

Conflict aware abstention does not improve selective accuracy over confidence ranking in FEVER claim verification

Venil Pandian

itsmevenil@gmail.com

Annai Teresa College of Engineering

Research Article

Keywords: evidence-aware question answering, selective prediction, abstention, fact verification, natural language inference, scientific claim verification

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11110025/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

A QA system that sorts its evidence into supporting, conflicting, and neutral piles before it answers anything has an obvious attraction: when it abstains, it can say why. What is less obvious is whether that reason buys any measurable reliability over ordinary confidence-based selective prediction, and that is the question this paper asks. We evaluated a conflict-aware abstention rule on a frozen, balanced 1,200-case FEVER sample and on an external 209-case SciFact holdout drawn from the scientific literature. On FEVER the evidence-aware rule reaches 81.58% overall accuracy (95% CI 79.27–83.74), answers 53.58% of cases (95% CI 50.71–56.43), and is 90.05% correct among the cases it does answer (95% CI 87.56–92.22). A confidence-margin baseline, thresholded to exactly the same coverage, reaches 90.36% selective accuracy (95% CI 88.02–92.53), so we do not observe any selective-accuracy advantage for evidence-aware selection. Removing conflict abstention lifts overall accuracy to 84.00% while lowering selective accuracy to 89.15%, and the paired overall accuracy difference is significant (exact McNemar $p = 3.7 \times 10^{-9}$). SciFact is considerably harder: 41.63% overall accuracy, 49.76% coverage, 83.65% selective accuracy. A manual pass through 81 SUPPORT claims pushed into abstention turned up direct-support misses, evidence mismatch, paraphrastic support, terminology mapping, indirect or compositional reasoning, and qualified scientific support as the recurring reasons. Majority-class and lexical/length baselines further show substantial dataset-level predictive signal. What comes out of all this is a controlled negative finding. Explicit conflict-aware selection gives an interpretable basis for abstaining, but it does not beat confidence ranking at matched coverage, and it carries over to scientific evidence rather poorly.

1 Introduction

QA systems built on modern NLP now routinely pull answers out of large text collections, and in academic settings relevance alone is not enough: the answer should actually be supported by the evidence the system has in hand. Reliability gets difficult here. Retrieved evidence can be incomplete, neutral, or internally contradictory, and scientific material adds its own trouble on top, qualifications, specialized terminology, results that do not reduce to a directional yes or no, so an unsupported conclusion in that setting is a substantive error, not a cosmetic one.

Fact-verification benchmarks give this problem a name. FEVER formalizes it with SUPPORTS, REFUTES, and NOT ENOUGH INFO labels attached to verifiable claims and their textual evidence [1, 14], and SciFact extends the same idea to specialized literature, pairing expert-written claims with evidence-bearing scientific abstracts and rationales [2, 15]. Both make it possible to ask not just whether a prediction is right, but whether the evidence behind it actually justifies the prediction.

The established route to reliability is selective prediction: the system answers only when some confidence criterion is met. Confidence and calibration matter because model probabilities are not, in general, aligned with the probability of being correct [3]. The trouble with a confidence value is that it says nothing about the shape of the evidence behind it. A retrieved set of sentences can hold support

and conflict at the same time, or nothing but neutral material, and a single scalar score cannot separate those situations, whereas naming the condition explicitly at least gives the abstention a reason a bare number does not.

That is the appeal of evidence-aware abstention. Classify the evidence first; answer when the directional evidence is consistent; abstain when it is mixed or absent. The extra mechanism only means something, though, if it is measured against the simpler alternative under comparable conditions, and a comparison run at different coverage levels is not comparable, since selective accuracy can always be inflated by answering fewer cases. A fair evaluation has to match coverage.

There is also the question of whether a decision rule built on general-domain fact verification carries over at all to scientific literature, where claims come loaded with specialized terminology, numbers, hedged conclusions, abbreviations, and evidence spread across several sentences, so a rule that behaves well on Wikipedia-derived claims cannot be assumed to behave the same way on scientific evidence. Dataset quirks compound the problem: class imbalance or surface features can hand a system predictive power that has nothing to do with the reasoning it was supposed to be doing.

This paper investigates a conflict-aware evidence framework on a frozen balanced sample of 1,200 FEVER cases and an external 209-case SciFact holdout. The framework sorts evidence sentences into supporting, conflicting, and neutral groups and abstains when the detected evidence structure does not justify a directional answer. The evaluation covers conventional accuracy, coverage, selective accuracy, an exact confidence-matched comparison, confidence intervals, ablations, paired significance tests, and risk–coverage analysis, plus a manual failure analysis of the scientific SUPPORT claims that got routed to abstention when they should not have been.

The evidence-aware rule does not show a selective-accuracy advantage over confidence ranking. At an identical 53.58% coverage, selective accuracy is 90.05% for the evidence-aware rule and 90.36% for the confidence-margin baseline. On SciFact the difficulty gets worse, and the dominant error is SUPPORT mapped to abstention. So this paper is not claiming a new state of the art. It aims to determine when evidence-aware abstention actually helps, what it costs in the accuracy-coverage trade-off, and where it breaks under scientific-domain transfer.

The contributions are: (i) a controlled matched-coverage comparison of conflict-aware evidence selection against confidence-margin selection; (ii) an ablation and risk–coverage analysis of conflict handling and evidence sufficiency; (iii) an external SciFact evaluation with a manual breakdown of the 81 SUPPORT-to-abstention failures; and (iv) majority-class and lexical/length artifact controls that bear on how scientific-domain performance should be read.

2 Related Work

2.1 Fact Verification and Evidence-Based QA

Automatic fact verification couples evidence retrieval with inference over what was retrieved. FEVER set up the benchmark version of this problem with 185,445 claims derived from Wikipedia, labeled SUPPORTED, REFUTED, or NOT ENOUGH INFO, with evidence annotations for the verifiable ones [1]. It gave the field a shared setting for evaluating evidence retrieval and textual entailment inside one pipeline, and systems built on it showed that performance depends heavily on both retrieval quality and the inference model. That is why we evaluate evidence explicitly rather than looking at claim labels alone. The evidence-selection problem is still very much alive in recent work: Li et al. formalize minimal evidence groups for claim verification and report gains on WiCE and SciFact, while Zerong et al. survey 198 recent claim-verification studies and identify persistent limitations in evidence annotation, pipeline design, and claim decomposition [4, 5].

Scientific verification is a more demanding version of the same problem, since claims lean on specialized terminology and often require interpretation of research findings rather than pattern matching. Wadden et al. built SciFact around that difficulty: expert-written scientific claims, evidence-bearing abstracts, and rationales [2]. Their finding that domain adaptation improves scientific verification over systems trained only on general-domain data is exactly why we do not assume a mechanism built on FEVER evidence transfers to scientific literature without checking. More recent scientific-verification benchmarks broaden this setting: SciVer evaluates foundation models on expert-annotated multimodal scientific claims, and Ho et al. show that correct claim labels do not necessarily imply faithful evidence localization in scientific tables [6, 7].

2.2 Confidence, Calibration, and Selective Prediction

Selective prediction lets a model abstain on whatever it is unsure about, and the usual way to describe the resulting trade-off is coverage and risk: coverage is the share of examples the system actually answers, risk the error rate among those. The split matters for reliability work for a simple reason, plain accuracy cannot tell a system that answers everything apart from one that quietly skips the hard cases.

Calibration studies in question answering have shown that model confidence is not automatically a reliable estimate of correctness. Jiang et al. examined calibration in language-model QA and reported that model probabilities can be poorly calibrated, which pushed the field toward post-hoc and learned fixes for the relation between confidence and correctness [3]. A confidence-based abstention mechanism is therefore an important baseline for evaluating any evidence-aware selection method, and we compare against a confidence-margin baseline at exactly matched coverage rather than against unmatched operating points, for the reasons given above. This concern is increasingly visible in contemporary verification research: ClaimDB reports that both closed- and open-source models still struggle to abstain when the available evidence is insufficient [8].

2.3 Evidence Conflict and Scientific-Domain Transfer

Evidence-aware reasoning differs from confidence-only selection in that the abstention decision can be tied to a detected relation between the claim and the individual evidence items, rather than to a score. In our formulation, evidence is split into supporting, conflicting, and neutral groups, so a claim ends up

unresolved either when support and conflict coexist or when nothing directional turns up at all. The rule is interpretable in the limited sense that each abstention corresponds to an observable evidence condition, not just a low number. Recent systems make evidence structure explicit in related ways: SciTrue emphasizes source-level traceability for scientific claims, and HARIS coordinates iterative evidence search with multi-step reasoning for claim verification [9, 10].

Interpretability, however, does not by itself establish better predictive performance. Scientific evidence supports a claim through paraphrase, terminology mapping, indirect composition, numerical relationships, or qualified statements, and any of those can lead a general-domain NLI model to classify the evidence as neutral even when the underlying statement does support the claim. That is why we pair the benchmark numbers with a manual failure taxonomy and dataset-artifact controls rather than reporting accuracy alone. Other recent approaches attack related weaknesses through domain-oriented synthetic verification data and jointly trained retrieval-reasoning workflows. SYNTHVERIFY targets zero-shot domain transfer, while Veri-R1 trains planning, retrieval, and reasoning jointly for faithful online verification [11, 12]. A 2026 survey of LLM-based claim verification further highlights retrieval, prompting, fine-tuning, and dataset choice as major design dimensions in current systems [13].

3 Experiments

3.1 Research Questions

The experiments address three questions: whether explicit evidence-aware abstention improves selective accuracy over confidence-based selection at matched coverage; how conflict handling and evidence sufficiency change the accuracy–coverage trade-off; and how well the decision rule transfers from FEVER to scientific claim verification in SciFact.

3.2 Evidence-Aware Decision Framework

Each claim's evidence sentences pass through a natural-language inference component based on the cross-encoder/nli-deberta-v3-small checkpoint. The model's entailment, contradiction, and neutral outputs are mapped to SUPPORT, CONFLICT, and NEUTRAL evidence relations, respectively. Write the resulting evidence counts as S, C, and N. The primary conflict-aware rule predicts SUPPORTS when $S > 0$ and $C = 0$, predicts REFUTES when $C > 0$ and $S = 0$, and abstains as NOT ENOUGH INFO whenever both directional classes appear or neither does.

A second quantity, the evidence-sufficiency score $(S + C)/(S + C + N)$, is computed alongside, but only for diagnostics and post-hoc threshold analysis. It plays no part in the final prediction of the primary 1,200-case FEVER run, and we keep that distinction visible throughout so the system's behaviour never gets credited to a threshold it did not use.

3.3 FEVER Evaluation

The primary in-domain evaluation runs on a frozen FEVER sample of 1,200 cases, balanced across 400 SUPPORTS, 400 REFUTES, and 400 NOT ENOUGH INFO. The frozen balanced subset was used to obtain a substantially larger evaluation than the development pilot while keeping the three FEVER labels equally represented and avoiding a majority-class advantage in the primary analysis. The subset was fixed before the reported large-scale analyses, so that subsequent ablations and matched-coverage comparisons were evaluated on the same cases.

The forced-decision baseline aggregates the NLI support and conflict scores and picks whichever direction scores higher. A confidence margin comes from the gap between the winning score and the runner-up. For the matched-coverage comparison, we rank the baseline's predictions by that margin and keep the top 643, which is exactly the number the evidence-aware rule answers.

3.4 Ablation and Risk–Coverage Analysis

Four decision configurations are examined: the primary conflict-aware rule; a no-conflict-abstention version that forces a directional call whenever support and conflict both appear; a rule requiring at least two informative evidence items; and a post-hoc evidence-coverage threshold set at 0.50. Alongside these, a full confidence risk–coverage sweep is generated from every unique confidence margin, giving 371 operating points. The evidence-sufficiency score yields 22 post-hoc operating points of its own, and the actual evidence-aware system sits plotted at its true, single operating point rather than anywhere along either sweep.

3.5 SciFact External Evaluation

The external evaluation draws on SciFact. After consolidation of the prepared holdout into unique claim-document cases, the frozen external set contains 209 cases: 138 SUPPORT and 71 CONTRADICT. This consolidated holdout was retained as an external scientific-domain test rather than rebalanced, so that the transfer analysis preserves its observed label distribution.

Because SciFact skews toward SUPPORT, we add two artifact controls. A majority SUPPORT baseline measures what class prevalence alone is worth, and a lexical/length feature baseline is evaluated under five-fold cross-validation. Neither is a matched selective baseline; they exist to show how much conventional accuracy can be extracted from the dataset's own surface characteristics rather than from anything the system is actually doing.

3.6 Evaluation Metrics and Statistical Analysis

Overall accuracy is the conventional classification number. Coverage is the share of cases getting a directional SUPPORTS or REFUTES call, selective accuracy the share correct among those. We report exact binomial 95% confidence intervals for overall accuracy and coverage, and bootstrap selective-accuracy intervals from 10,000 resamples. Paired overall comparisons use exact McNemar tests, and risk is defined as 1 minus selective accuracy. The matched-coverage comparison is reported descriptively: we make no equivalence or non-inferiority claim, since neither test was pre-specified.

4 Results

4.1 Main FEVER Performance

Table 1 gives the main FEVER results. The evidence-aware rule correctly classifies 979 of 1,200 cases, 81.58% overall accuracy (95% CI 79.27–83.74). It provides directional answers for 643 cases, 53.58% coverage (95% CI 50.71–56.43), at 90.05% selective accuracy (95% CI 87.56–92.22). The plain forced-decision baseline, which never abstains, reaches 91.83% overall accuracy (95% CI 90.14–93.32). In the paired comparison, 123 cases come out correct only for the forced-decision baseline, and none are correct only for the evidence-aware system. So conflict-aware abstention should not be read as an improvement to conventional FEVER accuracy.

Table 1
Main performance results on the frozen balanced FEVER evaluation.

System	Accuracy	Coverage	Sel. Acc.
Forced-decision baseline	91.83%	N/A (non-selective)	—
Evidence-aware	81.58%	53.58%	90.05%
Confidence baseline (matched)	—	53.58%	90.36%

4.2 Matched-Coverage Confidence Comparison

At exactly 643 answered cases, the confidence-margin baseline achieves 90.36% selective accuracy (95% CI 88.02–92.53), compared with 90.05% (95% CI 87.56–92.22) for the evidence-aware rule. The observed difference is 0.31 percentage points, in favour of confidence ranking. The two methods also pick nearly the same cases: 626 are answered by both, 17 only by the evidence-aware rule, 17 only by confidence selection, and 540 are rejected by both, an answer-set overlap of 97.36%. We observe no selective-accuracy advantage for the evidence-aware mechanism at matched coverage.

4.3 Ablation Results

Table 2
Ablation results on the frozen balanced FEVER evaluation.

Configuration	Accuracy	Coverage	Sel. Acc.
Conflict-aware	81.58%	53.58%	90.05%
No conflict abstention	84.00%	56.83%	89.15%
Min. 2 useful evidence	42.17%	9.58%	92.17%
Evidence coverage ≥ 0.50	80.33%	52.00%	90.38%

Table 2 gives the ablation results. Removing conflict abstention raises overall accuracy from 81.58% to 84.00% and coverage from 53.58% to 56.83%, while selective accuracy falls from 90.05% to 89.15%.

Thirty-nine cases in the frozen sample carry both supporting and conflicting signals at once; their gold labels are 33 SUPPORTS and 6 REFUTES. Forcing a directional prediction on these recovers 29 correct answers and introduces 10 directional errors. The paired comparison gives 29 improvements for the no-conflict rule and zero in the other direction, an exact McNemar p-value of 3.7×10^{-9} . At the other extreme, the minimum-two-evidence rule achieves the highest selective accuracy of any configuration, 92.17%, but its coverage collapses to 9.58%, which is the cost of excessive conservatism.

4.4 Risk–Coverage Behaviour

The confidence sweep produces 371 operating points and the post-hoc sufficiency sweep 22, with the actual evidence-aware system sitting at 53.58% coverage and 9.95% selective risk. Reliability therefore cannot be interpreted independently of coverage. We treat the sufficiency curve as diagnostic only, since the primary rule's abstention decision never actually used that score as a threshold.

4.5 SciFact External Results

Table 3
External evaluation results on the SciFact holdout.

Configuration	Accuracy	Coverage	Sel. Acc.
Sentence evidence	41.63%	49.76%	83.65%
Joint evidence	42.11%	50.24%	83.81%

Table 3 gives the SciFact transfer results. Sentence-level evaluation correctly classifies 87 of 209 cases and abstains on 105. Overall accuracy is 41.63% (95% CI 34.87–48.63), coverage 49.76% (95% CI 42.79–56.74), selective accuracy 83.65% (95% CI 75.96–90.38). Joint evidence correctly classifies one additional case, for 42.11% overall accuracy, 50.24% coverage, and 83.81% selective accuracy, and the paired comparison between the sentence and joint versions is McNemar $p = 1.00$: simply concatenating evidence does not materially change the observed performance.

4.6 Failure Analysis

Table 4
Manual taxonomy of SciFact SUPPORT-to-abstention failures.

SUPPORT → Abstain category	Cases
Direct Support Missed (DSE)	28
Evidence Mismatch (EVM)	14
Paraphrastic Support (PAR)	12
Terminology/Concept Mapping (TERM)	12
Indirect/Compositional (IND)	10
Qualified Scientific Support (QUAL)	5
Total	81

Table 4 gives the manual failure taxonomy. SUPPORT-to-abstention is the dominant error on SciFact. Of the 81 cases we hand-annotated, 28 are direct-support misses, 14 evidence mismatch, 12 paraphrastic support, 12 terminology or concept mapping, 10 indirect or compositional support, and 5 qualified scientific support. Secondary tags catch statistical expressions in 16 of them, numerical reasoning in 6, abbreviations in 6, multi-sentence reasoning in 4, and hedging in 3. The pattern indicates that scientific support is frequently missed even when the available evidence is directionally consistent with the claim.

4.7 Dataset Artifact Controls

A majority SUPPORT baseline achieves 66.03% conventional accuracy on the SciFact holdout by predicting only the majority label, and a lexical/length baseline achieves 53.59% five-fold accuracy using surface features alone. Neither number should be set against the evidence-aware system's 41.63% as if it measured the same thing, since the majority classifier answers every case while the evidence-aware system answers roughly half. What the controls do show is that label imbalance and surface features carry real predictive weight in this dataset, and any reading of the scientific verification numbers needs to account for that.

5 Discussion

The experiments yield a negative but informative result. The explicit evidence-aware mechanism does not improve selective accuracy over confidence ranking at matched coverage: the two methods select nearly identical answer sets and differ by 0.31 percentage points. In this FEVER setting, then, the extra evidence structure cannot be justified on predictive-performance grounds.

Still, the conflict-aware rule explains something a confidence threshold does not. A confidence threshold only says the model was not sure enough; the evidence-aware rule can point to the actual condition, co-occurring support and conflict, or nothing directional detected at all. That interpretability comes at

almost no cost in selective accuracy at the matched point, though we are not claiming formal equivalence.

The ablation makes the cost of conflict handling explicit. FEVER's gold labels demand a single final relation, and every one of the 39 mixed-signal cases in the frozen sample has a directional gold label, so abstaining on them is a reduction in conventional accuracy by construction. Forcing a call on those cases recovers 29 correct answers and introduces 10 new errors, and which rule you actually want depends on whether the application cares more about benchmark accuracy or about not handing out unsupported directional answers.

SciFact exhibits a different failure pattern. The problem there is not too much conflict detection but too little recognition of support in the first place: direct support, paraphrase, terminology mapping, compositional reasoning, and qualified statements frequently result in abstention. Joint-evidence concatenation produces only a negligible change, which suggests fragmentation across sentences is unlikely to be the primary cause; domain-adapted inference or something closer to scientific semantic representation appears to be a more promising direction.

The artifact controls also affect how the SciFact results should be read. A majority classifier gets a high score primarily because of class imbalance, and lexical and length features alone carry non-trivial predictive power. None of this makes SciFact an unsuitable testbed. It means raw accuracy can mislead when systems operate at different coverage, or when the dataset itself is doing some of the work.

5.1 Limitations

Several limitations apply. The primary FEVER evaluation uses 1,200 frozen cases rather than the full benchmark; the sample is balanced and large enough for the statistical analyses reported, but a bigger evaluation would tighten precision. The evidence-aware rule inherits the errors of the underlying NLI model, and evidence misclassification translates directly into abstention behaviour. The SciFact holdout has only 209 consolidated claim–document cases and is imbalanced toward SUPPORT. The manual taxonomy covers SUPPORT-to-abstention errors only and is not a complete taxonomy of every error type.

The matched-coverage confidence comparison is descriptive, not an equivalence or non-inferiority test. The close selective accuracies should be interpreted as an absence of an observed advantage, not as proof that the two methods are statistically the same thing.

The reported transfer results also depend on the chosen NLI backbone. Prompt engineering, domain-specific fine-tuning, or a larger language model could improve recognition of scientific paraphrases, terminology, quantitative relations, and qualified statements, and could therefore change the observed abstention pattern. Evaluating such alternatives is left for future work rather than being inferred from the present experiments.

6 Conclusion

The question we set out to answer was whether explicit evidence-aware abstention gives a measurable advantage over confidence-based selective prediction for academic QA. The present results show no such advantage in selective accuracy. On the balanced 1,200-case FEVER evaluation, the conflict-aware rule reaches 81.58% overall accuracy, 53.58% coverage, and 90.05% selective accuracy; at that exact coverage, confidence-margin selection reaches 90.36%. The ablation trades conventional accuracy for a small gain in selective accuracy, which is the accuracy–coverage–reliability trade-off in practice. Transfer to SciFact is considerably harder, with SUPPORT mapped to abstention as the dominant failure, and our manual analysis ties these failures to direct support misses, evidence mismatch, paraphrase, terminology mapping, indirect reasoning, and qualified scientific statements. The artifact controls add a further caution: class imbalance and simple lexical and length features already carry substantial predictive signal in the SciFact holdout. Taken together, these findings put evidence-aware abstention as useful mainly as an interpretable, evidence-level way of explaining an abstention decision rather than as a route to better prediction, and they argue for matched-coverage baselines, confidence intervals, artifact controls, and manual failure analysis as standard practice whenever a reliability mechanism for academic QA is evaluated.

Reproducibility Statement

The experiments were run on frozen evaluation sets, and the reported analyses use fixed case sets across the main evaluation, matched-coverage comparison, ablations, and statistical tests. The implementation records the evaluation outputs and the SciFact SUPPORT-to-abstention annotations used in the failure analysis. No public repository URL is claimed in this manuscript; code and evaluation artifacts can be released with the final publication subject to dataset licensing requirements.

Declarations

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used ChatGPT (OpenAI) to support manuscript organization, language refinement, and readability. The author reviewed and edited the content, verified the reported results and references, and takes full responsibility for the content of the publication.

Acknowledgments

Not applicable.

Author contributions

Venil Pandian conceived and designed the study, implemented the experiments, analysed and interpreted the results, and wrote and revised the manuscript.

Clinical trial number

Not applicable.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Conflicts of interest

The author declares no conflicts of interest.

Ethical approval

Not applicable. This computational study analysed publicly available benchmark datasets and did not recruit human participants or involve live animals.

Ethical Approval and accordance

Not applicable.

Data availability statements

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

The study uses publicly available FEVER and SciFact datasets. FEVER is available at <https://fever.ai/dataset/fever.html> and SciFact at <https://github.com/allenai/scifact>. The frozen evaluation subsets, case-level predictions, statistical outputs, and manual annotations generated for this study are available from the corresponding author on reasonable request.

References

1. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. Proceedings of NAACL-HLT, pp. 809–819.
2. D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. Proceedings of EMNLP, pp. 7534–7550.

3. Z. Jiang, J. Araki, H. Ding, and G. Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.
4. X. Li, S. Chen, R. Kapadia, J. Ouyang, and F. Zhang. 2025. Minimal Evidence Group Identification for Claim Verification. *Proceedings of the Fourth Workshop on Trustworthy Natural Language Processing (TrustNLP 2025)*, pp. 103–111. doi:10.18653/v1/2025.trustnlp-main.8.
5. Z. Zerong, C. Li, X. Liu, J.-h. Chen, and F. Xia. 2025. A Systematic Survey of Claim Verification: Corpora, Systems, and Case Studies. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21452–21474. doi:10.18653/v1/2025.findings-emnlp.1170.
6. C. Wang, Y. Shen, Z. Kuang, A. Cohan, and Y. Zhao. 2025. SciVer: Evaluating Foundation Models for Multimodal Scientific Claim Verification. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8562–8579. doi:10.18653/v1/2025.acl-long.420.
7. X. Ho, S. Kumar, Y.-A. Wu, F. Boudin, A. Takasu, and A. Aizawa. 2025. Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2509–2517. doi:10.18653/v1/2025.findings-emnlp.135.
8. M. Theologitis, P. P. S. Dammu, C. Shah, and D. Suci. 2026. ClaimDB: A Fact Verification Benchmark over Large Structured Data. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34428–34451. doi:10.18653/v1/2026.acl-long.1589.
9. N. Tan, M. Li, and M. Gahegan. 2026. SciTrue: Evidence-Grounded Claim Verification in Science. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 397–406. doi:10.18653/v1/2026.eacl-demo.27.
10. Q. Hu, Q. Long, and W. Wang. 2026. Coordinating Search-Informed Reasoning and Reasoning-Guided Search in Claim Verification. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2564–2585. doi:10.18653/v1/2026.acl-long.119.
11. R. Zhao and J. Flanigan. 2025. SYNTHVERIFY: Enhancing Zero-Shot Claim Verification through Step-by-Step Synthetic Data Generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3257–3274. doi:10.18653/v1/2025.findings-acl.170.
12. Q. He, C. Qian, X. Chen, B. He, Y. R. Fung, and H. Ji. 2026. Veri-R1: Toward Precise and Faithful Claim Verification via Online Reinforcement Learning. *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20372–20391. doi:10.18653/v1/2026.findings-acl.1019.
13. A. Dmonte, R. R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein. 2026. Claim Verification in the Age of Large Language Models: A Survey. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 15–29. doi:10.18653/v1/2026.acl-srw.2.

14. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER dataset. FEVER.
<https://fever.ai/dataset/fever.html>
15. D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. 2020. SciFact dataset. Allen Institute for AI. <https://github.com/allenai/scifact>