RESEARCH-ARTICLE

Detecting Health Misinformation by Leveraging LLM Models and Debunk List

MELIKA ROSTAMI, University of North Texas, Denton, TX, United States

K S HOSSAIN, University of North Texas, Denton, TX, United States

SULIMAN HAWAMDEH, University of North Texas, Denton, TX, United States

Open Access Support provided by:

University of North Texas

Published: 24 June 2025

Citation in BibTeX format

CHASE '25: ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies
June 24 - 26, 2025
NY, New York, USA

Conference Sponsors:
SIGBED

CHASE '25: Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (June 2025)
https://doi.org/10.1145/3721201.3725518
ISBN: 9798400715396

Detecting Health Misinformation by Leveraging LLM Models and Debunk List

Melika Rostami
MelikaRostami@my.unt.edu
Department of Information Science
University of North Texas
Denton, Texas, USA

K S M Tozammel Hossain
Tozammel.Hossain@unt.edu
Department of Information Science
University of North Texas
Denton, Texas, USA

Suliman Hawamdeh
Suliman.Hawamdeh@unt.edu
Department of Information Science
University of North Texas
Denton, Texas, USA

Abstract

Timely and accurate detection of information is essential for ensuring public health safety. COVID-19 pandemic highlighted challenges related to the spread of misinformation, which caused vaccine hesitancy, reduced trust in health authorities, and promoted the use of unapproved treatments. Several studies have used machine learning models to detect health misinformation on social media. However, most of these models rely heavily on large, labeled datasets, which are time-consuming, costly, and challenging to create during health crises. Large language models, like GPT, present a potential alternative to these solutions due to their generalizability and less dependency on annotated datasets. This study investigates the feasibility of GPT models, such as GPT-3.5-turbo, GPT-4o-mini, and GPT-4o, as well as BERT-based models, in detecting misinformation. The study assesses model performance by applying zero-shot and few-shot approaches using four different scenarios: 1) without a debunk list related to COVID-19; 2) with an internal COVID-19 debunk list derived from the Constraint dataset; 3) with an external debunk list collected from credible and authoritative sources; and 4) with a combination of internal and external debunk lists. The results show that GPT models consistently outperformed BERT-based models across different scenarios. Although BERT performance improves with domain-specific fine-tuning, it performs poorly when fine-tuned with an external debunk list. These findings highlight the adaptability of GPTs in detecting misinformation with minimal guidance, showing their potential to mitigate the risk of misinformation in real-world applications.

CCS Concepts

• **Computing methodologies → Artificial intelligence**; • **Information systems → Information retrieval**; • **Human-centered computing → Social media**.

Keywords

Health Misinformation Detection, Large Language Models (LLMs), BERT, Debunk List, COVID-19

CHASE '25, June 24–26, 2025, New York, NY, USA

ACM ISBN 979-8-4007-1539-6/2025/06
https://doi.org/10.1145/3721201.3725518

ACM Reference Format:
Melika Rostami, K S M Tozammel Hossain, and Suliman Hawamdeh. 2025. Detecting Health Misinformation by Leveraging LLM Models and Debunk List. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3721201.3725518

1 Introduction

The COVID-19 infodemic caused various challenges to public health, including delaying or avoiding COVID vaccination [1], losing individual trust in the health authorities [2], and even choosing unapproved treatments and medications that led into deaths [3]. Infodemic refers to the abundance of false and misleading information during a disease crisis [4]. Misinformation during an infodemic can spread across various media with social media platforms (SMPs) being particularly prone to propagating it widely and rapidly [5]. Social media platforms are user-friendly online resources for searching, distributing, and receiving health information [6, 7]. During COVID-19, these activities were highly prevalent, as individuals used online platforms as primary sources for seeking information about the virus [8].

In recent years, several studies have developed various models to automate misinformation detection [9, 10, 11]. Transformer-based models like BERT [12], a popular model, have been used in various studies for detecting COVID-19 misinformation [13, 14, 15]. Current misinformation detection solutions may not be applicable to situations like the novel pandemic, which requires rapid and accurate detection of misinformation. Machine learning models do not always generalize well, and their performance depends heavily on an annotated dataset. Manually annotating datasets is not efficient since it is costly and requires human labor [16]. To resolve these challenges, recent research has investigated the capabilities of large language models (LLMs) in text annotations [17, 18, 19]. LLMs are more efficient than human labeling since they can annotate data quickly and cheaply [16].

LLMs has further improved the capabilities of misinformation detection. These models perform well in few-shot learning, allowing them to have an effective performance with limited data [20], which is a valuable feature in a fast-paced environment of misinformation. Also, LLMs are capable of automatically applying feature engineering for few-shot tabular learning, which shows their skills in handling minimal data scenarios [21]. Previous research has investigated the effectiveness of LLMs in detecting misinformation [22, 23, 24, 25, 26, 27]. Chen and Shu [2024] found that LLM generated misinformation is more challenging to detect than human written misinformation [24]. Leite et al. [2023] showed that LLMs can effectively leverage credibility signals to achieve high performance in detecting misinformation tasks even in weakly supervised settings [28]. Furthermore, Wan et al. [2024] proposed a framework named DELL that integrates LLM to generate explanations for misinformation detection, helping to build trust in LLM classifications [29]. These studies collectively highlight the potential of LLMs in identifying and mitigating the risk of misinformation. Our study differs from these by focusing on how a debunk list can help improve model performance.

Debunk lists, which include verified authentic and false claims, have gained attention as an effective method to improve misinformation detection. Ng and Carley [2021] analyzed and categorized various fact-checked stories, collected from PolitiFact, Poynter and Snopes, into six clusters to better understand misinformation trends [30]. Shahi and Nandini [2020] created a multilingual annotated dataset collected from 92 fact-checking websites for training classifiers in detecting COVID-19 misinformation [31]. Due to challenges in label annotation, LLMs have the potential to serve as an automated, cost-effective alternative model for detecting misinformation. In this study, we aim to address this gap by evaluating the performance of GPT models, using BERT as a baseline, with different debunk configurations in both zero-shot and few-shot approaches, to explore scalable misinformation detection solutions for future health crises.

The main focus of this paper is to investigate the capabilities of LLMs, especially GPT models [32], including GPT-4o, GPT-4o-mini, GPT-3.5-turbo, in enhancing the detection of COVID-19 misinformation. In this paper, we address the following research questions:

RQ1: How do GPT models perform compared to BERT-based models in detecting COVID-19 misinformation across zero-shot and few-shot approaches?

RQ2: What is the impact of the configuration of debunk lists (derived internally or externally) on the performance of GPTs and BERT models?

RQ3: Does applying a debunk list with GPTs provide a cost effective alternative to the traditional human annotation process for detecting misinformation, especially during novel pandemic?

2 Methods

Different models, including BERT and LLMs, are evaluated for their applicability and performance in detecting COVID-19 misinformation. Two different approaches are used in the implementation of these models: the zero-shot approach and few-shot approach. These approaches are applied to evaluate the models' performance across four different scenarios as shown in Figure 1 and discussed below.

Figure 1 provides a comprehensive illustration of the model configuration and evaluation process for each approach. The top-left section (Fig. 1a) outlines the setup of the GPT models, including generating API requests and evaluating model responses under zero-shot (No Debunk List, Scenario 1) and a few-shot (With Debunk List, Scenarios 2-4) conditions. The debunk list used for both BERT and GPT frameworks are described in section 2.2. The bottom section (Fig. 1b) displays the BERT model framework, including text tokenization, fine-tuning, and validation. In Scenario 1, the base pre-trained BERT model is used without further training, whereas in Scenarios 2–4, the pre-trained BERT model is fine-tuned through three debunk lists. The top-right section (Fig. 1c) illustrates the various metrics used to evaluate model performance through all scenarios.

Table 1: Summary of the dataset used for COVID-19 misinformation detection.

Dataset	Real	Fake	Total
Constraint Dataset	5,600	5,100	10,700
Test Dataset	1,000	1,000	2,000

2.1 Datasets

We evaluate the proposed method using real-world datasets. For testing the models, we use a subset of the Constraint [33] dataset. This dataset includes 10,700 annotated COVID-19 posts and articles. Table 1 summarizes the Constraint dataset and the subset we use for this study. We randomly select 2,000 instances (1,000 authentic and 1,000 fake) out of these 10,700 posts to create an unbiased Test Dataset. This decision was made to keep the computational cost related to querying the LLMs model more manageable. Focusing on this subset ensures an unbiased and diverse dataset for a practical experiment. The processed data is fed into BERT and GPT models to evaluate their performance under zero-shot and few-shot approaches (see Fig. 2).

2.2 Debunk List

We develop debunk lists covering four different scenarios to evaluate the performance of the models in detecting COVID-19 misinformation (see Table 2). Each scenario is designed to test the performance of models when receiving zero-shot and few-shot support from internal or external information.

No Debunk List: The models are tested without additional information to obtain their baseline performance for Scenario 1.

Debunked List A (Internal Source): We randomly select 50 texts (25 real, 25 fake) from the Constraints dataset, ensuring that this selection is not part of 2,000 subsets used for test dataset. This scenario evaluates the model's performance when exposed to related (internal) label information.

Debunk List B (External Source): We collect 100 claims, labeled as true or false, from reputable sources (external), including WHO, CDC, Johns Hopkins Medicine, and Mayo Clinic. Then, we randomly select 50 facts (25 true, 25 false) to apply in Scenario 3.

Debunk List C (Combination of Internal and External Sources): For Scenario 4, we combine 50 texts (25 true, 25 false) from the Constraint dataset (excluding test dataset) along with 50 facts (25 true, 25 false) collected from reputable sources. We are interested in observing the models' performance when exposed to a combination of internal and external information.

2.3 Zero-shot Approach

This approach does not use a debunk to fine-tune the models. This allows us to capture the models' performance as a baseline. The pre-trained BERT model, trained only on its original dataset, was used without further fine-tuning. The BERT architecture consists of a dropout layer to mitigate overfitting problems and a binary classifier

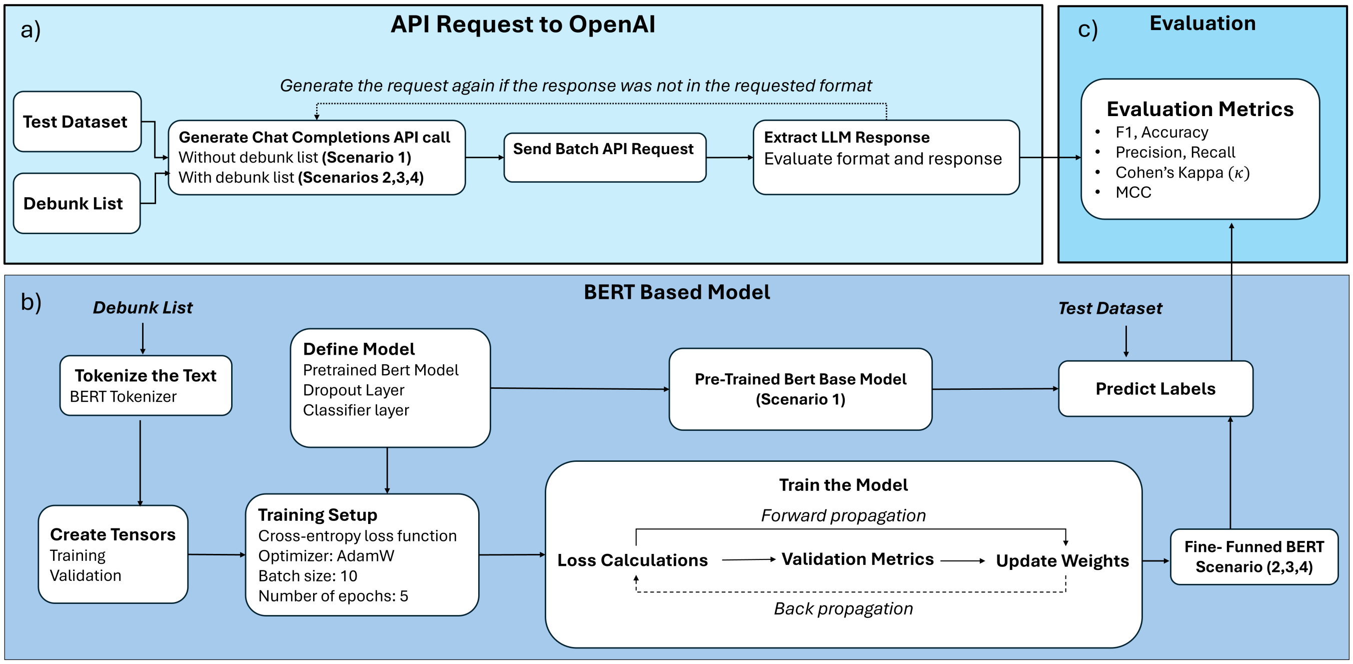

Figure 1: Model processing pipeline for misinformation detection using GPT-based models (via OpenAI) and BERT model. The top left section (a) shows GPT model setup that includes generating API requests, extracting, formatting and evaluating the GPT responses for both approaches: a zero-shot approach (without debunk list: scenario 1) and a few-shot approach (with debunk list: scenarios 2-4). The bottom section (b) shows the BERT model setup including text tokenization, fine-tuning and validation. The base pre-trained BERT model is used for scenario 1 without further training, while fine-tuning is applied with the debunk list for scenarios 2-4. The top right section (c) shows evaluation metrics that are used to assess model performance across all scenarios.

Table 2: Summary of debunk list setup for four scenarios.

Scenario	Source	Number of Texts
1. No Debunk List	None	0
2. Debunk List A	Derived from Constraints dataset	50
3. Debunk List B	Collected from Johns Hopkins Medicine, Mayo Clinic, CDC, WHO	50
4. Debunk List C	Combination of Debunk List A and Debunk List B	100

layer for labeling the test data as 1 or 0 (authentic vs. fake). Similarly, we evaluate the performance of GPT models without providing the debunk list as guidance to the models. For this study, we employ three versions of the GPT models: GPT-3.5-turbo, GPT-4o-mini, and GPT-4o. We then provide the test data to each model separately and capture their classification as a baseline. This setup allows us to evaluate the performance of BERT-based and GPT models in classifying COVID-19 misinformation without any additional context. This also provides a fair comparison between pre-trained BERT and GPT models as the baselines. Below is the prompt used for the zero-shot approach.

Zero-shot Prompt: Use your general knowledge to classify each text below as either misinformation (0) or truthful (1). You must return the response in the following format: 0 or 1 – Explanation. Classify the following texts: {test_dat}

2.4 Few-shot Approach

This approach uses a debunk list as minimal guidance to observe models' performance when exposes to the supplementary context. To make a fair comparison, we fine-tune the BERT model using the same debunk list we provide to the GPT models (See Section 2.2). This fine-tuning process enables BERT to adjust its parameters based on the additional information from the debunk list. The architecture of the fine-tuned BERT model remained the same as the pre-trained BERT model, including the dropout and classifier layers, with only one difference of utilizing the debunk list to improve model performance. The same debunk list that is used for fine-tuning the BERT model is provided to the GPT models for the few-shot approach. The debunk list is fed to GPT models as a supplementary context to classify the test data as authentic or fake information. We use the same API configuration across all

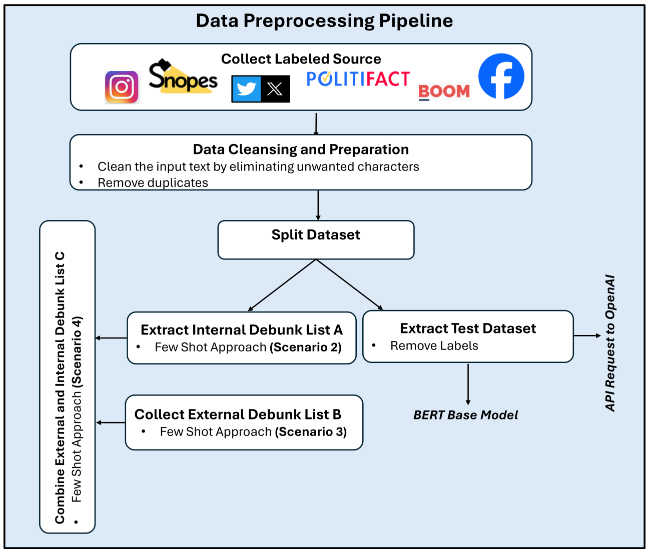
Data Preprocessing Pipeline
Collect Labeled Source
Snopes
POLITIFACT
BOOM
Data Cleansing and Preparation
• Clean the input text by eliminating unwanted characters
• Remove duplicates
Split Dataset
Extract Internal Debunk List A
• Few Shot Approach (Scenario 2)
Extract Test Dataset
• Remove Labels
Collect External Debunk List B
• Few Shot Approach (Scenario 3)
BERT Base Model
API Request to OpenAI
Combine External and Internal Debunk List C
• Few Shot Approach (Scenario 4)

Figure 2: Data preprocessing pipeline for misinformation detection. This pipeline shows the steps for collecting, cleansing and splitting the labeled data for model evaluation. The dataset is divided into 4 groups: 1) A test dataset for unbiased model evaluation, 2) Debunk List A, an internal debunk list for scenario 2, 3) Debunk List B, an external debunk list collected from reputable sources for scenario 3, 4) Debunk List C, a combination of internal and external debunk lists for scenario 4. The processed data is then fed into both BERT and GPT models to assess their performance under zero-shot and few-shot approach.

GPT models to ensure consistency throughout the experiment. Our goal is to evaluate whether the few-shot approaches improve the performance of GPTs compared to the fine-tuned BERT model. Below is the prompt used for the few-shot approach.

Few-shot Prompt: Classify each text below as either misinformation (0) or truthful (1). Use the debunk list as a reference when applicable. If the debunk list does not provide useful guidance, use your general knowledge. You must return the response in the following format: 0 or 1 – Explanation. Debunk List: {debunk_list}. Classify the following texts: {test_data}

2.5 Software and Tools

We use Python to execute experiments in our study. Hugging Face Transformers library is used for loading, fine tuning and testing the BERT model. We access the GPT models (GPT-3.5-turbo, GPT-4.o-mini, and GPT-4.o) through the OpenAI API platform [32].

3 Results

We evaluate the performance of BERT-based models and GPT models over four scenarios (see Section 2.2). We apply various evaluation metrics, including accuracy , precision, recall, F1 score, Cohen's Kappa (κ), and Mathew correlation coefficient (MCC) to provide a complete assessment of the models' performance [34, 35]. The goal of Scenario 1 is to establish a baseline for the models without any supplementary debunk list. As shown in Table 3: Scenario 1, the pre-trained BERT model does not perform well without additional context, which is expected. The GPT models perform better than pre-trained BERT without a debunk list. All these GPT models perform almost equally well in all measures.

In Scenario 2, Debunk List A is used as the training dataset to fine tune the pre-trained BERT model and as a few-shot approach to provide additional information to GPT models. Both BERT and GPT models show performance improvement with this debunked list (see Table 3: Scenario 2). When BERT is trained with an internal debunk list, the results show that it can be competitive with GPT models. The fine-tuned BERT model achieves accuracy of 0.81 and F1 score of 0.81, indicating a significant improvement compared to its baseline; Chone's Kappa (κ), and MCC values also increase to 0.62. The GPT models similarly show performance improvement with the debunk list. GPT-3.5-turbo achieves accuracy of 0.84, F1 score of 0.81 and Cohen's Kappa and MCC values of 0.67 and 0.69 respectively. GPT-4o-mini and GPT-4o reach a similar accuracy level of 0.83 and 0.84 respectively. GPT-4o achieves a slightly higher precision of 0.98 but a lower recall of 0.69. This scenario shows that providing minimal context from the dataset can improve models' performance in both BERT and GPT. The internal debunk list notably improved BERT performance in almost all metrics, compared to Scenario 1 (see Table 3). While GPT models also show improvement in their performance with the internal debunk list, this is less notable compared to BERT. This reflects that GPT models have stronger baseline capabilities even without additional context. Also, this indicates that minimal contextual guidance is specifically beneficial for models like BERT that rely heavily on domain-specific information.

In Scenario 3, we use Debunk List B which is derived from external sources. While the performance of the fine-tuned BERT model improves compared to the base model (pre-trained) in Scenario 1, its performance is notably lower than Scenario 2 where debunk list was captured from the same data distribution. Similarly, the GPT models show a slight decrease in performance compared to Scenario 2, although their overall performance stayed higher than BERT. The more notable drop in BERT's performance can contribute to its reliance on domain- specific data for fine- tuning. The external debunk list in Scenario 3, which has different data distribution and linguistic patterns compared to the test dataset, could have caused more challenges for BERT to generalize effectively. On the other hand, GPT models, which pre-trained on a larger range of data, show more resistance to this distribution shift (see Table 3, Scenario 3).

In Scenario 4, Debunk List C is used which is a combination of internal information and external information. The BERT model shows improved performance compared to Scenario 1 (no debunk list) and Scenario 3 (external only debunk list), but it still does not perform as well as scenario 2 (internal only debunk list). This highlights that combining internal and external information benefits BERT more than just relying only on external data, because BERT's fine tuning is highly dependent on domain specific data. The GPT models continue to outperform BERT in this scenario except for recall (see Table 3, Scenario 4). GPT-3.5-turbo reaches an accuracy of 0.85, F1 score of 0.83 and MCC of 0.71. GPT-4o-mini and GPT-4o have an accuracy of 0.83 and 0.84 respectively with precision of

Table 3: Summary of GPT models and BERT base models' performance in detecting COVID-19 misinformation across various scenarios.

Scenario 1- No Debunk List						
Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	MCC
BERT	0.50	0.80	0.00	0.01	0.00	0.03
gpt-3.5-turbo	0.80	0.83	**0.75**	**0.79**	0.60	0.60
gpt-4o-mini	**0.81**	0.90	0.69	0.78	**0.61**	0.63
gpt-4o	**0.81**	**0.92**	0.67	0.78	**0.61**	**0.64**
Scenario 2- Debunk List A						
Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	MCC
BERT	0.81	0.83	**0.78**	**0.81**	0.62	0.62
gpt-3.5-turbo	**0.84**	0.94	0.72	**0.81**	**0.67**	0.69
gpt-4o-mini	0.83	0.95	0.70	0.80	0.66	0.68
gpt-4o	**0.84**	**0.98**	0.69	**0.81**	**0.67**	**0.70**
Scenario 3- Debunk List B						
Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	MCC
BERT Finetuned	0.59	0.61	0.50	0.55	0.17	0.18
gpt-3.5-turbo	0.78	0.89	**0.65**	0.75	0.57	0.59
gpt-4o-mini	**0.80**	0.94	**0.65**	**0.77**	**0.61**	**0.64**
gpt-4o	0.79	**0.97**	0.60	0.74	0.58	0.63
Scenario 4- Debunk List C						
Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	MCC
BERT Finetuned	0.77	0.75	**0.82**	0.78	0.55	0.55
gpt-3.5-turbo	**0.85**	0.94	0.75	**0.83**	**0.70**	**0.71**
gpt-4o-mini	0.83	0.93	0.72	0.81	0.67	0.68
gpt-4o	0.84	**0.97**	0.69	0.81	0.67	0.70

over 0.92. Although the recall is slightly lower (0.72 and 0.69 respectively), they achieve the F1 scores of 0.81. While GPT models perform slightly better than scenario 3, their improved performance is as good as Scenario 2. This indicates that additional external information offers limited improvement for GPT models. This scenario shows that combining external and internal information improves the performance of both models compared to Scenario 3. However, the improvement is more notable for BERT, mainly due to its dependency on domain-specific data for effective fine-tuning, whereas GPT models benefit from their broader pre-training, making them less sensitive to specific data sources.

4 Discussion

This study evaluates the performance of BERT-based models and GPT models (GPT-3.5-turbo, GPT-4o-mini, and GPT-4o) in detecting COVID-19 misinformation using zero-shot and few-shot approaches. The results show that GPT models consistently outperform BERT models across all scenarios. Especially in the zero-shot setting where no debunk list is used, GPT models achieve notably higher performance and generalizability. This highlights their capability to detect misinformation effectively even without additional context. This makes GPT models suitable for real-world applications where minimal guidance is required. When a debunk list is used, both BERT and GPT models show performance improvements. The fine-tuned BERT model benefits most from an internal debunk list derived from the Constraint dataset, achieving notable improvement in most metrics, making it competitive with GPT models. However, its performance is particularly lower when fine-tuned with an external debunk list. This indicates BERT dependence on domain-specific data for effective adaptation. In contrast, GPT models show stronger adaptability. They are able to leverage both internal and external information more effectively, which results in maintaining more consistent performance across all scenarios. GPT models can be a cost-effective alternative to the human annotation process, which is costly and labor-intensive. These advantages of LLMs make them valuable resources for detecting misinformation, especially in future crises, when timely and accurate detection of misinformation is critical to public health safety.

Debunk lists can be used as additional domain-specific resources to enhance LLM performance by complementing their general training. However, their effectiveness depends on factors such as coverage and accuracy. If a debunk list is outdated, incomplete, or biased, it may negatively impact model performance. Future studies should evaluate both the benefits and limitations of using debunk lists in health misinformation detection. LLMs have explainability features that help in understanding how they detect misinformation and assess their trustworthiness. Further research can investigate the trustworthiness and stability of LLMs in health misinformation detection to strengthen their capability and robustness in real-world applications. GPT models have different knowledge cut-off dates, which may affect their performance. GPT-4o(-mini) has a knowledge cutoff of September 30, 2023, while GPT-3.5-turbo's cutoff is August 31, 2021 [32]. The GPT models' knowledge cutoffs occurred after COVID-19. Further analysis is needed to thoroughly investigate whether the knowledge cutoff impacted the models' performance. A future study could focus on evaluating a model whose knowledge cutoff is before December 2019 to assess its ability to detect health misinformation during a new pandemic.

4.1 Limitations

While this study provides valuable insights into the use of BERT and GPT models for detecting COVID-19 misinformation, there are several limitations to consider. Although the Constraint dataset is widely used and reputable, it may contain inherent biases that could impact model performance. Additionally, the externally collected debunk list, while collected from reputable organizations, may not fully represent the diversity of COVID-19 misinformation content found in real-world scenarios. This could limit the generalizability of the models to broader misinformation contexts. Furthermore, Constraint dataset focuses on English-language content may restrict the applicability of the findings to other languages and domains.

References

[1] Will Jennings, Gerry Stoker, Hannah Bunting, Viktor Orri Valgarðsson, Jennifer Gaskell, Daniel Devine, Lawrence McKay, and Melinda C. Mills. 2021. Lack of trust, conspiracy beliefs, and social media use predict covid-19 vaccine hesitancy. *Vaccines*, 9, 6. DOI: 10.3390/vaccines9060593.

[2] Sezer Kisa and Adnan Kisa. 2024. A comprehensive analysis of covid-19 misinformation, public health impacts, and communication strategies: scoping review. *J Med Internet Res*, 26, (Aug. 2024), e56931. DOI: 10.2196/56931.

[3] Md Saiful Islam et al. 2020. Covid-19–related infodemic and its impact on public health: a global social media analysis. *The American journal of tropical medicine and hygiene*, 103, 4, 1621.

[4] Sam Bradd. [n. d.] Infodemic. https://www.who.int/health-topics/infodemic#tab=tab_1.

[5] Ammara Malik, Faiza Bashir, and Khalid Mahmood. 2023. Antecedents and consequences of misinformation sharing behavior among adults on social media during covid-19. *Sage Open*, 13, 1, 21582440221147022. DOI: 10.1177/21582440221147022.

[6] Eric Afful-Dadzie, Anthony Afful-Dadzie, and Sulemana Bankuoru Egala. 2023. Social media in health communication: a literature review of information quality. *Health Information Management Journal*, 52, 1, 3–17. PMID: 33818176. DOI: 10.1177/1833358321992683.

[7] Maureen Olive Gallardo and Ryan Ebardo. 2024. Online health information seeking in social media. In *Soft Computing and Its Engineering Applications*. Kanubhai K. Patel, KC Santosh, Atul Patel, and Ashish Ghosh, (Eds.) Springer Nature Switzerland, Cham, 168–179. ISBN: 978-3-031-53731-8.

[8] Saira Hanif Soroya, Ali Farooq, Khalid Mahmood, Jouni Isoaho, and Shan-e Zara. 2021. From information seeking to information avoidance: understanding the health information behavior during a global health crisis. *Information processing & management*, 58, 2, 102440.

[9] Mohammed N. Alenezi and Zainab M. Alqenaei. 2021. Machine learning in detecting covid-19 misinformation on twitter. *Future Internet*, 13, 10. DOI: 10.3390/fi13100244.

[10] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2023. Towards covid-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274, 110642. DOI: https://doi.org/10.1016/j.knosys.2023.110642.

[11] Izzat Alsmadi, Natalie Manaeva Rice, and Michael J O'Brien. 2024. Fake or not? automated detection of covid-19 misinformation and disinformation in social networks and digital media. *Computational and Mathematical Organization Theory*, 30, 3, 187–205.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805. http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.

[13] Riccardo Cantini, Cristian Cosentino, Irene Kilanioti, Fabrizio Marozzo, and Domenico Talia. 2023. Unmasking covid-19 false information on twitter: a topic-based approach with bert. In *Discovery Science*. Albert Bifet, Ana Carolina Lorena, Rita P. Ribeiro, João Gama, and Pedro H. Abreu, (Eds.) Springer Nature Switzerland, Cham, 126–140. ISBN: 978-3-031-45275-8.

[14] Myeong Gyu Kim, Minjung Kim, Jae Hyun Kim, and Kyungim Kim. 2022. Fine-tuning bert models to classify misinformation on garlic and covid-19 on twitter. *International Journal of Environmental Research and Public Health*, 19, 9. DOI: 10.3390/ijerph19095126.

[15] Lwin Moe, Arghya Kundu, and Uyen Trang Nguyen. 2023. A bert-based explainable system for covid-19 misinformation identification. (2023).

[16] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. (2023). https://arxiv.org/abs/2306.00176 arXiv: 2306.00176 [cs.CL].

[17] Mohammed Aldeen, Joshua Luo, Ashley Lian, Venus Zheng, Allen Hong, Preethika Yetukuri, and Long Cheng. 2023. Chatgpt vs. human annotators: a comprehensive analysis of chatgpt for text annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, 602–609. DOI: 10.1109/ICMLA58977.2023.00089.

[18] Thi Huyen Nguyen and Koustav Rudra. 2024. Human vs chatgpt: effect of data annotation in interpretable crisis-related microblog classification. In *Proceedings of the ACM on Web Conference 2024*, 4534–4543.

[19] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. (2023). https://arxiv.org/abs/2304.10145 arXiv: 2304.10145 [cs.AI].

[20] Tom B. Brown et al. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165. https://arxiv.org/abs/2005.14165 arXiv: 2005.14165.

[21] Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. 2024. Large language models can automatically engineer features for few-shot tabular learning. (2024). https://arxiv.org/abs/2404.09491 arXiv: 2404.09491 [cs.LG].

[22] Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, KP Subbalakshmi, John R Wullert II, Chumki Basu, and David Shallcross. 2024. Can large language models detect misinformation in scientific news reporting? *arXiv preprint arXiv:2402.14268*.

[23] Vishnu S. Pendyala and Christopher E. Hall. 2024. Explaining misinformation detection using large language models. *Electronics*, 13, 9. DOI: 10.3390/electronics13091673.

[24] Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? (2024). https://arxiv.org/abs/2309.13788 arXiv: 2309.13788 [cs.CL].

[25] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

[26] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, 1302–1305.

[27] Jingwei Wang, Ziyue Zhu, Chunxiao Liu, Rong Li, and Xin Wu. 2024. Llm-enhanced multimodal detection of fake news. *PloS one*, 19, 10, e0312240.

[28] João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Weakly supervised veracity classification with llm-predicted credibility signals. (2024). https://arxiv.org/abs/2309.07601 arXiv: 2309.07601 [cs.CL].

[29] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: generating reactions and explanations for llm-based misinformation detection. (2024). https://arxiv.org/abs/2402.10426 arXiv: 2402.10426 [cs.CL].

[30] Lynnette Hui Xian Ng and Kathleen M Carley. 2021. "the coronavirus is a bioweapon": classifying coronavirus stories on fact-checking sites. *Computational and Mathematical Organization Theory*, 27, 2, 179–194.

[31] Gautam Kishore Shahi and Durgesh Nandini. 2020. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. ICWSM, US, (June 2020). DOI: 10.36190/2020.14.

[32] OpenAI. [n. d.] Models. https://platform.openai.com/docs/models.

[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] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. 2021. The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, 9, 78368–78381. DOI: 10.1109/ACCESS.2021.3084050.

[35] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21, 1–13.