Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments

Zhenrui Yue **Huimin Zeng** **Lanyu Shang** **Yifan Liu**
Yang Zhang **Dong Wang**
University of Illinois Urbana-Champaign
{zhenrui3, huiminz3, lshang3, yifan40, yzhangnd, dwang24}@illinois.edu

Abstract

The rapid propagation of misinformation poses substantial risks to public interest. To combat misinformation, large language models (LLMs) are adapted to automatically verify claim credibility. Nevertheless, existing methods heavily rely on the embedded knowledge within LLMs and / or black-box APIs for evidence collection, leading to subpar performance with smaller LLMs or upon unreliable context. In this paper, we propose retrieval augmented fact verification through the synthesis of contrasting arguments (RAFTS). Upon input claims, RAFTS starts with evidence retrieval, where we design a retrieval pipeline to collect and re-rank relevant documents from verifiable sources. Then, RAFTS forms contrastive arguments (i.e., supporting or refuting) conditioned on the retrieved evidence. In addition, RAFTS leverages an embedding model to identify informative demonstrations, followed by in-context prompting to generate the prediction and explanation. Our method effectively retrieves relevant documents as evidence and evaluates arguments from varying perspectives, incorporating nuanced information for fine-grained decision-making. Combined with informative in-context examples as prior, RAFTS achieves significant improvements to supervised and LLM baselines without complex prompts. We demonstrate the effectiveness of our method through extensive experiments, where RAFTS can outperform GPT-based methods with a significantly smaller 7B LLM[1].

1 Introduction

As the scope of social media and digital forums continue to expand, increasing amount of misinformation has been observed across multiple platforms (e.g., Twitter), posing risks to public interest (Chen et al., 2022). Therefore, fact-checking methods are proposed to prevent the spreading of false information before it leads to severe consequences (Litou et al., 2017; Hassan et al., 2017; Shu et al., 2017). For example, online fact-checking services (e.g., Snopes[2]) employ professional fact-checkers to identify instances of misinformation. Nevertheless, human fact-checking involves a significant amount of manual work, proving to be less efficient confronted with the vast volume of misinformation, particularly as it evolves and spreads online (Micallef et al., 2020; Nakov et al., 2021).

To perform fact-checking at scale, automated methods have emerged by leveraging large language models (LLMs) (Shu et al., 2022; Yang et al., 2022; Yue et al., 2023; Choi and Ferrara, 2024). For example, RARG proposes to train and align LLMs for generating faithful explanations upon detected misinformation (Yue et al., 2024). Despite their effectiveness, these methods typically require extensive training data and may demonstrate performance deterioration upon domain / concept shifts (Zhu et al., 2022; Nan et al., 2022; Gu et al., 2023; Shang et al., 2024a). Moreover, many models are unaware of external evidence / knowledge and must be frequently re-trained to incorporate up-to-date domain knowledge for accurate fact-verification (Izacard and Grave, 2021; Borgeaud et al., 2022; Yue et al., 2023).

As a solution, evidence-based fact-checking methods are proposed to collect evidence (e.g., documents, graphs etc.), followed by extracting relevant information and assessing the credibility of input claims through LLMs (Koloski et al., 2022; Kou et al., 2022b; Shang et al., 2022a; Wu et al., 2022b; Zhang and Gao, 2023; Wang and Shu, 2023; Liu et al., 2024). An example is FOLK, which leverages first-order-logic to construct sub-claims and perform question answering-based verification to generate predictions and explanations (Wang and Shu, 2023). Yet current approaches rely on the assumption that input claims can be decomposed

[1]Our implementation is publicly available at https://github.com/yueeeeeeee/RAFTS.

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

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10331–10343

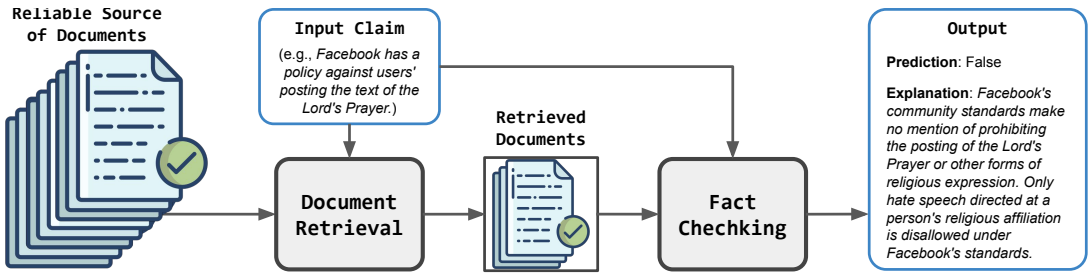

Figure 1: Our retrieval augmented generation framework for fact verification.

into a series of predicates (i.e., sub-claims) through complex prompts. Moreover, they depend on the embedded knowledge within LLMs and / or black-box APIs (e.g., SerpAPI[3]) to collect external information, leading to subpar performance with smaller LLMs or provided with unreliable evidence (Zhang and Gao, 2023; Wang and Shu, 2023).

Consequently, we consider a retrieval augmented generation (RAG) framework designed to extract relevant information from reliable documents (i.e., Wikipedia, scholarly articles etc.), where the extracted information can be used as supporting facts to assess the claim credibility through LLMs. That is, given the input statement, our first objective is to retrieve (and optionally re-rank) relevant documents among an extensive collection of documents from verifiable sources. Subsequently, we utilize the retrieved documents to fact-check the input claim, aiming to either confirm the input or uncover opposing information that identifies misinformation. Our framework is visually illustrated in Figure 1, where the RAG-based fact verification framework retrieves relevant documents, and then generates both prediction and explanation regarding the validity of the input statement.

To this end, we propose retrieval augmented fact verification through the synthesis of contrastive arguments (RAFTS), which effectively retrieves relevant documents and performs few-shot fact verification using pretrained LLMs. RAFTS is structured into three components: (1) demonstration retrieval, where informative in-context examples are collected to improve fact-checking performance; (2) document retrieval, in which we design a retrieve and re-rank pipeline to accurately identify relevant documents for input claims; and (3) few-shot fact verification through the synthesis of contrasting arguments. Unlike current approaches, RAFTS formulates supporting and opposing arguments derived from the facts within the collected documents. Combined the informative in-context examples, RAFTS demonstrates enhanced fact-checking performance and consistently generates high-quality explanations. To validate the effectiveness of RAFTS, we adopt multiple benchmark datasets and perform extensive experiments on both document retrieval and fact verification. Experiment results highlight the effectiveness of the proposed approach, where RAFTS can outperform state-of-the-art methods even with a significantly smaller LLM (e.g., Mistral 7B).

We summarize our contributions:

1. We propose a RAG-based framework, where relevant documents are retrieved from reliable sources to fact-check input claims.

2. We design RAFTS in three key components: demonstration retrieval, document retrieval and in-context prompting. RAFTS identifies informative examples and relevant documents, followed by synthesizing contrastive arguments for fine-grained fact-checking.

3. We show the effectiveness of RAFTS by experimenting on document retrieval and fact verification tasks. Both quantitative and qualitative results demonstrate that RAFTS can outperform state-of-the-art methods in fact verification and explanation generation.

2 Related Work

2.1 Large Language Models and Retrieval Augmented Generation

Recent advancements in large language models (LLMs) have shown significantly enhanced capabilities in language comprehension and generation (Raffel et al., 2020; Brown et al., 2020; Wei et al., 2021; Ouyang et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI, 2023; Jiang et al., 2024). Due to the vast number of parameters and extensive quantity of pretraining corpora,

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

LLMs can embed global knowledge within their parameters, and thus achieve significant performance improvements across diverse applications (OpenAI, 2023; Penedo et al., 2023; Sun et al., 2023). However, LLMs often fail to capture fine-grained knowledge and frequently generate inaccurate or fabricated information (also known as hallucination) (Peng et al., 2023; Rawte et al., 2023). To access up-to-date knowledge without costly retraining, retrieval augmented generation (RAG) has been proposed to generate text based on collected documents from verifiable sources (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Izacard et al., 2022; Shi et al., 2023; Ram et al., 2023; Wang et al., 2023a). For example, Self-RAG can dynamically fetch external documents to generate contents through the usage of special tokens for retrieval and reflection (Asai et al., 2023). Nevertheless, current RAG methods remain under-explored for fact verification, particularly regarding accurate evidence retrieval and fine-grained classification (Wang and Shu, 2023; Zhang and Gao, 2023). As such, our work studies retrieval augmented fact verification, which gathers evidence from reliable sources and integrates contrasting opinions to achieve fine-grained fact verification.

2.2 Fact Verification and Misinformation Detection

Fact verification methods can generally be divided into two main categories: (1) content-based approaches, where machine learning models predict and reason over input contents (e.g., text) to identify misinformation (Yue et al., 2022; Jiang et al., 2022; Yue et al., 2023; Chen and Shu, 2023a; Liu et al., 2023; Mendes et al., 2023; Huang et al., 2024). Incorporating additional attributes / modalities such as image and propagation paths can further enhance fact verification performance (Shang et al., 2021; Santhosh et al., 2022; Shang et al., 2022b; Wu et al., 2022c; Zhou et al., 2023; Yao et al., 2023; Qu et al., 2024); (2) evidence-based approaches, which involve gathering external knowledge (e.g., knowledge graphs or document pieces) as evidence to validate input claims and identify false information (Kou et al., 2021, 2022a; Wu et al., 2022a; Yang et al., 2022; Shang et al., 2022c; Xu et al., 2022; Zhao et al., 2023; Chen et al., 2023; Wang and Shu, 2023; Yue et al., 2024; Shang et al., 2024b). For example, HiSS adopts hierarchical step-by-step prompting with off-the-shelf LLMs and black-box question answering (QA) pipelines to perform few-shot fact verification (Zhang and Gao, 2023). However, state-of-the-art fact verification methods primarily concentrate on improving accuracy via sophisticated prompts and / or intrinsic knowledge of LLMs, causing performance degrade upon smaller LLMs or domain shifts (Wang and Shu, 2023; Pelrine et al., 2023; Chen and Shu, 2023b). Therefore, we concentrate on retrieval augmented fact verification by collecting relevant documents from reliable sources, enabling LLMs to augment their knowledge base for claim verification. Furthermore, we exploit in-context prompting by learning from demonstrations and synthesizing contrastive arguments, and thus significantly improves fact-checking performance.

3 Preliminary

We consider the following problem setup: given input claim x (with label y) and k-shot demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$, we aim to: (1) retrieve a set of m documents $\{d_i\}_{i=1}^{m}$ that provide relevant information to be used as supporting evidence; and (2) generate label $\hat{y}$ and explanation e based on the input x, k-shot examples $\{(x_i, y_i)\}_{i=1}^{k}$ and retrieved evidence $\{d_i\}_{i=1}^{m}$. For each input x, we leverage a pretrained embedding model f_{embed} to adaptively retrieve demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$, whereas a retrieval model is learnt to predict $\{d_i\}_{i=1}^{m}$ and provide relevant information from verifiable sources. Based on the retrieved examples and documents, the predicted $\hat{y}$ should ideally match the ground truth label y. In addition, the generated explanation e should demonstrate desirable properties (e.g., factuality), see example in Figure 1. We elaborate our settings in the following.

Input & Output: Given a dataset with train and test splits $\mathcal{X}^{\text{train}}$ and $\mathcal{X}^{\text{test}}$, we denote the document retrieval pipeline as f_{retrieve} and the LLM-based fact-checking model as f_{check}. Formally, our framework consists of two sub-problems in information retrieval (i.e., evidence collection) and fact verification (i.e., prediction and explanation), with each of the problem defined below:

- *Document Retrieval*: Given input claim x, human annotated document d and a collection of n documents $\{d_i\}_{i=1}^{n}$ (with $d \in \{d_i\}_{i=1}^{n}$), our objective is to learn a retrieval model f_{retrieve} that ranks the claim-document pair with the highest score ($f_{\text{retrieve}}(x, d) =$

$\max\{f_{\text{retrieve}}(x, d_i)\}_{i=1}^{n}$). During training, input x and d can be used to learn f_{retrieve}. In inference, we collect a subset of m documents $\{d_i\}_{i=1}^{m}$ for fact verification, where $m \ll n$.

- *Fact Verification*: Subsequently, we leverage both input x and collected documents $\{d_i\}_{i=1}^{m}$ from the previous step and utilize f_{check} to generate: (1) prediction $\hat{y}$ on the input credibility; and (2) explanation e on the reasoning of the prediction. To perform in-context prompting, we incorporate k-shot examples $\{(x_i, y_i)\}_{i=1}^{k}$ from $\mathcal{X}^{\text{train}}$ as input ($x_i \neq x$). In other words, $\hat{y}, e = f_{\text{check}}(\{(x_i, y_i)\}_{i=1}^{k}, \{d_i\}_{i=1}^{m}, x))$.

Learning: Our retrieval pipeline f_{retrieve} is parameterized by θ. To learn f_{retrieve}, we maximize the score of the sampled input-document pair (x, d). That is, we minimize the expected loss $\mathcal{L}$ over $\mathcal{X}^{\text{train}}$: $\min_\theta \mathbb{E}_{(x,d)\sim\mathcal{X}^{\text{train}}}[\mathcal{L}(\theta, (x, d))]$. Meanwhile, the fact-checking model f_{check} (i.e., pretrained LLM) remains unchanged to minimize training expenses. To optimize fact-checking performance of f_{check}, we employ a lightweight embedding model f_{embed} to select informative in-context demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$, we elaborate the details in Section 4.1.

4 Methodology

4.1 In-Context Demonstrations

Current LLM-based approaches for fact verification utilize sophisticated prompts to identify misinformation, but depend on carefully designed prompts and *static* in-context demonstrations (Wei et al., 2022; Zhang and Gao, 2023). Nevertheless, the classification criteria often vary from domain to domain, causing performance drops when identical prompts are applied across different contexts (as we show in Section 5). In addition, diverse and informative examples are found to be helpful for performance, in particular for smaller yet more efficient LLMs (Liu et al., 2021; Zhang et al., 2022; Levy et al., 2023; Li and Qiu, 2023). As such, we design a retrieval pipeline to select in-context demonstrations, thereby enhancing the fact-checking performance and mitigating performance deterioration issues across domains.

We formulate the in-context learning (ICL) problem as follows. Provided with k-shot examples $\{(x_i, y_i)\}_{i=1}^{k}$, we prompt a pretrained LLM with them as demonstrations to generate the fact-checking prediction $\hat{y}$ given input x:

$$\hat{y} = \arg\max_y f_{\text{check}}(y|\{(x_i, y_i)\}_{i=1}^{k}, x), \quad (1)$$

with f_{check} returning the output probabilities of the LLM. The prediction can be obtained by selecting the output with the highest probability conditioned on the provided in-context examples and input claim. In contrast to existing prompting methods, in RAFTS, LLM receives the task description via in-context examples. As a result, the performance of fact verification is highly sensitive to the selection of $\{(x_i, y_i)\}_{i=1}^{k}$. To this end, we design a simple and efficient example retrieval pipeline, which is designed to choose semantically similar examples from the training set to maximize the relevance and informativeness of demonstrations $\{(x_i, y_i)\}_{i=1}^{k}$ during in-context learning.

Specifically, we adopt a pretrained embedding model, denoted with f_{embed} (kept frozen in our RAFTS framework). The objective of our retrieval pipeline is to identify a set of k examples $\{(x_i, y_i)\}_{i=1}^{k}$ for each claim x, with:

$$\{(x_i, y_i)\}_{i=1}^{k} = \text{topk}(\{\text{sim}(f_{\text{embed}}(x), f_{\text{embed}}(x_i))\}_{i=1}^{|\mathcal{X}^{\text{train}}|}), \quad (2)$$

where topk returns k largest elements from the given set (i.e., claims with highest similarity to x), while sim represents the cosine similarity function (i.e., $\text{sim}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$). In essence, Equation (2) encodes the examples from the training set $\mathcal{X}^{\text{train}}$ (only needs to be performed once), and then identifies the top-k nearest elements by computing the highest cosine similarity scores. Overall, our in-context example retrieval pipeline performs similarity-based filtering to select semantically relevant examples, and thus optimizes the prior distribution for in-context learning. We additionally apply similarity thresholding by establishing a minimum cosine similarity of 0.5, and set $k = 10$ as the maximum number of demonstrations. In our implementation, SimCSE-RoBERTa is employed as the embedding function f_{embed} to encode input claims (Liu et al., 2019; Gao et al., 2021).

4.2 Document Retrieval

The majority of RAG and fact-checking methods utilize sparse retrieval algorithms, dense retrieval models or third-party APIs to collect relevant documents (Izacard and Grave, 2021; Izacard et al., 2022; Ram et al., 2023; Wang and Shu, 2023;

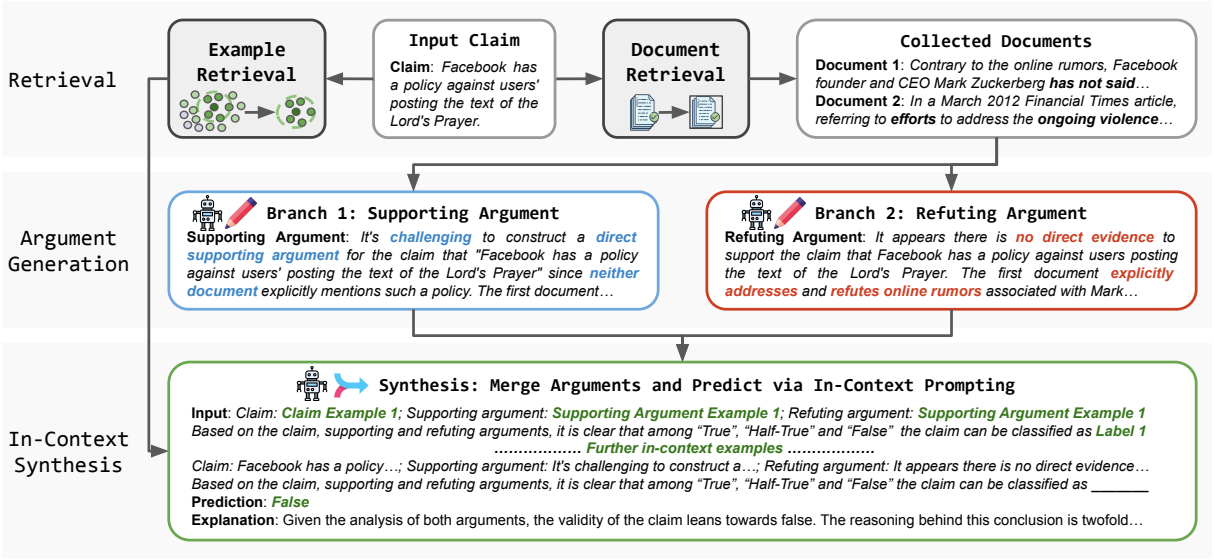

Figure 2: The proposed RAFTS, which performs few-shot fact verification by incorporating informative in-context demonstrations and contrastive arguments with nuanced information derived from the retrieved documents.

Zhang and Gao, 2023). While sparse retrieval methods are widely used, they often fall short in delivering optimal retrieval results for knowledge-intensive tasks like fact verification. On the other hand, dense retrieval methods suffer from efficiency issues in processing massive document collections and require extensive annotated data for optimal performance. These constraints render current retrieval approaches less effective for fact verification, where no / limited annotated claim-document pairs are available for training.

Therefore, we propose a two-stage pipeline f_{retrieve} in RAFTS that performs coarse-to-fine retrieval, which improves both computation efficiency and retrieval performance. Specifically, our pipeline includes: (1) sparse retrieval via BM25, which collects a subset $\{d_i\}_{i=1}^{\hat{m}}$ from a large collection of documents; and (2) an dense retrieval model (denoted with θ) that re-ranks and refines the selection of retrieved documents. Based on x, the first step narrows down to a subset from a much larger collection $\{d_i\}_{i=1}^{n}$, while the learnable dense retriever further selects the m most relevant documents $\{d_i\}_{i=1}^{m}$ to verify input validity. Although BM25 may retrieve less relevant or even irrelevant elements, we note that with proper selection of $\hat{m}$, the desired documents tend to be found within the retrieved set for most cases. In our implementation, we use $\hat{m} = 20$ and $m = 5$ to balance the document retrieval performance and efficiency.

Following the sparse BM25 retrieval, we elaborate the learning of our dense retrieval model. To enhance re-ranking performance with limited annotated data, we exploit the BM25 scores as a coarse estimation of claim-document relevance. That is, we utilize the BM25 scores from the previous retrieval stage, combined with a limited collection of annotated examples, to train the dense retriever model. Specifically for claim-document pair (x, d), we sample l positive documents $\{d_i^p\}_{i=1}^{l}$ and l negative documents $\{d_i^n\}_{i=1}^{l}$ based on the BM25 and inverse BM25 scores, which avoids introducing extensive noise in training. Using the sampled documents, we construct a ranking loss to expand the margin between document d and the highest ranked document from $\{d_i^p\}_{i=1}^{l}$ (i.e., $f_{\text{den}}(x,d) - \max(\{f_{\text{den}}(x,d_i^p)\}_{i=1}^{l})$). In addition, we enhance the relevance between input-document pairs by imposing a penalty when the margin is below threshold τ. Furthermore, our training objective incorporates a contrastive term derived from InfoNCE (Chen et al., 2020; Yue et al., 2024), which improves the relevance estimation between input-document pairs by 'pushing away' negative documents. Overall, the optimization objective is:

$$\mathbb{E}_{(x,d)\sim\mathcal{X}}[\max(0, \max(\{f(x,d_i^p)\}_{i=1}^{l}) - f(x,d) + \tau) - \lambda \frac{\exp(f(x,d))}{\exp(f(x,d)) + \sum_{2l}\exp(f(x,d_i))}], \tag{3}$$

where $\sum_{2l}\exp(f(x,d_i))$ represents the exponential sum of both positive examples $\{d_i^p\}_{i=1}^{l}$ and negative examples $\{d_i^n\}_{i=1}^{l}$. τ is the ranking margin threshold and λ is a scaling factor. For each pair of

x and d, the first term in Equation (3) becomes relevant when $f_{\text{den}}(x, d)$ does not exceed the highest ranked document (i.e., also known as hard negative) score by τ. Moreover, the contrastive term maximizes exponential score of the input-document pair in contrast to the sum of scores from the sampled documents. Hence, the dense retriever model learns to prioritize highly relevant documents while effectively filtering out those of less relevance to improve fact verification performance.

4.3 Fact Verification by Synthesizing Contrastive Arguments

To facilitate fact verification with LLMs, existing methods leverage intricate templates and techniques such as chain-of-thought (CoT), which decomposes input claims into sub-claims to verify (Wei et al., 2022; Wang and Shu, 2023). Yet when assessing the (sub)-claims, current approaches prompt LLMs to perform binary classification (i.e., true or false), and thus often fail to incorporate nuanced information for fine-grained fact-checking (Zhang and Gao, 2023; Pelrine et al., 2023). Moreover, the extended context created by retrieved demonstrations and documents can impair performance in LLMs with limited context windows or in smaller LLMs. Therefore, we propose a branching approach by generating and synthesizing contrastive arguments, in which we: (1) decompose the fact-checking task into generating supporting and refuting arguments upon input claim and retrieved documents; and (2) learn from informative in-context examples to synthesize the contrasting arguments, which incorporates adaptive prior knowledge and varying viewpoints.

Provided with claim and retrieved documents, our first sub-task involves creating two branches in parallel that generate independent yet varying arguments from two opposing perspectives. In particular, we leverage the text comprehension and summarization capabilities of LLMs and perform instruction prompting to extract relevant facts and generate supporting / refuting arguments. We adopt a simple task description and optimize it to obtain concise, yet accurate arguments within a few sentences. For input x and retrieved documents $\{d_i\}_{i=1}^m$, the generated supporting / refuting arguments are denoted with s and r. Therefore, for a specific example $(x, y) \sim \mathcal{X}$, we enrich the input to (s, r, x, y) by integrating both supporting and refuting arguments. Notably, if no pertinent evidence is found to form the argument, LLMs are instructed to recognize the absence of evidence, as illustrated in branch 1 of Figure 2. Consequently, this allows us to guide LLMs to take both arguments into consideration, facilitating a comprehensive analysis on the claim and its credibility.

Moving to the argument synthesis and inference phase (i.e., in-context synthesis) of our fact-checking framework, we aim to generate accurate prediction on the claim validity by leveraging in-context examples along with the contrasting arguments. Recall our in-context learning framework condition on the k-shot examples $\{(x_i, y_i)\}_{i=1}^k$, we also incorporate the generated arguments and reformulate our inference with:

$$\hat{y} = \arg\max_y f_{\text{check}}(y|\{s_i, r_i, x_i, y_i\}_{i=1}^k, s, r, x), \tag{4}$$

where s_i, r_i are the supporting and refuting arguments for the i-th demonstration. Note that the documents $\{d_i\}_{i=1}^m$ are implicitly included in the arguments and thus no longer used in the prompt. At this point, we adopt the following template for each example in the final prompt:

> Claim: `Claim`
> Supporting argument: `Supporting Arg`
> Refuting argument: `Refuting Arg`
> Based on the claim, its supporting and refuting arguments, it is clear that among `Classes`, the claim should be classified as `Label`.

Here, `Claim`, `Supporting Arg`, `Refuting Arg`, `Classes` are populated with the input claim, supporting and refuting arguments and the set of all classes. For in-context examples, `Label` is filled with the respective example's label, whereas for the target example, `Label` is left blank for prediction. Following the prediction, the explanation is generated in a similar fashion by integrating both arguments and prompting with instruction.

4.4 Summary of RAFTS

Overall, the proposed RAFTS has three components: (1) example retrieval; (2) document retrieval; and (3) in-context fact verification. The first two components are designed to collect relevant demonstrations and supporting documents that provide insightful context information. In the third component, we propose to generate contrasting arguments upon the retrieved documents, followed by incorporating these perspectives in inference to achieve

Model	MS MARCO					Check-COVID				
	N@1 ↑	N@3 ↑	R@3 ↑	N@5 ↑	R@5 ↑	N@1 ↑	N@3 ↑	R@3 ↑	N@5 ↑	R@5 ↑
TFIDF	0.419	0.531	0.613	0.562	0.687	0.266	0.363	0.427	0.385	0.480
BM25	0.665	0.746	0.801	0.760	0.836	0.292	0.395	0.467	0.426	0.545
DPR	0.738	0.793	0.850	0.797	0.903	0.324	0.411	0.477	0.457	0.588
E5	0.796	0.855	0.895	0.865	0.920	0.445	0.584	0.679	0.609	0.741
RAFTS	**0.802**	**0.858**	**0.896**	**0.868**	**0.920**	**0.513**	**0.631**	**0.712**	**0.646**	**0.750**

Table 1: Evaluation results on document retrieval, with best results in bold and second best results underlined.

fine-grained fact verification. With informative in-context examples featuring contrastive arguments, RAFTS can perform well regardless of the LLM size. To demonstrate the efficacy of RAFTS, we perform extensive experiments on multiple fact verification datasets, revealing that RAFTS can surpass state-of-the-art fact-checking methods even with a significantly smaller LLM.

5 Experiments

5.1 Experiment Design

Document Retrieval. Our example retrieval model f_{embed} uses the pretrained SimCSE-RoBERTa (Liu et al., 2019; Gao et al., 2021). The document retrieval model $f_{\text{retriever}}$ consists of BM25 and a dense retriever initialized with E5 (base) (Wang et al., 2022). We adopt MS MARCO and Check-COVID dataset for document retrieval (Nguyen et al., 2016; Wang et al., 2023b). The adopted metrics are NDCG and Recall (i.e., N@k and R@k) with $k \in [1, 3, 5]$. For baselines, we adopt the sparse TFIDF and BM25 and dense models DPR and E5 (Karpukhin et al., 2020; Wang et al., 2022).

Fact Verification. We adopt Mistral 7B and GPT-3.5 as our base LLM (Jiang et al., 2023; Ouyang et al., 2022). We adopt three datasets with varying granularity: LIAR (True / Mostly-true / Half-true / Barely-true / False / Pants-fire)), RAWFC (True / Half-true / False), and ANTiVax (True / False) (Wang, 2017; Yang et al., 2022; Hayawi et al., 2022). For LIAR and RAWFC, we adopt Wikipedia as document sources and use the MS MARCO trained retriever. The document collection for ANTiVax is collected from CORD and LitCOVID, thus we use the Check-COVID trained retriever (Karpukhin et al., 2020; Wang et al., 2020; Chen et al., 2021). Our supervised baselines are dEFEND, SentHAN, SBERT-FC and CofCED (Shu et al., 2019; Ma et al., 2019; Kotonya and Toni, 2020; Yang et al., 2022). GPT-3.5-based methods include GPT-3.5, CoT, ReAct and HiSS (Brown et al., 2020; Wei et al., 2022; Yao et al., 2022; Zhang and Gao, 2023). We adopt macro recall, precision and F1 scores to evaluate fact-checking performance. Automated evaluation is used for explanation quality, including politeness, factuality and claim-relevance following (He et al., 2023).

5.2 Document Retrieval

Our document retrieval results are reported in Table 1. In this table, rows represent retrieval methods and the columns represent different datasets / metrics. For top-1 scores, we use N@1 since top-1 NDCG and Recall scores are equivalent in this case. From the results we observe: (1) RAFTS retriever consistently outperforms baseline methods across all metrics, with an average performance improvement of 3.56% across metrics and datasets. (2) In contrast to sparse retrieval along, the additional dense retriever significantly improves the ranking performance. For example, RAFTS achieves 37.61% performance improvement in Recall@5 compared to BM25 on Check-COVID. (3) The performance gains through our retrieval pipeline are more significant on the Check-COVID dataset. For instance, the relative improvement of NDCG@5 shifts from 0.35% to 8.05% when moving from MS MARCO to Check-COVID. Overall, we find the proposed retrieval pipeline in RAFTS performs well in collecting relevant documents. In addition, the retrieval pipeline proves to be essential for specialized domains like healthcare (e.g., COVID), leading to notable performance improvements.

5.3 Fact Verification

We proceed to discuss our fact verification performance of RAFTS, with the results reported in Table 2. Similarly, methods are depicted in rows and datasets / metrics are represented in columns. The first group of baseline methods comprise supervised approaches (i.e. from dEFEND to CofCED), followed by methods built upon GPT-3.5 (i.e. from GPT-3.5 to HiSS), and the bottom row incorporate

Model	LIAR			RAWFC			ANTiVax		
	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
dEFEND	0.230	0.185	0.205	0.449	0.432	0.440	0.729	0.839	0.781
SentHAN	0.226	0.200	0.212	0.457	0.455	0.456	0.691	**0.984**	0.812
SBERT-FC	0.241	0.221	0.231	0.511	0.460	0.484	0.736	0.951	0.830
CofCED	0.295	0.296	0.295	0.530	0.510	0.520	0.731	0.956	0.828
GPT-3.5	0.291	0.251	0.270	0.485	0.485	0.485	0.771	0.850	0.808
CoT	0.226	0.242	0.237	0.424	0.466	0.444	0.816	0.877	0.845
ReAct	0.332	0.290	0.310	0.512	0.485	0.498	0.820	0.864	0.841
HiSS	0.468	0.313	0.375	0.534	**0.544**	0.539	0.823	0.887	0.853
RAFTS (w/ Mistral 7B)	**0.616**	0.305	0.408	0.626	0.516	0.566	0.839	0.873	0.854
RAFTS (w/ GPT-3.5)	0.471	**0.379**	**0.420**	**0.628**	0.526	**0.573**	**0.886**	0.908	**0.897**

Table 2: Evaluation results on fact verification, with best results in bold and second best results underlined.

RAFTS with Mistral 7B and GPT-3.5. We use P, R and F1 to abbreviate precision, recall and F1 scores[4], and we observe: (1) Both RAFTS variants demonstrates superior fact-checking performance across all datasets. For example, RAFTS with Mistral 7B outperforms the best baseline method in F1 by 8.8%, while RAFTS with GPT-3.5 achieves a significant 12.0% performance gain on F1. (2) RAFTS with GPT-3.5 delivers the best classification results overall. In particular, it leads in precision / recall on two of the three datasets and achieves the highest F1 for all datasets, averaging a 7.8% increase in F1 performance. (3) Notably, RAFTS w/ Mistral 7B backbone is superior than all baseline methods on F1 scores despite its significantly smaller size (7B) than GPT-3.5. This suggests that the proposed in-context synthesis can extract concise yet informative arguments and help LLMs generate accurate predictions on claim credibility. In summary, the RAFTS can outperform state-of-the-art fact verification methods by a substantial margin. Even when utilizing a notably smaller model (Mistral 7B), RAFTS consistently exhibits superior performance, highlighting its efficacy in fact verification.

5.4 Explanation Generation

Based on the fact verification results, the explanations for the prediction can be generated in a similar fashion. To evaluate explanation quality, we benchmark against GPT-3.5 and HiSS, as supervised and the rest LLM methods are not designed to generate fact-checking explanations. We report the explanation quality results in Table 3, with Po., Fa. and Rel. representing politeness, factuality and claim relevance. For RAFTS, we use (M) and (G) to denote the Mistral 7B and GPT-3.5 backbones. Our findings are: (1) both baselines and RAFTS perform well in generating explanations based on the fact-checking predictions, achieving average scores above 0.9 for both politeness and factuality. (2) GPT-3.5-based methods show similar performance regardless of prompting strategies. For instance, the average scores on ANTiVax across metrics are 0.898, 0.909 and 0.915 for GPT-3.5, HiSS and RAFTS (G). (3) Surprisingly, RAFTS with Mistral excels in explanation generation, achieving the highest politeness and factuality scores on all datasets, which may be attributed to the instruction-following capabilities of Mistral 7B. In sum, the explanation evaluation shows that RAFTS can consistently generate high-quality explanations regardless of the choice of the LLM.

Dataset	Method	Po. ↑	Fa. ↑	Rel. ↑
LIAR	GPT-3.5	0.947	0.943	0.846
	HiSS	0.967	0.964	0.848
	RAFTS (M)	**0.973**	**0.969**	**0.883**
	RAFTS (G)	0.969	0.969	0.852
RAWFC	GPT-3.5	0.965	0.949	0.856
	HiSS	0.971	0.955	0.861
	RAFTS (M)	**0.974**	**0.971**	0.757
	RAFTS (G)	0.970	0.960	**0.862**
ANTiVax	GPT-3.5	0.958	0.963	0.774
	HiSS	0.986	0.974	0.768
	RAFTS (M)	**0.987**	**0.976**	**0.800**
	RAFTS (G)	0.986	0.973	0.785

Table 3: Evaluation results on explanation quality, with best results in bold and second best results underlined.

6 Conclusion

In this paper, we propose RAFTS, a novel retrieval augmented fact verification framework. RAFTS

[4]In our experiments, F1 score is favored as it balances the trade-off between precision and recall, thereby offering a more comprehensive performance measure for fact verification.

consists of three key components: (1) example retrieval, which provides informative in-context demonstrations; (2) document retrieval that collects relevant documents from verifiable sources; and (3) in-context prompting, where few-shot fact-checking is performed by considering both informative examples and nuanced information from contrastive arguments. As a result, RAFTS achieves fine-grained fact verification without the need for complex prompting techniques and large-size LLMs. Our experiment results on benchmark datasets highlight the superiority of RAFTS, which consistently outperforms state-of-the-art methods methods in both fact-checking performance and the quality of generated explanations.

7 Limitations

Despite introducing RAFTS for retrieval augmented fact verification, we have not discussed the setting in which the document retrieval domain significantly differs from the fact-checking domain (e.g., using Wikipedia documents to fact-check COVID misinformation), which can cause performance deterioration for domain-generalized applications. Furthermore, we have not examined the robustness and reliability of our example retrieval and document retrieval, which could unlock additional improvements for fact verification. Consequently, we plan to explore a more generalized and domain-adaptive solution for retrieval augmented fact verification as future work.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Canyu Chen and Kai Shu. 2023a. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *arXiv preprint arXiv:2211.05289*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Eun Cheol Choi and Emilio Ferrara. 2024. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 883–886.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiawei Gu, Xuan Qian, Qian Zhang, Hongliang Zhang, and Fang Wu. 2023. Unsupervised domain adaptation for covid-19 classification based on balanced slice wasserstein distance. *Computers in Biology and Medicine*, page 107207.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Yue Huang, Kai Shu, Philip S Yu, and Lichao Sun. 2024. From creation to clarification: Chatgpt's journey through the fake news quagmire. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 513–516.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022a. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25.

Ziyi Kou, Lanyu Shang, Yang Zhang, Christina Youn, and Dong Wang. 2021. Fakesens: A social sensing approach to covid-19 misinformation detection on social media. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 140–147. IEEE.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022b. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *Proc. Int. Joint Conf. Artif. Intell.(IJCAI)*, pages 5087–5093.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media*, 2:19–31.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

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

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE international Conference on big data (big data)*, pages 748–757. IEEE.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. 2024. Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

Nikita Mariam Santhosh, Jo Cheriyan, and Lekshmi S Nair. 2022. A multi-model intelligent approach for rumor detection in social networks. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–5. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, Jin Chen, and Dong Wang. 2022a. A privacy-aware distributed knowledge graph approach to qois-driven covid-19 misinformation detection. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022b. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, pages 3623–3631.

Lanyu Shang, Yang Zhang, Bozhang Chen, Ruohan Zong, Zhenrui Yue, Huimin Zeng, Na Wei, and Dong Wang. 2024a. Mmadapt: A knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4653–4663.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2022c. A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 34–41. IEEE.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2024b. A domain adaptive graph learning framework to early detection of emergent healthcare misinformation on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1408–1421.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in Fake Media Generation and Detection*, pages 215–232. Springer.

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):22–36.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, et al. 2023a. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. *arXiv preprint arXiv:2304.06762*.

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

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv e-prints*, pages arXiv–2212.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

William Yang Wang. 2017. "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)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022a. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2022b. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks. *arXiv preprint arXiv:2210.05498*.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022c. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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 *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. *arXiv preprint arXiv:2403.14952*.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239, Toronto, Canada. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.