ICML 2026 WORKSHOP · PROGRAM SYNTHESIS & COMPOSITIONAL GENERALIZATION

Irreducible Supervision Enables Compositional Generalization in Post-Training

Ellen Ma & Nikhil Anand · Controlled program synthesis over `arc-dsl` | Qwen3-4B / 8B / 32B · Llama-3.1-8B

TLDR Compositional ability is **installed during SFT** and depends on the **irreducibility** of the data — not its quantity, and not RL. A handful of structurally & observationally irreducible exemplars is enough. An order of magnitude more shortcut-admitting data does nothing. **RL maintains but cannot deepen what SFT installs — no new compositional depth.**

THE QUESTION

Does RL *create* new capabilities, or only *sharpen* what is already there?

Unresolved for **compositional generalization** — combining learned primitives into novel multi-step functions. We isolate an **upstream** variable: the **structure of the SFT data**, before RL begins.

- **Domain:** `arc-dsl` — 119 typed primitives over integer grids; correctness verified by execution.
- **Metric — pass@64:** ≥1 of 64 sampled programs passes *all* held-out I/O pairs (execution-verified, no LLM judge). The base model scores **0%** here before fine-tuning, so *any* gain is attributable to post-training.

METHOD — TWO LEVELS OF IRREDUCIBILITY

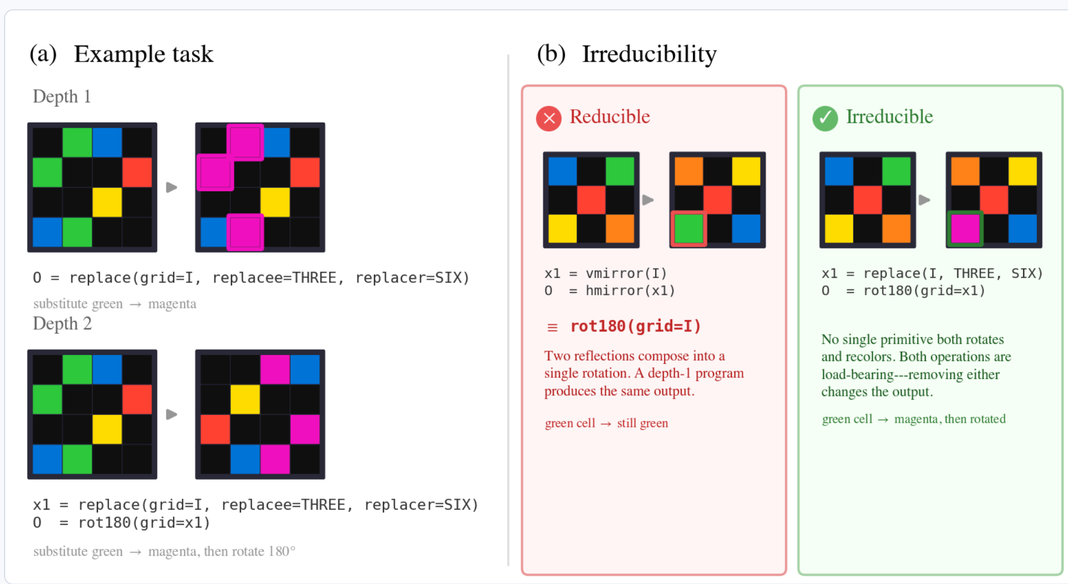


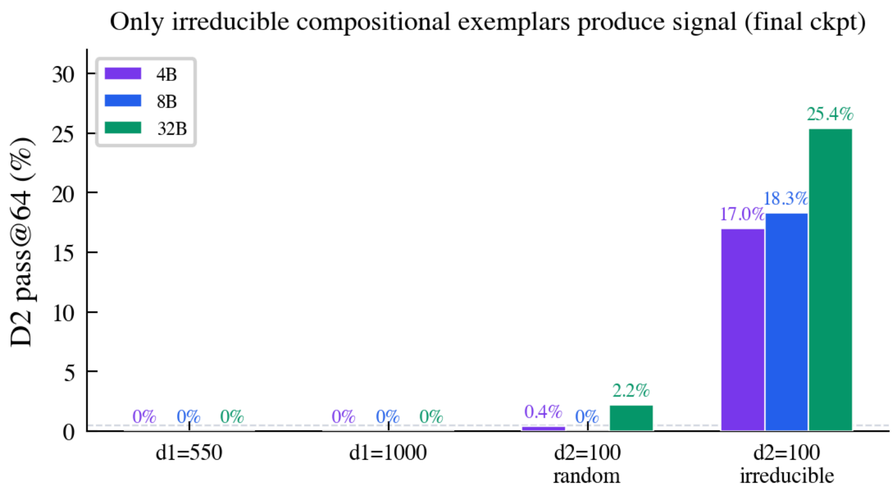
Fig 1. A depth-2 program is **reducible** if a single primitive computes the same function (`vmirror∘hmirror = rot180`), and **irreducible** otherwise (no primitive both recolors *and* rotates).

Structural
No shallower DSL *program* computes the same function on the validation pool.

Observational
The visible *I/O pairs* admit no shallower-program solution (adversarial pair selection).

Enforced by a multi-stage filter: algebraic pair-blocking, group-theoretic chain detection, and **matching each program's input→output behavior against tens of thousands of shallower programs** to discard any with a shallow equivalent — plus adversarial I/O selection. Audited residual contamination **<1.5%**.

RESULT 1 — BOTH LEVELS ARE NECESSARY

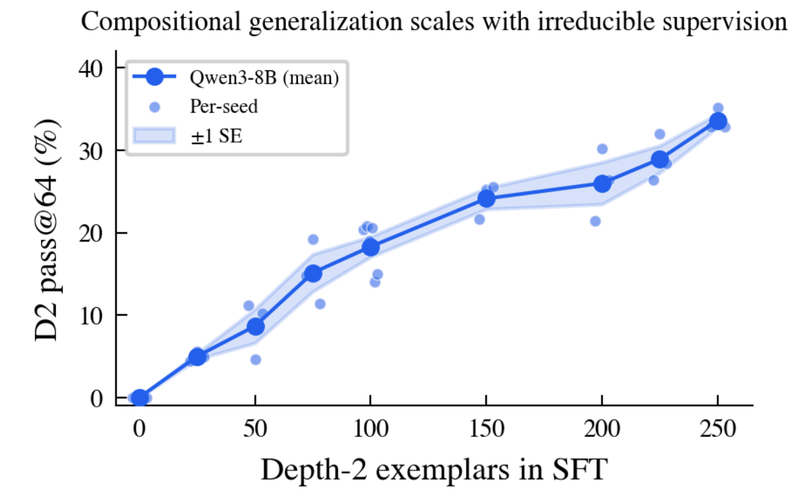


Quantity cannot substitute for structure: random depth-2 programs and pure depth-1 scaling (550→1000 examples) yield **0–2.2%**; only irreducible exemplars produce signal (17–25% at 4B/8B/32B, scaling with size).

D2 pass@64 (8B)	Adversarial I/O	Naive I/O
Irreducible programs	**18.3%**	6.9%
Reducible programs	**0.6%**	**0.1%**

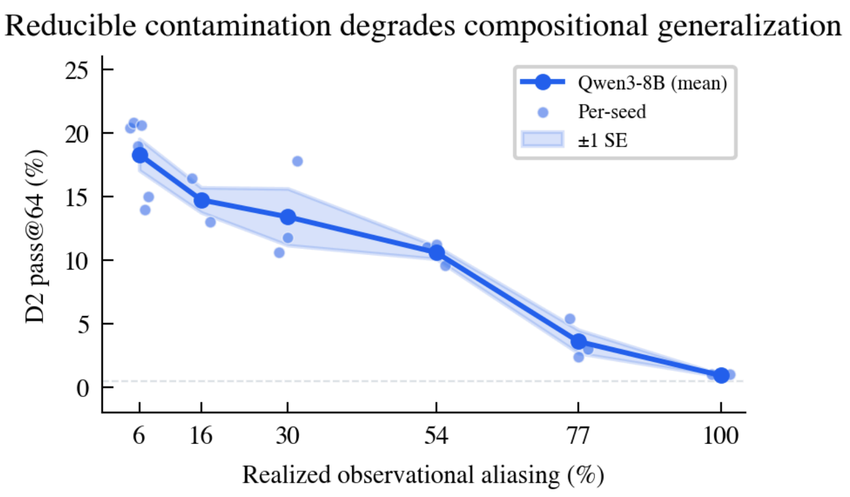
Rows = structural irreducibility · **columns = observational** irreducibility. Only the **both-irreducible** cell produces compositional signal — neither level alone suffices.

RESULT 2 — SCALES SMOOTHLY WITH SUPERVISION



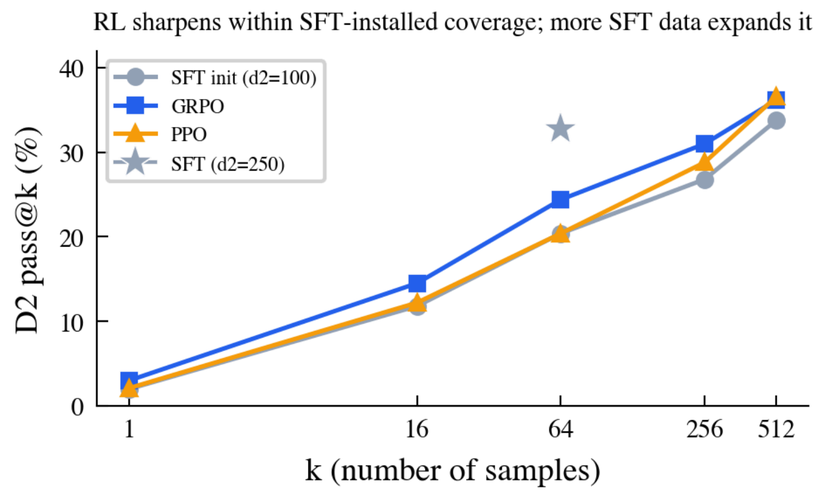
Given irreducible data, D2 pass@64 rises smoothly and monotonically **0% → 33.6%** from 0 to 250 exemplars. Success spreads across **139 / 427** distinct primitive pairs — genuine generalization, not pair memorization.

RESULT 3 — CONTAMINATION DEGRADES IT, WITH DOSE



Holding total exemplars fixed at 100 and swapping irreducible for reducible, D2 pass@64 falls **monotonically with dose** — realized aliasing of 16/30/54/77% → **14.7/13.4/10.6/3.6%** (low seed variance), hitting the **~1% noise floor only at ≥77%**. Even the control's nominally-clean arm carries 6% residual aliasing in its I/O pairs. Irreducibility verification is necessary, not optional.

RESULT 4 — RL MAINTAINS BUT DOES NOT EXTEND



0%
GRPO from d1-only SFT → no composition from scratch

+2.1pp
GRPO gain at k=64 (unbiased); **null at k≥256**

2.5×
more SFT data matches RL with **8×** rollouts

RL improves pass@k at **intermediate k** (peak ~k=16–32) but the gain **attenuates and is null for k≥256** — SFT already reaches the same solutions. GRPO's gains are **breadth** (more tasks solved, more distinct correct programs per task), **not new depth**: D3/D4 stay flat under both GRPO and PPO.

RESULT 5 — COMPOSITIONAL TRANSFER IS DEPTH-LOCAL

Depth-*d* generalization requires **explicit depth-*d* supervision**. Adding depth-3 exemplars lifts D3 (and even D2, via shared sub-trees), but transfer is **bidirectional yet asymmetric** — strong downward, weak upward. Capability propagates **one depth**, then plateaus — the ceiling RL cannot push past.

D3 EXEMPLARS IN SFT	D2 PASS@64	D3 PASS@64	D4 PASS@64
0	20.4%	1.0%	1.4%
250	24.4%	**9.2%**	3.0%
500	28.2%	**10.6%**	2.4%

WHAT THIS MEANS

- **Data curation > compute.** Targeted irreducible exemplars are necessary; RL compute and unstructured data don't substitute.
- **Verify irreducibility.** Shortcut-admitting "compositional" data actively destroys the signal.
- **RL sharpens, doesn't deepen.** The compositional ceiling is set by SFT-data structure, before RL begins.
- **Practical reading:** for multi-step tool-use / multi-hop tasks, ensure training examples genuinely require the target composition depth.

Key references: Yuan et al. 2025 (*From f(x) and g(x) to f(g(x))*) · Yue et al. 2025 (*Does RL incentivize reasoning?*) · Foster et al. 2025 (*Is a good foundation necessary?*) · Zhao et al. 2025 (*Echo Chamber*) · Hodel 2024 (*arc-dsl*) · Shao et al. 2024 (GRPO) · Min et al. 2019; Trivedi et al. 2021 (multi-hop reducibility).
Setup: dealiased eval, 500 tasks/depth, multi-seed, paired 95% bootstrap CIs · full-weight SFT + GRPO/PPO in VeRL on 8×H100.

arc-dsl · 119 primitives
Ellen Ma · Nikhil Anand
Paper & code: release pending review