When FEVER Meets Finance: A Controlled Benchmark Reveals Overlap-Driven Financial Claim Verification

Suprabhat Rapolu
Dirmacs, Hyderabad
suprabhat@dirmacs.com

Abstract

Realistic financial misinformation is typically a single-token edit of a true statement, yet no public financial dataset supports per-edit statements about *which* factual changes a verification system can detect. We introduce a controlled benchmark of 1,489 claim–evidence pairs derived from SEC 10-K filings via documented FactCC-style perturbations (number changes, negation flips, entity swaps, evidence mismatches), each pair tagged with its generation rule, and use it to evaluate eleven models spanning three technological generations—TF-IDF linear classifiers, tree ensembles, and neural models including a fine-tuned transformer—all trained on FEVER under one protocol. Per-edit attribution is our central result: the strongest model verifies verbatim financial claims at 98.8% accuracy but detects single-token falsifications at 33% accuracy or below, with number changes and negation flips detected at similar rates and entity swaps far the hardest (2.6%, on 38 pairs). Two diagnoses follow. A hypothesis-only probe shows the classical tier's in-domain competence is claim-surface artifact—a claim-only classifier reaches macro-F1 0.556 against 0.549 for full claim+evidence models—while our benchmark shows no such artifact under the same probe (0.358). And transfer to real financial claims collapses: on Fin-Fact, *no* learned model beats a tuned lexical-overlap heuristic (macro-F1 0.368), despite in-domain scores up to 0.841. We release the benchmark, generation code, and evaluation harness.[1]

1 Introduction

Financial misinformation has an unusually direct transmission channel into harm: prices respond to beliefs within seconds, and a fabricated earnings figure or inverted growth percentage can move money before any correction catches up. Because manual fact-checking is measured in hours per claim against a stream measured in millions per day (Guo et al., 2022), automated claim verification—given a claim and evidence, decide SUPPORTS, REFUTES, or NOT ENOUGH INFO (NEI)—is the natural response (Thorne et al., 2018).

In practice, verification models are trained on general-domain benchmarks such as FEVER and applied to domain text. Two questions then matter: how much competence transfers to financial claims, and—since realistic financial misinformation is usually a *small edit* of a true statement—which specific edits a model can detect. Public financial datasets cannot answer the second question: real claims vary in uncontrolled ways, so per-edit error attribution is impossible.

We address both questions with a controlled comparison, scoped deliberately to *compact, low-latency encoder architectures trained under the standard FEVER pipeline*—the class of models actually deployable as a high-throughput triage layer over a claim stream, and the class whose failure modes the retrieval-then-verify literature has assumed rather than measured. Generative LLM verifiers are complementary and out of scope here (Limitations). Our contributions are: **(1)** A controlled evaluation benchmark of 1,489 claim–evidence pairs built from SEC 10-K filings by documented FactCC-style perturbation (Kryściński et al., 2020), balanced across three classes, each pair tagged with its generation rule, and audited for claim-surface shortcuts with a hypothesis-only probe. **(2)** A systematic evaluation of eleven models across three modelling generations—sparse linear, tree ensemble, and neural including a fine-tuned DistilBERT—trained on FEVER under a single protocol and scored by one shared harness on FEVER, Fin-Fact (Rangapur et al., 2023), and our benchmark. **(3)** Evidence that general-domain benchmarks overestimate financial verification abil-

[1] Benchmark, generation code, predictions and harness: https://github.com/suprabhatrapolu/factverify-benchmark

ity, with a diagnosis of why: claim-surface shortcuts inflate in-domain scores, lexical overlap drives financial-domain decisions, and input-length limits are a measured constraint on transfer.

2 Related Work

FEVER standardised three-way claim verification at scale (Thorne et al., 2018), within the retrieval–selection–verification decomposition of Thorne and Vlachos (2018); we study the final verification stage. A parallel literature asks what NLI-style models actually learn: hypothesis-only baselines beat majority classes across NLI datasets (Poliak et al., 2018), traced to annotation artifacts (Gururangan et al., 2018) and demonstrated for FEVER specifically (Schuster et al., 2019). We adopt this diagnostic tradition directly. FactCC generates synthetic evaluation data by rule-based perturbation of source sentences (Kryściński et al., 2020), the method we port to SEC filings; MiniCheck extends efficient fact-checking to LLM grounding (Tang et al., 2024). In financial NLP, FinFact provides expert-annotated real claims (Rangapur et al., 2023), FinBERT domain-adapts pre-training (Araci, 2019), and FinDVer shows that numerical reasoning over financial evidence challenges even frontier LLMs (Zhao et al., 2024)—anticipating the numeric-failure dominance we measure. Three recent resources are complementary to ours. QuanTemp (Venktesh et al., 2024) collects naturally occurring numerical claims across domains with retrieved evidence and reports a best macro-F1 of 58.3, establishing numerical verification as hard on real claims but without per-edit control. FISCAL (Sharma et al., 2025) generates synthetic claim–document pairs to *train* a compact numerical-claim verifier and evaluates it on FinDVer and Fin-Fact; we generate controlled pairs to *diagnose* verifiers instead, and our finding that FEVER-trained compact models collapse on Fin-Fact is the failure mode such domain-specific training targets. FinVerBench (Panda, 2026) perturbs XBRL financial statements from 43 S&P 500 10-Ks along an arithmetic, cross-statement, year-over-year and magnitude taxonomy to test whether LLMs detect numerical inconsistency, and shows that rendering choices materially change measured recall; it targets statement-level consistency judged by LLMs, whereas we target sentence-level claims against prose evidence with compact encoders and single-edit attribution. What none of this work provides is a controlled cross-generation comparison on identical data with per-perturbation error attribution on financial prose claims; we fill that gap.

Generation rule	**n**
Verbatim SUPPORTS (own context)	497
Number change ($\pm$15–60% or digit swap)	230
Negation / directional flip	229
Entity swap (cross-issuer)	38
Evidence mismatch (NEI, overlap < 0.2)	495
Total	1,489

Table 1: Benchmark composition. Every pair is tagged with its rule, enabling per-perturbation accuracy.

3 The SEC-Synthetic Benchmark

From the latest 10-K filings of twelve cross-sector U.S. issuers (via SEC EDGAR), we extract prose sentences carrying a genuine quantity signal under a documented quality filter (8–60 tokens; must contain \$, %, a magnitude word, a fiscal year, or a comma-grouped number; cross-references and index fragments excluded). Pairs are then generated deterministically (seed 42) by one edit per claim (Table 1): SUPPORTS pairs a filing sentence verbatim with its three-sentence context window; REFUTES applies exactly one perturbation—a number scaled by $\pm$15–60% or digit-swapped, a directional/negation flip, or an entity swap; NEI pairs an unperturbed sentence with a context window from a *different* filing (token overlap < 0.2). Table 2 shows one released pair per rule.

Because SUPPORTS and REFUTES claims are drawn from the same sentence distribution and differ only in their relation to the evidence, the set is free of claim-surface shortcuts by construction—and, for FEVER-style artifacts, by measurement: a FEVER-trained claim-only classifier scores just 0.358 macro-F1 on it (Section 5.2). Any model scoring well must read the pair. Aggregate scores on the set partially reward overlap by design (the lexical baseline reaches 0.559); its intended reading is therefore per-perturbation accuracy, not the aggregate.

4 Experimental Setup

Data. All models train on FEVER's gold-evidence formulation (227,943 train / 15,873 val / 16,008 test pairs after dropping empty-evidence rows) and are evaluated zero-shot on Fin-Fact (3,367 expert-annotated financial claims, labels mapped true/false/neutral $\rightarrow$ SUPPORTS/REFUTES/NEI —

Rule	Label	Claim (edit in bold; the unedited filing sentence sits inside the evidence window)
Verbatim	SUPPORTS	Dealer inventory decreased less during 2025 than during 2024. sec_CAT_0108
Number change	REFUTES	Markets & Securities Services revenue was **$14.3** billion, up 19%. *(filing: $41.3 billion; adjacent digits swapped)* sec_JPM_0658
Negation flip	REFUTES	Search and news advertising revenue **decreased** $1.6 billion or 13%. *(filing: increased)* sec_MSFT_0779
Entity swap	REFUTES	Prior to December 9, 2025, **Caterpillar Inc.'s** common stock was listed on the New York Stock Exchange. *(filing: Walmart's)* sec_WMT_0912
Evidence mismatch	NEI	Stock options granted during 2024 and 2023 were not material. *(paired with a window on interest expense and tax rate from a different filing; token overlap < 0.2)* sec_BA_1052

Table 2: One released pair per generation rule. Each REFUTES claim differs from its source sentence by exactly the edit shown.

a one-to-one mapping, 1,271 SUPPORTS, 1,485 REFUTES and 611 NEI of the 3,367 scored pairs, with two further `true` rows dropped for empty fields; the audit table in the repository records every raw label) and our benchmark. Fin-Fact and the SEC set are evaluation-only.

Models. Eleven models in three tiers plus two diagnostics, all through one harness with fixed seed and splits: a tuned lexical-overlap baseline (two thresholds on stopword-free claim-token overlap, t_{low}=0.55 and t_{high}=0.80, fit once by grid search on FEVER validation and held fixed for every evaluation set, so its Fin-Fact and SEC scores are zero-shot); classical TF-IDF models (Multinomial Naive Bayes, Logistic Regression, Linear SVM; shared 100k uni+bigram vectorizer fitted on train only); ensembles (Random Forest on an SVD-300 projection, XGBoost (Chen and Guestrin, 2016) on raw sparse features, soft voting with and without a calibrated SVM); and deep models (BiLSTM with GloVe-100 (Pennington et al., 2014); DistilBERT (Sanh et al., 2019; Devlin et al., 2019) fine-tuned as a sentence pair truncated at 256 sub-word tokens, once on a 40k stratified subset and once on the full train set). Classical hyperparameters use 3-fold grid search (macro-F1) on a 60k stratified subsample with full-train refit; deep models early-stop on validation macro-F1. The diagnostic claim-only Logistic Regression reuses the winning configuration with evidence withheld end-to-end (Poliak et al., 2018; Schuster et al., 2019).

Metric. Macro-F1 (test sets are near-balanced; errors matter equally across classes), with accuracy and per-class metrics recorded.

Model	FEVER	Fin-Fact	SEC
Lexical overlap (tuned)	.410	**.368**	.559
Claim-only LogReg (diag.)	.556	.258	.358
Multinomial NB	.440	.108	.179
Logistic Regression	.549	.205	.228
Linear SVM	.542	.201	.234
Random Forest (SVD-300)	.372	.244	.212
XGBoost (sparse TF-IDF)	.536	.364	.346
Soft voting	.552	.181	.210
+ calibrated SVM	.551	.185	.214
BiLSTM + GloVe	.586	.151	.221
DistilBERT (40k)	.803	.337	.541
DistilBERT (full train)	**.841**	.313	.535

Table 3: Macro-F1 in-domain (FEVER test) and zero-shot on the financial sets. On Fin-Fact, no learned model surpasses the lexical baseline.

5 Results

5.1 In-Domain: Pre-Training Is the Step Change

Table 3 (FEVER column) shows the generational ordering: classical models plateau at macro-F1 ≈0.55, ensembling classical features adds at most +0.003, the BiLSTM reaches 0.586, and DistilBERT reaches 0.803 (40k pairs) and 0.841 (full train). Only the transformer encodes claim and evidence jointly with cross-attention; every other model encodes the pair as one bag of features, and the +0.22–0.25 jump dwarfs every other gap.

5.2 The Claim-Only Diagnostic

A Logistic Regression trained on claims alone—evidence withheld—scores **0.556** on FEVER test, marginally *above* the identical claim+evidence configuration (0.549). The classical tier's entire in-domain performance is thus attributable to claim-surface regularities, the annotation artifacts documented for NLI and FEVER (Poliak et al., 2018;

Gururangan et al., 2018; Schuster et al., 2019). Two corollaries: the transformer's +0.25–0.29 margin over the claim-only baseline is the study's true measure of evidence use, and in-domain leaderboard gaps between non-pre-trained models are largely artifact gaps. On our benchmark the same probe drops to 0.358. Because the probe is FEVER-trained, this certifies the absence of FEVER-style claim-surface artifacts on the benchmark, not the absence of every conceivable shortcut; an in-domain claim-only probe would require financial training data that the study holds out by design, and we leave it to future work.

5.3 Transfer to Financial Text Collapses

No FEVER-trained model of any generation approaches its in-domain score on real financial claims (Table 3). The headline is stark: on Fin-Fact, *every* learned system—the full-train transformer included—is outperformed by the two-threshold lexical heuristic (0.368), whose thresholds were never refit on financial data. The dominant failure is NEI-collapse under unfamiliar vocabulary and long evidence. Three mechanisms are measurable. **Representation governs transfer:** XGBoost on raw sparse TF-IDF is the best non-transformer (Fin-Fact 0.364)—its top features are domain-general negation and restriction cues—while the same tree family on an SVD projection collapses (0.244), and soft voting transfers *worse* than its best member (0.181) because averaging amplifies shared bias. **Input capacity is a measured constraint:** 62.0% of Fin-Fact evidence exceeds DistilBERT's 256-token input (vs. 0.14% for FEVER, 2.35% for SEC), so on the majority of Fin-Fact pairs the model cannot see the deciding sentence. We report this as an association rather than an established cause: isolating it requires an intervention (retraining at 512 tokens, or truncating FEVER evidence to match) that our compute budget did not permit. A cheaper test is available from the released predictions: restricting Fin-Fact to the 1,279 pairs (38.0%) whose untruncated pair encoding fits within 256 tokens. On that subset macro-F1 is 0.290 [95% CI 0.267–0.314] for DistilBERT-40k, 0.274 [0.250–0.297] for the full-train model and 0.283 [0.258–0.308] for the lexical baseline, versus 0.353, 0.321 and 0.404 on the 2,088 truncated pairs (Table 5). Removing the mechanical ceiling does not improve transformer transfer; the subset is not a controlled comparison—it is REFUTES-heavier (51% vs 40%) and shorter by construction—so it does not isolate causation, but it rules out truncation as the sole mechanism of the collapse. **More in-domain training did not buy transfer in our run:** the full-train DistilBERT, +0.038 in-domain over its 40k sibling, is slightly worse on both financial sets (0.313 vs. 0.337; 0.535 vs. 0.541), a gap within plausible seed variance that we report as suggestive—fine-tuning longer on FEVER appears to specialise toward FEVER.

Perturbation	n	LogReg	DB-40k	DB-full
Verbatim SUPP.	497	.024	.982	.988
Number change	230	.052	.265	.326
Negation flip	229	.114	.306	.279
Entity swap	38	.053	.105	.026
Evid. mismatch	495	.943	.448	.418

Table 4: Accuracy by perturbation type on the SEC benchmark (DB = DistilBERT). The classical model's mismatch "detection" is NEI-default collapse. Entity-swap cells rest on n=38 and are correspondingly imprecise (DB-full 95% Wilson CI $[.005, .135]$; DB-40k $[.042, .241]$); the larger cells are tight (DB-full number change $[.269, .389]$, negation $[.225, .341]$, verbatim $[.974, .995]$).

5.4 Which Edits Fool Which Models

Table 4 delivers the benchmark's purpose: controlled, per-edit attribution. The best transformer verifies verbatim claims at 98.8% yet detects single-edit falsifications at only 32.6% (numbers) and 28–31% (negation), with entity swaps lowest of all at 2.6%: that cell rests on n=38 and its interval is wide (95% CI 0.5–13.5%), but the gap to negation flips is not marginal (Fisher exact p=0.0003), so we state the ordering and give the n. The pattern is consistent across cells: a claim nearly identical to its evidence is called SUPPORTS even when its one differing token inverts the meaning. The classical champion is qualitatively different: 2.4% on verbatim SUPPORTS, a pure domain collapse into NEI that only *appears* as 94.3% mismatch detection. Number changes and negation flips are not distinguishable (32.6% vs. 27.9%, p=0.31), so the ordering we assert is entity swaps below both; it plausibly tracks pre-training exposure, and the numeric result echoes FinDVer's finding that numerical reasoning over financial evidence remains open at much larger scales (Zhao et al., 2024). Since fabricated financial claims are typically small edits of true statements, this is exactly the failure mode that matters in deployment.

Model	Full n=3,367	$\leq$256 n=1,279	$>$256 n=2,088
Lexical overlap (tuned)	.368	.283	.404
XGBoost (sparse TF-IDF)	.364	.376	.333
DistilBERT (40k)	.337	.290	.353
DistilBERT (full train)	.313	.274	.321

Table 5: Fin-Fact macro-F1 split by whether the untruncated DistilBERT pair encoding fits the 256-token input (evaluation of released predictions; no retraining). 95% bootstrap CIs are within $\pm$0.026 for every cell.

6 Conclusion

We contribute a controlled, shortcut-audited SEC benchmark with per-perturbation tags, an eleven-model cross-generation comparison under one protocol, and three transferable lessons: audit claim-only shortcuts before trusting in-domain scores; expect zero-shot transfer to finance to be bounded by representation and input length rather than model class; and measure single-edit robustness, where even strong verifiers largely fail. Natural next steps are a zero-shot generative-LLM baseline and a financial-domain-adapted encoder (e.g., FinBERT) on the same harness, long-evidence architectures, and numeric-aware verification.

Acknowledgments

The author thanks the three anonymous FinNLP reviewers, whose questions about threshold fitting, evidence truncation and perturbation validity are answered in this version.

Limitations

Our benchmark's SUPPORTS/NEI classes are partially separable by lexical overlap *by construction*; the aggregate lexical-baseline score there (0.559) is an artifact of that design, which is why per-perturbation accuracy is the intended reading. FEVER's NEI evidence is retrieved rather than gold, inherited from the benchmark's own protocol. Fin-Fact's editorial labels are not perfectly isomorphic to NLI classes; our documented mapping carries a register gap. All results are single-seed (42) within a bounded compute budget, so small margins ($\pm$0.01) should not be over-read, and the 256-token transformer input is a budget choice whose cost we measure (62% Fin-Fact truncation) rather than hide; on the untruncated subset the transformer does not improve (Section 5.3). The entity-swap cell contains only 38 pairs, because quantity-bearing filing prose usually self-refers as "the Company"; its accuracy estimates are correspondingly wide and we report intervals rather than point claims. Benchmark pairs are generated by deterministic rules and were not validated by human annotators. An automated scan of the released set finds that 94 of the 230 number-change edits fall on four-digit years and produce implausible values (e.g., 2025$\rightarrow$1692), which remain unambiguous REFUTES but are not realistic misinformation, and that 7 of the 229 negation edits yield double negatives; a constrained year rule and a manual audit are the natural next additions. Splitting the 230 number-change edits by whether the edited token is a four-digit year (95 year edits, 126 other numeric edits; 9 not localisable by an automated diff), the full-train DistilBERT detects year edits at 38.9% (95% CI 28.4–49.5) and other numeric edits at 28.6% (21.4–36.5); the 40k model gives 31.6% and 23.8%. Excluding the year edits therefore leaves number-change detection at 28.6%, still inside the "33% or below" we report, so the headline does not rest on the implausible edits. One ordering does: with the year edits removed, number changes (28.6%) and negation flips (27.9%) are detected at indistinguishable rates (Fisher exact p=0.90), which is why the abstract no longer ranks them. Entity swaps remain lowest by a wide margin (2.6%, p=0.0003 against negation flips). Our scope is compact encoder verifiers trained under the standard FEVER pipeline; generative LLM verifiers, whether zero-shot or finance-tuned, are not evaluated here, and establishing whether they clear the single-edit bar is the most important open question our benchmark is built to answer—FinDVer's results suggest the numeric failure mode persists at much larger scale, but we do not measure it. All experiments are English-only and text-only.

Ethics and Data Statement

All data sources are public: FEVER (Wikipedia-derived), Fin-Fact (published fact-checks), and SEC EDGAR filings (public disclosures). The benchmark contains deliberately falsified sentences; every pair is machine-tagged with its generation rule, and the set is released for evaluation of verification systems, not for content generation. No personal data is processed.

References

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint*

arXiv:1908.10063.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

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 (Short Papers)*, pages 107–112.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of EMNLP*, pages 9332–9346.

Silu Panda. 2026. FinVerBench: Benchmark validity and calibration in large language model financial statement verification. *arXiv preprint arXiv:2605.29586*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of EMNLP-IJCNLP*, pages 3419–3425.

Rishab Sharma, Iman Saberi, Elham Alipour, Jie JW Wu, and Fatemeh Fard. 2025. FISCAL: Financial synthetic claim-document augmented learning for efficient fact-checking. In *NeurIPS 2025 Workshop on Generative AI in Finance*. ArXiv:2511.19671.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of EMNLP*, pages 8818–8847.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of COLING*, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pages 809–819.

V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. QuanTemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ArXiv:2403.17169.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of EMNLP*, pages 14739–14752.