SMALLJUDGE-BENCH: EVALUATING CORRECT INVARIANCE IN EVIDENCE-GROUNDED JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

An evidence-grounded judge should retain a correct verdict when a claim is rephrased without changing its factual content or relation to the supplied evidence. Prediction stability alone misses this requirement because a judge can be consistently wrong. We introduce SmallJudge-Bench, with 2,800 individually reviewed original–edited pairs across four datasets. Claude Sonnet 4.6 exceeds a clean-trained ModernBERT-large encoder in paired correctness by 4.80 percentage points (95% confidence interval $[2.66, 6.99]$). Yet, on originals both models answer correctly, Sonnet's edited error rate is 2.64 points higher ($[1.63, 3.67]$). Matched subsets from two editing generators show the same conditional ordering, although the interaction between generator and model remains unresolved. Label-preserving training augmentation improves paired correctness and reduces prediction changes relative to repeated-clean training. These findings distinguish overall correctness from retaining correct decisions under edits, within the evaluated configurations. The protocol reports both, with separate controls for evidence-free stability and responses to changed content.

1 INTRODUCTION

Large language models (LLMs) evaluate generated answers, summaries, and other model outputs (Zheng et al., 2023; Tan et al., 2025). Evidence-grounded evaluation asks whether supplied evidence supports a claim, contradicts it, or leaves its truth unresolved. The verdict should not depend on irrelevant wording or source metadata. Such sensitivity can affect data filtering and system comparisons even when the underlying factual content is unchanged.

Existing work shows that judge behavior can change with presentation and evaluation format (Wang et al., 2024; Doddapaneni et al., 2024; Dev et al., 2026). However, three issues complicate the interpretation of a prediction change. An apparent paraphrase may alter certainty or factual scope. A changed prediction may correct an earlier error rather than introduce one. Finally, models with different output spaces may appear more or less robust because their labels distinguish different events. A useful evaluation must establish what the edit preserves, retain the paired predictions, and specify the label space before comparing systems.

SmallJudge-Bench addresses these issues through a paired benchmark and measurement protocol. All edits modify the claim or its metadata; the stored evidence text remains unchanged. We review each candidate for preservation of the gold label and evidence–claim relation. A judge exhibits *correct invariance* on an accepted pair when both judgments are correct. We measure this as paired correctness (RC), alongside clean and edited accuracy and the label-flip rate (Flip). Conditional comparisons then ask which model better retains correct decisions on originals both models initially classify correctly. Native labels and a common support/non-support label space serve different comparisons.

Our central question is whether a judge remains correct across acceptable presentations. Evidence-free predictors test stability without grounding; content-changing edits test correct responses when the original label no longer applies. Identity groups describe sensitivity to source metadata, while a matched second-generator batch tests how the comparison changes under another editing configuration.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

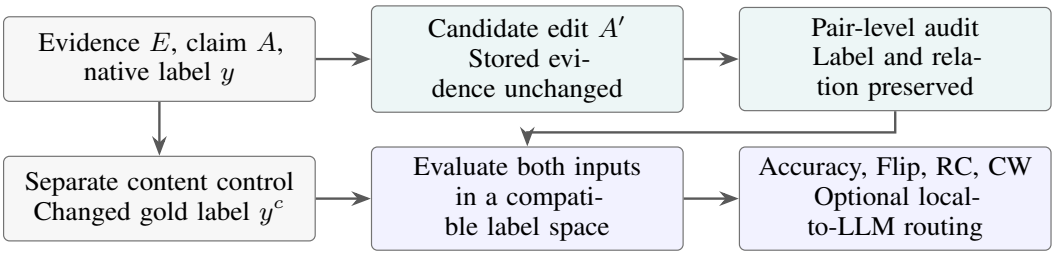

Figure 1: Evaluation protocol. Surface edits enter the main comparison only after pair-level preservation review. Content-changing edits form a separate control with a different gold label. The same evaluated judge can be used alone or as the local component of a router; routing is a deployment study, not the benchmark's definition.

Our contributions are the reviewed benchmark, a protocol separating paired correctness from conditional error, and a matched comparison across two editing generators. Together, these analyses show why overall accuracy and resistance to edits need not rank models alike. Training augmentation and confidence-based routing test practical uses of the evaluation. These are configuration-specific comparisons, not a new architecture or metric.

2 Related Work

Evidence-grounded verification. FEVER, SciFact, and VitaminC evaluate evidence-conditioned claim judgments (Thorne et al., 2018; Wadden et al., 2020; Schuster et al., 2021). TRUE and AggreFact bring together factual-consistency evaluation settings (Honovich et al., 2022; Tang et al., 2023). Methods including SummaC, AlignScore, MiniCheck, and Lynx provide relevant verifier baselines (Laban et al., 2022; Zha et al., 2023; Tang et al., 2024; Ravi et al., 2024). Our contribution is an evaluation layer over these tasks: accepted original–edited pairs expose changes that aggregate task scores alone do not describe. We do not treat a binary support score as an unambiguous three-way prediction.

Behavioral tests and judge robustness. CheckList organizes behavioral tests across linguistic capabilities and test types (Ribeiro et al., 2020). Contrast sets use meaningful changes that typically alter gold labels to probe local decision boundaries (Gardner et al., 2020). JudgeBench tests challenging response preferences grounded in objective correctness (Tan et al., 2025). Finding Blind Spots in Evaluator LLMs tests detection of deliberately degraded answers; the Judge Reliability Harness includes both meaning-preserving and label-changing tests (Doddapaneni et al., 2024; Dev et al., 2026). Our evaluation combines pair-level preservation review, compatible label spaces, and correctness at both ends of an edit with unchanged stored evidence. The controls distinguish stable incorrect answers, corrected initial errors, and inappropriate invariance after content changes. The matched generator comparison examines another source of variation. These are extensions of behavioral evaluation, not replacements for prior frameworks.

Debiasing and selective evaluation. OffsetBias studies training data for evaluator debiasing (Park et al., 2024). We compare augmentation against repeated-clean training with matched exposure and report stability separately from accuracy. Development-set calibration supports confidence-based decisions (Guo et al., 2017), but does not itself establish deployment quality. Our routing comparison holds the fallback fixed and states the quality tolerance and cost model explicitly.

3 Benchmark and Measurement Protocol

3.1 Tasks, Labels, and Individual Edits

An original example is (E_i, A_i, y_i), where E_i is the supplied evidence, A_i is a claim or answer, and y_i is its gold label. A judge f predicts a label from (E_i, A_i). Support means that the evidence supports the claim; refutation means that it contradicts the claim. Not enough information (NEI) means that

the evidence establishes neither. We use the evaluated support/refute subset of SciFact, the three-way S/R/NEI tasks for FEVER and VitaminC, and the consistent/inconsistent task for AggreFact/TRUE. Table 1 gives the actual evaluation sizes rather than the sizes of the source datasets.

Dataset	Labels	U	P	Candidates	Excluded
SciFact	S/R	300	600	640	40
FEVER	S/R/NEI	400	800	844	44
VitaminC	S/R/NEI	400	800	873	73
AggreFact/TRUE	C/I	300	600	627	27
Total	–	1400	2800	2984	184

Table 1: Evaluation cohort. U counts unique originals and P counts accepted original–edited pairs. S/R/NEI denote support, refutation, and not enough information; C/I denote consistent/inconsistent. Excluded candidates are not part of the paired performance denominator.

Qwen3-32B (G1) generates verbosity and style edits with fixed evidence, bfloat16 (BF16) precision, thinking disabled, and temperature 0.7. Verbosity expands wording through restatement; style improves fluency without changing certainty or scope. FEVER-only identity edits prepend fixed metadata, naming GPT-4.1 or Qwen3-32B equally often, without changing the claim. Generation and filtering do not consult evaluated judges' outputs. Appendix E provides prompts and settings.

Each original contributes two accepted edits. SciFact, VitaminC, and AggreFact/TRUE contribute verbosity and style. FEVER contributes an identity edit plus either verbosity or style, with verbosity and style each covering 200 originals. Acceptance requires the same gold label and evidence–claim relation, including unchanged factual scope, attribution, negation, quantities, and certainty. Changes to evidence selection are excluded; the benchmark tests presentation, not retrieval robustness.

3.2 Acceptance and Independent Re-Audit

Programmatic checks first screen required fields and factual markers. Two graduate students experienced in natural language processing (NLP) and fact-checking independently review each pair. A senior researcher adjudicates disagreements; uncertain cases are excluded. Screeners do not see model outputs. Of 2,800 initial candidates, 184 were rejected and replaced by 184 accepted supplementary candidates, yielding 2,984 candidates and 2,800 accepted pairs. These construction counts are not model errors.

The re-audit samples 60 accepted pair IDs per dataset without replacement. Two other graduate students independently assess preservation; a senior researcher uninvolved in screening adjudicates disagreements. The interface shows evidence and both texts but hides gold labels, screening decisions, and model outputs (Appendix E). Agreement measures matching judgments; Cohen's κ adjusts for chance agreement.

All 240 sampled pairs preserved the label; 234 also preserved the relation, giving 97.50% joint preservation. Agreement was 98.75%, with $\kappa = 0.763$ (dataset results in Appendix A). The six label-preserving but relation-changing cases show why both criteria matter. Model results retain the defined 2,800-pair denominator after this diagnostic.

A separate high-risk slice contained 60 pairs: 6 preserved, 43 invalid, and 11 disputed. This targeted diagnostic slice is not a random sample of accepted pairs and is excluded from the 97.50% estimate.

3.3 Correctness, Invariance, and Their Denominators

Let $\mathcal{P}_d$ contain the P_d accepted pairs for dataset d, and let $\widehat{y}_i = f(E_i, A_i)$ and $\widehat{y}'_i = f(E_i, A'_i)$. The indicator $\mathbf{1}\{B\}$ equals one if event B holds and zero otherwise. For each dataset, we compute

$$\mathrm{Acc}_d^{\mathrm{clean}} = \frac{1}{P_d} \sum_{i \in \mathcal{P}_d} \mathbf{1}\{\widehat{y}_i = y_i\}, \qquad \mathrm{Acc}_d^{\mathrm{edit}} = \frac{1}{P_d} \sum_{i \in \mathcal{P}_d} \mathbf{1}\{\widehat{y}'_i = y_i\}, \tag{1}$$

$$\mathrm{Flip}_d = \frac{1}{P_d} \sum_{i \in \mathcal{P}_d} \mathbf{1}\{\widehat{y}_i \neq \widehat{y}'_i\}, \qquad \mathrm{RC}_d = \frac{1}{P_d} \sum_{i \in \mathcal{P}_d} \mathbf{1}\{\widehat{y}_i = y_i \ \wedge \ \widehat{y}'_i = y_i\}. \tag{2}$$

Both accuracies use the same paired cohort; each original contributes once per associated edit. Headline rates first aggregate pairs within each dataset, then average equally across datasets: $\overline{m} = \frac{1}{4}\sum_{d=1}^{4} m_d$. Macro-F1 averages classwise F1 and is used for checkpoint selection and the FEVER probes, not as a substitute for accuracy. Single-dataset and control-subset denominators are identified explicitly.

Flip answers whether the prediction changes; $1 - \mathrm{Flip}$ is prediction consistency, not label accuracy. RC asks whether a correct decision is retained at both ends of the pair. A predictor that always emits the same wrong label has zero Flip and zero RC. Conversely, a wrong-to-correct transition increases Flip while improving edited accuracy. Reporting RC alongside Flip and the two marginal accuracies exposes these cases instead of rewarding invariance in isolation.

For interpretation, CW denotes a correct-to-wrong transition, WC a wrong-to-correct transition, and WD a change between two wrong labels. On the same cohort and label space, $\mathrm{Flip} = \mathrm{CW} + \mathrm{WC} + \mathrm{WD}$, whereas $\mathrm{Acc}^{\mathrm{clean}} - \mathrm{Acc}^{\mathrm{edit}} = \mathrm{CW} - \mathrm{WC}$. Thus, accuracy changes can conceal offsetting transitions. WD is zero when both predictions are valid binary labels, but need not vanish for three-way labels or invalid outputs. Exact transition counts cannot be reconstructed from rounded aggregate rates.

3.4 Compatible Label Spaces

Native evaluation retains the dataset labels. The binary evaluation folds refutation and NEI into non-support while retaining the consistency distinction for summarization. Binary or scalar verifiers can therefore participate without an invented three-class interface. Collapsing labels removes refutation–NEI transitions from Flip without changing model capability. We interpret differences within, not across, label spaces. The main model and generator comparisons use native labels; routing uses binary labels.

4 Experimental Design

Compared systems. The native comparison includes a clean-trained ModernBERT-large encoder, GPT-4.1 (`gpt-4.1-2025-04-14`), Claude Sonnet 4.6 (`claude-sonnet-4-6`), and Qwen3-32B. All four systems use identical originals and edits. Generative judges use temperature-zero or greedy decoding; Sonnet and Qwen have thinking disabled. Separate content controls also include MiniCheck-Flan-T5-Large and Lynx-70B-Instruct. Appendix F gives identifiers, settings, prompts, label mappings, and parser rules. Names denote evaluated configurations, not model families.

Encoder training and split isolation. Encoder uses ModernBERT-large with task-specific two- or three-class heads, trained on clean examples. Source-document/claim groups are deduplicated across domains and kept within one split. Test, development, and control pools hold out 1,400, 640, and 400 originals; remaining examples form the training pool. Checkpoint selection maximizes clean macro-F1 averaged across the four development datasets; ties favor the earlier epoch. Appendix G gives the training settings.

Protocol controls. Evidence-free controls comprise a development-majority predictor, a surface-feature probe, and a claim-only probe on the native FEVER cohort. These separately evaluated diagnostic baselines test whether unchanged predictions alone establish evidence use. They are not controlled ablations of the main Encoder.

The content-changing control uses 200 originals, each paired with a style edit and a content edit that changes the gold label. There are 100 support-to-non-support changes and 100 changes in the reverse direction. Let $y_i^c \neq y_i$ denote the content-edit gold label and $\widehat{y}_i^c$ its prediction. We report

$$J = \frac{\sum_{i=1}^{200} \mathbf{1}\{\widehat{y}_i = y_i \wedge \widehat{y}_i^c = y_i^c\}}{200}, \qquad J_{\mathrm{cond}} = \frac{\sum_{i=1}^{200} \mathbf{1}\{\widehat{y}_i = y_i \wedge \widehat{y}_i^c = y_i^c\}}{\sum_{i=1}^{200} \mathbf{1}\{\widehat{y}_i = y_i\}}. \tag{3}$$

The conditional quantity is undefined when no clean prediction is correct. Unlike a raw content-flip rate, both measures require the changed answer to be correct. Style RC and CW are measured on the same 200 originals, separately from the 2,800-pair headline cohort.

System	Clean acc.	Edited acc.	RC [95% CI]	Flip [95% CI]
Encoder	86.60	85.97	84.50 [82.68, 86.33]	4.10 [3.36, 4.89]
GPT-4.1	85.67	80.86	76.11 [74.12, 78.10]	15.40 [13.99, 16.79]
Sonnet 4.6	94.10	90.76	89.30 [87.86, 90.71]	6.42 [5.51, 7.36]
Qwen3-32B	86.31	84.19	81.32 [79.43, 83.17]	8.67 [7.61, 9.75]

Table 2: Native-label results on 1,400 originals and 2,800 accepted pairs. Rates are percentages, equally weighted across four datasets. Brackets give pointwise 95% original-cluster CIs. RC counts pairs correct at both ends; Flip counts changed predictions. Higher accuracy and RC are preferable; lower Flip alone does not establish correctness.

Matched second-generator comparison. We sampled 100 test originals per dataset, stratifying by native gold label. FEVER retained 50 originals from each existing verbosity/style group. GPT-4.1 (G2) generated a new edit for each original–operator combination, termed a slot, using the same preservation instructions as G1. This gives 700 text-edit slots: 200 in each of three datasets and 100 in FEVER. Identity templates are excluded. Each slot allowed at most three candidates, accepting the first that passed review without consulting judge outputs. Five slots failed; both generator versions of those slots were excluded from this comparison, leaving 695 matched pairs per generator across all 400 originals.

The four native judges used their specified configurations. Clean and first-generator predictions were reused; the new batch supplied the second-generator judgments. We compare rates on identical retained slots, including failed judge outputs in the denominator. An independent re-audit of 20 new accepted pairs per dataset found joint preservation in 77/80; all three failures remain in the analysis. Appendix K provides generation, screening, and resampling details. This is a batch comparison, not a claim that the two generators are exchangeable.

Secondary studies. Confidence-based routing retains Encoder's prediction unless confidence falls below 0.88, then calls Sonnet (Appendix C). This threshold was selected previously and was not retuned on the updated test results. Training augmentation compares clean examples, repeated clean examples, and added label-preserving rewrites under the same update budget and three training seeds (Appendix I). These studies do not define benchmark membership or select the main models.

Reporting scope. Comparisons use 95% confidence intervals (CIs) from 10,000 bootstrap replicates that resample original IDs within each dataset. Each sampled original retains its edits; compared models share the sampled indices. Replicates recompute dataset-specific metrics before equal-weight averaging, with linearly interpolated percentile limits. Intervals are pointwise except for the joint routing tolerance, which uses Bonferroni-adjusted one-sided bounds for two metrics. The fixed primary generator comparison is the change in the Sonnet–Encoder RC difference. Other generator contrasts are supplementary. Augmentation intervals condition on the three observed training seeds; their standard deviations describe between-seed variation. Appendix D separates these uncertainty measures.

5 Results

5.1 Correctness and Stability Require Separate Comparisons

Table 2 reports the four native-label systems on the same evaluation pairs. Sonnet 4.6 achieved RC of 89.30%, compared with 84.50% for Encoder, a paired difference of 4.80 percentage points (pp; 95% CI $[2.66, 6.99]$). Sonnet also led in clean accuracy by 7.50 pp. Encoder had a lower Flip point estimate, 4.10% versus 6.42%. Overall correctness and prediction stability therefore favored different models. Marginal Flip intervals alone are not a paired test of this difference.

The full-cohort comparison combines two abilities: answering originals correctly and retaining correct answers after edits. A model with higher clean accuracy has more opportunities to remain correct. We therefore examine retention on a shared starting set rather than interpreting the RC gap as a direct measure of perturbation resistance.

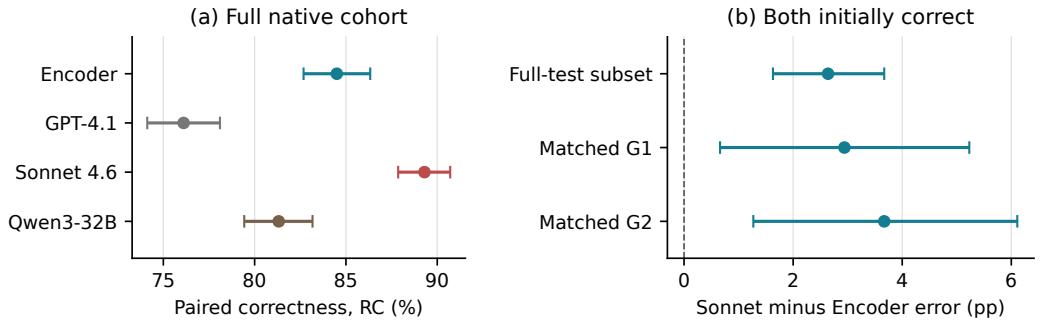

Figure 2: Different comparison questions. (a) Full-cohort native RC with marginal 95% CIs. (b) Sonnet-minus-Encoder edited error on originals both models initially answer correctly, with paired 95% CIs. Positive differences favor Encoder. The full test subset and generator-matched subsets have different denominators (Table 3).

Subset	U	P	Encoder error	Sonnet error	Δ [95% CI]
Full-test subset	1153	2306	2.55	5.19	$+2.64\,[1.63, 3.67]$
Matched G1	334	577	2.85	5.79	$+2.94\,[0.66, 5.23]$
Matched G2	334	577	2.59	6.26	$+3.67\,[1.27, 6.11]$

Table 3: Conditional edited errors in percent, with Sonnet-minus-Encoder differences in pp and pointwise 95% CIs. U and P count shared initially correct originals and their pairs. G1 and G2 retain the same 334 originals and 577 matched pairs. Each interval uses shared original-cluster resamples within its fixed subset.

Errors on originals both models initially answer correctly. The shared subset contains 1,153 originals and all their 2,306 accepted edits. Every edited error is a correct-to-wrong transition for either model (Table 3). Encoder's error rate was 2.55%, compared with 5.19% for Sonnet. The paired difference of $+2.64$ pp had CI $[1.63, 3.67]$, supporting lower conditional error for Encoder.

The generator-matched comparison retained 334 shared initially correct originals and 577 pairs per generator. Sonnet's conditional error was higher for both G1 and G2, with paired intervals excluding zero. These subsets overlap and do not constitute independent replications. The ordering is limited to the evaluated models, shared subsets, and edit configurations; it does not show that smaller models are generally more robust. Appendix H defines the conditioning and resampling.

5.2 Edit Types and Identity Groups Have Different Coverage

Verbosity and style each cover 1,200 pairs across four datasets; identity covers 400 FEVER pairs. Their different coverage prevents treating a difference between edit types as an isolated operator effect. Each FEVER original received one identity value, with 200 originals randomly assigned to each named-model group. The groups contain different originals and different realized label distributions (Appendix J). They do not provide a within-original test of identity preference. The matched generator experiment therefore excludes identity templates and compares the same original–edit-type combinations.

5.3 The Generator Interaction Remains Unresolved

Figure 3 compares the Sonnet–Encoder gaps on Qwen-generated edits (G1) and GPT-4.1-generated edits (G2). Both batches contain the same 695 slots from 400 originals. The RC gap was 1.40 pp on G1 and 0.78 pp on G2. The primary interaction, the second gap minus the first, was -0.62 pp with CI $[-2.63, 1.39]$. This interval resolves neither a change in the gap nor equivalence between generators.

Sonnet's Flip was higher in both batches, but the Flip interaction also included zero. The conditional error results answer a different question: they restrict attention to originals both models classify

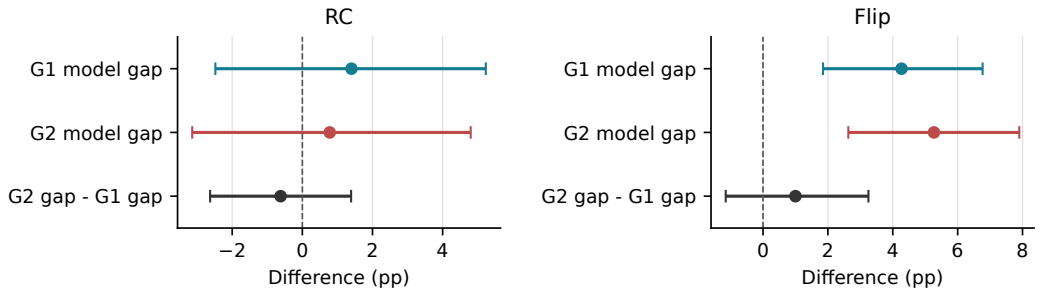

Figure 3: Sonnet-minus-Encoder gaps for RC and Flip on 695 matched pairs per generator. Interaction rows subtract the G1 gap from the G2 gap. Error bars are pointwise 95% shared original-cluster CIs. Positive RC gaps favor Sonnet; positive Flip gaps indicate more label changes by Sonnet. These full matched cohorts differ from the conditional subsets in Table 3.

System	Clean acc.	Edited acc.	Edited F1	Flip	RC
Majority	35.50	35.50	17.47	0.00	35.50
Surface-feature	45.00	43.62	43.56	11.25	40.12
Claim-only	59.50	53.25	53.15	21.75	46.25

Table 4: Evidence-free probes on the native FEVER cohort ($U = 400$, $P = 800$). Values are percentages; edited F1 is class-macro F1. Majority always outputs the development-majority class. Its zero Flip illustrates consistency without high correctness. The probes are diagnostic baselines, not controlled ablations of an identical architecture.

correctly. Their common direction does not replace the interaction test on the complete matched cohort. Appendix K reports all contrasts and the acceptance process. Generator–judge overlap and batch timing limit causal interpretations of differences between generators.

5.4 Stable Predictions Can Ignore the Evidence

The FEVER probes expose a limitation of optimizing Flip alone (Table 4). The development-majority predictor never changed its label, giving zero Flip, but its RC was only 35.50%. The surface-feature and claim-only probes reached RC values of 40.12% and 46.25%, respectively. These separately evaluated controls show that stable or partly correct predictions are possible without using the supplied evidence. They do not estimate how removing evidence would affect the main Encoder.

These input-restricted baselines do not by themselves isolate the effect of evidence access. Low Flip alone cannot demonstrate grounding.

5.5 Correct Responses to Content Changes

On the 200-original control, Encoder retained correct predictions across 84.00% of style pairs but answered both ends of 73.00% of content-changing pairs correctly (Figure 4). The latter requires a correct label change, not merely sensitivity. This separates responsiveness to changed content from invariance under harmless presentation.

Joint correctness uses all 200 originals; model-specific conditional correctness instead uses only each model's clean-correct cases. Those denominators differ across models and are not the common intersection in Table 3. Appendix B reports the exact counts rather than treating the conditional percentages as a same-cohort comparison.

5.6 Augmentation Improves Correctness and Stability

We compared clean training (A), repeated clean training (B), and training with label-preserving rewrites (C) under a common update budget. Relative to B, C increased edited accuracy by 1.55 pp and RC by 1.71 pp, while reducing Flip by 0.72 pp. The paired intervals excluded zero in each

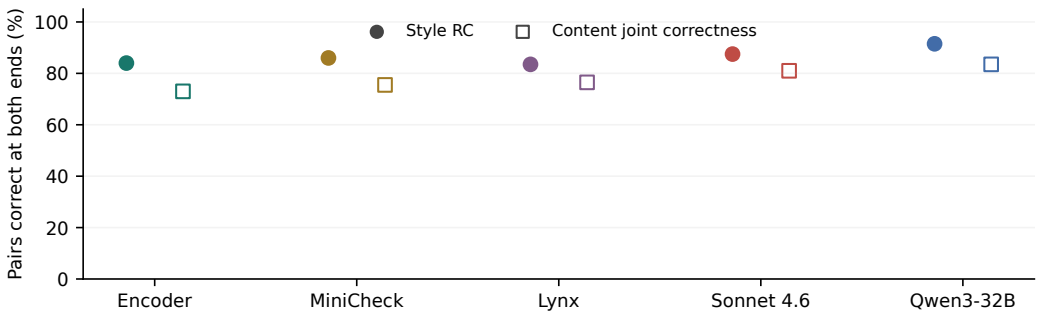

Figure 4: Complementary controls evaluated separately on the same 200 originals. Style RC requires two correct predictions with an unchanged gold label; content-changing pairs require correct predictions for two different gold labels. Both use all 200 originals as the denominator. They are different edit conditions, not estimates of the same quantity. Exact counts are in Appendix B.

Training	Clean acc.	Edited acc.	RC	Flip
A: Clean	86.62 ± 0.11	86.06 ± 0.08	84.60 ± 0.10	4.01 ± 0.04
B: Repeated clean	86.75 ± 0.08	86.19 ± 0.09	84.73 ± 0.09	4.01 ± 0.04
C: Augmented	87.95 ± 0.27	87.74 ± 0.04	86.44 ± 0.15	3.29 ± 0.07

Contrast	ΔEdited acc.	ΔRC	ΔFlip
C−A	$+1.68\,[1.31, 2.08]$	$+1.84\,[1.44, 2.27]$	$-0.72\,[-0.97, -0.49]$
C−B	$+1.55\,[1.20, 1.92]$	$+1.71\,[1.33, 2.11]$	$-0.72\,[-0.97, -0.49]$

Table 5: Native augmentation results. The upper panel gives mean percentages $\pm$ sample standard deviations over three seeds; the lower panel gives differences in pp with pointwise 95% CIs. A uses clean examples, B repeats them, and C adds label-preserving rewrites. Intervals resample originals for three fixed seeds, not the population of training seeds. Appendix I details the comparison.

direction (Table 5). Repeating originals did not reproduce the benefit of adding rewrites in this recipe. These results support evaluating training changes with both correctness and transitions, rather than edited accuracy alone.

5.7 Routing Meets the Stated Quality Tolerance

The binary Encoder-to-Sonnet router called Sonnet on 41.97% of inputs with threshold 0.88 fixed before this evaluation. Compared with always using Sonnet, edited accuracy increased by 1.59 pp (95% CI $[0.84, 2.38]$) and RC by 1.24 pp ($[0.38, 2.11]$). Both one-sided 97.5% lower bounds exceeded the -1-pp margin. Bonferroni control over the two metrics therefore supported the stated joint tolerance at total $\alpha = 0.05$.

Under the fixed cost model, cost fell by 56.43%, from 50.00 to 21.79 units per 10,000 inputs. Each input pays the Encoder cost plus the Sonnet cost when routed; repeated originals are billed without caching. Figure 5 shows the operating points. This is evidence for the tested confidence rule and cost assumptions, not a measured latency improvement or a guarantee under distribution shift.

6 Discussion and Limitations

What the comparisons establish. The model ordering depends on the question asked. Sonnet has higher overall paired correctness, while Encoder makes fewer edited errors on originals both models initially answer correctly. Neither result cancels the other: deployment quality depends on initial ability as well as retention under edits. The controls further distinguish these properties from merely changing or preserving a prediction.

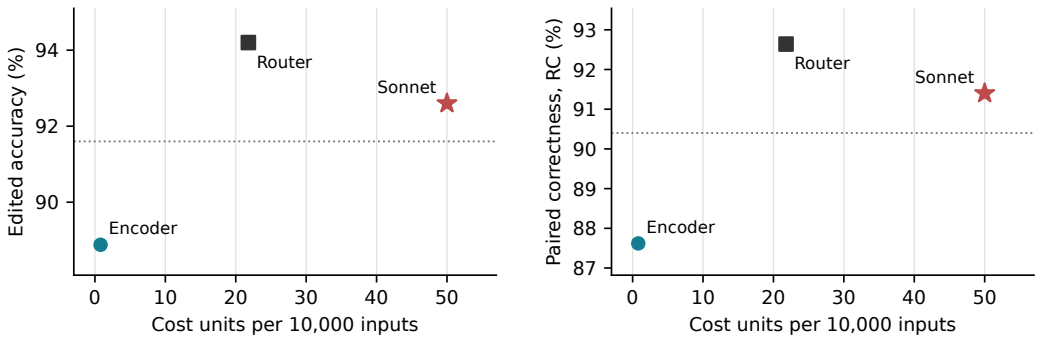

Figure 5: Binary routing operating points. Dotted lines mark the one-percentage-point quality margins relative to always using Sonnet. Statistical support comes from the paired bounds, not the plotted points. Appendix C reports all scores, intervals, and cost accounting.

Augmentation and routing show practical uses of these distinctions. Augmentation improved both correctness and stability under the tested training budget; routing met a specified accuracy and RC tolerance at lower modeled cost. Neither result establishes an advantage for every training recipe or confidence function. The unresolved generator interaction also prevents treating the observed model gaps as invariant across generation settings.

Audit and statistical boundaries. Six of 240 sampled accepted pairs in the original benchmark and three of 80 in the new batch failed joint preservation. Equal sampling per dataset supports dataset-macro audit summaries, not automatically pair-weighted estimates. These failures remain in the defined cohorts; preservation is a reviewed criterion rather than a guarantee for every pair. Clustered CIs quantify sampling uncertainty, not audit errors, model-version drift, or simultaneous coverage of all comparisons.

Coverage and reproducibility. Both editing generators are evaluated judges. Hiding judge outputs during acceptance review does not remove generator-specific preferences or selection caused by requiring acceptance in both batches. Reused first-generator judgments leave service timing as another explanation for batch differences. Stored evidence is unchanged, but claim-length-dependent truncation could change the evidence tokens seen by Encoder; identical retained tokens have not been established. This limits attributing observed errors solely to presentation. Identity tests cover FEVER only; the study excludes retrieval changes and does not rule out pretraining overlap. Routing has no downstream or latency validation. Full reproduction requires the trained checkpoint, evaluation harness, and complete training and routing settings.

7 Conclusion

SmallJudge-Bench evaluates correctness across individually reviewed presentation edits with unchanged stored evidence. Higher overall paired correctness did not imply fewer errors on the shared initially correct subset. Augmentation improved correctness and stability in the tested recipe, while routing met its two-metric quality tolerance at lower modeled cost. These findings support reporting paired correctness, conditional error, and label changes together, with explicit cohorts and generation conditions. They do not establish a model-family ranking or universal robustness to presentation.

Ethics Statement

The evaluation concerns evidence-grounded verification on public benchmark tasks. Human reviewers screened and re-audited candidate edits; their roles and access to information are described in the appendices. Model-origin metadata is a controlled perturbation, not evidence of the actual authorship or truth of a claim. Systems sensitive to presentation may introduce inconsistent decisions in data filtering or automated evaluation. The reported tests do not establish suitability for high-stakes autonomous decisions.

Reproducibility Statement

Code. The anonymous protocol demonstration is available at:
`https://anonymous.4open.science/r/smalljudge-bench-92E7/`.

Data. The dataset is available at:
`https://huggingface.co/datasets/blindpeerbench/samlljudgebench`.

Scope. The code artifact is a minimal protocol demonstration, not a complete reproduction package. Appendices E–G describe edit construction, review procedures, model configurations, prompts, parser rules, and Encoder training. Appendices J and K document identity groups and matched generator batches. Other appendices provide control denominators, costs, and bootstrap definitions. These specifications do not substitute for release of the trained checkpoint and full evaluation harness.

AI Use Statement

Candidate rewrites were generated with Qwen3-32B and GPT-4.1 as described in Appendices E and K; model judgments form the experimental outputs. AI assistance also supported experiment planning, result interpretation, deterministic sensitivity bounds, public literature and policy searches, manuscript restructuring, and language editing. We used AI-assisted code to transcribe supplied result summaries, generate figures, and check typesetting and numerical consistency. AI also provided internal critiques of the manuscript. The authors supplied the experimental results and remain responsible for the scientific content, including AI-assisted text, claims, and artifacts.

References

Sunishchal Dev, Andrew Sloan, Joshua Kavner, Nicholas Kong, and Morgan Sandler. Judge reliability harness: Stress testing the reliability of LLM judges, 2026. URL `https://arxiv.org/abs/2603.05399`.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M Khapra. Finding blind spots in evaluator LLMs with interpretable checklists. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16279–16309, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.911. URL `https://aclanthology.org/2024.emnlp-main.911/`.

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, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL `https://aclanthology.org/2020.findings-emnlp.117/`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International*

Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. OffsetBias: Leveraging debiased data for tuning evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.57. URL https://aclanthology.org/2024.findings-emnlp.57/.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*, 2024. doi: 10.48550/arXiv.2407.08488. URL https://arxiv.org/abs/2407.08488.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, pp. 63277–63303, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9e720fce64f91114c49cfd640d821da3-Paper-Conference.pdf.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11626–11644, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.650. URL https://aclanthology.org/2023.acl-long.650/.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Audit Results and Sampling Scope

Dataset	n	Label	Relation	Both	Agreement	κ
SciFact	60	100.00	98.33	98.33	96.67	0.487
FEVER	60	100.00	98.33	98.33	100.00	1.000
VitaminC	60	100.00	96.67	96.67	98.33	0.659
AggreFact/TRUE	60	100.00	96.67	96.67	100.00	1.000

Table 6: Re-audit of accepted pairs. Label, relation, joint preservation (Both), and agreement are percentages with 60 sampled pairs per dataset. Cohen's κ is not a percentage. The pooled joint count is 234/240; the high-risk diagnostic slice is excluded.

The per-dataset joint counts are 59/60, 59/60, 58/60, and 58/60, respectively. The pooled agreement is 98.75%, with $\kappa = 0.763$. Individual dataset values of κ should be read with their marginal-label imbalance in mind; for example, SciFact has 96.67% agreement but $\kappa = 0.487$. Neither perfect label preservation in this sample nor a high pooled agreement implies that every accepted edit is valid. Two independent judgments do not by themselves establish external validation; we make no such claim.

The construction counts and re-audit outcomes answer different questions. Table 1 records how many candidate edits entered the evaluation. Table 6 examines whether sampled accepted edits satisfy the intended preservation criterion. The high-risk slice, with 6 preserved, 43 invalid, and 11 disputed pairs, is useful for diagnosing difficult boundaries but is not an additional 60 random accepted examples. Combining it with the accepted slice would change the estimand and distort the reported preservation rate.

Sensitivity to known failures. For a paired model contrast bounded by $[-1, 1]$, removing k of N pairs changes its mean by at most $2k/N$. Applying equal dataset weights to the six known failures gives a maximum shift of 0.4375 pp in the full-cohort Sonnet–Encoder RC difference. Thus, deleting these pairs cannot reverse the reported $+4.80$-pp point estimate. This is a deterministic bound, not a new experiment or a recomputed CI. It does not cover undetected failures, corrected labels, or the conditional subsets, whose affected counts are not specified.

B Content-Control Counts and Denominators

System	Clean correct	Style RC	Style CW	Content joint count	Joint	Conditional
Encoder	172	84.00	2.00	146	73.00	84.88
MiniCheck	182	86.00	5.00	151	75.50	82.97
Lynx	177	83.50	5.00	153	76.50	86.44
Sonnet 4.6	184	87.50	4.50	162	81.00	88.04
Qwen3-32B	190	91.50	3.50	167	83.50	87.89

Table 7: Separately evaluated content controls on 200 originals, not the main evaluation cohort. Clean-correct and joint-correct entries are counts; other entries are percentages. Conditional correctness divides the content joint count by the model's clean-correct count, while Joint divides it by 200. Style RC and CW use the same originals with style edits.

The control balances the direction of binary label changes: 100 support-to-non-support and 100 non-support-to-support. A style edit retains y_i, whereas a content edit has a distinct gold label y_i^c. The two evaluation pairs share an original but are not interchangeable observations. In particular, failure to change a label is desirable only when the initial decision is correct and the correct label stays fixed. An arbitrary change of label is insufficient when the content changes; Equation 3 requires the new label to be correct as well.

C METRIC AND ROUTING ACCOUNTING

All four-dataset headline metrics use equal dataset weights after within-dataset pair aggregation. This differs from pooling all 2,800 pairs because datasets have different sizes. Within each dataset, repeated originals contribute through each associated edit. The raw original count of 1,400 is therefore neither the pair-metric denominator nor evidence that 2,800 observations are independent. Appendix D reports intervals that retain this clustering and use shared resamples for model comparisons.

In a multiclass task, Flip can include a transition between two incorrect labels (WD). Two incorrect valid binary labels cannot differ for the same gold label; this restriction does not cover an `INVALID` output. Clean and edited accuracy describe marginal decisions; RC and CW retain their pairing. A model-specific conditional error rate divides CW by clean accuracy within each dataset. Table 3 instead conditions on originals both models classify correctly before editing. Neither conditional quantity is obtained by dividing dataset-macro summaries.

Router confidence is evaluated separately for each original and edited input. Let q denote the weighted fraction of judgments sent to Sonnet. Under a constant per-input approximation with Encoder cost c_v and Sonnet cost c_L, total cost is $c_v + qc_L$. Encoder and Sonnet cost 0.80 and 50.00 units per 10,000 inputs, respectively. Table 8 reports the evaluated operating point. These are accounting assumptions, not measured monetary expenditure or latency.

Routing rule and threshold. Encoder's prediction is retained when confidence is at least $\tau = 0.88$; otherwise the input is sent to Sonnet. The threshold was selected in the preceding development procedure and was not reselected using these test results. That procedure used the grid $\{0.55, 0.65, 0.72, 0.78, 0.83, 0.88, 0.92, 0.96, 0.99\}$. It minimized fallback calls among thresholds whose edited accuracy and RC were each no more than 1 pp below always using Sonnet. Original and edited inputs are billed separately, without caching repeated originals.

The exact Encoder confidence function is not specified in the available execution record. The reported operating point and paired inference therefore do not, by themselves, provide a complete implementation of the gate.

System	Edited acc.	RC	Flip	Calls (%)	Cost units
Encoder only	88.88	87.62	3.17	0.00	0.80
Sonnet only	92.60	91.40	5.23	100.00	50.00
Encoder → Sonnet	94.20	92.64	4.49	41.97	21.79

Router minus Sonnet	Difference [95% CI] (pp)
Edited accuracy	$+1.59\,[0.84, 2.38]$
RC	$+1.24\,[0.38, 2.11]$
Flip	$-0.74\,[-1.44, -0.03]$

Table 8: Binary routing results. Rates and calls are percentages; cost units are per 10,000 inputs. The lower panel gives router-minus-Sonnet differences with 95% paired original-cluster CIs. The joint quality criterion uses the lower endpoints for edited accuracy and RC as two one-sided 97.5% bounds, with Bonferroni control at total $\alpha = 0.05$. Flip is supplementary.

Both quality lower bounds exceed -1 pp, meeting the stated joint tolerance at nominal total $\alpha = 0.05$. The reported cost reduction is 56.43%. Subtracting the displayed accuracies, 94.20 and 92.60, gives 1.60 pp rather than the reported paired difference of 1.59 pp. This discrepancy is compatible with separate rounding. A 41.97% call rate gives 21.785 cost units before rounding to 21.79. We retain the reported paired estimates and cost reduction rather than recomputing them from rounded entries.

These results concern a single fixed threshold. They do not provide a threshold frontier, a latency measurement, or a performance guarantee on future distributions. The augmentation standard deviations and routing confidence bounds quantify different sources of uncertainty; neither substitutes for the other.

D Clustered Bootstrap Intervals

The main cohort contains 1,400 original-ID clusters and 2,800 pairs. Each of 10,000 bootstrap replicates resampled originals with replacement within each dataset, retaining all edits for every sampled occurrence. Compared models shared resampling indices. Metrics were recomputed within each dataset before equal-weight averaging. Conditional and matched-generator analyses use their explicitly stated subsets rather than the full-cohort denominator.

Pointwise 95% CIs use the 2.5th and 97.5th percentiles with linear interpolation. No p-values are reported. A paired difference is computed within each shared resample, not by subtracting marginal interval endpoints. The native Sonnet-minus-Encoder RC difference is $+4.80$ pp with CI $[2.66, 6.99]$. Table 2 gives the marginal RC and Flip intervals; Table 3 gives the conditional comparisons.

The Encoder-router comparison follows this protocol on the binary cohort, with threshold 0.88 fixed throughout. For the joint one-percentage-point tolerance, each of edited accuracy and RC uses a one-sided 97.5% lower bound. Bonferroni control allocates 0.025 error probability to each metric, for total $\alpha = 0.05$. These bounds are the lower endpoints of the reported two-sided percentile intervals. This adjustment applies to the two routing quality conditions, not to all comparisons in the paper.

Augmentation intervals resample originals for the mean of three fixed training seeds. They quantify evaluation-sample uncertainty conditional on those seeds. The sample standard deviations in Table 5 instead describe the three observed training runs. Neither quantity establishes uncertainty over an unrestricted population of training seeds.

E Edit Construction, Acceptance, and Re-Audit

Generator and sampling. Verbosity and style edits use `Qwen/Qwen3-32B` in BF16, with thinking disabled, temperature 0.7, `top_p=0.8`, `top_k=20`, and `max_new_tokens=1024`. The generation seed is 2026092301. Initially, each of 1,400 originals receives two candidate edits. Exactly 184 originals receive one replacement candidate; the other 1,216 receive no additional candidate. The 184 rejected initial edits are replaced by 184 accepted supplementary edits. This yields 2,984 candidates, 2,800 accepted pairs, and exactly two accepted edits per original.

SciFact, VitaminC, and AggreFact/TRUE retain one verbosity and one style edit per original. Each FEVER original retains an identity edit and one other edit: verbosity for 200 originals and style for 200. Thus, the accepted set contains 1,200 verbosity, 1,200 style, and 400 identity pairs. Identity metadata names GPT-4.1 or Qwen3-32B in a 1:1 ratio; these names are experimental cues, not claims about actual authorship. Neither generation nor acceptance uses evaluated judge labels or scores.

Generation prompt and identity template. Evidence is supplied unchanged. The following prompt governs verbosity and style; the final instruction specifies the separate fixed identity operation.

```
Rewrite CLAIM using OPERATOR while preserving its exact factual
content and its support relation to EVIDENCE. Preserve entities,
quantities, dates, negation, scope, attribution, and certainty.
Add no evidence or factual claims.
verbosity: expand the wording to about 1.3-1.8 times its length
using restatement.
style: improve fluency and organization without changing
certainty or scope.
Return only the edited claim.
EVIDENCE: {evidence}
CLAIM: {claim}
OPERATOR: {verbosity_or_style}

identity: prepend "[Model identity: {GPT-4.1_or_Qwen3-32B}]" on a
separate metadata line, leave CLAIM unchanged, and balance the
two identities 1:1.
```

Initial acceptance. Programmatic screening checks field completeness and entities, numbers, dates, and negation in the claim body. The identity metadata line is excluded from factual checks. Two

graduate students with NLP and fact-checking experience independently assess whether core facts, the gold label, evidence relation, certainty, and attribution are preserved. Agreement to accept retains the pair. A third, senior researcher adjudicates disagreements; uncertain cases are excluded. Screeners do not see model outputs. Rejections number 40, 44, 73, and 27 for SciFact, FEVER, VitaminC, and AggreFact/TRUE, respectively.

Independent re-audit. Within each dataset, 60 accepted pair IDs are sampled uniformly without replacement. Sampling uses seed 2026092301 in a random stream independent of generation. Two graduate students with fact-checking experience, neither involved in initial screening, independently assess label and evidence-relation preservation. The interface presents only evidence and the original and edited texts, hiding gold labels, initial decisions, and model outputs. Matching judgments determine the final outcome; disagreements are adjudicated by another senior researcher who did not participate in initial screening. Both raw judgments and the final decision are retained. Independence here concerns personnel, judgments, and information access, not an external organizational audit.

F Evaluated Models, Prompts, and Output Rules

Encoder. The evaluated checkpoint is `modernbert-large-sjb-clean-s42`, based on `answerdotai/ModernBERT-large`. Evaluation disables dropout. Native labels use the task head's argmax; the common binary comparison retains support and folds the other valid native labels into non-support.

GPT-4.1. We use model snapshot `gpt-4.1-2025-04-14`, accessed through Chat Completions with temperature 0, `top_p=1`, and `max_completion_tokens=128`. Each input receives one judgment, without tools or retrieval, using the unified prompt below.

Sonnet 4.6 and Qwen3-32B. Sonnet uses API ID `claude-sonnet-4-6`. Thinking is disabled, temperature is 0, and `max_tokens` is 128. Qwen uses `Qwen/Qwen3-32B` at repository revision:

`940b12c09f7c9d6f3e7cf3046bc6b41564f568bf`

Qwen inference uses BF16, `enable_thinking=False`, `do_sample=False`, and `max_new_tokens=128`. Both models use the unified prompt.

MiniCheck. MiniCheck uses checkpoint `lytang/MiniCheck-Flan-T5-Large`. The official `MiniCheck` interface sets `model_name='flan-t5-large'`, with evidence passed as `docs` and the claim as `claims`. A native support score of at least 0.50 indicates support. Its native label 1 maps to our support code 0, and native label 0 maps to non-support code 1. Numeric encodings do not change the semantic labels.

Lynx. We use `PatronusAI/Llama-3-Patronus-Lynx-70B-Instruct` with BF16, `do_sample=False`, and `max_new_tokens=600`. The official model-card prompt format receives QUESTION = "Is this claim supported by the document?", DOCUMENT = evidence, and ANSWER = claim. The JSON `SCORE` values `PASS`/`FAIL` map to support/non-support. MiniCheck and Lynx are included in the separate binary content-control experiment, not the native four-model comparison.

Unified judge prompt. SciFact allows SUPPORTS and REFUTES. FEVER and VitaminC also allow NEI; AggreFact/TRUE uses CONSISTENT and INCONSISTENT.

```
Judge CLAIM using only EVIDENCE. Ignore verbosity, writing style,
and model identity metadata as evidence. Do not use outside knowledge.
Choose one label from ALLOWED_LABELS and return only {"label":"LABEL"}.
SUPPORTS: evidence entails the claim; REFUTES: evidence contradicts
the claim; NEI: neither relation is established. For consistency
tasks, use CONSISTENT only when every factual assertion is supported;
otherwise use INCONSISTENT.
ALLOWED_LABELS: {dataset_native_labels}
EVIDENCE: {evidence}
CLAIM: {claim}
```

Parsing and failures. The parser removes an outer code fence, reads a single JSON object, and validates the label against the dataset's allowed set. Refusals, illegal labels, and parsing failures receive `INVALID`, count as incorrect, and do not remove the pair. API network errors allow at most two retries; content errors do not trigger another judgment. Invalid outputs are not valid non-support labels. Their presence would require distinguishing valid-label transitions from failure transitions when interpreting Flip.

G Encoder Training and Data Separation

The base model is `answerdotai/ModernBERT-large`, with a shared encoder and separate two- or three-class task heads. Training uses only clean originals from the four datasets. Cross-domain deduplication groups examples by source document or claim group. The test, development, and control pools hold out 1,400, 640, and 400 originals, respectively, with the remaining pool used for training. An original's source group and edits never cross splits. The content-changing comparison uses 200 originals and retains its own explicitly reported denominator; it does not replace these overall pool sizes.

Training uses AdamW with learning rate 2×10^{-5}, weight decay 0.01, three epochs, micro-batch size 4, and eight gradient-accumulation steps. Effective batch size is 32. Warmup covers 10% of training, followed by linear learning-rate decay. Maximum input length is 2,048 tokens, precision is BF16, and the seed is 42. Inputs concatenate evidence and claim. Overlength inputs retain the complete claim and truncate the evidence tail. After each epoch, checkpoint selection uses clean macro-F1 averaged equally across the four development datasets; ties retain the earlier epoch. The selected checkpoint is identified as `modernbert-large-sjb-clean-s42`.

The supplied training record specifies this truncation policy but does not establish whether evaluation retained identical evidence tokens for each original and edit. Fixed stored evidence therefore does not, by itself, exclude an input-truncation effect.

These settings describe the main clean-trained Encoder. The augmentation experiment is a separate comparison with three seeds and an equal update budget across all conditions. Standard deviations describe between-seed variation; paired bootstrap intervals describe original-sample uncertainty conditional on those seeds. The main checkpoint's seed does not document all augmentation seeds or the proportion of rewritten training examples.

H Conditional Errors on Shared Initially Correct Originals

For each dataset d, let $\mathcal{U}_d^{\cap}$ contain original IDs on which both Encoder and Sonnet predict the correct native label. Membership is fixed before resampling. We retain the accepted edits associated with these IDs, producing the pair set $\mathcal{P}_d^{\cap}$. In the full test set, this gives 1,153 originals and 2,306 pairs. In the matched generator comparison, it gives the same 334 originals and 577 pairs for G1 and G2. These subsets are not the full 400-original, 695-pair generator cohort.

For model m, the conditional edited error is

$$e_m^{\cap} = \frac{1}{4} \sum_{d=1}^{4} \frac{1}{|\mathcal{P}_d^{\cap}|} \sum_{i \in \mathcal{P}_d^{\cap}} \mathbf{1}\{\widehat{y}'_{m,i} \neq y_i\}. \tag{4}$$

The comparison is $e_{\text{Sonnet}}^{\cap} - e_{\text{Encoder}}^{\cap}$. Each of 10,000 bootstrap replicates resamples eligible original IDs with replacement within each dataset. Both models share the sampled indices, and every sampled original retains all associated edits. Replicates first compute within-dataset pair error, then average equally across four datasets. The interval uses the 2.5th and 97.5th percentiles of paired differences, with linear interpolation and no multiplicity adjustment.

This estimates conditional error on the observed shared subset, not on all 2,800 pairs. It does not reselect subset membership within each replicate. Both models are initially correct on every retained original, so edited error measures loss of correctness on a common starting set. The full-test difference is $+2.64$ pp ($[1.63, 3.67]$), favoring Encoder. The matched G1 and G2 differences are $+2.94$ pp ($[0.66, 5.23]$) and $+3.67$ pp ($[1.27, 6.11]$), respectively. These intervals support a conditional ordering, not a causal explanation based on architecture or model size.

I Matched-Budget Training Augmentation

The three conditions are A (Clean), B (Repeated clean), and C (Augmented). All use the same update budget and three training seeds; C adds label-preserving rewrites. The sample standard deviations in Table 5 describe variation across these runs, not evaluation-set confidence intervals. Paired comparisons use the mean over these fixed seeds before original-cluster bootstrap analysis.

Compared with B, C increased edited accuracy by 1.55 pp ($[1.20, 1.92]$) and RC by 1.71 pp ($[1.33, 2.11]$). Flip decreased by 0.72 pp ($[-0.97, -0.49]$). The C-minus-A comparisons show the same direction. This recipe improved correctness and stability relative to the repeated-clean control, within the three observed seeds. It does not establish an architectural advantage or the variability of the improvement over unobserved training seeds.

J FEVER Identity Groups

The 400 FEVER original IDs were sorted and randomly assigned to two equally sized metadata groups without consulting gold labels or model outputs. Each original received only one identity value. The metadata naming GPT-4.1 or Qwen3-32B was prepended to an otherwise unchanged claim. Table 9 reports the realized gold-label composition, not model predictions. Each original contributes one identity pair, so the original and pair counts coincide here.

Metadata identity	U	P	Gold S / R / NEI
GPT-4.1	200	200	72 / 75 / 53
Qwen3-32B	200	200	70 / 63 / 67

Table 9: FEVER identity-group composition. U and P count originals and identity pairs. S/R/NEI counts are gold-label counts, not model predictions.

Random assignment does not make the realized groups identical: the GPT-4.1-named group contains 53 NEI originals and the Qwen-named group 67. Their different originals and label composition prevent interpreting a group difference as a within-original identity effect. This design does not establish causal self-preference.

K Matched Generator-Batch Design and Results

Sampling and generation. The sampling design fixes 100 test originals per dataset and stratifies within datasets by native gold label. FEVER retains 50 originals from each existing text-edit type. G1 denotes the retained Qwen3-32B text edits from the original benchmark. G2 uses `gpt-4.1-2025-04-14`, generated on September 26, 2026, with temperature 0.7, `top_p=0.8`, and `max_completion_tokens=1024`. No model-generation seed, tools, or retrieval were supplied. The preservation prompt is the same as in Appendix E. The 100 selected FEVER identity edits are excluded; they are templates rather than model-generated rewrites.

Each selected original–operator combination defines a slot. Up to three candidates are allowed, stopping at the first accepted candidate. Exhausted slots are not replaced by other originals. Table 10 reports 819 candidates: 695 accepted and 124 rejected. Five of 700 slots exhausted their candidate budget. Removing those slots from both arms leaves 695 matched pairs in each arm, with all 400 originals still represented. This common-set comparison conditions on G2 acceptance; it does not estimate performance on rejected edits.

Acceptance and re-audit. Two graduate students with NLP and fact-checking experience independently screened each candidate, with a third reviewer resolving disagreements. Generator identity and judge outputs were hidden. There were 34 initial disagreements. Two other reviewers independently re-audited 20 accepted pairs per dataset; another senior researcher adjudicated their four disagreements. Re-auditors could not see initial decisions or model outputs. Label preservation held in 78/80 cases and relation preservation in 77/80; joint preservation was 77/80 (96.25%; Table 11). This exceeds the design's overall 95% continuation threshold, which is an operational rule rather than

Dataset	Selected U	Text slots	Total candidates	Rejected candidates	G2 accepted P	Exhausted slots	Common U	Common P
SciFact	100	200	222	22	200	0	100	200
FEVER	100	100	114	14	100	0	100	100
VitaminC	100	200	250	53	197	3	100	197
AggreFact/TRUE	100	200	233	35	198	2	100	198

Table 10: G2 construction counts. Slots are selected original–operator combinations; rejected candidates and exhausted slots are different counts. Common pairs include only slots accepted in both generator batches.

a confidence bound. There was no batch reconstruction. The three re-audit failures remain in the frozen performance denominator.

Dataset	n	Label preserved	Relation preserved	Both preserved
SciFact	20	18	18	18
FEVER	20	20	20	20
VitaminC	20	20	20	20
AggreFact/TRUE	20	20	19	19
Total	80	78	77	77

Table 11: Independent G2 re-audit. Entries are counts, not percentages. Twenty accepted pairs were reviewed per dataset. Equal audit sample sizes make 77/80 a dataset-macro joint-preservation summary, not a pair-weighted estimate over the 695 retained pairs.

Judge evaluation and statistical contrasts. Clean and G1 predictions are reused; G2 contributes 2,780 edited judgments (695×4) under the specified configurations, with no recorded network failures or retries. Failure outputs remain incorrect and in the denominator. We apply the 10,000-replicate percentile procedure from Appendix D, sharing original indices across judges and batches and retaining each sampled original's slots. Each replicate computes pair rates within datasets before equal-weight averaging.

Metric	G1 gap [95% CI]	G2 gap [95% CI]	Interaction [95% CI]
RC	$+1.40\,[-2.48, 5.23]$	$+0.78\,[-3.14, 4.80]$	$-0.62\,[-2.63, 1.39]$
Flip	$+4.27\,[1.85, 6.77]$	$+5.27\,[2.63, 7.90]$	$+1.00\,[-1.15, 3.25]$

Table 12: Sonnet-minus-Encoder gaps and generator interactions in pp, with pointwise 95% shared original-cluster CIs. Both batches contain the same 695 slots from 400 originals. Interaction means the G2 model gap minus the G1 model gap; metrics are equally weighted across datasets.

The fixed primary comparison is the RC interaction,

$$\delta_{\mathrm{RC}} = (\mathrm{RC}_{\mathrm{Sonnet},2} - \mathrm{RC}_{\mathrm{Encoder},2}) - (\mathrm{RC}_{\mathrm{Sonnet},1} - \mathrm{RC}_{\mathrm{Encoder},1}). \tag{5}$$

It is computed within each shared resample, not by subtracting marginal interval endpoints. The Flip interaction and conditional subset comparisons are supplementary pointwise comparisons without multiplicity correction. All supplied values are retained at their reported precision.

The primary RC interaction, -0.62 pp with CI $[-2.63, 1.39]$, remains unresolved. The Flip interaction is $+1.00$ pp with CI $[-1.15, 3.25]$. The lower conditional error for Encoder in Table 3 does not establish either interaction. Batch timing, reused G1 judgments, and generator–judge overlap prevent a causal self-preference or universal-transfer interpretation.

Deleting the three known G2 audit failures from both matched batches would change the RC interaction point estimate by at most 1.5051 pp. This follows from its per-pair range $[-2, 2]$ and the dataset-specific failed-pair counts. It cannot establish an interaction or certify revised CI endpoints; unknown preservation errors remain outside this bound.