Claim Veracity Assessment for Explainable Fake News Detection

Bassamtiano Renaufalgi Irnawan[1] , Sheng Xu[1], Noriko Tomuro[3], Fumiyo Fukumoto[2], Yoshimi Suzuki[2]
[1]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences
[2]Graduate Faculty of Interdisciplinary Research
University of Yamanashi
[3]College of Computer and Digital Media
Depaul University
{g22dts03,fukumoto,ysuzuki}@yamanashi.ac.jp
tomuro@cs.depaul.edu

Abstract

With the rapid growth of social network services, misinformation has spread uncontrollably. Most recent approaches to fake news detection use neural network models to predict whether the input text is fake or real. Some of them even provide explanations, in addition to veracity, generated by Large Language Models (LLMs). However, they do not utilize factual evidence, nor do they allude to it or provide evidence/justification, thereby making their predictions less credible. This paper proposes a new fake news detection method that predicts the truth or false-hood of a claim based on relevant factual evidence (if exists) or LLM's inference mechanisms (such as common-sense reasoning) otherwise. Our method produces the final synthesized prediction, along with well-founded facts or reasoning. Experimental results on several large COVID-19 fake news datasets show that our method achieves state-of-the-art (SOTA) detection and evidence explanation performance. Our source codes are available online.[1]

1 Introduction

The rapid and massive spread of fake and misleading information during the COVID-19 pandemic has brought disbelief in information in human minds and led them to extremism (Balakrishnan et al., 2022; Ferreira Caceres et al., 2022). Several efforts have been made to detect fake news using pre-trained language models (PLMs) (Wani et al., 2021; Zhang et al., 2021; Kaliyar et al., 2021) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). However, they only focus on the textual features of the claim without external facts, which may lead to a misrepresentation of the truth. In more recent work, several researchers have attempted to use evidence to predict the veracity of a claim and generate explanations (Wang et al., 2023; Zhao et al., 2023; Yuan et al., 2023; Dementieva and Panchenko, 2021; Dementieva et al., 2022; Hu et al., 2023; Popat et al., 2018; Hu et al., 2022; Mosallanezhad et al., 2022). Some approaches use LLMs to generate veracity explanations (Hu et al., 2024; Wu et al., 2024). However, without utilizing evidence, explanations sometimes or even often hallucinate, deviating from the original line of reasoning and wandering into a labyrinth of confused gibberish. Most recent work in fact-checking these days uses evidence and LLMs to generate explanations (Gao et al., 2023). As illustrated in Figure 1, evidence-based explanations can cite the evidence source in addition to repeating the topic of

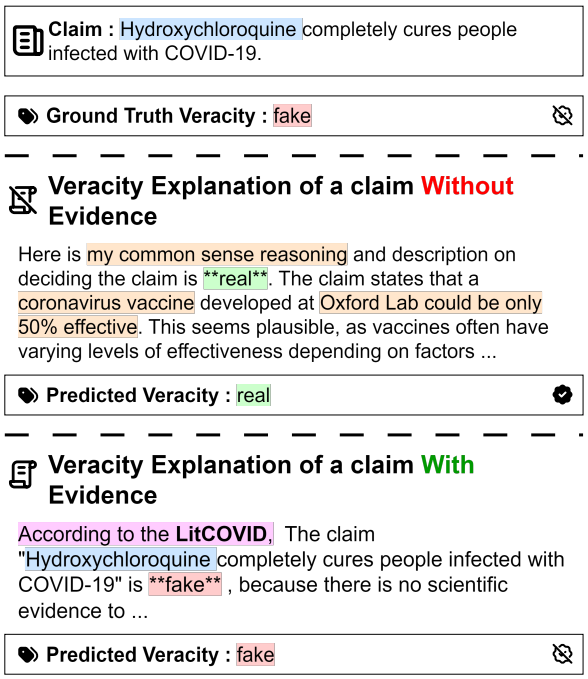

Figure 1: Examples of generated veracity explanations with and without utilizing evidence. The evidence-based method cites the evidence source (highlighted in pink). It references the keyword from the original claim (highlighted in blue), while the non-evidence explanation is hallucinating (highlighted in orange) and makes a wrong prediction.

[1]https://github.com/bassamtiano/covid_efnd

Proceedings of the 31st International Conference on Computational Linguistics, pages 4011–4029
January 19–24, 2025.

the claim to provide credibility to the user. On the other hand, non-evidence-based explanations may fail to mention the claim's topic and often make a wrong prediction.

In this paper, we focus on COVID-19 as the domain of fake news detection and propose a method to predict the claim's veracity with an explanation based on a large amount of grounded external factual evidence data as well as the logical reasoning capability facilitated in recent SOTA LLMs which are pre-trained with a vast amount of textual data. In addition, We incorporated a Retrieval Augment Generation (RAG) database and its mechanism, which stores a large amount of textual data in a vector format (for faster retrieval). We stored several publicly accessible COVID-19-related articles and news from trustworthy sources in the vector database to help us generate accurate and fact-grounded veracity predictions and explanations. We incorporated several modules in our system, utilizing LLMs (off the shelf) and PLMs, (which we fine-tuned) to generate partial predictions and explanations. Then, we employ an advisor model, which we extended from the work in (Hu et al., 2024), to sort out the candidate veracities and explanations generated by the sub-modules and produce the final prediction and explanation for a given claim. In summary, this work makes three contributions:

1. We propose a model for detecting fake COVID-19 claims and generating explanations that utilize factual evidence data and SOTA LLMs. The model also incorporates a trainable PLM model, which we fine-tune to choose the final model veracity and explanation.

2. We also propose an advisor model, which predicts the veracity of a claim to help the method select an appropriate explanation from multiple explanations: an evidence-based explanation, a commonsense-based explanation, and a textual-description-based explanation.

3. We conduct extensive experiments on publicly available COVID-19 fake news detection datasets and demonstrate that our model outperforms SOTA fake news veracity detection methods for detection and explanation generation.

2 Related Work

Fake News Detection Many COVID-19 fake news detection attempts utilize PLMs and LLMs. Their techniques can be classified into three groups: classification of claims using PLMs, classification and explanation generation using LLMs, and classification and explanation generation using external sources.

For PLM-based claim classification, (Wani et al., 2021) shows that PLMs perform fairly well in predicting veracity. Work by(Zhang et al., 2020; Kaliyar et al., 2021; Hu et al., 2024) used methods that bridge PLMs and LLMs by selectively injecting insight into the PLM from the generated LLM rationale. Similarly, Wu et al. proposed a technique to improve the semantic perception of evidence-aware fake news detection by injecting an LLM-generated semantic flip of a claim and LLM-generated claim invariance with the claim (Wu et al., 2024). Several strategies have been made to utilize external evidence in fake news detection: knowledge-based evidence (Popat et al., 2018; Hu et al., 2022), cross-lingual evidence (Dementieva and Panchenko, 2021; Dementieva et al., 2022), and CORD-19 evidence data for COVID-19 fake news detection (Wang et al., 2023).

Researchers have developed methods that utilize online and local evidence sources. For online-based evidence techniques, Yuan et al. developed a method to identify out-of-context multimedia misinformation by calculating the support-refutation score based on co-occurrence relations of named entities from claims and online search engine evidence (Yuan et al., 2023). Dementieva et al. present a cross-lingual evidence collection for the claim from different languages via machine translation and online search engines (Dementieva and Panchenko, 2021; Dementieva et al., 2022). For local-based evidence techniques, Wang et al. and Zhao et al. propose a method to predict claim veracity by utilizing claims and evidence retrieved from the evidence database using the BM25 algorithm (Wang et al., 2023; Zhao et al., 2023). Our proposed method uses a local-based evidence database via FAISS library (Douze et al., 2024; Johnson et al., 2019) to manage and provide the evidence for the claim.

Evidence-Based Explainable Fake News Detection Gao et al. proposed a method that utilizes multiple LLM instances to fix the claim context based on the evidence retrieved using the online search

engine (Gao et al., 2023). The research employs the query generation and agreement gate modules developed by (Gao et al., 2023) on COVID-19 fake news datasets. By modifying their prompt to a particular COVID-19 domain on query generation and agreement gate module, the proposed method is able to generate query and explanation inside the COVID-19 domain.

3 Explainable Fake News Detection

3.1 Overview

Our explainable fake news detection framework consists of three modules: (A) explanation generation without evidence, (B) explanation generation with evidence, and (C) explanation chooser, as depicted in Figure 2. Module (A) extends the work of (Hu et al., 2024) by employing two LLMs to produce explanations: one utilizes logical reasoning for commonsense explanations, and the other analyzes the claim's writing quality for textual description explanations. Module (B) functions as an RAG model, retrieving evidence pertinent to the claim from a stored database. This dual-module approach ensures that when Module (B) cannot find relevant evidence, Module (A) can still generate explanations, particularly for claims beyond the scope of COVID-19. Module (C), the explanation chooser, synthesizes the veracity assessments and explanations from Modules (A) and (B) and selects the final prediction and explanation. Overall, the goal of our proposed model is to provide veracity $\hat{y}$ with its explanation $\hat{y}_{desc}$. The veracity of a claim $\hat{y}$ will be real or fake. The explanation $\hat{y}_{desc}$ will contain a claim veracity explanation, e.g., *the claim "Hydroxychloroquine completely cures people infected with COVID-19" is **fake** because there is no scientific evidence that Hydroxychloroquine as a cure for COVID-19.*

3.2 Explanation without evidence

As shown in Figure 2, the commonsense-based analysis in module (A) utilizes LLMs LLM_{cs} to evaluate the plausibility of a claim against general knowledge, producing a veracity label $\hat{e}_{cs}$ and an explanation $\hat{y}_{cs}$. Conversely, the textual-description-based analysis LLM_{tx} examines the claim's writing style for indicators such as exaggeration or emotional language, generating its veracity label $\hat{e}_{tx}$ and explanation $\hat{y}_{tx}$. Detailed procedures are outlined in Algorithm 1, with prompt templates provided in Appendices Figures 5 and 6.

Algorithm 1 Explanation without Evidence

Require: Claim c
Ensure: Generated Veracity Explanation without Evidence $\hat{e}_{cs}, \hat{y}_{cs}, \hat{e}_{tx}, \hat{y}_{tx}$ for Claim c
Initialize Claim c
Run LLM Explanation Generation:
$\hat{e}_{cs}, \hat{y}_{cs} \leftarrow LLM_{cs}(c)$
$\hat{e}_{tx}, \hat{y}_{tx} \leftarrow LLM_{tx}(c)$
Return $\hat{e}_{cs}, \hat{y}_{cs}, \hat{e}_{tx}, \hat{y}_{tx}$

3.3 Explanation with evidence

Module (B) generates the claim's veracity explanation using evidence of the claim using RAG. It consists of three steps: query generation, evidence search, and explanation generation. The prompt for the explanation with evidence can be found in the Appendix Figure 10. The overall process of Evidence-Based Explanation is described in Algorithm 2.

Algorithm 2 Explanation with Evidence

Require: Claim c
Ensure: Generated Veracity Explanation with Evidence $\hat{e}_{db}, \hat{y}_{db}$ for Claim c
Query Generation:
$\cup Q = LLM(c)$
$q = sim(c, \cup Q)$
Evidence Search:
$\cup ev \leftarrow db(q)$
$\hat{ev} \leftarrow sim(c, \cup ev)$
Run LLM Explanation Generation:
$\hat{e}_{db}, \hat{y}_{db} \leftarrow f(c, q, \hat{ev})$
Return $\hat{e}_{db}, \hat{y}_{db}$

Query Generation The Query Generation step generates question sentences from input claims. The purpose of this step is to abstract away from the words and expressions in the claim and have multiple, generally worded sentences. We also formulate those sentences as questions to increase the accuracy of the RAG evidence retrieval. For the generalization technique, we design a generalization prompt by asking the LLM to remove the region name and famous people's names. We first feed the original claim to an LLM (LLAMA3 8 billion parameters) aligned to the COVID-19 domain by giving the example in the prompt to generate three to five-question queries. The prompt for the generalization technique and query generation can be found in the Appendix Figure 8 and 9, respec-

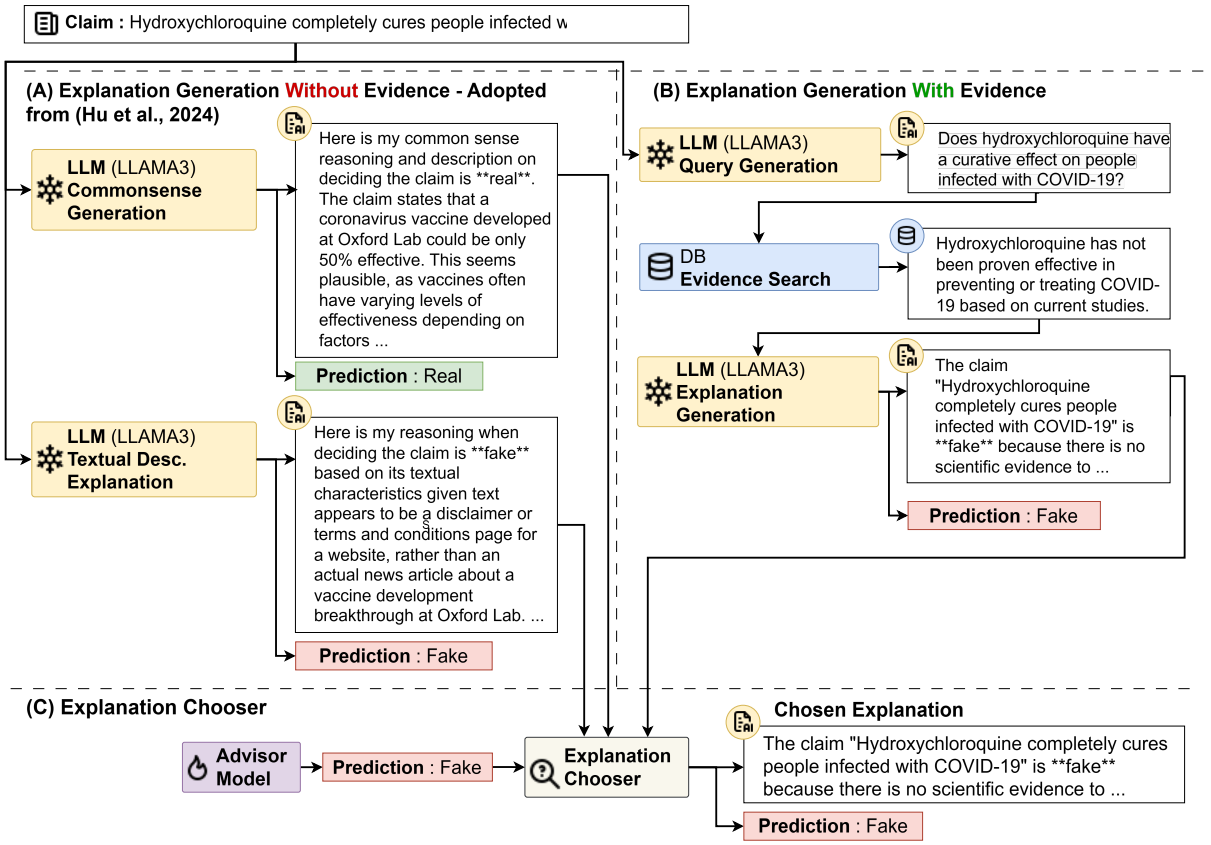

Figure 2: Overview of our proposed model that generates veracity and explanation for claims. For a given claim (at the top-left corner), module (A) generates two veracities and explanations using LLMs (without using factual evidence). The module (B) generates veracity and an explanation by considering evidence stored in a database. The module (C) is the explanation chooser, consisting of an Advisor Model that helps the Explanation Chooser select the final explanation for the claim.

tively. Generated queries are then ranked based on their similarity to the claim using cosine similarity between the claim and the query vectors (Reimers and Gurevych, 2019). Formally, given a claim c and multiple non-ranked LLM-generated queries $\cup Q$, the queries $\cup Q$ consist of multiple query $\cup Q = [\bar{q}_1, \bar{q}_2, \cdots \bar{q}\_n]$. Let q be the multiple LLM-generated queries that ranked based on their similarities to the claim.

$$q = sim(c, \cup Q), \tag{1}$$

Note that in the next Evidence search step, only the first query with the highest textual semantic similarity score to the claim is used. The Query Generation step is the first step of evidence-based explanation which is shown in Algorithm 2.

Evidence Search Evidence Search aims to provide evidence for the claim for the accuracy and credibility of the claim's veracity explanation. For each of the factual data resources we used (NIH, CDC, LitCOVID, Politifact, and the claim part of the COVID-19 dataset), we created an evidence database using FAISS (Douze et al., 2024; Johnson et al., 2019) and the sentence embedding vector library (Reimers and Gurevych, 2019). The generated query from the previous step is then presented to the evidence databases to retrieve the relevant evidence. We formulate the Evidence Search as follows:

$$\hat{ev} = sim(c, \cup ev), \tag{2}$$

where $\cup ev = [ev_1, ev_2, ...ev_n]$ refers to the sets of evidence retrieved from the databases $db()$. The evidences are then ranked once again using the same sentence similarity function. The top of the rank evidence $\hat{e}$ is selected and passed to the next step. The evidence search step is executed after generating the query, as described in Algorithm 2.

Explanation Generation Explanation Generation aims to generate predictions and explanations based on the evidence passed from the previous step. We use a LLM (LLAMA3 8 billion param-

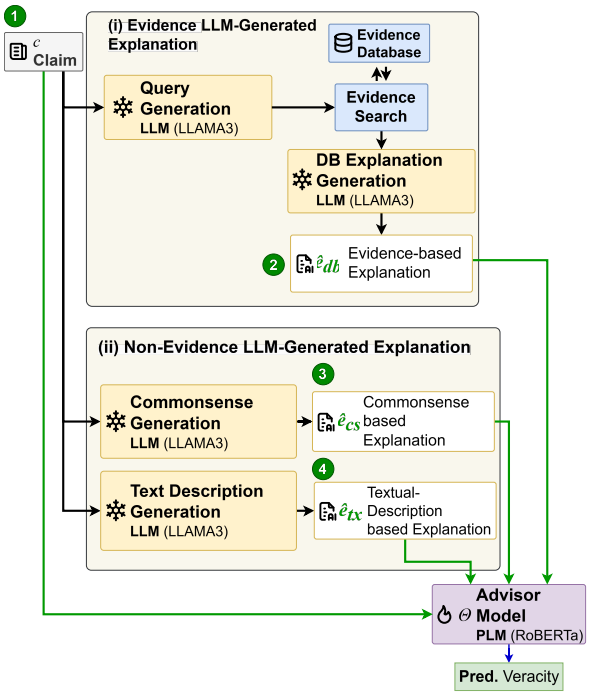

Figure 3: Information flow for the Advisor Model (on the right-bottom corner). It receives four inputs: (1) the original claim, (2) the evidence-based explanation, and two non-evidence-based explanations: (3) the commonsense-based explanation and (4) the textual-description-based explanation, and generates the final veracity-prediction.

eters) for this process. We send a prompt to the LLM with the query and evidence added with an instruction to generate explanation veracity. We formulate the explanation generation as follows. Given a query q and an evidence $\hat{ev}$

$$\hat{y}_{db}, \hat{e}_{db} = LLM_{db}(c, q, \hat{ev}) \quad (3)$$

where $\hat{y}_{db}$ is the veracity prediction using the databases, $\hat{e}_{db}$ is the explanation generated by the LLM, and $LLM_{db}()$ is the LLM runtime. Explanation generation is the last step, which is shown in Algorithm 1.

3.4 Advisor Model

Figure 3 shows the process of the advisor model Θ, which receives process processes four inputs: (1) original claim, (2) evidence-based explanation, (3) commonsense-based explanation, and (4) textual-description-based explanation to generate a final veracity prediction and explanation. The advisor model Θ refers to an extended model of (Hu et al., 2024) additional evidence-based explanation as an input of the model. The advisor model consists of two components: the RoBERTa PLM (Liu et al., 2019), which extracts the vector embedding from multiple inputs, and the cross-attention network bridging all vector embeddings. Note that the advisor model module notably differs from others in that it employs a trainable PLM model (in particular, RoBERTa (Liu et al., 2019)). Since several early works on fake news detection reported relatively good accuracy by fine-tuning pre-trained PLMs such as BERT (Wani et al., 2021; Zhang et al., 2021; Kaliyar et al., 2021), the advisor model fine-tuned on the COVID-19 fake news training datasets to predict the veracity of a claim. We process the original dataset through non-evidence-based and evidence-based explanation modules to prepare the advisor model's training data. The process generates explanations that include commonsense reasoning $\hat{e}_{cs}$, textual descriptions $\hat{e}_{tx}$, and evidence-based explanation $\hat{e}_{db}$. We then use training data and the corresponding claim c for the advisor model. We formulate the advisor model as follows:

$$\hat{y}_{adv} = \Theta(c, \hat{e}_{db}, \hat{e}_{cs}, \hat{e}_{tx}), \quad (4)$$

where $\hat{y}_{adv}$ indicates predicted veracity of a claim and Θ refers to the advisor model. The inputs to Θ are claim c, evidence-based explanation $\hat{e}_{db}$, commonsense-based explanation $\hat{e}_{cs}$, and textual-description based explanation $\hat{e}_{tx}$. The model output is the claim's real or fake veracity, which makes the fifth veracity to be sent to the final Explanation Chooser.

Explanation Chooser The Explanation Chooser selects the final veracity prediction and explanation from the ones generated and selected so far. It examines two aspects: the explanation that has the highest textual semantic similarity score to the claim, and the similarity of explanation veracity with the predicted advised veracity $\hat{y}_{adv}$.

The process, detailed in Algorithm 3, begins by consolidating three explanation types: database-based $\hat{e}_{db}$, commonsense-based $\hat{e}_{cs}$, and textual-description-based $\hat{e}_{tx}$ into a unified set, denoted as $\cup E$. $\cup E$ is ranked using a semantic similarity function, $sim()$, which orders the explanations based on their relevance to the original claim, resulting in a ranked set $\cup E_r$. An advisor model, $\Theta()$, then evaluates $\cup E_r$ against the claim c to produce an advised label, $\hat{y}_{adv}$. Subsequently, the veracity predictions from each explanation type: $\hat{y}_{db}$, $\hat{y}_{cs}$, and $\hat{y}_{tx}$ are compared with $\hat{y}_{adv}$ to select the final explanation. In this sequential selection process,

Algorithm 3 Explanation Chooser

Require:
Claim c
Database Explanation $\hat{e}_{db}$,
Database Veracity Prediction $\hat{y}_{db}$,
Commonsense Explanation $\hat{e}_{cs}$,
Commonsense Veracity Prediction $\hat{y}_{cs}$,
Textual Description Explanation $\hat{e}_{tx}$
Textual Veracity Prediction $\hat{y}_{tx}$
Ensure: Select Explanation from Claim c
$\cup E \leftarrow [\hat{e}_{db}, \hat{e}_{cs}, \hat{e}_{tx}]$
$\cup Er \leftarrow sim(\cup E)$
$\hat{y}_{adv} \leftarrow \Theta(c, \cup Er)$
if $\hat{y}_{db} == \hat{y}_{adv}$ **then**
    Return $\hat{e}_{db}$
else if $\hat{y}_{cs} == \hat{y}_{adv}$ **then**
    Return $\hat{e}_{cs}$
else if $\hat{y}_{tx} == \hat{y}_{adv}$ **then**
    Return $\hat{e}_{tx}$
else
    Return $\hat{e}_{db}$
end if

Datasets	Num. Claim
MMCoVaR (Chen et al., 2021)	2,593
ReCOVery (Zhou et al., 2020)	2,029
MM COVID-19 (Li et al., 2020)	9,457
Total	14,079

Table 1: The statistics of the three COVID-19 datasets.

Source	Num. Evidence
NIH	1,131
CDC	11,823
LitCOVID 19	407,982
Politifact	2,038
Total	422,974

Table 2: The statistics of evidence resources

the evidence-based explanation is prioritized if all three predicted veracity labels match $\hat{y}_{adv}$. If the evidence-based explanation's predicted label does not align with $\hat{y}_{adv}$, the system opts for either the commonsense-based or textual-description-based explanation that does. In scenarios where none of the predicted veracity labels correspond with $\hat{y}_{adv}$, $\hat{e}_{db}$, is selected.

4 Experiments

4.1 Experimental Setup

We use three common COVID-19 fake news datasets: MMCoVaR (Chen et al., 2021), ReCOVery (Zhou et al., 2020) and MM COVID-19 (Li et al., 2020) to evaluate our method. We use two datasets from three as the training data and use the remaining one as the test data. We repeated that three times, just in the same way as 3-fold cross-validation. The number of evidence in each dataset is shown in Table 1.

As for evidence, we use two types of sources: the first type is the (publicly available) published medical papers on COVID-19 from reliable, trusted sources (NIH, CDC, and LitCOVID) and articles from fact-checking sites (in particular, Politifact). The breakdowns of the data sizes are shown in Table 2. The second type is the collection of 'context' parts from the COVID-19 fake news detection datasets shown in Table 1. Note that, during training and evaluation, the data from the second type was properly controlled to avoid testing the data included in the training data. All those sources are stored separately in the FAISS vector database. We choose FAISS over other vector database libraries because implementation is simpler, and it scales well to large data.

In our method, we use the LLAMA 3 8B LLM for generating query and veracity explanation and the RoBERTa PLM model for the advisor model. We chose Llama 3 with 8 billion parameters because it is open-source and SOTA. We considered other models, such as GPT-4o and Claude Sonnet 3.5, but decided against them because they are closed-source (this risks privacy). Additionally, we aimed to experiment with whether and how much a smaller language model could effectively contribute to generating accurate explanations. We implemented our method with Pytorch and experimented on the Nvidia GeForce RTX A6000, (128GB memory).

For model parameters, we set the temperature and repetition penalty to 0.2 and 1.2, respectively, and the learning rate for PLM to be 1e-5. For evaluation metrics, we use accuracy and macro precision/recall/F1. To evaluate the generated explanations, we use three metrics: BERTscore (Zhang et al., 2020), which computes the vector similarity between the claim and generated explanation, FactCC score (Kryściński et al., 2019) which measures the factual consistency between a claim and

Methods	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
	MMCoVaR				ReCOVery				MM COVID-19			
RoBERTa (Wani et al., 2021)	0.630	0.315	0.500	0.386	0.672	0.336	0.500	0.402	0.447	0.246	0.421	0.309
Commonsense (Hu et al., 2024)	0.655	0.652	0.655	0.511	0.686	0.639	0.549	0.521	0.614	0.651	0.627	0.602
Textual Desc. (Hu et al., 2024)	0.682	0.695	0.589	0.570	0.712	0.709	0.583	0.569	0.621	0.620	0.620	0.620
Database (Ours)	0.576	0.521	0.518	0.513	0.599	0.535	0.533	0.534	0.533	0.523	0.516	0.486
Bad Actor Good Advisor (Hu et al., 2024)	0.811	0.816	0.772	0.784	0.832	0.812	0.802	0.807	0.691	0.345	0.500	0.408
Robust FND (Wu et al., 2024)	0.630	0.315	0.500	0.386	0.672	0.672	0.500	0.402	0.469	0.234	0.500	0.319
Check-COVID (Wang et al., 2023)	0.811	0.841	0.757	0.774	0.957	**0.964**	0.939	0.950	0.567	0.599	0.581	0.551
Full Framework (Ours)	**0.844**	**0.874**	**0.798**	**0.817**	**0.959**	0.961	**0.945**	**0.952**	**0.738**	**0.750**	**0.730**	**0.730**

Table 3: The classification performance of our proposed method on three datasets against seven baseline models.**Bold** is best, underline is second best.

an explanation, and ChrF (Popović, 2015) which compares the textual similarity of the claim and generated explanation.

4.2 Baselines

To evaluate our proposed method, we compare our model against several baseline models and approaches, which are classified into two groups:

Claim's Veracity Prediction consists of seven baselines: first **RoBERTa** predicts the claim's veracity with classification approach (Wani et al., 2021; Liu et al., 2019), **Commonsense**, which is another one of (Hu et al., 2024) modules, predicts the claim's veracity using LLM logical reasoning on the claim using LLM, **Textual description**, which is one of (Hu et al., 2024) modules, predicts the claim's veracity by analyzing the writing quality of the claim using LLM, **Database** predicts the claim's veracity by analyzing the gap between the claim and the evidence using LLM, **Bad Actor Good Advisor** integrates PLM and LLM by incorporating LLM-generated logic to evaluate the claim's veracity (Hu et al., 2024), **Robust FND** improves the semantic perception of evidence-aware fake news detection via LLM-generated semantic-flip and paraphrasing (Wu et al., 2024), **Check-COVID** utilizes COVID-19 evidence to predict the claim's veracity (Wang et al., 2023).

Claim's Veracity Explanation Generation consists of two baselines: first, **Bad Actor Good Advisor** generates its explanation based on commonsense reasoning and textual description characteristics of the input claim (Hu et al., 2024), and second, **Robust FND** generates an explanation based on the paraphrased explanation and semantic flip explanation of the input claim (Wu et al., 2024),

5 Results

5.1 Fake News Detection

As we can see in Table 3, our method outperforms most of the baselines except for one in the ReCOVery dataset, where precision is slightly lower than the method proposed by (Wang et al., 2023). Our method outperforms RoBERTa by achieving a 0.431 increase in F1 over MMCoVaR, 0.550 over ReCOVery, and 0.421 over MM COVID-19. This indicates that utilizing evidence and claims can enhance the performance of veracity prediction compared with methods that only rely on the claim. Our proposed method outperformed the three LLM commonsense, textual description, and evidence-based veracity prediction methods by achieving a 0.285 increase in F1 over MMCoVaR, 0.410 over ReCOVery, and 0.160 over MM COVID-19. This demonstrated the benefit of using evidence with an advisor model as a filter for selecting the suitable claim's veracity.

As shown in Table 3, the database method that only uses LLM to predict the veracity of a claim performs the worst compared to the other approaches, demonstrating the hallucination even after the corresponding evidence is provided. This happened because the evidence database did not cover some claims. However, Table 4 indicates that our method, which integrates a database for explanation generation with commonsense reasoning and textual descriptions generation to address claims not covered by the evidence database, outperformed baseline methods that do not utilize evidence database.

Compared with the Bad Actor Good Advisor and Robust FND baselines, the proposed method outperformed the veracity prediction by achieving a 0.232 increase in F1 over MMCoVaR, 0.3475 over

Methods	F1 BERT	FactCC	CHR F	F1 BERT	FactCC	CHR F	F1 BERT	FactCC	CHR F
	MMCoVaR			ReCOVery			MM COVID-19		
Bad Actor Good Advisor (Hu et al., 2024)									
a. Commonsense	0.827	0.963	12.195	0.800	0.955	11.040	0.828	0.958	24.352
b. Textual Desc.	0.813	0.960	10.762	0.797	0.957	10.450	0.823	0.960	24.417
Robust FND (Wu et al., 2024)									
a. Invariance	0.823	0.945	1.856	0.822	0.952	2.100	0.879	0.956	35.355
b. Semantic Flip	0.824	0.941	1.344	0.824	0.938	1.577	**0.934**	0.930	**75.401**
Ours	**0.828**	**0.972**	**12.285**	**0.828**	**0.970**	**14.520**	0.856	**0.990**	32.965

Table 4: The explanation generation performance of our proposed method on three datasets against two baseline models.**Bold** is best, underline is second best.

ReCOVery, and 0.3665 over MM-COVID-19. This shows the advantage of using relevant evidence as additional input for the models when predicting a claim's veracity. The proposed method that relies on an LLM-generated explanation can outperform the baseline Check-COVID, which uses claims and evidence created by humans. Our method outperforms Check-COVID, which uses human-created claims and evidence, by achieving a 0.043 increase in F1 over MMCoVaR, 0.002 over ReCOVery, and 0.179 over MM COVID-19. This demonstrates that our explanation generation technique as an input can be beneficial for improving model veracity prediction.

5.2 Generating Explanation

As shown in Table 4, our method outperforms all the baselines except for the MM COVID-19 dataset with the F1 BERT and ChrF scores. The method outperformed Bad Actor Good Advisor in two explanation types: Commonsense and Textual Description-based explanation. It outperforms the commonsense and textual-description-based explanation by achieving a 0.004 increase of F1 BERT (MMCoVaR), 0.029 (ReCOVery), and 0.030 (MM COVID-19). When the commonsense and textual-based description only relies on the LLM and the input claim, the proposed method demonstrates the use of evidence and advisor model when generating and selecting a suitable explanation for the claim can be beneficial.

The experimental results show that using evidence and an advisor model is beneficial when generating and selecting a suitable explanation for the claim. Our approach generates multiple explanations close to the claim and selects the suitable explanation. Table 4 shows that the F1 BERT score obtained by our method is better than the baselines. However, it does not work well when we use the MM COVID-19 dataset, indicating that it generates an explanation context closer to longer input claims that appear in the MMCoVaR and ReCOVery test dataset but not in the MM COVID-19. Our method generates longer explanations for veracity as it reflects claims and evidence. However, comparing longer explanations with shorter claims may result in a lower F1 BERT score due to the shorter claims only covering limited context.

Metrics	1	2	3	4	5	Majority Score
Query	0	0	2	13	15	**5**
Evidence	0	9	10	8	3	**3**
Explanation	0	3	3	20	4	**4**

Table 5: Human Evaluation results on randomly selected thirty samples from three test-dataset outputs: six on MMCoVaR, four on ReCOVery, and twenty on MM COVID-19.

We also found that evaluating a generated explanation using F1 BERT vector similarity is unreliable since slightly similar explanations can also represent the claim. To achieve that, FactCC is used to benchmark the consistency between the claim and context. With FactCC scoring metrics, the proposed method outperformed the (Liu et al., 2019) methods on three test datasets by a 0.02 increase in F1 over MMCoVaR, 0.024 over ReCOVery, and 0.046 over MM COVID-19. This indicates that our generated explanation is more consistent when explaining the claim's veracity by referring to the evidence.

5.3 Human Evaluation

We conducted a small-scale in-house human evaluation using three criteria: query quality, evidence correctness, and reasoning quality. Each criterion has a five score range from one to five, with five points being the best quality. Employing a five-point scale allows for nuanced differentiation between quality levels, facilitating a more detailed analysis of the generated outputs.

Table 5 shows that the proposed method gener-

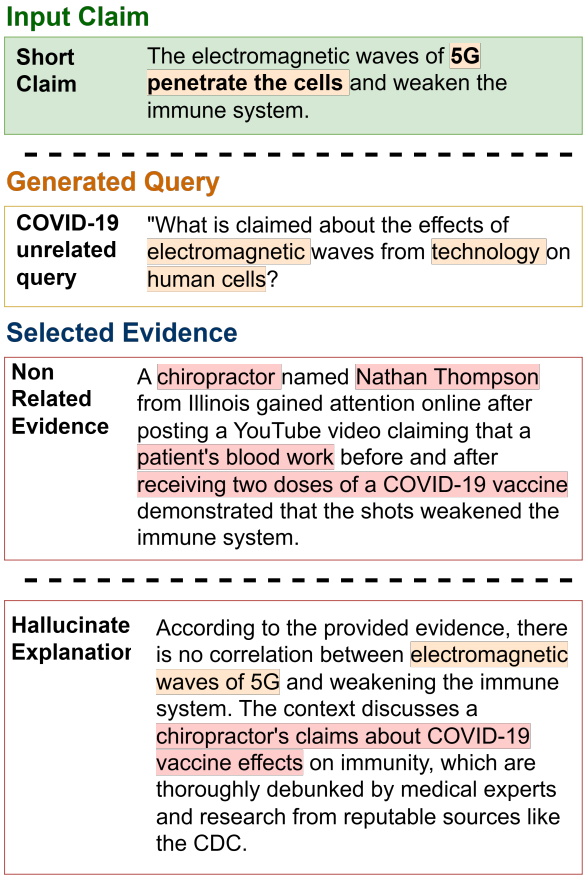

Figure 4: Example of an error on the non-related COVID-19 short claim. The method cannot generate COVID-19-related queries, resulting in non-relevant evidence and hallucinated veracity explanations.

ates queries for claims accurately with a majority score of five. The evidence search technique has a moderate majority score of three, indicating that evidence search may not capture appropriate evidence for some claims due to limited information. The explanation generation technique has a substantial majority score of four, demonstrating the ability of our proposed model to produce precise and reliable explanations.

5.4 Error Analyses

We conducted error analyses on the MMCoVaR, ReCOVery, and MM COVID-19 datasets. There are three major types of errors in the veracity explanation generation:

Non-related COVID-19 Topics. Our method struggles to explain claims that do not directly address COVID-19 as the main topic. As illustrated in Figure 4, the explanation is hallucinated when the provided claim is not directly related to the COVID-19 topics. We need further analysis on how to analyze the non-covid-19 related claims.

Short Claim. Our technique has difficulties generating explanations for short claims. As shown in Figure 4, the method struggles to generate a COVID-19 query for the short claim, leading to non-relevant evidence and hallucinated veracity explanation.

Prediction Latency. The proposed method takes approximately 30 to 50 seconds to process a single claim.

6 Conclusion

In this paper, we proposed a fake news detection method that provides a veracity explanation of a claim. The experimental results on three COVID-19 fake news datasets showed that our method achieved SOTA detection and evidence explanation performance. For future work, we plan to streamline the method without affecting the overall performance significantly to reduce the process of predicting the veracity of its explanation for one claim. To broaden and keep the information in our database up to date, we plan to collect new information periodically from reputable and open-domain sources and add it to the database. Lastly, we also plan to investigate strategies to improve the system's performance on short claims.

Limitations

One notable limitation of our method is its computational load. The process involves multiple runtime steps, including LLM generation for query, evidence-based explanation, commonsense-based explanation, and textual-description-based explanation. Consequently, processing each input claim requires approximately 40 to 50 seconds.

Ethical Statement

This research follows the standards in NLP research. The data used in the research is only from publicly available sources, and personally identifiable information was not included.

Acknowledgements

We would like to thank anonymous reviewers for their helpful comments and suggestions. This work is supported by the Kajima Foundation's Support Program for International Joint Research Activities and JKA. Bassamtiano Renaufalgi Irnamwan is funded by the MEXT scholarship, Grant Number 233203, and Sheng Xu by JST SPRING, Grant Number JPMJSP2133.

References

Vimala Balakrishnan, Wei Zhen Ng, Mun Chong Soo, Gan Joo Han, and Choon Jiat Lee. 2022. Infodemic and fake news – A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, 78:103144. Publisher: Elsevier.

Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021. Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21. ACM.

Daryna Dementieva, Mikhail Kuimov, and Alexander Panchenko. 2022. Multiverse: Multilingual evidence for fake news detection. *Preprint*, arXiv:2211.14279.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Maria Mercedes Ferreira Caceres, Juan Pablo Sosa, Jannel A Lawrence, Cristina Sestacovschi, Atiyah Tidd-Johnson, Muhammad Haseeb UI Rasool, Vinay Kumar Gadamidi, Saleha Ozair, Krunal Pandav, Claudia Cuevas-Lou, Matthew Parrish, Ivan Rodriguez, and Javier Perez Fernandez. 2022. The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, 9(2):262–277.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. *Preprint*, arXiv:2306.14728.

Xuming Hu, Zhijiang Guo, Guanyu Wu, Aiwei Liu, Lijie Wen, and Philip S. Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. *Preprint*, arXiv:2206.11863.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *Preprint*, arXiv:1910.12840.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *Preprint*, arXiv:2011.04088.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3632–3640, New York, NY, USA. Association for Computing Machinery.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *Preprint*, arXiv:1809.06416.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing. Association for Computational Linguistics.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-checking COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127, Toronto, Canada. Association for Computational Linguistics.

Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. *Evaluating Deep Learning Approaches for Covid19 Fake News Detection*, page 153–163. Springer International Publishing.

Yike Wu, Yang Xiao, Mengting Hu, Mengying Liu, Pengcheng Wang, and Mingming Liu. 2024. Towards robust evidence-aware fake news detection via improving semantic perception. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16607–16618, Torino, Italia. ELRA and ICCL.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. *Preprint*, arXiv:2311.01766.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3465–3476, New York, NY, USA. Association for Computing Machinery.

Runcong Zhao, Miguel Arana-catania, Lixing Zhu, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liakata, and Yulan He. 2023. PANACEA: An automated misinformation detection system on COVID-19. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212.

A Appendices

A.1 Example of Advisor Model Input

In the following Table 6, We show an example of an LLM-generated explanation proposed by (Hu et al., 2024) which does not consider evidence when generating the explanation of the claim's veracity. The explanation did not mention any evidence and mentioned the text that exists in the claim. In our proposed method, where we put the evidence when generating the explanation, we can see that LLM tries to compare and mention the evidence when deciding the claim's veracity. These three pieces of evidence will be fed to the Advisor Model together with the claim.

A.2 The Advisor Model Usability Demonstration

We demonstrate the effectiveness of the advisor model in selecting the most suitable explanation, as shown in Table 7. We found that relying solely on the textual similarity between the claim and the evidence is insufficient. The advisor model helps identify explanations that best align with the claim.

A.3 Example of Explainable Fake News Detection

Table 8 shows the outputs of our proposed methods: LLM-generated query, evidence selection, and explanation veracity in fake news detection. The output includes a claim with its corresponding evidence and an LLM-generated explanation that considers the claim, query, and evidence. The generated explanation that includes why the veracity is selected is listed as number 6 and the final veracity as number 7.

A.4 The Model LLM prompt

We illustrate the prompt that we implement in our proposed method. This includes the prompt for non-evidence explanation generation (Commonsense-based prompt: Figure 5 and textual-description-based prompt: Figure 6), and evidence explanation generation (Generalization Prompt: Figure 8, Query Generation Prompt: Figure 9, and Evidence-based Explanation prompt: Figure 10).

CLAIM:

Tray tables up: As vaccinations surge, so do airfares

If you got your two shots and you're looking to travel, fasten your seat belts: Airfares are climbing quickly and so are hotel rates Earlier this month, the Centers for Disease Control and Prevention issued new guidance easing coronavirus-related recommendations on travel throughout the U.S., saying fully-vaccinated Americans could safely travel again without needing to be tested or having to quarantine afterward. After more than a year of the pandemic, many Americans were clearly ready to hit the roads and take to the skies. Last weekend, the Transportation Security Administration reported screening more than 1.5 million travelers on a single day near a pandemic-level high and tons more than the 91,000 passengers screened on the day last year.But if you got your two shots and you're looking to travel, fasten your seat belts: Airfares are climbing quickly and so are hotel rates. Short-term bookings and long-term bookings are way up, CBS News Travel Editor Peter Greenberg told WTOP.

(A). LLM-generated explanation that considers evidence

4. From this source: https://www.politifact.com/factchecks/2021/aug/06/tom-tiffany/tiffany-misses-point-behind-why-prices-are-rising-/

5. I found this Evidence: During the COVID-19 pandemic, inflation increased by approximately 5% between May 2020 and May 2021. However, this does not fully explain the significant price hikes in travel-related expenses like car rentals, gas, and hotels. The actual causes stem from supply and demand imbalances. The rental car industry faced severe challenges during the pandemic, leading to the sale of roughly one-third of their fleets due to lack of demand. Additionally, the global chip shortage halted new car manufacturing, exacerbating the issue. As travel resumed, the high demand for rentals resulted in substantial price increases. Gas prices have also risen significantly since April 2020, following a decrease during the initial lockdowns. While some argue that Democratic policies are responsible, experts suggest that the primary driver of these increases is the lifting of restrictions and the subsequent surge in demand for travel.

6. By comparing the evidence and claim: According to the provided evidence, it highlights the factors contributing to price hikes in airfares, hotel rates, rental cars, and gas prices during the COVID-19 pandemic. The article mentions that airlines are expecting a ̈huge boom ̈in domestic travel due to vaccination growth, which could lead to increased demand for flights. The evidence suggests that supply and demand imbalances, rather than just vaccinations alone, drive price increases in these industries. It also explains how the rental car industry faced significant challenges during the pandemic, leading to high prices when travel resumed. The article's claim about airfares climbing quickly alongside vaccination growth is partially supported by the evidence but does not fully explain the causes of price hikes.

7. Therefore, the news is **real**.

(B). LLM-generated explanation that did not considers evidence - (Hallucinated Explanation)

The claim states that as vaccinations surge, so do airfares. The article provides evidence to support this statement by citing a travel editor who notes that short-term bookings are up 7% per week, with hotel rates increasing at a similar rate. Additionally, specific examples of high demand and prices in certain locations such as Hawaii and Florida are provided.

Using my common sense, I believe the claim is **real** because it seems to be based on factual information about travel trends during the pandemic era. The article cites real-world data from the Transportation Security Administration (TSA) and quotes a credible source, Peter Greenberg, who has expertise in travel news.

Additionally, the claims made are not absurd or impossible to happen; they seem plausible given the context of the pandemic's impact on air travel. Therefore, based on my common sense reasoning, I conclude that this claim is **real**.

Table 6: The example of Generated Explanation that considers Evidence and not

Similarity	Prediction	Ground Truth
Claim: Antiviral Drug Cocktail shows success in learning COVID-19 in 7 Days. For those waiting for more good news for this outbreak, recent reports may be just that as the three-drug cocktail consisting of the HIV medication lopinavir-ritonavir, the hepatitis therapy drug ribavirin, and the multiple sclerosis treatment interferon-beta is suggested to have successfully cleared COVID-19 in one week.		
Query: Antiviral drug cocktail shows success in learning COVID-19 within 7 days		
Chosen Explanation without Advisor Model		
0.714	Real	Fake
4. From this source: https://www.medicalnewstoday.com/articles/duo-of-antiviral-drugs-strongly-inhibits-sars-cov-2-in-the-lab 5. I found this Evidence: A recent study conducted by Norwegian and Estonian researchers has made a significant breakthrough in finding a potential treatment for COVID-19. A combination of two existing antiviral drugs, nelfinavir and amodiaquine, showed strong inhibition of SARS-CoV-2 in laboratory cell cultures. This combination has been shown to work synergistically, making it more potent than individual use of either drug. 6. By Comparing the evidence and claim: According to the https://www.medicalnewstoday.com/articles/duo-of-antiviral-drugs-strongly-inhibits-sars-cov-2-in-the-lab, a study found that a combination of two antiviral drugs showed strong inhibition of SARS-CoV-2 in laboratory cell cultures. This evidence suggests the possibility of finding effective treatments for COVID-19. 7. Therefore, the news is **real**		
Chosen Explanation with Advisor Model		
0.679	Fake	Fake
4. From this source: https://covid19.nih.gov/covid-19-treatments 5. I found this Evidence: The Centers for Disease Control and Prevention (CDC) recommends home treatment for those with mild COVID-19 symptoms using over-the-counter medication and practicing basic safety measures to prevent viral spread. The US Food and Drug Administration (FDA) has authorized the use of two oral antiviral treatments, Paxlovid (nirmatrelvir ritonavir) and Lagevrio (molnupiravir,) which are taken twice daily for five days starting as soon as symptoms appear. 6. By Comparing the evidence and claim: According to the provided evidence from the Centers for Disease Control and Prevention (CDC), there is no mention of a three-drug cocktail consisting of lopinavir-ritonavir, ribavirin, and interferon-beta being used or recommended by reputable medical organizations such as the CDC. The evidence suggests that FDA-approved treatments like Paxlovid and Lagevrio are available for mild COVID-19 symptoms. 7. Therefore, the news is **fake**.		

Table 7: The example of Evidence chosen with or Without Advisor Model, We also provide the result prediction and ground truth with or without advisor model

	Input Claim
1. Claim:	NEW DELHI – Indian Prime Minister Narendra Modi faced growing pressure Friday to impose a strict nationwide lockdown, despite the economic pain it will exact, as a startling surge in coronavirus cases that has pummeled the country‘s health system shows no signs of abating. Many medical experts, opposition leaders and even Supreme Court judges are calling for national restrictions, arguing that a patchwork of state rules is insufficient to quell the rise in infections.Indian television stations broadcast images of patients lying on stretchers outside hospitals waiting to be admitted, with hospital beds and critical oxygen in short supply. People infected with COVID-19 in villages are being treated in makeshift outdoor clinics, with IV drips hanging from trees.As deaths soar, crematoriums and burial grounds have been swamped with bodies, and relatives often wait hours to perform the last rites for their loved ones.The situation is so dramatic that among those calling for a strict lockdown are merchants who know their businesses will be affected but see no other way out.Only if our health is good, will we be able to earn, said Aruna Ramjee, a florist in the southern Indian city of Bengaluru...
	Evidence Search
2. Query:	Nationwide lockdown imposed in India due to surge in coronavirus cases.
3. Evidence:	As the COVID-19 pandemic continues to spread globally, India’s National Disaster Management Authority has ordered a nationwide lockdown extension until May 31 to combat the spread of the virus. This decision comes as the country has reported 90,927 cases and 2,872 deaths. Meanwhile, other countries around the world are also implementing various measures to contain the outbreak, including Egypt expanding its curfew and restricting travel during the Eid holiday, and Russia confirming 9,709 new cases. In Asia, Indonesia reported 489 new cases, Malaysia 22, and Thailand reopened shopping malls and restaurants as new cases decline. Globally, the number of confirmed cases has reached over 4.6 million, with 311,843 deaths.
	Generated Explanation
4. From this source:	https://www.cnbc.com/2020/05/17/coronavirus-live-updates-asia-europe.html
5. I found this Evidence:	As the COVID-19 pandemic continues to spread globally, India’s National Disaster Management Authority has ordered a nationwide lockdown extension until May 31 to combat the spread of the virus.
6. By Comparing the evidence and claim:	According to the provided evidence, it is stated that India imposed a nationwide lockdown extension until May 31 due to the surge in coronavirus cases. This information supports the context’s statement about pressure rising for an Indian lockdown. The evidence also confirms the record-breaking number of COVID-19 infections mentioned in the context.
7. Therefore,	the news is **real**.

Table 8: The Example of Input, Evidence Search, and Generated Explanation in our methods

```
<|begin_of_text|>
    <|start_header_id|>system<|end_header_id|>
        You are a helpful, straightforward, respectful, and honest assistant.
        Always respond as helpfully as possible, while being safe.
        I want you to follow the advised label when creating the description of the claim using
 your common sense.
        Do not make vague decisions such as slightly real or slightly fake.
        Do not make contradicted decision to the advised label such as if the advised label is **
 real** you need to make description using your commonsense as **real** and the opposite.

        I want you to follow this format when you create your response:
        'here is your common sense reasoning and description on deciding the input is **real** or
  **fake**. the description response only allowed here.' Therefore, The claim is 'here is your
 choice either **real** or **fake**. Do not put your desription here.'

    <|eot_id|>
    <|start_header_id|>user<|end_header_id|>
        You already know the claim is {advised_label} from the advised label.
        use your commonsense to create description for the claim following the status of the
 claim is {advised_label}.

        Given the following claim and the advised label, I want you to create commonsense
 description by following the advised label.
        Please refrain from providing ambiguous assessments such as undetermined: {claim}.
        Let`s think from the perspective of commonsense.
    <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 5: Commonsense-based explanation generation prompt

```
<|begin_of_text|>
    <|start_header_id|>system<|end_header_id|>
        You are a helpful, straight forward, respectful and honest assistant.
        Always answer as helpfully as possible, while being safe.
        I want you to follow the advised label when creating the description of the claim by
 analysing the text description of it.
        Do not make vague decisions such as slightly real or slightly fake.
        Do not make contradicted decision to the advised label such as if the advised label is **
 real** you need to make description using the textual description of the claim as **real** and
 the opposite.

        I want you to follow this format when you create your response:
        'here is your reasoning when deciding the input is **real** or **fake** based on the
 textual description. the description response only allowed here.'
        Therefore, The claim is 'here is your choice either **real** or **fake**. Do not put your
  desription here.'
    <|eot_id|>
    <|start_header_id|>user<|end_header_id|>
        You already know the claim is {advised_label} from the advised label.
        analyse the textual description of the claim and create description for the claim by
 following the status of the claim as {advised_label}.

        Given the following claim and the advised label, I want you to create textual description
  of the claim by following the advised label.
        Please refrain from providing ambiguous assessments such as undetermined: {claim}.
         Lets  think from the perspective of textual description.
    <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 6: Textual-Description-based explanation generation prompt

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
    You are a helpful, straight forward, respectful and honest assistant.
    Always answer as helpfully as possible, while being safe. Please ensure that your responses
 are socially unbiased and positive in nature.
    I want you to summarize the article without mentioning any region, person name, country name,
  state name, province name and time and put it as "Generalized Article".
    I also want you to convert the article into general and nonspecific domain articles. To do
 that, you must remove any details that mention the name of a person and the name of the location
  in your Generalized Article so that any specific domain information is removed.
    Do not change the value of number in your response.
    The Generalized Article must be shorter and compact with minimum of 3 and maximum of 5
 sentences inside one paragraph.
    Generalized mean summarized and transform specific domain sentence to more general domain
 sentence!.
    Do not remove the context that related to:
    1. covid or coronavirus topic
    2. vaccine type or vaccination information
    3. dose of vaccination.
    4. the location where the title and article happend.
    5. the type of coronavirus or covid variant.

    The following is example that you must follow in your response.

    1. Title: No, Joe -- You are Not a Nice Guy
    2. Article: Joe, you've been labeled as a "nice guy," but the truth reveals you as a
 pugilistic, plagiaristic, documented racist, and a creep towards women and children.
    Your pugilistic behavior and plagiarism have been well-documented, and your history is
 plagued with fraudulent actions, such as committing fraud to get into law school and
 plagiarizing leaders like Neil Kinnock and Hubert Humphrey.
    Your racism is evident in your smear campaigns against underpaid and overworked police forces
 , who risk their lives for American safety, and your not-so-veiled racism towards them is
 disgusting.
    The police force is not the only profession with bad actors, but your anti-cop rhetoric risks
  retirements and recruitment.
    Your intentions may be signaled as heroism on the hard left, but your racism goes way back.
    3. Generalized Title: Someone that not a nice guy.
    4. Generalized Article: Someone who labeled as nice guy, accused as pugilistic, plagiaristic,
  documunted racist, and a creep towards woman and children.

    You must follow the pattern of the example and do not response in XML or Any programming
 language format, only plain text is accepted.
    Arrange the summarized version of the article input with plain text format.

    Response in straight forward manner not like chat response, exclude your note regarding on
 How you do the process and only response the summarized version of the article.

<|eot_id|>
<|start_header_id|>user<|end_header_id|>
    1. Input Title: {input_title}
    2. Input Article: {input_article}

    Your task is now to generalized the Title and Article.
    Follow the example pattern but do not include the example in your response!. Instead process
 the Input Title and Input Article.

    make sure to follow this following format in your response:
    Put the generalized Title in the following format: 3. Generalized Title: 'Here is your
 generalized title response'
    Put the generalized article in the following format: 4. Generalized Article: 'Here is your
 generalized article response'

    3. Generalized Title:
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 7: Generalization prompt on Query Generation Module

```
<|begin_of_text|>
    <|start_header_id|>system<|end_header_id|>
        You are a helpful, respectful and honest assistant. Always answer as helpfully as
possible, while being safe.
        I will give you some example of pattern that you can use to generate short and compact
queries from the claim and context of an article.
        Build short and compact queries to confirm the claim and context about covid related
article in your response!.
        The ratio of query content is 60% from the claim, and 40% from the context.
        The queries contain the most important context in the claim and context.
        I want you to create at minimum 3 queries for each claim and context.
        Remove any word that mentioning any region, person name, country name, state name,
province name and time on your generated queries.
        Avoid using abbreviations, you need to include the full form of its abbreviations, for
example :
        1. J&J is Johnson & Johnson,
        2. CDC is Centers for Disease Control and Prevention
        3. NIH is National Institute of Health, etc.

        The following is the example of query that you use to verify the claim.

        Claim: Oregon: CDC investigating woman's death after J&J vaccine
        Context: Federal and state health authorities are investigating the death of a woman who
developed a rare blood clot and low platelets following the administration of the Johnson &
Johnson COVID-vaccine.
        The woman, whose identity remains confidential, received the vaccine prior to the CDC's
pause on the vaccine due to concerns over potential dangerous clots.
        Her symptoms, including severe headache, abdominal pain, leg pain, and shortness of
breath, were consistent with other reported cases.
        The investigation is ongoing, and it is unclear whether the woman's death is directly
related to the vaccine.
        The CDC has reported six cases of unusual blood clots, including one death, among the
approximately 8 million Americans who have received the one-dose vaccination.
        The decision to resume distribution of the J&J vaccine will depend on the health official
's recommendation. Health authorities emphasize that the potential benefits and risks of the
vaccine will be carefully considered throughout the investigation process.
        To verify the claim & context, you will use the following query,
        1. investigation into woman's death after receiving Johnson & Johnson COVID-vaccine.
        2. death because of blood clot and low platelets after receiving Johnson & Johnson COVID
-19 vaccine.
        3. investigation on rare blood clot and low platelets after getting Johnson & Johnson
COVID-vaccine.

        This is the end of the example
        Generate query from the input based on the example above!
    <|eot_id|>
<|start_header_id|>user<|end_header_id|>

    Claim: {input_claim}
    Context: {input_context}
    By referring to the example, To verify what you just said, you must make query.
    without adding example to your response, make only query and dont create a statements or
sentence.
    Do not includes reason or notes regarding to why you create the query.
    I want you to create at minimum 3 queries for each claim and context.
    dont include example on your answer!.

    make sure your questions follows this pattern:
    [Number]. [your questions]
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 8: Generalization prompt on Query Generation Module

```
<|begin_of_text|>
    <|start_header_id|>system<|end_header_id|>
        You are a helpful, respectful and honest assistant. Always answer as helpfully as
possible, while being safe.
        I will give you some example of pattern that you can use to generate short and compact
queries from the claim and context of an article.
        Build short and compact queries to confirm the claim and context about covid related
article in your response!.
        The ratio of query content is 60% from the claim, and 40% from the context.
        The queries contain the most important context in the claim and context.
        I want you to create at minimum 3 queries for each claim and context.
        Remove any word that mentioning any region, person name, country name, state name,
province name and time on your generated queries.
        Avoid using abbreviations, you need to include the full form of its abbreviations, for
example :
        1. J&J is Johnson & Johnson,
        2. CDC is Centers for Disease Control and Prevention
        3. NIH is National Institute of Health, etc.

        The following is the example of query that you use to verify the claim.

        Claim: Oregon: CDC investigating woman's death after J&J vaccine
        Context: Federal and state health authorities are investigating the death of a woman who
developed a rare blood clot and low platelets following the administration of the Johnson &
Johnson COVID-vaccine.
        The woman, whose identity remains confidential, received the vaccine prior to the CDC's
pause on the vaccine due to concerns over potential dangerous clots.
        Her symptoms, including severe headache, abdominal pain, leg pain, and shortness of
breath, were consistent with other reported cases.
        The investigation is ongoing, and it is unclear whether the woman's death is directly
related to the vaccine.
        The CDC has reported six cases of unusual blood clots, including one death, among the
approximately 8 million Americans who have received the one-dose vaccination.
        The decision to resume distribution of the J&J vaccine will depend on the health official
's recommendation. Health authorities emphasize that the potential benefits and risks of the
vaccine will be carefully considered throughout the investigation process.
        To verify the claim & context, you will use the following query,
        1. investigation into woman's death after receiving Johnson & Johnson COVID-vaccine.
        2. death because of blood clot and low platelets after receiving Johnson & Johnson COVID
-19 vaccine.
        3. investigation on rare blood clot and low platelets after getting Johnson & Johnson
COVID-vaccine.

        This is the end of the example
        Generate query from the input based on the example above!
    <|eot_id|>
<|start_header_id|>user<|end_header_id|>

    Claim: {input_claim}
    Context: {input_context}
    By referring to the example, To verify what you just said, you must make query.
    without adding example to your response, make only query and dont create a statements or
sentence.
    Do not includes reason or notes regarding to why you create the query.
    I want you to create at minimum 3 queries for each claim and context.
    dont include example on your answer!.

    make sure your questions follows this pattern:
    [Number]. [your questions]
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 9: Query Generation prompt

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
    You are a helpful, respectful and honest assistant.
    Always answer as helpfully as possible. Please ensure that your responses are socially
unbiased and positive in nature. I will provide you a COVID-19 related topic claim and context.
I want you to verify and decide if the claim and context about COVID-19 topic is **fake**, **
real**, or undecided based on the given evidence and your commonsense. You only can decide
either **fake**, **real**, or **undecided** in your response.
    When deciding the claim, do not make any vague decision such as partially fake or partially
real. If the evidence is commonsense or textual description, you need to follow the prediction
either **real** or **fake** as it is without changing it.
    Topic of the evidence must not be very specific to the claim, at least it has similar general
 context to it. If the claim and evidence in general is related but contradict each other, then
more likely the claim and context is **fake**. If the claim and evidence in general is related
and support each other, then more likely the claim and context is **real**. To generate your
response, you must follow the pattern of the example and not add example, prompt, and user input
 in your response!. I want you to follow the following pattern arrangement bellow!.
    1. Claim: Oregon: CDC investigating woman's death after J&J vaccine
    2. Context: Federal and state health authorities are investigating the death of a woman in
her who developed a rare blood clot and low platelets following the administration of the
Johnson & Johnson COVID-vaccine.
    The woman, whose identity remains undisclosed, developed symptoms such as severe headache,
abdominal pain, leg pain, and shortness of breath within two weeks of vaccination. The
investigation is ongoing, and it is unclear if the woman's death is directly related to the
vaccine.
    The CDC and state health departments are examining six cases of unusual blood clots,
including one death, among the approximately 8 million Americans who have received the one-dose
vaccination. The decision to resume distribution of the J&J vaccine will depend on the outcome
of the investigation and the recommendation of the CDC's advisory committee on vaccines.
    3. I Check: death investigation of a woman after receiving Johnson & Johnson COVID-vaccine
    4. From this source: https://www.cdc.gov/mmwr/volumes/70/wr/mm7018e2.htm
    5. I found this evidence:
    The Centers for Disease Control and Prevention (CDC) offers guidance on how to protect
oneself and others from the health risks associated with COVID-19 respiratory viruses.
    6. By Comparing the evidence and the claim: According to the CDC website, the organization
provides guidance on preventing the health risks associated with COVID-19 respiratory viruses.
The website mentions the importance of staying up-to-date with COVID-19 vaccines, practicing
good hygiene, wearing masks, maintaining social distancing, and staying home when sick. It also
emphasizes that vaccination reduces the risk of getting sick, hospitalization, or death from
COVID-19. The claim context discusses an investigation into a woman's death following the
administration of the Johnson & Johnson COVID-vaccine. The investigation is ongoing, and it is
unclear whether the woman's death is directly related to the vaccine.
    7. Therefore, the claim is **real**.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
    I want you to analyse the provided claim & context bellow by referencing the evidence.
    You must keep in mind that this claim and context is COVID-19 Related topic!, so use your
knowledge about COVID-19 when analyzing the evidence and claim. Check if the claim is **fake* or
 **real** based on the provided evidence. If the claim, context, and evidence is correlated and
the evidence debunk, describe or proof that the claim is **fake** then you need to treat the
claim as a **fake** news/claim. If the claim, context, and evidence is correlated and the
evidence describe and proof that the claim is **real**, then you need to treat the claim as a **
real** news/claim. If the claim, context in general have slightly similar topic with the
evidence, then you can extract the general idea of the evidence and use it as reference and use
your common sense to decide either the news/claim is **fake** or **real** based on extracted
evidence.

    1. Claim: {input_claim}
    2. Context: {input_context}
    3. I Check: {inputquery}
    4. From this source: {input_evidence_url}
    5. I found this Evidence: {input_evidence_text}
    6. By comparing the evidence and claim :
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 10: Evidence Based Explanation Generation prompt on Query Generation Module