NEURIPS 2026 · WORKSHOP ON AI FOR VERIFIABLE CODING

When Correctness Feedback Erases Evidence

Phase-Separated Completeness Auditing in Verus

Adarsh Agrawal

Independent Researcher

openreview.net/forum?id=N6yfUR70Uu

CERTIFIED WITNESSES FROM THE STUDY

Each verifies under a weakened contract, yet is wrong.

`9876`	**cf271A · minimality omitted** input 1987: expected 2013
`w >= 6`	**cf4A · output relation omitted** input 4: expected YES
`true`	**cf110A · output relation omitted** input 4: expected NO
`0`	**cf492A · maximality omitted** input 1: expected 1

MOTIVATION

A verified program can still be wrong

EXAMPLE · VERICONTEST TASK cf271A

Task: return the smallest year after y whose digits are all distinct.

Weakened contract (simplified)

```
ensures r > y
ensures digits_distinct(r)
~~ensures r is the smallest such year~~
```

omitted

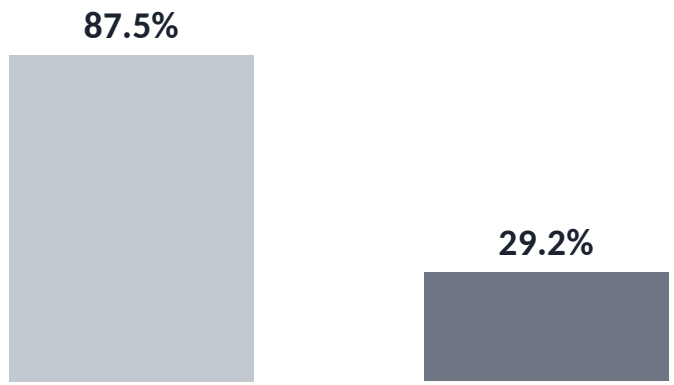
Witness

9876

Verifies under the weakened contract

Official case **y = 1987** expects 2013, not 9876

HOW COMMON IS THIS?



42 of 48 implementations verify against weakened contracts; only **14** still verify once the omitted postcondition is restored.

Downstream study, Original condition: 16 tasks × 3 attempts

Completeness witness: a program the contract accepts but the task rejects. It shows exactly what the contract forgot to forbid.

DEFINITION

A certified witness must pass four gates

1

Admitted

Verifies under the weakened contract.

2

Gap-exposing

Restoring the omitted postcondition makes verification fail.

3

Executable

Compiles and executes as Rust code.

4

Wrong

Disagrees with at least one official test case.

Selection

Each episode yields three candidates, ranked by parse success, then acceptance under the weakened contract, compilation, visible-test divergence, and call order. Only the top one is certified.

Certification

Replayed against the restored contract and the full official oracle bank. Gates 2 and 4 are separate evidence: failing the restored contract is not itself a behavioral counterexample.

Feedback that helps synthesis hurts auditing

Ordinary synthesis

Goal: a program that meets the intent

Candidate fails a visible test → Feedback: compile, verify, tests → Revised candidate passes

Feedback moves the search toward the goal

Completeness auditing

Goal: a program the contract admits but the task rejects

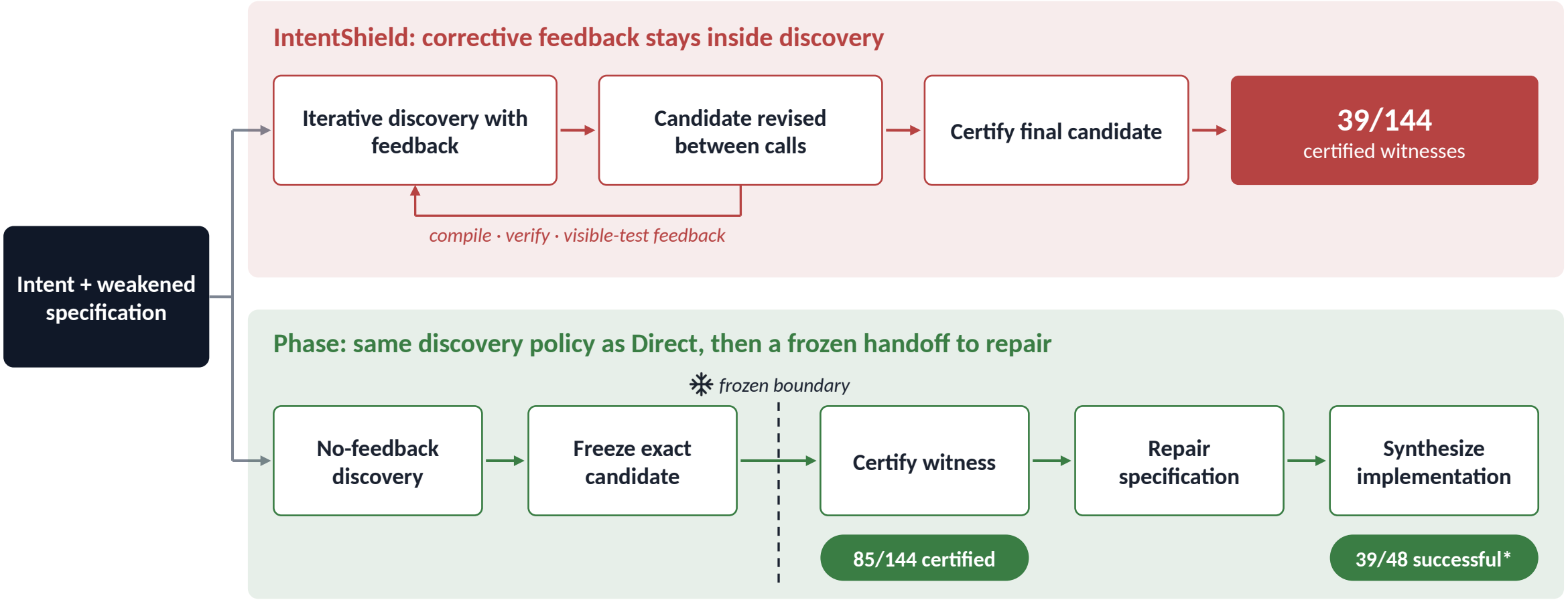
Admitted but wrong: the evidence → Same feedback: compile, verify, tests → Revised toward the intended behavior

Feedback erases the evidence

Same signal, opposite objective. Design rule: freeze and certify evidence of incompleteness before introducing correctness-oriented repair.

METHOD

Phase separation: freeze, then repair



The boundary is the method, not a new search algorithm: Phase reissues Direct's byte-identical prompts 48 h later.

*Separate 16-task downstream study

EXPERIMENTAL DESIGN

A matched-budget study on 12 Verus tasks

12

Verus tasks

VeriContest functions:
5 residual · 7 total-omission

4

model families

Claude Sonnet 5, DeepSeek V3.2, Devstral 2, Qwen3-Coder-480B

3 × 3

episodes × calls

Matched three-call budget for every policy

144

outcomes per policy

1,728 model calls in the four-policy campaign

Fixed before the first model query: tasks, omissions, and mutation examples. The task is the independent unit (task-cluster bootstrap intervals, exact sign tests).

Discovery policies

Primary contrast: Direct vs IntentShield

- **Direct**
  Three independent calls; no candidate, verifier result, or test result carries over.
- **IntentShield**
  Each call sees the previous candidate plus compiler, verifier, and visible-test feedback.
- **FMI (format-matched iterative)**
  Same template and candidate history, but every feedback field reads "[not evaluated]".
- **Model-Judgment**
  Secondary: the model self-assesses its previous candidate before revising it.
- **Phase**
  Direct's protocol, re-executed; the chosen candidate is frozen and certified before repair.

Feedback more than halves witness yield

59.0%

Direct · 85 of 144 certified

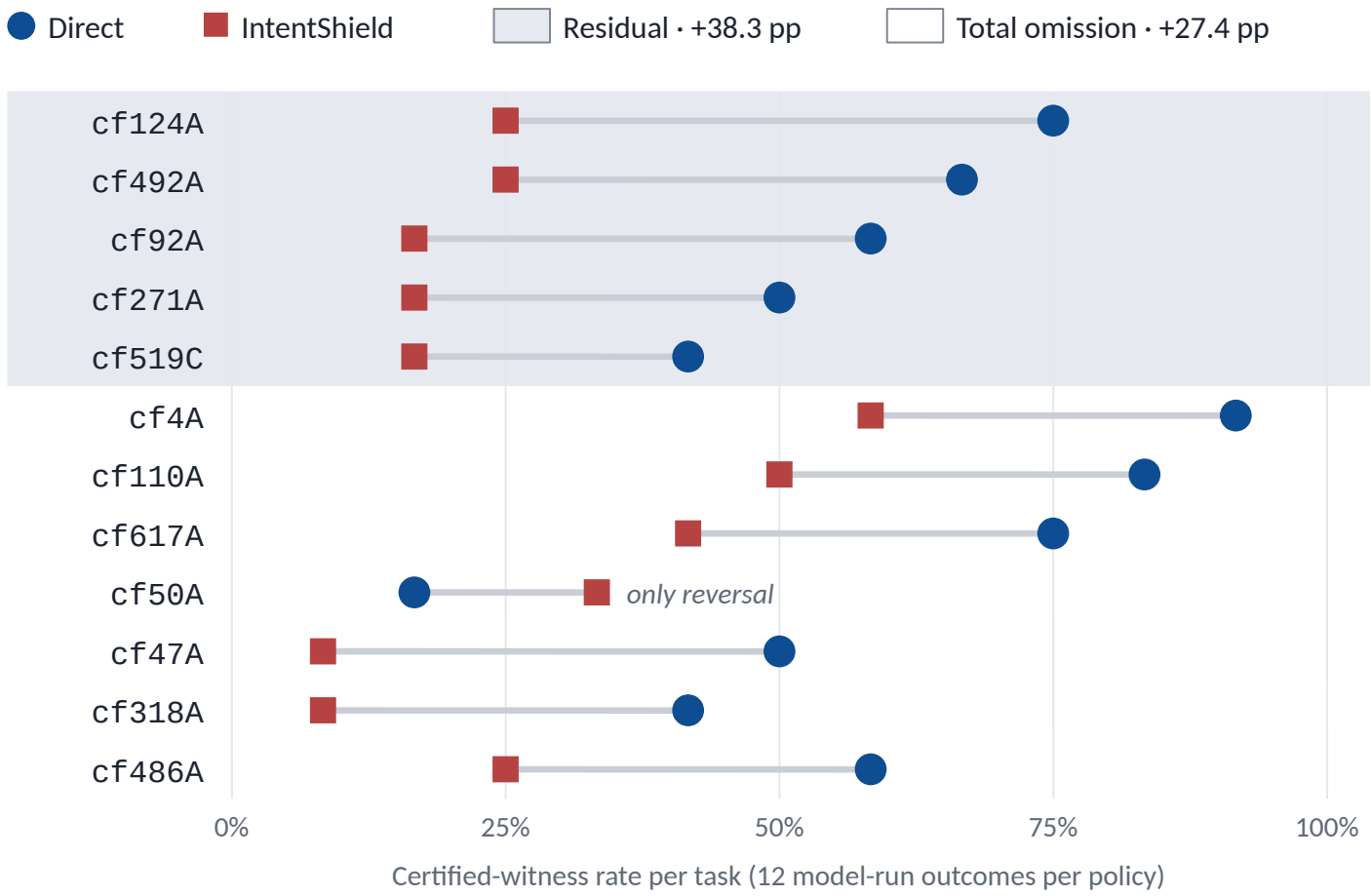
27.1%

IntentShield · 39 of 144 certified

+31.9 pp
mean task-level difference

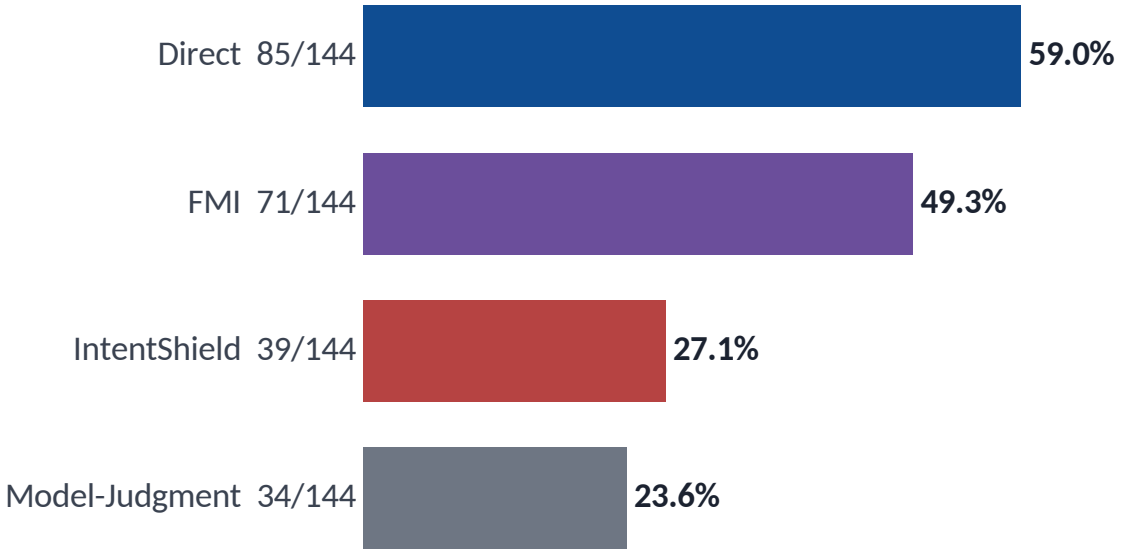
[21.5, 38.9]
95% task-bootstrap interval

11 of 12 tasks
favor Direct (exact sign test p = 0.006)



Feedback content drives most of the gap

CERTIFIED-WITNESS YIELD · 144 OUTCOMES PER POLICY



Pass@1 ≈ the full feedback loop. One Direct draw certifies 26.4% (38/144), close to IntentShield's three-call 27.1%. Three independent draws at that rate give $1 - (1 - 0.264)^3 \approx 60.1\%$, near the observed 59.0%.

DECOMPOSING THE 31.9-PP GAP



Restart diversity
Direct − FMI

Feedback content
FMI − IntentShield

What FMI changes

It keeps IntentShield's template and previous-candidate history but replaces the entire feedback block (prompt length within 3%):

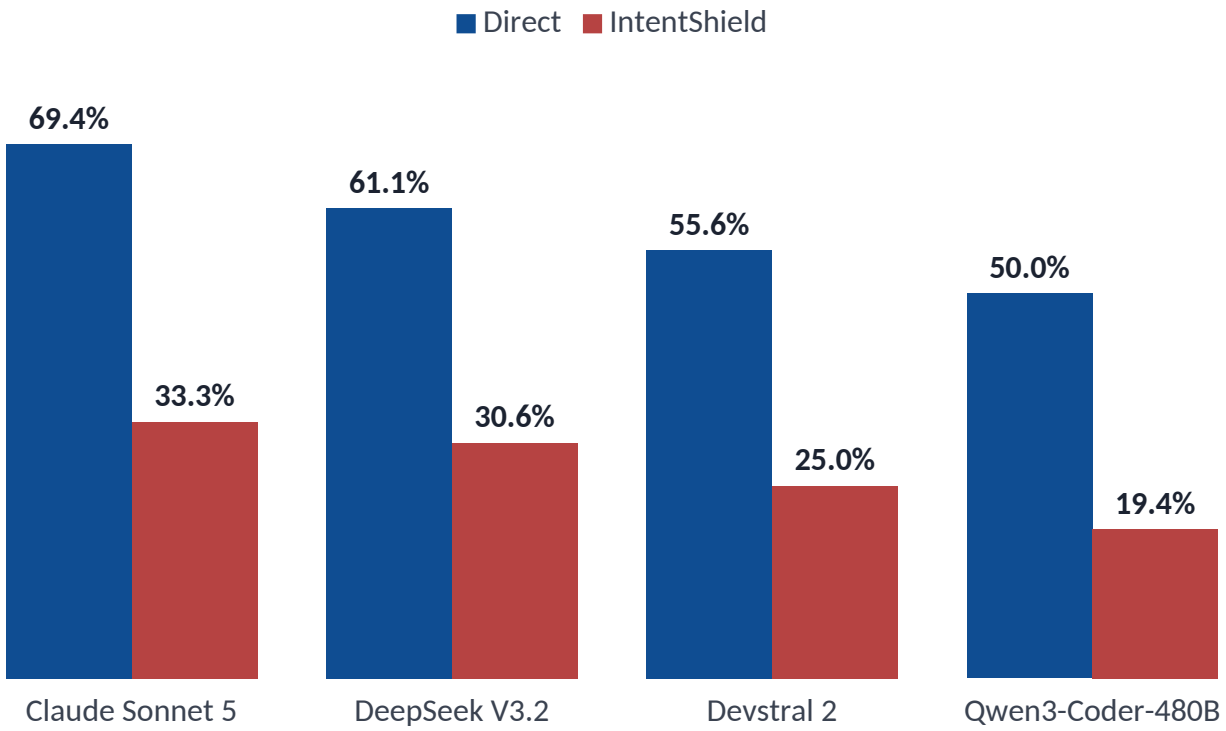
```
Feedback from Previous Round:
compilation:  [not evaluated]
verification: [not evaluated]
tests:        [not evaluated]
```

Observed-rate decomposition, not a formal mediation analysis.

RESULT 3 · ROBUSTNESS

Robust across model families and syntax repair

CERTIFIED-WITNESS YIELD BY MODEL FAMILY · 36 OUTCOMES PER POLICY



Gaps of +30.6 to +36.1 pp; within every family, Direct leads on 9–10 of the 12 tasks.

Syntax repair keeps the gap

	Direct	IntentShield	Gap (pp)
Raw	59.0%	27.1%	**+31.9**
Parser-repaired	63.2%	36.1%	**+27.1**
Parse-clean only*	58.9%	33.7%	**+25.2**

Parse failures fall from 34.0% to 11.8% (IntentShield) and 10.4% to 4.2% (Direct). *Post-hoc filter with policy-specific denominators.

Evidence holds on held-out cases

92.9%

of Direct witnesses (79/85) also disagree on a withheld case

IntentShield 92.3% · Phase 91.8%

A separate Phase run reproduces discovery

	Phase certified	Phase not certified	Total
Direct certified	**74** certified in both runs	**11** Direct only	85
Direct not certified	**11** Phase only	**48** neither run	59
Total	85	59	

144 paired task × model × repeat cells; Direct rows, Phase columns

85 = 85

certified outcomes per run, out of 144; task totals match on 10 of 12 tasks

84.7%

cell-level agreement, Cohen's κ = 0.68 (83.1% expected from independent draws)

43 of 74

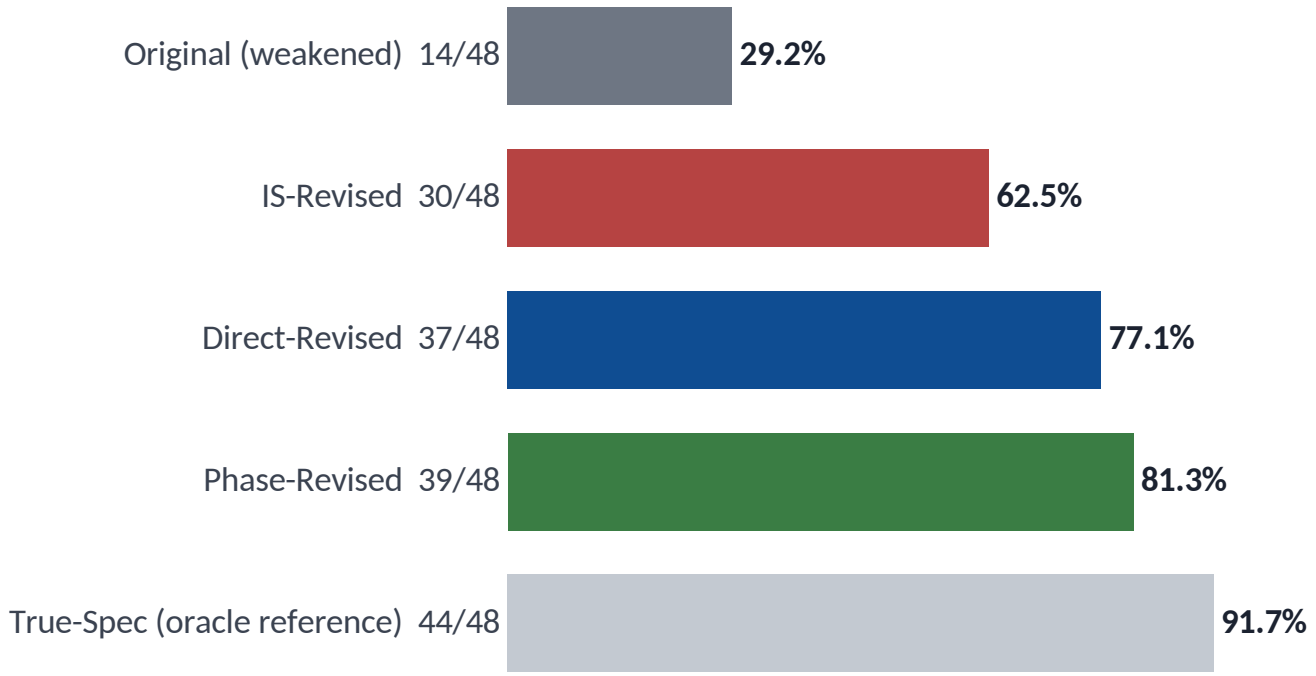
co-certified cells hold distinct expressions that each pass all four gates; 31 are string-identical

Fresh request IDs and timestamps, 48 h apart; no responses reused.

RESULT 5 · DOWNSTREAM REPAIR

Preserved witnesses yield better contracts

SUCCESSFUL IMPLEMENTATIONS · 16 TASKS × 3 ATTEMPTS = 48 PER CONDITION



Implementations: Claude Sonnet 5, three independent attempts per task and condition. For three low-yield tasks, the IS-Revised repair uses the best available, possibly uncertified, candidate.

+18.8 pp

Phase-Revised over IS-Revised (95% CI [10.4, 27.1]): wins 9 tasks, ties 7, loses none; sign test p = 0.004

88.6%

Phase-Revised reaches this share of the True-Spec success count (39 of 44)

2 attempts

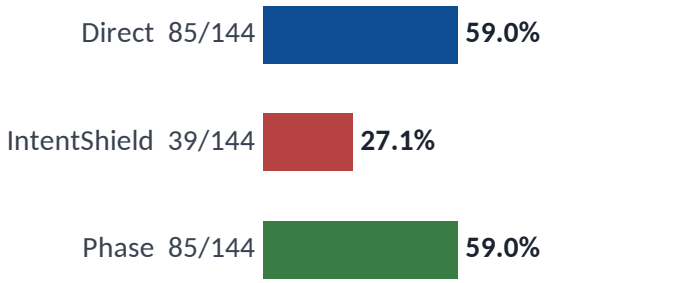
Phase-Revised 39 vs Direct-Revised 37. Witnesses found without feedback support repair either way.

SCOPE & LIMITATIONS

Bounded to Verus: no transfer to Dafny or Lean

Verus

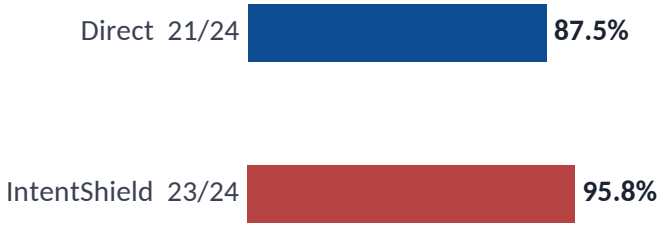
Primary matrix · n = 144 per method



Supported · D − IS = +31.9 pp

Dafny

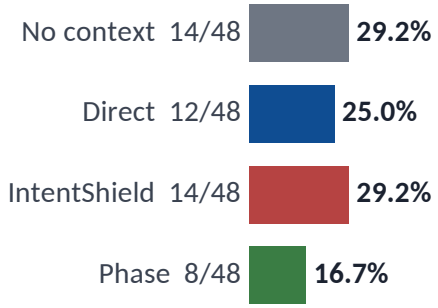
Qwen stratum with headroom · n = 24 per method



Descriptive · D − IS = −8.3 pp

Lean

Verina, 24 tasks · n = 48 per method



No support · Phase is lowest

Panels use different task matrices: compare direction and scope, not absolute difficulty.

Narrow benchmark

12 competitive-programming functions, expression-level Rust only. An 8-task extension agrees (51.0% vs 21.9%) but is preliminary.

Hosted models

Provider-default decoding and non-seeded repeat labels limit exact replay; all 1,728 primary prompts and responses are archived.

Why Verus may differ

Rust typing, ownership, and arithmetic obligations give feedback a rich correction surface. A hypothesis, not a measured mediator.

TAKEAWAY

Freeze the witness, certify the gap, then repair

1 Freeze

Discover without correctness feedback, then freeze the exact candidate.

59.0% vs 27.1%

certified-witness yield, Direct vs IntentShield

2 Certify

Require all four gates before any repair or implementation work begins.

92.9%

of Direct witnesses also disagree on at least one held-out case

3 Repair

Strengthen the contract from the preserved witness, then synthesize.

81.3% vs 62.5%

implementation success, Phase- vs IS-Revised

Scope: the studied Verus/Rust setting; the effect did not reproduce in Dafny or Lean.

In practice: triage witness reports like fuzzer findings, with the specification author.

Thank you — questions welcome

Paper and reviews on OpenReview

openreview.net/forum?id=N6yfUR70Uu