

AlphaEvolving sampling

(making the same pun for 4 months)

Angle of attack

Substrate dependence is a security vulnerability

- architectures might change
- adversarial AI might obfuscate itself

Possible approach: a meta-technique

mapping: substrate $\rightarrow$ specific interp technique

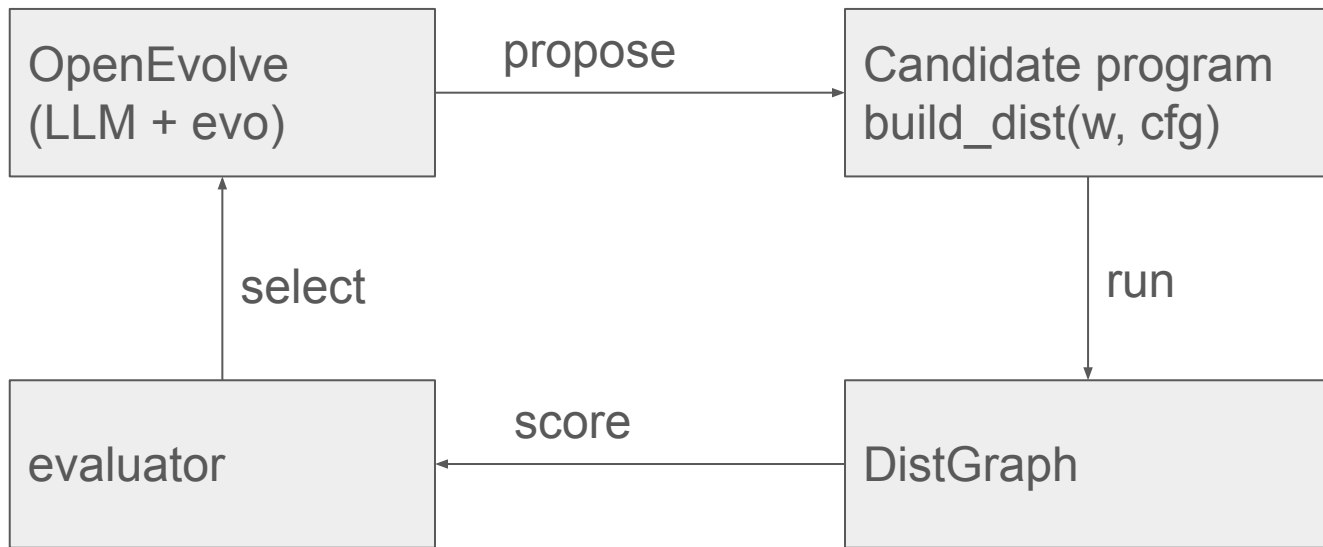
Optimization target? $\rightarrow$ ARC: competing with sampling

Competing with sampling

$P(\text{output} \in \text{region} \mid \text{input} \sim D) = ?$

- sampling: works but expensive, misses rare catastrophic events
- analytic from weights: if you match sampling, you understood something
- sampling is detectable from FLOPS!

The loop



- optimizing code, not weights
- DistGraph is declarative and describes entire distribution
- score = mean log prob on unseen samples

The lessons

Random MLPs, Gaussian input

→ moment propagation

(Generic!)

Sentiment transformer

→ mixture of low-rank Gaussians, FFN-SVD directions

(Okay, and does *this* count as understanding? Why / why not?)

Maybe interpretability doesn't scale that neatly with predictive power?

3 pressures

Pressure 1. Same program, many models.

39 modular-addition transformers (13 primes $\times$ 3 seeds)

Best result: +385 nats (over diagonal-Gaussian baseline)

Strategy: forward pass $\times$ per-class LowRank Gaussian $\times$ Fourier-aware variance

3 pressures

Pressure 2: Many tasks

90 models × 5 task families (mod-add, greater-than, Dyck, DFA, induction)

Task label hidden; can't branch on it

Best result: +24 nats (averaged across tasks, one program)

Strategy: generic forward pass + random inputs + k-means + diag-Gaussian mixture

3 pressures

Pressure 3. Pick tasks where generic methods are unhelpful

Modular addition $\rightarrow$ Fourier features on a torus

Hard to model a torus with a mixture of Gaussians in 64-dim

Okay, so what?

- On the one hand, this is kinda cool!
- On the other hand, bleh.

Some ideas:

- intervention prediction?
- extend to other architectures

What the hell is the connection between interpretability and predictive ability?