Title	ソーシャルニュース向けエンドツーエンド多角的事実検証システム
Author(s)	LUU, THANH SON
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	ETD
URL	https://hdl.handle.net/10119/20581
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

Doctoral Dissertation

END-TO-END MULTIMODAL FACT-CHECKING SYSTEM FOR SOCIAL NEWS

LUU Thanh Son

Supervisor NGUYEN Le Minh

Division of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

March 2026

Abstract

Fact-checking is the process by which journalists verify whether a claim is true. With the rapid spread of misinformation and the rapid growth of social media, it is crucial to automate fact-checking to help network moderators reduce its impact on social news platforms. Moreover, the output of the fact-checking task requires not only the stance for the claim but also the explanation for the stance of the claim. In addition, current trends combine data from multiple modalities, such as text and images; thus, the fact-checking system must be able to process multimodal data.

This thesis aims to construct an end-to-end Fact-checking system on a multimodal dataset, comprising three main stages: retrieving evidence, verifying the claim based on the evidence, and justifying the claim's truthfulness. There are several appropriate approaches for an automatic end-to-end Fact-checking system, such as deep learning with multimodal fusion techniques and a sequence-to-sequence model for generating results in natural language. However, the main challenge for the deep learning approach is the lack of large-scale human-annotated data and potential bias in the dataset. Large language models (LLMs) trained on vast amounts of data can exploit valuable latent patterns even with limited or no training data, enabling efficient instruction prompting for fact-checking tasks. The flexibility of LLMs can then be leveraged to construct an end-to-end Fact-checking system.

We proposed MCVE - a transformer model with fusion techniques that efficiently combine text and image feature vectors within a classification network to determine the truthfulness label of the claim. Besides, we introduce a simple idea that leverages the image caption as a feature to combine with textual evidence, forming the input sequence to a sequence-to-sequence model for generating a justification of the claim's veracity. The experimental results show the efficiency of our deep neural network-based fusion method for the Multimodal fact-checking task, as evaluated on two multimodal fact-checking datasets. Next, we propose three instruction prompting frameworks — ZeFAV, TabV4FC, and M-RAV — to leverage the robustness of LLMs for verifying claims across diverse evidence modalities: text, images, and tables. Then, we evaluate our frameworks on different fact-checking datasets with multimodal evidence, including text-only, text-image, and text-table, and the empirical results show the efficiency of our proposed frameworks in leveraging the LLMs for the Fact-checking task. Finally, we introduce FALCON - an end-to-end system that implements LLMs with a UI to demonstrate the fact-checking task in practice. The system can be hosted entirely locally with a single GPU, without any external API dependencies, demonstrating the potential for scalability and maintenance in a real system.

In the future, we would like to improve the performance of the end-to-end fact-checking system by optimizing inference runtime and integrating temporal information into LLMs to better understand time-based events in the data and verify the correctness of information, especially time-sensitive data such as financial, legal, and political.

Keywords: fact-checking, deep neural network, large language models, multimodal, end-to-end systems

Acknowledgment

I would like to express my heartfelt thanks to my supervisor, Prof. NGUYEN Minh Le, for his guidance and trust, as well as for his encouragement when I was facing difficulties in my research and life. I learned a lot from Prof. NGUYEN, Minh Le, not only about knowledge but also about skills in conducting research in academia and industry.

I would like to thank my advisors – Prof. INOUE Naoya, Prof. SHIRAI Kiyoaki, Prof. MATSUI Tomoko, Dr. RACHARAK Teeradaj, and Dr. TRAN Vu Duc– for their valuable advice during my doctoral studies.

Next, I would like to express my gratitude to my colleagues in NGUYEN's Lab for your warm support and insightful discussions during my doctoral course at the laboratory. I learn too much from you in my doctoral course. In particular, I would like to express my sincere thanks to my best colleague, Mr. Trung Thien Vo, for your kind support in my research and in daily life at JAIST.

Thank you to all JAIST Professors, lecturers, and staff for your support. Your lecture, your advice, and your kind support with administrative procedures helped me a lot during my studies at JAIST.

Thank you to all reviewers and examiners for your valuable comments to my research papers and my thesis. Your feedback helps me a lot to improve my research.

Finally, I would like to give special thanks to my father, my mother, and my younger brother for their support. Family is a great motivation for me to try my best every day.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Overview

In the digital era, an enormous amount of information is created and shared by many users on the internet. Since humans are inquisitive about what happens in their surroundings, the day-to-day facts and actions, which are often called "news", affect the daily life of humans in making decisions [1]. Since this "news" information affects the daily activities of humans, numerous attempts are made to create misleading information - a piece of information containing incorrect facts and distorted viewpoints - to cause people to believe in falsified information. The misleading information caused various bad results in actual life, especially in the COVID-19 pandemic, where the spreading of fake news about the COVID-19 situation made the panic worse [2]. Fighting with fake news and misleading information to reduce the bad impact on social communities will keep the social network clean and friendly for users to share and discuss information. Furthermore, the appearance of Language Models (LLMs) has recently led to a problem of hallucination, in which LLMs generate new content that is nonsensical or unfaithful to the provided source content [3]. Consequently, it is urgent to have a mechanism for automatically and timely verifying the information to prevent the bad effects on the communities.

Fact-checking is a task for assessing the truthfulness of a claim made in written or spoken language [4]. This task is a work in journalism, where a journalist first collects relevant evidence from potentially reliable sources, then evaluates the reliability of each source, and makes a comparison with the need-to-verify information to validate the veracity. This is a time-consuming and costly process, which is impossible to conduct when the size of the information increases manually. Hence, it is necessary to have a mechanism for automatically verifying information. Overall, a fact-checking pipeline consists of three main stages, as illustrated in Figure 1.1, which are described below.

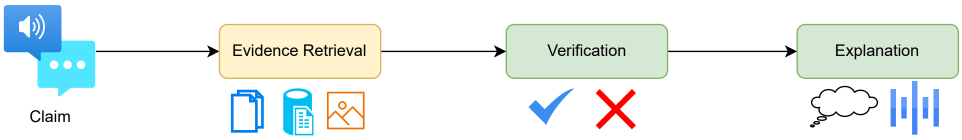


Figure 1.1: Overview of a pipeline in Fact-checking.

- **Stage 1 - Evidence Retrieval**: The purpose of evidence retrieval is to find relevant information beyond the given claim. According to [4], the veracity of the claim depends on the coherence with the evidence. Therefore, the evidence retrieval plays a vital role in providing quality information for verifying the truthfulness of the claim. In the case of a multimodal scenario, the evidence can include text, images, tables, and metadata.
- **Stage 2 - Claim Verification**: Giving a claim with corresponding pieces of evidence, the verification stage aims to assess the claim's veracity based on comparing its carried information with the facts and data from evidence. According to [4], the verification task can be classified as a classification task, which determines whether a claim is true or false, supported or refuted.
- **Stage 3 - Veracity Justification**: Since the verification stage determines the veracity of the claim, the justification step explains the decision-making process of verifying the claim to convince the audience in interpreting the reason for the veracity of the claim [4]. According to [5], the explanability in the Fact-checking task is the ability of a computer to deliver the rationales for the decision process of verifying the claim in human-readable languages. Since it is difficult to gather human explanations for Fact-checking via crowdsourcing, one potential approach is to extract the explanation from reliable Fact-checking platforms like Snopes[1] to PolitiFact[2], where expert journalists conduct the explanation and fact-checking process to make the explanation [5].

With the vast development of technologies, multimodal content is more credible compared to text-only because it carries rich visual information that can bias the readers to believe in a misleading claim represented by pictures [6, 7]. Additionally, nowadays multimodal content has become more popular on the internet, especially on social media platforms, where the posts contain text and images that spread rapidly and have high engagement by users, such as likes, reposts, and shares. Therefore, multimodal fact-checking becomes a crucial task in battling with misleading and falsified information spreading

[1] https://www.snopes.com/
[2] https://www.politifact.com/

on the internet. However, this task witnesses several challenges as described below [4, 8]:

- **Dataset issues**: Since the synthesis datasets are constructed through crowd-sourcing, the model trained on these datasets tends to rely on the dataset without learning the underlying task, causing a challenge for generalizability in the Fact-checking task. Additionally, constructing high-quality human-annotated datasets for Fact-checking is costly and time-consuming [9]. Finally, the temporal claim data that contains information over a timeline is a significant challenge for Fact-checking systems [10].
- **Multimodal Alignment**: The claim and evidence can be conveyed in various modalities such as text, image, table, and audio, indicating the challenge in efficiently aligning the information among modalities for extracting the necessary and valuable piece of information [4]. In addition, the Fact-checking model may be biased toward the specific manipulation models and struggle to generalize due to the rise of diffusion models for visual manipulation [8].
- **Evidence Sources**: The Fact-checking task relies on the truthworthiness of the evidence sources. However, not all information is equally trustworthy, and relying on a single source of evidence may create bias regardless of the intention of the claim, which may cause inaccuracy in verifying the information [4]. Moreover, the necessary information is not always available due to the incompleteness in the knowledge source and the timely updating of the evidence databases [11].
- **Explanability**: According to [5, 8], the multimodal fact-checking requires the ability to make a justification for a claim as explanability, such that the explanation must be clear, coherent, and concise in human-readable language.

The motivation of this doctoral thesis is to propose a methodology for efficiently performing fact-checking on multimodal data. The purpose is to construct an end-to-end system for verifying the information, in which retrieving relevant evidence for the claim, verifying the veracity of the claim, and showing the justification for the truthfulness of the claim in natural language. Additionally, the thesis presents a demonstration of the system, aiming to illustrate its practical potential for verifying internet information.

1.2 Related Works

Fact-checking is a task that verifies and detects the truthfulness of information. Following the term Fact-checking, there are several associated tasks, like fake news detection, disinformation, and misinformation detection. However, the scope of these tasks is different from fact-checking, as the fake news detection task focuses on assessing the aspect of information about the satire or hate speech, which is not relevant to the veracity [4, 12], and the misinformation detection focuses on detecting the aspect of information that is not accurate or complete [12]. Similarly, the disinformation detection task is designed to detect whether the information contains malicious motives to cause misleading to the audience rather than verify its veracity [12]. Therefore, the previous works in [4, 8, 12] use the term "fact-checking" or "fact-verification" to describe the process of detecting the veracity of the claim.

In this section, we conduct a literature review about current trends and challenges in the Fact-checking task. Since various works have been done on Multimodal Fact-checking, we categorize the previous works into three aspects: benchmark datasets, methods, and approaches for multimodal Fact-checking, as well as the current available Fact-checking systems.

1.2.1 Dataset for Fact-checking

The dataset plays an important role in constructing, fine-tuning, and evaluating the performance of an automatic Fact-checking system. Especially, benchmarking datasets in which the data are collected, labeled, and verified by humans is valuable because it reflects the actual features in the communities. According to [4], we divided the datasets into three dimensions corresponding to three stages of the Fact-checking pipelines, as illustrated in Figure 1.1. Table 1.1 shows the details about the current available dataset in Fact-checking.

As shown in Table 1.1, most of the datasets are constructed for the verification task. In contrast, fewer datasets serve for the justification, indicating the challenges for benchmarking in the justification task. Besides, the three labels for indicating the veracity of the claim include support (true), refute (false), and not enough information seems popular for the verification task, while some datasets like MultiFC and LIAR have fine-grained labels, which contain 9 labels for MultiFC and 6 labels for LIAR. Specifically, the Mocheg dataset is an end-to-end dataset, which contains three pipelines: retrieval, verification, and justification for benchmarking the fact-checking

Table 1.1: Available Fact-checking datasets with different modalities.

Dataset names	**Modalities**	**Source**	**Retrieval**	**Verification**	**Justification**	**# claim**
FEVER [13]	text	Wikipedia	x	x (3 labels)	-	185,445
HoVER [14]	text	Wikipedia	x	x (3 labels)	-	26,171
FEVEROUS [15]	text + table	Wikipedia	x	x (3 labels)	-	87,026
SEM-TAB-FACTS [16]	text + table	Wikipedia	-	x (3 labels)	-	180K
PubHealth [17]	text	Fact-checking sites	x	x (4 labels)	x	11,832
LIAR [18]	text	Politifact	-	x (6 labels)	x	12,836
TabFact [19]	table	Wikipedia	-	x (2 labels)	-	92,283
SCITAB [20]	table	arXiv	-	x (3 labels)	-	1,225
PubHealthTab [21]	table	Fact-checking sites		x (3 labels)	-	1,942
QuanTemp [22]	text	Fact-checking sites	x	x (3 labels)	-	15,514
MultiFC [23]	text + meta data	Fact-checking sites	x	x (9 labels)	-	36,534
FACTIFY [24]	text + image	Fact-checking sites	-	x (5 labels)	-	50,000
FACTIFY-5WQA [25]	text	Wikipedia and Fact-checking sites	-	x (3 labels)	-	391,041
Mocheg [26]	image + text	Politifact, Snopes	x	x (3 labels)	x	15,601
FINFACT [27]	image + text	Politifact, Snopes, FactCheck	-	x (3 labels)	x	3,369
FactDrill [28]	image + text	Fact-checking sites	x		-	22,435
MunMin [29]	image + text + meta data	Twitter		x (3 labels)		12,914
CLAIMDECOMP [30]	text	PolitiFact	x	-	x	1,200
Half-Truth [31]	audio	Real speakers	-	x (3 labels)	-	88,035
MAD [32]	audio	PolitiFact	-	x (2 labels)	-	4,915
SciVER [33]	text + image (charts, tables)	arXiv	x	x (2 labels)		3,000
Official-NV [34]	video	Xinhua	-	x (2 labels)	-	10,000
COVID-VTS [35]	video	Twitter	-	x (2 labels)	-	10,000

systems in multimodal settings. In general, the modalities among fact-checking datasets are often text, image, and table, while there are fewer datasets in the audio and video modalities. Instead, visual content like images and videos is popular on social media [8, 36] since most of the Fact-checking dataset focuses on text and image content.

1.2.2 Fact-checking Approaches

As discussed in [8], evidence retrieval, verification, and justification generation in the multimodal fact-checking task require the computer's ability to understand and process data in various modalities, i.e., text, image, audio, and video, to extract and align information across The motivation for approaches in multimodal is to find an efficient method for better aligning and integrating information from different modalities for each stage in the fact-checking pipeline. The details about current approaches are described below.

Evidence Retrieval: The retrieval task aims to retrieve relevant pieces of evidence for an input claim as a query, in which the evidence can be text, image, audio, or video, and so the claim is. To retrieve the relevant evidence, popular approaches include embedding-based search, in which the claim and evidence from multiple modalities are embedded into semantic vectors and used for ranking or entailment inference. Various robust approaches, such as

CLIP [37], BGE-Visualized [38], Jina Embedding [39], and C-RADIO [40], are introduced for encoding data from textual and visual content into semantic vectors. Also, large language models (LLMs) that support multimodal, like Gemma [41], LLaVa [42], LLama3.2 Vision [43], and Qwen2.5 Vision [44], are used to leverage the ability of language understanding and reasoning of LLMs for extracting the most relevant piece of evidence. Recent works like RAV [45] and FACTOR [46] are based on the transformer architecture, while HeterFC [47] and EGMMG [48] employ graph methodology for retrieving relevant evidence.

Claim Verification: This task is categorized as a classification task [4, 8]. Based on the number of classes used to determine veracity, the task can be a binary or multi-class classification. In multimodal classification, data fusion plays a vital role, as it efficiently combines and integrates key features from various modalities before fitting them to the primitive classifiers [49]. Previous works such as ECENet [50] and MAGE [51] leverage fusion to combine information from textual and visual modalities to support multimodal classification in fact-checking. Since fusion techniques have shown promising results, several challenges remain in accurately aligning information across modalities, and the evidence base for fact-checking is insufficient [11]. To enhance the ability to reason across modalities, recent approaches leverage LLMs for verifying the truthfulness of the claim, such as LVLM4FV [52], FinLLaMa [53], and CLIP + GPT-4o Guided [54]. Although achieving optimistic results, these works depend solely on golden evidence, which limits their performance in practice, where evidence is automatically retrieved without human supervision. Additionally, cross-modality alignment, reasoning, and the detection of manipulated visual content generated by the diffusion model remain challenging for LLMs in the verification task.

Justification: The task is categorized as a text generation task, whose purpose is to deliver the rationale of the decision on a verified claim [5]. Providing the explanation as a justification for a claim makes the fact-checking system more reliable and friendly to users, with the purpose of helping to interpret the process of verifying. To generate the explanation, there are various more for text-generation to consult this task, like encoder-decoder models such as T5 [55] and BART [56], or autoregressive LLMs via prompting. To generate a correct and coherent explanation that is interpretable and comprehensive, the model needs to understand the claim and its relation to evidence and veracity labels. Especially in multimodal settings, the models are required to perform cross-modality reasoning and understanding to generate a suitable explanation as justification for the claim. Since this task requires the generated explanation to be correct and comprehensive, the fine-tuning step is necessary to adjust the model according to human-labeled data.

Additionally, since human-annotated data are costly and time-consuming, it is necessary to build a fact-checking system that can work in the case of minimal or even no labeled data for training [9]. There are several proposed methods based on LLM-as-a-judge [57] that perform fact-checking tasks by leveraging the enormous knowledge within the LLMs and their reasoning ability through an efficient instruction pipeline or paradigm prompt templates. For example, the ProgramFC [9] decomposes the complex claim into "program"– consists of sequentially ordered reasoning steps, and verifies the claim based on the execution of those "programs". Similarly, FactScore [58] and FaithScore [59] decompose the complex claim into "atomic fact"– a short sentence that contains only single information [58], then verifying those atomic facts to exploit the truthfulness of the claim. The idea of decomposing complex claims into smaller pieces of information helps LLMs be more efficient at processing long sequences, as proposed by the VeriFactScore [60] and SAFE [61] pipelines. Additionally, InfoRE [62] proposes a simple idea: reorganizing the complex textual structure of evidence into a more compact and comprehensive structure, similar to a MindMap, to enhance LLMs' understanding and reasoning abilities on downstream tasks. In general, these methods rely on zero-shot reasoning in LLMs without human-annotated data.

1.2.3 Fact-checking Systems

With the rapid growth of social networks, it is necessary to deploy an automated system to verify and reduce the spread of fake news and misleading information. Many attempts have been made in constructing an automatic system for Fact-checking. To provide an overview of available Fact-checking systems, Table 1.2 describes the details of current end-to-end systems, including the retrieval and verification stages with UI support.

Table 1.2: Current available Fact-checking system

System name	**Modality**	**Retrieval**	**Verification**
LOKI [63]	Text	Web search APIs	GPT-4o
OpenFactCheck [64]	Text	Serper and SerpAPI	GPT-4
Holmes [65]	Text+Image	Web search APIs	GPT-4o
LiveFC [66]	Audio Stream	Web search APIs	XLM-Roberta-Large
Perplexity ai [3]	Text	Web search APIs	Sonar-online
QAChecks [9]	Text	Wikipedia Retrival	InstructGPT

According to Table 1.2, it can be seen that most of the automatic Fact-checking systems are employed on textual, i.e., LOKI [63], OpenFactCheck

[64], QAChecks [9], and Perplexity AI - a web platform. In addition, Holmes is employed on multimodal data, including image and text, while LiveFC is constructed to process audio streaming data. Overall, most systems rely on Web search APIs for evidence retrieval and commercial LLMs like GPT-4o for verification and decomposition. However, the trustworthiness of the evidence source can affect the performance of the fact-checking system [4], especially in multimodal scenarios where the visual data, such as images or videos, are at risk of being manipulated [8]. To summarize, there are two main concerns in constructing a fact-checking system in practice: the reliability of evidence sources (articles, images, text, etc.) and scalability in practical deployment, where reliance on API calls can lengthen response time [9]. To build an end-to-end Fact-checking system, it is better to use a locally hosted model rather than rely on API calls.

1.3 Contribution

The motivation of the dissertation is to propose a methodology for constructing a multimodal fact-checking system for practical use. To sum up, the dissertation has five main contributions:

Contribution 1: Proposed MCVE - a transformer-based fusion deep neural network for Fact-checking. The main idea is to efficiently combine the textual and visual information to extract valuable information for the verification task. Additionally, MCVE employed an image-to-text module that converts images into text to fit the encoder-decoder model for generating the explanation.

Contribution 2: Proposed ZeFAV - a method that leverages the ideas from relation extraction for filtering the most relevant pieces of evidence to the claim, and InfoRE for re-organizing the evidence into a compact and comprehensive form to instruct LLMs for verification.

Contribution 3: Introduce the TabV4FC for tabular Fact-checking by verbalizing the table into textual data and combining it with the table represented in the specification template to guide the LLMs for determining the veracity of the claim.

Contribution 4: Proposed the M-RAV - a multimodal Fact-checking framework that employs LLMs to generate alignment and consistency between text and image pairs in the evidence, thereby instructing the LLMs to verify the claim. M-RAV also shows potential for constructing an end-to-end system.

Contribution 5: Constructing the FALCON - an open-source Fact-checking system which employs small LLMs to perform end-to-end multimodal Fact-checking, including four stages: retrieving the relevant evidence for the input claim, making alignment between textual and visual pieces of evidence, giving the verification and corresponding justification to the veracity.

1.4 Dissertation Outline

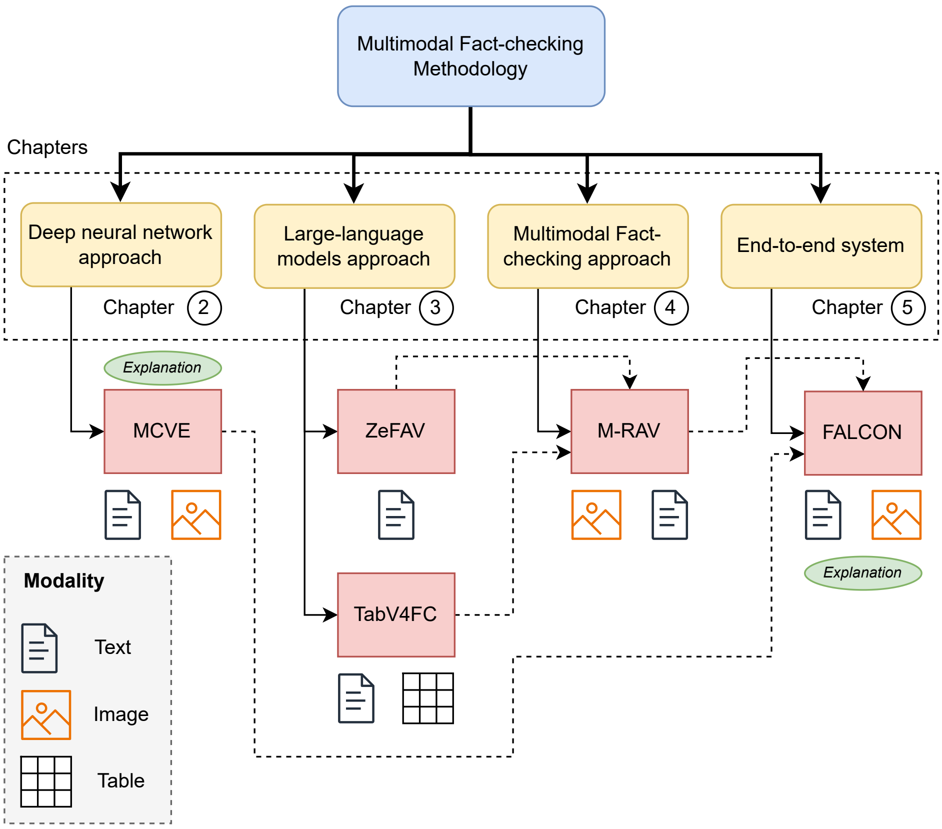


Figure 1.2: Outline of the dissertation.

Figure 1.2 shows the overview of the outline of this thesis. Chapter 1 introduces the end-to-end multimodal Fact-checking project and previous works. Chapter 2 describes the transformer-based fusion attention network for multimodal fact verification and explanation. Chapter 3 illustrates the usage of LLMs in fact-checking with instruction and efficient prompting techniques to enhance the ability of LLMs for verifying the claim. Next, chapter 4 introduces an approach for a multimodal fact-checking system that leverages the reasoning ability of LLMs to consult with system-retrieved evidence. Then, Chapter 5 shows the process and relevant methods for implementing an end-to-end multimodal fact-checking system for practical uses according to the methodology introduced in previous chapters. Finally, Chapter 6 concludes the contribution of the thesis and proposes future work.

Chapter 2

Deep Neural Network for Multimodal Fact-checking

2.1 Introduction

The rapid development of social media provides people with an open space for discussion and information sharing. However, this also creates a significant challenge: misleading information, in which people can create fake information and let it spread rapidly without authorization. Specifically, the appearance of large language models leads to hallucinated information, which is also a critical problem for truthfulness verification. Consequently, the fact-checking task is crucial to keep reliable information and stem the tide of misinformation.

In the work by [49], the authors introduced various multimodal classification taxonomies that efficiently concatenate information from different modalities information such as image and text, which is called data fusion techniques. Many works use the data fusion approach for the multimodal Fact-checking and fake news detection tasks [26, 67–70]. However, these works do not treat the claim that has multiple pieces of evidence and has no evidence. Besides, the length of the text evidence sequence is often much longer than 512 tokens, which exceeds the encoding ability of BERTology models. In addition, in the work by [26], the claim explanation task lacks visual information from image evidence. According to [71], multimodal information is more credible and has higher engagement than text only, which can provide valuable information that supports the truthfulness of the claim.

To solve the problem of fact-checking with multiple types of data such as textual, visual, and hearable, the learning model must have the ability to incorporate the data from various sources. The authors in [72] address the idea that multimodal data might improve the performance of the Fact-checking system because it can exploit the semantic relation between textual and visual information. The term multimodal is in terms of cases where the claim and/or supporting evidence are conveyed through multiple modalities

[49]. For building multimodal classification models, the fusion stage is utilized to concatenate and merge different features from multiple data sources to build the model. [49] proposed three data fusion strategies in multimodal including early fusion, late fusion, and cross-modality fusion. With the power of state-of-the-art transformer models for computer vision and natural language processing such as BERT [73], BERTology [74], Vision Transformer [75], and the early fusion are a potential approach because these models have a strong associate with the data source [49].

The automated fact-checking task has several research challenges. Nonetheless, according to [5, 71, 76], there exist two main challenges in constructing the model for the task. First, the multimodality method has made it difficult to combine the data from multiple forms such as text, image, and tables efficiently and flexibly. Second, there is insufficient evidence to verify the claims in the automated fact-checking models, especially multimodality data. With the availability of foundational and fine-tuned LLMs, fact-checking databases, and prior work on multimodal classification tasks, we propose a new methodology to solve the claim verification and claim justification (explanation) task. This chapter introduces **MCVE** - a new multimodal-fusion framework that can efficiently verify the claim that has multiple image and text evidence and has no evidence at all. First, we propose a new framework that leverages the transformer models for text like BERT [73] and RoBERTa [77], and images such as ViT [75], BEiT [78], and DeiT [79] for the claim verification task. The proposed method considers the evidence of each claim, whether it has multiple or no evidence. Then we use data fusion techniques to concatenate the features from both text and image before fitting them to the classification layer to determine the truthfulness label. Second, we propose a simple method that integrates the textual and visual information from the evidence of the claim to the encoder-decoder generation model like T5 [55] and BART [80] to generate the explanation of the claim verdict as the ruling sentence. MCVE is then evaluated on the two multimodal Fact-checking datasets, including the Mocheg [26] and FACTIFY [24], to explore the efficiency of proposed approaches for the multimodal fact-checking task as well as the current challenges. The experimental results show that the performance of the proposed method increases in comparison with the baseline from previous works, indicating MCVE's efficiency.

2.2 Background

Multimodal classification is the task of classifying data across multiple modalities, such as text, images, and speech, into a set of specific label classes. As defined by authors in [49], there are five stages for a multimodal classification pipeline: pre-processing, feature extraction, data fusion, primary learner, and final classification. The roles for each stage are described as follows:

- **Pre-process**: Preparing the clean data for the task, including removing noise, augmentation, or transforming the data into an appropriate form.
- **Feature Extraction**: Extracting valuable (or higher) features from the raw data in a suitable representation for machine learning algorithms or deep neural networks. For example, this stage employs the BERT encoder to encode text and contextualize embeddings, the Vision Transformer (ViT) to encode images into embeddings, TAPEX to encode tabular data, and HuBERT to encode speech. These embedding vectors are then fed into a deep neural network or machine learning model to estimate the label distribution for training or prediction.
- **Data fusion**: This stage concatenates the extracted features from multiple modalities in an efficient way to construct a single data representation that is suitable for mapping the valuable information from cross-modality information.
- **Primary Learner**: The main component or layer for extracting knowledge from the data represented as vectors. Primary Learner can be a set of CNN modules, LSTM layers, or transformers. Additionally, in multimodal classification, this stage can be leveraged to share information across modalities.
- **Final classification**: The final stage for class prediction, which mainly produces class likelihood scores. The final class is determined by the class distribution with the highest probability.

In multimodal classification, data fusion is a vital and unique stage that aims to extract relevant information across modalities [49]. There are three strategies for performing data fusion: early fusion, late fusion, and cross-modality fusion, and two techniques for combining the vectors from different modalities: concatenation and merging. According to [49], the late fusion techniques that extract features independently from each modality before combining in the fusion step to form one single vector benefit the final classifier than the two remaining. Also, concatenation techniques provide a simple yet efficient method for combining multimodal features into a single vector [49]. Therefore, the MCVE framework in this Chapter follows a late-fusion strategy with concatenation techniques to efficiently integrate text and image features for classification and sequence-to-sequence tasks in multimodal Fact-checking. The details of MCVE are described in the following section.

2.3 The MCVE framework

The MCVE framework (**M**ultimodal **C**laim **V**erification and **E**xplanation) involves two main tasks: claim verification and claim explanation. These tasks rely on multimodal evidence, including both textual and visual data. The claim verification task takes as input a textual claim $\mathbf{q}$, a set of textual evidence $\mathbf{T} = \{t_1, t_2, ..., t_n\}$, and image set of evidence $\mathbf{M} = \{m_1, m_2, ..., m_k\}$, where n and k represent the number of text and image evidence items, respectively. The output of this task is a classification label $\mathbf{c} \in$ C, where $C = \{\textit{supported}, nei, refuted\}$. For the claim explanation task, the system receives as input the same claim $\mathbf{q}$ and evidence $\mathbf{T}$ and $\mathbf{M}$, along with the label c from the verification step. The output is an explanatory sentence $\mathbf{e}$ that justifies the verification result, providing a reasoning for why the claim is supported, refuted, or undetermined based on the available evidence.

2.3.1 The claim verification task

In the Mocheg dataset [26], researchers proposed a framework that concatenates the embedding vector of the claim with the text and image evidence embedding by the attention mechanism [26]. Also, the authors have attempted to leverage CLIP [37] as a pre-trained model to extract the embedding vector from text and image. However, the lack of evidence relevant to the claim impacts the performance of CLIP because CLIP needs sufficient texts and images together to compute the similarity for exploiting valuable features. Therefore, in this section, we propose an architecture that uses the data fusion approach [49], which concatenates individual features from unimodal first and feeds the fusion features to the classifier. We have three modalities for extracting the claim, text evidence, and image evidence. On the other hand, according to [26], the claim itself has a significant performance effect since it achieved nearly 40% by F1-score. To leverage the information from the claim itself to determine the truthfulness, the VClaim checking system [81] proposed a simple but efficient approach by finding similar claims that have been verified and comparing them with the current claim to determine whether the claim is true or not. However, as mentioned by the authors in [81], the challenge for the VClaim checking system is that it has many false negative cases, which means that the current claim that needs to be verified is not relevant to those retrieved by the system. Referenced the idea from the VClaim system [81], we construct a pre-trained classifier based on BERTology for the claim only and integrate it with the fusion model to enhance the performance of the verification model. By leveraging the power of the BERT model in dynamic contextual learning from text, we can exploit the valuable semantic information from the claim and fine-tune the model for the claim verification task to enhance performance.

To the claim evidence, in [82–85], the author used the BERT [73] and RoBERTa [77] for extracting features from claim, and their performance was significantly better from the FACTIFY challenges. Therefore, in this work, we apply these two models (BERT [73] and RoBERTa [77]) for claim feature extraction. For the text evidence, we choose the Longformer [86] and BigBird [87] because they can process with long sequences efficiently [88]. Finally, for the image evidence, we exploit the ViT [75], BEiT [78], and DeiT [79] for image representation, according to [89].

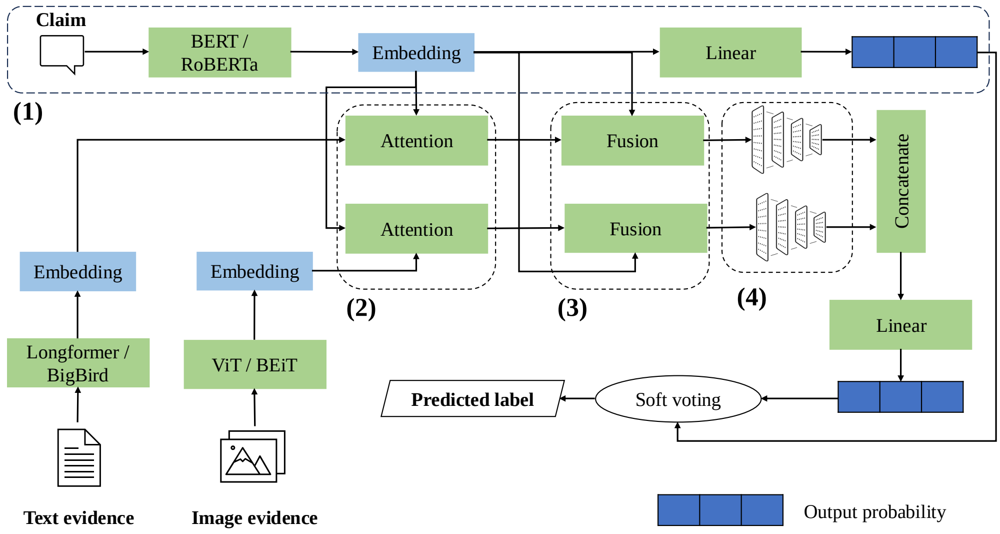


Figure 2.1: Proposed architecture of claim verification task

Figure 2.1 illustrates our proposed architecture for the claim verification task according to the task definition introduced in Section 1. There are four main components: (1) the claim features, (2) the attention modules, (3) the fusion modules, and (4) the convolutional modules. Let Hq, Ht, Hm be embedding vectors for the claim, text evidence, and image evidence, respectively.

The claim q is forwarded to the fully connected layer, and the softmax function is used to compute the distribution probability for each class. $\hat{y}_1$ is the vector that presents the probability for each class of claim, and the backward process updates the weight of the claim embedding vector according to the truthfulness label (Module (1) in Figure 2.1).

$$Hq = BERTology(q), \tag{2.1}$$

$$fc = fully\_connected(Hq), \tag{2.2}$$

$$\hat{y_1} = Softmax(fc) \tag{2.3}$$

Each claim has multiple text evidence $T = \{t_1, t_2, ..., t_n\}$ and image evidence $M = \{m_1, m_2, ..., m_k\}$ (n and k are maximum of text and image evidence, respectively), for that, we obtain all of the evidence of image and text to a single vector for computation. Following SentenceBERT [90], a mean operation is added to construct a fixed embedding vector for T and M.

Since we have the represented vector embedding for claims, text evidence, and image evidence as Hc, Ht, and Hm, in the next step, the attention is computed between the claim with text evidence and image evidence, according to [26]. In case the claim has missing text or image evidence, we arrange an empty string and for text evidence an empty image with a size of 50x50 for image evidence to generate the embedding Ht and Hm. Afterward, we obtain all the similarity information between the claim with text image and image evidence, respectively via the attention mechanism (Module (2) in Figure 2.1).

$$Att_{text} = Attention(Hq, Ht, Ht), \tag{2.4}$$

$$Att_{image} = Attention(Hq, Hm, Hm) \tag{2.5}$$

Next, we compute the fusion between the claim, the attention vectors Att_{text}, and Att_{image} to receive the stance representation for claim and evidence, according to [26]. After this step, we get the stance representation Gt and Gm (Module (3) in Figure 2.1).

$$Gt = [Att_{text}Hq : Att_{text} - Hq], \tag{2.6}$$

$$Gm = [Att_{image}Hq : Att_{image} - Hq] \tag{2.7}$$

To attain the important feature for the classification task, we apply the convolutional layer for Gt and Gm according to the TextCNN architecture [91]. We employ three convolutional layers with kernel sizes 3,5,7 respectively, and an output channel equal to 100 (Module (4) in Figure 2.1). For computing the $\hat{y_2}$ as the vector that presents the probability for each class, we concat two convolutional vectors $conv_{text}$ and $conv_{image}$ together and forward them to the fully connected layer. In the next step, the softmax function is employed to calculate the distribution probabilities for each class.

$$conv = Concat(conv_{text}, conv_{image}), \tag{2.8}$$

$$fc = fully\_connected(conv), \tag{2.9}$$

$$\hat{y_2} = Softmax(fc) \tag{2.10}$$

Finally, we use the soft voting in ensemble learning to build the final prediction vector $\hat{y}$. The predicted label is the class that has maximum probability. By using the ensemble model, we can leverage the effect of the claim in discriminating between classes while keeping the role of evidence.

$$\hat{y} = Soft\_voting(\hat{y_1}, \hat{y_2}), \tag{2.11}$$

$$output\_label = argmax_c(\hat{y}) \tag{2.12}$$

2.3.2 The explanation task

According to [5], the explanation of the veracity of the claim plays a vital role in the fact-checking process. In an automatic fact-checking system, the model must have the ability to explain how the model delivers its rationale for the decision on the truthfulness of a claim based on the given evidence. For example, in Figure 2.2, the fact-checking model verifies the claim as refuted through the learning of the embedding vector represented for the claim and the evidence in textual and visual form. Moreover, the model should generate an explanation that justifies the truthfulness of the claim in human-readable text. With explanation as justification for the truthfulness of the claim, the fact-checking model helps humans understand the process of giving results, making the model more reliable and friendly to users in practical use.

To satisfy the explainability, the model must have the ability to understand the full context from the facts, including the textual and visual context, which comes from the claim and its evidence. The fact-checking model has to give the ruling as the explanation in human-readable texts. In this article, we focus on the encoder-decoder text generation models such as T5 [55] and BART [80] for this task because of their robust ability in deep contextual understanding and flexible generation output from the input sequence. Additionally, we utilize LED - a text encoder-decoder version of

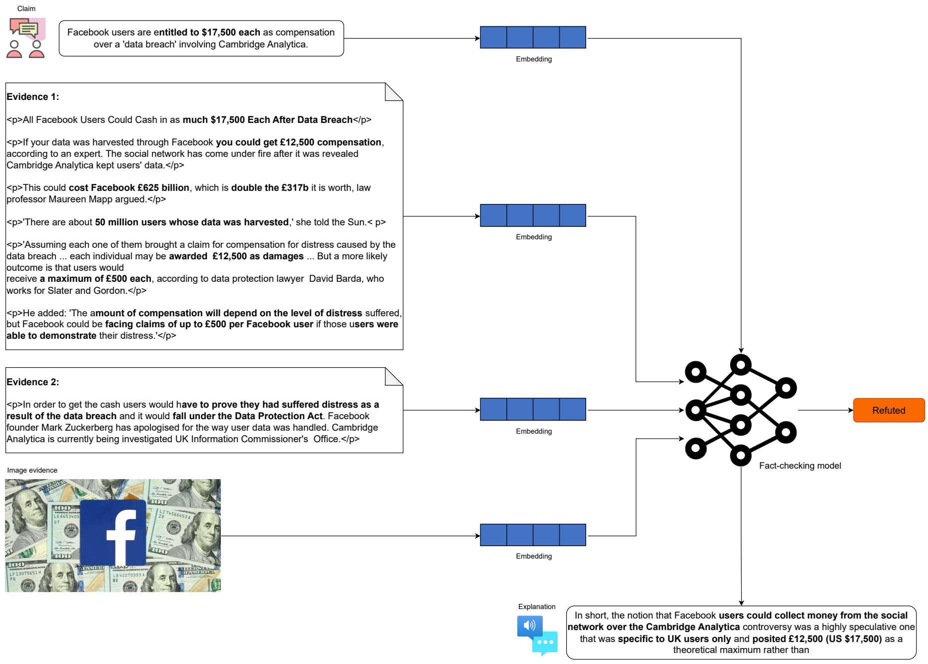


Figure 2.2: A sample of explanation about the truthfulness of a claim with given evidence

Longformer [86] to encode the input sequence because of its robustness in encoding long sequences.

For achieving better content interpretation of the claim, the authors [26] proposed a mechanism that concat the claim, its label, and its evidence to one sequence and fits this sequence to the BART model. Also, they exploit the truthfulness label as a reward function to control the generation model. Nevertheless, the authors have not integrated the visual information from image evidence, which is indicated as an essential feature for the task [24, 71]. To overcome this problem, we adopt a pre-trained image-to-text model to generate the text from image evidence. The image-generated text is concatenated to the input sequence, and then they are an input to the generation model. Finally, we fine-tune the generation model to fulfill the ruling statement as an explanation.

Figure 2.3 presents our proposed method for the explanation task. We denote the input claim q as text, a list of relevant text evidence $Ht = \{t_1, ..., t_n\}$, the predicted truthfulness label for claim $\hat{y}$, and the list of relevant image evidence $Hc = \{img_1, ..., img_k\}$, n and k are number of text and image in the evidence list, respectively. For each image evidence in Hc, we apply the *vit-*

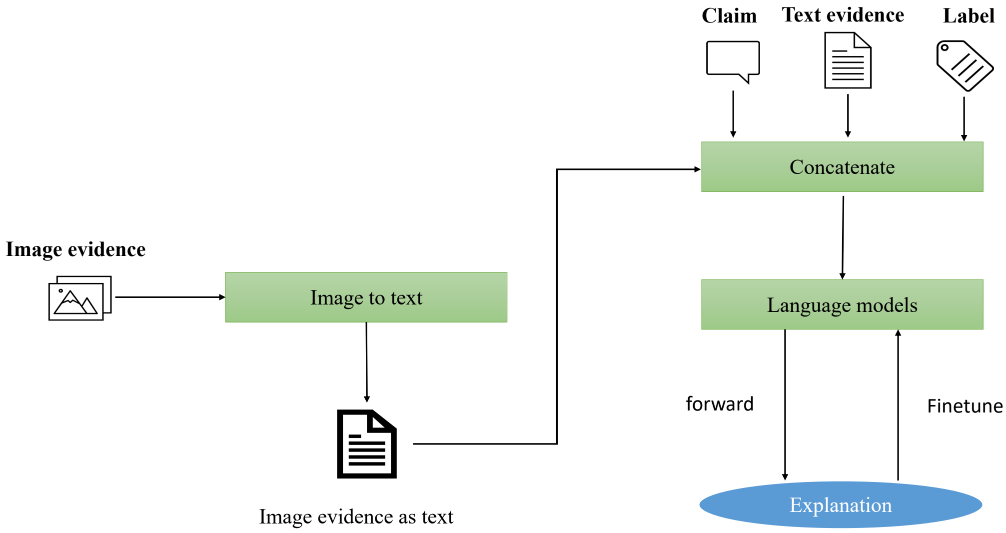


Figure 2.3: Proposed architecture of claim truthfulness explanation task

gpt2-image-captioning[1] as a pre-trained model to generate text from image. Besides, we also apply the InstructBLIP [92] to generate the description for the image as textual information by the prompting method. InstructBLIP takes an input comprising the image and the prompting with the following template: *"Please describe the detail facts, actions and people of this image according to the given claim: $< claimtext >$"*. Then it will generate the caption as the output. Table 2.1 illustrates several examples generated by both vit-gpt2-image-captioning and InstructBLIP.

$$Hc = [h_1, ..., h_k | h_i = image\_to\_text(image_i)] \tag{2.13}$$

Finally, the whole input is integrated with a single text sequence T, in which each sequence component is separated with the SEP token $< /s >$, and this sequence is taken into the generation models to generate the output explanation sequence $\hat{T}$.

$$T = q + < /s > + \sum_{i=1}^{n} t_i + < /s > + \hat{y} + < /s > + \sum_{j=1}^{k} h_j + < /s >, \tag{2.14}$$

$$\hat{T} = Generation\_model(T) \tag{2.15}$$

[1] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

To fine-tune the explanation model, let V be the set of vocabulary. We first denote q as the input query such that $q \in V^k$, in which k is the length of the input sequence. The gold label is a sequence $y \in V^z$, in which z is the length of the output sequence. First, the model encodes the input query q to construct an output vector h_E such that $h_E \in R^d$. Also, we transform the ground-truth label y into an embedding space–h_o ($h_o \in R^d$) by the encoder-decoder backbones, denoted as $Model_{encoder-decoder}$.

$$h_E = Encoder(q) \tag{2.16}$$

$$h_o = Model_{encoder-decoder}(o) \sim y \tag{2.17}$$

The encoded feature vector h_E is then fit into an autoregressive decoder model to make an output vector $h_D \in R^d$, which is the prediction sequence. The loss L is computed as the difference between the decoder output and the target text by Cross-entropy loss to measure the reconstruction of the output sequence.

$$h_D = Decoder(h_E) \tag{2.18}$$

$$L = Cross - entropy(Decoder(y, h_E), y) \tag{2.19}$$

As described in Table 2.1, it can be seen that the image description generated by InstructBLIP is more detailed than the vit-gpt2-image-captioning. For example, in sample No #1, the vit-gpt2-image-captioning only captures “a man in suit an tie”, “a microphone”, and “a man in suit an tie is holding a microphone, while the InstructBLIP can capture “a crowd of people” and “a robot successfully hit a hole-in-on”, which is relevant to the claim content. Moreover, InstructBLIP can capture the emotion of the crowd in the picture, which is “excited and cheering” and the perspective of the picture is in “a golf course”. However, in Example No #3 in Table 2.1, the InstructBLIP seems to generate incorrect information from the picture as “killing 211 people and injuring over 100 others.”. In the picture, the title shown that “Huge Molasses tank explodes in North End, 11 dead and 50 hurts”, which is unmatched with the generated description by InstructBLIP. For vit-gpt2-image-captioning, it can capture the information about “newspaper” and “local news-paper”, which is lack of in-depth information in comparison with InstructBLIP. In addition, as shown in Example No #2, both vit-gpt2-image-captioning and InstructBLIP can capture information about the “woman wearing black jacket and holding a microphone”. Nonetheless, the next information generated by InstructBLIP does not match either the claim or the image as it says that “she has been romantically linked to A$AP Rocky”. Overall, the descriptive information generated by image-to-text tools including vit-gpt2-image-captioning and InstructBLIP is relative. We leverage this information to enrich the evidence of claims to improve the generative model for the explanation task. Table 2.2 provides statistical information about the length of the generated description of both tools. The length of the description is calculated by token level. According to Figure 2.2, although the length of the generated caption of InstructBLIP is longer than vit-gpt2-image-captioning on average, InstructBLIP cannot generate a caption for most images (as illustrated in Example #5 and #6 in Table 2.1), which cause the median length of InstructBLIP equal to 0. Figure 2.4 illustrates the length distribution of the generated description for images by both vit-gpt2-image-captioning and InstructBLIP.

No.	Caption	*Claim*	*vit-gpt2-image-captioning*	*InstructBLIP*
1		“A video shows a golfing robot hitting a hole-in-one.”	“a man in a suit and tie is holding a microphone”	“The image depicts a crowd of people gathered around a golf course, watching a robot successfully hit a hole-in-one. The robot is equipped with a camera on its head, which captures the moment from a bird’s-eye view. The crowd is excited and cheering as they witness the impressive feat accomplished by the robot.”
2		“In May 2021, A$AP Rocky confirmed he and Rihanna were a couple.”	“a woman in a black jacket is talking on a cell phone”	“The image depicts a woman wearing a leather jacket and holding a microphone while singing into it. She is believed to be Rihanna, as she has been romantically linked to A$AP Rocky. The couple’s relationship was rumored to have started in late 2019, but they officially confirmed their relationship in May 2021.”
3		“A fatal wave of molasses swept through Boston in January 1919.”	“a newspaper advertisement for a local newspaper”	“The disaster occurred when a large tank of molasses exploded in the North End section of the city, killing 21 people and injuring over 100 others. The Boston Post, a local newspaper, covered the event with a front-page headline that reads ’Huge Molasses Explosion’ in North End; 21 Killed, 100 Injured.”
4		“Police shootings kill more white Americans than Black Americans.”	“a table with a pair of scissors and a pen”	“According to the claim, police shootings kill more white Americans than Black Americans in the United States. In the image, we can see a gun and bullet casings on a white surface. The presence of the gun and bullet casings implies that there has been a shooting or an event involving firearms. However, it is important to note that this image alone does not provide enough information to determine the ethnicity of the individuals involved in the shooting.”
5		“A newly discovered species of amphibian was named ’Dermophis donaldtrumpi’ as an unflattering reference to the U.S. President.”	“a close up of a close up of a man’s face”	
6		“Donald Trump has paid up to eight sexual partners to obtain abortions and sign nondisclosure agreements.”	“woman in a pink dress with a pink tie”	

Table 2.1: Image description generation by vit-gpt2-image-captioning and InstructBLIP

	vit-gpt2-image-captioning	InstructBLIP
Max length	19	187
Min length	5	0
Mean length	9.95	26.09
Median length	10	0

Table 2.2: Statistical length distribution of generated description

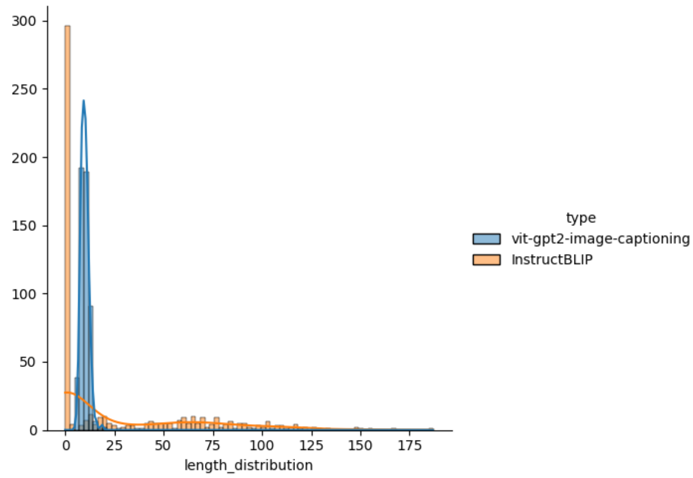


Figure 2.4: Distribution length of image description text

2.4 Dataset

In this section, we conduct an analysis of the two datasets used for our experiment in training and evaluating our proposed methodology: the Mocheg [26] and the FACTIFY [24] datasets. Both datasets are divided into three sets: the training, development, and test with proportion 75:10:15 for the Mocheg [26] and 70:15:15 for the FACTIFY [24]. Each claim in the dataset has relevant text and image evidence, the label indicates the truthfulness of the claim, and the ruling sentence is the explanation for the truthfulness of the claim. However, the evidence and the ruling sentence of the claim are not always sufficient. The evidence shown along with the claim is the gold evidence, which is constructed manually by annotators on both Mocheg and FACTIFY. To have a detailed look at the Mocheg dataset, Table 2.3 provides an overview of statistical information inside the dataset. For the FACTIFY dataset, the format of the dataset is similar to the Mocheg for the verification task. Nevertheless, the FACTIFY dataset is not employed for the claim explanation (ruling) task thus it has no explanation sentence. The overview information about the FACTIFY dataset is illustrated in Table 2.4.

Figure 2.5: The distribution of labels in the Mocheg dataset

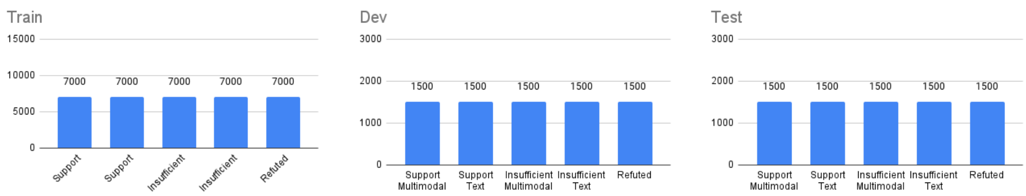


Figure 2.6: The distribution of labels in the FACTIFY dataset

From Table 2.3, the number of tokens in claims and evidence is significantly different. The maximum number of tokens in evidence is 888 and 600 for the explanation sequence (on the training set), which is a challenge for the BERT encoding model since its maximum token of the sequence is 512 [73]. In the FACTIFY dataset, since the maximum token in text evidence is 6,047, which is also more than 512 - the maximum length of the BERT model, we first present the first challenge that encodes the long sequence of text evidence.

The second challenge is insufficient evidence. According to Table 2.3, in the training set, the claim that has no text evidence accounts for nearly 64.83%. Same as text evidence, the claim that has no image evidence also takes a significant proportion, which is more than 70 percent. Furthermore, although suffering from insufficient evidence in the training data, the evidence in the development and test sets is sufficient. This leads to the fact that the Mocheg is really a challenging dataset to construct and evaluate the automated fact-checking model since it is not only evidence insufficient but also different between training and evaluating data.

In the FACTIFY dataset, we have modified a way that takes the input data to adapt to the fact-verification task. For the claim images, we append the input data with the image evidence. As shown in Table 2.4, FACTIFY also deals with the insufficient evidence problem for image evidence. Nonetheless, this proportion is dramatically lower than in the

	Train	Dev	Test
# Claim	18,583	600	2,001
# Text evidence	28,274	1,562	5,228
# Image evidence	13,200	519	1,649
Avg. token in claim [26]	20	20	21
Max token in claim [26]	81	77	89
Avg. token in explanation [26]	132	90	105
Max token in explanation [26]	600	521	600
Max token in evidence	888	611	664
# Claim has no text evidence	7,566	0	0
# Claim has no image evidence	8,419	184	624
# Claim has no evidence	4,727	0	0
# Claim has no ruling	8,775	400	312
Max text evidence per claim	38	16	30
Max image evidence per claim	17	6	8

Table 2.3: Overview statistic of the Mocheg

Mocheg dataset, which is only approximately 2%. There is no claim without evidence in the FACTIFY dataset, and the maximum number of evidence in FACTIFY is not as much as in Mocheg.

	Train	Dev	Test
# Claim [24]	35,000	7,500	7,500
# Text evidence	35,000	7,500	7,500
# Image Evidence	68,250	14,613	14,539
Max token in claim	154	145	141
Avg. token in claim	25	25	24
Max token in text evidence	6,047	6,047	6,047
Max text evidence per claim	1	1	1
Max image evidence per claim	2	2	2
# Claim has no text evidence	0	0	0
# Claim has no image evidence	618	133	138
# Claim has no evidence	0	0	0

Table 2.4: Overview statistic of the FACTIFY

Besides, it can be seen from Figure 2.5 that the distribution of labels on the Mocheg is not similar between the training, development, and test sets. On the training set, the number of refuted labels is more than supported and nei, while the NEI is the least. However, in the development and test set, the number of three labels is balanced. This indicates the challenge for the training model on the Mocheg dataset because the distribution of the training in comparison development and testing set is not the same, and the model should avoid the overfit when being trained on the training set. In contrast with the Mocheg model, the FACTIFY dataset has an ideal label distribution when it is balanced between five labels on three sets, according to Figure 2.6.

From the analysis of the two datasets, we indicate two main challenges for constructing the automated fact verification and explanation model. The first challenge is the long sequence encoding for the text evidence which exceeds the capacity of BERT models. The second challenge is the lack of evidence information relevant to the claim including lack of text evidence, image evidence, and both types. In the following section, we introduce our proposed methods aim to solve these problems to enhance the performance of models.

2.5 Experimental Results

2.5.1 Experiment Settings

For the verification task, we run the model with 30 epochs, batch_size equals 512, and the learning rate is 10^{-3}. To treat the imbalance labels in the training dataset, we use the Focal loss [93], which is an efficient method to solve the imbalance problem by enhancing the effect of the cross-entropy loss value on minority classes. According to the original work of [93], we choose the *gamma* values equal to 2. Finally, we employ the AdamW [94], which is an extended version of the Adam optimizer [95] by implementing the weight decay mechanism to avoid overfitting.

For the explanation task, we run the model with 10 epochs and the batch_size equals 16. We set the beam-search-based generate method with *num_beams* equals 3 and *max_length* is 600. Besides, to generate the description text for the image, we run the InstructBLIP with *num_beams* equals 5, *max_length* is 256, *temperature* is 1.0, *top_p* equals 0.9, and *do_sample* is False since we do not acquire the sampling strategy.

2.5.2 Empirical Results

Our experiments can be evaluated on two datasets: Mocheg and FACTIFY. As illustrated in Table 2.5, the combination of the RoBERTa, ViT, and Longformer achieves the highest score on the Mocheg dataset, which is 52.32% by micro F1, and 50.94% by macro F1-score on the test set. The performance of BERT is slightly worse than the RoBERTa's performance, which is 49.92% by macro F1-score on the test set. Observe that ELECTRA and AlBERT, the performance is significantly lower than RoBERTa. On the FACTIFY dataset, the BERT model achieved the best performance, which is 68.53% by micro F1 and 68.44% by macro F1 score on the test set, and BERT's performance BERT is approximately similar to RoBERTa's performance. Besides, the performance of the BEiT model is not much different from the ViT on the FACTIFY and Mocheg datasets.

In comparison with the baselines from Mocheg and FACTIFY, we achieve state-of-the-art performance with respect to the F1-score overall when comparing with baseline approaches, according to Table 2.5. Generally, the Longformer has a better performance in extracting features from long sequences than the BigBird. The ViT and BEiT have similar abilities in image representation, while DeiT performance is slightly lower than ViT and BEiT. Overall, the MCVE with RoBERTa for encoding the claim, Longformer for encoder the text evidence, and ViT for encoding the image achieved higher results than the baseline in the verification task, which is 52.32% by F1-score.

Feature Extractors			**Mocheg [26]**				**FACTIFY [24]**			
Claim	**Evidence**		**Test**		**Dev**		**Test**		**Dev**	
	text	image	F1-micro	F1-macro	F1-micro	F1-macro	F1-micro	F1-macro	F1-micro	F1-macro
BERT	**Longformer**	**ViT**	49.77	49.92	50.83	50.75	**68.53**	**68.44**	**70.86**	**70.91**
BERT	BigBird	ViT	50.27	49.35	51.16	50.29	68.06	68.00	70.97	71.02
BERT	**Longformer**	**BEiT**	51.57	50.73	51.66	50.97	**68.53**	**68.45**	**70.89**	**71.03**
BERT	BigBird	BEiT	51.02	49.60	52.50	51.34	68.38	68.27	71.26	71.30
RoBERTa	**Longformer**	**ViT**	**52.32**	**50.94**	**53.83**	**52.35**	68.16	68.13	71.90	72.01
RoBERTa	BigBird	ViT	51.27	50.28	50.66	49.63	67.92	67.95	71.61	71.64
RoBERTa	Longformer	BEiT	50.57	48.96	53.66	52.71	60.26	59.19	63.24	61.91
RoBERTa	BigBird	BEiT	48.27	46.93	51.66	50.57	68.41	68.38	72.13	72.23
BERT	Longformer	DeiT	49.77	48.05	50.33	48.83	68.26	68.18	71.10	71.14
BERT	BigBird	DeiT	47.47	46.65	51.00	49.19	67.92	67.94	71.29	71.54
RoBERTa	Longformer	DeiT	50.32	48.98	53.83	52.63	67.84	67.88	71.34	71.43
RoBERTa	BigBird	DeiT	50.17	49.48	51.83	51.51	45.04	36.83	45.54	37.48
	Baseline		50.78	-	-	-	-	53.09	-	-

Table 2.5: Empirical results of MCVE for claim verification task

Besides, we compared our proposed model's performance against state-of-the-art models using the Mocheg and FACTIFY datasets. As detailed in Table 2.6, our model outperformed the LLaVa-7B by approximately 8%, although it is significantly lighter in size. Although the LLaVa-7B model has hundreds of millions of parameters (for example, BERT, RoBERTa, Longformer, BigBird, ViT, BEiT, and DeiT ranges from 100M to 200M parameters), our model, even when combining three components for vector representation, totals fewer than 400M parameters (Table 2.7). Furthermore, our approach surpassed the CLIP-large-336 model, even when the latter was enhanced with GPT-4o explanations. However, the LLaVa-13B model demonstrated a significant performance boost, showcasing the advantages of large language models (LLMs) in multimodal understanding. Despite this, our methodology, which focuses on data fusion from multimodal representations, remains promising for practical applications, particularly in resource-constrained environments. With FACTIFY, ECENet, introduced by [50], achieved a notable improvement in the F1-score compared to our MCVE. This enhancement stems from ECENet's use of a deep reinforcement learning (DRL) paradigm, incorporating a sentence extractor for justification to identify key sentences for claim verification. Specifically, the authors in [50] design a reward model that continuously updates the confidence score of the document via the action (the action in [50] is a module that represents the sentence similarity in the document via Bernoulli distribution). Then, they continuously update the reward model with a policy gradient loss function to obtain the best sentence selection. In contrast, our model simply employs a fusion and attention mechanism that integrates claims with textual and visual evidence to construct a deep neural network for training. Despite its simpler architecture and smaller parameter size, our MCVE achieves competitive results against state-of-the-art models.

SOTAs	**Mocheg**	**FACTIFY**
ECENet [50]	-	81.50
CLIP-large-336 + GPT4o guided [54]	45.54	-
LLaVA-7B [96]	44.80	-
LLaVA-13B [96]	58.10	-
MCVE (Ours)	**52.32**	**68.53**

Table 2.6: Comparison between State-of-the-art (SOTA) models and MCVE on Mocheg and FACTIFY

For the explanation task, it can be seen that the LED model achieved the highest performance, which is 23.63 by BLEU, and 39.09 by ROUGE-L as shown in Table 2.8. LED, BART, and T5 models achieved higher performance than the baseline in Mocheg [26]. The results show the necessity for visual information integration into the generation model to generate high-quality human text explanations for the fact-checking task. Besides, the performance of generative models when using the caption of the image from vit-gpt2-image-captioning and InstructBLIP is not much different. The performance of LED with InstructBLIP is slightly higher than the result with vit-gpt2-image-captioning on the test set, which is 23.95 by BLEU and 38.95 by ROUGE-L, while it is significantly improved on the development set. In general, the enriched information from images helps improve the explanation model.

Feature Extractor			Total parameters
Claim	Evidence		
	text	image	
BERT	Longformer	ViT	350M
BERT	BigBird	ViT	329M
BERT	Longformer	BEiT	349M
BERT	BigBird	BEiT	328M
RoBERTa	Longformer	ViT	366M
RoBERTa	BigBird	ViT	345M
RoBERTa	Longformer	BEiT	366M
RoBERTa	BigBird	BEiT	344M
BERT	Longformer	DeiT	350M
BERT	BigBird	DeiT	329M
RoBERTa	Longformer	DeiT	366M
RoBERTa	BigBird	DeiT	345M

Table 2.7: The total parameters of MCVE by each pre-trained

	vit-gpt2-image-captioning				Instruct BLIP			
	Test		Dev		Test		Dev	
	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L
LED (allenai/led-base-16384)	**23.67**	**38.04**	**55.41**	**70.58**	**23.95**	**38.95**	**57.86**	**73.70**
BART (facebook/bart-base)	22.92	39.16	55.52	71.19	22.08	38.36	52.16	67.35
T5 (t5-base)	20.27	33.84	48.96	62.34	21.09	36.45	49.01	62.66
Baseline	21.84	35.41	-	-	21.84	35.41	-	-

Table 2.8: Empirical results of MCVE for claim explanation task on the Mocheg

2.5.3 Ablation Study

In this section, to clarify the impact of the claim features ((Module (1) in Figure 2.1)), attention module ((Module (2) in Figure 2.1)), fusion module ((Module (3) in Figure 2.1)) and the convolutional module ((Module (4) in Figure 2.1)) for the claim verification task, we conduct an ablation study on Mocheg and FACTIFY development sets with 4 cases described in Figure 2.1. Table 2.9 shows the results of an ablation study recorded in macro F1-score.

As shown in Table 2.9, without the Convolutional module, the claim feature brings optimistic performance on Mocheg and FACTIFY. From a technical perspective, the Convolutional module has a relatively minor impact on overall performance. Notably, the fusion features, even without convolutional components or claim integration, achieve strong results in the claim verification task. This highlights the critical role of claim features

	Dev macro F1	
	Mocheg	**FACTIFY**
w/o claim	21.13	35.91
w/o convolution	38.66	70.57
w/o attention	50.67	57.55
w/o fusion	46.59	70.69
w/o evidence (claim only)	46.84	70.42
Full	**52.35**	**70.91**

Table 2.9: Ablation study of claim verification models

in enhancing the model's effectiveness. Although the Convolutional module performs poorly in isolation on the Mocheg, its performance improves significantly when integrated with claim features, fusion, and attention. Additionally, the attention module has a substantial impact on the FACTIFY dataset, while its effect is less pronounced on Mocheg. Conversely, the fusion module proves more effective on Mocheg compared to FACTIFY. Overall, the four modules in MCVE complement each other, collectively improving claim verification performance across both datasets. Besides, it can be seen that the evidence also has a significant effect in verifying the truthfulness of the claim. Without the evidence, the performance of the model decreases in both Mocheg and FACTIFY.

	Dev BLEU	**Dev ROUGE-L**
w/o image evidence	53.60	68.38
w/o text evidence	50.87	65.33
w/o text and image evidence	53.75	68.12
Full	**55.41**	**70.58**

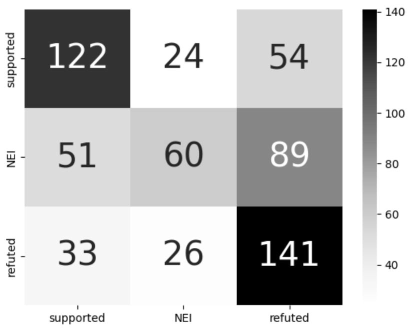
Table 2.10: Ablation study of claim explanation models

In addition, to demonstrate the impact of text and image evidence on the generation model for the explanation task, we show the ablation study results in Table 2.10. According to Table 2.10, the text evidence has a vital impact on the generation model to generate the ruling sentence. The image information, despite little effect when independent from the claim, helps increase the efficiency of the explanation model if integrated with text evidence and the claim. On the other hand, the claim itself also plays an important role in the ruling generation process when there is no text or image evidence.

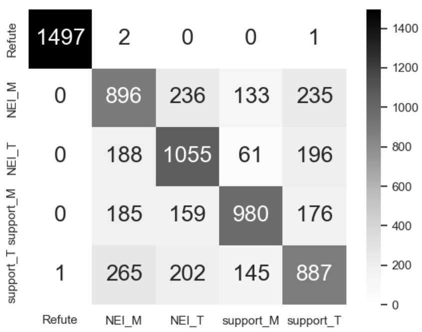
2.6 Error Analysis

2.6.1 Quantitative Analysis

Figure 2.7 illustrates the performance efficiency of the best model on the Mocheg and FACTIFY development sets (according to the empirical results from Table 2.5. As one can observe, the model on the FACTIFY dataset shows the highest classification ability when the number of true predicted labels is more than the number of incorrect labels. Specifically, the **Refuted** label has the least incorrect prediction. The **NEI** (not enough information) label is divided into the NEI multimodal **(NEI_M)** and NEI text **(NEI_T)** in the FACTIFY dataset since there is a misclassification between these two classes, as shown in Figure 2.7. Also, the proportion of incorrect classification in **NEI_M** and **NEI_T** is lower than in true classification. This is the same for the support label, while there is a little confusion between the support multimodal **(support_M)** and support text (**support_T**). In general, the proposed model can provide promising results in predicting the **Refuted** claim, while it is a little confused in discriminating between **support_M** with **support_T** and **NEI_M** with **NEI_T**. Further, the proposed model also has issues with the misclassification between the **NEI** (include **NEI_M** and **NEI_T**) and support (include **support_M** and **support_T**) labels.



Mocheg



Factify

Figure 2.7: The confusion matrix of claim verification on Mocheg and Factify development datasets. The x-axis is denoted as predicted values, and the y-axis is denoted as true label

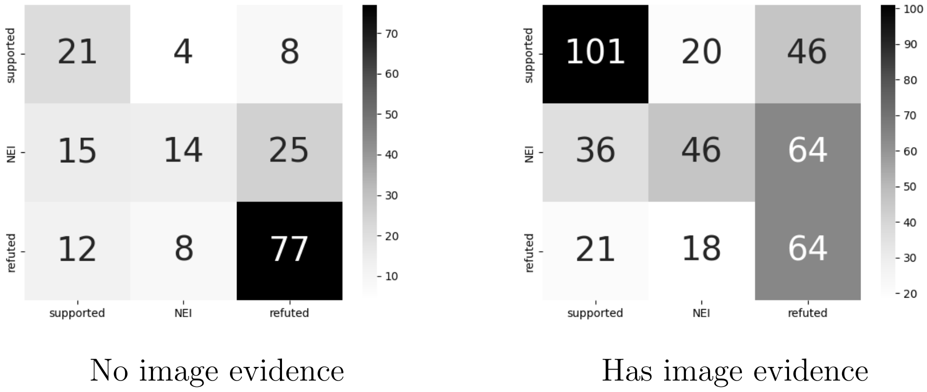


Figure 2.8: The confusion matrix of claim verification on Mocheg dataset with image and no image evidence. The x-axis is denoted as predicted values, and the y-axis is denoted as true label

For the Mocheg dataset, the misclassification instances focus on the **NEI** label. In Figure 2.7, the number of **NEI** claims that are predicted as **Refuted** is more than the number of correct **NEI** claims. In addition, the number of **NEI** claims predicted as support also accounted for a large proportion. According to Table 2.3, the number of claims that have no image evidence takes a significant proportion (nearly 12%). Hence, we compare the ability of the verification model in two cases: the claim has no image evidence and the claim has more than or equal to one image evidence. As shown in Figure 2.8, when the image evidence appears, the number of correct predictions of the **NEI** label increases. Notwithstanding that the misclassification between **NEI** and **Refuted** labels is still high, we can see that the current drawback of the model is the lack of information that integrates the feature from the claim and text evidence with the image evidence.

We investigate the ability of the explanation model to generate the ruling sentence for the claim. Figure 2.9 reports the word frequency distribution of the generated explanation and the ground-truth ruling sentences. It can be seen that approximately 44% of the generated ruling sentence in the development set begins with "We rate this Facebook post Pants on Fire", and 62.5% of the generated sentence has the pattern "Our ruling ....". In comparison with the ground-truth explanation, the generated ruling sentence is not diverse and coherent. This makes it a challenge for the explanation model to generate the correct and coherent ruling sentence for the claim based on its multimodal evidence.

Generated explanation Ground-truth ruling

Figure 2.9: Word cloud distribution of claim explanation generation on Mocheg dataset

In Table 2.11, we describe the **Supported** and the **NEI** has no ground-truth ruling sentence, and the ruling sentence is only in refuted classes. While the explanation model returns 8% of the ruling sentence for the claim that has no actual ruling sentence, the explanation model cannot generate about 5.5% of the claim that has ruling sentences for the **Refuted** label. Whereas the proportion of error ruling generated sentences is slight, there is a limitation of the explanation model on the hallucination problem in which the explanation model generates unfaithful ruling sentences for claims.

	Prediction		**Ground-truth**	
	Has ruling	No ruling	Has ruling	No ruling
Support	16	184	0	200
NEI	20	180	0	200
Refuted	180	20	200	0

Table 2.11: Statistic about the ruling sentence by classes in the development set of Mocheg

From the error analysis, we can conclude three current challenges for the multimodal fact-checking and explanation task:

- Integrate the information from image and text to extract the valuable features for the fact verification task.
- The explanation model that generates the ruling sentence is still not as diverse as the ground-truth ruling sentence.

- The explanation model generates the hallucination explanation for the non-ruling claim, and cannot generate the explanation for the claim that can be explained.

2.6.2 Qualitative Analysis

Table 2.12 illustrates several wrong prediction samples from the development set of the Mocheg. We categorized the error-predicting samples into 3 small groups corresponding to three truthfulness labels in the Mocheg dataset. Each sample has two misclassified to the remaining labels, as displayed in Table 2.12. For the first example, it can be seen that the image evidence seems to support the claim, with the fact that the car is burning, and the mooses playing in a wadding pool. However, text evidence contradicts the claim, where it is mentioned that the moose " gathered for the impromptu neighborhood bonfire". Hence, the truthfulness of this claim is refuted while the predicted label is supported because the two image evidence shows the relevancy. For the second example, text evidence 2 and 3 declare that the claim is false, while the image evidence does not provide enough relevant information to the claim (it is said that "Ninety pounds of cocaine was found on a boat owned by Mitch McConnell's family". However, the image does not describe any relevant information about McConnell's family or the finding of cocaine). Also, in the third example, the claim is true, and the image also shows support for the claim when it describes the 2 characters mentioned in the claim together (Alice Cooper and Jean Stapleton). However, the claim evidence seems to reject the claim where it is said that "... we don't have any info in that regard. Photographer is unknown to us". Moreover, for Example 5 and 6, the evidence contains many informal tokens and characters such as the hashtag token (#), mention token (@), and URL, which confuse the verification model. Specifically, for Example #4, the claim is correct (supported) if we infer the numeric fact from the evidence, which is "International Space Station orbits the Earth with the distance of 416 km, while the distance to the landmass is 2,700 km. So the distance from the Nemo Point to ISS is nearer than the landsides", we can easily clarify that the claim is true (supported). Nevertheless, the model predicts it as NEI while it can be reason the truthfulness from the numeric fact in the evidence thus it is a lack of reasoning ability in numerical fact and information from the evidence. Additionally, the inconsistency between textual and image as well as the characteristic in text evidence also impacts the performance of the verification model.

In addition, we also investigate several prediction samples on the explanation task, as shown in Table 2.13. In the first example, it can be seen that the generated explanation is similar to the ground truth where it describes sufficiently the truthfulness of the fact in the claim. For the second example, the explanation model just generated a short explanation mentioning "We rate this Facebook post Pants on Fire." for a refuted fact while there is no further information that supports the truthfulness verification as the ground-truth ruling. For example #3, this is the same for the generated explanation sentence. However, the result is nearly matched with the ground-truth ruling where it is only verbose the justification for the truthfulness of the claim without further explanation. Specifically, in Example #4, the explanation model cannot generate the explanation sentence when using the image description from InstructBLIP although the description generated by InstructBLIP is more detailed than the vit-gpt-2-image-captioning. As seen from the example, the description generated by InstructBLIP is detailed but it is not relevant to the claim and the text evidence. In contrast, the description generated by the vit-gpt-2-image-captioning is shorter and more general (it just mentioned a man in a suit with a cane) making the explanation model generate an explanation for the claim with enough information about the truthfulness of the label. Moreover, for Example #5 and #6, the explanation model generates the ruling sentence for the truthfulness of the claim with sufficient facts and details according to the evidence in the text and image. However, this explanation is not true because the ground truth is no explanation. The truthfulness of the claim can be defined directly from the evidence without explanation. Overall, the evidence and the claim itself play an essential role in the explanation of the truthfulness. However, in some cases, they generate the ruling sentence as an explanation that is reasonable and detailed but it is not the correct one.

Additionally, in Example #4 from Table 2.13, we can see that the Ground-truth ruling sentence for the truthfulness of the claim contains information from external knowledge. For example, to prove the claim that "Democratic nominee Joe Biden proposed the largest tax increase in modern U.S. history during his 2020 presidential campaign" is false (refuted), the moderator provided information about the meaning of GDP and how it impacts the economic and regulatory policies, investigating the federal tax revenues between 1940 and 2012, and compared with the five largest tax increases to show that "Biden's tax proposal is not the biggest hike in modern U.S. history". Those pieces of information are not directly retrieved from both text and image evidence since the generated explanation cannot mention this as humans do. The facts in the generated sentence by explanation model are just from the text evidence, despite its conclusion as "We rate it Mostly False.". In this case, the ground-truth ruling sentence combines the information from not only available evidence but also from external

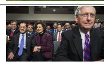
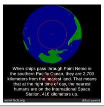
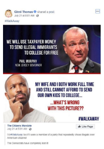
#	Claim	Text Evidence	Image Evidence	Ground Truthfulness	Predict Truthfulness
1	"A photograph shows moose enjoying a wading pool while watching a car burn."	**Evidence 1:** "< p >It seems as if they were trying to jump start it. Obviously they don't know their cars too well. The whole neighborhood has gathered for the impromptu neighborhood bonfire.< /p >"	 	refuted	supported
2	"Ninety pounds of cocaine was found on a boat owned by Mitch McConnell's family and that they are 'financing his Senate campaigns with their cocaine drug profits."	**Evidence 1:** "< p >Colombian authorities found cocaine hidden on a cargo ship belonging to a company owned by Mitch McConnell's family. This happened in 2014.< /p >" **Evidence 2:** "< p >No one was charged in the seizure.< /p >" **Evidence 3:** "< p >There is no evidence that members of McConnell's family are drug smugglers.< /p >"		refuted	NEI
3	"A photograph shows rocker Alice Cooper together with actress Jean Stapleton."	**Evidence 1:** "< p >'The photo is from a celeb-riddled charity gala at the Hollywood Bowl called 'A Shakespeare Cabaret' on August 19, 1973. Alice performed 'Gutter Cat Vs The Jets/Streetfight' (from the 'School's Out' album) that night. Stapleton may have also performed, but we don't have any info in that regard. Photographer is unknown to us.'< /p >"	 	supported	refuted
4	"There is a point in the middle of the Pacific Ocean so far from land that it's likely the closest humans to it are aboard the International Space Station".	**Evidence 1:** "< p >Point Nemo is so far from land, the nearest humans are often astronauts. The International Space Station orbits the Earth at a maximum of 258 miles (416km). Meanwhile the nearest inhabited landmass to Point Nemo is over 1,670 miles (2,700km) away.< /p >"	 	supported	NEI
5	"The #WalkAway Campaign used ads featuring stock photograph models and claimed they were Democrats leaving the party."	**Evidence 1:** "< p >I don't know who created the memes. I first saw them within the last week. Anything officially released by the #WalkAway Campaign will bear our 'branding' and trademark. Many people are excited and energized about #WalkAway, and in this excitement have created their own materials which are not approved or condoned by the official #WalkAway Campaign.< /p > < p >#WalkAway will exist beyond social media but currently does not. It is an LLC, and will exist soon as a non-profit, but currently we are building toward this.< /p >" **Evidence 2:** "< p >These memes have nothing 2do w/ the #WalkAway Campaign. They're being circulated by the left as evidence that #WalkAway is paid actors. So, in a rare moment of agreement, I am on the same page as those on the left- this is fake. These r not from the #WalkAway Campaign. pic.twitter.com/nN3kNlBAsr< /p > < p >- Brandon Straka (The Unsilent Minority) (@usminority) July 24, 2018< /p >" **Evidence 3:** "< p >Ginni Thomas, wife of Supreme Court Justice Clarence Thomas, shared images of black people who had allegedly left the Democratic Party.< /p > < p >They were actually models in royalty-free stock photos. pic.twitter.com/FjfpZzv8mN< /p > < p >- Mark Joseph Stern (@mjs_DC) July 23, 2018< /p >" **Evidence 4:** "< p >Walk Away is a scam created by Russian bots. In reality, I have never met a liberal who says, 'You know what, I've decided I hate people and their basic human rights. I think I'll become a Republican.' https://t.co/lb2VCeOVcm< /p > < p >- James Kosur (@JamesKosur) July 5, 2018< /p >"	 	NEI	refuted
6	"In the days leading up to the 2020 general election, social media users encouraged one another to wear pearls in honor of Democratic vice-presidential candidate Kamala Harris, not the late Justice Ruth Bader Ginsburg."	**Evidence 1:** "< p >I'm so ready for today!!! Ladies don't forget your pearls today In honor of RBG! #BidenHarris2020 #vote pic.twitter.com/ngjDunEu0m< /p > < p >- Christine (@christine_p_33) November 3, 2020< /p >"		NEI	supported

Table 2.12: Prediction examples for verification task

#	Claim	Text Evidence	Image evidence	Truthfulness	Ground-truth Ruling	Generated Explanation
1	"Says Biden supports a school choice ban that would impact 74,000 Wisconsin students."	**Evidence 1:** "< p >Biden does oppose one stripe of school choice programs, ones that apply public tax dollars to parents' private school tuition bills.< /p >" **Evidence 2:** "< p >And the president doesn't have the power to 'ban' voucher programs - they are authorized and funded by the state.< /p > < p >But the term 'school choice' also covers an array of other programs that Biden supports. < /p >"		refuted	"Our ruling A viral post on Facebook from a pseudo-news site says a 'Biden school choice ban would impact 74,000 Wisconsin students.' Biden does oppose one stripe of school choice programs, ones that apply public tax dollars to parents' private school tuition bills. But everything else in this headline is wrong. The term 'school choice' also covers an array of other programs that Biden supports. And the voucher program in Wisconsin is about 42,000 students, not 74,000. But most notably, the assertion that Biden has the power to 'ban' such voucher programs simply doesn't hold water. A president and Congress that opposed vouchers could put some hurdles in place, but the Wisconsin program is authorized and funded by the state. We rate this claim False."	"Our ruling A Facebook post claims that Biden supports a school choice ban that would impact 74,000 Wisconsin students. That's wrong. The term'school choice' covers a variety of programs that are authorized and funded by the federal government. And the president doesn't have the power to 'ban' voucher programs. We rate this claim False."
2	"The animated show Darkwing Duck will return to television in 2018."	**Evidence 1:** "< p >It looks like we, and TV Overmind, were tricked by an old April Fool's post dated April 2, 2015. Sorry about that.< /p >" **Evidence 2:** "< p >Darkwing Duck, which aired its last new episode in 1992, will return to TV in 2018. Plenty of time to get dangerous.< /p >"		refuted	"On 21 January 2016, several web sites such as i09 and TV Overmind published articles claiming that the popular animated show Darkwing Duck would return to television in 2018: Darkwing Duck, which aired its last new episode in 1992, will return to TV in 2018. Plenty of time to get dangerous. While both i09 and TV Overmind are both credible news sources, the story about Darkwing Duck's revival originated with an April Fool's joke that was published by Nerd Reactor on 1 April 2015. Shortly after realizing that they had been duped, i09 published a retraction: It looks like we, and TV Overmind, were tricked by an old April Fool's post dated April 2, 2015. Sorry about that. As of January 2016, Disney has not announced any plans to revive the Darkwing Duck television show."	"We rate this Facebook post Pants on Fire."
3	"'Leftists' sent 'Proud Boys' emails threatening Democratic voters."	**Evidence 1:** "< p >U.S. intelligence officials have said they believe Iran sent messages threatening some voters using a fake Proud Boys email address.< /p >"		refuted	"We rate this blog post False."	"We rate this Facebook post Pants on Fire."
4	"During his 2020 presidential campaign, Democratic nominee Joe Biden proposed the largest tax increase in modern U.S. history."	**Evidence 1:** "< p >Biden's proposal is primarily motivated by fairness. Many U.S. taxpayers are concerned that large corporations aren't paying their 'fair share.' While understandable, this position ignores the belief among many tax and economic policy experts that a significant portion of corporate taxes are ultimately paid by workers, shareholders, and consumers as opposed to companies themselves. < /p >" **Evidence 2:** "< p >They want to tax $4 trillion, going to be the biggest tax increase in history by far. It will triple up records, and they're big taxers. It's just something that won't work. We'll have - you will see a depression, the likes of which you have never seen.< /p > " **Evidence 3:** "< p >His plan would raise taxes on households in the top 1% of the income distribution (those with income more than $837,000) by an average of about $299,000. [...]< /p > < p >By contrast, taxpayers in the middle-income quintile (those with income between $52,000 and $93,000) would experience an average tax increase of just $260. [...]< /p > < p >Taxpayers in the bottom quintile (those with income less than $26,000) would see an average tax increase of only $30.< /p >"	 **Description by vit-gpt2-image-captioning**: "a man in a suit with a cane" **Description by InstructBLIP**: "his image depicts a man dressed in a suit and tie, wearing a surgical mask on his face. He is making a gesture with his hand, which could be interpreted as either a wave or a fist bump. In the background, an American flag can be seen, emphasizing the patriotic aspect of the image."	refuted	"In sum, the overwhelming majority of Americans - those who earn $400,000 or less annually - would not see their taxes increase as a direct result of Biden's tax proposal, and the top 1% earners would shoulder the brunt of his proposed hike. But low-to-middle-class workers would see slightly lower investment returns and wages as a result of corporate tax increases, according to economists' analysis including a a budgetary model by the Wharton School of the University of Pennsylvania. How Does the Proposal Rank Historically? Biden's plan would raise about $3.38 trillion in additional tax revenue over 10 years, according to the university's model. Concurrently, per Gleckman's analysis, that would amount to an increase in federal revenue by about 1.5% of the country's gross domestic product (GDP). GDP is the economy's total output of goods and services, and economists rely on it greatly to measure the economy's health and how fiscal or regulatory policies could impact it - not raw dollar amounts that don't consider inflation and population changes. To determine the legitimacy of Trump's claim, we considered federal tax proposals and their impact on GDP over decades. Of almost two dozen congressional tax bills that increased federal tax revenues between 1940 and 2012, for example, the five largest tax increases raised annual revenue between 1.33% and 5.04% of GDP, according to a 2013 compilation of such records by the U.S. Department of Treasury. The following table compiled by the Tax Foundation, another independent public policy think tank, showed where Biden's plan would stand in comparison to those other proposals: According to those findings, Biden's tax proposal, if enacted by Congress, would amount to the fifth-largest tax increase since the 1940s - not the biggest hike in modern U.S. history."	**Result with vit-gpt2-image-captioning**: "Our ruling A Facebook post claimed that during his 2020 presidential campaign, Democratic nominee Joe Biden proposed the largest tax increase in modern U.S. history. Biden's plan would raise taxes on households in the top 1% of the income distribution by an average of about $299,000. While understandable, this position ignores the belief among many tax and economic policy experts that a significant portion of corporate taxes are ultimately paid by workers, shareholders, and consumers as opposed to companies themselves. The statement is partially accurate but leaves out important details or takes things out of context. We rate it Mostly False." **Result with InstructBLIP**: ""
5	"President Trump signed a bill blocking Obama-era background checks that would have made it easier to flag mentally ill potential gun purchasers."	**Evidence 1:** "< p >The Obama rule would have prevented an estimated 75,000 people with mental disorders from being able to purchase a firearm. It was crafted as part of Obama's efforts to strengthen the federal background check system in the wake of the 2012 massacre of 20 young students and six staff at Sandy Hook Elementary School in Newtown, Connecticut.< /p >"	 **Description by vit-gpt2-image-captioning**: "a man in a suit and tie speaking into a microphone" **Description by InstructBLIP**: ""	supported		**Result with vit-gpt2-image-captioning**: "Our ruling Trump signed a bill blocking Obama-era background checks that would have made it easier to flag mentally ill potential gun purchasers. It was crafted as part of Obama's efforts to strengthen the federal background check system in the wake of the 2012 massacre of 20 young students and six staff at Sandy Hook Elementary School in Newtown, Connecticut. The Obama rule would not have prevented an estimated 75,000 people with mental disorders from being able to purchase a firearm. We rate this claim Mostly True. MOSTLY TRUE - The statement is accurate but needs clarification or additional information. Click here for more on the six PolitiFact ratings and how we select facts to check." **Result with InstructBLIP**: "Our ruling Clinton said Trump signed a bill blocking Obama-era background checks that would have made it easier to flag mentally ill potential gun purchasers. Clinton's claim is accurate, but needs clarification or additional information. We rate her statement Mostly True."
6	"Police shootings kill more white Americans than Black Americans."	**Evidence 1:** "< p >According to the most recent census data, there are nearly 160 million more white people in America than there are black people. White people make up roughly 62 percent of the U.S. population but only about 49 percent of those who are killed by police officers. African Americans, however, account for 24 percent of those fatally shot and killed by the police despite being just 13 percent of the U.S. population. As The Post noted in a new analysis, that means black Americans are 2.5 times as likely as white Americans to be shot and killed by police officers.< /p >" **Evidence 2:** "< p >An analysis shows that more white people died at the hands of law enforcement than those of any other race in the last two years, even as the Justice Department, social-justice groups and media coverage focus on black victims of police force.< /p >" **Evidence 3:** "< p >I think it's pretty obvious that black people are killed at much higher rates than white people. I'm not going to say that white people are underrepresented in these numbers, since I think all people are overrepresented in this data, but it's clear that black people are highly overrepresented.< /p >"	 **Description by vit-gpt2-image-captioning**: "a table with a pair of scissors and a pen" **Description by InstructBLIP**: "According to the claim, police shootings kill more white Americans than Black Americans in the United States. In the image, we can see a gun and bullet casings on a white surface. The presence of the gun and bullet casings implies that there has been a shooting or an event involving firearms. However, it is important to note that this image alone does not provide enough information to determine the ethnicity of the individuals involved in the shooting."	NEI		**Description by vit-gpt2-image-captioning**: "Our ruling Trump said police shootings kill more white Americans than Black Americans. The most recent census data shows that black people are 2.5 times as likely to be shot and killed by police than white people. However, that doesn't mean black Americans account for 24 percent of the U.S. population. We rate Trump's statement Half True. HALF TRUE - The statement is partially accurate but leaves out important details or takes things out of context. Click here for more on the six PolitiFact ratings and how we select facts to check." **Description by InstructBLIP**: "Our ruling The Washington Post said police shootings kill more white Americans than Black Americans. That's not entirely accurate. According to the most recent census data, there are nearly 160 million white people in America, but only about 49 percent of the U.S. population. We rate this claim Half True. HALF TRUE - The statement is partially accurate but leaves out important details or takes things out of context. Click here for more on the six PolitiFact ratings and how we select facts to check."

Table 2.13: Examples for explanation task

knowledge sources that are relevant to the truthfulness verification of the claim.

2.6.3 Discussion

Multimodal Contextual Alignment: Although the fusion techniques combine the valuable features from textual and visual data, in some cases, there is an inconsistency between the image and text that affects the performance of the claim verification model (as shown in the Qualitative Analysis). To overcome this issue, the claim verification model must have the ability to deeply understand both textual and visual information, and then align them together contextually.

Multimodal Reasoning: There are several cases in practice in which the evidence represents the relevant fact about the claim implicitly. The verification model must have the inference in both text and image to attain the correct truthfulness about the claim. In addition, the visual information also contains the scene text, which is a valuable information source for claim verification. Thus it requires the ability of science text recognition and visual-semantic representation to exploit this semantic feature.

Correctness and Coherence in Explanation: From the error analysis, there are two main challenges for the explanation sentence. First, although the generated explanation sentence is correct in terms of truthfulness and evidence, the descriptive information in the generated claim is not sufficient, as humans do. Second, for some claims that can be directly verified without the explanation (as shown in Qualitative and Quantitative Analysis), the explanation generated the ruling sentence that caused the hallucination in the explanation. In addition, evaluating the ruling sentence of the explanation model by BLEU and ROUGE is not enough to evaluate the performance of the Fact-checking model since the explanation model requires three main criteria: readability, plausibility, and faithfulness [97].

The role of Evidence and External Knowledge: It can be seen that the evidence plays a vital role in both verification and explanation tasks for the truthfulness of a claim. External knowledge is also necessary to enhance the ability of models to verify and explain the truthfulness of the claim. However, in this paper, we have not yet treated the Evidence retrieval task since we use available gold evidence (evidence that is collected and annotated by humans) from the dataset for the claim verification and claim truthfulness explanation tasks.

2.7 Summary

In this chapter, we introduce the **MCVE** framework that focuses on efficiently combining the claim, image, and text evidence to solve the claim verification and explanation tasks for fact-checking. The **MCVE** framework is flexible for claim that has multiple text and image evidence. Compared with the original baseline, **MCVE** provides an improved approach in both the verification and explanation tasks. Nonetheless, the limitation of our methodology is the alignment of information between the image and text to extract the most valuable information that can represent both text and image. In addition, our explanation model has a lack of diversity in the generation of the ruling sentence for the claim.

Based on the limitation of **MCVE**, one potential approach for enhancing the alignment among different modality data is LLMs, which is robust and has strong reasoning ability in language understanding. Therefore, the next Chapter describes the methodology that leverages the ability of LLMs in the fact-checking task.

Declaration: Parts of this chapter have been published in Publication [2].

Chapter 3

Large Language Models for Fact-verification

3.1 Background

Large Language Models (LLMs) generate new sequences using an autoregressive process. Given an input context x—a question Q paired with a prompt template T—the model produces a final answer A through intermediate reasoning content z (in practice, z is instantiated by prompting). We use the standard factorization shorthand:

$$p(A \mid x{=}(T, Q)) \; = \; p(A \mid x, z) \, \cdot \, p(z \mid x), \tag{3.1}$$

Following [98], the generation of an answer A proceeds via a sequence of reasoning steps C conditioned on exemplars. Let $T = \{(Q_i, C_i, A_i)\}_{i=1}^{K}$ be K in-context instructions (each with a question Q_i, steps C_i, and answer A_i). The aim is to maximize the likelihood of answer A given the instruction T and the question Q. The likelihood then decomposes as:

$$\begin{aligned} p(A \mid x{=}(T, Q)) = & \prod_{i=1}^{|C|} p_{\mathrm{LM}}(c_i \mid x, c_{<i}) \\ & \cdot \prod_{j=1}^{|A|} p_{\mathrm{LM}}(a_j \mid x, C, a_{<j}), \end{aligned} \tag{3.2}$$

where $C = (c_1, \ldots, c_{|C|})$ denotes the reasoning steps and p_{LM} is the model's conditional distribution.

Since LLMs have been trained on large-scale datasets, they can capture the nuances and patterns in natural language and generate meaningful text for verification and justification in fact-checking [99]. Leveraging the ability of LLMs can help improve the performance and accuracy of the Fact-checking task. However, with complex and diverse data modalities, it acquires an efficiency in prompting to instruct the LLMs for better understanding the claim and reasoning on the given evidence to determine the veracity of the information. In this chapter, we propose three frameworks that leverage the LLMs for fact-checking with an efficiency prompting method:

- **ZeFAV**: This method employs the relation extraction with reorganizing information (InfoRE) to instruct the LLMs for fact checking on textual data. The details are described in Section 3.2.
- **TabV4FC**: This method enhances the performance of fact-checking on tabular data by leveraging the description of the table through a table-to-text model like TAPEX [100] to instruct the LLMs. TabV4FC are described in Section 3.3.

3.2 Fact-Verification via Relation Extraction and Re-organizing information

3.2.1 Introduction

The in-context learning ability of large language models demonstrated the effectiveness in verifying the truthfulness of complex claims. ProgramFC [9] recently used the in-context learning ability of large language models to break down the input claims into reasoning sub-tasks and then obtain the verdicts by tackling each reasoning sub-task. Using few-shot learning, it decomposes claims into Python-like programs with functions to question, verify, and predict facts. ProgramFC demonstrates that decomposing claims into smaller tasks is more effective than a one-step prediction approach, reducing the cognitive load on the language model and enhancing its fact-checking capability. Another approach, the QACheck system [101], includes five components, each component is a LLM with a specific task for the verification process. These LLMs leverage their ability to generate text and learn from context to ask and answer key questions that determine the claim's truthfulness. In addition, InfoRE [62] improves these models' reasoning abilities by restructuring evidence into MindMap forms, which effectively represent knowledge and concepts [102]. This reorganized evidence used alone or with original data, improves claim verification. Experiments show that InfoRE achieves comparable results in fact-checking, demonstrating the effectiveness of this information reorganization method. Finally, a recent study from [103] shows that large language models excel in few-shot relation extraction (FSRE) by producing linearized strings that encode entity pairs and their relations.

To address these challenges, applying Large Language Models (LLMs) in fact verification tasks is promising due to their strong comprehension of human language [104]. Previous works like ProgramFC [9], QAChecker [101], and InfoRE [62] utilize LLMs for fact verification and show potential results. However, LLM outputs are not always consistent [104], and the reasoning process needs support for the chain of thought (CoT) [105]. Additionally, hosting and executing LLMs is costly and inefficient [106]. The authors in [101] recommend using locally run and open-source LLMs like LLaMa [107] instead of external API-based LLMs like InstructGPT or GPT-4 [108]. Therefore, our work aims to enhance LLMs for fact-checking by enriching and guiding their understanding and reasoning processes. We propose ZeFaV, a framework that utilizes few-shot relation extraction and reorganized information (InfoRE) to boost the performance of LLMs for the fact verification task. Few-shot Relation extraction identifies entities in the claim and evidence and their relationships, using these relations to guide LLMs in understanding and verifying the claim, while InfoRE reconstructs the evidence provided with the claim to help LLMs better comprehend it. Our empirical results show that our proposed method improves the ability of LLMs for the fact verification task. All of our experiments are based on zero-shot learning on LLMs.

3.2.2 Datasets

In this work, we evaluate the HOVER [109] and FEVEROUS-S [110] datasets, both of which contain multi-hop claims requiring evidence from multiple sources for verification. We utilize the validation sets from both datasets for our experiments. The HOVER dataset includes multi-hop claims derived from Wikipedia articles. We employed the division by Pan et al. [9], which separated the validation set into three subsets: 1,126 two-hop claims, 1,835 three-hop claims, and 1,039 four-hop claims.

Table 3.1: Number of claims on each challenge in FEVEROUS dataset

Challenge	**Number of claims**
Multi-hop Reasoning	459
Combining Tables and Texts	7
Numerical Reasoning	103
Entity Disambiguation	112
Search Terms not in Claims	46
Others	2,235

The FEVEROUS dataset features complex claims supported by evidence in sentences or a combination of sentences and cells from Wikipedia tables. We extracted 2,962 claims requiring sentence evidence from the FEVEROUS dataset for our experiments. Each claim in the FEVEROUS dataset was annotated to classify the anticipated challenges that the fact-checking model might encounter during verification. These challenges were categorized into the following: *Multi-hop Reasoning*, *Combining Tables and Texts*, *Numerical Reasoning*, *Entity Disambiguation*, *Search terms not in claim*, and *Others*. Table 3.1 describes the number of samples per challenge in the FEVEROUS-S dataset.

3.2.3 Method

3.2.3.1 Task Formulation

Giving a sample *(c, E)*, where c is a claim sentence, and $E = \{e_1, ..., e_n\}$ is a set of evidence relevant to c. The fact verification task aims to obtain the result v such that $v \in \{$*True, False*$\}$. The *True* indicates the claim is correct (supports), and *False* indicates the claim is not correct (refutes). To solve the task, we proposed the ZeFaV - our zero-shot prompting technique leverages the text re-organizing and relation extraction to enhance the reasoning ability of the LLMs for the Fact verification task. Figure 3.1 illustrates the overview of ZeFaV.

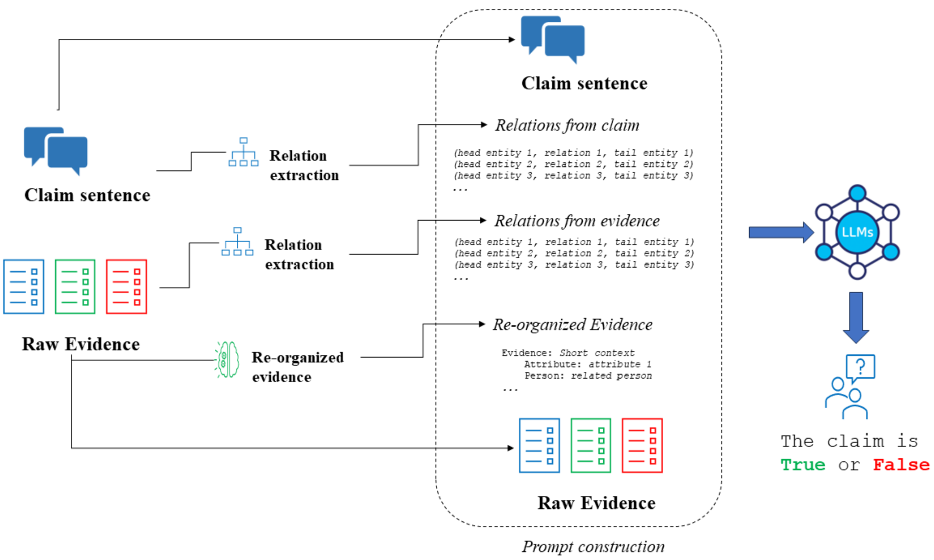


Figure 3.1: Overview of the ZeFaV framework

3.2.3.2 Extract Relation Triplet from LLMs

Recent advances in LLMs have created new possibilities for the RE task. One approach involves inputting a sentence and a predefined set of relation types to the LLM. The model is then tasked with identifying the most likely relation from this set and pinpointing the corresponding head and tail entities within the sentence. This is achieved through carefully crafted prompts that highlight the specific information needed. For instance, a prompt might instruct the LLM to "Identify the relation between *entity1* and *entity2* in the *sentence* and return the triplet *(head, relation, tail)*." Leveraging their understanding of sentence structures and language patterns, LLMs efficiently manage relevant entities and their connections. This method offers advantages such as adaptability to novel relations and the capability to discern subtle relationships that traditional rule-based methods may overlook. Our approach harnesses the LLM's contextual comprehension and ability to generalize from extensive datasets, presenting a robust technique for extracting relational facts crucial in automated fact-checking.

```
### Instruction: Given a sentence, please identify the head and
    tail entities in the sentence, and classify the relation type
    into one of the appropriate categories; The collection of
    categories is: [<list_relations_in_FewRel>];
Sentence: [CLAIM_SENTENCE]
### Response: (claim_head, claim_relation, claim_tail)
```

Algorithm 3.1 Finding a closure of evidence relation from the claim relation

```
Function FindingEvidenceRelations(Claim_Rels, Evidence_Rels):
    Evidence_Rels_News ← {}
    Hypos ← {}
    is_found ← True
    Hypos ← Hypos ∪ head(claim_rel) ∪ tail(claim_rel) | claim_rel ∈ Claim_Rels
    while is_found is True do
        for evidence_rel ∈ Evidence_Rels do
            if head(evidence_rel) ∈ Hypos then
                Evidence_Rels_News ← Evidence_Rels_News ∪ evidence_rel
                Hypos ← Hypos ∪ tail(evidence_rel)

            if Cannot expand the Hypos anymore then
                is_found ← False

return Claim_Rels, Evidence_Rels_New
```

After fine-tuning the model, we use the prompt below to extract the relation for an input sentence (we leave the ***### Response:*** as blank thus the LLM can generate the following output to the prompt). The result is a set of relations of the claim (Claim_Rels) and evidence by sentence (Evidence_Rels). Then, we find the evidence relations that are related to the claim and remove the redundant relations in the evidence. We referred to the closure definition, which is a function that combines the bundled with their surrounding states. By using this, we can find the evidence relations that are relevant to the claim. Algorithm 3.1 shows the procedure of finding the closure of evidence relations from the claim relations.

3.2.3.3 Re-organizing the Evidence

Re-organizing information enhances large language models' performance across NLP tasks [62]. To leverage this capability, we apply the InfoRE method to structure evidence into a more informative textual format, boosting the in-context learning of these models. InfoRE summarizes and reconstructs information to reveal logical relationships and connections among entities within the evidence, facilitating easier access to knowledge compared to traditional text formats. In our study, we aim to preserve both interpretability and hidden details by combining original and re-organized evidence forms as context for large language models. From evidence set E related to claim c, we derive set E' containing re-organized evidence, structured using few-shot learning techniques by these models. The prompt template used is as follows:

```
Transform the following text into a hierarchical structure that
    organizes the information in the text into levels. The same
    level can reflect parallel relationships and indented levels
    reflect causal relationships. Here are some examples:

The evidence: [EXAMPLE EVIDENCE 1]
The hierarchical structure:
    [EXAMPLE RE-ORGANIZED EVIDENCE 1]
### The evidence: [EXAMPLE EVIDENCE 2]
The hierarchical structure:
    [EXAMPLE RE-ORGANIZED 2]
### The evidence: [EVIDENCE]
The hierarchical structure:
```

3.2.3.4 Integrating InfoRE and Relation extraction for Zero-shot

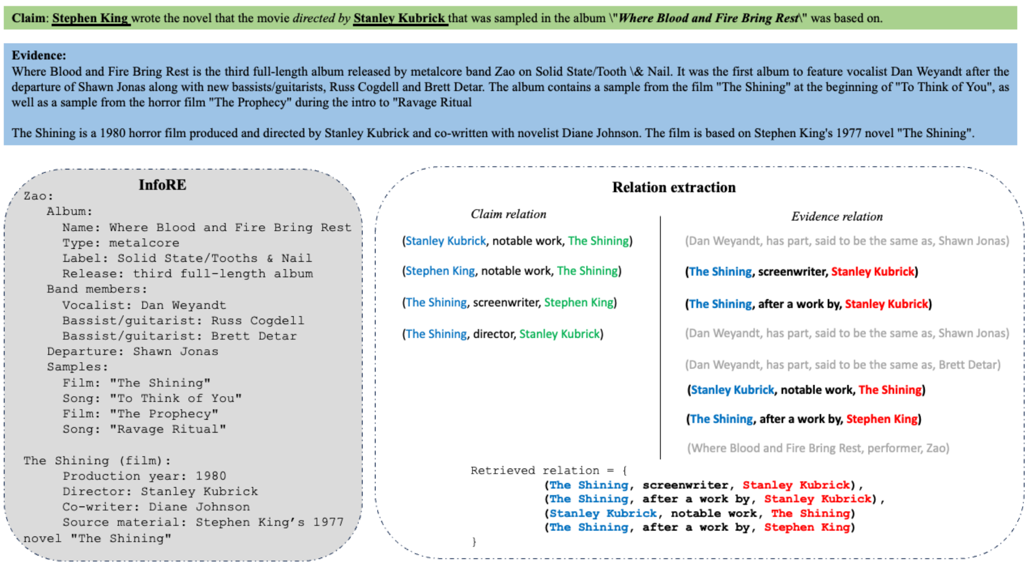


Figure 3.2: Example about the construction of InfoRE and relation extraction in ZeFAV

It can be shown that the InforRE represented the evidence in a concise form. For the example in Figure 3.2, based on the evidence on the left, the InforRe has reorganized the evidence in an easy-to-understand form. The information from the InforRE is more compact than in the raw context. In addition, the relation between the claim and evidence helps extract more meaningful semantic information used to verdict the truthfulness of the claim. As shown from the example in Figure 3.2, these relations proposed the connection between the claim and the evidence and thus help the LLMs in defining whether the claim is true or not. In addition, the relation between the claim and evidence helps extract more meaningful semantic information used to verdict the truthfulness of the claim. For example, given a claim as *"Stephen King wrote the novel that the movie directed by Stanley Kubrick that was sampled in the album Ẅhere Blood and Fire Bring Restẅas based on."*, the following extracted relation describes the inner relation in the claim and evidence. From the claim, we found evidence as *(The Shining, director, Stanley Kubrick), (Stanley Kubrick, notable work, The Shining), (Stephen King, notable work, The Shining), and (The Shining, screenwriter, Stephen King)*. From the raw context (same as the example above), we extracted relevant relations as: *(The Shining, after a work by, Stanley Kubrick),* *(Stanley Kubrick, notable work, The Shining), (The Shining, said to be the same as, film), (The Shining, after a work by, Stephen King).* These relations proposed the connection between the claim and the evidence, and thus help the LLMs in defining whether the claim is true or not, as shown in Figure 3.2.

To boost the performance of LLMs in understanding the evidence context and the relation between the claim and its evidence, we integrate the Relation representation of the claim and evidence and InforRE with the CoT technique, as described in the following prompt.

```
Documents:
    [The InforRE content]
Context:
    [The full context evidence in textual form]
Question: [The claim]?
Please answer the question based on Documents, Context, and the
    following relations. The answer must belong to one of two
    values: True or False.
    1. The question mentioned the relation between <head entity>
    and <tail entity> as <type of relation> *
    2. ...
    3. <head> and <tail> has relation with <type of relation> **
    4. ...
    Let's think step-by-step.
###The answer is:
```

*Note: * representing for claim relations, ** representing for evidence relations*

3.2.4 Results

We performed our empirical study with ZeroFaV on *Meta-Llama-3-70B-Instruct*[1]. We run the LLMs with LoRA technique [111] quantized to 4 bits, and maximum length for generation model is 2048. Our ZeFaV can run on one NVIDIA A40 GPU with 49GB memory. The F1-score computes all results reported. For the ProgramFC [9], we run the proposed methodology with N=1 on *Meta-Llama-3-70B-Instruct* for comparison with our ZeFaV. Besides, we also compare our proposed methodology with QACheck [101] (run on InstructGPT) and InfoRE [62] (run on Llama-2-70B). Table 3.2 describes our empirical results with ZeroFAV compared to other state-of-the-art models.

[1]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

Table 3.2: Emprical results of ZeFaV on the HoVer and FEVEROUS dataset

	HoVer (2-hop)	**HoVer (3-hop)**	**HoVer (4-hop)**	**FEVEROUS-S**
ZeFaV (w context)	**77.85**	**70.61**	**67.47**	**86.74**
ZeFaV (w/o context)	71.31	67.41	61.62	75.97
ProgramFC [9]	76.53	70.84	66.88	86.34
QACheck [101] *(InstructGPT)*	55.67	54.67	52.35	59.47
InfoRE [62] *(llama2-70B)*	52.40	51.21	50.07	67.96

According to Table 3.2, ZeFaV obtained better results than other methodologies on the HoVer dataset. On the FEVEROUS-S dataset, ZeFaV achieved 86.54% by F1-score, which is better than ProgramFC [9], InfoRE [62] and QACheck [101]. Since ProgramFC is a few-shot learning method for about 20 examples, according to [9], ZeFaV performs efficiently on the HoVer dataset. In comparison with InfoRE, ZeFaV outperforms InfoRE with the LLama architecture on both HoVer and FEVEROUS-S datasets. Besides, the context evidence plays a vital role in the ZeFaV when it increases the performance of the Fact verification task on both HoVer and FEVEROUS-S. It can be seen that ZeFaV showed efficient performance for the Fact verification task with zero-shot learning since it outperforms other methodologies on the same LLama architecture.

3.2.5 Error Analysis

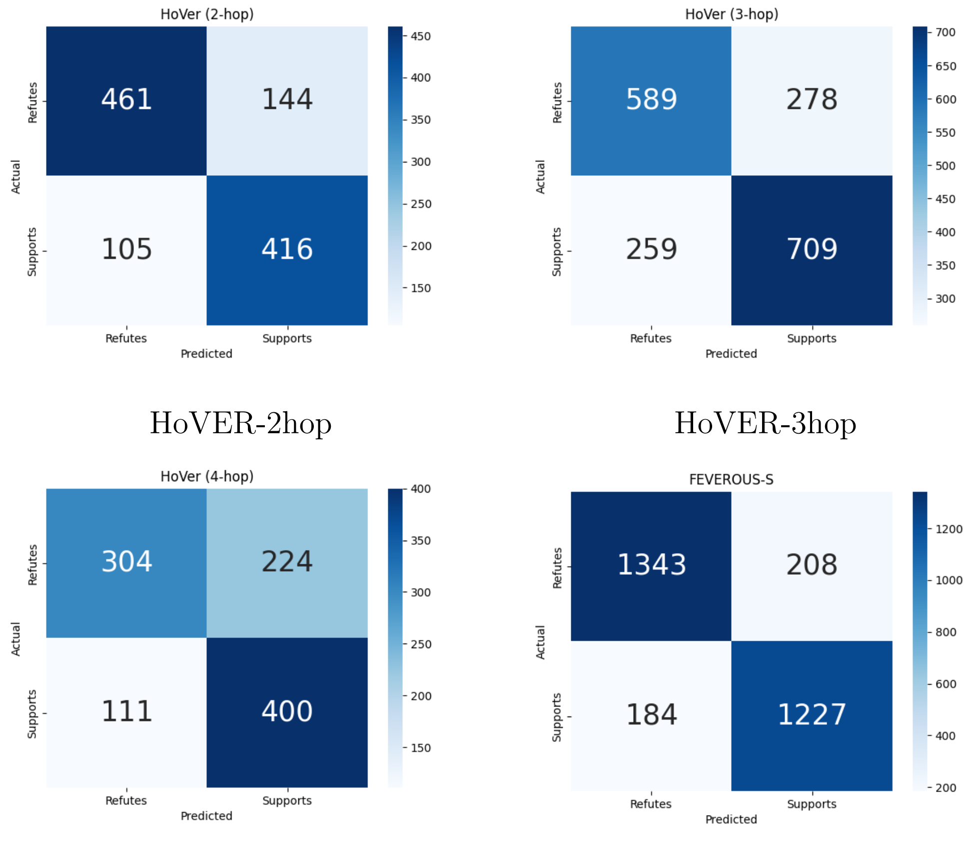
To investigate the impact of Relation and InfoRE on the performance of the LLM for Fact verification, we conduct the ablation study on two scenarios: ZeFaV with **full evidence context** and ZeFaV with **no evidence context**. Table 3.3 illustrates our ablation study of ZeFaV on the two datasets: HoVer and FEVEROUS-S.

As shown in Table 3.3, it can be seen that both InfoRE and Relation help increase the accuracy of LLMs on 2-hop and 4-hop of HoVer. Specifically, on the 3-hops of HoVer, InfoRE helps increase the accuracy while it is slightly decreased when combined with the relation. This is similar to the FEVEROUS-S where the performance slightly decreases when combining InfoRE with relation. In general, InfoRE helps the LLMs in reasoning and understanding the information by re-organizing the data in a compact and concise form. For the relation, it helps increase the ability of LLMs when combined with InfoRE. However, the performance of ZeFaV when there is only a relation is not as good as InfoRE. This is more clear when there is a lack of evidence context. The accuracy of ZeFaV with InfoRE is better than ZeFaV with relation only. Overall, both InfoRE and Relation help increase the performance of LLMs, and the evidence context also plays a vital role in the zero-shot Fact verification task.

Table 3.3: Ablation study on the performance of LLMs for Fact verification

	Relation	InfoRE	HoVer (2-hop)	HoVer (3-hop)	HoVer (4-hop)	FEVEROUS-S
Has Evidence context	✓	✓	77.85	70.61	67.47	86.74
	✗	✓	76.78	72.03	67.64	86.89
	✓	✗	76.16	69.72	66.70	85.11
No Evidence context	✓	✓	71.31	67.41	61.62	75.97
	✗	✓	71.57	68.30	63.59	76.11
	✓	✗	61.61	56.64	55.16	59.80

In addition, according to Figure 3.3, the model tends to predict the "refutes" claim to be the "support" claim rather than predicting the "supported" claim to "refutes" claim. The challenge remains in the claim that 4-hops in the HoVer dataset, where the number of wrong predictions in which refuted claims are predicted as supported claims is significant.



HoVER-2hop HoVER-3hop

HoVER-4hop FEVEROUS-S

Figure 3.3: Confusion matrix of ZeFaV on HoVer and FEVEROUS-S

Besides, we analyze the performance of ZeFaV on the FEVEROUS-

S dataset. There are six types of challenges including Search terms not in the claim, Multi-hop Reasoning, Combining Tables and Text, Entity Disambiguation, Numerical Reasoning, and others. Table 3.4 describes the performance of ZeFaV based on each type of challenge by F1-score. It can be seen that the ZeFaV performs well on almost all types of challenges. Moreover, for the Numerical Reasoning, ZeFaV is still inefficient where the accuracy by F1-score is 72.83%.

Table 3.4: Performance of ZeFaV on each type of challenge on the FEVEROUS-S dataset

Type of challenge	**F1-score (%)**
Search terms not in claim	82.47
Multi-hop Reasoning	82.72
Combining Tables and Text	100.00
Entity Disambiguation	82.80
Numerical Reasoning	72.83
Other	86.77

Additionally, Table 3.5 presents error cases from HoVer and FEVEROUS-S involving two specific challenges: 4-hop reasoning in HoVer and Numerical Reasoning in FEVEROUS-S. A frequent error example from ZeFaV involves predicting a refuted claim as supported. For instance, a claim erroneously states that “a star” in the sitcom “Then Came You” also voiced “Randall Boggs” in “Monsters, Inc.”. However, the evidence identifies “Susan Floyd” as the star of “Then Came You”, with no mention of her voicing “Randall Boggs”. Furthermore, “Randall Boggs” is a character in “Monsters, Inc.”, which does not align with the claim. This lack of semantic coherence between the claim and evidence confuses large language models (LLMs) during verification. In another example, a claim asserts that “Tracy” settled in Montreal, whereas the evidence shows he lived in London, Ottawa, and Quebec, settling in Quebec until age 18. The evidence, although mentioning places Tracy lived or worked, fails to provide crucial details such as the final residence until age 18 and the timeline of moves. Moreover, the complex structure of the evidence poses challenges for InfoRE in fully capturing the context required for accurate verification.

Finally, three main challenges remain in the ZeFaV. First, from the analysis in Figure 3.3, the model tends to miss the “refutes” claim, where it predicts the “refutes” claim as a “support” claim. Second, the model is mostly confused between predicting the “refutes” claim as a “support” claim when there are 4 hops in the evidence. Third, in the FEVEROUS dataset, the model is difficult to perform verification with Numerical Reasoning information.

Table 3.5: Several error examples

Claim	Evidence	Retrieved Relation	InfoRE	Ground truth	Predicted
The star of the sitcom Then Came You had a featured role in a motion picture. The movie stars an actor who voiced Randall Boggs in "Ghost Roaster" Source: HoVer Num of hops: 4	Steven Vincent Buscemi (born December 13, 1957) is an American actor and film director. Buscemi has starred and supported in successful Hollywood and indie films, including "Parting Glances", "New York Stories", "Mystery Train", "Reservoir Dogs", "Desperado", "Con Air", "Armageddon", "The Grey Zone", "Ghost World", "Big Fish", and "The Sopranos". He is also known for his appearances in many films by the Coen brothers: "Miller's Crossing", "Barton Fink", "The Hudsucker Proxy", "Fargo", and "The Big Lebowski". Buscemi provides the voice of Randall Boggs in the "Monsters, Inc." franchise. Susan Floyd (born May 13, 1968) is an American actress who has appeared in many episodes of "Law & Order", as well as numerous other television series. She has also had featured roles in several motion pictures, including "Domestic Disturbance" and "Forgiven", and starred opposite Al Pacino and Jerry Orbach in "Chinese Coffee". Along with mainstream films, she has also appeared in a 2003 indie film "Particles of Truth". Then Came You is a half-hour sitcom that aired on ABC for two months from March 22, 2000 to April 26, 2000. The show dealt with the romantic relationship between a young man and an older woman. It starred Susan Floyd, Thomas Newton, and Desmond Askew. Domestic Disturbance is a 2001 American psychological thriller film directed by Harold Becker (his last film to date) and starring John Travolta, Vince Vaughn, Teri Polo, Steve Buscemi, and Matt O'Leary.	**Claim relation:** (Ghost Roaster, characters, Randall Boggs) (Then Came You, has part, motion picture) (Then Came You, has part, sitcom) (Then Came You, said to be the same as, Then Came You) **Evidence relations:** (Then Came You, genre, sitcom) (Then Came You, country of origin, United States) (Then Came You, original network, ABC)	**Steven Vincent Buscemi:** Date of birth: December 13, 1957 Profession: Actor and film director Career: Film appearances: Film 1: "Parting Glances" Film 2: "New York Stories" Film 3: "Mystery Train" Film 4: "Reservoir Dogs" Film 5: "Desperado" Film 6: "Con Air" Film 7: "Armageddon" Film 8: "The Grey Zone" Film 9: "Ghost World" Film 10: "Big Fish" TV series: "The Sopranos" Collaborations: Director: Coen brothers Films: Film 1: "Miller's Crossing" Film 2: "Barton Fink" Film 3: "The Hudsucker Proxy" Film 4: "Fargo" Film 5: "The Big Lebowski" Voice acting: Character: Randall Boggs Franchise: "Monsters, Inc." **Susan Floyd:** Date of birth: May 13, 1968 Nationality: American Profession: Actress TV appearances: "Law & Order" TV appearances: numerous other television series Film appearances: "Domestic Disturbance" Film appearances: "Forgiven" Film appearances: "Chinese Coffee" Film appearances: "Particles of Truth" (2003) Co-stars: Name: Al Pacino Name: Jerry Orbach	refutes	support
Tracy Howe, who finally settled down in Montreal, lived in three different cities early in life. Source: FEVEROUS-S Challenge: Numerical Reasoning	Tracy Howe In 1952 Tracy and his family moved to London, England. Tracy Howe In 1956, when Tracy was 4 they moved to Ottawa, Canada. Tracy Howe In 1957 the family moved to Ville St-Laurent, Quebec and in 1959 moved again to Pointe-Claire, Quebec where Tracy lived until he left home in 1970 at age 18.	**Claim relations:** (Tracy Howe, work location, Montreal) (Tracy Howe, language of work or name, Montreal) (Tracy Howe, work location, three different cities) (Tracy Howe, country of citizenship, three different cities) (Tracy Howe, country of citizenship, Montreal) (Tracy Howe, language of work or name, three different cities) **Evidence relations:** (Tracy, language of work or name, English) (Tracy, work location, London) (Tracy, language of work or name, French) (Tracy, residence, London) (Tracy, country of citizenship, Canada) (Tracy, instance of, person) (Tracy, residence, Pointe-Claire) (Tracy, country of origin, England) (Tracy, country of citizenship, England)	Tracy Howe: Early Life: Location: London, England (1952) Location: Montreal, Canada (settled down)	refutes	support

3.2.6 Discussion

ZeFaV addresses the challenge of improving LLM performance in fact verification using zero-shot learning. This method leverages relation extraction and InfoRE to enhance LLM fact verification through zero-shot learning with improved evidence representation for CoT learning. The empirical study shows that the ZeFaV performance is efficient when evaluated on the HoVer and FEVEROUS-S datasets. However, from the empirical results, we found that the challenge for this task is the reasoning ability on multiple and complex information, such as non-textual and numerical contexts. Therefore, future work focuses on improving the ability to reason in complex contexts such as numeric, table, and temporal to increase the accuracy of the Fact verification task.

3.3 Fact-Verification via Table-to-text Verbalization

3.3.1 Introduction

Although the text is a primary modality for the AFC task, other modalities like tables or images provide more valuable information to enrich the ability of the Fact-checking model in reasoning for exploiting the veracity of claims based on provided information as evidence. Among various forms of information besides textual, tabular information plays an important role in Fact-checking since it is an efficient and compact method to store and represent complex scenario information in the actual world, such as financial or scientific reports [112]. However, the main challenge of the Fact-checking task on Tabular data is the numerical representation and reasoning [112]. Furthermore, the motivation for the AFC task is to construct a Fact-checking model that can be performed with minimal or even no training data [9]. Therefore, in this work, we propose a simple method that involves converting tables into textual information and using this information to generate a prompt that guides large language models (LLMs) to verify a claim.

Large-language models (LLMs) show significant potential results in performance for the automated fact-checking task because they have an enormous size of parameters and are trained on large-scale datasets, thus they can perform reasoning on complex tasks [99]. Many approaches based on zero-shot Fact-checking on LLMs show impressive performance such as ProgramFC [9] decomposing the complex claims into single ones via program-guide reasoning on LLMs, FACTSCORE [58] decomposing the long-text to fine-grained atomic samples for verification, PROTRIX [113] use Plan-then-Reason strategy in which decompose the question into atomic sentences and then reasoning on table and textual context to find the correct answer, and TART [114] providing a chain-of-thought reasoning process to augment the ability of LLMs in understanding tabular data. Although these methods reach state-of-the-art accuracy, they rely on several LLMs and multi-stage pipelines, resulting in high inference latency and GPU memory usage. Such resource footprints conflict with the AFC objective of verifying a wide range of claims accurately yet cost-effectively. To address this, we present TabV4FC, a single-shot framework that leverages a lightweight table-verbalization module and just one off-the-shelf LLM, thereby reducing computational cost while preserving competitive accuracy for tabular fact-checking.

At present, Large Language Models (LLMs) show an impressive performance on various NLP tasks such as information extraction, data-to-text generation, summarization, classification, and language understanding [115]. For the Fact-checking task, the LLMs approach is a potential method for improving the efficiency and accuracy of the Fact-checking task because LLMs are trained on a very large-scale dataset and contain billions of parameters to capture the nuances and patterns in natural language [99]. In [116], the authors show that the LLMs can understand basic information from the table, such as the table format representation (i.e., JSON, XML, YAML). In addition, the work [117] shows that the representation of the table does not have a significant effect on the performance of LLMs. Since LLMs are constructed mostly based on large-scale and rich textual data, they perform better if the input data is represented in textual form. Moreover, the LLMs are sensitive to the prompt since a minor modification in the prompt can significantly impact the performance of LLMs, especially with complex heterogeneous data like graphs or tabular [118]. To leverage the ability of LLMs in reasoning and understanding the textual form, we propose TabV4FC (Tabular Verbalization for Fact Checking) - a simple but efficient framework for fact-checking tasks on tabular evidence by verbalizing the table to textual data. Then, the verbalized data is combined with the table represented in text-specific format and fed to the LLMs via prompting methods to exploit the veracity label for the Fact-checking task. Specifically, we first fine-tuned TAPEX (Table Pre-training via Learning a Neural SQL Executor) - a robust method that varies from the BART architecture for text generation to adapt for table-to-text generation. To better fine-tune the TAPEX for table-to-text generation, we use two gold annotated datasets, including QTSUM [119] and the SCIGEN [120]. Then, we use the generated text as the verbalization for the table to prompt the LLMs to verify the truthfulness of the claim. Finally, we evaluate our proposed method on various robust LLMs like LLama3 [121], Qwen2.5 [122], and DeepSeek-R1 distilled [123] based on three different tabular Fact-checking datasets, including TabFACT [19], SCITAB [20], and PubHealthTAB [21]. The empirical results show that TabV4FC helped improve the performance of LLMs for the zero-shot Fact-checking model on tabular data, and also obtained comparative results with other SOTA models.

3.3.2 Datasets

In this section, we analyze two kinds of datasets: the datasets for fine-tuning the TAPEX model for text generation from tabular data, and the datasets for evaluating the performance of LLMs for tabular fact-checking tasks. Table 3.6 shows the analysis of the QTSUMM and SCIGEN datasets for table-to-text generation. The QTSUMM is constructed on the Wikipedia article domain [119], and the SCIGEN is constructed on scientific domain text, including scientific papers in Computation and Language, Computational Geometry, Machine Learning, Networking, Distributed, Parallel, and Cluster Computing [120]. It can be seen that most of the table in QTSUMM has the size of (10,6) while the SCIGEN tables mostly have the size of (4,4). In contrast, the maximum number of rows and columns in the SCIGEN is larger than in the QTSUMM. According to [119], the QTSUMM constructed the description text according to the query from users since the description text focused on the main information targeted by the query. Nevertheless, the description texts in the SCIGEN dataset are the caption for the table [120] since they are trying to exploit as much information from the table. Therefore, SCIGEN also has a longer description for the table than QTSUMM. We combined the training set of two datasets and used the combined training set to fine-tune the TAPEX for the table-to-text generation model.

Besides, we use three datasets as shown in Table 3.7 to evaluate the performance of LLMs for fact-checking. Since our method is based on zero-shot prompting, we use only the test sets of the three datasets for evaluation through the F1 score. As shown in Table 3.7, it can be seen that the number of support and refute samples in the TabFACT and SCITAB is balanced (in the TabFact, the author used *entailed*, which is similar to support). Meanwhile, the data seems skewed to the support label in the PubHealthTab. In addition, the description generated by the table-to-text model on SCITAB is slightly longer than TabFACT and PubHealthTAB on average. As shown in Figure 3.4, the generated description on three datasets has a length mostly from 50 to 100 tokens, indicating a medium length.

	QTSumm			SCIGEN		
	Train	***Dev***	***Test***	***Train***	***Dev***	***Test***
Num. samples	4,980	1,050	1,080	13,600	3,450	492
Max columns	15	15	14	38	21	19
Max rows	50	50	38	101	97	52
Min columns	2	3	2	1	1	2
Min rows	5	5	5	0	0	1
Mode columns	6	6	5	4	4	5
Mode rows	10	10	10	4	4	5
Max description len	136	132	138	418	450	702
Min description len	1	1	1	10	15	12
Avg. description len	58.54	59.06	58.45	103.49	107.48	96.37

Table 3.6: Overview of the QTSUMM and SCIGEN datasets

Next, we investigate the relation between the size of the table and the generated description. According to Table 3.7, it can be seen that although the average number of columns and rows of the table in SCITAB and PubHealthTab is lower than that of TabFACT, the average length of the generated description is significantly longer, indicating that there is potentially more information represented in the table of SCITAB (Scientific domain) and PubHealthTab (Wikipedia and Medical domain) than TabFACT (Wikipedia domain).

	TabFACT		SCITAB		PubHealthTAB	
Domains		Wikipedia		Scientific Articles		Wikipedia and medical sources
Num. samples		12,779		1,224		194
Label	**entailed**	6,425	**support**	457	**support**	103
	refutes	6,354	**NEI**	356	**NEI**	45
	-	-	**refutes**	411	**refutes**	46
Max claim length		56		85		27
Min claim length		4		1		5
Avg. claim length		14.06		22.14		11.67
Max description length[1]		277		324		309
Min description length[1]		9		9		15
Avg. description length[1]		50.70		64.94		70.60
Max columns		20		16		11
Max rows		48		32		18
Min columns		5		2		0
Min rows		4		1		0
Average columns		6.22		6.11		3.43
Average rows		13.34		7.52		7.15

Table 3.7: Overview of the three datasets for tabular fact-checking

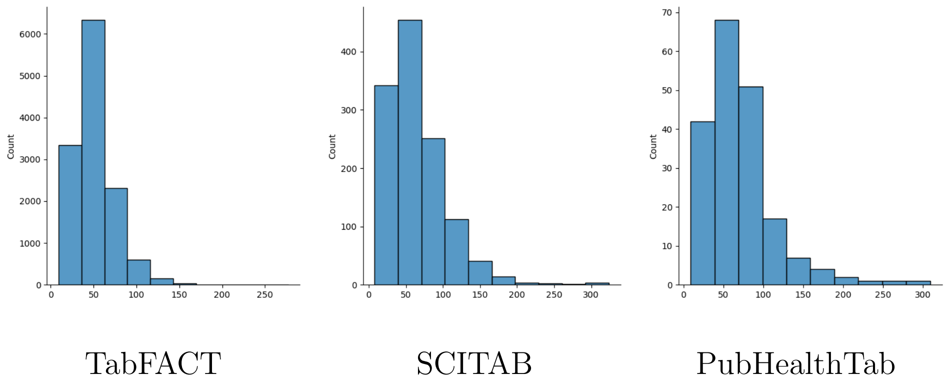


Figure 3.4: Distribution of generated description[1] on the three datasets.

Overall, the three benchmark datasets, including TabFACT, SCITAB, and PubHealthTab, are suitable for constructing and evaluating our proposed TabV4FC framework since they have a large size and diversity in domains and tabular data representation. The next section describes the performance and evaluation of the fine-tuned table-to-text model.

3.3.3 Method

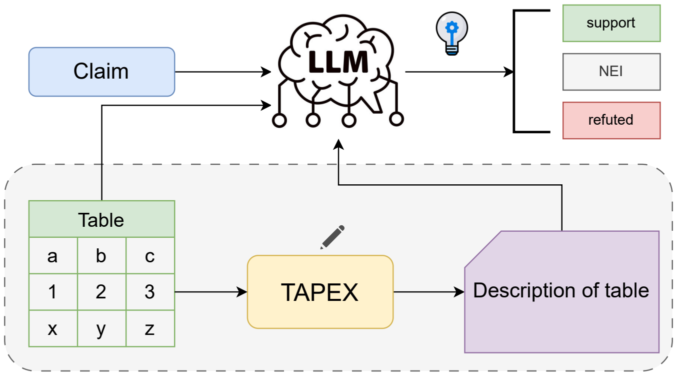


Figure 3.5: Overview about the TabV4FC framework

Given a claim c in textual form and a table t in tabular format, TabV4FC aims to determine the truthfulness of the claim c and respond to a label lb, which $lb \in \{\text{support, NEI, refuted}\}$. In practice, the table represents the data in a structured organization, including rows and columns in which each column represents a specific feature and each row represents a set of values

[1]The description is generated by the fine-tuned table-to-text model TAPEX

corresponding to the features. Although tabular data is familiar and friendly to humans since it stores the data in a structured form, tabular data is a challenge with LLMs, as LLMs are better at understanding textual data rather than tabular [124]. On the other hand, one of the limitations of LLMs in tabular understanding is the hallucination that the LLMs generate inconsistent data with real-world facts and the input queries [125]. Therefore, we use a fine-tuned table-to-text language model to generate the description for the table and use this description as a verbalization for the table to guide the LLMs for the Fact-checking task instead of directly feeding the table to the LLMs. Because the table-to-text model has been fine-tuned on the gold-annotated datasets, it can generate a high-quality description of the table as evidence that helps boost LLMs in performing the fact-checking task. Figure 3.5 illustrates our TabV4FC framework. It consists of two main phases: Generating a description for the table through TAPEX - a sequence-to-sequence (seq2seq) language model that generates text from tabular input, and verifying the claim based on zero-shot prompting for Large-language models (LLMs). The detail of TabV4FC is described below.

3.3.3.1 Table-to-text Generation

This task aims to generate a detailed description that captures the context of the given table. Let V be the set of vocabularies in natural language. Giving a table with a specific format $T \in V^{r \times c}$, where r denotes the number of lines and c denotes the number of columns of a table and a query or instruction as a sequence of words $Q \in V^k$, in which k is the length of the query or instruction, the output is a sequence $y \in V^z$, in which z is the length of the output sequence describing the table. Figure 3.6 illustrates the overview of a table-to-text model. The model consists of two main components: the first component encodes the table and the input query to feature vectors, and the second component maps the feature vectors to the vocabulary V to extract the most relevant words for constructing the output sequence.

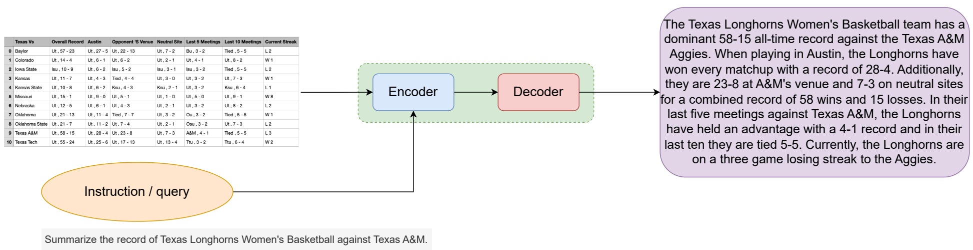


Figure 3.6: Overview about the table-to-text models

In [126], the empirical results shows that the fine-tuned model for table-to-text generation shows better performance than the zero-shot LLMs. Moreover, table-to-text generation is a supplement stage that generates the description from the table to enhance the reasoning ability of LLMs on tabular fact-checking tasks. Therefore, we fine-tuned a table-to-text model instead of using zero-shot generation from LLMs to avoid calibration. In this paper, we fine-tune the TAPEX, a seq2seq model based on the BART architecture, for generating text from the table. First, the model encodes the tables represented in data frame format as $T \in V^{r \times c}$ (r denotes the number of lines and c denotes the number of columns) and the query in the textual form $Q \in V^k$ (k is the length of the query to construct an output vector $h_E \in R^d$. Besides, we embed the ground-truth output sequence $o \in V^m$ (m is the length of the output sequence) as a label using the BART since it is the backbone of the TAPEX model [100], which transforms the ground-truth text into embedding space, denoted as $h_o \in R^d$.

$$h_E = TAPEX - Encoder(T, q) \tag{3.3}$$

$$h_o = BART(o) \sim y \tag{3.4}$$

Next, we use the TAPEX decoder to decode the encoder feature vector h_E using an autoregressive decoder to make an output vector $h_D \in R^d$ as the prediction. To fine-tune TAPEX, we compute the loss using the cross-entropy as the reconstruction loss to measure the difference between the decoder output and the target text.

$$h_D = TAPEX - Decoder(h_E) \tag{3.5}$$

$$loss = Cross - entropy(TAPEX - Decoder(y, h_E), y) \tag{3.6}$$

Since we use the base version of TAPEX and BART, the d, which indicates the dimension of space R, is equal to 768. We then fine-tune the TAPEX by optimizing the loss function in Eq. 3.6 by the ADOPT optimizer [127]. We run the fine-tuning process with epochs in the range of $\{50, 100, 200\}$, a batch size equal to 8, and a learning rate equal to 10^{-5}. We use two annotated datasets for fine-tuning the table-to-text task, including the QTSum [119] and SCIGEN [120].

3.3.3.2 Zero-shot Fact-checking on Tabular Data

Giving a claim $Q \in V^k$ as the query (k is the length of the query), the table $T \in V^{r \times c}$ (r determines the number of rows, and c denotes the number of columns) is represented in textual format and the verbalization $S \in V^m$ (m is the length of the verbalization sequence) that describes the summarization from the table. The S sequence is the output sequence extracted in the *Table-to-text Generation* stage. Figure 3.7 illustrates the verification stage using LLMs via the prompt instruction. We combine the three components (Q, T, S) to a prompt $P \in V^{k+r\times c+m}$ such that it provides an instruction for verifying the truthfulness of the claim Q. According to [98], the reasoning process of an LLM is a process that maximizes the likelihood of the answer according to a prompt and a question as the query. We denote A as the answer for the veracity of the claim Q such that $A \in \{\texttt{support}, \texttt{not enough information (NEI)}, \texttt{refute}\}$. To be clear, the answer that denotes the truthfulness of the claim is the answer generated by a parameterized probabilistic model p_{LM} through a maximized likelihood process as:

$$P(A|P(Q,T,S)) = \prod_{i=1}^{|A|} p_{LM}(a_i|P(Q,T,S), a_{<i}) \tag{3.7}$$

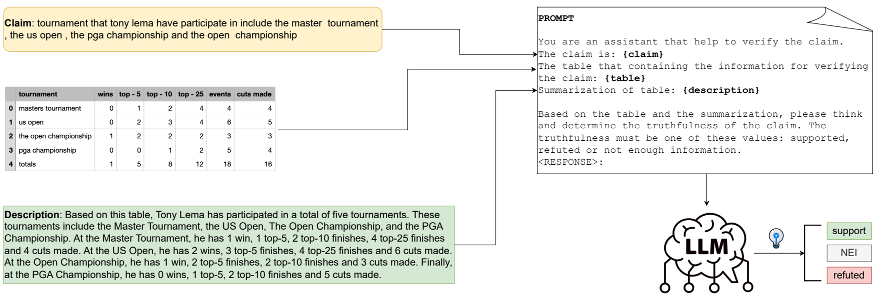


Figure 3.7: Overview of zero-shot prompting for LLMs

In Eq. 3.7, the a_i denotes the i-th token and $|A|$ denotes the length of the final answer. The prompts for generating the instruction that is used to instruct the LLMs is defined as:

```
You are an assistant that help to verify the claim.
The claim is: {claim}
The table that containing the information for verifying the claim:
    {table}
Summarization of table: {description}

Based on the table and the summarization, please think and
   determine the truthfulness of the claim. The truthfulness must
```

```
    be one of these values: supported, refuted or not enough
    information.
<RESPONSE>:
```

We run the zero-shot learning for verification tasks with robust LLMs such as LLama3.3 [121], Qwen2.5 [122], and DeepSeek-R1 Distill [123]. We use three different tabular fact-checking datasets, including TABFACT [19], SCITAB [20], and PubHealthTAB [21]. Also, we provide the example prompt on each benchmark dataset, TABFACT, SCITAB, and PubHealthTAB in Appendix B.

3.3.4 Results

3.3.4.1 Empirical Configuration

Table 3.8 shows the information about the used models in our experiment and hyperparameter settings. We run the LLMs with 4-bit quantization on LLMs can be hosted and run locally on a single GPU with about 49Gb of memory, which is suitable for deployment on a practical system. To perform the running of LLMs on a 4-bit setting, we use the BitsansBytes library[2].

Model	**Pre-train**	**#parameters**	**Hyper-parameter**
	Table-to-text generation		
TAPEX	TAPEX-based (encoder)	139M	max_length=400, do_sample=True, temperature=0.2, num_beams=4
	LLMs for verification		
LLama 3.3	Llama-3.3-70B-Instruct	70B	max_new_tokens = 10, do_sample=False
Qwen 2.5	Qwen2.5-72B-Instruct	72B	
	Qwen2.5-32B-Instruct	32B	
	Qwen2.5-14B-Instruct	14B	
DeepSeek-R1	DeepSeek-R1-Distill-Qwen-32B	72B	
	DeepSeek-R1-Distill-Llama-70B	70B	

Table 3.8: LLMs information and configurations in experimental results. "M"" denotes for "Million", and "B" denotes for "Billion" in model parameter size.

3.3.4.2 Results of Table-to-text generation task

This section illustrates our empirical results for the table-to-text generation task. We use four metrics for evaluating the quality of the generated text

[2] https://github.com/bitsandbytes-foundation/bitsandbytes

in comparison with the ground-truth one, including BLEU [128], METEOR [129], ROUGE [130], and BERTScore [131]. The BLEU [128] and METEOR [129] measure the alignment level of n-gram matching between the generated output and the ground truth, while ROUGE [130] and BERTScore measure the semantic similarity through n-gram and BERT contextual embedding, respectively, between the generated and the ground truth. According to the experiments from [132], BLEU and METEOR are more strict than BERTScore or ROUGE since BLEU and METEOR capture the exact-match error between the generated and the ground truth sequences. Table 3.9 shows the results of the fine-tuned TAPEX model on the development sets of QTSUMM and SCIGEN.

Epoch	50		100		200	
Datasets	QTSUMM	SCIGEN	QTSUMM	SCIGEN	QTSUMM	SCIGEN
BLEU	15.75	1.80	16.18	2.48	16.09	2.52
ROUGE	39.86	15.76	40.32	17.02	40.12	17.05
METEOR	42.37	15.72	43.56	18.79	43.56	18.30
BertScore	90.17	82.64	90.39	84.20	90.45	84.61

Table 3.9: Empirical results of the TAPEX model on development sets.

		BLEU	ROUGE	METEOR	BERTScore
TAPEX-finetuned	QTSUM	16.65	40.93	44.46	90.57
	SCIGEN	2.59	17.41	19.09	84.56
Baseline	QTSUM	23.10	42.10	45.60	90.60
	SCIGEN	5.30	-	23.00	84.00

Table 3.10: Results of finetuned TAPEX model on the test sets.

According to Table 3.9, the performance of TAPEX on both QTSUMM and SCIGEN is low on the first 50 epochs. When the epoch increases to 100, the performance of the TAPEX model improves slightly. If the epoch reached 200, the performance of the model would not seem much different. For the TAPEX model with 200 epochs trained, the model seems to slightly decrease on the QTSUMM dataset, while the BLEU score and BERTScore increased on the SCIGEN dataset. Overall, the TAPEX model with 200 epochs obtained good results on the development set since we evaluated it on the test set to investigate the performance. According to Table 3.10, the fine-tuned TAPEX model obtains competitive results with the baseline finetuned on each dataset. This indicates that the fine-tuned TAPEX model can efficiently learn the features from the table and generate corresponding description texts, even when the training data has been mixed between two different domains. Finally, we use the fine-tuned TAPEX model to generate description text for the table in TabFACT, SCITAB, and PubHealthTAB. The next section describes the empirical results of different LLMs for the Tabular Fact-checking task evaluated on three datasets.

3.3.4.3 Results of Fact-checking on Tabular data task

Table 3.11 illustrates our empirical results of the tabular Fact-checking task on three different datasets, including SCITAB, TabFact, and PubHealthTab. To better investigate the performance of LLMs for the task, we also conducted an ablation study in which we removed the description or table from the prompt (w/o stands for "without" in Table 3.11). The underline indicates the best settings in individual LLMs.

According to Table 3.11, it can be seen that the Qwen2.5 models (both 72B and 14B) perform well on the tabular fact-checking task. In addition, the results of LLama3.3-70B, Qwen2.5-14B, and Qwen2.5-72B seem stable between the F1-macro and F1-micro, while the models distilled from DeepSeek, including LLama-70B and Qwen-32B, have a significant gap between the F1-micro and F1-macro. In addition, it can be seen that the summary plays an important role in boosting the ability of the model than the table. As shown from Table Table 3.11, the LLama3.3-70B, Qwen2.5-14B, and Qwen2.5-72B have significantly better performance when running without a table than without a description on SCITAB and TabFact. For the PubHealTab, the results are similar on LLama3.3-70B and Qwen2.5-72B, where the information from the description better enhances the performance of checking the truthfulness of the claim than the table. For Qwen2.5-14B and Qwen2.5-32B, the gap between the model without description and the table is not much. Specifically, the information from the table in the PubHealthTab impacts the performance of DeepSeek, where the performance of both models distilled from DeepSeek dramatically decreased in the case of a lack of table information. Overall, the Qwen 2.5-72B has the best performance on three datasets when enhanced by the TabV4FC, which obtained by F1-macro of 45.20%, 57.35%, and 58.60% on SCITAB, TabFact, and PubHealthTab, respectively. In addition, the description that verbalizes the table helps improve the ability of the table to understand LLMs to help verify the veracity of the claim.

Besides, we compare our method with baselines and state-of-the-art models (SOTAs) on Tabular Fact-checking, including TART [114] and ProTrix [113]. The baselines consist of TAPEX-Zero-XL on SCITAB [20], Table-BERT on TabFact [19], and TAPAS on PubHealthTab [21]. On the SCITAB and PubHealthTab, our method TabV4FC with Qwen2.5-72B model outper-

	SCITAB		TabFact		PubHealthTab	
	F1-micro	F1-macro	F1-micro	F1-macro	F1-micro	F1-macro
TabV4FC (Qwen 2.5 - 72B)	**47.22**	**45.20**	**62.32**	**57.35**	**57.21**	**58.60**
w/o description	32.76	23.29	49.83	33.52	40.72	39.26
w/o table	45.26	40.60	55.97	54.10	44.32	41.90
TabV4FC (LLama 3.3 - 70B)	37.90	34.49	59.42	54.75	43.29	43.00
w/o description	32.76	24.40	51.29	38.95	23.71	19.05
w/o table	41.58	35.33	54.63	54.63	40.20	34.65
TabV4FC (Qwen 2.5 - 32B)	28.26	17.39	49.92	34.79	30.92	29.13
w/o description	29.98	19.30	53.65	45.35	23.71	13.24
w/o table	28.83	16.43	49.64	33.93	23.71	13.24
TabV4FC (Qwen 2.5 - 14B)	43.79	37.43	53.47	41.58	51.54	47.51
w/o description	30.31	17.96	49.83	33.62	39.69	33.22
w/o table	41.83	37.08	53.04	46.23	38.65	33.96
TabV4FC (DeepSeekR1-LLama 70B)	33.25	26.98	55.12	51.54	32.47	30.73
w/o description	32.67	23.53	49.86	33.69	57.21	49.94
w/o table	42.81	30.21	54.08	52.79	52.06	43.07
TabV4FC (DeepSeekR1-Qwen 32B)	39.13	32.23	49.87	33.55	46.90	37.75
w/o description	40.03	32.23	49.72	33.20	59.79	49.83
w/o table	41.25	35.66	49.87	34.02	37.62	31.63
Baselines [19–21]	-	34.30	65.10	-	-	48.00
TART [114] (Llama3-8b)	47.20	-	69.70	-	68.50	-
ProTrix [113] (Llama-3-8B)	41.30	-	79.40	-	-	-

Table 3.11: Empirical results of Tabular Fact-checking tasks on different LLMs by TabV4FC.

forms the baseline. On the TabFact dataset, the performance of our method is competitive with the baseline fine-tuned on the training dataset, despite we use of the zero-shot settings. For the SOTAs, both TART and ProTrix achieved outstanding results on SCITAB, TabFact (TART and ProTrix), and PubHealthTab (TART). As shown in Table 3.11, our method obtains competitive results on SCITAB in comparison with TART, and the F1 score is higher than Protrix (47.22% for our and 41.30% for Protrix). On TabFact, our methods are slightly lower than TART, while it is significantly lower than the ProTrix model. On PubHealthTab, TART also obtains higher results than TabV4FC (57.21% of our method and 68.50% of TART). However, TART and ProTrix require many computational steps based on LLMs, such as formatting and representing the table in a compact form, then generating the answer via a reasoning process (introduced in TART) or decomposing the question into atomic facts and reasoning based on the table context to form the final answer (introduced in ProTrix). Our method, TabV4FC, simply employs a fine-tuned table-to-text model to convert the table into natural language text to leverage the ability of language understanding in verifying the claim of LLMs without training or fine-tuning them.

3.3.4.4 Inference time analysis

In this section, we analyze the inference time of TabV4FC. Since state-of-the-art (SOTA) methods such as TART [114] and ProTrix [113] do not report LLM inference times, a direct comparison is not possible. To evaluate TabV4FC, we randomly sample 100 instances from each of the three benchmark tabular fact-checking datasets—TabFACT, SCITAB, and PubHealthTAB—and report the total inference time in Table 3.12. This time includes both the table-to-text generation and verification stages. As shown in Table 3.12, the table-to-text generation stage is significantly faster on TabFACT than on SCITAB and PubHealthTAB, despite TabFACT tables being larger on average (Table 3.7). This is likely because SCITAB and PubHealthTAB tables contain denser or more complex information, which increases generation time. In contrast, the verification stage—performed by LLMs—scales directly with table size: TabFACT incurs the highest verification time, followed by SCITAB and then PubHealthTAB, consistent with their average table sizes reported in Table 3.7.

	Total Time for Generation (s)	**Total Time for Verification (s)**	**Total Time / Question (s)**
TabFact	75.01	347.84	4.22
SCITAB	113.42	311.03	4.24
PubHealthTab	111.09	285.59	3.96

Table 3.12: Inference time of TabV4FC with fine-tuned TAPEX for table-to-text generation and *Qwen2.5-72B-Instruct* for verification.

.

Similar to the previous experiment, we examine the inference time of various LLMs across their model sizes using 100 samples from each of the three benchmarks. The results are summarized in Table 3.13. As expected, inference time scales with model size: larger LLMs require more time to complete the task. Combining these findings with the accuracy results in Table 3.11, we observe a clear trade-off between inference time and task accuracy: larger LLMs achieve higher accuracy at the cost of increased computational latency.

3.3.5 Error Analysis

3.3.5.1 Table-to-text errors

We investigate several prediction samples generated by the TAPEX for the table-to-text task as shown in Figure 3.8, 3.9, 3.10, and 3.11 (Figure 3.8 and

Datasets	LLMs	Total Time For Verification (s)	Time / Question (s)
TabFACT	Qwen2.5-72B-Instruct	347.84	3.47
	Llama3.3-70B-Instruct	325.30	3.25
	Qwen2.5-32B-Instruct	193.12	1.93
	Qwen2.5-14B-Instruct	111.70	1.11
SCITAB	Qwen2.5-72B-Instruct	311.03	3.11
	Llama3.3-70B-Instruct	294.62	2.94
	Qwen2.5-32B-Instruct	175.55	1.75
	Qwen2.5-14B-Instruct	104.18	1.04
PubHealthTab	Qwen2.5-72B-Instruct	285.59	2.85
	Llama3.3-70B-Instruct	273.07	2.73
	Qwen2.5-32B-Instruct	161.45	1.61
	Qwen2.5-14B-Instruct	99.71	0.99

Table 3.13: Inference time of different LLMs on three benchmark datasets.

3.9 are extracted from the QTSUMM, and Figure 3.10 and 3.11 are extracted from the SCIGEN).

According to Figure 3.8 and 3.9, the table-to-text models generate the correct information in the early part of the paragraph. For example, in Figure 3.8, the model captured the topic information about the object "Texas Longhorns Women's Basketball" as it is mentioned in the summary query. The model can extract the correct information about the "overall record of 58-15" win against the Texas A&M and "28-4 against Texas A&M" (line 9 of the table in Figure 3.8). Also, the model retrieves correct information about "won 23 out of the 28 games" (columns 4, line 9 of the table in Figure 3.8). However, the model gives the wrong description for "3-3 in the neutral site games" since the correct fact is "7-3" (column 5, line 9 of the table in Figure 3.8). The information about the results of "Texas Longhorns Women's Basketball" in last 5 and 10 Meetings is also not extracted by the model. Similar to the phenomenon in Figure 3.8, the information extracted in the sample in Figure 3.9 is also partly correct in the early sentence where it is shown that "Brazil won two gold medals and one bronze medal in tennis at the 1999 Pan American Games" (line 0 of the table in Figure 3.9). However, the information about "United States won two gold medals" and "Argentina took three gold medals and one bronze medal" is not correct since the United States won only 1 gold medal (line 1 of the table in Figure 3.9) and Argentina won 3 bronze medal (line 5 of the table in Figure 3.9). Finally, the model misses the information about 2 nations that won bronze medals, including Venezuela and Canada.

	Texas Vs	Overall Record	Austin	Opponent 'S Venue	Neutral Site	Last 5 Meetings	Last 10 Meetings	Current Streak
0	Baylor	Ut , 57 - 23	Ut , 27 - 5	Ut , 22 - 13	Ut , 7 - 2	Bu , 3 - 2	Tied , 5 - 5	L 2
1	Colorado	Ut , 14 - 4	Ut , 6 - 1	Ut , 6 - 2	Ut , 2 - 1	Ut , 4 - 1	Ut , 8 - 2	W 1
2	Iowa State	Isu , 10 - 9	Ut , 6 - 2	Isu , 5 - 2	Isu , 3 - 1	Isu , 3 - 2	Tied , 5 - 5	L 2
3	Kansas	Ut , 11 - 7	Ut , 4 - 3	Tied , 4 - 4	Ut , 3 - 0	Ut , 3 - 2	Ut , 7 - 3	W 1
4	Kansas State	Ut , 10 - 8	Ut , 6 - 2	Ksu , 4 - 3	Ksu , 2 - 1	Ut , 3 - 2	Ksu , 6 - 4	L 1
5	Missouri	Ut , 15 - 1	Ut , 9 - 0	Ut , 5 - 1	Ut , 1 - 0	Ut , 5 - 0	Ut , 9 - 1	W 8
6	Nebraska	Ut , 12 - 5	Ut , 6 - 1	Ut , 4 - 3	Ut , 2 - 1	Ut , 3 - 2	Ut , 8 - 2	L 2
7	Oklahoma	Ut , 21 - 13	Ut , 11 - 4	Tied , 7 - 7	Ut , 3 - 2	Ou , 3 - 2	Tied , 5 - 5	W 1
8	Oklahoma State	Ut , 21 - 7	Ut , 11 - 2	Ut , 7 - 4	Ut , 2 - 1	Osu , 3 - 2	Ut , 7 - 3	L 2
9	Texas A&M	Ut , 58 - 15	Ut , 28 - 4	Ut , 23 - 8	Ut , 7 - 3	A&M , 4 - 1	Tied , 5 - 5	L 3
10	Texas Tech	Ut , 55 - 24	Ut , 25 - 6	Ut , 17 - 13	Ut , 13 - 4	Ttu , 3 - 2	Ttu , 6 - 4	W 2

Query: Summarize the record of Texas Longhorns Women's Basketball against Texas A&M.

Prediction:
The Texas Longhorns Women's Basketball Team has an **overall record of 58-15** and an overall record of **28-4 against Texas A&M**. They have **won 23 out of the 28 games** played against A&M. Texas is also 3-3 in the neutral site games. Lastly, the series is tied at 5-5.

Ground truth:
The Texas Longhorns Women's Basketball team has a dominant **58-15 all-time record against the Texas A&M Aggies**. When playing in Austin, **the Longhorns have won every matchup with a record of 28-4**. Additionally, they are **23-8 at A&M's venue** and 7-3 on neutral sites for a combined record of 58 wins and 15 losses. In their last five meetings against Texas A&M, the Longhorns have held an advantage with a 4-1 record and in their last ten they are tied 5-5. Currently, the Longhorns are on a three game losing streak to the Aggies.

Figure 3.8: Error prediction sample 1 - Wrongly extracted numerical facts and missing information. The model wrongly extracts "3-3 in the neutral site games" and misses "Texas Longhorns Women's Basketball" in the last 5 and 10 Meetings.

	Nation	Gold	Silver	Bronze	Total
0	Brazil	2	0	1	3
1	United States	1	2	2	5
2	Venezuela	1	0	1	2
3	Mexico	0	1	0	1
4	Chile	0	1	0	1
5	Argentina	0	0	3	3
6	Canada	0	0	1	1

Prediction:
Brazil **won two gold medals** and **one bronze medal in tennis** at the 1999 Pan American Games. The United States won **two gold medals** and **two bronze medals in tennis** as well. Argentina took **three gold medals** and **one bronze medal** from the tournament.

Ground truth:
At the **1999 Pan American Games**, **five nations won bronze medals** in Tennis, with varying amounts. Brazil, Venezuela, Canada each won **one bronze medal**, and the United States **won two**, while Argentina **won three**.

Query: How many nations won Bronze medal in Tennis at the 1999 Pan American Games? Summarize the medal tally of these nations.

Figure 3.9: Error prediction sample 2 - Wrongly extracted entities and corresponding numerical facts. In the example, "United States won two gold medals" and "Argentina took three gold medals and one bronze medal" are both incorrect.

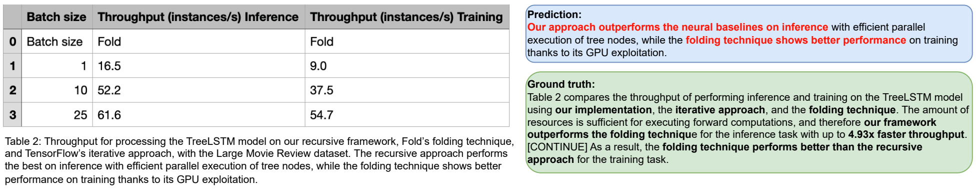
	[BOLD] Relation	[BOLD] best F1 (in 5-fold) without sdp	[BOLD] best F1 (in 5-fold) with sdp	[BOLD] Diff.
0	USAGE	60.34	80.24	+ 19.90
1	MODEL-FEATURE	48.89	70.00	+ 21.11
2	PART_WHOLE	29.51	70.27	+40.76
3	TOPIC	45.80	91.26	+45.46
4	RESULT	54.35	81.58	+27.23
5	COMPARE	20.00	61.82	+ 41.82
6	macro-averaged	50.10	76.10	+26.00

Table 1: Effect of using the shortest dependency path on each relation type.

Prediction:
The result table shows that **SSDP significantly improves the performance** of all relations. Compared to the baseline model, the improvement of SSDP is more significant on *"part_whole"*, *"topic"*, *"combined"* and *"regime change"* relations. In particular, SSDP improves the accuracy on *"combined"*, *"part_whole"*, *"topic"* and *"regime change"* relations by **45.46**, **27.23** and **41.82** respectively.

Ground truth:
We find that the **effect of syntactic structure varies** between the **different relation types**. However, the **sdp information has a clear positive impact** on all the relation types (Table 1).

Figure 3.10: Error prediction sample 3 - Incorrect table understanding and wrongly extracted facts. "combined" and "regime change" values do not appear in the "Relation" columns. Consequently, the facts about "...45.46, 27.23, and 41.82, respectively, according to these two values, are not correct.



	Batch size	Throughput (instances/s) Inference	Throughput (instances/s) Training
0	Batch size	Fold	Fold
1	1	16.5	9.0
2	10	52.2	37.5
3	25	61.6	54.7

Table 2: Throughput for processing the TreeLSTM model on our recursive framework, Fold's folding technique, and TensorFlow's iterative approach, with the Large Movie Review dataset. The recursive approach performs the best on inference with efficient parallel execution of tree nodes, while the folding technique shows better performance on training thanks to its GPU exploitation.

Figure 3.11: Error prediction sample 4 - Lack of qualitative comparison information extraction. The model showed that it "outperforms the neural baseline" and "folding technique shows better performance" correctly, but does not provide a sufficient comparison, like as "4.93x faster throughput" between the proposed framework and the folding technique.

In addition, in the sample in Figure 3.10, the first sentence correctly summarizes the fact from the table that "SSDP significantly improves the performance". However, the following sentences contain error information about the relations such as "combined" and "regime change" (these two "new" relation does not appear in the table shown in Figure 3.10). Also, the accuracy extracted as shown in the last sentence is not correct, corresponding to the relations in the table. Furthermore, in the sample in Figure 3.11, although the model can extract the information about the improvement of the approach as "Our approach outperforms the neural baselines on inference" and " folding technique shows better performance", the information seems too shallow since it does not contain the qualitative comparison as "4.93x faster throughput" between the proposed framework and the folding techniques (mentioned in the ground truth in Figure 3.11). To have an in-depth analysis, we categorized the errors in table-to-text generation into four types as follows:

- **Missing information**: The generated description lacks essential details in the table, including entities, facts, the relationship between facts and entities, and qualitative and quantitative comparisons.
- **Numerical extraction errors**: The generated description extracts incorrect numerical facts, such as figures or proportions, or incorrect figures, according to entities or facts from the table.
- **Entity extraction errors**: The generated description extracts the wrong entities or objects from the table.
- **Redundant facts**: The generated description extracts facts, figures, or objects that do not come from the tables.

Based on these types of error, we annotate the error types for each generated sample from TAPEX, and employ the LLMs-as-a-judge [133] paradigm for labeling the error types. By leveraging the strong ability of LLMs in language understanding and reasoning, we annotate the error types for each predicted sample, which is more cost-efficient and less time-consuming than human annotation. We select 50 prediction samples each from the QTSUMM and SCIGEN datasets and submit them to three commercial AI platforms—ChatGPT[3], QwenChat[4], and Grok[5]. Each system reads the model predictions and compares them against the ground truth to classify any discrepancies into one of the four error types defined earlier. The final error annotation for each sample is determined by majority vote among the three AI systems to ensure objectivity. Table 3.14 presents the resulting error-type statistics for TAPEX on both QTSUMM and SCIGEN.

Type of Error	**QSUMM**	**SCIGEN**
Missing information	41	43
Numerical extraction errors	29	15
Entity extraction errors	34	9
Redundant facts	17	11

Table 3.14: Statistical information about the error types of TAPEX on two datasets.

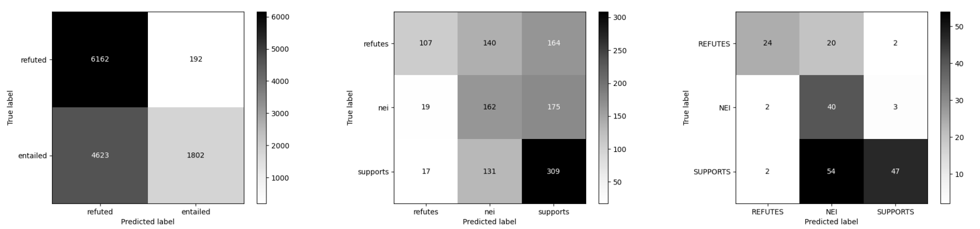
According to Table 3.14, the table-to-text generation stage in TabV4FC primarily suffers from omitting essential information present in the original table. On QTSUMM, errors in numerical extraction are less frequent than those in entity extraction. However, the trend reverses in SCIGEN, where numerical extraction errors significantly outnumber entity extraction errors—highlighting the particular difficulty of accurately understanding and extracting numerical content in scientific contexts (recall that SCIGEN is constructed from scientific data). Additionally, the generation process tends to produce redundant facts, which can inadvertently lead to hallucinations, further degrading output reliability.

Overall, it can be seen that the table-to-text model - TAPEX can extract partly correct information from the table. From the error analysis, the model lacks in extracting numerical information and performing reasoning across the values from the table to exploit latent information, such as qualitative computation and measurement.

[3]https://chatgpt.com/
[4]https://chat.qwen.ai/
[5]https://grok.com/

3.3.5.2 Tabular Fact-checking errors

Next, as shown in Figure 3.12, most of the mispredictions remain between entailed and refuted on TabFact, NEI, and support in SCIFACT and PubHealthTab. Specifically, on the SCITAB, the model is difficult to discriminate in refuted labels. According to Table 3.15, we found that the challenge of the model focuses on the NEI and refuted labels on SCITAB (which are 38.63% and 41.06%, respectively, by F1), refuted labels on TabFact and PubHealthTab (which are 42.81% and 50.31%, respectively). This shows that there is still room for future development in the extraction of latent information from the table to help improve the performance of the tabular fact-checking task using LLMs. Additionally, the empirical results in Table 3.15 show that the information from the description of the table is valuable in improving the ability of LLMs for zero-shot tabular fact-checking. As a result, LLMs are better at understanding and processing natural language text than the table. When lack of description in the table, the performance decreased dramatically on three datasets.



TabFACT SCITAB PubHealthTab

Figure 3.12: Confusion matrix of TabV4FC on three tabular fact-checking datasets.

3.3.6 Discussion

TabV4FC is a simple but efficient framework for boosting zero-shot tabular fact-checking by verbalizing the table into natural language text as a description. The TabV4FC framework integrates the TAPEX fine-tuned for table-to-text generation with LLMs based on zero-shot prompting to solve tabular fact-checking tasks. We evaluate our framework on various robust LLMs, and the empirical shows the efficiency of TabV4FC in enhancing the ability of LLMs via representing the table to natural language text. Nevertheless, the current table-to-text module in TabV4FC cannot extract

	SCITAB			TabFact			PubHealthTab		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
Full									
support (entailed)	47.69	67.61	55.93	57.13	96.98	71.91	90.38	45.63	60.65
NEI	74.83	26.03	38.63	-	-	-	85.71	52.17	64.86
refuted	37.41	45.51	41.06	90.37	28.05	42.81	35.09	88.89	50.31
w/o description									
support (entailed)	65.79	10.94	18.76	49.78	99.94	66.46	92.86	25.24	39.69
NEI	100	2.92	5.67	-	-	-	90.91	21.74	35.09
refuted	29.84	95.22	45.44	82.61	0.30	0.59	27.74	95.56	43.00
w/o table									
support (entailed)	42.27	80.74	55.49	54.04	76.60	63.37	70.00	40.78	51.53
NEI	72.73	23.36	35.36	-	-	-	47.83	23.91	31.88
refuted	40.64	25.00	30.96	60.59	35.58	44.83	29.73	73.33	42.31

Table 3.15: Performance of TaV4FC with Qwen2.5-72B-Instruct on the three datasets.

the latent information from the table, especially numerical values, qualitative computation, and comparison among columns and rows in the table.

Future works focus on improving the ability of table understanding for table-to-text generation to construct a sufficient and correct description for a table by deeply extracting the information from the table features, such as data format, column alignment, and types of values in the table [114]. Besides, we apply advanced prompting techniques to boost the performance of LLMs for tabular fact-checking, like Self-prompting [134] and Plan-then-Reason [113]. Last but not least, TabV4FC has the potential to integrate into an automatic Fact-checking system to verify the truthfulness of the claim.

3.4 Summary

This chapter demonstrates the use of LLMs for Fact-checking tasks by leveraging their robust ability to reason via prompting. Specifically, this chapter introduces two frameworks—ZeFAV and TabV4FC —that efficiently align information across data modalities to verify the claim. Although demonstrating impressive performance, ZeFAV and TabV4FC consulted with gold evidence only, which is limited in practical application when there is no available human-annotated evidence for the claim. Following the paradigm of leveraging LLMs for Fact-checking, the next chapter introduces an automatic framework that can automatically retrieve relevant evidence and verify the truthfulness of the claim in a multimodal setting.

Declaration: Parts of this chapter have been published in Publications [1,3].

Chapter 4

An approach for Multimodal Fact-verification

4.1 Introduction

Multimodal fact-checking is increasingly vital due to the prevalence of misleading content combining text and images [8]. However, it faces two main obstacles. First, training fact-checking models requires large, human-annotated datasets, which are costly, time-consuming, and prone to bias [9]. Second, models must reason across modalities, integrating text and visual information to identify features that support or refute a claim [26]. For example, consider the claim: *"920 women lose their unborn babies after getting vaccinated"*. As shown in Figure 4.1, evidence includes text stating, *"There is no confirmation about miscarriages,"* and *"Public health experts suggest COVID-19 vaccines are safe for pregnant women,"* alongside an image of a pregnant woman wearing a mask during the COVID-19 pandemic.[1]. The evidence suggests the claim is false, but the image alone cannot confirm vaccine effects, and manipulated visuals can introduce bias [8]. Thus, effective multimodal fact-checking requires combining text and image evidence clearly and reliably.

Multimodal fact-checking addresses claims involving both text and images, requiring methods to integrate and align these data types. Recent approaches have achieved strong results. For example, Lee et al. [54] use external LLMs (e.g., GPT-4, GPT-3.5) to explain the context between images and text, enhancing verification. Geng et al. [96] employ in-context learning with samples generated by GPT-4V to guide a system (LLaVA) in verifying new claims. Cekinel et al. [135] use a Vision Language Model (VLM) to combine image and text information into feature vectors, which are then processed by a feed-forward network to predict a claim's truthfulness. The key to effective multimodal fact-checking lies in aligning image and text

[1]The image originates from a 2021 Washington Post article about pregnant women during the pandemic: https://www.washingtonpost.com/health/2021/08/11/pregnant-covid-vaccine/

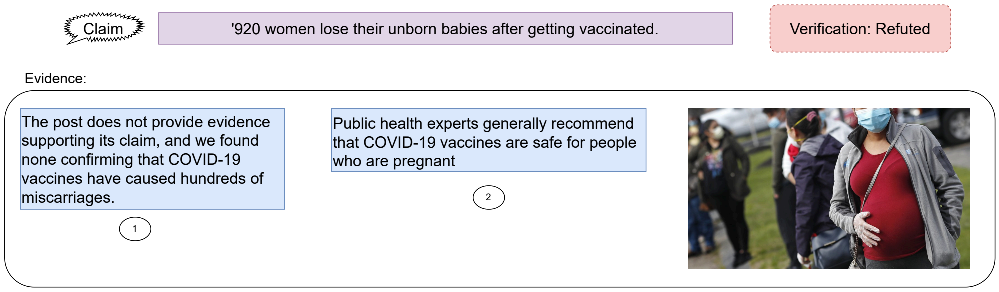


Figure 4.1: Example of a claim with text and image evidence. The example is extracted from the Mocheg dataset.

evidence to uncover consistent information. However, methods like [135] require significant computational resources, making them less efficient for zero-shot settings. Additionally, the explanations in [54,96] do not emphasize consistency between images and text, a critical factor for accurate verification. These methods also rely on commercial LLMs (e.g., GPT-4, GPT-4V), which depend on costly API services and are less practical for independent systems [101]. Our work addresses these limitations by proposing an efficient zero-shot fact-checking method that leverages open-source LLMs. By avoiding reliance on commercial APIs, our approach is designed for practical, standalone systems. We focus on generating consistent alignments between text and image evidence to enhance verification, improving performance in multimodal settings without excessive computational demands.

To address these challenges, large language models (LLMs) like GPT-4 offer a promising solution. LLMs store vast knowledge and excel at reasoning, enabling claim verification with minimal or no training data [96, 136]. Their reasoning capabilities support logical deductions based on claims and evidence, improving verification accuracy [137]. Prompting techniques, such as zero-shot or few-shot prompting, leverage LLMs' knowledge in data-scarce scenarios [99]. However, ambiguous prompts can reduce performance, necessitating clear, precise instructions. Additionally, establishing explicit connections between claims and multimodal evidence is crucial, as discrete evidence complicates verification. Briefly, we outline two main challenges for LLMs in multimodal fact verification as follows.

- **Challenge 1: Cross-modality aligning between text and image**: The heterogeneity between text and image reduces the accuracy of the model because the information from each modality, such as text or image, can overlap and be redundant, which leads to inconsistency and complicates features (Figure 3.2 illustrates an example of the confusing between text and image evidence in verifying a claim). Addressing the heterogeneity helps improve the performance of the model through multimodal integration and reasoning [138]. To tackle this problem for the multimodal Fact-checking task, we form the first research question as **RQ1**: Does aligning text and image evidence enhance LLM reasoning for claim verification?

- **Challenge 2: Fact-verification with system-retrieved evidence**: In the end-to-end Fact-checking system, the retrieval stage provides curated evidence for verifying the claim [8]. The evidence retrieval plays a vital role in providing sufficient knowledge to verify the truthfulness of the claim, since most fact-checking methods use external retrieved evidence to verify the claim [45]. In addition, most of the current multimodal Fact-checking datasets like Fin-Fact [139] and FACTFY [24] consult the evaluation results on gold evidence only. Hence, we address this challenge to investigate the performance of the fact-checking model with the system-retrieved evidence scenario. We form this problem as the second research question: **RQ2** How effective is the fact-checking system with automatically retrieved evidence in practical, unsupervised fact-checking?

One potential approach for solving these challenges in the Retrieve-Augmented-Generation (RAG) strategy, which is effective for Fact-checking, a knowledge-intensive task [140]. By leveraging reliable information from validated external sources, RAG helps improve the accuracy of LLMs and reduce factual inconsistencies. However, for the multimodal task, the challenges are indicated in the evidence where there is a potential risk of cross-modality inconsistency that affects the process of verifying the truthfulness of the claim (as shown in Figure 4.1). Our motivation is leverage the RAG paradigm with efficient cross-modal instruction for supporting the verification process by LLMs. To sum up, we address these challenges by proposing **M-RAV**, a multimodal fact-checking framework that harnesses LLMs to verify claims using text and image evidence. M-RAV operates in three stages: **Retrieve** relevant evidence using the claim as a query, **Augment** evidence by aligning text and image content for consistency, and **Verify** the claim's truthfulness by assessing evidence relevance. By employing zero-shot or few-shot prompting, M-RAV minimizes reliance on annotated data [99]. It supports both manually curated (gold) and automatically retrieved (system) evidence, making it adaptable to real-world scenarios.

4.2 The M-RAV Frameworks

The M-RAV framework is a zero-shot learning approach designed to verify the truthfulness of claims using large language models (LLMs) and multimodal evidence (text and images). Inspired by fact-checking pipelines [4], M-RAV aims to create a practical, portable system for real-world applications. Given a claim, the framework retrieves relevant evidence, aligns text and image evidence, evaluates their relevance, and uses LLMs to determine the claim's veracity. The process consists of four stages, as illustrated in Figure 4.2. Each stage is described as follows:

Stage 1: Evidence Retrieval Given a claim c, the system retrieves relevant evidence from a database $\mathcal{DB}$, built from fact-checking platforms like Snopes[2] and PolitiFact[3]. We use CLIP [37] to fetch the top-k images $I = \{img_1, \ldots, img_n\}$ and FlagEmbedding BGE [141] to retrieve the top-k texts $T = \{text_1, \ldots, text_m\}$, where n and m depend on the chosen k. The optimal value of k is found from experiments. In our implementation, $\mathcal{DB}$ consists of texts and images from the Mocheg database constructed on top of PolitiFact and Snopes [26].

Stage 2: Evidence Alignment This stage aligns text and image evidence to extract meaningful connections. For each pair of image $img_i \in I$ and text $text_j \in T$, we create a set of pairs $P = \{(img_i, text_j), \ldots\}$, covering all $n \times m$ combinations. An LLM generates a short explanation for each pair, describing whether the image and text are consistent, using the following prompt template:

```
Please generate a short paragraph describing the consistency of
    the image based on the given text following this template:
    <HYPOTHESIS>: Please determining whether the image is
    consistent with the text or not.
    <EXPLANATION>: Explanation the aligment between the image
    hypothesis and the text.
    <FINAL ANSWER>: Give one paragraph describing the consistency
    of the image and text based on the explanation.
    <RESPONSE>:
```

The output is a list P', where each element is a triple (h, e, c): h states the hypothesis (consistent or not), e provides the explanation, and c summarizes the consistency. This verbalized alignment, as shown in Figure 4.2, enhances the LLM's ability to reason about multimodal evidence. We implement the alignment step with *Llama-3.2-90B-Vision-Instruct*, which ingests each

[2] https://www.snopes.com/

[3] https://www.politifact.com/

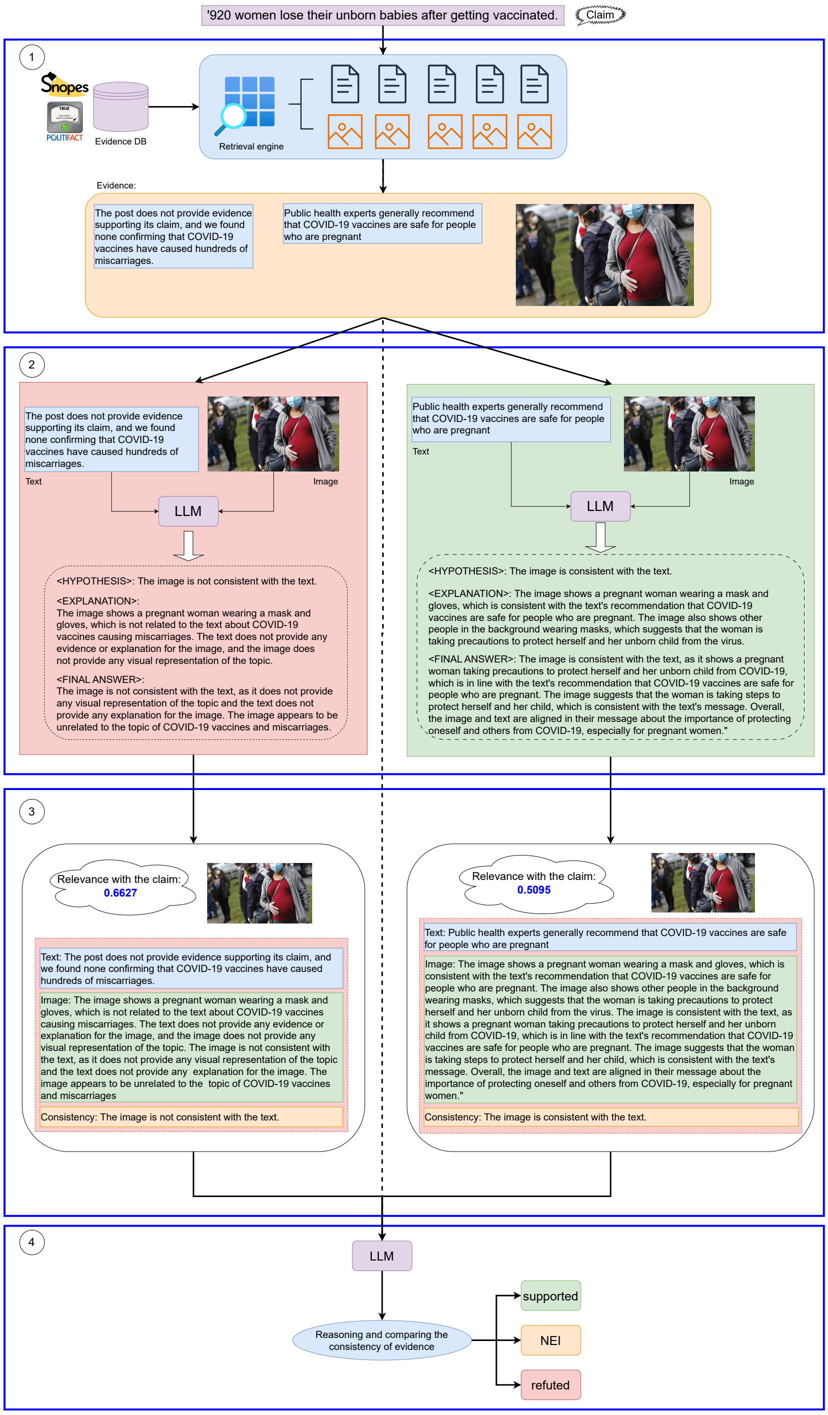


Figure 4.2: Overview of the M-RAV framework.

(img_i, $text_j$) pair and produces the structured triple (h, e, c) as specified above.

Stage 3: Relevance Scoring To assess the relevance of evidence to the claim, we compute similarity scores between the claim c and each verbalized evidence pair in P'. We use LLM2Vec [142] to obtain sentence-level embeddings by reusing a large language model as an encoder. In our setup, we embed the claim c and the concatenated verbalization v_{ij} for each pair and compute cosine similarity. LLM2Vec has been reported to provide strong text embeddings across tasks [143], making it a practical choice for robust relevance scoring.

Stage 4: Claim Verification In the final stage, the LLM verifies the claim's truthfulness based on the claim c and the evidence set E (including text, images, alignment explanations, consistency conclusions, and relevance scores). The LLM outputs a label $lb \in$ {supported, refuted, nei} (where "nei" stands for not enough information), treating the task as a multiclass classification problem. The following prompt guides the LLM:

```
Is it true that: <the claim>?
Here are the evidence for checking:
    Text: <text evidence>
    Image: <image evidence>
    Description: <the explanation for the alignment between the
 text and image evidence>
    Consistency: <the conclusion about the consistency between
the text and image evidence>
    Relevance score: <relevant score of claim and evidence>
    ....
To verify the truthfulness of the claim, please following these
 steps:
    STEP 1: Consult the relevance between the claim and each
given evidence based on the relevance score.
    STEP 2: Think and conclude the truthfulness of the claim
based on the relevance and logical of the evidence.

    The truthfulness must be only one of three value: supported
, refuted, or not enough information. Please think step-by-step
 carefully and response only the truthfulness of the claim.
<RESPONSE>:
```

From the prompt template, the "Description" indicates the explanation for the alignment between the text and image evidence, along with "Consistency" showing whether the text and image in the evidence are consistent or not. This information is extracted from Stage 2 in the M-RAV framework as shown in Figure 4.2. Finally, the "Relevance score" describes the level of similarity between the claim and the evidence, which is extracted from Stage 3. A visual depiction of the operation of our M-RAV strategy on the Mocheg benchmark, with the additive information of each component, is displayed in Figure 4.2. We provide a corresponding algorithmic description in Algorithm 4.1.

Algorithm 4.1 The algorithm of the M-RAV framework

Input: Claim C

Retrieval (Stage 1): Using CLIP to retrieve top-k images $\mathcal{I}$ and FlagEmbedding-BGE to retrieve top-k relevant texts $\mathcal{T}$ from $\mathcal{DB}$.

Augmentation (Stage 2):

for *text t in* $\mathcal{T}$ **do**

  for *image i in* $\mathcal{I}$ **do**

    Generate an explanation e to describe the consistency between t and i with the LLM.

    Store the explanation e, image i, and text t. Denote this triple as raw consistent explanation p such that $p = \langle e, i, t \rangle$.

Relevance Computation (Stage 3): Compute the semantic similarity for all triples $\{p_j\}_{j=1}^{\|\mathcal{I}\|*\|\mathcal{T}\|}$. Denote the tuple as raw consistent text explanation ex such that $ex_j = \langle e_j, t_j \rangle$, we compute the semantic similarity to C by using LLM2Vec with the distance function as the cosine distance s_{c_j} as follows:

$$s_{c_j} = 1 - \frac{\langle ex_j, C \rangle}{\|ex_j\|\|C\|}$$

Denote this quadruple as the evidence v_j such that $v_j = \langle s_{c_j}, e_j, i_j, t_j \rangle$.

Verification (Stage 4): A function $inference\_LLM$ takes $\langle C, v_j \rangle$ as an input, and generates a final label $lb \in \{$`refuted, nei, supported`$\}$. Mathematically, the function $inference\_LLM$ representing the **Verification** module, is expressed as follows:

$$lb = inference\_LLM(C, v_j)$$

4.3 Datasets

To evaluate the M-RAV framework, we use three large-scale multimodal datasets designed for fact-checking: Mocheg, FACTIFY, and Fin-Fact. These datasets, sourced from reliable platforms like PolitiFact and Snopes, are manually annotated by humans. Since M-RAV operates in a zero-shot setting, we use these datasets solely for testing, without fine-tuning the model. Each dataset has unique characteristics, making them suitable for assessing the framework's ability to handle diverse fact-checking challenges.

Mocheg The Mocheg dataset supports end-to-end fact-checking on social media platforms like Twitter [26]. Built from PolitiFact and Snopes, it includes tasks for evidence retrieval, claim verification, and truthfulness justification. We use its test set, containing 2,001 claim instances, to evaluate M-RAV and compare it with state-of-the-art (SOTA) methods.

FACTIFY FACTIFY is a dataset focused on fact-checking social news, primarily from Indian sources like Vishwas, Times of India, AFP India, and BOOM [24]. It is designed specifically for claim verification. We use its test set, with 7,500 claim instances, to assess M-RAV's performance and compare it with other approaches.

Fin-Fact Fin-Fact targets fact-checking financial information on social platforms [139]. Sourced from PolitiFact, Snopes, and FactCheck, it supports claim verification and truthfulness justification. We use the entire dataset, comprising 3,369 claim instances, to test M-RAV's performance across different large language models (LLMs).

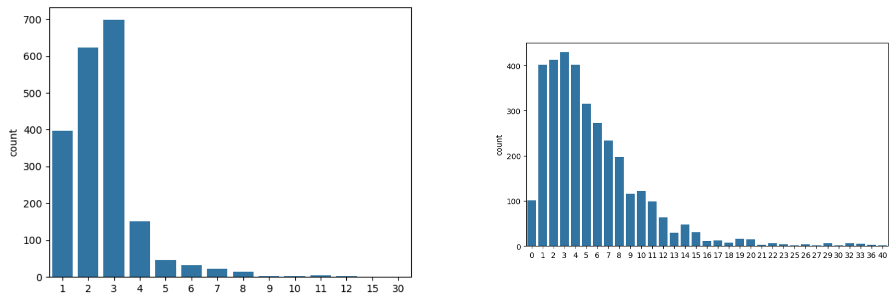
The datasets differ in structure and challenges. Mocheg and Fin-Fact provide multiple pieces of evidence per claim, while FACTIFY offers a single piece of evidence per modality (text or image). However, FACTIFY's text evidence is significantly longer than that in Mocheg or Fin-Fact. Additionally, Mocheg often lacks sufficient image evidence, while Fin-Fact includes claims with missing image or text evidence. These unique characteristics test the robustness of M-RAV and the ability of LLMs to verify claims effectively. In the following section, we analyze each dataset's characteristics, including evidence distribution, text evidence length, claim length, and label distribution, to further understand their impact on fact-checking performance.

4.3.1 Distribution of evidence

According to Table 4.1, it can be seen that the FACTIFY dataset has a single evidence for each modality per claim. For Mocheg, there is a claim with the maximum number of text evidence equal to 30. This is similar to the FINFACT with a maximum of 40 pieces of text evidence for a claim. For image evidence, the maximum number of evidence per claim on Mocheg and FINFACT is 8 and 18, respectively. This indicates the challenge of multiple pieces of evidence in the claim on Mocheg and FINFACT. On average, the number of text evidence in FINFACT is more than in the Mocheg dataset (2.61 pieces of text evidence per claim for Mocheg and 5.44 pieces of text evidence per claim for FINFACT), while the number of images is nearly the same (approximately 1 image per claim on both datasets).

Table 4.1: Distribution of evidence in three datasets

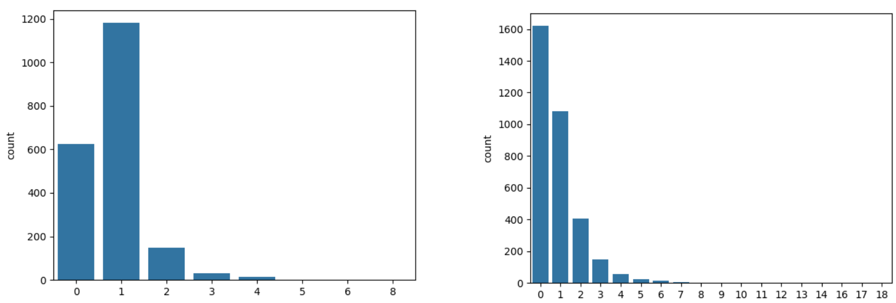
	Mocheg	**FACTIFY**	**FINFACT**
Max text evidence	30	1	40
Average text evidence	2.61	1	5.44
Median text evidence	2	1	4
Min text evidence	1	1	0
Max image evidence	8	1	18
Average image evidence	0.82	1	0.88
Median image evidence	1	1	1
Min image evidence	0	1	0



Mocheg

FINFACT

Figure 4.3: Distribution of Text evidence per claim in Mocheg and FINFACT dataset.



Mocheg

FINFACT

Figure 4.4: Distribution of Image evidence per claim in Mocheg and FIN-FACT dataset.

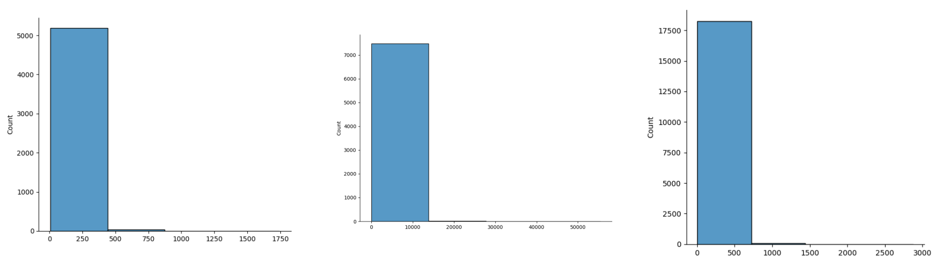
Investigating the distribution of text evidence per claim, it can be seen that most of the claims in Mocheg and FINFACT have about 3 to 4 pieces of evidence supporting the veracity of the claim, as shown in Table 4.1. There are no claims with missing text evidence in the Mocheg, while the FINFACT has about 100 claims with missing text evidence. For the image evidence distribution, as shown in Figure 4.4, most of the claims in Mocheg and FINFACT have about 1 to 2 image evidence. Specifically, the claim with no image in both datasets is significantly high.

4.3.2 Distribution of text evidence length

Table 4.2 shows the text evidence length distribution of Mocheg, FACTIFY, and FINFACT datasets. We utilize the spaCy[4] for tokenizing the text into tokens, and computing statistics based on the token level. It can be seen that FACTIFY has the longest text evidence than Mocheg with gold evidence and FINFACT. As shown in Figure 4.5, most of the text evidence in Mocheg has a length of around 250 tokens, and around 500 tokens for the FINFACT. Instead, most of the text evidence in FACTIFY is quite long, which is about 10,000 tokens.

Table 4.2: Length distribution of text evidence

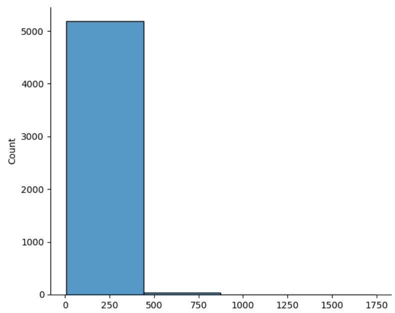
	Mocheg (gold evidence)	**FACTIFY**	**FINFACT**	**Mocheg (system evidence)**
Max length	1,745	55,459	2,882	142,214
Average length	67.06	1,107.63	63.27	481.67
Median length	33	524	51	70
Min length	5	3	0	4



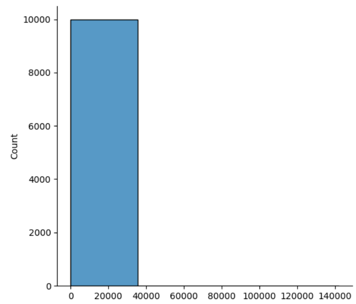
Mocheg FACTIFY FINFACT

Figure 4.5: Distribution of Text evidence length.

[4] https://spacy.io/



Mocheg - gold evidence



Mocheg - system evidence

Figure 4.6: Distribution of text evidence length on Mocheg.

Besides, as shown in Figure 4.6, there is a significant difference between the gold evidence and the system evidence in the Mocheg, where most of the system evidence is much longer than the gold evidence, since the system evidence retrieved by the retrieval model is at the article level.

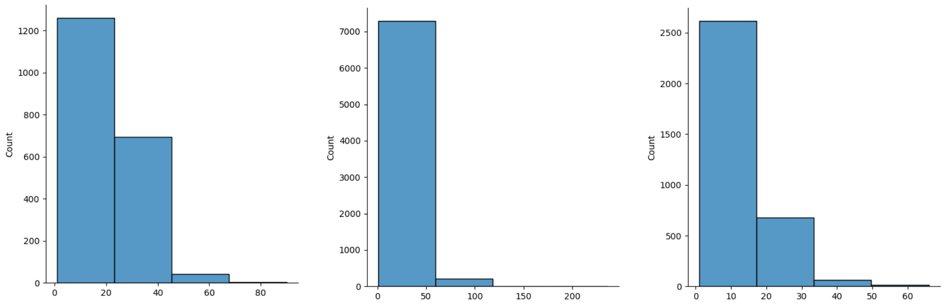
4.3.3 Distribution of claim length

Table 4.3 describes the maximum and minimum length of claims in three datasets. We also use spaCy[5] for tokenizing the text into tokens and computing statistics based on the token level. It can be seen that the length of the claim in Mocheg and FINFACT is similar, while the claim in the FACTIFY dataset is slightly longer. As shown in Figure 4.7, most of the claim in Mocheg and FINFACT has a length of around 20 tokens, while the claim in FACTIFY contains around 50 tokens. In comparison with the text evidence, the number of tokens in the claim is much less than the evidence.

Table 4.3: Length distribution of claim.

	Mocheg	**FACTIFY**	**FINFACT**
Max length	90	236	66
Average length	22.38	30.71	13.59
Median length	21	28	13
Mode length	16	19	12
Min length	1	1	1

[5] https://spacy.io/



Mocheg FACTIFY FINFACT

Figure 4.7: Distribution of claim length.

4.3.4 Distribution of labels

Table 4.4 and Figure 4.8 illustrate the label distribution in the three datasets. For Mocheg and FACTIFY, the statistic is conducted on the test set. Since FINFACT does not split the data into train, dev, or test sets, we use the whole dataset to analyze and test our methods. In general, the Mocheg has a balanced distribution on three labels, while FACTIFY takes a minority proportion at the "refuted" label, and FINFACT has fewer samples in the "NEI".

Table 4.4: Distribution of label in Mocheg, FACTIFY, and FINFACT.

	Mocheg	**FACTIFY**	**FINFACT**
supported	667	3,000	1,275
NEI	667	3,000	602
refuted	667	1,500	1,492

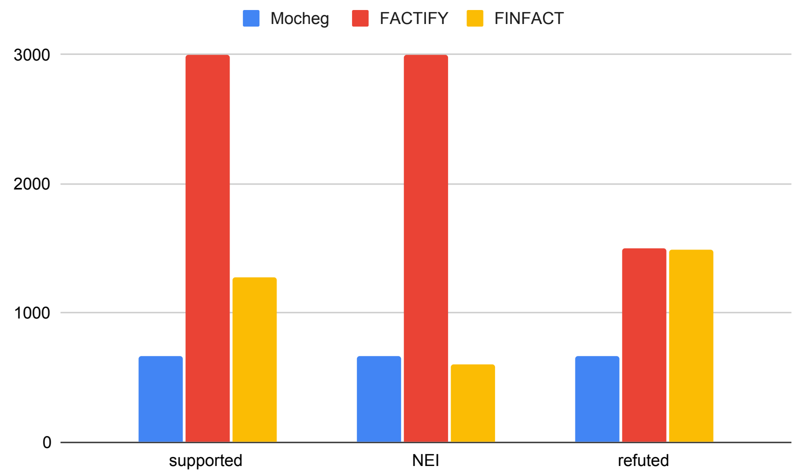


Figure 4.8: Visualization of the distribution of labels in Mocheg, FACTIFY, and FINFACT.

4.4 Results

4.4.1 Empirical Preparation

Table 4.5 shows the pre-trained LLMs and their configurations used in our empirical study about the M-RAV. We run the LLMs with a 4-bit quantization setting using the *BitandBytes*[6] library. All models are run using the HuggingFace library. We run the experiment on a single NVIDIA RTX A6000 with 49 GB of memory.

Table 4.5: Model configurations

Task	Model name	Pre-trained	Configuration
Making aligment	LlaMa 3.2 Vision	meta-llama/Llama-3.2-90B-Vision	max_new_tokens=512, temperature=0.2, top_p=0.5, do_sample=True
Relevance score computation	LLM2Vec	McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp	default configuration
Retrieval	FlagEmbedding CLIP	BAAI/bge-m3 (text evidence) ViT-L/14@336px (image evidence)	default configuration
Verification	LlaMa 3.2	meta-llama/Llama-3.2-3B-Instruct	do_sample=False, max_new_tokens=20
	LlaMa 3.1	meta-llama/Llama-3.1-8B-Instruct meta-llama/Llama-3.1-70B-Instruct	
	Qwen-2.5	Qwen/Qwen2.5-32B-Instruct Qwen/Qwen2.5-72B-Instruct	
	Mixtral	mistralai/Mixtral-8x7B-Instruct-v0.1	
	LLaVaNext	llava-hf/llava-v1.6-vicuna-7b-hf llava-hf/llava-v1.6-vicuna-13b-hf llava-hf/llava-v1.6-34b-hf	

[6] https://pypi.org/project/bitsandbytes/

To assess the performance of the retrieval task, we use the F1@k and MAP metrics based on the ground truth provided from the Mocheg dataset [26]. The F1@K metric combines precision and recall into a single score, measuring the proportion of correctly identified relevant items relative to both the system's predicted positives and the actual ground truth positives [144]. This metric offers an efficient way to evaluate the performance of retrieval systems in practical applications. To further assess both the presence and ranking of relevant items, we incorporate Mean Average Precision (MAP) alongside F1@K, providing a comprehensive evaluation of the evidence retrieval task's performance. For the verification task, we use both micro and macro F1 scores to evaluate the performance of various models based on the gold labels in three benchmark datasets.

4.4.2 Gold-evidence results

In this experiment, we investigate the ability of different LLMs to verify the claim. In this scenario, the gold evidence means that the claim has a set of evidence that has been created and verified by humans. We run the evaluation by two categories of LLMs: the text-only LLMs, including LLaMA [107], Qwen [122], and Mixtral [145], and the vision-language LLM LLaVA-NeXT [146], as shown in Table 4.6. The details about the prompt for LLMs are shown in Appendix A.

Table 4.6: Empirical results on gold evidence

	Mocheg		**FACTIFY**		**Fin-Fact**	
	F1-micro	**F1-macro**	**F1-micro**	**F1-macro**	**F1-micro**	**F1-macro**
llama-3.2-3B + M-RAV	32.28 ↓	25.76 ↓	**39.12** ↑	**33.36** ↑	**29.77** ↑	**26.11** ↑
llama-3.2-3B	**32.43**	**27.51**	38.30	27.54	29.71	20.99
llama-3.1-8B + M-RAV	**49.92** ↑	46.12 ↓	**66.86** ↑	**69.54** ↑	**49.24** ↑	**38.23** ↑
llama-3.1-8B	49.72	**49.14**	64.94	68.73	38.79	29.73
Qwen-2.5-32B + M-RAV	**57.37** ↑	**56.77** ↑	**66.86** ↑	**70.13** ↑	**49.33** ↑	**48.60** ↑
Qwen-2.5-32B	53.47	49.82	65.22	67.54	45.71	43.99
llama-3.1-70B + M-RAV	**55.52** ↑	**54.19** ↑	**72.33** ↑	**74.87** ↑	**63.52** ↑	**57.23** ↑
llama-3.1-70B	55.27	52.49	71.22	74.44	61.14	59.61
Mixtral-8x-7b + M-RAV	56.77 ↓	56.03 ↓	**60.06** ↑	**59.79** ↑	36.12 ↓	35.40 ↓
Mixtral-8x-7b	**57.97**	**57.58**	54.86	51.02	**44.43**	**43.85**
LLaVaNext-7B + M-RAV	40.32 ↓	34.95 ↓	**40.61** ↑	**21.49** ↑	**30.18** ↑	**29.31** ↑
LLaVaNext-7B	**42.02**	**38.06**	40.05	19.78	28.10	26.54
LLaVaNext-13B + M-RAV	**43.02** ↑	**42.48** ↑	**41.48** ↑	**36.61** ↑	34.66 ↓	35.34 ↓
LLaVaNext-13B	40.07	34.24	41.29	25.21	**37.37**	**38.03**
LLaVaNext-34B + M-RAV	**50.07** ↑	**49.15** ↑	**45.94** ↑	**38.99** ↑	40.48 ↓	40.49 ↓
LLaVaNext-34B	45.42	41.59	43.64	30.35	**46.83**	**46.89**

From Table 4.6, it can be seen that the M-RAV helps to increase the performance of LLaVA-34B and LLaVA-13B (vision-language LLMs) on the

Mocheg dataset. However, on the Fin-Fact dataset, the M-RAV shows efficiency when it increases the accuracy of both text-only and vision-language LLMs (except LLaVANeXT-34B and Mixtral-8x-7b). The M-RAV especially increases the performance significantly of all LLMs on the FACTIFY dataset. In general, the M-RAV helps to improve the ability of LLMs to reason and understand the evidence context, which determines the truthfulness of the claim. In addition, as shown in Table 4.6, the performance of text-only LLMs such as LLaMA, Qwen, and Mixtral is better than vision-language LLMs like LLaVA. The *llama-3.1-70B* attains the best performance on both Fin-Fact and FACTIFY datasets when enhanced with M-RAV, while it performs well with Mocheg in the case of no M-RAV. On the other hand, the *Qwen-2.5-32B* obtains good results on the Mocheg when enhanced with M-RAV. Moreover, from Table 4.6, it can be seen that the M-RAV method shows a significant impact with 32 billion models or more since it leverages the ability of reasoning on large models for boosting the accuracy for the zero-shot fact verification task. Overall, the augmentation step, which integrates the explanation of the consistency between the image and the text evidence to the LLMs, helps improve the performance of the multimodal fact-checking task (answered for the **RQ1**).

Table 4.7: Comparison between M-RAV with SOTAs methods.

Dataset	Methods	F1-micro	F1-macro
Mocheg	Baseline (CLIP-large) [26]	50.78	-
	CLIP + GPT-4o Guided [54]	54.22	-
	LVLM4FV [147]	53.40	53.55
	M-RAV -*Qwen-2.5-32B* (our)	**57.37**	**56.77**
FACTIFY	Baseline (ResNET+SBERT) [24]	-	53.09
	Logically [148]	-	76.71
	ECENet [50]	-	81.50
	M-RAV - *llama-3.1-70B* (our)	**72.33**	**74.87**
Fin-Fact	Baseline (Gemini Vision pro) [139]	58.00	59.00
	Baseline (LLaVA-13B) [139]	50.00	45.00
	FinLLaMA [149]	34.62	-
	M-RAV - *llama-3.1-70B* (our)	**63.52**	**57.23**

Next, comparing the baseline shown in Table 4.7, our proposed method M-RAV shows the efficiency in boosting the LLMs for verifying the claim when the accuracy computed by F1 scores is much better than the results of baselines on three datasets. In the Mocheg dataset, the performance of *Qwen-2.5-32B* with M-RAV obtained 57.37% by F1-score, which is a better performance than CLIP+GPT-4o Guided with about 54.22% by F1-score [54] and LVLM4FV with 53.40% by F1-score [147]. Besides, the *llama-3.1-70B* with M-RAV also obtains potential results on Fin-Fact when compared with FinLLaMA, a series of LLMs trained on financial data [149] (M-RAV with *llama-3.1-70B* obtained 74.87% by F1-score in comparison with FinLLaMA that attains 34.62%). On the FACTIFY dataset, the Logically [148] and ECENet [50] show high results since they were carefully trained and fine-tuned on the training set of FACTIFY. Moreover, the ECENet [50] employed Deep Reinforcement Learning (DRL) to extract the most significant sentence that impacts the veracity of the claim and then adjust the weight through a reward model. From the empirical results, our M-RAV method generally yields competitive results with SOTAs, even without training or fine-tuning steps for the LLMs.

4.4.3 Ablation Study

To investigate the impact of the relevance score, text-image alignment, and evidence in the M-RAV framework for fact-checking, we conducted an ablation study with three configurations: (1) excluding the relevance score (w/o relevance score), (2) excluding the text-image alignment while also omitting the relevance score (w/o alignment), and (3) using only the claim without supporting evidence (w/o evidence). Table 4.8 presents the results of this ablation study, based on the *Qwen-2.5-32B-Instruct* model, evaluated across three benchmarks.

Table 4.8: Ablation Study of M-RAV

	Mocheg		FACTIFY		Fin-Fact	
	F1-micro	**F1-macro**	**F1-micro**	**F1-macro**	**F1-micro**	**F1-macro**
w/o relevance score	57.37	56.67	64.57	67.15	50.60	50.46
w/o alignment	47.32	47.30	69.41	72.46	31.49	29.64
w/o evidence	37.63	27.81	41.64	27.38	18.84	11.97
Full	57.37	56.77	66.86	70.13	49.33	48.60

As shown in Table 3.3, alignment plays a crucial role in improving the verification model's performance on Mocheg and Fin-Fact, the two benchmarks that involve multiple pieces of evidence per claim. Without alignment, both macro and micro F1 scores drop significantly on these datasets. In contrast, alignment has minimal impact on FACTIFY, as this benchmark provides only a single piece of evidence—combining text and image—for each claim. Overall, the performance difference with and without alignment remains small. In practical scenarios with multiple evidence items per claim, aligning each text-image pair is essential to strengthen the verification model's ability to assess claim veracity. Furthermore, evidence is central to the verification process, as model performance declines sharply across all three benchmarks when evidence is absent.

Finally, the relevance score has less impact on the performance of verification models, as shown in Table 3.3. However, the relevance score plays a key role in filtering the most relevant alignment evidence for the claim, especially in a system-evidence scenario where the evidence is automatically retrieved. The next section presents an empirical study on system-evidence results, where the relevance score is used to identify the top-k most relevant alignments for the input claim.

4.4.4 System-evidence results

This section evaluates our M-RAV on different LLMs in the system evidence scenario. In this experiment, we only give the claim. The system retrieves the relevant evidence, including text and images, from the evidence database. Then, these pieces of evidence are augmented by M-RAV to extract the veracity of the claim.

Table 4.9: Performance of the retrieval module on the development set of Mocheg by different top-k

top-k	**Modality**	**prec@K**	**rec@K**	**F1@K**	**MAP**
k=5	Image	**7.50**	27.13	**11.35**	22.47
	Text	**23.70**	51.46	**30.19**	**57.86**
k=7	Image	5.76	29.22	9.35	22.65
	Text	18.88	56.21	26.48	57.17
k=10	Image	4.25	31.13	7.31	22.80
	Text	14.23	59.80	21.75	56.58
k=15	Image	3.09	34.61	5.60	22.71
	Text	10.04	63.15	16.59	55.68
k=20	Image	2.44	36.77	4.52	22.81
	Text	7.79	65.07	13.43	55.04
k=25	Image	2.02	38.08	3.80	22.84
	Text	6.39	66.46	11.31	54.42
k=30	Image	1.74	39.54	3.32	22.81
	Text	5.42	67.58	9.78	54.14
k=40	Image	1.35	41.12	2.61	22.78
	Text	4.19	69.60	7.74	53.46
k=50	Image	1.10	41.75	2.14	22.78
	Text	3.42	70.73	6.42	52.97

Table 4.10: Result of retrieval module on the test set of Mocheg

Method	**Modality**	**prec@K**	**rec@K**	**F1@K**	**MAP**
BGE+CLIP (k=5)	Image	4.26	14.89	6.31	**12.32**
	Text	**22.45**	**47.68**	**29.06**	**58.71**
Baseline (k=5)	Image	4.71	17.01	7.37	11.93
	Text	*14.92*	*19.72*	*16.98*	14.34
Baseline (k=10)	Image	3.02	21.44	5.29	12.58
	Text	*9.79*	*23.99*	*13.90*	15.34

In the retrieval step, we employ the FlagEmbedding BGE-M3 [141] for retrieving the text and CLIP [37] for retrieving the image based on the claim as the query. We get the *top-k* most relevant evidence to the claim. To choose an efficient *top-k*, we test the performance of the retrieval model on *top-k* $\in \{5, 7, 10, 15, 20, 25, 30, 40, 50\}$, as shown in Table 4.9. We obtain that the ***top-5*** achieves the best performance according to the F1@k metric. Finally, we use the ***top-5*** for the retrieval model and attain the competitive results with the baseline in Mocheg [26], as shown in Table 4.10.

Table 4.11: The impact of choosing the top-k most relevant alignment for verifying the claim

top-K aligment	**F1-micro**	**F1-macro**
5	**49.00**	**49.42**
10	46.83	45.16
15	45.83	41.87
20	45.83	40.20
25	44.50	38.95

Table 4.12: Empirical results on system evidence

	F1-micro	**F1-macro**
llama-3.2-3B + M-RAV	**32.63** ↑	28.34 ↓
llama-3.2-3B	31.48	**29.96**
llama-3.1-8B + M-RAV	**49.02** ↑	**41.22** ↑
llama-3.1-8B	45.32	42.66
Qwen-2.5-32B + M-RAV	**50.02** ↑	**50.14** ↑
Qwen-2.5-32B	42.62	38.06
llama-3.1-70B + M-RAV	**49.17** ↑	**46.33** ↑
llama-3.1-70B	45.07	43.55
Mixtral-8x-7b + M-RAV	**49.52** ↑	**47.55** ↑
Mixtral-8x-7b	47.42	45.66
LLaVaNext-7B + M-RAV	33.28 ↓	16.74 ↓
LLaVaNext-7B	**33.38**	**16.79**
LLaVaNext-13B + M-RAV	33.23 ↓	16.73 ↓
LLaVaNext-13B	**33.43**	**16.90**
LLaVaNext-34B + M-RAV	**38.73** ↑	**30.91** ↑
LLaVaNext-34B	33.38	21.17
Baseline	44.06	-

Additionally, the alignment between text and image pairs explains their consistency. When selecting the top-N images and top-M texts, the total number of possible alignments is roughly $N \times M$. For example, choosing the top-5 images and top-5 texts generates 25 alignment pairs. This large number of pairs can create lengthy input for large language models (LLMs), potentially reducing their performance during claim verification. To address this issue, we conducted an empirical study to identify the most relevant text-image alignments using relevance scores from Stage 3 of the M-RAV framework. We sampled data from the Mocheg development set and used retrieval models to select the top-5 evidence items (both texts and images). Then, we calculated relevance scores for the alignments and selected the k most relevant pairs, where k was set to 5, 10, 15, 20, or 25. These pairs were used to verify claims with the *Qwen-2.5-32B-Instruct* model. As shown in Table 4.11, the top-5 alignments ($k = 5$) yielded the best performance. Therefore, for each claim, we use the top-5 most relevant text-image alignments based on their relevance scores.

Next, we augmented the retrieved evidence with the M-RAV method to get a list of verbalized evidence between text and image (as described in Figure 4.2). Then, based on the relevance score with the claim, we choose *top-5* most relevant evidence to guide the LLMs in verifying the veracity.

As shown in Table 4.12, the M-RAV boosted the performance of all text-only LLMs in the verification of the claim with system evidence. To the vision-language LLMs, the M-RAV enhanced the performance of LLaVaNext-34B. The Qwen-2.5-32B obtains the best performance with 50.02% by F1-micro score, which is approximately 6% better than the baseline, and slightly reaches the performance of the baseline trained on gold evidence of Mocheg [26]. Therefore, the results from the empirical study indicate that the M-RAV is efficient for applying to the fact-checking system, where the evidence is automatically retrieved (answered for the **RQ2**).

4.5 Error Analysis

Table 4.13 describes the detailed performance of the classification of LLMs on the three datasets. For text-only LLMs, we choose the results of the model that has the best performance on each dataset by M-RAV according to Table 4.6 and Table 4.12, in which the *Qwen2.5-32B* for Mocheg, the *llama-3.1-70B* for FACTIFY, and Fin-Fact. Similarly, we choose the *LLaVaNext-34B* as the vision-language LLM for Mocheg, FACTIFY, and Fin-Fact.

Table 4.13: Performance results of text-only and vision-language LLMs on the three datasets. F1 represents the F1-macro score, P is Precision, and R is Recall.

	Macro F1	*Macro P*	*Macro R*	*Support F1*	*Support P*	*Support R*	*NEI F1*	*NEI P*	*NEI R*	*Refute F1*	*Refute P*	*Refute R*
	Mocheg - gold evidence											
Text-only LLMs + M-RAV	**56.78**	58.83	**57.37**	**68.66**	52.90	**43.03**	70.99	64.05	**79.61**	46.45	43.77	49.48
Text-only LLMs	49.83	**59.48**	53.47	26.94	**70.70**	16.64	**71.99**	**66.54**	78.41	**50.55**	41.21	**65.37**
Vision-language LLMs + M-RAV	**49.16**	57.16	**50.07**	**62.11**	53.40	**74.21**	**45.12**	81.57	**31.18**	40.24	36.51	44.83
Vision-language LLMs	41.59	**63.00**	45.43	28.50	**68.57**	17.99	43.93	**83.26**	29.84	**52.35**	**37.18**	**88.46**
	Mocheg - system evidence											
Text-only LLMs + M-RAV	**50.15**	**55.40**	**50.02**	**43.50**	**61.71**	33.58	**59.90**	**66.61**	**54.42**	**47.05**	**37.88**	62.07
Text-only LLMs	38.06	56.24	42.63	23.92	67.36	14.54	39.50	65.17	28.34	50.76	36.18	**85.01**
Vision-language LLMs + M-RAV	**30.92**	**58.63**	**38.73**	**54.12**	**43.20**	**72.41**	**1.49**	**100**	**0.75**	**37.15**	**32.69**	43.03
Vision-language LLMs	21.18	35.00	33.38	14.93	43.80	0.90	0.59	28.57	0.30	48.02	32.63	**90.85**
	FACTIFY											
Text-only LLMs + M-RAV	**74.87**	**79.18**	**73.78**	**73.17**	62.97	**87.33**	89.31	**99.51**	81.00	62.13	**75.07**	53.00
Text-only LLMs	74.44	77.13	72.93	70.06	**64.01**	77.37	**89.52**	99.35	**81.47**	**63.75**	68.04	**59.97**
Vision-language LLMs + M-RAV	**38.99**	61.78	**40.44**	**59.25**	48.71	**75.60**	**22.89**	**99.49**	**12.93**	34.84	37.15	32.80
Vision-language LLMs	30.36	**68.45**	38.28	12.75	**65.23**	7.07	20.55	98.85	11.47	**57.77**	**41.27**	**96.30**
	Fin-Fact											
Text-only LLMs + M-RAV	57.23	57.74	57.03	69.18	72.45	**66.20**	**75.50**	76.35	**74.66**	27.02	24.43	30.23
Text-only LLMs	**59.62**	**64.04**	**61.82**	**69.29**	**73.46**	65.57	69.04	**88.91**	56.43	**40.51**	**29.75**	**63.46**
Vision-language LLMs + M-RAV	40.49	49.22	43.49	**49.85**	54.91	**45.65**	43.09	73.39	30.50	28.55	19.36	54.32
Vision-language LLMs	**46.90**	**56.26**	**49.73**	49.05	**68.01**	38.35	**59.24**	**78.99**	**47.39**	**32.41**	**21.77**	**63.46**

In general, it can be seen that text-only LLMs have better performance than vision-language LLMs. Although the efficiency of M-RAV is faint on text-only LLMs, it is significant on vision-language LLMs since it helps improve performance. Especially, on the Mocheg with system evidence, the M-RAV helps improve almost both text-only and vision-language models, indicating the contribution of M-RAV in a practical fact-checking system when the evidence is retrieved automatically instead of manually created as gold evidence.

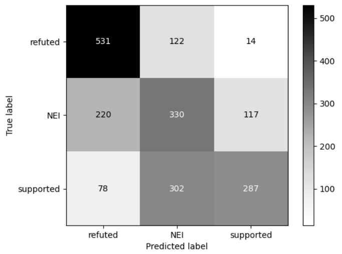
In addition, from Table 4.13, the Mocheg (both gold and system evidence) is mostly misclassified on the Refute and NEI labels. As shown by the confusion matrices in Figures 4.9,4.10,4.11,4.12, most of the time the Refutes is predicted as NEI, while the NEI tends to be predicted as Support. For the FACTIFY dataset, the Refuted model is mostly misclassified to the NEI label and Support label on text-only LLMs, while it is focused mostly on the NEI label when testing with vision-language LLMs. Finally, on the Fin-Fact dataset, the Refute label tends to be predicted as NEI. Overall, it can be seen that the Support and Refute are clearly discriminated against, while the remaining challenge falls into the NEI labels.

To investigate more about the error, we categorize the misclassification sample into 4 types: **Error type 1:** The "support" is being predicted as "NEI", **Error type 2:** The "refute" is being predicted as "NEI", **Error type 3:** The "NEI" is being predicted as "support", and **Error type 4:** The "NEI" is being predicted as "refute". The details of the errors are described as follows, with corresponding examples.

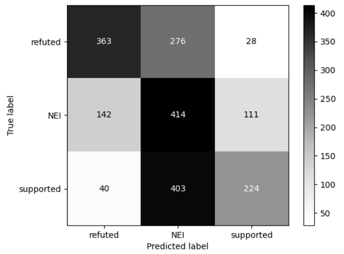
- For **Error type 1**: it can be seen that the model cannot capture the latent information from the image and text. For example, in Figure 4.13, the LLMs understand the *"tall building"* and the *"Burj Khalifa"* in the explanation between the image and the text. Besides, it can exploit the information as *"a person working on a tall building with a cityscape in the background"* from the explanation between the consistency of the image and text. However, the model does not understand *"Tom Cruise"* as a person in the picture thus, it cannot determine whether *"Tom Cruise"* sits on the top of *"a tall building"* named *"Burj Khalifa"*. Therefore, the model predicted "NEI" while the correct label is "support".
- For **Error type 2**: It can be seen from Figure 4.15 that in the explanation, the information about *"a group of sheep passing through a gate"* and *"a fence in the background"* has been correctly extracted. With the first piece of evidence, it is enough to determine the truthfulness of the claim as "refuted" because the explanation and the text image said that *"sheep passing through a gate at Brenda Station"* and *"the image accurately represents the scene described in the text"* is matching with the information in the text evidence as *"fence posts are clearly visible in a higher resolution version"*. However, in the second piece of evidence, there is a little confusion since the information in the image shows that *"fence posts visible in the background"*, which contradicts the information in the text evidence as *"encircling fence may be hard to see"*. Thus, the model predicts the claim as "NEI" instead of

"Refuted".

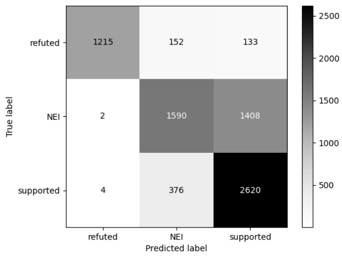
- For **Error type 3**: As shown in Figure 4.16, the text evidence at first seems to support the claim since it says that *"The Tripura government will promote all students from Class 1 to 4 and Class 6 and 7 without examinations this year. 'We have decided to promote students from classes 1 to 4, 6, and 7 to the next classes except Classes 5 and 8.'*. Also, the image is consistent with the text as the explanation shows the information about the *"discussion of promoting students without examinations and extending summer vacations"*. However, in the text evidence, the remaining information seems uncertain about whether the government decided for *"No examination for Class 1 to 4 and Class 6 to 7"* because the text evidence mentioned that *"But the students need to sit for their examinations after opening of schools if the situation becomes normal."*. Moreover, from the evidence, it is said that *"The exams would be held to evaluate their educational loss in the Covid-19 pandemic"* and *"Regarding Class 5 and 8, Nath said that they would send the matter for approval from the state cabinet and for Class 9 and 11 examinations"*, which means it is not sure that *Class 1 to 4 and Class 6 to 7* do not have to take the exam, and *Class 5 and Class 8* must have to take the exam. Therefore, the actual truthfulness of this claim is "NEI". Nevertheless, the model was predicted as "support".
- For **Error type 4**: As illustrated in Figure 4.17, it can be seen that, as with the evidence extracted from the image via explanation and from the text evidence, the claim must be refuted. However, no information showed that the *"Facebook user Rosemary Thomas"* is a customer who found the Rat Head in the food from Popeyes Chicken since this is *"the iteration of a very old urban legend known as the 'Kentucky Fried Rat'"*. Consequently, the actual label of the claim is "NEI".



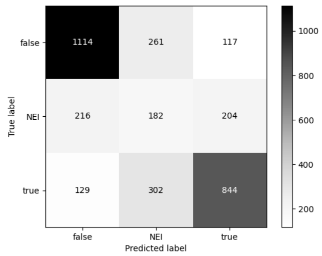
Mocheg - gold



Mocheg - system

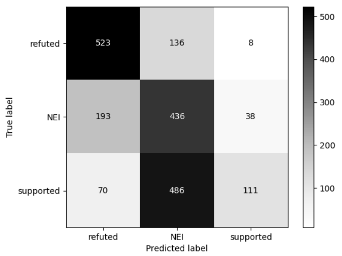


FACTIFY

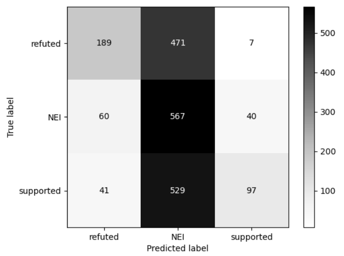


Fin-Fact

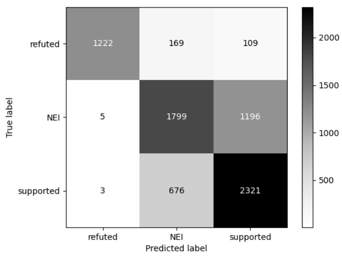
Figure 4.9: Performance of text-only LLMs with M-RAV.



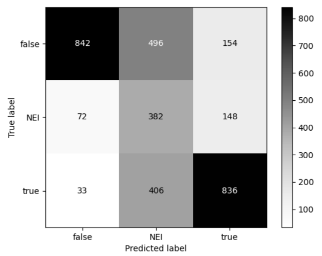
Mocheg - gold



Mocheg - system

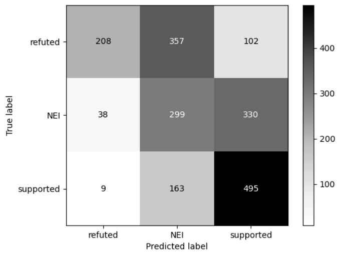


FACTIFY

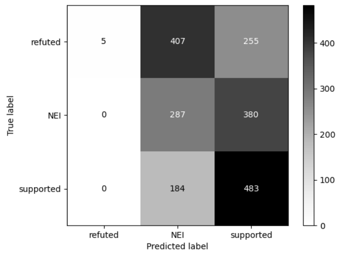


Fin-Fact

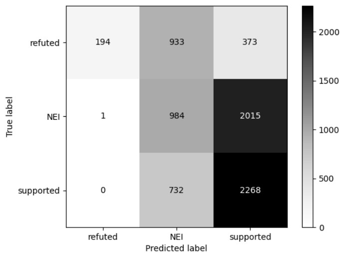
Figure 4.10: Performance of text-only LLMs without M-RAV.



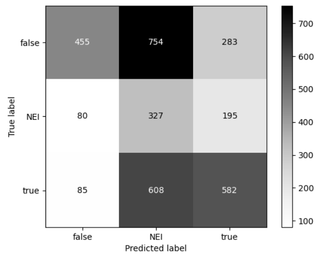
Mocheg - gold



Mocheg - system

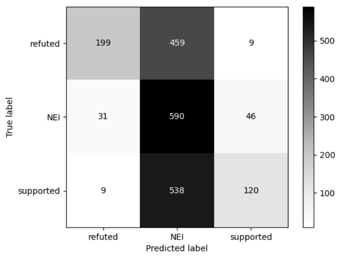


FACTIFY

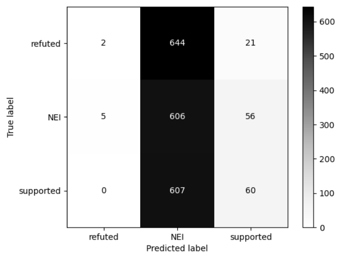


Fin-Fact

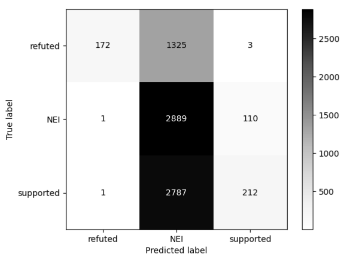
Figure 4.11: Performance of vision-language LLMs with M-RAV



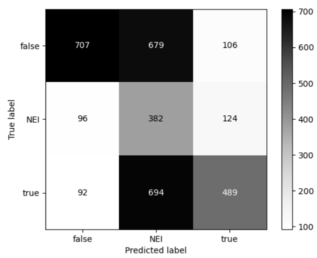
Mocheg - gold



Mocheg - system



FACTIFY



Fin-Fact

Figure 4.12: Performance of vision-language LLMs without M-RAV

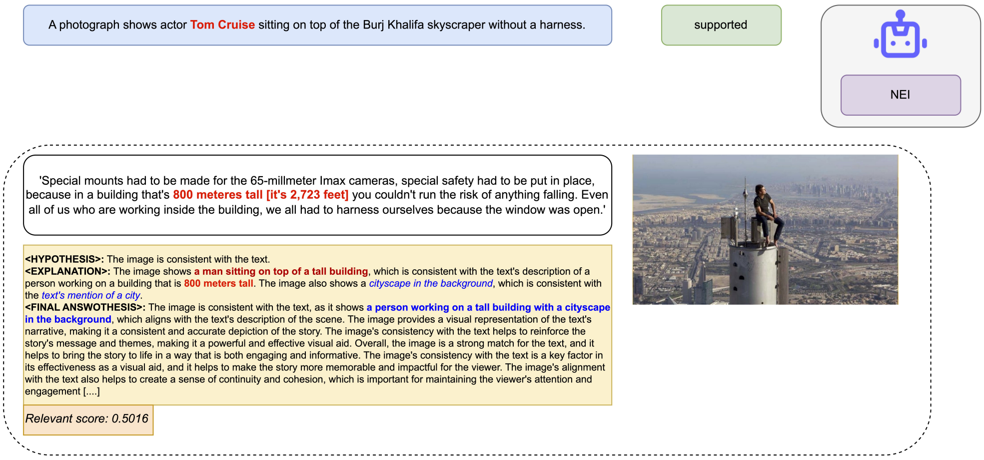


Figure 4.13: Error prediction sample 1. The sample is extracted from the Mocheg dataset.

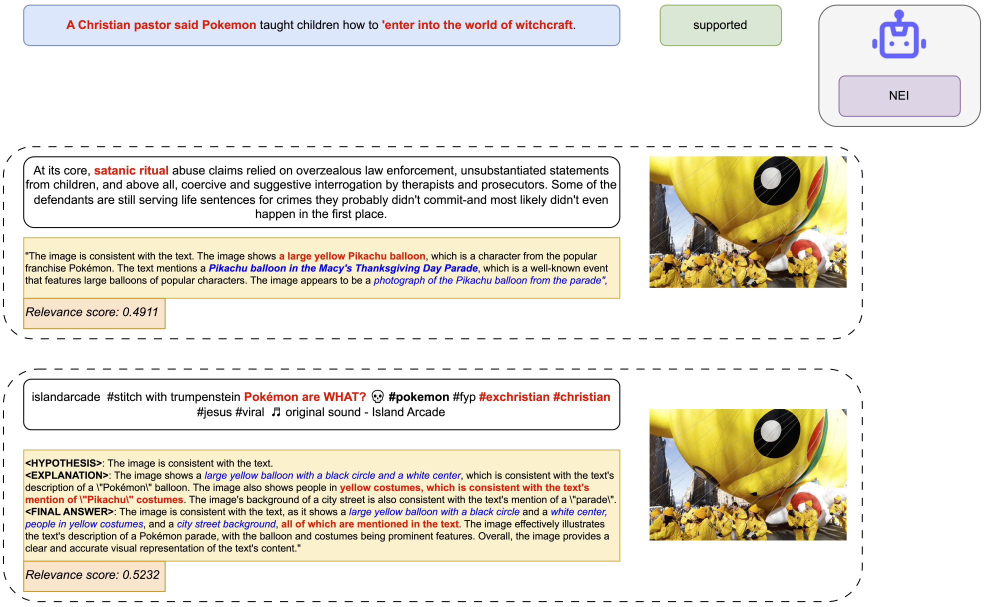


Figure 4.14: Error prediction sample 2. The sample is extracted from the Mocheg dataset.

Figure 4.15: Error prediction sample 3. The sample is extracted from the Fin-Fact dataset.

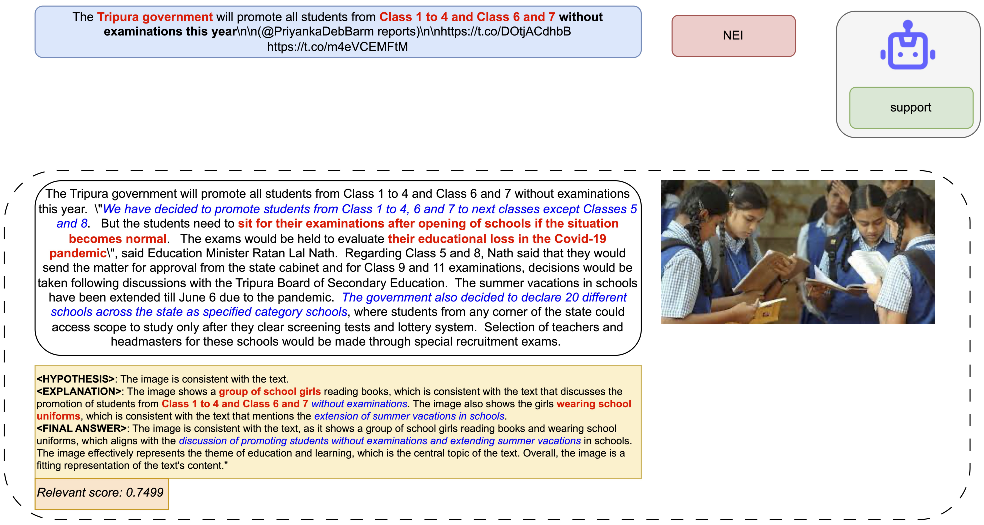


Figure 4.16: Error prediction sample 4. The sample is extracted from the FACTIFY dataset.

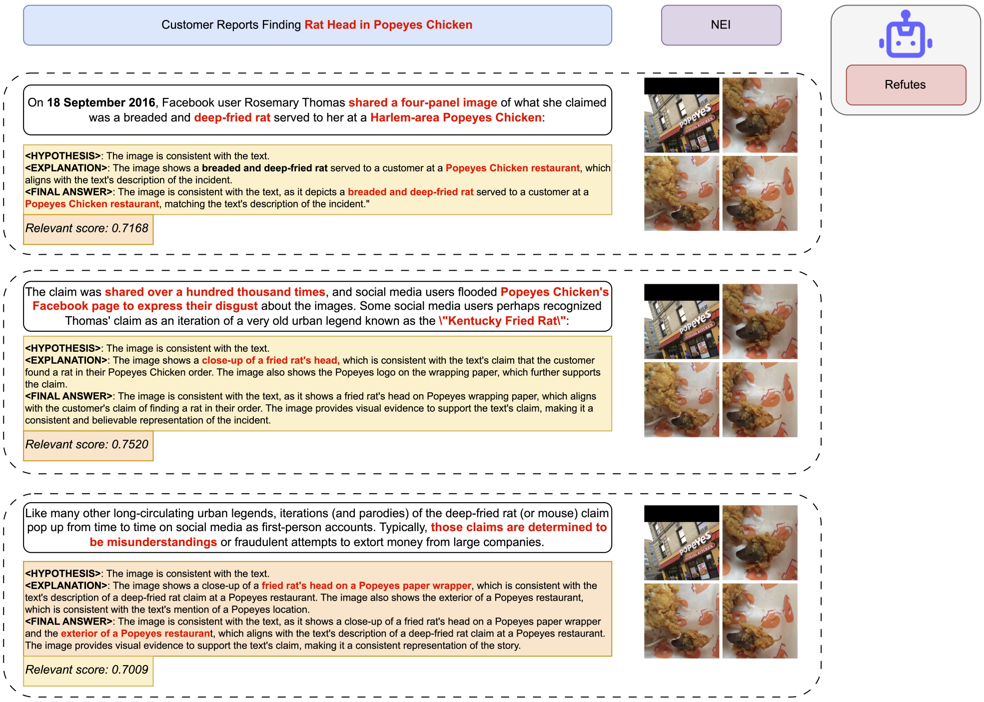


Figure 4.17: Error prediction sample 5. The sample is extracted from the Fin-Fact dataset.

Overall, it can be seen that the evidence in some cases is not enough to determine the veracity of the claim, even if it is correct and clearly represented. To exactly verify the claim, the LLM needs external knowledge about specific cultures, politics, and human behavior to capture the latent contextual information along with the evidence in verifying the claim. To improve how LLMs adapt to new information, several techniques can be used. CAMeL [150] and MemoryLLM [151] help LLMs integrate new knowledge into their internal parameters. Alternatively, approaches like LLatrieval [152] enable LLMs to iteratively validate external knowledge, enhancing their ability to verify information and augment their performance.

4.6 Discussion

The M-RAV (**M**ultimodal **R**etrieve-**A**ugment-**V**erify) is designed to enhance the ability of large language models (LLMs) to understand evidence and verify the truthfulness of claims. This framework significantly improves the performance of open-source text-only and vision-language LLMs in zero-shot fact-checking tasks. Empirical results demonstrate that M-RAV enables open-source LLMs to surpass baseline models and achieve performance comparable to state-of-the-art (SOTA) models, such as commercial LLMs and fine-tuned models on specific datasets. Moreover, M-RAV proves effective in practical fact-checking systems by automatically retrieving evidence, eliminating the need for manually provided ground-truth evidence. These findings highlight M-RAV's potential for robust multimodal fact-checking in real-world scenarios.

Nevertheless, error analysis reveals that LLMs struggle to verify claims labeled as "not enough information" (NEI), even when sufficient evidence is provided. This challenge arises because LLMs often fail to interpret latent social factors, such as cultural or political contexts, which are critical for accurate verification. To address this limitation, our future work will explore techniques like MemoryLLM [151] and MemoryBank [153] to dynamically update LLMs' knowledge bases with external environmental information or iteratively validate new external knowledge like LLatrieval [152]. These advancements aim to enhance LLM performance in fact-checking and other complex tasks requiring contextual understanding.

4.7 Summary

This chapter presents M-RAV (Multimodal Retrieve-Augment-Verify) for a multimodal Fact-checking system. M-RAV demonstrated a potential performance in practical application since it can process system-retrieved evidence efficiently for verifying the truthfulness of the claim and attains competitive results with the gold evidence approach. However, since M-RAV is implemented on large-scale LLMs (e.g, 70B and 72B of parameter size), it poses a challenge for the practical deployment of an end-to-end system. Hence, the next Chapter introduces the methodology for constructing FALCON—an end-to-end Fact-checking system based on LLMs that can be entirely hosted on a single GPU and has no external API dependencies.

Declaration: Parts of this chapter have been published in Publication [4].

Chapter 5

End-to-end Multimodal Fact checking System Construction

5.1 Introduction

The rapid spread of misleading information online has become a critical societal challenge. Manual verification of such content is infeasible due to the sheer volume, velocity, and multimodal diversity of digital information. This necessitates automated systems capable of verifying the truthfulness of claims at scale. Automated Fact-checking (AFC), grounded in Natural Language Processing (NLP), offers a promising solution by modeling verification as a machine learning task [4].

As noted in [4, 5, 8], the performance of AFC models critically depends on high-quality labeled datasets. However, human-annotated data for fact-checking is expensive and time-intensive to produce, motivating the need for systems that can operate effectively under data-scarce conditions [9]. Large Language Models (LLMs), trained on vast corpora with billions of parameters, have demonstrated remarkable ability to infer truthfulness from linguistic patterns, making them compelling candidates for AFC [154]. Recent approaches such as ProgramFC [9] and QAChecks [101] leverage proprietary LLMs like GPT-4 or InstructGPT via external APIs to perform fact-checking. Yet, this reliance on third-party services introduces significant limitations in scalability, cost, and control [101].

According to [155], Fact verification is a task that automatically predicts the truthfulness of the claim based on the collected evidence. On the other hand, as defined in the work [4], claim verification is a process in the fact-checking task to verify whether the claim is correct or not. It consists of two parts: predicting the verdict of the claim by providing truthfulness labels, and justifying the verdict by giving a sufficient explanation. Both these tasks play a vital role in the end-to-end fact-checking pipeline. For the claim verification, the MCVE that was demonstrated in Chapter 2 indicates the efficiency of image and text concatenation in verdict prediction and justification explanation. Instead, the ZeFAV and TabV4FC, which were introduced in Chapter 3 and Chapter 4 shows the robust ability of LLMs in verifying the claim based on given evidence. Also, the M-RAV acquired the LLM's ability to perform fact-verification under the system-evidence settings, showing the potential of LLMs' involvement in practical fact verification. However, the current proposed LLM-based method does not consult with the claim verification yet, where the justification task is lacking.

To address these challenges, we propose **FALCON**—an end-to-end, open-source, and locally deployable multimodal fact-checking system that eliminates dependence on external APIs. **FALCON** operates through four integrated stages: (1) retrieval of multimodal evidence (text and images), (2) interpretability augmentation through explanatory reasoning, (3) truth verification grounded in the retrieved evidence, and (4) justification of the final claim's veracity. For evidence retrieval, we utilize the curated database from Snopes and Politifact, as provided in the MOCHEG dataset [26].

FALCON employs three lightweight, open-source Vision Large Language Models (VLLMs) and Large Language Models (LLMs)—each with approximately 4 billion parameters—to perform end-to-end multimodal fact-checking. To enhance performance on downstream tasks, we fine-tune these models using a combination of human-labeled data and synthetic datasets generated by larger-scale VLLMs. These fine-tuned models are applied to each stage of the pipeline: evidence augmentation, verification, and justification.

We evaluate **FALCON**'s performance on two annotated multimodal fact-checking benchmarks: MOCHEG [26] and FINFACT [27], measuring accuracy in both claim verification and justification tasks.

5.2 System Design

This section provides a detailed description of the FALCON system's architecture. FALCON consists of four modules corresponding to four stages: Retrieval, Augmentation, Verification, and Justification. Figure 5.1 illustrates the overview of the FALCON architecture.

Retrieval stage This stage employs CLIP [37] to retrieve the top-5 relevant images, and BGE-M3 [141] to retrieve the top-5 relevant documents based on users' claims as queries. For the retrieved document, we employed an encoder-decoder model to summarize the key information from the document based on the claim. According to the Mocheg dataset [26], each piece of evidence is an excerpt from the original document used to verify the

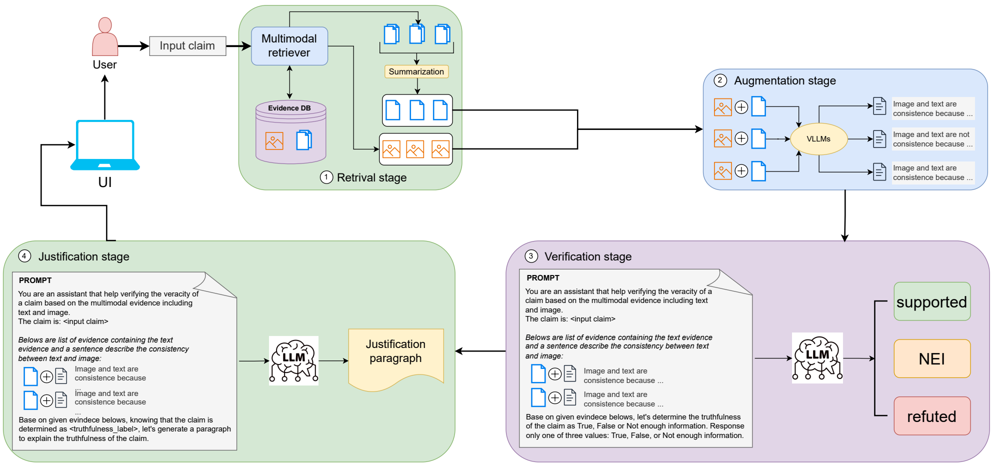


Figure 5.1: Overview of the FALCON system.

claim. Therefore, we leverage the sample in Mocheg to fine-tune the encoder-decoder models as a summarization step to reduce the length of the original document. We use the BART [80] and T5 [55] models for constructing the summarization module.

Augmentation stage This stage aims to verbalize the consistency between the text and image in pieces of retrieved evidence in order to instruct the LLMs in interpreting the evidence and verifying the truthfulness of the claim. In this stage, we employ a light vision–Qwen2.5VL-3B-Instruct [156] LLM to generate an explanation about the consistency between text and image. To better instruct the Qwen2.5VL-3B-Instruct for this stage, we use LLama3.2-Vision-90B-Instruct [121] to generate the consistency explanation between text and image evidence as synthesis data. Then, we use the generated data to fine-tune Qwen2.5VL-3B-Instruct and distill knowledge from a large model to a small one via Supervised Fine-Tuning (SFT). The prompt template for fine-tuning and inference is described in Appendix C. Since we have top-5 relevance documents and top-5 relevance images from the retrieval step, the total combination when performing augmentation is 25, which could exceed the LLM's capacity and increase computational cost. Therefore, we select the top 5 most relevant evidence based on the similarity between the generated explanation of consistency and the claim computed by BGE-M3 [141].

Verification stage With the evidence and augmentation explanation from the retrieval and augmentation steps, we use them for verifying the claim's

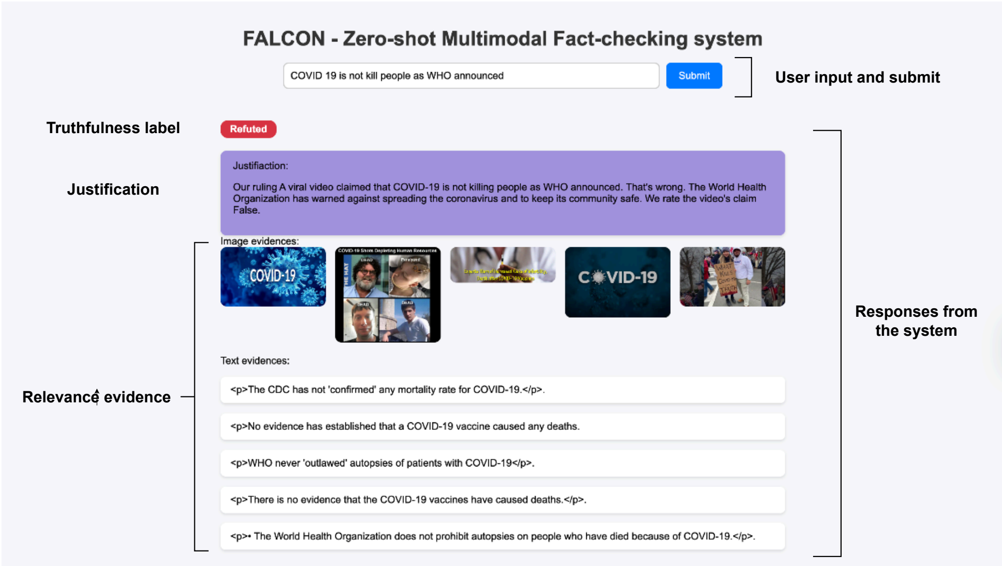


Figure 5.2: The UI of the FALCON system. Users input the claim and click the submit button. The system then performs checking the veracity and returns the label, the justification, and relevant text and image evidence for verifying the claim.

veracity via prompt instruction. We employ two small models: Qwen2.5VL-3B-Instruct [156] as VLLM and Qwen3-4B-Instruct-2507 [157] in this stage. To better train the LLMs to perform verification, we fine-tune two small LLMs using SFT with supervised samples from Mocheg [26] and FINFACT [27] with the prompt template as described in Appendix C.

Justification stage This stage provides an explanation in natural language as justification for the veracity of the claim based on the evidence and the truthfulness label. Similar to the Verification stage, we employ two small models: Qwen2.5VL-3B-Instruct [156] as VLLM and Qwen3-4B-Instruct-2507 [157] with prompt instruction for generating the justification sentence after fine-tuning the two models using SFT according to supervised samples from Mocheg [26] and FINFACT [27]. The prompt template is described in Appendix C.

The **FALCON** can be hosted on a single NVIDIA A6000 GPU after quantization to 4 bits. Additionally, Figure 5.2 shows an example of the user interface (UI) of **FALCON** for user interaction.

5.3 Experiment preparation

Table 5.1 shows the details about the training dataset used to construct the baseline for FALCON. Besides, we run the SFT tuning with the hyperparameters as follows:

Augmentation stage We run the SFT on *Qwen2.5-VL-3B-Instruct* with 5 epoches and batch_size equals 2. We run the inference mode with do_sample is False and max_new_token equals 2,048.

Verification stage We run the SFT on *Qwen2.5-VL-3B-Instruct* with 5 epoches and batch_size equals 2 and *Qwen3-4B-Instruct-2507* with 10 epoches and batch_size equals 1. We run the inference mode with do_sample is False and max_new_token equals 10.

Justification stage We run the SFT on *Qwen2.5-VL-3B-Instruct* with 10 epoches and batch_size equals 2 and *Qwen3-4B-Instruct-2507* with 15 epoches and batch_size equals 1. We run the inference mode with do_sample is False and max_new_token equals 2,048.

	MOCHEG	**FINFACT**	**Total**
train	7,319	1,123	8,442
dev	416	94	510
test	1,377	436	1,813

Table 5.1: Overview of the dataset used for training and evaluation in FALCON.

All the LLMs are run by 4-bit quantization so that FALCON can be hosted entirely on a single NVIDIA A6000 GPU.

5.4 Results

5.4.1 Performance of the summarization step

We conducted an experiment on the performance of two encoder-decoder models, including T5 [55] and BART [80], for summarizing the text evidence. The BART and T5 models are both fine-tuned on the training set of Mocheg to summarize the main content of the evidence that is used to determine the truthfulness of the claim. According to Table 5.2, the fine-tuned BART model achieved better results than T5 on the summary task across both Mocheg sets. Therefore, we employ fine-tuned BART for summarization in FALCON (as illustrated in the retrieval stage in Figure 5.1).

	Development set			
	BLEU	**ROUGE**	**METEOR**	**BertScore**
T5 (base)	56.24	77.95	68.30	93.59
BART (base)	70.28	81.90	80.59	95.81
	Test set			
T5 (base)	63.97	84.69	74.60	94.77
BART (base)	79.93	88.77	87.93	97.20

Table 5.2: Performance of summarization model on the development and test sets of Mocheg dataset.

5.4.2 Performance with gold-evidences

This experiment aims to investigate the performance of various LLMs and VLLMs, as well as the efficiency of a fine-tuned model for integration into the FALCON system. Specifically, we compare the performance of the FALCON system with gold-standard samples from the Mocheg [26] and FINFACT [27] datasets across two stages: verification and justification. For the verification stage, we evaluate performance using micro and macro-F1 scores based on the dataset's gold classes. For the justification stage, we employ four metrics —ROUGE [130] (including 1-gram, 2-gram, and Longest Common Subsequence configurations), BLEU [128], METEOR [129], and BERTScore [131] — to evaluate the match between the system-generated justification and the gold justification.

According to Table 5.3, our fine-tuned models in FALCON outperformed the original models. Even with larger models like GPT-oss-20B or Qwen3VL-32B-Instruct, the fine-tuned models with only 3B (Qwen2.5VL-Instruct-3B) and 4B (Qwen3-4B-Instruct-2507) achieve better performance on both micro and macro F1 scores, indicating the efficiency of fine-tuning smaller models for the verification task. Although the performance of fine-tuned models on Mocheg is slightly lower than that of Qwen2.5VL-Instruct-7B, they show a significant improvement on the FIN-FACT dataset. Overall, the fine-tuned model with a small size of about 3-4B performs better than the original models with large parameter sizes, indicating greater efficiency in practical implementation. As with the verification task, the results of the fine-tuned 3B and 4B models in the justification task are also significantly better than those of other original models, as illustrated in Table 5.4, even with large models like GPT-oss-20B or Qwen3VL-32B-Instruct. In general, the fine-tuning stage can improve performance on small models by distilling knowledge from human-labeled data.

	MOCHEG		FINFACT	
	F1 Micro	**F1 macro**	**F1 Micro**	**F1 macro**
Qwen2.5VL-Instruct-3B (fine-tuned)	39.21	34.58	63.99	50.05
Qwen3-4B-Instruct-2507 (fine-tuned)	40.08	35.84	**66.97**	**57.95**
Qwen2.5VL-Instruct-3B	37.54	21.53	30.50	23.99
Qwen2.5VL-Instruct-7B	**42.99**	**40.51**	44.95	42.23
Qwen3-4B-Instruct-2507	36.89	18.46	24.31	13.83
Qwen2.5-Instruct-7B	37.61	23.05	25.91	21.14
GPT-oss-20B	36.89	17.96	23.62	12.73
Qwen3VL-Instruct-32B	36.89	17.96	23.62	12.73

Table 5.3: Results of FALCON on verification stage with gold-evidence

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore
			MOCHEG			
Qwen2.5VL-Instruct-3B (fine-tuned)	**44.42**	**14.07**	**37.93**	**25.42**	**40.03**	**87.60**
Qwen3-4B-Instruct-2507 (fine-tuned)	**44.09**	**14.07**	**37.93**	**25.12**	**38.51**	**87.52**
Qwen2.5VL-Instruct-3B	11.23	1.14	10.29	7.43	10.21	79.68
Qwen2.5VL-Instruct-7B	12.23	4.68	7.85	2.74	9.81	77.97
Qwen3-4B-Instruct-2507	15.58	7.19	10.50	3.80	20.08	81.59
Qwen2.5-Instruct-7B	21.62	9.74	13.63	6.21	22.96	82.04
GPT-oss-20B	1.11	0.01	1.02	0.40	1.46	70.62
Qwen3VL-Instruct-32B	13.63	4.15	10.82	5.56	13.71	80.28
			FINFACT			
Qwen2.5VL-Instruct-3B (fine-tuned)	**42.27**	**22.21**	**26.33**	**17.32**	**32.28**	**84.98**
Qwen3-4B-Instruct-2507 (fine-tuned)	**44.73**	**26.25**	**30.62**	**19.29**	**32.25**	**85.88**
Qwen2.5VL-Instruct-3B	4.17	1.68	2.63	0.09	1.65	76.98
Qwen2.5VL-Instruct-7B	13.83	5.71	8.46	1.05	6.34	78.97
Qwen3-4B-Instruct-2507	32.92	15.35	19.65	9.46	24.24	82.50
Qwen2.5-Instruct-7B	23.36	12.24	14.46	5.72	12.71	80.66
GPT-oss-20B	1.16	0.02	1.01	0.05	1.13	71.02
Qwen3VL-Instruct-32B	6.38	2.21	3.09	0.96	3.57	75.42

Table 5.4: Results of FALCON on justification stage with gold-evidence

5.4.3 Performance in practical system

This experiment illustrates the practical performance of the FALCON in a real scenario, including retrieving relevant pieces of image and text evidence, augmenting the evidence, verifying the claim, and providing justification. We take 100 samples from the test sets of the Mocheg [26] and FINFACT [27] datasets for evaluation. In FALCON, two fine-tuned LLMs are employed, including the *Qwen3-4B-Instruct-2507* as text LLM and *Qwen2.5VL-Instruct-3B* as vision LLM. As shown in Table 5.5, the text LLM outperforms the vision LLM by 40.34% in macro F1 on Mocheg and 44.28% on FINFACT in the practical scenario. For the justification task, as shown in Table 5.6, the vision LLM seems slightly better than the text LLM on the Mocheg dataset. In contrast, the text LLM shows significantly better results than the vision LLM on the FINFACT dataset for the justification task. The text LLM performs better on verification and justification than the vision LLM in the FALCON system.

	MOCHEG		FINFACT	
	F1 Micro	**F1 macro**	**F1 Micro**	**F1 macro**
FALCON (vision LLM)	36.00	27.92	50.00	35.56
FALCON (text LLM)	**44.00**	**40.34**	**54.00**	**44.28**

Table 5.5: Performance of FALCON in a practical scenario for verification with 100 samples on each Mocheg and FINFACT datasets

	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BLEU**	**METEOR**	**BERTScore**
	MOCHEG					
FALCON (vision LLM)	**34.23**	6.15	**30.94**	**22.12**	**31.80**	**84.96**
FALCON (text LLM)	32.52	**6.64**	29.13	20.41	30.09	84.56
	FINFACT					
FALCON (vision LLM)	1.85	0.28	1.52	0.84	1.30	74.39
FALCON (text LLM)	**10.41**	**2.11**	**5.42**	**1.37**	**5.75**	**76.37**

Table 5.6: Performance of FALCON in a practical scenario for justification with 100 samples on each Mocheg and FINFACT datasets

	# Empty	**Avg. length**
MOCHEG (vision LLM)	63	34.35
MOCHEG (text LLM)	61	29.05
FINFACT (vision LLM)	93	6.92
FINFACT (text LLM)	68	99.56

Table 5.7: Statistics of generated justification by the FALCON. The length of justification is computed at the token level.

	Retrieval (s)	**Augmentation (s)**	**Verification (s)**	**Justification (s)**	**Total (s)**	**Time per samples (s)**
MOCHEG (vision LLM)	532.51	25,347.64	241.50	446.57	26,568.23	265.68
MOCHEG (text LLM)	536.21	25,622.38	62.19	248.02	26,468.82	264.68
FINFACT (vision LLM)	622.81	26,706.93	258.18	297.23	27,885.16	278.85
FINFACT (text LLM)	619.35	26,311.35	73.09	733.17	27,736.97	277.36

Table 5.8: Running time evaluation on 100 samples

Additionally, we conduct an experiment to evaluate FALCON's runtime. We measure the execution time of each of the four stages in FALCON and report the time required to run 100 samples on Mocheg and FINFACT. Table 5.8 shows the total running time for retrieval, augmentation, verification, and justification stages, and the total time and average time for one sample. Overall, the augmentation stage takes a long time to run, since it tries to generate consistency between a pair of retrieved text and image to align the evidence (we have the top 5 images and texts, respectively, so the model has to generate 5x5=25 samples). During the verification stage, the text LLM demonstrates significantly lower execution time than the vision LLM. Compared with the accuracy performance in Table 5.5, the text LLM demonstrates its efficiency in the verification stage of FALCON. For the justification task, the runtime of the text LLM on Mocheg is less than that of the vision LLM, while it takes more time to run with the text LLM than the vision LLM on the FINFACT dataset. To investigate the behavior of justification modules in this case, Table 5.7 briefly illustrates the statistics of the generated justification by FALCON. According to Table 5.7, it can be seen that the number of empty justifications and the length of the justifications are similar in Mocheg between the text and vision LLMs. On the other hand, the number of empty justifications generated by vision LLM on FINFACT is significantly more than that of text LLM, and the average length of the generated justification by vision LLM is also dramatically shorter than that of text LLM. Since the number of empty justifications by vision LLM takes a large proportion, the performance of vision LLM is worse than text LLM in FALCON (as illustrated in Table 5.6), although it takes less running time than text LLM, as shown in Table 5.8

5.5 Summary

We introduces **FALCON** - an end-to-end Multimodal Fact-checking system that follows the practical Fact-checking pipeline: retrieve relevant evidence, verify the information, and justify the veracity of the information. Through the experimental results on both gold and system-retrieved evidence, **FALCON** demonstrates exemplary performance in a balance between the accuracy and the computational cost. As an open-source tool that does not depend on external APIs, **FALCON** is suitable for improving accuracy and runtime, as well as for practical deployment. Finally, **FALCON** aims to support communities in fighting misinformation and verifying information.

Although **FALCON** shows potential in practical deployment with entirely locally hosted systems, there are two main challenges for the system:

Inference time execution As shown in the Experimental section, one sample takes approximately 4.5 minutes to run, specifically in the augmentation stage. This will help the system handle increased user numbers in the future. Possible methods for improving the execution time are SPECTRA [158] – a speculative decoding framework and VLLM [159].

Up-to-date evidence database **FALCON** based on the evidence database from Mocheg [26], which is constructed on Snope and Politifact data around 2023. To maintain an up-to-date database, the system may need to crawl new data, which can leverage the crawling script available at https://github.com/UKPLab/conll2019-snopes-crawling.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

Automatically verifying information on social media is a challenge in the digital era, where vast amounts of information exist in diverse modalities. To assist the automatic fact-checking process, this thesis presents research on the end-to-end Fact-checking problem. In summary, the main findings in this thesis are:

- First, the thesis introduces MCVE - an approach that combines the features of multiple texts and images via a fusion technique to train a transformer model for verifying the truthfulness of the claim and to integrate information from textual and visual data into encoder-decoder models for explanation. MCVE demonstrates its efficiency in handling multiple evidence and no-evidence claims in the verification and explanation task. The main challenge in MCVE is the requirement for a human-annotated dataset for training the deep neural network, which limits practical performance when there are fewer or even no datasets, and the alignment across modalities to exploit latent information.
- Second, the thesis proposed three frameworks that leverage LLMs' robustness in language understanding and reasoning for fact-checking tasks, including ZeFAV, TabV4FC, and M-RAV. These frameworks proposed an efficient prompt-based technique that concatenates and integrates multimodal data to verify the information in the claim. Since this framework uses zero-shot learning, which requires no training data, it is flexible to implement in practice. However, the LLMs - especially large-sized LLMs require much computational resources, which is a challenge when running on real systems.
- Third, the thesis introduces FALCON. This end-to-end Fact-checking system employs the whole pipeline: it retrieves the evidence, verifies the claim based on that evidence, and explains the verification process as justification for the claim's veracity. FALCON implements small LLMs that can be hosted and run inference on a single GPU. To train

small LLMs for the Fact-checking task, FALCON distilled knowledge from large LLMs and available annotated data, and each small LLM was trained on the task. FALCON demonstrates the potential for a practical application system. Nevertheless, FALCON needs a technique to frequently update the knowledge in the evidence database to keep up with the latest events and expertise in the real world, and a method to improve LLM inference time.

6.2 Future Works

Future work of this thesis focuses on two aspects. For research purposes, the following study aims to improve the ability to automatically update knowledge from the real world by proposing a technique for continuously updating evidence databases or transferring the latest expertise to LLMs via knowledge elicitation [160]. Besides, Fact-checking also needs to consider temporal information via temporal representation to capture the correct event that happened, thereby varying the veracity of the information. Additionally, for practical implementation, further studies should focus on improving the system's efficiency by increasing LLM inference speeds to enhance user experience in system interaction. Last but not least, the system should integrate human-in-the-loop into the Fact-checking process by implementing reinforcement learning to capture user behavior and boost the performance of the Fact-checking system.

Appendix A

Prompt for LLMs with M-RAV

A.1 Prompt for text-only LLMs with sufficient text and image

This sample is from FACTIFY with *llama3.1-70B*.

```
Is it true that: The Tripura government will promote all students from
    Class 1 to 4 and Class 6 and 7 without examinations this year\n\n(
    @PriyankaDebBarm reports)\n\nhttps://t.co/DOtjACdhbB https://t.co/
    m4eVCEMFtM?
The evidence:
  Evidence 1:
  Text: The Tripura government will promote all students from Class 1 to
      4 and Class 6 and 7 without examinations this year. \"We have
      decided to promote students from Class 1 to 4, 6 and 7 to next
      classes except Classes 5 and 8. But the students need to sit for
      their examinations after opening of schools if the situation
      becomes normal. The exams would be held to evaluate their
      educational loss in the Covid-19 pandemic\", said Education
      Minister Ratan Lal Nath. Regarding Class 5 and 8, Nath said that
      they would send the matter for approval from the state cabinet and
      for Class 9 and 11 examinations, decisions would be taken following
       discussions with the Tripura Board of Secondary Education. The
      summer vacations in schools have been extended till June 6 due to
      the pandemic. The government also decided to declare 20 different
      schools across the state as specified category schools, where
      students from any corner of the state could access scope to study
      only after they clear screening tests and lottery system. Selection
       of teachers and headmasters for these schools would be made
      through special recruitment exams.

  Image: The image shows a group of school girls reading books, which is
      consistent with the text that discusses the promotion of students
      from Class 1 to 4 and Class 6 and 7 without examinations. The image
       also shows the girls wearing school uniforms, which is consistent
      with the text that mentions the extension of summer vacations in
      schools. The image is consistent with the text, as it shows a group
       of school girls reading books and wearing school uniforms, which
```

```
      aligns with the discussion of promoting students without
      examinations and extending summer vacations in schools. The image
      effectively represents the theme of education and learning, which
      is the central topic of the text. Overall, the image is a fitting
      representation of the text's content.
  Consistency: The image is consistent with the text.
  Relevance score: 0.7499760389328003

To verify the truthfulness of the claim, please following these steps:
    STEP 1: Consult the relevance between the claim and each given
        evidence based on the relevance score.

  STEP 2: Think and conclude the truthfulness of the claim based on the
      relevance and logical of the evidence. If the evidence does not
      help concluding the claim is supported or refuted, it may be not
      enough information.

The truthfulness must be only one of three value: supported, refuted, or
    not enough information. Please think step-by-step carefully and
    response only the truthfulness of the claim.

<RESPONSE>:
```

A.2 Prompt for text-only LLMs without image evidence

This sample is from Mocheg with *mixtral-8x-7b*.

```
Is it true that: Image depicts a huge, inexplicably half-eaten shark
    found on a Florida beach.?
The evidence:
  Evidence 1:
  Text: People taking surf lessons on Butler Beach were in a for big
      surprise when they entered the water Saturday. At around 12:30 p.m
      ., Dakota Dodson, a surf instructor for the St. Augustine Surf
      School, noticed a fin sticking out of the water and went to
      investigate. 'At first, I only saw the fin, so I just assumed it a
      live shark. But then I realized it was only half of one,' he said.
      The half-eaten carcass, measuring about 2.5 feet, was found
      floating in knee-deep water, Dodson said. The circumstances of how
      the shark died remain unclear, but experts believe it was likely
      eaten by another shark. 'It's most likely it was bitten in half by
      another shark - sharks have to eat, too,' said Tara Dodson,
      environmental supervisor for St. Johns County.
  Relevance score: 0.5962506532669067
```

```
To verify the truthfulness of the claim, please following these steps:
  STEP 1: Consult the relevance between the claim and each given evidence
       based on the relevance score.
  STEP 2: Think and conclude the truthfulness of the claim based on the
      relevance and logical of the evidence. If the evidence does not
      help concluding the claim is supported or refuted, it may be not
      enough information.
The truthfulness must be only one of three value: supported, refuted, or
    not enough information. Please think step-by-step carefully and
    response only the truthfulness of the claim.
<RESPONSE>:
```

A.3 Prompt for vision-language LLMs

This sample is from FINFACT with *llava1.6-34B*. in the Tokenizer, the LLaVA accepts multiple image as list of sequence since the *¡image¿* tokens in the prompt corresponds to each image in the evidence.

```
Is it true that: 'So Far No One Has Found Another Number' Walmart Gift
    Card Facebook Scam? Here are the evidence for checking:
  Evidence 1 :
    Text: In late July 2023, readers pointed us to several posts that had
         been made inside of Facebook groups. The users who made these
        posts, whose profiles all indicated they may have been from
        Bangladesh, claimed to be offering Walmart gift cards to anyone
        who could find a special number in an image. For example, one
        post read, \"So Far No One Has Found Another Number Apart From
        86, No Winners Yet (Walmart Gift Card). We Still Have 24 More
        Wins. \"\n
    Image: <image>
    Description: The image is consistent with the text, as it shows a
        hand holding a Walmart gift card and a grid of numbers, which
        aligns with the text's mention of a Walmart gift card and finding
         a number. The image and text also have a similar tone and style,
         with a playful and promotional feel. Overall, the image
        effectively supports the text and helps to convey the message in
        a visually engaging way.
    Consistency: The image is consistent with the text.
    Relevance score: 0.6894003748893738
  Evidence 2 :
    Text: Survey scam websites usually promise cash prizes, pricey
        electronics, and other interesting purported \"rewards, \" all
        supposedly if the user takes a few minutes to answer some
        questions. However, as we've reported for the last two decades or
         so, survey scam websites have historically proven to be a waste
        of time. They often ask users to provide personal and financial
        information on various websites, as well as to sign up for trials
```

```
     of unfamiliar streaming services. All of this appeared to be an
    attempt at receiving affiliate-marketing commission based on the
    amount of information given away when providing personal and
    financial data to these websites.
  Image: <image>
  Description: The image is consistent with the text, as it shows a
    Facebook post with a photo of a Walmart gift card and a caption
    that reads, \"So Far No One Has Found Another Number Apart From
    86, No Winners Yet (Walmart Gift Card). We Still Have 24 More
    Wins. \" The image also includes a yellow circle with a white \"S
    \" inside it, which is likely a logo for a scam alert. This
    suggests that the image is a screenshot of a Facebook post that
    is warning people about a scam. Therefore, the image is
    consistent with the text.
  Consistency: The image is consistent with the text.
  Relevance score: 0.6600311994552612
Evidence 3 :
  Text: Sometimes, these kinds of scammers might instead provide a link
     sending users to hidden subscription scams that supposedly offer
     \"free\" prizes. However, such scams hide monthly fees in the
    fine print, much like a Cash App scam we once reported about.
  Image: <image>
  Description: The image is consistent with the text, as it shows a
    Facebook post from Isabelle Freya, a scammer, and the text is a
    scam. The image and text are consistent because they both have
    the same scammer and the same scam. This suggests that the image
    and text are likely to be from the same source and are intended
    to deceive users. The consistency between the image and text
    highlights the importance of being cautious when encountering
    suspicious content online. It is essential to verify the
    authenticity of information and be aware of potential scams to
    avoid falling victim to fraudulent activities. By recognizing the
     consistency between the image and text, users can take steps to
    protect themselves from scams and maintain a safe online
    experience.
  Consistency: The image is consistent with the text.
  Relevance score: 0.5112748146057129
Evidence 4 :
  Text: These sorts of scammers might also sometimes direct users to
    phishing websites that claim a gift card or other prize could be
    ordered for \"free, \" only with a small shipping and handling
    charge. Of course, there would be no real gift card or other
    prize. This simply would be an attempt to obtain a victim's
    financial information for criminal activities, such as a credit
    card number or PayPal login. This kind of a scam was similar to
    another one we previously reported about concerning the U. S.
    Postal Service.
  Image: <image>
  Description: The image is consistent with the text, as it shows a
```

```
        Facebook post from Isabelle Freya that matches the text in the
        image. The image also shows a hand holding a Walmart gift card
        and a grid of numbers with the number 86 repeated multiple times,
         which is consistent with the text in the post. Overall, the
        image appears to be a screenshot of the Facebook post, and it is
        consistent with the text.
    Consistency: The image is consistent with the text.
    Relevance score: 0.6907112002372742
  Evidence 5 :
    Text: If readers are looking for legitimate promotions for Walmart
        gift cards, we recommend our previous reporting that found the
        company truly does give away $80,000 in prizes every three months
        . Such promotions are offered by the company in official email
        correspondence and on receipts handed out in its brick-and-mortar
         stores.
    Image: <image>
    Description: The image is consistent with the text, as it shows a
        Facebook post from Isabelle Freya that mentions a Walmart gift
        card and the number 86, which is also depicted in the image. The
        image also includes a yellow \"SCAM\" logo, which suggests that
        the post may be a scam. Overall, the image and text are
        consistent in their content and message.
    Consistency: The image is consistent with the text.
    Relevance score: 0.6813938617706299
To verify the truthfulness of the claim, please following these steps:
  STEP 1 : Consult the relevance between the claim and each given
      evidence based on the relevance score.
  STEP 2 : Think and conclude the truthfulness of the claim based on the
      relevance and logical of the evidence.
The truthfulness must be only one of three value: true, false, or not
    enough information. Please think step-by-step carefully and response
    only the truthfulness of the claim.
<RESPONSE>:
```

A.4 Prompt for vision-language LLMs to generate an explanation for the consistency between text and image evidence

This sample is from the Mocheg dataset for generating the consistency explanation between the text and image evidence. The *¡image¿* token in the prompt corresponds to the image evidence. We used the *Llama-3.2-90B-Vision-Instruct* for generation.

```
<|image|>
```

```
Agriculture Secretary Tom Vilsack today released the following statement
     regarding the language in the omnibus bill repealing the country of
     origin labeling requirements for beef and pork products. 'The
    omnibus bill repealed the country of origin labeling (COOL)
    requirements for muscle cuts of beef and pork, and ground beef and
    pork. Effective immediately, USDA is not enforcing the COOL
    requirements for muscle cut and ground beef and pork outlined in the
     January 2009 and May 2013 final rules.' USDA will be amending the
    COOL regulations as expeditiously as possible to reflect the repeal
    of the beef and pork provisions. In addition, all imported and
    domestic meat will continue to be subject to rigorous inspections by
     USDA to ensure food safety.

Please generate a short paragraph describing the about the consistency
    of the image based on the given text following this template:
<HYPOTHESIS>: Please determining whether the image is consistent with
    the text or not.
<EXPLANATION>: Explanation the aligment between the image hypothesis and
     the text.
<FINAL ANSWER>: Give one paragraph describing the consistency of the
    image and text based on the explanation.

<RESPONSE>
```

Appendix B

Prompt for LLMs on TabV4FC

B.1 Prompt with Qwen2.5-72B-Instruct for TabFACT

```
You are an assistant that help to verify the claim.
The claim is: the yugoslavian national team fail to score only 1 time ,
    drop a world cup qualify match 2:1 against denmark
The table that containing the information for verifying the claim:

+----+--------------+---------------------+------------+-----------
| | date | city | opponent | results | type of game |
+====+==============+=====================+============+===========
| 0 | march 22 | sarajevo | uruguay | 2:1 | friendly |
+----+--------------+---------------------+------------+-----------
| 1 | march 30 | belgrade | romania | 2:0 | balkan cup |
+----+--------------+---------------------+------------+-----------
| 2 | april 26 | borovo | poland | 2:1 | friendly |
+----+--------------+---------------------+------------+-----------
| 3 | august 27 | bucharest , romania | romania | 1:4 | balkan cup |
+----+--------------+---------------------+------------+-----------
| 4 | september 10 | luxembourg | luxembourg | 5:0 | 1982 wcq |
+----+--------------+---------------------+------------+-----------
| 5 | september 27 | ljubljana | denmark | 2:1 | 1982 wcq |
+----+--------------+---------------------+------------+------------+
| 6 | november 15 | torino , italy | italy | 0:2 | 1982 wcq |
+----+--------------+---------------------+------------+-----------

Summarization of table: The claim made by the table is that the
    Yugoslavia national team failed to score only one time in their 1982
    World Cup Qualifying match against Denmark. The result of the match
    was 2-1 in Ljubljana.

Based on the table and the summarization, please think and determine the
    truthfulness of the claim. The truthfulness must be one of these
    values: entailed or refuted.
<RESPONSE>:
```

B.2 Prompt with Qwen2.5-72B-Instruct for SCITAB

```
You are an assistant that help to verify the claim.
The claim is: Table 4: Comparison of per-document accuracy (% ) by
    different systems for top 1, 3 and 5 words of abstractive sentences.
The table that containing the information for verifying the claim:
+----+----------------------------+----------+-------+-------+-------+
| | System | Reward | R-1 | R-2 | R-L |
+====+============================+==========+=======+=======+=======+
| 0 | Kryscinski e tal. ( 2018 ) | R-L | 40.2 | 17.4 | 37.5 |
+----+----------------------------+----------+-------+-------+-------+
| 1 | Narayan e tal. ( 2018b ) | R-1,2,L | 40 | 18.2 | 36.6 |
+----+----------------------------+----------+-------+-------+-------+
| 2 | Chen and Bansal ( 2018 ) | R-L | 41.5 | 18.7 | 37.8 |
+----+----------------------------+----------+-------+-------+-------+
| 3 | Dong et al. ( 2018 ) | R-1,2,L | 41.5 | 18.7 | 37.6 |
+----+----------------------------+----------+-------+-------+-------+
| 4 | Zhang et al. ( 2018 ) | [EMPTY] | 41.1 | 18.8 | 37.5 |
+----+----------------------------+----------+-------+-------+-------+
| 5 | Zhou et al. ( 2018 ) | [EMPTY] | 41.6 | 19 | 38 |
+----+----------------------------+----------+-------+-------+-------+
| 6 | Kedzie et al. ( 2018 ) | [EMPTY] | 39.1 | 17.9 | 35.9 |
+----+----------------------------+----------+-------+-------+-------+
| 7 | (ours) NeuralTD | Learned | 39.6 | 18.1 | 36.5 |
+----+----------------------------+----------+-------+-------+-------+
Summarization of table: Table 4 compares the results of different systems
     for top 1, 3 and 5 words of abstractive sentences. Our model (
    NeuralTD) shows the best performance among all of the comparison
    systems. It outperforms all of the state-of-the-art systems by
    meaningful margins in terms of ROUGE-1, ROUGE-2 and ROUGE-L. This
    indicates that our model captures important information contained in
    the sentences.

Based on the table and the summarization, please think and determine the
    truthfulness of the claim. The truthfulness must be one of these
    values: supported, refuted or not enough information.
<RESPONSE>:
```

B.3 Prompt with Qwen2.5-72B-Instruct for PubHealthTab

```
You are an assistant that help to verify the claim.
```

```
The claim is: Depending on the fasting process in a 24/7 food culture,
    the followers of periodic fasting increases.
The table that containing the information for verifying the claim:
+----+----------------------------------------------------+-------+
| | Planning a Fast | 16 |
+====+====================================================+=======+
| 1 | The Physiological Process of Fasting | 37 |
+----+----------------------------------------------------+-------+
| 3 | Complementary and Alternative Medicine and Fasting | 64 |
+----+----------------------------------------------------+-------+
| 5 | Research on Fasting and Mental Health | 86 |
+----+----------------------------------------------------+-------+
| 7 | Fasting and Transpersonal Psychology | 106 |
+----+----------------------------------------------------+-------+
| 9 | Fasting and Depression | 120 |
+----+----------------------------------------------------+-------+
| 11 | The Practice of Asceticism | 173 |
+----+----------------------------------------------------+-------+
| 13 | Fasting in Religious and Spiritual Traditions | 213 |
+----+----------------------------------------------------+-------+
Summarization of table: The claim makes use of the physiological process
    of fasting described in section 3.4 and applies to a wider range of
    domains, including complementary and alternative medicine and fasting
    . We argue that, in a cultural culture, the type of thinking about
    food that is related to fasting is more general and less restricted
    than that of a natural language processing application. We observe
    that the frequency of fasting increases in the 21st century, as
    people become more aware of the concept of fast and then consider
    taking a break from fasting during their religious and spiritual
    traditions.

Based on the table and the summarization, please think and determine the
    truthfulness of the claim. The truthfulness must be one of these
    values: supported, refuted or not enough information.
<RESPONSE>:
```

Appendix C

Prompt on FALCON

C.1 Augmentation Prompt

```
System prompt:
Please generate a short paragraph describing the about the consistency of
     the image based on the given text following this template:
    <HYPOTHESIS>: Please determining whether the image is consistent with
        the text or not.
    <EXPLANATION>: Explanation the aligment between the image hypothesis
       and the text.
    <FINAL ANSWER>: Give one paragraph describing the consistency of the
       image and text based on the explanation.
User prompt:
<input_image>
<input_text>
Assistant prompt (used for fine-tunning):
<sample_aligment>
```

C.2 Verification Prompt (vision LLM)

```
System prompt:
You are an assistant that help verifying the veracity of a claim based
    on the multimodal evidence including text and image.
     The claim is: <input_claim>
     Belows are list of evidence containing the image, the text
         evidence, and a sentence describe the consistency between text
         and image:
 User prompt: }
     [(<image_evidence>, <text_evidence>, <consistency_explanation>),
         ... ]
 System prompt:
 Base on given evindece belows, let's determine the truthfulness of the
      claim as True, False or Not enough information. Response only one
      of three values: True, False, or Not enough information.
 Assistant prompt (used for fine-tunning):
<truthfulness_label_of_claim>
```

C.3 Verification Prompt (text LLM)

```
You are an assistant that help verifying the veracity of a claim based
    on the multimodal evidence including text and image.
     The claim is: <input_claim>
     Belows are list of evidence containing the text evidence and a
         sentence describe the consistency between text and image:
 User prompt
       [(<text_evidence>, <consistency_explanation>), ... ]
 System prompt:
 Base on given evindece belows, let's determine the truthfulness of the
      claim as True, False or Not enough information. Response only one
      of three values: True, False, or Not enough information.
 Assistant prompt (used for fine-tunning)
<truthfulness_label_of_claim>
```

C.4 Justification Prompt (vision LLM)

```
 System prompt:
 You are an assistant that help verifying the veracity of a claim based
      on the multimodal evidence including text and image.
     The claim is: <input_claim>
     Belows are list of evidence containing the image, the text
         evidence, and a sentence describe the consistency between text
         and image:
 User prompt:
    [(<image_evidence>, <text_evidence>, <consistency_explanation>),
       ...]
  System prompt:
  Base on given evindece belows, knowing that the claim is determined
      as <truthfulness_label>, let's generate a paragraph to explain
      the truthfulness of the claim.
Assistant prompt (used for fine-tunning):
<example_justification_of_claim>
```

C.5 Justification Prompt (text LLM)

```
System prompt:
 You are an assistant that help verifying the veracity of a claim based
      on the multimodal evidence including text and image.
```

```
    The claim is: <input_claim>
    BBelows are list of evidence containing the text evidence and a
        sentence describe the consistency between text and image:
User prompt:
 [(<text_evidence>, <consistency_explanation>), ...]
System prompt:
 Base on given evindece belows, knowing that the claim is determined
     as <truthfulness_label>, let's generate a paragraph to explain
     the truthfulness of the claim.
Assistant prompt (used for fine-tunning):
<example_justification_of_claim>
```

References

[1] W. Ansar and S. Goswami, "Combating the menace: A survey on characterization and detection of fake news from a data science perspective," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100052, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096821000458

[2] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish *et al.*, "Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 913–922.

[3] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Jan. 2025. [Online]. Available: https://doi.org/10.1145/3703155

[4] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.11/

[5] N. Kotonya and F. Toni, "Explainable automated fact-checking: A survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5430–5443. [Online]. Available: https://aclanthology.org/2020.coling-main.474/

[6] C. Pulm, A. Gast, and J. Rummel, "A picture corrects a thousand words–the effect of photos on veracity feedback," *Consciousness and Cognition*, vol. 125, p. 103758, 2024.

[7] E. J. Newman, M. Garry, D. M. Bernstein, J. Kantner, and D. S. Lindsay, "Nonprobative photographs (or words) inflate truthiness," *Psychonomic Bulletin & Review*, vol. 19, no. 5, pp. 969–974, 2012.

[8] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, "Multimodal automated fact-checking: A survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5430–5448. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.361/

[9] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, and P. Nakov, "Fact-checking complex claims with program-guided reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6981–7004. [Online]. Available: https://aclanthology.org/2023.acl-long.386/

[10] A. M. Barik, W. Hsu, and M. L. Lee, "Chronofact: Timeline-based temporal fact verification," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, J. Kwok, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2025, pp. 8031–8039, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2025/893

[11] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Fact checking with insufficient evidence," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 746–763, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.43/

[12] J. Thorne and A. Vlachos, "Automated fact checking: Task formulations, methods and future directions," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3346–3359. [Online]. Available: https://aclanthology.org/C18-1283/

[13] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819. [Online]. Available: https://aclanthology.org/N18-1074/

[14] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3441–3460. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.309/

[15] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, "Feverous: Fact extraction and verification over unstructured and structured information," *arXiv preprint arXiv:2106.05707*, 2021.

[16] N. X. R. Wang, D. Mahajan, M. Danilevsky, and S. Rosenthal, "SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS)," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 317–326. [Online]. Available: https://aclanthology.org/2021.semeval-1.39/

[17] N. Kotonya and F. Toni, "Explainable automated fact-checking for public health claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online, Nov. 2020.

[18] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 422–426. [Online]. Available: https://aclanthology.org/P17-2067/

[19] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, "Tabfact: A large-scale dataset for table-based fact verification," *arXiv preprint arXiv:1909.02164*, 2019.

[20] X. Lu, L. Pan, Q. Liu, P. Nakov, and M.-Y. Kan, "SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association

for Computational Linguistics, Dec. 2023, pp. 7787–7813. [Online]. Available: https://aclanthology.org/2023.emnlp-main.483/

[21] M. Akhtar, O. Cocarascu, and E. Simperl, "PubHealthTab: A public health table-based dataset for evidence-based fact checking," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1–16. [Online]. Available: https://aclanthology.org/2022.findings-naacl.1/

[22] V. V, A. Anand, A. Anand, and V. Setty, "Quantemp: A real-world open-domain benchmark for fact-checking numerical claims," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 650–660. [Online]. Available: https://doi.org/10.1145/3626772.3657874

[23] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, and J. G. Simonsen, "MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4685–4697. [Online]. Available: https://aclanthology.org/D19-1475/

[24] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. N. Reganti, P. Patwa, A. Das, T. Chakraborty, A. P. Sheth, A. Ekbal *et al.*, "Factify: A multi-modal fact verification dataset." in *DE-FACTIFY@AAAI*, 2022.

[25] A. Rani, S. T. I. Tonmoy, D. Dalal, S. Gautam, M. Chakraborty, A. Chadha, A. Sheth, and A. Das, "FACTIFY-5WQA: 5W aspect-based fact verification through question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 421–10 440. [Online]. Available: https://aclanthology.org/2023.acl-long.581/

[26] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, "End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2733–2743. [Online]. Available: https://doi.org/10.1145/3539618.3591879

[27] A. Rangapur, H. Wang, L. Jian, and K. Shu, "Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation," in *Companion Proceedings of the ACM on Web Conference 2025*, ser. WWW '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 785–788. [Online]. Available: https://doi.org/10.1145/3701716.3715292

[28] S. Singhal, R. R. Shah, and P. Kumaraguru, "Factdrill: A data repository of fact-checked social media content to study fake news incidents in india," in *Proceedings of the international AAAI conference on web and social media*, vol. 16, 2022, pp. 1322–1331.

[29] D. S. Nielsen and R. McConville, "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*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3141–3153. [Online]. Available: https://doi.org/10.1145/3477495.3531744

[30] J. Chen, A. Sriram, E. Choi, and G. Durrett, "Generating literal and implied subquestions to fact-check complex claims," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3495–3516. [Online]. Available: https://aclanthology.org/2022.emnlp-main.229/

[31] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," *arXiv preprint arXiv:2104.03617*, 2021.

[32] M. Soldan, A. Pardo, J. L. Alcázar, F. Caba, C. Zhao, S. Giancola, and B. Ghanem, "Mad: A scalable dataset for language grounding in videos from movie audio descriptions," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2022, pp. 5026–5035.

[33] C. Wang, Y. Shen, Z. Kuang, A. Cohan, and Y. Zhao, "SciVer: Evaluating foundation models for multimodal scientific claim verification," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8562–8579. [Online]. Available: https://aclanthology.org/2025.acl-long.420/

[34] Y. Wang, L. Chen, Z. Qian, and P. Li, "Official-nv: An llm-generated news video dataset for multimodal fake news detection," *arXiv preprint arXiv:2407.19493*, 2024.

[35] F. Liu, Y. Yacoob, and A. Shrivastava, "COVID-VTS: Fact extraction and verification on short video platforms," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia, May 2023.

[36] S. Rani and M. Kumar, "Multi-modal topic modeling from social media data using deep transfer learning," *Applied Soft Computing*, vol. 160, p. 111706, 2024.

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[38] J. Zhou, Z. Liu, S. Xiao, B. Zhao, and Y. Xiong, "VISTA: Visualized text embedding for universal multi-modal retrieval," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3185–3200. [Online]. Available: https://aclanthology.org/2024.acl-long.175/

[39] M. Günther, S. Sturua, M. K. Akram, I. Mohr, A. Ungureanu, B. Wang, S. Eslami, S. Martens, M. Werk, N. Wang *et al.*, "jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval," *arXiv preprint arXiv:2506.18902*, 2025.

[40] G. Heinrich, M. Ranzinger, H. Yin, Y. Lu, J. Kautz, A. Tao, B. Catanzaro, and P. Molchanov, "Radiov2. 5: Improved baselines for agglomerative vision foundation models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 487–22 497.

[41] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[42] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[43] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[44] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[45] L. Zheng, C. Li, X. Zhang, Y.-M. Shang, F. Huang, and H. Jia, "Evidence retrieval is almost all you need for fact verification," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9274–9281. [Online]. Available: https://aclanthology.org/2024.findings-acl.551/

[46] S.-I. Papadopoulos, I. Beňová, S. Kula, M. Gregor, G. Karantaidis, T. Javůrek, M. Šimko, and S. Papadopoulos, "Multimodal and multilingual fact-checked article retrieval," in *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ser. ICMR '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1063–1071. [Online]. Available: https://doi.org/10.1145/3731715.3733402

[47] H. Gong, W. Xu, S. Wu, Q. Liu, and L. Wang, "Heterogeneous graph reasoning for fact checking over texts and tables," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 100–108.

[48] S. Duwal, M. N. S. Shopnil, A. Tyagi, and A. M. Proma, "Evidence-grounded multimodal misinformation detection with attention-based gnns," *arXiv preprint arXiv:2505.18221*, 2025.

[49] W. C. Sleeman, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *ACM Comput. Surv.*, vol. 55, no. 7, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3543848

[50] F. Zhang, J. Liu, Q. Zhang, E. Sun, J. Xie, and Z.-J. Zha, "Ecenet: Explainable and context-enhanced network for muti-modal fact verification," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1231–1240. [Online]. Available: https://doi.org/10.1145/3581783.3612183

[51] L. Hu, Z. Wang, J. Zhu, Y. Hu, and X. Wang, "Mage-fend: Multimodal adaptive fusion with guidance from llm expertise for fake news detection on short video platforms," *Knowledge-Based Systems*, p. 114298, 2025.

[52] S. Tahmasebi, E. Müller-Budack, and R. Ewerth, "Multimodal misinformation detection using large vision-language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 2189–2199. [Online]. Available: https://doi.org/10.1145/3627673.3679826

[53] J. Huang, M. Xiao, D. Li, Z. Jiang, Y. Yang, Y. Zhang, L. Qian, Y. Wang, X. Peng, Y. Ren *et al.*, "Open-finllms: Open multimodal large language models for financial applications," *arXiv preprint arXiv:2408.11878*, 2024.

[54] J. Lee, X. Lu, J. Hessel, F. Brahman, Y. Yu, Y. Bisk, Y. Choi, and S. Gabriel, "How to train your fact verifier: Knowledge transfer with multimodal open models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 13 060–13 077. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.764/

[55] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning

with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[56] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[57] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu *et al.*, "A survey on llm-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.

[58] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 076–12 100. [Online]. Available: https://aclanthology.org/2023.emnlp-main.741/

[59] L. Jing, R. Li, Y. Chen, and X. Du, "FaithScore: Fine-grained evaluations of hallucinations in large vision-language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5042–5063. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.290/

[60] X. Liu, L. Zhang, S. Munir, Y. Gu, and L. Wang, "Verifact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts," *arXiv preprint arXiv:2505.09701*, 2025.

[61] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang *et al.*, "Long-form factuality in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 80 756–80 827, 2024.

[62] X. Cheng, Z. Tan, W. Xue, and W. Lu, "Information re-organization improves reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 130 214–130 236, 2024.

[63] H. Li, X. Han, H. Wang, Y. Wang, M. Wang, R. Xing, Y. Geng, Z. Zhai, P. Nakov, and T. Baldwin, "Loki: An open-source

tool for fact verification," in *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, B. Mather, and M. Dras, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 28–36. [Online]. Available: https://aclanthology.org/2025.coling-demos.4/

[64] Y. Wang, M. Wang, H. Iqbal, G. N. Georgiev, J. Geng, I. Gurevych, and P. Nakov, "OpenFactCheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and LLMs," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 11 399–11 421. [Online]. Available: https://aclanthology.org/2025.coling-main.755/

[65] H. Ou, G. Deng, X. Han, J. Zhang, X. He, H. Qiu, S. Guo, and T. Zhang, "Holmes: Automated fact check with large language models," *arXiv preprint arXiv:2505.03135*, 2025.

[66] V. V and V. Setty, "Livefc: A system for live fact-checking of audio streams," in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1060–1063. [Online]. Available: https://doi.org/10.1145/3701551.3704128

[67] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal *et al.*, "Benchmarking multi-modal entailment for fact verification," in *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.

[68] A. M. Luvembe, W. Li, S. Li, F. Liu, and X. Wu, "Caf-odnn: Complementary attention fusion with optimized deep neural network for multimodal fake news detection," *Information Processing & Management*, vol. 61, no. 3, p. 103653, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645732400013X

[69] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information Processing &*

Management, vol. 60, no. 1, p. 103120, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457322002217

[70] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, p. 102437, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457320309304

[71] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, "Multimodal automated fact-checking: A survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5430–5448. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.361

[72] T. Chakraborty, V. La Gatta, V. Moscato, and G. Sperlì, "Information retrieval algorithms and neural ranking models to detect previously fact-checked information," *Neurocomputing*, vol. 557, p. 126680, 2023.

[73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[74] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

[75] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[76] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[77] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[78] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[79] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[80] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[81] W. Mansour, T. Elsayed, and A. Al-Ali, "This is not new! spotting previously-verified claims over twitter," *Information Processing & Management*, vol. 60, no. 4, p. 103414, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457323001516

[82] N. Hulke, B. R. Siva, A. Raj, and A. A. Saifee, "Tyche at factify 2022: Fusion networks for multi-modal fact-checking," 2021.

[83] W. Bai, "Greeny at factify 2022: Ensemble model with optimized roberta for multi-modal fact verification," in *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection*, 2022.

[84] Y. Zhuang and Y. Zhang, "Yet at factify 2022: Unimodal and bimodal roberta-based models for fact checking," in *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.

[85] B. Chaitanya, P. Prathyush, and V. Rutvik, "Truthformers at factify 2022: Evidence aware transformer based model for multimodal fact checking," 2021.

[86] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[87] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17 283–17 297, 2020.

[88] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," *arXiv preprint arXiv:2011.04006*, 2020.

[89] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, sep 2022. [Online]. Available: https://doi.org/10.1145/3505244

[90] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[91] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[92] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[93] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[94] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[95] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[96] J. Geng, Y. Kementchedjhieva, P. Nakov, and I. Gurevych, "Multimodal large language models to support real-world fact-checking," *arXiv preprint arXiv:2403.03627*, 2024.

[97] A. Das, H. Liu, V. Kovatchev, and M. Lease, "The state of human-centered nlp technology for fact-checking," *Information Processing & Management*, vol. 60, no. 2, p. 103219, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645732200320X

[98] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5368–5393. [Online]. Available: https://aclanthology.org/2023.acl-long.294/

[99] I. Vykopal, M. Pikuliak, S. Ostermann, and M. Šimko, "Generative large language models in automated fact-checking: A survey," *arXiv preprint arXiv:2407.02351*, 2024.

[100] Q. Liu, B. Chen, J. Guo, M. Ziyadi, Z. Lin, W. Chen, and J.-G. Lou, "Tapex: Table pre-training via learning a neural sql executor," *arXiv preprint arXiv:2107.07653*, 2021.

[101] L. Pan, X. Lu, M.-Y. Kan, and P. Nakov, "QACheck: A demonstration system for question-guided multi-hop fact-checking," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 264–273.

[102] T. Buzan, B. Buzan, and J. Harrison, "The mind map book: Unlock your creativity, boost your memory, change your life," 2010.

[103] J. Li, Y. Wang, S. Zhang, and M. Zhang, "Rethinking document-level relation extraction: A reality check," in *Findings of the Association for Computational Linguistics*, 2023, pp. 5715–5730.

[104] H. Cao, L. Wei, M. Chen, W. Zhou, and S. Hu, "Are large language models good fact checkers: A preliminary study," *arXiv preprint arXiv:2311.17355*, 2023.

[105] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[106] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

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

[108] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[109] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3441–3460. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.309/

[110] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, "The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 2021, pp. 1–13.

[111] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[112] N. Jin, J. Siebert, D. Li, and Q. Chen, "A survey on table question answering: recent advances," in *China Conference on Knowledge Graph and Semantic Computing*. Springer, 2022, pp. 174–186.

[113] Z. Wu and Y. Feng, "ProTrix: Building models for planning and reasoning over tables with sentence context," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 4378–4406. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.253/

[114] X. Lu, L. Pan, Y. Ma, P. Nakov, and M.-Y. Kan, "Tart: An open-source tool-augmented framework for explainable table-based reasoning," *arXiv preprint arXiv:2409.11724*, 2024.

[115] L. Qin, Q. Chen, X. Feng, Y. Wu, Y. Zhang, Y. Li, M. Li, W. Che, and P. S. Yu, "Large language models meet nlp: A survey," *arXiv preprint arXiv:2405.12819*, 2024.

[116] Y. Sui, M. Zhou, M. Zhou, S. Han, and D. Zhang, "Table meets llm: Can large language models understand structured table data? a benchmark and empirical study," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, ser. WSDM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 645–654. [Online]. Available: https://doi.org/10.1145/3616855.3635752

[117] N. Deng, Z. Sun, R. He, A. Sikka, Y. Chen, L. Ma, Y. Zhang, and R. Mihalcea, "Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 407–426. [Online]. Available: https://aclanthology.org/2024.findings-acl.23/

[118] A. Remadi, K. El Hage, Y. Hobeika, and F. Bugiotti, "To prompt or not to prompt: Navigating the use of large language models for integrating and modeling heterogeneous data," *Data & Knowledge Engineering*, vol. 152, p. 102313, 2024.

[119] Y. Zhao, Z. Qi, L. Nan, B. Mi, Y. Liu, W. Zou, S. Han, R. Chen, X. Tang, Y. Xu, D. Radev, and A. Cohan, "QTSumm: Query-focused summarization over tabular data," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1157–1172. [Online]. Available: https://aclanthology.org/2023.emnlp-main.74/

[120] N. S. Moosavi, A. Rücklé, D. Roth, and I. Gurevych, "Scigen: a dataset for reasoning-aware text generation from scientific tables," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[121] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[122] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[123] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Ding, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Chen, J. Yuan, J. Tu, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. You, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Zhou, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, p. 633–638, Sep. 2025. [Online]. Available: http://dx.doi.org/10.1038/s41586-025-09422-z

[124] T. Liu, F. Wang, and M. Chen, "Rethinking tabular data understanding with large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 450–482. [Online]. Available: https://aclanthology.org/2024.naacl-long.26/

[125] X. Fang, W. Xu, F. A. Tan, J. Zhang, Z. Hu, Y. Qi, S. Nickleach, D. Socolinsky, S. Sengamedu, and C. Faloutsos, "Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey," *arXiv preprint arXiv:2402.17944*, 2024.

[126] Y. Zhao, H. Zhang, S. Si, L. Nan, X. Tang, and A. Cohan, "Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, M. Wang and I. Zitouni, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 160–175. [Online]. Available: https://aclanthology.org/2023.emnlp-industry.17/

[127] S. Taniguchi, K. Harada, G. Minegishi, Y. Oshima, S. C. Jeong, G. Nagahara, T. Iiyama, M. Suzuki, Y. Iwasawa, and Y. Matsuo, "Adopt: Modified adam can converge with any \beta_2 with the optimal rate," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., pp. 72 438–72 474.

[128] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[129] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/

[130] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[131] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[132] H. Saadany and C. Orasan, "BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text," in *Proceedings of the Translation and Interpreting Technology Online Conference*, R. Mitkov, V. Sosoni, J. C. Giguère, E. Murgolo, and E. Deysel, Eds. Held Online: INCOMA Ltd., Jul. 2021, pp. 48–56. [Online]. Available: https://aclanthology.org/2021.triton-1.6/

[133] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu, "Llms-as-judges: a comprehensive survey on llm-based evaluation methods," *arXiv preprint arXiv:2412.05579*, 2024.

[134] J. Li, J. Wang, Z. Zhang, and H. Zhao, "Self-prompting large language models for zero-shot open-domain QA," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 296–310. [Online]. Available: https://aclanthology.org/2024.naacl-long.17/

[135] R. F. Cekinel, P. Karagoz, and Ç. Çöltekin, "Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, 2025, pp. 4622–4633. [Online]. Available: https://aclanthology.org/2025.coling-main.310/

[136] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. Da San Martino, "Automated fact-checking for assisting human fact-checkers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence,*

IJCAI-21, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4551–4558, survey Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/619

[137] J. Chen, L. Liu, and F. Zhou, "Do not wait: Preemptive rumor detection with cooperative llms and accessible social context," *Information Processing & Management*, vol. 62, no. 3, p. 103995, 2025.

[138] Z. Li, J. Yang, X. Wang, J. Lei, S. Li, and J. Zhang, "Uncertainty-aware disentangled representation learning for multimodal fake news detection," *Information Processing & Management*, vol. 62, no. 5, p. 104190, 2025.

[139] A. Rangapur, H. Wang, L. Jian, and K. Shu, "Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation," in *Companion Proceedings of the ACM on Web Conference 2025*, ser. WWW '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 785–788. [Online]. Available: https://doi.org/10.1145/3701716.3715292

[140] D. Russo, S. Menini, J. Staiano, and M. Guerini, "Face the facts! evaluating RAG-based pipelines for professional fact-checking," in *Proceedings of the 18th International Natural Language Generation Conference*. Hanoi, Vietnam: Association for Computational Linguistics, Oct. 2025, pp. 846–865. [Online]. Available: https://aclanthology.org/2025.inlg-main.50/

[141] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 2318–2335. [Online]. Available: https://aclanthology.org/2024.findings-acl.137/

[142] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "Llm2vec: Large language models are secretly powerful text encoders," *arXiv preprint arXiv:2404.05961*, 2024.

[143] H. Cao, "Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark," *arXiv preprint arXiv:2406.01607*, 2024.

[144] J. Huang, J. Chen, J. Lin, J. Qin, Z. Feng, W. Zhang, and Y. Yu, "A comprehensive survey on retrieval methods in recommender systems," *arXiv preprint arXiv:2407.21022*, 2024.

[145] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[146] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024.

[147] S. Tahmasebi, E. Müller-Budack, and R. Ewerth, "Multimodal misinformation detection using large vision-language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2189–2199.

[148] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, and A. Bandhakavi, "Logically at factify 2022: Multimodal fact verification," *arXiv preprint arXiv:2112.09253*, 2021.

[149] J. Huang, M. Xiao, D. Li, Z. Jiang, Y. Yang, Y. Zhang, L. Qian, Y. Wang, X. Peng, Y. Ren, R. Xiang, Z. Chen, X. Zhang, Y. He, W. Han, S. Chen, L. Shen, D. Kim, Y. Yu, Y. Cao, Z. Deng, H. Li, D. Feng, Y. Dai, V. Somasundaram, P. Lu, G. Xiong, Z. Liu, Z. Luo, Z. Yao, R.-L. Weng, M. Qiu, K. E. Smith, H. Yu, Y. Lai, M. Peng, J.-Y. Nie, J. W. Suchow, X.-Y. Liu, B. Wang, A. Lopez-Lira, Q. Xie, S. Ananiadou, and J. Tsujii, "Open-finllms: Open multimodal large language models for financial applications," *arXiv preprint arXiv:2408.11878*, 2024.

[150] H. Yu, J. Wen, and Z. Zheng, "Camel: Cross-modality adaptive meta-learning for text-based person retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4651–4663, 2025.

[151] Y. Wang, Y. Gao, X. Chen, H. Jiang, S. Li, J. Yang, Q. Yin, Z. Li, X. Li, B. Yin, J. Shang, and J. McAuley, "Memoryllm: towards self-updatable large language models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[152] X. Li, C. Zhu, L. Li, Z. Yin, T. Sun, and X. Qiu, "LLatrieval: LLM-verified retrieval for verifiable generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 5453–5471. [Online]. Available: https://aclanthology.org/2024.naacl-long.305/

[153] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 724–19 731.

[154] I. Vykopal, M. Pikuliak, S. Ostermann, and M. Šimko, "Generative large language models in automated fact-checking: A survey (2024)," *arXiv preprint arXiv:2407.02351*.

[155] W. Xu, Q. Liu, S. Wu, and L. Wang, "Counterfactual debiasing for fact verification," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6777–6789. [Online]. Available: https://aclanthology.org/2023.acl-long.374/

[156] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[157] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[158] N.-K. Le, T. D. Do, and L.-M. Nguyen, "SPECTRA: Faster large language model inference with optimized internal and external speculation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 14 015–14 034. [Online]. Available: https://aclanthology.org/2025.acl-long.685/

[159] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[160] R. Zhang, Y. Tian, P. Wei, D. D. Zeng, and W. Mao, “An LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance identification,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 253–12 266. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.714/

Publications

[1] Luu, S.T., Nguyen, H., Vo, T., Nguyen, LM. (2025). **ZeFaV: Boosting Large Language Models for Zero-Shot Fact Verification**. In: *PRICAI 2024: Trends in Artificial Intelligence. PRICAI 2024. Lecture Notes in Computer Science, vol 15282. Springer, Singapore.* https://doi.org/10.1007/978-981-96-0119-6_28

[2] Luu, S.T., Vo, T. & Nguyen, LM. **MCVE: multimodal claim verification and explanation framework for fact-checking system.** *Multimedia Systems 31, 242 (2025).* https://doi.org/10.1007/s00530-025-01804-7

[3] Luu, S.T., Vo, T. & Nguyen, LM. **Boosting Large Language Models for Fact-Checking: Leveraging Verbalized Tabular Data as Evidence.** *International Journal of Data Science and Analytics (accepted)*

[4] Luu, S.T., Vo, T. & Nguyen, LM. **M-RAV: Multimodal Retrieve-Augment-Verify Framework for Boosting Zero-shot Fact Verification System with Large Language Models.** *Information Processing & Management (under review at 3rd revision)*

[5] Chu, H., Chu, H., Nguyen, T.-M., Luu, S. T., Hoang, C., Nguyen, H., Tran, V., & Nguyen, L.-M. (2025). **DeepSIX at ACM MM 2025 Grand Challenge: Enhancing context text processing for multimodal hallucination detection and fact verification**. *In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 13874–13880). Association for Computing Machinery.* https://doi.org/10.1145/3746027.3762061