Multimodal Misinformation Detection using Large Vision-Language Models

Sahar Tahmasebi
sahar.tahmasebi@tib.eu
TIB – Leibniz Information Centre for
Science and Technology
Hannover, Germany

Eric Müller-Budack
eric.mueller@tib.eu
TIB – Leibniz Information Centre for
Science and Technology;
L3S Research Center, Leibniz
University Hannover
Hannover, Germany

Ralph Ewerth
ralph.ewerth@tib.eu
TIB – Leibniz Information Centre for
Science and Technology;
L3S Research Center, Leibniz
University Hannover
Hannover, Germany

ABSTRACT

The increasing proliferation of misinformation and its alarming impact have motivated both industry and academia to develop approaches for misinformation detection and fact checking. Recent advances on large language models (LLMs) have shown remarkable performance in various tasks, but whether and how LLMs could help with misinformation detection remains relatively underexplored. Most of existing state-of-the-art approaches either do not consider evidence and solely focus on claim related features or assume the evidence to be provided. Few approaches consider evidence retrieval as part of the misinformation detection but rely on fine-tuning models. In this paper, we investigate the potential of LLMs for misinformation detection in a zero-shot setting. We incorporate an evidence retrieval component into the process as it is crucial to gather pertinent information from various sources to detect the veracity of claims. To this end, we propose a novel re-ranking approach for multimodal evidence retrieval using both LLMs and large vision-language models (LVLM). The retrieved evidence samples (images and texts) serve as the input for an LVLM-based approach for multimodal fact verification (LVLM4FV). To enable a fair evaluation, we address the issue of incomplete ground truth for evidence samples in an existing evidence retrieval dataset by annotating a more complete set of evidence samples for both image and text retrieval. Our experimental results on two datasets demonstrate the superiority of the proposed approach in both evidence retrieval and fact verification tasks and also better generalization capability across dataset compared to the supervised baseline.

KEYWORDS

Multimodal misinformation detection, large language models, social media, news analytics

1 INTRODUCTION

Misinformation and fake news contain false information to deliberately deceive readers [2] and have become a pressing challenge since they can cause severe consequences for society. Typically, misinformation is conveyed in different modalities such as images, text, and videos to provide a stronger story line and attract attention from readers. The situation has become even more complicated with the emergence of large language models (LLMs) like Generative Pre-trained Transformer (GPT) [29], since they can be intentionally misused to generate [12] or spread misinformation due to the hallucination issue [53]. Thus, automated solutions for
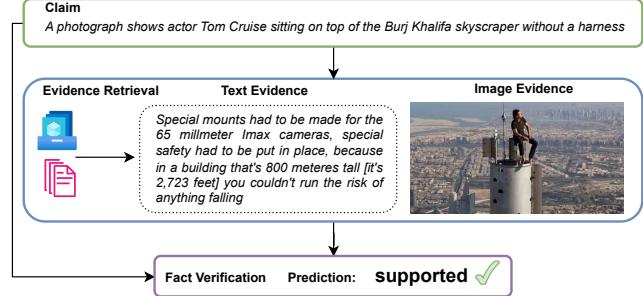

Figure 1: Example of multimodal misinformation detection.

fact checking and multimodal misinformation detection are in great need. The multimodal misinformation detection is generally structured into distinct stages including claim detection and extraction, check-worthiness prediction, verdict prediction and explanation generation. This work focuses on the critical stage of verdict prediction which comprises an evidence retrieval component followed by the prediction of the targeted claim regarding its truthfulness (see Figure 1).

Recently, researchers have started to investigate automatic misinformation detection by developing various benchmark datasets [27, 44, 50]. State-of-the-art approaches [42, 47, 50] have mainly used deep learning techniques to extract features from text and image. However, detecting misinformation on (multimodal) social media is a challenging research problem since it is difficult to maintain applicability and good performance for unforeseen events. We found the following limitations with the current studies. (1) Most of them only consider text [5, 37, 46] while ignoring the multimodal nature (e.g., combination of images and texts) of online articles, which are useful to predict the truthfulness of claims. (2) While evidence retrieval is a core sub-task in detecting misinformation, most multimodal-based approaches either do not use evidences and rely only on the claim-related features from the associated news [42, 47, 52], or assume that evidences are already provided [10, 46] based on which the models can directly predict the truthfulness of the target claim. However, this is not realistic in practice as the claim does not typically come with evidence. Instead evidences should be retrieved from a knowledge base or the Internet. (3) Only few approaches (e.g., [50]) consider evidence retrieval as part of multimodal misinformation detection. However, these approaches have not leveraged powerful generative LLMs (e.g., [17, 45]) and LVLMs (e.g., [7, 23] for

retrieval and the classification of claims regarding misinformation. These LLMs and LVLMs have demonstrated superior comprehension capabilities in tasks like image captioning and visual question answering. Their strong zero-shot and few-shot generalization opens doors to myriad of applications across many domains.

In this paper, we introduce a pipeline for multimodal misinformation detection including a novel re-ranking method for evidence retrieval using LLMs and LVLMs followed by a fact verification step. Overall, we make the following contributions. (1) In contrast to related work that either does not consider evidence [42, 47, 52] or assume it to be provided [10, 46, 54], we incorporate an evidence retrieval component into our misinformation detection approach, to make an informed decision regarding the veracity of claims. (2) Unlike other approaches that consider evidence retrieval as part of multimodal misinformation detection using supervised models, we propose a novel unsupervised re-ranking approach leveraging LLMs and LVLMs for evidence retrieval called LVLM4EV. For this purpose, we propose effective prompting strategies and extract ranking scores directly from both LLMs and LVLMs for text and image retrieval. The retrieved image and text evidences serve as input for fact verification using L(V)LMs as a classifier through a prompting strategy that classifies misinformation using majority voting on the answers. An example is shown in Figure 1. (3) Existing datasets [26, 50] create ground-truth evidences for a claim by retrieving (single) expert-verified evidences from sources such as *PolitiFact*[1] or *Snopes*[2]. However, in a large database, this might overlook other pertinent evidence and thus some relevant evidences for a claim remain unlabeled. We identify this issue of *incomplete ground-truth labels* in the *MOCHEG* dataset for the evidence retrieval task and resolve it by annotating a more complete evaluation subset, consisting of top-10 relevant candidate evidences for both image and text retrieval to ensure fair and accurate assessment of system performance. (4) We perform experiments on two datasets to evaluate the effectiveness of the proposed pipeline and also explore generalization capability across datasets, which is critical for real-world applications. The findings underscore the effectiveness of our unsupervised, multimodal approach for misinformation detection and an improved generalization capability compared to supervised baselines.

The remainder of this paper is structured as follows. Section 2 reviews related work on misinformation detection and large generative AI models. Our pipeline for multimodal misinformation detection, including a novel re-ranking approach based on LLM and LVLM for evidence retrieval, as well as an approach for fact verification is presented in Section 3. The experimental setup and results including details of annotation for a clean evaluation set to address incomplete ground truth issue are presented in Section 4. Section 5 concludes this paper and provides future work directions.

2 RELATED WORK

In this section, we review the related work on misinformation detection approaches, including pattern-based and evidence-based methods, as well as large generative AI models.

2.1 Misinformation Detection

Misinformation detection methodologies can be divided into two main categories: pattern-based and evidence-based [47].

Pattern-based methods treat the misinformation detection as a feature classification task, where language/vision models are employed to determine the integrity of news content. This evaluation encompasses various aspects such as writing style, sentiment analysis, user credibility and other relevant features. Majority of previous research in this category heavily depends on features from text and user metadata [30, 32]. Recent research has begun incorporating images [4, 9, 42, 52] and videos [24, 31] into the detection of misinformation due to the inherent multimodality of information. Most of these approaches rely on verifying cross-modal consistency [40, 43] or generating a combined representation of textual and visual features for classification [18, 42, 52]. For example, Wang et al. [47] introduced an *Event Adversarial Neural Network (EANN)*, which extracts event-invariant features to detect fake news on unforeseen events. From the same perspective, *Spotfake* [39] combined features from both modalities and showed the effectiveness of pre-trained language models like *BERT* [8] and computer vision approaches like Visual Geometry Group model (*VGG-19*) [38].

Evidence-based approaches are designed to meticulously verify claims by integrating a wealth of external information. Early attempts on fact verification were based on textual information [11, 44, 46]. For example, The *Liar* dataset and framework [46] collected relevance evidences from *PolitiFact* website and explored automatic fake news detection based on surface-level linguistic patterns. They proposed a hybrid convolutional neural network to integrate evidences with text and showed the improvement compared to evidence free deep learning model. Textual datasets and methods are no longer enough in the social media age as the claim and evidence could be in any modalities like image [1, 26, 50] or videos [25, 34]. Gao et al. [10] framed the task as multimodal entailment task and proposed two baseline approaches including an ensemble model which combines two unimodal models and a multimodal attention network that models the interaction between image and text pair from claim and evidence document. Yao et al. [50] proposed an end-to-end Multimodal fact-CHecking and Explanation Generation (*MOCHEG*) benchmark dataset. To set the baseline for this benchmark, they have fine-tuned SBERT (Sentence Bidirectional Encoder Representations from Transformers, [35]) for text retrieval and CLIP (Contrastive Language–Image Pre-Training [33]), for image retrieval with a contrastive loss. For the verification task, they used CLIP for claim and evidence embedding following an attention layer to compute distribution between them and a classification layer to get final labels.

While pattern-based approaches (e.g., [42, 47, 52]) do not consider retrieving evidences, most evidence-based approaches (e.g. [10, 46, 54]) operate under the assumption that relevant evidence is already provided alongside with the content under scrutiny. However, this assumption diverges from real-world scenarios since evidence is often required to verify claims but rarely provided. Only few approaches (e.g., [50]) consider evidence retrieval as part of multimodal misinformation detection and generally fine-tune pre-trained Small Language Models (SLMs) like BERT [8] to understand news content and provide fundamental representation. SLMs do bring
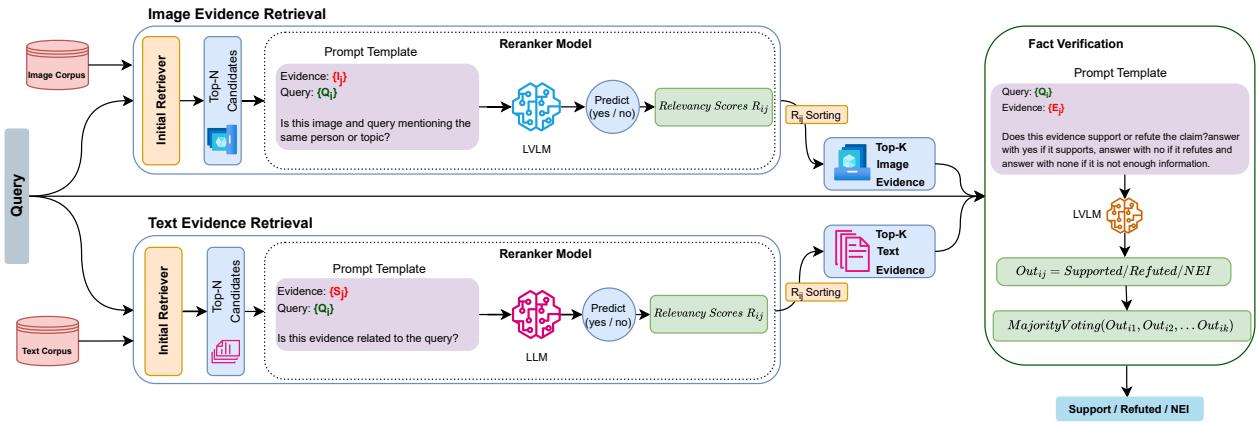

Figure 2: Overview of our misinformation detection approach. Blue border (LVLM4EV): Based on a textual input claim, evidence texts and images are initially retrieved from a corpus using a state-of-the-art approach (e.g., *MOCHEG* [50]). Generative LLMs (e.g., Mistral-7B [17]) and LVLMs (e.g., InstructBLIP [7]) are used to re-rank the top-N text and image evidences. Green border (LVLM4FV): Based on the re-ranked evidences, we finally employ a LVLM (e.g., LLaVA [22]) for misinformation detection.

improvements, but their knowledge and capability limitations also compromise further enhancement of misinformation detectors. For example, BERT was pre-trained on text corpus like Wikipedia and thus struggled to handle news items that require knowledge not included [36]. More powerful LLMs and LVLMs that have achieved impressive performance for many tasks, still remain to be explored for the task of evidence-based misinformation detection.

2.2 Generative AI Models

Large Language Models (LLMs) such as GPT [28], LLaMA (Large Language Model Meta AI, [45] and Mistral-7B [17] which are usually trained on the larger-scale corpus and aligned with human preferences, have shown impressive emergent abilities on various tasks [48] and are considered promising as general task solvers. For example, in the similar task of fact checking, Hoes et al. [14] find that ChatGPT accurately classifies 69% of text-only statements in a dataset built from the factchecking platform PolitiFact. *FactL-LaMA* [6] integrates external evidence into the instruct-tuning process to enhance the model's ability to leverage evidence for predictions. Hu et al. [15] also studied how to utilize LLMs help in improving the performance of text-based fake news detection and showed that LLMs like *GPT-3.5* underperforms the task-specific SLM but could provide informative rationales and complement SLMs in news understanding.

LLMs are expanding beyond text-only functions to include multiple modalities like image and videos. With the advancement of LVLMs like Bootstrapping Language-Image Pre-training (BLIP-2) [21], *InstructBLIP* [7] and Large Language-and-Vision Assistant (*LLaVA*) [23] the systematic evaluation of their capabilities has become increasingly critical. A broad array of benchmarks [1, 49, 51] has been established to estimate the LVLMs performance across a variety of tasks, such as image captioning, optical character recognition and chart fact checking. However, there is an absence of research on evaluating the ability of LVLMs to conduct multimodal misinformation detection.

3 MULTIMODAL MISINFORMATION DETECTION

In this section, we propose a pipeline that exploits LLMs and LVLMs for multimodal misinformation detection. The task is to predict whether a post is *refuted* ($y = 0$), *supported* ($y = 1$) or the evidence provides not enough information *NEI* ($y = 2$). As illustrated in Figure 2, it consists of two main components: (1) An evidence retrieval component which automatically retrieves evidences from a large collection of web sources including articles and images for a given claim. (2) A multimodal misinformation detection approach that predicts the veracity of the claim given the retrieved evidences.

3.1 Problem Definition

We define the task of multimodal misinformation detection as follows. Given a textual claim (or query) Q and a corpus with m evidences $\mathbb{C} = \{E_1, E_2, \ldots, E_m\}$ as input, we aim to predict whether the claim is refuted ($y = 0$), supported ($y = 1$) or evidences do not provide enough information (NEI, $y = 2$). Each evidence $E = (\mathbb{I}, \mathbb{S})$ is defined as a set of v images $\mathbb{I} = \{I_1, I_2, \ldots, I_v\}$ and t sentences $\mathbb{S} = \{S_1, S_2, \ldots, S_t\}$.

3.2 Evidence Retrieval

Most existing methods for misinformation detection prioritize claim verification and directly compute a fused representation of multimodal (textual and visual) information for final classification. However, retrieving high-quality evidences is the foundation of misinformation detection especially in terms of new events.

Given a textual claim Q and a corpus of evidences $\mathbb{C}$ as input, the goal of initial text and image retrieval is to find a sorted subset of the N most related sentences $\mathbb{S}' = \{S_1, S_2, \ldots, S_N\} \subset \mathbb{C}$ and images $\mathbb{I}' = \{I_1, I_2, \ldots, I_N\} \subset \mathbb{C}$. For both subtasks, we initially retrieve text and image evidences using an appropriate state-of-the-art model. Since existing supervised ranking methods can suffer from weak generalizability to new domains, and restricted commercial use [16], we aim to exploit the excellent generalization

capabilities of LLMs [3] and LVLMs [19] in zero-shot settings to improve evidence retrieval. To this end, we propose to use LLMs and LVLMs to re-rank text and images evidences, respectively. In the following, we provide details for our proposed re-ranking approach, which we denote as LVLM4EV (Large Vision-Language Models for Evidence Retrieval).

3.2.1 Initial Retriever. In information retrieval tasks, the utilization of an initial retriever is imperative due to the computational complexity of directly applying a computational-heavy model (especially LLMs or LVLMs) to the entire corpus [41]. The initial retriever efficiently narrows down the search space by quickly identifying a subset $\mathbb{C}' \subset \mathbb{C}$ of potentially relevant evidences, ensuring that computational resources are utilized effectively while maintaining the performance. As shown in Figure 2, we employ an initial retrieval approach to retrieve text and image evidences from a large corpus. For thus purpose, we apply MOCHEG [50], which is a state-of-the-art evidence retrieval approach, to get top-N most related evidences.

Text Retrieval. For text evidence retrieval, documents are segmented into sentences, and the fine-tuned *SBERT* model [35] from MOCHEG [50] is applied to generate contextual representations for input claims and sentences from the evidences. We compute the ranking score $r_j = \text{cossim}(\text{SBERT}(Q), \text{SBERT}(S_j))$ using the cosine similarity between the claim Q and each sentences $S_j \in \mathbb{C}$ from all evidences in the corpus to create an initial ranking and filter the top-N evidences to output a sorted subset $\mathbb{S}' = \{S_1, S_2, \ldots, S_N\}$ used for re-ranking (Section 3.2.3).

Image Retrieval. Similarly, for image evidence retrieval, we apply the text CLIP^T and image encoder CLIP^I from the CLIP model [33] fine-tuned from MOCHEG [50] to encode both textual input claim and images. We use the cosine similarity as ranking score $r_j = \text{cossim}(\text{CLIP}^T(Q), \text{CLIP}^I(I_j))$ to rank all images $I_j \in \mathbb{C}$ from each evidences in the corpus. As a result, a sorted subset of N images $\mathbb{I}' = \{I_1, I_2, \ldots, I_N\}$ is created for re-ranking (Section 3.2.3).

3.2.2 Prompting. We use prompting to re-rank the results for both image and text retrieval. For this purpose, we use a prompt template to input the query claim Q and a initially retrieved sentence $S \in \mathbb{S}'$ or image evidence $I \in \mathbb{I}'$ into a generative AI model for relevance assessment. The models are instructed to output "Yes" if a candidate evidence is relevant to the claim and "No" otherwise. Given that prompt design significantly influences the response of an LLM or LVLM, different prompts have been used. Examples are shown in Figure 2 and are evaluated in the experiments.

3.2.3 Re-ranking using Generative AI Models. We intent to improve the original retrieved evidences by a re-ranking method $f(\mathbb{C})' \mapsto \mathbb{C}''$ based on generative AI models, e.g., *Mistral* [17] for text and *InstructBLIP* [7] for image evidences. To this end, we first apply the prompt templates mentioned in Section 3.2.2 to the Top-N text $\mathbb{S}'$ and image evidences $\mathbb{I}'$ from the initial retriever and the input claim Q to get an answer which is "Yes" or "No". These answers then allow us to perform a re-ranking of the top-N evidences $\mathbb{C}'$. For this purpose, we consider the following strategies.

*Initial Ranking Scores (*IRS*).* In this approach, we re-rank evidences solely based on the answer of the LLM or LVLM. For this

purpose, we rank all evidences that are considered relevant (answer: "Yes") by the LLM or LVLM higher than the irrelevant ones (answer: "No"). The remaining ranking, i.e., the ranking within relevant and irrelevant evidences, is created based on the ranking score r_j computed during the initial retrieval step (Section 3.2.1) using the cosine similarity. Finally, we select the top-K re-ranked evidences from $\mathbb{C}''$ as final evidence for fact verification (Section 3.3).

*Generative AI Scores (*GAIS*).* Each generative AI model uses a tokenizer for text generation with a specific dictionary of possible tokens. We use the log probability of the generative AI model to extract the scores for all pairs in order to re-rank them. For this purpose, we re-rank the evidences using the probability of model generating the token 'Yes' or 'No' as follows:

$$p_j = \begin{cases} p_j(Yes) & \text{if output Yes} \\ \lambda\,(1 - p_j(No)) & \text{if output No} \end{cases} \tag{1}$$

where λ is a small value that ensures ranking irrelevant (answer "No") below relevant evidences (answer "Yes"). The probability p_j is calculated with three different settings.

GAIS-ALL: We select Softmax normalization across all tokens in the dictionary and extract the probability of the generated token ($p(Yes)$ if the output is "Yes" and $1 - p(No)$ if the output is "No"). The P_j then will be calculated according to Eq. 1.

GAIS-YN: This setting employs prompt-based classification which involves categorizing tokens into "Yes" and "No" classes, each with all existing variants of yes (e.g Yes, yes, YES) and no (e.g no, No, NO) token IDs in the dictionary. The Softmax is applied exclusively on the cumulative sum of the tokens within these two classes. The class with the highest probability determines the generated token, and its probability serves as the extracted score p_j.

GAIS-YNO: As in the second approach, we categorize "Yes" and "No" tokens but also sum the probabilities of the remaining tokens for the class "other" for Softmax normalization. The final probability p_j is calculated depending on the class with the highest probability, only considering the outputs for "Yes" and "No".

Using one of the strategies above, we re-rank the candidate evidence based on p_j and select the top-K as the retrieved evidence samples $\mathbb{C}''$ for fact verification.

3.3 Fact Verification

Based on the retrieved text and image evidence samples, we suggest a fact verification component to predict the veracity of an input claim (Figure 2, right). We consider two models, *Mistral* [17] for text-only evidence, and *LLaVA* [23] for multimodal evidence.

For *text-based misinformation detection*, we pair each retrieved evidence sentence with the input claim (Q, S_j), $j \in [1, K]$. For *multimodal misinformation detection*, we first form a set of multimodal pairs $\mathbb{P}$ by augmenting each retrieved sentence S_j with its corresponding images $\mathbb{I}_j$ from the same document, and likewise, each retrieved image I_j with its corresponding sentences in the text. Then, we create prompts using the claim and each of these multimodal pairs.

3.3.1 Prompting. We follow two strategies to create a prompt template for fact verification: one-level prompting (an example is shown in Figure 2, right) and two-level prompting. In one-level prompting,

models are instructed to answer 'Yes' if the evidence support the claim, 'No' if it refutes it and "None" if it does not provide enough information. This enables us to directly obtain all three labels from the model. However, this strategy might lead to a bias towards 'Yes' or 'No' decision as the definition of NEI is not very explicit. Thus, in two-level prompting, the idea is to first explicitly ask the model if the evidence is enough to support or refute the claim to determine samples of "NEI" class. Then, we input all remaining samples with enough information (answer "Yes") to the second level and ask the model whether a sample supports or refutes the claim. In this way, the models are instructed to answer 'Yes' or 'No' in both levels.

3.3.2 Classification. We use prompt-based classification to ensure the model output is robust and we avoid hallucination. To this end, the normalization and extraction of the generated token are based on `GAIS-YN`, as mentioned in Section 3.2.3 for both prompting strategies. Note that for one-level prompting, as we have three classes (Yes/No/None tokens), Softmax is applied across all of them. Once we have the generated label for each claim-evidence pair ($Q, P \in \mathbb{P}$), we apply majority voting to determine whether the claim is supported, refuted, or NEI. In case of equal number of maximum votes for multiple labels, we use the one with maximum extracted probability from the model.

4 EXPERIMENTS

This section provides details on the experimental setup (Section 4.1), results for both evidence retrieval and fact verification tasks (Section 4.2 and 4.3 respectively), and a generalization study on *Factify* dataset (Section 4.4).

4.1 Experimental Setup

In this section, we describe the setting of experiments including the multimodal datasets used in the experiments (Section 4.1.1), baseline methods (Section 4.1.2) and implementation details (Section 4.1.3).

4.1.1 Dataset. To evaluate the performance and generalization across various domains of the proposed pipeline, we conduct experiments on two datasets, i.e., *MOCHEG* [50] and *Factify* [26].

MOCHEG is a multimodal fact checking benchmark. This dataset is originally based on textual claims from *Snopes* and *PolitiFact*, with associated text and image evidences. It is divided into a train, validation and test set. Since we propose a zero-shot approach, we only use the validation set for hyperparameter selection and test set for evaluation.

Factify is a fact-checking benchmark featuring news from India and United States, with both claim and evidence being multimodal. The data is originally labeled in five categories of *Support_Text*, *Support_Multimodal*, *Insufficient_Text*, *Insufficient_Multimodal* and *Refute*. In order to effectively apply our approach and compare it to the baselines, we convert it to the broad 3-class categories including *Support* (Support_Multimodal & Support_Text), *Insufficient* (Insufficient_Multimodal & Insufficient_Text) and *Refute*. Since the test set labels are not publicly available due to the dataset's origin from an open challenge, we utilize the development set for hyperparameter selection and the validation set for evaluation as proposed by [10].

4.1.2 Baselines. To compare our approach to the state of the art, we use two baselines. *MOCHEG* [50] is the first end-to-end multimodal fact-checking method which considers both tasks of evidence retrieval and verification. As mentioned in the Section 2, it fine-tunes SBERT model [35] for text retrieval and CLIP [33] for image retrieval in a contrastive manner and based on cosine similarity between the input claim and candidate evidence. For the verification task, it uses the CLIP model to encode both claim and evidence and then an attention layer to compute the distribution between them. The combined representation is subsequently passed through a classification layer to predict the final labels. We have used the official implementation of *MOCHEG* provided on GitHub[3] with the best parameters reported.

We also compare the approaches to *Logically* [10] on *Factify* validation set as this method was ranked first on the challenge leaderboard. *Logically* treats the challenge as a multimodal entailment task and proposes an ensemble model which combines predictions of two uni-modal models (a transformer architecture for text and a ResNet-50 [13] for image) fine-tuned on the task dataset. As *Logically* is fully-supervised and trained specifically on this dataset, it has an advantage over our proposed approach, which operates in a zero-shot setting.

4.1.3 Implementation Details. For the initial retriever (Section 3.2.1), the input length of claims and textual evidences is truncated to 77 tokens and evidence images are resized to 224×224×3 pixels. We use Mistral-OpenOrca (7B)[4] [17] for text retrieval and fact verification with text evidence due to its demonstrated superior performance over many advanced LLMs, such as LLaMA2, and its exceptional capabilities across various natural language processing tasks [17]. For image retrieval, we use InstructBLIP (Flan-T5-xl)[5] [7] as it has ranked first in SEED-Bench [20] leaderboard by achieving the best performance based on the averaged results across nine evaluation dimensions [20]. Finally, we use the LLaVA v1.6-Mistral (7B)[6] [22] model for fact verification with multimodal evidence. LLaVA was among the top-ranked models on the SEED-Bench leaderboard for question answering tasks [19]. Additionally, since LLaVA v1.6-Mistral utilizes the same language model (Mistral), it makes fact verification with text and multimodal evidence comparable. All models are loaded in 8-bit quantization with a maximum token length of 2048 for Mistral and LLaVA and 512 for InstructBLIP. All experiments were conducted on two NVIDIA A3090, 24-GB GPUs. We set the size of the initial retriever $N = 100$ and choose the best working prompt according to Section 4.2.3 for the experiments. We will make our code publicly available for further research.

4.2 Results for Evidence Retrieval

In this section, we report and discuss the performance of proposed pipeline in comparison to state-of-the-art baseline on evidence retrieval using *MOCHEG* dataset (Section 4.2.1) and annotated data (Section 4.2.2) that addresses the issue of incomplete ground truth

---

Table 1: Precision (Pre), Recall (Rec) and mean Average Precision (mAP) of models for evidence retrieval on MOCHEG dataset with N=100 in percent [%].

Method	K	Text Retrieval			Image Retrieval		
		Pre	*Rec*	*mAP*	*Pre*	*Rec*	*mAP*
MOCHEG	1	27.14	8.63	**27.14**	8.56	6.46	8.56
	2	21.81	13.13	**20.39**	6.89	10.01	8.91
	5	13.70	18.77	**16.23**	4.18	15.40	10.17
	10	9.18	23.59	**15.79**	2.73	19.95	10.89
LVLM4EV	1	26.34	7.44	26.34	9.65	7.13	**9.65**
	2	20.64	10.68	19.00	7.29	10.18	**9.50**
	5	14.10	16.40	14.77	4.19	14.80	**10.42**
	10	9.58	21.04	13.97	2.54	18.01	**10.96**

Table 2: Precision (Pre), Recall (Rec) and mean Average Precision (mAP) of models for evidence retrieval on union annotated set in percent [%]. Mistral-Open-Orca and InstructBlip models are used for text and image retrieval respectively.

Method	K	Text Retrieval			Image Retrieval		
		Pre	*Rec*	*mAP*	*Pre*	*Rec*	*mAP*
MOCHEG	1	80.00	19.96	80.00	80.00	34.02	80.00
	5	64.00	45.17	64.23	56.00	67.05	72.07
	10	47.00	57.83	58.40	43.00	89.27	75.29
LVLM4EV	1	90.00	20.79	**90.00**	90.00	35.45	**90.00**
	5	70.00	42.46	**69.80**	56.00	67.05	**72.77**
	10	60.00	76.75	**70.48**	43.00	89.27	**75.83**

evidences in the retrieval step. We then discuss the impact of different prompts (Section 4.2.3) and re-ranking strategies (Section 4.2.4) for evidence retrieval.

4.2.1 Results on MOCHEG. Similar to *MOCHEG*, for each claim, we evaluate the retrieval performance based on precision, recall and mean average precision (mAP) scores using the top-K retrieved text and image evidences. The results on the test set of *MOCHEG* with $K \in \{1, 2, 5, 10\}$ are shown in Table 1.

The results indicate that mAP decreases by increasing K which is expected since a larger number of retrieved evidences may lead to the inclusion of noisy or redundant documents, reducing the precision at higher ranks. Besides, for example in image retrieval, the average number of gold evidences in the retrieved top-100 is 1.26 which means most of the claims have only one labeled gold image evidence. The results also demonstrate that LVLM4EV outperforms the *MOCHEG* even in zero-shot setting for image retrieval highliting its ability to retrieve images with higher contextual relevance. For text retrieval, LVLM4EV is comparable but falls slightly below the baseline. However, we believe that the assessment of evidence retrieval may not accurately reflect the true capabilities of the models under scrutiny as not all relevant pieces of evidences are appropriately labeled. This is because the ground truth only considers evidence used in *Snopes* and *PolitiFact*. But the large evidence collection might contain more relevant evidences that are disregarded. In the following section, we investigate this issue in more detail.

4.2.2 Results on Annotated Data. While *MOCHEG* [50] is one of the pioneering datasets for evaluating multimodal evidence retrieval, a significant challenge arises from incomplete annotation of relevant ground truth evidence samples. For example, in image retrieval each claim is associated with a set of labeled images meant to serve as ground-truth image evidence, while there exist other visually similar images within the dataset that are left unlabeled (Figure 3). So the results can be misleading as the actual relevant evidences sometimes are not considered due to missing labels. Also, some relevant images are indeed relevant but sometimes do not serve as direct evidence to the claim. This is also the case for text retrieval as shown in Figure 3. This inconsistency undermines the

integrity of the dataset, introduces ambiguity and noise and leads to a diminished performance in identifying relevant evidence samples.

To address these aforementioned issues in multimodal evidence retrieval and ensure fair and accurate assessments of system performance, we provide a subset of *MOCHEG* test set with more complete and accurate annotations for both text and image evidence retrieval. The idea is to ensure all the top-10 relevant candidate evidences are labeled for the sampled claims by considering both baselines. For each modality, we consider three evidence sets: (1) the top-10 evidences $\mathbb{C}'$ from the initial retrieval using *MOCHEG* (*Initial-top10*), (2) the re-ranked evidences $\mathbb{C}''$ (*Re-ranked-top10*) and (3) their union $\mathbb{C}' \cup \mathbb{C}''$ (*Union*). Please note that we only retain distinct candidate evidence when there is overlap for a claim.

Annotation Process. For a fair assessment of the evidence retrieval, an expert manually performed annotations for the *Initial-top10* and *Re-ranked-top10* using ten randomly selected queries from the *MOCHEG* test dataset. Note that the ground-truth labels for the *Union* set can be derived from these annotations. The annotation process is performed on three levels. (1) Entity-level: This level denotes when the candidate evidence is relevant but solely shares the same entity or topic without offering direct support or refutation of the claims. (2) Evidence-level: This is the case if the candidate provides direct evidence to support or reject the claim. (3) Overall: If either of the preceding two levels is deemed true, we designate the total relevancy as true. Figure 4 demonstrates an example of the annotation.

Results. We evaluate the performance of models on the annotated union set while considering the "Overall" labels for each claim. Results for other two types of annotated labels (evidence-level and entity-level) will be additionally provided on GitHub but they generally lead to the same conclusions. The result are shown in Table 2. It clearly shows that LVLM4EV has a significant improvement over the baseline in both image and text retrieval using the revised labels. This confirms our hypothesis that the model retrieves more relevant pieces of evidence compared to the baseline, as illustrated in Figure 3. However, this improvement is not reflected in the numbers for the *MOCHEG* test set.

4.2.3 Prompt Impact. As mentioned in Section 3.2.2, we also experiment with different prompts to investigate the performance for

Figure 3: Qualitative example for evidence retrieval component which also shows the incomplete ground truth issue in the *MOCHEG* dataset. Green border shows the labeled ground truth, Blue border shows evidences with similar content to the ground truth which are left unlabeled and red border shows irrelevant content.

Figure 4: An example of data annotation at entity level and evidence level for image retrieval (top) and text retrieval (bottom).

both image and text retrieval. We report results for the prompts shown in Table 3 using the following template:

- Text prompt for *Mistral*:
  ```
  [prompt] \n
  ### Query: [Q_i]\n
  ### corpus: [S_j]\n
  ### Answer:
  ```

- Image prompt for *InstructBLIP*:
  ```
  [prompt]\n
  ### Query: [Q_i]
  ```

For image retrieval, the prompt asking whether the image and text mentions the same person or topic performs the best. This is because most gold image evidence in the dataset shares the same entity or topic rather than providing direct evidence for the claim (only 1.5% of images were identified as evidence-level in the annotated data).

Table 3: Results of LVLM4EV for both text and image retrieval with different prompts. Best results are highlighted in bold.

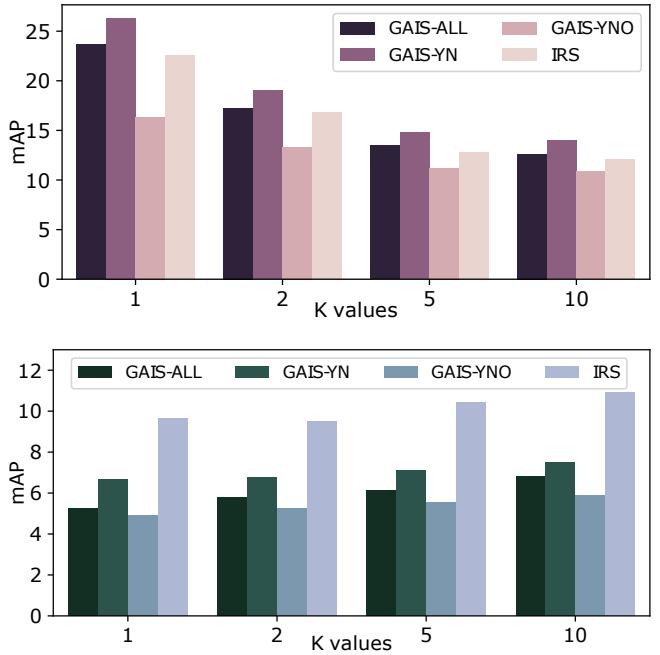| Media	Prompts	K	Pre	Rec	mAP
txt	Is this corpus related to the query? Answer with yes or no.	1	26.19	7.59	**26.19**
		5	14.62	18.31	**15.82**
		10	10.81	24.60	**15.40**
txt	Is query and corpus mentioning the same person or topic? Answer with yes or no.	1	19.94	5.32	19.94
		5	10.81	13.09	11.23
		10	7.71	17.91	10.78
txt	Is this corpus an evidence for the query? Answer with yes or no.	1	21.39	5.83	21.39
		5	17.84	9.22	16.08
		10	7.91	17.95	11.58
img	Does this query describe the image?	1	7.71	5.36	7.18
		5	3.66	13.27	8.54
		10	2.54	17.92	9.22
img	Based on the query below, is it related to the image?	1	6.89	5.20	6.89
		5	3.73	13.47	8.46
		10	2.59	18.90	9.26
img	Is this image and text query mentioning the same person or topic?	1	9.65	7.13	**9.65**
		5	4.19	14.80	**10.42**
		10	2.54	18.16	**10.96**

Figure 5: Mean average precision (mAP) of LVLM4EV using different re-ranking strategies (Section 3.2.3) for top: text retrieval and bottom: image retrieval.

While other similar prompts may also be useful in practice, we use this one for our experiments due to its superior results. In text retrieval, prompt that directly assess the relevance of the corpus to the claim demonstrate better performance compared to others as text evidence usually carries key relevancy information and serve as the evidence to the claim. However, given that documents are segmented into sentences and evidence typically comes from the comprehension of long paragraphs rather than single sentence, the prompt directly asking if a corpus is an evidence for the claim performs worse than the previous one.

4.2.4 Impact of Re-ranking Strategies. Figure 5 illustrates a comparison of the mAP for various re-ranking strategies, as referenced in Section 3.2.3. The results indicate that the CLIP model score proves superior for image retrieval, while the GAIS-YN scores from LLM outperforms others in text retrieval. This discrepancy is likely attributed to the inherent hallucination issue associated with LLMs, wherein the model may generate outputs that do not necessarily match the intended prompts. Accordingly, the model may assign probabilities to irrelevant tokens that are not part of the correct answer sets. Incorporating these redundant probabilities during the normalization process (as in GAIS-All and GAI-YNO) introduces noise, thereby affecting the final score and degrading the final ranking performance.

4.3 Results for Fact Verification

Table 4 presents the performance of various models across the *MOCHEG* test set on fact verification task based on micro F1-score (F1). To evaluate the impact of each type of evidence for classification, we design ablated models by considering the text evidence only or multimodal evidence. We have used one-level prompting for text evidence and two-level prompting for multimodal evidence

as mentioned in Section 3.3.1 Our findings reveal that fact verification utilizing multimodal gold evidences outperformed its text-only counterparts. This suggests that while image evidences alone may only share the same entity or topic and have less impact in the final verification task, together with text, they contain a wealth of information and serve as more concise indicators within the dataset. For example, claims that start with phrases like *"A photograph of…"* or *"The image shows…"* are multimodal, meaning an important part of the information is contained within the image modality. Therefore, fact verification based on multimodal pairs provides more significant gain than text only. The results also show the various model performances in comparison to the state of the art. Our proposed approach achieves an improvement over the *MOCHEG* baseline in the overall F1-score, surpassing it by 4.3% with respect to the multimodal gold evidences, despite operating in zero-shot manner. This suggests the potential of proposed method for verification task compared to *MOCHEG* method which relies on fine-tuning for this specific dataset and incorporates a dense information retrieval component. One explanation for this superior performance is that the instruction-tuning data of LLaVA contains totally 158K samples and covers a wide range of multimodal tasks, including question answering and visual reasoning data, which allows it to effectively capture semantic relation between images and text. After conducting an error analysis of the fact verification outcomes, we have observed that the majority of failure cases are due to a model's maximum length constraint. As we need to input the prompt template, claim, and complete textual evidence into the model, this might result in loss of content, particularly

Table 4: Precision (Pre), Recall (Rec) and F1-score (F1) of models for fact verification task on MOCHEG test set. Mistral-Open-Orca model is used for text evidence and LLaVA model is used for multimodal evidence. All retrieved evidences are with K=5

Method	Evidence	Modality	Supported			Refuted			NEI			micro F1
			Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	
LVLM4FV	Retrieved	text	0.639	0.380	0.477	0.393	0.896	0.547	0.352	0.044	0.079	0.440
	Gold	text	0.500	0.677	0.575	0.816	0.406	0.542	0.411	0.472	0.439	0.518
LVLM4FV	Retrieved	multimodal	0.622	0.491	0.549	0.460	0.398	0.428	0.451	0.316	0.372	0.451
	Gold	multimodal	0.453	0.800	0.578	0.858	0.426	0.569	0.422	0.501	0.457	**0.534**
MOCHEG	Gold	multimodal	0.502	0.479	0.490	0.481	0.811	0.604	0.533	0.191	0.282	0.491

Table 5: Precision (Pre), Recall (Rec) and F1-score (F1) of models for fact checking task on Factify validation set. Mistral-Open-Orca model is used for text evidence and LLaVA model is used for multimodal evidence. All retrieved evidences are with K=5

Method	Evidence	Modality	Supported			Refuted			NEI			micro F1
			Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	
LVLM4FV	Retrieved	text	0.533	0.634	0.579	0.507	0.620	0.558	0.492	0.344	0.405	0.515
	Gold	text	0.578	0.608	0.593	0.455	0.806	0.581	0.778	0.438	0.560	0.580
LVLM4FV	Retrieved	multimodal	0.636	0.446	0.524	0.489	0.696	0.575	0.495	0.480	0.487	0.521
	Gold	multimodal	0.636	0.726	0.678	0.485	0.804	0.605	0.521	0.496	0.508	0.595
MOCHEG	Gold	multimodal	0.454	0.688	0.547	0.578	0.670	0.621	0.475	0.193	0.275	0.456
Logically	Gold	multimodal	0.830	0.860	0.850	1.00	1.00	1.00	0.850	0.830	0.840	**0.870**

with lengthy evidence passages. We believe that an increase of the model's maximum length would boost the results, although this adjustment would also cause longer run times. Another type of failure occurs with multimodal claims where the evidence is in a chart image. This demonstrates that LLaVA struggles with text recognition and information extraction from charts, as mentioned in [20] as a common issue within similar models.

4.4 Generalization Study

We conduct evaluations on the *Factify* dataset to analyze the robustness and generalizability of employed methods in a cross-domain context which is crucial for real-world misinformation detection. Misinformation varies widely across platforms, topics, and languages and a model that excels only on specific data may fail with new examples.

Table 5 presents the performance of various models across the *Factify* validation set on fact verification task with different types of evidence. The results demonstrate that while the performance of proposed method may fall short of *Logically*, which is fine-tuned specifically for the dataset, it notably surpasses *MOCHEG* that has been trained on a different benchmark and fails to generalize to the new domains, topics, and characteristics in *Factify*. This indicates the robustness and efficacy of our approach in adapting to unseen data domains without the need for extensive fine-tuning.

5 CONCLUSIONS

In this paper, we have presented a pipeline for multimodal misinformation detection, introducing a novel re-ranking approach for evidence retrieval using both LLMs and LVLMs (LVLM4EV). The claim and retrieved evidence samples (texts and images) serve as the input for LVLM-based fact verification (LVLM4FV). We demonstrated the capabilities of LLMs and LVLMs for evidence retrieval and fact verification in zero-shot settings, employing them as rankers for text and image retrieval and used a prompting strategy which allows us to extract ranking scores and determine the veracity of claims. Furthermore, we tackled the issue of incomplete and inaccurate ground-truth labels for evidence retrieval tasks and provided an improved annotation to enable a more meaningful system evaluation. Our experiments on two datasets have demonstrated that our proposed zero-shot approach outperforms a supervised baseline for multimodal misinformation detection. Moreover, it has shown much better generalization capabilities as it is not fine-tuned on relatively small domain- and topic-specific datasets.

In future work, we will focus on the interpretability of the pipeline. We aim to achieve this by obtaining explanations directly from the model itself, enabling us to discern the specific contributions of different evidence components towards the final decision-making process. Furthermore, Given the lack of high-quality annotations in current multimodal datasets, a comprehensive, well-annotated, large-scale dataset is crucial for advancing research in multimodal evidence retrieval for misinformation detection. Finally, fine-tuning suitable L(V)LMs for evidence retrieval and fact verification can further improve performance.

REFERENCES

[1] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023. ChartCheck: An Evidence-Based Fact-Checking Dataset over Real-World Chart Images. *arXiv preprint* abs/2311.07453 (2023). https://doi.org/10.48550/ARXIV.2311.07453 arXiv:2311.07453

[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[3] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. *Hugging Face* (2023).

[4] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2015. Verifying Multimedia Use at MediaEval 2015. (2015). http://ceur-ws.org/Vol-1436/Paper4.pdf

[5] Lia Bozarth and Ceren Budak. 2020. Toward a Better Performance Evaluation Framework for Fake News Classification. In *International AAAI Conference on Web and Social Media, ICWSM 2020, Virtual Event, June 8-11, 2020.* AAAI Press, 60–71. https://ojs.aaai.org/index.php/ICWSM/article/view/7279

[6] Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2023, TICC, Taipei, Taiwan, October 31-November 3, 2023.* IEEE, 846–853.

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.* http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019.* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[9] Yi R. Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Virtual Event, August 1-6, 2021.* Association for Computational Linguistics, 1683–1698. https://doi.org/10.18653/V1/2021.ACL-LONG.133

[10] Jie Gao, Hella-Franziska Hoffmann, Stylianos Oikonomou, David Kiskovski, and Anil Bandhakavi. 2021. Logically at the Factify 2022: Multimodal Fact Verification. *arXiv preprint* abs/2112.09253 (2021). arXiv:2112.09253 https://arxiv.org/abs/2112.09253

[11] Daniel Gerber, Diego Esteves, Jens Lehmann, Lorenz Bühmann, Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, and René Speck. 2015. DeFacto - Temporal and multilingual Deep Fact Validation. *The Journal of Web Semantics* 35 (2015), 85–101. https://doi.org/10.1016/J.WEBSEM.2015.08.001

[12] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint* abs/2301.04246 (2023). https://doi.org/10.48550/ARXIV.2301.04246 arXiv:2301.04246

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[14] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging ChatGPT for efficient fact-checking. *PsyArXiv* 3 (2023).

[15] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *AAAI Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, Vancouver, Canada, February 20-27, 2024.* AAAI Press, 22105–22113. https://doi.org/10.1609/AAAI.V38I20.30214

[16] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint* abs/2112.09118 (2021). arXiv:2112.09118 https://arxiv.org/abs/2112.09118

[17] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint* abs/2310.06825 (2023). https://doi.org/10.48550/ARXIV.2310.06825 arXiv:2310.06825

[18] Manvi Kamboj, Christian Hessler, Priyanka Asnani, Kais Riani, and Mohamed Abouelenien. 2021. Multimodal Political Deception Detection. *IEEE MultiMedia* 28, 1 (2021), 94–102. https://doi.org/10.1109/MMUL.2020.3048044

[19] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023. SEED-Bench-2: Benchmarking Multimodal Large Language Models. *arXiv preprint* abs/2311.17092 (2023). https://doi.org/10.48550/ARXIV.2311.17092 arXiv:2311.17092

[20] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint* abs/2307.16125 (2023). https://doi.org/10.48550/ARXIV.2307.16125 arXiv:2307.16125

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA.* PMLR, 19730–19742. https://proceedings.mlr.press/v202/li23q.html

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.* http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

[24] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir D. Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022.* AAAI Press, 651–662. https://ojs.aaai.org/index.php/ICWSM/article/view/19323

[25] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir D. Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022.* AAAI Press, 651–662. https://ojs.aaai.org/index.php/ICWSM/article/view/19323

[26] Shreyash Mishra, Suryavardan S, Amrit Bhaskar, Parul Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, and Asif Ekbal. 2022. FACTIFY: A Multi-Modal Fact Verification Dataset. In *Workshop on Multi-Modal Fake News and Hate-Speech Detection co-located with the AAAI Conference on Artificial Intelligence, DE-FACTIFY @ AAAI, 2022, Virtual Event, Vancouver, Canada, February 27, 2022.* CEUR-WS.org. https://ceur-ws.org/Vol-3199/paper18.pdf

[27] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *The International Conference on Language Resources and Evaluation, LREC 2020,*. European Language Resources Association, 6149–6157. https://aclanthology.org/2020.lrec-1.755

[28] OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. (2022). https://openai.com/blog/chatgpt/

[29] OpenAI. 2023. GPT-4 Technical Report. *arXiv* abs/2303.08774 (2023). https://doi.org/10.48550/ARXIV.2303.08774 arXiv:2303.08774

[30] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content Based Fake News Detection Using Knowledge Graphs. In *The International Semantic Web Conference, ISWC 2018, Monterey, CA, USA, October 8-12, 2018.* Springer, 669–683. https://doi.org/10.1007/978-3-030-00671-6_39

[31] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online Information Review* 43, 1 (2019), 72–88. https://doi.org/10.1108/OIR-03-2018-0101

[32] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018.* Association for Computational Linguistics, 231–240. https://doi.org/10.18653/V1/P18-1022

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning, ICML 2021, Virtual Event, 18-24 July, 2021.* PMLR, 8748–8763. http:

//proceedings.mlr.press/v139/radford21a.html

[34] Frédéric Rayar, Mathieu Delalandre, and Van-Hao Le. 2022. A large-scale TV video and metadata database for French political content analysis and fact-checking. In *International Conference on Content-based Multimedia Indexing, CBMI 2022, Graz, Austria, September 14 - 16, 2022*. ACM, 181–185. https://doi.org/10.1145/3549555.3549557

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/V1/D19-1410

[36] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning. In *The ACM International Conference on Information and Knowledge Management, CIKM '21, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. ACM, 1640–1650. https://doi.org/10.1145/3459637.3482440

[37] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[38] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1409.1556

[39] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*. IEEE, 39–47. https://doi.org/10.1109/BIGMM.2019.00-44

[40] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management* 58, 1 (2021), 102437. https://doi.org/10.1016/J.IPM.2020.102437

[41] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 14918–14937. https://aclanthology.org/2023.emnlp-main.923

[42] Sahar Tahmasebi, Sherzod Hakimov, Ralph Ewerth, and Eric Müller-Budack. 2023. Improving Generalization for Multimodal Fake News Detection. In *ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*. ACM, 581–585. https://doi.org/10.1145/3591106.3592230

[43] Reuben Tan, Bryan A. Plummer, and Kate Saenko. 2020. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 2081–2106. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.163

[44] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018*. Association for Computational Linguistics, 809–819. https://doi.org/10.18653/V1/N18-1074

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint* abs/2302.13971 (2023). https://doi.org/10.48550/ARXIV.2302.13971 arXiv:2302.13971

[46] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, 2017*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 422–426. https://doi.org/10.18653/V1/P17-2067

[47] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2018, London, UK, August 19-23, 2018*. ACM, 849–857. https://doi.org/10.1145/3219819.3219903

[48] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* 2022 (2022). https://openreview.net/forum?id=yzkSU5zdwD

[49] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *arXiv preprint* abs/2306.09265 (2023). https://doi.org/10.48550/ARXIV.2306.09265 arXiv:2306.09265

[50] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. In *International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, 2733–2743. https://doi.org/10.1145/3539618.3591879

[51] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint* abs/2306.13549 (2023). https://doi.org/10.48550/ARXIV.2306.13549 arXiv:2306.13549

[52] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In *International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206973

[53] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. *arXiv preprint* abs/2301.12867 (2023). https://doi.org/10.48550/ARXIV.2301.12867 arXiv:2301.12867

[54] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-Checking Meets Fauxtography: Verifying Claims About Images. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 2099–2108. https://doi.org/10.18653/V1/D19-1216