Evidence Completeness Governs Verifier Gains in Fact Verification

Anonymous ACL submission

Abstract

Evidence-aware fact verification rewards both a verdict and supporting evidence. We ask when verifier-stage gains become evidence-aware score gains. In FEVEROUS, fixed public DCUF retrieved evidence gives an oracle-label strict-score ceiling of 0.4322; replacing it with precomputed UnifEE evidence raises that ceiling to 0.5507, and gold evidence with oracle labels reaches 0.9992. A locked four-seed hybrid verifier improves FEVEROUS label macro-F1 over the DCUF public verifier from 0.4934 to 0.5416 and NEI-F1 from 0.0000 to 0.1767, but strict score falls from 0.3516 to 0.3478. A transition analysis explains the gap: the hybrid predicts NEI on 597 examples; 54 replace DCUF labels that were strict-correct, while 24 become newly strict-correct. In a FEVER fixed/gold-evidence contrast, stream cross-attention improves both macro-F1 and strict-style score. The results show that evidence completeness governs whether verifier gains affect evidence-aware metrics.

1 Introduction

Evidence-linked verification is usually decomposed into evidence acquisition and verdict prediction, but benchmark scores combine the two. FEVER introduced Wikipedia claim verification with sentence evidence (Thorne et al., 2018a), and the FEVER shared task formalized evidence-aware scoring for retrieved-evidence systems (Thorne et al., 2018b). FEVEROUS extends this setting to mixed sentence, table-cell, header, caption, and list evidence, and its shared-task score requires both correct labels and correct evidence (Aly et al., 2021). This creates a practical question: when does a better verifier improve the evidence-aware metric?

We study this question by holding evidence fixed and changing the verifier. The evidence regimes differ. In FEVEROUS, we use fixed public DCUF retrieved evidence (Hu et al., 2022), yielding an incomplete-evidence setting in which strict score can be capped before the verifier is changed. We also compute oracle ceilings with a second FEVEROUS evidence source from precomputed UnifEE outputs (Hu et al., 2023). In FEVER, we use fixed/gold evidence, producing a high-sufficiency contrast where evidence-aware score should track label behavior more directly.

The resulting pattern is two-regime. On FEVEROUS, a locked hybrid verifier improves label macro-F1 and NEI-F1 over the DCUF public verifier, yet loses strict score because added NEI behavior replaces some strict-correct SUPPORTS/REFUTES predictions. Oracle diagnostics show that the DCUF evidence itself caps strict score at 0.4322, while UnifEE evidence raises the ceiling to 0.5507. On FEVER fixed/gold evidence, a stream cross-attention verifier improves both macro-F1 and strict-style score. Verifier improvements can buy score gains, but only when supplied evidence is sufficiently complete.

2 Setup

Evidence-aware scores. For FEVEROUS, a strict prediction requires a correct label and coverage of a complete gold evidence set (Aly et al., 2021). We use a local official-style scorer that reproduces the stored DCUF dev-verdict values used here: strict score 0.3516, label accuracy 0.7205, and macro-F1 0.4934. Paired example-level bootstrap intervals use 10,000 resamples over the 7,890 official-dev examples with seed 13.

For FEVER, we report a strict-style diagnostic motivated by the FEVER scoring setup (Thorne et al., 2018b). It is the fraction of examples with both a correct label and a supplied fixed evidence set containing a complete gold evidence group. Under fixed/gold evidence, this score is dominated by label behavior; NEI examples are treated as evidence-sufficient by FEVER convention.

Verifier heads and systems. The controlled FEVER contrast compares three ways of combining the same frozen claim and evidence embeddings from an MPNet/Sentence-BERT-style encoder (Song et al., 2020; Reimers and Gurevych, 2019). The concat verifier contextualizes the claim and evidence embeddings together as a short sequence. The stream verifier keeps evidence units separate and lets the claim representation attend to them before classification. As a third baseline, we use a gated evidence-summary verifier, which compresses evidence units into a learned summary before combining it with the claim. All three heads receive the same fixed/gold evidence; only the fusion rule changes.

On FEVEROUS, the locked hybrid keeps the DCUF label unless a trainable NLI verifier crosses a train-confirm-selected NEI gate; the reported row is a four-seed mean-probability ensemble. Frozen-head audits use weak concat, parameter-matched concat, contextualized stream, router, and matched-budget DCUF-style heads on frozen MPNet embeddings. UnifEE is used as a precomputed evidence source for oracle-bound diagnostics, not as a verifier retraining run.

Study scope and artifact. The experiments isolate verifier-stage behavior by keeping evidence fixed, selecting thresholds on train/confirm screens before official-dev audit, and reporting development-split diagnostics. The artifact contains text-free prediction files, scorer outputs, paired bootstrap outputs, per-seed rows, table CSVs, scripts, and manifests. We use third-party datasets, pretrained encoders, and public system outputs only under their published research access conditions, licenses, or terms of use. Our released artifact is intended for research reproducibility and contains text-free predictions, scorer/bootstrap outputs, CSV summaries, scripts, and manifests; raw claim/evidence/wiki text, checkpoints, logits, and large caches are not redistributed and remain governed by their original providers' terms.

Implementation details. Frozen-head audits and the FEVER fixed/gold contrast use frozen MPNet/Sentence-BERT-style embeddings and train only lightweight fusion heads. The FEVEROUS hybrid uses a trainable NLI verifier only to supply the locked NEI gate over fixed DCUF evidence; no retrieval model is trained and no pretrained encoder is fine-tuned for the frozen-head controls. Experiments were run on Ubuntu 22.04 with Python 3.12, PyTorch CUDA, and an NVIDIA RTX A6000. Exact seeds, selected thresholds, configurations, checksums, compact prediction artifacts, and run-level wall-clock times are provided in the released manifests.

Evidence / labels	Strict	Acc.	Evid. rec.	Evid-F1
DCUF evidence / DCUF labels	.3516	.7205	.4322	.2234
DCUF evidence / oracle labels	.4322	1.000	.4322	.2234
UnifEE evidence / oracle labels	.5507	1.000	.5507	.2830
Gold evidence / oracle labels	.9992	1.000	.9992	.9996

Table 1: FEVEROUS strict-score ceilings. Oracle-label strict score equals the complete-evidence recall ceiling, and that ceiling changes with evidence source.

System	Macro	Acc.	NEI	Strict
DCUF public verifier	.4934	.7205	.0000	.3516
Locked hybrid	**.5416**	.6901	**.1767**	.3478
Delta, hybrid-DCUF	+.0482	-.0304	+.1767	-.0038
95% CI	[.0367,.0597]	[-.0356,-.0252]	[.1454,.2077]	[-.0061,-.0016]

Table 2: FEVEROUS official-dev comparison under fixed DCUF evidence. Confidence intervals are paired, example-level bootstrap intervals.

3 Evidence ceilings

Table 1 shows the FEVEROUS score ceiling before and after changing evidence. With fixed DCUF evidence, oracle labels reach 0.4322 strict score because complete-evidence recall is 0.4322. With UnifEE precomputed evidence, the oracle-label strict ceiling rises to 0.5507. Evidence-F1 rises more modestly, from 0.2234 to 0.2830, because strict scoring rewards whether at least one complete gold evidence group is covered; the ceiling change is driven primarily by recall of complete evidence groups. With gold evidence and oracle labels, the ceiling reaches 0.9992. The upper bound moves with evidence source.

4 Verifier gains under incomplete evidence

Figure 1a and Table 2 report the locked FEVEROUS comparison. The hybrid improves macro-F1 by +0.0482 over DCUF, with paired example-level bootstrap CI [0.0367, 0.0597], and improves NEI-F1 by +0.1767, with CI [0.1454, 0.2077]. Across the four locked seeds, macro-F1 ranges from 0.5365 to 0.5456, NEI-F1 from 0.1574 to 0.1866, and strict score from 0.3418 to 0.3487.

The strict-score decrease is exactly explained by the NEI transition. DCUF has 2,774 strict-correct examples, while the hybrid has 2,744, a difference of 30 examples. The hybrid predicts NEI on 597 examples. Among those, 54 replace DCUF predictions that were strict-correct under the fixed evi-

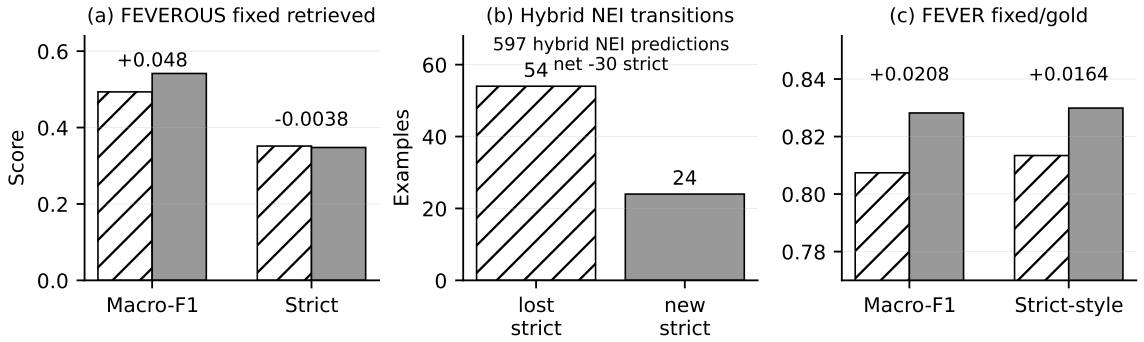

Figure 1: Evidence completeness governs how verifier gains affect evidence-aware scores. (a) On FEVEROUS fixed retrieved evidence, the hybrid improves label macro-F1 but loses strict score. (b) The strict loss comes from the hybrid's 597 NEI predictions: 54 replace DCUF strict-correct labels while 24 become newly strict-correct. (c) On FEVER fixed/gold evidence, the stream verifier improves both macro-F1 and strict-style score.

dence, while 24 become newly strict-correct. Thus the same behavior that improves NEI-F1 also trades against strict correctness under incomplete fixed evidence.

Frozen-head experiments show the same separation between verifier behavior and evidence completeness. Weak concat reaches only 0.3380 macro-F1 and zero NEI-F1. Stronger frozen heads cluster between 0.3636 and 0.3710 macro-F1, with the matched-budget DCUF-style head, contextualized stream, and router all close. On the train/confirm screens used before dev, contextualized stream beats weak concat by +0.0441 macro-F1 ($p = 0.00549$), but beats parameter-matched concat by only +0.0031 ($p = 0.285$). The point is not that one frozen head dominates; it is that verifier design changes label behavior without changing the fixed evidence ceiling.

A simple sufficiency head also does not move the ceiling. We tested a three-way label head plus a binary sufficiency head, trained with $CE_{3way} + \lambda CE_{suff}$ for $\lambda \in \{0.25, 0.5, 1.0\}$. Coarse empty-evidence counterfactuals were used for the binary sufficiency loss only. The best strict score among these variants is 0.3183; the best NEI-F1 is 0.0293. Calibration shows that the sufficiency head still predicts "sufficient" on most gold NEI examples.

5 A fixed/gold contrast on FEVER

The FEVER rerun tests the complementary regime. Here evidence sufficiency is controlled by fixed/gold evidence: the deterministic first-K evidence set is sufficient for 0.9918 of dev examples. In this setting, the stream head's label gain also appears in the evidence-aware score. Figure 1c and Table 3 show that stream improves macro-F1 by +0.0208 over concat, with CI [0.0156, 0.0261], and strict-style score by +0.0164, with CI [0.0126, 0.0203], positive on all five seeds.

FEVER model	Macro	Acc.	Strict-style
Concat	.8074	.8160	.8134
Stream	**.8282**	**.8328**	**.8299**
Gated summary	.8085	.8162	.8135
Delta, stream-concat	+.0208	+.0168	+.0164
95% CI	[.0156,.0261]	[.0129,.0207]	[.0126,.0203]

Table 3: FEVER fixed/gold evidence contrast. With evidence sufficiency controlled, verifier label gains also improve strict-style score. Confidence intervals are paired over matched seeds.

This contrast reframes the FEVEROUS result. Verifier gains are not intrinsically disconnected from evidence-aware metrics. Rather, under incomplete retrieved evidence, the metric has limited headroom. When evidence is nearly sufficient, as in the FEVER fixed/gold diagnostic, a verifier gain can translate into a score gain.

6 Discussion and conclusion

The evidence regime determines what a verifier improvement can buy. In FEVER fixed/gold evidence, evidence sufficiency is high and stream gains improve both labels and strict-style score. In FEVEROUS fixed retrieved evidence, oracle labels still hit a ceiling, and a stronger verifier's NEI behavior trades against strict-correct SUPPORTS/REFUTES examples. A second FEVEROUS evidence source raises the oracle ceiling from 0.4322 to 0.5507, reinforcing that strict-score headroom is controlled by evidence completeness.

The practical implication is simple: component-level verifier wins are meaningful, but evidence-

aware progress requires enough evidence for the verifier to act on. Evidence-extraction approaches such as DCUF and UnifEE target the complementary problem of improving the evidence source itself (Hu et al., 2022, 2023). The oracle ceilings are model-agnostic: an LLM verifier receiving the same incomplete evidence would still face the same evidence ceiling, while its label behavior could be analyzed with the same transition method. Whether the same governance pattern holds on open-web verification benchmarks such as AVeriTeC (Schlichtkrull et al., 2023) or on scientific-claim benchmarks remains an open question. Future FEVEROUS progress should treat retrieval, evidence sufficiency, and verdict modeling as coupled problems rather than expecting verdict swaps alone to overcome incomplete evidence.

Limitations

This study uses fixed-evidence diagnostics. FEVEROUS results use dev verdict artifacts and a local official-style scorer validated on the stored DCUF dev output used here. FEVER is a fixed/gold contrast, not a retrieval evaluation. The locked hybrid improves NEI-F1 but reduces accuracy and strict score. The sufficiency experiment covers one auxiliary formulation with coarse counterfactuals. The compact artifact releases text-free prediction and bootstrap outputs; large data, wiki text, checkpoints, and caches remain external.

Ethical Considerations

The experiments analyze component behavior in fact verification. The main risk is mistaking verifier-stage label gains for end-to-end evidence-aware progress; the results show why that interpretation can be wrong.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.fever-1.1. URL https://aclanthology.org/2021.fever-1.1/.

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. Dual-channel evidence fusion for fact verification over texts and tables. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.384. URL https://aclanthology.org/2022.naacl-main.384/.

Nan Hu, Zirui Wu, Yuxuan Lai, Chen Zhang, and Yansong Feng. UnifEE: Unified evidence extraction for fact verification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1150–1160, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.82. URL https://aclanthology.org/2023.eacl-main.82/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=fKzSz0oyaI.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, 2020.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium, November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/W18-5501. URL https://aclanthology.org/W18-5501/.

A Additional provenance

Rows in Table 4 summarize earlier controlled fixed-evidence runs used as provenance for the evidence-regime motivation. These rows are not additional leaderboard claims; they locate the present FEVEROUS/FEVER diagnostics in a broader fixed-evidence trajectory.

Dataset/protocol	Evidence	Main result	Use in this paper
AVeriTeC verifier mode (Schlichtkrull et al., 2023)	fixed evidence	stream-concat +10.29 pp; mechanism slices pass	stream-fusion motivation
FEVER fixed/gold evidence (Thorne et al., 2018a)	fixed/gold	stream-concat +1.357 pp; primary metric positive, mechanism slice below gate	contrast setting; rerun reported in main text
FEVER longer-evidence-cap stress	fixed/gold, longer evidence cap	stream-concat +1.797 pp; primary metric positive, mechanism slice remains below gate	truncation-stress context
FEVEROUS fixed/gold structured evidence (Aly et al., 2021)	fixed/gold structured	stream-concat -0.410 pp; NEI weak	structured-evidence boundary context
FEVEROUS fixed DCUF retrieved evidence (Hu et al., 2022)	fixed retrieved	hybrid label/NEI gain, strict loss	main result
FEVEROUS fixed UnifEE retrieved evidence (Hu et al., 2023)	fixed retrieved	oracle strict ceiling 0.5507	evidence-source diagnostic

Table 4: Provenance for additional controlled runs. Earlier rows are internal controlled experiments summarized only to situate the present FEVEROUS/FEVER diagnostics.