Humans vs. LLMs on Open Domain Scientific Claim Verification: A Baseline Study

Benjamin Curtis[1,†], Stefania Dzhaman[2,†], Matthew Maisonave[3] and Jian Wu[3,*]

[1]James Madison University, Harrisonburg, VA, United States
[2]Computer Science & Engineering, Lehigh University, Lehigh, PA, United States
[4]Computer Science, Old Dominion University, Norfolk, VA, United States

Abstract

Verifying scientific claims is challenging for the general public because most people lack domain knowledge. Manual verification by subject domain experts is accurate, but it is obviously not scalable to meet the rising number of scientific claims on the Web. Whether the emerging large language models and large reasoning models can be used for scientific claim verification, and how their performances compare to humans, are still research questions. To this end, we developed a new benchmark MSVEC2 that consists of 138 claims from credible fact verification websites and science news outlets. Two tasks were given to both human and LLM participants. Task 1 requests the tester (LLMs or humans) to discern the truthfulness of claims using only prior knowledge. Task 2 requests testers to determine the stance of a scientific claim relative to an abstract of a research paper. The LLMs that were evaluated include GPT-3.5, GPT-4, GPT-4o, GPT-o1, and DeepSeek-R1. We recruited 23 college students in various majors to participate in the human study. We found that all LLMs score higher in F1 and accuracy compared to human testers in truthfulness classification (Task 1), with GPT-4o achieving the highest F1 score among all the models. The performance of LLMs in stance classification (Task 2) depended on the prompting configuration, with Chain-of-thought prompting yielding consistent improvements for all LLMs except GPT-o1. However, the best performance of LLMs is still not sufficient for reliable scientific claim verification under standard prompt settings.

Keywords

scientific claim verification, large language model, large reasoning model, prompt engineering

1. Introduction

Online scientific disinformation misrepresents the findings of scientific papers and disseminates misleading or even malicious information to internet users. The prevalence of scientific misinformation online has become rampant in the news and on social media sites. Fact verification websites such as `Reuters.com` and `Snopes.com` use teams of professionals to fact-check claims from multiple sources before judging their truthfulness. However, manually verifying scientific claims is time-consuming and often requires extensive domain knowledge (e.g., to read and digest scientific literature), and therefore does not scale to the massive number of claims spread on the internet. This leaves a pressing need for tools that can automatically verify scientific claims by assessing their credibility and providing a rationale for the assessment. Large language models (LLMs) and their variants, large reasoning models (LRMs), have been shown to have exceptional skills in text parsing and reasoning tasks, e.g., [1]. For convenience, we call both types LLMs. Although LLMs have been evaluated in their fact-checking capabilities against benchmark datasets such as FEVER [2], the majority of existing datasets focus on verifying *general claims*. Whether contemporary LLMs' capabilities on *scientific claim verification* (SCV) have reached the level of human beings has not been systematically investigated. It also remains unclear whether their performance is sufficient for reliable deployment in SCV applications.

AI4SciSci'25: Workshop on the Artificial Intelligence and the Science of Science, December 15, 2025, DeKalb, Illinois, USA
*Corresponding author.
†These authors contributed equally.
curtisbw@dukes.jmu.edu (B. Curtis); dzhamastefa@gmail.com (S. Dzhaman); Manfred.Jeusfeld@acm.org (M. Maisonave); j1wu@odu.edu (J. Wu)
https://www.cs.odu.edu/~jwu/ (J. Wu)
0000-0003-0173-4463 (J. Wu)

In this paper, we aim to fill this gap by evaluating the performance of five widely used LLMs on a carefully curated dataset containing 138 scientific claims compiled from credible fact verification websites. As a pilot study, we explore baseline prompting methods, including zero-shot, one-shot, and Chain-of-Thought (CoT; [3]) on two tasks. Task 1 requires testers (i.e., LLMs or human respondents) to judge the truthfulness of scientific claims. Task 2 requires testers to classify the stances of an abstract from a scientific paper relative to a claim. To evaluate the human performance on the same tasks, we recruited 23 college students and asked them the same questions. The results allow us to perform a comparative study across LLMs and between LLMs and college students.

To evaluate the performance, we developed a new dataset by carefully selecting a subsample of scientific claims from an existing dataset, MSVEC [4]. The new dataset consists of 138 scientific claims, each of which is annotated as true or false based on its original labels in the fact verification websites or credible science news outlets, and paired with a reference abstract that supports the claim, refutes the claim, or does not have enough information to determine the truthfulness of the claim.

We perform extensive experiments by varying the prompts and evaluate performances in multiple settings. For Task 1, we find that all LLM versions outperform humans in either discerning true or false claims. We also observe that the two LRMs (GPT-o1 and DeepSeek-R1) tend to assign the "false" label to claims they struggle with. For Task 2, we find that CoT prompting used on all GPT versions outperforms humans in nearly all trials; the few-shot prompting method generally outperforms humans, and the zero-shot prompting method yielded mixed results. We find that for Task 2, the *refutes* stances achieve relatively low performance by 10% in both human and LLM experiments, suggesting that contradictory relationships are the hardest to correctly identify compared with *support* or *NEI* (not enough information) stances.

2. Related Works

2.1. SCV Datasets

Early SCV datasets for evidence-based fact-checking were created from general fact-checking sources rather than scientific sources. The FEVER dataset [2], laid the groundwork for claim-evidence alignment. The dataset contains 185,000 claims sourced from Wikipedia and uses the claim labeling structure Supports/Refutes/NEI, which FEVER helped define. The SciFact dataset [5] contains 1409 scientific claims, which are human-coded citation contexts, supported by 5,183 abstracts of papers, mostly in biomedical science domains, which are also labeled as Supports/Refutes/NEI.

The SCitance [6] and RECV [1] datasets emphasize the reasoning process. The aim behind SCitance was to manually rewrite FEVER-style claims with "citances", described as naturally occurring citation sentences. The RECV benchmark introduced either deductive or abductive reasoning-type labels, which span across multiple datasets, including VitaminC [7], CLIMATE-FEVER [8], and PHEMEPlus [9]. Other datasets are developed in specific domains, such as CliVER [10] (biomedical sciences), HealthVer [11] (health-related claims), and NLI4CT [12] (clinical trial), These datasets contributed reasoning awareness and naturalistic text to the space; however, they lack human baselines and cross-domain diversity and the claims and claim labels are not collected from credible fact-verification websites.

As SCV datasets grew, researchers focused their attention on expanding the size of the datasets by automatically generating claims and evidence. The datasets SciClaimHunt and SciClaimHuntNum [13] were built using this methodology. Synthetic datasets achieve impressive scalability, wide domain diversity, and a meaningful inclusion of numerical reasoning. However, synthetic negations can misrepresent reasoning due to a lack of human nuance, and they may embed generator bias.

Our dataset distinguishes itself from existing datasets in several key aspects. First, instead of rewriting citation contexts in scientific papers as claims, our claims are collected from fact-checking websites or credible science news outlets, making them closer to the scientific claims seen in the real world. The global truthfulness has been verified by experts or science news editors instead of being inferred from the citation relationships in scientific papers. Furthermore, the dataset contains 9 distinct domains with a stance distribution balanced as 35.5% Supports, 21% Refutes, and 43.5% NEI. For the binary truthfulness task, the stances are balanced as 53.6% True and 46.4% False. In our dataset, a True claim does not necessarily have to be associated with a supportive abstract. Table 1 summarizes the properties of selected SCV datasets and ours.

Table 1
A Comparison of Selected Scientific Claim Verification (SCV) Datasets.

Dataset	Size (#claims)	Task	Topics	Source
FEVER [2]	185K claims	Supports / Refutes / NEI	General fact-checking (Wikipedia)	Wikipedia-based claims; foundation for SCV schema
SciFact [14]	1,409 claims	Supports / Refutes / NEI	Biomedical / scientific papers	PubMed abstracts paired with expert-verified claims
SCITANCE [6]	1,400 pairs	Supports / Refutes	Scientific research (citation-based)	Derived from SciFact citation sentences + abstracts
VitaminC [7]	125K pairs	Supports / Refutes	General / evidence sensitivity	Wikipedia revisions introducing factual perturbations
PHEMEPlus [9]	6,000 claims	True / False / Unverified	Social media / news claims	Extended PHEME rumor dataset with stance labels
Climate-FEVER [8]	5,300 claims	Supports / Refutes / NEI	Environmental science	Climate-related claims with evidence from scientific sources
SciClaimHunt [13]	300K synthetic claims	Supports / Refutes / NEI / Numerical	Multi-domain scientific literature	Generated via **LLaMA-2** from paper discussions; tests factual and numerical reasoning
RECV [1]	2,000 claim–evidence pairs	Supports / Refutes (Reasoning)	Multi-domain reasoning tasks	Derived from SCITANCE with added deductive/abductive reasoning annotations
MSVEC2 (ours)	**138 claims**	**True / False and Supports / Refutes / NEI**	**Open-domain scientific claims (9 domains)**	Human-verified benchmark updated from MSVEC (2023); curated claim–abstract pairs for LLM evaluation

2.2. SCV Methods

The two mainstream SCV methods include named retrieval-based systems and LLMs.

FactDetect [15] introduced a modular pipeline that performs claim decomposition, evidence retrieval, fact-level evaluation, and aggregation. Both the lexical retriever (BM25 [16]) and the dense retriever (ColBERT [17]) were used to locate relevant sentences for evidence retrieval. The CliVER framework consists of document collection in which a hybrid lexical and dense retrieval from PubMed was used, document retrieval, sentence selection, label prediction, and training and evaluation. The ensemble of RoBERTa [18], PubMedBERT [19], and T5 [20] models predicts whether the rationale Supports, Refutes, or is Neutral to the claim. Recently, CoVERt [21] was introduced along with the PICO [22] structured evidence framework. This approach emphasizes scalability and domain specialization. Both FactDetect and CliVER highlight retrieval and decomposition for accurate verification. Limitations to these systems include a supervised data dependency, domain-specific design, and limited reasoning depth.

Recently, researchers shifted their focus to the improvement of SCV using LLMs, which provides a generalizable solution for open-domain claim verification. For example, ProToCo [23] is a prompt-based consistency training framework that uses three claim variants: affirmation, negation, and uncertainty. The framework trains LLMs to keep answers logically coherent across variants, as well as improving factual reliability in few and zero-shot settings. MAPLE [24] models micro-language evolution between claims and evidence, as well as capturing subtle semantic shifts that signal factual entailment. A T5 model with LoRA [25] is trained to generate claims from evidence and vice versa.

This paper focuses on providing a baseline comparison of SCV performance between commonly used LLMs and humans (represented by college students), which has not been done by any of the previous studies.

Table 2
The domain distribution of our SCV dataset.

Domain	#Claims	%Claims	#Support	#Refute	#NEI
Environment	16	11.6%	6	6	4
Health	61	44.2%	21	12	28
Humans	14	10.1%	2	2	10
Nature	11	8.0%	6	0	5
Opinion	6	4.3%	1	3	2
Physics	7	5.1%	5	1	1
Society	8	5.8%	1	1	6
Space	5	3.6%	2	1	2
Tech	7	5.1%	5	0	2
Uncategorized	3	2.2%	1	2	0
Total	**138**	**100%**	**50**	**28**	**60**

Table 3
Examples of claims removed from MSVEC [4] and the reasons.

Reason to remove	Examples
non-scientific	`New Florida scheme allows veterans - not their spouses - to a temporary teaching certificate without having completed a college degree.`
lack context	`A viral animation shows a myosin molecule transporting endorphins.`
compound	`A third of us can no longer see the Milky Way, which negatively impacts our health.`

3. The MSVEC2 Dataset

3.1. Dataset Construction and Properties

Our SCV dataset, named MSVEC2, is derived from the original MSVEC dataset consisting of 200 labeled claims and claim–abstract pairs [4]. The claims were sourced from fact-checking websites and credible news outlets, and the abstracts are from peer-reviewed scientific articles. We removed 62 claims in the following categories. (1) Non-scientific claims ; (2) claims that are not self-contained (i.e., needing more context); (3) compound claims (i.e., a claim composed of multiple sub-claims) (see Table 3 for examples).

In addition to the removal of the above unqualified claims, each claim-abstract pair is also manually inspected by two undergraduate researchers independently against the source to ensure the paper was actually used to support/refute the claims. In certain cases, the MSVEC data may identify a different paper from the correct paper in the claim-abstract pair because the original news article cites several papers. The consensus rate is 99%. Pairs with misidentified papers or the lack of reviewer consensus are removed, leaving 138 scientific claims in the final dataset. Each claim is labeled with `True` or `False` and an abstract that either supports, refutes, or does not provide enough information (NEI) relative to the claim. Covering nine distinct scientific domains (Table 2), MSVEC2 was designed as a multi-domain benchmark dataset rather than a domain-specific corpus. The distribution of stance labels is shown in Table 2. In total, 53.6% of the claims were labeled `True` and 46.4% `False`.

3.2. Research Tasks

MSVEC2 supports two tasks. Task 1 evaluates the ability to determine the truthfulness of a scientific claim. The tester, either an LLM or a human respondent, is presented with the claim text only and asked to judge whether it is true or false. Task 2 evaluates the ability to classify the stance of a scientific abstract relative to a claim. Given a claim and an abstract, the tester, either an LLM or a human respondent, selects one of three stances: Supports, Refutes, or NEI.

4. Evaluation

4.1. Evaluation Metrics

Both tasks can be treated as classification problems, we adopt precision P, recall R, and $F1$ score as the evaluation metrics. We also calculate the Accuracy to evaluate the overall performance. For Task !, we calculate the P, R, and $F1$ of the `True` and `False` claims. For Task 2, we calculate the P, R, and $F1$ for the `support`, `refute`, and `NEI` stances.

4.2. Human Study

Because we aim to compare humans' performance against LLMs', we selected human participants with *reasonable educational backgrounds to understand scientific claims and make independent decisions*. Although varied across countries, the general academic goal of K-12 education is to equip students with foundational knowledge and critical thinking skills. The majority of college students have finished K-12 education, so they should possess a reasonable educational background to make independent decisions about scientific claims. The goal of graduate school is to achieve a deep, specialized education in a chosen field. Therefore, choosing graduate students will significantly narrow the range of the represented population of this study. According to the US Census Bureau, more than 90% of US population aged 18 and above have finished secondary education. Obtaining a large-scale human subject sample with diverse ages and backgrounds is beyond our capability and will be reserved for future study. Therefore, we chose to focus on college students because they meet our educational level criteria, and we can draw meaningful conclusions based on a reasonably sized human subject sample.

We recruited a total of 23 college students from the 1st through the 4th year from an R1 university according to the Carnegie classification system. The participants include 12 females and 11 males, with an average GPA of 3.57. Among the participants, 69.6% majored in engineering disciplines, 17.4% in the nursing, biological, and chemistry sciences, and 13.0% in other disciplines. Each participant took part in a survey to carry out Task 1 and Task 2. Qualtrics, an online survey platform, was used to pose the queries. Participants took the surveys on their own devices and on their own time. Participants were shown a five-minute instructional video before beginning the surveys, which gave examples of questions they would encounter and explained the protocol for answering them.

The whole survey was divided into 5 sessions, each covered 14 claims. Each claim had 2 corresponding questions corresponding to Task 1 and Task 2 (see Section 3.2). Each session generally took participants between 30 and 60 minutes to complete, and they were asked to complete all 5 sessions within 10 days. A limit of 10 days was given to balance the workload and reduce the possibility of acquiring external knowledge relevant to the claims through school education or life experience, so their performance stayed relatively consistent across all sessions. Participants were required not to refer to any external sources when working on the tasks. Each participant was awarded an Amazon gift card worth $80 upon completion as compensation for their time. The human study results were micro-averaged, or pooled together and evaluated as one participant, and the F1-score was compared to the F1-score observed in the LLM trials.

4.3. Large Language Model Study

Here, we evaluate 5 commonly used LLMs, including GPT-3.5, GPT-4, GPT-4o, GPT-o1, and DeepSeek-R1 on Tasks 1 and 2. GPT-3.5, GPT-4, and GPT-4o were selected due to their strong performance on many general tasks and popularity to be used for baseline comparison, e.g., [6]. GPT-o1 and DeepSeek-R1 are usually considered LRMs [26].

For Task 1, we only test the zero-shot prompting method because the claims are ad hoc and thus do not need examples or an articulation of the reasoning process. For Task 2, we test three prompting methods for each LLM (including LRMs), zero-shot, few-shot, and chain-of-thought (CoT; [3]). In few-shot prompts, we provide an LLM with examples of correctly answered queries before posing the test query. In the CoT prompts, we provide examples of correctly answered queries before posing a question. In the examples, an abstract and a claim were first given, followed by a four-step reasoning process, shown below.

```
Read the claim and abstract below, then reason step by step before answering the
question:

Claim: [example claim]
Abstract: [example abstract]
Question: Does the abstract of the scientific paper support the claim, refute the claim,
or is there not enough information?

Answer: Step 1: Read the whole abstract and extract information relevant to the question:
[relevant information]
Step 2: Identify the relevant statement: [relevant statement]
Step 3: Give reasoning to rationalize your decision: [rationale]
Step 4: Conclusion: [conclusion]
Now, read the new claim and abstract below and answer the question at the end:

Claim: [target claim]
Abstract: [target abstract]
Question: Does the abstract of the scientific paper support or refute the claim, or is
there not enough information?
Answer with one of the following labels: SUPPORTS, REFUTES, or NOT ENOUGH INFORMATION
```

4.3.1. Experimental Settings

All model runs were performed with temperature 0 using standardized prompt templates to ensure consistency across models and tasks. Model outputs were normalized to canonical labels (i.e., `True/False` for Task 1 and `Supports/Refutes/NEI` for Task 2) before scoring. 15 entries in the dataset were reserved as examples for few-shot and CoT prompting (Task 2 only). The remaining entries are used as test samples for Tasks 1 and 2. For each claim of Task 2, three examples, corresponding to three stance labels (i.e., `Supports/Refutes/NEI`) were given. We experimented with up to 3 shots. The examples were selected by prioritizing an even distribution of claims from different domains.

5. Results

5.1. Task 1

The results of Task 1 are summarized in Table 4. The results suggest that all LLMs outperform humans in terms of F1 scores and accuracy at determining the truthfulness of scientific claims, whether the original claim is true or false. The discrepancy of accuracy ranges $0.10 - 0.19$. The discrepancies of $F1_{true}$ and $F1_{false}$ range $0.10 - 0.22$ and $0.09 - 0.19$, respectively. All LLMs achieve slightly better F1 scores for `False` claims compared with `True` claims.

The two LRM models (GPT-o1 and DeepSeek-R1) favored recall on `False` claims, showing that they prefer rejecting uncertain statements rather than incorrectly affirming them. In contrast, the three general LLMs (GPT-3.5, GPT-4, and GPT-4o) favored recall on `True` claims, indicating an opposite bias. The results from Task 1 demonstrate that statistically state-of-the-art LLMs are more accurate than humans for discerning true or false scientific claims. However, even the best performing LLM (an accuracy of 0.90 and an $F1_{true}$ of 0.91 for GPT-4o) has significant room to improve.

5.2. Task 2

The F1 scores of humans and LLMs with various prompting methods are shown in Table 5 (the detailed results are shown in the Appendix.), suggesting that depending on the version and prompting method,

Table 4
The human and LLM evaluation results for Task 1. The highest F1-score for all LLMs for either label and the highest accuracy for all LLMs are shown in bold.

Label	Metric	Human	GPT-3.5	GPT-4	GPT-4o	GPT-o1	DeepSeek-R1
True Claims	Precision	0.59	0.81	0.66	0.83	0.98	0.90
	Recall	0.72	0.87	0.91	0.92	0.61	0.68
	F1	0.65	0.84	0.76	**0.87**	0.75	0.77
False Claims	Precision	0.81	0.89	0.96	0.94	0.74	0.76
	Recall	0.69	0.84	0.81	0.88	0.99	0.93
	F1	0.75	0.87	0.88	**0.91**	0.85	0.84
All Claims	Accuracy	0.71	0.85	0.84	**0.90**	0.81	0.81

Table 5
The F1 scores of LLMs and humans for Task 2. The highest F1-score for each LLM is shown in bold. Note that humans are not provided with any examples, so the experiments are presented as 0-shot.

Stance	Configuration	Human	GPT-3.5	GPT-4	GPT-4o	GPT-o1	DeepSeek-R1
Support	0-shot	0.66	0.64	**0.87**	0.76	0.59	0.62
	1-shot	-	0.68	0.76	0.75	0.71	0.57
	2-shot	-	0.67	0.76	0.72	**0.77**	0.56
	3-shot	-	0.68	0.73	0.72	0.75	0.67
	1-CoT	-	0.78	0.70	**0.80**	0.64	0.57
	2-CoT	-	**0.82**	0.72	0.78	0.64	0.64
	3-CoT	-	**0.82**	0.72	0.78	0.73	**0.67**
Refute	0-shot	0.53	0.43	0.61	0.55	0.60	0.49
	1-shot	-	0.36	0.49	0.60	0.65	0.52
	2-shot	-	0.38	0.52	0.67	0.67	0.57
	3-shot	-	0.33	0.51	0.64	**0.68**	0.48
	1-CoT	-	**0.65**	**0.65**	**0.74**	0.61	0.57
	2-CoT	-	0.49	0.62	0.70	0.64	0.62
	3-CoT	-	0.64	0.61	0.70	0.62	**0.70**
NEI	0-shot	0.68	0.52	**0.79**	0.64	0.71	0.69
	1-shot	-	0.58	0.72	0.73	0.75	0.71
	2-shot	-	0.64	0.77	0.74	**0.77**	0.71
	3-shot	-	0.64	0.74	0.74	**0.77**	0.71
	1-CoT	-	0.75	0.76	**0.82**	0.73	0.72
	2-CoT	-	0.79	0.76	0.80	0.74	0.73
	3-CoT	-	**0.80**	0.74	0.77	0.75	**0.76**

the LLMs may outperform humans at classifying the stance of a scientific paper abstract with respect to a given claim. LLMs outperformed human participants in most stance categories, with the amount of improvement depending on the prompting method and model type. CoT prompting produced the most consistent performance gains and was especially beneficial for Refute stances, which is likely due to examples and step-wise reasoning instructions aiding in models resolving contradictions between the claim and the abstract. Few-shot prompting performed well on `Support` stances, but was less effective on `Refute` and `NEI` stances. GPT-4o achieved the most balanced results across all stances. GPT-o1 and DeepSeek-R1 reached similar accuracy and performed particularly well on `NEI` and `Refute` classes. The human group achieves relatively low performance on the `Refute` stance.

6. Discussion

6.1. Performance Discussion

The implications of our study shed light on possibilities in using the state-of-the-art LLMs to discern the truthfulness of scientific claims seen on the web. Our results suggest that LLMs have powerful reasoning and text parsing capabilities that allow them to outperform humans, here represented by college students, at scientific claim verification tasks. However, the overall performance of the best LLMs is still unsatisfying for being deployed as a service. For example, the best performance of Task 1 was achieved by GPT-4o with an accuracy. The best F1-score is 0.87 and 0.91 for `True` and `False` claims, respectively. This indicates that a significant fraction of claims are still mislabeled.

In contrast, the low F1 scores of the Refutes stance in Task 2 (Table 5) for both humans and LLMs suggest that contradictory relationships between the claims and abstracts are likely to be more difficult to discern than support or NEI relationships.

The limitations of our study are the size of the human study and the participant population being limited to college students. In future work, a larger and more diverse human participant pool will be constructed to better represent the web content consumers.

6.2. Reasoning-Optimized Model Behavior

GPT-o1 and Deepseek-R1 are reasoning-optimized models that generate intermediate reasoning traces before outputting answers and use deliberative reasoning over direct factual recall. The results of Task 1 (Table 4) indicate that GPT-o1 and DeepSeek-R1 lean toward cautious labeling like `False`, which is seen from the high recall compared with low precision values. This pattern aligns with findings from previous benchmarks, which show that explicit reasoning and structured explanation traces can lead models to over-reject partially supported claims, e.g., [1]. Our findings show that producing longer reasoning chains does not always improve F1, which could be attributed to generated rationales being occasionally self-contradictory or disconnected from evidence. If so, this suggests that explicit reasoning may introduce error propagation when intermediate steps are not grounded in fact.

GPT-o1 performed best with concise few-shot prompts in Task 2, indicating that GPT-o1 likely performs implicit internal reasoning that explicit CoT disrupts. Interestingly, DeepSeek-R1 showed the opposite trend, with 3-CoT prompting yielding the best performance. This may be due to a difference in structure between the two models, so DeepSeek-R1 benefits from explicit external reasoning traces that reinforce stance alignment and coherence. The difference suggests that reasoning-optimized design can manifest differently and that it shapes the decision style of the model rather than uniformly improving factual accuracy.

7. Conclusion

We developed a new benchmark dataset MSVEC2, consisting of 138 scientific claims from credible fact-verification websites and science news outlets, including truthfulness labels and an abstract that supports, refutes, or does not contain enough information with respect to the claim. We benchmarked humans, represented by 23 college students and 5 state-of-the-art LLMs, through two tasks, namely truthfulness classification and stance classification. We found that all LLMs score higher in F1 scores and accuracy compared to humans in truthfulness classification, with GPT-4o achieving the highest F1 score among all the models. The performance of LLMs in stance classification depends on the prompting configuration, with Chain-of-thought yielding consistent improvements for all LLMs except GPT-o1. However, the performance of LLMs is still not sufficient for reliable scientific claim verification under standard prompt settings.

Generative AI Declaration

The authors used ChatGPT and Grammarly to perform grammar and spelling checks, paraphrase, and reword. After using the tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Acknowledgments

This project is partially supported by the US National Science Foundation Award #2149607.

References

[1] Dougrez-Lewis, John and Akhter, Mahmud Elahi and Ruggeri, Federico and Löbbers, Sebastian and He, Yulan and Liakata, Maria, Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification, arXiv preprint arXiv:2402.10735 (2024).

[2] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074/. doi:10.18653/v1/N18-1074.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[4] Evans, Michael and Soós, Dominik and Landers, Ethan and Wu, Jian, MSVEC: A multidomain testing dataset for scientific claim verification, in: Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, 2023, pp. 504–509.

[5] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609/. doi:10.18653/v1/2020.emnlp-main.609.

[6] Alvarez, Carlos and Bennett, Maxwell and Wang, Lucy Lu, Zero-shot scientific claim verification using LLMs and citation text, in: Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), 2024, pp. 269–276.

[7] Schuster, Tal and Fisch, Adam and Barzilay, Regina, Get Your Vitamin C! Robust Fact Verification with Contrastive Evidenc, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 624–643. URL: https://aclanthology.org/2021.naacl-main.52/. doi:10.18653/v1/2021.naacl-main.52.

[8] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, M. Leippold, Climate-fever: A dataset for verification of real-world climate claims, arXiv preprint arXiv:2012.00614 (2020).

[9] Dougrez-Lewis, John and Kochkina, Elena and Arana-Catania, Miguel and Liakata, Maria and He, Yulan, PHEMEPlus: enriching social media rumour verification with external evidence, arXiv preprint arXiv:2207.13970 (2022).

[10] H. Liu, A. Soroush, J. G. Nestor, E. Park, B. Idnay, Y. Fang, J. Pan, S. Liao, M. Bernard, Y. Peng, et al., Retrieval augmented scientific claim verification, JAMIA open 7 (2024) ooae021.

[11] Sarrouti, Mourad and Abacha, Asma Ben and Mrabet, Yassine and Demner-Fushman, Dina,

Evidence-based fact-checking of health-related claims, in: Findings of the association for computational linguistics: EMNLP 2021, 2021, pp. 3499–3512.
[12] Jullien, Maël and Valentino, Marco and Frost, Hannah and O'Regan, Paul and Landers, Donal and Freitas, André, SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data, arXiv preprint arXiv:2305.02993 (2023).
[13] Kumar, Sujit and Sharma, Anshul and Khincha, Siddharth Hemant and Shroff, Gargi and Singh, Sanasam Ranbir and Mishra, Rahul, Sciclaimhunt: A large dataset for evidence-based scientific claim verification, arXiv preprint arXiv:2502.10003 (2025).
[14] D. Wadden, K. Lo, B. Kuehl, A. Cohan, I. Beltagy, L. L. Wang, H. Hajishirzi, Scifact-open: Towards open-domain scientific claim verification, arXiv preprint arXiv:2210.13777 (2022).
[15] Jafari, Nazanin and Allan, James, Robust claim verification through fact detection, arXiv preprint arXiv:2407.18367 (2024).
[16] Robertson, Stephen and Zaragoza, Hugo and others, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.
[17] Khattab, Omar and Zaharia, Matei, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020, pp. 39–48.
[18] Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin, RoBERTa: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[19] Gu, Yu and Tinn, Robert and Cheng, Hao and Lucas, Michael and Usuyama, Naoto and Liu, Xiaodong and Naumann, Tristan and Gao, Jianfeng and Poon, Hoifung, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, ACM Trans. Comput. Healthcare 3 (2021). URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.
[20] j. Raffel, Colin and Shazeer, Noam and Roberts, Adam and Lee, Katherine and Narang, Sharan and Matena, Michael and Zhou, Yanqi and Li, Wei and Liu, Peter J, Exploring the limits of transfer learning with a unified text-to-text transformer 21 (2020) 1–67.
[21] Liu, Hao and Soroush, Ali and Nestor, Jordan G and Park, Elizabeth and Idnay, Betina and Fang, Yilu and Pan, Jane and Liao, Stan and Bernard, Marguerite and Peng, Yifan and Weng, Chunhua, Retrieval augmented scientific claim verification, JAMIA Open 7 (2024) ooae021. URL: https://doi.org/10.1093/jamiaopen/ooae021. doi:10.1093/jamiaopen/ooae021. arXiv:https://academic.oup.com/jamiaopen/article-pdf/7/1/ooae021/56904263/ooae021.pdf.
[22] S. A. Miller, J. L. Forrest, Enhancing your practice through evidence-based decision making: Pico, learning how to ask good questions, Journal of Evidence Based Dental Practice 1 (2001) 136–141. URL: https://www.sciencedirect.com/science/article/pii/S1532338201700243. doi:https://doi.org/10.1016/S1532-3382(01)70024-3.
[23] F. Zeng, W. Gao, Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models, in: Findings of the Association for Computational Linguistics, ACL 2023, Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2023, pp. 4555–4569. Publisher Copyright: © 2023 Association for Computational Linguistics.; 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023 ; Conference date: 09-07-2023 Through 14-07-2023.
[24] X. Zeng, A. Zubiaga, MAPLE: Micro analysis of pairwise language evolution for few-shot claim verification, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1177–1196. URL: https://aclanthology.org/2024.findings-eacl.79/.
[25] Hu, Edward J and Shen, Yelong and Wallis, Phillip and Allen-Zhu, Zeyuan and Li, Yuanzhi and Wang, Shean and Wang, Lu and Chen, Weizhu and others, LoRA: Low-rank Adaptation of Large Language Models, ICLR 1 (2022) 3.
[26] Li, Zhong-Zhi and Zhang, Duzhen and Zhang, Ming-Liang and Zhang, Jiaxin and Liu, Zengyan

and Yao, Yuxuan and Xu, Haotian and Zheng, Junhao and Wang, Pei-Jie and Chen, Xiuyi and others, From System 1 to System 2: A Survey of Reasoning Large Language Models, arXiv preprint arXiv:2502.17419 (2025).

Appendix

lllll Extended Task 2 Metrics by Model and Configuration

Configuration	Stance	Accuracy	Precision	Recall	F1

Table 5 – continued from previous page

Configuration	Stance	Accuracy	Precision	Recall	F1

Continued on next page

o1-zeroShot	Refutes	0.845528	0.666667	0.538462	0.595745
o1-zeroShot	Supports	0.788618	0.95	0.431818	0.59375
o1-zeroShot	NEI	0.682927	0.585366	0.90566	0.711111
o1-oneShot	Refutes	0.869919	0.75	0.576923	0.652174
o1-oneShot	Supports	0.837398	0.961538	0.568182	0.714286
o1-oneShot	NEI	0.739837	0.636364	0.924528	0.753846
o1-twoShot	Refutes	0.878049	0.789474	0.576923	0.666667
o1-twoShot	Supports	0.861789	0.965517	0.636364	0.767123
o1-twoShot	NEI	0.756098	0.653333	0.924528	0.765625
o1-threeShot	Refutes	0.886179	0.833333	0.576923	0.681818
o1-threeShot	Supports	0.853659	0.964286	0.613636	0.75
o1-threeShot	NEI	0.756098	0.649351	0.943396	0.769231
o1-1CoT	Refutes	0.853659	0.7	0.538462	0.608696
o1-1CoT	Supports	0.804878	0.954545	0.477273	0.636364
o1-1CoT	NEI	0.707317	0.604938	0.924528	0.731343
o1-2CoT	Refutes	0.869919	0.777778	0.538462	0.633664
o1-2CoT	Supports	0.804878	0.954545	0.477273	0.636364
o1-2CoT	NEI	0.707317	0.60241	0.943396	0.735294
o1-3CoT	Refutes	0.861789	0.736842	0.538462	0.622222
o1-3CoT	Supports	0.845528	0.962963	0.590909	0.732394
o1-3CoT	NEI	0.739837	0.636364	0.924528	0.753846
r1-zeroShot	Refutes	0.813008	0.578947	0.423077	0.488889
r1-zeroShot	Supports	0.788618	0.875	0.477273	0.617647
r1-zeroShot	NEI	0.666667	0.575	0.867925	0.691729
r1-oneShot	Refutes	0.837398	0.6875	0.423077	0.52381
r1-oneShot	Supports	0.780488	0.947368	0.409091	0.571429
r1-oneShot	NEI	0.666667	0.568182	0.943396	0.70922
r1-twoShot	Refutes	0.853659	0.705882	0.48	0.571429
r1-twoShot	Supports	0.772358	0.9	0.409091	0.5625
r1-twoShot	NEI	0.674797	0.581395	0.925926	0.714286
r1-threeShot	Refutes	0.821138	0.625	0.384615	0.47619
r1-threeShot	Supports	0.804878	0.916667	0.5	0.647059
r1-threeShot	NEI	0.682927	0.585366	0.90566	0.711111
r1-1CoT	Refutes	0.853659	0.75	0.461538	0.571429
r1-1CoT	Supports	0.780488	0.947368	0.409091	0.571429
r1-1CoT	NEI	0.682927	0.579545	0.962264	0.723404
r1-2CoT	Refutes	0.869919	0.8125	0.5	0.619048
r1-2CoT	Supports	0.804878	0.954545	0.477273	0.636364
r1-2CoT	NEI	0.691057	0.588235	0.943396	0.724638
r1-3CoT	Refutes	0.894309	0.882353	0.576923	0.697674
r1-3CoT	Supports	0.813008	0.92	0.522727	0.666667
r1-3CoT	NEI	0.739837	0.62963	0.962264	0.761194
gpt4-0shot	Supports	0.910569	0.923076	0.818182	0.867469
gpt4-0shot	Refutes	0.853658	0.666666	0.56	0.608695
gpt4-0shot	NEI	0.796747	0.730158	0.851851	0.786324
gpt4-1shot	Supports	0.853658	0.933333	0.636363	0.756756
gpt4-1shot	Refutes	0.796747	0.5	0.48	0.489795
gpt4-1shot	NEI	0.715447	0.637681	0.814814	0.715447
gpt4-2shot	Supports	0.853658	0.933333	0.636363	0.756756
gpt4-2shot	Refutes	0.837398	0.647058	0.44	0.523809

gpt4-2shot	NEI	0.756097	0.657894	0.925925	0.76923
gpt4-3shot	Supports	0.837398	0.9	0.613636	0.729729
gpt4-3shot	Refutes	0.829268	0.611111	0.44	0.511627
gpt4-3shot	NEI	0.731707	0.64	0.888888	0.744186
gpt4-1CoT	Supports	0.821138	0.866666	0.590909	0.702702
gpt4-1CoT	Refutes	0.878048	0.777777	0.56	0.651162
gpt4-1CoT	NEI	0.747967	0.653333	0.907407	0.759689
gpt4-2CoT	Supports	0.829268	0.870967	0.613636	0.72
gpt4-2CoT	Refutes	0.869918	0.764705	0.52	0.619047
gpt4-2CoT	NEI	0.747967	0.653333	0.907407	0.759689
gpt4-3CoT	Supports	0.821138	0.823529	0.636363	0.717948
gpt4-3CoT	Refutes	0.853658	0.666666	0.56	0.608695
gpt4-3CoT	NEI	0.739837	0.661764	0.833333	0.737704
gpt4o-0shot	Supports	0.853658	0.90625	0.659091	0.763157
gpt4o-0shot	Refutes	0.788617	0.484848	0.64	0.551724
gpt4o-0shot	NEI	0.674796	0.620689	0.666666	0.642857
gpt4o-1shot	Supports	0.853658	0.964285	0.613636	0.75
gpt4o-1shot	Refutes	0.837398	0.6	0.6	0.6
gpt4o-1shot	NEI	0.723577	0.642857	0.833333	0.725806
gpt4o-2shot	Supports	0.845528	1	0.568181	0.724637
gpt4o-2shot	Refutes	0.878048	0.75	0.6	0.666666
gpt4o-2shot	NEI	0.723577	0.628205	0.907407	0.742424
gpt4o-3shot	Supports	0.845528	1	0.568181	0.724637
gpt4o-3shot	Refutes	0.853658	0.64	0.64	0.64
gpt4o-3shot	NEI	0.731707	0.643835	0.87037	0.740157
gpt4o-1CoT	Supports	0.869918	0.911764	0.704545	0.794871
gpt4o-1CoT	Refutes	0.910569	0.888888	0.64	0.744186
gpt4o-1CoT	NEI	0.813008	0.718309	0.944444	0.816
gpt4o-2CoT	Supports	0.861788	0.90909	0.681818	0.77922
gpt4o-2CoT	Refutes	0.902439	0.933333	0.56	0.7
gpt4o-2CoT	NEI	0.780487	0.68	0.944444	0.790698
gpt4o-3CoT	Supports	0.861788	0.885714	0.704545	0.78481
gpt4o-3CoT	Refutes	0.894308	0.833333	0.6	0.697674
gpt4o-3CoT	NEI	0.772357	0.685714	0.888888	0.774193
gpt3.5-0shot	Supports	0.612903	0.483146	0.955556	0.641791
gpt3.5-0shot	Refutes	0.830645	0.666667	0.32	0.432432
gpt3.5-0shot	NEI	0.701613	0.869565	0.37037	0.519481
gpt3.5-1shot	Supports	0.707317	0.557143	0.886364	0.684211
gpt3.5-1shot	Refutes	0.764228	0.4	0.32	0.355556
gpt3.5-1shot	NEI	0.699187	0.757576	0.462963	0.574713
gpt3.5-2shot	Supports	0.682927	0.534247	0.886364	0.666667
gpt3.5-2shot	Refutes	0.788618	0.470588	0.32	0.380952
gpt3.5-2shot	NEI	0.747967	0.848485	0.518519	0.643678
gpt3.5-3shot	Supports	0.707317	0.557143	0.886364	0.684211
gpt3.5-3shot	Refutes	0.772358	0.411765	0.28	0.333333
gpt3.5-3shot	NEI	0.739837	0.805556	0.537037	0.644444
gpt3.5-1CoT	Supports	0.845528	0.790698	0.772727	0.781609
gpt3.5-1CoT	Refutes	0.861789	0.695652	0.615385	0.653061
gpt3.5-1CoT	NEI	0.772358	0.719298	0.773585	0.745455
gpt3.5-2CoT	Supports	0.869919	0.818182	0.818182	0.818182
gpt3.5-2CoT	Refutes	0.829268	0.666667	0.384615	0.487805
gpt3.5-2CoT	NEI	0.796748	0.71875	0.867925	0.786325
gpt3.5-3CoT	Supports	0.868852	0.804348	0.840909	0.822222
gpt3.5-3CoT	Refutes	0.860656	0.681818	0.6	0.638298
gpt3.5-3CoT	NEI	0.827869	0.796296	0.811321	0.803738