InFact: A Strong Baseline for Automated Fact-Checking

Mark Rothermel[*] Tobias Braun[*] Marcus Rohrbach Anna Rohrbach
TU Darmstadt & hessian.AI, Germany

Abstract

The spread of disinformation poses a global threat to democratic societies, necessitating robust and scalable Automated Fact-Checking (AFC) systems. The AVeriTeC Shared Task Challenge 2024 offers a realistic benchmark for text-based fact-checking methods. This paper presents *Information-Retrieving Fact-Checker (*InFact*)*, an LLM-based approach that breaks down the task of claim verification into a 6-stage process, including evidence retrieval. When using GPT-4o as the backbone, InFact achieves an AVeriTeC score of 63 % on the test set, outperforming all other 20 teams competing in the challenge, and establishing a new strong baseline for future text-only AFC systems. Qualitative analysis of mislabeled instances reveals that InFact often yields a more accurate conclusion than AVeriTeC's human-annotated ground truth.

1 Introduction

The weaponization of disinformation poses a critical threat to global stability. The World Economic Forum, in its January report (World Economic Forum, 2024), identified mis- and disinformation as the most significant global risk for the next 24 months, surpassing even extreme weather events and military conflicts. As such, the development and deployment of Automated Fact-Checking (AFC) is essential in safeguarding the integrity of democratic societies worldwide.

Schlichtkrull et al. (2023) introduced the *Automated VERIfication of TExtual Claims* (AVeriTeC) benchmark, consisting of 4,568 real-world claims subject to fact-checks by 50 organizations. AVeriTeC classifies each claim as either ✓ `Supported`, ✗ `Refuted`, ? `NEI` (**N**ot **E**nough **I**nformation) or ⚡ `C/CP` if there is **c**onflicting evidence or the claim is technically true but misleading due to the exclusion of important context (**c**herry-**p**icking). The benchmark expects the fact-check to be structured as a set of questions and answers, comparing them against the gold QA pairs using the Hungarian METEOR metric in order to ensure that the predicted veracity is sufficiently justified. It further provides a Knowledge Base (KB), a collection of scraped web pages. Each claim is associated with the resources used to fact-check it (gold evidence) and ca. 1,000 unrelated resources to simulate open web search.

Several early works suggest that LLMs and LLM prompting techniques such as Chain-of-Thought could be used for AFC (Geng et al., 2024; Khaliq et al., 2024; Zhang and Gao, 2023; Wei et al., 2024; Zhou et al., 2024). Following these works, we present an approach that is customized for the AVeriTeC challenge (Schlichtkrull et al., 2024) and incorporates intermediate question generation and evidence retrieval to provide answers.

We propose **In**formation-Augmented **Fact**-Checker (InFact), an AFC system with the capability of retrieving evidence. InFact achieves an AVeriTeC score of 62.6% on the test set and yields an accuracy of 72.4% on the development dataset. Qualitative analysis shows that our method's retrieval process and reasoning capabilities provide a powerful baseline for text-only AFC. Further details will be provided on https://github.com/multimodal-ai-lab/InFact.

2 The InFact System

Open-domain, text-only claim verification requires world and commonsense knowledge and some degree of reasoning. Due to their remarkable success in both of these skills, we chose to drive the fact-check by an LLM, supplemented with a custom evidence retrieval module. While our approach is agnostic to the choice of the LLM, the LLM's abilities influence the quality and accuracy of the resulting fact-check. Since the task of fact-checking

[*] These authors contributed equally to this work.

Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 108–112
November 15, 2024 ©2024 Association for Computational Linguistics

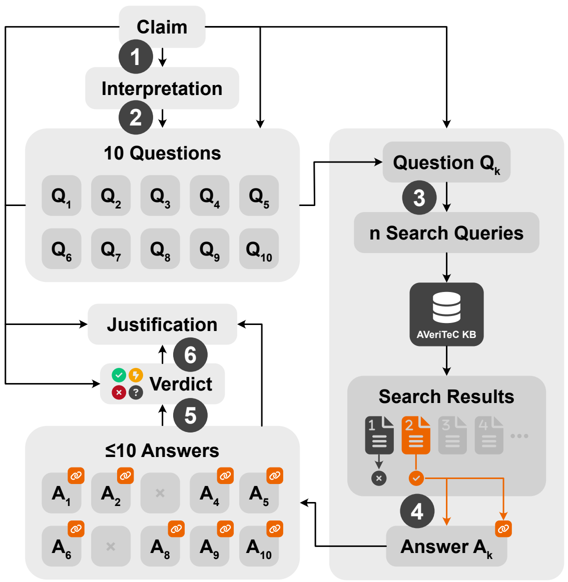


Figure 1: The **InFact System**. (1) Interpret the claim, (2) pose 10 questions, (3) for each question individually, generate search queries and retrieve potentially relevant evidence from the AVeriTeC Knowledge Base, (4) answer the corresponding question using the found evidence, (5) after completing all questions, predict a verdict and (6) generate a justification.

is broad and complex, we subdivide the process into six stages, as shown in Figure 1.

In short, InFact addresses the task with a static, single-pass pipeline that poses critical questions which are answered through evidence retrieved from the AVeriTeC KB. Each of the six stages corresponds to an engineered prompt, applying prompting best practices including Chain-of-Thought (Wei et al., 2022) and In-Context Learning (Min et al., 2021), whenever applicable.

Stage 1: Interpret the Claim. The pipeline begins with an augmentation of the claim text with its author, date, and origin URL. Subsequently the LLM is prompted to reformulate the claim, considering the supplied metadata. This step is helpful when the time frame is unclear as in *“Joe Biden’s income has increased recently.”* We also expect the interpretation to help when the claim misses context as in *“Tourism, lockdown key to deep New Zealand recession.”*

Stage 2: Pose Questions. Next, InFact produces a list of 10 questions that it deems essential for fact-checking. To facilitate the question generation, we provide the LLM with manually selected in-context examples. Furthermore, the instructions are inspired by fact-checking best practices from Silverman (2014).

Stage 3: Retrieve Evidence. For each generated question, InFact iteratively retrieves a list of evidence resources. InFact approaches this by letting the LLM propose one or multiple search queries, which are submitted to the AVeriTeC KB, yielding a list of 5 search results per query.

The AVeriTeC KB contains a collection of about $1,000$ resources per claim. A resource is a scraped URL, ranging from news articles over social media posts to PDF documents. We decided to use the AVeriTeC KB over open-web search for two main reasons: First, it guarantees to contain the gold evidence (possibly erased from the open web) and, second, it yields reproducible results (in contrast to open-web search).

To retrieve the most relevant resources from the KB, we implement a semantic search mechanism. For each resource, we compute its document-level embedding by employing a text embedding model. We chose `gte-base-en-v1.5` (Alibaba-NLP, 2024) due to its competitive FEVER score at time of the challenge given its manageable size. We compute the query’s text embedding and use it to perform k-nearest neighbor search w.r.t. the Euclidean Distance in the document embedding space. This outputs a list of the semantically closest 5 resources. We drop resources that were found in previous searches and end up with a list of $\leq 5n$ evidences per question. We found this approach qualitatively superior to the common BM25 ranking method.

Stage 4: Answer Questions. Taking all the search results, InFact iterates from the semantically most similar to the least similar, instructing the LLM to either answer the question using the information from the result or respond with `NONE` if the result is deemed unhelpful. If the LLM returned an answer to the question, InFact saves the answer along with the evidence URL, and the Q&A process continues with the next question. However, if the LLM returned `NONE` for all search results, the question is dropped for the remainder of the fact-check.

Stage 5: Predict a Verdict. Once all the questions are processed, the LLM judges the claim’s veracity based on the recorded QA pairs in a single prompt as follows: First, it summarizes the key insights from the Q&A. Second, it identifies any pending, missing information. Third, it writes a brief conclusion, including the final verdict.

Stage 6: Justify the Verdict. In this last stage, InFact generates a brief justification for the verdict through summarization of the previous findings.

System	METEOR Q-Only	METEOR Q&A	AVERITEC Score
INFACT (Ours)	45	34	**63**
HERO	48	**35**	57
AIC	46	32	50
DUN-FACTCHECKER	**49**	**35**	50
PAPELO-TEN	44	30	48
Challenge Baseline	24	20	11

Table 1: Top-5 systems and the baseline on the AVERITEC challenge test set, ranked by AVERITEC score (in %) as defined in Schlichtkrull et al. (2023).

The LLM takes the claim, all the QA pairs, the verdict, and any in-between reasoning (e.g., from stage 5) and creates a summary, focusing on the main reasons for the verdict. This stage is not required by the AVERITEC task and does not affect any of the metrics.

3 Experimental Results

Experimental Setup. We evaluate INFACT on the development set which consists of 305 ✖ Refuted, 122 ✔ Supported, 35 ❓ NEI and 38 ⚡ C/CP claims, 500 claims in total. As our LLM backbone, we test three models: (a) the open-source LLAMA 3 (70B), (b) the closed-source GPT-4O MINI, and (c) the more expensive GPT-4O model. We use the models without any finetuning and set the temperature to 0.01 and top-p to 0.9. Additionally, we truncate each resource to about 8 k tokens, which is the maximum input length of the embedding model. We compare INFACT with the following baseline and ablations: The *Naive* baseline instructs the LLM to predict the verdict right away in a single prompt, skipping evidence retrieval entirely and relying solely on the LLM's parametric knowledge; the *No Interpretation* ablation omits stage 1; *No Evidence* answers the questions by leaving out stage 3 (evidence retrieval); *No Q&A* generates search queries based on the claim instead of a Q&A, gathers 10 results and proceeds to make a verdict based on those; *No Query Generation* skips the step of query generation by using the question as the search query directly.

Challenge Results. Table 1 presents the top-5 entries from the challenge leaderboard, sorted by the AVERITEC score on the test set. INFACT achieves the best score with a significant margin to the second-best system. Yet, it is not the best in terms of the retrieval metrics.

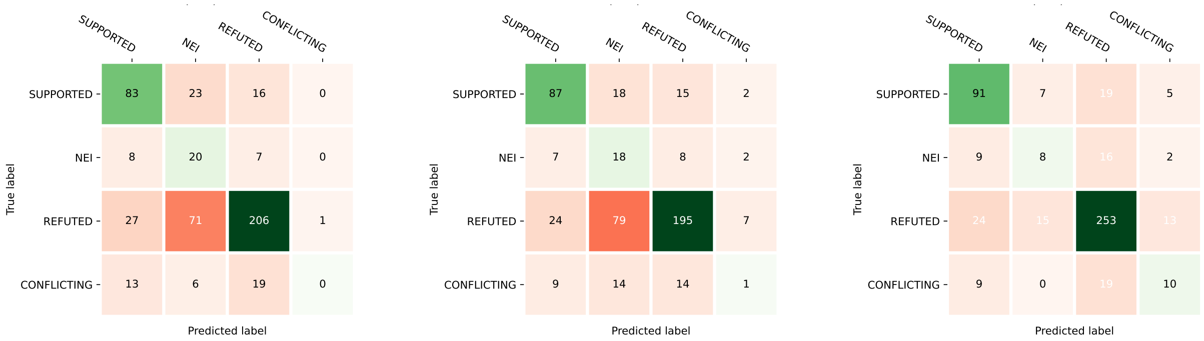
Metric	LLAMA 3	GPT-4O MINI	GPT-4O	Naive	No Interpret.	No Evidence	No Q&A	No Query Gen.	INFACT
AVERITEC Score	✓			-	42.4	40.8	-	40.4	40.2
		✓		-	48.2	36.4	-	41.6	47.2
			✓	-	**59.8**	53.0	-	56.4	58.8
Accuracy	✓			67.0	63.2	67.0	52.9	65.0	61.8
		✓		36.2	61.6	56.8	54.8	59.6	60.2
			✓	52.6	71.8	71.0	68.8	70.2	**72.4**
Q-Only METEOR	✓			-	39.5	41.8	-	37.8	39.6
		✓		-	43.0	44.3	-	42.1	43.3
			✓	-	**46.2**	45.7	-	44.3	45.8
Q&A METEOR	✓			-	29.6	28.7	-	28.4	29.5
		✓		-	31.2	29.1	-	30.9	31.5
			✓	-	**33.5**	32.0	-	32.8	33.2

Table 2: Results in % on the AVERITEC development dataset, showing four metrics for INFACT and the five ablation variants, all tested with three different LLMs.

Analysis. The ablation comparison is shown in Table 2. GPT-4O almost consistently outperforms both other LLMs. INFACT and *No Interpretation* score best in terms of AVERITEC score and accuracy. Their similarity hints at a potential redundancy of the interpretation step in the case of AVERITEC. While our experiments show that generating search queries is superior to searching the literal question, the optimal value for the number of queries per question n remains unknown. Moreover, and surprisingly, leaving out all evidence does not lead to a drastic decline of the METEOR scores, showing its insensitivity to generated (thus potentially hallucinated) evidence vs. actually retrieved evidence.

Judging by the confusion matrices (cf. Fig. 2, the most distinct confusion for LLAMA 3 and GPT-4O MINI happens between ❓ NEI (predicted) and ✖ Refuted (true), which is less critical than confusion between ✔ Supported and ✖ Refuted. At the same time, GPT-4O predicts much fewer ❓ NEI in favor of ✖ Refuted, which could be attributed to its stronger reasoning capabilities.

Surprisingly, in the *Naive* setting, LLAMA 3 outperforms the GPT models by a large margin. As opposed to the GPT models, LLAMA 3 commits more often to either ✔ Supported or ✖ Refuted rather than choosing ❓ NEI, showing a "self-confident" behavior despite having little evidence. In the *No Evidence* variant, the GPT models



(a) LLAMA 3

(b) GPT-4O MINI

(c) GPT-4O

Figure 2: Confusion matrices of INFACT on the AVERITEC development set for three different LLMs.

achieve a higher accuracy and predict ? NEI much less, while still having no access to any external information. This indicates that structured reasoning elevates GPT models' confidence, regardless of the knowledge source.

Qualitative analysis of 20 failure cases reveals that, in more than half of the cases, the ground truth was at least debatable or INFACT delivered a valid alternative fact-check. E.g., the ground truth of "While serving as Town Supervisor on Grand Island, Nebraska, US Nate McMurray voted to raise taxes on homeowners" is ✓ Supported, however McMurray served on Grand Island, **New York**. In two cases, the gold fact-check considered a *different* claim than the one presented. E.g., the claim: "Scientific American magazine warned that 5G technology is not safe" is about the magazine issuing a warning about 5G. However, the gold fact-check analyzed the safety of 5G itself.

In only 6 of the analyzed 20 failure cases, the cause for the mislabeling can be clearly attributed to INFACT. The cases include the usage of unreliable evidence sources, misinterpretations of the claim, the missing ability to process non-textual evidence, and the confusion between clearly refuted and merely unsupported claims. In a nutshell, the analysis implies that the model performs better than the metrics might reflect.

4 Discussion & Conclusion

INFACT establishes a robust baseline for information-augmented fact-checking without requiring fine-tuning. Its LLM-agnostic design ensures that it benefits from advancements in the reasoning capabilities of LLMs, making it adaptable to future developments. Additionally, INFACT provides justifications, enhancing interpretability and trust in its outputs. However, INFACT also has limitations. The inclusion of closed-source models limits transparency, reproducibility, and incurs high cost with about $ 0.46 per claim when using GPT-4O. While GPT-4O MINI is much cheaper (about $ 0.01 per claim), it exhibited lower performance. The open-source alternative LLAMA 3 resulted in 8 times longer computation times and reduced effectiveness. Also the number of retrievals was relatively high (7 searches per claim). Increasing INFACT's efficiency by reducing searches and skipping and/or combining steps in the pipeline are a great opportunity for future work. All LLMs evaluated in this study were pre-trained on datasets that extend into 2023, likely covering many of AVERITEC's claims and evidence available online.

Moreover, the AVERITEC dataset comes with its own limitations. The wording of the QA pairs is crucial when using the METEOR score to evaluate them against gold-standard QA pairs. The automated comparison method is limited in capturing semantically similar statements, and it is infeasible to provide an exhaustive list of all potentially relevant evidence. Moreover, we found many questionable ground truth answers, cf. Section 3. We suspect that these inaccuracies stem from layperson annotations. Addressing these limitations and refining the dataset/metric will benefit measuring progress in this challenging task.

Acknowledgements

The research was partially funded by a LOEWE-Start-Professur (LOEWE/4b//519/05.01.002-(0006)/94), LOEWE-Spitzen-Professur (LOEWE/4a//519/05.00.002-(0010)/93) and an Alexander von Humboldt Professorship in Multimodal Reliable AI, sponsored by Germany's Federal Ministry for Education and Research.

References

Alibaba-NLP. 2024. Gte base en v1.5. Accessed: 2024-08-14.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *arXiv:2403.03627*.

M. Abdul Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić. 2024. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *Preprint*, arXiv:2404.12065.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *CoRR*, abs/2110.15943.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (averitec) shared task. In *Proceedings of the Seventh Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics (ACL). To appear.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. *arXiv preprint*. ArXiv:2305.13117 [cs].

Craig Silverman, editor. 2014. *Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage*. European Journalism Centre, Maastricht, the Netherlands. Copyeditor: Merrill Perlman, The American Copy Editors Society (ACES).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *arXiv preprint*. ArXiv:2403.18802 [cs].

World Economic Forum. 2024. The global risks report 2024. Page 18.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *Preprint*, arXiv:2310.00305.

Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model. *arXiv preprint*. ArXiv:2403.11169 [cs].