Exploring Human-AI Synergy for Complex Claim Verification

Mukherjee, Shubhalaxmi; Jonker, Catholijn M.; Murukannaiah, Pradeep K.

DOI
10.3233/FAIA250620

Publication date
2025

Document Version
Final published version

Published in
Frontiers in Artificial Intelligence and Applications

Citation (APA)
Mukherjee, S., Jonker, C. M., & Murukannaiah, P. K. (2025). Exploring Human-AI Synergy for Complex Claim Verification. *Frontiers in Artificial Intelligence and Applications*, *408*, 2-15. https://doi.org/10.3233/FAIA250620

HHAI 2025
D. Pedreschi et al. (Eds.)

doi:10.3233/FAIA250620

Exploring Human-AI Synergy for Complex Claim Verification

Shubhalaxmi MUKHERJEE [a,1], Catholijn M. JONKER [a] and
Pradeep K. MURUKANNAIAH [a]
[a] *Delft University of Technology*
ORCiD ID: Shubhalaxmi Mukherjee
https://orcid.org/https://orcid.org/0009-0004-0856-4107, Catholijn M. Jonker
https://orcid.org/https://orcid.org/0000-0003-4780-7461, Pradeep K. Murukannaiah
https://orcid.org/https://orcid.org/0000-0002-1261-6908

Abstract. Combating widespread misinformation requires scalable and reliable fact-checking methods. Fact-checking involves several steps, including question generation, evidence retrieval, and veracity prediction. Importantly, fact-checking is well-suited to exploit hybrid intelligence since it requires both human expertise and AI's large-scale information processing abilities. Thus, constructing an effective fact-checking pipeline requires a systematic understanding of the relative strengths and weaknesses of humans and AI in different steps of the fact-checking process. We investigate the ability of LLMs to perform the first step of the process, i.e., to generate pertinent questions for analyzing a claim. To evaluate the quality of the LLM-generated questions, we crowdsource a dataset in which 150 claims are annotated with questions (1) a novice fact-checker would ask and (2) a professional fact-checker would ask when fact-checking those claims. We study the effects of the human- and LLM-generated questions on evidence retrieval and veracity prediction. We find that LLMs are able to generate nuanced questions to verify a complex claim, but the final label prediction depends on the quality of the evidence corpus. However, the evidence collected by automated methods yields lower accuracy in the veracity prediction task than the evidence curated by experts.

Keywords. Fact checking, Claim decomposition, Large Language Models

1. Introduction

The prevalence of fake news and misinformation is a major societal problem. Traditionally, detecting misinformation has been carried out exclusively by expert fact-checkers. However, fact-checking by experts is a slow and expensive process that does not scale with the rapidly growing volume of online misinformation. To address this challenge, researchers are working on automated methods to identify misinformation, with substantial progress being made in developing fast, scalable, and cutting-edge Artificial Intelligence (AI) methods [1,2,3,4]. However, the performance of fully automated methods remains subpar compared to experts, especially for complex real-world claims [5,6].

[1] Corresponding Author: Shubhalaxmi Mukherjee, S.Mukherjee-2@tudelft.nl.

We explore fact-checking as a human-AI collaborative process as opposed to a fully manual or fully automated process. As Figure 1 shows, fact checking involves several steps, requiring different skills or capacities. For example, question generation requires creativity and domain understanding, whereas evidence retrieval requires large-scale information processing. Accordingly, some tasks in a fact-checking pipeline are better suited for humans and others for AI. Thus, fact-checking is, inherently, a process that requires hybrid intelligence [7]. However, there is a lack of systematic studies investigating the relative strengths of humans and AI in fact-checking steps.

To fact check a claim, one must first know what evidence to look for. It requires knowledge of the subject matter and ability to decompose the implicit and explicit sub claims within the original claim. Recent works [8,9] propose to decompose complex claims into logical forms using LLMs before verification. But this doesn't address the implicit claims or intention of the speaker. Consider, for example, the claim in Figure 1. To verify this claim, a fact checker must ask (1) whether the national debt really increased by $3 trillion dollars under President Trump, (2) whether the tax cut exclusively benefited the wealthiest few, and (3) whether the tax cut is solely to blame for the increase in the debt. A decomposition of this claim text yields the first two questions but not the third, which is not explicit in the claim text. As it turns out, in this case, the answers to the first two questions support the claim, but the answer to the third (i.e., the implicit blame assigned by the speaker) is untrue, making the claim 'half true.' As this example demonstrates, the ability to ask fine-grained questions is the crux of complex fact checking.

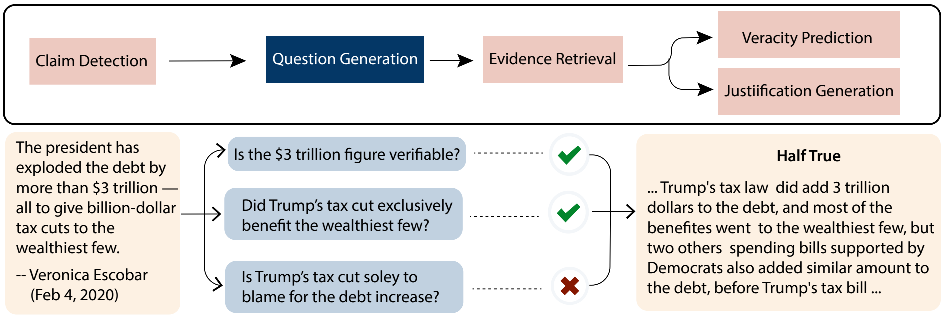
Claim Detection
Question Generation
Evidence Retrieval
Veracity Prediction
Justiification Generation
The president has exploded the debt by more than $3 trillion — all to give billion-dollar tax cuts to the wealthiest few.
-- Veronica Escobar (Feb 4, 2020)
Is the $3 trillion figure verifiable?
Did Trump's tax cut exclusively benefit the wealthiest few?
Is Trump's tax cut soley to blame for the debt increase?
Half True
... Trump's tax law did add 3 trillion dollars to the debt, and most of the benefites went to the wealthiest few, but two others spending bills supported by Democrats also added similar amount to the debt, before Trump's tax bill ...

Figure 1. An example showing the importance of question generation in a complex claim verification pipeline.

In this work, we compare the performance of humans and LLMs to ask nuanced questions and the effect of these questions on the veracity prediction accuracy. To this end, first, we contribute a novel dataset in which 150 claims are annotated with questions necessary to fact-check the claims. The analysis includes three types of questions for each claim: (1) questions generated by an LLM, (2) questions a novice fact-checker would ask (crowdsourced), and (3) questions a professional fact-checker would ask (crowdsourced after the crowd workers read a journalist's fact-checking article of the claim).

Then, we perform a comprehensive analysis of how the three question sets differ, and importantly, how they affect the downstream task of veracity prediction. We find that LLMs are capable of generating questions that capture both implicit and explicit aspects of a claim. Prompting with examples helps to reduce the unnecessary questions, which leads to a more concise evidence set and better veracity prediction. Yet, the questions generated by professional fact-checkers yield the best results by a significant margin.

Once the questions are generated, answering these questions requires gathering evidence from different documents. This is a very challenging task to automate. For real-world claims, such as political claims, evidence retrieval is challenging even for journalists. We study two variations of the evidence retrieval task. (1) We use the analysis article written by the journalists at Politifact as the evidence corpus. We assume that all the necessary information is contained in the article, so questions from all generation methods are answered from this article. This helps us compare the effect of question generation, without the confounding effect of evidence retrieval from search engines. (2) We extract keywords from each claim and collect the relevant results from Google search to compile the evidence set. Then the questions are answered from this corpus.

We find that noisy evidence collected from Google leads to a significant drop in veracity prediction accuracy. Thus, although LLMs have impressive abilities to analyze large volumes of text, they struggle with reasoning when presented with noisy evidence. Subtle differences in the evidence set, such as ordering of supporting and refuting evidence, interferes with veracity prediction. Even when all the necessary information is collected in the question-answering phase, the presence of redundant or repeated information in the evidence paragraph throws off the reasoning process.

Contributions. First, we contribute a novel dataset. (1) It contains questions generated by novice fact-checkers, establishing a new baseline for a priori reasoning abilities for human fact checkers. (2) We reverse engineer questions from professional fact-checkers based on the full article published by Politifact. These two question sets serve as a benchmark for evaluating LLM-generated questions. Second, we evaluate a complete fact-checking pipeline using GPT-4o and Llama-8b, providing key insights on LLM's capabilities to ask and reason about nuanced questions about complex claims. We perform our evaluation both in perfect and noisy evidence settings. Our findings—on question generation and evidence retrieval—suggest that fully automating fact-checking is far from ideal and that a hybrid human-AI approach is essential for reliable fact-checking.

Data and Code The dataset, prompts, code, and crowdsourcing study materials used in this paper are available at https://github.com/ShubhalaxmiM/HHAI_2025.git.

2. Related Work

Many fact-checking datasets have been proposed to evaluate claim verification accuracy of models, such as FEVER [10], HoVeR [11], and Fact or Fiction [12]. However, retrieving evidence to verify the claims in these datasets is easy—often, there is direct evidence to support or refute a claim. The evidence for claims in these datasets can often be found simply in Wikipedia. However, real-world fact checking is rarely that easy. It requires subjective judgment and expertise.

Early fact-checking systems simply used the claim as the query to retrieve evidence. Many datasets have been proposed [13,14,15,16], where only the text of the claim is used to query a search engine. However, in such cases, the search engine may access fact-checking websites that have published verdicts about the claims. This is not viable for new claims. In our work, we query Google using relevant keywords only and filter out fact checking websites from Google search results to prevent data leakage.

Further, natural claims, like those made by public figures or the ones found on social media, tend to be complex. Therefore, claim decomposition has become a focus of research in natural language processing [5]. Chen et al. [17] showed that generative models can learn to decompose claims into explicit and implicit questions. Compiling background information about various parts of a claim has been shown to help crowd workers analyze complex claims [18]. Question answering also helps to trace the reasoning path to reach a verdict [19]. However, these works restrict the type of questions that can be asked and answered in the annotation process, to simplify the reasoning process. In contrast, we ask annotators to generate as many questions as they can think of.

Finally, a comprehensive review highlights that despite advancements, NLP-based fact-checking tools still face challenges in effectively assisting human fact-checkers, particularly with complex claims [6]. Recently, Human-In-The-Loop approaches have gained importance in NLP research [20,21,22,23,24]. However, there is a lack of insights on the relative strengths and weaknesses of humans and AI in different parts of a fact-checking pipeline. We seek to fill this gap by systematically evaluating the performance of humans (novices and experts) and LLMs in the question generation and evidence retrieval phases, and how this affects downstream veracity prediction.

3. Dataset

We select claims for question generation from PolitiHop [16]. We exclude claims that require verifying the content of a photograph, video, or someone's social media post as that makes crowdsourced question generation difficult. We replace these claims with other claims from Politifact collected in a similar timeframe as PolitiHop. Similar to Wang et al. [9], we consider three classes: True, Half True, and False. Ultimately, we have a balanced dataset of 50 claims in each class.

3.1. Crowdsourced Task

We employed annotators from the Prolific platform. Our data collection process was approved by a Human Research Ethics Committee.

We restricted the pool of annotators to those whose employment role in Prolific was 'Journalist'. This was done in order to enhance the quality of the questions generated. However, we could not establish the nature of the workers' journalistic experience. Further, these workers spent a short amount of time in fact-checking a claim compared to professional fact-checkers. Thus, we treat the crowd workers as novice fact-checkers.

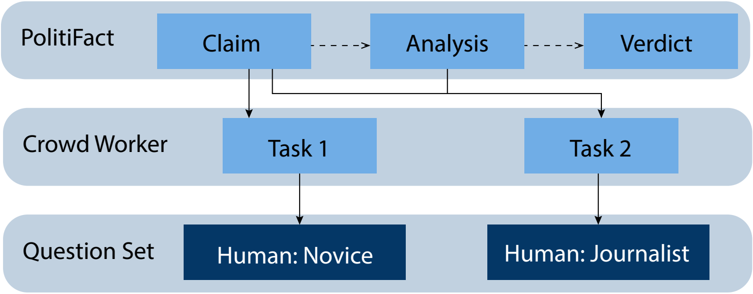

Figure 2. Crowdsourcing task workflow

Figure 2 shows an overview of our data collection process involving two tasks.

Task 1 The annotator was shown claim, speaker or source of the claim, and the date when the claim was made. They were asked to start with a neutral stance on the claim and generate questions they would ask in order to prove or disprove the claim.

Task 2 The annotator (who has completed Task 1) is now shown the fact-checking article written by the Politifact journalists, and asked to reverse engineer the questions the original journalists had considered when writing the analysis.

The Politifact label was never shown to the annotators as it might bias their questions. Each annotator was given three claims to annotate, and each claim was annotated by at least two annotators. The average duration for an annotator to complete both tasks was approximately 50 minutes. Additional details, including the instructions provided to the annotators, are included in the repository.

4. Experiments

4.1. Question Generation

We experiment with the following variants for question generation.

LLM: Direct The LLM is prompted to come up with questions. No examples are provided. Figure 3 shows the prompt used.

LLM: Few Shot The LLM is prompted with examples of claims and the questions that were asked by professional journalists (obtained in the crowdsourcing task) to verify the claim. Figure 4 shows the prompt used.

Human: Novice We use the novice questions obtained from Task 1 (Figure 2).

Human: Journalists We use the journalists' questions obtained from Task 2 (Figure 2).

LLM: Direct Prompt

Please tell me the necessary questions that need to be answered in order to verify the following claim. Take into account the speaker and the date of the claim, if necessary. Generate as many questions as you can think of:

—–

Claim: 90% of policemen are for expanding background checks to all gun sales.
Speaker/Source: Richard Durbin
Date: September 23, 2019

Figure 3. Direct Prompt used to generate questions, without any examples.

We experiment with GPT-4o and Llama-3.1-8b-Instruct in this work, for question generation and other steps of the pipeline. We set the temperature to 0 for GPT-4o and 0.1 for LLama-3.1-8b-Instruct (the LLama model we used did not allow setting temperature to 0). We use low temperature settings to reduce variability in the model's responses. We set the maximum token length to 1024 for both models.

LLM: Few Shot Prompt

Please tell me the necessary questions that need to be answered in order to verify the following claim. Take into account the speaker and the date of the claim, if necessary. Generate as many questions as you can think of:

Claim: Back when I was studying it, two out of three families that ended up in bankruptcy after a serious medical problem had health insurance.
Speaker: Elizabeth Warren
Date: October 15, 2019

»»»

Followup Question: Is there any published research on the relationship between medical debt and bankruptcy?
Followup Question: What do people cite as reasons for bankruptcy in bankruptcy applications?
Followup Question: When was Warren studying the issue of medical debt and bankruptcy?
Followup Question: Has something changed in health insurance policy between Warren's research period and when this claim was made?
Followup Question: How many families ended up in bankruptcy after a serious medical problem having health insurance?

——

Claim: {}

Figure 4. Few shot prompt used to generate questions. Whereas this prompt shows one claim as in-context example, the full prompt consists of six in-context examples.

4.2. Evidence Retrieval

To answer the questions, we experiment with two methods of evidence collection.

Closed Domain: We use the full analysis article published by Politifact as the source of all evidence. This is article was written after thorough research by professional journalists. Therefore, we assume that all the necessary information to verify a claim are available.

Open Domain: We first extract keywords from each claim using GPT-4o and submit them to Google through SERP API. We set the date filter to search results only available before the date of the claim, to prevent temporal leakage. We collect the top 10 results per claim and strictly filter out any results from over 150 fact-checking domains to prevent any leakage from their justification. Then, we scrape the full text from each of the URLs to compile the evidence set.

The same set of questions generated in Section 4.1 are used in both cases. Thus, we study the importance of evidence curated by experts versus automated evidence collection.

4.3. Veracity Prediction

Each question from the question generation step is answered by the LLM based on the evidence corpus. The answer to each question is concatenated to the prompt as a single paragraph. Since the rationale for the label must be the same in all circumstances, the same prompt is used for label prediction, regardless of the variations in the question generation prompts.

The different question sets obtained in the question generation set leads to different evidence sets. We measure the effect of the different evidence sets by the label accuracy. We also determine the highest achievable result for the 'entailment' task, i.e., the model predicts the veracity label based on the full evidence article from Politifact. We call this the **Full Evidence** setting. We also measure the RMSE between the actual (y_i) and predicted label ($\hat{y}_i$), $\sqrt{\sum_{i=1}^{150}(y_i - \hat{y}_i)^2)/150}$ where $y_i, \hat{y}_i \in \{0,1,2\}$.

5. Results and Discussion

First, we perform a quantitative evaluation on the effect of different question generation methods on the downstream task of veracity prediction in closed and open domain settings. However, the quantitative evaluation does not reveal the complexities in the intermediate steps of the fact-checking pipeline. Thus, we present some notable observations obtained by manual inspection of the results of each step of the pipeline as qualitative evaluation.

5.1. Quantitative Evaluation: Closed Domain Evidence Retrieval

GPT-4o: Table 1 shows the F1 scores for veracity prediction for GPT-4o. We see that the GPT-4o based pipeline is most effective for True claims, for all variations of the prompt. On overall F_1 score, LLM: Few Shot and Human: Novice perform similarly but slightly better than LLM: Direct. However, the question set from Human: Journalist outperforms the other variations with a large margin. This pattern also holds for RMSE.

Method	Per-Class F_1 ↑			Macro F_1 ↑	RMSE ↓
	True	Half True	False		
GPT-4o: Direct	72.05	43.67	70.58	62.10	0.699
GPT-4o: Few Shot	71.60	49.90	68.02	63.17	0.710
Human: Novice	68.31	51.77	68.83	62.97	0.717
Human: Journalist	74.95	58.95	70.85	68.25	0.628
Topline: Full Evidence	71.7	65.81	75.05	70.85	0.579

Table 1. F1 scores of the label prediction task with GPT-4o.

The confusion matrices in Table 2 confirm that there is a significant True bias in the predictions made by GPT-4o. However, this bias is reduced when questions are not generated by GPT-4o, as demonstrated in the confusion matrices 2c and 2d.

Actual \ Predicted	T	H	F
T	42	7	1
H	21	18	11
F	6	10	33

(a) GPT-4o: Direct

Actual \ Predicted	T	H	F
T	39	8	3
H	16	22	12
F	4	13	33

(b) GPT-4o: Few Shot

Actual \ Predicted	T	H	F
T	37	9	4
H	15	26	9
F	4	13	33

(c) Human: Novice

Actual \ Predicted	T	H	F
T	35	13	2
H	7	29	14
F	2	13	35

(d) Human: Journalist

Table 2. Confusion matrices showing the True bias and label mismatches for GPT-4o

Llama-8b: Table 3 shows the F1 scores for veracity prediction for Llama-3.1-8b-Instruct. The Llama based pipeline performs significantly worse than GPT-4o in overall F1 score, likely due to the smaller size of the model. However, unlike GPT-4o based pipeline, in the Llama based pipeline, LLM: Few Shot performs better than the other variations of the question generation prompt including the Human variants. However, since the overall performance is quite low, we cannot draw strong conclusions on the relative performances.

Method	Per-Class F_1 ↑			Macro F_1 ↑	RMSE ↓
	True	Half True	False		
Llama-3.1-8b-Instruct: Direct	31.03	53.63	26.23	36.96	0.764
Llama-3.1-8b-Instruct: Few Shot	33.33	52.98	53.93	46.64	0.779
Human: Novice	21.05	51.69	36.93	36.56	0.757
Human: Journalist	32.79	55.49	39.39	43.65	0.753
Topline: Full Evidence	36.06	58.14	50.75	48.32	0.693

Table 3. F1 scores of the label prediction task with Llama-8b.

The confusion matrices in Table 4 also reveal a different pattern than Table 2. Whereas GPT-4o showed a True bias, Llama-3.1-8b-Instruct shows a significant bias to predict most claims as Half True. However, similar to GPT-4o, the Few shot prompt and Human: Journalist prompts reduce this bias, as seen in matrices Tables 4b and 4d.

Number of Questions: To further analyze the differences between the performances different question generation pipelines, we analyze the number of questions generated as shown in Table 5. First, it is striking that Llama-8b generates a large number of questions in the LLM: Direct variation compared all other variants. This provides an intuition on the Half-True bias of this variant. That is, when the Llama-8b model cannot find answers to all the questions asked from the given evidence corpus, it has a tendency to label the claim as Half True. Second, we notice that the Human: Novice yields the least number of questions among the variants. This shows that question generation is difficult for non-professionals.

		Predicted		
		T	H	F
Actual	T	9	39	2
	H	0	48	2
	F	0	42	8

(a) Llama-3.1-8b-Instruct: Direct

		Predicted		
		T	H	F
Actual	T	10	35	5
	H	0	40	10
	F	0	26	24

(b) Llama-3.1-8b-Instruct: Few Shot

		Predicted		
		T	H	F
Actual	T	6	44	0
	H	1	46	3
	F	0	38	12

(c) Human: Novice

		Predicted		
		T	H	F
Actual	T	10	38	2
	H	1	48	1
	F	0	37	13

(d) Human: Journalist

Table 4. Confusion matrices showing the True bias and label mismatches for Llama-8b

Question Generation Source	Number of Questions
GPT-4o: Direct	21.6 ± 3.1
GPT-4o: Few Shot	15.1 ± 1.9
Llama-3.1-8b-Instruct: Direct	56.3 ± 49.2
Llama-3.1-8b-Instruct: Few Shot	8.6 ± 7.7
Human: Novice	5.8 ± 2.4
Human: Journalist	13.5 ± 0.03

Table 5. Number of questions generated per claim.

Since GPT-4o's accuracy is substantially better than Llama-3.1-8b-Instruct in the closed domain setting, we conduct the remaining experiments only on the GPT-4o based pipeline.

5.2. *Quantitative Evaluation: Open Domain Evidence Retrieval*

Table 6 shows the F1 scores for veracity prediction for GPT-4o in the open domain setting. The overall F1 scores are significantly lower in the open than the closed domain setting. These results show that evidence retrieval has an important role in fact-checking.

Method	Per-Class F_1 ↑			Macro F_1 ↑	RMSE ↓
	True	Half True	False		
GPT-4o: Direct	46.58	54.55	56.25	52.46	0.983
GPT-4o: Few Shot	50.63	51.49	63.33	55.15	0.883
Human: Novice	55.26	49.54	57.39	54.06	0.860
Human: Journalist	48.64	50.45	64.35	54.48	0.828

Table 6. F1 scores of the label prediction task with GPT-4o with open domain question evidence retrieval.

Table 7 shows the confusion matrices for the open domain setting. In this case, we find a greater tendency to predict False. While the overall performance is poorer than closed domain, the classification accuracy for Half True class is higher in this setting.

		Predicted		
		T	H	F
Actual	T	17	12	21
	H	2	27	21
	F	4	10	36

(a) GPT-4o: Direct

		Predicted		
		T	H	F
Actual	T	20	16	14
	H	6	26	18
	F	3	9	38

(b) GPT-4o: Few Shot

		Predicted		
		T	H	F
Actual	T	21	16	13
	H	4	27	19
	F	1	16	33

(c) Human: Novice

		Predicted		
		T	H	F
Actual	T	18	21	11
	H	17	28	17
	F	1	12	37

(d) Human: Journalist

Table 7. Confusion matrices for GPT-4o based pipeline and open domain evidence retrieval.

5.3. *Qualitative Evaluation*

We begin our analysis by examining the LLM-generated questions. We manually analyzed the questions generated for 50 claims and categorized each question as ‘relevant’ if it is pertinent to the verification rationale and is answerable from the Politifact article, or ‘irrelevant’, otherwise. We find the number of relevant questions in the LLM: Direct and LLM: Few Shot settings as 55.6% and 59.6%, respectively. The problem with a large number of irrelevant questions is that, when a question cannot be answered from the available evidence, the LLM sometimes hallucinates the answer (as also observed by Ji et al. [25]), which in turn leads to inaccurate veracity prediction.

By further examining the LLM-generated questions, we find that some questions assume a part of the claim to be true. For example, for the claim: *“The Machinists Union does not (support the USMCA trade deal). And every environmental organization in this country(also) opposes it,”* one of the questions asked by LLM is *What specific reasons does the Machinists Union have for not supporting the USMCA trade deal?* This assumes that the former part of the claim is true.

We took a closer look at the misclassified claims in the Direct and Few Shot settings. We find that more questions retrieve the supporting evidence than the falsifying evidence. Therefore, at the reasoning step, the supporting evidence is presented more number of times, while the falsifying evidence is lost in the middle. With a more concise question set, as in Human: Journalist, there is lesser repetition of evidence, which we conjecture yields better accuracy, especially for Half True claims.

We also observe that, when the questions are answered by a generative model like GPT-4o, it introduces judgments about the evidence in the sentence structure. For example, for the claim *“Gov. Tony Evers proposed raising taxes on the agriculture industry to pay for expanded welfare programs,”* a question from the Direct set is answered as *“Specifically, he proposed limiting the credit for manufacturing to $300,000 per tax year, which would have reduced the tax credit for agriculture as well.*” However, the Politifact article specifically says that the agricultural portion of the Manufacturing and Agriculture Tax Credit was unchanged. So, this sentence in the LLM’s answer introduces incorrect reasoning in the evidence.

Finally, we take a closer look at some of the claims individually. Specifically, we focus on some examples that are consistently misclassified across all the runs. First, we observe that a large majority of these examples are Half True claims, demonstrating the complexity of this class. Then, we identify the potential reason for the mistakes in LLM reasoning going by the justification produced. Consider the claim *"Lead levels in Milwaukee's water are higher than Flint, MI."* Journalists of Politifact rate this as Half True, because, while lead levels in Milwaukee's water is true strictly in terms of the numbers, but the reasons are different in the two cities. This reason is picked up in the evidence retrieval phase, but the LLM does not consider this, therefore labeling it True. In another claim, *"One in three women is sexually assaulted on the dangerous journey north,"* the claim is supported by a single survey. When evidence for a claim is survey data, journalists look at the parameters of the survey to see if there are severe limitations. The survey that supports this claim has a severe limitation that the majority of 400 migrants surveyed were men. The LLM's justification mentions this limitation, yet the label is predicted as True.

6. Conclusion

We study the utility of Large Language Models in the question generation phase in a fact-checking pipeline. To facilitate research on this phase, we contribute a novel dataset of 150 complex political claims, annotated with a comprehensive set of questions. We also evaluate a fact-checking pipeline, from question generation to label prediction, that exploits human-LLM synergy. As LLMs are being utilized in every profession, it is crucial that we understand the relative strengths of these models and where they could augment human intelligence. Our experiments show that the LLMs perform quite well on the question generation, but their reasoning ability is well below that of journalists'. Different LLMs have different biases, so one must be cautious while using results from any of them. So, LLMs could play a role in compiling evidence, but expert human judgment is indispensable when it comes to determining truth and falsehood.

6.1. Limitations

We identify four key limitations of our work.

Diversity of Claims The claims in our dataset are all from one source (Politifact), which mainly includes claims from politicians in the United States. This serves our purpose as most such claims are complex, but limits the diversity of the claims. Claims from other domains, e.g., health claims or scientific claims, and claims from other countries and languages, may bring nuances to complexity that we have not considered in this paper. A systematic analysis of how different sources of diversity influence fact-checking systems, e.g., as done in other NLP tasks [26,27], is an interesting avenue for future work.

Explainability and Faithfulness Explainability is a serious issue with black box models by GPT-4o. "Chain-of-Thought" reasoning [28] has been very successful in improving LLM reasoning and explainability. Yet, faithfulness of the model's explanation on grounding documents is questionable [29]. In fact checking, even though we provide external knowledge sources, LLMs may introduce unfaithful artefacts in the question answering step, which affect the label prediction.

Relevance Analysis The relevancy of a question, as defined in section 4 is decided by manual evaluation by a single individual. Automated methods like cosine similarity did not yield meaningful results. For example, the cosine similarity between the questions *"What is the purpose of the United Nations' supposed plan to implant everyone with a biometric ID?"* and *"What is the aim for the United Nation precisely?"* is 0.83, which is very high for two questions are not similar at all. Due to the limitations of evaluation metrics like cosine similarity, fine grained evaluation of LLM outputs is a very challenging task, and manual evaluation is often unfeasible.

Justification Analysis In addition to veracity prediction, our fact-checking pipeline also generates justifications for the predictions. In the qualitative analysis, we examined some of these justifications to understand why some claims were misclassified. However, we do not systematically evaluate the quality of the justifications. We conjecture that the question-generation phase not only affects veracity prediction but also the justifications produced. We defer the analysis of this conjecture to future work.

6.2. Ethics Statement

AI is being used in nefarious ways to drive political opinion. Automated fact checking has the potential to be very useful for political fact-checking. We have demonstrated the ability of LLMs to ask nuanced questions about a claim, but the overall accuracy is still low. The tendency of LLMs to label claims as True is particularly alarming. We have highlighted examples where LLMs do not appropriately consider falsifying evidence that should have led to a False prediction. In some cases, the LLM even hallucinates answers for questions that cannot be answered from the knowledge base provided. These insights suggest that LLMs should be used cautiously in practical fact-checking applications.

Acknowledgements

This research was partially funded by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022).

References

[1] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NeuRIPS '20. Vancouver, BC, Canada; 2020. p. 1877-901.

[2] Press O, Zhang M, Min S, Schmidt L, Smith N, Lewis M. Measuring and Narrowing the Compositionality Gap in Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 5687-711. Available from: https://aclanthology.org/2023.findings-emnlp.378/.

[3] Zhang X, Gao W. 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). Nusa Dua, Bali: Association for Computational Linguistics; 2023. p. 996-1011. Available from: https://aclanthology.org/2023.ijcnlp-main.64/.

[4] Althabiti S, Alsalka MA, Atwell E. Generative AI for Explainable Automated Fact Checking on the FactEx: A New Benchmark Dataset. In: Disinformation in Open Online Media: 5th Multidisciplinary International Symposium, MISDOOM 2023, Amsterdam, The Netherlands, November 21–22, 2023, Proceedings. Berlin, Heidelberg: Springer-Verlag; 2023. p. 1–13. Available from: https://doi.org/10.1007/978-3-031-47896-3_1.

[5] Guo Z, Schlichtkrull M, Vlachos A. A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics. 2022;10:178-206. Available from: https://aclanthology.org/2022.tacl-1.11.

[6] Nakov P, Corney D, Hasanain M, Alam F, Elsayed T, Barrón-Cedeño A, et al. Automated Fact-Checking for Assisting Human Fact-Checkers. IJCAI International Joint Conference on Artificial Intelligence. 2021:4551-8.

[7] Dell'Anna D, Murukannaiah PK, Dudzik B, Grossi D, Jonker CM, Oertel C, et al. Toward a Quality Model for Hybrid Intelligence Teams. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. Auckland; 2024. p. 434-43.

[8] Pan L, Wu X, Lu X, Luu AT, Wang WY, Kan MY, et al. 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). Toronto, Canada: Association for Computational Linguistics; 2023. p. 6981-7004. Available from: https://aclanthology.org/2023.acl-long.386.

[9] Wang H, Shu K. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 6288-304. Available from: https://aclanthology.org/2023.findings-emnlp.416.

[10] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. 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). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 809-19. Available from: https://aclanthology.org/N18-1074/.

[11] Jiang Y, Bordia S, Zhong Z, Dognin C, Singh M, Bansal M. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 3441-60. Available from: https://aclanthology.org/2020.findings-emnlp.309/.

[12] Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 7534-50. Available from: https://aclanthology.org/2020.emnlp-main.609.

[13] Wang WY. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 422-6. Available from: https://aclanthology.org/P17-2067/.

[14] Alhindi T, Petridis S, Muresan S. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER). Brussels, Belgium: Association for Computational Linguistics; 2018. p. 85-90. Available from: https://aclanthology.org/W18-5513.

[15] Augenstein I, Lioma C, Wang D, Chaves Lima L, Hansen C, Hansen C, et al. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 4685-97. Available from: https://aclanthology.org/D19-1475/.

[16] Ostrowski W, Arora A, Atanasova P, Augenstein I. Multi-Hop Fact Checking of Political Claims. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization; 2021. p. 3892-8. Main Track. Available from: https://doi.org/10.24963/ijcai.2021/536.

[17] Chen J, Sriram A, Choi E, Durrett G. Generating Literal and Implied Subquestions to Fact-check Complex Claims. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 3495-516. Available from: https://aclanthology.org/2022.emnlp-main.229/.

[18] Fan A, Piktus A, Petroni F, Wenzek G, Saeidi M, Vlachos A, et al. Generating Fact Checking Briefs. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 7147-61. Available from: https://aclanthology.org/2020.emnlp-main.580/.

[19] Schlichtkrull M, Guo Z, Vlachos A. AVERITEC: A dataset for real-world claim verification with evidence from the web. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NeuRIPS '23. New Orleans, LA, USA; 2023. p. 65128-67.

[20] Barrón-Cedeño A, Alam F, Caselli T, Da San Martino G, Elsayed T, Galassi A, et al. The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2023;13982 LNCS:506-17. Available from: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_59.

[21] Pennycook G, Rand DG. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences. 2019;116(7):2521-6. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1806781116.

[22] Mendes E, Chen Y, Xu W, Ritter A. Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 15817-35. Available from: https://aclanthology.org/2023.acl-long.881/.

[23] Liscio E, Siebert LC, Jonker CM, Murukannaiah PK. Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. Journal of Artificial Intelligence Research. 2025 Feb;82:1-32. Available from: https://doi.org/10.1613/jair.1.14958.

[24] van der Meer M, Liscio E, Jonker C, Plaat A, Vossen P, Murukannaiah P. A Hybrid Intelligence Method for Argument Mining. Journal of Artificial Intelligence Research. 2024 Sep;80. Available from: https://doi.org/10.1613/jair.1.15135.

[25] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 2023 Mar;55(12). Available from: https://doi.org/10.1145/3571730.

[26] van der Meer M, Vossen P, Jonker CM, Murukannaiah PK. An Empirical Analysis of Diversity in Argument Summarization. In: Graham Y, Purver M, editors. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). St. Julian's, Malta: Association for Computational Linguistics; 2024. p. 2028-45. Available from: https://aclanthology.org/2024.eacl-long.123/.

[27] van der Meer M, Falk N, Murukannaiah PK, Liscio E. Annotator-Centric Active Learning for Subjective NLP Tasks. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 18537-55. Available from: https://aclanthology.org/2024.emnlp-main.1031/.

[28] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NeuRIPS '22. New Orleans, LA, USA; 2022. p. 24824-37.

[29] Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NeuRIPS '23. New Orleans, LA, USA; 2023. p. 74952-65.