ICPR 2026 · ORAL PRESENTATION

SigRef

Verification-Driven Reflection for Faithful Paper-to-Code Development

[1,2]Mingyang Zhou · [1]Quanming Yao · [2]Lun Du · [2]Lanning Wei · [2]Da Zheng

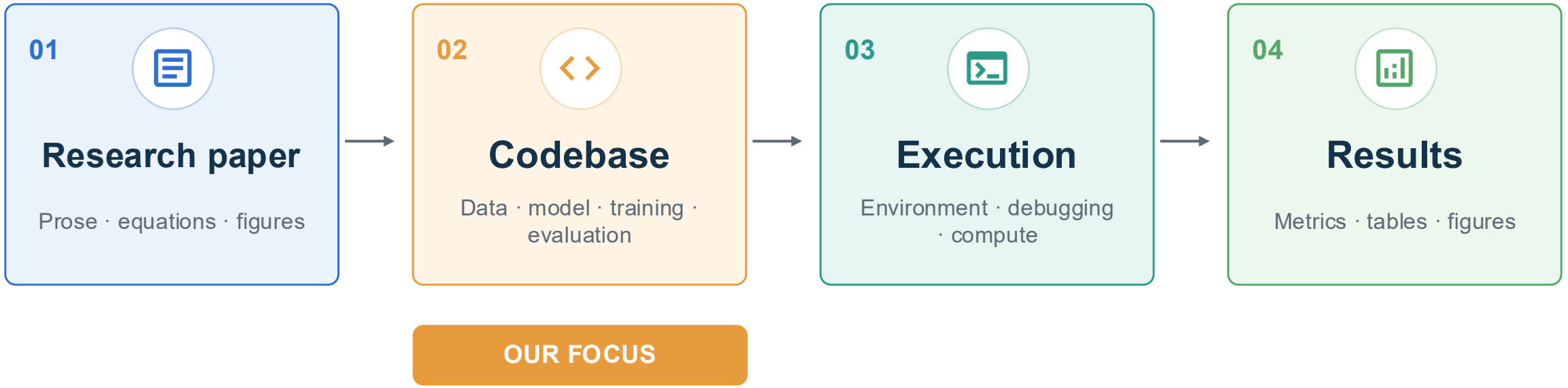
Mingyang Zhou

[1]Department of Electronic Engineerging, Tsinghua University, Beijing, China
[2]Ant Research, Beijing, China

Work done during an internship at Ant Research.

MOTIVATION

Paper-to-code is a long-horizon engineering task



If the implementation is not faithful, execution and result matching cannot rescue it.

PROBLEM

A small value can encode a non-trivial scope

PAPER SPECIFICATION

2D experiment families $\sigma_{min} = 0.01$

SIR · Two Moons · related 2D settings

All other experiments $\sigma_{min} = 0.05$

A different experimental regime

The values are tied to experiment scope.

GENERATED CODE

One shared setting

```
self.sigma_min = 0.01
```

Applied to every experiment

✓ valid code ✗ lost value-to-scope mapping

In long-horizon paper reproduction, it is difficult for an agent to get every fine-grained requirement right.

KEY INSIGHT

Turn the paper into a checklist

PAPER

15 pages

prose · equations · tables · captions

Requirements are scattered and entangled.

extract · ground · atomize

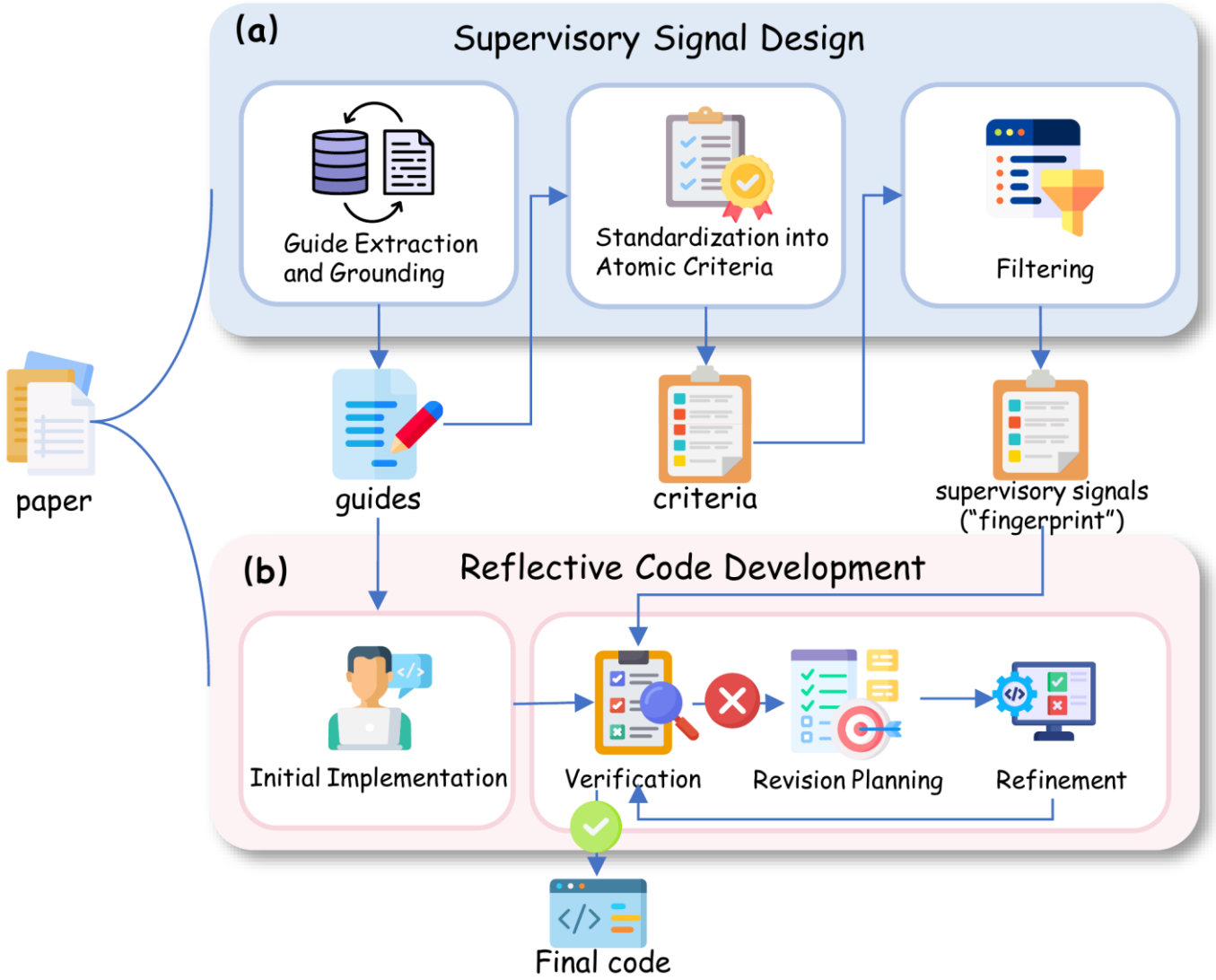
PAPER FINGERPRINT

Hundreds of requirements

- □ one paper-grounded fact
- □ one explicit scope
- □ one atomic criterion

Make every implementation requirement explicit and verifiable.

Framework of SigRef



METHOD

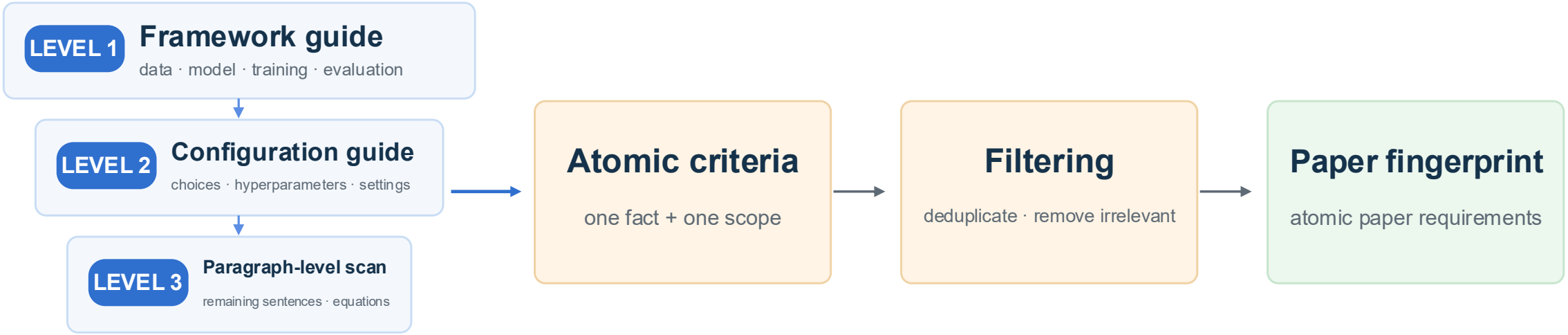
Building the paper fingerprint

Comprehensive & Accurate

Cover relevant implementation details and ground every unit in source text.

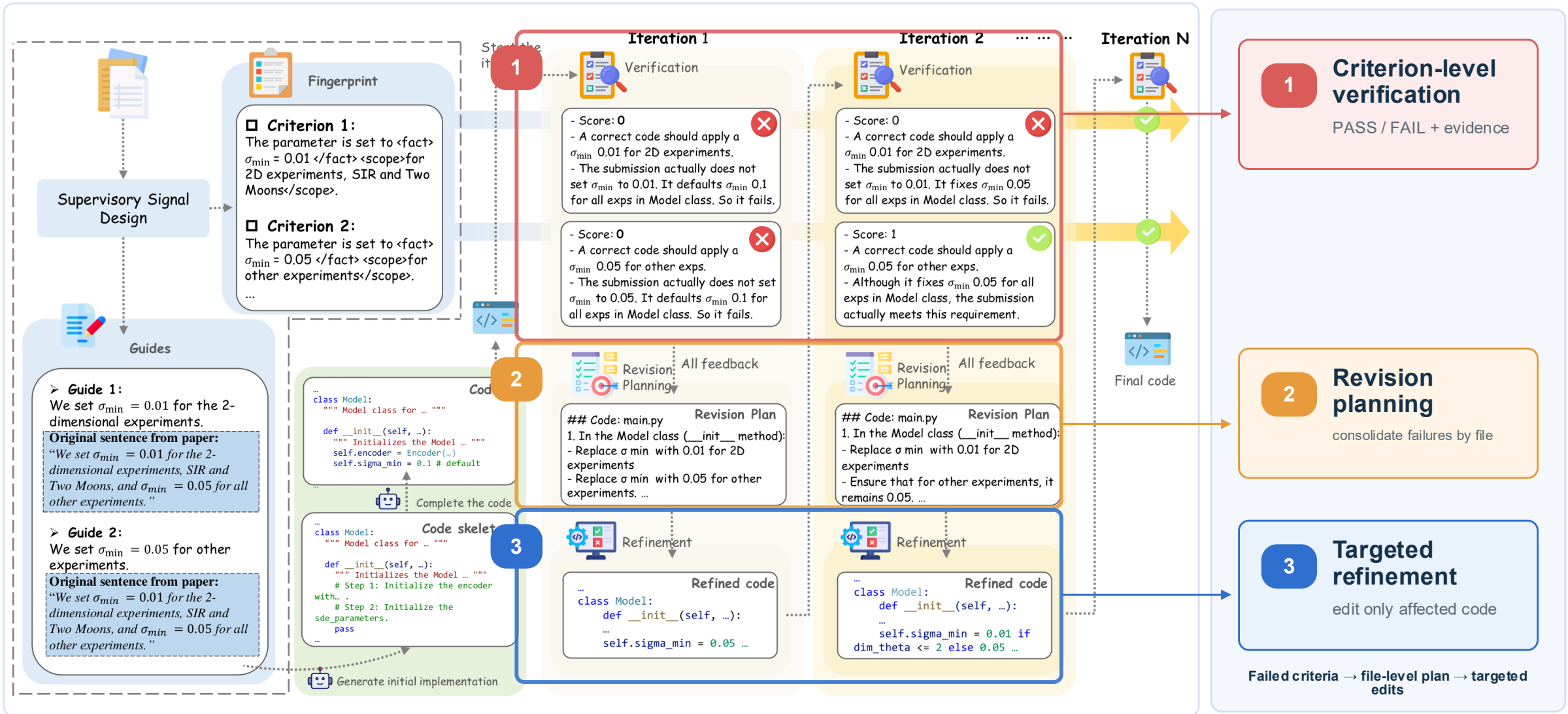
Atomic

One criterion = one paper-grounded fact + one explicit scope.



Average over 20 papers: 237 guide units → 895 atomic facts → 164 final criteria

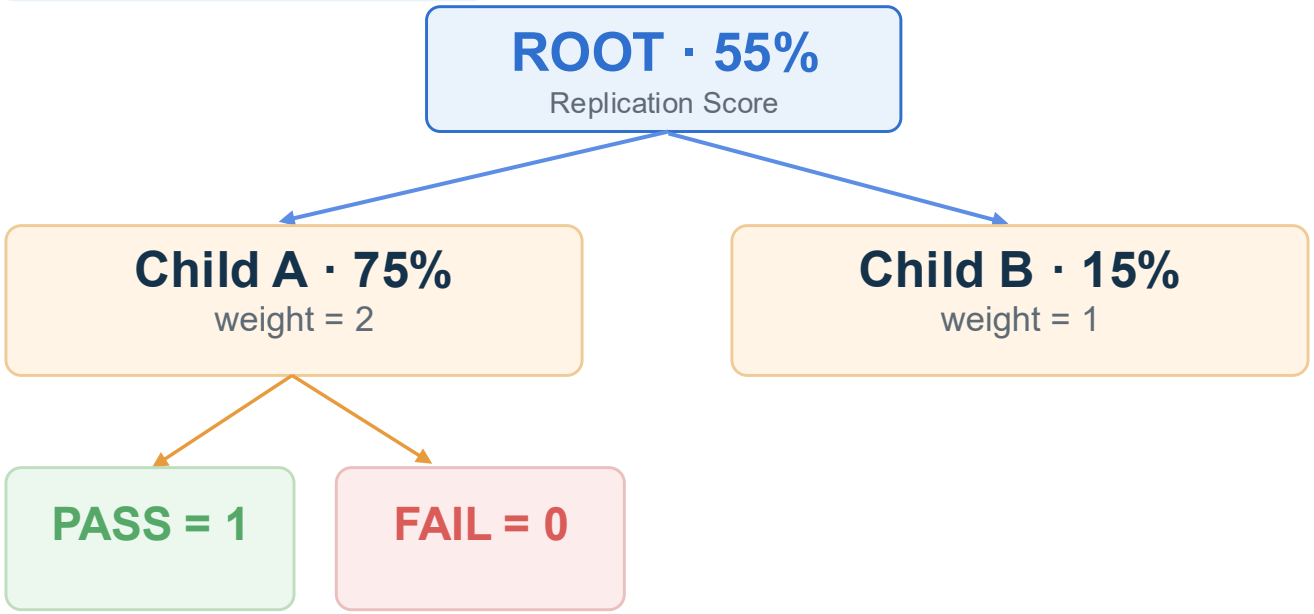
Verification-driven reflection



EVALUATION

How PaperBench computes the score

RECURSIVE WEIGHTED TREE



Every leaf is binary; every parent is the weighted average of child scores.

Example root: (2 × 75% + 1 × 15%) / 3 = 55%

$$S(v) = \Sigma[w(c) \times S(c)] / \Sigma w(c)$$

METRIC MAPPING

PaperBench Replication Score

↓

= SigRef PRroot

PRleaf = Npass / Ntotal

SigRef's unweighted pass rate over Code-Dev leaf requirements

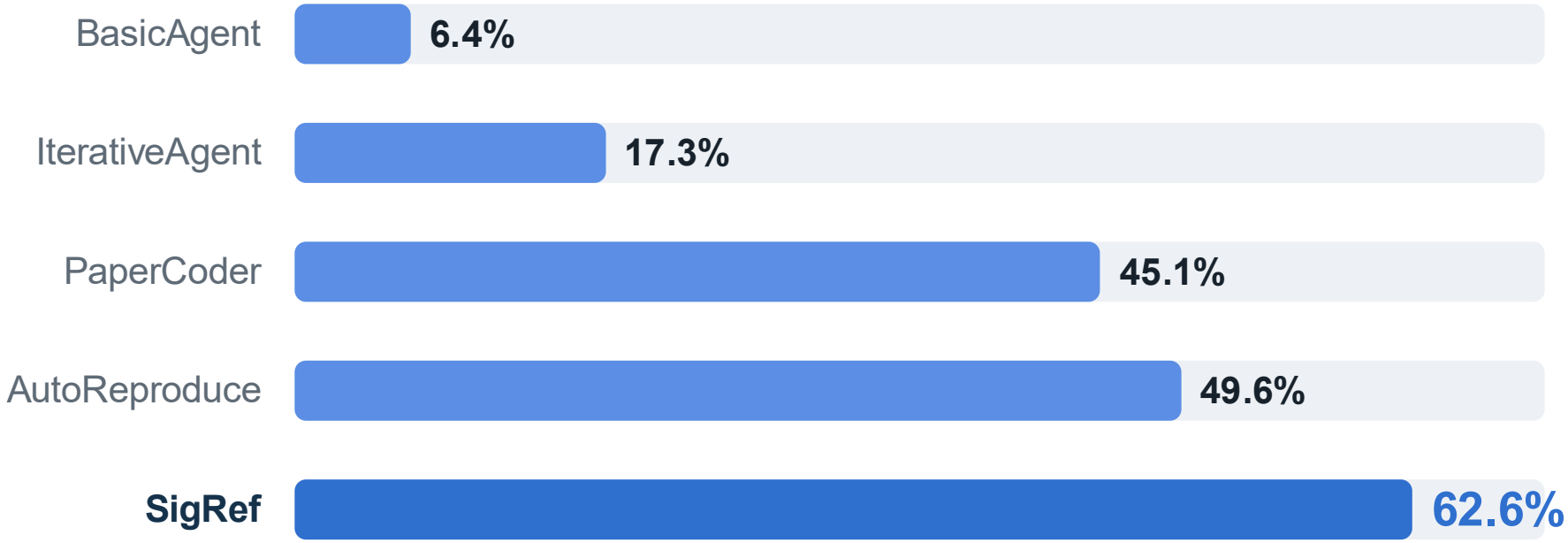
Graded **Code Development** Skipped **Execution** **Result Match**

Code-Dev isolates implementation fidelity.

Source: PaperBench: Evaluating AI's Ability to Replicate AI Research. ICML 2025.

RESULTS

SigRef reaches 62.6%



+13.0 pp

over the strongest baseline

+30

requirements per paper vs. PaperCoder

Experimental setup

20 papers · up to 4 refinement rounds · early stop when all criteria pass

Signal design & verifier	Code generation, planning & refinement	PaperBench judge	Retrieval encoder
DeepSeek-V3	o3-mini-high	o3-mini-high	all-MiniLM-L6-v2

RESULTS

What does SigRef improve over PaperCoder?

Representative rubric leaves changed from FAIL to PASS

PAPER · PRleaf GAIN	E.g. PaperCoder · FAIL		E.g. SigRef · PASS	Capability gain
PINN +65.1 pp	Phase 2 uses damped Newton / PCG	→	Switches Adam → L-BFGS after 1k steps	Training-flow fidelity
Mechanistic Understanding +52.8 pp	Filters with a fixed threshold	→	Computes cosine similarity, ranks, selects top 128	Selection-logic fidelity
Bridging Data Gaps +41.8 pp	Adaptor misses down-pooling, attention, and 3×3 conv	→	Restores the full adaptor module	Module-design fidelity

These gains are paper-level PRleaf improvements. Each row shows one representative FAIL-to-PASS criterion, rather than the sole cause of the total gain.

ABLATION

Fingerprint design matters

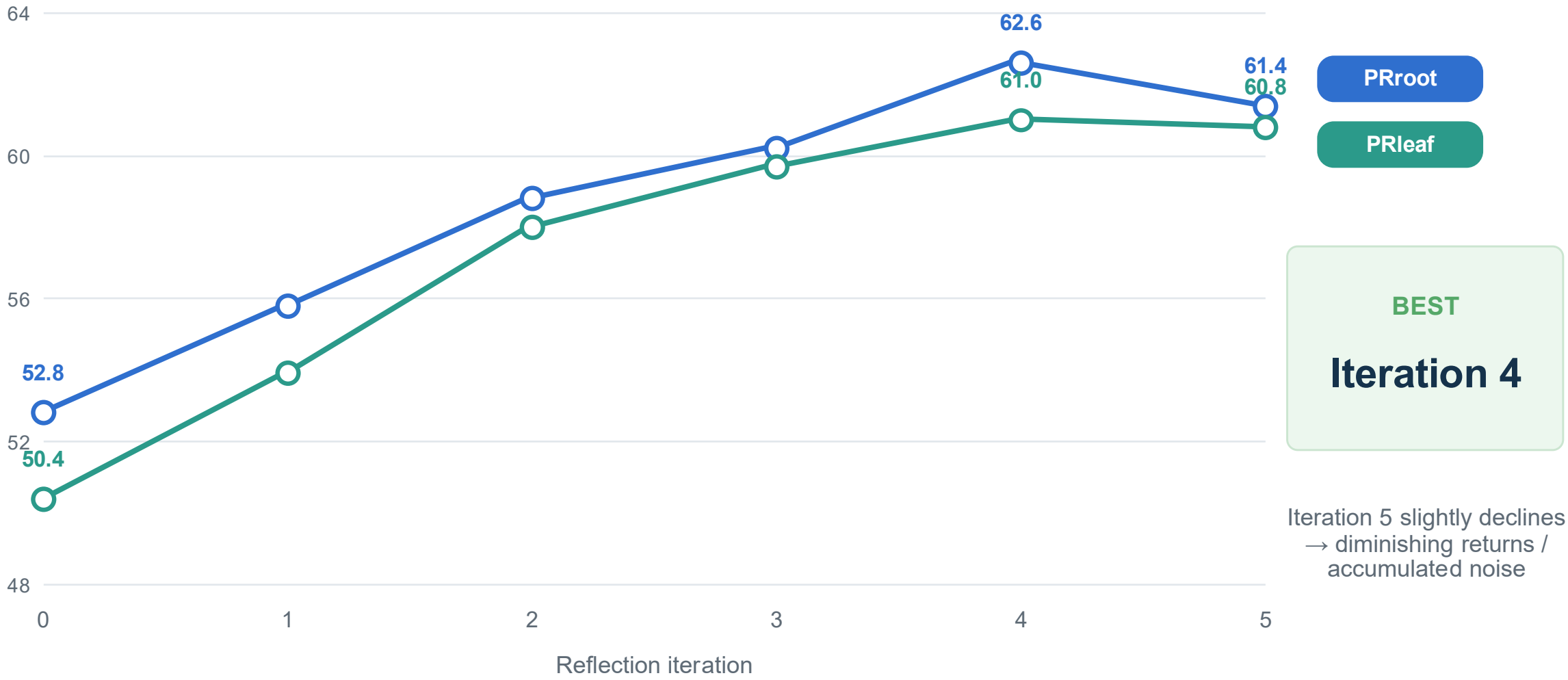
−6.7 pp without comprehensiveness

−4.4 pp without atomicity

Coverage finds the mismatches; atomicity makes them actionable.

ABLATION

Reflection progressively improves fidelity



BEST

Iteration 4

Iteration 5 slightly declines → diminishing returns / accumulated noise

CONCLUSION

Takeaways

01

Fine-grained alignment is a key bottleneck.

Agents can produce plausible code while missing conditional, mathematical, or procedural requirements.

02

Paper fingerprints make supervision explicit.

Comprehensive, grounded, atomic criteria turn a long paper into verifiable targets.

03

Verification makes reflection actionable.

Criterion-level feedback drives planning and targeted refinement, reaching 62.6% PRroot.

Scope: faithful code development—not complete execution and result reproduction.

Thank you for your attention

SigRef: Verification-Driven Reflection for Faithful Paper-to-Code Development

Mingyang Zhou · Tsinghua University · Ant Research

ICPR 2026 Oral Presentation