Evidence Absence Is Not Evidence Insufficiency: Diagnosing NEI Construction Artifacts in Fact Verification

Jingxi Qiu[1,2], **Zeyu Han**[2], **Cheng Huang**[1,†]
[1]ZenWeave AI, [2]Georgetown University, [†]Corresponding Author
{jingxi,chenghuang}@zenweaveai.com

Abstract

Evidence absence is not evidence insufficiency, but fact verification benchmarks can make them observationally similar. The NOT ENOUGH INFORMATION (NEI) label is often operationalized through constructed evidence conditions, and that choice silently determines what a verifier learns. We introduce NEI-CAP, a construction-aware diagnostic protocol for insufficient-evidence evaluation. Each NEI example carries the construction family that produced it; NEI-CAP audits shortcut cues, validates hard cases through human adjudication, and tests whether competence transfers across constructions. We instantiate the protocol on SciFact, with FEVER and HoVer as bounded external controls. Across these settings, NEI competence does not transfer reliably: encoder verifiers and an instruction-tuned decoder trained on shortcut-prone constructions fail to recognize semantically related insufficient evidence, and mixed-construction training narrows but does not close the gap. Fixed-claim diagnostics further show that the evidence condition shifts confidence in the reference SUPPORT/REFUTE label, not only NEI recall, so an aggregate NEI score can hide which problem a model has actually solved. We therefore recommend reporting the construction family alongside the score, and distill the results into a checklist for benchmarks that carry an insufficient-evidence label.

1 Introduction

A fact verification system labels a claim as supported, refuted, or NOT ENOUGH INFORMATION (NEI) when the available evidence is inconclusive (Thorne et al., 2018; Wadden et al., 2020; Jiang et al., 2020). The NEI label is meant to be evidence-conditioned: for a claim c and an evidence set E, it means that E does not establish c either way. Building those negative evidence sets, however, is itself a design step that no formal definition covers. An empty field, an off-topic passage, a high-overlap retrieval miss, and a non-rationale sentence drawn from a cited document can all be labeled NEI, and a verifier trained on one of these constructions can predict NEI for reasons that have little to do with whether the evidence is actually sufficient.

Prior work on fact-verification artifacts has focused on the claim side. Schuster et al. (2019) show that FEVER can be partially solved by claim-only classifiers, and adversarial or contrastive verification resources further show that standard evidence-aware accuracy can hide brittle decision rules (Thorne and Vlachos, 2019; Schuster et al., 2021). A separate line studies evidence sufficiency by removing parts of otherwise-valid evidence (Atanasova et al., 2022). We study a complementary failure mode—how the negative evidence condition is built in the first place—and argue that this construction silently determines what a verifier learns and what an aggregate NEI-F1 can hide.

Figure 1 illustrates the mechanism. Easy NEI can be solved by recognizing absence, format, or topic mismatch; hard NEI keeps the evidence related to the claim but incomplete, which can induce false SUPPORT when the model mistakes overlap for sufficiency. NEI-CAP targets this gap by making construction family part of the evaluation record rather than treating NEI as a construction-free class.

NEI-CAP makes the construction explicit. Each NEI example carries the family of evidence condition that produced it, which lets us audit each family for shortcut features and stress-test whether a model trained on one family recognizes insufficiency in another. Across these variants, only the NEI evidence condition changes; the label space and claim distribution stay the same, so the comparisons measure construction sensitivity rather than domain transfer. We instantiate the protocol on SciFact-style scientific verification (Wadden et al., 2020), with FEVER (Thorne et al., 2018),

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 16862–16885
October 24-29, 2026 ©2026 Association for Computational Linguistics

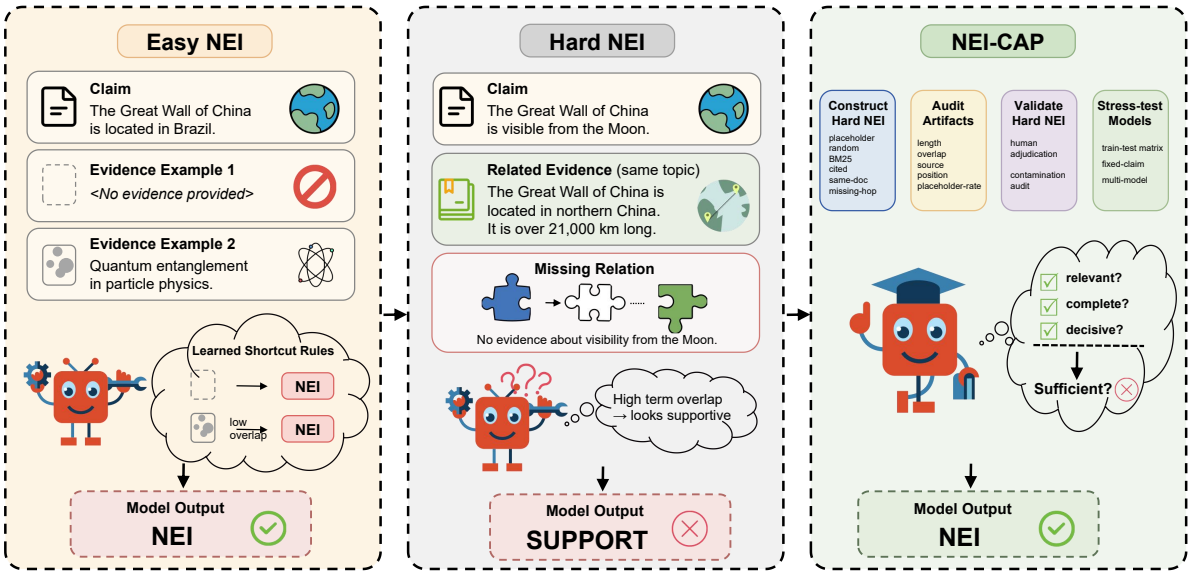


Figure 1: Conceptual illustration of NEI construction artifacts. Easy NEI constructions, such as placeholders or unrelated passages, can teach absence, format, or topic-mismatch shortcuts. Hard NEI keeps the evidence semantically related but incomplete; a verifier may therefore overpredict SUPPORT from overlap rather than recognize insufficiency. NEI-CAP records the construction family, audits shortcuts, validates hard examples, and stress-tests whether NEI competence transfers. Examples are schematic; experiments use the SciFact, FEVER, and HoVer constructions described in Sections 3–4.

HoVer (Jiang et al., 2020), and the broader fact-verification literature as context (Augenstein et al., 2019; Aly et al., 2021).

The headline finding is a transfer failure that aggregate NEI-F1 cannot detect. A DeBERTa verifier trained on placeholder NEI reaches perfect matched-placeholder NEI-F1 across five seeds, yet scores zero NEI-F1 on BM25 near-miss and cited non-rationale evaluation; the collapse replicates on RoBERTa and SciBERT. Probability mass shifts to SUPPORT and REFUTE rather than to NEI, so the failure reflects high-confidence misclassification rather than abstention. Training on random-irrelevant NEI fares only marginally better, showing that the problem extends beyond placeholder detection. Mixed-construction training narrows but does not close the gap. The failure is not specific to encoders. An instruction-tuned decoder handles hard NEI well zero-shot, but after fine-tuning on placeholder NEI alone it collapses in the same way, in every seed; fine-tuning on mixed constructions avoids most of the damage. A fixed-claim diagnostic further shows that swapping the evidence shifts confidence in the reference SUPPORT or REFUTE label, not only NEI recall. The construction choice therefore affects the verifier on the full three-way task, not only the NEI corner.

We make three contributions. *First*, we recast NEI as a construction-sensitive evidence condition rather than a single negative label. *Second*, we introduce NEI-CAP: a diagnostic protocol that treats the construction family as an explicit evaluation variable, audits its shortcut surface, and validates hard cases through human adjudication. *Third*, our SciFact, FEVER, and HoVer experiments show that shortcut-prone training fails to transfer to semantically related insufficient evidence, and that multi-seed and mixed-construction protocols do not remove the need for construction-stratified reporting.

2 Related Work

2.1 Fact Verification Benchmarks

Fact verification is commonly formulated as a three-way classification problem: given a claim and an evidence set, predict whether the evidence supports, refutes, or is insufficient to verify the claim. FEVER introduced a large-scale Wikipedia benchmark with SUPPORTED, REFUTED, and NOTENOUGHINFO labels (Thorne et al., 2018); SciFact extended the formulation to expert-written scientific claims that require retrieving evidence-containing abstracts and rationales (Wadden et al., 2020); and HoVer added many-hop evidence retrieval, where verification can depend on facts spread across multiple Wikipedia articles (Jiang

et al., 2020). MultiFC broadens fact-checking to real-world multi-domain claims (Augenstein et al., 2019); HealthFC focuses on evidence-backed medical claims (Vladika et al., 2024); FEVEROUS adds structured table evidence (Aly et al., 2021); VitaminC creates contrastive claim–evidence pairs that require sensitivity to small factual changes (Schuster et al., 2021); AVeriTeC collects real-world fact-checks whose verdict set includes a not-enough-evidence category (Schlichtkrull et al., 2023); and SciFact-Open extends scientific claim verification to an open corpus of abstracts (Wadden et al., 2022). Rationale-centered resources further ask whether systems identify the evidence used to support predictions (DeYoung et al., 2020). Across these benchmarks the NEI label is treated as a fixed third class, but how its evidence side is built is left to each benchmark's discretion—and that choice is not made part of the evaluation protocol.

2.2 Dataset Artifacts and Behavioral Evaluation

NLP benchmarks routinely contain artifacts that let models score well without learning the intended capability, so aggregate scores can overstate progress (Bowman and Dahl, 2021). In natural language inference, hypothesis-only classifiers can recover the label from premise-free input (Gururangan et al., 2018; Poliak et al., 2018), and controlled challenge sets such as HANS show that high benchmark accuracy can mask reliance on lexical or syntactic heuristics (McCoy et al., 2019). Similar shortcut effects appear beyond NLI, including argument reasoning artifacts (Niven and Kao, 2019); more broadly, shortcut learning is a known failure mode of modern neural systems (Geirhos et al., 2020). Multi-hop datasets show the same pattern: compositional questions can often be answered in a single hop (Min et al., 2019), and models can exploit disconnected reasoning instead of combining facts across passages (Trivedi et al., 2020). In fact verification, Schuster et al. (2019) show analogous claim-side cues in FEVER and demonstrate that claim-only baselines remain competitive against evidence-aware models; FEVER2.0-style adversarial work further stresses robustness to perturbations (Thorne and Vlachos, 2019). A broader line of work uses contrast sets, counterfactually augmented data, and behavioral testing to expose brittle shortcut reliance (Kaushik et al., 2020; Gardner et al., 2020; Ribeiro et al., 2020). These studies mostly examine claim-side, hypothesis-side, or local decision-boundary artifacts. NEI-CAP carries the same diagnostic stance to the evidence side and asks what shortcuts the construction of NEI evidence itself can teach.

2.3 Evidence Sufficiency and Missing Evidence

A claim can be true or false in the world while the available evidence is still insufficient to settle it, so evidence sufficiency is a distinct question from veracity prediction. Atanasova et al. (2022) make this question operational by removing parts of otherwise-valid evidence and asking whether fact-checking models notice the omission, and rationale evaluation benchmarks ask whether models identify supporting passages rather than only predicting labels (DeYoung et al., 2020). Work on missing counter-evidence similarly argues that evidence availability and sufficiency are assumptions built into fact-checking datasets (Glockner et al., 2022). The same question appears in factuality evaluation for generation, where long-form claims are decomposed into atomic facts and checked against supporting sources (Min et al., 2023) and the granularity of that decomposition itself changes verification outcomes (Lu et al., 2025), and in retrieval-augmented generation or grounding evaluation, where generated claims must be supported by provided context (Niu et al., 2024; Jacovi et al., 2025). Sufficiency has also been studied on the retrieval side, where the question is whether the retrieved context alone could answer the query (Joren et al., 2025). It can further be treated as a property of the evidence set as a whole: missing hops and unresolved conflicts are not detectable when passages are scored independently (Qiu et al., 2026). These studies motivate evidence-sensitive evaluation, but they usually treat the negative or unsupported condition as already given or derived from a valid one. NEI-CAP works in the opposite direction: it asks how the insufficient evidence set was built in the first place, and how that construction determines what a verifier can be said to have learned.

3 NEI-CAP: Construction-Aware NEI Evaluation

Figure 1 motivates NEI-CAP as a way to separate shortcut recognition from evidence-insufficiency recognition. This section formalizes that idea with a construction variable, a compact taxonomy of evidence conditions, and the operational workflow

in Protocol 1.

3.1 Evidence-Conditioned NEI

A verification instance is (c, E, y), where c is a claim, $E = \{e_1, \ldots, e_k\}$ is its evidence set, and $y \in \{\text{SUPPORT}, \text{REFUTE}, \text{NEI}\}$. The NEI label is a property of the pair (c, E) rather than of the claim alone (Thorne et al., 2018; Wadden et al., 2020): whether the evidence is insufficient depends on what evidence is provided. We make this dependence explicit by extending each example with a construction variable,

$$x = (c, E, y, z, g),$$

where z records the family of NEI evidence condition that produced E and g is a grouping identifier that keeps variants of the same claim within a single split. The model never receives z or g; they are diagnostic interventions, used only for auditing, splitting, and stratified reporting.

3.2 NEI Construction Families

NEI-CAP separates shortcut-prone constructions from semantically related insufficient evidence. The former expose format, topic, position, or retrieval shortcuts; the latter test whether evidence that remains related to the claim is still recognized as insufficient. This follows the same motivation as contrast-set and behavioral evaluation: a benchmark should expose when a model succeeds through an unintended decision rule rather than the intended capability (Gardner et al., 2020; Ribeiro et al., 2020; Geirhos et al., 2020). We keep five evidential roles distinct: shortcut anchors, candidate hard constructions, human-adjudicated hard constructions, paired diagnostics, and external controls. Each result is reported with the family that produced it. Table 1 lists the compact taxonomy used in the rest of the paper; full definitions, metadata fields, and shortcut-risk dimensions are in Appendix A.

3.3 Diagnostic Protocol

Protocol 1 lists the five stages that produce the construction-stratified evidence reported in Sections 5–6. The stages share a common output discipline: each one returns a typed artifact that the next stage can consume without re-deriving anything from the raw text.

Family	Evidence condition	Role
Placeholder	Fixed/empty no-evidence marker	Format shortcut anchor
Random irrelevant	Unrelated evidence	Topic-mismatch anchor
Position-biased	Predictable non-rationales	Position/source audit
BM25 near-miss	High-overlap insufficient evidence	Hard NEI
Cited non-rationale	Cited but non-rationale evidence	Hard NEI
Same-document	Same-source non-rationale evidence	Source-controlled NEI
Fixed-claim	Same claim, changed evidence	Evidence-substitution diagnostic
Missing-hop	Multi-hop evidence, required fact removed	External multi-hop control

Table 1: Compact NEI-CAP construction taxonomy.

Protocol 1: NEI-CAP diagnostic workflow

Input: claim–evidence examples (c, E, y) and construction rules.
Output: audit tables, adjudicated subsets, construction-stratified metrics, and claim boundaries.
1. Construct. Assign construction family z and group ID g.
2. Audit. Measure evidence-side shortcut features by label and construction.
3. Validate. Adjudicate candidate hard NEI used for central claims.
4. Stress-test. Evaluate single- and mixed-construction training under construction-stratified tests.
5. Report. Release construction-specific metrics, uncertainty, and claim boundaries.

4 Data, Validation, and Experimental Setup

SciFact is our primary setting (Wadden et al., 2020): it requires evidence-sensitive verification over scientific abstracts, where reference evidence and semantically related but insufficient evidence routinely coexist in the same document. FEVER and HoVer serve as bounded external controls (Thorne et al., 2018; Jiang et al., 2020).

4.1 SciFact Construction Suite

The main SciFact suite instantiates the construction families in Table 1: placeholder, random irrelevant, position-biased, BM25 near-miss, and cited non-rationale. BM25 near-miss examples are obtained with BM25 retrieval (Robertson and Zaragoza,

2009) and then filtered to retain high-overlap but insufficient evidence. The SUPPORT and REFUTE portions of the task are held comparable across variants—every test variant contains 76 SUPPORT, 40 REFUTE, and 61 NEI examples—while only the NEI evidence condition changes. This lets us train a verifier under one NEI construction and evaluate it under another. Splits are group-disjoint, keyed by the original claim or claim–document grouping, so construction variants of the same claim never straddle train and test partitions.

4.2 Human-Adjudicated Hard NEI

Candidate hard NEI examples are useful only if they are actually insufficient: BM25 near-miss, cited non-rationale, and same-document non-rationale evidence can carry implicit support or refutation that the construction rule does not catch. NEI-CAP therefore separates *candidate* hard NEI from *human-adjudicated* hard NEI.

We use two validation resources (Table 2). The SciFact hard-NEI audit adjudicates candidate BM25/cited near-miss examples from the construction suite; the fixed-claim/same-document audit covers examples where the claim or source document is held fixed while the evidence condition changes. Two annotators produced consensus adjudications; five-way primary-label agreement is 94.4% (Cohen's $\kappa = 0.732$), and agreement on the valid-NEI-versus-contamination gate that defines the hard subsets is 94.8% ($\kappa = 0.743$). AI-assisted checks were used only as secondary support. The full protocol, label schema, and agreement analysis are in Appendix D.

The primary human-hard model evaluation uses the held-out test split of the SciFact human-adjudicated hard-NEI audit: 54 validated hard-NEI examples after group-disjoint splitting. The larger audit pool estimates label validity, while the held-out subset supports model evaluation. Appendix G reports the full evaluation table and Appendix B documents the sampling trail.

Validation resource	n	Valid NEI	Contam.	Hard
BM25/cited hard-NEI audit	250	89.2%	10.8%	195 (87.4%)
Fixed-claim/same-doc audit	122	96.7%	3.3%	114 (96.6%)

Table 2: Human validation summary. Contamination includes examples adjudicated as actually supported, actually refuted, ambiguous, or invalid. The Hard column reports the absolute number of human-adjudicated hard NEI examples that pass the audit, with the in-parentheses rate computed against valid NEI. Full adjudication details are in Appendix D.

4.3 Models and Metrics

Our primary verifier is DeBERTa-v3-base (He et al., 2021), with RoBERTa-base and SciBERT as secondary backbones (Liu et al., 2019; Beltagy et al., 2019). We use these as diagnostic probes rather than proposed architectures: training variants differ only in the NEI construction family used during training. Mixed-construction regimes combine the single-construction variants into an easy mixture (placeholder, random irrelevant, position-biased), a hard mixture (BM25 near-miss, cited non-rationale), and a balanced mixture over all five families. Multi-seed experiments use seeds 13, 17, 23, 29, and 37. Qwen2.5-7B-Instruct (Yang et al., 2024) serves as a decoder probe, evaluated zero-shot and under controlled five-seed LoRA fine-tuning; its design is described in Section 5.6 and Appendix N.

For three-way verification, we report Macro-F1 and class-specific F1, especially NEI-F1. For human-validated hard-NEI subsets, every example is adjudicated as NEI, so Macro-F1 is not informative; we instead report NEI recall, false SUPPORT rate, false REFUTE rate, and mean predicted class probabilities. Hyperparameters, checkpoint selection, seed aggregation, and metric details are in Appendices E and L.

4.4 External Controls

FEVER and HoVer are Wikipedia verification datasets; all core experiments run on SciFact. FEVER has a native NOTENOUGHINFO label (Thorne et al., 2018) and serves as a non-toy subset control. HoVer requires multi-hop evidence, so sufficiency can depend on more than one supporting fact (Jiang et al., 2020); it serves as a candidate missing-hop control. Full details are in Appendix J.

5 Results

5.1 Matched NEI Performance Does Not Imply Transfer

Figure 2 shows the SciFact train/test construction matrix. A placeholder-trained verifier obtains perfect matched-placeholder NEI-F1, but falls to 0.000 on BM25 near-miss and cited non-rationale NEI. Position-biased training shows the same hard-construction collapse, while random-irrelevant training is less extreme but still shortcut-prone: it is nearly solved under matched evaluation yet transfers poorly to BM25 and cited hard constructions. Full numeric matrices are in Appendix F.

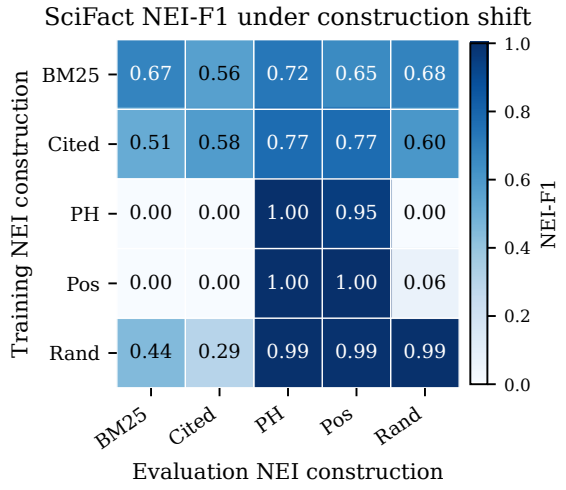


Figure 2: SciFact NEI-F1 under train/test construction shifts. Cell values are seed-aggregated means from the primary construction matrix; Appendix F reports full numeric matrices and Macro-F1. Matched easy NEI can yield high scores without transferring to semantically related insufficient evidence.

The point is not that placeholder evidence is realistic; it is an intentionally shortcut-prone anchor. The broader finding is that different easy constructions teach different shortcuts—format absence, position/source bias, or topic mismatch—whereas the target of NEI-CAP is hard NEI: semantically related evidence that remains insufficient.

5.2 Human-Adjudicated Hard NEI Exposes Easy-Training Failure

Human validation does not rescue easy NEI training. Table 3 shows that placeholder and position-biased training yield 0.000 recall on human-adjudicated hard NEI. Random-irrelevant training also performs poorly (0.216 recall), confirming that topic-unrelated NEI does not teach hard-insufficiency recognition. Errors are dominated by false SUPPORT, and BM25 near-miss and cited non-rationale training partially recover recall.

Train NEI	Recall	False SUP	False REF
BM25 near-miss	0.691	0.142	0.167
Cited non-rationale	0.679	0.074	0.247
Random irrelevant	0.216	0.574	0.210
Placeholder	0.000	0.870	0.130
Position-biased	0.000	0.907	0.093

Table 3: Evaluation on the held-out SciFact human-adjudicated hard-NEI test split ($n = 54$). Since all examples are validated as NEI, we report NEI recall and error rates rather than Macro-F1; Appendix B gives the sampling trail from the larger audit pool.

Placeholder training yields 0/54 recall, with a Clopper–Pearson 95% upper bound of 0.066 (Clopper and Pearson, 1934); BM25 near-miss training gives a bootstrap interval of $[0.593, 0.790]$. Under placeholder training, mean $P(\text{NEI})$ on the validated hard examples is 0.013, with the remaining probability mass on SUPPORT and REFUTE, so the failure is high-confidence misclassification rather than abstention (Appendix G).

5.3 Mixed Training Helps but Remains Construction-Stratified

Mixed training is a stronger test because real benchmarks rarely contain a single NEI construction. Table 4 summarizes DeBERTa under an easy mixture, a hard BM25/cited mixture, and a balanced all-family mixture. Easy-mixture training remains weak on hard constructions; hard and balanced mixtures improve recall but still produce construction-stratified profiles. Mixed training changes the failure profile—it does not make NEI construction-free.

Train regime	BM25	Cited	Hard recall	Takeaway
Easy mixture	0.379	0.273	0.178	Shortcut-heavy mixture remains weak
Hard mixture	0.662	0.652	0.770	Hard training improves recall
Balanced mixture	0.802	0.803	0.915	Best, but still stratified

Table 4: DeBERTa mixed-construction training summary. BM25 and Cited are construction-stratified NEI-F1; Hard recall is one-class recall on the SciFact human-adjudicated hard-NEI subset. Easy mixture includes random-irrelevant and position-biased NEI, so it improves over placeholder-only training but remains weak on hard insufficient evidence.

The balanced mixture improves the hard-side metrics, but the full stratified matrix in Appendix H still shows different performance profiles across placeholder, random-irrelevant, BM25, and cited conditions; mixed training is therefore a stronger stress test, not a replacement for construction-aware reporting.

5.4 Placeholder-to-Hard Collapse Is Stable Across Seeds and Backbones

The placeholder-to-hard collapse is not specific to DeBERTa or to one seed. Across five seeds for DeBERTa, RoBERTa, and SciBERT, each backbone reaches perfect matched-placeholder NEI-F1 and zero NEI-F1 on BM25 near-miss and cited non-rationale evaluation when trained on placeholder NEI (Table 5). That the same exact 1.000→0.000 pattern reproduces across three encoder backbones and fifteen training runs rules out interpreting the collapse as a training-dynamic artifact of any one backbone.

Model	Seeds	PH→PH	PH→BM25	PH→Cited	Drop
DeBERTa	5	1.000	0.000	0.000	1.000
RoBERTa	5	1.000	0.000	0.000	1.000
SciBERT	5	1.000	0.000	0.000	1.000

Table 5: Five-seed placeholder-to-hard robustness. PH denotes placeholder NEI. Scores of 0.000 are exact: in all five seeds, placeholder-trained models never predicted NEI on the BM25 or cited hard test sets, yielding zero NEI-F1.

5.5 Source-Controlled Diagnostics Are Useful but Bounded

The fixed-claim diagnostic is the closest available paired control: across 114 paired cases, the same claim is paired first with reference evidence and then with human-adjudicated insufficient evidence. Holding the claim fixed removes claim-side variation, so behavioral changes arise from evidence substitution; because the paired passages are not feature-matched, this diagnostic does not isolate semantic sufficiency alone. The probability that the verifier assigns to the reference SUPPORT/REFUTE label drops on the insufficient side; construction choice therefore affects how confidently the model commits to the non-NEI labels, not only whether it predicts NEI. Same-document hard NEI provides an additional source-controlled view that retains human adjudication; a shallow-feature audit shows it remains constructionally distinct from BM25/cited hard NEI, so we report it as its own diagnostic family rather than as a universal hard-NEI score. Full fixed-claim and same-document diagnostics are in Appendix I, with row-level examples in Appendix K.

5.6 Zero-Shot Robustness Does Not Survive Placeholder Fine-Tuning

We evaluate Qwen2.5-7B-Instruct (Yang et al., 2024) zero-shot with deterministic conditional label scoring, then fine-tune it with LoRA (Hu et al., 2022) across five seeds, changing only the training NEI realization between regimes and holding everything else fixed. Appendix N reports the design, data controls, and per-regime details.

Qwen regime	PH	Hard	Hum. recall	REF-F1
Zero-shot base	0.847	0.793	0.963	0.532
Placeholder-only SFT	1.000 $\pm$0.000	0.054 $\pm$0.054	0.011	0.645
Hard-mixture SFT	0.920 $\pm$0.013	0.834 $\pm$0.008	0.900	0.784
Balanced-all SFT	0.955 $\pm$0.012	0.817 $\pm$0.018	0.856	0.795

Table 6: Qwen2.5-7B decoder diagnostic. PH is placeholder-test NEI-F1; for the placeholder-only regime this equals the matched-construction score. Hard is mean NEI-F1 over BM25 near-miss and cited non-rationale evaluation; Hum. recall is NEI recall on the held-out human-adjudicated hard-NEI split ($n = 54$). SFT rows are five-seed means with sample standard deviations.

Zero-shot Qwen does not reproduce the encoder collapse (Table 6), but its high human-hard recall comes partly from overpredicting NEI: it predicts SUPPORT/REFUTE/NEI at 45.4/9.3/45.3%, against a gold distribution of 42.9/22.6/34.5%, with REFUTE-F1 of 0.532. Placeholder-only SFT then induces the encoder-style failure in every seed: matched-placeholder NEI-F1 stays at 1.000 while hard NEI-F1 falls to 0.054 and human-hard recall to 0.011. Hard-mixture and balanced SFT recover most of the loss, and they do so by discriminat-

ing the three classes better, not by predicting NEI more often: balanced SFT moves the predicted distribution to 43.0/22.7/34.3%, raises REFUTE-F1 to 0.795, and keeps hard NEI-F1 at 0.817.

A verifier that is robust when prompted can therefore lose that robustness during fine-tuning. Llama-3.1-8B-Instruct (Grattafiori et al., 2024) failed our predeclared label-order stability gate and was not fine-tuned.

5.7 External Controls Mirror the SciFact Pattern

On FEVER and HoVer, a placeholder-trained baseline reaches NEI-F1 of 1.000 on placeholder evaluation and 0.000 on BM25 (FEVER) or missing-hop (HoVer) evaluation, mirroring the SciFact pattern at a different model scale and on a different domain. Full external-control results are in Appendix J.

6 Analysis and Discussion

6.1 NEI Is a Family of Evidence Conditions

Hard NEI should not be treated as one uniform category. BM25 near-miss, cited non-rationale, same-document non-rationale, fixed-claim hard NEI, and missing-hop evidence remove different shortcuts and stress different behaviors. Construction-level audits show that these families differ in evidence length, claim–evidence overlap, coverage, and retrieval metadata (Appendix C). These audits separate evidence-insufficiency evaluation from construction recognition: the families differ in surface features, so results are reported per family rather than folded into aggregate NEI-F1. Random irrelevant evidence is especially important: it can teach topic mismatch rather than insufficiency, demonstrating that the problem is broader than placeholder detection.

6.2 Mixed Training Does Not Remove the Reporting Requirement

Mixed-construction training does not remove construction sensitivity (Section 5.3), so construction-aware reporting remains necessary even for models trained on diverse constructions. NEI-CAP's output is a reporting discipline rather than a better aggregate score: every result carries the construction family that produced it, the audit metrics that describe its shortcut surface, and the validation status of any hard examples it relies on.

6.3 What Fixed-Claim Diagnostics Can and Cannot Show

Fixed-claim evidence substitution separates two abilities: recognizing semantically related evidence as insufficient, and assigning the reference label when decisive evidence is present. The probability-drop metric shows that replacing reference evidence with insufficient evidence lowers confidence in the reference label, so construction affects SUPPORT/REFUTE verification confidence as well as NEI recall.

Same-document hard NEI is human-adjudicated, yet shallow features nearly separate it from BM25/cited hard NEI and other construction families. Human validity and constructional distinctness are different properties: even human-valid hard NEI can carry construction-specific signatures.

6.4 A Reporting Checklist for Practitioners

We distill the diagnostic results into a reporting checklist for any benchmark or system evaluation that includes an insufficient-evidence label. Where dataset documentation records provenance for a resource as a whole (Gebru et al., 2021), this checklist records it per NEI construction. A report should record: (1) *construction provenance*—the family of evidence condition that produced each NEI example; (2) a *shortcut audit* of evidence-side features by label and construction family; (3) *validation status and uncertainty*—which hard examples are human-adjudicated, with intervals when validated subsets are small; (4) *aggregate and stratified metrics*—per-family and worst-family NEI scores alongside any aggregate; (5) *false* SUPPORT *and false* REFUTE *rates*, since confidently accepting a claim on insufficient evidence is the central downstream risk; and (6) the *model and training regime*, because construction sensitivity can be training-induced even in models that look robust zero-shot (Section 5.6). Protocol 1 operationalizes items (1)–(4); Appendix M packages the checklist together with claim boundaries.

Family selection follows the same logic as the diagnosis: the NEI constructions in a benchmark should mirror the negative evidence conditions its retrieval setting actually produces. In SciFact-style scientific verification, related-but-insufficient abstracts motivate the BM25 near-miss and cited non-rationale families; in multi-hop settings such as HoVer, incomplete evidence chains motivate

missing-hop constructions. Whatever the domain, the results support a minimum: one shortcut-prone anchor, kept as an artifact control rather than as competence evidence, and one semantically related hard family validated by human adjudication, since these examples carry the competence claims.

Because NEI is evidence-conditioned, high-stakes deployment calls for construction-stratified regression testing, uncertainty communication, and human escalation paths, not a single aggregate acceptance score.

7 Conclusion

NEI is often treated as a single negative label, but the way it is built decides what a verifier can actually learn. In SciFact-style verification, a placeholder-trained verifier earns a perfect matched-placeholder NEI-F1 yet scores zero on evidence that is topically related but insufficient—a pattern that reproduces across three encoder backbones, five seeds, and two external datasets.

NEI-CAP makes the construction explicit. By attaching every NEI example to the family of evidence condition that produced it and validating the hard cases through human adjudication, the protocol turns insufficient-evidence evaluation from a single number into a construction-stratified report.

The construction of insufficient evidence belongs alongside the labels and metrics that define any fact-verification benchmark. An aggregate NEI-F1 that hides how its NEI examples were built cannot tell a reader whether the model has learned to recognize insufficiency or simply to recognize the artifact.

Limitations

NEI-CAP is a diagnostic protocol. Strong results under it do not guarantee general-purpose sufficiency reasoning, and other domains may need additional construction families and annotation guidelines. The suite also favors internal validity: each construction family isolates one NEI evidence condition, whereas production retrieval mixes missing, irrelevant, partial, stale, and conflicting evidence.

Our strongest evidence comes from SciFact-style scientific verification. FEVER and HoVer show that the placeholder-to-hard pattern is not unique to SciFact, but they are narrower designs: the construction matrix, human adjudication, and training experiments were not repeated on them. Multi-domain expert adjudication of hard NEI remains future work.

The held-out human-adjudicated split is small ($n = 54$). Adjudication is expensive, and the audit pipeline shrinks the pool at every stage: 250 candidates, 223 adjudicated valid NEI, 195 hard, and 54 after group-disjoint splitting. At this size the split can separate large effects, but it cannot rank insufficiency subtypes: 54 examples spread over broad-topic, near-miss, and partial subtypes leave too few per subtype. It also cannot estimate subtype prevalence, because the audit pool was selected for hardness, not sampled to be representative. Cross-domain conclusions would require comparable audits outside SciFact.

Human validation reduces label-validity risk but does not eliminate it. Near-miss and non-rationale evidence can carry implicit support, refutation, or ambiguity that two annotators may both miss. Agreement was weakest on fine-grained insufficiency subtypes (54.9% conditional agreement, $\kappa = 0.363$): annotators reliably agreed on whether an example was hard, but often differed on which kind of insufficiency it exhibited, which is consistent with the ambiguity-driven disagreement documented for claim–evidence pairs by Glockner et al. (2024). Subtype-level analyses should therefore be read as descriptive (Appendix D).

Fixed-claim and same-document diagnostics are decomposition tests. The paired evidence passages are not feature-matched: the reference and insufficient passages can differ in length, style, or specificity, so paired behavioral changes cannot be attributed to sufficiency alone. Same-document hard NEI also remains constructionally distinct under shallow-feature audits, and the matched length/coverage audit was too small to settle how much these surface differences matter.

Our model probes are encoder cross-encoders plus one decoder. The decoder evidence covers Qwen2.5-7B on SciFact under one adaptation method (LoRA); Llama-3.1-8B failed a predeclared label-order stability gate and was not fine-tuned, so we cannot say how far the training-induced failure extends across decoder families, scales, or adaptation methods. Extending NEI-CAP to generative factuality and grounding evaluation (Min et al., 2023; Niu et al., 2024; Jacovi et al., 2025) is future work.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. 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*, pages 4843–4855, Online. Association for Computational Linguistics.

C. J. Clopper and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 541 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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 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, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

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*.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, and 7 others. 2025. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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*.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. 2025. Optimizing decomposition for optimal claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5114, Vienna, Austria. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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*, pages 3428–3448. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664. Association for Computational Linguistics.

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*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Jingxi Qiu, Zeyu Han, and Cheng Huang. 2026. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *arXiv preprint arXiv:2605.03534*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

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 36: Datasets and Benchmarks Track*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina

Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

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, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A NEI-CAP Construction Taxonomy

This appendix expands the compact taxonomy in Table 1. NEI-CAP treats NEI not as a construction-free negative label, but as an evidence-conditioned label whose interpretation depends on how the evidence set is paired with the claim.

A.1 Taxonomy Axes

We characterize each NEI construction family along six axes: evidence availability, topical relatedness, source control, lexical or entity overlap, evidential completeness, and validation status. These axes separate absence-based shortcuts from semantically related insufficient evidence.

Evidence availability. Some NEI examples contain no substantive evidence or use a fixed placeholder. Others contain non-empty evidence that is related to the claim but still insufficient.

Topical relatedness. Random irrelevant evidence may be insufficient because it is off topic. BM25 near-miss, cited non-rationale, same-document non-rationale, and missing-hop examples are instead related to the claim but incomplete.

Source control. Same-document non-rationales and fixed-claim diagnostics reduce source and topic shortcuts by holding the source document or claim fixed.

Lexical and entity overlap. Near-miss examples preserve lexical or entity overlap while removing decisive support, testing whether models mistake overlap for sufficiency.

Evidential completeness. Partial and missing-hop examples contain some relevant information while omitting a necessary relation, condition, or reasoning step.

Validation status. NEI-CAP distinguishes constructed examples, candidate hard NEI, and human-adjudicated hard NEI. The strongest hard-NEI claims are restricted to human-adjudicated subsets.

A.2 Construction Families

Family	Construction rule	Shortcut risk	Validation status	Paper role
Placeholder	Replace evidence with a fixed or empty no-evidence marker.	Format, length, absence.	Constructed only.	Shortcut baseline.
Random irrelevant	Pair claim with evidence from unrelated claims or documents.	Topic mismatch, low overlap.	Constructed only.	Topic-mismatch baseline.
Position-biased	Select non-rationale evidence from predictable locations.	Sentence position, source distribution.	Not primary human-validated.	Artifact stress test.
BM25 near-miss	Retrieve high-overlap evidence that remains insufficient.	Lexical overlap mistaken for support.	Human-adjudicated in SciFact audit.	Hard NEI.
Cited non-rationale	Use evidence from claim-associated cited documents that is not the rationale.	Source relevance mistaken for sufficiency.	Human-adjudicated in SciFact audit.	Hard NEI.
Same-document non-rationale	Use non-rationale evidence from the same source document.	Residual topic/source cues.	Human-adjudicated in fixed-claim/same-document audit.	Source-controlled hard NEI.
Fixed-claim hard NEI	Pair the same claim with reference evidence and validated insufficient evidence.	Evidence-side differences may still contain shallow cues.	Human-adjudicated in fixed-claim/same-document audit.	Evidence-subst. diagnostic.
Missing-hop control	Remove a required supporting fact from multi-hop evidence.	Partial evidence over-interpreted as sufficient.	Candidate-only in HoVer.	External multi-hop control.
FEVER subset control	Construct bounded Wikipedia verification controls with alternative NEI evidence.	Dataset-specific lexical or retrieval artifacts.	Subset control, not human-validated hard NEI.	External shortcut control.

Table 7: Full NEI-CAP construction taxonomy. The taxonomy separates shortcut-prone baselines, candidate hard NEI, human-adjudicated hard NEI, and bounded external controls.

A.3 Construction Strength

We use three construction-strength categories.

Easy NEI. Placeholder, random irrelevant, and position-biased examples are shortcut-prone by design. They are useful probes of artifact sensitivity but should not be treated as strong evidence-sufficiency tests.

Candidate hard NEI. Candidate hard examples are constructed to be semantically related but insufficient. Before adjudication, they may contain semantic contamination.

Human-adjudicated hard NEI. Human-adjudicated hard examples are candidate hard examples labeled as truly insufficient rather than actually supportive, actually refuting, ambiguous, or invalid. These subsets support the strongest hard-NEI claims in the paper.

A.4 Terminology and Claim Boundaries

Avoid	Use instead	Reason
hard-negative	hard NEI example	Avoids retrieval/contrastive ambiguity.
gold-truth	gold label / reference label	Standard terminology.
gold-side	reference-evidence side	More precise.
hard-side	insufficient-evidence side	Defines evidence condition.
counterfactual proof	fixed-claim evidence-substitution diagnostic	Avoids causal overclaim.
HoVer human-validated	HoVer candidate missing-hop control	No HoVer human audit.
FEVER full validation	FEVER subset control	Bounded external probe.

Table 8: Terminology mapping used throughout the paper.

The taxonomy supports three bounded claims: NEI evaluation is construction-sensitive; easy NEI can inflate apparent competence; and hard insufficient evidence should be validated when it supports central claims.

B Dataset Construction and Manifests

This appendix documents the dataset assets and manifest requirements behind NEI-CAP. The main paper reports only compact descriptions; the appendix records the construction metadata needed for reproducible construction-aware evaluation.

B.1 Example Representation

Each example is represented as:

$$x_i = (c_i, E_i, y_i, z_i, g_i, m_i),$$

where c_i is the claim, E_i is the evidence set, $y_i \in \{\text{SUPPORT}, \text{REFUTE}, \text{NEI}\}$ is the label, z_i is the construction family, g_i is a group identifier, and m_i stores provenance and audit metadata. The construction variable z_i is not given to the model; it is used for splitting, auditing, and reporting.

B.2 Manifest Schema

Field	Type	Description
example ID	string	Unique claim–evidence instance ID.
claim ID	string	Original claim ID.
group ID	string	Group key for group-disjoint splitting.
source data	string	SciFact, FEVER, HoVer, or derived resource.
claim	string	Claim text.
evidence	string/list	Evidence text or list of evidence units.
label	categorical	SUPPORT, REFUTE, or NEI.
construction	categorical	Placeholder, BM25 near-miss, same-document, etc.
split	categorical	Train, development, test, or audit.
document ID	string/list	Source document IDs when available.
sentence IDs	string/list	Sentence or rationale IDs when available.
retrieval method	string	Retrieval or sampling method.
retrieval rank	integer	Rank of retrieved evidence.
BM25 score	float	Retrieval score, when applicable.
sentence position	integer/list	Sentence position within document.
validation status	categorical	Not validated, candidate, valid NEI, contaminated, or ambiguous.
adjudicated label	categorical	Final adjudicated label if available.

Table 9: Recommended NEI-CAP manifest schema. Construction metadata is required to report results by evidence condition rather than by label alone.

B.3 SciFact Suite

The core SciFact suite contains five variants: placeholder, random irrelevant, position-biased, BM25 near-miss, and cited non-rationale. The SUPPORT and REFUTE portions remain comparable across variants; only the NEI evidence condition changes. This design supports train/test construction-shift evaluation.

B.4 Group-Disjoint Splitting

Construction variants derived from the same claim or claim–document group must not leak across train and test. NEI-CAP therefore uses group-disjoint splitting keyed by claim, claim–document pair, or fixed-claim substitution group where applicable. This prevents a model from exploiting claim memorization across evidence variants.

B.5 Human-Adjudicated Assets

Human-adjudicated assets are separated from automatically constructed candidate data. The SciFact audit validates BM25/cited candidate hard NEI; the fixed-claim/same-document audit validates same-claim and same-document diagnostics. These assets estimate semantic contamination and define human-adjudicated hard-NEI subsets for model evaluation.

B.6 Human-Hard Evaluation Sampling Trail

The SciFact human-audit pool estimates label validity, while the model evaluation uses only the held-out group-disjoint test split. Table 10 records the paper-facing sampling trail used to interpret the $n = 54$ hard-NEI evaluation in Table 3.

Stage	Count
Candidate BM25/cited hard-NEI audit pool	250
Human-valid NEI after adjudication	223
Human-adjudicated hard-NEI subtype	195
Held-out group-disjoint model-evaluation split	54

Table 10: Sampling trail for the SciFact human-hard model-evaluation subset. The larger audit pool estimates label validity; the held-out split supports model evaluation.

B.7 External Controls

FEVER and HoVer are included as bounded external controls. FEVER provides a non-toy Wikipedia subset control. HoVer provides a candidate missing-hop control for multi-hop insufficiency. Neither is treated as human-validated hard NEI.

B.8 Construction Split Statistics and Leakage Audit

Construction-level split statistics are recorded for each paper-facing family: label counts, claim and document group counts, evidence length, sentence count, placeholder rate, claim–evidence overlap, coverage, retrieval metadata, duplicate counts, and missing-field checks. Table 11 gives a compact test-split view for the core SciFact suite; full CSV artifacts will be released with the accompanying code and data package.

Variant	n	SUP/REF/NEI	Avg. tok.	Coverage
Placeholder	177	76/40/61	174.7	0.444
Random irrelevant	177	76/40/61	236.8	0.474
Position-biased	177	76/40/61	181.7	0.501
BM25 near-miss	177	76/40/61	251.4	0.630
Cited non-rationale	177	76/40/61	246.9	0.590

Table 11: Compact test-split statistics for the SciFact construction suite. Each variant keeps comparable SUPPORT/REFUTE/NEI label counts while changing the NEI evidence condition.

The leakage audit reports zero claim-group overlap across train/development/test for the core SciFact construction variants and zero construction-variant cross-split leakage. Document overlap can occur because the same scientific paper may be relevant to multiple claims; we report it as source-distribution metadata rather than claim leakage.

B.9 Asset Map

A future public release will link each reported result to its source dataset, construction family, split policy, model configuration, seed set, prediction artifact, and evaluation output. The full machine-readable asset map will be provided with the released manifests.

C Artifact Audit Statistics

NEI-CAP audits construction families before interpreting model performance. The audit asks whether NEI examples can be separated using superficial evidence-side features rather than evidence-sufficiency reasoning.

C.1 Audited Features

We audit evidence length, number of evidence sentences, claim–evidence lexical overlap, claim–evidence coverage, placeholder rate, sentence position, source concentration, and discourse markers. These features do not determine semantic validity; they identify shortcut risk.

C.2 Construction-Level Summary

Construction group	n	Avg. sent.	Avg. tok.	Coverage	Status
SciFact BM25/cited hard NEI	195	11.49	216.56	0.483	Human-adjudicated
Same-document hard NEI	114	1.04	23.93	0.233	Human-adjudicated
Fixed-claim hard NEI	114	1.23	23.93	0.233	Human-adjudicated
Placeholder NEI	61	1.00	2.00	0.005	Constructed only
Random irrelevant NEI	61	9.70	182.46	0.092	Constructed only

Table 12: Artifact audit summary by construction group. Placeholder NEI has extremely short evidence and near-zero coverage; human-adjudicated hard NEI contains substantive evidence.

C.3 Overlap and Marker Statistics

Construction group	Overlap count	Jaccard	Context marker	Method marker
SciFact BM25/cited hard NEI	6.24	0.053	0.210	0.226
Same-document hard NEI	3.06	0.098	0.316	0.140
Fixed-claim hard NEI	3.06	0.098	0.316	0.140
Placeholder NEI	0.05	0.005	0.000	0.000
Random irrelevant NEI	1.31	0.013	0.148	0.213

Table 13: Claim–evidence overlap and marker statistics. Random irrelevant NEI contains long evidence but low overlap, indicating a topic-mismatch shortcut rather than evidence insufficiency.

C.4 Hard-NEI Subtypes

Group	Broad topic	Near-miss	Partial
SciFact BM25/cited hard NEI	62	102	31
Fixed-claim/same-document hard NEI	50	49	15

Table 14: Subtype distribution for human-adjudicated hard NEI. Hard NEI is heterogeneous rather than a single condition.

C.5 Interpretation

The audit supports three conclusions. First, placeholder NEI exposes strong format and absence cues. Second, random irrelevant NEI mainly tests topic mismatch. Third, human-adjudicated hard NEI is heterogeneous across BM25/cited, same-document, and fixed-claim settings. These findings motivate construction-specific reporting rather than aggregate NEI-F1 alone.

D Human Validation Protocol and Adjudication

Hard insufficient evidence can be noisy. Retrieved near-misses, cited non-rationales, and same-document non-rationales may contain implicit support, implicit refutation, or ambiguity. NEI-CAP therefore separates automatically constructed candidate hard NEI from human-adjudicated hard NEI.

D.1 Validation Goal

Human validation determines whether a candidate hard-NEI example is truly insufficient. The resulting labels are used to estimate semantic contamination and to define human-adjudicated hard-NEI subsets for evaluation.

D.2 Annotation Labels

Label	Definition
`truly_insufficient`	Evidence does not support or refute the claim.
`actually_supported`	Evidence supports the claim.
`actually_contradicted`	Evidence refutes the claim; reported as REFUTE in paper-facing terminology.
`ambiguous`	Status cannot be determined reliably.
`invalid_or_unreadable`	Claim or evidence is malformed, missing, unreadable, or out of scope.

Table 15: Human validation label schema. Contamination includes actually supported, actually contradicted/refuted, ambiguous, invalid, and unreadable cases.

D.3 Reporting Metrics

We report:

$$\text{valid NEI} = \frac{N_{\text{insufficient}}}{N_{\text{audited}}},$$

and:

$$\text{contam.} = \frac{N_{\text{sup}} + N_{\text{ref}} + N_{\text{amb/invalid}}}{N_{\text{audited}}}.$$

AI-assisted checks, where used, are treated only as triage or secondary support, not as human validation.

D.4 SciFact Hard-NEI Audit

Metric	Observed	95% low	95% high
Valid NEI rate	0.892	0.852	0.928
Contamination rate	0.108	0.072	0.148
Actually supported rate	0.056	0.028	0.088
Actually refuted rate	0.052	0.028	0.084
Ambiguous/invalid rate	0.000	0.000	0.000
Hard subtype rate among valid NEI	0.874	0.826	0.919
Topic-unrelated rate among valid NEI	0.117	0.076	0.164

Table 16: SciFact hard-NEI audit validation over 250 candidate hard-NEI examples. Intervals are bootstrap 95% intervals.

D.5 Fixed-Claim and Same-Document Validation

Subset	Rows	Contam.	Hard rows	Hard rate
Final adjudicated labels	122	0.0328	114	0.966
Human-validated NEI	118	0.0000	114	0.966
Human-validated hard NEI	114	0.0000	114	1.000
Same-claim hard NEI	114	0.0000	114	1.000
Same-document hard NEI	114	0.0000	114	1.000

Table 17: Fixed-claim/same-document validation summary. The 114 hard-NEI examples support same-claim and same-document diagnostics.

D.6 Annotator and Adjudication Protocol

Two annotators participated in human validation, and final labels are consensus adjudications. Annotators independently evaluated claim plus candidate evidence for semantic sufficiency. Blinded packets removed LLM labels, gold labels, model predictions, and construction provenance from primary annotation fields. AI-assisted checks were used only as secondary support or triage and are never reported as standalone human validation.

D.7 Inter-Annotator Agreement

Table 18 reports formal agreement on the 250-example SciFact audit, computed from the two independent pre-consensus annotation streams with paired-row bootstrap intervals (10,000 resamples). Each decision is reported separately rather than pooled into a single coefficient (Artstein and Poesio, 2008).

Agreement metric	n	Raw	Cohen's κ	95% CI
Five-way primary label	250	94.4%	0.732	[0.585, 0.854]
Valid NEI vs. contaminated	250	94.8%	0.743	[0.589, 0.866]
Subtype (cond., both NEI)	215	54.9%	0.363	[0.268, 0.454]
Hard vs. non-hard (cond., both NEI)	215	93.0%	0.714	[0.563, 0.837]

Table 18: Inter-annotator agreement on the SciFact hard-NEI audit. The conditional rows are computed only on the subset where both annotators independently assigned NEI; they are never pooled with the $n = 250$ figures.

The decisions the paper relies on are stable across annotators: five-way primary-label agreement is 236/250 (94.4%), and agreement on the binary valid-NEI-versus-contamination gate that defines the hard subsets is 94.8%. Conditional on both annotators assigning NEI ($n = 215$), exact insufficiency-subtype agreement is lower (54.9%, $\kappa = 0.363$), while hard-versus-non-hard agreement remains high (93.0%, $\kappa = 0.714$). Of the 97 subtype disagreements, 82 remain within the same hard/non-hard category; only 15 cross the evaluation gate. Fine-grained subtype boundaries are therefore comparatively subjective, which is consistent with treating hard NEI as heterogeneous (Section 6.1), while the coarser decisions remain reliable. The 14 primary-label disagreements were resolved through joint adjudication; the resulting consensus labels are the labels used throughout the paper.

D.8 Boundary

The paper may claim that the SciFact hard-NEI audit and the fixed-claim/same-document audit provide human-adjudicated hard-NEI subsets. It should not claim that human validation eliminates all ambiguity, that candidate-only examples are valid without adjudication, or that FEVER/HoVer controls are human-validated hard NEI.

E Experimental Setup and Model Details

NEI-CAP uses models as diagnostic probes of construction sensitivity, not as proposed architectures.

E.1 Task Format

All neural experiments are three-way claim–evidence classification:

$$y \in \{\textsc{Support}, \textsc{Refute}, \text{NEI}\}.$$

The model receives claim and evidence text. Construction family is not provided as model input; it defines train/evaluation variants and reporting groups.

E.2 Primary SciFact Configuration

Setting	Value
Primary model	`deberta-v3-base`
Input root	NEI-CAP SciFact suite
Train variants	placeholder, cited, random irrelevant, BM25 near-miss, position-biased
Evaluation variants	placeholder, cited, random irrelevant, BM25 near-miss, position-biased
Evaluation split	test
Max length	384
Epochs	3
Batch size	16
Learning rate	2.0×10^{-5}
Weight decay	0.01
Warmup ratio	0.10
Class weighting	enabled
Early stopping metric	development Macro-F1
Patience	1
Precision	bf16 if available, otherwise fp32
Seeds	13, 17, 23
Checkpoint policy	save best checkpoint
Prediction logging	enabled

Table 19: Primary SciFact construction-matrix configuration.

E.3 Construction Matrix

The primary matrix trains one verifier for each NEI construction and evaluates it on all five constructions. Each cell asks whether performance transfers across evidence conditions while SUPPORT and REFUTE portions remain comparable.

E.4 Secondary Backbones

RoBERTa-base and SciBERT are secondary robustness probes. They reuse the same construction suite and train/evaluation variants. Secondary models supplement, but do not replace, the primary DeBERTa matrix.

E.5 Metrics

For three-way verification, we report accuracy, Macro-F1, class-specific F1, and especially NEI-F1. For one-class human-hard evaluation, Macro-F1 is not meaningful. We report NEI recall, false SUPPORT rate, false REFUTE rate, and mean predicted class probabilities.

E.6 Robustness Classification

For secondary-backbone SciFact probes, construction sensitivity is considered replicated when placeholder matched NEI-F1 is high, placeholder-to-hard drop is large, and at least one hard construction has non-trivial matched NEI-F1. These thresholds are diagnostic for this study, not universal standards.

E.7 Five-Seed and Mixed-Construction Configuration

The five-seed robustness and mixed-construction experiments use seeds {13,17,23,29,37}. Mixed-construction regimes include an easy mixture over shortcut-prone constructions, a hard mixture over BM25/cited constructions, and a balanced mixture over all five SciFact construction families. These experiments test whether construction sensitivity persists under more realistic training distributions.

E.8 Result Provenance

Each reported result should be traceable to a dataset manifest, construction family, model configuration, seed set, prediction file, evaluation output, and paper table or figure. Aggregate scores without construction metadata are insufficient for NEI-CAP.

F Full SciFact Construction Matrices

This appendix reports the full DeBERTa SciFact construction-shift matrices.

F.1 NEI-F1 Matrix

Train NEI	BM25	Cited	Placeholder	Position	Random
BM25 near-miss	0.675	0.561	0.725	0.652	0.676
Cited non-rationale	0.508	0.578	0.768	0.768	0.599
Placeholder	0.000	0.000	1.000	0.947	0.000
Position-biased	0.000	0.000	1.000	1.000	0.063
Random irrelevant	0.445	0.294	0.995	0.992	0.995

Table 20: Full SciFact construction-shift matrix for NEI-F1. Rows are train constructions; columns are evaluation constructions.

F.2 Macro-F1 Matrix

Train NEI	BM25	Cited	Placeholder	Position	Random
BM25 near-miss	0.529	0.479	0.559	0.530	0.532
Cited non-rationale	0.357	0.386	0.474	0.474	0.399
Placeholder	0.277	0.277	0.684	0.657	0.276
Position-biased	0.262	0.262	0.665	0.665	0.285
Random irrelevant	0.474	0.415	0.716	0.715	0.716

Table 21: Full SciFact construction-shift matrix for Macro-F1. Aggregate performance can appear strong under easy matched evaluation while hiding hard-NEI failure.

F.3 Drop Summary

BM25 and cited training do not yield perfect matched scores, but they transfer more stably across hard constructions.

Train NEI	Matched	BM25	Cited	Hard drop
Placeholder	1.000	0.000	0.000	1.000
Position-biased	1.000	0.000	0.000	1.000
Random irrelevant	0.995	0.445	0.294	0.625
BM25 near-miss	0.675	0.675	0.561	0.057
Cited non-rationale	0.578	0.508	0.578	0.035

Table 22: Matched and hard-evaluation NEI-F1 comparisons. Hard drop compares matched performance against average BM25/cited performance.

G Human-Validated Hard-NEI Evaluation

This appendix reports full evaluation on the held-out SciFact human-adjudicated hard-NEI test split ($n = 54$). The larger audit pool and sampling trail are documented in Appendix B; since all evaluated examples are validated as NEI, Macro-F1 is not reported.

G.1 Recall and Error Rates

Train NEI	Recall	95% low	95% high	False SUP	False REF	n
BM25 near-miss	0.691	0.593	0.790	0.142	0.167	54
Cited non-rationale	0.679	0.562	0.790	0.074	0.247	54
Random irrelevant	0.216	0.123	0.321	0.574	0.210	54
Placeholder	0.000	0.000	0.000	0.870	0.130	54
Position-biased	0.000	0.000	0.000	0.907	0.093	54

Table 23: Primary DeBERTa evaluation on human-adjudicated hard NEI. False REF maps internal contradiction errors to the paper-facing REFUTE label.

Bootstrap intervals degenerate for the exact 0/54 outcomes: resampling a subset with zero successes always yields zero. For placeholder and position-biased training we therefore additionally report the Clopper–Pearson exact 95% interval (Clopper and Pearson, 1934), $[0.000, 0.066]$. This interval does not overlap the BM25 near-miss bootstrap interval of $[0.593, 0.790]$.

G.2 Predicted Probabilities

Train NEI	Mean P(NEI)	95% low	95% high	Mean P(SUP)	Mean P(REF)	n
BM25 near-miss	0.398	0.366	0.431	0.324	0.277	54
Cited non-rationale	0.416	0.371	0.458	0.318	0.266	54
Random irrelevant	0.212	0.129	0.304	0.417	0.371	54
Placeholder	0.013	0.012	0.013	0.540	0.448	54
Position-biased	0.014	0.013	0.014	0.551	0.435	54

Table 24: Mean predicted probabilities on human-adjudicated hard NEI. Easy training assigns near-zero NEI probability to validated hard insufficient evidence and reallocates probability mass to SUPPORT and REFUTE.

H Multi-Model Robustness

This appendix documents secondary-backbone robustness, mixed-construction training, and same-document diagnostics. RoBERTa and SciBERT are diagnostic probes that supplement the primary DeBERTa matrix.

H.1 Secondary Backbones on Human-Adjudicated Hard NEI

Model	Train NEI	Recall	False SUP	False REF	Mean P(NEI)
SciBERT	BM25 near-miss	0.838	0.070	0.092	0.710
SciBERT	Cited non-rationale	0.875	0.063	0.062	0.692
SciBERT	Placeholder	0.000	0.574	0.426	0.011
SciBERT	Position-biased	0.007	0.838	0.156	0.024
SciBERT	Random irrelevant	0.099	0.520	0.381	0.114
RoBERTa	BM25 near-miss	0.703	0.116	0.181	0.404
RoBERTa	Cited non-rationale	0.638	0.109	0.253	0.435
RoBERTa	Placeholder	0.000	0.867	0.133	0.004
RoBERTa	Position-biased	0.009	0.932	0.060	0.014
RoBERTa	Random irrelevant	0.186	0.586	0.227	0.191

Table 25: Secondary-backbone evaluation on human-adjudicated hard NEI. Placeholder-trained RoBERTa and SciBERT both collapse on validated hard insufficient evidence.

H.2 Mixed-Construction Training

Table 26 summarizes DeBERTa mixed-construction training. Easy mixtures remain weak on hard conditions, while hard and balanced mixtures improve hard-NEI recognition.

Regime	BM25	Cited	SciFact hard
Easy mixture	0.379	0.273	0.178
Hard mixture	0.662	0.652	0.770
Balanced mixture	0.802	0.803	0.915

Table 26: DeBERTa mixed-construction training summary. BM25 and Cited are construction-stratified NEI-F1; SciFact hard is one-class NEI recall on human-adjudicated hard NEI. Same-document results are reported separately because that family is constructionally distinct.

Regime	PH	Position	Random	BM25	Cited
Easy mixture	0.989	0.989	0.960	0.379	0.273
Hard mixture	0.754	0.729	0.694	0.662	0.652
Balanced mixture	0.903	0.903	0.901	0.802	0.803

Table 27: Full DeBERTa mixed-construction stratified NEI-F1 matrix. The balanced mixture improves all conditions but still yields a construction-specific profile: easy conditions remain higher than BM25/cited hard conditions. PH denotes placeholder.

H.3 Same-Document Robustness

Model	Train NEI	Recall	False SUP	False REF
SciBERT	BM25 near-miss	0.482	0.231	0.287
SciBERT	Cited non-rationale	0.731	0.143	0.126
SciBERT	Placeholder	0.000	0.509	0.491
SciBERT	Position-biased	1.000	0.000	0.000
SciBERT	Random irrelevant	0.216	0.415	0.368
RoBERTa	BM25 near-miss	0.816	0.023	0.161
RoBERTa	Cited non-rationale	0.871	0.000	0.129
RoBERTa	Placeholder	0.526	0.412	0.061
RoBERTa	Position-biased	1.000	0.000	0.000
RoBERTa	Random irrelevant	0.950	0.029	0.020

Table 28: Secondary-backbone same-document evaluation. Same-document hard NEI behaves differently from SciFact BM25/cited hard NEI, reinforcing hard-NEI heterogeneity.

H.4 Boundary

The secondary models support the claim that construction sensitivity persists beyond DeBERTa. They do not show that all model families behave identically, and they do not replace the primary SciFact matrix.

I Same-Claim and Same-Document Diagnostics

Fixed-claim/same-document diagnostics control the claim or source document while varying evidence. They are decomposition tests, not proofs of clean counterfactual evidence use.

I.1 Setup

The same-claim diagnostic pairs the same claim with reference evidence and with human-adjudicated insufficient evidence. The same-document diagnostic draws insufficient evidence from the same source document as the reference evidence. We report hard-NEI recall, false SUPPORT/REFUTE rates, reference-side accuracy, probability-drop success, strict swap success, and the mean drop in the reference-label probability. The reference label is SUPPORT or REFUTE, so this metric tracks non-NEI verification confidence rather than only NEI recall.

I.2 Primary DeBERTa Fixed-Claim Diagnostics

Train NEI	Hard recall	Swap success	Mean Δ ref
BM25 near-miss	0.661	0.535	0.017
Cited non-rationale	0.988	0.743	0.023
Placeholder	0.898	0.737	0.112
Position-biased	1.000	0.728	0.034
Random irrelevant	0.947	0.807	0.124

Table 29: Primary DeBERTa fixed-claim diagnostics. The same claim is paired with reference evidence and with insufficient evidence, so swap success and mean reference-label probability drop are defined.

I.3 Primary DeBERTa Same-Document Hard-NEI Diagnostics

Train NEI	Hard recall	False SUP	False REF
BM25 near-miss	0.661	0.310	0.029
Cited non-rationale	0.988	0.009	0.003
Placeholder	0.898	0.085	0.018
Position-biased	1.000	0.000	0.000
Random irrelevant	0.947	0.053	0.000

Table 30: Primary DeBERTa same-document hard-NEI diagnostics. The hard subset is shared with the fixed-claim diagnostic, so hard recall matches Table 29; same-document evaluation does not define swap success or reference-label probability drop, and instead reports error rates on the insufficient-evidence side.

I.4 Secondary-Backbone Same-Claim Diagnostics

Model	Train NEI	Ref. acc.	Hard recall	Prob. drop	Strict swap
SciBERT	Placeholder	0.599	0.000	0.494	0.000
SciBERT	Cited non-rationale	0.488	0.731	0.827	0.365
SciBERT	Random irrelevant	0.556	0.216	0.602	0.140
SciBERT	BM25 near-miss	0.529	0.482	0.734	0.289
SciBERT	Position-biased	0.006	1.000	0.883	0.006
RoBERTa	Placeholder	0.401	0.503	0.643	0.137
RoBERTa	Cited non-rationale	0.108	0.833	0.614	0.053
RoBERTa	Random irrelevant	0.091	0.956	0.760	0.079
RoBERTa	BM25 near-miss	0.135	0.763	0.605	0.050
RoBERTa	Position-biased	0.009	1.000	0.538	0.009

Table 31: Secondary-backbone same-claim diagnostics. Results motivate the bounded interpretation: hard-side recognition and reference-side verification can diverge.

I.5 Same-Document Artifact Audit

We audit whether the same-document hard-NEI family has a shallow construction signature. A shallow feature classifier nearly separates same-document hard NEI from several other construction groups using evidence length, sentence count, overlap, and coverage features. Table 32 reports the compact audit result. The matched length/coverage audit is underpowered, so matched-subset comparisons are not conclusive.

Comparison	Acc.	Macro-F1	Status
Same-doc vs BM25/cited hard	0.997	0.997	completed
Same-doc vs placeholder	1.000	1.000	completed
Same-doc vs random irrelevant	1.000	1.000	completed
Same-doc vs constructed BM25/cited	1.000	1.000	completed
Matched feature audit	–	–	underpowered

Table 32: Same-document artifact audit. Same-document hard NEI is human-adjudicated but constructionally distinct; it should be reported as a source-controlled diagnostic family, not as artifact-free universal hard NEI.

I.6 Interpretation

The positive mean reference-label deltas in Table 29 show that substituting insufficient evidence lowers confidence in the original SUPPORT/REFUTE label, and the secondary-backbone rows show that a model can succeed on the hard-NEI side while remaining weak on the reference-evidence side.

J External Controls: FEVER and HoVer

FEVER and HoVer are bounded external controls: they test whether construction-aware diagnostics are useful beyond SciFact.

J.1 FEVER Subset Control

Baseline	Train NEI	Placeholder	BM25	Random	PH→BM25 drop
Claim+evidence TF-IDF	Placeholder	1.000	0.000	0.000	1.000
Claim+evidence TF-IDF	BM25 near-miss	0.614	0.373	0.327	0.241
Claim+evidence TF-IDF	Random irrelevant	0.237	0.414	0.476	-0.177
Evidence-only TF-IDF	BM25 near-miss	0.826	0.313	0.274	0.513
Length/overlap logistic regression	BM25 near-miss	0.772	0.406	0.536	0.366

Table 33: FEVER external subset control. Placeholder-trained shallow models solve placeholder NEI while failing on BM25 near-miss and random-irrelevant evaluation.

J.2 HoVer Candidate Missing-Hop Control

Train	Eval	NEI-F1	NEI recall	False SUP	n_{eval}
Missing-hop	Missing-hop	0.835	0.756	0.097	7,978
Missing-hop	Placeholder	0.973	1.000	0.000	7,978
Placeholder	Missing-hop	0.000	0.000	0.431	7,978
Placeholder	Placeholder	1.000	1.000	0.000	7,978

Table 34: HoVer candidate missing-hop control. Values are rounded to three decimals. HoVer rows are candidate-only and are not human-validated hard NEI.

J.3 Boundary

FEVER supports a non-toy subset-level shortcut-control claim. HoVer supports a candidate-only multi-hop construction-sensitivity claim. Neither supports full cross-dataset human-validated hard-NEI generalization.

K Diagnostic Case Studies

This appendix provides qualitative examples from row-level artifacts. We include only cases with available claim/evidence text and, where model behavior is discussed, saved predictions and probabilities. We do not reconstruct unavailable cases from aggregate statistics.

K.1 Case Selection Policy

Cases are selected only when: the claim and evidence are available, prediction and probability fields are available when model behavior is discussed, and human validation status is available when the case is described as human-adjudicated hard NEI. FEVER and HoVer cases are construction examples or candidate controls unless human-adjudicated.

K.2 Fixed-Claim Hard-Side NEI Recognition

Field	Value
Dataset	SciFact fixed-claim/same-document
Claim	Cholesterol loading induces KLF4 expression in vascular smooth muscle cells, resulting in pro-inflammatory cytokine expression.
Construction	Same-claim / same-document human-validated hard NEI
Train/seed	Random irrelevant / 17
Ref. pred.	SUPPORT
Hard pred.	NEI
Prob.	$P(\text{reference}) = 0.594$; $P(\text{NEI} \mid E_{\text{hard}}) = 0.978$
Why it matters	The model changes prediction when evidence changes while the claim is fixed, supporting hard-side insufficiency recognition.

Table 35: Case study: fixed-claim hard-side NEI recognition.

K.3 Reference-Side Weakness

Field	Value
Dataset	SciFact fixed-claim/same-document
Claim	T cell receptor/CD3 microdomains are required to induce the immunologic synapse.
Construction	Same-claim reference-side weakness
Train/seed	Position-biased / 13
Ref. pred.	NEI
Hard pred.	NEI
Prob.	$P(\text{reference}) = 0.001$; $P(\text{NEI} \mid E_{\text{hard}}) = 0.998$
Why it matters	The hard side is recognized as NEI, but the reference-evidence side is missed, motivating the same-claim caveat.

Table 36: Case study: fixed-claim reference-side weakness.

K.4 Same-Document Human-Validated Hard NEI

Field	Value
Dataset	SciFact fixed-claim/same-document
Claim	Removal of H3K9me3 improves reprogramming efficiency in human somatic cell nuclear transfer experiments.
Evidence excerpt	Aberrant epigenetic reprogramming can cause developmental defects in somatic cell nuclear transfer embryos.
Status	Human-validated hard NEI
Prediction	NEI
Prob.	$P(\text{NEI}) = 0.998$; $P(\text{Support}) = 0.001$; $P(\text{Refute}) = 0.001$
Why it matters	Same-document evidence controls source/topic better than random irrelevant evidence while remaining insufficient.

Table 37: Case study: same-document human-validated hard NEI.

K.5 HoVer Candidate Missing-Hop Control

Field	Value
Dataset	HoVer
Claim	Brett Herron's team competes in the Pro14 and the European Rugby Champions Cup.
Construction	Missing-one-supporting-fact candidate control
Status	Candidate-only, not human-validated
Train/seed	HoVer placeholder / 13
Prediction	SUPPORT
Prob.	$P(\text{Support}) = 0.981$; $P(\text{NEI}) = 0.000$; $P(\text{Refute}) = 0.018$
Why it matters	Placeholder-trained behavior can fail on topically related missing-hop evidence.

Table 38: Case study: HoVer candidate missing-hop control.

K.6 Unavailable Row-Level Cases

Case type	Unavailable reason
SciFact hard-audit placeholder failure on human-validated BM25/cited hard NEI	Row-level model predictions for final hard-NEI examples are unavailable; aggregate bootstrap tables exist only.
SciFact BM25/cited recovery case	Row-level paired recovery predictions for placeholder versus BM25/cited hard NEI are unavailable in primary outputs.

Table 39: Unavailable case-study reasons. We do not invent row-level cases when prediction artifacts are missing.

L Statistical Testing and Uncertainty Estimation

This appendix defines metrics and uncertainty procedures for construction matrices, one-class hard-NEI evaluation, fixed-claim diagnostics, and prediction coverage checks.

L.1 Three-Way Verification Metrics

For standard verification, models predict:

$$y \in \{\text{SUPPORT}, \text{REFUTE}, \text{NEI}\}.$$

We report accuracy, Macro-F1, class-specific F1, and especially NEI-F1. Construction-shift matrices emphasize NEI-F1 because the experimental manipulation changes the NEI evidence condition.

L.2 One-Class Human-Hard Metrics

For human-validated hard-NEI subsets, all evaluated examples are adjudicated as NEI. Macro-F1 is therefore not informative. We report:

$$\text{NEI recall} = \frac{\#\text{predicted NEI}}{\#\text{validated NEI}},$$

$$\text{False Support} = \frac{N_{\text{pred=SUP}}}{N_{\text{validated NEI}}},$$

$$\text{False Refute} = \frac{N_{\text{pred=REF}}}{N_{\text{validated NEI}}}.$$

We also report mean predicted probabilities for NEI, SUPPORT, and REFUTE.

L.3 Seed Aggregation and Bootstrap Intervals

The primary construction matrix uses seeds 13, 17, and 23. The five-seed robustness and mixed-construction experiments use seeds 13, 17, 23, 29, and 37. Where seed-level outputs are available, we report mean, standard deviation, minimum, maximum, and bounded intervals. Human validation and human-hard evaluations report bootstrap 95% intervals where available, computed by resampling evaluation examples or groups and recomputing the metric.

L.4 Fixed-Claim Metrics

For same-claim diagnostics, reference-label probability drop is:

$$\Delta_i = P_\theta(y_{\text{ref}} \mid c_i, E_i^{\text{ref}}) - P_\theta(y_{\text{ref}} \mid c_i, E_i^{\text{hard}}).$$

Probability-drop success is the fraction of pairs with $\Delta_i > 0$. Strict swap success additionally requires the reference-evidence side to be predicted as the reference label and the insufficient-evidence side to be predicted as NEI.

L.5 Prediction Coverage

Prediction coverage is:

$$\text{coverage} = \frac{n_{\text{predicted}}}{n_{\text{expected}}}.$$

Rows with missing or duplicate predictions are not treated as complete.

Scope group	Expected	Predicted	Coverage
Fixed-claim reference side	114	114	1.000
Fixed-claim insufficient side	114	114	1.000
Same-document hard side	114	114	1.000
SciFact matrix regression	885	885	1.000

Table 40: Prediction coverage summary. Values summarize the repeated complete-coverage pattern across train variants and seeds.

L.6 Reporting Rules

NEI-CAP follows these rules: report construction family for every NEI result; report Macro-F1 only when all three labels are meaningful; do not report Macro-F1 on one-class hard-NEI subsets; document prediction coverage before interpreting fixed-claim diagnostics; and mark FEVER/HoVer controls as bounded unless human validation is available.

M Reproducibility, Release, and Claim Boundary Checklist

This appendix records release readiness, the output-freeze policy, and paper-facing claim boundaries.

M.1 Release Audit

Check	Passed	Critical	Detail
Key reports exist	yes	yes	Experiment reports and summaries are present.
Paper tables exist and are non-empty	yes	yes	Paper-facing tables are generated.
Run registry entries exist	yes	yes	Core experiment entries are tracked.
Fixed-claim labels are preserved	yes	yes	Final fixed-claim/same-document labels path exists.
No LLM labels marked human	yes	yes	Reports use boundary language.
No candidate-only HoVer rows marked human-validated	yes	yes	HoVer marked human_validated=no.
No Macro-F1 on one-class human subsets	yes	yes	One-class hard-NEI evaluations exclude Macro-F1.
Primary SciFact matrix primary outputs unchanged	yes	yes	Primary matrix results preserved unchanged.
External dependencies documented	yes	no	Upstream data dependencies documented.
Known limitations documented	yes	yes	Limitations recorded.
Unavailable cases documented	yes	no	Missing row-level case reasons tracked.

Table 41: NEI-CAP release audit checklist. Critical checks concern result traceability, label provenance, metric validity, and claim boundaries.

M.2 Primary Outputs and Traceability

The primary evidence set includes the SciFact construction matrix, SciFact human-audit and hard-NEI evaluation, fixed-claim/same-document adjudicated labels, secondary-backbone robustness, FEVER subset controls, HoVer candidate missing-hop controls, and derived paper tables/figures. Each reported result should be traceable to a source dataset, construction family, split policy, model configuration, seed set, prediction artifact, and evaluation output.

M.3 Release Package Contents

The accompanying release will include construction manifests, group-disjoint split files, human-adjudicated labels, paper-facing prediction logs, evaluation scripts, artifact-audit scripts, and table-generation scripts. Upstream datasets are referenced and documented but not redistributed. Rows or case studies unavailable in primary outputs are documented rather than reconstructed from aggregate statistics.

M.4 Claim Strength Matrix

Claim	Strength	Allowed wording	Forbidden wording
NEI-CAP is construction-aware	Strong	Protocol for constructing, auditing, and stress-testing NEI evaluation.	Do not claim it solves fact verification.
SciFact NEI is construction-sensitive	Strong	Scores vary substantially across construction conditions.	Do not claim all errors have the same cause.
Easy NEI inflates performance	Moderate	Shortcut-like constructions can inflate apparent competence.	Do not claim every model uses only shortcuts.
Human validation confirms hard subsets	Strong	Audited BM25/cited and Fixed-claim/same-doc hard-NEI subsets are human-adjudicated.	Do not call LLM triage human validation.
Same-claim is diagnostic	Limited	Supports hard-side recognition with reference-side caveats.	Do not claim clean counterfactual evidence-use success.
HoVer is external control	Limited	Candidate-only multi-hop construction-sensitivity probe.	Do not call HoVer human-validated.
FEVER is external control	Limited	Non-toy subset-level shortcut control.	Do not call FEVER full-data validation.
Hard NEI is heterogeneous	Moderate	Different construction families stress different behaviors.	Do not collapse topic-unrelated cases into hard NEI.
Release is reproducible with limitations	Strong	Includes reproducibility instructions and claim boundaries.	Do not report toy, failed, or smoke outputs as paper-facing.

Table 42: Claim strength and wording boundaries.

M.5 Final Release Audit

The primary matrix results were frozen before the follow-up analyses and are unchanged, and no analysis created new human labels. The final audit covers split statistics, mixed training, multi-seed checks, the same-document artifact audit, and human-protocol documentation; it confirms no Macro-F1 on one-class hard subsets, no LLM-only human validation, and no candidate-only HoVer human validation. NEI-CAP is an audit and reporting protocol, not a blanket declaration that a dataset is invalid.

N Decoder Zero-Shot and Controlled SFT Diagnostics

This appendix documents the decoder diagnostic summarized in Section 5.6: a deterministic zero-shot evaluation of Qwen2.5-7B-Instruct and a controlled five-seed LoRA fine-tuning comparison in which only the training NEI realization changes. The question is whether fine-tuning changes a decoder verifier's sensitivity to the NEI construction used during training.

N.1 Zero-Shot Evaluation and Class Bias

The zero-shot evaluation uses deterministic conditional label scoring over the frozen SciFact construction tests (five constructions $\times$ 177 examples), the held-out human-adjudicated hard-NEI split ($n = 54$), and the fixed-claim paired diagnostic ($n = 114$ pairs). On the frozen core set, zero-shot Qwen reaches accuracy 0.775 and Macro-F1 0.725, with SUPPORT-F1 0.818, REFUTE-F1 0.532, and NEI-F1 0.824. Its easy-family mean NEI-F1 (over placeholder, random-irrelevant, and position-biased evaluation) is 0.845 and its BM25/cited hard mean is 0.793, so the zero-shot construction gap is modest. Human-hard recall is high (0.963), but the predicted SUPPORT/REFUTE/NEI rates are 45.4/9.3/45.3% against a gold distribution of 42.9/22.6/34.5%, so part of that recall comes from NEI overprediction and REFUTE underprediction.

Claim-only and shuffled-evidence controls caused the model to predict almost exclusively NEI, and we did not observe direct claim-only recovery of the original SUPPORT/REFUTE labels. These controls are diagnostic proxies: they neither prove contamination nor decontamination, and benchmark pretraining exposure remains unknown.

N.2 Controlled SFT Design and Data Gates

Across the three SFT regimes, only the training NEI evidence realization changes. The base model and revision, the SUPPORT/REFUTE multiset, claim groups, aligned NEI units, training size and label counts, the shared balanced development set, the checkpoint-selection metric, and the conditional-label evaluator are identical under explicit equality gates. After removing two aligned NEI groups that overlapped evaluation or shared-development assets, every regime trains on $n = 831$ examples (774 claim groups; 343 SUPPORT, 194 REFUTE, 294 NEI).

Regime	Training NEI realization	NEI n
Placeholder-only	Placeholder	294
Hard mixture	BM25 near-miss (148) + cited non-rationale (146)	294
Balanced-all	BM25 (51) + cited (59) + placeholder (60) + position-biased (58) + random irrelevant (66)	294

Table 43: Decoder SFT training regimes. Each aligned NEI unit contributes exactly one construction realization, so SUPPORT/REFUTE rows and NEI claim units are not multiplied across regimes.

All regimes share one byte-identical balanced development set ($n = 179$; 166 groups; 89 SUPPORT, 31 REFUTE, 59 NEI spanning all five families). One global learning rate was selected from mean shared-development Macro-F1 across all three regimes; test, human-hard, and paired-diagnostic data were never used for learning-rate, checkpoint, or hyperparameter selection. Leakage gates all pass with zero intersections: train/development/core-test claim groups, normalized claim texts against all evaluation assets, normalized prompt contents across candidate NEI realizations, and source example IDs. Five seeds are reported per regime; the expansion beyond the three-seed primary evaluation followed a predeclared rule, triggered by placeholder-only seed variability above threshold, and test results never fed back into hyperparameters.

N.3 Per-Regime Results

Qwen regime	Easy NEI-F1	Hard NEI-F1	Easy–hard gap	Hum. recall	REF-F1
Zero-shot base	0.845	0.793	0.052	0.963	0.532
Placeholder-only SFT	0.559 ± 0.192	0.054 ± 0.054	0.505 ± 0.140	0.011	0.645
Hard-mixture SFT	0.916 ± 0.014	0.834 ± 0.008	0.082 ± 0.010	0.900	0.784
Balanced-all SFT	0.953 ± 0.014	0.817 ± 0.018	0.135 ± 0.019	0.856	0.795

Table 44: Full decoder regime results. Easy NEI-F1 is the mean over placeholder, random-irrelevant, and position-biased evaluation and is not the matched-placeholder score: under placeholder-only SFT, matched-placeholder NEI-F1 remains 1.000 in every seed even though the broader easy-family mean is 0.559, because the placeholder-specific rule also degrades the other easy families. SFT rows are descriptive five-seed means with sample standard deviations.

All five placeholder-SFT training-induced gap deltas are positive, and their paired group-bootstrap 95% intervals (Koehn, 2004) exclude zero. The fixed-claim strict-pair success rate shows the same qualitative training dependence: 0.535 for the zero-shot base, 0.126 after placeholder-only SFT, and 0.653/0.654 after hard-mixture/balanced SFT. The balanced regime is not uniformly bet-

ter than the zero-shot base: its human-hard recall is lower (0.856 vs. 0.963). It does, however, move the predicted label distribution from 45.4/9.3/45.3% to 43.0/22.7/34.3%, close to the gold 42.9/22.6/34.5%, and raises REFUTE-F1 from 0.532 to 0.795.

N.4 Llama Exclusion Gate

Llama-3.1-8B-Instruct failed the predeclared fully permuted label-order stability gates: its maximum construction NEI-F1 delta across label orders was 0.222 under corrected generation (Qwen: 0.032) and 0.164 under conditional label scoring, with maximum predicted-label-rate deltas of 0.254 and 0.243 (Qwen: 0.040). Llama was therefore not fine-tuned; its zero-shot behavior is reported only as an exploratory prompt-sensitivity observation.

N.5 Boundary

The bounded claim is that training-induced construction sensitivity occurs in Qwen2.5-7B under the frozen SciFact setup, and that construction-diverse SFT is a practical mitigation. The evidence does not support a universal decoder collapse, a claim that other LLMs behave like Qwen, decontamination, production readiness, or LLM outputs as human validation.