Multi-Agent Cross-Lingual Veracity Assessment for Explainable Fake News Detection

Bassamtiano Renaufalgi Irnawan[1], Fumiyo Fukumoto[2], Noriko Tomuro[3], 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
{g23dtsa2,fukumoto,ysuzuki}@yamanashi.ac.jp
tomuro@cs.depaul.edu

Abstract

The spread of fake news during the COVID-19 pandemic era triggered widespread chaos and confusion globally, causing public panic and misdirected health behavior. Automated fact checking in non-English languages is challenging due to the low availability of trusted resources. There are several prior work that attempted automated fact checking in multilingual settings. However, most of them fine-tune pre-trained language models (PLMs) and only produce veracity prediction without providing explanations. The absence of explanatory reasoning in these models reduces the credibility of their predictions. This paper proposes a multi-agent explainable cross-lingual fake news detection method that leverages credible English evidence and Large Language Models (LLMs) to verify and generate explanations for non-English claims, overcoming the scarcity of non-English evidence. The experimental results show that the proposed method performs well across three non-English written multilingual COVID-19 datasets in terms of veracity predictions and explanations. Our source code is available online.[1]

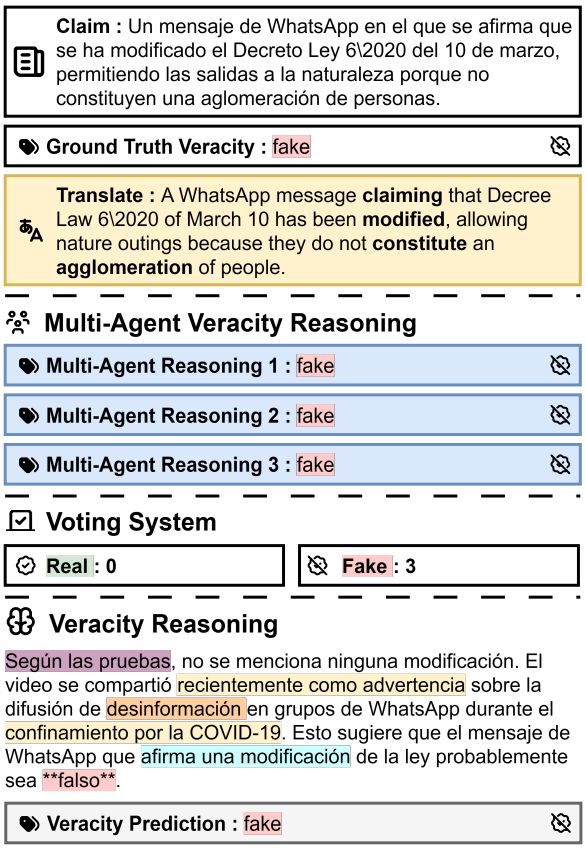

Figure 1: Example of voting-based ensemble reasoning outputs system consists of three predictions (represented as blue boxes). Voting systems decide the final veracity prediction and reasoning based on all predictions. The detail is shown in Appendix A

1 Introduction

During the COVID-19 pandemic, the spread of fake news caused a wide range of chaos and confusion around the world, for example, panic buying (Sarraf et al., 2024), stock market volatility (Olakoyenikan, 2024), and misguided public health that led to higher infection rates (Caceres et al., 2022). Many attempts have been made to automate the detection of fake news. Some of them have tackled the problem in a multilingual setting to leverage the known fake news in one language in detecting potential fake news in other languages (Zhang et al., 2021; Kaliyar et al., 2021; Hasanain and Elsayed, 2022). Despite operating in multilingual settings, most of those approaches rely solely on the textual patterns of claims learned by fine-tuning Pre-Trained Language Models (PLMs) and do not incorporate external facts, which may lead to misrepresentation of the truth.

Several studies have proposed multilingual fact-checking methods that integrate non-English claims and evidence into verification pipelines (Hammouchi and Ghogho, 2022; Dementieva et al., 2022). However, these approaches often strug-

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

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2195–2213

gle due to limited trusted-source evidence in non-English languages. Cross-lingual techniques address this gap by utilizing English as the intermediate language: they verify non-English claims using English resources (Kazemi et al., 2021; Huang et al., 2022; Zhang et al., 2024; Dementieva and Panchenko, 2021; Subramanian et al., 2023). While effective, these pipelines generally lack explanation generation, limiting transparency and user trust.

Some prior work has explored LLM-based veracity explanation generation using zero- and few-shot prompting (Boyina et al., 2024; Kumar et al., 2024; Kasim, 2022; Cekinel et al., 2024), or by generating commonsense and textual explanations (Hu et al., 2024). However, these approaches often did not use grounded factual evidence. Recent methods incorporate grounded evidence (Irnawan et al., 2025; Tan et al., 2025), though primarily for English claims and evidence. Extending such approaches to non-English claims that utilize English evidence must be investigated for cross-lingual misinformation detection.

In this paper, we propose a cross-lingual, explainable fake news detection framework for non-English COVID-19 claims that leverages English evidence through a vectorized Retrieval Augment Generation (RAG) database and employs a multi-agent LLM voting mechanism, as illustrated in Figure 1. Agents here are independent LLM veracity reasoners operating in parallel, with outputs aggregated by majority voting as a lightweight coordination mechanism. The original non-English claim is first translated into English by three current state-of-the-art MT systems—M2M-100 (Fan et al., 2021), LLAMA 3.1 (Feng et al., 2024), and Google Translate. From the three English translations, we choose the best one based on the COMET scores (Chen et al., 2021). Then the best English claim is transformed into three queries/questions (in English) using an LLM.

We integrated the translation component into the system, but did not apply it to the evidence, since machine translation often distorts context, especially for low-resource languages (Nakazawa et al., 2023). To avoid risking unreliable translated evidence, we chose to rely only on original-text evidence, given that evidence credibility and informational integrity are critical.

Each query retrieves the most relevant evidence from the knowledge base. For each query–evidence pair, an LLM generates a veracity prediction and explanation. These pairs form agents in a multi-agent ensemble, and the final claim veracity (Fake or Real) is decided by majority voting across agents. Additionally, the same aforementioned three MT models are used to translate the reasoning of the chosen veracity, in English, back to the original language of the user. Note that, although there are several multilingual LLMs available such as LLAMA 3.1, we chose to adapt a cross-lingual approach, mediated by English, due to the abundant, high-quality English COVID-19 fact resources. In summary, this work makes three contributions:

1. We propose a cross-lingual explainable fake news detection that generates veracity prediction and explanation reasoning for non-English claims by capturing English-written evidence.

2. We introduce a fact-based multi-agent explainable framework that generates veracity predictions along with explanations in the claim of source language, enabling a coherent and interpretable fake-news detection system applicable across diverse languages.

3. We conduct extensive cross-lingual fake news detection experiments on publicly available multilingual non-English COVID-19 datasets and demonstrate that our method performs well against existing cross-lingual approaches in both veracity prediction and explanation generation.

2 Related Work

Recent cross-lingual fake news detection approaches leverage both PLMs and LLMs. It can be classified into two groups: veracity prediction without evidence and veracity prediction with evidence.

Fake News Detection without evidence Veracity prediction without evidence, further categorized into two approaches: PLM and LLM. (Popat et al., 2018; Hasanain and Elsayed, 2022) proposed a PLM-based fake news detection approach that fine-tunes the claim in cross-lingual settings covering five languages and demonstrated that PLMs can achieve reasonably good performance even without external evidence. Although PLMs achieve fair veracity-prediction results using only the claim as input, they do not provide explanatory outputs. Monolingual few-shot and zero-shot prompting techniques address this by adding instructions in

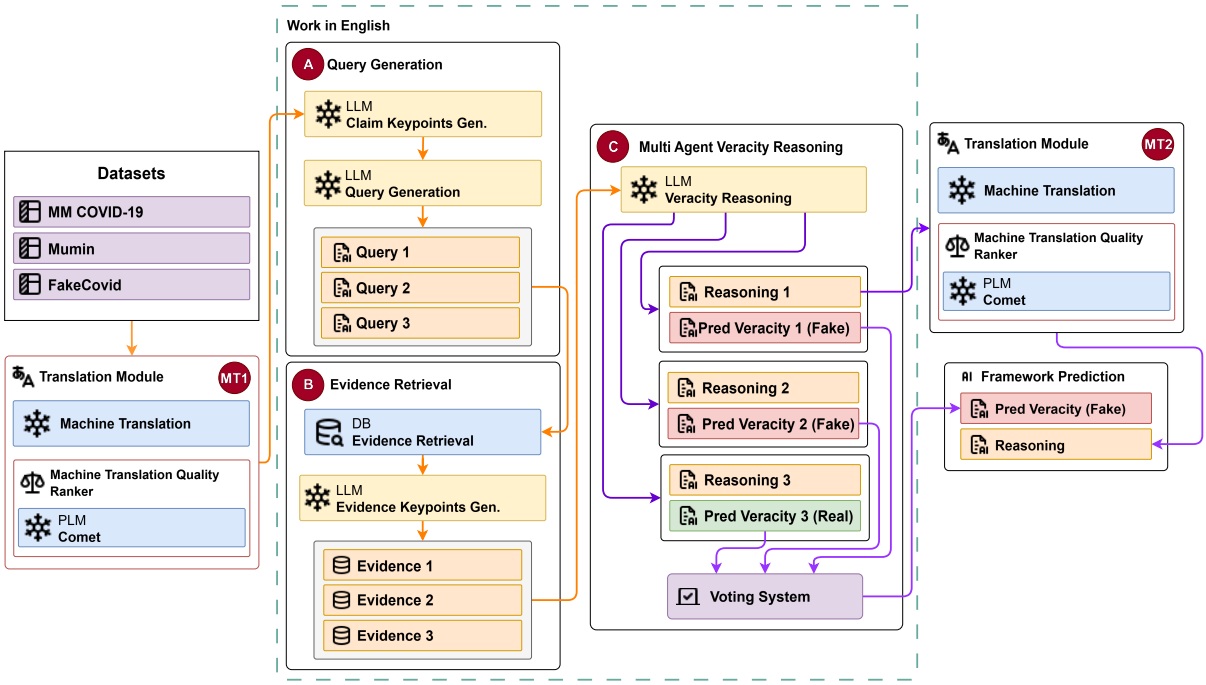

Figure 2: Our framework comprises three LLM-based modules: (A) query generation, (B) evidence retrieval, and (C) multi-agent veracity reasoning, and two machine translation modules (MT1 and MT2) that translate the claim into English and revert the generated English reasoning back into the source claim language, respectively.

the prompt to generate a reasoning explanation (Boyina et al., 2024; Kumar et al., 2024; Kasim, 2022). (Hu et al., 2024) further enhances the approach by using RoBERTa with LLM-generated commonsense and textual justifications. To inject domain-specific knowledge, Cekinel et al. (2024) leverages QLoRA fine-tuning on LLMs, enabling veracity prediction based on newly learned evidence.

Fake News Detection with Evidence. Veracity prediction without factual grounding is unreliable. To address this, several cross-lingual methods incorporate external evidence. (Huang et al., 2022; Dementieva and Panchenko, 2021; Dementieva et al., 2022) process multilingual claims and evidence using a multilingual PLM. Subramanian et al. (2023) further improves retrieval by selecting the evidence with the highest LaBSe multilingual semantic similarity score (Feng et al., 2020) before feeding it into the PLM. Irnawan et al. (2025); Tan et al. (2025) utilize RAG that first retrieves relevant external evidence, which is then passed to an LLM to generate a veracity explanation. This explanation is used by a PLM to predict the veracity. In this work, we adopt their LLM veracity explanation prompting strategy. The detail prompt is described in appendix E.1

3 Cross-Lingual Explainable Fake News Detection

3.1 Overview

The cross-lingual explainable fake news detection framework is illustrated in Figure 2. It consists of three modules: (A) Query Generation, (B) Evidence Retrieval, and (C) Multi-Agent LLM Veracity Reasoning. Module (A) generates the number of i English queries based on the LLM-generated summarized claim keypoints, and Module (B) evidence retrieval uses multiple generated queries to retrieve the evidence from the vector-based evidence database to create three pairs of query-evidence. Lastly, Module (C) multi-agent veracity reasoning generates veracity explanation from three pairs of query-evidence, then a voting system decides the output based on three veracity prediction outputs. Additionally, we attached the Machine Translation method indicated as (MT1), which translates non-English into English, and module (MT2), which translates English to the source claim languages.

3.2 Translation Method

Although the state-of-the-art LLMs support multilingual input, they frequently default to English output, limiting their effectiveness for low-resource languages (Zhang et al., 2023; Guo et al., 2024). To

ensure consistent multilingual claim–explanation generation, we integrate a translation module into our framework. This module leverages three complementary machine translation systems—M2M-100 (Fan et al., 2021), LLAMA 3.1 (Feng et al., 2024), and Google Translate—to reduce the weaknesses of the individual model. Specifically, M2M-100 is a many-to-many encoder–decoder model, LLAMA 3.1 is a decoder-only model used via prompting (details in Appendix E.2), and Google Translate provides online translation via API calls. We select the optimal translation output using COMET (Rei et al., 2020), a metric that evaluates translation quality based on semantic similarity aligned with human judgments.

3.3 Query Generation Modules

As illustrated in Figure 2 module (A), Query Generation Modules consist of two LLMs that generate Claim Keypoints c and the number of i queries q.

Claim Keypoints Generation This module generates claim keypoints c from the English translation of the non-English claim, removing the unnecessary details to retain only the core content and ensure consistent input size and prevent issues with LLMs failing to process overly long claims. The key points represent either a summary of a long claim or a collection of atomic details of a shorter claim extracted by LLMs, and they are stored in a list. Although there are several traditional sentence-truncation or summarization approaches, we decided not to adopt them as they often retain unnecessary information and require ad-hoc tuning of token limits, potentially truncating or distorting the claim (Cuconasu et al., 2024; Hwang et al., 2025).

Question Generation This module generates sentences that support evidence retrieval modules by extracting the fact-assessment questions from the claim. Specifically, the module requires generating at least three questions from the input claim keypoints. We designed it using a few-shot prompting strategy by providing examples of COVID-19 queries that need to be generated by the LLM. The generated query was then ranked based on its textual semantic similarity $cos()$ towards the claim keypoints using the sentence transformers and cosine similarity method (Reimers and Gurevych, 2019). Formally, given a claim keypoints c and multiple non-ranked LLM-generated queries $\cup\overline{Q}$, the queries $\cup\overline{Q}$ consist of multiple LLM generated queries $\cup Q = [q_1, q_2, \cdots q_n]$. Let $Q = [q_1, q_2, \cdots, q_i]$ be a set of queries ranked based on their similarities to the claim, and i represents the number of top-ranked queries used in the evidence retrieval modules.

$$Q = \left[cos(c, \cup Q)\right]_i, \tag{1}$$

3.4 Evidence Retrieval Modules

As illustrated in Figure 2 module (B), Evidence Retrieval modules consist of one vector-semantic evidence database that provides the evidence and one LLM that summarizes the evidence in a compact format.

Evidence Retrieval Evidence Retrieval aims to retrieve evidence based on the input query. The evidence is the English-written facts collected from factual, credible, and open-access sources. The FAISS (Douze et al., 2024; Johnson et al., 2019) and sentence-embedding vector library provided by the sentence transformers library (Reimers and Gurevych, 2019) are applied to the evidence data.

Translating evidence is one possible direction for evidence. However, machine translation often distorts context (Nakazawa et al., 2023), especially for low-resource languages, thereby jeopardizing the credibility of the evidence when it is most critical. Instead, we choose to preserve the evidence in its original English form.

The i generated queries were sent to the Evidence Retrieval sub-module $db()$ to collect i evidence $EV = [ev_1, ev_2, \cdots ev_i]$, where i is the maximum number of ranked queries used in this module. For one query q_i, we retrieve one evidence ev_i by ranking multiple candidate evidence using cosine similarity (Reimers and Gurevych, 2019) between the embeddings of the query and the claim and the retrieved evidence, where the embeddings are derived from XLM-RoBERTa (Conneau et al., 2019) multi-lingual PLM. The reason for choosing the one-to-one query evidence design is to avoid evidence redundancy and overlapping that may mislead the reasoning (Cuconasu et al., 2024). The example of a one-to-one query evidence design choice can be found in the Appendix D.

$$ev_i = [sim(db(c, q_i))]_1, \tag{2}$$

where ev_i refers to one of the evidence from a set of evidence EV for one query q. The evidence ranker defined in equation 2 is applied iteratively for each of the top i generated queries to construct the evidence set $EV = [ev_1, ev_2, \cdots ev_i]$. After

the evidence is collected, evidence keypoints $\hat{ev}_n$ for each collected evidence are generated and then applied to make the evidence more compact.

Evidence Keypoints Generation uses an LLM to compress multiple retrieved evidence items EV, often long, redundant, or cluttered, into concise keypoint-style summaries. This allows the LLM to do veracity reasoning more efficiently during veracity inference. It also filters noise such as redundancy, typos, and merged sentences.

3.5 Multi-Agent Veracity Reasoning

Multi-agent veracity reasoning leverages diverse evidence and reasoning perspectives, helping to reduce bias associated with single-veracity reasoning bias and enhancing the overall robustness of claim assessment. In our approach, multiple LLMs operate concurrently, doing veracity reasoning on the claim keypoints c with i number of queries $Q = [q_1, q_2, \cdots q_i]$ and evidence keypoints $\hat{EV} = [\hat{ev}_1, \hat{ev}_2, \cdots \hat{ev}_i]$. The i number of veracity prediction outputs will be used by the voting system to decide the veracity of the input claim.

Evidence-Based Veracity Reasoning works by doing veracity reasoning to predict whether the claim is real or fake and generates veracity explanation on a claim's keypoints c with i groups of queries Q and evidence keypoints $\hat{EV}$, where i is the number of ranked queries and evidence used in the module. We employ LLAMA 3.1 (8B) on all multi-agent LLMs. claim's keypoints c and each of the query q_i and evidence keypoints $\hat{ev}_i$ will be applied to the veracity reasoning prompt and sent to the LLM to assess the claim c whether the evidence $\hat{ev}_n$ is supported or refuted. If the evidence debunks the claim, most likely the claim is fake, and when the evidence supports the claim, most likely the claim is real. We formulate the claim veracity reasoning as follows.

$$\hat{y_{dbi}}, \hat{e_{dbi}} = LLM_{db}(c', q_i, \hat{ev}_i), \quad (3)$$

where $\hat{y}_{db}$ refers to the evidence-based veracity prediction and $\hat{e}_{db}$ indicates the evidence-based explanation reasoning generated by the evidence-based veracity-reasoning LLM $LLM_{db}()$. Both $\hat{y_{dbi}}$ and $\hat{e_{dbi}}$ only represent the veracity prediction and explanation for a query q and an evidence keypoint $\hat{ev}$. Veracity predictions and explanations are generated for each of the i query and evidence, producing multi-agent outputs $y_{db}, e_{db} = [[\hat{y_{db1}}, \hat{e_{db1}}], [\hat{y_{db2}}, \hat{e_{db2}}], \cdots [\hat{y_{dbi}}, \hat{e_{dbi}}]]$. The multi-agent veracity reasoning outputs y_{db}, e_{db} will be passed to the voting system to decide the final veracity predictions.

Datasets	Num. Claim
Training Dataset	
MMCoVaR (Chen et al., 2021)	2,593
ReCOVery (Zhou et al., 2020)	2,029
Total	4,622
Testing Dataset	
MM COVID-19 (Li et al., 2020)	5001
MuMiN (Nielsen and McConville, 2022)	2,897
FakeCovid (Shahi and Nandini, 2020)	7,723
Total	14,079

Table 1: The statistics of the training dataset consisting of two English-language COVID-19 datasets, and the test dataset which includes three Multilingual COVID-19 datasets after removing the English portion to facilitate cross-lingual experiments.

The voting system The voting system mitigates bias that may happen in an LLM veracity reasoning by aggregating the multi-agent veracity reasoning outputs into one final veracity prediction. Specifically, we apply majority voting over the three individual veracity prediction outputs $[\hat{y_{db1}}, \hat{y_{db2}}, \cdots \hat{y_{dbi}}]$. For example, if two predict 'fake' and one predicts 'real', the ensemble decision is 'fake'. Predictions with unknown veracity (attributable to unreliable evidence) are excluded from the voting process; only the remaining predictions are considered in the vote.

We chose not to use a PLM for weighted prediction (as suggested in (Irnawan et al., 2025)) because its performance under cross-lingual fine-tuning remains unsatisfactory. Instead, we adopt a binary voting system, which offers greater simplicity and robustness across a variety of languages. The minority-veto strategy proposed by Jain et al. (2025) could have been incorporated into our framework, yet we chose not to implement it because it risked discarding credible evidence and may lead to premature rejection.

4 Experiments

4.1 Experimental Setup

We use three publicly accessible multilingual COVID-19 fake news datasets: MM COVID-19 (Li et al., 2020), MuMin (Nielsen and McConville, 2022), and Fake Covid (Shahi and Nandini, 2020)

Source	Num. Evidence
NIH	1,131
CDC	11,823
LitCOVID 19	407,982
Politifact	2,038
CORD-19	368,618
Total	791,592

Table 2: The statistics of English language evidence resources

to evaluate the proposed framework. To make the dataset cross-lingual, the English language in all three datasets is removed. The breakdowns of the data sizes are shown in Table 2 in the Testing data section. The baseline methods that are based on a trainable model, including the method proposed by (Hasanain and Elsayed, 2022; Huang et al., 2022; Subramanian et al., 2023; Hu et al., 2024; Cekinel et al., 2024; Irnawan et al., 2025), use two English-written COVID-19 datasets: MMCoVaR (Chen et al., 2021) and ReCOVery (Zhou et al., 2020). The breakdowns of the data sizes are shown in Table 2 in the Training data section.

We compile the evidence from peer-review COVID-19 medical publications (NIH, CDC, CORD-19, LitCOVID) and fact-check news-sites (e.g., PolitiFact) described in Table 2, along with annotated claim–context–veracity data from English fake news datasets listed in the Training section Table 1, and store all entries in a scalable FAISS (Douze et al., 2024; Johnson et al., 2019) vector database. The framework is implemented in Python 3.10 using the open-source LLAMA 3.1 (8B) LLM for efficiency and privacy protection, and evaluated on a multi-core CPU, with 128 GB RAM and NVIDIA RTX 6000Ada GPU.

We empirically set the number of generated queries and retrieved evidence to three results in three veracity predictions and explanations. The details are in the appendix C. For fair comparison, we utilize MT on baselines that work on English only. We evaluate veracity prediction accuracy using three classification-based metrics: macro-precision, macro-recall, and macro-F1. To assess the quality of generated explanations, we employ two metrics: BERTScore (Zhang et al., 2020) with XLM-RoBERTa embeddings (Conneau et al., 2019), enabling multilingual semantic comparison between the claim and its explanation, and ChrF (Popović, 2015), which measures character-level n-gram overlap to quantify textual similarity.

4.2 Baselines

To evaluate our proposed method, we use identical evidence corpora across all experiments. We compare our model against several baselines, which we classify into two groups:

Claim's Veracity Prediction consists of ten baselines: (1) **XLM-RoBERTa**, a multilingual PLM-based classifier (Hasanain and Elsayed, 2022); (2) **Zero-shot** and (3) **Few-shot** prompting strategy by (Boyina et al., 2024); (4) **Commonsense** and (5) **Textual Description** modules by (Hu et al., 2024), which apply LLM-based logical reasoning directly on claims; (6) **Bad Actor Good Advisor** (Hu et al., 2024), which integrates PLM and LLM reasoning; (7) **FCTR** by (Cekinel et al., 2024) utilizing LLM fine-tuning; (8) **CONCRETE** (Huang et al., 2022), which leverages translation and cross-lingual evidence retrieval; (9) **Cross-Lingual FC** (Subramanian et al., 2023), which combines multilingual semantic similarity with evidence retrieval; and (10) **Covid EFND** (Irnawan et al., 2025), which uses RAG-based veracity reasoning paired with PLM inference.

Claim's Veracity Explanation Generation. Consist of four baselines: (1) **Zero-/Few-Shot** Prompting using LLM prompts (Boyina et al., 2024). (2) **Bad Actor Good Advisor**, which applies commonsense and textual reasoning (Hu et al., 2024). (3) **FCTR**, which fine-tunes an LLM via QLoRA for explanation generation in cross-lingual settings (Cekinel et al., 2024). (4) **Covid EFND**, a RAG-based approach combining retrieval and PLM reasoning for veracity explanations in COVID-19 contexts (Irnawan et al., 2025).

4.3 Results

4.3.1 Fake News Detection

As shown in Table 3, our framework performs better veracity prediction than most baseline methods. The only exception is on the Fake COVID dataset, where our F1-macro score is 0.5% lower than the method by (Hu et al., 2024), and precision is 3.2% lower than the method by (Subramanian et al., 2023).

Our method outperforms the non-evidence-based fine-tuning approach by (Hasanain and Elsayed, 2022), with F1-macro improvements of +45.4% on MM COVID-19, +46% on MuMiN, and +43.9% on Fake COVID, highlighting the effectiveness of incorporating evidence in veracity prediction. Compared to non-evidence-based LLM

Methods	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec
	MM COVID-19			MuMIN			Fake Covid		
XLM RoBERTa (Hasanain and Elsayed, 2022)	0.328	0.244	0.500	0.037	0.022	0.118	0.009	0.005	0.031
Zero-Shot Reasoning (Boyina et al., 2024)	0.672	0.727	0.689	0.340	0.357	0.437	0.329	0.343	0.444
Few-Shot Reasoning (Boyina et al., 2024)	0.648	0.701	0.667	0.333	0.354	0.427	0.325	0.325	0.494
Commonsense (Hu et al., 2024)	0.460	0.468	0.457	0.299	0.350	0.376	0.147	0.335	0.191
Textual Desc. (Hu et al., 2024)	0.445	0.450	0.450	0.311	0.349	0.412	0.169	0.334	0.267
Bad Actor Good Advisor (Hu et al., 2024)	0.665	0.500	0.251	0.434	0.500	0.385	**0.453**	0.500	0.415
FCTR (Cekinel et al., 2024)	0.333	0.249	0.500	0.489	0.478	0.500	0.323	0.329	0.318
CONCRETE (Huang et al., 2022)	0.339	0.256	0.500	0.043	0.022	0.500	0.012	0.506	0.500
Cross-Lingual FC (Subramanian et al., 2023)	0.605	0.643	0.646	0.279	0.456	0.253	0.413	**0.537**	0.357
Covid EFND (Irnawan et al., 2025)	0.167	0.462	0.106	0.285	0.500	0.208	0.439	0.522	0.385
Full Framework (Ours)	**0.782**	**0.783**	**0.782**	**0.497**	**0.509**	**0.530**	0.448	0.505	**0.590**

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

reasoning methods by (Boyina et al., 2024) and (Hu et al., 2024), our approach achieves an average F1-macro of +22.6% on MM COVID-19, +17.6% on MuMiN, and +20.6% on Fake COVID, indicating that external evidence remains critical even when leveraging LLM internal knowledge. Furthermore, against evidence-based fine-tuning methods by (Huang et al., 2022; Subramanian et al., 2023), our method yields average improvements of +31.0% on MM COVID-19, +33.6% on MuMiN, and +23.6% on Fake COVID, suggesting that incorporating LLM-based reasoning in addition to evidence is essential for achieving accurate prediction. Our method outperforms the LLM fine-tuning approach proposed by (Cekinel et al., 2024), with F1-macro improvement of +44.9% on MM COVID-19, +0.08% on MuMiN, and +12.5% on Fake Covid. These results suggest that leveraging external evidence through retrieval-augmented generation (RAG) is more effective for cross-lingual veracity prediction than relying solely on knowledge embedded via LLM fine-tuning. Compared to (Hu et al., 2024), our method achieves higher F1-macro scores by +11.7% on MM COVID-19 and +6.3% on MuMiN, and is narrowly outperformed on Fake COVID by just 0.5%. This indicates that combining two LLM commonsense and textual descriptions reasoning for PLM fine-tuning remains insufficient for achieving optimal accuracy. Our method outperformed that (Irnawan et al., 2025) approach employ RAG, veracity reasoning, and PLM fine-tuning by +61.5% on MM COVID-19, +21.2% on MuMiN, and +0.9% on Fake Covid, showing that in crosslingual settings, evidence-based reasoning and PLM fine-tuning with this approach is not suitable. Instead, multi-agent, which relies on a non-fine-tuning approach.

Lastly, our method significantly outperforms the method proposed by (Irnawan et al., 2025), which combines RAG, veracity reasoning, and PLM fine-tuning, achieving F1-macro gains of +61.5% on MM COVID-19, +21.2% on MuMiN, and +0.9% on Fake COVID. These improvements demonstrate that, in cross-lingual settings, conventional evidence-based reasoning with PLM fine-tuning is less effective. Instead, our results indicate that a multi-agent, non-fine-tuning approach is more suitable. Furthermore, To assess whether our framework generalizes across languages, we also evaluate our framework on four languages: French, Spanish, Hindi, and Arabic, which include both alphabetic and non-alphabetic scripts. Detailed experimental results can be found in the Appendix B

4.3.2 Generating Explanation

As shown in Table 4, our framework achieves better quality compared to most baselines. The only exception is the Fake COVID dataset, where our BERTScore-based F1 is 0.2% lower and ChrF is 21.149 points lower than the method proposed by (Boyina et al., 2024). When investigating veracity prediction shown in Table 3, our method exceeds (Boyina et al., 2024) by +12.1% F1-macro. This suggests that even when explanations accurately reflect the claim, they may still contain hallucinations, which can reduce veracity prediction accuracy compared to our approach.

As shown in Table 4, our method outperforms the commonsense and textual-description reason-

Methods	F1 BERT	CHR F	F1 BERT	CHR F	F1 BERT	CHR F
	MM COVID-19		MuMiN		Fake Covid	
Bad Actor Good Advisor (Hu et al., 2024)						
a. Commonsense	0.731	0.000	0.727	0.005	0.727	0.000
b. Textual Desc.	0.728	0.000	0.723	0.019	0.725	0.000
Zero and Few Shot (Boyina et al., 2024)						
a. Zero Shot	0.783	18.760	0.840	25.152	0.824	21,648
b. Few Shot	0.781	17.137	0.841	26,156	**0.827**	**27.287**
FCTR (Cekinel et al., 2024)	0.786	0.000	0.786	0.001	0.761	0.000
COVID EFND (Irnawan et al., 2025)	0.748	0.063	0.744	5.150	0.733	0.266
Ours	**0.847**	**20.725**	**0.856**	**27.081**	0.825	6.138

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

ing proposed by (Hu et al., 2024), achieving F1-BERT improvements of +12.1% on MM COVID-19, +13% on MuMiN, and +9.9% on FakeCovi. These indicate that incorporating evidence when generating veracity explanations enables more accurate and comprehensive claim representation. Furthermore, we find that fine-tuning LLMs, as proposed by (Cekinel et al., 2024), remains insufficient to produce veracity explanations. In contrast, leveraging evidence yields explanations that more effectively integrate the claim and its related evidence.

When compared with the RAG-based veracity approach proposed by (Irnawan et al., 2025), which integrates RAG-evidence retrieval and PLM fine-tuning to choose among commonsense, textual, and evidence-based reasoning, our method yields higher F1-BERT scores: +9.9% on MM COVID-19, +11.1% on MuMiN, and +9.2% on Fake COVID. These results indicate that (Irnawan et al., 2025) method is less effective in multilingual settings. In contrast, our multi-agent system, featuring multiple LLMs performing evidence-based reasoning without PLM fine-tuning, can choose more accurate veracity based on multiple evidence-based veracity-reasoning.

4.4 Ablation Study

To better understand which module contributes the most within our framework, we conducted an ablation study using the MM COVID-19 dataset by systematically removing or combining components. We categorized the analysis into two:

Contribution of each Module We evaluated seven configurations based on three primary LLM modules (query-generation, evidence-retrieval, and multi-agent veracity reasoning) and additionally assessed the removal of claim and evidence keypoint modules to determine their effect on veracity prediction. Table 5 indicates that the configuration combining both claim and evidence keypoints within the multi-agent framework achieved the best performance with F1-macro = 0.782, significantly outperforming alternatives that either omitted keypoints or used non-multi-agent reasoning, thus demonstrating the crucial contribution of both components. Lastly, Table 6 shows a framework that incorporates multi-agent veracity reasoning achieves a +9% improvement over a non-multi-agent framework (linear LLM reasoning: w/o CK and MA, w/o EK and MA, and w/o MA). The gains are consistent across configurations for both Claim Keypoints (w/o EK) +8% and Evidence Keypoints (w/o CK) +1%. These results highlight the benefit of incorporating the multi-agent veracity reasoning in the architecture over linear-veracity-reasoning.

Methods	F1	Prec	Rec
w/o EK and MA	0.679	0.728	0.690
w/o CK and MA	0.658	0.750	0.682
w/o CK and EK	0.729	0.776	0.739
w/o CK	0.761	0.780	0.767
w/o EK	0.761	0.774	0.764
w/o MA	0.691	0.731	0.702
Full (CK + EK + MA)	**0.782**	**0.783**	**0.782**

Table 5: The ablation study of our framework on MM COVID-19 datasets. CK = Claim Keypoints, EK = Evidence Keypoints, and MA = Multi-Agent Veracity reasoning. **Bold** is best, underline is second best.

Impact of Translation module Table 6 shows that implementing full translation (Full-TL) in our framework increases claim–explanation language consistency by 99% and improves evidence align-

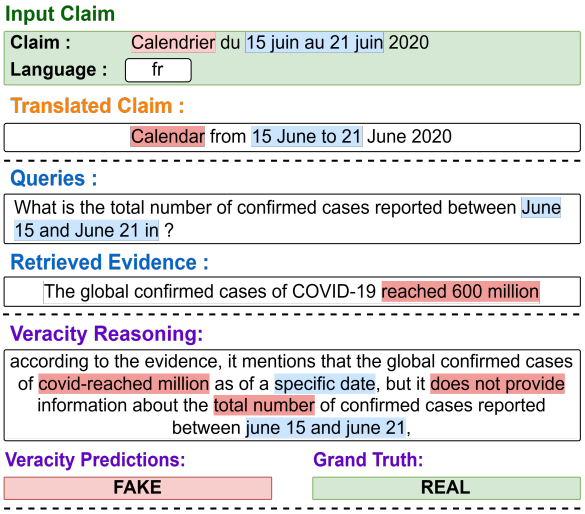

Figure 3: Example of an error on the meaningless COVID-19 short claim. result in a meaningless query, hallucinated queries, and incorrect veracity prediction.

ment—measured via F1-BERT semantic similarity—by about 7% across all datasets. Although Full-TL slightly lowers veracity-prediction performance (macro-F1 drops by approximately 5%), it substantially enhances language alignment. When translation is applied only to the explanation module (w/o R-TL), performance suffers by around 3% on explanation metrics and 7% on retrieval metrics. Conversely, translation applied only to the retrieval module (w/o E-TL) boosts macro-F1 by about 7%, but fails almost entirely in claim–explanation language alignment (99%) and shows only a slight (1%) gain in explanation quality. In sum, translation incurs a modest cost in classification accuracy but delivers significant improvements in language alignment, explanation fidelity, and evidence credibility.

4.5 Error Analyses

We conducted error analyses on the MM COVID-19 (Li et al., 2020), MuMiN (Nielsen and McConville, 2022), and Fake Covid (Shahi and Nandini, 2020) datasets. There are three major types of errors in the veracity explanation generation:

Meaningless and Short Claim: As illustrated in Figure 3, short or meaningless claims impair effective query generation for evidence retrieval, which can lead to hallucinations and incorrect veracity judgments.

Translation Method: Although local machine translation is faster and works offline, its quality is sometimes inferior to online translation. Reliance on online tools introduces latency and inconsistency in veracity performance.

Regional Language. Languages unsupported by translation systems (e.g, regional language) lead to mistranslations, which in turn degrade evidence retrieval and reasoning accuracy, causing misrepresentation of veracity.

Methods and Datasets	Retrieval	Explanation		
	F1 BERT	Lang Acc	F1-BERT	macro-F1
MM COVID-19				
w/o R-TL	0.835	**0.994**	**0.853**	**0.790**
w/o E-TL	**0.840**	0.002	0.850	0.782
No-TL	0.840	0.000	0.850	**0.790**
Full-TL (Ours)	**0.835**	**0.994**	0.848	0.782
MuMiN				
w/o R-TL	0.805	**0.994**	0.850	**0.500**
w/o E-TL	**0.811**	0.003	0.837	0.497
No-TL	0.805	0.003	0.839	**0.500**
Full-TL (Ours)	**0.811**	**0.994**	**0.856**	0.497
Fake Covid				
w/o R-TL	0.791	**0.965**	0.823	0.425
w/o E-TL	**0.801**	0.046	**0.825**	**0.448**
No-TL	0.791	0.032	0.832	0.425
Full-TL (Ours)	**0.801**	**0.965**	**0.825**	**0.448**

Table 6: Ablation study on the translation-module contribution in the overall framework. R-TL denotes translation applied in the retrieval step; E-TL denotes translation applied in explanation generation. "No-TL" means no translation is applied; "Full-TL" means translation is applied in both steps. **Bold** indicates the best result and underline indicates the second best.

5 Conclusion

In this paper, we proposed a cross-lingual fake news detection framework that leverages grounded factual evidence to generate veracity explanations in the claim's source language. The experimental results showed that our approach achieves strong performance across three multilingual COVID-19 datasets. Future work includes removing translation dependencies, expanding multilingual evidence coverage via LLMs fine-tuning, and improving support for underrepresented regional languages.

Limitations

Our framework has three notable limitations: it cannot reliably translate regional or low-resource languages, leading to mistranslations that degrade the reasoning accuracy. It struggles to generate coherent veracity explanations for very short or semantically vague claims, resulting in hallucinations and incorrect predictions. The reliance on multiple translation steps and quality validations, often switching between offline and online tools, signifi-

cantly increases processing time for each claim.

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.

Acknowledgments

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. Bassamtiano Renaufalgi Irnamwan is funded by the MEXT scholarship, Grant Number 233203.

References

Kamal Boyina, Gujja Manaswi Reddy, Gunturi Akshita, and Priyanka C Nair. 2024. Zero-shot and few-shot learning for telugu news classification: A large language model approach. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

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, and 1 others. 2022. The impact of misinformation on the covid-19 pandemic. *AIMS public health*, 9(2):262.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2024. Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4127–4142, Torino, Italia. ELRA and ICCL.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 719–729, New York, NY, USA. Association for Computing Machinery.

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.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement. *arXiv preprint arXiv:2402.16379*.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2024. Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms. *arXiv preprint arXiv:2410.15956*.

Hicham Hammouchi and Mounir Ghogho. 2022. Evidence-aware multilingual fake news detection. *Ieee Access*, 10:116808–116818.

Maram Hasanain and Tamer Elsayed. 2022. Cross-lingual transfer learning for check-worthy claim identification over twitter. *arXiv preprint arXiv:2211.05087*.

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.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C. Park. 2025. EXIT: Context-aware extractive compression for enhancing retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4895–4924, Vienna, Austria. Association for Computational Linguistics.

Bassamtiano Renaufalgi Irnawan, Sheng Xu, Noriko Tomuro, Fumiyo Fukumoto, and Yoshimi Suzuki. 2025. Claim veracity assessment for explainable fake news detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4011–4029.

Suryaansh Jain, Umair Z Ahmed, Shubham Sahai, and Ben Leong. 2025. Beyond consensus: Mitigating the agreeableness bias in llm judge evaluations. *arXiv preprint arXiv:2510.11822*.

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.

Samra Kasim. 2022. One true pairing: evaluating effective language pairings for fake news detection employing zero-shot cross-lingual transfer. In *International Conference on Soft Computing and its Engineering Applications*, pages 17–28. Springer.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A. Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

Raghvendra Kumar, Bhargav Goddu, Sriparna Saha, and Adam Jatowt. 2024. Silver lining in the fake news cloud: Can large language models help detect misinformation? *IEEE Transactions on Artificial Intelligence*.

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.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondřej Bojar, Akiko Eriguchi, Yusuke Oda, Chenhui Chu, and Sadao Kurohashi. 2023. Overview of the 10th workshop on Asian translation. In *Proceedings of the 10th Workshop on Asian Translation*, pages 1–28, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3141–3153.

Oluwasegun Olakoyenikan. 2024. The economic consequences of misinformation an analysis of the impact of fake news on stock market volatility during the covid19 pandemicl international journal of innovative science and research technology. *International Journal of Innovative Science and Research Technology (IJISRT)*, pages 667–674.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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.

Shagun Sarraf, Amit Kumar Kushwaha, Arpan Kumar Kar, Yogesh K Dwivedi, and Mihalis Giannakis. 2024. How did online misinformation impact stockouts in the e-commerce supply chain during covid-19–a mixed methods study. *International Journal of Production Economics*, 267:109064.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

Shivansh Subramanian, Ankita Maity, Aakash Jain, Bhavyajeet Singh, Harshit Gupta, Lakshya Khanna, and Vasudeva Varma. 2023. Cross-lingual fact checking: Automated extraction and verification of information from Wikipedia using references. In *Proceedings of the 20th International Conference on Natural*

Language Processing (ICON), pages 828–831, Goa University, Goa, India. NLP Association of India (NLPAI).

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving explainable fact-checking with claim-evidence correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1600–1612, Abu Dhabi, UAE. Association for Computational Linguistics.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024. Do we need language-specific fact-checking models? the case of Chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1914, Miami, Florida, USA. Association for Computational Linguistics.

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

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatgpt when your question is not in english: a study of multilingual abilities and types of llms. *arXiv preprint arXiv:2305.16339*.

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.

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.

Appendix

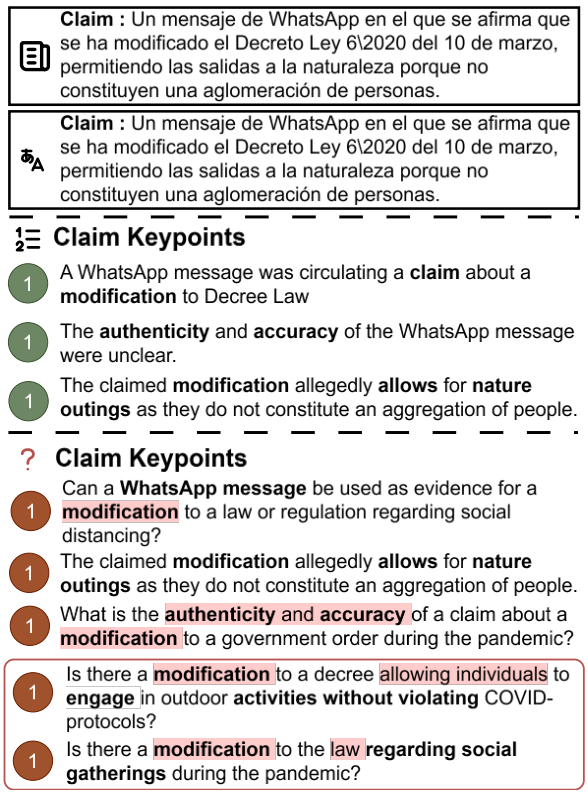

Figure 5: Analysis on Three and Five Query Generation

A Example of Each Proposed Method outputs

Figure 4 illustrates our modular pipeline example output: (1) translate the claim, (2) extract claim keypoints, (3) generate three queries, (4) retrieve evidence for each query, (5) extract keypoints from each retrieved evidence, (6) perform multi-agent veracity reasoning, each agent uses the translated claim, all three generated queries, and its corresponding evidence to produce three prediction and explanation. (7) Use majority voting among the three agents to determine the final veracity label. Lastly, (8) select one of the agent explanations matching the final label and translate it back into the original claim language and present it together with the final veracity label.

B Cross-lingual Experiments by Languages

Table 7 shows the experimental results of the generalization performance of our proposed framework towards alphabetical and non-alphabetical based languages against three baselines: Zeroshot (Boyina et al., 2024), COVID-EFND (Irnawan et al., 2025), and Cross-LingualFC (Subramanian et al., 2023). The results show strong macro-F1 gains on alphabetical languages by +17% with drops on non-alphabetical ones by -0.06%. This highlights effectiveness in cross-lingual settings while leaving further study on non-alphabetical claims as future work.

Methods and Datasets	Language			
	Spanish	French	Hindi	Arabic
MM COVID-19				
Zeroshot	0.649	0.591	**0.443**	0.463
COVID-EFND	0.143	0.143	0.197	0.299
Cross-LingualFC	0.507	0.548	0.411	**0.498**
Ours	**0.805**	**0.694**	0.441	0.440
MuMiN				
Zeroshot	0.346	0.463	0.423	0.446
COVID-EFND	0.261	0.294	0.158	0.314
Cross-LingualFC	0.275	0.267	0.179	0.304
Ours	**0.510**	**0.517**	**0.475**	**0.501**
Fake Covid				
Zeroshot	0.354	0.301	0.489	**1.000**
COVID-EFND	0.442	0.383	0.447	1.000
Cross-LingualFC	0.395	0.291	**0.756**	0.524
Ours	**0.458**	**0.465**	0.431	0.434

Table 7: The classification performance of our proposed method on three datasets against ten baseline models. **Bold** is the best, underline is the second best.

C Analysis on the Suitable Number for Queries

Figure 5 shows that using exactly three queries effectively covers all aspects of the claim, whereas generating five queries often leads to redundancy. queries four and five typically mirror earlier queries and focus again on regulatory modifications related to outdoor activity. This overlap results in redundant evidence and irrelevant information, ultimately degrading performance. Therefore, three queries achieve both robust multi-agent aggregation and efficient, high-quality evidence retrieval. In addition, we set the number of generated queries to three because our multi-agent LLM predicts claim veracity by aggregating assessments from multiple queries, using an even number raises the risk of tied veracity prediction decisions.

D Analysis on one claim-query-evidence design

Figure 6 illustrates that utilizing a single query to retrieve multiple pieces of evidence can return both

redundant and off-topic items. Evidence one (E1) and evidence three (E3) are redundant, while evidence two (E2) discusses the UK social distancing policy and is unrelated to the query about misinformation circulating on WhatsApp. This mixture can mislead the veracity-reasoning module and degrade prediction accuracy.

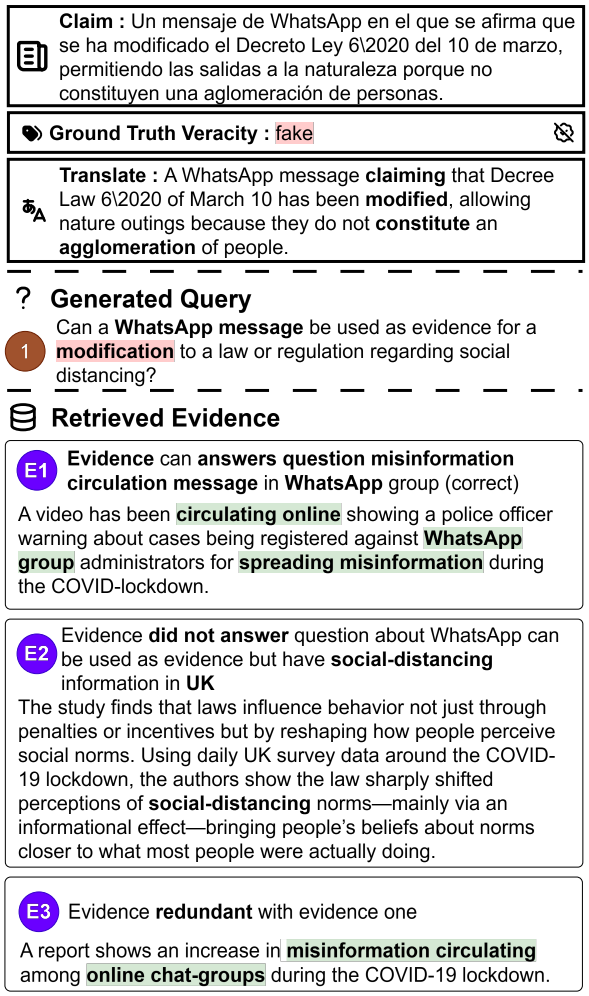

Figure 6: Analysis on One Query Multiple Evidence

E Framework Prompts

E.1 Evidence Keypoints Prompt

The following is the prompt applied in all modules of our framework as illustrated in Figure 2, including: (1) Figure 7, which is utilized in LLM Claim Keypoints Generation in modules A and LLM Evidence Keypoints Generation in modules B; (2) Figure 8 that illustrates the prompt use in Query Generation module A; and (3) Figure 9 that illustrate the prompt to generate multiple evidence-based veracity-explanation-reasoning use in Multi-Agent Veracity Reasoning module (c).

E.2 LLM Translation Prompt

Figure 10 illustrates the prompt used in our LLM translation module, assigned module **MT1** and **MT2** in Figure 2. The prompt variables are defined as follows: $lang\_src$ denotes the source language, $lang\_tgt$ specifies the target language for translation, and $text$ represents the input sentence to be translated.

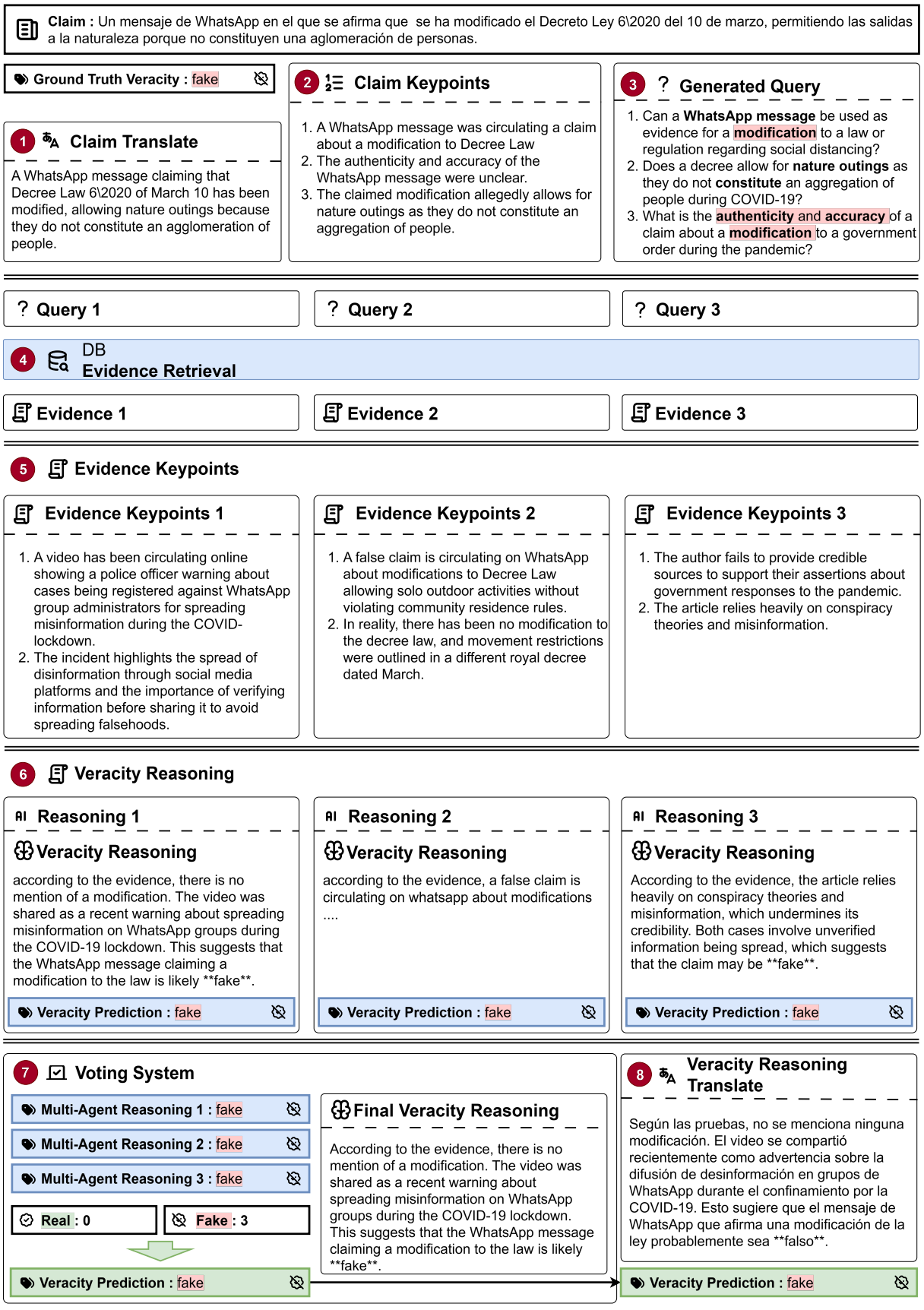

Figure 4: The example of each proposed method module outputs

```
<|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 want you to generate the keypoint summarization of the input sentence.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
    Your task is to create a concise summary of the long article by listing its key points. Each
key point should be listed on a new line and numbered sequentially.

    ### Requirements:
    - The key points should be brief and focus on the main ideas or events.
    - Ensure that each key point captures the most critical and relevant information from the
article.
    - Maintain clarity and coherence, making sure the summary effectively conveys the essence of
the article.

    The following is the article:
    {claim}
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 7: Prompt used on Claim and Evidence Keypoints generation

```
<|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 examples of patterns 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!.
        I want you to create a query by considering the key points and the the COVID-relatedinput
 claim.
        Remove any word that mentions any region, person name, country name, state name, province
 name, and time in your generated queries.
        Avoid utilizing abbreviations when creating the query; instead, use the full version of
the abbreviated word, for example:
        1. J&J is Johnson & Johnson,
        2. CDC is the Centers for Disease Control and Prevention
        3. NIH is the National Institute of Health, etc.

        The following is an example of a query that you use to verify the claim.
        - Claim: Oregon health authorities have reported that federal officials are investigating
 the death of a woman who developed a rare blood clot and low platelets after receiving the
Johnson & Johnson COVID-19 vaccine.
        - Query:
        1. What vaccine is associated with a woman's death in Oregon due to a clot?
        - Claim: The woman received the dose before the CDC ordered a pause on the vaccine due to
 concerns about dangerous clots, but it's unclear if her death is related to the vaccine until
the investigation is complete.
        - Query:
        1. Did the woman receive the vaccine before the Centers for Disease Control and
Prevention ordered a pause on the 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|>
    I want you to create a concise covid-19 related question with a maximum of 5 queries with
each of them have fifteen words from the following paragraphs:
    {claim}.
    Refrain from use abbreviations while crafting the question; instead, utilize the complete
form of the abbreviated term, for instance:
    1. J&J is Johnson & Johnson,
    2. CDC is Centers for Disease Control and Prevention
    3. NIH is National Institute of Health, etc.
    The main topic of the sentence are as follows:
    {claim_keywords_string}

    don't 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: Prompt used on Query Generation Modules

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
    You are a helpful, respectful, and honest assistant. Your goal is to provide the most helpful
 and unbiased response possible.
    Ensure that your responses remain socially unbiased and positive.
    The following content is to help us to proof the claim whether it is fake or real news.
    I will provide a claim and context related to COVID-19. Your task is to verify the
information and determine whether the claim is **fake**, **real**, or **undecided**, based on
the given evidence and common sense.
    Your response must be one of these three options: fake, real, or undecided. Do not use vague
classifications such as "partially fake" or "partially real."
    You must follow the response pattern from the example provided below. Do not include
additional examples, prompts, or user input in your response.

    1. Claim: Oregon: CDC investigating woman's death after J&J vaccine
    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 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.
    2. I check the claim with the following query: death investigation of a woman after receiving
 Johnson & Johnson COVID-vaccine
    3. From this source: https://www.cdc.gov/mmwr/volumes/70/wr/mm7018e2.htm
    4. 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.
    Core prevention strategies include staying up-to-date with COVID-19 vaccines, practicing good
 hygiene, wearing masks, maintaining social distancing, and staying home when sick.
    Vaccination reduces the risk of getting sick, hospitalization, or death from COVID-19.
    5. 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.
    6. 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.
    You must follow the following format when generating your response:

    3. From this source: [you put the provided evidence_url here!].
    4. I found this Evidence : [you put the evidence_summary here!].
    5. By Comparing the evidence and claim: According to the <|you put your evidence here|>, <|
you put your analysis why the news / claim and context is **fake**, **real**, or **undecided**
here!|>
    6. Therefore, the news is <|your final decision based on your analysis on the news whether is
 **fake**, **real**, or **undecided!**. The answer must be either **fake**, **real**, or **
undecided!**, and you don`t need to describe your decision here!|>
    [you must end your response here!]

    1. Claim: {claim}
    2. I check the claim with the following query: {query}
    3. From this source: {evidence_url}
    4. I found this evidence: {evidence_text}
    5. By comparing the evidence and claim :
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 9: Prompt used on Evidence-Based Veracity-Explanation-Reasoning module

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
    You are a helpful, respectful, and honest assistant. Your goal is to provide the most helpful
 , correct, and unbiased response possible.
    I want you to translate the input sentence into the target language.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
    I want you to translate the following sentence from {lang_src} to {lang_tgt}.
    Please do not include the opening word in your response, such as I am happy to help.
    Make sure your translated response is in {lang_tgt} language and do not respond in English if
  the language target is not English.
    Response only to the translated version of the text.

    The following is the sentence that I want you to translate:
    {text}
    """
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Figure 10: Prompt used on LLM based Machine Translation module