

# Empowering LLMs to Synthesize AI and Human Intelligence for Explainable Public Health Misinformation Detection on Social Media

Ruohan Zong, Yang Zhang, Dong Wang

School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA  
{rzong2, yzhangnd, dwang24}@illinois.edu

## Abstract

This paper studies a critical problem of explainable public health misinformation detection on social media, where clear explanations are essential for enhancing user understanding and trust, surpassing the limitations of black-box misinformation detection results. To tackle this problem, there is a growing trend of leveraging collective intelligence from diverse intelligence sources, such as deep neural networks (DNNs), human intelligence, and large language models (LLMs). However, integrating hybrid intelligence from different sources remains a challenge: DNNs excel in accurate and efficient classification, crowd workers provide contextual understanding and readable explanations, and LLMs offer extensive domain knowledge and advanced language generation. Moreover, current crowdsourcing and human-AI collaboration methods mainly focus on aggregating misinformation detection labels using traditional measures like consistency, often overlooking more complex and challenging inputs like textual explanations. We propose *SynthX*, a collective intelligence framework that incorporates a holistic prompting design to harness the language and reasoning capabilities of LLMs for synthesizing diverse detection and explanation results. It also integrates a novel estimation theory-LLM hybrid approach to assess the varying reliability of detection results from different intelligence sources. Our evaluation on a real-world social media misinformation dataset demonstrates that *SynthX* consistently outperforms a rich set of state-of-the-art baselines in both detection accuracy and explanation quality.

## Introduction

Public health misinformation on social media platforms has become a pervasive issue, threatening both individual and collective well-being (Chen et al. 2022). Effective detection of online misinformation is critical for debunking false claims and preventing their spread, which can lead to serious societal consequences, such as health complications from following inaccurate prevention and treatment advice during the COVID pandemic. However, high-quality misinformation detection is often time-consuming and labor-intensive, relying primarily on dedicated fact-checking websites or manual annotations by experts who verify sources and evidence. Moreover, these fact-checking processes often fall short in addressing multi-class misinformation de-

tection, which involves more nuanced and challenging forms such as conspiracy theories and sarcasm that extend beyond binary true-or-false classifications (Diab, Nefriana, and Lin 2024). These categories are important to consider because they are often more difficult to interpret or more harmful than straightforward false claims. Sarcasm is challenging to identify due to its reliance on subtle cues and implicit meaning, while conspiracy theories are rooted in deeply held beliefs and emotionally charged narratives that can amplify public distrust. Tackling these complex misinformation types enables public health campaigns to not only correct inaccurate information but also address underlying skepticism and fear, fostering trust and reducing resistance to accurate public health information (Kumar and Bhat 2021).

DNNs have been leveraged to enhance misinformation detection accuracy and efficiency (Islam et al. 2020). Additionally, holistic representation and ensemble frameworks have been developed for multi-class misinformation detection (Shang et al. 2024; Rezaei et al. 2022). While highly effective at predicting categories, DNNs often operate as black-box models, offering little insight into decision-making process and limited ability to provide understandable explanations (Yang et al. 2023). This challenge is more pronounced in multi-class misinformation detection, where nuanced categories, such as conspiracy theories and sarcasm, require a deeper interpretability of both the factuality and the subtle tones or contexts that may alter the intended meaning. In contrast, humans tend to comprehend information better when explanations and reasoning accompany the content (McCain 2015). Therefore, providing explanations to users is vital during intervention efforts, especially when correcting misinformation may alter users' original beliefs.

To achieve explainable misinformation detection, eXplainable AI (XAI) approaches focusing on feature importance and representation investigation have been developed (Athira, Kumar, and Chacko 2023). For example, attention mechanisms have been applied to capture and highlight key input patterns, allowing users to understand the most influential parts of the data (Ni, Li, and Kao 2021). While these methods offer valuable insights for AI practitioners, they are less suited for general users during misinformation interventions, who prefer human-like explanations with clear reasoning, evidence, and accessible language. This creates a gap between the technical focus of

current XAI methods and the public needs in misinformation correction. Crowdsourced human intelligence offers a scalable and efficient alternative to domain experts in manual fact-checking, leveraging properly guided general participants to verify claims and formulate explanations effectively (Kim et al. 2018). This approach leverages the diverse knowledge backgrounds and perspectives of a large population, generating explanations that resonate more effectively with online users.

Emerging research also explores the use of state-of-the-art LLMs for online misinformation detection (Chen and Shu 2023; Xu et al. 2024). LLMs possess a vast repository of world knowledge, including societal common sense and general domain expertise, and demonstrate strong reasoning and language generation capabilities with a reasonable quality and readability. Despite the strengths, LLMs are prone to issues like hallucinations and vagueness (OpenAI 2023). For instance, LLMs have been shown to inadvertently produce highly convincing but fabricated health misinformation, such as false health articles that appear authentic and credible (De Angelis et al. 2023).

Motivated by the limitations of solutions that rely on a single source of intelligence, *we propose a novel approach that synthesizes misinformation detection and explanation results from DNNs, crowd workers, and LLMs*. Our approach explores the complementary strengths of DNNs for accurate and efficient classification, crowd workers for their ability to grasp contextual nuances and provide readable explanations, and LLMs for their extensive domain knowledge, language generation, and reasoning abilities. Traditional aggregation methods typically focus on consolidating labels predicted by AI or annotated by humans (Zhang et al. 2014). In contrast, LLMs excel in processing and synthesizing complex and diverse textual information. Moreover, by aggregating inputs from diverse sources like DNNs and humans, LLM reduces risks such as hallucination and vagueness by cross-validating information and learning from concrete inputs. Unlike Retrieval-Augmented Generation (Gao et al. 2023), which retrieves information from static external databases to mitigate hallucination, our approach integrates AI and human inputs to gain context-specific and user-readable insights for each post. Hence, this paper leverages LLMs to synthesize insights from multiple sources, achieving readable, reliable, and explainable misinformation detection. There are two key technical challenges:

The first challenge lies in *leveraging LLMs to synthesize answers of varying quality from intelligence sources with diverse levels of reliability*. First, *source reliability* refers to the varying trustworthiness of different intelligence sources. It has been observed that different sources often exhibit different levels of reliability, which can be estimated by measuring the consistency of their predictions across various samples. Second, unlike label aggregation, assessing *explanation quality* is more difficult as it requires evaluating evidence provided, factual accuracy, and logical coherence of reasoning. An explanation backed by solid evidence and clear reasoning should carry more weight than one based on vague or speculative conclusions. For instance, an explanation of “Scientific research shows that 5G does not harm the

immune system. The WHO confirms there is no biological link between 5G radiation and immune function. COVID-19 is caused by the SARS-CoV-2 virus, spread through respiratory droplets, not through radio waves” provides concrete and well-supported information. In contrast, an explanation like “This tweet seems to spread a conspiracy about the connection between 5G and the coronavirus” lacks substantiation and clarity, and therefore should be weighted less in its quality. Hence, synthesizing explainable detection results is challenging, as it requires accounting for both the reliability of each source in detection accuracy and the factual accuracy and logical rigor of each explanation.

The second challenge is *how to leverage LLMs to synthesize heterogeneous and diversified answers from three distinct types of intelligence sources*. Traditional aggregation methods, such as expectation maximization (EM) and Bayesian aggregation, typically focus on homogeneous data types like categorical or numerical labels. However, this paper tackles a more complex scenario of synthesizing *heterogeneous* data types: DNNs provide only categorical labels, while crowd workers and LLMs offer categorical labels with textual explanations. The diversity in the responses from these sources further complicates the synthesis process. Crowd workers’ explanations are typically concise and straightforward. For example, “COVID-19 is not a bioweapon, no credible evidence supports this claim” is easy to understand but lacks depth in addressing the origin of such conspiracy or providing scientific counterpoints. On the other hand, LLM-generated explanations tend to be more thorough, such as elaborating on the origins of the virus and referencing virology studies. However, LLM also introduces the risk of hallucination, such as speculating on unrelated theories or providing fabricated sources. Ensuring LLM functions as a synthesizer is crucial for cross-referencing diverse inputs, reducing over-reliance on its own outputs, and emphasizing reliable sources. Balancing strengths and weaknesses of heterogeneous responses remains a significant challenge for producing trustworthy results.

To address the outlined challenges, we introduce *SynthX*, an LLM-driven framework designed to synthesize both misinformation detection and *eXplanation* outputs. For the first challenge, we introduce a hybrid estimation-theoretical approach that accurately estimates source reliability while leveraging the LLM’s capabilities to assess explanation quality. For the second challenge, we propose a prompting strategy to synthesize categorical and explanatory detection results from diverse sources, while accounting for their varying levels of reliability. Evaluation results on a real-world social media dataset show that SynthX significantly improves both misinformation detection and explanation compared to a diverse set of DNN, LLM, and human-AI baselines. We summarize our key contributions as follows:

- We tackled the challenge of synthesizing heterogeneous detection and explanation results from diverse types of intelligence for explainable misinformation detection.
- We developed SynthX, a hybrid system that evaluates both source reliability and explanation quality, and synthesizes detection and explanation outcomes to improve

misinformation detection accuracy and explainability.

- Our experiments on a real-world social media dataset show that SynthX consistently outperforms state-of-the-art DNN, LLM, and human–AI baselines in both detection accuracy and explanation quality.

## Related Work

### Explainable Misinformation Detection

AI-driven approaches have gained prominence in online misinformation detection due to their ability to enhance accuracy, efficiency, and scalability (Paraschiv et al. 2022; Zong et al. 2023; Papanikou et al. 2024). However, many of these AI-based methods lack explainability, as they focus solely on predicting misinformation labels without offering easily accessible explanations. To address this gap, explainable AI (XAI) techniques have emerged as a solution for making misinformation detection more understandable (Epstein et al. 2022). For instance, Ni, Li, and Kao (2021) introduced attention networks to capture the meaning of social media posts and their propagation structure, providing attention visualization interpretability. Mohseni et al. (2021) investigated an XAI assistant in news review platforms, showing how algorithmic transparency aids users in building trust and combating misinformation propagation. Chien, Yang, and Yu (2022) combined an LSTM model, a relevance propagation algorithm, and a transparency model to enhance understanding and trust of social media users in system decisions. While these methods offer valuable explainable insights of model predictions for AI practitioners, they may be less accessible to general users during misinformation interventions, where more intuitive explanations are needed. In contrast, we introduce a hybrid system that integrates the strengths of human reasoning and LLM processing to generate more readable and well-grounded explanations, tailored for general users in online environments.

### LLMs to Automate Human-Centric Tasks

Recent advancements in generative AI and LLMs are transforming traditional tasks conducted by humans such as misinformation detection, content annotation, and explanation generation (Xu et al. 2024). In particular, LLMs have been used to generate high-quality textual outputs and to assess the quality of human-generated content, significantly reducing reliance on human annotators (Yenduri et al. 2024). For example, Ostyakova et al. (2023) compared experts, crowd workers, and GPT in annotating open-domain conversations, showing that, with the help of prompt engineering, GPT can sometimes reach human-level performance in complex discourse annotation tasks. He et al. (2024) examined GPT’s performance in a data annotation pipeline alongside a well-structured crowdsourcing system, finding that GPT slightly outperformed crowd workers in labeling accuracy, with further accuracy gains when combining GPT’s labels with those from the crowd using advanced aggregation methods. Jiang et al. (2024) explored the evolving challenge of misinformation detection with LLMs, investigating how LLMs could be used both as a threat and a defense mechanism, while also addressing the limitations of current detec-

tion methods. However, challenges like LLM hallucinations and the need for nuanced human judgment remain key areas of focus for further improvement. We introduce a framework that positions LLM as a synthesizer, rather than merely a generator, cross-referencing inputs from multiple sources to ensure grounded and concrete explanations.

### Human–AI Collaboration

Human–AI collaboration is critical in leveraging the complementary strengths of collective intelligence, especially in complex tasks like online misinformation detection. On one side, human intelligence is employed to troubleshoot and improve AI systems. For instance, Zhan et al. (2022) explored active learning strategies like uncertainty sampling and query-by-committee, which enhance model performance by focusing on the most informative data samples, thus reducing the need for extensive manual labeling. Ouyang et al. (2022) demonstrated that fine-tuning large language models with reinforcement learning from human feedback significantly improves their alignment with user preferences. On the other side, AI is used to support human decision-making by pre-processing information, generating recommendations, and producing preliminary results. For example, Zhang, Liao, and Bellamy (2020) found that confidence scores and explanations can help calibrate human trust in AI models, improving decision-making by facilitating collaboration between human expertise and AI. However, these methods often focus on a single direction—either using humans to help AI or using AI to assist humans. We synthesize the complementary strengths of both types of intelligence by leveraging LLMs’ advanced information processing and synthesis capabilities to bridge the gap between the two directions of human–AI collaboration.

## Problem

In this paper, we tackle the problem of *explainable misinformation detection*, which focuses on both identifying the category of misinformation and providing an accompanying explanation for each social media post. The task involves classifying posts into categories such as true information, false information, sarcasm, or conspiracy, and justifying these classifications with clear explanations. Formally, let the input be a set of social media posts  $P_i$  for  $i \in \{1, \dots, N\}$ , where  $N$  is the total number of post samples. The objective is twofold: 1) to classify each post  $P_i$  into a category  $C_i$ , where  $C_i \in \{\text{true, false, sarcasm, conspiracy}\}$  and  $K$  is the number of categories, and 2) to generate an explanation  $E_i$  that provides reasoning or evidence supporting the classified category. Illustrative examples of misinformation detection category and explanation are shown in Figure 1.

We focus on synthesizing both misinformation detection categories and corresponding explanations from multiple sources, which can be categorized into three types: 1) *Misinformation categories detected by DNNs*: For each social media post  $P_i$ , we obtain predicted categories  $\hat{C}_i^{DNN_a}$  from a set of DNNs, where  $a \in \{1, \dots, A\}$  and  $A$  is the number of distinct DNN models used (e.g., BERT, Roberta). These models classify posts into predefined categories based on

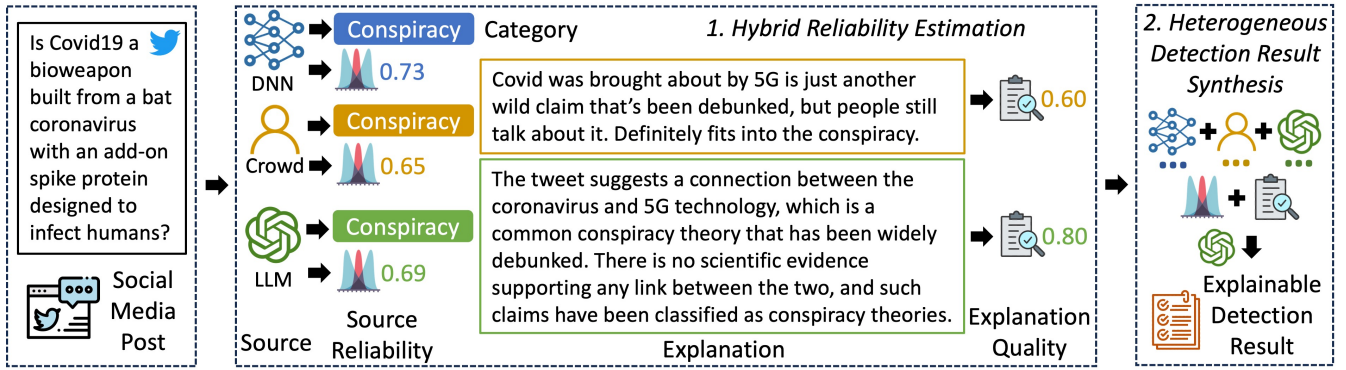


Figure 1: Overview of SynthX Framework

the content of the post. 2) *Misinformation categories and explanations from crowd workers*: For each post  $P_i$ , we collect both a predicted category  $\hat{C}_i^{Crowd_b}$  and an explanation  $\hat{E}_i^{Crowd_b}$  from crowd workers, where  $b \in \{1, \dots, B\}$  and  $B$  is the number of crowd workers assigned to annotate each post. Crowd workers provide not only the classification but also a textual explanation justifying their choice. 3) *Misinformation categories and explanations from LLMs*: For each post  $P_i$ , we also gather categories  $\hat{C}_i^{LLM_c}$  and explanations  $\hat{E}_i^{LLM_c}$  from LLMs, where  $c \in \{1, \dots, C\}$  and  $C$  represents the number of different LLM instances utilized.

Our objective is to synthesize both misinformation detection categories and corresponding explanations from all three sources:  $\hat{C}_i^{DNN_a}$ ,  $\hat{C}_i^{Crowd_b}$  and  $\hat{E}_i^{Crowd_b}$ , and  $\hat{C}_i^{LLM_c}$  and  $\hat{E}_i^{LLM_c}$ . The aim is to produce a final reliable (e.g., factual and evidence-based) misinformation detection result,  $\hat{C}_i$ , along with a corresponding explanation,  $\hat{E}_i$ , for each post  $i \in \{1, \dots, N\}$ . This process is formally defined as follows for  $\forall i \in \{1, \dots, N\}$ :

$$\begin{aligned} \max P(\hat{C}_i = C_i) \\ \max F(\hat{E}_i), D(\hat{E}_i) \end{aligned} \quad (1)$$

where  $P(\hat{C}_i = C_i)$  represents the probability that the synthesized detection matches the true category,  $F(\hat{E}_i)$  and  $D(\hat{E}_i)$  denote the factuality and evidence measurements of the generated explanation, respectively. These evaluation dimensions are designed to counter common challenges faced by LLMs, such as hallucinations and vague reasoning. Our goal is to synthesize the diverse predictions and explanations from DNNs, crowd workers, and LLMs into a coherent and reliable output for each post. The synthesis process must account for the varying levels of reliability across different sources, as well as the differing quality of explanations, to generate a final, trustworthy misinformation detection result along with a high quality accompanying explanation.

### Solution

SynthX is a collective intelligence framework that harnesses the advanced language reasoning and processing capabilities of LLMs to synthesize multiple forms of intelligence,

producing high-quality, explainable misinformation detection results. An overview of SynthX framework is shown in Figure 1. It comprises two key modules:

- *Hybrid Reliability Estimation*: This module integrates estimation theory-based source reliability assessment with LLM-driven explanation quality evaluation, enabling the system to assess inputs with diverse data types with varying source reliability and explanation quality.
- *Heterogeneous Detection Result Synthesis*: This module synthesizes both categorical and explanatory results from different sources, together with the source reliability and explanation quality estimated by the first module, to generate final explainable misinformation detection result.

### Hybrid Reliability Estimation

Our solution integrates three key types of intelligence: DNNs, crowd workers, and a LLM. This hybrid approach ensures that we synthesize diverse insights while addressing the inherent limitations of each source. First, we trained three DNN models that have consistently demonstrated strong performance in language classification tasks: BERT (Devlin et al. 2019), RoBERTa (Liu 2019), and XLNet (Yang et al. 2020). These models provide predictions of the potential misinformation categories, denoted as  $\hat{C}_i^{DNN_1}$ ,  $\hat{C}_i^{DNN_2}$ , and  $\hat{C}_i^{DNN_3}$  for  $i \in \{1, \dots, N\}$ . DNNs were chosen for the ability to efficiently produce highly accurate predictions, making them a reliable foundation for detecting misinformation categories. However, we do not include explanations from DNNs because, while they excel at classification tasks, they typically function as black-boxes and lack the inherent ability to produce human-readable explanations. Although post-hoc explanation techniques exist for DNNs such as feature importance or attention, they tend to offer less user-friendly explainability compared to the other two sources. In contrast, crowd workers and LLMs naturally produce textual explanations, offering insights that are more intuitive and accessible for general users. These sources are better suited to provide understandable explanations that DNNs alone cannot deliver as effectively.

Second, we leveraged the collective intelligence of crowd workers by collecting misinformation detection results and

explanations from three individuals for each social media post via crowdsourcing platforms. In addition to selecting the category they believed a social media post belonged to, crowd workers were required to provide an explanation justifying their selection, based on their personal understanding and reasoning. Their provided categories are denoted as  $\hat{C}_i^{Crowd_1}$ ,  $\hat{C}_i^{Crowd_2}$ ,  $\hat{C}_i^{Crowd_3}$ , and explanations are denoted as  $\hat{E}_i^{Crowd_1}$ ,  $\hat{E}_i^{Crowd_2}$ ,  $\hat{E}_i^{Crowd_3}$ . The inclusion of crowd-sourced explanations introduces human-like reasoning and evidence-based explanations, enhancing the robustness of misinformation detection and improving the readability of the explanations. However, crowd worker contributions are often limited by individual knowledge gaps and less polished language, which can result in incomplete or less comprehensive responses.

Third, we gathered detection and explanation results from GPT-4 (OpenAI 2023), a state-of-the-art LLM. Through carefully designed prompts, GPT-4 was tasked with both classifying misinformation and providing explanations to justify its conclusions, denoted as  $\hat{C}_i^{LLM_1}$  and  $\hat{E}_i^{LLM_1}$ . For consistency in implementation and comparison across models, we selected a relatively small temperature value of 0.3. This value minimizes randomness in outputs, enhancing reliability—a critical requirement in misinformation detection and explanation tasks, where accuracy and consistency are prioritized over creativity. The slight allowance for variability ensures explanations remain grounded and precise without sacrificing the flexibility needed for nuanced reasoning. GPT-4’s strengths lie in its vast integration of common sense and domain knowledge, alongside its advanced reasoning and language generation capabilities. Despite these advantages, the model is prone to issues like hallucination and vagueness in its responses, which necessitates careful consideration of its outputs.

Motivated by the complementary strengths and weaknesses of the three distinct sources of intelligence—DNNs, crowd workers, and LLMs—we introduce a hybrid reliability estimation module to assess both *source reliability* and *explanation quality*. We define *source*  $s$  to represent an entity in the set of all aforementioned DNNs, crowd workers, and LLM as follows:  $s \in \{DNN_a, Crowd_b, LLM_1 | a = 1, \dots, A; b = 1, \dots, B\}$ .

First, different types of intelligence exhibit varying levels of reliability. For instance, DNN models typically achieve a relatively high accuracy in detecting misinformation categories due to their training on large-scale dedicated datasets, enabling them to detect complex patterns and nuances that individual crowd workers might miss. Crowd workers, while less accurate in some cases, bring a different perspective, especially in detecting contextual subtleties and understanding the emotions behind certain pieces of misinformation. GPT, on the other hand, generally provides accurate predictions by leveraging vast amounts of domain knowledge and advanced reasoning capabilities. However, it may also encounter issues such as hallucination, where it generates convincing but fabricated information in certain cases. To address the problem of source reliability, we employ the Expectation Maximization (EM) algorithm. The EM algorithm

operates as an iterative optimization method that alternately updates the predicted misinformation category and reliability of each source based on the likelihood function:

$$L = \prod_{i=1}^N \prod_{k=1}^K \left( \prod_{a=1}^A P\left(\hat{C}_i^{DNN_a} = k \mid C_i = k, R^{DNN_a}\right)^{\mathbb{1}(C_i=k)} \cdot \prod_{b=1}^B P\left(\hat{C}_i^{Crowd_b} = k \mid C_i = k, R^{Crowd_b}\right)^{\mathbb{1}(C_i=k)} \cdot P\left(\hat{C}_i^{LLM_1} = k \mid C_i = k, R^{LLM_1}\right)^{\mathbb{1}(C_i=k)} \right) \quad (2)$$

where  $P(\hat{C}_i^s = k \mid C_i = k, R^s)^{\mathbb{1}(C_i=k)}$  represents the probability that source  $s$  predicts  $\hat{C}_i^s = k$ , given the true category  $C_i = k$  and the reliability  $R^s$  of source  $s$ .  $N$  is the total number of social media posts and  $K$  is the number of misinformation categories.

In the iterative process, the reliability of each member, whether a DNN, crowd worker, or LLM, is inferred based on its consistency with other sources in the absence of ground truth labels. For instance, if a particular source frequently agrees with others in predicting the same misinformation category, its reliability is considered higher. Conversely, a source that consistently deviates from the majority would be deemed less reliable. This allows us to assess the reliability of sources when the true labels are unknown. Thus, we ran the EM algorithm to evaluate the reliability of each of the three DNN models, three crowd workers, and the GPT-4 model. The reliability scores are denoted as  $\hat{R}^{DNN_1}$ ,  $\hat{R}^{DNN_2}$ ,  $\hat{R}^{DNN_3}$ ,  $\hat{R}^{Crowd_1}$ ,  $\hat{R}^{Crowd_2}$ ,  $\hat{R}^{Crowd_3}$ , and  $\hat{R}^{LLM_1}$ . The update rule for reliability  $R^s$  of source  $s$  at iteration  $t + 1$  is given by:

$$R_{(t+1)}^s = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \mathbb{1}(C_i = k) \cdot P\left(C_i = k \mid \hat{C}_i^s = k, \hat{C}_i^{-s}, R_{(t)}^s\right) \quad (3)$$

where  $\hat{C}_i^{-s}$  denotes the predictions of all other sources except  $s$  for post  $i$ . Term  $P(C_i = k \mid \hat{C}_i^s = k, \hat{C}_i^{-s}, R_{(t)}^s)$  is the posterior probability of the true category  $C_i = k$ , given the predictions from source  $s$  and other sources at iteration  $t$ .

The final reliability score for each source is calculated based on the consistency of their predictions with all other sources as follows:

$$R^s = \frac{\sum_{i=1}^N \sum_{k=1}^K \mathbb{1}(\hat{C}_i^s = k) \cdot \mathbb{1}(C_i = k)}{N} \quad (4)$$

where  $\mathbb{1}(\hat{C}_i^s = k) \cdot \mathbb{1}(C_i = k)$  is the indicator function that returns 1 if the prediction  $\hat{C}_i^s$  from source  $s$  matches the true category  $C_i$ , and 0 otherwise. These reliability estimates are then incorporated into the second module of misinformation detection and explanation synthesis, ensuring that more

trustworthy sources have a greater influence on the final results. Although these reliability scores depend on comparative consistency among sources, our holistic framework integrates an additional dimension of reliability assessment, mitigating the risk of relying solely on majority consensus.

Another critical dimension considered by our framework is explanation quality, which varies significantly across different sources. For example, crowd workers or LLMs may possess different levels of expertise or familiarity with the content of social media posts. Explanations that include clear evidence and well-structured reasoning naturally have higher quality than those that are vague or speculative as illustrated in the Introduction Section. To tackle the problem of varying explanation quality across crowd workers and LLMs for different social media posts, we utilize the language understanding and reasoning capabilities of the GPT-4 model. For each social media post, we provide GPT-4 with the detection task, along with the categories and explanations generated by each crowd worker and LLM. We carefully prompt GPT-4 to prioritize evaluating the reasoning behind the explanations rather than the correctness of the final category, thus distinguishing this step from the earlier reliability assessment, which focuses on category detection. To ensure an unbiased evaluation of explanation quality, we exclude any information about the source of each explanation (i.e., whether it comes from a crowd worker or LLM). This allows GPT-4 to assess the explanation purely based on its contents, avoiding any potential bias or preconceptions about the different sources of intelligence. The explanation quality for the crowd workers, and the LLM on each social media post are denoted as  $\hat{Q}_i^{Crowd_1}$ ,  $\hat{Q}_i^{Crowd_2}$ ,  $\hat{Q}_i^{Crowd_3}$ , and  $\hat{Q}_i^{LLM_1}$  for  $i \in \{1, \dots, N\}$ . This method helps in providing a fair comparison of explanation quality, ensuring that the focus remains on the strength of the reasoning and clarity of the explanation.

## Heterogeneous Detection Result Synthesis

Building on the source reliability and explanation quality scores, we developed a comprehensive approach to synthesize both misinformation detection categories and explanations from the various types of intelligence sources. Unlike traditional aggregation methods that focus solely on label predictions, our method addresses the more complex scenario of heterogeneous data types and the diversity of inputs from different sources, as illustrated in the Introduction Section. Specifically, DNN models provide only misinformation detection categories, with their reliability estimated through the EM algorithm. In contrast, both crowd workers and LLMs provide misinformation detection categories along with explanations that justify their choices, each accompanied by its source reliability and explanation quality. The challenge lies in effectively synthesizing these three distinct types of intelligence—DNNs, crowd workers, and LLMs—which provide different kinds of inputs, such as labels and explanations, with varying levels of reliability and quality. Following the synthesis of heterogeneous detection results from various intelligence sources, our objective is to generate the final estimated misinformation detection cat-

egory and the synthesized explanations. This process leverages the source reliability scores for predicted categories and the explanation quality scores for explanatory text.

First, the final estimated category  $\hat{C}_i$  for an item  $i$  can be estimated by integrating the predicted categories from DNN models, crowd workers, and LLMs, weighted by their respective source reliability scores as follows:

$$\begin{aligned} \hat{C}_i = & \sum_{a=1}^A \hat{R}^{DNN_a} \cdot f(\hat{C}_i^{DNN_a}) \\ & + \sum_{b=1}^B \hat{R}^{Crowd_b} \cdot f(\hat{C}_i^{Crowd_b}) \\ & + \hat{R}^{LLM_1} \cdot f(\hat{C}_i^{LLM_1}) \end{aligned} \quad (5)$$

where  $f(\cdot)$  indicates a numerical encoding or weighting function that translates predicted categories into quantitative values (Hryniewska et al. 2024). It aligns with ensemble learning methods, where model predictions are combined based on reliability scores to improve decision-making by normalizing and weighting outputs from multiple sources.

Second, the explanations  $\hat{E}_i$  for each item  $i$  can be synthesized by taking into account both the reliability and quality scores of the explanations provided by the crowd workers and LLMs as follows:

$$\begin{aligned} \hat{E}_i = & g\left(\{\hat{Q}_i^{Crowd_b}, \hat{R}^{Crowd_b}, \hat{E}_i^{Crowd_b}\}, \right. \\ & \left. \{\hat{Q}_i^{LLM_1}, \hat{R}^{LLM_1}, \hat{E}_i^{LLM_1}\}\right) \quad \forall b \in \{1, \dots, B\} \end{aligned} \quad (6)$$

where  $g(\cdot)$  focuses on synthesizing explanations by scoring explanation quality, emphasizing factuality or evidence (Wang 2023). Tools like GPTScore assess these aspects to evaluate how well generated explanations align with source content, using metrics like factuality and evidence to ensure clear and accurate outputs, critical in misinformation detection settings.

To synthesize heterogeneous detection categories and explanations while accounting for their varying reliability and qualities, we carefully prompt GPT to perform the synthesis by following these steps: 1) We remove any indication of which type of intelligence (DNN, crowd worker, or LLM) each answer originates from to eliminate potential preconceptions that could influence GPT’s synthesis process. 2) We explicitly specify the source reliability score for the predicted categories and the explanation quality score for each explanation. This ensures that GPT can weigh the contributions of each source appropriately, factoring in both the accuracy of the category prediction and the quality of the explanation provided.

An illustrative example of our synthesizing prompt design is presented in the following gray prompt box. Specifically, we use an example tweet, where the ground truth label for its misinformation detection category is conspiracy. We provide sample answers from three types of intelligence sources: Source 1 represents a DNN, Source 4 a crowd worker, and Source 7 an LLM. As shown in the example, DNN mod-

els often achieve higher reliability scores than crowd workers and LLM (e.g., 0.73 vs. 0.65 vs. 0.69) because they are trained on large-scale datasets specifically curated for the misinformation detection task. Crowd workers, while generally having lower prediction reliability scores, can provide easily understandable explanations, as demonstrated in Source 4, “is just another wild claim that’s been debunked.” LLMs, on the other hand, are capable of generating more comprehensive and polished explanations, such as the statement in Source 7: “There is no scientific evidence supporting any link between the two, and such claims have been classified as conspiracy theories.” This prompt design guides GPT to synthesize these diverse inputs, generating a balanced and objective final classification along with a concise explanation. A pseudocode summary of SynthX is in Appendix.

### Synthesizing Prompt Design

**Task:** Categorize the following tweet into one of the four categories: True Information, False Information, Conspiracy, or Sarcasm. Provide a justification for your classification decision.

**Tweet:** Is Covid19 a bioweapon built from a bat coronavirus with an add-on spike protein designed to infect humans? If true, then who built it, and was it released accidentally or deliberately?

We gathered responses from seven different sources, with respective reliability scores for predictions and quality scores for explanations. The prediction reliability and explanation quality are evaluated using different methods, leading to differences in scale or range. Your task is to synthesize these results and provide the final classification along with an explanation that justifies your decision. This final answer should be clear and focused on helping the general public mitigate misinformation, without mentioning the multiple sources involved.

**Source 1: Answer:** Conspiracy.  
**Prediction Reliability:** 0.73

... ..

**Source 4: Answer:** Conspiracy. Covid was brought about by 5G is just another wild claim that’s been debunked, but people still talk about it. Definitely fits into the conspiracy.  
**Prediction Reliability:** 0.65  
**Explanation Quality:** 0.60

... ..

**Source 7: Answer:** Conspiracy. The tweet suggests a connection between the coronavirus and 5G, which is a common conspiracy theory that has been widely debunked. There is no scientific evidence supporting any link between the two, and such claims have been classified as conspiracy theories. Hence, this tweet falls into the conspiracy category.  
**Prediction Reliability:** 0.69  
**Explanation Quality:** 0.80

## Evaluation

### Datasets and Crowdsourcing Settings

We use a publicly available dataset focused on misinformation detection, sourced from Twitter/X using public health-

related keywords during the COVID pandemic (Memon and Carley 2020). The data collection targeted tweets combining terms like “COVID” with health misinformation topics, such as unproven remedies (e.g., bleach, garlic), controversial treatments (e.g., hydroxychloroquine, steroids), conspiracy theories (e.g., 5G, bioweapon), and claims about immunity, cures, and vaccines. The dataset includes ground truth labels for 17 categories, which were manually annotated and cross-validated by six annotators following the detailed guidelines outlined in the annotation codebook<sup>1</sup>.

To tailor the dataset to the misinformation detection task, we excluded categories irrelevant to information validity (e.g., politics, panic buying). Next, we consolidated overlapping categories: 1) *true treatment*, *true prevention*, *true public health response*, *emergency response*, and *news* were merged into the category of *true information*; 2) *false fact or prevention*, *fake cure*, *fake treatment*, and *false public health response* were combined into the category of *false information*; 3) *conspiracy* and *sarcasm* were retained as distinct categories. In compliance with Twitter/X’s terms of service, we retrieve the text of tweets accessible on the platform using the tweet IDs provided in the dataset. After processing, the dataset comprises 1,607 tweets, divided into four categories: *true information* (24.6%), *false information* (20.8%), *conspiracy* (32.9%), and *sarcasm* (21.7%). A 7:3 split between training and testing data is used for evaluating all methods. Aligned with standard practices in language applications, fine-tuning pre-trained language models on downstream tasks has proven effective with relatively small datasets, as opposed to training models from scratch. For example, Grieshaber, Maucher, and Vu (2020) demonstrated BERT’s adaptability by fine-tuning it on datasets with fewer than 1,000 training samples. Therefore, our dataset size is sufficient to effectively fine-tune and evaluate all baselines.

Our crowdsourcing interface design is presented in Figure 2. We employ the interface to gather misinformation predictions and explanations from crowd workers on Amazon Mechanical Turk, a widely-used crowdsourcing platform that provides around-the-clock access to a large global workforce. Crowdsourcing offers a more efficient and cost-effective approach compared to relying solely on domain experts, enabling us to rapidly collect a high volume of annotations at an reduced expense. For each explainable misinformation detection task, we assign three crowd workers, compensating each \$0.05 per task. The process adheres to the guidelines established by the Institutional Review Board (IRB) protocol to ensure ethical compliance. To ensure the quality of the labels, we set qualification criteria, requiring crowd workers to have completed at least 10,000 approved tasks and to maintain an approval rating of 98% or higher before participating in our project. The inter-worker agreement, measured by a Kappa score of 0.667, indicates a solid level of consistency, as a Kappa score above 0.6 reflects good agreement between individuals (Cohen 1960).

<sup>1</sup><https://zenodo.org/records/4024154>

**Please select the category you believe this tweet belongs to and provide a corresponding explanation.**

Select the category that best fits this tweet from the four options provided and write an explanation. You are **not** required to search for external sources to verify your answer, but your explanation should be **based on your understanding and reasoning**.

Please note that your response may be **rejected** if **no explanation** is given or if the explanation is **irrelevant**.

**Tweet:** Is Covid19 a bioweapon built from a bat coronavirus with an add-on spike protein designed to infect humans? If true, then who built it, and was it released accidentally or deliberately?

**Category:** Select the category you believe this tweet belongs to.

- Conspiracy
- Sarcasm or Satire
- True Information
- False Information

**Explanation:** Provide an explanation justifying your selected category. Responses may be rejected if no explanation is given or if the explanation is irrelevant.

**Submit**

Figure 2: Illustration of Crowdsourcing Interface

## Baselines and Evaluation Settings

To thoroughly evaluate our SynthX approach, we include a diverse set of DNN, LLM, and human-AI baselines in the assessment process: **BERT** (Devlin et al. 2019), **RoBERTa** (Liu 2019), **XLNet** (Yang et al. 2020), **GPT-4** (OpenAI 2023), **LLaMA3** (Touvron et al. 2023), **Mistral** (Jiang et al. 2023), **TD** (Wang et al. 2012), **DAL** (Zhan et al. 2022), **StreamCollab** (Zhang et al. 2021), **MEGAnno+** (Kim et al. 2024). Detailed descriptions of the compared baselines are presented in Appendix. DAL, StreamCollab, and MEGAnno+ are crowd-AI collaborative baselines, where the AI components of them are trained using the training set. To evaluate the explanation generation capability of our SynthX framework, we generate misinformation explanations for all baseline models that produce textual outputs: **GPT-4**, **LLaMA3**, **Mistral**, and the **Crowd**. We prompted the LLMs for both the misinformation detection category and its corresponding explanation. Additionally, we used the crowdsourcing interface shown in Figure 2 to gather both detection categories and explanations from crowd workers. The evaluation results for the Crowd were obtained by averaging the scores for each metric (factuality and evidence) across three crowd workers collected for each social media post. We did not use crowdsourcing aggregation baselines for explanation aggregation, as they are specifically designed for label aggregation and not for synthesizing explanations. The DNN models in our framework, as well as the DNN baselines, were implemented using PyTorch and trained on NVIDIA A16 GPUs. We optimized hyperparameters with the Adam optimizer, using a learning

rate of  $1 \times 10^{-5}$  and a batch size of 100, for 50 epochs until convergence. GPT-4 was integrated into both our framework design and as a baseline model, with both implementations carried out via the GPT-4 API<sup>2</sup>.

We use four commonly adopted multi-class textual classification metrics to assess misinformation detection accuracy (Artstein and Poesio 2008): 1) Accuracy, 2) F1-Score, 3) Cohen’s Kappa (Kappa), and 4) Matthews Correlation Coefficient (MCC). Higher scores in these metrics reflect stronger model performance. To evaluate the quality of misinformation explanations, we employ two key metrics (He, Ahamad, and Kumar 2023): 1) Factuality: does the explanation accurately and truthfully justify the tweet’s classification into the misinformation category? 2) Evidence: does the explanation provide clear and convincing evidence to support the misinformation category? We prompted GPT-4 as the evaluator by providing it with the definitions of Factuality and Evidence outlined above and instructing it to assess each explanation accordingly. GPT-4 can effectively evaluate explanation quality because the task of evaluation is distinct from detection, and prior research has shown that LLMs can effectively judge the quality of generated responses in various tasks. For example, GPT-4 has been successfully employed in evaluating language generation tasks, where it demonstrates a strong correlation with human judgment, particularly in evaluating factual accuracy (Liu et al. 2023; OpenAI 2023). These evaluation dimensions are designed to counter common LLM challenges, such as hallucinations and vagueness. Hallucinations occur when the model generates information that is not grounded in reality or fact, leading to explanations that may seem coherent but are incorrect. By focusing on factuality, we aim to ensure that the generated explanations are both accurate and aligned with the true nature of the misinformation. Similarly, vagueness can obscure the reliability of explanations, as they may lack the necessary specificity or substantiation. The evidence metric addresses this by requiring concrete support for each classification, thereby reinforcing clarity and mitigating the risk of ambiguous or unsupported explanations.

## Evaluation Results

**Detection Accuracy Comparison** In this set of experiments, we compare the classification accuracy of our SynthX framework with all baseline models for misinformation detection. The results, as shown in Table 1, clearly demonstrate SynthX’s performance improvements over a wide range of baseline methods. For example, compared to the best-performing baseline, MEGAnno+, SynthX achieves substantial improvements across all evaluation metrics, including 3.75% in Accuracy, 3.75% in F1-Score, 5.48% in Kappa, and 5.46% in MCC. These improvements are driven by SynthX’s ability to integrate DNNs, crowd intelligence, and LLMs, leveraging the efficient and large-scale classification strengths of deep models, the contextual understanding from human crowd input, and the advanced reasoning capabilities of LLMs. This approach effectively handles heterogeneous input types with varying reliability, ensuring robust

<sup>2</sup><https://platform.openai.com/>

Algorithm	Accuracy	F1-Score	Kappa	MCC
BERT	0.7785	0.7787	0.6989	0.6996
RoBERTa	0.8012	0.7950	0.7251	0.7290
XLNet	0.7743	0.7731	0.6925	0.6932
GPT-4	0.8054	0.8077	0.7359	0.7363
LLaMA3	0.7371	0.7313	0.6465	0.6594
Mistral	0.6936	0.6964	0.5868	0.5893
TD	0.7909	0.7859	0.7083	0.7183
DAL	0.8240	0.8210	0.7579	0.7594
StreamCollab	0.8260	0.8204	0.7597	0.7631
MEGAnno+	0.8261	0.8259	0.7634	0.7643
<b>SynthX</b>	<b>0.8571</b>	<b>0.8569</b>	<b>0.8052</b>	<b>0.8060</b>

Table 1: Detection Accuracy Comparison

Algorithm	GPT-4	LLaMA3	Mistral	Crowd	<b>SynthX</b>
Factuality	0.8845	0.7727	0.7895	0.7493	<b>0.9086</b>
Evidence	0.8914	0.7245	0.7115	0.7038	<b>0.9164</b>

Table 2: Detection Explanation Comparison

misinformation detection.

**Detection Explanation Comparison** We further evaluate the explanation quality of misinformation detection produced by our SynthX framework in comparison to all baseline models capable of generating textual explanations, as detailed in the Baselines and Evaluation Settings Subsection. The evaluation results presented in Table 2 show that SynthX consistently outperforms the baseline models. For instance, compared to the best-performing baseline, GPT-4, SynthX achieves notable performance improvements of 2.72% in Factuality and 2.80% in Evidence. The superior performance of SynthX can be attributed to its ability to synthesize insights from multiple intelligence sources, such as crowd workers and LLMs, combining their respective strengths. By integrating these complementary sources, SynthX produces explanations that are not only more factual but also better grounded in evidence.

We present the explanation results generated by each type of intelligence integrated into SynthX in Table 3, using an example tweet classified under the conspiracy category. Crowd-generated explanation was overly brief, lacking the depth and analytical rigor necessary to effectively debunk misinformation and ensure clarity for online users. GPT-4, despite offering detailed analyses, misclassified the misinformation by focusing solely on identifying falsehoods while overlooking the key of conspiracy theories, and provided an explanation that was relatively repetitive in expression. In contrast, SynthX synthesized the strengths of these complementary approaches: it correctly identified the conspiracy category, balanced clarity with sufficient detail, and maintained concise language to enhance readability and user comprehension. A detailed comparison across all baselines for this tweet is provided in the Appendix.

Tweet	<b>True Category: Conspiracy</b> <i>#CoronaVirusSA #covid19sa #bioweapon Based on current epidemiological mathematical models and govt's inability to fix anything, SA would by mid-June have 100,000 infected persons, 3,000 dead, prison riots, economic collapse, countrywide riots.</i>
Crowd	<b>Detection Category: Conspiracy</b> This tweet is spreading conspiracy about the aftermath of coronavirus.
GPT-4	<b>Detection Category: False Information</b> The tweet presents a series of alarming predictions regarding the COVID-19 situation in South Africa that are not substantiated by credible data or evidence. While it references mathematical models, it does not provide any sources or context for these claims. The use of hashtags like #bioweapon suggests a conspiratorial framing, but the primary content of the tweet focuses on specific predictions that are exaggerated and unfounded. The predictions about infection rates, deaths, and societal collapse are alarmist and not reflective of the actual situation as understood by public health experts. Therefore, it falls under the category of false information rather than conspiracy or satire.
SynthX	<b>Detection Category: Conspiracy</b> The tweet presents a series of dire predictions regarding the COVID-19 situation in South Africa, framing them in a sensational manner that suggests a hidden agenda or malicious intent, particularly with the use of hashtags like #bioweapon. While it references mathematical models, it lacks credible evidence or data to support its claims. The alarmist tone and the speculative nature of the predictions contribute to a conspiratorial narrative rather than factual reporting, making it fit conspiracy category.

Table 3: Detection Explanation Example Comparison

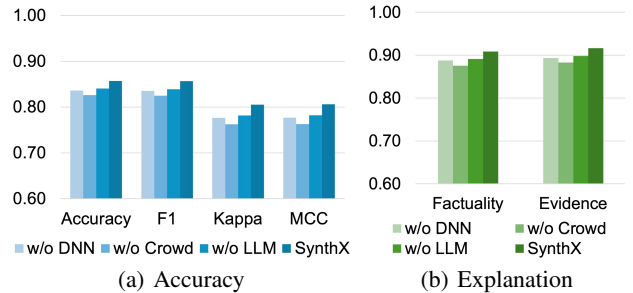


Figure 3: Ablation Study of SynthX

**Ablation Study** We conducted an ablation study to assess the contribution of each type of intelligence in our framework, aiming to validate the effectiveness of our collective intelligence design. Specifically, we removed each of the following: DNNs, crowd workers, and the LLM from our synthesis framework, resulting in the configurations *w/o DNN*, *w/o Crowd*, and *w/o LLM*, respectively. The evaluation results are shown in Figure 3. We observe that the exclusion of any one type of intelligence leads to a noticeable decline in both detection accuracy and explanation quality. These

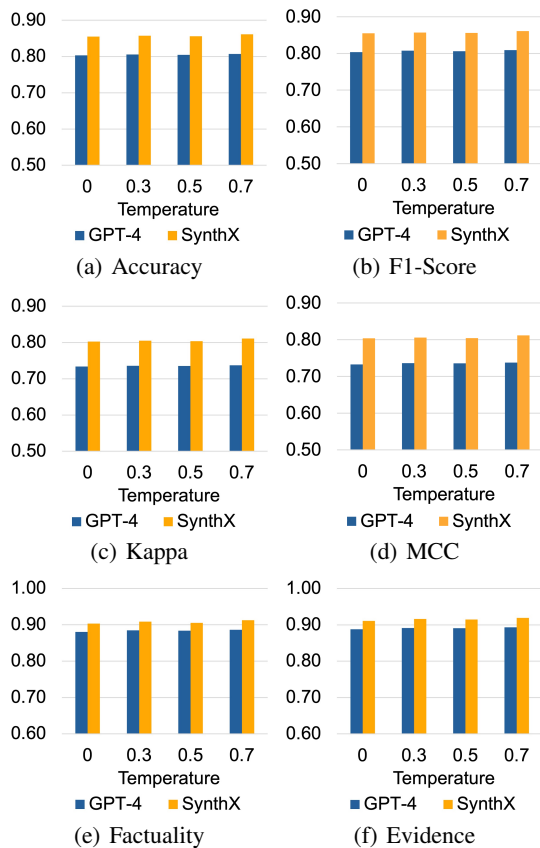


Figure 4: Robustness Study of SynthX

findings highlight the effectiveness of our design in harnessing the complementary strengths of DNNs, crowdsourcing, and LLMs. The results provide insights on the potential of integrating LLMs with traditional crowdsourcing or human–AI collaboration frameworks for enhanced collective intelligence. Additionally, we observe that removing the crowd component results in the largest drop in detection accuracy. This is because, while individual crowd workers may not achieve the same level of accuracy as AI models, the collective accuracy derived from aggregating multiple crowd workers notably enhances the model’s performance.

**Robustness Study** In the final set of experiments, we conducted a robustness study to assess the performance stability of our framework when varying the temperature of the GPT-4 model, which serves as the LLM component in our framework. Model temperature controls the randomness of the output, with lower values making predictions more deterministic and higher values introducing more variability and creativity. We explored temperature settings of 0, 0.3, 0.5, and 0.7 to investigate the trade-offs between output diversity and consistency. We compared the performance of our SynthX framework with GPT-4 across different temperature settings, as GPT-4’s performance is also changing based on this hyperparameter. The evaluation results, depicted in Figure 4, indicate that SynthX consistently outperforms GPT-4

across all temperature settings. This consistent performance gain demonstrates the robustness of our SynthX framework, maintaining high accuracy and explanation quality as the LLM’s temperature varies.

## Conclusion

We introduced SynthX, a novel collective intelligence framework designed for explainable misinformation detection. We developed a hybrid system that estimates both the reliability of DNN, crowd worker, and LLM sources and the quality of explanations provided by crowd workers and LLMs. We developed an innovative framework that transcends conventional label aggregation by synthesizing heterogeneous detection and explanation outcomes. It incorporates a hybrid reliability estimation method that assesses both source reliability and explanation quality, ensuring accurate and explainable detection. Experiments on a real-world social media dataset proved that SynthX consistently outperforms state-of-the-art baselines in both detection accuracy and explanation quality. This paper highlights the potential of using LLMs to synthesize collective intelligence beyond traditional label aggregation, accommodating complex and diverse textual information, such as explanations, opinions, and ethical considerations.

## Acknowledgements

This research is supported in part by the National Science Foundation under Grant No. CNS-2427070, IIS-2331069, IIS-2202481, IIS-2130263, CNS-2131622. 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

- Artstein, R.; and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4): 555–596.
- Athira, A.; Kumar, S. M.; and Chacko, A. M. 2023. A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122: 106087.
- Chen, C.; and Shu, K. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.
- Chen, C.; Wang, H.; Shapiro, M.; Xiao, Y.; Wang, F.; and Shu, K. 2022. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *arXiv preprint arXiv:2211.05289*.
- Chien, S.-Y.; Yang, C.-J.; and Yu, F. 2022. XFlag: Explainable fake news detection model on social media. *International Journal of Human–Computer Interaction*, 38(18-20): 1808–1827.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

- De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G. P.; Ferragina, P.; Tozzi, A. E.; and Rizzo, C. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in public health*, 11: 1166120.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.
- Diab, A.; Nefriana, R.; and Lin, Y.-R. 2024. Classifying Conspiratorial Narratives at Scale: False Alarms and Erroneous Connections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 340–353.
- Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 183–193.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Grieblhaber, D.; Maucher, J.; and Vu, N. T. 2020. Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1158–1171.
- He, B.; Ahamad, M.; and Kumar, S. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, 2698–2709.
- He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–25.
- Hryniewska, W.; et al. 2024. EnsembleXAI: Simplifying aggregation of explanations, normalization, and assessment for enhanced model explainability. Accessed: 2024-10-13.
- Islam, M. R.; Liu, S.; Wang, X.; and Xu, G. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1): 82.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jiang, B.; Tan, Z.; Nirmal, A.; and Liu, H. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 427–435. SIAM.
- Kim, H.; Mitra, K.; Chen, R. L.; Rahman, S.; and Zhang, D. 2024. MEGAnno+: A Human-LLM Collaborative Annotation System. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 168–176.
- Kim, J.; Tabibian, B.; Oh, A.; Schölkopf, B.; and Gomez-Rodriguez, M. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 324–332.
- Kumar, R.; and Bhat, A. 2021. An analysis on sarcasm detection over twitter during COVID-19. In *2021 2nd international conference for emerging technology (INCET)*, 1–6. IEEE.
- Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- McCain, K. 2015. Explanation and the nature of scientific knowledge. *Science & Education*, 24(7): 827–854.
- Memon, S. A.; and Carley, K. M. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.
- Mohseni, S.; Yang, F.; Pentylala, S.; Du, M.; Liu, Y.; Lupfer, N.; Hu, X.; Ji, S.; and Ragan, E. 2021. Machine learning explanations to prevent overtrust in fake news detection. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 421–431.
- Ni, S.; Li, J.; and Kao, H.-Y. 2021. MVAN: Multi-view attention networks for fake news detection on social media. *IEEE Access*, 9: 106907–106917.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774*.
- Ostyakova, L.; Smilga, V.; Petukhova, K.; Molchanova, M.; and Kornev, D. 2023. Chatgpt vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 242–254.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Papanikou, V.; Papadakos, P.; Karamanidou, T.; Stavropoulos, T. G.; Pitoura, E.; and Tsaparas, P. 2024. Health Misinformation in Social Networks: A Survey of IT Approaches. *arXiv preprint arXiv:2410.18670*.
- Paraschiv, M.; Salamanos, N.; Iordanou, C.; Laoutaris, N.; and Sirivianos, M. 2022. A unified graph-based approach to disinformation detection using contextual and semantic relations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 747–758.
- Rezaei, S.; Kahani, M.; Behkamal, B.; and Jalayer, A. 2022. Early multi-class ensemble-based fake news detection using content features. *Social Network Analysis and Mining*, 13(1): 16.
- Shang, L.; Zhang, Y.; Chen, B.; Zong, R.; Yue, Z.; Zeng, H.; Wei, N.; and Wang, D. 2024. MMAdapt: A Knowledge-guided Multi-source Multi-class Domain Adaptive Framework for Early Health Misinformation Detection. In *Proceedings of the ACM on Web Conference 2024*, 4653–4663.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, D.; Kaplan, L.; Le, H.; and Abdelzaher, T. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 233–244.

Wang, e. a. 2023. GPTScore: Evaluate as You Desire. *arXiv preprint arXiv:2302.04166*.

Xu, J.; Han, L.; Sadiq, S.; and Demartini, G. 2024. On the Role of Large Language Models in Crowdsourcing Misinformation Assessment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1674–1686.

Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. 2023. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3(3): 161–188.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2020. XLNet: Generalized Autoregressive Pre-training for Language Understanding. *arXiv:1906.08237*.

Yenduri, G.; Ramalingam, M.; Selvi, G. C.; Supriya, Y.; Srivastava, G.; Maddikunta, P. K. R.; Raj, G. D.; Jhaveri, R. H.; Prabadevi, B.; Wang, W.; et al. 2024. Gpt-a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*.

Zhan, X.; Wang, Q.; Huang, K.-h.; Xiong, H.; Dou, D.; and Chan, A. B. 2022. A comparative survey of deep active learning. *arXiv preprint arXiv:2203.13450*.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. I. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*, 27.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.

Zhang, Y.; Shang, L.; Zong, R.; Wang, Z.; Kou, Z.; and Wang, D. 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*, volume 9, 179–190.

Zong, R.; Zhang, Y.; Shang, L.; and Wang, D. 2023. Contrastfaux: Sparse semi-supervised fauxtography detection on the web using multi-view contrastive learning. In *Proceedings of the ACM Web Conference 2023*, 3994–4003.

## Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, please see the Introduction.**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes, please see the Abstract and Introduction.**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, please see the Solution.**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, please see the Evaluation.**
- (e) Did you describe the limitations of your work? **Yes, please see the Ethics Statement.**
- (f) Did you discuss any potential negative societal impacts of your work? **Yes, please see the Ethics Statement.**
- (g) Did you discuss any potential misuse of your work? **Yes, please see the Ethics Statement.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, please see the Ethics Statement.**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, we have read the ethics review guidelines to ensure our paper conforms to them.**

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
- (b) Have you provided justifications for all theoretical results? **NA**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
- (e) Did you address potential biases or limitations in your theoretical framework? **NA**
- (f) Have you related your theoretical results to the existing literature in social science? **NA**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? **NA**
- (b) Did you include complete proofs of all theoretical results? **NA**

4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes, please see the Evaluation and Appendix.](#)
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes, please see the Evaluation and Appendix.](#)
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes, please see the Appendix.](#)
  - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes, please see the Evaluation.](#)
  - (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? [Yes, please see the Ethics Statement.](#)
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? [Yes, please see the Evaluation.](#)
  - (b) Did you mention the license of the assets? [Yes, please see the Evaluation.](#)
  - (c) Did you include any new assets in the supplemental material or as a URL? NA
  - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes, please see the Evaluation.](#)
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes, please see the Evaluation and Ethics Statement.](#)
  - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
  - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? [Yes, please see the Appendix.](#)
  - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [Yes, please see the Appendix.](#)
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes, please see the Appendix.](#)
  - (d) Did you discuss how data is stored, shared, and de-identified? [Yes, please see the Ethics Statement.](#)

## Ethics Statement

This paper focuses on explainable misinformation detection on social media, a critical task with significant societal implications for preventing the spread of false information and misconceptions. To avoid potential misuse, we employed rigorous data anonymization, ensuring that all datasets used are publicly available, ethically sourced, and contain only post content without any personally identifiable information. Evaluation settings are comprehensively documented to facilitate reproducibility while maintaining controlled access. Misclassification in this domain carries the potential to either amplify misinformation or unnecessarily flag accurate information. We incorporate hybrid intelligence and reliability assessments to enhance fault tolerance.

## Appendix

### Summary of SynthX Framework

We summarize our SynthX framework in Algorithm 1. The inputs include social media posts  $P$  and different types of sources  $s \in \{DNN_a, Crowd_b, LLM_1 | a = 1, \dots, A; b = 1, \dots, B\}$ . The framework synthesizes these inputs to produce the explainable misinformation detection outputs: the synthesized categories  $\hat{C}$  and explanations  $\hat{E}$ . The synthesis process is designed to leverage the complementary strengths of different types of intelligence for more accurate and explainable results.

---

#### Algorithm 1: Summary of SynthX Framework

```
1: input:  $P, s \in \{DNN_a, Crowd_b, LLM_1 | a = 1, \dots, A; b = 1, \dots, B\}$ 
2: output:  $\hat{C}, \hat{E}$ 
3: for each post  $P_i$  do
4:   for each source  $s \in \{DNN_a | b = 1, \dots, B\}$  do
5:     predict category  $\hat{C}_i^s$ 
6:   end for
7:   for each source  $s \in \{Crowd_b, LLM_1 | b = 1, \dots, B\}$  do
8:     collect category  $\hat{C}_i^s$  and explanation  $\hat{E}_i^s$ 
9:   end for
10: end for
11: for each post  $P_i$  do
12:   update source reliability  $R^s$  by EM,
   where  $s \in \{DNN_a, Crowd_b, LLM_1 | a = 1, \dots, A; b = 1, \dots, B\}$ 
13:   compute explanation quality  $Q_i^s$  by LLM,
   where  $s \in \{Crowd_b, LLM_1 | b = 1, \dots, B\}$ 
14: end for
15: for each post  $P_i$  do
16:   generate  $\hat{C}_i$  from  $\hat{C}_i^s, \hat{R}^s$  by LLM,
   where  $s \in \{DNN_a, Crowd_b, LLM_1 | a = 1, \dots, A; b = 1, \dots, B\}$ 
17:   generate  $\hat{E}_i$  from  $\hat{E}_i^s, \hat{R}^s, \hat{Q}_i^s$  by LLM,
   where  $s \in \{Crowd_b, LLM_1 | b = 1, \dots, B\}$ 
18: end for
```

---

### Baselines

To thoroughly evaluate our SynthX approach, we include a diverse set of DNN, LLM, and human–AI baselines in the assessment process:

- **BERT** (Devlin et al. 2019): a powerful representation learning model that leverages a bidirectional transformer architecture to process language context in both directions, allowing it to deliver high classification accuracy across various complex natural language tasks.
- **RoBERTa** (Liu 2019): an enhanced deep learning model that employs a refined transformer network to deeply capture and learn semantic representations from text, improving classification performance through extensive pretraining and sophisticated fine-tuning techniques.
- **XLNet** (Yang et al. 2020): a representative language model using a permutation based transformer architecture to capture bidirectional context, offering enhanced performance across various language understanding and classification tasks.

- **GPT-4** (OpenAI 2023): a state-of-the-art large language model utilizing a transformer architecture, fine-tuned with human feedback to generate highly coherent and contextually aware text, excelling in diverse language tasks.
- **LLaMA3** (Touvron et al. 2023): an open-source large language model designed for efficiency and scalability, leveraging advanced transformer architectures to perform competitively across a wide range of language tasks.
- **Mistral** (Jiang et al. 2023): a highly efficient foundation language model that builds on transformer-based architectures, emphasizing compactness and performance, and delivering strong results across diverse language benchmarks.
- **TD** (Wang et al. 2012): an expectation maximization-based crowdsourcing algorithm that evaluates worker expertise by modeling each worker’s reliability, and iteratively refines both misinformation category estimates and worker accuracy through its two-step process.
- **DAL** (Zhan et al. 2022): a deep active learning-based crowd–AI collaborative approach that integrates crowd-sourced labels with AI predictions for samples exhibiting high prediction entropy to enhance overall classification accuracy.
- **StreamCollab** (Zhang et al. 2021): a human–AI collaborative framework that estimates AI uncertainty to detect and correct potentially incorrect AI predictions via human intelligence using estimation theoretical approaches to achieve optimal collective classification performance.
- **MEGAnno+** (Kim et al. 2024): a human–LLM collaborative annotation system that combines LLM efficiency with human contextual understanding to produce high-quality classification results.

### Qualitative Analysis of Detection Explanation

We present an example comparing the misinformation explanation results of SynthX with all baseline methods in Table 4. The example tweet promotes a conspiracy theory about South Africa’s COVID-19 spread predictions, lacking any evidence to support its claims. Crowd-generated explanations tend to be overly brief, lacking sufficient detail and analysis to ensure reliability and clarity for debunking misinformation when presented to online users. Both GPT-4 and LLaMA incorrectly categorized the misinformation, focusing only on identifying false information without recognizing the underlying conspiracy of exaggerating infections and consequences to spread fear without evidence. While Mistral provided an accurate category and explanation for this specific example, its performance was less robust and consistent overall, which explains its lower scores in Tables 1 and 2. In contrast, our synthesized detection explanation identified the correct category and provided a significantly improved explanation. It balanced clarity and detail while maintaining concise language and enhancing readability for better user comprehension.

Tweet and Label	Algorithm	Detection Category	Detection Explanation
<p>#CoronaVirusSA #covid19sa #bioweapon</p> <p>Based on current epidemiological mathematical models and govt's inability to fix anything, SA would by mid-june have</p> <ul style="list-style-type: none"> <li>- 100 000 infected persons</li> <li>- 3000 dead</li> <li>- prison riots</li> <li>- economic collapse</li> <li>- countrywide riots</li> </ul> <p><b>True Category: Conspiracy</b></p>	Crowd	Conspiracy	This tweet is spreading conspiracy about the aftermath of coronavirus.
	GPT-4	False Information	The tweet presents a series of alarming predictions regarding the COVID-19 situation in South Africa that are not substantiated by credible data or evidence. While it references mathematical models, it does not provide any sources or context for these claims. The use of hashtags like #bioweapon suggests a conspiratorial framing, but the primary content of the tweet focuses on specific predictions that are exaggerated and unfounded. The predictions about infection rates, deaths, and societal collapse are alarmist and not reflective of the actual situation as understood by public health experts. Therefore, it falls under the category of false information rather than conspiracy or satire.
	LLaMA3	False Information	The tweet makes specific predictions about the number of infected persons and deaths due to COVID-19 in South Africa by mid-June, which is a claim that can be verified or disproven with actual data. Although it's possible that the author was basing their predictions on available epidemiological models at the time, without knowing the exact date of the tweet and comparing it with actual data, it's difficult to confirm its accuracy. However, given that such predictions often turn out to be incorrect or exaggerated, I lean towards categorizing it as False Information. It's worth noting that while some parts of the tweet might have been based on real concerns (e.g., economic collapse), others seem speculative or alarmist (e.g., prison riots and countrywide riots). This mix of potentially valid concerns with unsubstantiated claims further supports my classification as False Information rather than True Information.
	Mistral	Conspiracy	The user is suggesting that the COVID-19 outbreak in South Africa was deliberately engineered as a bioweapon (an unproven claim), and then predicting specific outcomes based on this assumption (e.g., high infection rates, deaths, prison riots, economic collapse, countrywide riots). These predictions are not supported by empirical evidence or accepted scientific models for disease spread. Furthermore, the use of hashtags like "#CoronaVirusSA" and "#covid19sa" implies that the user is discussing a real event but is presenting an unfounded interpretation or theory about it.
	SynthX	Conspiracy	The tweet presents a series of dire predictions regarding the COVID-19 situation in South Africa, framing them in a sensational manner that suggests a hidden agenda or malicious intent, particularly with the use of hashtags like #bioweapon. While it references mathematical models, it lacks credible evidence or data to support its claims. The alarmist tone and the speculative nature of the predictions contribute to a conspiratorial narrative rather than factual reporting, making it fit the category of conspiracy.

Table 4: Qualitative Comparison of Detection Explanation