Small LLMs for Biomedical Claim Verification: Cost-Effective Fine-Tuning, Structural Dataset Shortcuts, and Cross-Domain Generalization

Gaurav Kumar[1,2]
[1]University of California San Diego
[2]Moveworks AI
gkumar@ucsd.edu

Abstract

Large Language Models such as GPT-4o and GPT-5 achieve strong zero-shot performance on biomedical claim verification, but cost and opacity limit scalable use. We fine-tune three small LLMs: Phi-3-mini (3.8B), Qwen2.5-3B, and Mistral-7B, via QLoRA on SciFact and HealthVer, providing the first study of QLoRA models against GPT-4o and fine-tuned BioLinkBERT encoders. Mistral-7B QLoRA achieves higher F1 than both GPT-4o and GPT-5 (up to 12% gain) at 44.5× lower cost using just 1,008 training examples, representing a compelling cost-quality trade-off. We conduct extensive in-domain and cross-domain evaluation: models trained on SciFact tested on HealthVer and vice versa, at matched sizes to isolate dataset structure from data quantity. We identify a previously unreported structural artifact in SciFact that inflates in-domain scores, and show through bidirectional out-of-domain evaluation that training on structurally sound data enables robust cross-domain transfer. We release all code and adapter checkpoints.

1 Introduction

Automated biomedical claim verification determines whether a claim is SUPPORTED, REFUTED, or undetermined by evidence—a form of Natural Language Inference (NLI) applied to the biomedical domain. It is increasingly critical as health misinformation increases (Vladika and Matthes, 2023; Guo et al., 2022; Kotonya and Toni, 2020). Large language models (LLMs) such as GPT-4o offer strong zero-shot performance (Nori et al., 2023; Košprdić et al., 2024). However, API costs scale linearly with volume, they cannot be locally deployed in privacy-sensitive clinical environments, and silent provider updates undermine reproducibility. QLoRA (Dettmers et al., 2023) enables 4-bit fine-tuning of billion-parameter models on a single GPU in under an hour, yet no systematic comparison of QLoRA-adapted small LLMs against both proprietary and encoder baselines exists for this task under conditions that jointly test in-domain performance and out-of-domain generalization.

We fine-tune Qwen2.5-3B (Qwen Team, 2025), Phi-3-mini 3.8B (Abdin et al., 2024), and Mistral-7B-Instruct (Jiang et al., 2023) on SciFact (Wadden et al., 2020) and HealthVer (Sarrouti et al., 2021) separately, evaluating them bidirectionally alongside GPT-4o, GPT-5, and fine-tuned BioLinkBERT (Yasunaga et al., 2022). Our contributions are:

1. QLoRA fine-tuning beats both GPT-4o and GPT-5 on biomedical claim verification at 44.5× lower cost. Fine-tuning on just 1,008 examples enables Mistral-7B to achieve 88.4% macro-F1 on SciFact and 65.2% on HealthVer, surpassing GPT-4o (85.6%, 53.2%) and GPT-5 (77.9%, 42.4%) on both datasets, as well as earlier reported fine-tuned encoder results (Košprdić et al., 2024).

2. We identify a previously unreported structural artifact in SciFact. All 243 NEI training examples have empty evidence fields, making NEI trivially detectable from evidence absence rather than genuine reasoning. Fine-tuned models exploit this cue, achieving 100% in-domain NEI F1, inflating macro-F1 in a way unreported by prior work.

3. Bidirectional out-of-domain evaluation: Training on SciFact and testing on HealthVer, and vice versa at matched sizes demonstrates robust cross-domain adaptability: fine-tuning on just 1,008 HealthVer examples enables Mistral-7B to achieve 74.3% NEI F1 on SciFact OOD, outperforming BioLinkBERT trained on 10× more data (60.8%). The reverse direction confirms the shortcut mechanism: SciFact-trained models collapse on HealthVer, with the asymmetry ruling out domain shift and data quantity as explanations.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 704–712

2 Related Work

Biomedical claim verification. SciFact (Wadden et al., 2020) formalized verification with 1,409 expert-written claims annotated as SUPPORTS, REFUTES, or NEI. HealthVer (Sarrouti et al., 2021) extends this to real-world health queries, while PubHealth (Kotonya and Toni, 2020) targets public health misinformation. Most systems treat verification as NLI problem and fine-tune encoders such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), or DeBERTa (He et al., 2021); MultiVerS (Wadden et al., 2022) advanced SciFact state of the art via full-document modeling. Košprdić et al. (2024) showed DeBERTa achieves 88% F1 on SciFact, outperforming GPT-4 zero-shot, but only 48% on HealthVer OOD. We revisit this benchmark with instruction-tuned decoder LLMs.

LLMs for fact verification. GPT-4 has been evaluated on clinical QA (Singhal et al., 2023; Nori et al., 2023) and claim verification (Zheng et al., 2024), achieving strong but costly zero-shot performance. Prior work on NLI artifacts (Gururangan et al., 2018) shows models readily exploit spurious dataset correlations.

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) injects trainable rank-decomposition matrices $\Delta W = BA$ into frozen transformer weights. QLoRA (Dettmers et al., 2023) adds 4-bit NF4 quantization and paged optimizers, enabling single-GPU fine-tuning. QLoRA remains underexplored for biomedical NLI despite its practical fit for constrained label schemas and small datasets.

3 Method

3.1 Datasets

SciFact (Wadden et al., 2020) contains 1,409 expert-written claims paired with PubMed evidence, annotated as SUPPORTS, REFUTES, or NEI. We use an 80/20 stratified train/val split with the official dev set as test (450 examples). Label distribution: 48.8% SUPPORTS, 27.1% REFUTES, 24.1% NEI. Critically, all NEI examples have empty evidence fields by construction. Following standard practice for the claim verification task (Wadden et al., 2020), we use the annotated evidence sentences as model input rather than full abstracts; NEI claims have no annotated evidence by definition, resulting in empty evidence fields.

HealthVer (Sarrouti et al., 2021) provides 14,330 evidence–claim pairs from real-world health queries verified against PubMed, with official splits (10,590/1,917/1,823). Unlike SciFact, NEI in HealthVer requires reasoning over present-but-inconclusive evidence, eliminating the absence shortcut. Label distribution: 35.7% SUPPORTS, 22.8% REFUTES, 41.5% NEI.

Both datasets use unified (claim, evidence, label) triples with deterministic label normalization (Appendix A). For controlled bidirectional experiments, we sample 1,008 HealthVer training examples to match SciFact training size exactly, isolating dataset structure effects from data quantity.

3.2 Models

Zero-shot baselines. We run GPT-4o and GPT-5 at temperature 0, minimal reasoning with no chain-of-thought prompting (Appendix B).

QLoRA fine-tuned models. We fine-tune Phi-3-mini-4k-instruct (3.8B) (Abdin et al., 2024), Qwen2.5-3B-Instruct (Qwen Team, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) models. Each model is loaded in 4-bit NF4 quantization with LoRA adapters (r=16, α=32) on all attention and feed-forward layers. We perform supervised fine-tuning with lr 2×10^{-4}, cosine schedule, 3 epochs and AdamW 8-bit. We select hyperparameters through grid search (Appendix D). We train multiple models across SciFact (1,008) and HealthVer subset (1,008) separately. Full config is added in Appendix E.

Encoder baseline. We fine-tune BioLinkBERT-base (Yasunaga et al., 2022) with a three-class classification head on SciFact, HealthVer subset (1,008), and HealthVer full (10,590) to provide matched and ceiling comparisons.

3.3 Evaluation

We report macro-averaged F1, accuracy, and per-class F1 for all three labels. Every model trained on SciFact is evaluated on both SciFact (in-domain) and HealthVer (out-of-domain) test sets, and vice versa, enabling direct bidirectional comparison under controlled conditions.

4 Experiments and Results

4.1 In-Domain Performance

Table 1 shows in-domain results for all models.

SciFact. All three QLoRA models surpass GPT-4o macro-F1 on only 1,008 training examples. Mistral-7B achieves 88.4%, beating GPT-4o

Model	Train	Acc	F1	SUP	REF	NEI
SciFact test set						
GPT-4o	—	85.8	85.6	86.8	87.7	82.3
GPT-5	—	76.9	77.9	75.1	**89.1**	69.5
BioLinkBERT	SF	87.1	87.5	**86.8**	75.6	100.0
Phi-3-mini	SF	86.2	86.4	86.0	73.3	100.0
Qwen2.5-3B	SF	85.3	86.8	82.4	78.0	100.0
Mistral-7B	SF	**87.6**	**88.4**	86.5	78.6	100.0
HealthVer test set (1,008-example training subset)						
GPT-4o	—	56.0	53.2	39.7	56.8	63.1
GPT-5	—	50.4	42.4	15.9	49.4	61.9
BioLinkBERT	HV_{sub}	65.5	62.4	62.7	47.7	76.6
Phi-3-mini	HV_{sub}	**66.1**	65.1	64.1	59.8	71.3
Qwen2.5-3B	HV_{sub}	57.0	53.6	46.9	49.6	64.5
Mistral-7B	HV_{sub}	66.0	**65.2**	**71.1**	**61.7**	62.7
HealthVer test set (full training set, reference ceiling)						
BioLinkBERT	HV_{full}	82.7	81.9	81.8	77.4	86.3

Table 1: In-domain results. SF = SciFact (1,008). HV_{sub} = HealthVer 1,008-sample subset. HV_{full} = full HealthVer training set (10,590), shown as reference ceiling. Metrics are percentages.

Model	Train→Test	Acc	F1	SUP	REF	NEI
SciFact-trained → HealthVer						
GPT-4o	—→HV	56.0	53.2	39.7	56.8	63.1
GPT-5	—→HV	50.4	42.4	15.9	49.4	61.9
BioLinkBERT	SF→HV	40.3	31.5	53.5	40.9	0.3
Phi-3-mini	SF→HV	42.8	36.0	55.1	44.9	8.0
Qwen2.5-3B	SF→HV	45.4	43.2	54.7	47.3	27.5
Mistral-7B	SF→HV	**48.9**	**44.4**	**60.1**	**52.9**	**20.1**
HealthVer-trained → SciFact (matched 1,008 examples)						
GPT-4o	—→SF	85.8	85.6	86.8	87.7	82.3
GPT-5	—→SF	76.9	77.9	75.1	89.1	69.5
BioLinkBERT	HV_{sub}→SF	61.1	53.3	70.8	21.9	67.1
Phi-3-mini	HV_{sub}→SF	63.3	59.2	71.3	43.3	63.1
Qwen2.5-3B	HV_{sub}→SF	41.3	37.9	39.1	27.8	46.7
Mistral-7B	HV_{sub}→SF	**72.4**	**69.3**	**77.8**	**55.7**	**74.3**
HealthVer full → SciFact (reference ceiling, 10,590 examples)						
BioLinkBERT	HV_{full}→SF	68.9	60.8	77.9	29.1	75.3

Table 2: Bidirectional OOD results. SF = SciFact; HV = HealthVer. Top: SciFact-trained models on HealthVer test. Middle: HealthVer subset (1,008)-trained models on SciFact test. Bottom: BioLinkBERT trained on full HealthVer (reference). GPT-4o shown for reference

(85.6%) by 2.8 points and BioLinkBERT (87.5%) by 0.9 points. McNemar's test confirms statistical indistinguishability between Mistral and GPT-4o (p=0.46) and across QLoRA models (p=0.54). GPT-4o holds an edge on REFUTES, reflecting broad pre-training for detecting subtle directional contradictions. GPT-5 zero-shot achieves only 77.9% macro-F1 on SciFact, below GPT-4o and all fine-tuned models, suggesting newer proprietary models do not automatically improve on structured verification tasks. Moreover, we discuss the mechanism behind perfect NEI score in Section 4.2.

HealthVer. Fine-tuning on just 1,008 HealthVer examples, Mistral-7B QLoRA achieves 65.2% macro-F1, surpassing GPT-4o, GPT-5, and BioLinkBERT (53.2%, 42.4%, 62.4%). It demonstrates that QLoRA-adapted decoders beat both the proprietary models and the encoder approach on real-world health queries with minimal data. The full-training BioLinkBERT ceiling (81.9%) shows that adding 9,582 examples gains 19.5 macro-F1 points, confirming data quantity effects are real but do not close the architectural gap at matched scale. Beyond performance, decoder models allow free-form explanation, zero-shot prompting, and do not require task-specific classification heads.

4.2 The SciFact NEI Structural Shortcut

High NEI F1 achieved by fine-tuned models on SciFact is not genuine epistemic reasoning. We find that all NEI examples have empty evidence fields, while every SUPPORTS and REFUTES example contains evidence. The label is perfectly separable from evidence length alone, without reading the claim content. Full stats in Appendix C, Table 8.

SciFact assigns NEI when no cited evidence exists, inadvertently creating a structural signal any expressive model will learn. Zero-shot GPT-4o attempts genuine reasoning on NEI instances. This shortcut inflates macro-F1 for all fine-tuned models and has gone unreported in prior work.

4.3 Bidirectional Out-of-Domain Generalization

Table 2 presents the full cross-dataset evaluation.

SciFact → HealthVer. All SciFact-trained models suffer catastrophic degradation as macro-F1 drops from 86-88% in-domain to 36-44% OOD. NEI collapse is most severe as HealthVer NEI has non-empty, topically relevant evidence and the absence shortcut does not fire. Table 2 shows the results. Qwen2.5-3B is the most NEI-resilient SciFact-trained model, suggesting broader pre-training provides some generalization advantage.

HealthVer → SciFact. Reverse direction tells a strikingly different story. Mistral-7B trained on just 1,008 HealthVer examples achieves 69.3% macro-F1 OOD on SciFact with 74.3% NEI F1, far above any SciFact-trained model on HealthVer. Remarkably, this outperforms BioLinkBERT trained on the full HealthVer set (60.8%), despite using 10× less

Model	Cost/1K	Fine-tune	vs. GPT-4o
GPT-4o (API)	$1.3000	—	1.0×
GPT-5 (API)	$0.5750	—	2.4× cheaper
Phi-3 (QLoRA)	$0.0292	$0.35	44.5× cheaper
Qwen2.5 (QLoRA)	$0.0292	$0.35	44.5× cheaper
Mistral (QLoRA)	$0.0292	$0.35	44.5× cheaper
BioLinkBERT	$0.0292	$0.17	44.5× cheaper

Table 3: Inference cost per 1,000 preds and one-time fine-tuning cost (T4 GPU). API pricing as of early 2026.

training data. BioLinkBERT at matched size shows meaningful transfer and achieves 53.3% macro-F1 and 67.1% NEI F1. Together, these results demonstrate that training on structurally sound data enables robust cross-domain transfer and model architecture matters more than data quantity.

Asymmetric OOD generalization. Symmetric degradation would indicate distributional shift; instead, collapse is strictly one-directional. The asymmetry holds at matched training sizes (1,008 examples), ruling out data quantity. Models trained on genuine epistemic reasoning (HealthVer) transfer; models trained on a structural proxy (SciFact) do not.

Per-class analysis. REFUTES F1 degrades most severely OOD in both directions: BioLinkBERT (HealthVer$_{full}$→SciFact) achieves only 29.1% despite strong SUPPORTS (77.9%) and NEI (75.3%) transfer. Directional contradiction detection requires domain-specific reasoning that generalizes poorly across biomedical claim types. This is a critical gap given that refutation detection is essential for any clinical task.

GPT-4o and GPT-5: paradoxical OOD behavior. GPT-4o achieves the highest HealthVer NEI F1 (63.1%) among zero-shot models despite the lowest SciFact NEI F1 (82.3%). Its in-domain weakness reflects genuine uncertainty reasoning that transfers. GPT-5 shows more pronounced NEI over-prediction: 95.5% NEI recall on SciFact but only 8.8% SUP recall on HealthVer, suggesting over-cautious RLHF training that predicts NEI regardless of evidence. In-domain scores are poor proxies.

4.4 Cost Analysis

GPT-4o costs $1.30 per 1,000 predictions versus $0.03 for local open models (44.5× drop). The $0.35 one-time fine-tuning cost amortizes to negligible overhead, making fine-tuned small models strictly preferable on both performance and cost grounds. For OOD tasks, training data choice is critical. SciFact-trained models collapse OOD while HealthVer-trained models transfer robustly, achieving strong cross-domain performance.

4.5 Qualitative Error Analysis

We examined 48 cases where Mistral-7B QLoRA is correct and GPT-4o is wrong, and 40 reverse cases, on the SciFact test set.

Fine-tuning wins. Mistral-7B outperforms GPT-4o on NEI: for "Statins increase blood cholesterol," GPT-4o predicts REFUTES from pharmacological knowledge while Mistral correctly withholds judgment given no retrieved evidence. Fine-tuning teaches models to attend to evidence rather than prior knowledge.

GPT-4o wins. GPT-4o retains an edge on directional REFUTES: for "The risk of male prisoners harming themselves is ten times that of female prisoners," the evidence states the opposite direction. GPT-4o correctly predicts REFUTES; Mistral latches onto "ten times" without detecting the subject inversion. Mistral nonetheless closes this gap vs. Phi-3-mini (11 GPT-4o wins vs. 29), suggesting larger capacity helps with contradiction detection.

Shared failure modes. Both models err on claims where evidence describes a related but experimentally distinct condition, and on logical negations introduced by experimental targeting. The annotation ambiguity is documented in SciFact inter-annotator studies (Wadden et al., 2020).

5 Conclusion

Fine-tuning small open-weight LLMs via QLoRA on just 1,008 examples surpasses both GPT-4o and GPT-5 on SciFact and HealthVer at a fractional cost, establishing parameter-efficient fine-tuning as a practical alternative to APIs. Through extensive OOD evaluation, we show both the promise and limits of this approach: HealthVer-trained models transfer robustly to SciFact, outperforming BioLinkBERT trained on 10× more data, while SciFact-trained models collapse cross-domain due to a structural artifact in SciFact. REFUTES remains the hardest class to transfer in both directions, pointing to cross-domain contradiction detection as an open challenge. Our findings motivate structural auditing of biomedical NLI benchmarks and bidirectional evaluation as standard practice.

Limitations

Our bidirectional experiments use matched 1,008-example training sets, but scaling to full HealthVer data would further improve in-domain performance, as evidenced by BioLinkBERT's 81.9% ceiling vs. 62.4% at matched size. We evaluated GPT-5 as an additional zero-shot baseline; its lower performance (77.9% SciFact, 42.4% HealthVer) confirms that QLoRA fine-tuning advantages hold against the latest proprietary systems. Cost estimates reflect T4/A100 and API pricing as of early 2026. We evaluate English claims only; whether analogous structural shortcuts exist in PubHealth, MedNLI, or other benchmarks remains an open question. Data augmentation and debiasing strategies for improving cross-domain robustness are left for future work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Amin, Jyoti Aneja, Ahmed Awadalla, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Beber, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

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 Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Miloš Košprdić, Adela Ljajić, Darija Medvecki, Bojana Bašaragin, and Nikola Milošević. 2024. Scientific claim verification with fine-tuned NLI models. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)*, pages 15–25.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhoujun Li, Dacheng Li, Eric Xing, and 1 others. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Label Normalization

Refer to Table 4.

Raw output variants	Label
SUPPORTS, SUPPORT, SUPPORTED, ENTAILMENT	SUPPORTS
REFUTES, REFUTE, REFUTED, CONTRADICT(S)	REFUTES
NEI, NOT_ENOUGH_INFO, NOT ENOUGH INFO, NEUTRAL	NEI
(unrecognized)	NEI (fallback)

Table 4: Label normalization mapping

B Prompt Templates

All instruction-tuned LLMs use the following prompt. Phi-3-mini uses ChatML format; Mistral-7B and Qwen2.5 use `[INST]/[/INST]` format.

```
System: You are a biomedical claim verification expert. Given a claim and an evidence passage from a scientific abstract, determine whether the evidence SUPPORTS the claim, REFUTES the claim, or whether there is NOT ENOUGH INFO. Respond with exactly one of: SUPPORTS, REFUTES, NEI.

User: Claim: {claim}
Evidence: {evidence}
What is the verdict?

Assistant: {label}
```

For GPT-4o, only the system and user turns are used at inference. For fine-tuning, the full exchange including the gold label is the SFT training target. For SciFact NEI examples, the evidence field is empty (`Evidence: `), the structural signal described in Section 4.2. We run GPT-4o with temperature 0 and GPT-5 with minimal reasoning.

C Dataset Statistics

Tables 5, 6 present dataset split sizes and label distributions. Tables 7, 8 provide representative input examples and evidence field statistics respectively

Dataset	Train	Val	Test	Total
SciFact	1,008	253	450	1,711
HealthVer (subset)	1,008	—	1,823	—
HealthVer (full)	10,590	1,917	1,823	14,330

Table 5: Dataset splits. SciFact uses stratified 80/20 train/val split; official dev set as test. HealthVer subset sampled with random_state=42 to match SciFact training size.

Dataset (train)	SUP%	REF%	NEI%
SciFact	48.8	27.1	24.1
HealthVer	35.7	22.8	41.5

Table 6: Label distributions. HealthVer's higher NEI proportion and structurally distinct NEI definition contribute to the generalization asymmetry.

Dataset	Claim	Evidence	Label
SciFact	Statins increase blood cholesterol.	*(empty)*	NEI
SciFact	Metformin reduces HbA1c in diabetic patients.	Metformin significantly reduced HbA1c levels compared to placebo...	SUP
SciFact	Aspirin prevents colorectal cancer.	Aspirin use was associated with reduced risk...	REF
HealthVer	Vitamin C prevents the common cold.	Some studies suggest marginal effects on duration but evidence remains inconclusive...	NEI
HealthVer	Exercise reduces depression symptoms.	Aerobic exercise significantly reduced depressive symptoms across trials...	SUP

Table 7: Representative input examples from SciFact and HealthVer. SciFact NEI examples have empty evidence fields by construction; HealthVer NEI examples contain present-but-inconclusive evidence, eliminating the absence shortcut. Evidence text truncated for brevity. This structural difference explains the asymmetric OOD generalization reported in Section 4.3.

Label	Count	Mean chars	Zero-length
NEI	243	0.0	243/243 (100%)
SUPPORTS	492	214.5	0/492 (0%)
REFUTES	273	227.4	0/273 (0%)

Table 8: Evidence field length by label in SciFact training split. All NEI examples have empty evidence; no SUP/REF example does. This holds in val and test splits as well.

D Hyperparameter Grid Search

We perform multi configuration grid search over $r \in \{8, 16, 32\}$ and lr $\in \{10^{-4}, 2\times10^{-4}, 5\times10^{-4}\}$ using Mistral-7B on SciFact validation set. Selected config (r=16, lr= 2×10^{-4}) in bold.

E Hyperparameter Configuration

We list down the hyperparameters used for fine-tuning the LLMs in Table 10.

r	LR	Val Macro-F1
16	$\mathbf{2\times10^{-4}}$	**88.4**
8	5×10^{-4}	87.4
16	5×10^{-4}	85.7
32	5×10^{-4}	84.6
8	2×10^{-4}	81.3
32	2×10^{-4}	79.0
16	1×10^{-4}	77.8
32	1×10^{-4}	75.7
8	1×10^{-4}	74.5

Table 9: Hyperparameter grid search results on SciFact validation set (Mistral-7B). Selected configuration in bold. Higher learning rates generally outperform lower ones; r=16 at lr= 2×10^{-4} achieves the best validation macro-F1 of 88.4%.

Hyperparameter	Value
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Quantization	NF4, double quantization
Learning rate	2×10^{-4}
LR schedule	Cosine, warmup ratio 0.05
Effective batch size	16 (batch 4, grad. accum. 4)
Max sequence length	1,024 tokens
Training epochs	3
Precision	bf16 (fp16 fallback on T4)
Optimizer	Paged AdamW 8-bit
Checkpoint selection	Best validation macro-F1
Trainable params (Phi-3)	8.9M / 3.83B (0.23%)
Trainable params (Mistral/Qwen)	41.9M / 7.29B (0.58%)
LoRA target modules:	
`q_proj, k_proj, v_proj, o_proj,`	
`gate_proj, up_proj, down_proj`	

Table 10: QLoRA configuration, identical across all architectures and datasets.

F Training Efficiency

We measure training efficiency as a function of training time and cost. All models were trained on Google Colab Pro. Phi-3-mini, Qwen2.5-3B, and BioLinkBERT were trained on a T4 GPU (16 GB); Mistral-7B required an A100 (40 GB) due to memory requirements at 7B scale. Times are reported for SciFact (1,008 examples, 3 epochs); HealthVer subset training is approximately equivalent. The detailed breakdown is in Table 11.

G McNemar's Test Details

McNemar's test statistic: $\chi^2 = (b_{01} - b_{10})^2/(b_{01} + b_{10})$, where b_{01} = model A correct, B wrong; b_{10} = model B correct, A wrong. Evaluated on SciFact test set (450 examples).

Model	GPU	Time	Cost
Phi-3-mini	T4 (16 GB)	∼50 min	$0.15
Qwen2.5-3B	T4 (16 GB)	∼70 min	$0.21
Mistral-7B	A100 (40 GB)	∼35 min	$0.35
BioLinkBERT	T4 (16 GB)	∼15 min	$0.04

Table 11: Training times and costs per model (SciFact, 1,008 examples, 3 epochs). T4: ≈$1.76 CU/hr; A100: ≈$10–15 CU/hr.

Model A	Model B	b_{01}	b_{10}	χ^2	p
Mistral QLoRA	GPT-4o	48	40	0.56	0.46
Mistral QLoRA	Phi-3 QLoRA	36	30	0.38	0.54

Table 12: McNemar's test on SciFact test set. Neither comparison is significant at $p < 0.05$.

H Qualitative Error Analysis: Extended Examples

Claim	Gold	Mistral	GPT-4o
Pattern A: Fine-tuned wins on NEI			
Statins increase blood cholesterol.	NEI	NEI	× REF
Venules have larger lumen than arterioles.	NEI	NEI	× SUP
Pattern B: GPT-4o wins on REFUTES			
Male prisoner self-harm risk is 10× female.	REF	× SUP	REF
Nicotine combo therapy yields significantly higher abstinence.	REF	× SUP	REF
Pattern C: Mistral wins on REFUTES (new vs. Phi-3)			
NIV use should be decreased for poor responders.	REF	REF	× NEI
Tirasemtiv has no effect on fast-twitch muscle.	REF	REF	× NEI
Pattern D: Both wrong			
TNFAIP3 is a tumor suppressor in glioblastoma.	REF	× SUP	× SUP
Low nucleosome occupancy correlates with low methylation.	REF	× SUP	× SUP

Table 13: Representative error examples. REF = REFUTES, SUP = SUPPORTS.

A. Fine-tuned models withhold judgment without evidence; GPT-4o applies world knowledge.

B. GPT-4o detects subject inversions and non-significant p-values; Mistral closes this gap vs. Phi-3-mini (11 wins vs. 29).

C. Mistral infers directional implications from indirect evidence where GPT-4o hedges to NEI.

D. Both models fail on logical negations and directional correlation signs (Wadden et al., 2020).

I Confusion Matrices

Figure 1 visualizes the NEI shortcut directly. BioLinkBERT trained on SciFact achieves perfect

NEI separation in-domain (left). The NEI column is completely clean with zero misclassifications. When evaluated on HealthVer OOD (right), the model predicts NEI exactly once across 727 true NEI examples, instead routing all NEI instances to SUPPORTS (647) or REFUTES (79). This collapse is the behavioral signature of shortcut learning: the absence signal that perfectly predicted NEI in SciFact is absent in HealthVer, and the model has learned nothing else.

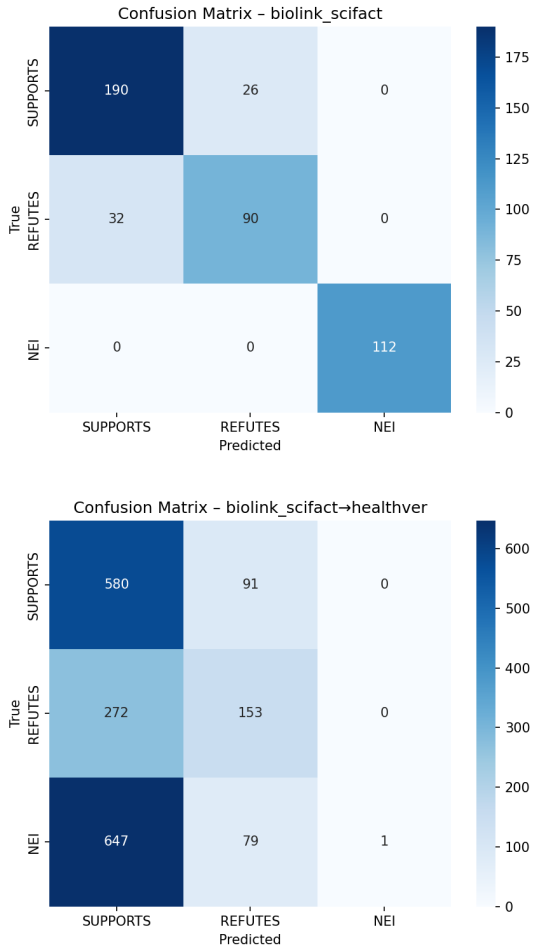

Figure 1: Confusion matrices for BioLinkBERT trained on SciFact, evaluated in-domain (top) and OOD on HealthVer (bottom). In-domain NEI is perfectly classified (112/112); OOD NEI is nearly never predicted (1/727), with true NEI examples overwhelmingly misclassified as SUPPORTS. This is the behavioral signature of structural shortcut learning.