RESEARCH-ARTICLE

Tripartite Intelligence: Synergizing Deep Neural Network, Large Language Model, and Human Intelligence for Public Health Misinformation Detection (Archival Full Paper)

YANG ZHANG, University of Illinois Urbana-Champaign, Urbana, IL, United States

RUOHAN ZONG, University of Illinois Urbana-Champaign, Urbana, IL, United States

LANYU SHANG, University of Illinois Urbana-Champaign, Urbana, IL, United States

ZHENRUI YUE, University of Illinois Urbana-Champaign, Urbana, IL, United States

HUIMIN ZENG, University of Illinois Urbana-Champaign, Urbana, IL, United States

YIFAN LIU, University of Illinois Urbana-Champaign, Urbana, IL, United States

View all

Published: 27 June 2024

Citation in BibTeX format

CI '24: Collective Intelligence Conference
June 27 - 28, 2024
MA, Boston, USA

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3643562.3672613
ISBN: 9798400705540

Tripartite Intelligence: Synergizing Deep Neural Network, Large Language Model, and Human Intelligence for Public Health Misinformation Detection

Yang Zhang
University of Illinois Urbana-Champaign
Champaign, IL, USA
yzhangnd@illinois.edu

Ruohan Zong
University of Illinois Urbana-Champaign
Champaign, IL, USA
rzong2@illinois.edu

Lanyu Shang
University of Illinois Urbana-Champaign
Champaign, IL, USA
lshang3@illinois.edu

Zhenrui Yue
University of Illinois Urbana-Champaign
Champaign, IL, USA
zhenrui3@illinois.edu

Huimin Zeng
University of Illinois Urbana-Champaign
Champaign, IL, USA
huiminz3@illinois.edu

Yifan Liu
University of Illinois Urbana-Champaign
Champaign, IL, USA
yifan40@illinois.edu

Dong Wang
University of Illinois Urbana-Champaign
Champaign, IL, USA
dwang24@illinois.edu

ABSTRACT

The threat of rapidly spreading health misinformation through social media during crises like COVID-19 emphasizes the importance of addressing both clear falsehoods and complex misinformation, including conspiracy theories and subtle distortions. This paper designs a novel tripartite collective intelligence approach that integrates deep neural networks (DNNs), large language models (LLMs), and crowdsourced human intelligence (HI) to collaboratively detect complex forms of public health misinformation on social media. Our design is inspired by the collaborative strengths of DNNs, LLMs, and HI, which complement each other. We observe that DNNs efficiently handle large datasets for initial misinformation screening but struggle with complex content and rely on high-quality training data. LLMs enhance misinformation detection with improved language understanding but may sometimes provide eloquent yet factually incorrect explanations, risking misinformation mislabeling. HI provides critical thinking and ethical judgment superior to DNNs and LLMs but is slower and more costly in misinformation detection. In particular, we develop *TriIntel*, a tripartite collaborative intelligence framework that leverages the collective intelligence of DNNs, LLMs, and HI to tackle the public health information detection problem under a novel few-shot and uncertainty-aware maximum likelihood estimation framework. Evaluation results on a real-world public health misinformation detection application related to COVID-19 show that TriIntel outperforms representative DNNs, LLMs, and human-AI collaboration baselines in accurately detecting public health misinformation under a diverse set of evaluation scenarios.

CI '24, June 27–28, 2024, Boston, MA, USA

ACM ISBN 979-8-4007-0554-0/24/06
https://doi.org/10.1145/3643562.3672613

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

Human-AI Collaboration, Collective Intelligence, Large Language Model, Misinformation

ACM Reference Format:
Yang Zhang, Ruohan Zong, Lanyu Shang, Zhenrui Yue, Huimin Zeng, Yifan Liu, and Dong Wang. 2024. Tripartite Intelligence: Synergizing Deep Neural Network, Large Language Model, and Human Intelligence for Public Health Misinformation Detection. In *Collective Intelligence Conference (CI '24), June 27–28, 2024, Boston, MA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3643562.3672613

1 INTRODUCTION

In the digital age, the rapid spread of public health misinformation, especially through social media, poses significant risks to global health and societal stability [18]. The COVID-19 pandemic exacerbated this issue, where prevalent misinformation related to COVID-19 (e.g., unfounded virus origin theories, inaccurate prevention methods) has led to widespread confusion, harmful practices, and hindered public health efforts [42]. The detection and mitigation of public health misinformation, particularly in the context of the COVID-19 pandemic, have underscored the importance of addressing not just clear-cut factual inaccuracies [54] but also more nuanced and complex forms of misinformation such as conspiracy

theories, sarcasm, and subtly misleading content that go beyond simple true-or-false dichotomies [9]. By incorporating sophisticated misinformation types such as sarcasm and conspiracy theories into detection tasks, we can deepen our understanding of misinformation's complexities, allowing public health campaigns not only to correct inaccurate public health information but also directly address the roots of skepticism and fear on scientific evidence and health guidelines [22]. In this paper, we design a novel tripartite collective intelligence approach that integrates deep neural networks (DNNs), large language models (LLMs), and crowdsourced human intelligence (HI) to collaboratively detect complex forms of public health misinformation on social media.

Previous works have shown that DNNs, LLMs, and HI can play different yet complementary roles in public health misinformation detection [4]. Illustrative examples of the complementary intelligence are shown in Figure 1. In particular, DNNs are adept at swiftly processing a vast amount of datasets, making them ideal for initial misinformation screening [43]. However, they often lack context understanding of more complex, subtly misleading content and depend heavily on the quality of their training data, leading to possible misclassifications [56]. For example, a sarcastic post exclaiming, "Absolutely brilliant idea! Let's all forego mask-wearing and embrace one another at a large gathering to celebrate our liberation from the virus", which DNNs erroneously identify as misinformation. LLMs, like GPT-4, LLaMA, and Vicuna, offer a more refined understanding of language and context, capable of detecting misinformation that might be missed by DNNs and providing comprehensible explanations for their decisions on misinformation detection [58]. Yet, they too have limitations, such as the risk of generating eloquent explanations but factually incorrect answers to the prejudiced misinformation labels [34]. For example, LLMs may struggle to differentiate between conspiracy theories and sarcasm in nuanced expressions of questioning Bill Gates on social media (e.g., "Why wasn't @Billgates, one of the biggest advocates for vaccine and public health, a covid-19 vaccine test subject?"). HI, on the other hand, brings critical thinking, deep contextual and cultural insights, and ethical judgment, capabilities still beyond the reach of current AI models [64]. Humans can evaluate the credibility of information, considering its source, context, and potential impact [25]. However, utilizing human intelligence for misinformation detection is slower and more costly compared to DNNs and LLMs [63]. For example, in Figure 1, humans can clearly identify the tweet as a conspiracy that is missed by both DNNs and LLMs. Therefore, an effective strategy against public health misinformation involves integrating DNNs' speed and data-processing capabilities with LLMs' refined nature language understanding and HI's critical oversight and contextual judgment. However, developing such a tripartite collaborative intelligence system faces two key challenges.

The first challenge is how to effectively identify inaccurate misinformation detection by DNNs without prior knowledge of the correct misinformation labels. In particular, we can use DNNs to rapidly analyze the content of social media posts and initially estimate their misinformation labels, but DNNs can also make certain errors in their estimations [18]. Previous works have aimed to tackle this challenge by prioritizing social media posts that exhibit complex textual characteristics (e.g., intricate syntax and grammar and sophisticated contents) for expert review, based on the assumption

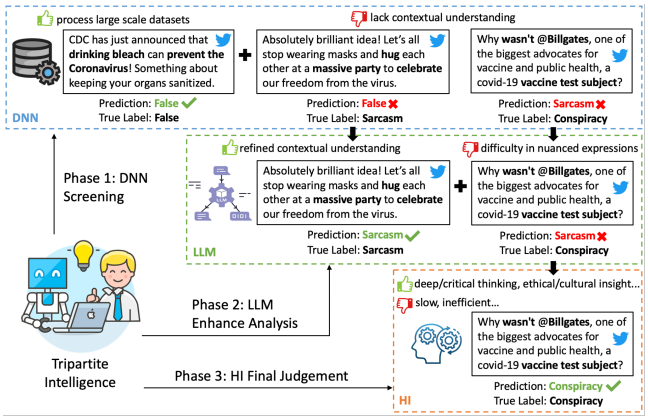


Figure 1: Illustrative examples of the complementary tripartite intelligence.

that DNNs are more prone to errors when analyzing texts with intricate details [2]. Nonetheless, this assumption is not always valid, as DNNs can also falter with seemingly straightforward texts if they contain subtle misinformation cues or sophisticated rhetorical strategies [1]. For example, during the COVID-19 pandemic, DNNs struggled to distinguish between legitimate news articles and those spreading conspiracy theories about vaccine microchipping, primarily because both utilized similar factual tones and authoritative references. Recent advances in uncertainty-aware DNN methods (e.g., uncertainty quantification or active learning) have been explored to identify DNN detection failures by leveraging a group of diverse DNN models or multiple instances of the same model to pinpoint discrepancies through the consensus of group members' outputs [40]. Yet, these methods might not be effective if all models in the group concurrently err in their analysis of the same social media posts. This is because all models may share the same inherent biases or lack sufficient contextual understanding of non-trivial misinformation, such as scams and conspiracies [13]. Therefore, it remains a challenging question on how to effectively detect the failure cases of DNNs in the absence of ground truth misinformation labels.

The second challenge lies in how to minimize the inaccurate predictions from LLMs, which may exhibit high uncertainty in their misinformation predictions. This is intended to effectively address failures in DNNs and reduce the need for human intervention. Unlike DNNs, which offer more direct misinformation classification, LLMs can produce uncertain and ambiguous responses [44]. For instance, a tweet that ambiguously discusses COVID-19 vaccines might be interpreted by an LLM in a way that leaves room for confusion, with the response states that "This post could be either conspiracy or sarcasm", reflecting the model's uncertainty. This is because of LLMs' reliance on pattern recognition without the ability to verify factual accuracy or the latest scientific findings in real time [50]. Moreover, as generative models, LLMs may provide detailed yet vague explanations instead of a direct prediction of the concrete category of misinformation [26]. For example, an LLM can respond, "The post suggests doubts about COVID-19 vaccines' effectiveness in a vague way," which shows the LLM's uncertainty

regarding the misinformation labels given the ambiguous language in the LLM's explanation. To tackle these challenges, emerging strategies like Contextual Calibration and Reinforcement Learning (RL) have been developed [32, 62]. Contextual Calibration aims to enhance the precision of LLM predictions by adjusting confidence levels based on historical data, whereas RL focuses on improving model responses through feedback learning. However, these methods face challenges, including difficulty in adapting to recent factual knowledge due to the dynamic nature of misinformation and the potential reinforcement of biases from the training data [11]. Therefore, it remains to be a key challenge on how to derive the accurate misinformation labels by exploring the insightful yet potentially uncertain intelligence from LLMs to address the failure cases of DNNs.

To address the aforementioned challenges, this paper develops *TriIntel*, a tripartite collaborative intelligence framework that leverages the collective intelligence of DNNs, LLMs, and HI to tackle the public health information detection problem. We first develop a deep learning based misinformation detection approach to efficiently identify potentially misclassified social media posts by DNNs through a principled few-shot learning network optimization design. We then forward the identified posts to LLMs for further misinformation analysis. Here, we design a transformer-based conditional probabilistic learning approach to interpret the uncertainty of misinformation labels and explanations from LLMs, thereby obtaining accurate labels for failure cases from DNN. Finally, we develop a crowdsourced human intelligence acquisition model and a principled maximum likelihood estimation approach to derive high-quality human intelligence for accurately detecting misinformation in posts that both DNNs and LLMs were unable to address. To the best of our knowledge, TriIntel is the first *tripartite* collective intelligence framework designed to address the complex context-based text classification problem from social media like public health misinformation detection. We also envision that TriIntel can be extended to a wide range of real-world applications beyond misinformation detection (e.g., disaster damage assessment, healthcare and medical diagnosis, environmental monitoring). We evaluate TriIntel in a real-world public health misinformation detection application related to COVID-19. The evaluation results demonstrate that TriIntel consistently outperforms state-of-the-art DNNs, LLMs, and human-AI collaboration baselines, significantly improving public health misinformation detection accuracy under a diverse set of evaluation scenarios. This underscores TriIntel's effectiveness in integrating diversified intelligence from DNNs, LLMs, and HI to address the complex problem of public health misinformation detection. We summarize our main contributions as follows:

- We address the novel problem of detecting public health misinformation by integrating the comprehensive and distinct capabilities of DNNs, LLMs, and HI.
- We design an innovative tripartite collective intelligence framework, TriIntel, that effectively combines the initial data-processing power of DNNs, the refined language understanding of LLMs, and the critical oversight and contextual judgment of HI to improve misinformation detection.
- TriIntel is the first tripartite collaborative intelligence framework specifically designed to tackle complex social media context-based text classification problems, such as public health misinformation detection.
- We conduct extensive evaluations of TriIntel on real-world public health misinformation detection applications related to COVID-19. The results demonstrate that TriIntel consistently outperforms state-of-the-art DNNs, LLMs, and human-AI collaboration baselines in accurately detecting public health misinformation across various scenarios.

2 RELATED WORK

2.1 Collective Intelligence

Collective intelligence refers to the concept of leveraging the knowledge, skills, and contributions of a group of individuals, AI models, or computer systems to solve complex problems, make decisions, or achieve specific goals [23, 49]. It involves combining the capabilities of multiple entities, such as humans, AI, and machines, to create a collective system that is more intelligent, efficient, and effective than any individual component [57]. Collective intelligence has been applied in various domains, such as crisis informatics [47], medical research [37], knowledge management [19], decision-making [15], and crowdsourcing [17]. For example, Uchino *et al.* introduced a deep learning and nephrologist–AI collective intelligence model to classify glomerular pathological findings to achieve efficient and objective diagnosis in renal pathology [52]. Hao *et al.* designed a multimodal neural network-based disaster damage assessment framework that utilizes crowdsourcing data from large crowds to classify disaster damage types in hurricane events [16]. Gregg *et al.* developed a web-based student performance-sharing platform for special education, aiming to facilitate collaboration among educators, parents, and students to enhance individualized education plans and improve learning outcomes for students with special needs [14]. However, it remains a critical challenge to leverage the collective intelligence of DNNs, LLMs, and HI to solve the public misinformation detection problem while considering the unique strengths of each of them. In this paper, we develop a novel tripartite collective co-learning framework to accurately detect public misinformation by exploring the collective tripartite intelligence.

2.2 Public Health Misinformation

Public health misinformation has emerged as a widespread issue in the world, posing significant threats to individual and community well-being [48]. AI techniques have been increasingly applied to detect and combat the spread of false or misleading public health information [30]. For example, Upadhyay *et al.* developed a hybrid representation learning approach that jointly extracted content and context information for health misinformation detection on web pages [53]. Cui *et al.* proposed a meta-path learning strategy that designs a graph neural network to incorporate social interaction for online health misinformation detection [7]. More recently, with the advancement of LLMs, there has been a significant shift towards leveraging these models for misinformation detection [4]. LLMs, such as GPT-3.5 and GPT-4, have shown remarkable capabilities in understanding context and generating human-like text, which

makes them particularly useful for detecting nuanced misinformation [59]. For instance, Leite *et al.* designed a weakly supervised method that incorporated credibility estimation from LLMs to detect misinformation[24]. In another example, Pendyala *et al.* [36] demonstrated the effectiveness of fine-tuning LLMs on misinformation datasets to enhance detection accuracy. However, previous approaches often only focus on leveraging DNNs or LLMs for public health misinformation separately, but they might encounter undesirable performance limitations given the complexity of context in social media and the complex forms of misinformation. Human-AI collaboration has also seen advancements in misinformation detection. For example, Sharma *et al.* [45] developed a framework that integrates human feedback into the training process of LLMs, improving their ability to detect misinformation by learning from human expertise. Another notable work by Mcgrath *et al.* [28] involved a collaborative approach where human auditors work alongside LLMs to identify and correct misinformation, thus leveraging the strengths of both human intuition and machine learning. These approaches highlight the growing trend of combining human intelligence with advanced AI models to tackle the complex problem of misinformation. Human-AI collaboration, while beneficial in integrating human judgment and AI efficiency, also faces critical challenges. These collaborations can be slow and resource-intensive, relying heavily on the availability and expertise of human annotators. The process of integrating human feedback into AI systems is often complex and can introduce new biases if not managed carefully. Additionally, ensuring consistent quality and scalability in human-AI collaboration can be difficult, especially when dealing with large volumes of data and the need for rapid responses. Our approach, TriIntel, addresses these limitations by integrating the strengths of DNNs, LLMs, and HI into a cohesive framework. Unlike previous methods that rely solely on one type of intelligence, TriIntel leverages DNNs for efficient initial processing, LLMs for refined contextual understanding, and HI for critical oversight and ethical judgment. This tripartite collaboration ensures a more balanced and comprehensive approach to misinformation detection, mitigating the individual weaknesses of each component while enhancing their collective strengths.

2.3 Human-AI Collaboration

With the recent advancement of AI, human-AI collaboration has become an effective method that aims to jointly utilize the complementary strengths of human intelligence and AI to address complex real-world problems [12, 21, 46, 61]. For example, Reverberi *et al.* developed a Bayesian-based collaborative framework that integrates human judgments and AI predictions to improve the accuracy of medical decision-making [41]. Fan *et al.* proposed a Human-AI collaborative system to enhance the explanations and synchronization in AI-assisted user experience (UX) evaluation [10]. More recently, as generative AI become more powerful and versatile, researchers start to explore the opportunities of involving generative AI in human-AI collaboration [12]. For example, Rastogi *et al.* designed a human-AI collaborative approach to jointly incorporate human intelligence and generative models to collaboratively audit LLMs [39]. To the best of our knowledge, our paper is the first *tripartite* collective intelligence framework designed to jointly leverage three forms of intelligence, i.e., DNNs, LLMs, and HI, to collectively address the complex social media context-based text classification problem, such as public health misinformation detection.

3 PROBLEM FORMULATION

In this section, we formally define our public health misinformation detection problem with tripartite collective intelligence. We first introduce the input data, **health-related social media posts ($\boldsymbol{P}$)**, to detect online public health misinformation in our problem. The health-related social media posts $\boldsymbol{P} = \{P_1, P_2, ..., P_N\}$ are defined as a set of posts collected from online social media platforms that are related to the public health domain (e.g., COVID-19). N represents the total number of social media posts in the studied application, where P_n is the n^{th} post in $\boldsymbol{P}$.

In online misinformation detection, the social media posts $\boldsymbol{P}$ can be categorized into different categories based on the characteristic of each post, where we refer to such categories as **misinformation classes ($\boldsymbol{C}$)** (e.g., true information, false information, conspiracy, sarcasm). The objective of our problem is to identify the **misinformation label ($\boldsymbol{L}$)** for each of the health-related social media posts $\boldsymbol{P}$. The misinformation labels $\boldsymbol{L} = \{L_1, L_2, ..., L_N\}$ are defined as a set of the misinformation class categories that each social media post belongs to, where L_n refers to the misinformation label for the n^{th} social media post P_n.

Determining the truthfulness of public health information is inherently more straightforward than discerning the presence of more complex forms of misinformation, such as conspiracy theories or sarcasm. This distinction arises because the assessment of true or false claims can be based on tangible evidence and scientific research. In contrast, more sophisticated misinformation classes, such as conspiracy theories and sarcasm, are imbued with subtleties and subjectivity, often reliant on the context and underlying meanings. These elements make them significantly more challenging to classify, necessitating an in-depth comprehension of the intent and the social intricacies involved – a stark contrast to the direct method of comparing statements against recognized health documents and scientific findings for truthfulness verification.

Motivated by these challenges, our paper focuses on a scenario where only annotations for the two predominant categories, true and false information, are available for social media posts in the training dataset. We observe that obtaining training data for those categories involves merely verifying individual claims against credible sources, while labeling data from conspiracy theories and sarcasm requires untangling complex narratives, emotional manipulation, and social dynamics [33]. Therefore, our goal is to investigate whether the patterns of misinformation identified within these available categories can extend to new, typically inaccessible categories that are inherently more complex and difficult to pinpoint. This exploration aims to enhance our understanding and detection of misinformation beyond the straightforward true and false separation, encompassing the broader spectrum of misinformation phenomena. Specifically, we delineate **known misinformation classes ($\boldsymbol{C^K}$)** as those categories present in the training dataset for health-related social media posts, namely true and false information, which are more straightforward and commonly used in

misinformation detection. For example, a claim stating "The vaccine has been proven safe in clinical trials" belongs to the known class of true, while "The vaccine contains microchips" belongs to the known class of false. Conversely, **unknown misinformation classes (C^U)** refer to categories not present in the training dataset, such as conspiracy and sarcasm. These unknown classes are typically inaccessible in real-world application scenarios and require more nuanced understanding due to their inherent complexity and the subtleties involved in their detection. An example post of the unknown conspiracy class could be "The vaccine is part of a global depopulation agenda," while a post of the unknown sarcasm class might be "Sure, because injecting unknown substances into our bodies is always a great idea." Mathematically, we define all misinformation classes C as a combination of both known and unknown classes, represented by: $C = C^U \bigcap C^K$.

The objective of our problem is to accurately identify the misinformation labels L for the studied social media posts from both known and unknown misinformation classes C, leveraging the training data from the known misinformation classes C^K. To achieve such an objective, we propose to take advantage of the complementary strengths of the DNNs, LLMs, and HI. We refer to the predictions on misinformation labels from the DNN, LLM, and HI as **DNN classification results ($\widehat{L^{DNN}}$)**, **LLM classification results ($\widehat{L^{LLM}}$)**, and **human classification results ($\widehat{L^{HI}}$)**, respectively. We define the overall collaborative predictions on misinformation labels L of our TriIntel framework generated from the DNN, LLM, and human classification results to be the **collaborative classification results ($\widehat{L}$)**. We formally define the objective of our tripartite collective public health misinformation detection problem as follows:

$$\arg\max_{\widehat{L}_n}(\Pr(\widehat{L}_n = L_n \mid \boldsymbol{P}, C^K)), \forall 1 \le n \le N \quad (1)$$

where $\widehat{L}_n$ represents the collaborative classification results for the n^{th} social media post P_n.

4 SOLUTION

Our TriIntel is a tripartite collaborative framework that collectively integrates the diversified yet complementary intelligence from DNNs, LLMs, and HI to solve the public health misinformation detection problem. The overview of the TriIntel framework is presented in Figure 2. The TriIntel framework consists of three key modules:

- *Few-Shot-based Unknown Class Discovery (FUCD)*: The FUCD module develops a DNN model based on few-shot learning scheme to identify whether social media posts are from known or unknown misinformation classes. The module then predicts the labels of social media posts if they are from known misinformation classes. Furthermore, it forwards the identified posts from unknown misinformation classes to the LLM to determine the concrete misinformation labels.
- *LLM-based Misinformation Detection and Interpretation (LMDI)*: The LMDI module designs an LLM-based misinformation detection model to effectively provide misinformation labels and associated explanations for social media posts identified by FUCD. We also develop a transformer-based conditional probabilistic learning approach to interpret the uncertainty of the misinformation labels and explanations from LLMs, identifying cases where LLMs might exhibit high uncertainty and necessitate human intervention.
- *Tripartite Collaborative Intelligence Fusion (TCIF)*: The TCIF module develops an LLM-assisted, crowdsourced human intelligence acquisition model and a principled estimation approach to obtain high-quality human intelligence for accurately detecting misinformation in posts that exhibit high uncertainty in LLMs' misinformation predictions.

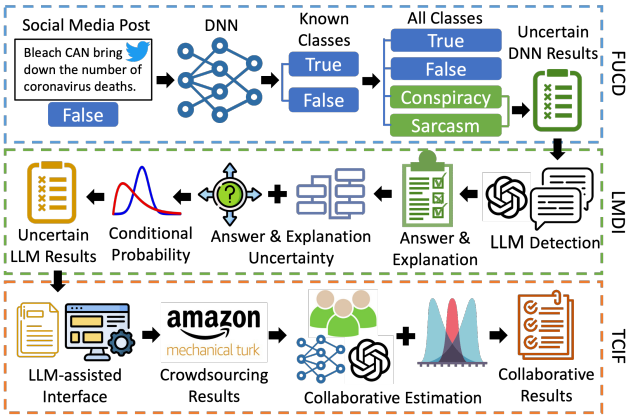


Figure 2: Overview of the TriIntel Framework

4.1 Few-Shot-based Unknown Class Discovery

The FUCD module focuses on developing a few-shot learning scheme that 1) identifies whether a social media post is from known or unknown classes; 2) predicts the misinformation labels of the social media posts from known classes by leveraging the available training data; and 3) forwards the identified posts to an LLM for further estimation of the misinformation labels if the social media posts are identified as belonging to unknown classes.

In our FUCD module, we first define a DNN that is capable of extracting contextual features related to public health misinformation from the input social media posts. To select the DNN model, we analyze and compare the strengths and weaknesses of multiple state-of-the-art models. BERT [8] provides deep bidirectional representations; however, its pre-training includes the next sentence prediction task, which has been found somewhat ineffective for nuanced context understanding needed in misinformation detection. XLNet [60] excels at learning diverse text representations by using permutation-based training, which enables a better understanding of sentence structure and context; however, its complexity can lead to difficulties in tuning for specific tasks like ours. GPT-2 [38] generates coherent and contextually rich text outputs, making it excellent for generative tasks; but it does not perform as well in discerning the veracity of information due to its unidirectional nature. In contrast, RoBERTa [27] stands out for its robust training on a larger corpus and longer duration, and the removal of the next sentence prediction task, which enhances its ability to understand and process the complex language used in social media posts. Therefore,

we utilize the transformer-based RoBERTa network as the DNN in our TriIntel framework. Such a network provides efficient contextual feature extraction capabilities for desirable misinformation detection. In particular, we have:

$$\widehat{L}_n^{DNN} = DNN(P_n), \forall 1 \leq n \leq N \tag{2}$$

where P_n is the representation of input social media posts.

The DNN can accurately classify misinformation from known classes they have been trained on but it struggles to label misinformation from unknown classes correctly due to the lack of training data. To overcome such a challenge, our FUCD module introduces a few-shot network optimization mechanism to supervise DNN that jointly identifies whether a social media post is from a known or unknown class and predicts the misinformation label of the post if it is identified to be from known class. In particular, the few-shot network optimization function is defined as follows:

$$\begin{aligned} &\sum_{L_n \in \boldsymbol{C}^{\boldsymbol{K}}} \mathcal{L}_{\mathcal{CE}}(L_n, DNN(P_n | L_n \in \boldsymbol{C}^{\boldsymbol{K}})) \\ &+ \sum_{L_n \in \boldsymbol{C}^{\boldsymbol{U}}} \mathcal{L}_{\text{Few-shot}}(DNN(P_n | L_n \in \boldsymbol{C}^{\boldsymbol{U}})) \end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathcal{CE}}(\cdot)$ indicates the cross-entropy loss function, which is used to supervise our DNN to accurately estimate the misinformation labels of social media posts from the known class by minimizing the difference between the misinformation label estimated by DNN and the actual misinformation label in the training dataset. $\mathcal{L}_{\text{Few-shot}}(\cdot)$ is the similarity-based few-shot loss function [5]. It supervises the DNN to distinguish social media posts of known classes from those of unknown classes. It does so by maximizing the intraclass similarity and interclass dissimilarity in terms of the misinformation-related contextual features. After our DNN is optimized by the above loss function and makes predictions on the testing data, we will forward the social media posts to LLM for further estimation of the misinformation label if the post is predicted to come from the unknown classes. We will keep the misinformation labels provided by DNN as the final output for our TrilIntel if the predicted labels by DNN are from one of the known classes. Note that the DNN is sufficient to provide misinformation labels for the known classes because the training data for these classes is adequate for the DNN to learn all relevant contextual features, ensuring accurate detection of misinformation within the known classes [43].

4.2 LLM-based Misinformation Detection and Interpretation

In this section, we design an LLM-based misinformation detection and interpretation module that focuses on two tasks: 1) designing an LLM-based misinformation detection model to effectively provide misinformation labels and associated explanations, and 2) developing a transformer-based conditional probabilistic learning approach to interpret the uncertainty of the misinformation labels and explanations generated by LLMs. This approach identifies cases where LLMs might exhibit high uncertainty, necessitating human intervention.

To select the appropriate LLM for the first task, we assessed several leading LLMs and their relative strengths and weaknesses. GPT-3.5 offers impressive language understanding and generation capabilities but struggles with consistency in longer contexts and nuanced contextual understanding, which can be crucial when handling misinformation in complicated social media texts. LLaMA [51] excels at understanding and generating responses based on a leaner model architecture, providing efficiency; however, its smaller training corpus compared to others may limit its effectiveness in nuanced misinformation contexts. Another typical LLM, Claude[1], demonstrates adaptability and user-friendly interaction capabilities, but it may not consistently handle the depth and complexity of misinformation due to its design primarily for general consumer applications rather than specialized analytical tasks. In contrast, GPT-4 stands out due to its advanced capabilities in handling complex and nuanced language tasks. GPT-4 is trained on a broader range of Internet text and structured data, enabling it to generate more accurate and contextually appropriate responses. Furthermore, GPT-4 exhibits improved performance in terms of reasoning and coherence over extended interactions, which is vital for analyzing and interpreting the intricate narratives found in public health misinformation. Therefore, we incorporate the advanced LLM model, GPT-4, to provide misinformation labels and associated explanations for the studied social media posts.

To adapt GPT-4 for our specific need of detecting healthcare misinformation, we adopt a prompting strategy that involves feeding LLM with precise definitions of complex misinformation categories, such as conspiracy theories and sarcastic comments. To ensure the precision of misinformation class definitions, we prompt LLM with the definitions provided in the annotation codebook [29] that are designed to guide ground truth misinformation label annotation. For example, a tweet is considered sarcastic if it humorously or exaggeratedly mocks false information, especially about cures, prevention, or conspiracies, to critique ignorance or wrongdoing in current affairs. The tweet "Do you want to earn more from a commodity? Just link it to a coronavirus cure, and its value will skyrocket" uses exaggeration, fitting the criteria of presenting false information through humor or ridicule. We present the LLM prompting design based on misinformation class definitions and examples in Figure 3. Such a precise calibration of LLM significantly improves the precision of misinformation detection, especially in cases involving complex or nuanced expressions.

While LLMs provide sufficient capacity to classify unknown classes after our prompting strategy, as generative models, they could still offer detailed yet vague explanations when tasked for misinformation detection. Therefore, our second task focuses on interpreting the uncertainty of the misinformation labels and explanations provided by LLMs. This approach aims to identify cases where LLMs exhibit high uncertainty in the misinformation labels, necessitating human intervention. Consequently, we develop an uncertainty-based approach to identify situations where LLMs are unable to provide definitive misinformation classification results.

In particular, the task of measuring **LLM uncertainty (ϵ)** of text-based LLM classification results is challenging and multifaceted,

[1] https://www.anthropic.com/claude

A tweet shall be classified as a conspiracy if it *endorses a conspiracy story*. Some examples of conspiracy themes related to COVID-19 include:
1. It is a *bioweapon*.
2. Electromagnetic fields and the introduction of *5G wireless technologies* led to COVID-19 outbreaks.
3. This was a plan from *Gates Foundation* to increase the Gates' wealth.
4. It leaked from the *Wuhan Labs* or Wuhan Institute of Virology in China.
5. It was *predicted* by Dean Koontz.

A tweet shall be classified as sarcasm if any of the following conditions are met:
1. The tweet contains *clear signs of a satire* calling out a fake cure, a fake prevention or a conspiracy.
2. The tweet includes a clear *joke* about a fake cure, a fake prevention or a conspiracy.
Concretely, this is a tweet where the information in the post is false but is presented using humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, particularly in the context of contemporary politics and other topical issues. This kind of post is used to ridicule other false statements or people.

Based on the definitions provided, please categorize each of the following tweets into one of the four categories: true information, false information, conspiracy, or sarcasm/satire. Additionally, include your justifications for each classification alongside your answers.

Figure 3: LLM prompting design based on misinformation class definitions and examples.

encompassing two dimensions: **answer uncertainty (ϵ^A)** and **explanation uncertainty (ϵ^E)**. Specifically, the answer uncertainty refers to the level of ambiguity in GPT-4's answer of misinformation classes for each social media post. For instance, a response indicating multiple possible classes (e.g., "This post could be either conspiracy or sarcasm") exhibits a higher degree of uncertainty compared to a response that suggests a single class (e.g., "This post is a conspiracy"). Besides, the explanation uncertainty represents the degree of uncertainty in the language expressions used by LLM in its responses, which is particularly aimed at discerning subtle differences in textual responses. For example, if LLM explains, "The post suggests doubts about COVID-19 vaccines' effectiveness in a vague way", it shows a higher degree of uncertainty due to the ambiguous language used in the explanation, indicating a lack of clarity in identifying misinformation. Conversely, a more confident response, "The post spreads false claims about COVID-19 vaccines", demonstrates a clear understanding of the post's misleading intent.

Motivated by the above observations, we develop a novel probabilistic method to identify uncertain misinformation classification results from LLMs and measure its uncertainty. First, we define the answer uncertainty to be proportional to the number of possible misinformation classes indicated in the LLM's response. For example, we assign a higher answer uncertainty if LLM suggests posts may involve both conspiracy and sarcasm instead of solely conspiracy. Second, we generate the value of LLM explanation uncertainty by leveraging a transformer-based sentence-level language uncertainty measurement network [35], which quantifies the ambiguous words and contexts in LLM's response. Finally, we combine the explanation uncertainty with the answer uncertainty as the LLM uncertainty as follows:

$$\epsilon_m = \epsilon_m^E \cdot \epsilon_m^A, \ \forall 1 \leq m \leq M \tag{4}$$

where M is the total number of social media posts being classified by GPT-4. ϵ_m represents the LLM uncertainty for the m^{th} post studied. ϵ_m^A and ϵ_m^E denote the answer and explanation uncertainties for the same post, respectively. The above equation is designed to generate ϵ_m, taking into consideration of both answer and explanation uncertainties. Given the computed LLM uncertainty for each social media post, we identify the posts with an uncertainty score ϵ_m greater than an application-specific threshold θ. The actual value of θ depends on the trade-off between misinformation classification accuracy and the number of posts needing human review. The selected uncertain samples of the LLM will be further refined by crowdsourced human intelligence, which will be discussed in the next subsection.

4.3 Tripartite Collaborative Intelligence Fusion

In this section, we describe how we use human intelligence to accurately identify misinformation categories of social media posts which LLMs are uncertain about. We select posts that LLMs exhibit high uncertainty and task crowdsourcing platform workers to annotate their misinformation labels. The design of our crowdsourcing interface, shown in Figure 4, includes explanations generated by GPT-4 to help workers by providing relevant context and background information. However, we caution workers that these explanations might not always be accurate. Our approach focuses on using only the LLM-generated explanations, not their direct misinformation classifications, to avoid biasing workers towards the potentially incorrect LLM outputs and to also encourage independent critical thinking.

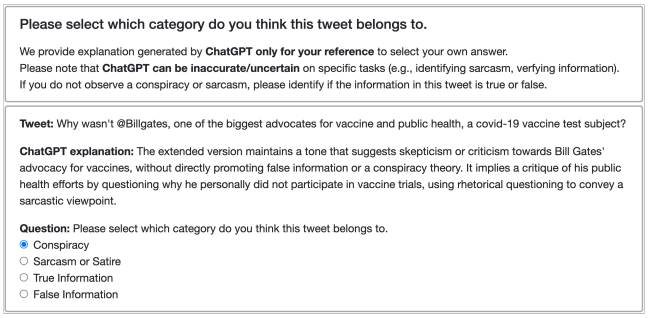


Figure 4: LLM-assisted Crowdsourcing Interface Design

We observe that it remains challenging to directly leverage crowdsourcing results to address LLM uncertainty (e.g., replacing uncertain LLM results with crowd labels) because the results collected from crowdsourcing platforms can be imperfect compared to domain expert annotations. To overcome this challenge, we effectively integrate the misinformation classification results from the DNN, LLM, and crowdsourced human workers to synergize their complementary advantages in misinformation detection. A straightforward method to integrate classification results from DNN, LLM, and crowdsourced human workers is majority voting. However, the misinformation classification accuracy of the DNN, LLM, and crowd workers can differ, especially considering the uncertainty observed in DNN and LLM classification results. In this case, we develop a principled maximum likelihood estimation (MLE) model to estimate the collaborative classification results $\widehat{\boldsymbol{L}}$ while evaluating the misinformation classification accuracy of each member in the integration. In particular, our likelihood function $\mathbb{L}(\Phi; \Delta, Z)$ is defined as:

$$\mathbb{L}(\Phi; \Delta, Z) = \mathbb{L}(\Phi; (\widehat{L^{DNN}}, \widehat{L^{LLM}}, \widehat{L^{HI}}), \widehat{L}) \tag{5}$$

where $\widehat{L^{DNN}}$, $\widehat{L^{LLM}}$, and $\widehat{L^{HI}}$ indicate the predicted misinformation labels from DNN, LLM, and HI, respectively. $\Delta = (\widehat{L^{DNN}}, \widehat{L^{LLM}}, \widehat{L^{HI}})$ is the observed variable of the MLE model. Φ is the estimated variable of the MLE model. Z is the latent variable of the model, which indicates the misinformation label $\widehat{L}$ for all studied social media posts. In particular, the formulated problem can be solved using the expectation maximization (EM) algorithm [55] to obtain the estimated misinformation label $\widehat{L}$, which serves as the final output of the misinformation labels that were identified by the LMDI module.

5 EVALUATION

5.1 Dataset and Crowdsourcing Settings

To evaluate our TriIntel framework, we utilize a publicly available social media dataset on public health misinformation detection [29]. This dataset also includes ground truth labels for misinformation classes, which have been manually annotated according to the comprehensive and detailed definitions and examples provided in the misinformation annotation codebook[2]. The gold standard labels were created through a rigorous and structured manual annotation process, guided by this codebook. Several key steps were taken to ensure annotation accuracy and minimize individual annotator bias. First, different misinformation categories were identified for classifying the tweets after extensive discussions and revisions, with definitions and examples detailed in the publicly available codebook. Tweets were randomly and uniformly sampled from the data collection to maintain diversity in terms of topics covered, ensuring a comprehensive representation of topics and misinformation types in the dataset. The first phase of annotation was conducted by an annotator, providing an initial broad categorization of the dataset. In the second phase, a subset of the annotated tweets were randomly assigned to six additional annotators to verify consistency and reliability. The consistency of the annotations was assessed through inter-annotator agreement, ensuring that the labels were reliable and accurately reflected the defined categories. Following Twitter's terms and conditions, we employ the tweet IDs provided in the dataset to retrieve the contents of each tweet that are still available on Twitter. After the aforementioned processing, the dataset used in our experiments comprises a total of 1,607 tweets. Our dataset encompasses four classes: true information (24.6%), false information (20.8%), conspiracy (32.9%), and sarcasm (21.7%). The ratio of training to testing data is set at 7:3 for the evaluation of all models.

We note that the aforementioned existing annotations in the misinformation dataset serve as a gold standard for *evaluation purposes only*, and such annotations do not exist in real-world deployment scenarios of misinformation detection frameworks. In practice, deploying these frameworks requires dynamic and ongoing assessment of content, which cannot rely on pre-existing annotations. In contrast, crowdsourcing is more efficient and readily available 24/7 compared to the option of relying on domain expert annotators, enabling us to gather a large volume of annotations quickly and cost-effectively. Therefore, this approach is particularly useful for addressing mistakes made by DNN and LLM on difficult cases. In our experiments, we leverage our crowdsourcing interface design (as shown in Figure 4) to collect human intelligence on misinformation labels for selected social media posts via Amazon Mechanical Turk (MTurk), a prominent crowdsourcing platform known for its expansive and cost-effective freelance workforce. To ensure high-quality labels from crowd workers, we set qualification standards requiring a worker to have completed over 10,000 approved tasks and to maintain an approval rating above 95% before engaging in our project. We engage five crowd workers for each classification task. Each participant receives a compensation of $0.05 per task. The inter-worker agreement between different crowd workers is 0.7310 in terms of the Kappa score, where a Kappa score above 0.6 indicates a good agreement between different individuals [6]. We adhere to the guidelines established by the Institutional Review Board (IRB) protocol for this project.

[2]https://zenodo.org/records/4024154

5.2 Baselines and Experimental Settings

To comprehensively evaluate our TriIntel, we include a rich set of DNN, LLM, and human-AI baseline methods in our evaluation:

DNN Baselines:

- **BERT** [8]: a representation learning model that harnesses a bidirectional transformer architecture to effectively process words in both directions, thereby achieving high classification accuracy in complex language tasks.
- **RoBERTa** [27]: a robust deep learning model utilizing an advanced transformer network to intricately learn and extract deep semantic representations from textual data, which enhances performance on a wide range of classification tasks through extensive training and advanced fine-tuning mechanisms.
- **CloserLook** [5]: A representative few-shot text-based classification framework that combines deep augmentation techniques with cosine similarity-based learning to achieve an efficient deep classification network for limited training data.

LLM Baselines:

- **GPT-4** [31]: a highly sophisticated transformer-based architecture that generates remarkably coherent and contextually relevant text by learning from a vast corpus of diverse data, which sets new benchmarks in natural language understanding tasks.
- **LLaMA** [51]: an innovative chat-based framework that incorporates both foundational language models and fine-tuned conversational models to enhance its interactive capabilities and contextual understanding.

Human-AI Baselines:

- **LL++** [46]: a human-AI hybrid method that includes a versatile and task-agnostic loss function to effectively aggregate AI and human intelligence to achieve precise classification results.
- **StreamCollab** [61]: a human-AI collaborative framework that designs uncertainty estimation and fusion of AI and human intelligence to ensure desirable collective classification performance.

- **MEGAnno+** [21]: A human-LLM collaborative learning approach designed to integrate human input with LLM reasoning for reliable and efficient data labeling, ensuring reliable collaboration and classification results.

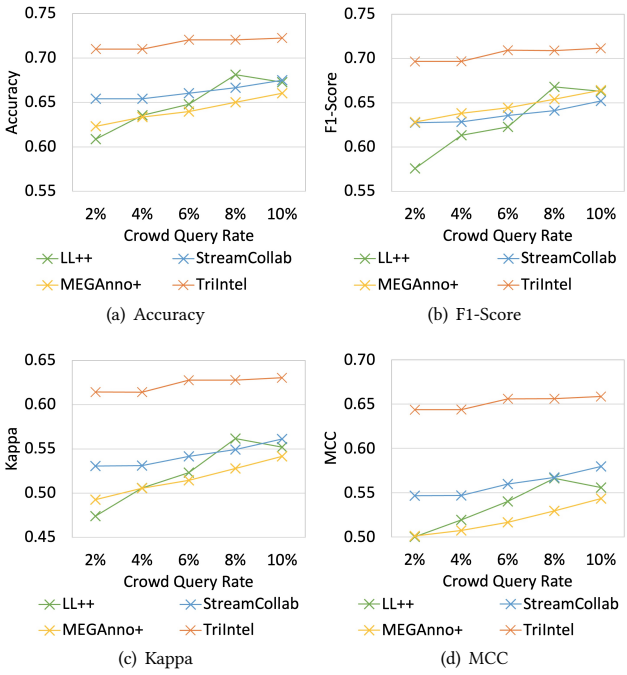
To ensure a fair comparison across all compared approaches, we set the input data for all approaches to be the same. This includes: 1) all collected tweets, 2) ground truth misinformation class labels for tweets in the training set, and 3) annotated tweets provided by crowd workers. In particular, we retrain DNN baseline models using the crowdsourced annotations. For human-AI baselines, we follow their schemes to query the same amount of crowd labels and integrate the collected inputs with AI models for misinformation detection tasks. Additionally, we incorporated a random baseline as a reference, which classifies each tweet's misinformation label by randomly selecting from all candidate categories. We implemented our TriIntel and baselines in PyTorch and trained these models on NVIDIA RTX 6000 GPUs. We optimized all hyperparameters using the Adam optimizer. Additionally, the learning rate was set to 1×10^{-5}, with a batch size of 100, over a training period of 100 epochs.

To evaluate the performance of all compared approaches, we employ four widely recognized metrics for multi-class textual classification: 1) Accuracy, 2) F1-score, 3) Cohen's Kappa Score (Kappa) [3], and 4) Matthews Correlation Coefficient (MCC) [20]. Higher values in Accuracy, F1-Score, Kappa, and MCC are indicative of better performance in misinformation classification.

5.3 Experiment Results

5.3.1 Performance Comparison on Misinformation Classification Accuracy. In this subsection, we compare the performance of our TriIntel with all baselines in terms of misinformation classification accuracy. The evaluation results are presented in Table 1. These results indicate that TriIntel achieves clear performance gains compared to all baselines from different categories across various metrics. For example, when compared to the best-performing baseline, StreamCollab, TriIntel exhibits clear performance improvements in terms of Accuracy, F1-Score, K-Score, and MCC, with increases of 9.1%, 11.6%, 15.9%, and 17.2%, respectively. Such performance gains can be attributed to our novel tripartite collective intelligence framework design. Our framework effectively integrates the data-processing abilities of DNNs, the refined natural language understanding of LLMs, and the critical oversight and contextual judgment provided by HI in a holistic solution. This integration ensures desirable misinformation classification accuracy under our principled few-shot and uncertainty-aware maximum likelihood estimation framework design.

5.3.2 Robustness Study of TriIntel on Crowd Inputs. Subsequently, we conduct a robustness study to examine TriIntel's performance when the percentage of social media posts selected for crowd labeling (which we refer to as the *crowd query rate*) varies. In this experiment, we vary the crowd query rate from 2% to 10%. Additionally, we compare TriIntel with all human-AI baselines (LL++, StreamCollab, and MEGAnno+) that are designed to jointly utilize the collective intelligence of humans and AI. The evaluation results are presented in Figure 5. We observe that TriIntel maintains stable performance and achieves consistent performance gains over these baseline models across different crowd query rates. This consistency validates TriIntel's robustness, attributed to its effective integration of predictions from DNN, LLM, and HI.



(a) Accuracy (b) F1-Score (c) Kappa (d) MCC

Figure 5: Robustness Study of TriIntel

5.3.3 Ablation Study of TriIntel. Finally, we perform an ablation study to evaluate the contributions of the three key modules of our TriIntel framework (i.e., FUCD, LMDI, and TCIF) to the overall misinformation classification accuracy. We present the performance evaluation results obtained by eliminating each of these modules individually. Specifically, we replace the FUCD module by uniformly sampling the same number of social media posts for further analysis by LLM and humans. The LMDI module is replaced by substituting the misinformation detection outputs generated by the LLM with those from DNNs. Furthermore, we exclude the TCIF module by skipping the misinformation labels provided by human intelligence. The evaluation results are shown in Figure 6. We observe a clear decrease in performance across all evaluation metrics when any of the FUCD, LMDI, or TCIF modules is removed. The results clearly demonstrate that all three modules make critical contributions to the TriIntel framework in terms of misinformation classification accuracy.

5.3.4 In-depth Analysis of TriIntel's Performance . TriIntel's evaluation highlights several key strengths that establish its efficacy in public health misinformation detection. The framework's integration of DNNs, LLMs, and HI leverages the unique advantages of each component, resulting in superior performance across all evaluated metrics. As shown in Table 1, TriIntel outperforms baseline models in terms of Accuracy, F1-Score, Kappa, and MCC, indicating its robust capability in handling diverse misinformation types. The

Table 1: Performance Comparisons on Misinformation Classification Accuracy

Category	Algorithm	Accuracy	F1-Score	Kappa	MCC
Random	Random	0.2547	0.2589	0.0077	0.0078
DNN	BERT	0.6211	0.6049	0.4872	0.4934
	RoBERTa	0.6460	0.6200	0.5200	0.5358
	CloserLook	0.5797	0.5513	0.4358	0.4490
LLM	GPT-4	0.6190	0.6239	0.4871	0.4896
	LLaMA	0.5528	0.5610	0.3990	0.4174
Human-AI	LL++	0.6480	0.6226	0.5233	0.5401
	StreamCollab	0.6605	0.6356	0.5415	0.5598
	MEGAnno+	0.6398	0.6444	0.5145	0.5165
Ours	**TriIntel**	**0.7205**	**0.7093**	**0.6276**	**0.6559**

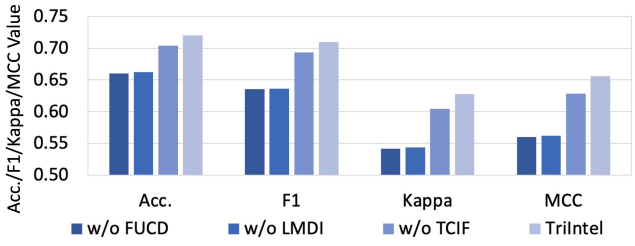


Figure 6: Ablation Study of TriIntel

FUCD module's few-shot learning mechanism effectively identifies new misinformation categories with minimal training data, allowing for rapid initial screening and efficient data processing by DNNs. The LLMs, particularly GPT-4, provide advanced contextual understanding and nuanced language interpretation through the LMDI module, which is critical for detecting subtle forms of misinformation such as conspiracy theories and sarcasm. This capability is demonstrated by TriIntel's high F1-Score and Kappa values, reflecting its precision and reliability. The LMDI module's probabilistic approach to uncertainty measurement ensures that high-risk cases are accurately flagged for human review, significantly enhancing the framework's reliability, as evidenced by the robustness study in Figure 5. Additionally, the TCIF module's integration of crowdsourced human intelligence adds a crucial layer of critical thinking and contextual analysis, ensuring accurate classification of complex cases that automated methods alone might misclassify. The ablation study in Figure 6 further validates the essential contributions of each module to the overall performance of TriIntel, demonstrating that the synergistic combination of DNNs, LLMs, and HI results in a comprehensive and effective misinformation detection system. Overall, TriIntel's design ensures scalability, adaptability, and robustness, making it a powerful and reliable solution for tackling complex public health misinformation challenges.

6 DISCUSSION

Our research significantly advances the state-of-the-art public health misinformation detection solutions in the field of collective intelligence. By integrating DNNs, LLMs, and HI within the TriIntel framework, we offer a novel approach that leverages the complementary strengths of each component to address the complex challenges of misinformation detection. This tripartite intelligence framework not only improves detection accuracy but also addresses the nuanced and sophisticated forms of misinformation that are often overlooked by conventional DNN methods. In the broader context of collective intelligence research, our results demonstrate the efficacy of combining automated and human insights to tackle intricate problems that require both computational efficiency and human judgment. The TriIntel framework's ability to synergize these diverse types of intelligence provides a robust foundation for future research and applications in various domains, including disaster response, healthcare, and environmental monitoring, thereby enriching the collective intelligence research community's understanding and methodologies.

Scalability is a crucial challenge in public health misinformation detection, especially when handling large-scale social media data. Our framework achieves robust scalability primarily through the use of DNNs, which excel at processing and analyzing vast amounts of data rapidly. In the TriIntel framework, DNNs act as the primary filter, efficiently scanning extensive social media feeds to identify potential cases of misinformation for further verification by LLMs and HI. This capacity to manage and analyze large datasets not only accelerates the detection process but also enhances the framework's ability to respond in real-time during public health crises, where the volume of information and speed of spread can be overwhelming. Furthermore, the integration of DNNs with LLMs and HI ensures that the scalability does not compromise the quality of misinformation detection. While DNNs process the bulk of the data, LLMs add a sophisticated layer of contextual understanding, analyzing the nuances of language that may indicate misinformation. This synergy allows the system to expand its detection capabilities without sacrificing the depth of analysis required for complex forms of

misinformation, such as subtle distortions or conspiracy theories. Additionally, human insights in the TriIntel framework provide a critical review layer, ensuring that automated processes uphold accuracy and ethical standards, crucial in public health contexts. This multi-layered approach significantly bolsters the TriIntel framework's scalability and reliability, making it an effective tool for monitoring and combating health misinformation across extensive social media landscapes.

The generalizability of the TriIntel framework extends beyond the specific context of COVID-19 misinformation to other public health domains. For instance, during vaccination campaigns for diseases like measles or influenza, DNNs can be retrained with relevant data from previous outbreaks to identify misinformation trends. LLMs contribute significantly by utilizing their advanced contextual understanding to analyze the nuances of emerging health discussions, refining the detection process by correlating real-time data with historical health trends and guidelines. Human intelligence plays a crucial role in this process by applying ethical judgment and interpreting cultural contexts that may affect the spread and reception of misinformation, thereby enhancing the effectiveness of detection and correction strategies. Similarly, in managing misinformation related to chronic diseases such as diabetes, DNNs are employed to sift through extensive patient data and public discourse to detect prevalent myths regarding treatment options or dietary recommendations. LLMs excel by providing a deep contextual analysis, linking these findings to the latest clinical research and nutritional guidelines. This ensures a highly nuanced detection of misinformation. Human insights are vital, allowing the identification of subtle, culturally-specific misinformation and ethically questionable claims that might be overlooked by automated systems. This integrated approach not only improves the accuracy of misinformation detection but also ensures that interventions are culturally sensitive and ethically sound, underscoring TriIntel's capability to address a variety of public health challenges effectively.

The TriIntel framework offers a robust approach to public health misinformation detection, but it is essential to consider the potential limitations and ethical implications of using LLMs and crowdsourcing. A key challenge with LLMs is their dependency on the quality of their training data. If this data contains biases or inaccuracies, LLMs may inadvertently perpetuate these issues, leading to misclassified posts. Additionally, LLMs can generate plausible yet factually incorrect information, which may further complicate the detection process. To address these challenges, one potential solution involves integrating additional debiasing and fact-checking through fine-tuning and prompting to refine LLM outputs, coupled with the implementation of more robust validation frameworks to regularly assess and reinforce these outputs. Furthermore, the involvement of human contributors in reviewing and interpreting misinformation may lead to emotional or psychological challenges, particularly when dealing with sensitive or distressing content. Future efforts will focus on establishing a robust support system for these contributors. The potential support system could include access to professional mental health resources, training in emotional resilience, and regular psychological assessments to ensure their well-being. Additionally, creating a supportive community where contributors can share experiences and strategies for coping with the stresses of the work could prove highly beneficial.

7 CONCLUSION

This paper presents a TriIntel framework to address the public health misinformation detection problem. In particular, we design a tripartite collaborative learning framework that integrates DNNs, LLMs, and HI to collaboratively derive accurate misinformation labels for studied social media posts. Our TriIntel is the first design that integrates the three forms of intelligence to address the complex social media context-based text classification problem, effectively integrating the distinct yet diversified strengths of DNNs, LLMs, and HI. We observe that TriIntel significantly improves misinformation classification accuracy compared to a rich set of DNNs, LLM, and human-AI collaboration baselines in real-world COVID misinformation detection applications. We believe that TriIntel offers valuable insights for exploring hybrid collective intelligence to address a wide range of real-world applications beyond misinformation detection (e.g., disaster damage assessment, healthcare and medical diagnosis, environmental monitoring).

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Sara Abdali. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities. *arXiv preprint arXiv:2203.13883* (2022).
[2] Theodora Alexopoulou, Marije Michel, Akira Murakami, and Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67, S1 (2017), 180–208.
[3] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
[4] Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656* (2023).
[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.
[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[7] Jian Cui, Kwanwoo Kim, Seung Ho Na, and Seungwon Shin. 2022. Meta-path-based fake news detection leveraging multi-level social context information. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 325–334.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[9] Adam M Enders, Joseph E Uscinski, Casey Klofstad, and Justin Stoler. 2020. The different forms of COVID-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review* (2020).
[10] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. 2022. Human-ai collaboration for ux evaluation: Effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–32.
[11] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*. 595–602.
[12] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. , 277–304 pages.
[13] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on*

machine learning. 1050–1059.
[14] Dawn Gregg. 2009. Developing a collective intelligence application for special education. *Decision Support Systems* 47, 4 (2009), 455–465.
[15] Carina Antonia Hallin, Julian Johannes Umbhau Jensen, Oded Koren, Nir Perel, and Sigbjørn Landazuri Tveteraas. 2019. Testing Smart Crowds for the Economy. In *Proceedings of the ACM Collective Intelligence 2019*. Association for Computing Machinery.
[16] Haiyan Hao and Yan Wang. 2020. Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction* 51 (2020), 101760.
[17] Drew Hemment, Mel Woods, Raquel Ajates Gonzalez, Andrew Cobley, and Angelika Xaver. 2019. Enhancing collective intelligence through citizen science: The case of the GROW citizens' observatory. In *Collective Intelligence 2019*. Association for Computing Machinery (ACM), 1–4.
[18] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining* 10 (2020), 1–20.
[19] YUCHAO Jiang, MARCOS Báez, and BOUALEM Benatallah. 2021. Understanding how early-stage researchers perceive external research feedback. In *ACM Collective Intelligence Conference 2021*.
[20] Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. 2012. A comparison of MCC and CEN error measures in multi-class prediction. *PloS one* 7, 8 (2012), e41882.
[21] Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A Human-LLM Collaborative Annotation System. *arXiv preprint arXiv:2402.18050* (2024).
[22] Raju Kumar and Aruna Bhat. 2021. An analysis on sarcasm detection over twitter during COVID-19. In *2021 2nd International Conference for Emerging Technology (INCET)*. IEEE, 1–6.
[23] Jan Marco Leimeister. 2010. Collective intelligence. *Business & Information Systems Engineering* 2 (2010), 245–248.
[24] João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601* (2023).
[25] Stephan Lewandowsky. 2020. The 'post-truth'world, misinformation, and information literacy: a perspective from cognitive science. *Informed Societies: Why information literacy matters for citizenship, participation and democracy* (2020), 69–88.
[26] Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large language models for generative recommendation: A survey and visionary discussions. *arXiv preprint arXiv:2309.01157* (2023).
[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[28] Melanie J. McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. 2024. Collaborative Human-AI Trust (CHAI-T): A Process Framework for Active Management of Trust in Human-AI Collaboration. *arXiv preprint arXiv:2404.01615* (2024). https://doi.org/10.48550/arXiv.2404.01615
[29] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791* (2020).
[30] Muhammad F Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE Access* 9 (2021), 156151–156170.
[31] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[33] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 21–29.
[34] Bohdan M Pavlyshenko. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704* (2023).
[35] Jiaxin Pei and David Jurgens. 2021. Measuring Sentence-Level and Aspect-Level (Un) certainty in Science Communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9959–10011.
[36] Vishnu S. Pendyala and Christopher E. Hall. 2024. Explaining Misinformation Detection Using Large Language Models. *Electronics* 13, 9 (2024), 1673. https://doi.org/10.3390/electronics13091673
[37] Kate Radcliffe, Helena C Lyson, Jill Barr-Walker, and Urmimala Sarkar. 2019. Collective intelligence in medical decision-making: a systematic scoping review. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 1–11.
[38] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[39] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 913–926.
[40] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)* 54, 9 (2021), 1–40.
[41] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.
[42] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.
[43] Pradeep Kumar Roy, Asis Kumar Tripathy, Tien-Hsiung Weng, and Kuan-Ching Li. 2023. Securing social platform from misinformation using deep learning. *Computer Standards & Interfaces* 84 (2023), 103674.
[44] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems* 36 (2024).
[45] M. Sharma and P. Bhattacharya. 2023. Leveraging Human-AI Collaboration for Improved Misinformation Detection. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023)*. ACM, 190–200. https://doi.org/10.1145/3593011.3593031
[46] Megh Shukla and Shuaib Ahmed. 2021. A mathematical analysis of learning loss for active learning in regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3320–3328.
[47] LWL Simonse. 2022. Social foresights on Covid-19 futures: developing a NLP tool for Strategic Design Research. In *ACM Collective Intelligence conference 2021 (online)*. ACM CI, e1.
[48] Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research* 23, 1 (2021), e17187.
[49] Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. 2020. Frameworks for collective intelligence: A systematic literature review. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–36.
[50] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503* (2021).
[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[52] Eiichiro Uchino, Kanata Suzuki, Noriaki Sato, Ryosuke Kojima, Yoshinori Tamada, Shusuke Hiragi, Hideki Yokoi, Nobuhiro Yugami, Sachiko Minamiguchi, Hironori Haga, et al. 2020. Classification of glomerular pathological findings using deep learning and nephrologist–AI collective intelligence approach. *International journal of medical informatics* 141 (2020), 104231.
[53] Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. 2021. Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on web2vec. In *Proceedings of the conference on information technology for social good*. 19–24.
[54] Toni GLA Van der Meer and Yan Jin. 2019. Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health communication* (2019).
[55] Dong Wang, Lance Kaplan, and Tarek F Abdelzaher. 2014. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)* 10, 2 (2014), 1–27.
[56] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. Evaluating deep learning approaches for covid19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 153–163.
[57] Daniel S Weld, Christopher H Lin, and Jonathan Bragg. 2015. Artificial intelligence and collective intelligence. *Handbook of collective intelligence* (2015), 89–114.
[58] Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, and Nikos Komninos. 2022. Evaluation of fake news detection with knowledge-enhanced language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1425–1429.
[59] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9291–9298.
[60] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
[61] Yang Zhang, Lanyu Shang, Ruohan Zong, Zeng Wang, Ziyi Kou, and Dong Wang. 2021. StreamCollab: A streaming crowd-AI collaborative system to smart urban

infrastructure monitoring in social sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 179–190.

[62] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 12697–12706.

[63] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.

[64] Bahman Zohuri and Farhang Mossavar Rahmani. 2020. Artificial intelligence versus human intelligence: A new technological race. *Acta Scientific Pharmaceutical Sciences (ISSN: 2581-5423)* 4, 5 (2020).