Metadata Predictability Is Not Evidence Dependence: An Intervention-Based Audit for Weak-Label Benchmarks

Kan Shao [1]

Abstract

We study a protocol-level test for weak-label benchmarks: whether benchmark outputs change when the provided evidence is intervened on. Metadata-only shortcut checks answer a different question, namely whether outputs are predictable from metadata priors. We therefore combine a metadata statistic, the Metadata Prior Dominance Score (MPDS), with an evidence-intervention statistic, ΔEvi, measuring sensitivity to evidence identity under cross-item shuffling. Synthetic HotpotQA gives a constructed counterexample to metadata-only screening: MPDS is only moderate (0.643), yet ΔEvi is zero. Stronger-reader reruns show why calibration belongs in the test procedure: SNLI shows a calibration reversal, reconstructed HotpotQA occupies a question-dominant warning region, and FEVER is a strongly evidence-sensitive positive control across four transformers. The practical lesson is simple: benchmark audits should report metadata-only screening, evidence intervention, and reader-strength calibration together.

1. Introduction

Consider a weak-label protocol that generates benchmark labels from heuristics over query types, answer forms, or claim patterns. The usual audit question is whether those labels are predictable from metadata alone. For evidence-based evaluation, the sharper question is: *does the protocol actually depend on the provided evidence?* This concern fits broader critiques of benchmark validity and evaluation practice in NLP (Ethayarajh & Jurafsky, 2020; Bowman & Dahl, 2021; Kiela et al., 2021). It also matches recent work that treats robustness and factuality evaluation as statistical testing problems (Rauba et al., 2025; Nie et al., 2025) and develops testing tools such as e-values for prediction-assisted inference (Csillag et al., 2025).

We frame this mismatch as a hypothesis-testing problem. Metadata-only predictability is informative, but it does not test the null hypothesis that protocol behavior is invariant to evidence identity. We therefore combine MPDS, a metadata-prior accuracy ratio, with the evidence-shuffling statistic ΔEvi, which measures how much performance changes when evidence identity is broken. The audit therefore returns two decision statistics: MPDS for metadata predictability and ΔEvi for evidence dependence.

This framing adds a protocol-level layer to familiar dataset-artifact and model-shortcut analyses. Dataset-artifact work asks whether data collection leaves shortcut cues (Gururangan et al., 2018). Model-shortcut analyses ask whether systems exploit heuristic cues (McCoy et al., 2019; Ribeiro et al., 2020). We ask a third question: whether the *evaluation protocol itself* rewards metadata recovery rather than evidence use. This connects weak-label benchmark construction to data programming, weak-supervision systems, and weak-supervision benchmark suites (Ratner et al., 2016; 2017; Zhang et al., 2021; 2024), but focuses on evaluation integrity rather than only label efficiency. It also complements reproducibility-oriented benchmark audits and partial-identification approaches to weak-supervision evaluation (Calamai et al., 2025; Polo et al., 2024).

The empirical result is a compact diagnostic map. Synthetic HotpotQA yields a constructed latent-coupling counterexample: MPDS is moderate, yet ΔEvi is zero. Real-benchmark audits show why stronger-reader calibration belongs in the test procedure: SNLI shows a calibration reversal, and reconstructed HotpotQA highlights a warning region. FEVER provides a positive control: its labels are clearly evidence-sensitive under calibrated readers. OOD and counterfactual analyses show that unstable protocol behavior can also have downstream consequences.

Contributions. We make three contributions: we separate metadata predictability from evidence dependence, define a two-statistic audit using MPDS and ΔEvi, and show a compact diagnostic map with four illustrative cases: constructed counterexample, calibration reversal, positive control, and

[1]Jinglue Technology Development (Nanjing) Co., Ltd., Nanjing, China. Correspondence to: Kan Shao <shaokan1991@gmail.com>.

Proceedings of the ICML 2026 Workshop on Hypothesis Testing, PMLR, 2026. Copyright 2026 by the author(s).

warning case.

2. A Two-Statistic Test for Protocol Dependence

The test has two axes. Let $\mathrm{Acc}_{\mathrm{meta}}$ be the accuracy of a metadata-majority predictor and $\mathrm{Acc}_{\mathrm{full}}$ be the accuracy of the full audited system. MPDS normalizes metadata predictability by full-system accuracy. For evidence dependence, the null hypothesis is that behavior is invariant to evidence identity:

$$H_0 : \mathrm{Acc}_{\mathrm{full}} = \mathrm{Acc}_{\mathrm{shuf}},$$
$$\mathrm{MPDS} := \mathrm{Acc}_{\mathrm{meta}}/\mathrm{Acc}_{\mathrm{full}},$$
$$\Delta\mathrm{Evi} := \mathrm{Acc}_{\mathrm{full}} - \mathrm{Acc}_{\mathrm{shuf}},$$

where $\mathrm{Acc}_{\mathrm{shuf}}$ uses a cross-item evidence permutation while queries and labels are fixed. This is a paired intervention on evidence identity: the question and target label remain the same, but the evidence attached to the item is replaced. Near-zero ΔEvi indicates invariance to evidence identity; positive ΔEvi indicates evidence-sensitive behavior. In practice we estimate $\mathrm{Acc}_{\mathrm{shuf}}$ over K independent evidence permutations and report the mean and per-permutation population standard deviation σ_{shuf} (population SD, not standard error); it is small because the intervention collapses predictions to a near-constant distribution across permutations (under shuffled evidence the $K = 8$ FEVER model abstains on 1980–1994 of 2000 items in *every* permutation), not by construction. An independent $K = 50$ BERT rerun gives $\sigma_{\mathrm{shuf}} \approx 0.004$ for both FEVER and SNLI—roughly $3\times$ the ≈ 0.0015 that a standard error of the mean would give, and below the i.i.d. single-shuffle bound (≈ 0.010)—so the reported deviations are genuine per-permutation SDs, not mislabeled SEMs. We use $K = 8$ for the reported sweep and recommend $K \geq 20$ for production audits.

Together, these statistics define a diagnostic map with three illustrative regions: **direct coupling** (high MPDS, near-zero ΔEvi), **latent coupling** (moderate MPDS, near-zero ΔEvi), and **evidence-sensitive protocols** (clearly positive ΔEvi). These are not exhaustive categories but useful signposts. The latent-coupling region is the critical one: the metadata screen is not dominant enough to look trivial, but the evidence intervention still shows no dependence on evidence identity. MPDS as a ratio conflates metadata strength with task difficulty (e.g., $(0.5, 0.5)$ and $(0.8, 0.8)$ both give 1.0). A chance-corrected form $\mathrm{MPDS}_+ = (\mathrm{Acc}_{\mathrm{meta}} - c)/(\mathrm{Acc}_{\mathrm{full}} - c)$ (c the majority rate) need not track it: synthetic HotpotQA has ratio 0.643 but near-zero lift over chance ($\mathrm{MPDS}_+ = 0.03$), while NQ is 1.0 under both, and MPDS_+ is itself ill-conditioned when $\mathrm{Acc}_{\mathrm{full}} \approx c$. We report the ratio as the screen and read both only alongside ΔEvi. For real benchmarks, we treat lightweight TF-IDF+LR as a *screening layer*, stronger transformer reruns as a *calibration layer*, and OOD or counterfactual analyses as *consequence evidence*. The decision rule is deliberately operational: a near-zero ΔEvi at the screening layer triggers calibration; persistent near-zero ΔEvi after calibration is a warning region; consistently positive ΔEvi rejects evidence invariance for the audited reader family.

We instantiate the audit on controlled synthetic HotpotQA (a constructed counterexample built to exhibit latent coupling) and three evidence-bearing benchmark settings: SNLI (Bowman et al., 2015), FEVER (Thorne et al., 2018), and reconstructed HotpotQA (Yang et al., 2018). Reconstructed HotpotQA uses the HuggingFace `fullwiki` config (train = 2000, eval = 600) with all retrieved Wikipedia paragraphs as evidence; labels use a heuristic over question type, answer type, and supporting-fact count. Calibration uses four transformer families where shown, and input ablations separate evidence sensitivity from residual query- or hypothesis-side signal. For a new weak-label benchmark, the audit is a four-step packet: specify the metadata schema used by the protocol, compute MPDS as a metadata-only screen, estimate ΔEvi under paired evidence shuffles, and rerun near-zero cases with stronger readers plus input ablations.

3. Results

3.1. Controlled and Real-Benchmark Test Outcomes

The constructed counterexample is synthetic HotpotQA, where MPDS is only 0.643 but ΔEvi is zero. This is the key latent-coupling case: metadata-only screening would look moderately reassuring, yet evidence identity has no measured effect. The synthetic suite also supplies endpoints for the decision map: a synthetic NQ-style task is a direct-coupling stress test (MPDS $= 1.0$, ΔEvi $= 0$), while a synthetic TriviaQA-style task is evidence-sensitive (ΔEvi $= 0.808$).

The lightweight real-benchmark layer is mixed. SNLI and reconstructed real HotpotQA both show near-zero ΔEvi under TF-IDF+LR, but for different reasons: a weak-reader limitation in SNLI, and question-dominant collapse under severe skew in reconstructed HotpotQA. FEVER, by contrast, is already positive under LR. The lightweight layer is therefore a screening stage rather than the final decision.

These four cases span the diagnostic spectrum, from the latent-coupling counterexample through calibration reversal and the question-dominant warning region to the positive control.

3.2. Calibration Changes the Test Conclusion

SNLI is the clearest calibration example. Under the lightweight reader, ΔEvi is near zero (-0.010 ± 0.005 over $K = 8$ premise-level shuffles). Under stronger read-

Case	Test outcome	Why it matters
HotpotQA (syn.)	MPDS $= 0.643$, ΔEvi $= 0$	metadata screen can miss evidence independence
SNLI	LR ≈ 0; transformers $= 0.26$–0.37	weak-reader false negative; calibration changes the conclusion
FEVER	LR ΔEvi $= 0.13$; transformers $= 0.63$–0.68	evidence-sensitive positive control
HotpotQA (recon.)	q-only $= 0.975$; BERT-like ≈ 0	query dominance, skew, and reader collapse

Table 1. Decision view of the audit outcomes. MPDS and ΔEvi separate metadata predictability, evidence sensitivity, and reader calibration effects.

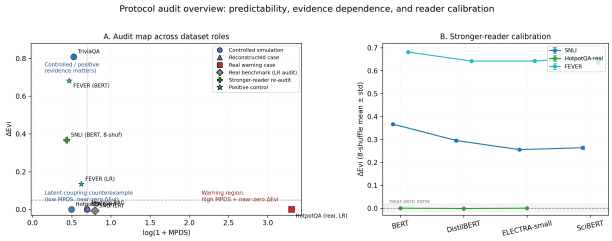

Figure 1. Diagnostic map under the intervention-based audit view. Left: MPDS and ΔEvi separate controlled positive cases, the synthetic HotpotQA latent-coupling counterexample, the FEVER positive control, and the reconstructed HotpotQA warning region. Right: stronger-reader reruns show that SNLI and FEVER are positive after calibration, whereas reconstructed real HotpotQA stays in the warning region.

ers, that conclusion is overturned consistently: multishuffle BERT yields ΔEvi$= 0.3671 \pm 0.0036$, DistilBERT 0.2954 ± 0.0051, ELECTRA-small 0.2558 ± 0.0039, and SciBERT 0.2639 ± 0.0061. Under calibrated readers, SNLI is therefore not evidence-independent. At the same time, SciBERT input ablations show that hypothesis-only signal remains strong (accuracy 0.5975; premise-only 0.3365), so the right conclusion is *evidence-sensitive after calibration, with residual hypothesis-side signal*.

FEVER (Thorne et al., 2018) is the positive-control case. Lightweight LR already has ΔEvi$= 0.13$, and the transformer sweep is strongly positive: BERT 0.6813 ± 0.0022, DistilBERT 0.6423 ± 0.0028, ELECTRA-small 0.6428 ± 0.0042, and SciBERT 0.6580 ± 0.0025 with a second SciBERT seed at 0.6338 ± 0.0038. Thus the audit identifies both evidence-insensitive and strongly evidence-sensitive protocols.

Reconstructed HotpotQA remains near-zero across BERT, DistilBERT, and ELECTRA-small (ΔEvi ≤ 0.002, $\sigma_{\text{shuf}} \leq 0.002$). The label distribution is severely skewed (578 FULL vs. 22 CONFLICT in eval), and the question-only baseline reaches 0.975, so near-zero ΔEvi reflects question-side collapse rather than clean evidence independence. It remains a warning case.

3.3. Consequences Under Distribution Shift

Protocol coupling has measurable downstream consequences. Under OOD answer-type shift, the synthetic NQ-style task collapses completely; SNLI and both HotpotQA variants show degradation tied to metadata dimensions driving protocol behavior. Counterfactual metadata flips are strongest for the synthetic NQ task (flipping answer-type changes every held-out label with evidence fixed), weaker for HotpotQA due to FULL-label dominance. MPDS-gated filtering is not a reliable fix: on synthetic HotpotQA, removing the dominant high-risk group worsens the OOD gap, showing post-hoc deletion is insufficient once the shortcut is built into the protocol.

4. Discussion

The main lesson is methodological: metadata predictability is not evidence dependence, so an audit should report a metadata screen, an evidence-intervention statistic, and a calibrated stronger-reader rerun together—stating the metadata schema, shuffle count, and reader family—rather than a single shortcut baseline.

5. Limitations

Our sweep is budget-limited (four transformers, $K = 8$ permutations; $K \geq 20$ is preferable). Metadata features are hand-designed, so higher-order couplings may escape detection. MPDS as a simple ratio conflates metadata strength with task difficulty. The three-region diagnostic map is illustrative, not exhaustive. The synthetic HotpotQA counterexample is constructed, and we found no natural benchmark in the latent-coupling region; reconstructed HotpotQA is shaped by severe label skew (96% majority). The framework targets evidence-identity sensitivity rather than semantic reasoning quality.

References

Bowman, S. R. and Dahl, G. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015. URL https://aclanthology.org/D15-1075.pdf.

Calamai, T., Balalau, O., and Suchanek, F. M. Benchmarking the benchmarks: Reproducing climate-related nlp tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17967–18009, 2025. doi: 10.18653/v1/2025.findings-acl.925. URL https://aclanthology.org/2025.findings-acl.925/.

Csillag, D., Struchiner, C. J., and Goedert, G. T. Prediction-powered e-values. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rkegUc8d0c.

Ethayarajh, K. and Jurafsky, D. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, 2020. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393/.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. URL https://aclanthology.org/N18-2017.pdf.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1334.pdf.

Nie, F., Hou, X., Lin, S., Zou, J., Yao, H., and Zhang, L. Facttest: Factuality testing in large language models with finite-sample and distribution-free guarantees. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=tuKwODJ08b.

Polo, F. M., Maity, S., Yurochkin, M., Banerjee, M., and Sun, Y. Weak supervision performance evaluation via partial identification. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f4c6bec746b0aeca8c2cd15096f1ad1f-Paper-Conference.pdf.

Ratner, A., De Sa, C., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, 2016. URL https://papers.neurips.cc/paper_files/paper/2016/file/6709e8d64a5f47269ed5cea9f625f7ab-Paper.pdf.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11 (3):269–282, 2017. doi: 10.14778/3157794.3157797. URL https://www.vldb.org/pvldb/vol11/p269-ratner.pdf.

Rauba, P., Wei, Q., and van der Schaar, M. Statistical hypothesis testing for auditing robustness in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ECayXPDoha.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. 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*, 2018. URL https://aclanthology.org/N18-1074.pdf.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. URL https://aclanthology.org/D18-1259/.

Zhang, J., Yu, Y., Li, Y., Wang, Y., Yang, Y., Yang, M., and Ratner, A. Wrench: A comprehensive benchmark for weak supervision. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Zhang, T., Cai, L., Li, J., Roberts, N., Guha, N., and Sala, F. Stronger than you think: Benchmarking weak supervision on realistic tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2024.