VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning

Cheng Niu[1], **Yang Guan**[1], **Yuanhao Wu**[1], **Juno Zhu**[1], **Juntong Song**[1], **Randy Zhong**[1],
Kaihua Zhu[1], **Siliang Xu**[1], **Shizhe Diao**[2], and **Tong Zhang**[3]

[1]NewsBreak
[2]Hong Kong University of Science and Technology
[3]University of Illinois Urbana-Champaign
cheng.niu@newsbreak.com

Abstract

The proliferation of fake news poses a significant threat not only by disseminating misleading information but also by undermining the very foundations of democracy. The recent advance of generative artificial intelligence has further exacerbated the challenge of distinguishing genuine news from fabricated stories. In response to this challenge, we introduce VeraCT Scan, a novel retrieval-augmented system for fake news detection. This system operates by extracting the core facts from a given piece of news and subsequently conducting an internet-wide search to identify corroborating or conflicting reports. Then sources' credibility is leveraged for information verification. Besides determining the veracity of news, we also provide transparent evidence and reasoning to support its conclusions, resulting in the interpretability and trust in the results. In addition to GPT-4 Turbo, Llama-2 13B is also fine-tuned for news content understanding, information verification, and reasoning. Both implementations have demonstrated state-of-the-art accuracy in the realm of fake news detection[1].

1 Introduction

The contemporary digital landscape is rife with the proliferation of fake news, presenting a multi-faceted challenge that undermines public discourse, affects democratic processes, and incites real-world consequences (Vasu et al., 2018). Fake news, characterized by the deliberate dissemination of misinformation, exploits the rapid spread of information online, often outpacing the verification processes that traditional media outlets adhere to.

Fake news detection is defined as the process of identifying and verifying the veracity of news content, employing various computational and manual methods. This process involves distinguishing between true and false information, considering the intent behind the information dissemination, whether it be to mislead, harm, or manipulate public opinion.

Traditional approaches in fake news detection have primarily focused on the linguistic features, also called content-based detection (Castillo et al., 2011; Pérez-Rosas et al., 2018; Giachanou et al., 2019; Przybyla, 2020; Giachanou et al., 2020; Sheikhi, 2021; Kirchknopf et al., 2021; Zhou et al., 2020), which demands laborious feature engineering and is ineffective when the fake news is written by imitating the real news to mislead intentionally. Another line of research is the social context-based method (Qazvinian et al., 2011; Baly et al., 2018; Shu et al., 2020; Monti et al., 2019; Nan et al., 2023), which analyzes the interactions among users, publishers, and posts. However, the feasibility of obtaining user information is challenging for the real-world application. A more recent research approach is the knowledge-based method (Hu et al., 2021; Saeed et al., 2022; Pan et al., 2023; Chen et al., 2023; Liao et al., 2023; Zhang and Gao, 2023; Li et al., 2024), which discerns the veracity of a factual claim by comparing against the evidence retrieved from external knowledge base. However, current approaches often do not fully utilize external resources like the Internet. Additionally, there is a lack of development and optimization of a comprehensive end-to-end pipeline that includes news comprehension, search optimization, verification, and reasoning.

In this paper, we introduce VeraCT Scan, a novel retrieval-augmented system for fake news detection. VeraCT Scan initiates this process by identifying key factual claims across multiple levels of granularity. For each identified factual claim, a comprehensive internet search is conducted to gather relevant information. Then, the veracity of the news is determined by combining this typically disparate and conflicting information, taking into

---

account the varying degrees of source credibility. To increase the trustworthiness of our approach, we underscore the necessity of a transparent reasoning process and provide rationales for each supporting or conflicting judgment.

In summary, our main contributions are:

(i) We introduce VeraCT Scan, that operates across multiple levels of information granularity, employing optimized information retrieval techniques to enhance fake news detection performance.

(ii) We investigate the generation of verification rationales as a means to increase the system's transparency and trustworthiness. Additionally, we address the management of conflicting evidence by leveraging the credibility of sources, thereby improving the reliability of the verification process.

(iii) We conduct a comprehensive evaluation of VeraCT Scan using several fake news detection datasets. Our results demonstrate that the system achieves state-of-the-art performance in news verification tasks, employing both prompted and fine-tuned LLMs.

2 Related Work

In this section, we first review the progress of fake news detection and then discuss the retrieval-augmented generation methods.

2.1 Fake News Detection

Existing fake news detection methods can be categorized into three types: 1) Content-Based Methods (Sheikhi, 2021; Pérez-Rosas et al., 2018; Castillo et al., 2011; Przybyla, 2020; Giachanou et al., 2019; Huang et al., 2023; Giachanou et al., 2020; Kirchknopf et al., 2021; Nakamura et al., 2020; Chen et al., 2023; Zhou et al., 2020) which analyze articles' linguistic features (e.g., text length, punctuation usage, emotion symbols) to differentiate fake news from real ones. However, these methods demand laborious feature engineering and are often ineffective when fake news is written to intentionally mislead readers. 2) Social Context-Based Methods (Shu et al., 2020; Nan et al., 2023; Baly et al., 2018; Monti et al., 2019; Qazvinian et al., 2011) which analyze the interactions among users, publishers, and posts to detect fake news. However, the feasibility of obtaining user information in the news propagation process presents challenges for the real-world applicabil-

ity of this method. 3) Fact-Based Methods (Saeed et al., 2022; Pan et al., 2023; Hu et al., 2021; Xu et al., 2023; Chen et al., 2023; Cheung and Lam, 2023) which focus on factual claim verification by comparing against external knowledge. These methods fall short in providing an end-to-end solution that considers information seeking and the management of conflicting evidence.

Recently, Wang and Shu (2023) leverage large language models (LLMs) to decompose complex claims into sequences of first-order logic, and then guide the search and information verification. Different from their work, we propose a pipeline that includes full steps to classify fake news. Liao et al. (2023) outlines a multi-step process for detecting fake news, which consists of news summarization, searching, and verification. In contrast to their method, we employ LLMs instead of specifically trained encoder-decoder transformers for these natural language processing tasks. In addition, we leverage source credibility to differentiate conflicting evidences, a common challenge in real-world news verification that has rarely been explored in previous research.

2.2 Retrieval-Augmented Generation

The integration of retrieval-augmented generation (RAG) allows LLMs to extend beyond the limits of the training corpus by retrieving information from external knowledge bases before the generative process (Lewis et al., 2020; Chen et al., 2024). RAG has emerged as a solution to overcome the limitations of LLMs including the challenge of out-of-date knowledge and the tendency to produce hallucinations or irrelevant and factually incorrect content. By integrating external, up-to-date documents into the generation process, LLMs can generate more reliable responses across a broad spectrum of tasks, including open-domain question answering (Izacard and Grave, 2021; Trivedi et al., 2023; Li et al., 2023; Xu et al., 2024), dialogue systems (Cai et al., 2019; Peng et al., 2023), and code generation (Zhou et al., 2023b). RAG is also commonly integrated into commercial chatbot products to provide updated information, e.g Perplexity[2] and Gemini[3]. In this paper, we leverage RAG for fake news detection by generating both verdicts and justifications.

---

[2] https://www.perplexity.com
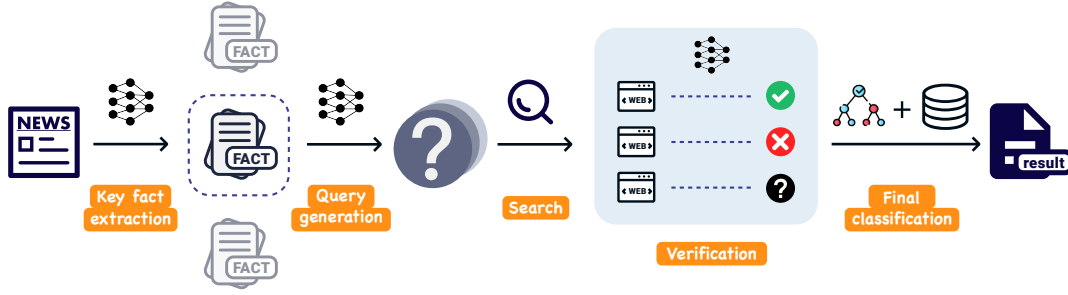[3] https://gemini.google.com

Figure 1: Main workflow of VeraCT Scan. VeraCT Scan includes the following steps: 1) extract key facts from the news to verify; 2) generate search queries for each extracted fact; 3) search; 4) verify the fact based on each search result; 5) aggregate all verifications with a final classification model.

3 Approach

In this paper, the term "claim" refers to the fact stated in a news article. The terms "factual claim extraction" and "fact extraction" are used interchangeably throughout the paper.

Figure 1 shows the main workflow of VeraCT Scan. We prompt GPT-4 Turbo for key fact extraction, query generation, verification, and rationale generation (See Appendix A for prompts being used). These individual components can be easily exchange to other LLMs or search engines. In this work, the outputs from GPT-4 Turbo, supplemented with manual reviews, serve as training data to fine-tune Llama-2 13B (Touvron et al., 2023), enabling it to support these tasks as well. Regarding the search component, we employ both Google and our proprietary in-house news search engine for comprehensive information retrieval.

3.1 Key Fact Extraction

In this paper, we focus on identifying facts at two levels of granularity: (i) the primary fact reported by the news story and (ii) all the salient facts being reported in the news article.

Given that the internet search operates as a stateless module, we instruct the LLM in the prompt to ensure each key fact is self-contained with its information. This approach allows the search function to generate queries for each key fact independently, without relying on additional context.

In line with the previous research (Shahandashti et al., 2024), our manual review has confirmed the high quality of key facts being identified by GPT-4 Turbo.

3.2 Query Generation and Search

When verifying a fact, we prompt GPT-4 Turbo to generate search queries. We allow up to three queries per fact to search the Internet. Subsequently, GPT-4 Turbo assesses the relevance of the results returned by each query. The goal is to identify the shortest sequence of queries that can retrieve all the relevant information. This optimal query sequence is then utilized to fine-tune Llama-2 13B, enabling its query generation capabilities.

We have developed a proprietary search engine designed to support news searches for articles published within the last six months. This search engine is especially effective in searching articles hosted on NewsBreak platform and can be used in NewsBreak APP. To ensure comprehensive search results, we also utilize the Google search API [4].

3.3 Fact Verification and Rationale Generation

Once the search results are retrieved, each fact is evaluated against them. GPT-4 Turbo is prompted to iterate each of the search results, and determine whether the search result supports, conflicts with, or is unrelated to the fact. If the search result aligns with the fact, it is labeled as "support". If it contradicts the fact, it is labeled as "negate". If the fact is not mentioned or only partially mentioned in the search result, the label "baseless" is applied. Besides, a rationale is generated to justify the judgment. A concrete example of our pipeline is shown in Appendix B.

3.4 Source Credibility and Final Decision

When researching a given topic, it is common to encounter conflicting information on the Internet. To avoid bias from single source, multiple sources are used to corroborate each other. Therefore, assessing the credibility of each information source

---

[4]https://developers.google.com/custom-search/v1/overview

is crucial. Mediabiasfactcheck.com is one of the most comprehensive resources for assessing media bias on the internet, offering credibility ratings for over 8,000 news publishers. Similarly, News-Break has developed a proprietary 5-level credibility rating system for more than 30,000 publishers. While NewsBreak's ratings are also based on the credibility of source content, unlike mediabiasfactcheck.com, NewsBreak does not identify the political bias of the sources.

In this paper, NewsBreak's rating systems serves as features to train a LightGBM(Ke et al., 2017) classifier that determines the likelihood of a fact claim being true. Besides, domain and verification flags (i.e. support, negate, or baseless) from each search result are also used as classification features.

3.5 Llama-2 13B Fine Tuning

To enhance service stability, response speed, and reduce costs, Llama-2 13B is fine-tuned to support our fake news detection pipeline.

Dataset   Following previous studies(Zhou et al., 2023a; Taori et al., 2023), we utilize a mixed dataset of diverse tasks for supervised fine-tuning (SFT). Outputs of GPT-4 Turbo from the tasks described above are used as part of the training data. Specifically, we purposely modify some key factual claims being extracted from news articles into fake ones when generating claim verification data set. Besides, the following datasets have also been incorporated into the training set:

1. QA with RAG: GPT-4 generated answers to questions in NewsBreak search logs using knowledge retrieved from our proprietary search engine.
2. WebGLM(Liu et al., 2023): web-enhanced question-answering dataset.
3. No robots(Rajani et al., 2023): a diverse instruction fine-tuning dataset created by skilled human annotators.

The training data distribution is shown in Table 1. This design allows a single model to handle both general question-answering and specialized news verification tasks, resulting in significant reductions in inference costs.

Hyper parameters   To enhance the capability of processing long inputs, we trained the model with RoPE scaling(Su et al., 2021; Liu et al., 2024). Specifically, we adjusted the context window size in SFT to be twice as large as that in the original

Task/Dataset	# Samples	% Samples
Key Fact Extraction	10299	18.52
Query Generation	3000	5.39
Fact Verification	23429	42.12
QA with RAG	8091	14.55
No robots(Rajani et al., 2023)	9500	17.08
WebGLM(Liu et al., 2023)	1300	2.34
Total	55619	100.0

Table 1: The distribution of the fine-tuning data from different tasks/datasets.

Key Task	ROUGE-1	ROUGE-2	ROUGE-L
Key Fact Extraction	0.678	0.497	0.655
Query Generation	0.690	0.503	0.662
Rationale Generation	0.637	0.449	0.600

Table 2: Performance of key tasks.

Llama-2 model, setting it to 8192 tokens, and we set the scaling factor at 2.0. We employed full training with an initial learning rate of 1e-5, and limited the training to 1 epoch. The training process was executed on four NVIDIA A100 GPUs.

3.6 Key Task Evaluations

The end-to-end metrics will be present in Section 5. In this section, we present the performance metrics for the critical components.

With GPT-4 Turbo outputs as the gold standard, we benchmarked the finetuned Llama-2 model on key fact extraction, query generation, and rationale generation. ROUGE scores (Lin, 2004) were employed as the metrics, as shown in Table 2.

For the fact verification accuracy, micro-F1 score was employed as the metric. According to human review, GPT-4 Turbo achieved a score of 0.805, while the finetuned Llama-2 model achieved 0.759.

4 Experimental Settings

In this section, we conduct comprehensive fake news detection benchmarks using multiple datasets.

4.1 Datasets

BuzzFeedNews(Silverman et al., 2016)   This dataset consists of news articles shared on Facebook during the week surrounding the 2016 U.S. election. It includes data collected from nine different news agencies, spanning from September 19 to 23, and then September 26 and 27. Each article was fact-checked by a team of five BuzzFeed journalists. The articles are categorized under four labels: mostly true, mostly false, a mix of true and false, and no factual content. In line with Shu et al. (2019), we utilize the subset of 182 news articles

for our benchmark. Each article in this subset has been assigned one of two binary labels (true or fake news), making it suitable for our binary classification setting.

Fakenewsnet (Shu et al., 2017a,b, 2018) A fake news dataset characterized by its rich diversity, including news articles and social context. The contents have been sourced from PolitiFact[5] and GossipCop[6], with most of them dating back to before 2018. In this paper, we have chosen to utilize the PolitiFact portion due to its high quality, as all the facts have been verified by domain experts.

LLMFake (Chen and Shu, 2024) A misinformation dataset is further modified by LLMs such as ChatGPT. These models utilize various techniques, including paraphrasing, rewriting, etc. for information manipulation. The information within this dataset traces back to 2020 or earlier.

PolitiFact-Snopes-2024 The dataset was manually collected from the prestigious fact-checking organizations PolitiFact and Snopes[7]. It includes approximately 1,200 verifiable claims along with the fact-check rating labels that determine the level of truthfulness for each claim. The clarifications for the labels and the additional detailed analysis reports were not collected. Non-text-based claims were filtered out, and exclusive fact-checks with supporting sources specific to these organizations were also filtered out.

FakeNews2024 This dataset consists 46 real news and 63 fake news articles. All the news articles are less than one year old, and are confirmed by NewsBreak moderation team.

The first three datasets were selected to enable a comparison of our system against three distinct fake news detection methods: content-based, LLMs-based, and retrieval-augmented approaches. The last two datasets are used to demonstrate our approach's ability to detect the latest fake news.

4.2 Evaluation Metrics

For the existing datasets, we strive to employ the same evaluation metrics that have been utilized in prior studies to enable direct comparisons.

For BuzzFeedNews, we report the precision, recall, and F1 scores related to fake news, as well as

---

[5] https://www.politifact.com
[6] https://www.gossipcop.com is now closed
[7] https://www.snopes.com

Method	Accuracy	Precision	Recall	F1
Pérez-Rosas et al. (2018)	75.5	74.5	76.9	75.7
Shu et al. (2019)	86.4	84.9	89.3	87.0
Zhou et al. (2020)	**87.9**	**85.7**	**90.2**	**87.9**
Ours (GPT)	79.1	81.2	75.8	78.4
Ours (Llama)	73.6	71.3	79.1	75.0

Table 3: Detection performance on BuzzFeedNews.

the accuracy for the entire dataset. For Fakenewsnet, PolitiFact-Snopes-2024, and FakeNews2024, we report the precision (P-F), recall (R-F), and F1 score (F1-F) of the fake news, the precision (P-T), recall (R-T), and F1 score (F1-T) of the real news, as well as the Micro F1 score (F1) of the overall dataset. For LLMFake, we report the detection success rate, which is calculated by the percentage of successfully identified fake news (Chen and Shu, 2024).

4.3 Implementation Details

To aid in the verification of news articles, the main factual claim of each news article is identified and then compared against internet search results. To ensure a fair comparison, we have developed heuristics to carefully filter out fact-checking content from search engine results in all the experiments below.

The datasets above except LLMFake are each aggregated to train the final LightGBM classifier, utilizing the features outlined in Section 3.4, and subsequently report the end-to-end accuracy. Both the training and testing processes are conducted using a 5-fold cross-validation approach. We also provided baseline benchmarks for comparison.

5 Experimental Results

The performance with the BuzzfeedNews dataset is detailed in Table 3. The baseline methods being reported in Zhou et al. (2020) utilize features from article content, and outperform our approach. This outcome is expected since BuzzfeedNews dataset focuses primarily on a limited range of topics, specifically the 2016 US election. The nature of the fake news within this dataset allows it to be effectively modeled through content features. Furthermore, the fake news articles are approximately 7 years old, posing additional challenges for search engines in retrieving relevant evidences.

In Table 4, we present a performance comparison between VeraCT Scan and another retrieval-augmented system, utilizing the FakeNewsNet

Method	F1	F1-T	R-T	P-T	F1-F	R-F	P-F
Liao et al. (2023)	72.9	75.7	78.0	73.5	70.2	68.1	72.8
Ours (GPT)	**80.3**	**81.9**	**85.9**	**78.2**	**78.3**	**74.1**	**83.0**
Ours (Llama)	77.3	79.0	82.3	75.9	75.3	71.9	79.1

Table 4: Detection performance on Fakenewsnet.

Dataset	Written	Paraphrasing	Rewriting	Generating
GPT-4-based Zero-shot Detector (COT) (Chen and Shu, 2024)				
Politifact	62.6	56.0	53.6	41.6
Gossipcop	26.3	30.0	25.0	25.7
CoAID	81.0	82.2	73.3	52.7
Ours (GPT)				
Politifact	**63.7**	**62.2**	**60.0**	**60.7**
Gossipcop	**42.9**	**42.0**	**40.3**	**39.4**
CoAID	**83.7**	**86.0**	**77.9**	**69.8**
Ours (Llama)				
Politifact	56.3	55.9	55.5	51.1
Gossipcop	31.2	30.3	34.6	28.6
CoAID	74.4	75.6	70.9	60.5

Table 5: Detection performance on LLMFake.

Method	F1	F1-T	R-T	P-T	F1-F	R-F	P-F
Ours (GPT)	91.7	91.7	90.7	92.8	91.7	92.8	90.7
Ours (Llama)	85.6	85.9	86.4	85.3	85.3	84.8	85.9

Table 6: Detection performance on PolitiFact-Snopes-2024.

Method	F1	F1-T	R-T	P-T	F1-F	R-F	P-F
Ours (GPT)	89.9	87.6	84.8	90.7	91.5	93.7	89.4
Ours (Llama)	82.9	80.0	78.3	81.8	85.9	87.3	84.6

Table 7: Detection performance on FakeNews2024.

6 Conclusion and Future Work

In this paper, we present VeraCT Scan, a novel retrieval-augmented system for fake news detection. Two of our implementations, properly prompted GPT-4 Turbo and fine-tuned Llama-2 13B demonstrated notable accuracy in detection. Specifically, the GPT-4 Turbo implementation exhibited state-of-the-art performance in several datasets. VeraCT Scan is especially successful in identifying the latest instances of fake news. This emphasizes the critical role of search result relevance in gathering compelling evidence.

Our observations reveal that the rationales generated by LLMs offer rich insights into potentially dubious aspects with a high degree of details. As a future work, we plan to investigate the potential of using these rationales as input features for the final verification classifier. And throughout our evaluations, Llama-2 13B consistently lags behind GPT-4 Turbo in terms of detection accuracy. We will explore more effective fine-tuning strategies to narrow this performance gap.

Furthermore, we observe that within the entire system, the majority of errors occur during the verification stage, with a smaller fraction arising during the claim extraction phase. The causes of these errors include: (i) Irrelevant search results used for verification. (ii) Updated news events leading to outdated reports being used for verification. (iii) Each report only supporting a part of the claim, necessitating the proper merging of relevant information from multiple news reports for full verification. (iv) Improper normalization of named entities or temporal expressions during the claim extraction stage, making alignment difficult during verification (e.g., "last weekend" vs. an exact date). We hope to address these issues in future work.

dataset. Our two implementations, GPT-4 Turbo and the fine-tuned version of Llama-2 13B, both exhibit superior accuracy. This comparison underscores the efficacy of using either prompted or fine-tuned LLMs over specialized encoder-decoder transformers that have been specifically trained for this task.

Table 5 presents the detection performance using LLMFake. Notably, although the news articles in LLMFake are from 2020 or earlier—falling within GPT-4's inherent knowledge base, VeraCT Scan significantly outperforms GPT-4 in verification accuracy. Notably, the Llama-2 13B implementation also wins 7 out of 12 benchmarks. This underscores the benefits and efficacy of incorporating knowledge from the Internet. It is important to note that LLMFake verification is not straightforward. According to Chen and Shu (2024), the accuracy of human annotations falls well below 40%.

In Tables 6 and 7, we present the detection accuracy of our system when tested against the latest news articles. Unlike BuzzFeedNews, these two datasets consist of a wide variety of topics, including politics, entertainment, international warfare, and more. Both implementations of our system present relatively high detection accuracy, and underscores the effectiveness in verifying the latest news. Our approach benefits significantly from the enhanced efficiency of both Google and our proprietary search engine in sourcing relevant evidences for recent news.

7 Limitations

News events are inherently dynamic, and the truth surrounding them can evolve over time. When verifying a news article being published in 2015 that discusses the average income increase ratio since 2001, it is crucial to obtain accurate data spanning from 2001 to 2015. This task presents challenges not only to search engines but also to LLMs. We have observed that our system performs more effectively when verifying more recent news articles. To close the gap, it requires truly understanding of timestamps by LLMs and the ability to accurately perform time sensitive calculations.

It has been noted that low-quality news articles frequently mix facts with opinions. In addition to verifying facts, it's important to distinguish the opinion segments within a news report. To accomplish this, it is crucial to integrate article-level linguistic features with retrieval-augmented fact verification methods.

Fake news can be deliberately created on a large scale. Beyond verifying individual articles, checking the authenticity of clusters of articles, can significantly enhance the detection effectiveness.

For practical considerations such as enhancing service robustness, reducing latency, and cutting costs, it is desirable to develop a smaller-sized LLM specifically for fake news detection. We plan to significantly invest in creating high-quality training data and explore advanced fine-tuning technologies to bridge the performance gap with GPT-4 in this area.

8 Ethical Discussion

Detecting fake news is a critical task with significant consequences. The effectiveness of this detection depends on various factors, such as the quality of searches, the impartial assessment of source credibility, and the language understanding capabilities of large language models (LLMs), among others. Our system aims to gather pertinent evidence from reputable sources, thereby aiding users in making informed decisions but not making those decisions for them. This approach is clearly outlined on our demo site.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228, Minneapolis, Minnesota. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA. Association for Computing Machinery.

Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Tsun Hin Cheung and Kin Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023*, 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023, pages 846–853. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2023 IEEE.; 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023 ; Conference date: 31-10-2023 Through 03-11-2023.

Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 877–880.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Armin Kirchknopf, Djordje Slijepčević, and Matthias Zeppelzauer. 2021. Multimodal detection of information disorder from social media. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4461–4472, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4549–4560, New York, NY, USA. Association for Computing Machinery.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. 2024. Scaling laws of roPE-based extrapolation. In *The Twelfth International Conference on Learning Representations*.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *CoRR*, abs/1902.06673.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, Jintao Li, and Kai Shu. 2023. Exploiting user comments for early detection of fake news prior to users' commenting. *arXiv preprint arXiv:2310.10429*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

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*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Przybyla. 2020. Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. 2023. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots.

Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1736–1746, New York, NY, USA. Association for Computing Machinery.

Kimya Khakzad Shahandashti, Mithila Sivakumar, Mohammad Mahdi Mohajer, Alvine B. Belle, Song Wang, and Timothy C. Lethbridge. 2024. Evaluating the effectiveness of gpt-4 turbo in creating defeaters for assurance cases.

Saeid Sheikhi. 2021. An effective fake news detection method using woa-xgbtree algorithm and content-based features. *Applied Soft Computing*, 109:107559.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8 3:171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017a. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2017b. Exploiting tri-relationship for fake news detection. *ArXiv*, abs/1712.07709.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 312–320, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2020. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 436–439, New York, NY, USA. Association for Computing Machinery.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. https://github.com/BuzzFeedNews/2016-10-facebook-fact-check.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Norman Vasu, Benjamin Ang, Terri-Anne Teo, Shashi Jayakumar, Muhammad Faizal, and Juhi Ahuja. 2018. Fake news: National security in the post-truth era. Technical report, S. Rajaratnam School of International Studies.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Weizhi Xu, Q. Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Annual Meeting of the Association for Computational Linguistics*.

Xin Xu, Shizhe Diao, Can Yang, and Yang Wang. 2024. Can we verify step by step for incorrect answer detection?

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, and Graham Neubig. 2023b. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*.

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. *Digital Threats*, 1(2).

A Prompts

Here we list the prompts used in the pipeline:

MAIN CLAIM EXTRACTION
Given the input content below, please summarize the single key claim. Input content: {content} Please output with the follow json format {{"key_claim": XXX}}. Please output now:
KEY CLAIMS EXTRACTION
Given the input content below, please extract distinct key claims. The key claims should be concrete enough containing clear context so that it can be efficiently verified. Input content: {content} Please output with the follow json format {{"key_claims": [{{"claim": XXX}}, ...]}}. Please output now:
QUERY GENERATION
Given the claim below, please generate a Google query which can be used to search content to verify this claim. Claim: {claim} Please output with the following JSON format {{"query": "XXX"}} Please output now:
CONTENT CLAIM VERIFICATION
Below is one web search result Search Result: {search_result} Below is a claim to be verified Claim: {claim} Please perform the following rules to generate an output with this json format : {{"support_or_negate_or_baseless": "support" or "negate" or "baseless", "confidence": "high" or "medium" or "low", "rationale": "XXX"}} Rule 1: if the search result content support the claim, set the "support_or_negate_or_baseless" field as "support", and offer a confident score and a rationale. Rule 2: if the search result content negate the claim, set the "support_or_negate_or_baseless" field as "negate", and offer a confident score and a rationale. Rule 3: if the search result content cannot either support or negate the claim, set the "support_or_negate_or_baseless" field as "baseless", and offer a confident score and a rationale. To clarify: if the content of the search results does not contradict the claim, but lacks some or all of the information presented in the claim, please use the label "baseless" rather than "negate". Please output now:
SAME NEWS/RELEVANT VERIFICATION
Below is one web search result. Search Result: {search_result} Below is a claim: Claim: {claim} Please make the following two investigations: 1. Please check if the news article and the search result is about the same news story. 2. Please check if the search result contains content (facts, opinions, or claims) related to the news article. Please output with the following json format : {{"about_the_same_news_story": "yes" or "no", "contains_related_content": "yes" or "no"}} Please output now:

Table 8: Prompts used for key tasks.

B Sample Results

We provide an illustration of the process involved in verifying a news article below.

NEWS ARTICLE: Scientists Warn Eggs Are Causing Thousands of People to 'Suddenly' Form Blood Clots
In what appears to be another example of the global elite attempting to distract the public from the real cause of the surge in heart problems since the jab rollout, scientists now want us to believe that a nutrient found in eggs increases the risk of blood clotting.
Express.co.uk reports: The study conducted by Cleveland Clinic, suggested that choline could make the blood more prone to clotting.
Choline is sometimes sold in over-the-counter dietary supplements.
It is deemed an essential nutrient that is widely recommended for optimal health.
According to Medical News Today, it supports various bodily functions, including cellular growth and metabolism.
...

CLAIMS: 1. A study conducted by Cleveland Clinic suggested that choline could make the blood more prone to clotting.

2. Choline is deemed an essential nutrient that supports various bodily functions, including cellular growth and metabolism.

3. Researchers found that consuming choline in high concentrations could lead to blood clotting.

4. The interaction between choline and gut bacteria produces TMAO, which has been linked to an increased risk of blood clots, heart attack, and stroke.
...(8 claims extracted)

VERIFY: A study conducted by Cleveland Clinic suggested that choline could make the blood more prone to clotting.
GENERATED QUERIES:

1. Cleveland Clinic study choline blood clotting

2. Researchers found that consuming choline in high concentrations could lead to blood clotting.

NUM OF SEARCH RESULTS: 18

VERIFICATION & RATIONALE:
Title: Study from 2017 misrepresented to link eggs to blood clots.
Result: supported
Domain: apnews.com
Rationale: The search result explicitly mentions a study conducted by the Cleveland Clinic in 2017, which found that taking supplements in capsule form of choline 2014 a nutrient present in red meat, egg yolks, and other foods 2014 raised levels of a compound in the body (trimethylamine N-oxide, or TMAO) that could make the blood more prone to clotting. This directly supports the claim that consuming choline in high concentrations could lead to blood clotting.

—

Title: Cleveland Clinic Researchers First to Show Dietary Choline and Gut Bacteria Byproduct Linked with Increased Blood...
Result: supported
Domain: clevelandclinic.org
Rationale: The search result explicitly states that Cleveland Clinic researchers have shown for the first time in humans that choline is directly linked to increased production of a gut bacteria byproduct that increases the risk of blood-clotting events like heart attack and stroke. This directly supports the claim that choline could make the blood more prone to clotting.

—

Title: Cleveland Clinic study did not conclude eggs are causing...
Result: baseless
Domain: reuters.com
Rationale: The provided search result does not contain information about the effect of choline intake on blood clotting. It only mentions a study done by Cleveland Clinic regarding eggs, which is unrelated to the claim about choline.

...

FINAL DECISION: Supported

Table 9: An example of our pipeline ouput. Given that multiple claims can be extracted from a single article, we only exemplify the verification details of the first claim. The system generated two search queries related to the claim, resulting in the retrieval of 18 documents from the web. Based on the analysis of these documents, 14 documents are marked *baseless* (irrelevant or not fully support the claim), whereas the remaining 4 documents *support* the claim. By considering the sources credibility, the claim is classified as supported.