Gold Label Errors in the SciFact Benchmark: An LLM-Assisted Annotation Audit

Julien Sylvestre
BXCT Partners Ltd
js@bxct.co.uk

Abstract

SciFact is a widely-used benchmark for scientific claim verification (645 citations, included in the BEIR evaluation suite). We present, to our knowledge, the first systematic annotation audit of its development and training sets, combining automated screening with a small language model ($0.11 in API fees) and exhaustive manual verification against source publications. We identify 11 gold-label errors in the development set (5.3%, 95% CI 2.7–9.2%, of 209 audited claim–document pairs) and 13 in the training set (2.3%, 95% CI 1.2–3.9%, of 564 audited pairs). The dev errors exhibit a directional asymmetry—9 of 11 mislabel a claim as SUPPORT (one-sided binomial p=0.033, two-sided p=0.065)—and fall into four recurring types. Correcting the dev labels raises binary macro-F1 by 1.7–3.8 points across GPT-5.4 (mini, nano) and Claude Haiku 4.5; gains are larger in 3-way evaluation when mislabeled evidence is recast as NEI (e.g., +9.2 with Haiku 4.5). The binary range is comparable in magnitude to inter-system margins on the SciFact leaderboard. A simple claim-only probe with Haiku 4.5 does not support label memorization as the main explanation for these gains. We release corrected annotations and a blind annotator packet, and recommend that benchmark users prefer the corrected release going forward.

1 Introduction

Benchmark quality is foundational to NLP progress: if gold labels contain systematic errors, model comparisons become unreliable and optimization targets become misaligned. This concern is especially acute for small, expert-annotated datasets where a handful of label errors can shift rankings by several F1 points.

SciFact (Wadden et al., 2020), introduced at EMNLP 2020 and subsequently incorporated into the BEIR retrieval evaluation suite (Thakur et al., 2021), is the standard benchmark for scientific claim verification (645 citations on Semantic Scholar as of April 2026). The task requires systems to retrieve relevant abstracts and classify scientific claims as SUPPORT, CONTRADICT, or NOT ENOUGH INFO (NEI) given evidence sentences from PubMed abstracts. The dataset contains 300 dev claims, 809 training claims, and 300 hidden test claims; of these, 209 dev and 564 training claim–document pairs have gold evidence annotations that we audit here. Gold labels were assigned by domain experts (reported Cohen's $\kappa = 0.75$). Despite wide adoption—over 600 citations and use in recent work at EMNLP, SIGIR, and JAMIA Open—to our knowledge no prior study has audited the quality of SciFact's gold annotations. A review of the dataset's GitHub repository, Hugging Face page, and citing literature (conducted March 2026) reveals zero reported label corrections, no errata, and no corrected release in six years of use. This is notable given that the annotated pairs are small enough to verify manually in a few hours and that the original inter-annotator agreement ($\kappa = 0.75$) leaves room for disagreement.

We contribute: (1) identification of 11 gold label errors in the dev set (5.3%, 95% CI 2.7–9.2%) and 13 in the training set (2.3%, 95% CI 1.2–3.9%), showing errors contaminate both model training and evaluation; (2) a taxonomy of error types and analysis of directional asymmetry, with statistical framing; (3) cross-model verification of the F1-impact and a contamination probe using a non-GPT model family (Claude Haiku 4.5); and (4) an efficient, replicable audit methodology costing $0.11 in API fees for the automated screening stage.

2 Related Work

Annotation errors in benchmarks. Northcutt et al. (2021) found 3–6% label errors across 10 pop-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 97–103

ular benchmarks (including ImageNet and MNIST) and showed that correcting these errors changes model rankings. More recently, Vendrow et al. (2025) created "platinum" corrected versions of 15 benchmarks and found that over 50% of model errors on the originals were attributable to label noise—GSM8K alone had ∼5% problematic questions. Nahum et al. (2025) used LLM ensembles to detect 6–21% error rates across factual consistency benchmarks, showing that many reported LLM "mistakes" are correct responses to mislabeled examples. In NER, Rucker and Akbik (2023) re-annotated CoNLL-03 and found 7% of labels were incorrect; nearly half of state-of-the-art model "errors" were actually correct predictions penalized by noisy ground truth.

Annotation artifacts and disagreement. Gururangan et al. (2018) identified annotation artifacts in SNLI; Pavlick and Kwiatkowski (2019) and Nie et al. (2020) showed many NLI items lack a single correct label. Weber-Genzel et al. (2024) separated genuine label variation from error in MNLI. Swayamdipta et al. (2020) found that hard-to-learn instances often correspond to labeling errors, and Klie et al. (2023) surveyed 18 error detection methods, finding the field lacks standardized evaluation.

SciFact. Wadden et al. (2022b) showed 15+ F1 drops when expanding SciFact to open-domain retrieval, and Košprdić et al. (2024) found fine-tuned DeBERTa outperformed GPT-4 zero-shot by ∼7%. Despite this active research community, to our knowledge no prior work has audited SciFact's gold labels. The dataset's GitHub repository contains 27 issues—none concerning annotation quality—and neither the original authors nor any of the 600+ citing papers have published corrections or errata.

3 Method

3.1 Stage 1: Automated LLM Screening

We processed all 209 dev claim–document pairs using GPT-5.4-mini (temperature 0.3) with a structured audit prompt. For each pair, the model received the claim text, the gold label, the gold rationale sentences, and the full abstract. The model returned structured JSON: assessment (AGREE/DISAGREE/DEBATABLE), reasoning, error type (from a predefined taxonomy), and suggested correct label.

The automated screening flagged 57 claims (27.3%): 43 as DISAGREE and 14 as DEBATABLE. This stage cost $0.11 in API fees (approximately 440K tokens at March 2026 pricing; cached-input rates applied).

3.2 Stage 2: Manual Expert Verification

The automated screening has a high false-positive rate: the LLM frequently confuses claim direction with evidence direction (Section 4.3). Each of the 57 mini-flagged disagreements was therefore adjudicated by a single annotator working interactively with a *frontier* LLM (GPT-5.4) in chat mode: the frontier model surfaced its own reading of the claim and evidence, the annotator probed and challenged it, and the annotator made the final verdict. This human-in-the-loop protocol provides an additional model-based second opinion on every disputed pair while keeping final adjudication with the human; the screening model (mini) and the verification model (frontier) are different scales of the same family, and the human breaks ties. The 152 non-flagged claims were exhaustively reviewed by the same annotator without the chat-mode loop, since the screen had already implicitly endorsed the gold label. Corrected labels are determined solely by the assigned evidence paper; later literature is consulted only for context (see footnotes in §4.1).

Of the 57 flagged claims, 8 were ultimately confirmed as annotation errors, 8 remained debatable but defensible, and 41 were false alarms (after a second-pass review on the borderline cases: 2 claims initially classified as debatable were upgraded to confirmed errors and are listed in the appendix). Of the 152 non-flagged claims, exhaustive review identified 3 additional errors that the automated screening missed entirely (rating them AGREE), bringing the total to 11.

3.3 Source-Paper Verification

All 11 dev errors were verified against the original source publications (retrieved via PubMed), confirming that each correction is consistent with the assigned evidence paper's own findings. This verification is independent of the benchmark labels, not of the annotator: a second independent rater was not used (see Limitations).

4 Results

4.1 Confirmed Annotation Errors

Table 1 presents the 11 confirmed errors, falling into four types: **(1) Label reversal** (4 cases): gold

label directly contradicts the evidence. ID 593 claims incidence "decreased by 10%" but the rationale states "did not change over time." ID 879 is labeled CONTRADICT, yet the abstract concludes "the large majority of lincRNAs do not function through encoded proteins" (sentence 6), agreeing with the claim.[1] ID 1385's rationale states "enhancing cSMAC formation *reduced* stimulatory capacity," the opposite of the claim's "enhances," yet gold says SUPPORT.[2] ID 808's rationale is a methodology sentence; the abstract's actual finding—"does not rely on cis-acting sequences" (sentence 3)—directly contradicts the claim, yet gold says SUPPORT. **(2) Outcome mismatch** (3 cases): claim and evidence address different clinical outcomes. ID 343 claims "bleeding events" but the rationale reports cardiovascular mortality, MI, stroke, and CHF—bleeding is never mentioned. ID 1368 claims effect on "term of delivery" (preterm birth) but the rationale reports gestational diabetes, pre-eclampsia, and SGA—preterm birth is absent from the abstract. ID 770 claims "reduced efficacy and lower quality of life" for single-agent fluoropyrimidines vs. oxaliplatin, but PFS was non-significant (p=0.07) and the trial's QoL comparison was between two fluoropyrimidines (fluorouracil vs. capecitabine), not single-agent vs. combination. **(3) Entity mismatch** (3 cases): claim and evidence refer to different biological entities. ID 1216 claims cleavage in "human beta cells" but the paper reports exclusively on mouse Tmem27 (Tcf1$^{-/-}$ mice, transgenic mice).[3] ID 1274 attributes T6SS mechanisms to *E. coli* but the paper demonstrates them in *V. cholerae* and *A. baylyi*—organisms from different taxonomic orders. ID 847 claims "new drugs" fail to penetrate necrotic lesions; the rationale cites rifampicin (a 1960s drug, not new) accumulating in caseum, while the same abstract reports that moxifloxacin (a newer drug) "does not diffuse well." **(4) Overgeneralization** (1 case): ID 208 claims "CHEK2 is not associated with breast cancer," but the paper's abstract states "Previous investigation has established a role for the CHEK2 gene in breast cancer aetiology" (sentence 1) and reports OR 2.26 for the rare 1100delC mutation (sentence 8).

ID	Gold→Corr.	Type	Key issue
593	S→C	Reversal	"decreased 10%" vs "did not change"
879	C→S	Reversal	Abstract [6] agrees with claim
1385	S→C	Reversal	"enhances" vs "reduced capacity"
808	S→C	Reversal	Abstract [3] contradicts claim
343	S→N	Outcome	CV events, not bleeding
1368	S→N	Outcome	Diabetes/SGA, not preterm birth
770	S→N	Outcome	PFS p=0.07; QoL compared wrong arms
1216	S→N	Entity	Mouse study, claim says "human"
1274	S→N	Entity	*V. cholerae* study, claim says *E. coli*
847	C→N	Entity	Old drug in rationale; claim says "new"
208	S→C	Overgen.	Abstract [1,8] contradict claim

Table 1: The 11 confirmed gold label errors (n=209 dev claim–document pairs). S=SUPPORT, C=CONTRADICT, N=NEI. 9/11 are mislabeled SUPPORT. One additional per-document error (ID 597, mortality≠incidence) is not shown.

4.2 Impact on System Evaluation

We ran two zero-shot configurations on all 209 dev claims (oracle setting) and computed macro F1 before and after correcting the 11 errors. GPT-5.4-mini gained +3.8 F1 (84.9→88.7); GPT-5.4-nano gained +1.7 (79.5→81.2). For context, the top systems on the SciFact test-set leaderboard are separated by just 1.25 F1 points (MultiVerS at 72.5 vs. ARSJoint at 71.2; Wadden et al., 2022a; Zhang et al., 2021), and dev-set gaps among top models range from 0.7 to 2.7 F1 (Košprdić et al., 2024). The correction-induced shifts of 1.7–3.8 F1 are of the same order as these gaps, introducing noise that can plausibly perturb leaderboard ordering—though our zero-shot estimates do not establish reordering of fine-tuned systems, whose predictions we do not have. The stronger model benefits more because the mislabeled claims were among its "errors"—it correctly rejected claims that the

[1] Subsequent studies have identified functional micropeptides encoded by some lncRNAs, complicating the blanket claim. However, the *cited paper's own conclusion* supports the claim; the gold label is wrong relative to the assigned evidence regardless of later findings.

[2] A follow-up by the same group (Cemerski et al., 2008) refined this finding, reporting that cSMAC can enhance signaling from weak agonists under certain conditions. Nevertheless, the *assigned evidence paper* unambiguously states "reduced," making the SUPPORT label incorrect for this claim–evidence pair.

[3] Later work confirmed TMEM27 cleavage in human beta cells (Esterházy et al., 2011). The claim is likely true, but the *assigned evidence* is a mouse-only study and cannot support the species-specific claim.

gold standard incorrectly accepted.

Cross-model verification. To address the concern that the F1 gain might reflect GPT-family preferences leaking into our corrections, we reran the same evaluation with Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) using identical inputs (claim, paper title, full abstract, gold rationale), structured tool-use output, and prompt caching. On the original labels, Haiku reaches a binary macro F1 of 92.77; applying the corrections raises it to 95.75—a $+2.98$ F1 gain, directionally consistent with GPT-5.4-mini's $+3.8$. The 3-way (S/C/NEI) macro F1 gain is larger ($61.85 \rightarrow 71.09$, $+9.24$), reflecting the 6 pairs whose evidence licenses NEI rather than a binary verdict. Across both model families and both binary and 3-way evaluation, every configuration gains F1 from the corrections; the effect is not confined to the GPT family in our tests.

Contamination probe. We then asked Haiku 4.5 to classify the same 209 dev claims given *only* the claim text—no abstract, no rationale, no title. Haiku reaches 70.81% accuracy, only 4.78 points above the 66.03% majority-class baseline (SUPPORT), with 18% of predictions being NEI. A model that had memorized SciFact's binary gold labels would be expected to commit to a side; the small gap above majority class is consistent with claim-plausibility reasoning rather than label memorization. This argues against dataset contamination as the source of the F1 gain.

4.3 Automated Auditor Performance

Of 57 flagged claims, 8 were errors (14% precision). Most false alarms were *direction confusion*: the auditor flagged CONTRADICT labels as wrong when the claim reversed the evidence, not recognizing that CONTRADICT is correct *because* the claim reverses the evidence. Crucially, three errors (IDs 1385, 770, 1216) were missed entirely—the LLM rated them AGREE despite clear mismatches. ID 1385 is a direct claim–evidence contradiction; ID 770 requires checking statistical significance and QoL directionality; ID 1216 requires noticing a species mismatch (mouse vs. human). The auditor's recall was 8/11 (73%); automated screening alone is insufficient.

4.4 Per-Document Error

One additional per-document error (not counted above): ID 597's third evidence document discusses cervical cancer *mortality* but is labeled SUPPORT for an *incidence* claim.

4.5 Training Set Audit

To assess whether errors also contaminate model training, we manually audited all 564 claim–document pairs in the training set. We identified 13 errors (2.3%, 95% CI 1.2–3.9%), exhibiting the same error types as the dev set; the dev and train CIs overlap, so we do not claim that dev is strictly noisier than train. Seven are clear label reversals: e.g., ID 164 labels SUPPORT for "bariatric surgery increases colorectal cancer" when the rationale states "no association was detected between bariatric surgery and … cancer"; ID 933 labels SUPPORT for "increased morphine use" when the evidence reports significantly *less* morphine use ($P{=}.03$). Three are entity mismatches: e.g., ID 621 claims "Individuals with Alzheimers" but the study explicitly recruited adults who "did not meet criteria for dementia"; ID 657 attributes a function to signal peptide peptidase but the evidence is about RHBDL4, a different enzyme. One is a causal misread: ID 416 claims APOE4 carriers "have longer lifetime exposure to estrogen due to an increased reproductive period," but the evidence only shows that APOE4 modifies the effect of reproductive period length on dementia risk (effect modification), not that APOE4 causes longer reproductive periods.[4] Two involve modality or numerical mismatches between claim and evidence (ID 1145 treats a microsimulation projection as observed data; ID 1297 claims 30 million cases when the evidence reports $\sim$15 million hospitalizations). All 13 training errors with claim IDs, corrected labels, rationale indices, and one-sentence justifications are provided in the supplementary materials.[5] The lower error rate in training (2.3% vs. 5.3% in dev) may reflect differences in annotation workflow, but the key finding is that errors are present in *both* splits: models fine-tuned on SciFact are optimized toward wrong labels during training, tuned against wrong labels during development, and—if the hidden test set has similar rates—evaluated against wrong labels at submission. This means label noise

[4] A 1.8-year delay in menopause among APOE3/4 carriers has been reported in one Chinese cohort (Meng et al., 2012), but findings are mixed across populations, and the *assigned evidence paper* reports only effect modification, not a causal relationship.

[5] Full training-set error table in supplementary materials and at `https://github.com/Kefez/scifact-audit-bionlp2026`.

can affect every stage of the SciFact pipeline, from retrieval model training (learning which documents to retrieve for a claim) to inference model training (learning which label to assign) to final evaluation.

5 Discussion

Our 5.3% dev error rate (95% CI 2.7–9.2%) falls within the 3–7% range reported for other benchmarks (Northcutt et al., 2021; Rucker and Akbik, 2023; Vendrow et al., 2025), suggesting that expert-annotated scientific datasets are not immune to label noise. The directional asymmetry (9/11 mislabeled SUPPORT) is, on its own sample size, suggestive but not conclusively significant: an exact two-sided binomial test against p=0.5 yields p=0.065 (one-sided p=0.033). The 95% Clopper–Pearson CI on the SUPPORT-error proportion is wide ([48%, 98%]), reflecting n=11. However, the asymmetry is robust to reasonable alternative classifications: if all 8 final "debatable" borderline cases (Appendix A) were reclassified as errors, the SUPPORT proportion remains 15/19 (79%) with two-sided p=0.019; under the expected 4/8 reclassification scenario, p=0.035. Nahum et al. (2025) observed a similar SUPPORT-skew pattern across factual-consistency benchmarks, lending external plausibility to the direction. We therefore describe this as a directional asymmetry warranting follow-up audit on a larger sample, rather than a confirmed systematic bias. With leaderboard gaps of 0.7–2.7 F1 between top systems, the binary correction-induced shifts of 1.7–3.8 F1 we observe across GPT-5.4 and Haiku 4.5 introduce noise of the same order as inter-system margins; the larger 3-way Haiku gain (+9.24) reflects 6 corrections that the binary schema cannot represent and is best read as an NEI-sensitive analysis rather than a leaderboard-comparable number. Whether the original-label rankings would actually reorder under corrected labels depends on whether top fine-tuned systems succeed and fail on the same items as our zero-shot models, which we do not establish here.

Six years without verification. Despite 600+ citations, our review of the dataset's GitHub repository (27 issues, none on annotation quality), Hugging Face page, and citing literature found zero prior label corrections or errata. This contrasts with other benchmarks: CoNLL-03 was re-annotated after two decades (Rucker and Akbik, 2023), and Vendrow et al. (2025) corrected 15 benchmarks simultaneously. We hypothesize that the "expert-annotated" label and biomedical domain specificity may have discouraged the kind of systematic verification that crowdsourced benchmarks routinely receive.

6 Conclusion

SciFact remains a useful benchmark, but its audited train and dev portions contain nontrivial annotation errors that can distort optimization targets and reported system comparisons. We document a directional asymmetry (9/11 dev errors mislabel SUPPORT; one-sided p=0.033, n=11) and show that correcting the dev labels raises binary zero-shot macro-F1 by 1.7–3.8 points across GPT-5.4 (mini, nano) and Claude Haiku 4.5—comparable to inter-system margins on the leaderboard. The 3-way Haiku gain is larger (+9.24) when the 6 mislabeled-as-NEI pairs are admitted; we report it as an additional NEI-sensitive analysis rather than as a direct leaderboard-comparable number. A claim-only contamination probe with Haiku 4.5 does not support label memorization as the main explanation for these gains. Rather than positioning our corrections as a definitive replacement, we recommend that the community treat them as a starting point and pursue an independent multi-annotator re-audit. We release the audit corrections, a blind-annotator packet for that review, and our reproduction scripts as a starting point.[6]

Limitations

We audit the dev and training sets only; the hidden test set may differ in error rate and error-type distribution, and we cannot directly verify it. *Manual verification is by one human annotator*, who was assisted on the 57 mini-flagged disagreements by a chat-mode frontier-LLM second opinion (§3.2) but not by an independent second human rater—a real concern given the original Cohen's κ=0.75 on the same task. To address this, we release a fully blind re-annotation packet for the entire dev set (no original gold, no LLM signal in the file the annotator sees) so any subsequent reviewer can independently re-audit and report inter-annotator agreement. Following Pavlick and Kwiatkowski (2019), some identified "errors"—particularly cross-species generalizations, borderline p=0.07 readings, and overgeneralization-type

[6] Corrected annotations, blind packet, and audit scripts are available at https://github.com/Kefez/scifact-audit-bionlp2026.

cases—may reflect legitimate methodological disagreement among clinical experts rather than objective labeling failures; the entity-mismatch errors (IDs 1216, 1274, 847) rely on the principle that findings in one organism do not automatically transfer to another, which some annotators may consider acceptable. The 8 final debatable cases (10 originally identified at first-pass review, with 2 subsequently upgraded to confirmed errors and counted in 11/209) are enumerated in Appendix A with a sensitivity analysis; case-by-case justifications are in the supplementary materials. The headline error rate stays under 10% even if all 8 were reclassified as errors. The impact analysis uses zero-shot models only on a single non-GPT family (Claude Haiku 4.5); fine-tuned systems and additional open-weight families are not evaluated and cannot be ruled out as differing in their response to the corrections. The contamination probe is a single negative test (claim-only with Haiku 4.5); a more extensive probe across open-weight models with known training data would strengthen the conclusion. Finally, our scope is a single benchmark; whether the directional SUPPORT-skew generalizes to other scientific-claim-verification corpora is an open empirical question.

Acknowledgments

We thank the BioNLP 2026 reviewers for detailed feedback that materially improved the statistical framing, cross-model verification, and contamination analysis in this paper.

References

Saso Cemerski, Jayajit Das, Emanuele Giurisato, Mary A. Markiewicz, Paul M. Allen, Arup K. Chakraborty, and Andrey S. Shaw. 2008. The balance between T cell receptor signaling and degradation at the center of the immunological synapse is determined by antigen quality. *Immunity*, 29(3):414–422.

Daria Esterházy, Ina Stützer, Haiyan Wang, Markus P. Rechsteiner, Jeremy Beauchamp, Heinz Döbeli, Hans Hilpert, Hugues Matile, Michael Prummer, Alexander Schmidt, and 1 others. 2011. Bace2 is a β cell-enriched protease that regulates pancreatic β cell function and mass. *Cell Metabolism*, 14(3):365–377.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pages 107–112.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Miloš Košprdić, Adela Ljajić, Darija Medvecki, Bojana Bašaragin, and Nikola Milošević. 2024. Scientific claim verification with fine-tuned NLI models. In *Proceedings of IC3K/KMIS*.

Fan-Tao Meng, Yan-Li Wang, Juan Liu, Jing Zhao, Rui-Yuan Liu, and Jiang-Ning Zhou. 2012. ApoE genotypes are associated with age at natural menopause in Chinese females. *Age*, 34(3):653–660.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are LLMs better than reported? Detecting label errors and mitigating their effect on model performance. In *Proceedings of EMNLP*.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of EMNLP*.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of NeurIPS*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the ACL*, 7:677–694.

Susanna Rucker and Alan Akbik. 2023. CleanCoNLL: A nearly noise-free named entity recognition dataset. In *Proceedings of EMNLP*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS Datasets and Benchmarks*.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of EMNLP*, pages 7534–7550.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022a. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of NAACL*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of EMNLP*.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of ACL*.

Zhiwei Zhang, Jihua Fang, and Barbara Plank. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of EMNLP*.

A Debatable Cases and Sensitivity Analysis

We enumerate the **8** SciFact dev claims that survived all review passes as "debatable but defensible" rather than confirmed errors. At an earlier triage pass, 10 cases were initially classified as debatable in the manual-review spreadsheets (`my_verdict == "DEBATABLE"` in the two false-alarm review files); on a second-pass review, 2 of those (claims 847 and 1274) were upgraded to confirmed errors and are part of the 11/209 figure. The 8 final debatable claim IDs are 57, 163, 431, 540, 982, 1041, 1132, 1150. Original gold-label split among the 8: 6 SUPPORT, 2 CONTRADICT (claims 57 and 1041 are CONTRADICT).

Sensitivity analysis (8 final debatables). If we instead reclassified all 8 as confirmed errors, the dev error rate rises to 19/209 = 9.1% (95% CI 5.6–13.8%). Under the expected 4/8 reclassification (3 of 4 added are SUPPORT, matching the 75% gold-split rate), the rate is 15/209 = 7.2% (95% CI 4.1–11.6%). The directional asymmetry persists in every scenario, and *strengthens* as more debatables are admitted: 9/11 SUPPORT-mislabel (82%, two-sided p=0.065) at 0/8 reclassified; 12/15 (80%, p=0.035) at the expected 4/8 case; 15/19 (79%, p=0.019) at 8/8. Per-case justifications for all 8 (and the 2 upgrades) are in the supplementary materials.

B Reproduction Artifacts

We release: (1) corrected gold labels for all 11 dev and 13 training errors as JSON; (2) a blind re-annotation packet for the 209 dev pairs (no original gold, no LLM signal) for inter-annotator agreement studies; (3) a 30-pair random sample of LLM-rejected pairs for false-negative-rate audit of the screen; (4) the Haiku 4.5 cross-model and contamination-probe scripts; (5) the statistical-analysis script that reproduces every CI and binomial test in this paper.