Faithful Reasoning over Scientific Claims

Neşet Özkan Tan[1], Niket Tandon[2], David Wadden[2], Oyvind Tafjord[2], Mark Gahegan[1], Michael Witbrock[1]

[1]University of Auckland
[2]Allen Institute for Artificial Intelligence
neset.tan@auckland.ac.nz

Abstract

Claim verification in scientific domains requires models that faithfully incorporate relevant knowledge from the ever-growing, vast existing literature. Unfaithful claim verifications can lead to misinformation such as those observed during the COVID-19 pandemic. Fact-checking systems often fail to capture the complex relationship between claims and evidence, especially with ambiguous claims and implicit assumptions. Relying only on current LLMs poses challenges due to hallucinations and information traceability issues. To address these challenges, our approach considers multiple viewpoints onto the scientific literature, enabling the assessment of contradictory arguments and implicit assumptions. Our proposed inference method adds faithful reasoning to large language models by distilling information from diverse, relevant scientific abstracts. This method provides a verdict label that can be weighted by the reputation of the scientific articles and an explanation that can be traced back to sources. Our findings demonstrate that humans not only perceive our explanation to be significantly superior to the off-the-shelf model, but they also evaluate it as faithfully enabling the tracing of evidence back to its original sources.

Introduction

The vast amount of published research (including papers, studies, and data available) poses a challenge for experts and the public in staying abreast of the latest research advancements (Knoth et al. 2023). In particular, differentiating trustworthy sources from lesser sources, interpreting evidence by considering the context and assumptions across multiple documents, and synthesizing these assumptions to arrive at a conclusion are difficult tasks even for expert humans. This difficulty can create opportunities for distortion within the scientific field such as experienced with COVID-19 (Loomba et al. 2021). In order to ensure reliable verification of claims, scientists and the public alike must delve into the contextual nuances of claims and uncover specific assumptions, taking into account the latest advances in their respective fields. Consequently, ensuring reliable decision-making in the realm of science becomes a formidable task, creating an opportunity for effective methods for claim verification that can alleviate the burden on experts and facilitate accurate and timely analysis of scientific claims against the relevant literature.

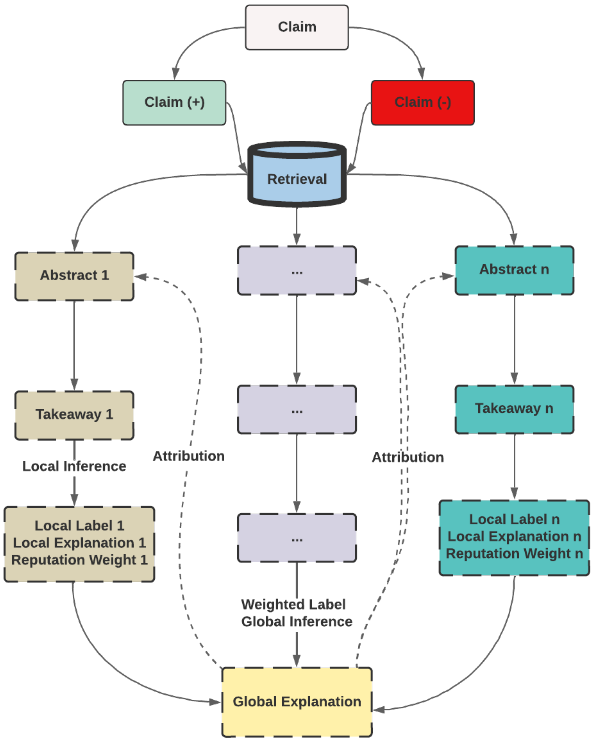

Figure 1: Through multiple stages, our framework enables faithful reasoning over scientific claims by providing a traceable provenance from global explanations to evidence abstracts.

This increasing need for trustworthy and accurate claim verification motivates the development of new methods that can enhance the contextual-understanding abilities of existing tools, such as Large Language Models (LLMs), by incorporating relevant knowledge from the existing literature. Relying solely on LLMs, even those with a substantial number of parameters and capabilities like GPT, presents a challenge due to the hallucination problem (Alkaissi and McFarlane 2023), which hampers the ability to trace the origins of information (Weidinger et al. 2021), (Ray 2023). This requirement is especially critical in scientific domains such as evidence-based medicine, where making decisions about patient care and outcomes relies on the utilization of up-to-date knowledge (Zakka et al. 2023; Alkaissi and McFarlane 2023).

As articulated by various researchers (e.g. (Sarrouti et al. 2021), (Wadden et al. 2020a)), scientific claim verification is aimed at facilitating the determination of the truth of scientific statements through an examination of a corpus of scholarly literature. The number of collections of documents that current fact-checking systems verify claims against is usually small (Tan et al. 2023a), and the limited label sets (such as "support" and "refute") may not be sufficient for determining the relationship between a claim and its evidence. This is particularly true for claims that contain ambiguous elements, such as "Incidence rates of cervical cancer have increased over time." which is taken from the SciFact dataset (Wadden et al. 2020a). In fact, this rate is increasing in some countries and while decreasing in others. In this case, it is unclear which population, time frame, or type of cervical cancer the claim is referring to, so the answer is uncertain. Labeling this claim as supported or refuted means ignoring all of these implicit assumptions, which can be critical for the assessment of the claim since these implicit assumptions can change the verdict of the claim. In other words, two aspects missing from prior work on evidence synthesis, including tasks like fact-checking, are: (1) does not explicitly examine assumptions, and (2) does not attempt to synthesise a global explanation as a compact and rich explanation of a claim with respect to the multiple perspectives in the literature.

Within the medical domain, there are numerous examples that require a careful analysis of the nuances in claims by experts. For example, there are over 50 nutrients that have been analysed in the literature for their carcinogenic properties, and while some studies may suggest that certain nutrients are carcinogens, others indicate that they can actually aid in the prevention of specific types of cancer (Schoenfeld and Ioannidis 2013).

To effectively address the complexities inherent in such cases, it is crucial to develop accessible, transparent, and explainable methods that incorporate external knowledge. Taking into account the metadata associated with publications, such as citation numbers and journal impact factors, can introduce biases based on consensus-driven publications with higher impact. By leveraging this information, we can mitigate the issue of hallucination and enhance the capabilities of existing language models, thereby enabling us to make more informed and accurate decisions biased by reliable sources.

The primary objective of this paper is to introduce a novel method that facilitates the clarification of scientific claims by incorporating multiple viewpoints from published scientific literature. This method involves reasoning over these diverse viewpoints while considering their metadata, including impact factor and citation count. Instead of merely classifying claims as true or false based on a closed-domain setting, our approach allows for the inclusion of "contradictory" arguments from scientific papers that stem from different assumptions and perspectives. Our method provides an explanation of the input statement that reflects reasoning based on multiple viewpoints derived from many scientific articles. By considering different perspectives, guided by the literature, we contribute to a more comprehensive understanding of complex topics. In summary, the key contributions of our work can be summarized as follows:

- We demonstrate the capability of systematically grounding LLMs with retrieved abstracts to uncover the underlying assumptions associated with a scientific claim.
- We introduce an inference method that effectively accounts for the underlying assumptions in papers. This method incorporates: (i) Global explanations that encompass reasoning elements from various viewpoints, guided by scientific articles. Each sentence in the explanation can be traced back to its source in the original abstract. (ii) Global inference labels, which combine language model inference with metadata measures such as the impact factor of the journal and citation count of the article.

Framework

The framework encompasses a cohesive integration of multiple essential components, as depicted in Figure 1 and more detailed in Figure 2. In this section, we provide a detailed description of the framework and associated workflow. While our methodology remains adaptable to any retrieval component and language model, our current implementation incorporates a combination of Google Scholar [1] and Semantic Scholar[2] APIs to retrieve relevant document candidates. For the model, we used the gpt-3.5-turbo[3] language model for generation. For all the generation and inference sub-tasks, we have employed instruction prompt approaches. In the remaining part of this section, we will present descriptions of these sub-tasks.

Claim Opposition and Retrieval

The primary objective of the first step is to enhance the retrieval process by incorporating more comprehensive sources. (Step 1 and Step 2 in Figure 2). This is achieved by considering arguments for and against the claim. A crucial component in this process is the abstract retriever, which plays a pivotal role in retrieving relevant information. It takes into account a given statement and its opposite to retrieve multiple scientific article abstracts. The quantity of abstracts can be specified using a user-specified parameter.

To illustrate the effectiveness of this approach, consider again the statement, "Incidence rates of cervical cancer have increased over time." Using an instruction learning approach, we can derive the opposite meaning as "Incidence rates of cervical cancer have decreased over time." By using both of these statements as queries in the retrieval system,

[1] https://serpapi.com/google-scholar-api

[2] https://www.semanticscholar.org/product/api

[3] https://platform.openai.com/docs/models/gpt-3-5

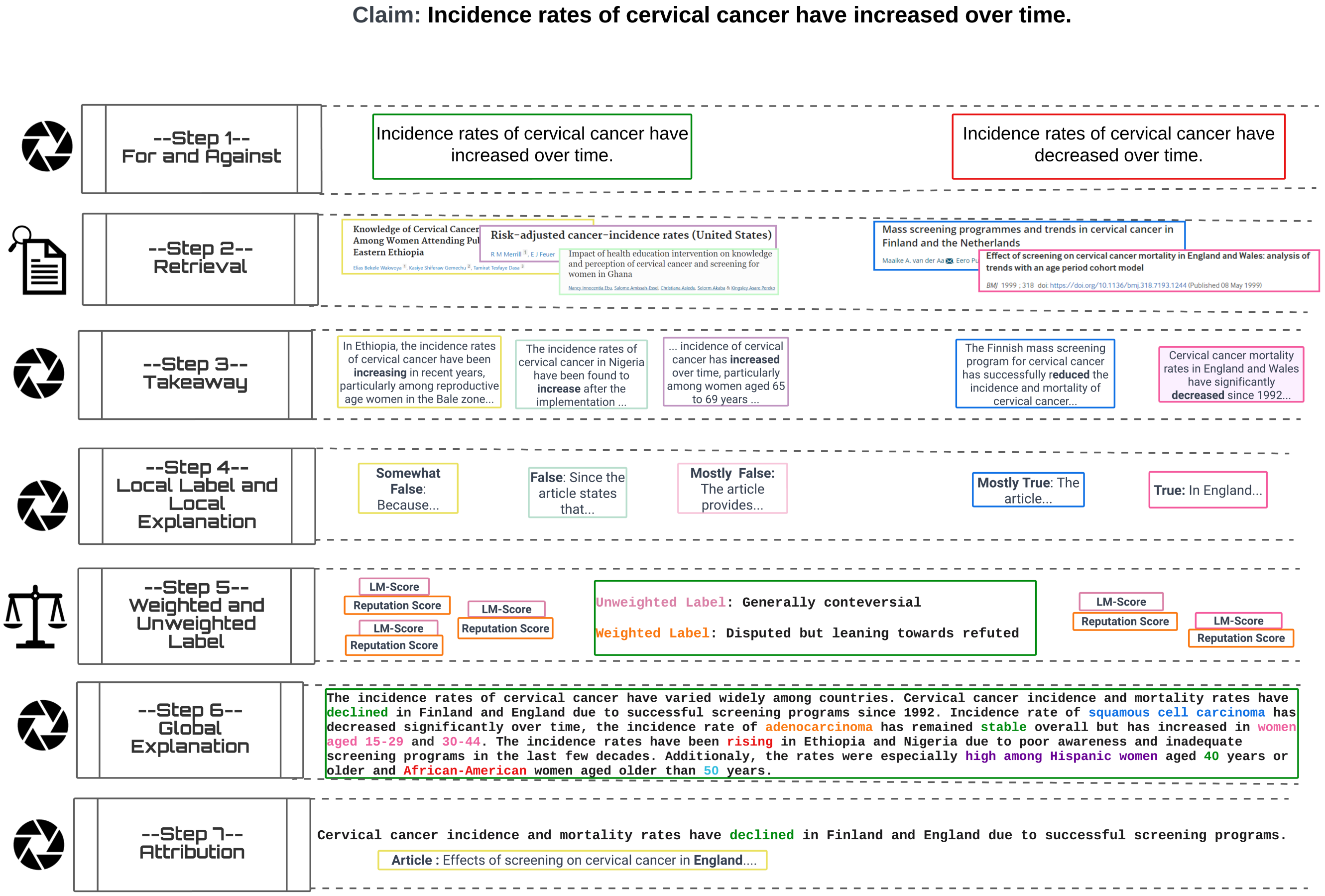
Claim: Incidence rates of cervical cancer have increased over time.
--Step 1-- For and Against
Incidence rates of cervical cancer have increased over time.
Incidence rates of cervical cancer have decreased over time.
--Step 2-- Retrieval
Knowledge of Cervical Cancer Among Women Attending Pul Eastern Ethiopia
Elias Bekele Wakwoya, Kasiye Shiferaw Gemechu, Tamirat Tesfaye Dasa
Risk-adjusted cancer-incidence rates (United States)
R M Merrill, E J Feuer
Impact of health education intervention on knowledge and perception of cervical cancer and screening for women in Ghana
Nancy Innocentia Ebu, Salome Amissah-Essel, Christiana Asiedu, Selorm Akaba & Kingsley Asare Pereko
Mass screening programmes and trends in cervical cancer in Finland and the Netherlands
Maaike A. van der Aa, Eero Pu
Effect of screening on cervical cancer mortality in England and Wales: analysis of trends with an age period cohort model
BMJ 1999 ; 318 doi: https://doi.org/10.1136/bmj.318.7193.1244 (Published 08 May 1999)
--Step 3-- Takeaway
In Ethiopia, the incidence rates of cervical cancer have been increasing in recent years, particularly among reproductive age women in the Bale zone...
The incidence rates of cervical cancer in Nigeria have been found to increase after the implementation ...
... incidence of cervical cancer has increased over time, particularly among women aged 65 to 69 years ...
The Finnish mass screening program for cervical cancer has successfully reduced the incidence and mortality of cervical cancer...
Cervical cancer mortality rates in England and Wales have significantly decreased since 1992...
--Step 4-- Local Label and Local Explanation
Somewhat False: Because...
False: Since the article states that...
Mostly False: The article provides...
Mostly True: The article...
True: In England...
--Step 5-- Weighted and Unweighted Label
LM-Score
Reputation Score
Unweighted Label: Generally conteversial
Weighted Label: Disputed but leaning towards refuted
--Step 6-- Global Explanation
The incidence rates of cervical cancer have varied widely among countries. Cervical cancer incidence and mortality rates have declined in Finland and England due to successful screening programs since 1992. Incidence rate of squamous cell carcinoma has decreased significantly over time, the incidence rate of adenocarcinoma has remained stable overall but has increased in women aged 15-29 and 30-44. The incidence rates have been rising in Ethiopia and Nigeria due to poor awareness and inadequate screening programs in the last few decades. Additionaly, the rates were especially high among Hispanic women aged 40 years or older and African-American women aged older than 50 years.
--Step 7-- Attribution
Cervical cancer incidence and mortality rates have declined in Finland and England due to successful screening programs.
Article : Effects of screening on cervical cancer in England....

Figure 2: Our framework has seven steps. In step 1 and 2, we retrieve relevant scientific abstracts for the claim and the anti-claim. In Step 3, we generate claim-specific takeaways from these retrieved abstracts, and prune abstracts with weak takeaways. Step 4 generates a local label and its explanation for each takeaway. In Step 5, local labels are aggregated and weighted based on the label score and article reputation. In Step 6, a detailed summary of the local explanations with respect to the given claim is generated, which we call the global explanation. Finally, in Step 7, we trace back each sentence in the global explanation to sentences in the abstracts to show its attribution.

we can effectively retrieve candidate abstracts that closely align with the original claim as well as its opposing viewpoint.

For the retrieval activity (Step 2 in Figure 2), we leverage the titles obtained from the Google Scholar search engine and match them with the Semantic Scholar database, facilitating the retrieval of abstracts from the corresponding articles. It is important to acknowledge that this retrieval component does have limitations, which are explained in the limitation section.

Takeaway Generation

Our objective in this task (Step 3 in the Figure 2) is to extract claim-specific information from closely related documents. The input consists of an abstract candidate retrieved from the previous step and the original claim. We aim to achieve two objectives. Firstly, we ensure the relevance of the abstract candidate to the claim. Secondly, if the abstract candidate is deemed relevant by the language model, we proceed to extract the main message of the abstract from the perspective of the claim.

For example, consider the statement “Incidence rates of cervical cancer have increased over time.” The retrieved article candidate may contain information about various types of cancer, not just cervical cancer. However, our focus is specifically on information related to cervical cancer while disregarding any other irrelevant details. We refer to this main message as the takeaway. Thus, the output of this task is a set of text summaries that contain the main message of each abstract in terms of the claim, only if the abstract is determined to be related to the original claim.

Local Explanation and Local Label

In this task (Step 4), our objective is to evaluate the claim by utilizing the corresponding takeaway. We achieve this by assigning a verdict label to the claim, known as a ‘local la-

Global Label	Description
Generally supported	Average score ≥ 0.66
Generally refuted	Average score ≤ -0.66
Disputed but leaning towards refuted	$-0.66 <$ Average score < -0.33
Disputed but leaning towards supported	$0.33 <$ Average score < 0.66
Generally controversial	Otherwise

Table 1: Global labels and their descriptions.

bel', based on the information presented in the takeaway. Furthermore, we aim to express the rationale behind the assigned label, which we refer to as the local explanation for the verdict.

To facilitate this process, we employ a 7-graded scientific credibility labeling space inspired by fact-checking organizations[4] [5]. The label space includes the following categories: True, Mostly True, Somewhat True, No Evidence, Somewhat False, Mostly False, and False. These labels help us evaluate the veracity of the claim based on the information extracted from the takeaway.

Unweighted and Weighted Label

In this step (Step 5), we assign scores to each local (verdict) label using the following simple scale: True (1.0), Mostly True (0.66), Somewhat True (0.33), No Evidence (0.0), Somewhat False (-0.33), Mostly False (-0.66), and False (-1.0). We refer to these as language model-assigned scores since the labels are determined by the language model itself during the labeling process.

Additionally, we calculate the reputation score by normalizing the average of three reputation metrics specific to each claim: the citation count, the journal impact factor, and the Scimago Journal Rank.

Using the local labels and reputation scores, we compute the weighted global label score. This score considers both the article's reputation and the language model's assessment of the claim's veracity, resulting in a balanced evaluation. Based on this weighted average, we assign global labels, which provide overall verdicts about a claim and are less granular compared to local labels (see Table 1). We also calculate an unweighted label, which only considers the average of the language model-assigned scores (local labels).

Global Explanation

The main output of our system is the global explanation (Step 6), which serves as a comprehensive summary of the local explanations, excluding cases where the local labels indicate "No Evidence." This process of generating the global explanation plays a crucial role in uncovering implicit assumptions within claims by assimilating information from a variety of perspectives found in different literature sources.

To illustrate this, consider the example of cervical cancer. When examining the claim related to cervical cancer, the global explanation provides a condensed yet informative overview that encompasses a range of factors. These factors may include changing populations, geographic variations in immunisation rates, and the specific types of cervical cancer that have evolved over time. While these components might not be explicitly evident in the initial claim itself, they are explicitly stated in the global explanation, offering a comprehensive understanding of the topic–see the example explanation shown in Step 5 of Figure 2.

By uncovering and incorporating these literature-supported perspectives, the global explanation sheds light on previously unexplored aspects of a claim, enhancing its instructive nature. Furthermore, the global explanation may contain critical information that directly impacts the assessment of a claim's veracity as the example in (Figure 2) shows. Through this comprehensive approach, our system aims to provide a holistic evaluation that goes beyond the surface-level interpretation of claims.

Attribution

In this step (Step 6), we undertake an analysis of the explanation, focusing primarily on identifying the relevant supporting sentences that can be traced back to the pipeline's original article abstracts. This process involves transparently tracing each sentence of the global explanation, reaching back to the precise sentences in the source, and thus making transparent the evidence used.

Our analysis in this step has two main objectives. Firstly, we aim to determine whether a sentence in the global explanation originates from hallucinations, false information, or an interpretation that is not directly supported by any abstract. Secondly, we seek to assess whether a sentence incorporates information obtained from multiple sources, indicating the possibility of composition or potential misrepresentation.

Evaluation

Evaluation Setup

We conducted evaluations using human participants and automated methods to assess the quality of our generated explanations and labels.

We compared our explanations with those generated by GPT-3.5 with and without grounding retrieval component (the absence and presence of retrieved relevant abstracts.) in a prompt tuning setting and employed human evaluators to determine which explanation offered a more informative and in-depth understanding of the subject. Evaluators assessed the level of detail and instructiveness provided by each explanation, with the goal of identifying the explanation that contributed to a deeper understanding of the claim.

[4] https://healthfeedback.org/process

[5] https://climatefeedback.org/process/

We also focused on evaluating pipeline components, specifically the attribution results of our explanations and the accuracy of the global labels. Regarding the attribution results, we aimed to assess whether the global explanation accurately reflected the information presented in the articles. In terms of global label accuracy, we sought to evaluate both weighted and unweighted labels, as well as the alignment between them based on the information provided in the global explanation.

Dataset

Dataset	**Claims**
Climate	Burning fossil fuels are the main cause of greenhouse gas emissions.
Climate	Climate change is causing more extreme weather events and sea level rise.
Covid	Covid vaccine leads to infertility.
Covid	Vaccines cause autism.
Myths	Drinking green tea burns fat.
Myths	Eating carrots can improve your eyesight.
Myths	Using a cell phone may cause brain cancer.
Nutrients	Ice cream may cause sickness.
Nutrients	Sausages may cause cancer.
SciFact	Obesity is determined in part by genetic factors.
SciFact	Bariatric surgery has a deleterious impact on mental health.

Table 2: Examples of Claims from Climate, Covid, Myths, Nutrients, and SciFact Datasets

Existing datasets in the fact-checking literature primarily focus on measuring verdict labels for claims, typically with binary options of “supports” or “refutes.” However, we aim to assess the capabilities of our method using a more graduated set of claims (existing datasets typically exist as synthetically generated claims). To achieve this, we created a combination of diverse topics, including technical claims from the biomedical domain, climate change-related claims, COVID-19-related claims, nutrition-related claims, and general myth-related claims.

In the biomedical domain, we utilized claims from the SciFact dataset (Wadden et al. 2020a), which consists of claims for which there are both supporting and refuting arguments in the scientific literature. To gather nutrition-related claims, we followed the nutrition guidelines outlined in (Schoenfeld and Ioannidis 2013), which involved conducting a comprehensive survey of the scientific literature on nutrition and its relationship to cancer. For climate change and COVID-19, we sourced data from various fact-checking websites and forums. We also found that including the gpt-3.5-turbo generated claims allowed us to incorporate real-world queries and statements related to general myths, and we included those claims as well.

Due to the high costs and time-consuming nature of annotating scientific claims, we evaluated a total of 50 scientific claims (please refer to Table 2 for some examples), ensuring equal representation across the mentioned topics. On average, we retrieved 5.8 articles per claim, and the evaluators annotated a total of 291 scientific articles relevant to these 50 claims[6]. Although the number of retrieved articles is an adjustable parameter, we maintained a moderate value (5.8 average per claim) to keep the annotation process feasible. To automate all the steps outlined in the methodology, we developed a demo that generates evaluation samples. Each claim in the dataset is evaluated by two evaluators, and the results presented here are based on their agreed evaluation (see details in Table).

Results

Despite the fact that GPT-3.5 explanations do not provide links to consistent, reliable sources, we evaluate how humans find them in terms of deep understanding of the subject compared to our method. Figure 3 presents the results of this question. For the attribution measurement, we followed the approach of (Bohnet et al. 2022), which was originally defined in (Rashkin et al. 2021). This approach allowed us to assess the extent to which the explanation provided by our system can be attributed to the source document in the context of a given claim. We asked evaluators the following two questions:

1. “Is all of the information in the global explanation interpretable to you?”
2. “Do the sources support the target sentences in global explanation?”

Datasets	Explanation		Attribution	
	Ours	GPT-3.5	Ours	GPT-3.5
Climate	**80%**	20%	80%	-NA-
Covid	**70%**	30%	70%	-NA-
Myths	**80%**	20%	90%	-NA-
Nutrition	**90%**	10%	80%	-NA-
Sci-fact	**70%**	30%	80%	-NA-

Table 3: Across different domains, our global explanation is preferred over GPT-3.5’s. Human evaluators also find 80% of our global explanations correctly attributed to the abstracts. X% means two annotators agreed on the explanation and attribution question options for x out of 10 claims. For example, out of 10 Covid claims, two annotators agreed on the explanation options for 7 claims.

A sample is considered ’attributable’ if both of these questions are answered affirmatively. As in (Rashkin et al. 2021), for the claims $c_1, c_2, ..., c_n$, and our system g, we defined the value of the function $h(c_i, g(c_i), r_i)$ to be 1 if the answer to both questions above is affirmative for the evaluator r_i, or 0 otherwise. The test accuracy is then defined as:

$$\frac{1}{n}\sum_{i=1}^{n} h(c_i, g(c_i), r_i).$$

[6] All 50 claims and a demo sample for evaluation can be seen at https://taneset.github.io/frosc.github.io/

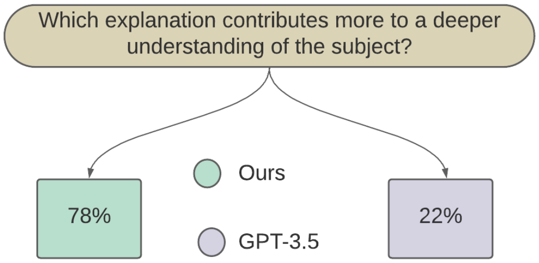

Figure 3: Human judges find our global explanation richer than an off-the-shelf LLM generated explanation.

We refer to this number as the attribution accuracy. We obtained 80 percent of the test samples were found attributable. Both explanation preferences and attribution results per dataset are represented in Table 3. The comparison figures between the accuracy of weighted and unweighted labels reveal that the weighted labels exhibit a 20% superiority over the unweighted labels. To see the results for each domain, refer to Figure 4. (in response to the question: "According to the global explanation, which label accurately reflects the verdict of the given claim? Unweighted, Weighted, Either, or Neither").

Datasets	Global Label	
	Weighted	Unweighted
Climate	**90%**	60%
Covid	**90%**	80%
Myths	**90%**	60%
Nutrition	**70%**	60%
Sci-fact	**80%**	60%
Average	**84%**	64%

Table 4: Aggregating local labels based on their reputation score makes the resulting global label more accurate. X% means two annotators agreed on the global label option for x out of 10 claims. For example, out of 10 nutrition claims, 7 out of 10 weighted labels were agreed to be correct.

Additionally, we conducted an experiment comparing the global explanations and labels generated by GPT-3.5 for all retrieved raw documents. We focused on samples where annotators had designated their weighted labels as 'true' based on the global explanation. In this context, we found an exact match accuracy of 0.6 for labels and an average BLEU Score of 0.12 for explanations. These findings indicate that 40% of the labels assigned by GPT-3.5 based on raw documents do not align with labels validated as accurate by humans.

We also conduct experiments to understand the contribution of each pipeline component. We measure how retrieval results change with and without claim opposition. We examined 50 claims and compared them against 10 articles each, both in favor of and against the claim. We found that, on average, 51.7 percent of the documents were additionally retrieved as a result of querying with opposing claims. We also analyzed these results per dataset, which can be seen in Table 6. For the remaining pipeline components, we present results in Table **??**, with each component's accuracy and its main purpose/impact with respect to the end task.

Error Analysis

In our error analysis, we focus on examples that are classified as non-attributable, either because the global explanation is not interpretable or the attribution is erroneous. We comprehensively analyze these samples and have identified three types of typical errors see Table (7): inference-based errors, retrieval-based errors, and parsing-based errors.

Retrieval Errors (60%)

- When the documentation of an article from Semantic Scholar is mistakenly inputted. In this case, a short article-related text, such as the title and keywords of the article or the introduction part of the article, is erroneously treated as an abstract.
- When weakly related articles are retrieved. For example, consider the query "Flour may cause cancer." Articles dominated by the relationship between corn and cancer may be included, as they were not filtered out due to the possibility that flour can be made with corn flour. However, the results may indicate ”Corn may cause cancer" instead of ”Flour may cause cancer."
- When the query is too specific, such as ”MEK inhibitors are not effective in RAS-driven mouse models of cancer." Although there are numerous relationships between MEK inhibitors and cancer, none of them specifically relate to RAS-driven mouse models.

Inference Errors (30%)

- Sentences in global explanations may provide reasons to attribute too many relationships. For example, the sentence ”The relationship between autophagy and aging is complex and depends on various factors" could lead to the retrieval of 20 attributions, as ”various factors" is a vague term. This ambiguity may cause human evaluators to consider this attribution as incorrect. In some cases, having too many attributions can result in the exact opposite direction, leading evaluators to believe that the target sentence is not adequately covered.
- The language model tends to soften attributions. For instance, consider the target sentence: ”Additionally, the study only shows an association, not a causal relationship, between masturbation and negative outcomes." The corresponding source sentence is: ”Masturbation is significantly associated with fatigue, soreness, and weakness in the lumbar region, memory decline, immunity decline, insomnia, dreaminess, and an increase in related symptoms that accompany an increase in masturbation frequency."

Parsing Errors (10%)

- When the sentence number or abstract number in the attribution part refers to the wrong one.

We also examined sentences that were labeled as interpretation or hallucination. We discovered that approximately 11% of the target sentences fell into this category. These

Component	Accuracy	Expected Impact on the Overall System
Takeaway	87%	Uncovers implicit assumptions, removes irrelevant documents and information.
Local explanation and label	85%, 81%	Answers claim (query) with explanation w.r.t. the local takeaway (in context of its assumptions).
Global Explanation	78%	Final answer explanation that combines multiple perspectives from local explanations in light of the claim (query) and leverages redundancy but reduces duplicates and accurately answers the claim globally
Verdict label	84%	Final answer label weighted by the reputation of the source abstract (this accounts for local label and reputation. If we drop reputation accuracy of verdict label drops to 64%)
Attribution	80%	For each sentence in global explanation, give a reference citation/ supporting sentence in source abstracts

Table 5: Accuracy of each individual component. Reported accuracy is computed by checking whether the component output was correct/ appropriate or not)

Datasets	% Different
Climate	43.7%
Covid	49.5%
Myths	57.3%
Nutrition	62%
Sci-fact	45.8%
Average	51.7%

Table 6: Contribution of opposing claim to the retrieval component.

types of sentences generally do not specifically address any argument, for example, sentences like ‘However, there is no evidence to support this claim.’ In such instances, our attribution method identifies them as interpretations or hallucinations.

Human Evaluation and Inter-Annotator Agreement

The human evaluation process for this study adheres to ethical guidelines and has been conducted at the University of Auckland. All participants in the study provided informed consent, and their privacy and rights have been carefully protected throughout the research process. The call for annotators was extended through a well-attended Slack channel, which comprises individuals with an interest in machine learning, including those with master’s, PhDs, and academic professionals. The recruitment process was voluntary, and participation was based on responses to the call. It’s worth highlighting that none of the evaluators has any hierarchical or personal affiliations with the paper’s co-authors, and none of them are co-authors of the paper themselves. For each claim, two individuals conduct the evaluation; the tables presented in the results section reflect mutually agreed-upon scores.

Retrieval	Inference	Parsing
60%	30%	10%

Table 7: Error Analysis

Related Work

Claim verification In recent years, there has been an increase in fact-checking research, largely driven by the data available on general fact-checking websites (Kotonya and Toni 2020a). However, the availability of such data across multiple scientific domains still remains a critical bottleneck. This bottleneck is particularly challenging for scientific claims due to the expensive annotations required from domain experts. There have been only limited attempts to create scientific fact-checking datasets (Wadden et al. 2020b; Wright et al. 2022; Tan et al. 2023a), so not surprisingly there remains a scarcity of available data.

Furthermore, existing models and datasets designed for less technical language and public releases have limitations when applied to scientific domains. Nevertheless, recent studies have started exploring specific domains. These include work on fact-checking COVID-19 claims on social media platforms (Roozenbeek et al. 2020; Saakyan, Chakrabarty, and Muresan 2021), health-related claims in science releases (Sarrouti et al. 2021), and climate change claims using evidence from Wikipedia (Diggelmann et al. 2020). Another common characteristic of existing models and datasets is that they are designed for closed-domain settings, with only two recent exceptions (Wadden et al. 2022a; Stammbach, Zhang, and Ash 2023). However, these works resume the three-label verdict tradition, without considering the nuances and contested aspects of the claims.

To enhance fact-checking capabilities, various techniques have been developed over the past decade. These include employing multi-layer perceptron models (Vlachos and Riedel 2014), incorporating attention mechanisms (Parikh et al. 2016), utilizing Graph Neural Networks (Liu et al. 2020), employing semantic role labeling and logical reasoning tools (Chen et al. 2020). Transformer-based language models, particularly BERT models, have gained significant attention in claim verification (Soleimani, Monz, and Worring 2019; Portelli et al. 2020; Chernyavskiy and Ilvovsky 2019; Nie, Chen, and Bansal 2019; Tokala et al. 2019; Tan et al. 2023b). Additionally, methods (Tan et al. 2023b) and models (Wadden et al. 2022b; Stammbach, Zhang, and Ash 2023) have been developed to increase the performance of fact-checking tasks for longer input lengths, which is usually the case for scientific fact-verification tasks. To the best of our knowledge, we are not aware of any zero-shot attempts for this task, except (Wright et al. 2022), which was applied as a demonstration of how artificially generated data can enhance the performance of the verdict prediction task.

Claim explanation In their work, (Atanasova et al. 2020) tackled the challenge of generating explanations by treating them as a text summarization task. They employed two models for this purpose: one model generated explanations (both extractive and abstractive) after the claim, where the prediction and explanation models were trained independently. Meanwhile, the second model was trained jointly to handle both tasks. In a similar vein, (Kotonya and Toni 2020b) focused on extractive summaries for the fact verification task. It is important to note that both of the aforementioned approaches require a fine-tuning and supervised training process, along with annotated data for each sub-task, such as summarization and verdict prediction.

In the field of document summarization, there is another notable line of work focused on summarizing multiple documents in the medical domain (DeYoung et al. 2021; Wallace et al. 2021) by treating the survey paper sections as summaries of papers that are discussed within the same paper. Additionally, in the biomedicine domain, researchers have explored zero-shot summarization techniques for both single and multi-document scenarios (Shaib et al. 2023). However, none of the existing summarization approaches synthesize information to support a specific claim or provide a verdict label. it is also worth noting that although there are two commercial products, elicit[7] and consensus[8], which seem to retrieve relevant articles based on scientific questions, their development processes are not transparent, and limited information is available regarding their functioning. These products also lack an overall explanation regarding a given scientific claim and neither provide a verdict to evaluate the claim accurately as of the date of our submission.

Conclusion and Outlook

Existing claim verification methods operate in a closed-domain context, and need supervised training data; however scientific knowledge constantly evolves with new perspectives that necessitate a supervision-free, open-domain perspective that we address in this work. We present an inference method that aggregates and reasons over multiple evidence abstracts using zero-shot instruction prompt to generate explanations that are traceable and thus faithful. Human evaluators find that our claim verification labels are of high quality and the corresponding explanations are traceable and preferred over an LLM explanation. As future work, we want to improve the components of our pipeline such as retrieval, and enforce consistency in the resulting graph of takeaways, explanations, abstracts and attributions. Thus, our framework creates new avenues for claim reasoning research and enables time-saving, accurate analysis of scientific claims.

Limitations

Enhancing retrieval systems with external knowledge beyond lexical similarities often requires extensive training or fine-tuning (Karpukhin et al. 2020). In the retrieval process, we initially gather titles from Google Scholar API and then match these titles using the Semantic Scholar API. However, we frequently encounter the issue of mismatched titles. Furthermore, there are instances when API calls to Semantic Scholar return empty results or provide unrelated article sections instead of abstracts. These problems adversely affect the quality and quantity of the search space. Further, incomplete, outdated, or inaccurate information from the search engine can pose significant problems, potentially leading to erroneous conclusions or decisions.

Ethics Statement

This research paper emphasizes the significance of automated scientific fact-checking in combating the dissemination of scientific misinformation within the community. Our primary objective is to enhance the effectiveness and practicality of scientific fact-checking systems. A key aspect of our approach is transparency, emphasizing the importance of providing explanations and traceability while utilizing language models in our reasoning process. We rely on scientific articles, and there are factual and ethical risks often against a particular community when relying on less trustworthy articles. We mitigate this risk through a weighted scoring based on reputation scores, and we will continue to improve this weighting. Further, the search engine results can be biased in retrieving documents aligned or misaligned to a certain community's beliefs. For this, we recognize the need for more sophisticated retrieval methods as highlighted in the limitations section. While current LMs have limited guardrails and thus inherently carry risks of incorrect information, our paper uses scientific literature-guided explanations to address this issue.

References

Alkaissi, H.; and McFarlane, S. I. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2): e35179.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. Online: Association for Computational Linguistics.

[7] https://elicit.org

[8] https://consensus.app

Bohnet, B.; Tran, V. Q.; Verga, P.; Aharoni, R.; Andor, D.; Soares, L. B.; Ciaramita, M.; Eisenstein, J.; Ganchev, K.; Herzig, J.; Hui, K.; Kwiatkowski, T.; Ma, J.; Ni, J.; Saralegui, L. S.; Schuster, T.; Cohen, W. W.; Collins, M.; Das, D.; Metzler, D.; Petrov, S.; and Webster, K. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models.

Chen, J.; Bao, Q.; Sun, C.; Zhang, X.; Chen, J.; Zhou, H.; Xiao, Y.; and Li, L. 2020. LOREN: Logic-Regularized Reasoning for Interpretable Fact Verification.

Chernyavskiy, A.; and Ilvovsky, D. 2019. Extract and Aggregate: A Novel Domain-Independent Approach to Factual Data Verification. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 69–78. Hong Kong, China: Association for Computational Linguistics.

DeYoung, J.; Beltagy, I.; van Zuylen, M.; Kuehl, B.; and Wang, L. L. 2021. MSˆ2: Multi-Document Summarization of Medical Studies. In *EMNLP*.

Diggelmann, T.; Boyd-Graber, J.; Bulian, J.; Ciaramita, M.; and Leippold, M. 2020. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-T. 2020. Dense Passage Retrieval for Open-Domain Question Answering.

Knoth, P.; Herrmannova, D.; Cancellieri, M.; Anastasiou, L.; Pontika, N.; Pearce, S.; Gyawali, B.; and Pride, D. 2023. CORE: A Global Aggregation Service for Open Access Papers. *Sci Data*, 10(1): 366.

Kotonya, N.; and Toni, F. 2020a. Explainable Automated Fact-Checking: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5430–5443. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Kotonya, N.; and Toni, F. 2020b. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740–7754. Online: Association for Computational Linguistics.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7342–7351. Online: Association for Computational Linguistics.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav*, 5(3): 337–348.

Nie, Y.; Chen, H.; and Bansal, M. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *AAAI*, 33(01): 6859–6866.

Parikh, A.; Täckström, O.; Das, D.; and Uszkoreit, J. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2249–2255. Austin, Texas: Association for Computational Linguistics.

Portelli, B.; Zhao, J.; Schuster, T.; Serra, G.; and Santus, E. 2020. Distilling the Evidence to Augment Fact Verification Models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, 47–51. Online: Association for Computational Linguistics.

Rashkin, H.; Nikolaev, V.; Lamm, M.; Aroyo, L.; Collins, M.; Das, D.; Petrov, S.; Tomar, G. S.; Turc, I.; and Reitter, D. 2021. Measuring Attribution in Natural Language Generation Models.

Ray, P. P. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3: 121–154.

Roozenbeek, J.; Schneider, C. R.; Dryhurst, S.; Kerr, J.; Freeman, A. L.; Recchia, G.; Van Der Bles, A. M.; and Van Der Linden, S. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science*, 7(10): 201199.

Saakyan, A.; Chakrabarty, T.; and Muresan, S. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic.

Sarrouti, M.; Ben Abacha, A.; Mrabet, Y.; and Demner-Fushman, D. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3499–3512. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Schoenfeld, J. D.; and Ioannidis, J. P. A. 2013. Is everything we eat associated with cancer? A systematic cookbook review. *Am. J. Clin. Nutr.*, 97(1): 127–134.

Shaib, C.; Li, M.; Joseph, S.; Marshall, I. J.; Li, J. J.; and Wallace, B. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). *ArXiv*, abs/2305.06299.

Soleimani, A.; Monz, C.; and Worring, M. 2019. BERT for Evidence Retrieval and Claim Verification.

Stammbach, D.; Zhang, B.; and Ash, E. 2023. The Choice of Textual Knowledge Base in Automated Claim Checking. *ACM Journal of Data and Information Quality*, 15: 1 – 22.

Tan, N.; Nguyen, T.; Bensemann, J.; Peng, A.; Bao, Q.; Chen, Y.; Gahegan, M.; and Witbrock, M. 2023a. Multi2Claim: Generating Scientific Claims from Multi-Choice Questions for Scientific Fact-Checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2652–2664. Dubrovnik, Croatia: Association for Computational Linguistics.

Tan, N. Ö.; Peng, A. Y.; Bensemann, J.; Bao, Q.; Hartill, T.; Gahegan, M.; and Witbrock, M. 2023b. Input-length-shortening and text generation via attention values.

Tokala, S.; G, V.; Saha, A.; and Ganguly, N. 2019. AttentiveChecker: A Bi-Directional Attention Flow Mechanism for Fact Verification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2218–2222. Minneapolis, Minnesota: Association for Computational Linguistics.

Vlachos, A.; and Riedel, S. 2014. Fact Checking: Task definition and dataset construction. 18–22.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020a. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550. Online.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020b. Fact or Fiction: Verifying Scientific Claims.

Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Beltagy, I.; Wang, L. L.; and Hajishirzi, H. 2022a. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4719–4734. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wadden, D.; Lo, K.; Wang, L. L.; Cohan, A.; Beltagy, I.; and Hajishirzi, H. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 61–76. Seattle, United States: Association for Computational Linguistics.

Wallace, B. C.; Saha, S.; Soboczenski, F.; and Marshall, I. J. 2021. Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. In *AMIA Summits on Translational Science Proceedings*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and social risks of harm from Language Models.

Wright, D.; Wadden, D.; Lo, K.; Kuehl, B.; Cohan, A.; Augenstein, I.; and Wang, L. L. 2022. Generating Scientific Claims for Zero-Shot Scientific Fact Checking.

Zakka, C.; Chaurasia, A.; Shad, R.; Dalal, A. R.; Kim, J. L.; Moor, M.; Alexander, K.; Ashley, E.; Boyd, J.; Boyd, K.; Hirsch, K.; Langlotz, C.; Nelson, J.; and Hiesinger, W. 2023. Almanac: Retrieval-Augmented Language Models for Clinical Medicine.