SciTrue: Evidence-Grounded Claim Verification in Science

Neşet Özkan TAN
Minghao Li
Mark Gahegan
Department of Computer Science
University of Auckland
Auckland, New Zealand
neset.tan@auckland.ac.nz, minghao.lee2017@outlook.com,
m.gahegan@auckland.ac.nz

Abstract

Large language models (LLMs) have expanded the potential for AI-assisted scientific claim verification, yet existing systems often exhibit unverifiable attributions, shallow evidence mapping, and hallucinated citations. We present SciTrue, a claim verification system providing source-level accountability and evidence traceability. SciTrue links each claim component to explicit, verifiable scientific sources, enabling users to inspect and challenge model inferences, addressing limitations of both general-purpose and search-augmented LLMs. In a human evaluation of 300 attributions, SciTrue achieves high fidelity in summary traceability, attribution accuracy, and context alignment, substantially outperforming RAG-based baselines such as GPT-4o-search-preview and Perplexity Sonar Pro. These results underscore the importance of principled attribution and context-aware reasoning in AI-assisted scientific verification. A system demo is available at www.scitrue.org.

1 Introduction

The increasing adoption of large language models (LLMs) such as GPT-4 (OpenAI, 2023), Gemini 2.5 (Google DeepMind, 2024), and Llama-3 (Meta, 2025) has transformed the landscape of information access, reading comprehension, and scientific literature summarization. Despite their impressive capabilities, LLMs exhibit a significant tendency to generate content that is inconsistent with established knowledge or the provided input context, a phenomenon widely referred to as hallucination (Mittelstadt et al., 2023). This issue manifests in various ways within the scientific domain, including the fabrication of scientific references and the generation of seemingly accurate citations that, upon closer inspection, do not support the claims being made (Tilwani et al., 2024). This is particularly problematic in high-stakes domains such as biomedicine or policy, where unverified or misattributed evidence can propagate misinformation or erode trust (Chen et al., 2025).

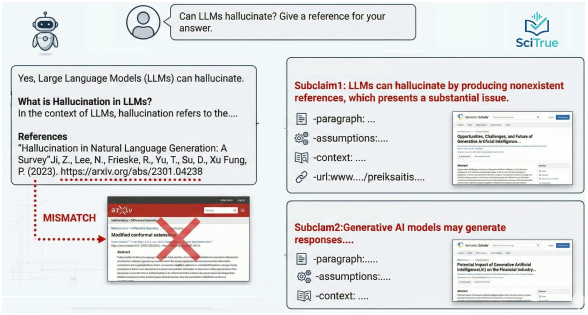

Figure 1: Scientific claim verification with web-access LLMs versus SciTrue on real-world examples. Web-access LLMs often respond with overconfident answers and cite incorrect references (e.g., linking to a differential geometry paper for a claim about LLMs). In contrast, SciTrue provides end-to-end scientific traceability for each attribution. To view the full hallucination example, see: https://chatgpt.com/share/6867bbc2-73e0-8002-9bda-4eb1223041b0

Conventional "research assistant" models, such as OpenAI's Deep Research[1] and Gemini Deep Research[2], integrate access to the web and scholarly databases to enhance factual grounding. Nonetheless, use cases reveal persistent deficiencies. These systems often take more than five minutes to process a single claim, ranging from 5 to 30 minutes according to the official website[1], when reviewing multiple documents, making them impractical for time-sensitive applications such as medicine or journalism. Their outputs tend to be excessively long, which hinders the extraction of concise, actionable insights. Attribution is often shallow: although paragraphs are linked to sources, the synthesis does not clearly indicate which information comes from which source. Moreover, many of these systems follow rigid, predefined formats, such as systematic reviews, that limit their

[1] https://openai.com/index/introducing-deep-research/
[2] https://gemini.google/overview/deep-research/?hl=en

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 397–406
March 24-29, 2026 ©2026 Association for Computational Linguistics

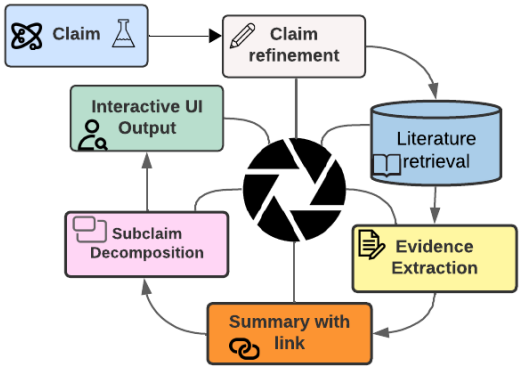

Figure 2: Workflow of the SciTrue agentic pipeline. A user submits a scientific claim, which is then refined. Relevant articles are retrieved, and supporting evidence is extracted. The system generates a summary, decomposes it into subclaims, and presents the results in an interactive interface.

adaptability to diverse, real-world claim verification tasks.

In science, transparency and accountability are foundational principles (Binz et al., 2025). Accurately interpreting and verifying scientific claims necessitates a thorough understanding of the contextual factors and underlying assumptions that frame these claims (Kanavouras and Coutelieris, 2020). Scientific findings are inherently contingent upon specific experimental conditions, population characteristics, and methodological choices (Bottesini et al., 2023). Neglecting such context can lead to overgeneralization or misapplication of results, thereby compromising the validity and utility of the claim. Failure to properly attribute claims can lead to serious issues, including plagiarism, the spread of misinformation, and an erosion of user trust.

To address these challenges, we present **SciTrue**, a scientific claim verification system specifically designed for domains where auditability is critical, such as scientific research. SciTrue provides end-to-end traceability by explicitly attributing verifiable sources, along with the associated context, assumptions, and credibility. Human evaluations demonstrate that SciTrue's attributions significantly outperform those of leading systems in terms of summary traceability, attribution accuracy, contextual and assumption alignment, and scholarly credibility.

[1] https://openai.com/index/introducing-deep-research/

2 Related Work

The automation of claim verification has received increasing attention due to the rise of misinformation and the need for auditability in scholarly communication (Wadden et al., 2020; Augenstein et al., 2019; Thorne et al., 2018). Benchmarks such as SciFact (Wadden et al., 2020), MultiFC (Augenstein et al., 2019), and FEVER (Thorne et al., 2018) have advanced the development of systems for evidence retrieval and claim stance classification, typically pairing passage retrieval with textual entailment models. However, most approaches focus on closed-domain settings and do not provide sufficient granularity, such as context or assumptions, in associating specific components of a claim with the underlying supporting text.

In the domain of document summarization, a significant line of research has explored multi-document summarization in the medical field (DeYoung et al., 2021; Wallace et al., 2021). More recently, the biomedical community has investigated zero-shot summarization methods for both single- and multi-document settings (Shaib et al., 2023). However, existing techniques generally lack fine-grained provenance for subclaims, such as their source documents, surrounding context, and reasoning trail and often operate at the abstract level (Tan et al., 2024). While commercial tools like (Elicit) and (Consensus) appear to retrieve relevant literature in response to scientific queries, their development processes are opaque. As of 2025, they typically attribute claims only at the document or paragraph level, without clear mappings to the context or assumptions underlying each individual claim.

A promising avenue of research focuses on developing source-aware training methods for LLMs. This approach aims to imbue LLMs with the ability to intrinsically link the knowledge they possess to the specific source documents from their pre-training data (Khalifa et al., 2024). This method of intrinsic source citation offers a potential way to trace the origins of information directly within the model's parameters, providing an alternative to relying solely on external retrieval mechanisms. However, it faces challenges related to the practicalities of collecting and managing the vast amounts of source data and the specificity required for document identifiers (Khalifa et al., 2024).

Feature	GPT-4.1	Gemini 2.5	LLaMA 3-70B
Link generated	✓ Yes	✓ Yes	✓ Yes
Link accurate	✗ Halluc.	✗ Halluc.	✗ Halluc.

Table 1: Link generation and accuracy for leading LLMs. "Halluc." indicates that the title or content is often hallucinated.

Positioning SciTrue A range of research prototypes and commercial products now provide scientific summarisation, literature review, or evidence retrieval, often integrating large language models (LLMs) to improve fluency and coverage. However, most outputs remain overly verbose, inflexible (typically adhering to systematic review templates), and lack precise mappings between synthesized claims and underlying evidence. Attribution is frequently surface-level or error-prone, and users are given limited support to verify individual inference steps. In contrast, **SciTrue** implements principled, context-sensitive, source-level attribution for each decomposed claim component. It explicitly links each subclaim to a specific, verifiable passage in the scientific literature, enabling user-centric audit and challenge. To our knowledge, SciTrue is among the first openly available systems to operationalize fine-grained attribution through an interface purpose-built for scientific accountability and transparent verification.

3 SciTrue: Transparent Scientific Claim Verification

SciTrue is an interactive, end-to-end platform for verifying scientific claims by combining large language models (LLMs) with retrieval from scientific literature (see the high-level workflow in Figure 2). This section describes the system's data flow, user experience, and transparency features.

3.1 Motivation and Overview

SciTrue is designed to assist scientists, journalists, students, and the general public in appraising scientific claims by grounding each claim in the scientific literature and making all evidence and reasoning steps explicit. Users interact through a web-based interface that allows them to submit claims and receive synthesized, well-attributed explanations directly linked to the original scientific sources.

Claim Refinement: A user begins by entering a scientific claim along with the desired number of scientific articles for analysis. If the query is unclear or contains abbreviations, SciTrue's query refinement agent transforms it into a well-formed scientific statement. If refinement is not possible, SciTrue prompts the user to revise their query, ensuring high-quality retrieval and synthesis in subsequent stages.

Literature Retrieval: Leveraging the Semantic Scholar API, which indexes over 214 million scientific papers across all fields of science, SciTrue retrieves articles relevant to each claim from a large database of scientific publications. For each selected article, SciTrue agents extract metadata (e.g., title, authors, journal, year, citation count) and identify candidate abstracts and paragraphs. If the article is deemed relevant, SciTrue extracts the most pertinent sentence and analyzes it using the rubric from Wei (2023), which categorizes evidence as follows:

- Declares something is better.
- Proposes something new.
- Describes a new finding or cause-effect relationship.

A SciTrue agent then assesses whether the evidence fully or conditionally supports or refutes the claim, identifies key assumptions or conditions underlying this support or refutation (such as a limited sample size, the study being conducted on mice rather than humans, or findings specific to a particular demographic), and determines its rhetorical role (i.e., the context in which the evidence is used in the article—such as a main finding, background information, or limitation). It also evaluates the relationship between the claim and the retrieved article by considering the article's title, abstract, and relevant paragraphs, categorizing the strength of this relationship as strong, medium, or weak. This enables SciTrue to create context-aware, interpretable, and fine-grained links between scientific claims and supporting literature.

Summary and Verdict Generation : A SciTrue agent synthesises all supporting and contradictory evidence into a coherent, concise summary by considering all evidence parameters identified in the previous step, including underlying assumptions. Each statement in the summary is precisely linked to a unique source article via a clickable citation,

Scientific Claim	Domain
Artificial sweeteners are healthier than sugar.	Health and Nutrition
Vitamin D supplements prevent respiratory infections.	Health and Nutrition
Nuclear waste can't be made safe for the long term.	Environment and Climate
Renewable energy deployment requires more mining overall.	Environment and Climate
Universal basic income will eliminate poverty.	Social Science
School uniforms reduce bullying and misbehaviour.	Social Science
Artificial intelligence will inevitably lead to widespread unemployment.	Technology and Policy
AI-based hiring tools reduce human biases in recruitment.	Technology and Policy

Table 2: Examples of scientific claims used in our study and their associated domains.

ensuring transparency and traceability. The system also generates a structured verdict and succinct justification, indicating whether the claim is fully, mostly, partially, or not supported by the aggregated evidence.

Subclaim and Evidence Interface: Alongside the summary and verdict, a SciTrue agent automatically decomposes the main claim into underlying subclaims. For each subclaim, the agent retrieves and aligns relevant supporting or refuting passages from primary sources. Where available, citation metrics and journal impact factors are included to contextualize the strength of the evidence. For each subclaim, the system provides the following information:

- **Subclaim expression**: The specific subclaim derived from the summary.
- **Relevant sentence**: The sentence from the source that supports or refutes the subclaim.
- **Exact paragraph**: The paragraph in which the subclaim appears.
- **Contribution label**: The extent to which the subclaim corroborates or contrasts with the main claim.
- **Supporting and refuting assumptions**: Key assumptions that underpin support for or refutation of the subclaim.
- **Rich metadata**: Authors, year, title, journal/venue, section, and a clickable source link.
- **Explicit evidence label**: Indicates whether the evidence fully or conditionally supports or refutes the claim.
- **Context**: The role the evidence plays within the article (e.g., main finding, background information, or limitation).
- **Claim-article relationship**: The relationship between the claim and the retrieved article, categorizing the strength of this relationship as strong, medium, or weak.
- **Quantitative indicators (where available)**: Citation counts and impact metrics.

Interactive Exploration and History: Users can browse previously analyzed claims in a personalized history, inspect executive summaries, subclaims, and all evidence details with expandable views, and revisit or reevaluate prior results. This workflow supports longitudinal and collaborative usage.

Language Model, Retrieval, and UI Choices While our methodology is compatible with any retrieval system or large language model, our current implementation utilizes the Semantic Scholar API[3] for its comprehensive coverage. For text generation, we employ GPT-4o[4], chosen after evaluating alternatives such as GPT-4.1, Gemini 2.5, and LLaMA 3 (70B). We selected GPT-4o due to its favorable balance of parsing accuracy, response speed, and lower API costs in small-scale experiments. To further manage costs, we limited the number of processed articles to a maximum of 15. For the user interface, we combined the Streamlit[5] library with front-end technologies including CSS, JavaScript, and HTML.

4 Evaluation

4.1 Dataset

Existing datasets in the scientific fact-checking domain, such as SciFact (Wadden et al., 2020) and Multi2Claim (Tan et al., 2023), are primarily designed for closed-domain setups. In contrast, SciTrue is designed for an open-domain setting, enabling it to track and incorporate the most recent

[3] https://www.semanticscholar.org/product/api
[4] https://openai.com/index/gpt-4o-system-card/
[5] https://streamlit.io/

Feature	GPT-4o-search-preview	Perplexity	SciTrue
Summary Traceability	86.36%	85.71%	98.50%
Overall Verdict	88.00%	85.71%	97.80%
Title Information	93.00%	86.00%	99.00%
Author Information	68.51%	67.78%	96.50%
Scientific Validity	48.18%	43.81%	94.20%
Factual Accuracy of Attribution	74.80%	80.95%	96.70%
Context & Assumptions	54.80%	51.43%	95.30%
Contribution Label Attribution	54.80%	71.43%	94.00%
Source Credibility	5.6%	7.5%	92.80%

Table 3: Human evaluation results comparing SciTrue with GPT-4o-Search-Preview and Perplexity Sonar Pro across key scientific attribution measures. The table reflects inter-annotator agreement, and the scores represent accuracy, as all evaluation questions were binary (Yes/No).

developments in scientific research. This open-domain approach not only allows for greater adaptability but also facilitates the identification and synthesis of 'grey areas' in science, namely, contested claims where multiple perspectives may coexist. Addressing such areas is essential, as the existence of debate, uncertainty, and evolving viewpoints is a fundamental aspect of the scientific process (Kuhn, 1962).

To this end, we tasked large language models (LLMs) with generating scientific claims across four key areas: health and nutrition, environment and climate, social science, and technology and policy. Representative examples of these claims are provided in Table 2, which also contains the curated subset used for our evaluation. The full dataset, comprising more than 3,000 scientific claims, is available in the project repository for community use.

4.2 Human Evaluation

We conducted a human evaluation comparing with two leading search-augmented LLMs: GPT-4o-Search-Preview and Perplexity Sonar Pro. Two independent annotators, each holding at least a bachelor's degree in science, assessed approximately 300 attributions, analyzing over 900 full-text sources (300 per model across three models), generated from 60 scientific claims. The claims were evenly distributed across four domains: health and nutrition, environment and climate, social science, and technology and policy. Five articles were retrieved per claim (60 claims × 5 articles = 300 attributions per model), a limit imposed due to current constraints in the capabilities of search-augmented LLMs (see an example at https://chatgpt.com/share/6858d793-00fc-8002-9333-434f10b41166).

In summary, we evaluated the three systems: SciTrue, GPT-4o-search-preview, and Perplexity Sonar Pro—across the following perspectives:

- **Summary Traceability**: Whether the summary is traceable to a specific, verifiable source and accurately reflects that source.
- **Overall Verdict**: Whether the model's overall conclusion and reasoning correctly reflect the summary information.
- **Title and Author Information**: Whether the cited source exists and the title and the listed authors matches the linked content.
- **Scientific Validity**: Whether the cited source is an academic paper, including key metadata such as title, abstract, authors, and publication year.
- **Factual Accuracy of Attribution**: Whether the cited sentence, or a semantically equivalent statement, appears in the referenced article.
- **Context and Assumption Awareness**: Whether the system detects when a source is used out of context or based on misleading assumptions.
- **Contribution Label Attribution**: Whether the assigned contribution type (e.g., corroborating, contrasting) is accurate within the sub-claim's context.
- **Source Credibility**: Whether the cited source's credibility (e.g., citation, impact factor) is accurately reflected.

Annotators followed a detailed rubric (see Appendix A.1). A screenshot of the full evaluation

interface is shown in Figure 3 in the appendix and is also available via the live demo for voluntary evaluation by future users. Inter-annotator agreement was 90%. Any remaining discrepancies were resolved through discussion.

5 Results and Discussion

SciTrue consistently outperformed across every perspective assessed. A closer analysis revealed key shortcomings of the competing systems. Notably, GPT-4o-search-preview and Perplexity Sonar Pro frequently retrieved blog posts, news sites, and general web pages as evidence rather than authoritative scientific articles; both systems included less than fifty percent scientific articles among their sources. These less rigorous sources often lack transparent vetting and replicability, severely limiting their usefulness for scientific verification. Moreover, despite relying on these sources, the models frequently hallucinated citations or misattributed findings to unreliable content, further undermining trust and underscoring the critical need for careful source selection in scientific fact-checking.

In contrast, SciTrue consistently prioritized primary scientific literature from reputable publication venues, ensuring that every claim was linked to a verifiable, high-credibility source. Its transparent audit trail not only links claims to sources, but also explicitly documents the attribution process—empowering users to replicate or challenge system decisions. This level of accountability is a foundational requirement for trustworthy AI in both research and journalism.

The results underscore key challenges faced by retrieval augmented LLM-based scientific assistants, particularly in guaranteeing the credibility and appropriateness of retrieved evidence. SciTrue's ability to outperform mainstream models highlights the need for rigorous source filtering and transparent attribution mechanisms. Looking forward, future research should explore source-aware training methods via prevention of non-scholarly source retrieval, robustness to adversarial or ambiguous cases, and interfaces that support community-driven auditing and feedback. Such advances are critical for the safe and effective integration of AI into scientific, journalistic, and public knowledge environments.

6 Conclusion

SciTrue addresses critical gaps in scientific claim verification by emphasizing evidence traceability, rigorous attribution, and context-aware reasoning. This is vital for claims in gray areas of science, where demographic factors, time period, and location shape interpretation. By incorporating these contexts, SciTrue provides nuanced, reliable assessments.

Targeted at researchers, journalists, and policymakers who require trust and auditability, SciTrue offers transparent, source-linked outputs that enable informed debate and scrutiny foundations of trustworthy science and public discourse. It overcomes limitations of both standalone and retrieval-augmented LLMs through granular source mapping.

Following a stepwise pipeline of claim refinement, literature retrieval, evidence extraction, linked summarization, subclaim decomposition, and interactive UI output, SciTrue outperforms retrieval-augmented LLM baselines in human evaluations across multiple metrics, delivering accurate, verifiable, and contextually aware scientific claim assessments.

7 Limitations and Ethical Considerations

Limitations Our approach depends on the coverage and metadata quality of the Semantic Scholar API. Although it is a broad and openly accessible resource, representation varies across disciplines, and some relevant publications may be missing, paywalled, or not digitally available. As a result, SciTrue is currently limited to the literature indexed by Semantic Scholar, and its effectiveness may be reduced in areas where key source documents are not accessible. These constraints reflect the current state of scholarly data infrastructure, and future integrations with additional corpora may help broaden SciTrue's reach.

Ethical Considerations SciTrue aims to reduce hallucinated attributions and improve verifiability, but automated systems can still inherit biases or overlook contextual nuances present in the underlying data. For this reason, SciTrue should be viewed as an assistive tool, supporting rather than replacing expert judgment, close reading, and peer review. Responsible use involves interpreting its recommendations critically and in conjunction with domain expertise.

References

Isabelle Augenstein, Sebastian Riedel, Pontus Stenetorp, Leon Derczynski, Georgios Spithourakis, Alex Dutton, and Yang Ji. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *NAACL-HLT*.

Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, Richard M Shiffrin, Samuel J Gershman, Vencislav Popov, Emily M Bender, Marco Marelli, Matthew M Botvinick, Zeynep Akata, and Eric Schulz. 2025. How should the advancement of large language models affect the practice of science? *Proc. Natl. Acad. Sci. U. S. A.*, 122(5):e2401227121.

Julia G Bottesini, Christie Aschwanden, Mijke Rhemtulla, and Simine Vazire. 2023. How do science journalists evaluate psychology research? *Adv. Methods Pract. Psychol. Sci.*, 6(3).

Lihu Chen, Shuojie Fu, Gabriel Freedman, Cemre Zor, Guy Martin, James Kinross, Uddhav Vaghela, Ovidiu Serban, and Francesca Toni. 2025. Pub-guard-LLM: Detecting fraudulent biomedical articles with reliable explanations. *arXiv [cs.CL]*.

Consensus. `https://consensus.app/`.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms^2: Multi-document summarization of medical studies. In *EMNLP*.

Elicit. `https://elicit.org/`.

Google DeepMind. 2024. Gemini 1.5 technical report. Accessed June 2025.

Antonios Kanavouras and Frank Coutelieris. 2020. Similarity among physical phenomena recognized on the basis of the classification of existing knowledge. *Journal of Mathematical Sciences and Modelling*, 3(2):47–54.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. *arXiv [cs.CL]*.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Meta. 2025. Gemini 2.5 report.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use LLMs as zero-shot translators. *Nat. Hum. Behav.*, 7(11):1830–1832.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain James Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *ArXiv*, abs/2305.06299.

Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. 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*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics.

Neşet Özkan Tan, Niket Tandon, David Wadden, Oyvind Tafjord, Mark Gahegan, and Michael Witbrock. 2024. Faithful reasoning over scientific claims. *Proceedings of the AAAI Symposium Series*, 3(1):263–272.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Majid Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pages 809–819.

Deepa Tilwani, Revathy Venkataramanan, and Amit P Sheth. 2024. Neurosymbolic AI approach to attribution in large language models. *arXiv [cs.CL]*.

David Wadden, Shanchan Lin, Kyle Lo 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 7534–7546.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2021. Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. In *AMIA Summits on Translational Science Proceedings*.

Xin Wei. 2023. ClaimDistiller: Scientific claim extraction with supervised contrastive learning. pages 65–77.

A Appendix

A.1 Instructions for Human Evaluators

This section describes the human evaluation interface and protocol used for claim and summary evaluations, as implemented in our Streamlit annotation tool.

User Identification Before beginning any evaluation, you must enter your **username** in the sidebar. This is required to personalize and track your progress.

System Selection & Claim Loading Using the sidebar, select the system output you wish to evaluate (*e.g.*, Model1, Model2, or Model3). The interface will then display a list of available claims, marking those you have already evaluated. Select a claim to open its evaluation form.

Summary and Claim Presentation The main panel presents:

- The **Claim** (as provided by the system).
- The system-generated **Summary**.
- **Q1:** *Is the summary traceable to the cited sources (i.e., when you click sources, do you get the relevant page)?*
- Choose "Yes" or "No." Optionally, provide a brief justification.
- If you select "No", you **do not need to complete the rest of the form for this claim**; clicking Save will record all subclaim fields as "No" and end the evaluation for this entry.

Overall Verdict Consistency (Q2)

- **Q2:** *Does the system's overall verdict/label correspond correctly to the summary?*
- Select "Yes" or "No", and optionally state your reasoning.

Subclaims Evaluation For each listed subclaim, you are presented with detailed information such as citation, title, journal, authors, abstract, section, contribution, relevant sentence(s), and credibility scores if provided. For each subclaim, you must answer:

1. **Does the given URL open an article page?** (Select "Yes" or "No")
2. **Does the title align with the URL?** ("Yes" or "No")
3. **Do the authors align with the URL?** ("Yes" or "No")
4. **Is the cited source an academic paper?** (i.e., does it have plausible title, abstract, authors, and year? "Yes"/"No")
5. **Factuality:** *Does the subclaim (or its semantic equivalent) appear in both the summary and the cited article?* ("Yes"/"No")
   Note: If you select "No" here, the following three questions will be automatically locked to "No":
6. **Contribution Label:** Is the contribution label accurate in context? ("Yes"/"No")
7. **Context & Assumption:** Does the model correctly detect context and assumptions? ("Yes"/"No")
8. **Credibility/Impact Representation:** Is the cited source's impact accurately represented (± 1 for impact factor, ± 10 for citation count)? ("Yes"/"No")

You may optionally provide reasoning for each subclaim evaluation.

Saving and Error Checking Before saving, the interface checks that all required fields are filled. Save your evaluation when complete. If this is the second time you are evaluating the same claim, your previous answers will be updated.

Notes and Best Practices

- **Carefully check** each source, title, and author match. Do not assume correctness based on citation style alone.
- **For factuality:** If the subclaim is not justified by the cited article, answer "No" even if it seems plausible.
- Use the "Show Paragraph" and "Show Abstract" toggles to see more context if needed.

Support For ambiguities or technical issues, contact the lead researcher.

A.2 Evaluation User Interface

A.3 LLM-as-judge evaluation

In addition to human annotation, we conducted an LLM-as-judge evaluation on a small subset of scientific claims (10 in total). For each claim, a state-of-the-art language model (GPT-4.1) was prompted with the same annotation criteria used in the human evaluation and asked to assess system outputs based on the provided evidence. The results of this evaluation showed only weak correlation with human annotators' judgments, as accessing source articles often requires additional clicks, and evaluating information from full articles remains beyond the current capabilities

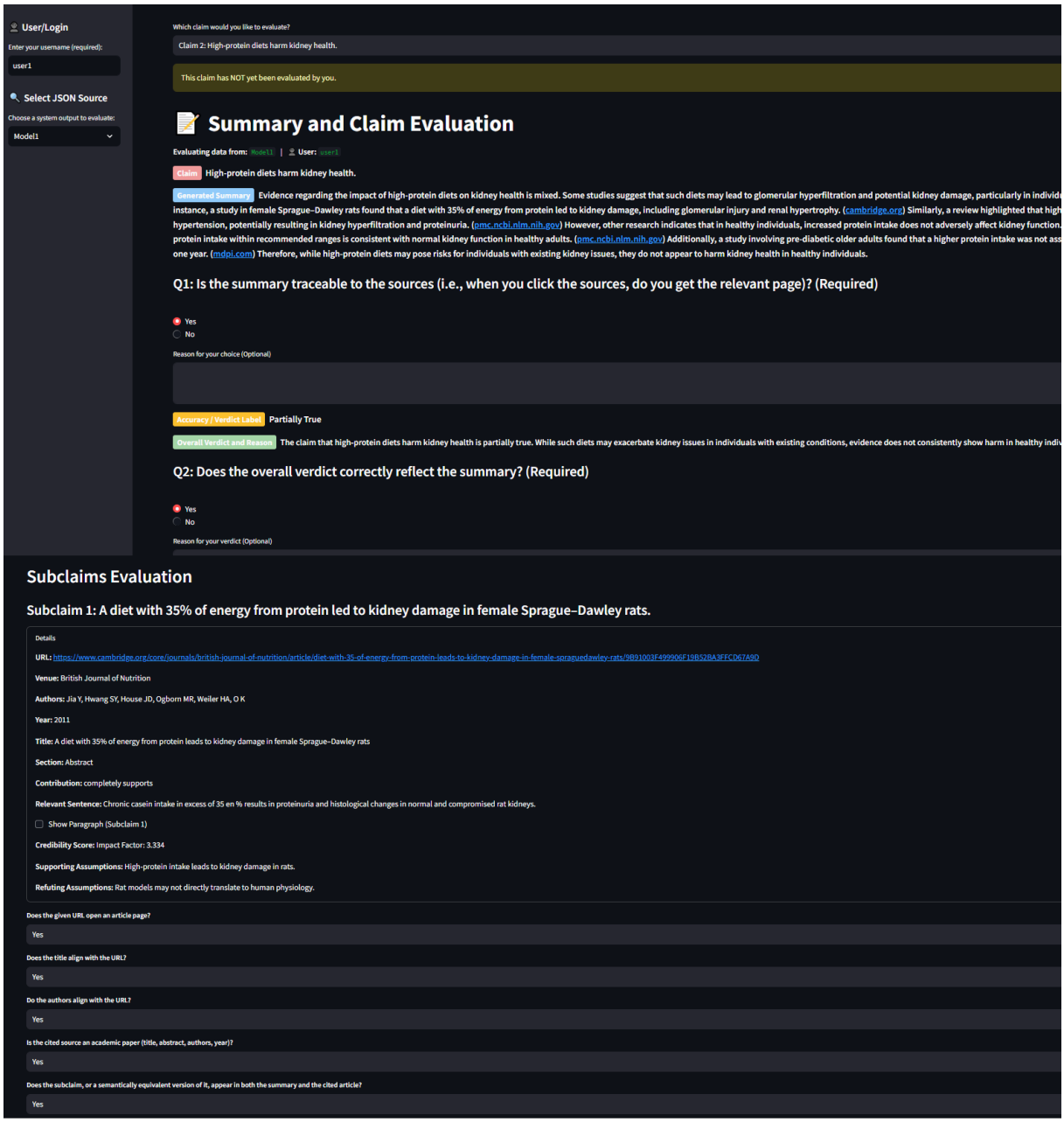

Figure 3: Screenshot of the evaluation interface. This interface is also available via the live demo.

A.4 Automated System Prompt for Claim Evaluation (GPT-4o-search-preview)

To generate system outputs for benchmark comparison, we used the following prompt and protocol with the GPT-4o-search-preview model. Each claim was presented alongside an explicit instruction to use exactly five scientific articles as evidence. The system was asked to produce a structured JSON object (and no extra text), suitable for downstream human annotation.