NEURIPS 2026 · WORKSHOP ON AI FOR VERIFIABLE CODING

# Three Agents Are Not Three Verifiers

## Measuring Common-Mode Failure in LLM Verification

**Adarsh Agrawal**

Independent Researcher

openreview.net/forum?id=Klbp3xuEkw

ALL THREE VERIFIERS WRONG

*Held-out candidates, cross-model trio*

**Observed** **8 of 47**



**If errors were independent** **≈ 1.4 of 47**

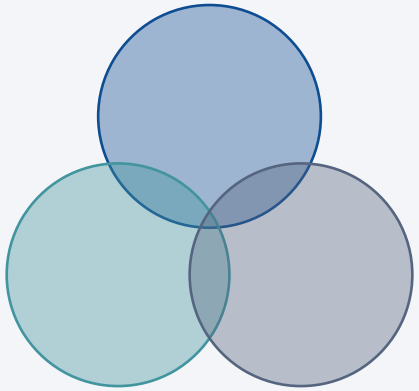


Claude Sonnet 5 + Qwen3-Coder + DeepSeek V3.2
17.0% observed vs 3.0% expected (5.7×)

MOTIVATION

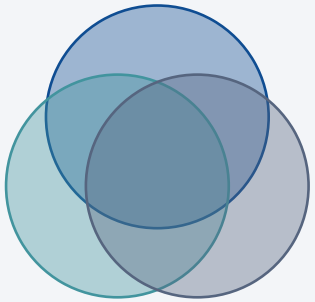
# Agreeing verifiers are not independent evidence

**If errors were independent**



**3.0%**
all three wrong

**What we measure**



**17.0%**
all three wrong

Held-out cross-model trio. Circles are schematic; the rates are measured.

THE MEASUREMENT

**All-member failure**
Share of candidates on which every verifier in panel S is wrong

$$JF(S) = (1/n)\ \Sigma_x\ \Pi_i\ e_i(x)$$

**Independence product**
What joint failure would be if member errors were independent

$$JF_{ind}(S) = \Pi_i\ p_i$$ ($p_i$ = error rate of verifier i)

Ratio JF / $JF_{ind}$ = 1 under independence
**> 1 means common-mode failure**

**Votes add evidence only when errors rarely overlap.** LLM verifiers share training sources, conventions, prompt interpretations, and blind spots.

EXPERIMENTAL DESIGN

# Three linked studies, six verifier models

### CONTROLLED EXECUTABLE STUDY

**1,152** judgments

64 artifacts from 23 task groups: 16 Verus, 24 Lean, 12 SMT, and 12 repository patches with pinned pytest outcomes.

47 held-out and 17 calibration candidates, split by task group.

Six models × three runs; no parse failures.

### INDEPENDENT REPOSITORY STUDY

**141** candidates

47 popular PyPI packages × 3: an accepted PR, a submitted-then-revised PR, and a targeted mutation.

Labels: full tests, downstream tests, and two blinded expert reviews (κ = 0.73).

130 resolved labels: 71 correct, 59 wrong. An external reviewer settled the 11 disputes.

### TARGETED MUTATION SEARCH

**3** mutation generators

Three generators: Claude Sonnet 4.7, DeepSeek V3.2, and Qwen3-Coder.

Return substitutions, boundary shifts, branch flips, off-by-one.

500-attempt cap; stop at 20 accepted.

Blinded evaluation, hidden regression tests.

VERIFIERS

**Claude Sonnet 5** | **Claude Opus 5** | **Qwen3-Coder** | **DeepSeek V3.2** | **Devstral 2** | **GPT-5.6 Terra**

All panels, including the calibrated trio, are fixed on the 17 calibration candidates before held-out and repository evaluation.

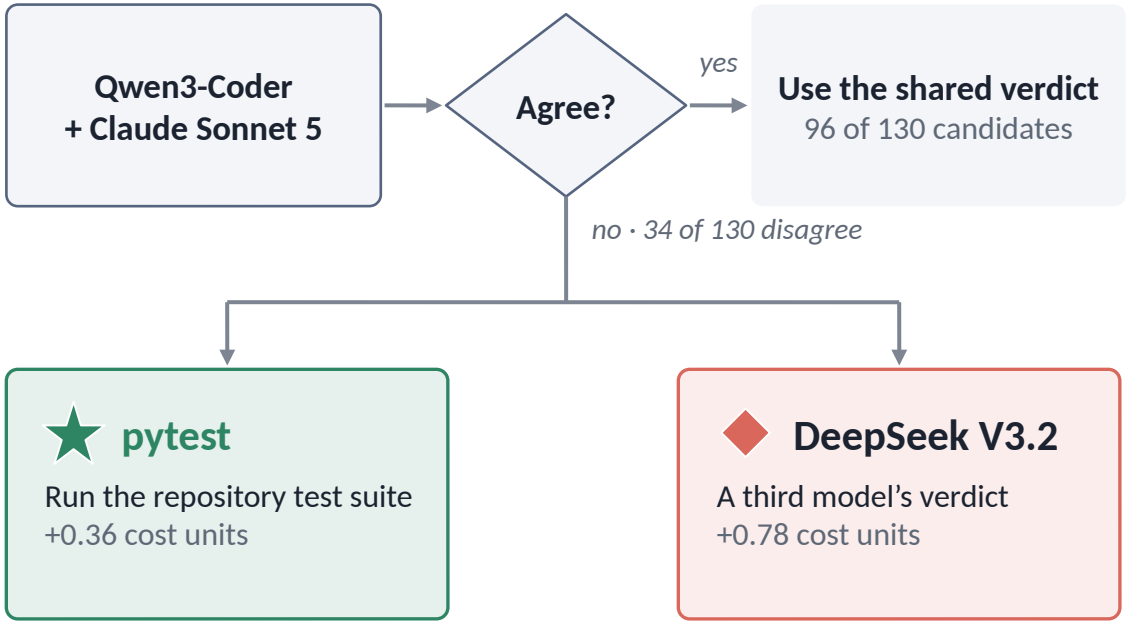
# More votes, or a different checker?

## Model-only policies

- **Single judge**
  GPT-5.6 Terra or Qwen3-Coder, one call
- **3× Repeated Sonnet**
  Claude Sonnet 5 three times; accept with ≥ 2 votes
- **Uncalibrated 3-Model Vote**
  Fixed panel: Sonnet + Qwen + DeepSeek, majority
- **Calibrated 3-Model Vote**
  Best of 816 trios on calibration: DeepSeek, Devstral, Opus
- **Calibrated Vote + Conf. Filter**
  Also requires every member's confidence ≥ 0.6

$$\text{Calibration score} = 0.4\,A_{\text{maj}} + 0.3\,(1 - JF) + 0.2\,\bar{A} + 0.1\,F$$

## Disagreement routing: same trigger, different tiebreaker




Expected cost per candidate: 1.91 with pytest vs 2.02 with DeepSeek, in units of one median Claude Sonnet 5 call.

# Joint failure is 3.3–9.7× the independent rate

**5.7×**

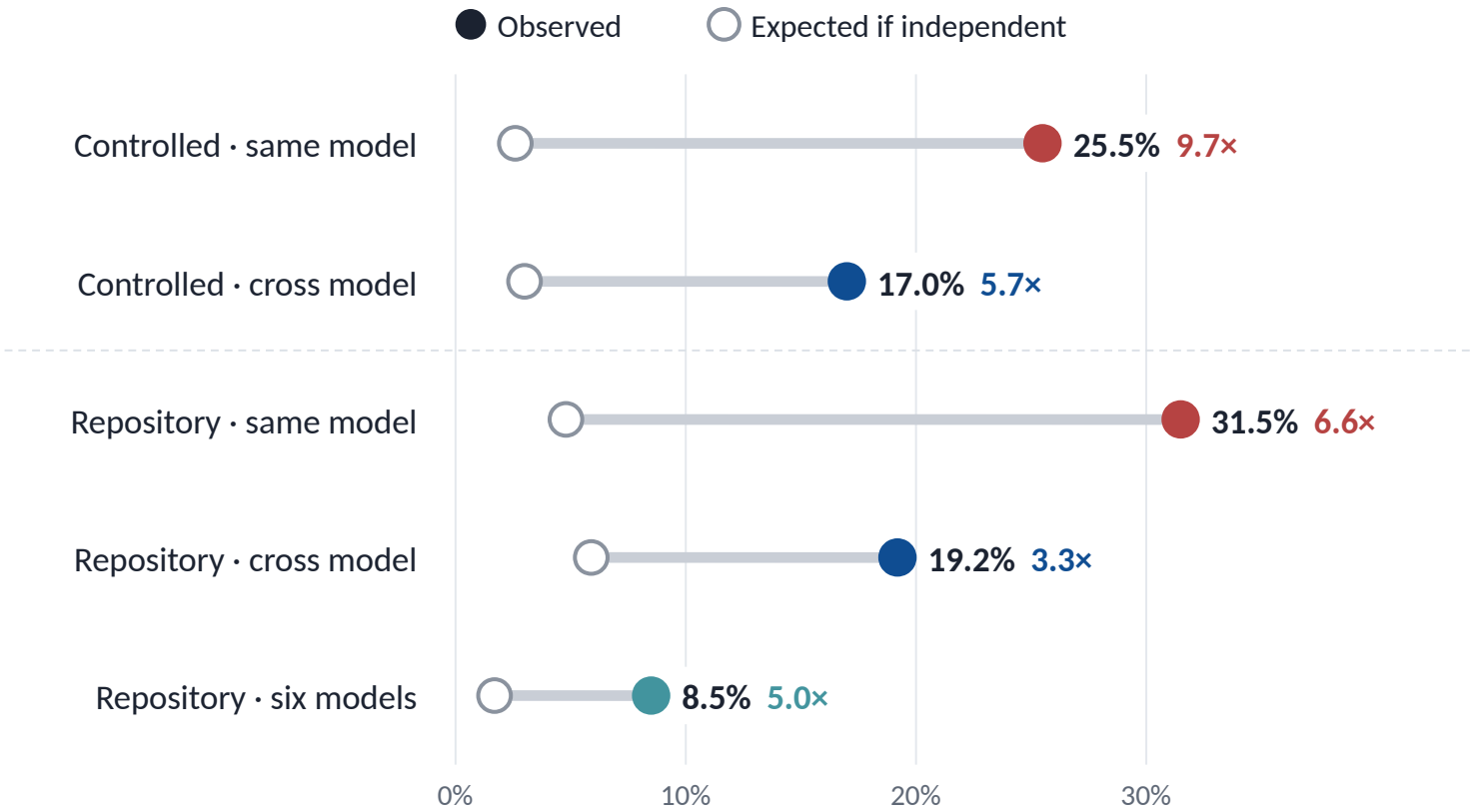
**Held-out cross-model trio**

17.0% of candidates all wrong (8 of 47) vs 3.0% if errors were independent

**[2.4, 13.8]×**
95% delta-method interval for the ratio

**p = 0.002**
cluster-preserving permutation test

**p < 0.001**
all four exact binomial tests (Bonferroni)




All-member failure rate: observed vs. product of member error rates

# Shared errors span every model pair

PAIRWISE ERROR CORRELATION (φ) · 47 HELD-OUT CANDIDATES




Sonnet = Claude Sonnet 5 · Opus = Claude Opus 5 · Terra = GPT-5.6 Terra
Qwen = Qwen3-Coder · DeepSeek = DeepSeek V3.2 · Devstral = Devstral 2 · nominal p = 0.003–0.048

**15 of 15**

pairs positively correlated: φ = 0.29–0.58, mean 0.40
(0.38 without the same-provider Sonnet/Opus pair)

**1.62–1.78×**

error lift: when one model is wrong, the others' mean error rate rises from 27.1% to 43.8–48.2%

**3.7–6.0×**

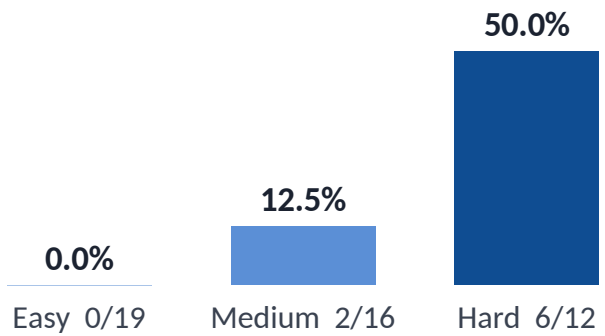
excess over independence for alternative model pairs on the same 47 held-out candidates

Joint failures in every source: 2/10 VeriContest, 2/8 FVSpec, 2/21 ITPEval, 2/8 repository

RESULT 1 · WHERE SHARED ERRORS COME FROM

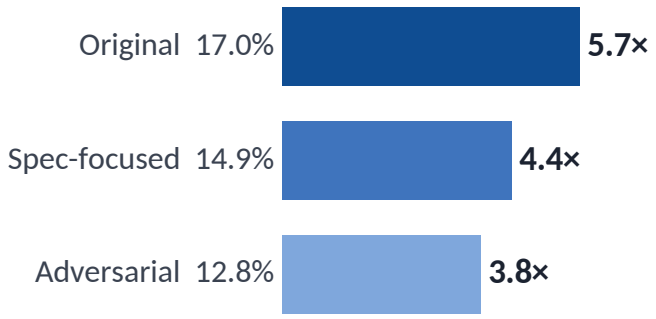
# Shared failures concentrate on hard cases

JOINT FAILURE BY DIFFICULTY




Band = how many of the six models are wrong (easy 0–1, medium 2–3, hard 4–6); descriptive

EXCESS UNDER PROMPT VARIANTS




Labels: joint failure · bars: ratio to independence

**61%**

of cross-model error variance is explained by one shared factor (loadings 0.54–0.79)

Devstral 2 loads lowest (0.54), matching its lower pairwise correlations.

WHAT SHARED FAILURES LOOK LIKE

### Valid but unfamiliar

Correct Lean and Verus artifacts are rejected because valid constructs look unfamiliar to a general code model.

### Convincing approval

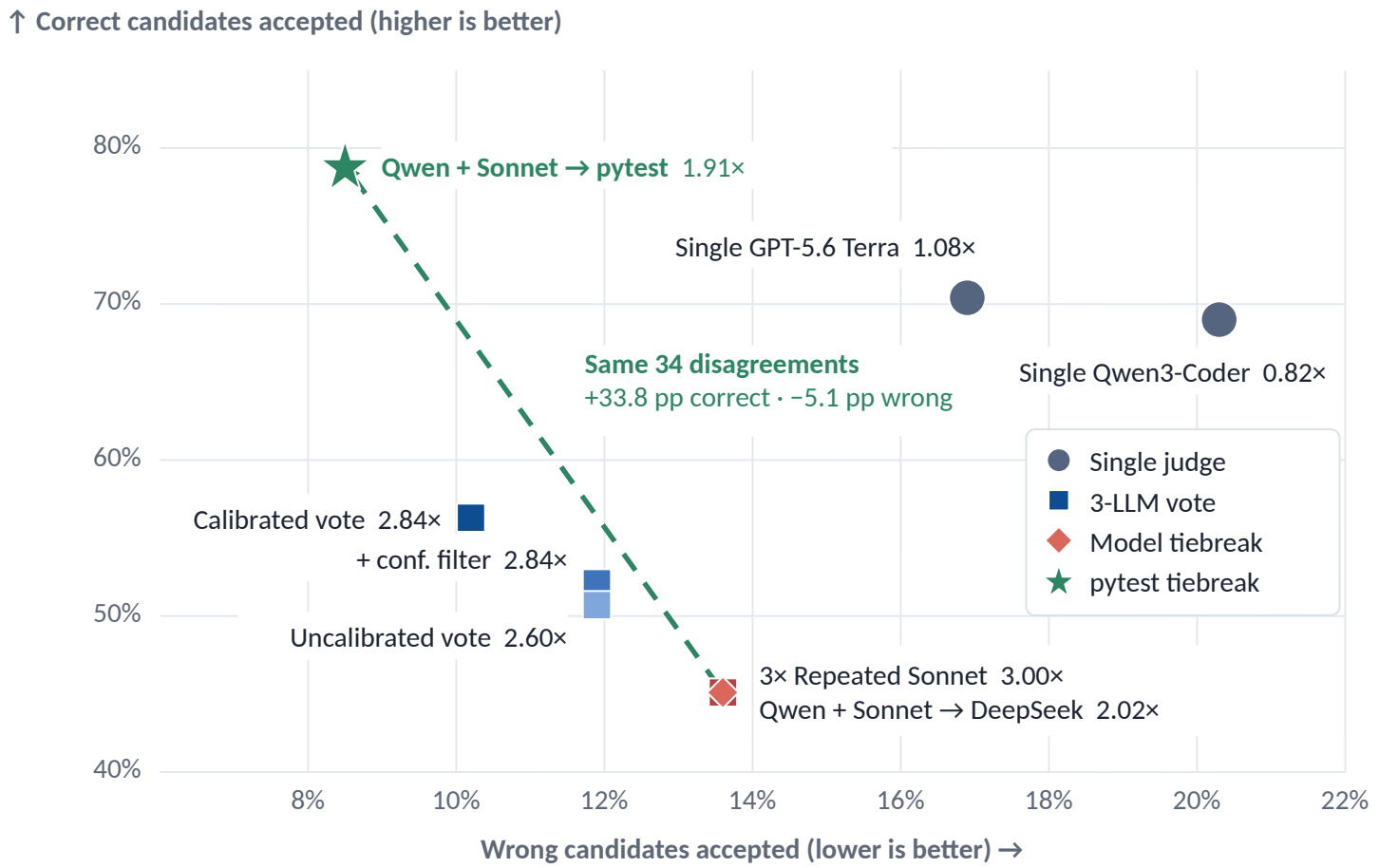
Repository patches get persuasive verbal approval while violating behavior that full or downstream tests expose.

### One reading of intent

Models often share an interpretation of the intended contract, even when their explanations differ.

RESULT 2 · OPERATING POINT

# A different checker beats another vote




130 repository candidates (59 wrong, 71 correct). Cost in units of one median Claude Sonnet 5 call.

PYTEST INSTEAD OF DEEPSEEK

**+33.8 pp**
correct acceptance (45.1% → 78.9%)

**−5.1 pp**
wrong acceptance (13.6% → 8.5%)

**84.6%**
accuracy, vs 63.8%; cost 1.91× vs 2.02×

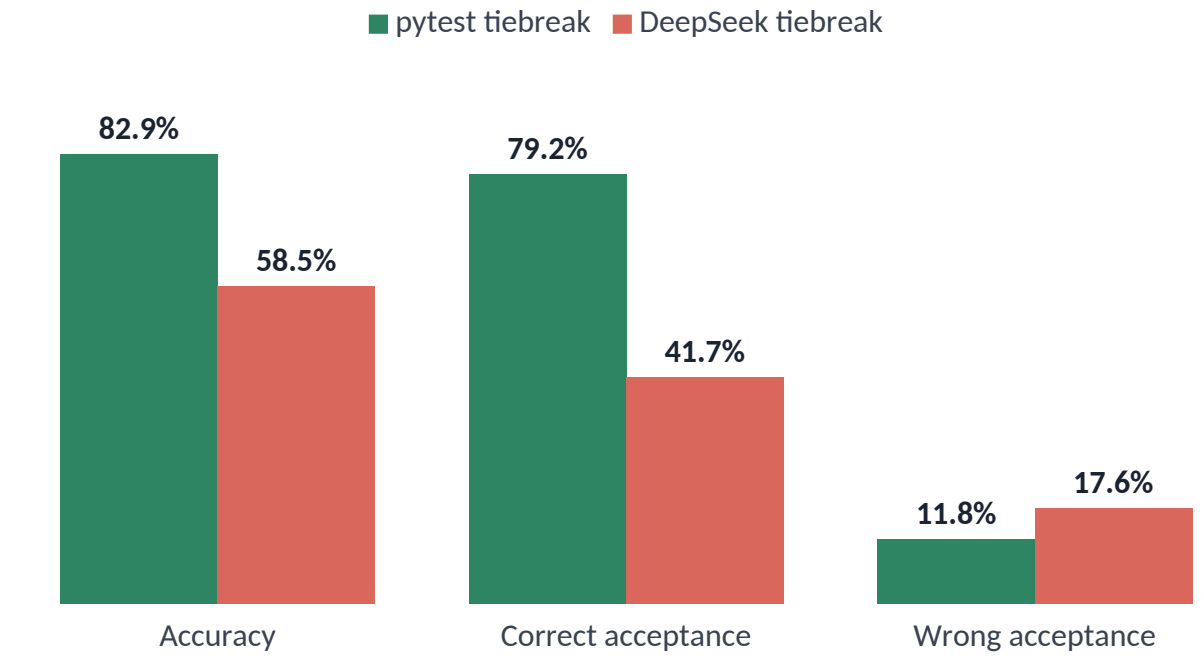
MODEL-ONLY TRADEOFF

Versus Single GPT-5.6 Terra, the calibrated vote cuts wrong acceptance from 16.9% to 10.2%, but correct acceptance falls from 70.4% to 56.3%. Terra keeps the best model-only accuracy (76.2%).

RESULT 2 · LABEL ROBUSTNESS

# The routing gain survives independent labels

41 CANDIDATES WITH PYTEST-INDEPENDENT LABELS (17 WRONG, 24 CORRECT)




Lower is better for wrong acceptance. Labels here come only from expert review and downstream dependent-package tests. On 28 of the 34 escalations, pytest agrees with the composite reference label.

**+24.4 pp**

accuracy on the pytest-independent subset (34/41 vs 24/41)

**83.0%**

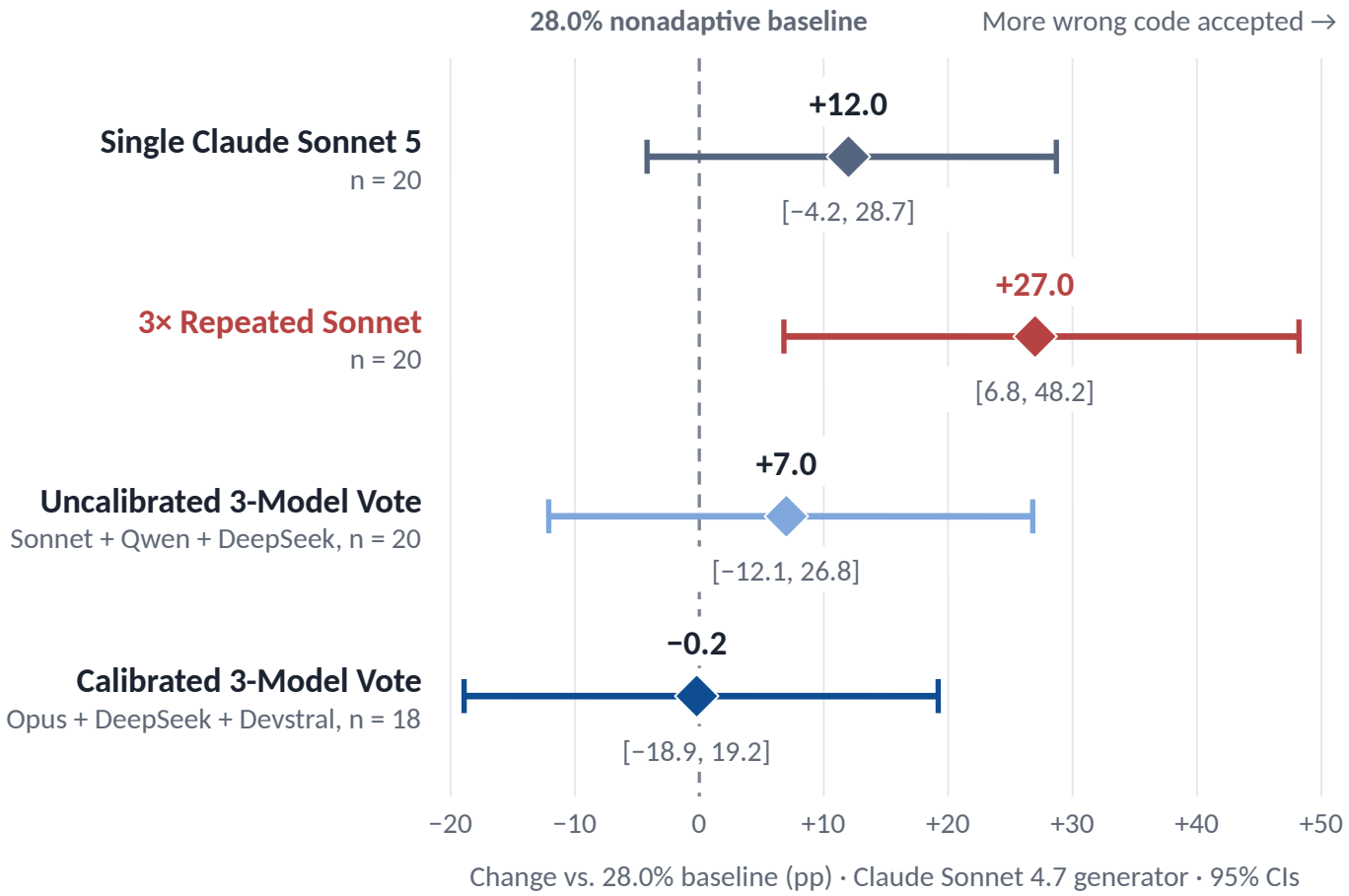
accuracy on all 141 candidates after external adjudication of the 11 disputes (wrong acceptance 9.5%)

**83.7–84.6%**

accuracy in every disputed-label scenario; wrong acceptance stays at 8.5–11.4%

*pytest also contributes to the composite labels, so the 41-candidate subset is supportive rather than confirmatory.*

# Targeted search exploits repeated voting




ACROSS THREE GENERATORS

**+17 to +27 pp**

3× Repeated Sonnet, above the 28.0% baseline

**within 3 pp**

Calibrated 3-Model Vote, all three generators

| Generator | 3× Repeated | Calibrated |
|---|---|---|
| Claude Sonnet 4.7 | 55.0% +27.0 | 27.8% −0.2 |
| DeepSeek V3.2 | 50.0% +22.0 | 25.0% −3.0 |
| Qwen3-Coder | 45.0% +17.0 | 30.0% +2.0 |

*Accepted-but-wrong share of up to 20 accepted mutations. Intervals spanning zero do not establish immunity.*

# Measure joint error on your own distribution

## Recommended workflow

1. **Same labeled examples**
   Evaluate every candidate verifier on one shared, labeled candidate set.

2. **Joint, not marginal, error**
   Report all-member failure against the independence product.

3. **Cost-aware operating point**
   Weigh wrong acceptance against correct rejection for your application.

4. **A different mechanism**
   Route disagreements to execution or another checker type; re-measure after prompt, provider, or model changes.

### LIMITATIONS

**Scope**
Executable formal artifacts and Python repositories with pytest suites; rates are distribution-specific.

**Small samples**
17 calibration candidates; adaptive search covers three generators, four operators, and up to 20 accepted mutations per policy.

**Shared pytest**
pytest informs both labels and routing; mitigated by independent labels and external adjudication.

**Hosted endpoints**
Provider-specific serving; in an 11-day repeat check, 46 of 48 decisions held stable.

TAKEAWAY

# Count independent checks, not agreeing models

## 1 Measure

Compare joint failure with the independence product, on your data.

**3.3–9.7×**

all-member failure vs. the independence rate

## 2 Diversify

Send model disagreements to a different checking mechanism.

**+33.8 pp**

correct acceptance with pytest tiebreaks; wrong acceptance falls 5.1 pp

## 3 Don’t repeat

Repeated votes from one model are a targetable blind spot.

**+17 to +27 pp**

accepted-but-wrong under targeted search, vs. a 28.0% baseline

**Scope:** Python repositories with pytest suites and formal artifacts from Verus, Lean, and SMT.

**In practice:** measure shared errors before treating another model vote as independent evidence.

**Thank you — questions welcome**

Paper and reviews on OpenReview

openreview.net/forum?id=Klbp3xuEkw