

Relevant Is Not Enough: Expert-Audited Decision-Grade Web Evidence for Financial Fact Verification

Hyunkyung Park¹, Euihyun (Ian) Kim², Wooah Choi¹, Arkaitz Zubiaga¹

¹Queen Mary University of London, London, UK

²Deutsche Bank AG, Düsseldorf, Germany

{hyunkyung.park, bb24276, a.zubiaga}@qmul.ac.uk ian.kim@db.com

Abstract

Fact verification systems rank web evidence by relevance, yet relevant passages may lack verdict-critical facts. We define *decision-grade evidence* as credible, temporally valid, and claim-matched information independently sufficient to support or refute a claim. Using 750 expert-audited claim–passage pairs for 150 financial claims, we compare relevance and decision-grade supervision within two rerankers while holding candidate evidence pools and verifiers fixed. Within these pools, 66.5% of relevant pairs are insufficient, and claim-level decision-grade coverage trails relevance coverage by 40.0 percentage points. Compared with relevance supervision, decision-grade supervision improves the ranking of sufficient evidence in one reranker but shows no clear advantage in the other. Frozen (untuned) rerankers remain competitive, and downstream verification gains are modest and input-dependent. Thus, relevance is an unreliable proxy for verdict sufficiency; evidence availability, ordering, and verifier use require separate evaluation.

1 Introduction

Evidence retrieval for fact verification typically prioritises relevance, yet a relevant passage may not provide enough information to justify a verdict (Zhang et al., 2023). We define a passage as *decision-grade* if it is credible, temporally valid, claim-matched, and sufficient on its own to support or refute the claim. For a claim about a firm’s FY2023 operating margin, a passage reporting only the prior-year margin or that year’s revenue is relevant but insufficient. Financial evidence selection must therefore account for the claim’s specific metric, reporting period, and other constraints (Zhou et al., 2026).

Prior work has examined retrieval guided by verifier feedback, evidence-quality evaluation, and the ranking of sufficient evidence sets (Zhang et al.,

2023; Akhtar et al., 2026; Alt et al., 2026). However, usefulness to a verifier, agreement with reference evidence, and a passage’s standalone sufficiency are distinct criteria. It remains unclear whether replacing relevance labels with expert-audited sufficiency labels improves financial evidence selection and downstream verification when candidate evidence and verifiers are held fixed.

We address this gap through two comparisons. First, we compare the ranks assigned to verdict-sufficient passages under relevance and decision-grade supervision (RQ1). Second, we compare the performance of fixed verifiers given the evidence selected under each target (RQ2). Within each reranker backbone, we match training conditions and hold candidate pools and verifiers fixed to isolate the effect of the *supervision target*.

For 150 financial claims, we freeze pools of five pre-cutoff web passages before annotation and obtain expert-audited four-class labels for the resulting 750 claim–passage pairs. Figure 1 summarises the study: the same pools are reordered under relevance (REL) or decision-grade (DG) supervision, and the top-1 or top-2 passages are supplied to fixed verifiers. This design distinguishes whether sufficient evidence is available (*availability*), whether it is ranked early (*ordering*), and whether it helps the verifier (*use*). Within these bounded pools, relevance-based coverage overstates the availability of verdict-sufficient evidence. The benefits of decision-grade supervision vary across rerankers, and downstream verification gains remain modest and input-dependent (§4).

Contributions.

- A 750-pair diagnostic set with four-class decision-grade labels, annotated under a blinded protocol and fully audited by a financial risk analyst, released on GitHub with schema documentation.¹

¹<https://github.com/dweebie/dg-evidence-750>

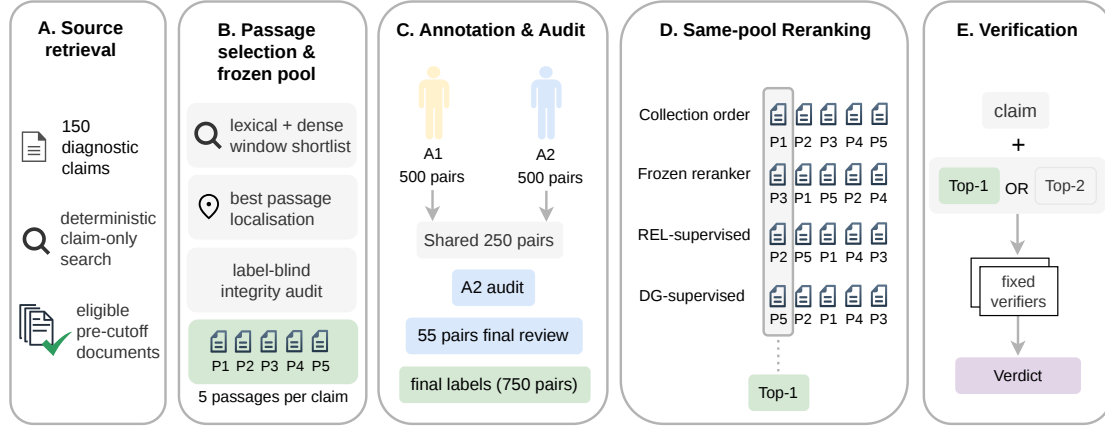


Figure 1: Study pipeline. (A) Claim-only search retrieves eligible pre-cutoff documents for 150 Fin-Fact claims. (B) Five-passage pools pass a label-blind integrity audit and are frozen. (C) Blinded annotation with full expert audit yields 750 four-class labels. (D) Frozen, REL-, and DG-supervised rerankers reorder the same fixed pools. (E) Top-1/top-2 passages feed the fixed verifiers.

- A matched comparison of REL and DG supervision within two reranker backbones, using identical candidate pools and fixed verifiers.
- An analysis of evidence availability, ordering, and use that identifies the limits of relevance as a proxy for verdict sufficiency.

2 Related Work

Financial verification and evidence constraints.

Fin-Fact provides financial claims with verdicts, ruling statements, and multimodal evidence (Ranapur et al., 2025); FinDVer evaluates claim verification over long hybrid-content documents (Zhao et al., 2024). Capybara combines search-derived evidence with financial chain-of-thought prompting (Cao et al., 2025), while TSVer studies temporal and numerical reasoning over time-series evidence (Strong and Vlachos, 2025). These studies examine verification across different evidence sources rather than isolating the effect of the passage-level supervision target. RealFin tests the effect of missing premises by modifying financial questions (Dai et al., 2026); we instead assess evidence sufficiency for unchanged claims. FinCARDS explicitly aligns metric, period, and scope constraints through zero-shot intra-document reranking (Zhou et al., 2026). Its constraint-aware selection motivates our sufficiency criteria, but differs from testing expert-audited REL/DG training targets on independently retrieved web passages.

Verifier utility and evidence evaluation. FER trains an evidence retriever using feedback from a fixed verifier’s output distributions (Zhang et al.,

2023). Its utility signal is verifier-dependent, whereas our passage labels are assigned without model outputs or original claim verdicts. The distinction is therefore the training signal—verifier feedback versus audited standalone sufficiency—not the use of a fixed verifier. Ev2R combines alignment with reference evidence and agreement with the reference verdict, validating evidence scores against human judgements and adversarial tests (Akhtar et al., 2026). This evaluates retrieved evidence quality but does not itself test the effect of replacing relevance supervision with passage-sufficiency supervision. We use audited labels for that controlled contrast and evaluate verifier outcomes separately.

Sufficient sets and evidence availability. User-centric evidence ranking prioritises the earliest prefix containing a sufficient evidence set (Alt et al., 2026). Its benchmark supplies sufficient gold sets, making it suitable for studying ordering and complementary evidence. Our labels instead assess each passage’s standalone sufficiency, and our pools retain cases with no decision-grade passage; this exposes availability failures that reranking alone cannot repair. Our passage-level labels do not measure the joint sufficiency of individually insufficient passages. AVeriTeC represents web evidence through question–answer pairs and justifications connecting them to a verdict (Schlichtkrull et al., 2023), addressing a broader evidence-combination problem.

Temporal control and diagnostic scope. Pre-cutoff eligibility is distinct from evidential suffi-

Label	Operational meaning	R	D	n
0	Irrelevant	0	0	302
1	Relevant but insufficient	1	0	298
2	Decision-grade support	1	1	71
3	Decision-grade refute	1	1	79

Table 1: Four-class passage labels and binary supervision mappings. R and D are the relevance and decision-grade reranking targets; n counts claim–passage pairs in the frozen 750-pair set.

ciency, and search-engine date filters can still admit post-cutoff content (El Lahib et al., 2026). Our analysis fixes candidate membership but does not imply that web retrieval is exhaustive or free of temporal leakage. Our contribution is a matched diagnostic of the supervision target, complementing evidence scorers, sufficient-set ranking, and end-to-end verification rather than replacing them.

3 Decision-Grade Evidence

Fixed pools and label space. For each of 150 claims sampled from Fin-Fact, we freeze five web passages that passed our metadata-based pre-cutoff eligibility filter, using the claim’s posting date as the cutoff, so every selector reorders the same 750 claim–passage pairs. Claims were preselected from training items that were self-contained and searchable. Up to three claim-only queries retrieved eligible pre-cutoff documents; verdict-bearing, inaccessible, and duplicate sources were excluded.

Ranking models served only to localise a passage window within each eligible document and could not add or reorder sources; source membership is thus model-independent. A label-blind integrity audit replaced 59 provisional pairs before the pools were frozen. Availability estimates are therefore conditional on this bounded pool; details appear in Appendix A.

Table 1 operationalises the decision-grade definition (§1): the relevance target $R(y)$ is positive for labels 1–3, the decision-grade target $D(y)$ only for labels 2–3. Labels 2 and 3 preserve support-versus-refute semantics but share the positive DG reranking target; pair labels supervise reranking, whereas Fin-Fact claim verdicts supervise verification.

Annotation and expert audit. A trained MSc annotator (A1) and a financial risk analyst (A2) each labelled 500 pairs, independently annotating the 250 pairs shared between them. Both annotators were blinded to the original verdicts and evidence,

model and rank information, and each other’s labels. A2 then audited all A1 labels and resolved remaining queries and disagreements. Initial agreement on the binary decision-grade target was 87.2% ($\kappa = .604$).

A decision-grade label required alignment with the claim’s entity, metric, fiscal period, geography, currency, and comparison, with sufficient evidence to establish or contradict the claimed value or relation. For multi-part claims, support required evidence for all necessary components, whereas refutation required a contradiction sufficient to invalidate the claim. Edge-case guidance and audit details appear in Appendix B.

Matched comparison. The relevance condition (REL) trains on $R(y)$ and the decision-grade condition (DG) on $D(y)$. Within each backbone, candidate pools, evaluation folds (§4.1), training pairs, initialisation, optimisation budget, and seeds are matched; only the target mapping and resulting class balance differ. This design isolates the effect of switching supervision targets under the matched training protocol. Frozen rerankers, used as released without fine-tuning, serve as references; all selected passages are evaluated by the same fixed verifiers.

4 Experiments and Results

4.1 Experimental Protocol

We evaluate BGE-Gemma (Liu et al., 2024; Chen et al., 2024) and GTE-base (Zhang et al., 2024) under frozen, REL, and DG conditions; BGE2.5-Gemma2 is an additional frozen reference, while collection order (the original retrieval order), BM25, and BM25–Nemotron are label-free references used only for downstream verification. Supervised predictions use a fixed claim-grouped five-fold cross-fitting split: each claim is scored only by models trained on the other folds, and results are averaged over seeds 7, 14, and 35. The split was built without claim verdicts or model outputs; claims without a decision-grade passage remain as all-negative examples under DG supervision, and class weights are estimated per training fold. BGE uses LoRA and GTE full fine-tuning, so only within-backbone REL/DG contrasts are controlled and cross-backbone differences are descriptive.

RoBERTa and FinBERT verifiers are trained once after removing diagnostic claims, duplicates, and linked records, then held fixed across all selector conditions; pair-level evidence labels are

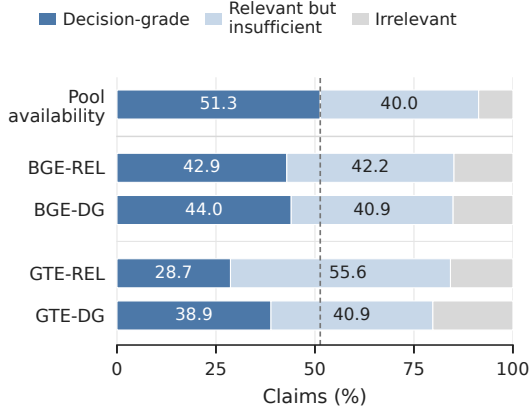


Figure 2: The top row shows claim-pool availability, and the remaining rows show Top-1 evidence composition under matched supervision.

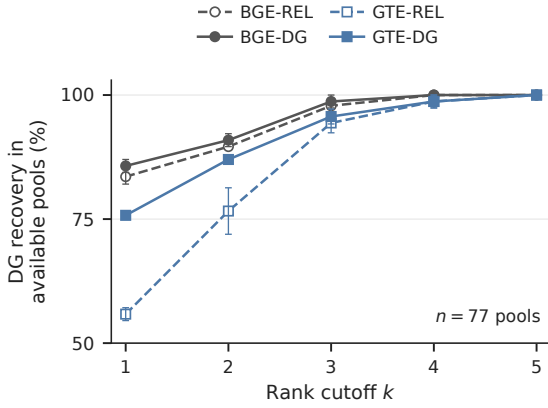


Figure 3: The curves show Top- k decision-grade recovery among the 77 pools containing at least one decision-grade passage.

never used as claim verdicts. Coverage@5 measures availability. RQ1 uses decision-grade P@1 over all 150 claims, regardless of availability, as its primary (*end-to-end*) metric, with MRR@5 and nDCG@5 as secondary metrics. Claim-level bootstrap confidence intervals (CIs) quantify uncertainty in matched P@1 differences. RQ2 is descriptive and reports three-class Macro-F1 with one or two passages. Full experimental details appear in Appendix C.

4.2 Results

Evidence availability. In the fixed candidate pools, 298 of 448 relevant pairs (66.5%) are insufficient for a verdict. Figure 2 contrasts pool availability with Top-1 evidence composition. The top row partitions the 150 pools into 77 containing decision-grade evidence, 60 containing relevant evidence but no decision-grade passage, and 13 con-

(a) Decision-grade				
System	P@1	MRR@5	nDCG@5	Δ P@1 (DG-REL)
BGE-REL	42.89	46.13	47.02	—
BGE-DG	44.00	46.83	47.74	+1.11 [-0.22, 2.89]
GTE-REL	28.67	37.73	40.98	—
GTE-DG	38.89	43.78	45.20	+10.22 [5.11, 15.78]
<i>Frozen references</i>				
BGE-Frozen	45.33	47.27	47.79	—
GTE-Frozen	38.00	43.13	44.51	—
BGE2.5-Frozen	42.67	46.47	47.26	—

(b) Relevance			
System	P@1	MRR@5	nDCG@5
BGE-REL	85.11	87.86	87.85
BGE-DG	84.89	87.40	87.09
GTE-REL	84.22	87.38	87.72
GTE-DG	79.78	84.61	85.35
<i>Frozen references</i>			
BGE-Frozen	85.33	87.89	87.73
GTE-Frozen	80.67	85.44	85.92
BGE2.5-Frozen	86.00	88.11	87.83

Table 2: Matched relevance and decision-grade reranking results with frozen references. Bold marks the higher REL/DG value per backbone; brackets are 95% claim-level bootstrap CIs for Δ P@1.

taining no relevant evidence. Thus, relevance Coverage@5 (137/150) exceeds decision-grade Coverage@5 (77/150) by 40.0 points. The dashed line marks the resulting 51.3% ceiling on end-to-end decision-grade P@1; the remaining 73 pools cannot yield a sufficient passage through reranking alone. The lower rows show mean Top-1 composition over three reranker training seeds, where the decision-grade and relevant-but-insufficient segments together constitute relevance P@1. Across these conditions, 40.9–55.6% of Top-1 selections are relevant but insufficient, exposing the gap between relevance and verdict sufficiency.

Decision-grade reranking (RQ1). Table 2 compares matched REL/DG conditions with frozen references. For BGE-Gemma, DG improves decision-grade P@1 over REL by only +1.11 points, with an interval crossing zero, while BGE-Frozen remains highest at 45.33. For GTE-base, DG exceeds REL by +10.22 points, with the interval entirely above zero. For GTE, REL training increases relevance P@1 relative to Frozen (84.22 vs. 80.67) but lowers decision-grade P@1 (28.67 vs. 38.00); DG training raises decision-grade P@1 to 38.89.

Figure 3 isolates ordering from availability by restricting evaluation to the 77 pools containing at

(a) RoBERTa		
Verifier input	Top-1	Top-2
Claim only	23.74	
+ Collection order	38.62	—
+ BM25	45.13	—
+ BM25–Nemotron	42.13	—
+ GTE–Frozen	36.21	37.88
+ BGE–Frozen	42.80	44.48
+ BGE2.5–Frozen	42.40	44.87
+ GTE–REL	39.85 ± 2.24	41.96 ± 2.49
+ GTE–DG	41.53 ± 0.87	41.29 ± 1.19
+ BGE–REL	41.79 ± 0.96	44.78 ± 0.67
+ BGE–DG	43.80 ± 0.67	46.35 ± 1.92

(b) FinBERT		
Verifier input	Top-1	Top-2
Claim only	38.70	
+ Collection order	32.61	—
+ BM25	38.32	—
+ BM25–Nemotron	38.25	—
+ GTE–Frozen	37.87	40.79
+ BGE–Frozen	35.50	39.36
+ BGE2.5–Frozen	36.29	43.03
+ GTE–REL	34.96 ± 0.62	41.20 ± 0.96
+ GTE–DG	35.40 ± 1.33	39.50 ± 1.43
+ BGE–REL	36.86 ± 0.47	39.97 ± 1.76
+ BGE–DG	37.48 ± 0.61	43.91 ± 0.89

Table 3: Fixed-verifier three-class Macro-F1 under claim-only, top-1, and top-2 evidence inputs; supervised rows report mean $\pm$ sd over three reranker training seeds. Bold marks the higher matched REL/DG mean per comparison.

least one decision-grade passage. At each cutoff k , recovery is the proportion of these pools with a decision-grade passage among the top- k candidates. Curves report means over reranker training seeds 7, 14, and 35, and error bars show $\pm$ one standard deviation across seeds, not confidence intervals.

At $k = 1$, BGE recovers 83.5% of these pools under REL and 85.7% under DG, while GTE rises from 55.8% to 75.8%. BGE–Frozen reaches 88.3% on the same subset but is not plotted. The GTE gap is concentrated at early ranks, whereas the BGE curves remain close. All curves reach 100% at $k = 5$ by construction, since the full five-passage pool is considered; the 73 pools without decision-grade evidence remain excluded.

Thus, within the bounded pools, BGE is primarily coverage-limited, whereas GTE–REL exhibits an ordering failure even when a sufficient candidate is available. RQ1 therefore yields a backbone-conditional answer: DG improves GTE’s decision-grade P@1 relative to REL, but the BGE contrast remains inconclusive and gains over the corresponding frozen rerankers are not uniform.

Fact verification (RQ2). Table 3 reports claim-only and evidence-conditioned performance.

At Top-1, DG yields higher mean Macro-F1 than matched REL in all four backbone–verifier comparisons, with gains of 0.45–2.01 points computed before rounding. BM25 yields the highest Top-1 Macro-F1 for RoBERTa, whereas FinBERT’s claim-only score exceeds those of all Top-1 evidence inputs. The competitive performance of label-free inputs is consistent with the *use* bottleneck: better-ranked evidence does not necessarily improve verification performance. At Top-2, both verifiers achieve higher mean Macro-F1 with DG than REL for BGE, whereas the reverse holds for GTE. RQ2 therefore shows small mean gains across all matched Top-1 comparisons, but not a uniform advantage across input depths; these descriptive results do not establish robust transfer to unchanged verifiers.

5 Conclusion

We introduced an expert-audited resource and a matched evaluation protocol for studying verdict-sufficient web evidence in financial fact verification. Holding candidate pools and verifiers fixed separates evidence availability, ordering, and use. Within these pools, topical relevance does not reliably indicate that a passage can independently justify a verdict. Decision-grade supervision can improve the early selection of sufficient evidence, but its benefit depends on the reranker and does not consistently transfer to fixed verifiers. These findings motivate evaluating sufficiency at each stage rather than treating relevance or ranking gains alone as indicators of verification quality. The resource supports further study of how evidence collection, selection, and use jointly shape verification.

Limitations

Our diagnostic set covers 150 self-contained English financial claims from the Fin-Fact training split, each with five pre-cutoff passages retrieved through a single search pipeline; availability and class-distribution figures are conditional on these bounded pools, not prevalence estimates for Fin-Fact or live open-web retrieval. Temporal filtering depends on source date metadata and may retain residual post-cutoff content. 250 of 750 pairs were independently double-annotated; the remainder received direct annotation or audit by one financial risk analyst and may reflect analyst-specific judge-

ment.

Supervised models train on 120 claims per fold with one training configuration per backbone, and only within-backbone REL–DG contrasts are controlled. RQ2 is descriptive—two encoder-based verifiers, point estimates without paired inference. Confidence intervals are conditional on the selected claims, pools, labels, and seeds, and do not propagate collection or annotation uncertainty.

Ethical Considerations

All passages are short excerpts from publicly accessible web pages that passed our metadata-based pre-cutoff eligibility filter; verdict-bearing, inaccessible, and high-risk social or video sources were excluded, and claims derive from the published FinFact dataset (Rangapur et al., 2025). Annotators worked under a blinded protocol without external browsing. The labels are context-dependent expert judgements of evidence sufficiency, not financial advice; neither they nor systems trained on them should be the sole basis for investment or other consequential decisions.

References

- Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2026. [Ev2R: Evaluating evidence retrieval in automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 14:530–561.
- Guy Alt, Eran Hirsch, Serwar Basch, Ido Dagan, and Oren Glickman. 2026. [User-centric evidence ranking for attribution and fact verification](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7237. Association for Computational Linguistics.
- Yupeng Cao, Haohang Li, Yangyang Yu, and Shashidhar Reddy Javaji. 2025. [Capybara at the financial misinformation detection challenge task: Chain-of-thought enhanced financial misinformation detection](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 321–325, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Yuyang Dai, Yan Lin, Zhuohan Xie, and Yuxia Wang. 2026. [RealFin: How well do LLMs reason about finance when users leave things unsaid?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25050–25080. Association for Computational Linguistics.
- Ali El Lahib, Ying-Jieh Xia, Zehan Li, Yuxuan Wang, and Xinyu Pi. 2026. [Temporal leakage in search-engine date-filtered web retrieval: A retrospective forecasting case study](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 787–795. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Allen H. Huang, Hui Wang, and Yi Yang. 2023. [FinBERT: A large language model for extracting information from financial text](#). *Contemporary Accounting Research*, 40(2):806–841.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. [Cumulated gain-based evaluation of IR techniques](#). *ACM Transactions on Information Systems*, 20(4):422–446.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024. [Llama2Vec: Unsupervised adaptation of large language models for dense retrieval](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. [Fin-Fact: A benchmark dataset for multimodal financial fact-checking and explanation generation](#). In *Companion Proceedings of the ACM Web Conference 2025*, pages 785–788. Association for Computing Machinery.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. *AVeriTeC: A dataset for real-world claim verification with evidence from the web*. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Marek Strong and Andreas Vlachos. 2025. *TSVer: A benchmark for fact verification against time-series evidence*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29906–29926. Association for Computational Linguistics.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. *From relevance to utility: Evidence retrieval with feedback for fact verification*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. *mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. *Qwen3 Embedding: Advancing text embedding and reranking through foundation models*. *arXiv preprint arXiv:2506.05176*.

Yilun Zhao, Yitao Long, Yuru Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Yiming Zhang, Xiangu Tang, Chen Zhao, and Arman Cohan. 2024. *FinDVer: Explainable claim verification over long and hybrid-content financial documents*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752. Association for Computational Linguistics.

Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, and Zhuohan Xie. 2026. *FinCARDS: Card-based analyst reranking for financial document question answering*. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 24836–24852. Association for Computational Linguistics.

A Data Construction and Integrity Control

Claim selection. The 150 claims were selected before any new passages were retrieved or any model was run, so claim selection could not be shaped by retrieval or model behaviour. We retained self-contained financial or economic claims

that could be assessed primarily from text and contained a usable search anchor (an entity, figure, or event specific enough to query), then normalised and deduplicated them. Fin-Fact’s TRUE, FALSE, and NEI labels are normalised to SUPPORTS, REFUTES, and NEI.

Candidate construction. Table 4 records the label-blind search, eligibility, passage, diversity, and repair settings. BM25 (Robertson and Zaragoza, 2009), Nemotron, and Qwen3-Reranker-4B (Zhang et al., 2025) scores were used only to localise windows within eligible documents and could not admit or reorder source URLs.

Before annotation, a label-blind audit checked temporal validity, extraction quality, verdict-page leakage, and duplication. It replaced 59 of the 750 provisional pairs across 49 claims: 55 replacements came from eligible documents or windows not yet in use, and four from a limited round of additional claim-only search. The audit used no gold label, original evidence or justification, or annotator judgement.

B Annotation and Expert Audit

Allocation and blinding. The claims formed three balanced 50-claim blocks—shared, A1-only, and A2-only—each containing 20 SUPPORTS, 20 REFUTES, and 10 NEI claims. Annotators saw the claim, cutoff date, passage, and minimal source metadata (title, domain, and date), but not gold labels, original evidence or justifications, model scores or ranks, URLs, or the other annotator’s labels. External browsing was prohibited, and claim and passage orders were randomised.

Additional label guidance. Beyond the core matching criteria in §3, annotators distinguished forecasts, proposals, estimates, and targets from realised outcomes and matched numerator/denominator and nominal/real metric formulations. Verdict-like source wording alone was not treated as evidence.

Expert audit. After both annotators’ initial labels were frozen, A2 audited all 500 of A1’s labels, querying 19 pairs and changing 17 labels. A2 then issued 55 final decisions: 53 resolved the remaining disagreements on shared pairs, and two addressed pairs where the annotators agreed but a query had been raised.

Table 5 traces the four-class label counts through these stages; the audit reduced label 1 by 13

Component	Setting
Search	Google Organic Search via SerpApi; at most three deterministic claim-only query variants and 20 results per query
Cutoff and eligibility	Fin-Fact posting date as cutoff; unknown-date, post-cutoff, verdict-bearing, inaccessible, high-risk social/video, and duplicate sources rejected
Passages	55–230 words; BM25 shortlist; nvidia/Nemotron-3-Embed-8B-BF16 hybrid shortlist; and Qwen/Qwen3-Reranker-4B within-source selection
Diversity	At most two passages per registered domain; five-token-shingle Jaccard threshold ≥ 0.86
Repair and freeze	Unused eligible documents or windows first; bounded claim-only supplementary search only if necessary

Table 4: Fixed label-blind settings for the frozen candidate pool.

Stage	Annotation set	n	0	1	2	3
<i>Stage 0</i>	A1	500	205	218	31	46
	A2	500	192	205	48	55
<i>Stage 1</i>	A1 after A2 audit	500	210	205	34	51
	Audit Δ	—	+5	−13	+3	+5
<i>Stage 2</i>	A2 final review	55	10	14	16	15
Final		750	302	298	71	79

Table 5: Four-class label counts across initial annotation, expert audit, and final review. The final row gives the frozen 750-pair set used in all experiments.

Unit	Formulation	n	Exact (%)	κ
Pair	Four-class	250	74.0	.606
Pair	REL	250	87.6	.738
Pair	DG	250	87.2	.604
Claim	REL Coverage@5	50	96.0	.811
Claim	DG Coverage@5	50	80.0	.610

Table 6: Initial pre-audit agreement on the independently labelled shared set.

pairs and shifted labels toward the irrelevant and decision-grade classes.

Agreement was computed from the two original labels on the 250 shared pairs, before audit or discussion. Table 6 reports exact agreement and nominal Cohen’s κ (Cohen, 1960).

C Ranking and Verification Protocol

C.1 Rerankers and Matched Training

Collection order was reconstructed from collection phase, query index, and organic-result position, with automatic-query candidates preceding supplementary-search candidates; annotation order was never used. BM25–Nemotron ranks by $0.7s_{\text{dense}} + 0.3s_{\text{BM25}}$. The Qwen3 reranker used for passage localisation was excluded from evaluation. Supervised rerankers use pointwise binary cross-entropy with logits. In addition to the controls listed in §3, matched REL and DG runs use the

same training-pair order.

Configurations. BGE-Gemma uses q/k/v/o LoRA (Hu et al., 2022) ($r = 32$, $\alpha = 64$), learning rate 2×10^{-4} , and the raw Yes-token logit. GTE-base is fully fine-tuned with learning rate 2×10^{-5} and a raw scalar logit. Both use three epochs, effective batch size 16, AdamW, weight decay 0.01, warm-up ratio 0.06, maximum gradient norm 1.0, and no early stopping. The exact checkpoints are BAAI/bge-reranker-v2-gemma, Alibaba-NLP/gte-multilingual-reranker-base, and BAAI/bge-reranker-v2.5-gemma2-lightweight.

C.2 Claim-Level Cross-Fitting

One immutable five-fold split was created with seed 42, keeping all five passages for a claim together and balancing pair labels and decision-grade availability. Each fold has 30 claims and 150 pairs; pair-label counts differ by at most two, and the 77 claims with a decision-grade candidate are distributed 15/15/15/16/16. For each fold, the model is trained on the other 120 claims and scores only the 30 held-out claims. Held-out predictions are concatenated into one 150-claim result per training seed; folds are diagnostic partitions, not independent repetitions.

C.3 Fixed Verifiers and Two-Passage Inputs

The fixed verifiers are initialised from ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli (Liu et al., 2019) and yiyanghust/finbert-pretrain (Huang et al., 2023). Both use the same source-grounded Fin-Fact training and development records after removing the 150 diagnostic claims, normalised duplicates, records linked by claim or evidence identifiers, and records without usable text. Original justifications and verdict-bearing summaries are excluded.

Training and model selection. The fixed split contains 2,339 training and 586 development records, grouped by normalised claim, canonical URL, and exact evidence text using fold 0 of a five-fold stratified group split (split seed 42). RoBERTa and FinBERT are trained with AdamW (Loshchilov and Hutter, 2019) for up to five epochs using learning rates 2×10^{-5} and 3×10^{-5} , training batch sizes 8 and 16, and gradient accumulation 4 and 2, respectively; both therefore use effective batch size 32. Both use a linear scheduler, warm-up ratio 0.06, weight decay 0.01, maximum length 512, and training seed 7.

Evidence precedes the claim and only evidence is truncated. Training uses class-balanced cross-entropy weights computed from the training split, early-stopping patience 2, and checkpoint selection by development Macro-F1. The RoBERTa NLI head is remapped to the common label order, whereas the FinBERT classification head is reinitialised. For two-passage evaluation, selected passages are concatenated in rank order under the 512-token limit while preserving the claim. Because no joint gold label exists for the pair, this condition tests input robustness rather than complementary sufficiency.

C.4 Metrics and Uncertainty

End-to-end decision-grade P@1 over all 150 claims is the pre-specified primary ranking metric. MRR@5 and nDCG@5 use binary gains $D(y)$ for decision-grade metrics and $R(y)$ for relevance metrics (Järvelin and Kekäläinen, 2002). Frozen and label-free systems are evaluated once.

We compute 95% percentile claim-level bootstrap confidence intervals for paired DG-REL P@1 differences using 10,000 replicates (seed 42). Each claim is resampled together with its five passages and predictions from all three seeds, preserving within-claim dependence and pairing between conditions. Other ranking metrics are descriptive.