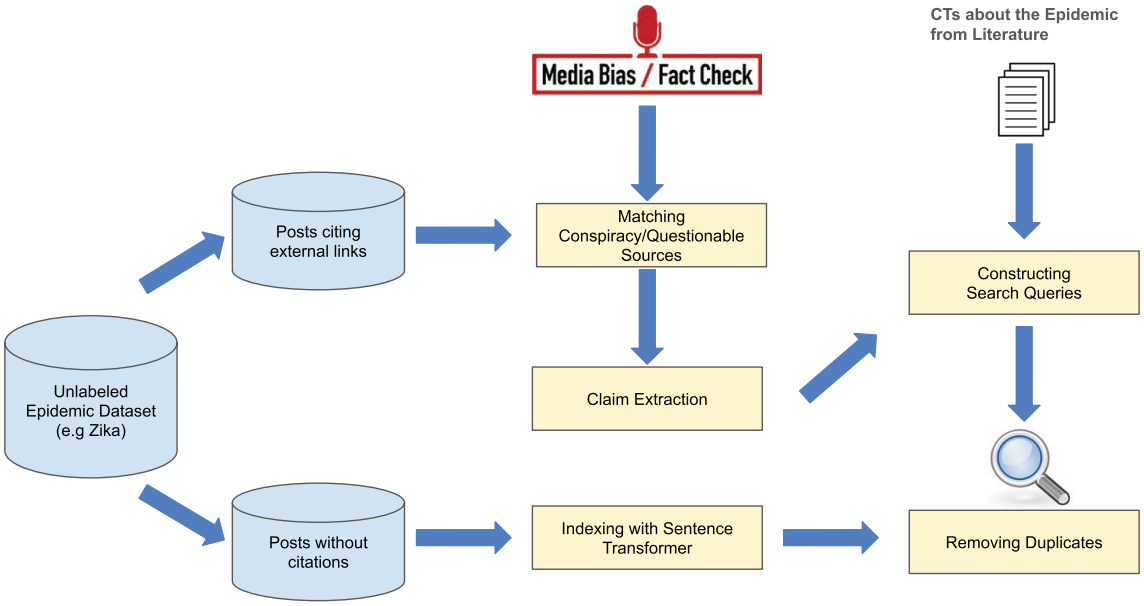

FIGURE 2 | Construction of search queries for annotating a collection about one epidemic.

described in its paper (Langguth et al. 2023) that includes only tweets without links for conspiracy analysis since those tweets reflect user intent. After applying preprocessing on the selected tweets, we embed them with Sentence Transformer (Reimers and Gurevych 2019) for obtaining their semantic representations. Finally, the embeddings and cleaned samples were subsequently indexed into the search engine.

3.1.2 | Search Query Construction

As illustrated in Figure 2, for each disease, we use two sources to construct search queries. In this way, we aim to obtain more conspiracy samples and to prevent information transfer between training and test sets. These sources are as follows:

- *Conspiracy/Questionable Sources*: To identify conspiracies that have the potential to reach a broad audience, we identified questionable or conspiracy publishers from the tweets citing external links. We first extracted the domain names from these links and compared them against the database from MBFC by using fuzzy match.[6] Next, we selected the samples that contain domains labelled as conspiracy-pseudoscience or questionable sources by assuming that they may share conspiracy news. Lastly, to get candidate queries for the search engine, we adopted the automatic claim identification framework (Yang et al. 2021) by excluding the summarization step for clustering tweets and identifying irrelevant samples. The final queries are the tweets detected by the framework.
- *Literature*: As the last source of conspiracies, we investigated the literature on conspiracies seen during Ebola (Ouattara and Århem 2021; Feuer 2014), Zika (Klofstad et al. 2019) and Monkeypox (Zenone and Caulfield 2022) to identify disease-specific conspiracies. These research studies provide a list of conspiracies, we use the lists with respect to a disease collection for constructing queries. The claims can be found at our code repository.

Once we had the final search queries for each disease, we retrieved the top-10 similar samples by using the search engine and removed duplicates on the corresponding unlabelled corpora. The subset of the resulted collection was labelled by in-house and external annotator (details will be given in Section 3.2).

We present the statistics of the unlabelled collections for each aforementioned step to construct the search queries in Table 1. Due to the size of Ebola and Zika, the number of claims from publishers were higher than Monkeypox, however we found more claims about Monkeypox CTs in the literature.

3.2 | Dataset Labelling

Following the prior studies (Douglas and Sutton 2023; Langguth et al. 2023), we define CTs as complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful and malicious groups. In this study, our focus is to investigate how well can language models detect CTs circulating on social media during epidemics. We formulate the task as a binary classification task.

Annotators received the following precise instructions created by a social psychologist:

- Social media posts that explicitly or implicitly support a CT are labelled as CT.
- All other posts, including those that mention or criticize conspiracy theories, are labelled as not-CT.

A post explicitly supports a conspiracy theory if it claims that a secret agent or group with a hidden agenda (and usually with a malicious intent) is responsible for an event. Not all elements (agent, agenda, intent) need to be explicitly present, but some should be. For example: 'Bill Gates is behind all of this' (agent), 'Of course this massive vaccination is done to plan mind-control

microchips into people' (agenda) and 'Those awake know that the endgame of this whole charade is a new world order' (intent). A post implicitly supports a conspiracy theory if it suggests these elements without stating them directly, such as using the term 'plandemic' to describe the situation.

Based on these instructions, we constructed gold-standard datasets for Ebola, Zika and Monkeypox. The annotation was performed by two authors and one research assistant from the Computer Science department of the Universitat Politècnica de València, all of them with previous experience of working with CT texts. Any uncertainties that occurred during the annotation process were resolved through team discussions.

Next, we calculated Fleiss' Kappa scores (Fleiss 1971) by using Pyirr[7] to measure agreement among the annotators. As shown in Table 2, we obtained substantial agreement on the Zika and Ebola datasets and moderate agreement on the Monkeypox dataset. We consider these levels of agreement as acceptable given the complexity of the task, and comparable to the agreement reported in prior work on CT detection (Phillips et al. 2022). Through discussions during and after the annotation process, we identified that the task is inherently subjective and challenging. A prominent challenge is distinguishing between conspiracy content and politically critical content, often containing rhetorical devices such as irony and satire, particularly in the Monkeypox dataset. Additionally, annotating very short tweets (frequent in Ebola and Zika datasets) is difficult because the context is often unclear, with cryptic or obscure references. Language can also be unclear in such tweets, containing errors, slang and abbreviations. These tweets are difficult to parse and understand, which adversely affects the quality of the final labelling decision.

The examples of the difficult cases are presented in Table 3. We attempted to resolve decisions about labelling such tweets by carefully parsing the text and trying to understand the semantics, on a per-case basis. Of all the Monkeypox tweets in Table 3, the first two were labelled as CT since they seem to imply a group (someone within WHO) with a hidden agenda (elections or profit), while the rest was considered as legitimate political criticism. We note that considering the first two tweets as legitimate criticism is also a plausible interpretation.

We determined the gold labels using majority voting, which also served as a mechanism for resolving any remaining disagreements among the annotators. This approach aligns with the methodology used in the COCO dataset, where each sample was annotated by three in-house annotators, and the final label was assigned based on the majority vote (Langguth et al. 2023).

3.3 | Final Dataset

The details of the final dataset are shown in Table 4. The time coverage of the datasets does not overlap, which makes them suitable for the transfer learning experiments. The average text lengths of the datasets for earlier diseases are shorter than COVID-19 and Monkeypox, since the character limit of Twitter was expanded in 2017.[8] The COCO texts are on average more lengthy than the Monkeypox texts. The samples were written by many social media users, resulting in diverse text styles. Furthermore, readability scores of Zika and Monkeypox are lower than COCO that, which might make CT detection on these datasets challenging.

TABLE 1 | The statistics of the unlabelled collections for each step of search query construction.

	Ebola	Zika	Monkeypox
# of Raw samples	1,607,411	4,446,572	493,969
# of Tweets citing links	864,179	2,295,020	90,857
# of Tweets citing questionable sources	5294	13,749	2421
# of Tweets citing conspiracy sources	11,836	137,353	591
# of Tweets without links (after cleaning)	45,600	142,774	62,163
# of Claims from publishers	208	323	29
# of Claims from literature	15	10	31

TABLE 2 | According to Fleiss' Kappa (Fleiss 1971), we obtained substantial agreements on the Ebola and Zika datasets and moderate agreement on the Monkeypox dataset.

Dataset	Fleiss' Kappa	Explanation
Ebola	60.56%	Substantial
Zika	63.80%	Substantial
Monkeypox	57.57%	Moderate

Table 4 shows that the COVID-19 COCO dataset is more balanced than the texts related to other epidemics, with the number of CT and not-CT tweets being almost equal. Since, the dataset authors (Langguth et al. 2023) compiled a list of conspiracy-related keywords to filter the collected tweets and get more CT tweets. During the data collection period the authors developed the list which was based on previous knowledge, widely discussed conspiracy keywords and data analysis of the tweets (Langguth et al. 2023). As discussed earlier, in our case of filtering large datasets related to three previous epidemics the laborious process of developing keyword lists was not feasible. Therefore, we opted for automatic techniques based on semantic similarity with conspiratorial and misinformation content.

The dataset creation pipeline is designed to increase the prevalence of rare CT tweets (Section 3.1), which raises the question of its effect on the recall of CTs. Additionally, the filtering techniques that were used could potentially introduce selection biases and make the final datasets non-representative of the original stream of social media posts. To quantify recall and characterize the selection bias, we sampled 100 tweets

TABLE 3 | Examples of posts that annotators found challenging to label due to their unclear language, unclear context, or because of the conspiracy theory versus political criticism ambiguity.

Epidemic	Example	Challenge
Ebola	Every article on the transmission of EBOLA, reminds one about Dr. —'s theory on how HIV/AIDS was transmitted.	Unclear context
Ebola	Dis got me wondrin; so —'s Wife n d Stupid Liberian Minister knew — had Ebola? Which means they purposely brot Ebola to Nigeria??	Unclear language, unclear context
Ebola	fb bigwigs are making to mainstream media interviews and you wonder why ebola is killing Africans	Unclear language, unclear context
Zika	dere cn b exceptions bt zika was created n a lab by da US, just like AIDS virus	Unclear language
Zika	Dems want taxpayers to pay for Planned Parenthood aka baby killing machine via Zika funding. We will not be part of Dems murders	Political criticism, unclear context
Monkeypox	Well, the WHO has just declared Monkeypox a global health crisis...just in time for the midterms	Political criticism
Monkeypox	I hear the WHO wants to remain Monkeypox to combat racism; stigma against monkeys. I stand by their decision and believe they need to call it Monkeypox without the K, Moneypox. Keeping the name relevant to what the virus is all about	Political criticism
Monkeypox	The cdc allocated 60 doses in their most recent allotment of monkey pox vaccines to Baltimore city. 60 doses. They want us to suffer.	Political criticism
Monkeypox	Y'all pushed this 'Monkey pox is a gay/bi men disease ONLY' rhetoric so much that you WILL, mark my words, have straight men catch MP and to avoid being seen/called gay hide the fact they have it + not get vaccinated and continue to infect many others in their vicinity.	Political criticism
Monkeypox	WHO declared MONKEYPOX VIRUS outbreak PUBLIC HEALTH EMERGENCY OF INTERNATIONAL CONCERN! HIGHEST LEVEL!I watched live...one great concern 'UNCERTAINTY'! I AGREE! LET ME HELP! ...I included the 200,000,000 BUBONIC PLAGUE GLOBAL DEATH TO the 600,000,000 HUMAN DEATHS prevented 'IF'!	Political criticism, unclear language

Note: The names mentioned in the tweets are masked with —. For example, the first tweet mentions a name of a conspiracy theorist.

TABLE 4 | The details of the final epidemic dataset where COVID-19 is the COCO dataset (Langguth et al. 2023).

			Samples			Users		Readability	
Dataset	Time coverage	Text length	%CT	CT	Not-CT	CT	Not-CT	All	CT
Ebola	2014.07–2014.08	103.08	4.67	142	2898	118	1678	64.33	69.13
Zika	2015.11–2016.10	96.60	9.04	225	2263	170	1341	61.21	60.35
COVID-19	2020.01–2021.07	256.90	50.91	1753	1690	N/A	N/A	67.09	68.12
Monkeypox	2022.05–2022.10	160.31	19.43	252	1045	226	830	66.72	64.84

Note: CT stands for conspiracy theory, not-CT stands for non-conspiracy theory. The readability scores were calculated based on Flesch Reading Ease (Kincaid et al. 1975).

from the tweets discarded by the pipeline, for each of the three collections (Ebola, Zika and Monkeypox). Two of the authors annotated the samples and settled labelling disagreements by discussion.

Newly annotated samples enable us to approximate the pipeline's recall and its effect on the prevalence of CTs. The statistics, displayed in Table 5, show that the prevalence of CTs among pipeline-retrieved tweets is over three times higher than their prevalence among discarded tweets. The pipeline-retrieved percentages of CTs are 4.7%, 9.0% and 19%, respectively, for Ebola, Zika and Monkeypox, while for discarded tweets these percentages are 1%, 3% and 6%. For Ebola, Zika and Monkeypox the estimated pipeline recall is 23.8%, 5% and 6.7%, respectively, while random recalls stand at 6.3%, 1.7% and 2.1%. Due to the relatively small size of the labelled datasets in comparison to the large initial collections (Table 1), total recall of CT tweets is expected to be small. However, previous statistics demonstrate that both the

TABLE 5 | Recall analysis across epidemics: proportion of CTs in the labelled datasets created using the pipeline, proportion of CTs in the sample of data discarded by the pipeline, estimated total CT recall of the pipeline, expected CT recall when creating datasets with random sampling (with sample size equal to the dataset created using the pipeline).

Epidemic	CT—pipeline (%)	CT—discard (%)	Estimated recall	Random recall (%)
Ebola	4.67	1.00 [0.00, 5.40]	23.80 [5.40, 63.80]	6.30
Zika	9.04	3.00 [1.00, 8.50]	5.00 [1.80, 13.30]	1.70
Monkeypox	19.43	6.00 [2.80, 12.50]	6.70 [3.10, 12.70]	2.10

Note: 95% confidence intervals are calculated using the Wilson score interval (Wilson 1927).

recall of CTs and their proportion in the learning dataset are increased by the pipeline. Without applying the pipeline, the datasets would be even more imbalanced and they would contain a much smaller subset of the CT tweets from the original data.

The proposed pipeline could potentially introduce biases and skew the original data and class distribution. Ideally, we would want both the subset of not-CT and the subset of CT tweets in the final collections to be representative of the corresponding subsets in the original data. In order to examine potential biases, we perform an analysis of semantic embeddings of tweets using a high-performing `gte-large` model for semantic embedding of short texts (Li et al. 2023). For each disease, we use UMAP to project, to a common 2D space, all the labelled tweets (the final dataset), the tweets randomly sampled from the unlabelled (discarded) collection and the remainder of the unlabelled tweets.

The projections, displayed in Figure 3, show that the discarded tweets generally align with the labelled tweets in the semantic space. This alignment is observed for both CT and not-CT tweets. For Ebola, the space of the discarded tweets (green, grey) is very well covered by the pipeline-selected ones (pink). Only a single Ebola CT tweet was found in the sample of discarded tweets, and it is not an outlier—there are pipeline-selected CT tweets more isolated from main CT clusters. For Zika, the space of all discarded tweets is well covered by the pipeline-selected tweets, and all the discarded CT tweets are in the vicinity of pipeline-selected CT tweets. For Monkeypox, there is a subspace of discarded tweets (left side) that is relatively sparsely covered by the pipeline-selected tweets. For the rest of the discarded tweets the coverage is improved. The CT tweets from the discarded subsample are well covered, with all the tweets being in the vicinity of or within clusters of pipeline-selected CT tweets. The visual analysis of the semantic space indicates no large discrepancies between the semantic spaces of discarded and pipeline-selected tweets, for both the CT and not-CT tweets.

3.4 | Exploratory Analysis

In order to explore characteristics of the epidemic datasets we use Scattertext (Kessler 2017), an open-source Python library for visualizing linguistic variation between document categories. For all four datasets we visualized association scores between terms (unigrams and bi-grams) and CT and not-CT categories. Scattertext uses Pointwise Mutual Information (Church and Hanks 1989) (PMI) as the measure of association. The scatterplots for all datasets are presented in Figure 4. The plots also show the terms that are characteristic for a dataset, in contrast to Brown corpora (Francis and Kucera 1979) which Scattertext (Kessler 2017) uses to represent generic English words, based on a term's rank in the list of words ordered by frequency. The measure used to distinguish characteristic terms is the difference between the rank within the dataset and the rank within the generic corpus. For each dataset, the plot is accompanied by ranked lists of terms most associated with the conspiracy and non-conspiracy classes, and the list of characteristic terms. Upper left corner of each plot contains words characteristic of conspiracy texts, that is, words associated with the CT class and not associated with the not-CT class.

As can be seen in the plots, characteristic words associated with Ebola CTs include 'population', 'weapon', 'patent', 'white people' and 'big pharma'. For Zika CTs, characteristic words include 'foundation', 'Rockefeller', 'agenda', 'created' and 'hoax'. The COVID-19 CTs are characterized by terms like 'patent', 'globalist', 'UN', 'foundation' and 'fake pandemic', while the Monkeypox CTs are characterized by 'Bill Gates', 'plandemic', 'money', 'research' and 'test'. Regarding the characteristics of the datasets, for all the diseases except COVID-19, the characteristic terms are mostly related to the disease itself, whereas for COVID-19, they are predominantly linked to conspiracy theories. This is likely due to the large proportion of conspiracy-related texts in the COVID-19 dataset. When analysing the commonality across the datasets from the perspective of a narrative motif (a story where agents and patients are connected by an action), we observe the recurring characterization of CTs by terms related to an agent which is a power structure. For example, we see 'deep state' and 'nwo'[9] in COVID-19 CTs, 'big pharma' in Ebola CTs, 'Rockefeller' and 'Bigpharma' in Zika CTs and 'Billgates' in Monkeypox CTs. However, the commonality across not-CTs in the four datasets are terms which from the perspective of a narrative motif refer to a patient or an object that suffers the effects of the disease: 'patient', 'infected' and 'hospital' in Ebola not-CTs, 'quarantine' and 'life' in COVID-19 not-CTs, 'pregnant', 'microcephaly' and 'health' in Zika not-CTs, and 'cases', 'children' and 'gays' in Monkeypox not-CTs. The top-right corner of the plots primarily contains stop-words and commonly used generic terms. The bottom-right corner consists of event-specific or medical terms which likely reflects standard public health discourse.

In addition to the Scattertext analysis, we examine the semantic structure of CT and not-CT texts across epidemics using SBERT embeddings and UMAP projection. Figure 5 shows the semantic projection of texts from all four epidemics. The centroids on the visualizations indicate the average position of each dataset. While

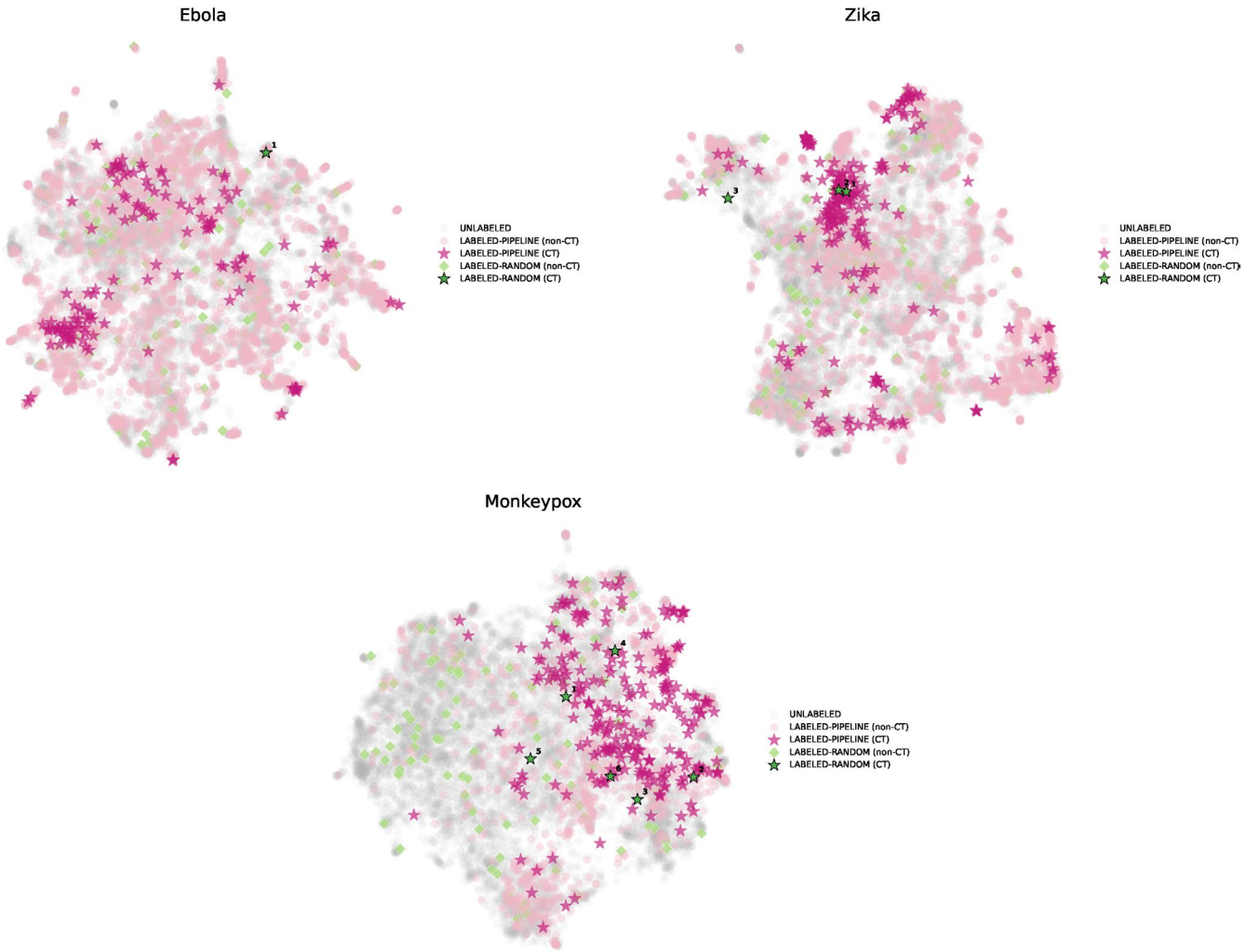

FIGURE 3 | UMAP projections of embedded tweet using SBERT (`thenlper/gte-large`), for Ebola, Zika and Monkeypox (from top to bottom). The 'pipeline' data-points correspond to tweets from the final collection created using the pipeline and labelled by humans. The 'random' data-points correspond to a labelled subsample of tweets from the collection discarded by the pipeline. The 'unlabelled' tweets correspond to the rest of the discarded collection.

texts cluster broadly by epidemic, CT texts of different epidemics tend to overlap more than not-CT texts of different epidemics. Specifically, CT texts about COVID-19 overlap with Monkeypox more than with the CT texts of other epidemics. Ebola and Zika CTs are semantically closer to Monkeypox CTs than to COVID-19 CTs. In contrast, not-CT texts tend to be more specific to each epidemic, with each epidemic having a significant number of texts that are not overlapping with texts of other epidemics.

4 | Models

Here we describe the language models used to perform CT detection in subsequent experiments. For both BERT-like models and LLMs we describe the choice of models and their application to the classification problem.

4.1 | BERT-Like Models

BERT-like models (Devlin et al. 2019) based on the transformer architecture (Vaswani et al. 2017) have since their inception proved as powerful classifiers (Devlin et al. 2019) that are becoming increasingly cost-effective due to the raising prevalence and power of GPUs. For problems related to CSS they tend to outperform LLMs in the majority of use-cases (Ziems et al. 2024). Regardless of the relative performance, BERT-like models can be still preferred by fact-checking or health organizations due to advantages such as much lower development and deployment costs. Additionally, data privacy is ensured when the BERT-like models are deployed internally, which is in contrast with LLMs such as ChatGPT-3.5 available only through web APIs.

We experiment with BERT (Devlin et al. 2019), XLNet (Yang et al. 2019) and RoBERTa (Liu et al. 2019), three widely used encoder transformers. BERT is the original application of the transformer architecture for classification and sequence labelling tasks (Devlin et al. 2019), while RoBERTa is an optimized BERT approach that maintains the same architecture as BERT but introduces several training improvements (Liu et al. 2019). XLNet is a transformer-based architecture that features relative positional encoding and segment recurrence, and is able to better capture long-range dependencies across

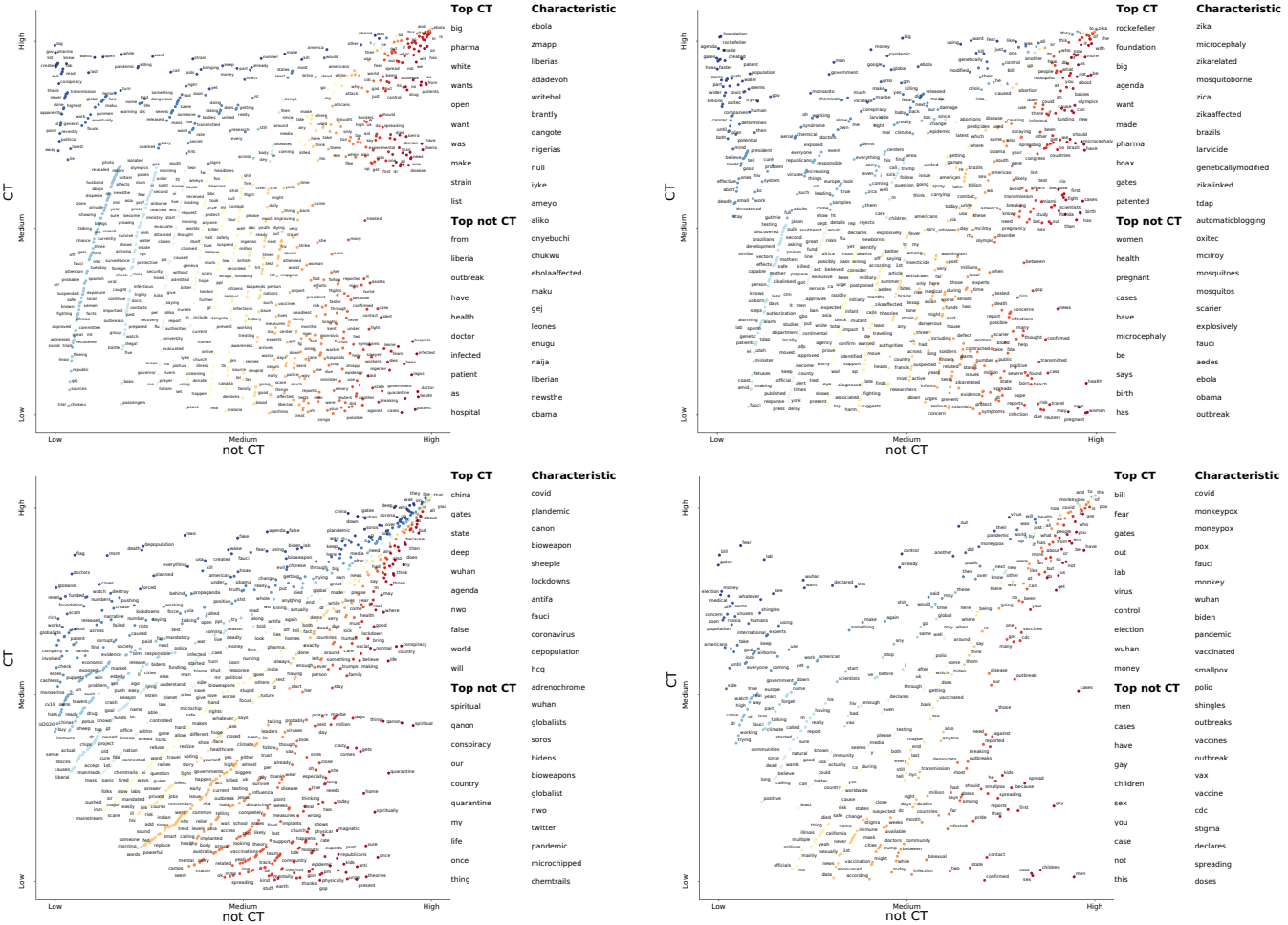

FIGURE 4 | Content of the epidemic datasets. The top uni/bi-grams are characteristic words per labels, the uni/bi-grams under 'Characteristic' title shows the terms differentiates the corpus from Brown Corpora Francis and Kucera (1979) which is used by Scattertext to represent generic English words (Kessler 2017). Top-left: Ebola, top-right: Zika, bottom-left: COVID-19 and bottom-right: Monkeypox.

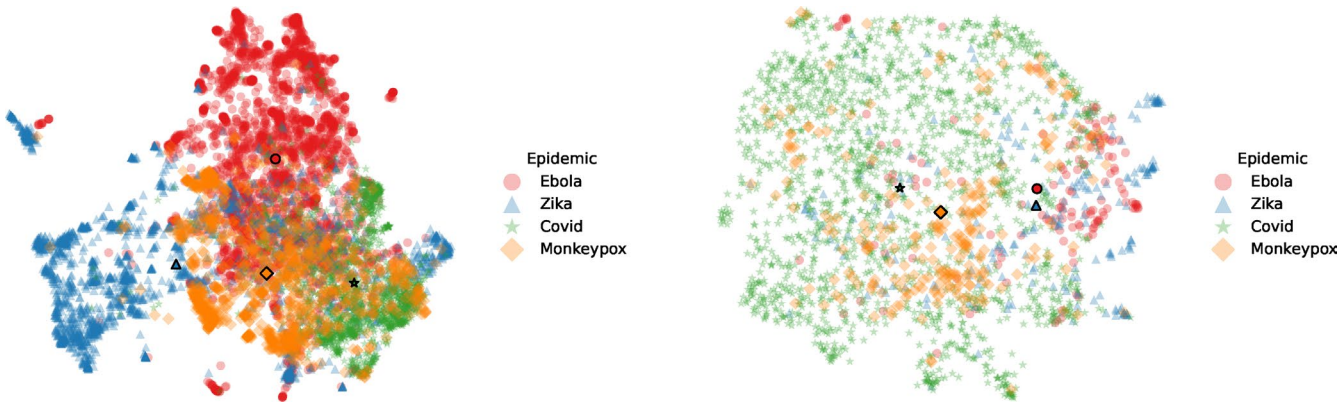

FIGURE 5 | Semantic structure of the texts using SBERT embeddings (`thenlper/gte-large`) projected into a shared space via UMAP across Ebola, Zika, COVID-19 and Monkeypox. Left: Projection on the not-CT texts, right: Projection on the CT texts.

text sequences (Yang et al. 2019). Table 6 shows the information about the models' pre-trained datasets.

All the models across the experiments were trained using the same method. In order to obtain more robust performance estimates, for each combination of the train and test dataset we fine-tuned five different models by using the 5-fold cross-validation approach. Our experiments encompass both the standard in-domain training with data from a single epidemic, and the transfer training where the models are fine-tuned on data from a previous epidemic and tested on a later epidemic. For the in-domain case, the train and test splits are generated

TABLE 6 | The BERT-like models and the datasets that the models were pre-trained on.

Model	Training data	Knowledge cutoff
BERT	BooksCorpus, English Wikipedia	2018
RoBERTa	BooksCorpus, English Wikipedia, CC-NEWS, OPENWEBTEXT, STORIES	2019
XLNet	BooksCorpus, English Wikipedia, Giga5, ClueWeb and Common Crawl	2019

Note: The publication years of the CCNews English News articles in 2016–2019. Therefore, the news may cover topics such as Ebola and Zika.

in the standard way. In the transfer case, standard train splits of the dataset from the earlier epidemic are used, while the test split is fixed to the dataset from the later epidemic. If more than one previous epidemic dataset is used for training, train splits from each previous dataset are concatenated in order to create the final train split. We always use a random 10% subsample of the training split as the development set. All the sampling was done in a stratified way to preserve the proportion of CT- and not-CT-labelled texts. Given a training set, the models were fine-tuned for a maximum of 100 epochs. To prevent over-fitting we used early stopping with a patience of 20 epochs, based on the macro-$F1$ calculated on the development set. We chose this training procedure since it works well for similar tasks of misinformation detection (Cui and Jia 2025; Caselli et al. 2021; Pelrine et al. 2023).

We fine-tuned `google-bert/bert-base-cased` (BERT), `roberta-base` (RoBERTa) and `xlnet-base-cased` (XLNet) from the HuggingFace repository[10] using the `transformers` library (Wolf et al. 2020). Each model was fine-tuned using the default hyperparameters of the library's `Trainer` method,[11] except for a batch size of 16, a learning rate of 2e−5, a weight decay of 0.01 and warmup steps set as a ratio of the training data length to the batch size.

4.2 | Large Language Models

LLMs (Zhao et al. 2023), generative models based on the transformer architecture (Vaswani et al. 2017) and pre-trained on huge amounts of natural language data, have proved useful for a variety of NLP and Information Retrieval tasks (Zhao et al. 2023). LLMs also show promise for automatization of content analysis in Computational Social Science (Ornstein et al. 2025; Chae and Davidson 2023; Korenčić, Chulvi, Bonet-Casals, et al. 2024).

We analyse the performance on LLMs on our use-case of detecting conspiracy theories about new and emerging epidemics since their zero-shot classification abilities (Brown et al. 2020; Zhao et al. 2023) could enable practitioners and researchers to quickly bootstrap conspiracy classifiers.

TABLE 7 | Properties of the LLMs used in the experiments.

LLM	Open weights	# of params	Knowledge cutoff
Llama3-70b	Yes	70 billion	End of 2023
Mixtral-47b	Yes	47 billion	December, 2023
GPT-4o	No	Unknown	October 2023

We experiment with three LLMs: Llama3-70b, Mixtral-47b and GPT-4o. The former two models are mid-sized open-source models and the latter is a popular model from OpenAI. All of the models we use are instruction-tuned (Zhang et al. 2026), that is, fine-tuned on examples of instruction-following on various language tasks, which is expected to improve zero-shot performance (Zhang et al. 2026).

Llama3-70b (Grattafiori et al. 2024) is part of a newest version of open-source models created by Meta. Mixtral-47b (Jiang et al. 2024) is an open-source Sparse Mixture of Experts (SMoE) language model offering competitive performance. In both cases, model weights and code are publicly accessible, which enables both model analysis and the private deployment of the models. The OpenAI model is GPT-4o (OpenAI et al. 2024) from the GPT-4 family, accessible through the company's web-based API.[12]

Table 7 contains the information about the LLMs used, including the knowledge cutoff dates. These dates designate the end dates of the training datasets of the LLMs. With Monkeypox epidemic, occurring during 2022 and 2023, being the most recent one we can see that all of the LLMs have 'seen' the texts discussing all of the epidemics. However, the mappings between the texts and the associated conspiracy labels of the datasets have been unavailable to the models, since our datasets were not public at the time of the experiments.

Implementation Details: We experimented with two zero-shot classification prompts, each of which was used in combination with all of the LLM models. The short version of the prompt briefly and generally describes the classification task. The long version of the prompt contains the full definition of what supporting a CT means. The definition, constructed by an expert on conspiracy theories, was the same as the one given to the annotators (Section 3.2). The short and long prompts are shown, respectively, in Figures A1 and A2.

For each prompt, the value of `disease` was either 'Zika', 'Ebola', 'COVID-19' or 'Monkeypox', and the `text` variable was replaced with the text of the tweet being classified. Only this prompt was sent to a model, as a 'user' message, and no additional 'system' messages were used to configure model behaviour. Greedy decoding was used to generate the text of the answer. The resulting model-produced text was lower-cased, whitespace and non-alphanumeric characters were removed from both borders, after which the first word (whitespace-delimited) was compared to 'yes' and 'no' strings to determine the predicted class. If the comparison failed, the default result 'no' was returned (the majority class).

We used the LangChain framework[13] to facilitate the zero-shot learning. In case of the OpenAI model, the model variant `gpt-4o-2024-05-13` was used. The Llama3-70b model was accessed via the HuggingFace Inference Pro platform,[14] while the Mixtral-47b model was deployed on a FriendliAI dedicated endpoint.[15] The HuggingFace IDs of the two open models are `meta-llama/Meta-Llama-3-70B-Instruct` and `mistralai/Mixtral-8x7B-Instruct-v0.1`, respectively.

5 | Results and Analysis

In this section we describe the experiments aimed at answering research questions *RQ1–RQ4*. First we tackle the case of in-domain CT detection with BERT-like models, and we examine the cross-epidemic transfer of labels. The second set of experiments examines the performance of zero-shot LLM classifiers, the performance of LLMs as annotators and the performance of BERT-like models on LLM-annotated data. The reported classification results include Precision, Recall and the binary *F*1 score, with the CT as the positive class.

5.1 | In-Domain Prediction Results

In this experiment we tackle *RQ1* by examining the performance of CT detection with BERT-like classifiers trained and tested on the same epidemic. These classifiers represent a strong baseline for all the other approaches which either use out-of-domain training data or rely on zero-shot CT detection with LLMs. The models were fine-tuned using the procedure described in Section 4.1.

In order to compare the performance of models with expected human agreement, we computed the agreement between annotators in terms of the binary *F*1 score. First, the annotations (Section 3.2) were used to extract binary *CT* versus *not-CT* labels for each annotator. Then these labels were viewed as class predictions and evaluated against two definitions of ground truth: labels of other annotators (human-vs-human), and the labels obtained by the majority vote (human-vs-gold). In each case, the final score was formed by averaging over all possible (prediction, ground truth) pairs. Note that the human-vs-human scores are pessimistic estimates of human performance, since labels of a single annotator represent a ground truth that is more variable than the averaged gold standard. On the other hand, the human-vs-gold scores are optimistic estimates since the individual predictions are merged into the average, and can therefore influence it. In the case of the COVID-19 dataset the human *F*1 scores could not be calculated since we had no access to labels of individual annotators.

Results displayed in Table 8 show that medical conspiracy detection is in general a challenging problem, with the exception of the COVID-19 dataset (Langguth et al. 2023) where *F*1 score of above 80 is achieved. For other epidemics, *F*1 scores are either slightly below or slightly above (in the case of Zika dataset) the pessimistic estimates of human performance. The performance on COVID-19 dataset is likely the consequence of it being a balanced dataset with a proportion of positive (conspiracy) examples much higher than for the other datasets (see Table 1). Similarly, worst results are achieved on the Ebola dataset, where the proportion of conspiracy examples is only 4.66%. This demonstrates that a larger proportion of positive examples tends to enable the supervised models to make a better approximation of the conspiracy class. Regarding the transformer models, XLNet and RoBERTa consistently outperform BERT, while the difference between XLNet and RoBERTa is within two *F*1 points and their relative performance depends on the dataset.

High performance of models on COVID-19 data could also be explained by the method of dataset creation. Namely, unlike other datasets, the COVID-19 COCO dataset was curated using 'targeted COVID-19 related keywords' Langguth et al. (2023). This approach may have resulted in a more semantically coherent CT class, or in a set of not-CT tweets that are easier to distinguish from the conspiracies. Both cases could lead to a better decision boundary, and the semantic projections of COVID-19 tweets in Figure 5 suggest that the latter case is more likely.

5.2 | Transferring Knowledge From the Prior Epidemics

To answer *RQ2*, we investigate the impact of labelled data from prior epidemics on identifying CTs during emerging epidemics of new diseases. To this end, we fine-tune the BERT-like models (BERT, RoBERTa and XLNet) on datasets from prior epidemics to predict CTs about a target epidemic. This corresponds to the scenario when models for detection are needed during an emerging epidemic, but only CT-labelled data from a previous epidemic is available. We compare the performance of these models against the following baselines: (1) Logistic Regression model trained on TF-IDF features (denoted as TfIdf-LogReg, a simple baseline for assessing the relative gains of the BERT-like models), (2) BERT-like models fine-tuned on the target epidemic dataset (Section 5.1, strong baseline) and (3) BERT-like models fine-tuned on a

TABLE 8 | Performance of the classifiers (*F*1): trained and tested on the data from the same epidemic, Human-Pair (human-vs-human) and Human-Majority (human-vs-gold).

Test-epidemics	BERT	RoBERTa	XLNet	Human-Pair	Human-Majority
Ebola	56.33	57.83	59.12	61.19	80.24
Zika	62.56	69.70	70.78	65.35	82.41
COVID-19	78.53	81.87	81.03	N/A	N/A
Monkeypox	59.92	66.09	64.66	66.59	81.41

TABLE 9 | The label distribution of the non-epidemic dataset (Phillips et al. 2022).

Collection	CT	Not-CT
Climate change	110	255
Maxwell-Epstein	146	139

non-epidemic CT detection dataset (weak baseline). The non-epidemic dataset (Phillips et al. 2022) contains CT-labelled texts on conspiracies related to Maxwell-Epstein (ME) and Climate Change (CC). The ternary annotation scheme codifying 'stance towards the conspiracy theory {support, neutral, against}' (Phillips et al. 2022) was adapted by merging the 'neutral' and 'against' labels into the not-CT label. The details of the non-epidemic dataset are displayed in Table 9.

In order to faithfully simulate the new epidemic scenario, that is, to evaluate transformers on transferring knowledge about preceding epidemics to identifying CTs about emerging epidemics, we opted to use transformer models pretrained on data that has no time overlap with the test epidemics. Namely, if the transformers are pretrained on a dataset collected in a period that overlaps with an epidemic, there is a chance that some texts about the epidemic had been 'seen' by the model. However, when a new epidemic breaks out the existing pretrained transformers have not been exposed to texts about it. For the stated reasons, we test only on the COVID-19 and Monkeypox datasets, since Ebola and Zika datasets overlap with the time period of the pretraining datasets, while the COVID-19 and Monkeypox epidemics happened after the pretraining datasets of the transformers were compiled. The time periods of the epidemic-related CT datasets and the transformer pretraining datasets are reported in Tables 4 and 6, respectively.

For similar reasons, we excluded Maxwell-Epstein non-medical CTs, originating in 2021 (Phillips et al. 2022), from the non-medical training set used for models tested on identifying COVID-19 CTs. Namely, there is a chance that some COVID-19-related info is contained in this dataset, while no such possibility exists for Monkeypox that occurred in 2022. Table 10 contains all the tested scenarios defined by the selection of the training set, with the in-domain results included for comparison. In the non-epidemic scenario, an organization only has CT-labelled data on non-medical topics. In the other scenarios, an organization has access to labelled data from one or more past epidemics. For instance, when COVID-19 started, an organization might have had data on Zika, or data on both Zika and Ebola.

All the models were trained and evaluated using the procedure described in Section 4.1. If applicable, the training splits were created in the 5-fold cross-validation manner by combining training splits from multiple training datasets. The results, displayed in Table 10, show that the BERT-like models perform better than the TfIdf-LogReg, as the samples might not contain enough keyword-based information to differentiate the CT content. Additionally, the strong baseline of in-domain trained models expectedly outperforms the transfer models for both COVID-19 and Monkeypox. However, the relative performance of transfer models varies greatly for these two datasets. Their performance for COVID-19 is below the strong baseline by at

TABLE 10 | Precision, recall and *F*1 scores of BERT-like models (BERT, RoBERTa and XLNet) across different datasets.

Training data	Model	Precision	Recall	*F*1
COVID-19				
In-domain	TfIdf-LogReg	74.12	76.92	75.52
	BERT	77.83	79.25	78.53
	RoBERTa	81.10	82.67	81.87
	XLNet	80.98	81.07	81.03
Non-epidemic (CC)	TfIdf-LogReg	59.20	38.53	45.34
	BERT	55.82	59.06	57.40
	RoBERTa	62.96	68.81	65.76
	XLNet	66.32	64.65	65.47
Zika	TfIdf-LogReg	69.55	45.35	54.65
	BERT	66.31	59.92	62.95
	RoBERTa	72.23	57.70	64.15
	XLNet	73.77	57.07	64.35
Ebola, Zika	TfIdf-LogReg	38.36	58.05	46.03
	BERT	67.77	55.02	60.73
	RoBERTa	70.01	65.62	67.75
	XLNet	70.60	63.51	66.87
Monkeypox				
Training data	**Model**	**Precision**	**Recall**	***F*1**
In-domain	TfIdf-LogReg	51.24	57.58	54.00
	BERT	63.96	56.35	59.92
	RoBERTa	71.96	61.11	66.09
	XLNet	65.45	63.89	64.66
Non-epidemic (CC, ME)	TfIdf-LogReg	21.56	53.97	28.04
	BERT	27.11	53.57	36.00
	RoBERTa	27.53	57.14	37.16
	XLNet	29.37	51.98	37.54
COVID-19	TfIdf-LogReg	35.72	68.25	46.63
	BERT	36.97	61.90	46.29
	RoBERTa	43.34	71.03	53.83
	XLNet	45.20	71.03	55.25
Zika, COVID-19	TfIdf-LogReg	48.50	55.91	51.67
	BERT	56.61	54.37	55.47
	RoBERTa	56.21	68.25	61.65
	XLNet	57.19	66.27	61.40

(Continues)

TABLE 10 | (Continued)

Monkeypox				
Training data	**Model**	**Precision**	**Recall**	***F*1**
Ebola, Zika, COVID-19	TfIdf-LogReg	54.44	46.45	48.70
	BERT	54.86	62.70	58.52
	RoBERTa	56.96	71.43	63.38
	XLNet	57.88	67.06	62.13

Note: The top plot shows performance for detecting COVID-19 CTs, and the bottom plot shows performance for detecting Monkeypox CTs. Models trained on prior epidemic datasets (e.g., Ebola and Zika for COVID-19; COVID-19, Ebola and Zika for Monkeypox) outperformed those trained on non-epidemic datasets but underperformed compared to models trained on in-epidemic datasets.

least 15 *F*1 points, while in the case of Monkeypox the difference is only 3 *F*1 points. Surprisingly, for COVID-19 the models trained on a much smaller non-epidemic dataset are competitive with the transfer models. However, in the case of Monkeypox the non-epidemic models are outperformed by a large margin of approximately 25 *F*1 points. For both test epidemics, it is clear that adding more training data from previous epidemics is beneficial for all transformer models. For Monkeypox, this benefit is more pronounced as adding Zika and Ebola helps a lot, and as models trained only on COVID-19 perform poorly despite COVID-19 being the largest dataset. Model-wise performance is consistent, as more advanced RoBERTa and XLNet consistently outperform BERT across all training setups.

The results show that re-using CT-labelled data from previous epidemics can lead to models with performance close to that of in-domain models. However, the effectiveness of transfer models in not guaranteed, as exemplified by COVID-19. Previous research (Samory and Mitra 2018; Klein et al. 2019) found that certain motifs of CT narratives—defined as a story where agents and patients are connected by an action—are similar across different domains and events. For example, in the medical domain, some CTs are recycled in subsequent diseases (Miguel 2022). The data-points demonstrating effectiveness of both non-epidemic data and data from previous epidemics support these claims. Furthermore, the results indicate that the 'narrative motifs' occurring in COVID-19 differ from motifs occurring in Ebola and Zika, while the Zika and Ebola motifs are closer to Monkeypox motifs. One plausible explanation is that the set of COVID-19 motifs is larger than that of other diseases, due to the huge amount of attention induced by COVID-19. Namely, models trained on Zika and Ebola and tested on COVID-19 clearly have higher precision than recall, which can be explained by their ability to detect only those motifs of COVID-19 conspiracies present in these diseases. On the other hand, COVID-19-trained models have low precision and high recall when tested on Monkeypox, which can be explained by their tendency to detect as (falsely) positive a number of COVID-19 motifs not present in Monkeypox CTs.

5.2.1 | Further Analysis

To better understand the impact of the training dataset on the performance of the BERT-like models, we analysed the texts from the test datasets (COVID-19 and Monkeypox) that contain one or more of the top 10 terms characteristic for the CT and not-CT classes. The characteristic terms were ranked by Scattertext based on PMI (introduced in Section 3.4), and represent dataset-specific terms highly correlated with the two classes. The number of samples per term is given in Table A1. Next, we examined the predictions of the BERT-like models on these samples and grouped the predictions based on the training datasets and characteristic terms. For each group, we computed binary *F*1 scores. To assess model predictions on the samples containing CT terms, we calculated *F*1 scores for CT class, while for samples with the not-CT terms, we computed *F*1 scores for not-CT terms. To obtain an overall perspective on the impact of training datasets, we averaged these scores across model architectures. We presented the average *F*1 scores as heatmaps in Figure 6.

The BERT-like models trained on prior epidemics outperform non-epidemic models, particularly in classifying as conspiracy the texts containing CT terms such as agenda, NWO and Bill Gates even though proportion of the CT texts are higher in non-epidemic dataset. These models also generally demonstrate superior performance in recognizing not-CT samples compared to their non-epidemic counterparts in classifying as non-conspiracy the texts containing not-CT terms. This improvement is likely due to the fact that prior epidemic datasets contain conspiracy-related narratives similar to those in the test datasets, which makes the models better at distinguishing between CT and not-CT patterns in language.

Additionally, the results underline the differences and similarities in the conspiracy narratives across epidemics. From the scores in the top left corner of Figure 6 it can be seen that the classifiers trained on the Ebola and Zika datasets underperform for the texts containing CT-related terms of the COVID-19 dataset, except for the 'agenda' and 'nwo'. This effect is reversed for COVID-19-trained classifier applied to texts containing Monkeypox-specific CT terms (Figure 6, bottom left). These observations indicate that the conspiracy narratives of Zika and Ebola differ from that of COVID-19, while COVID-19 and Monkeypox narratives are similar, possibly due to the time differences of the outbreaks of the epidemics in question.

Previous observations are supported by the semantic analysis of epidemic narratives (Figure 5). Namely, CTs from COVID-19 occupy a broader and more heterogeneous region of the semantic space, and they align better with Monkeypox CTs than with the Ebola and Zika CTs. Ebola an Zika overlap only with a limited subspace of COVID-19, but their coverage of Monkeypox CTs is much better. These semantic differences and similarities can be related to the performance of transfer learning. Specifically, the performance of transfer from Ebola and Zika to COVID-19 stays well below in-domain COVID-19 performance. However, when data of Ebola and Zika is added to the data of COVID-19, the transfer performance on Monkeypox improves markedly.

5.3 | Zero-Shot Classification With LLMs

In order to answer *RQ3*, we examined how well can LLMs detect CT in text. We use the LLMs as zero-shot classifiers. Each

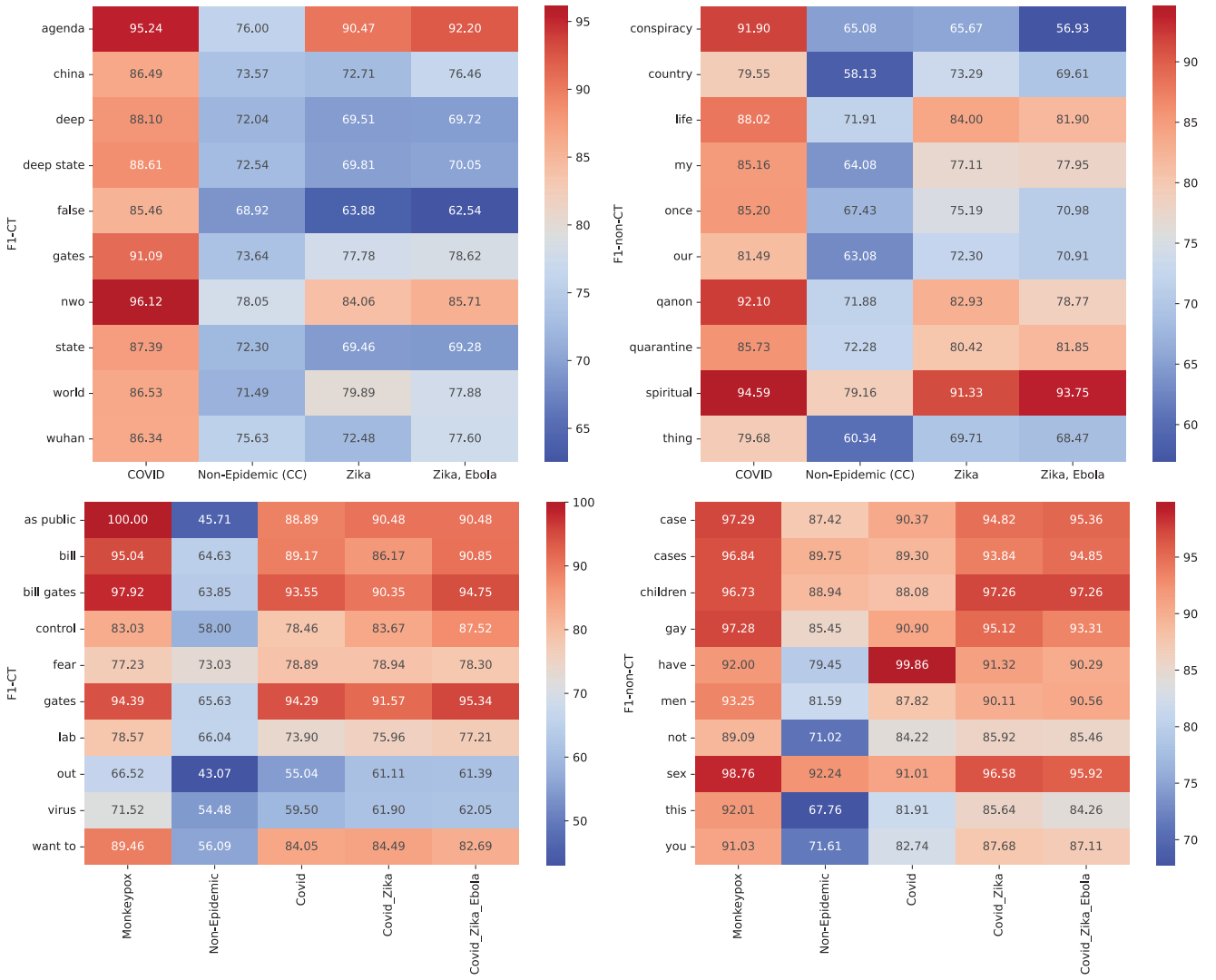

FIGURE 6 | Heatmaps illustrating the models' performance in identifying the labels of samples that contain class-specific characteristic words. The top row presents results for COVID-19-related texts, while the bottom row corresponds to Monkeypox-related texts. Epidemic-focused models generally outperform non-epidemic models in predicting CT/not-CT theories when characteristic words are present.

model prompt is composed only of the text being classified and additional instructions about the task (Section 4.2). Prompts come in two variants: 'simple' and 'definition'. The 'simple' variant contains the instruction to 'determine weather the tweet supports a CT about' about a disease, and the 'definition' variant includes the description of what it means to support a CT. We experiment with three popular LLMs (Mixtral-47b, Llama3-70b and GPT-4o), each in combination with both prompts, as well as with a majority-voting ensemble of LLMs. The ensemble consists of all three LLMs combined with the definition prompt, that is, the best-performing prompt variant. We also include a prompt ensemble labelled GPT-4o-PE, based on the GPT-4o model in combination with three reformulations of the definition prompt, since it proved to be a robust and well-performing solution (Section 5.3.1). As a baseline we use BERT-like models trained on in-domain disease datasets (Section 4.1).

The results, displayed in Table 11, reveal that for most datasets the best-performing LLMs compare favourably to supervised baselines. For Ebola and Monkeypox the best LLMs outperform BERT-like models by ca. 3–6 $F1$ points, while for COVID-19 their performance is comparable. The definition prompt variant is clearly superior, resulting in better performance for all LLM models and all datasets except COVID-19. This demonstrates the benefit of including precise instructions in the prompt. The definition prompt improves the models' precision in the majority of cases, often by a large margin. Two largest models (GPT-4o and Llama3-70b) perform comparably in most cases, and they almost always outperform the Mixtral-47b model. While the LLMs tend to have high recall and lower precision, for the supervised models' precision tends to be higher and the difference between precision and recall tends to be smaller.

However, the relative and performance of individual LLM models varies across datasets. Llama3-70b achieves top performance for Zika, COVID-19 and Monkeypox, and GPT-4o for Ebola, with the differences between the two being small, ca. 1–3 $F1$ points. The individual LLMs are outperformed by supervised models

TABLE 11 | Zero-shot performance of the LLMs in comparison with the in-domain BERT-like models.

Ebola					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
Baseline	TfIdf-LogReg	NA	53.29	45.05	48.02
BERT-like	BERT	NA	66.99	48.59	56.33
	RoBERTa	NA	68.63	49.30	57.83
	XLNet	NA	61.36	57.04	59.12
Open	Llama3-70b	Simple	26.64	97.18	41.82
		Definition	45.71	90.14	60.66
	Mixtral-47b	Simple	28.18	73.24	40.70
		Definition	42.55	84.51	56.60
Closed	GPT-4o	Simple	23.99	96.48	38.43
		Definition	50.43	82.39	<u>62.57</u>
Ensemble	All 3	Definition	51.21	89.44	**65.13**
	GPT-4o-PE	Definition	47.71	88.03	61.88
Zika					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
Baseline	TfIdf-LogReg	NA	59.89	64.89	60.59
BERT-like	BERT	NA	64.32	60.89	62.56
	RoBERTa	NA	71.50	68.00	69.70
	XLNet	NA	72.77	68.89	**70.78**
Open	Llama3-70b	Simple	36.09	96.89	52.59
		Definition	50.48	92.89	65.41
	Mixtral-47b	Simple	42.59	71.56	53.40
		Definition	50.38	88.44	64.19
Closed	GPT-4o	Simple	37.34	91.11	52.97
		Definition	55.63	76.89	64.55
Ensemble	All 3	Definition	53.51	91.56	67.54
	GPT-4o-PE	Definition	55.94	94.22	<u>70.20</u>
COVID-19					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
Baseline	TfIdf-LogReg	NA	74.12	76.92	75.52
BERT-like	BERT	NA	77.83	79.25	78.53
	RoBERTa	NA	81.10	82.67	81.87
	XLNet	NA	80.98	81.97	81.03
Open	Llama3-70b	Simple	76.26	93.96	**84.19**
		Definition	75.38	91.28	<u>82.57</u>
	Mixtral-47b	Simple	73.89	87.91	80.29
		Definition	69.53	88.08	77.72

(Continues)

TABLE 11 | (Continued)

COVID-19					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
Closed	GPT-4o	Simple	71.39	95.32	81.64
		Definition	75.43	90.48	82.27
Ensemble	All 3	Definition	74.94	91.22	82.28
	GPT-4o-PE	Definition	73.82	91.79	81.83
Monkeypox					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
Baseline	TfIdf-LogReg	NA	51.24	57.58	54.00
BERT-like	BERT	NA	63.96	56.35	59.92
	RoBERTa	NA	71.96	61.11	66.09
	XLNet	NA	65.44	63.89	64.66
Open	Llama3-70b	Simple	43.11	95.63	59.43
		Definition	60.47	91.67	**72.87**
	Mixtral-47b	Simple	49.37	77.78	60.40
		Definition	51.72	83.73	63.94
Closed	GPT-4o	Simple	39.38	95.63	55.79
		Definition	58.76	86.51	69.98
Ensemble	All 3	Definition	60.22	88.89	71.79
	GPT-4o-PE	Definition	57.49	94.44	71.47

Note: NA denotes not applicable. The ensemble of all three LLMs and the prompt ensemble (GPT-4o-PE) achieve results close to the top-performing (bold) results.

for Zika, but not for other datasets. This unstable behaviour makes the choice of the optimal model difficult. However, LLM ensembles achieve both stable and predictably good performance as they achieve nearly optimal performance for each dataset. This finding demonstrates that one of the ensemble approaches should be the method of choice in practical application. The potential downside of these approaches is the need to run inference on three LLM models and, in the case of the prompt ensemble, to construct additional prompt variations.

Our results indicate that for the use-case of conspiracy detection in medical texts, zero-shot LLM models are a good option, especially when used as an ensemble, and that they can match or outperform supervised BERT-like models. This finding does not contradict existing experiments—although in general zero-shot LLMs perform worse than BERT-like models (Ziems et al. 2024; Bucher and Martini 2024; Mu et al. 2024), they can match or outperform them on a number of benchmarks (Ziems et al. 2024; Mu et al. 2024). Moreover, they can perform well (*F*1 scores of 70 or more) on conspiracy classification (Pustet et al. 2024; Corso et al. 2025) tasks.

LLMs are prone to data contamination (Golchin and Surdeanu 2024), a phenomenon that occurs when labelled text instances from the training or test set are also contained in the dataset used to pre-train an LLM. Contamination is a consequence of LLMs being trained on huge datasets that contain large amounts of text from the web (Zhao et al. 2023). In our experiment the contamination is not possible since the conspiracy labels were kept private until the time of the experiment, that is, they could not have been a part of the pre-training datasets. Furthermore, the labelled data was not exposed to the possibility of test-time leaking (Balloccu et al. 2024) since the LLM APIs, and not their web interfaces (Balloccu et al. 2024), were used to invoke the LLMs.

To investigate other approaches with a potential to improve LLM classifiers, we assess the influence of non-deterministic decoding on LLM performance (Appendix A.5), and examine the effect of two calibration methods Zhou et al. (2024) (Appendix A.6). The results suggest that switching from greedy decoding to temperature-based sampling has no influence of performance. The Batch Calibration method Zhou et al. (2024) increases precision at the cost of recall, but it also decreases performance in the majority of cases.

5.3.1 | Sensitivity to Prompt Variation

Performance of LLMs on classification tasks is sensitive to variations in prompt formulation (Sclar et al. 2024; Sivarajkumar et al. 2024; Xia et al. 2025). In order to obtain more robust performance assessment for our use-case, we examine the sensitivity of the LLM performance to variation of the original prompts. We vary the prompts by applying reformulations that preserve prompt semantics, and we keep the high-level prompt structure

intact. We opt for this approach to keep the evaluation feasible (given the infinite space of viable prompt variants), and to provide a robust 'recipe' for future applications that might require a degree of prompt adaptation.

The prompts (Section 4.2, Appendix A.2) are based on instructions for human annotators designed by a conspiracy expert. From the high-level structural perspective, they are composed of an instruction that solicits a judgement weather a text 'supports a conspiracy theory', and that restricts the answer to 'yes' or 'no' class labels. After the instruction, the input text is inserted into the prompt, and the prompt is finalized with a query string that solicits a yes/no answer: 'supports a conspiracy (YES or NO):'. The definition prompt is prefixed with a full definition of 'supporting a conspiracy theory', bracketed in special delimiter strings (Appendix A.2).

We manually construct semantically equivalent variations of the prompts while respecting the described structure. Specifically, we moderately vary the wording of the instructions, the definition and its delimiters, and the class labels. We preserve the expert-defined 'supports a conspiracy theory' formulation, and, in order to avoid potential confusion, we consistently use the same designator of the text being classified (ex. 'text', 'tweet', 'social media post', ...). The structure of the definition text is also preserved: (a) explicit support, (b) elements and examples, (c) implicit support and how to detect it. We construct 5 variations of both the simple prompt and the definition prompt. Figure A3 exemplifies one variation of the definition prompt, while the other variations can be found in the code repository.[16]

Prompt variations enable us to examine the performance of majority-voting LLM ensembles composed of different prompt variations. This approach can perform competitively for LLM-based classification (Sivarajkumar et al. 2024), and it can raise performance above that of low-performing prompts (Sivarajkumar et al. 2024).

For each combination of a model and a prompt variant (simple or definition), we compute classification results for each of the five reformulations and calculate averages and distribution statistics. The statistics for prompt ensembles (designated by '(ens)') are based on all 10 three-out-of-five combinations of prompt variations.

The results in Table 12 show that the performances remain relatively stable for most of the LLM models and for most of the diseases. Specifically, for Llama3-70b and GPT-4o models the difference between the minimum and maximum $F1$ score remains close to or below 5 $F1$ points. The exception is GPT-4o with the definition prompt on Ebola and Zika, where the difference increases to 7 $F1$ points. On the other hand, the Mixtral-47b model is much less stable, with performance differences being close to 15 $F1$ points in some cases. The simple prompt variant is more stable then the definition variant for Llama3-70b and GPT-4o models, while for Mixtral-47b the opposite is the case.

When the prompt variation is taken into account, the average performance of the largest LLM models (Llama3-70b and GPT-4o) combined with definition prompts stays comparable to or above the BERT-like models' performance for all diseases except Zika. The average performance of LLM ensembles (with definition prompts) is better than that of individual LLMs, and it comes close to or exceeds the BERT-like performance for all diseases.

The results show that, in almost all cases, prompt ensembling is beneficial for increasing minimum performance and for decreasing the variation of performance. While the cross-model ensembles raise the minimum performance in most cases, the gains in stability, as measured by the reduction of standard deviation, are neither large nor guaranteed. This could be, at least in part, because of the instability of Mixtral-47b models. On the other hand, prompt ensembling raises minimum performance and reduces standard deviation across the board, with the largest gains being achieved for the most unstable Mixtral-47b models. This demonstrates that practitioners should seriously consider this approach if the added cost (crafting three prompts, increased inference time) is acceptable.

5.3.2 | Expected Performance of LLMs on Emerging Epidemics

In Section 5.2 we examine the potential of applying BERT-like models to detect conspiracies on an emerging epidemics, that is, in a scenario when only labelled data on previous epidemics is available. Zero-shot LLMs are also a viable approach in this scenario, but in order for the LLM experiments to be faithful to the emerging epidemic use case, we need LLMs with pre-training datasets that have no overlap with the test epidemic's dataset. This is hard to achieve since the state-of-art LLMs are expected to be trained on data that is nearly up to date. In 2023, almost all large models overlapped with the COVID-19 period and most of them overlapped with the Monkeypox period (Zhao et al. 2023). Additionally, the majority of the teams that build LLMs do not disclose the details on the pre-training dataset (the 'open' label in most cases refers only to model parameters), which makes it hard to determine whether an LLM has 'seen' data on certain epidemic.

In order to assess the performance of zero-shot LLMs on truly 'unseen' conspiracies, we apply an older ChatGPT-3.5 model (knowledge cutoff in September 2021[17]) to the tweets about Monkeypox, the most recent of all four epidemics (outbreak in 2022). ChatGPT-3.5 is a version of GPT-3 (Brown et al. 2020) model with 175B parameters, fine-tuned on supervised data for improved instruction-following (Ye et al. 2023). The model's architecture, size and training procedures make it representative of modern LLM models, and its knowledge cutoff data guarantees zero overlap between training data and posts about Monkeypox. We apply the model to the classification task using the same techniques (Section 4.2) that are used with other LLMs.

Results in Table 13 show that, on the detection of conspiracy tweets about Monkeypox, ChatGPT-3.5 is comparable to models whose training data overlaps with the period when the tweets were created. When used with the definition prompt, the model outperforms BERT-like models and comes close to the performance of two largest LLM models, Llama3-70b and GPT-4o.

TABLE 12 | Robustness of classification performance to prompt variation, for individual LLMs and their ensembles. Statistics (minimum, average, maximum and standard deviation) are based on 5 different prompt variations/reformulations.

Ebola													
		Precision				**Recall**				***F*1**			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	26.64	27.75	29.37	1.14	94.37	96.20	97.18	1.05	41.82	43.06	44.96	1.28
	Simple (ens.)	27.01	28.02	28.69	0.46	95.77	**96.48**	97.18	0.55	42.27	43.42	44.16	0.52
	Definition	39.27	42.74	45.71	2.41	86.62	89.58	91.55	1.75	54.97	57.81	60.66	2.04
	Definition (ens.)	40.95	43.33	45.32	1.31	88.03	90.70	92.96	1.54	56.46	58.62	60.00	1.12
Mixtral-47b	Simple	23.83	34.85	52.48	9.78	37.32	67.32	76.06	15.03	36.21	43.18	48.29	4.38
	Simple (ens.)	28.65	33.30	39.85	3.47	71.83	73.31	75.35	1.24	41.41	45.68	51.61	3.16
	Definition	28.29	34.85	42.55	4.61	84.51	88.73	90.85	2.36	43.14	49.80	56.60	4.36
	Definition (ens.)	32.40	34.80	37.09	1.47	87.32	88.94	91.55	1.26	47.57	50.00	52.19	1.43
GPT-4o	Simple	20.73	23.45	24.95	1.48	91.55	94.93	96.48	1.80	34.12	37.58	39.53	1.87
	Simple (ens.)	22.50	23.68	24.73	0.75	94.37	95.56	96.48	0.71	36.48	37.94	39.24	0.93
	Definition	39.58	45.25	50.43	3.88	82.39	88.03	92.25	3.36	55.39	59.57	62.57	2.66
	Definition (ens.)	43.00	**45.34**	48.06	1.62	87.32	89.15	90.85	1.19	58.37	60.08	62.00	1.19
Ensemble	Simple	26.13	28.89	31.48	1.83	91.55	93.94	96.48	1.58	40.86	44.15	46.85	2.05
	Definition	41.85	45.16	51.21	3.63	88.73	90.56	92.25	1.31	57.52	**60.15**	65.13	2.93
XLNet	NA		61.36				57.04				59.12		
Zika													
		Precision				**Recall**				***F*1**			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	34.26	36.62	37.80	1.32	96.44	**97.07**	98.22	0.67	50.80	53.15	54.32	1.32
	Simple (ens.)	36.33	37.07	37.78	0.45	96.89	97.02	97.33	0.20	52.85	53.64	54.36	0.47
	Definition	45.73	48.32	50.60	2.09	91.11	93.16	95.11	1.28	61.19	63.60	65.62	1.83
	Definition (ens.)	47.02	48.56	49.88	0.93	93.33	94.27	94.67	0.46	62.83	64.09	65.23	0.76

(Continues)

TABLE 12 | (Continued)

Zika

		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Mixtral-47b	Simple	40.35	45.53	50.94	3.99	36.00	66.31	81.33	15.65	42.19	52.57	57.88	5.69
	Simple (ens.)	43.57	45.89	49.84	2.13	64.44	70.44	77.33	4.07	52.56	55.47	57.84	1.35
	Definition	39.66	42.95	50.38	3.79	88.44	91.64	93.78	1.83	55.75	58.34	64.19	2.98
	Definition (ens.)	41.22	42.97	44.81	1.21	91.11	92.09	93.33	0.79	57.10	58.58	60.26	1.01
GPT-4o	Simple	35.75	36.75	37.87	0.76	91.11	95.56	96.89	2.23	52.16	53.07	54.39	0.73
	Simple (ens.)	36.45	37.10	37.74	0.34	96.00	96.62	96.89	0.29	52.98	53.62	54.25	0.35
	Definition	48.75	53.10	57.70	3.18	76.89	90.67	95.56	7.04	64.55	66.73	70.79	2.28
	Definition (ens.)	51.32	**53.48**	56.42	1.76	93.78	94.53	95.11	0.45	66.67	**68.29**	70.45	1.34
Ensemble	Simple	39.31	40.10	40.79	0.49	94.67	95.82	96.44	0.67	55.86	56.53	57.33	0.51
	Definition	47.77	50.49	53.51	2.02	91.56	94.22	95.56	1.57	63.60	65.70	67.54	1.38
XLNet	NA		72.77				68.89				70.78		

COVID-19

		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	73.23	**75.42**	76.61	1.18	93.16	94.13	95.61	0.81	82.94	83.73	84.19	0.45
	Simple (ens.)	74.85	**75.42**	76.04	0.39	93.67	94.17	94.64	0.31	83.51	**83.76**	84.03	0.16
	Definition	72.74	74.48	75.84	1.30	90.94	91.52	92.36	0.59	81.39	82.11	82.71	0.56
	Definition (ens.)	73.60	74.61	75.60	0.67	91.33	91.78	92.30	0.33	81.85	82.30	82.73	0.28
Mixtral-47b	Simple	69.09	73.33	79.58	3.49	65.11	83.91	91.51	9.52	71.62	77.74	80.29	3.11
	Simple (ens.)	71.06	72.94	75.04	1.34	85.18	87.38	89.62	1.45	78.82	79.49	80.05	0.40
	Definition	65.75	67.16	69.53	1.28	88.08	90.65	92.36	1.44	76.62	77.14	77.72	0.40
	Definition (ens.)	66.27	67.10	67.93	0.57	90.08	90.86	91.68	0.52	76.75	77.19	77.48	0.24

(Continues)

TABLE 12 | (Continued)

COVID-19

		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
GPT-4o	Simple	68.54	69.48	71.39	1.01	95.32	95.87	96.18	0.32	79.90	80.56	81.64	0.58
	Simple (ens.)	68.87	69.39	69.75	0.31	95.72	**95.91**	96.24	0.16	80.17	80.52	80.77	0.20
	Definition	71.37	73.24	75.43	1.35	90.48	91.96	92.93	0.89	80.73	81.52	82.27	0.50
	Definition (ens.)	72.65	73.31	74.10	0.46	91.68	92.17	92.76	0.39	81.31	81.66	81.96	0.18
Ensemble	Simple	72.09	72.84	74.16	0.76	94.24	94.83	95.04	0.30	81.97	82.39	83.00	0.39
	Definition	71.49	72.80	74.94	1.22	91.22	92.12	92.93	0.60	80.81	81.32	82.28	0.53
RoBERTa	NA		81.10				82.67				81.87		

Monkeypox

		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	41.74	43.07	43.53	0.68	94.84	95.71	97.22	0.81	58.40	59.40	59.80	0.51
	Simple (ens.)	42.98	43.35	43.63	0.18	94.84	95.75	96.43	0.47	59.39	59.68	60.07	0.22
	Definition	53.94	58.67	61.52	2.83	89.29	90.95	92.46	1.13	68.13	71.27	73.11	1.89
	Definition (ens.)	56.72	**59.14**	61.33	1.48	90.48	91.79	93.25	0.77	70.20	**71.92**	73.37	0.93
Mixtral-47b	Simple	44.63	52.24	62.72	6.00	42.06	73.81	85.71	16.14	50.36	59.37	64.00	4.90
	Simple (ens.)	48.85	52.04	56.25	2.38	73.02	79.64	84.52	3.58	60.82	62.84	65.56	1.25
	Definition	40.24	44.47	51.72	3.93	83.73	89.92	92.46	3.22	56.08	59.32	63.94	2.71
	Definition (ens.)	42.25	44.48	46.45	1.46	89.68	91.03	92.46	0.93	57.68	59.74	61.48	1.24
GPT-4o	Simple	38.27	39.04	39.97	0.61	95.63	96.75	97.62	0.68	54.79	55.63	56.65	0.62
	Simple (ens.)	38.43	39.22	39.84	0.38	96.83	**97.10**	97.22	0.18	55.02	55.88	56.52	0.41
	Definition	51.16	55.32	58.76	2.80	86.51	93.57	97.22	3.78	66.85	69.41	71.73	1.58
	Definition (ens.)	53.51	55.56	57.99	1.43	93.65	95.12	96.83	1.14	68.93	70.13	71.62	0.90
Ensemble	Simple	42.53	44.60	45.73	1.16	94.44	95.08	95.63	0.40	58.72	60.71	61.87	1.14
	Definition	53.22	56.26	60.22	2.93	88.89	92.94	95.24	2.19	68.19	70.00	72.42	1.77
RoBERTa	NA		71.96				61.11				66.09		

Note: Majority-voting ensembles composed of three prompt variations are designated with '(ens.)', and their statistics are based on all choose-three-out-of-five combinations. The highest average scores are put in boldface.

TABLE 13 | Comparison of the ChatGPT-3.5 model, published before the onset of Monkeypox, with other LLMs and their ensembles, and with top-performing BERT-like models (Table 11).

Monkeypox					
Model category	**Model**	**Prompt**	**Precision**	**Recall**	***F*1**
BERT-like	RoBERTa	NA	71.96	61.11	66.09
	XLNet	NA	65.44	63.89	64.66
Open	Llama3-70b	Simple	43.11	95.63	59.43
		Definition	60.47	91.67	**72.87**
	Mixtral-47b	Simple	49.37	77.78	60.40
		Definition	51.72	83.73	63.94
Closed	GPT-4o	Simple	39.38	95.63	55.79
		Definition	58.76	86.51	69.98
	ChatGPT-3.5	Simple	44.55	90.87	59.79
		Definition	62.95	76.19	68.94
Ensemble	All 3	Definition	60.22	88.89	<u>71.79</u>
	GPT-4o-PE	Definition	57.49	94.44	71.47

Note: Other LLM models have training data cutoffs that overlap with the Monkeypox epidemic. The highest score is shown in bold and the runner up score is underlined.

When used with the simple prompt, the model is comparable to other LLMs.

We use this data-point to support the claim that zero-shot LLMs will perform competitively on new epidemics occurring after their knowledge cutoff. If ChatGPT-3.5 remains competitive on truly unseen data, we don't expect future LLMs to fall behind, since their performance should profit from improved architectures, training techniques and datasets.

Additionally, we provide two arguments for why the zero-shot LLM performance in Table 11 could be a realistic, and not overly optimistic, approximation of their performance on the texts from an emerging epidemic. Firstly, since the social posts supporting conspiracies are expectedly a rare minority class (Langguth et al. 2023), it is not likely that the models have seen much of the conspiratorial posts during pretraining. Second, large models applied to a new epidemics are likely to have seen some data related to it. Namely, of all the epidemics in our experiments only COVID-19 was a new disease, but even it was preceded by outbreaks of respiratory coronaviruses SARS (2002) and MERS (2012). Therefore, it is reasonable to expect that an LLMs was exposed to some relevant texts able to influence it's understanding of the epidemic. However, the question of the generalization ability of LLMs in this case, and the question how to measure the models' knowledge relevant to detection of medical conspiracies, are both open research questions.

5.4 | Performance of BERT-Like Models Trained on LLMs Labelled Datasets

To address *RQ4*, which investigates the use of LLM-labelled datasets in the absence of expert labelling, we fine-tuned the BERT-like models on LLM-labelled datasets and compared their performance to those fine-tuned on expert-labelled datasets. First, we examine the agreement between LLM labels and expert labels in Section 5.4.1, and then we elaborate on BERT-like models trained on LLM labels in Section 5.4.2.

5.4.1 | Comparing LLMs to Expert Annotators

Here we examine the question of feasibility of using the LLMs as data annotators. First we annotate each dataset by computing the CT labels using the models and prompts as in Section 5.3. To compare the LLM annotations with human ones we use Fleiss kappa, a chance-corrected measure of inter-annotator agreement. Table 14 contains the kappa values between LLM annotations and gold-standard human annotations (Section 3.2).

The kappa agreement values vary significantly, between fair and substantial (Landis and Koch 1977) agreement, depending on the configuration of the LLM serving as an annotator. Prompts including a precise definition of supporting a CT perform better in most cases, as do the largest LLM models Llama3-70b and GPT-4o. Llama3-70b and GPT-4o with definition prompt yield substantial or moderate agreement in all cases and, therefore, can be used as a feasible replacement for human annotators. Additionally, the ensemble of three LLMs achieves substantial agreement for Ebola, Zika and Monkeypox.

These results are in line with the related work that confirms that LLMs are reasonably reliable annotators (Ziems et al. 2024; Kuzman and Ljubeši'c 2025; Ornstein et al. 2025). A large study found that kappa scores of LLMs on 6 out of 16 classification tasks (Ziems et al. 2024) are above 0.5 (moderate agreement), with one task having kappa of 0.65 (substantial agreement). For the misinformation detection tasks which is

TABLE 14 | Comparison of the open and closed LLMs on conspiracy detection and closeness their results to the gold labels based on Fleiss Kappa.

Category	Model	Prompt	Fleiss Kappa			
			Ebola	Zika	COVID-19	Monkeypox
Open	Llama3-70b	Simple	0.347	0.431	0.634	0.410
		Definition	0.577	0.603	0.601	0.641
	Mixtral-47b	Simple	0.353	0.470	0.555	0.472
		Definition	0.534	0.501	0.473	0.516
Closed	GPT-4o	Simple	0.303	0.443	0.547	0.337
		Definition	0.601	0.603	0.597	0.605
Ensemble	All 3	Definition	0.627	0.630	0.593	0.629
	GPT-4o-PE	Definition	0.592	0.661	0.576	0.616

Note: The colour code for the Fleiss Kappa: Fair agreement, moderate agreement, substantial agreement. Definition prompts produced results that are more aligned with the experts (in terms of Fleiss Kappa coefficient).

most similar to conspiracy detection the kappa score is 0.55 (Ziems et al. 2024).

5.4.2 | BERT-Like Models Trained on LLM Labels

In this section we answer *RQ4* by analysing the performance of BERT-like models fine-tuned on LLM annotations. All the LLMs were combined with the 'definition' prompts (Section 5.4.2) due to their superior performance. For each dataset and each source of annotations, BERT-like models are trained using the method described in Section 4.1. We use BERT-like models fine-tuned on expert labels (Section 5.1) as a strong baseline.

Results, displayed in Table 15, show that BERT-like models trained on LLM labels often perform competitively with those trained on expert labels. Best results are achieved for two largest LLM models (Llama3-70b and GPT-4o), which also consistently achieve moderate or substantial agreement with human annotators (Section 5.4.2). Although the results vary depending on the BERT-like model used, LLM labels can lead to models with performance equal or superior to the performance of models based on expert labels. For the Monkeypox dataset the best performance is achieved for LLM labels, while for COVID-19 the performance is comparable. However, the Zika dataset, where expert labels obtain better results with a margin of up to eight $F1$ points, demonstrates that competitiveness of LLM labels can perhaps be expected but is not guaranteed. Performance of individual LLMs varies across datasets and even BERT-like models, but the ensembling strategy proves as an effective stabilizer that yields more consistent performance. This is in line with the findings from the LLM classification experiments (Section 5.3).

As observed in our experiments, the LLMs tend to misclassify not-CT texts as CT texts. This can be seen by observing their recall scores which are consistently higher than the precision scores (Table 11). Another set of statistics, displayed in Table A2, shows that the percentage of the CTs for the LLM labels is higher than for the expert labels. The performance of the models might be affected by an imbalance in label distribution, and it can lead to lower performance when the percentage of CT texts are underrepresented.

To investigate this issue, we experimented with two additional approaches. The first method is to apply a modified loss function that diminishes the loss score calculated on the examples labelled with the majority class. The second approach addresses the imbalance by undersampling the texts labelled with the majority class.[18] Under this approach, for each train-test split we randomly reduced the dataset-level majority class in the train split by removing between 10% and 50% of the examples. The results of the custom loss experiments are presented in Table A3, while the undersampling results are shown in Figure A4. Both approaches show that this approach does not lead to significant and consistent improvements and that it can degrade performance in a non-negligible number of cases. Under-performance of balancing methods points to systematic mislabelling by the LLMs, focused on specific categories of text. Namely, if the misclassifications are non-random, balancing methods can amplify these systematic errors by putting more weight on the misclassified CT examples. This seems to be the case, as the LLMs are prone to high recall and low precision, producing falsely positive learning examples, while the analysis in Section 5.5 suggests that they are also prone to mislabelling specific types of text, such as texts containing political or critical rhetoric.

5.5 | Error Analysis

The classifiers that we use for the detection of conspiracies, based on BERT-like models (Section 4.1) and on LLM models (Section 4.2), represent two distinct approaches to text classification. BERT-like models represent a discriminative framework, with contextual representations optimized on labelled data. LLMs are much larger models pre-trained on massive datasets, with the ability to interpret instructions and generate answers. In this section we perform a contrastive analysis of performance of these two model types.

As the first step, we examine how many tweets are correctly classified by both model types, and how many pose challenges to

TABLE 15 | Comparison of BERT-like models fine-tuned on LLM-generated labels and expert labels.

Ebola				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	66.99	48.59	56.33
	Mixtral-47b	44.33	63.57	50.97
	Llama3-70b	44.14	66.90	52.95
	GPT-4o	56.08	52.22	53.82
	GPT-4o-PE	50.33	67.98	**56.67**
	All 3	48.98	60.62	54.10
RoBERTa	Expert	68.63	49.30	**57.83**
	Llama3-70b	44.78	74.66	55.29
	Mixtral-47b	45.68	77.59	55.97
	GPT-4o	45.55	69.75	54.88
	GPT-4o-PE	47.80	71.92	56.79
	All 3	49.51	66.92	56.47
XLNet	Expert	61.36	57.04	**59.12**
	Mixtral-47b	40.77	72.64	51.39
	Llama3-70b	41.53	73.74	52.18
	GPT-4o	56.41	58.47	56.81
	GPT-4o-PE	44.34	72.54	54.45
	All 3	47.30	70.74	56.27
Zika				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	64.32	60.89	**62.56**
	Mixtral-47b	47.76	76.00	51.48
	Llama3-70b	49.03	74.22	58.77
	GPT-4o	54.89	64.89	59.22
	GPT-4o-PE	47.21	76.00	58.17
	All	49.29	70.22	56.97
RoBERTa	Expert	71.50	68.00	**69.70**
	Mixtral-47b	45.13	88.00	59.58
	Llama3-70b	53.32	80.44	64.03
	GPT-4o	54.33	74.67	62.63
	GPT-4o-PE	53.08	83.11	64.31
	All 3	48.02	84.44	61.18
XLNet	Expert	72.77	68.89	**70.78**
	Mixtral-47b	41.11	85.78	55.50
	Llama3-70b	48.16	80.89	59.89
	GPT-4o	52.37	79.11	62.90
	GPT-4o-PE	49.34	81.33	61.08
	All 3	50.22	82.67	62.38

(Continues)

TABLE 15 | (Continued)

COVID-19				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	77.83	79.25	**78.53**
	Mixtral-47b	66.15	84.78	74.23
	Llama3-70b	72.88	81.77	74.16
	GPT-4o	70.65	85.46	77.26
	GPT-4o-PE	69.52	87.40	<u>77.42</u>
	All 3	69.04	86.77	76.81
RoBERTa	Expert	81.10	82.67	**81.87**
	Mixtral-47b	66.88	86.09	75.25
	Llama3-70b	74.25	87.23	<u>80.19</u>
	GPT-4o	72.60	85.35	78.42
	GPT-4o-PE	71.50	86.83	78.35
	All 3	73.39	86.94	79.46
XLNet	Expert	80.98	81.97	**81.03**
	Mixtral-47b	66.53	87.11	75.41
	Llama3-70b	74.28	86.94	<u>80.05</u>
	GPT-4o	71.27	86.14	77.97
	GPT-4o-PE	71.86	87.45	78.82
	All 3	71.98	87.80	79.06
Monkeypox				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	63.95	56.35	<u>59.92</u>
	Mixtral-47b	43.20	69.48	52.89
	Llama3-70b	54.06	69.42	**60.76**
	GPT-4o	45.29	68.26	53.88
	GPT-4o-PE	48.13	76.98	58.99
	All 3	53.59	67.88	59.86
RoBERTa	Expert	71.96	61.11	66.09
	Mixtral-47b	52.27	78.96	62.17
	Llama3-70b	64.91	81.22	**71.53**
	GPT-4o	51.41	81.00	62.73
	GPT-4o-PE	54.11	80.97	64.36
	All 3	59.68	80.96	<u>68.42</u>
XLNet	Expert	65.45	63.89	<u>64.66</u>
	Mixtral-47b	50.95	73.07	59.50
	Llama3-70b	53.73	76.21	62.78
	GPT-4o	53.61	76.96	63.10
	GPT-4o-PE	50.82	79.36	61.83
	All 3	55.45	81.34	**65.36**

Note: The top (bold) and runner-up (underline) *F*1 scores for each model are highlighted. In most cases, models trained on expert labels achieve the best performance, while LLM-labelled datasets, particularly those labelled by Llama3-70b, GPT-4o and by the ensemble of all three LLMs, can lead to competitive results.

one or both types. To calculate the statistics we use per-instance categorizations of the three different models of each type, on a union of all four per-epidemic datasets. BERT-like classifiers are BERT, RoBERTa and XLNet (Section 5.1), while the LLM classifiers are Llama3-70b, Mixtral-47b and GPT-4o in combination with the definition prompt (Section 5.3).

The results, displayed in Table 16, show that BERT-like models make more errors than LLMs in 30.7% of all instances, whereas the opposite pattern occurs for 19.6% of instances. For 49.0% of instances, all models produce correct predictions. Interestingly, only a small fraction of instances on which models errors exhibit identical performance for both model types: for 0.3% of instances, BERT-like models and LLMs make the same number of errors, while for 0.4% of instances all models make a prediction error.

The minimal overlap in errors—with all models failing simultaneously on just 0.4% of instances—suggests that BERT-like models and LLMs are prone to different types of failure. To investigate this divergence, we focus on extreme cases: instances where all LLMs make an error but all BERT-based models succeed (52 instances), and vice versa (38 instances). We conducted a Chi-square test of independence to examine the relationship between the type of model failure and the ground-truth label (CT vs. not-CT). The test reveals a highly significant association, $\chi^2(1) = 71.88, p < 0.001$, confirming that the ground-truth label of the text strongly dictates which model type is likely to fail in these boundary cases. When LLMs fail and BERT-based models succeed, 51 out of 52 instances (98.1%) are not-CT, with only one being a true CT. Conversely, when BERT-based models fail and LLMs succeed, 34 out of 38 instances (89.5%) are true CTs, with only 4 being not-CT. These results indicate that the model types have different blind spots: in these boundary cases, LLMs overwhelmingly produce false positives, while BERT-like models are prone to false negatives. This contrast reflects the overall precision-recall trade-offs on the entire dataset: LLMs exhibit high recall but lower precision (Table 11), while BERT-like models have higher precision at the cost of lower recall (Table 8).

Next we perform a qualitative analysis of tweets for two contrasting error categories, one where all LLMs make an error and all BERT-based models succeed, and the other where the opposite is the case. We examine all the tweets from both categories in order to determine types of texts for which most of the error happens. Three most prevalent error categories for LLMs, together with representative examples, are presented in Table 17. LLMs frequently misclassify as CTs tweets containing political content, and the tweets being critical of main-stream medicinal truths. In addition, 11.0% of LLM errors involve failures to recognize irony, leading to the misclassifications of ironic statements as conspiracy theories.

For the texts where all BERT-based models make an error but all LLMs succeed, three most prevalent categories, together with representative examples, are displayed in Table 18. For these texts, where BERT-like models fail to identify CTs, majority of the errors involve conspiracies with race-related or political objectives (34.2%), conspiracies aimed at population control (10.5%), or claims about the dissemination of false information by the media or political actors (10.5%).

The qualitative analysis suggests that LLMs tend to falsely detect as CTs texts with critical and political character. The BERT-like models, that tend to miss true CTs, also have a tendency to err on tweets with explicitly political content. The former error type demonstrates the potential of LLMs for filtering out legitimate critical and political content.

6 | Discussion

The results of our experiments, summarized in Table 19, outline the comparative performance of approaches used to tackle the classification of CTs. We selected the models with consistently high performance across datasets and evaluation setups: the RoBERTa model represents BERT-like models while the ensemble of LLMs represents the LLMs. All the experiments except the in-domain baseline correspond to low resource scenarios where either annotations or computing resources are scarce. The medical CT detection problem is challenging and the performance varies widely across epidemics, as demonstrated by the results of BERT-like classifiers trained and tested on the same dataset (in-domain). The transfer of labels from previous epidemics yields mixed results, but in the case of Monkeypox its performance is competitive. LLM-based zero-shot classifiers tend to outperform all the other approaches, surpassing the performance of in-domain supervised classifiers for all epidemics except Zika, which makes them the approach to choose for CT detection. However, as the results of Section 5.3 show, individual LLM models have unreliable performance, which shows that the creation of an LLM ensemble is necessary for a predictable level of performance. Additionally, expert-crafted definitions of CT, such as the one provided in this paper, are necessary for optimal performance. Since the reliance on LLMs can be resource-intensive, we also examine the alternative of using the LLMs to create a labelled dataset for BERT-like models. This approach is below the zero-shot performance, but if can perform competitively or even surpass the in-domain models, which makes it a viable alternative when the available compute is limited.

LLMs are known to be prone to different kinds of biases (Gallegos et al. 2024). For our use case of detection of medical conspiracies, the qualitative analysis in Section 5.5 suggests that both LLMs and BERT-like models misclassify text containing political or critical speech. Additionally, LLM-based classifiers have a general tendency of classifying posts as CTs, as is demonstrated by their low precision scores (Table 11). It follows that misclassification of social media posts could have negative implications for the freedom of speech if AI models are aggressively used to automatize content moderation. Therefore, we believe that models for conspiracy detection should be used for automatic filtering of social media posts only with humans in the loop.

On the other hand, there are viable applications of CT classifiers to automatic data analysis. One is monitoring of trends in the social media streams (Srikanth et al. 2021). Findings from such analyses could lead to useful insights and guide human experts to examine phenomena of particular interest. Text classifiers are often applied to automatic content analysis in CSS, where scrutinizing their performance and using multiple methods to

validate the results is considered good practice (Grimmer and Stewart 2013).

Lastly, Table 20 summarizes which approach is most appropriate under different combinations of annotation availability and inference budget. In general, fine-tuned BERT-like models are suitable when target-epidemic labels are available and inference resources are limited, while cross-epidemic transfer is useful when only labels from previous epidemics are available, provided that narratives are sufficiently similar. When labels are scarce or unavailable, zero-shot LLM ensembles are preferable if inference budget permits, whereas using LLM-labelled data to train a BERT-like model can reduce repeated LLM inference costs under tighter budgets. It is worthy to note that this table is intended as practical guidance for selecting CT detection methods rather than as a fixed rule, since real-world deployment may involve additional factors beyond annotation and computational constraints.

TABLE 16 | Number of tweets for each of the categories describing the relations between errors made by all BERT-like models, and all individual LLM-models.

Error category	Number of tweets	Percentage
More errors BERT-like	3149	30.7
More errors LLMs	2010	19.6
All models correct	5032	49.0
All models incorrect	41	0.4
Same number of errors	35	0.3
Total	10,267	100.0

7 | Limitations

Platform Specificity: The datasets focus exclusively on posts from Twitter. Linguistic style, content moderation practices and community norms differ across platforms (e.g., Telegram, TikTok), which may affect the prevalence and expression of conspiratorial language. As a result, the findings may not fully generalize to other social media platforms. Nevertheless, our source code is publicly available, enabling replication on alternative platforms. While experiments for platforms with short-form content (e.g., Bluesky) can be directly executed, platforms characterized by longer texts (e.g., Reddit) may require additional preprocessing or adaptation of the input representations (e.g., use of generative AI for summarizing posts).

Language Scope: All experiments are conducted on posts in English and the BERT-like models used in this study are primarily pretrained on English-centric corpora. Therefore, the findings may not generalize to other languages or cultural contexts. CT narratives in other languages may employ different rhetorical strategies, and discourse patterns shaped by local sociopolitical and cultural norms. Extending this work to multilingual or cross-lingual settings remains an important direction for future research.

TABLE 17 | Most prevalent categories among texts where LLMs make an error but BERT-like models succeed.

Category	% of errors	Example
Texts that do not contain a conspiracy but are rich in political or ideological content.	28.8%	What if this whole thing has turned in favour of the white hats but it's also a great way to teach all nations including America what it feels like living under a communist regime...? Never waste a good pandemic...[LAUGHING] Vote NO to socialism.
Texts critical of vaccines or of the severity of the disease.	25.0%	Zika virus has been around since. Suddenly it causes microcephaly? No. Check the mandatory vaccines fo... The microcephaly cases in Brazil were caused by larvicide chemicals, not by Zika.
Texts critically discussing remedies or pharmaceutical practices without proposing a conspiracy.	19.2%	The cure has been around since the first Ebola outbreak, but major pharmaceutical companies refused to mass produce it.

TABLE 18 | Most prevalent categories among texts where BERT-like models make an error but LLMs models succeed.

Category	% of errors	Example
Conspiracy with objectives related to social class or race	34.2%	GMO mosquitos are a weapon against the poor. Zika did not appear till after they were released.
Conspiracy to control the population	10.5%	Please Tom understand who owns the Zika virus threatened to use mosquitoes to deliver agents to reduce populat
Conspiracy texts implying dissemination of lies by the politicians or the media	10.5%	Certainly looks possible that polio and Monkey Pox may well be Covid jab in disguise. 'They' are not to be trusted as we have found ALL politicians to be either ill informed and deceived (like rest of us) or bloody liars, the lot of them

TABLE 19 | Performance of best classification approaches in each category: BERT-like models trained on the labelled text from the same epidemic and from previous epidemics, LLM models used as zero-shot classifiers and BERT-like models trained on LLM-labelled texts from the same epidemic.

	Resource constraint	Ebola	Zika	COVID-19	Monkeypox
BERT-like in-domain		59.12	70.78	81.03	64.66
BERT-like transfer	Annotation	N/A	N/A	67.75	63.38
All 3 LLM ensemble	Annotation	65.13	67.54	82.28	71.79
All 3 BERT-like LLM-labels	Compute, annotation	56.81	64.31	80.19	68.42

TABLE 20 | Practical guidance for selecting CT detection methods under different annotation and inference-budget constraints.

Recommended approach	Target labels	Prior-epidemic labels	LLM budget
Fine-tuned BERT-like model	✓	✗	Limited
Cross-epidemic transfer with BERT-like models	✗	✓	Limited
Zero-shot LLM ensemble	✗	✗	Sufficient
LLM-labelled data + BERT-like model	✗	✗	Limited

Other Medical Topics: While our datasets cover four epidemics, our framework is not inherently limited to infectious disease outbreaks. Conspiracy narratives related to big pharma, attribution of malicious intent, rejection of official explanations and anti-vaccination, are largely topic-independent. Therefore, they may be observed across a wide range of health-related CTs, including the ones about non-viral health scares. In this context, LLMs are particularly promising. Because they rely less on topic-specific training data and more on general semantics, LLMs can be more readily adapted when labelled data is initially unavailable. At the same time, we acknowledge that non-viral health scares may introduce domain-specific vocabulary or context that could limit the performance of both BERT-like models and LLMs.

Fairness and Societal Impact: This work focuses on the technical performance of CT detection models across epidemics and it does not include a dedicated audit of fairness or bias. The datasets that we use do not contain geographic, demographic or other user-level metadata, which limits our ability to evaluate potentially disparate performance of classifiers across population groups. However, CT detection is a socially sensitive task, and model errors may disproportionately affect certain communities or political viewpoints. A comprehensive fairness assessment, including the analysis of performance on population subgroups, and harm evaluation, remains an important direction for future work.

Annotation Subjectivity: Distinguishing CTs from politically motivated, ironic, or satirical content is inherently subjective. Although a consistent annotation schema was applied and final labels were determined through majority voting, borderline cases inevitably require subjective judgement and may introduce annotator-related bias. While the Section 3.2 on data annotation considers the subjectivity and provides examples, a more systematic examination of annotator disagreement patterns and alternative labelling strategies remains an important direction for future work.

Real-World Deployment Considerations: Our experiments are conducted in controlled research settings and they do not consider downstream deployment scenarios such as content moderation pipelines, human-AI decision-making workflows, or policy enforcement contexts. Practical deployment would require additional analyses, focused on robustness to adversarial input, interpretability, transparency and on mechanisms for appealing the content removal. Therefore, our findings should be interpreted as methodological insights rather than deployment-ready solutions.

8 | Conclusion and Future Work

In this paper, we investigated methods for detecting conspiracy theories in epidemic-related social media posts under resource-constrained conditions. We introduced a novel multi-epidemic dataset containing CT-labelled texts and covering four recent epidemics: Ebola, Zika, COVID-19 and Monkeypox. Our experiments explored the performance of supervised CT classifiers across epidemics (*RQ1*), the viability of cross-epidemic label transfer (*RQ2*), the performance of zero-shot LLMs for CT classification (*RQ3*) and the use of LLMs as annotators of training data (*RQ4*).

The results of the experiments show that the performance of fine-tuned BERT-like models varies widely across epidemics (*RQ1*), which suggests that CT-related narratives and text-based signals for their detection are epidemic-specific. Cross-epidemic label transfer resulted in a suboptimal performance for COVID-19 and in a good performance for Monkeypox, which demonstrates the potential viability of the method and also supports the claim of epidemic-specific narrative structure (*RQ2*). Additionally, the transfer experiments showed that the performance increases as multiple prior epidemics are added to the training data. Zero-shot classification with LLMs proved to be the most successful approach for the detection of CTs (*RQ3*), yielding competitive or optimal performance across datasets (Table 11). However, it is crucial to use an ensemble of LLMs as individual models' performance can vary

significantly. BERT-like models trained on LLM-annotated data show potential and come close to or exceed the performance of models trained on gold data for three out of four epidemics (*RQ4*). Therefore, such models are a viable solution in cases when a large-scale inference is required and the annotation budget is scarce.

Data analysis of CT-related terms in Section 3 offers more detailed insights into the narratives of medical CTs, showing for example that CTs are commonly characterized by terms related to power structures and influential entities. The analysis of cross-epidemic transfer performed in Section 5.2 reveals that the CT narratives are epidemic-specific, as the CT-related terms vary across epidemics and the classifiers' performance depends on the source and the target epidemic.

The best-case performance of the classifiers reaches an $F1$ of approximately 70 for all diseases except COVID-19, which demonstrates that the classification of conspiracies in social media posts is a challenging problem. Having in mind the cross-epidemic variation in performance and potential for bias (Section 5.2), we believe that models for CT detection should not be used for filtering of social media posts without human supervision.

This research provides a useful foundation for future studies and practical applications in epidemic-related CT detection, by providing data analyses and experimental findings on the viability of solutions for low-resource scenarios. There exists a number of future work directions, mainly along the line of increasing the classification performance and of performing more elaborate data analyses.

Methods such as Adapter Fusion (Pfeiffer et al. 2021) might improve performance of BERT-like models in scenarios considered in this work. Namely, these methods have shown superior performance in handling imbalanced datasets for tasks similar to CT detection (Schlicht et al. 2023), and may improve both the transfer of labels from multiple previous epidemics, and the fusion of labels generated by several LLMs.

The performance of LLM classifiers might be improved with few-shot learning approaches based on the targeted selection of a small number of relevant CT examples, or with the inclusion of event-specific contextual information in the prompt. However, these approaches would increase the amount of needed resources, in terms of human time required to annotate the examples or to find relevant contextual information. The chain-of-thoughts approach (Wei et al. 2022) has shown potential to increase LLM performance on a variety of tasks, including text classification (Wu et al. 2025). It would be interesting to examine the benefit of these methods for our use-case, and to establish the tradeoff between performance and the added inference-time cost. Additionally, LLMs could be used to augment the training data with LLM-generated examples.

It would be interesting to perform an analysis of the amount and the nature of CT- and epidemic-specific knowledge contained in the LLMs, and to gauge the related generalization capabilities. Such analyses could provide insight about the LLM capabilities for zero-shot classification, and might answer the question whether the observed levels of LLM performance will hold for an unseen epidemics (Section 5.3).

Analysis, based on explainability techniques (Zhao et al. 2024), of text features that influence the predictions of LLM-based CT classifiers, has the potential to both improve the classifiers and detect their blind spots. Investigation of robustness of LLM-based CT classifiers to adversarial inputs (Przybyła et al. 2025), that is, CT posts designed to avoid detection, would provide important data on the feasibility of deployment of these classifiers in real-world setting.

Despite their importance, methods for quantifying and mitigating semantic biases occurring as a result of dataset creation (Ousidhoum et al. 2020; Elsafoury 2023), remain an underresearched topic. We plan to use our data, containing both large unfiltered collections and final filtered datasets, to investigate detection and mitigation of semantic biases, such as changes in the prevalence of specific conspiracy narratives. We attempt to extend methods from Section 3.3, and to examine the viability of the LLM-as-a-Judge approach (Li et al. 2024) for large-scale annotation of text with semantic categories related to bias quantification.

We plan to extend the data analysis experiments performed in this work in order to: (1) gain more insight into the similarities and differences of CT narratives across different epidemics and their temporal dynamic, and (2) explore the type of errors and biases that language models introduce when detecting CT posts. Such studies, employing approaches based on LLMs and topic models as well as qualitative analysis and coding by human experts, are expected to both improve the performance of AI detection methods and improve the understanding of CT-related discourse.

We also plan to investigate the transferability of fine-grained CT identification using a fine-grained labelling schema (Langguth et al. 2023), as such experiment could offer more insights into differences and similarities among epidemic-related CTs. Since our experiments are based on X posts in the English language, the performance and robustness of the methods should be investigated using data from other social platforms and in other languages.

Author Contributions

I.B.S. led the research and paper writing, wrote the first draft, contributed to conceptualization and methodology, curated and annotated the dataset, implemented the experiments and analysed the results. D.K. contributed to writing and conceptualization, experiments (implementing text de-duplication and LLM experiments), data annotation and result analysis. B.C. supervised data annotations and contributed to conceptualization. L.F. and P.R. supervised the work and provided financial support for the experiments. Lastly, all authors participated in editing and reviewing the paper.

Acknowledgements

The work of I.B.S. and L.F. was supported by DynSoDA (funded by BMBF) and the Lamarr Institute for Machine Learning and Artificial Intelligence. The work of D.K. was supported in part by the Croatian Science Foundation under the grant agreement UIP-2025-02-7498

(AGENTS). The work of P.R. was supported by the MARTINI project on Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence funded by MCIN/AEI/10.13039/501100011033 and by EU NextGenerationEU/PRTR. Views and opinions expressed are however those of the author(s) only and do not reflect those of the funding organizations, EU or HaDEA. We also thank Benedetta Togni for the annotations. The authors used ChatGPT for proof-reading. All generated content were reviewed and edited. Authors declare full responsibility for the manuscript content.

Funding

This work was supported by Bundesministerium für Bildung und Forschung; Lamarr Institute for Machine Learning and Artificial Intelligence; European Union NextGenerationEU/PRTR, PCI2022-135008-2; Croatian Science Foundation, UIP-2025-02-7498 (AGENTS); MCIN/AEI/10.13039/501100011033, PCI2022-135008-2.

Disclosure

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The source code, the IDs of the tweets and the trained models are publicly available at https://github.com/isspek/cross-epidemic-ct-detection. The full tweets are not publicly available due to platform restrictions, they can be collected by using the platform API.

Endnotes

[1] https://github.com/isspek/cross-epidemic-ct-detection.

[2] https://developer.x.com/en/docs/x-api.

[3] https://github.com/DocNow/hydrator.

[4] https://mediabiasfactcheck.com/.

[5] The semantic search engine is built upon Elasticsearch (https://www.elastic.co/) and DocArray (https://docarray.jina.ai/).

[6] https://github.com/seatgeek/thefuzz.

[7] https://github.com/sophiedkk/pyirr.

[8] https://www.bbc.com/news/technology-41900880.

[9] New World Order.

[10] https://huggingface.co/models.

[11] https://huggingface.co/docs/transformers/main_classes/trainer.

[12] https://openai.com/api.

[13] https://python.langchain.com/.

[14] https://huggingface.co/blog/inference-pro.

[15] https://friendli.ai/docs/guides/dedicated_endpoints/introduction.

[16] https://github.com/isspek/cross-epidemic-ct-detection.

[17] https://www.allmo.ai/articles/list-of-large-language-model-cut-off-dates.

[18] The majority class across the k-fold splits is not-CT across Zika, Monkeypox and Ebola. For COVID-19, it is CTs.

References

Alam, F., S. Shaar, F. Dalvi, et al. 2021. "Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society." In *Findings of the ACL: EMNLP 2021*, 611–649. ACL. https://doi.org/10.18653/v1/2021.findings-emnlp.56.

Balloccu, S., P. Schmidtová, M. Lango, and O. Dusek. 2024. "Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs." In *Proceedings of the 18th Conference of the EACL (Volume 1: Long Papers)*, edited by Y. Graham and M. Purver, 67–93. ACL.

Beer, E. M., and V. B. Rao. 2019. "A Systematic Review of the Epidemiology of Human Monkeypox Outbreaks and Implications for Outbreak Strategy." *PLoS Neglected Tropical Diseases* 13, no. 10: e0007791. https://doi.org/10.1371/journal.pntd.0007791.

Brown, T., B. Mann, N. Ryder, et al. 2020. "Language Models Are Few-Shot Learners." In *NIPS*, vol. 33, 1877–1901. Curran Associates, Inc.

Bucher, M. J. J., and M. Martini. 2024. "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification." arXiv. https://doi.org/10.48550/arXiv.2406.08660.

Buchholz, M. G. 2023. "Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset." arXiv. https://doi.org/10.48550/arXiv.2306.08190.

Caselli, T., O. Mutlu, A. Basile, and A. Hürriyetoğlu. 2021. "PROTEST-ER: Retraining BERT for Protest Event Extraction." In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events From Text (CASE 2021)*, 12–19. ACL. https://doi.org/10.18653/v1/2021.case-1.4.

Chae, Y., and T. Davidson. 2023. *Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning*. Open Science Foundation.

Chang, C., K. Ortiz, A. Ansari, and M. E. Gershwin. 2016. "The Zika Outbreak of the 21st Century." *Journal of Autoimmunity* 68: 1–13. https://doi.org/10.1016/j.jaut.2016.02.006.

Chen, C., and K. Shu. 2024. "Combating Misinformation in the Age of LLMs: Opportunities and Challenges." *AI Magazine* 45, no. 3: 354–368. https://doi.org/10.1002/aaai.12188.

Chong, Y. Y., H. Y. Cheng, H. Y. L. Chan, W. T. Chien, and S. Y. S. Wong. 2020. "Covid-19 Pandemic, Infodemic and the Role of Ehealth Literacy." *International Journal of Nursing Studies* 108: 103644.

Church, K. W., and P. Hanks. 1989. "Word Association Norms, Mutual Information, and Lexicography." In *Proceedings of the 27th ACL*, 76–83. ACL. https://doi.org/10.3115/981623.981633.

Ciotti, M., M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini. 2020. "The COVID-19 Pandemic." *Critical Reviews in Clinical Laboratory Sciences* 57, no. 6: 365–388. https://doi.org/10.1080/10408363.2020.1783198.

Corso, F., F. Pierri, and G. De Francisci Morales. 2025. "Conspiracy Theories and Where to Find Them on TikTok." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, 8346–8362. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.408.

Cui, C., and C. Jia. 2025. "Towards Real-World Rumor Detection: Anomaly Detection Framework With Graph Supervised Contrastive Learning." In *Proceedings of the 31st COLING*, edited by I. O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, 7141–7155. ACL.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the NAACL-HLT, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio, 4171–4186. ACL. https://doi.org/10.18653/v1/N19-1423.

Dharawat, A., I. Lourentzou, A. Morales, and C. Zhai. 2022. "Drink Bleach or Do What Now? Covid-Hera: A Study of Riskinformed

Health Decision Making in the Presence of Covid-19 Misinformation." *Proceedings of the ICWSM* 16, no. 1: 1218–1227. https://doi.org/10.1609/icwsm.v16i1.19372.

Di Sotto, S., and M. Viviani. 2022. "Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach." *International Journal of Environmental Research and Public Health* 19, no. 4: 2173. https://doi.org/10.3390/ijerph19042173.

Douglas, K. M., and R. M. Sutton. 2023. "What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication." *Annual Review of Psychology* 74, no. 1: 271–298. https://doi.org/10.1146/annurev-psych-032420-031329.

Du, J., S. Preston, H. Sun, et al. 2021. "Using Machine Learning–Based Approaches for the Detection and Classification of Human Papillomavirus Vaccine Misinformation: Infodemiology Study of Reddit Discussions." *Journal of Medical Internet Research* 23, no. 8: e26478. https://doi.org/10.2196/26478.

Elsafoury, F. 2023. "Thesis Distillation: Investigating the Impact of Bias in NLP Models on Hate Speech Detection." In *Proceedings of the Big Picture Workshop*, edited by Y. Elazar, A. Ettinger, N. Kassner, S. Ruder, and N. A. Smith, 53–65. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bigpicture-1.5.

Fast, E., B. Chen, and M. S. Bernstein. 2016. "Empath: Understanding Topic Signals in Large-Scale Text." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACL. https://doi.org/10.1145/2858036.2858535.

Feuer, A. 2014. "The Ebola Conspiracy Theories." https://www.nytimes.com/2014/10/19/sunday-review/the-ebola-conspiracy-theories.html.

Fleiss, J. L. 1971. "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin* 76, no. 5: 378–382. https://doi.org/10.1037/h0031619.

Fort, M., Z. Tian, E. Gabel, et al. 2023. "Bigfoot in Big Tech: Detecting Out of Domain Conspiracy Theories." In *Proceedings of the Conference RANLP*, 353–363. INCOMA Ltd. https://doi.org/10.26615/978-954-452-092-2_040.

Francis, W. N., and H. Kucera. 1979. "Brown Corpus Manual." *Letters to the Editor* 5, no. 2: 7.

Gallegos, I. O., R. A. Rossi, J. Barrow, et al. 2024. "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics* 50, no. 3: 1097–1179. https://doi.org/10.1162/coli_a_00524.

George, A. R., M. Ahrens, J. B. Pierrehumbert, and M. McMahon. 2024. "Conspiracy Detection Beyond Text: Exploring the Feasibility of Adding Psycho-Linguistic Features to Enhance Conspiracy Detection Models." In *Disinformation in Open Online Media*, 32–45. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71210-4_3.

Ghenai, A., and Y. Mejova. 2018. "Fake Cures: User-Centric Modeling of Health Misinformation in Social Media." *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW: 1–20. https://doi.org/10.1145/3274327.

Golchin, S., and M. Surdeanu. 2024. *Time Travel in Llms: Tracing Data Contamination in Large Language Models*. ICLR. OpenReview.net.

Grattafiori, A., A. Dubey, A. Jauhri, et al. 2024. "The Llama 3 Herd of Models." arXiv. https://doi.org/10.48550/arXiv.2407.21783.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21, no. 3: 267–297. https://doi.org/10.1093/pan/mps028.

He, H., X. Shi, J. Mueller, S. Zha, M. Li, and G. Karypis. 2021. "Distiller: A Systematic Study of Model Distillation Methods in Natural Language Processing." In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, edited by N. S. Moosavi, I. Gurevych, A. Fan, et al., 119–133. ACL. https://doi.org/10.18653/v1/2021.sustainlp-1.13.

Jacob, S. T., I. Crozier, W. A. Fischer, et al. 2020. "Ebola Virus Disease." *Nature Reviews Disease Primers* 6, no. 1: 1–31. https://doi.org/10.1038/s41572-020-0147-3.

Jiang, A. Q., A. Sablayrolles, A. Roux, et al. 2024. "Mixtral of Experts." arXiv. https://doi.org/10.48550/arXiv.2401.04088.

Kessler, J. 2017. "SCattertext: A Browser-Based Tool for Visualizing How Corpora Differ." In *Proceedings of ACL 2017, System Demonstrations*, 85–90. ACL. https://doi.org/10.18653/v1/P17-4015.

Kincaid, J. P., R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom. 1975. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel." Branch Report 8–75.

Kinsora, A., K. Barron, Q. Mei, and V. V. Vydiswaran. 2017. "Creating a Labeled Dataset for Medical Misinformation in Health Forums." In *2017 IEEE ICHI*, 456–461. IEEE. https://doi.org/10.1109/ICHI.2017.93.

Klein, C., P. Clutton, and A. G. Dunn. 2019. "Pathways to Conspiracy: The Social and Linguistic Precursors of Involvement in Reddit's Conspiracy Theory Forum." *PLoS One* 14, no. 11: e0225098. https://doi.org/10.1371/journal.pone.0225098.

Klofstad, C. A., J. E. Uscinski, J. M. Connolly, and J. P. West. 2019. "What Drives People to Believe in Zika Conspiracy Theories?" *Palgrave Communications* 5, no. 1: 36. https://doi.org/10.1057/s41599-019-0243-8.

Korenčić, D., B. Chulvi, X. Bonet-Casals, M. Taulé, P. Rosso, and F. Rangel. 2024. "Overview of the Oppositional Thinking Analysis PAN Task at CLEF 2024." Working Notes of CLEF.

Korenčić, D., B. Chulvi, X. B. Casals, A. Toselli, M. Taulé, and P. Rosso. 2024. "What Distinguishes Conspiracy From Critical Narratives? A Computational Analysis of Oppositional Discourse." *Expert Systems* 41, no. 11: e13671. https://doi.org/10.1111/exsy.13671.

Kou, Y., X. Gui, Y. Chen, and K. H. Pine. 2017. "Conspiracy Talk on Social Media: Collective Sensemaking During a Public Health Crisis." *Proceedings of the ACM on Human Computer Interaction* 1, no. CSCW: 21–61. https://doi.org/10.1145/3134696.

Kuzman, T., and N. Ljubeši'c. 2025. "LLM Teacher-Student Framework for Text ClassificationWith no Manually Annotated Data: A Case Study in IPTC News Topic Classification." *IEEE Access* 13: 35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33, no. 1: 159–174. https://doi.org/10.2307/2529310.

Langguth, J., D. T. Schroeder, P. Filkuková, S. Brenner, J. Phillips, and K. Pogorelov. 2023. "COCO: An Annotated Twitter Dataset of COVID-19 Conspiracy Theories." *Journal of Computational Social Science* 6, no. 2: 443–484. https://doi.org/10.1007/s42001-023-00200-3.

Li, H., Q. Dong, J. Chen, et al. 2024. "Llms-As-Judges: A Comprehensive Survey on Llm-Based Evaluation Methods." https://arxiv.org/abs/2412.05579.

Li, Z., X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. 2023. "Towards General Text Embeddings With Multi-Stage Contrastive Learning." arXiv Preprint arXiv:2308.03281.

Liu, Y., M. Ott, N. Goyal, et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv. https://doi.org/10.48550/arXiv.1907.11692.

Medina Serrano, J. C., O. Papakyriakopoulos, and S. Hegelich. 2020. "NLP-Based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, edited by K. Verspoor, K. B. Cohen, M. Dredze, et al. ACL.

Miani, A., T. Hills, and A. Bangerter. 2022. "LOCO: The 88-Million-Word Language of Conspiracy Corpus." *Behavior Research Methods* 54, no. 4: 1794–1817. https://doi.org/10.3758/s13428-021-01698-z.

Miguel, R. 2022. "Monkeypox Versus COVID-19: Can Recurrent Hoaxes Help Us Predict and Tackle the Next Infodemic?" https://www.disinfo.eu/publications/monkeypox-versus-covid-19-can-recurrent-hoaxes-help-us-predict-and-tackle-the-next-infodemic/.

Mu, Y., B. P. Wu, W. Thorne, et al. 2024. "Navigating Prompt Complexity for Zero-Shot Classification: A Study of Large Language Models in Computational Social Science." In *Proceedings of the 2024 Joint International Conference on LREC-COLING 2024*, 12074–12086. ELRA and ICCL.

Mukherjee, S., G. Weikum, and C. Danescu-Niculescu-Mizil. 2014. "People on Drugs: Credibility of User Statements in Health Communities." In *Proceedings of the 20th ACM SIGKDD KDD*, 65–74. ACM. https://doi.org/10.1145/2623330.2623714.

OpenAI, J. Achiam, S. Adler, et al. 2024. "GPT-4 Technical Report." arXiv. https://doi.org/10.48550/arXiv.2303.08774.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13, no. 2: 264–281. https://doi.org/10.1017/psrm.2024.64.

Ouattara, S., and N. Århem. 2021. "Fighting Ebola in the Shadow of Conspiracy Theories and Sorcery Suspicions:Reflections on the West African EVD Outbreak in Guinea-Conakry (2013–2016)." *Cahiers D'études Africaines* 241, no. 1: 9–39. https://doi.org/10.4000/etudesafricaines.33151.

Ousidhoum, N., Y. Song, and D.-Y. Yeung. 2020. "Comparative Evaluation of Label-Agnostic Selection Bias in Multilingual Hate Speech Datasets." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Emnlp)*, edited by B. Webber, T. Cohn, Y. He, and Y. Liu, 2532–2542. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.199.

Pelrine, K., A. Imouza, C. Thibault, et al. 2023. "Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4." In *Proceedings of the 2023 Conference on EMNLP*, edited by H. Bouamor, J. Pino, and K. Bali, 6399–6429. ACL. https://doi.org/10.18653/v1/2023.emnlp-main.395.

Peskine, Y., D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, and P. Rosso. 2023. "Definitions Matter: Guiding GPT for Multi-Label Classification." In *Findings of the ACL: EMNLP 2023*, edited by H. Bouamor, J. Pino, and K. Bali, 4054–4063. ACL. https://doi.org/10.18653/v1/2023.findings-emnlp.267.

Pfeiffer, J., A. Kamath, A. Rücklé, K. Cho, and I. Gurevych. 2021. "AdapterFusion: Non-Destructive Task Composition for Transfer Learning." In *Proceedings of the 16th Conference of the EACL: Main Volume*, edited by P. Merlo, J. Tiedemann, and R. Tsarfaty, 487–503. ACL. https://doi.org/10.18653/v1/2021.eacl-main.39.

Phillips, S. C., L. H. X. Ng, and K. M. Carley. 2022. "Hoaxes and Hidden Agendas: A Twitter Conspiracy Theory Dataset: Data Paper." In *Companion Proceedings of the Web Conference 2022*, 876–880. ACM. https://doi.org/10.1145/3487553.3524665.

Pogorelov, K., D. T. Schroeder, S. Brenner, and J. Langguth. 2021. "FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021." In *Mediaeval*, vol. 3181. CEUR-WS.org.

Pruss, D., Y. Fujinuma, A. R. Daughton, et al. 2019. "Zika Discourse in the Americas: A Multilingual Topic Analysis of Twitter." *PLoS One* 14, no. 5: e0216922. https://doi.org/10.1371/journal.pone.0216922.

Przybyła, P., E. McGill, and H. Saggion. 2025. "Attacking Misinformation Detection Using Adversarial Examples Generated by Language Models." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, edited by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, 27626–27642. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1405.

Pustet, M., E. Steffen, and H. Mihaljevic. 2024. "Detection of Conspiracy Theories Beyond Keyword Bias in German-Language Telegram Using Large Language Models." In *Proceedings of the 8th WOAH 2024*, edited by Y.-L. Chung, Z. Talat, D. Nozza, et al., 13–27. ACL. https://doi.org/10.18653/v1/2024.woah-1.2.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. *Language Models Are Unsupervised Multitask Learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rahman, M. M., D. Balakrishnan, D. Murthy, M. Kutlu, and M. Lease. 2021. "An Information Retrieval Approach to Building Datasets for Hate Speech Detection." NeurIPS Datasets and Benchmarks.

Reimers, N., and I. Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, 3980–3990. ACL. https://doi.org/10.18653/v1/D19-1410.

Sakketou, F., J. Plepi, R. Cervero, H. J. Geiss, P. Rosso, and L. Flek. 2022. "FACTOID: A New Dataset for Identifying Misinformation Spreaders and Political Bias." In *Proceedings of the Thirteenth LREC*, edited by N. Calzolari, F. Béchet, P. Blache, et al., 3231–3241. ELRA.

Samory, M., and T. Mitra. 2018. "'The Government Spies Using Our Webcams': The Language of Conspiracy Theories in Online Discussions." *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW: 1–24. https://doi.org/10.1145/3274421.

Schlicht, I. B., E. Fernandez, B. Chulvi, and P. Rosso. 2024. "Automatic Detection of Health Misinformation: A Systematic Review." *Journal of Ambient Intelligence and Humanized Computing* 15, no. 3: 2009–2021. https://doi.org/10.1007/s12652-023-04619-4.

Schlicht, I. B., L. Flek, and P. Rosso. 2023. "Multilingual Detection of Check-Worthy Claims Using World Languages and Adapter Fusion." In *Advances in Information Retrieval*, edited by J. Kamps, vol. 13980, 118–133. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28244-7_8.

Sclar, M., Y. Choi, Y. Tsvetkov, and A. Suhr. 2024. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting." In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net. https://openreview.net/forum?id=RIu5lyNXjT.

Sharifpoor, E., M. Okhovati, M. Ghazizadeh-Ahsaee, and M. Avaz Beigi. 2025. "Classifying and Fact-Checking Health-Related Information About COVID-19 on Twitter/X Using Machine Learning and Deep Learning Models." *BMC Medical Informatics and Decision Making* 25, no. 1: 73. https://doi.org/10.1186/s12911-025-02895-y.

Sivarajkumar, S., M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang. 2024. "An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study." *JMIR Medical Informatics* 12: e55318. https://doi.org/10.2196/55318.

Song, Y., G. Wang, S. Li, and B. Y. Lin. 2025. "The Good, the Bad, and the Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by L. Chiruzzo, A. Ritter, and L. Wang, 4195–4206. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.211.

Srikanth, M., A. Liu, N. Adams-Cohen, J. Cao, R. M. Alvarez, and A. Anandkumar. 2021. "Dynamic Social Media Monitoring for Fast-Evolving Online Discussions." In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3576–3584. ACM. https://doi.org/10.1145/3447548.3467171.

Stefanov, P., K. Darwish, A. Atanasov, and P. Nakov. 2020. "Predicting the Topical Stance and Political Leaning of Media Using Tweets." In *Proceedings of the 58th ACL*, edited by D. Jurafsky, J. Chai, N. Schluter,

and J. Tetreault, 527–537. ACL. https://doi.org/10.18653/v1/2020.acl-main.50.

Tamine, L., L. Soulier, L. Ben Jabeur, et al. 2016. "Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations." In *Proceedings of the 27th ACM Hypertext*, 159–168. ACM. https://doi.org/10.1145/2914586.2914589.

Tausczik, Y. R., and J. W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29, no. 1: 24–54. https://doi.org/10.1177/0261927X09351676.

Thakur, N. 2022. "MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 Monkeypox Outbreak, Findings From Analysis of Tweets, and Open Research Questions." *Infectious Disease Reports* 14, no. 6: 855–883. https://doi.org/10.3390/idr14060087.

Thapa, S., K. Rauniyar, H. Veeramani, A. Shah, I. Razzak, and U. Naseem. 2024. "Did You Tell a Deadly Lie? Evaluating Large Language Models for Health Misinformation Identification." In *WISE*, edited by M. Barhamgi, H. Wang, and X. Wang, 391–405. Springer Nature. https://doi.org/10.1007/978-981-96-0576-7_29.

Ullah, I., K. Khan, M. Tahir, A. Ahmed, and H. Harapan. 2021. "Myths and Conspiracy Theories on Vaccines and COVID-19: Potential Effect on Global Vaccine Refusals." *Vacunas* 22, no. 2: 93–97. https://doi.org/10.1016/j.vacun.2021.01.001.

Van Prooijen, J.-W., and K. M. Douglas. 2017. "Conspiracy Theories as Part of History: The Role of Societal Crisis Situations." *Memory Studies* 10, no. 3: 323–333. https://doi.org/10.1177/1750698017701615.

Vaswani, A., N. Shazeer, N. Parmar, et al. 2017. "Attention Is All You Need." In *NIPS*, vol. 30. Curran Associates, Inc.

Vosoughi, S., D. Roy, and S. Aral. 2018. "The Spread of True and False News Online." *Science* 359, no. 6380: 1146–1151. https://doi.org/10.1126/science.aap9559.

Wei, J., X. Wang, D. Schuurmans, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–24837.

Wilson, E. B. 1927. "Probable Inference, the Law of Succession, and Statistical Inference." *Journal of the American Statistical Association* 22, no. 158: 209–212.

Wolf, T., L. Debut, V. Sanh, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, edited by Q. Liu and D. Schlangen, 38–45. ACL. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Wu, H., Y. Zhang, Z. Han, et al. 2025. "Cot-Driven Framework for Short Text Classification: Enhancing and Transferring Capabilities From Large to Smaller Model." *Knowledge-Based Systems* 311: 113057. https://doi.org/10.1016/j.knosys.2025.113057.

Xia, M., Z. Jiang, H. Wang, J. Zhao, T. Hu, and G. Chen. 2025. "Ensembling Prompting Strategies for Zero-Shot Hierarchical Text Classification With Large Language Models." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, edited by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, 18189–18208. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.918.

Yang, J., D. Vega-Oliveros, T. Seibt, and A. Rocha. 2021. "Scalable Fact-Checking With Human-in-the-Loop." In *2021 IEEE WIFS*, 1–6. IEEE. https://doi.org/10.1109/WIFS53200.2021.9648388.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In *NIPS*, vol. 32. Curran Associates, Inc.

Ye, J., X. Chen, N. Xu, et al. 2023. "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models." arXiv. https://doi.org/10.48550/arXiv.2303.10420.

Zenone, M., and T. Caulfield. 2022. "Using Data From a Short Video Social Media Platform to Identify Emergent Monkeypox Conspiracy Theories." *JAMA Network Open* 5, no. 10: e2236993. https://doi.org/10.1001/jamanetworkopen.2022.36993.

Zhang, S., L. Dong, X. Li, et al. 2026. "Instruction Tuning for Large Language Models: A Survey." *ACM Computing Surveys* 58, no. 7: 1–36.

Zhao, H., H. Chen, F. Yang, et al. 2024. "Explainability for Large Language Models: A Survey." *ACM Transactions on Intelligent Systems and Technology* 15, no. 2: 1–38. https://doi.org/10.1145/3639372.

Zhao, W. X., K. Zhou, J. Li, et al. 2023. "A Survey of Large Language Models." *arXiv* 1, no. 2: 1–124.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh. 2021. "!Calibrate Before Use: Improving Few-Shot Performance of Language Models." In *Proceedings of the 38th International Conference on Machine Learning*, 12697–12706. PMLR. https://proceedings.mlr.press/v139/zhao21c.html.

Zhou, H., X. Wan, L. Proleev, et al. 2024. "Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering." In *International Conference on Learning Representations*, edited by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, vol. 2024, 49–70. IEEE. https://proceedings.iclr.cc/paper_files/paper/2024/file/003e438cf04e3caf0a5c908495e484fe-Paper-Conference.pdf.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50, no. 1: 237–291. https://doi.org/10.1162/coli_a_00502.

Appendix A

Dataset

Table A1 shows the distribution of the characteristic words to CT and not-CT in the Monkeypox and COVID-19 datasets.

TABLE A1 | Number of samples per characteristic words specific to CT and not CT in the Monkeypox and COVID-19 datasets.

		COVID-19			Monkeypox	
		Label distribution			Label distribution	
Label	Keyword	CT	Not-CT	Keyword	CT	Not-CT
CT	agenda	336	33	as public	12	3
	china	576	231	bill	63	21
	deep	918	447	bill gates	51	—
	deep state	885	384	control	60	78
	false	234	117	fear	60	38
	gates	606	249	gates	57	—
	nwo	375	30	lab	78	72
	state	1077	531	out	759	189
	word	996	453	virus	162	285
	wuhan	546	183	want to	33	30
Not-CT	conspiracy	495	102	case	477	24
	country	186	150	cases	360	24
	life	240	120	children	165	15
	my	870	588	gay	309	18
	once	168	96	have	630	123
	our	1318	1137	men	528	96
	qanon	531	78	not	555	174
	quarantine	255	132	sex	240	9
	spiritual	765	87	this	471	129
	thing	819	780	you	549	144

Task Prompts

We interact with the LLMs by using the short prompt in Figure A1 and the long prompt in Figure A2.

Short Prompt

```
Determine whether the tweet supports a conspiracy theory about {disease}.
Answer yes or no.

tweet: {text}
supports a conspiracy (YES or NO):
```

FIGURE A1 | The prompt describes only the classification task.

Long Prompt

```
Here is the precise definition of what 'supporting a conspiracy theory' means:
--- START OF DEFINITION ---
A post explicitly supports a conspiracy theory when it claims that a secret
agent or group with a hidden agenda is behind an event, usually with malicious
intent. It is not necessary that every element is present explicitly in the text,
but only some of them.
These are examples posts supporting a conspiracy theory (with
the element in parentheses): "Bill Gates is behind all of this" (agent),
"Of course this massive vaccination is done to plan mind-control microchips
into people" (agenda),
"Those awake know that the endgame of this whole charade is
a new world order" (intent).

A post implicitly supports a conspiracy theory if it hints at it,
for example by using the word plandemic.
The problem with detection of such conspiracy claims is that one needs to know
the hints such as short claims and keywords, which can
require knowing about the broader context.
If an broader context is referenced, use your knowledge
in order to understand the reference.
--- END OF DEFINITION ---

Determine, based on the provided definition, whether the tweet
supports a conspiracy theory about {disease}.
Answer yes or no.

tweet: {text}
supports a conspiracy (YES or NO):
```

FIGURE A2 | The prompt has the full definition of what supporting a CT means.

Long Prompt

```
This is the definition that describes
what 'supporting a conspiracy theory' means:
--- DEFINITION START ---
A text explicitly supports a conspiracy theory when it claims that
either a secret agent or group with a hidden agenda is behind an event,
usually with malicious intent. Not every element needs to be present
explicitly in the text, only some of them.
These are examples of texts that support a conspiracy theory
(the element is in parentheses): "Bill Gates is behind all these events" (agent),
"Of course this massive vaccination is done to put mind-control
microchips into people" (agenda),
"Those awake know that the purpose of this whole charade
is a new world order" (intent).

A text can implicitly support a conspiracy theory if it hints at it,
for example by using the word plandemic.
The problem with the detection of this type of conspiracy claims
is that one needs to know the hints such as short claims and keywords,
which can require knowing about the broader context.
If a broader context is referenced, use your knowledge
to interpret the reference.
--- DEFINITION END ---

By applying the above definition, decide if the social media text
supports a conspiracy theory about {disease_desc}.
Answer Yes or No.

social media text: {text}
supports a conspiracy (yes or no):
```

FIGURE A3 | A semantically equivalent variation of the long prompt in Figure A2 that respects its structure.

Additional Experiments for Tackling Imbalanced Labels

TABLE A2 | Number of samples labelled as CT in the training sets for each labelling strategies.

Training set	Expert	Llama3-70b	Mixtral-47b	GPT-4o	All 3
Ebola	102.6	200.80	204.60	168.20	178.80
Zika	162.0	300.60	355.60	224.80	279.20
COVID-19	1262.2	1527.0	1598.60	1514.00	1535.40
Monkeypox	181.6	275.60	293.80	267.40	268.60

Note: The average of the number in each training fold was used as the number of labels. The LLMs have a tendency to label more posts as CTs than the experts.

TABLE A3 | Results of the BERT-like models trained on LLM-labelled data using weighted loss.

Ebola				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	67.68	47.18	**55.60**
	Mixtral-47b	45.88	49.38	46.83
	Llama3-70b	45.48	62.64	51.81
	GPT-4o	53.13	43.08	46.37
	All 3	44.95	57.68	49.68
RoBERTa	Expert	59.35	51.41	55.09
	Mixtral-47b	50.99	69.06	58.17
	Llama3-70b	42.24	73.97	53.70
	GPT-4o	43.32	58.42	48.83
	All 3	49.41	71.87	**58.43**
XLNet	Expert	70.37	53.52	**60.80**
	Llama3-70b	37.38	71.85	48.88
	Mixtral-47b	45.28	69.06	52.50
	GPT-4o	49.27	61.31	53.88
	All 3	53.16	67.00	58.02
Zika				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	64.66	60.00	**62.21**
	Mixtral-47b	36.99	85.33	51.41
	Llama3-70b	43.76	75.56	55.23
	GPT-4o	50.71	64.00	56.25
	All 3	50.96	72.89	59.79
RoBERTa	Expert	74.22	74.22	**74.22**
	Mixtral-47b	43.91	89.33	58.79
	Llama3-70b	51.38	83.11	63.31
	GPT-4o	57.53	75.56	64.72
	All 3	53.03	84.89	65.26
XLNet	Expert	69.13	70.67	**69.89**
	Mixtral-47b	46.06	81.78	58.69
	Llama3-70b	47.77	83.11	60.16
	GPT-4o	49.74	75.56	59.53
	All 3	47.12	80.00	59.00

(Continues)

TABLE A3 | (Continued)

COVID-19				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	79.53	80.39	**79.95**
	Mixtral-47b	65.55	85.75	74.26
	Llama3-70b	73.14	84.44	78.31
	GPT-4o	70.82	84.43	76.94
	All 3	70.75	85.97	77.56
RoBERTa	Expert	81.34	81.01	**81.18**
	Mixtral-47b	67.84	85.12	75.48
	Llama3-70b	74.53	86.83	80.18
	GPT-4o	73.26	86.15	79.16
	All 3	73.40	87.00	79.61
XLNet	Expert	81.31	78.39	79.83
	Mixtral-47b	67.58	84.43	74.98
	Llama3-70b	74.26	87.29	**80.19**
	GPT-4o	72.40	84.61	77.97
	All 3	72.54	85.23	78.36
Monkeypox				
Model	**Labelling strategy**	**Precision**	**Recall**	***F*1**
BERT	Expert	62.11	46.83	53.39
	Mixtral-47b	42.45	66.64	51.76
	Llama3-70b	53.35	74.64	**61.82**
	GPT-4o	46.10	71.03	55.81
	All 3	54.76	66.24	59.90
RoBERTa	Expert	64.34	73.02	**68.40**
	Mixtral-47b	54.01	78.59	63.61
	Llama3-70b	55.97	82.16	66.33
	GPT-4o	56.04	79.39	65.27
	All 3	59.79	78.94	67.78
XLNet	Expert	69.09	60.32	**64.41**
	Mixtral-47b	48.99	75.39	59.24
	Llama3-70b	53.14	75.79	62.31
	GPT-4o	54.45	67.85	60.16
	All 3	53.94	74.65	62.29

Note: The top (bold) and runner-up (underline) *F*1 scores per model group are highlighted. Overall, models trained with LLM-generated labels—particularly those labelled by Llama3-70b or the ensemble of all three LLMs often achieve performance comparable to models trained on expert-labelled data.

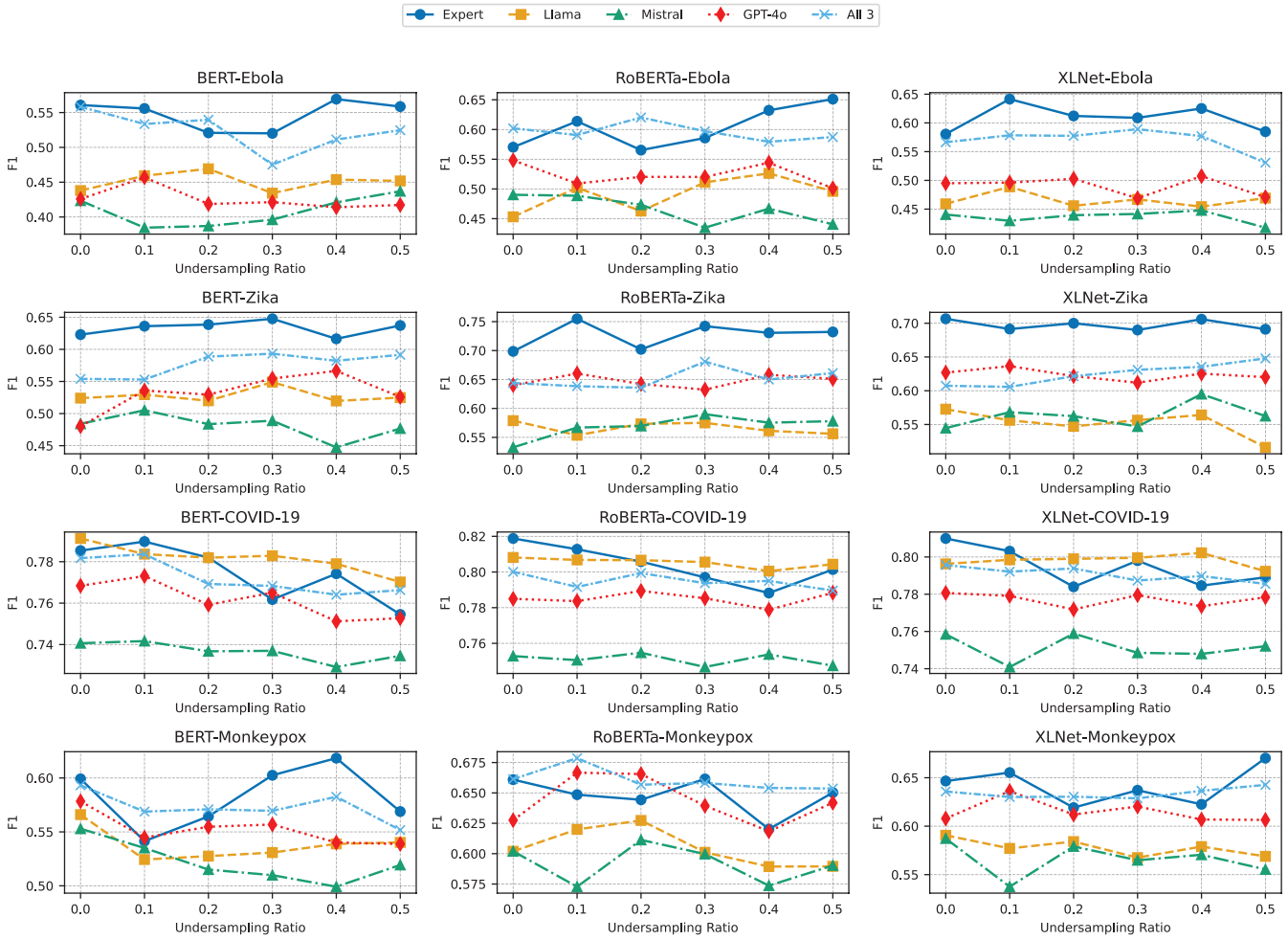

FIGURE A4 | Comparison of *F*1 scores for BERT-like fine-tuned on LLM-labelled datasets with various under-sampling ratios versus expert-labelled datasets. When LLM-labelled datasets closely align with expert labels, the performance of BERT-like performs similarly to those trained on expert-labelled datasets.

This section presents the results of additional experiments for tackling imbalance in labels of the LLM-labelled training samples. Table A2 presents the average number of samples labelled as CT by the LLMs. Table A3 shows the results when the models were trained with a weighted loss on the LLM-labelled training samples. Lastly, Figure A4 compares the BERT-like models under different under-sampling ratios.

Training the Simple Baselines

To train a TF-IDF + Logistic Regression classifier we perform a grid search over n-gram ranges (uni-gram, bi-gram), regularization strengths (0.1, 1, 10) and apply class-weights. The hyperparameters are selected based on macro-*F*1 performance on the development set. The configuration achieving the highest development *F*1 score is selected as the final baseline model.

TABLE A4 | Statistics of the probabilities of the most likely candidate for the first generated token: mean, standard deviation, median and the 1st percentile (1% of the values are lower and 99% are higher).

Ebola					
Model	**Prompt**	**Mean**	**Std**	**Median**	**Q01**
Llama3-70b	Simple	97.52	6.62	99.75	62.08
	Definition	95.71	8.50	98.91	52.26
Mixtral-47b	Simple	97.28	9.00	100.00	51.75
	Definition	92.40	13.39	99.59	48.40
GPT-4o	Simple	92.83	13.35	99.75	49.78
	Definition	98.31	6.13	99.93	62.21
Zika					
Model	**Prompt**	**Mean**	**Std**	**Median**	**Q01**
Llama3-70b	Simple	97.34	6.81	99.77	62.20
	Definition	95.34	9.31	99.03	52.23
Mixtral-47b	Simple	97.30	8.91	100.00	52.18
	Definition	93.70	12.21	99.74	49.89
GPT-4o	Simple	92.90	12.92	99.59	49.96
	Definition	97.59	7.43	99.91	56.20
COVID-19					
Model	**Prompt**	**Mean**	**Std**	**Median**	**Q01**
Llama3-70b	Simple	96.38	8.23	99.75	56.22
	Definition	92.42	13.54	99.15	49.91
Mixtral-47b	Simple	98.49	6.55	100.00	61.26
	Definition	97.08	9.32	100.00	51.39
GPT-4o	Simple	93.02	12.95	99.81	50.00
	Definition	97.60	6.85	99.92	61.30
Monkeypox					
Model	**Prompt**	**Mean**	**Std**	**Median**	**Q01**
Llama3-70b	Simple	95.79	9.17	99.64	54.30
	Definition	92.37	12.40	97.81	47.74
Mixtral-47b	Simple	96.59	9.76	100.00	52.89
	Definition	93.04	13.22	99.91	48.72
GPT-4o	Simple	92.33	13.55	99.59	49.87
	Definition	97.10	8.06	99.88	56.02

Note: For each combination of a model and prompt variant, probabilities are collected over all tweets in the dataset and all prompt variations/rephrasing. The statistics show that the models' certainty about the first generated token, one that determines the classification decision, is very high.

TABLE A5 | Conspiracy detection performance, on Monkeypox data, for LLM classifiers with both greedy decoding (default) and non-deterministic (sampling) decoding, for temperatures of 0.8 and 1.0.

Monkeypox				
Model	**Prompt**	**Precision**	**Recall**	***F*1**
Llama3-70b	Simple	43.11	95.63	59.43
	Simple-T0.8	42.93	96.43	59.41
	Simple-T1.0	43.19	95.63	59.51
	Definition	60.47	91.67	72.87
	Definition-T0.8	60.96	90.48	72.84
	Definition-T1.0	61.02	90.08	72.76
Mixtral-47b	Simple	49.37	77.78	60.40
	Simple-T0.8	49.62	78.57	60.83
	Simple-T1.0	49.87	78.17	60.90
	Definition	51.72	83.73	63.94
	Definition-T0.8	51.10	82.94	63.24
	Definition-T1.0	51.60	83.33	63.73
GPT-4o	Simple	39.38	95.63	55.79
	Simple-T0.8	39.53	94.44	55.74
	Simple-T1.0	39.44	95.63	55.85
	Definition	58.76	86.51	69.98
	Definition-T0.8	60.05	88.89	71.68
	Definition-T1.0	60.56	86.51	71.24

In order to assess the influence of temperature on our zero-shot classification setup, we first examine the distribution of probability mass over most likely candidates for the first token generated by an LLM (the token that basically decides the yes/no answer). Token probabilities are extracted from the LLM APIs by soliciting top-K tokens with their log-probabilities (by setting `logprobs=True`, `top_logprobs=K`). We solicited the maximum number of 20 tokens and confirmed that the remaining probability mass is negligible, for all models and prompts.

Table A4 contains the statistics of the top-token probabilities, collected over all tweets and all prompt variations. It shows that a large majority of the probability mass is concentrated in the top token, for all diseases and prompt variants. This suggests that sampling the first token is very unlikely to deviate from the selecting the token with the highest probability.

In order to confirm this empirically, we calculated the performances of LLM classifiers for Monkeypox, with temperature set to 0.8 and 1.0, and using the basic sampling strategy (Top-k and Top-p sampling disabled). We focus on Monkeypox since its top-token probabilities, in comparison with probabilities for other diseases, tend to have lower values and larger standard deviations, which makes sampling more likely to select other tokens. The results in Table A5 show that the impact of temperature on classification performance is small. The largest performance differences, obtained for the GPT-4o model with the Definition prompt variant, stay well below the differences caused by the variation of prompt formulations (Table 12).

Based on the above findings, we conclude that sampling-based decoding should have a negligible impact on our approach to the zero-shot detection of CT texts. These results are in line with a recent experiment (Song et al. 2025) showing that for tasks with a 'constrained output space', greedy decoding is comparable to non-deterministic decoding (Song et al. 2025).

The Influence of Calibration on LLM Performance

LLM calibration methods are designed to mitigate inherent LLM biases that influence the underlying probability of the generated tokens, and that can cause performance instability in prompting and in-context learning (Zhao et al. 2021; Zhou et al. 2024). These methods work by estimating a model's baseline token probabilities, for example by feeding the model neutral content-free input (Zhao et al. 2021), or by averaging over texts in a sample (Zhou et al. 2024). During inference, calibration methods adjust the model's token probabilities by taking into account the unbiased baseline.

Calibration methods have shown promise in improving LLM performance on zero-shot and few-shot classification tasks (Zhao et al. 2021; Zhou et al. 2024). In order to estimate the effect of calibration on the performance of zero-shot LLMs for CT detection, we experiment with the seminal 'contextual calibration' method (Zhao et al. 2021), and the more recent 'batch calibration' method (Zhou et al. 2024).

Contextual Calibration

Contextual calibration starts by inserting content-free inputs, such as the string 'N/A' or an empty string, into the prompt template in order to obtain a set of baseline probability distributions over tokens (Zhao et al. 2021). It then calculates the parameters of an affine transformation that shifts baseline distributions to the uniform distribution—the expected result for a content-free input. During inference, the transformation is applied to the model's uncalibrated output token probabilities.

A hidden assumption of the contextual calibration is that a model, when combined with a prompt, will produce a stable set of tokens regardless of the input. Concretely, the set of tokens capturing the majority of probability mass should fall to the tokens corresponding to the variations of the class labels (`YES` and `NO` in our case).

However, we observed that the output tokens of content-free inputs significantly differ from the output tokens of real CT posts. We observed this for all LLM models, both prompt variants, and for all five content-free strings that we experimented with (`'N/A'`, `'[MASK]'`, `'EMPTY TEXT'`, `'…'`). Table A6 contains representative examples of top candidates for a first generated token, and the start of the generated text, for two content-free inputs and two CT posts. Basically, the models detects when the provided text is semantically empty and, instead of providing a yes/no answer, it tries to communicate that a meaningful input should be provided. Therefore, content-free inputs provide no useful data on the probabilities of relevant yes/no tokens, on which the affine transformation, used for calibration, is to be trained. With the described model outputs, we expect the probabilities of yes/no tokens to be vastly overestimated (mapped-up to match the uniform distribution) after transformation. This would happen since their probability, given content-free input, is in most cases negligible. Therefore we decided not to continue with the application of the contextual calibration.

Our explanation for this occurrence is the key difference between LLM models that we used, and the models used in the original experiment (Zhao et al. 2021). Namely, we experiment with models that are larger, trained on more data, aligned using RLHF and instruction-tune. In the original experiment, the authors used GPT-3 (2.7B, 13B and 175B parameters) (Brown et al. 2020) and GPT-2 (Radford et al. 2019) (1.5B parameters) (Zhao et al. 2021). These models are very capable next-token predictors that were neither instruction-tuned nor aligned to act as conversational assistants. Since they were trained to continue a text pattern, if they were fed a prompt ending in `'Input: N/A Sentiment:'`, they would tend to output the most likely next token from their pre-training distribution, often defaulting to a variation of `Positive` or `Negative`.

Batch Calibration

The Batch Calibration method (Zhou et al. 2024) works by estimating a baseline by averaging the next-token probability distributions over a sample (batch) of unlabelled data. For each text in the batch, it is inserted into the prompt, the model is prompted, and the probabilities are collected. The method then calibrates the inference-time predictions, for an input text, by dividing the next-token probabilities by the estimated baseline probabilities. In effect, the method adjusts the next-token probabilities of an input text based on next-token probabilities averaged across the batch.

In our experiments, we used the batch size of 100. Since the LLM APIs expose only top-20 tokens and log-probabilities per response, rather than the distribution over the full vocabulary, we adapted the Batch Calibration algorithm as follows. During the estimation phase, we define the 'batch vocabulary' as the union of lists of top-20 tokens across all batch samples. For each sample, tokens present in its top-20 list retain their reported log-probabilities, and we distribute the remainder of the probability mass uniformly among the remaining tokens of the batch vocabulary. We then proceed to average, over the entire batch, the probabilities of all tokens in the batch vocabulary. At inference time, only the response tokens present in the batch vocabulary are calibrated, and the tokens not seen during estimation are left uncalibrated. The final label is derived from the token with the highest calibrated score.

TABLE A6 | Examples of 10 top-probability candidates for a first generated token, and the start of the generated text, for two content-free inputs and two CT posts.

		Top-10 first-token candidates and log-probabilities									
Input type	**Response**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**
Content-free	`I'm ready! Please`	`I`	`Since`	`Please`	`Unfortun`	`You`	`To`	`Waiting`	`This`	`The`	`There`
		− 0.01	− 5.26	− 5.88	− 8.26	− 9.26	− 9.88	− 10.38	− 10.63	− 10.76	− 10.88
	`To determine whether`	`To`	`The`	`Please`	`There`	`Could`	`It`	`I`	`I'm`	`Based`	`In`
		− 1.32	− 1.57	− 1.57	− 2.57	− 2.57	− 3.32	− 3.57	− 3.82	− 3.82	− 4.32
Real tweets	`YES`	`YES`	`Based`	`According`	`Yes`	`The`	`NO`	`_ YES`	`I`	`Answer`	`After`
		− 0.06	− 2.93	− 6.18	− 7.43	− 8.68	− 9.31	− 9.56	− 9.68	− 11.18	− 11.43
	`NO`	`NO`	`No`	`Based`	`**`	`YES`	`The`	`_ NO`	`supports`	`Given`	`Yes`
		− 0.00	− 8.00	− 9.50	− 12.00	− 12.13	− 12.75	− 12.88	− 14.13	− 14.50	− 14.50

TABLE A7 | Performance of LLM classifiers adapted using the batch calibration method (with the batch size of 100).

		Ebola											
		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	26.64	27.75	29.37	1.14	94.37	**96.20**	97.18	1.05	41.82	43.06	44.96	1.28
	Simple (BC)	27.25	28.92	30.43	1.26	83.09	88.68	91.18	3.11	41.96	43.57	45.16	1.30
	Definition	39.27	42.74	45.71	2.41	86.62	89.58	91.55	1.75	54.97	57.81	60.66	2.04
	Definition (BC)	46.50	49.04	50.53	1.63	68.38	78.68	86.03	7.85	55.36	**60.29**	63.24	2.96
Mixtral-47b	Simple	23.83	34.85	52.48	9.78	37.32	67.32	76.06	15.03	36.21	43.18	48.29	4.38
	Simple (BC)	23.04	40.39	56.67	11.55	25.00	44.41	66.18	16.47	26.91	40.30	52.66	10.47
	Definition	28.29	34.85	42.55	4.61	84.51	88.73	90.85	2.36	43.14	49.80	56.60	4.36
	Definition (BC)	45.76	**59.20**	67.44	7.69	17.65	39.85	59.56	17.20	27.75	44.10	57.81	11.80
GPT-4o	Simple	20.73	23.45	24.95	1.48	91.55	94.93	96.48	1.80	34.12	37.58	39.53	1.87
	Simple (BC)	20.66	23.36	26.59	1.93	50.00	67.35	80.15	10.91	30.98	34.51	39.93	3.01
	Definition	39.58	45.25	50.43	3.88	82.39	88.03	92.25	3.36	55.39	59.57	62.57	2.66
	Definition (BC)	45.33	53.23	57.93	4.33	56.62	67.79	75.00	7.26	54.80	59.25	63.46	3.20
		Zika											
		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	34.26	36.62	37.80	1.32	96.44	**97.07**	98.22	0.67	50.80	53.15	54.32	1.32
	Simple (BC)	37.66	39.00	40.97	1.11	82.57	89.45	92.66	4.02	51.72	54.30	56.82	1.64
	Definition	45.73	48.32	50.60	2.09	91.11	93.16	95.11	1.28	61.19	63.60	65.62	1.83
	Definition (BC)	41.03	46.70	51.25	3.45	44.04	65.78	84.40	15.12	42.48	54.25	63.78	7.54
Mixtral-47b	Simple	40.35	45.53	50.94	3.99	36.00	66.31	81.33	15.65	42.19	52.57	57.88	5.69
	Simple (BC)	44.51	51.27	57.21	4.99	25.69	44.95	66.97	14.49	35.00	46.77	59.47	9.42
	Definition	39.66	42.95	50.38	3.79	88.44	91.64	93.78	1.83	55.75	58.34	64.19	2.98
	Definition (BC)	55.10	60.28	64.41	3.34	10.09	29.63	50.00	14.26	17.19	37.56	55.19	13.50
GPT-4o	Simple	35.75	36.75	37.87	0.76	91.11	95.56	96.89	2.23	52.16	53.07	54.39	0.73
	Simple (BC)	36.84	39.79	44.23	2.88	51.38	65.78	73.85	8.85	42.91	49.39	55.33	4.03
	Definition	48.75	53.10	57.70	3.18	76.89	90.67	95.56	7.04	64.55	**66.73**	70.79	2.28
	Definition (BC)	58.87	**62.48**	66.23	3.01	46.79	58.99	66.97	8.13	54.84	60.28	65.75	4.23
		COVID-19											
		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	73.23	75.42	76.61	1.18	93.16	94.13	95.61	0.81	82.94	**83.73**	84.19	0.45
	Simple (BC)	76.06	77.18	78.31	0.92	82.84	86.23	89.28	2.41	79.30	81.45	83.44	1.58
	Definition	72.74	74.48	75.84	1.30	90.94	91.52	92.36	0.59	81.39	82.11	82.71	0.56
	Definition (BC)	76.56	77.75	79.02	0.98	53.95	67.74	80.73	9.82	63.52	71.99	78.59	5.44
Mixtral-47b	Simple	69.09	73.33	79.58	3.49	65.11	83.91	91.51	9.52	71.62	77.74	80.29	3.11
	Simple (BC)	73.17	76.28	82.13	3.20	57.88	74.77	85.76	11.83	67.90	74.76	79.87	5.05
	Definition	65.75	67.16	69.53	1.28	88.08	90.65	92.36	1.44	76.62	77.14	77.72	0.40
	Definition (BC)	69.59	73.37	79.07	3.17	61.75	72.52	83.25	7.87	69.34	72.55	75.84	2.74

(Continues)

TABLE A7 | (Continued)

		COVID-19											
		Precision				Recall				*F*1			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
GPT-4o	Simple	68.54	69.48	71.39	1.01	95.32	**95.87**	96.18	0.32	79.90	80.56	81.64	0.58
	Simple (BC)	72.41	74.97	78.17	1.86	43.35	58.22	65.91	8.18	54.23	65.24	70.14	5.78
	Definition	71.37	73.24	75.43	1.35	90.48	91.96	92.93	0.89	80.73	81.52	82.27	0.50
	Definition (BC)	76.98	**78.84**	81.99	1.80	33.98	44.66	55.24	6.99	47.37	56.67	64.32	5.61
		Monkeypox											
		Precision				**Recall**				***F*1**			
Model	**Prompt**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**	**Min**	**Avg**	**Max**	**Std**
Llama3-70b	Simple	41.74	43.07	43.53	0.68	94.84	95.71	97.22	0.81	58.40	59.40	59.80	0.51
	Simple (BC)	44.42	46.49	47.72	1.32	93.22	94.49	96.19	1.14	60.78	62.30	63.28	1.00
	Definition	53.94	58.67	61.52	2.83	89.29	90.95	92.46	1.13	68.13	71.27	73.11	1.89
	Definition (BC)	61.95	64.12	66.19	1.68	72.88	80.93	88.98	6.04	68.80	**71.38**	73.04	1.49
Mixtral-47b	Simple	44.63	52.24	62.72	6.00	42.06	73.81	85.71	16.14	50.36	59.37	64.00	4.90
	Simple (BC)	45.45	55.95	68.55	7.66	35.59	55.42	77.54	17.31	42.21	53.75	64.55	9.05
	Definition	40.24	44.47	51.72	3.93	83.73	89.92	92.46	3.22	56.08	59.32	63.94	2.71
	Definition (BC)	53.33	65.41	72.88	7.18	18.22	43.81	65.25	17.66	29.15	49.19	63.90	12.42
GPT-4o	Simple	38.27	39.04	39.97	0.61	95.63	**96.75**	97.62	0.68	54.79	55.63	56.65	0.62
	Simple (BC)	38.81	42.38	44.60	2.31	62.71	76.02	89.83	8.85	50.45	54.16	56.97	2.49
	Definition	51.16	55.32	58.76	2.80	86.51	93.57	97.22	3.78	66.85	69.41	71.73	1.58
	Definition (BC)	62.61	**66.51**	68.91	2.09	55.93	60.76	69.49	4.78	60.70	63.43	69.20	3.13

Note: The statistics for the results are calculated over 5 prompt variations (Section 5.3.1). The highest average scores are shown in bold.

Table A7 displays the performance of batch-calibrated zero-shot classifiers, averaged over all five prompt variations (Section 5.3.1), alongside the performance of non-calibrated models. For each calibrated model, the performance scores are calculated by randomly sampling a batch of 100 tweets for calibration, and using the rest of the data to estimate the performance. In contrast to the uncalibrated models, the calibrated classifiers have increased precision, decreased recall and higher variance. The shift in the precision-recall tradeoff suggests that the calibration manages to soften the tendency of LLMs to over-classify tweets as conspiracies. However, the overall performance, in terms of the *F*1 score, is lower in the majority of cases. The relative performance between calibrated and original classifiers depends on the model and on the dataset. For example, for the Ebola dataset the calibration is beneficial for both Llama3-70b and GPT-4o, while in the case of Monkeypox, only Llama3-70b profits from calibration. For an LLM, calibration gains can vary depending on the prompt variant, as is the case for GPT-4o on Monkeypox. Only the Mixtral-47b models consistently fails to profit from calibration.

Because of the lack of consistent gains, and of the increase in *F*1 variance, we do not recommend using batch calibration for our approach and use-case. However, it might be a useful tool if boosting the models' precision is important.