Step-by-Step Fact Verification System for Medical Claims with Explainable Reasoning

Juraj Vladika, Ivana Hacajová, Florian Matthes
Technical University of Munich, Germany
School of Computation, Information and Technology
Department of Computer Science
`{juraj.vladika, ivana.hacajova, matthes}@tum.de`

Abstract

Fact verification (FV) aims to assess the veracity of a claim based on relevant evidence. The traditional approach for automated FV includes a three-part pipeline relying on short evidence snippets and encoder-only inference models. More recent approaches leverage the multi-turn nature of LLMs to address FV as a step-by-step problem where questions inquiring additional context are generated and answered until there is enough information to make a decision. This iterative method makes the verification process rational and explainable. While these methods have been tested for encyclopedic claims, exploration on domain-specific and realistic claims is missing. In this work, we apply an iterative FV system on three medical fact-checking datasets and evaluate it with multiple settings, including different LLMs, external web search, and structured reasoning using logic predicates. We demonstrate improvements in the final performance over traditional approaches and the high potential of step-by-step FV systems for domain-specific claims.

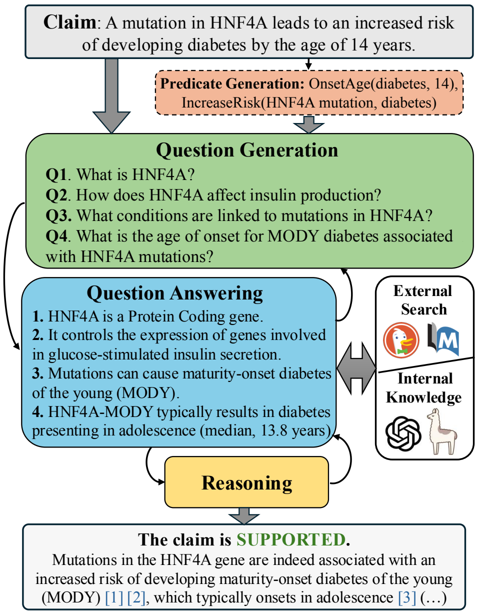
Claim: A mutation in HNF4A leads to an increased risk of developing diabetes by the age of 14 years.
Predicate Generation: OnsetAge(diabetes, 14), IncreaseRisk(HNF4A mutation, diabetes)
Question Generation
Q1. What is HNF4A?
Q2. How does HNF4A affect insulin production?
Q3. What conditions are linked to mutations in HNF4A?
Q4. What is the age of onset for MODY diabetes associated with HNF4A mutations?
Question Answering
1. HNF4A is a Protein Coding gene.
2. It controls the expression of genes involved in glucose-stimulated insulin secretion.
3. Mutations can cause maturity-onset diabetes of the young (MODY).
4. HNF4A-MODY typically results in diabetes presenting in adolescence (median, 13.8 years)
External Search
Internal Knowledge
Reasoning
The claim is SUPPORTED.
Mutations in the HNF4A gene are indeed associated with an increased risk of developing maturity-onset diabetes of the young (MODY) [1] [2], which typically onsets in adolescence [3] (…)

Figure 1: The step-by-step fact verification system used in our study iteratively collects additional knowledge and evidence until it can predict a veracity verdict.

1 Introduction

The digital age has been marked by the rise and spread of online misinformation, which has negative societal consequences, especially when related to public health (van der Linden, 2022). Fact verification (FV) has emerged as an automated approach for addressing the increasing rate of deceptive content promulgated online (Das et al., 2023; Schlichtkrull et al., 2023a). On top of that, FV can help improve the factuality of generative large language models (Augenstein et al., 2024) and help scientists find reliable evidence for assessing their research hypotheses (Eger et al., 2025).

The common pipeline for automated fact verification consists of document retrieval, evidence extraction, veracity prediction, and optionally justification production (Guo et al., 2022). In such a setup, document retrieval is usually done with a method like BM25 or semantic search, evidence selected using sentence embedding models, and the final verdict predicted with an encoder-only model like DeBERTa (He et al., 2021). In fact, most state-of-the-art FV systems for the popular FEVER dataset (Thorne et al., 2018) and other recent real-world misinformation datasets rely on this pipeline (Zhang et al., 2024; Glockner et al., 2024).

Similarly, most previous work relies on providing pre-selected evidence to the final inference model. A more realistic setting is *open-domain* fact verification, where evidence first has to be discovered in large knowledge bases before the system produces the verdict. Recent FV work has explored this setting, but most of them also rely on the traditional pipeline, utilizing BM25, sentence embeddings, and encoder-only inference model for producing their verdicts (Wadden et al., 2022; Stammbach et al., 2023; Vladika and Matthes, 2024b).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 805–816
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

The recent advent of large language models (LLMs) has transformed the field of NLP (Fan et al., 2024). LLMs have many properties that positively benefit the fact-verification process (Dmonte et al., 2025). First, their long context window means a lot more evidence can be provided than to encoder-only models. Furthermore, the multi-turn nature of instruction-tuned LLMs has enabled addressing FV as a step-by-step problem where new questions inquiring for more evidence are generated in subsequent iterations before there is enough information to produce a verdict on a claim's veracity (Dhuliawala et al., 2024). This also makes the verification process interpretable since the reasoning steps can be traced through the question-answer pairs, thus justifying the verdict (Eldifrawi et al., 2024).

These step-by-step LLM systems for FV have been shown to work well on complex, multi-hop claims found in datasets like HOVER (Jiang et al., 2020). Intuitively, complex synthetic claims from these datasets, like "*Yao Ming's wife's alma mater is in Texas*", have to be broken down into subunits to be verified effectively. Nevertheless, we posit that more realistic but simple claims such as "*Honey can cure a common cold*" also necessitate generating follow-up questions and collecting deeper knowledge before producing a verdict. To the best of our knowledge, no research has been conducted to test how well can these step-by-step FV systems perform on domain-specific claims.

To bridge this research gap, in this study, we develop a step-by-step LLM system, shown in Figure 1, and apply it on three medical fact-checking datasets. We contrast the results to the previous work on open-domain scientific fact verification based on a traditional system, showcasing significant improvements in the final predictive performance of the system. We outline additional findings regarding the influence of the base LLM, evidence source, and reasoning with predicate logic on the final verification performance, highlighting the great potential of these systems for diverse claims.

We make our data and code available in a public GitHub repository.[1]

2 Related Work

There have been many synthetic FV datasets constructed from Wikipedia, such as FEVER (Thorne et al., 2018). While FEVER focuses on simple claims, datasets like HOVER (Jiang et al., 2020) and FEVEROUS (Aly et al., 2021) introduced complex claims requiring multi-hop reasoning. Apart from synthetic datasets, there are also datasets focusing on more realistic claims and real-world misinformation (Schlichtkrull et al., 2023b; Glockner et al., 2024). Increasingly popular are also domain-specific datasets focusing on scientific fact-checking (Vladika and Matthes, 2023), especially for the domains of medicine (Saakyan et al., 2021; Sarrouti et al., 2021), climate (Diggelmann et al., 2020), and computer science (Lu et al., 2023).

Most FV approaches follow the traditional three-part pipeline (Bekoulis et al., 2021). In recent years, approaches incorporating LLMs and iterative reasoning into the process have achieved great performance on multi-hop FV. This includes FV through varifocal questions (Ousidhoum et al., 2022) or *wh*-questions to aid verification (Rani et al., 2023), step-by-step prompting (Zhang and Gao, 2023), and program-guided reasoning (Pan et al., 2023b).

Most studies with iterative FV systems focus on multi-hop encyclopedic claims. To the best of our knowledge, our study is among the first to explore the step-by-step FV systems for real-world claims rooted in scientific and medical knowledge.

3 Foundations

In this section, we describe in more detail the two FV approaches: the conventional three-part pipeline, serving as a baseline, and the step-by-step LLM-based system, which we mainly use.

3.1 Three-Part Pipeline for Fact Verification

The traditional three-part pipeline consists of: (1) document retrieval; (2) evidence extraction; (3) verdict prediction. It was used in the study by Vladika and Matthes (2024a), whose results we use as the baseline. Since it is an open-domain FV system, evidence documents have to be retrieved first. For that, step (1) was modeled with semantic search (similarity of query and corpus embeddings) over a large document corpus (PubMed and Wikipedia). In another experiment, evidence was sought with Google search. After selecting the top documents, step (2) again used a sentence embedding model to compare the claim to passages from the documents, selecting the most relevant evidence snippets. Finally, step (3) is modeled as the task of Natural Language Inference (NLI), where the goal is to predict the logical entailment relation between the claim and evidence, i.e., whether the claim is supported

[1] https://github.com/jvladika/StepByStepFV

by evidence (entailment), refuted by evidence (contradiction), or there is not enough information (neutral). The model was DeBERTa-v3 fine-tuned on various NLI datasets from Laurer et al. (2024).

3.2 Step-by-Step LLM System

The recent LLM advancements have brought a lot of features that can enhance the FV process. With their generative capabilities and multi-turn nature, LLMs can generate follow-up questions that aim to collect deeper background evidence related to claims. They are able to produce verdicts for claims over multiple pieces of evidence with mechanisms like chain-of-thought reasoning (Ling et al., 2023).

The system we develop in this work is mainly inspired by QACheck (Pan et al., 2023a) and its FV components. We expand that system by introducing novel prompts, additional chain-of-thought reasoning, amplify evidence retrieval with an online search engine, and experiment with structured reasoning in the form of logic predicates. The idea of this system is, given the claim c being verified, to generate up to five follow-up questions $q_1, ..., q_5$, which try to gather more evidence related to the claim. This is generated using a base LLM M_q and a prompt. Afterward, evidence for each question q is retrieved from the source s (web search or internal knowledge) using the method $R(q, s)$. This collected evidence is summarized with model M_s and together with original c posed to a reasoning model M_r. This reasoning module determines whether it should continue generating new questions or if there is enough evidence. If there is enough, it predicts a final verdict label v, one of SUPPORTED or REFUTED, and generates an explanation e.

On top of the described approach, we also experiment with a setting incorporating *predicate logic* into the process. Given the claim c, a predicate is generated by an LLM in the form of *verb(subject, object)*, such as *Treats(aspirin, headache)*, and used to generate better questions q_i and verdict v. Inspired by FOLK (Wang and Shu, 2023), the idea behind this is that the structured nature of predicates can help in finding more accurate evidence and introduce structured reasoning for the final verdict prediction (Strong et al., 2024).

4 Experiments and Setup

In the experiments, our main research question is **RQ:** *Does the iterative LLM approach outperform the traditional three-part pipeline for domain-specific fact verification?* On top of that, we test three further aspects of the system: (a) knowledge source, (b) structured reasoning, and (c) base LLM.

The knowledge sources include: internal knowledge of the LLM and the online search of the whole web. Our search engine of choice is DuckDuckGo, an open-source tool focused on privacy. We use it through a dedicated Python library.[2] This search engine provided a smooth search experience with no interruptions, and we deemed the quality of the retrieved results similar to the more popular Google or Bing for our use case. We take the provided *snippets* from the first 5 results and give them as input evidence to the reasoner LLM. The structured reasoning in (b) refers to using logic predicates, as described in the previous section. All the experiments in (a) and (b) were done using *GPT-4o-mini-2024-07-18* as the base LLM, the model from OpenAI with good reasoning capabilities (OpenAI, 2024).

In experiment round (c), we additionally test normal reasoning with internal knowledge and online search using Mixtral 8x7B (Jiang et al., 2024), a highly performing open-weights model based on a mixture-of-experts architecture, and LLaMa 3.1 (70B) (Meta, 2024), a recent advanced open-weights model from Meta. We use GPT through the OpenAI API and the two other models through the Together AI API,[3] setting temperature to 0 for best reproducibility and maximum tokens to 512. We use these LLMs for all parts of the fact verification process, i.e. for all steps M_q, M_s, M_r as described in the previous section. All the used prompts are in the Appendix. All experiments were run on one Nvidia V100 GPU with 16 GB VRAM.

4.1 Datasets and Evaluation

We choose three English datasets of biomedical and healthcare claims, designed for different purposes:

SCIFACT (Wadden et al., 2020) is a dataset with expert-written biomedical claims originating from citation sentences found in medical paper abstracts. The subset we use contains 693 claims, of which 456 are supported, and 237 are refuted.

HEALTHFC (Vladika et al., 2024a) is a dataset of claims concerning everyday health and spanning various topics like nutrition, the immune system, and mental health. The claims originate from user inquiries and they were checked by a team of medical experts. The subset we use contains 327 claims, of which 202 are supported, and 125 are refuted.

[2] https://pypi.org/project/duckduckgo-search/

[3] https://www.together.ai

verification system	evidence source	HealthFC			CoVERT			SciFact		
		P	R	F1	P	R	F1	P	R	F1
Three-part pipeline	PubMed	62.6	84.6	72.0	75.6	76.8	76.2	73.7	80.0	76.8
(with semantic search	Wikipedia	65.2	92.6	76.5	78.5	86.8	82.5	68.8	83.6	75.4
and DeBERTa)	whole web	62.3	92.6	74.5	76.4	68.7	72.3	75.5	91.5	82.7
GPT 4o-mini system	whole web	71.4	90.1	79.6	88.7	83.3	**85.9**	87.7	87.5	**87.6**
	internal	72.3	91.6	80.8	87.4	80.8	84.0	83.5	82.5	83.0
GPT 4o-mini system	whole web	74.9	88.6	81.2	90.1	68.7	77.9	88.2	82.2	85.1
(with predicates)	internal	73.7	91.6	**81.7**	89.1	70.2	78.5	84.9	77.9	81.2
Mixtral 8x7B system	whole web	68.2	78.7	73.1	79.8	81.8	80.8	82.0	86.2	84.1
	internal	68.5	74.3	71.3	86.9	77.3	81.8	80.9	83.3	82.1
LLaMa 3.1 (70B) system	whole web	74.3	88.6	80.8	79.1	89.9	84.2	86.1	82.7	84.3
	internal	64.7	86.1	73.9	74.3	81.8	77.9	80.0	87.5	83.6

Table 1: The results of the study. The first three rows come from a related study using the three-part pipeline. The further rows are from this study, using a consistent system with varying base LLM, structured reasoning type, and evidence source. The best F1 score for each dataset is in **bold**, while the second best is underlined.

CoVERT (Mohr et al., 2022) is a dataset of health-related claims, which are all causative in nature (such as "*vaccines cause side effects*"). All the claims originate from Twitter, which brings an additional challenge of informal language and provides a real-world scenario of misinformation checking. The subset we use contains 264 claims, of which 198 are supported, and 66 are refuted.

We find these three datasets to be well suited for our study because they are representative of three different applications of fact verification: helping researchers in their work (SciFact), verifying everyday user questions (HealthFC), and misinformation detection on social media (CoVERT).

We take claims from these datasets and use them as input to our system. To evaluate if the prediction is correct, we use the original veracity gold label. We do not give the system any original gold evidence documents from the datasets, as we are studying an open-domain setting. In essence, we evaluate the performance of the whole system by looking at its final classification performance as a "proxy" and observing how it changes when varying different parameters (Chen et al., 2024). While an important class in datasets is *not enough information* (NEI), we simplify the problem to only the *supported* and *refuted* classes and leave NEI for future work. Therefore, we use binary precision, recall, and F1 score as the evaluation metrics.

5 Results and Discussion

The first three rows of Table 1 show the results of the traditional three-part pipeline (described in Section 3.1) taken from the related study by Vladika and Matthes (2024a). It compared the performance over three knowledge sources: PubMed, Wikipedia, and online search. The results in further rows are from the experiments done in this study.

Improvement. As seen in Table 1, the step-by-step verification systems considerably improved the final F1 performance on all three datasets, especially precision values. The first GPT system improved the F1 performance by +4.3 on HealthFC, +3.4 on CoVERT, and +4.9 on SciFact, which is a major improvement when compared to the traditional pipeline using single-turn verification. This answers our main research question.

Internal vs. External Knowledge. Utilizing web search improved the performance in all cases for SciFact, showing that this dataset worked better when grounded to biomedical studies found online. For the other two datasets, which contain common health claims, there were instances where internal knowledge of LLMs even outperformed the web search. This is a very noteworthy finding, demonstrating how LLMs already encode a lot of internal medical knowledge that can be useful in knowledge-rich tasks, as observed by Singhal et al. (2023) and Vladika et al. (2024b). Similarly, Frisoni et al. (2024) showed how using LLM-generated evidence passages can improve medical QA performance more than retrieved passages.

Predicate Logic. The next experiment incorporated first-order-logic predicates into the FV process. In the GPT system, this resulted in the best overall performance for HealthFC, ending at 81.7 F1 (+5.2 improvement to baseline, +1 to without predicates). This is because predicates, like *Outcomes(Tamoxifen, Breast Cancer)*, led to more precise and targeted evidence, as indicated by the

	GPT 4o-mini	LLaMa 3.1	Mixtral 8x7B
Claim (Dataset)	EMDR works similarly to exposure therapy, another well-studied form of psychotherapy, for the symptoms of post-traumatic stress disorder. (**HEALTHFC**)	The relationship between a breast cancer patient's capacity to metabolize tamoxifen and treatment outcome is independent of the patient's genetic make-up. (**SCIFACT**)	Severe Covid is a systemic disease not just confined to resp system... causes hyper inflammation,clotting, encephalopathy etc..we need all hands on deck from generalists to specialists, nurses, ward attendants! (**COVERT**)
Generated Questions	**Q1:** How does EMDR therapy work? **Q2:** What are the key components of exposure therapy for PTSD? **Q3:** What are the similarities between EMDR and exposure therapy for PTSD? **Q4:** What are the differences between EMDR and exposure therapy for PTSD? **Q5:** What are the outcomes of EMDR compared to exposure therapy for PTSD?	**Q1:** What is tamoxifen used for in breast cancer treatment? **Q2:** Is the metabolism of tamoxifen influenced by genetic factors? **Q3:** What does research suggest about tamoxifen metabolism in breast cancer patients with different genotypes? **Q4:** Does the genetic variation in tamoxifen metabolism affect the treatment outcome in breast cancer patients?	**Q1:** What are some of the problems beyond respiratory issues that severe Covid causes? **Q2:** What types of medical professionals are needed to treat severe Covid? **Q3:** What medical equipment is used for patients with severe Covid in intensive care? **Q4:** How is hyperinflammation treated in Long Covid patients? **Q5:** How is anticoagulation managed in Long Covid patients to prevent clotting?

Table 2: Examples of three claims (all supported) from the three datasets used in the study, with generated verification questions from the three different LLMs. GPT generates the most general questions with wider coverage, while LLaMa and Mixtral generate more specific and in-depth questions.

increase in precision scores. On the other hand, while precision also increased for the other two datasets, it led to large drops in recall, resulting with a lower F1. This was especially seen with informal language in CoVERT claims, where produced predicates included underspecified instances like *Has(Person, Covid)*, which only degraded the evidence retrieval process. Therefore, predicates are better suited for clearly written queries and for complex claims.

Choice of LLMs. Comparative analysis of different LLMs was the last round of experiments. Overall, GPT-4o-mini came out on top as the best LLM for the task. Table 2 shows an example of generated questions for all three LLMs for different claims. It is evident that GPT gives the most general and simplest questions, whereas LLaMa and Mixtral provide more specific and detailed questions. The specific questions can be a strength but also complicate the evidence retrieval process with noisy retrieved passages. GPT was the best at following the style of few-shot example questions. Also, Mixtral produces the most questions on average per claim, followed by GPT, and then LLaMa. Finally, we observed the reasoning capabilities of models to be on a similar level, showing the final performance is often dependent on the quality of question generation and answering.

Qualitative Analysis. As evident in Table 2, a lot of generated questions were asking for definitions of the diseases, symptoms, drugs, and other terms found in claims. Once such complex terms were described, the FV process was well-equipped to continue with the verification. This explains *why* the step-by-step systems worked so well for medical claims, similarly to multi-hop claims in previous studies – they inherently contain complex concepts and relations that shall be clarified first before making the final decision.

A common reason for errors in the system was the generated questions going too in-depth about a certain point with its follow-up questions and not collecting wider evidence about other parts of the claim. Moreover, another issue were *knowledge conflicts* – when the LLM would predict an incorrect label even when shown evidence to the contrary because of its encoded internal knowledge.

Future work could expand the system to leverage structured knowledge sources like knowledge graphs (Kim et al., 2023) or use methods like formal proof generation (Strong et al., 2024). The final step of the system focusing on explanation generation should ideally include different user perspectives in the process (Warren et al., 2025).

6 Conclusion

In this study, we develop a step-by-step system for fact verification based on iterative question generation and explainable reasoning. We apply the system on three medical fact-checking datasets and test different settings. We show that by utilizing LLMs, this system can create follow-up questions on complex concepts and relations from the claims in order to gather background evidence, reason over newly discovered evidence, and finally lead to predictions that achieve higher results when compared to traditional pipelines. We hope that our study encourages more exploration of advanced systems for domain-specific fact verification.

Limitations

Since all modules of the step-by-step verification system rely on using LLMs, they come with their own set of challenges and limitations. The generated follow-up questions are not always perfect or precise, the generated evidence snippets can be off point, and the reasoning over long chains of evidence can, of course, lead to logical errors and mistakes. We observed certain instances where even though all the evidence was pointing towards one of the verdicts (*refuted*), the system would still mistakenly output the other one (*supported*).

Another limitation comes from the high complexity of the system and reliance on calls to external APIs, including LLM APIs and search engine APIs. This inevitably led to some challenges in terms of slower processing speed of this system when compared to traditional approaches that use an out-of-the-box NLI model like DeBERTa. Still, we were forced to rely on API calls for LLMs due to hardware resource limitations, but models like Mixtral and LLaMa showed decent performance and are open-weights, so they can be downloaded and run locally to speed up the performance.

Lastly, for easier evaluation we disregard claims annotated with *Not Enough Information* due to different definitions of this label across different datasets (e.g., the definition from SciFact does not serve the open-domain setting well). This is an important label in fact verification, since not all claims can be conclusively assessed for their veracity. This is especially important in the scientific domain considering the constantly evolving nature of scientific knowledge, and sometimes conflicting evidence from different research studies. Future work should find a way to effectively include this label into model predictions.

Ethics Statement

Our dataset and experiments deal with the highly sensitive domain of healthcare and biomedical NLP. While we observed good scores when verifying health-related question using responses directly generated by language models, this is not a recommended way of using them by end users or patients. Responses can still contain hallucinations or misleading medical advice that should always be manually verified within reliable sources. Similarly, experiments using online search results did not go through any manual quality filtering, which means not all of them will be trustworthy or approved by experts. One should always consult with medical professionals when dealing with health-related questions and advice.

Acknowledgements

This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München). We would like to thank the anonymous reviewers for their valuable feedback.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6(8):852–863.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. Claim verification in the age of large language models: A survey. *Preprint*, arXiv:2408.14317.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *Preprint*, arXiv:2502.05151.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692, Bangkok, Thailand. Association for Computational Linguistics.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023a. QACheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 264–273, Singapore. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. 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.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023a. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023b. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Dominik Stammbach, Boya Zhang, and Elliott Ash. 2023. The choice of textual knowledge base in automated claim checking. *ACM Journal of Data and Information Quality*, 15(1):1–22.

Marek Strong, Rami Aly, and Andreas Vlachos. 2024. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17021–17035, Miami, Florida, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28:460 – 467.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024a. Comparing knowledge sources for open-domain scientific claim verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2103–2114, St. Julian's, Malta. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024b. Improving health question answering with reliable and time-aware evidence retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4752–4763, Mexico City, Mexico. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024a. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024b. MedREQAL: Examining medical knowledge recall of large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14459–14469, Bangkok, Thailand. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. 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.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. *arXiv preprint arXiv:2502.09083*.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024. Do we need language-specific fact-checking models? the case of Chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1914, Miami, Florida, USA. 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.

A Appendix

In the appendix, we provide the prompts used for the systems (Figures 2–7).

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, the first simple question we need to ask is:
Question = Is Superdrag a rock band?

Claim = Jimmy Garcia lost by unanimous decision to a professional boxer that
challenged for the WBO lightweight title in 1995.
To validate the above claim, the first simple question we need to ask is:
Question = Who is the professional boxer that challenged for the WBO
lightweight title in 1995?
```

Figure 2: Two out of ten few-shot examples used in the prompt for generating the first verification question.

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we need to ask the following simple questions
sequentially:
Question 1 = Is Superdrag a rock band?
Answer 1 = Yes
Question 2 = Is Collective Soul a rock band?

Claim = Jimmy Garcia lost by unanimous decision to a professional boxer that
challenged for the WBO lightweight title in 1995.
To validate the above claim, we need to ask the following simple questions
sequentially:
Question 1 = Who is the professional boxer that challenged for the
WBO lightweight title in 1995?
Answer 1 = Orzubek Nazarov
Question 2 = Did Jimmy Garcia lose by unanimous decision to Orzubek Nazarov?
```

Figure 3: Two out of ten few-shot examples used in the prompt for generating the follow-up questions (after the first one had been generated).

```
Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we have asked the following questions:
Question 1 =to explainAnswer 1 = Yes
Can we know whether the claim is true or false now?
Prediction = No, we cannot know.

Claim = Superdrag and Collective Soul are both rock bands.
To validate the above claim, we have asked the following questions:
Question 1 = Is Superdrag a rock band?
Answer 1 = Yes
Question 2 = Is Collective Soul a rock band?
Answer 2 = Yes
Can we know whether the claim is true or false now?
Prediction = Yes, we can know.
```

Figure 4: Two out of ten few-shot examples for the verifier module. In this step, the LLM decides if there is enough evidence to make the final veracity prediction or if question generation shall continue.

```
Claim: Superdrag and Collective Soul are both rock bands.

To validate the above claim, we need to ask the first question with predicate:
Question:
Is Superdrag a rock band?
Predicate:
Genre(Superdrag, rock) ::: Verify Superdrag is a rock band

Claim : Jimmy Garcia lost by unanimous decision to a professional boxer that
challenged for the WBO lightweight title in 1995.

To validate the above claim, we need to ask the first question with predicate:
Question:
Who is the professional boxer that challenged for the WBO lightweight title
in 1995?
Predicate:
Challenged(player, WBO lightweight title in 1995) ::: Verify name of the
professional boxer that challenged for the WBO lightweight title in 1995.
```

Figure 5: Two out of ten few-shot examples for question generation in the predicate pipeline. Each generated question is accompanied by a predicate defining the question and a simple instruction on what to verify.

```
Claim: Superdrag and Collective Soul are both rock bands.

Question 1:
Is Superdrag a rock band?
Predicate 1:
Genre(Superdrag, rock) ::: Verify Superdrag is a rock band
Answer 1:
Yes

To validate the above claim, we need to ask the follow-up question with predicate:
Follow-up Question:
Is Collective Soul a rock band?
Predicate:
Genre(Collective Soul, rock) ::: Verify Collective Soul is a rock band
```

Figure 6: One out of then few-shot examples of follow-up question generation for the predicate system. The already gathered evidence and predicates from previous questions are given.

```
Question:
Is it true that The writer of the song Girl Talk and Park So-yeon have both
been members of a girl group.?
Context:
Write(the writer, the song Girl Talk) ::: Verify that the writer of the song
Girl Talk
Member(Park So-yeon, a girl group) ::: Verify that Park So-yeon is a member
of a girl group
Member(the writer, a girl group) ::: Verify that the writer of the song Girl
Talk is a member of a gril group

Who is the writer of the song Girl Talk? Tionne Watkins is the writer of the
song Girl Talk.
Is Park So-yeon a member of a girl group? Park Soyeon is a South Korean singer.
She is a former member of the kids girl group I& Girls.
Is the writer of the song Girl Talk a member of a girl group? Watkins rose to
fame in the early 1990s as a member of the girl-group TLC
Prediction:
Write(Tionne Watkins, the song Girl Talk) is True because Tionne Watkins is the
writer of the song Girl Talk.
Member(Park So-yeon, a girl group) is True because Park Soyeon is a South Korean
singer. She is a former member of the kids girl group I& Girls.
Member(Tionne Watkins, a girl group) is True because Watkins rose to fame in the
early 1990s as a member of the girl-group TLC
Write(Tionne Watkins, the song Girl Talk) && Member(Park So-yeon, a girl
group) && Member(Tionne Watkins, a girl group) is True.
The claim is [SUPPORTED].
Explanation:
Tionne Watkins, a member of the girl group TLC in the 1990s, is the writer of
the song "Girl Talk."
Park Soyeon, a South Korean singer, was formerly part of the girl group I& Girls.
Therefore, both Watkins and Park Soyeon have been members of girl groups in
their respective careers.
```

Figure 7: One example used in the prompt for the reasoning module using predicates.