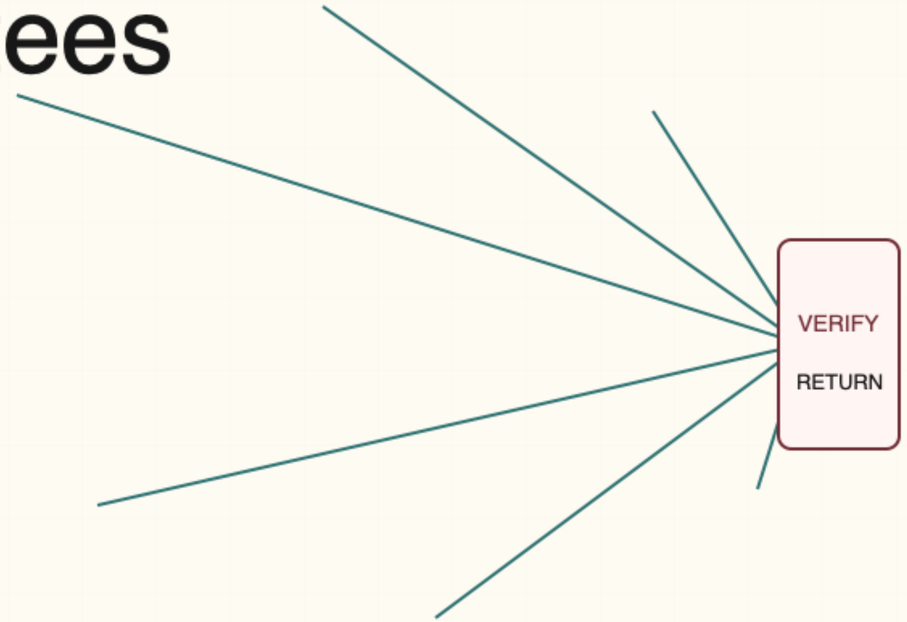
TEST-TIME SCALING

PART 2 01:11-15:34

More Inference Compute Creates Options, Not Guarantees

Stanford CS329A Part 2 - generation, allocation, and verification under a fixed model

FIXED
MODEL
weights locked



Generate

Allocate

Verify

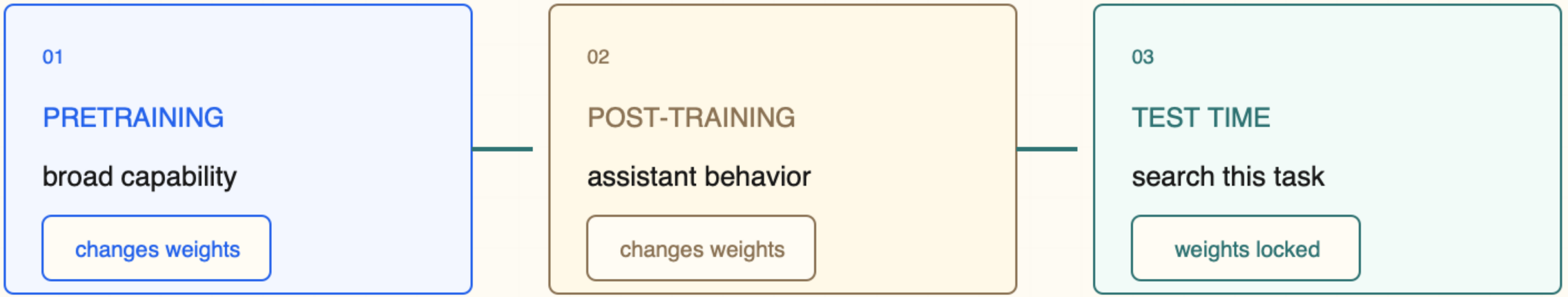
Stop

WHERE COMPUTE LIVES

PART 2 00:05-01:11 + arXiv:2408.03314

Test-Time Scaling Searches Without Changing the Weights

Training changes the model; inference spends a recurring per-request budget.



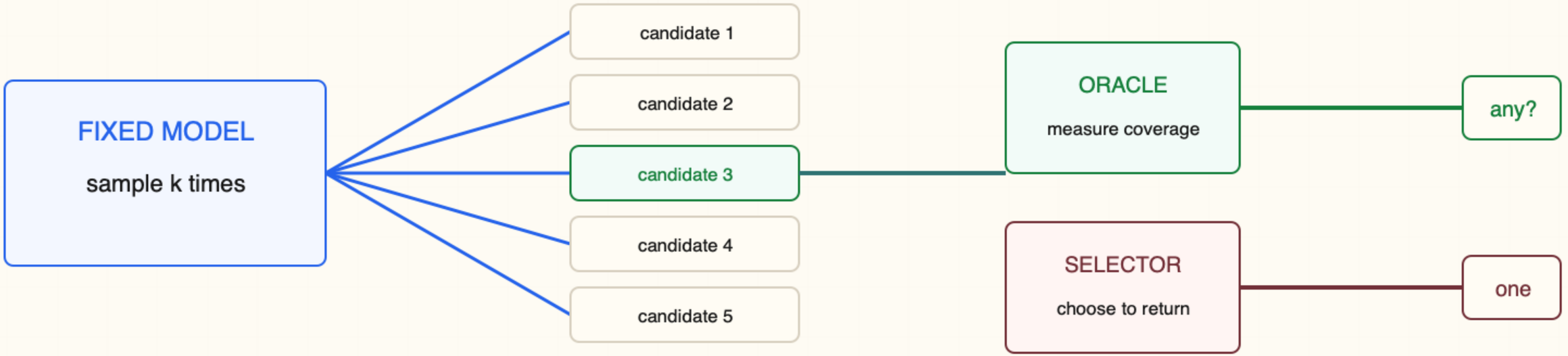
TRAIN ONCE -> AMORTIZE | INFER PER REQUEST -> PAY AGAIN

CANDIDATE PRODUCTION

PART 2 01:29-03:36 + arXiv:2407.21787

Repeated Sampling Buys More Chances

One fixed model produces many candidates; a verifier determines whether the gain is usable.



CAVEAT

Oracle coverage can rise even when a deployable selector cannot find the correct candidate.

METRIC BOUNDARY

PART 2 01:29-03:36 / 12:20-26:55

A Correct Candidate Can Exist and Still Never Reach the User

Coverage belongs to the candidate set; reliability belongs to the whole pipeline.

COVERAGE / PASS@K

Does any candidate contain the correct answer?

candidate-set metric

CAVEAT

A benchmark-to-product claim must name both the candidate metric and the selection mechanism.

PER-PROBLEM SCALING

PART 2 05:27-07:33 + arXiv:2502.17578

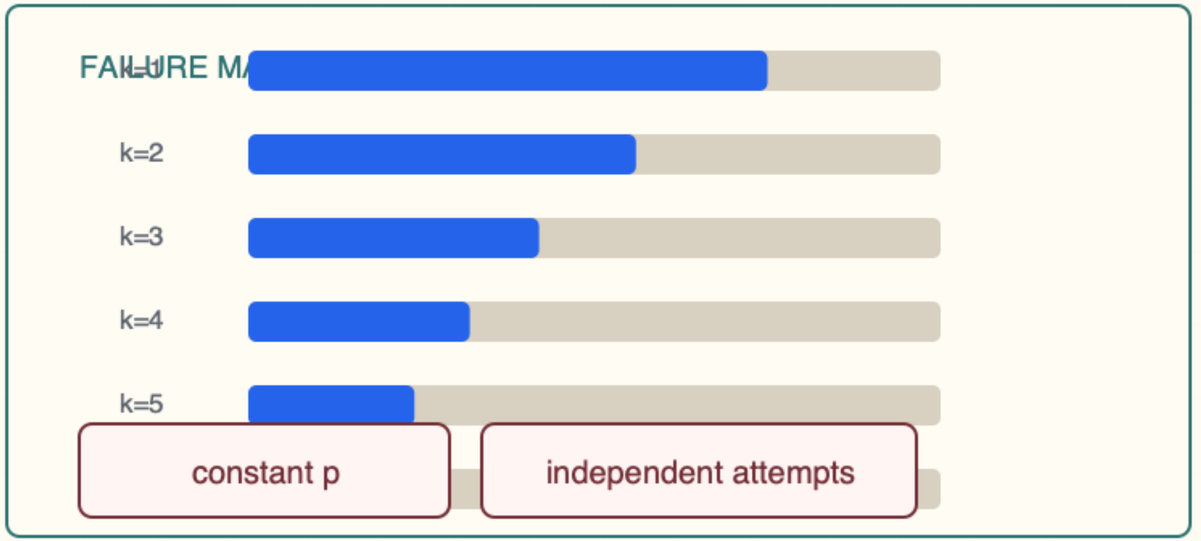
One Problem Has an Exponential Failure Curve

The familiar pass@k formula is an idealized model with visible assumptions.

pass@k

$$1 - (1 - p)^k$$

failure after k = (1 - p)^k



CAVEAT

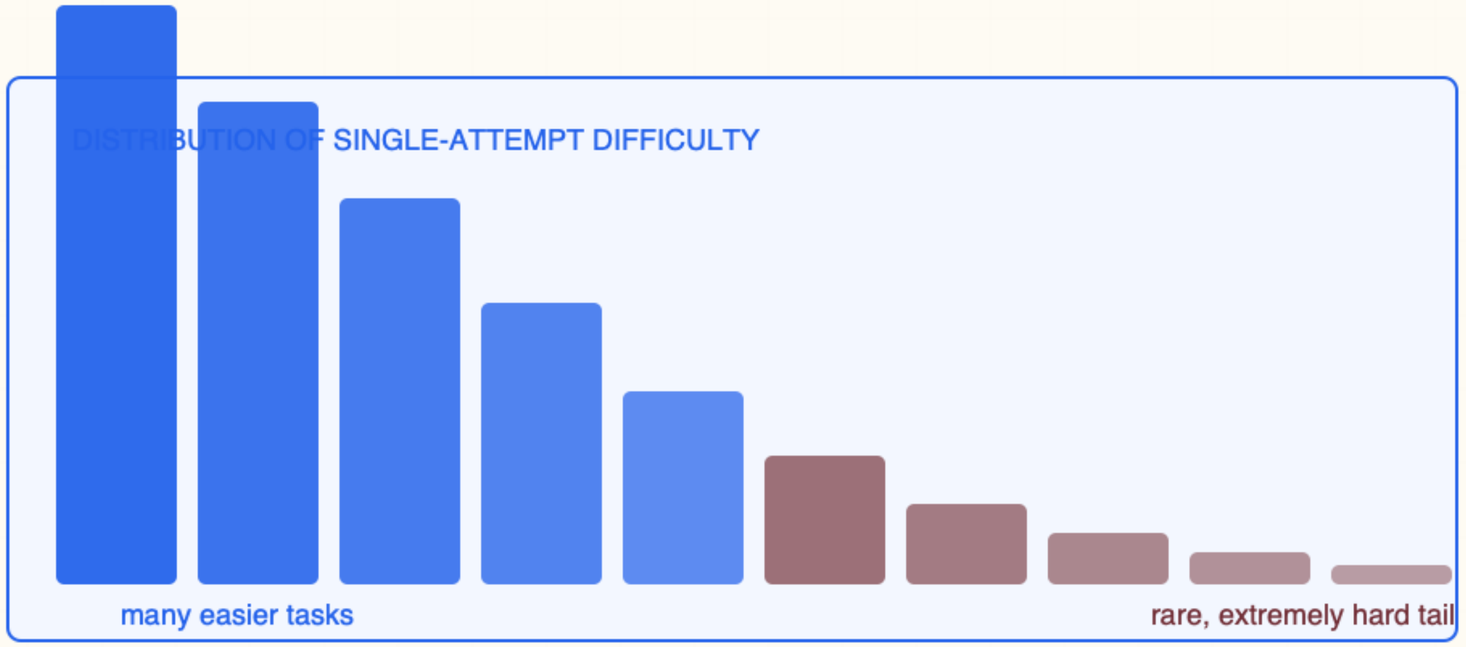
Samples from one model can be correlated, so effective diversity may be much lower than k.

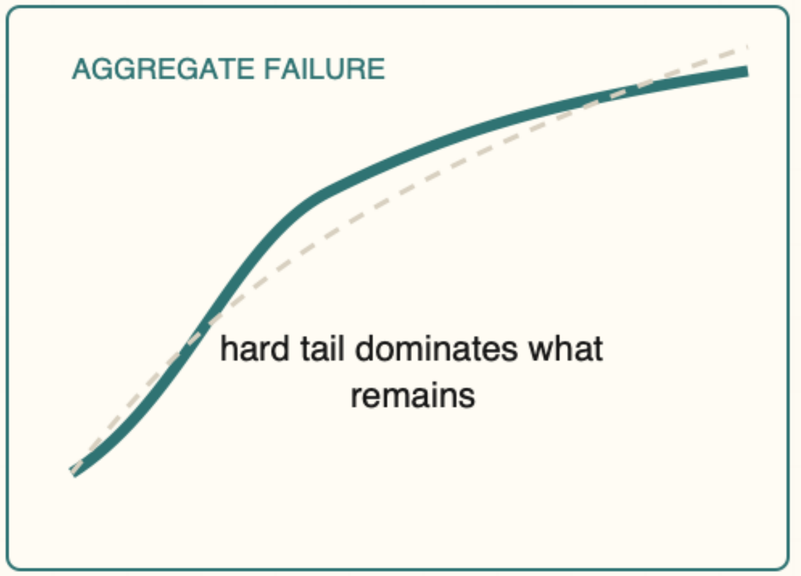
AGGREGATE SCALING

PART 2 07:33-11:20 + arXiv:2502.17578

A Few Extremely Hard Problems Bend the Benchmark Curve

Per-problem exponential scaling and aggregate power laws can coexist.





CAVEAT

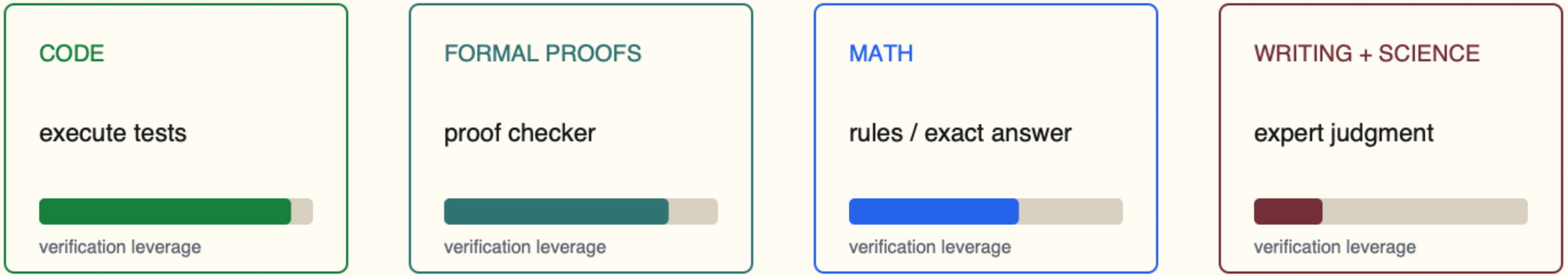
The aggregate curve can look power-law scaled even though each fixed problem decays exponentially.

VERIFICATION BOTTLENECK

PART 2 12:20-26:55 + arXiv:2407.21787

Generation Scales Only as Far as Verification Can Follow

More candidates help when evidence can distinguish success from polished error.

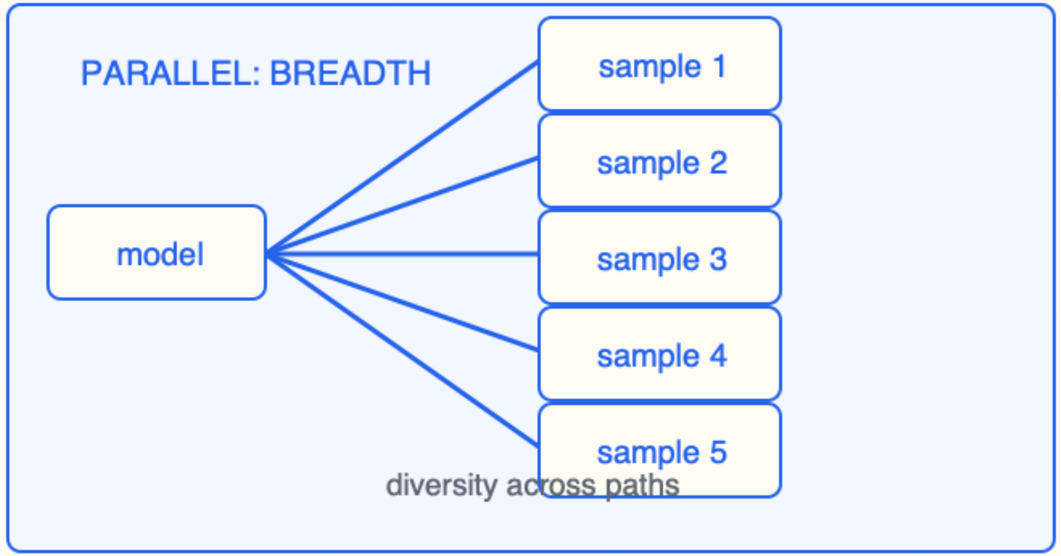


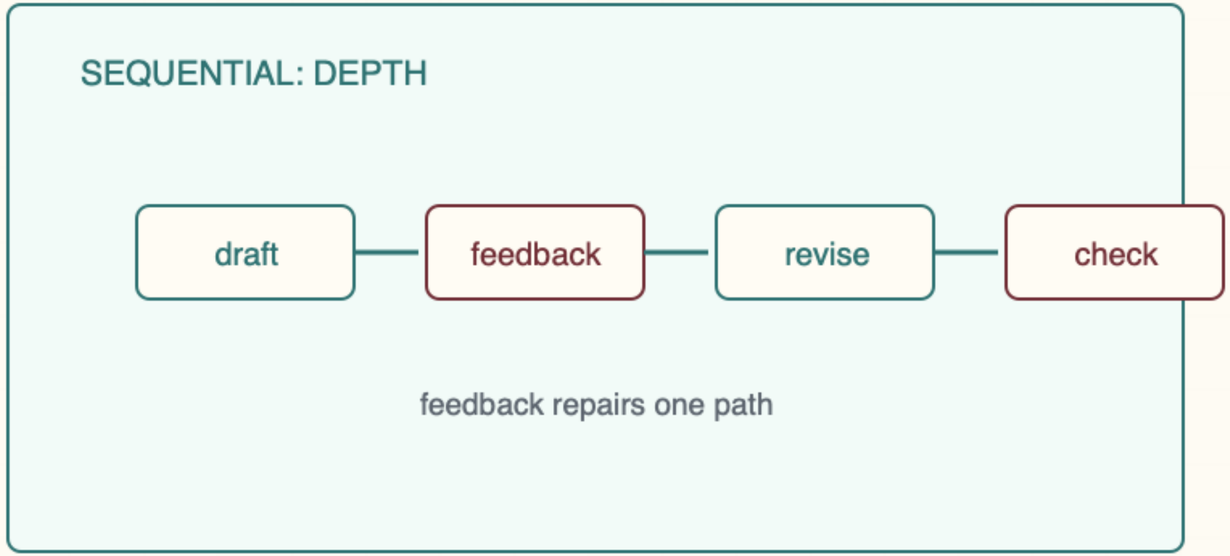
BUDGET ALLOCATION

PART 2 26:55-30:30

The Same Budget Can Explore Breadth or Refine Depth

Parallel samples diversify; sequential revisions use feedback to repair a path.





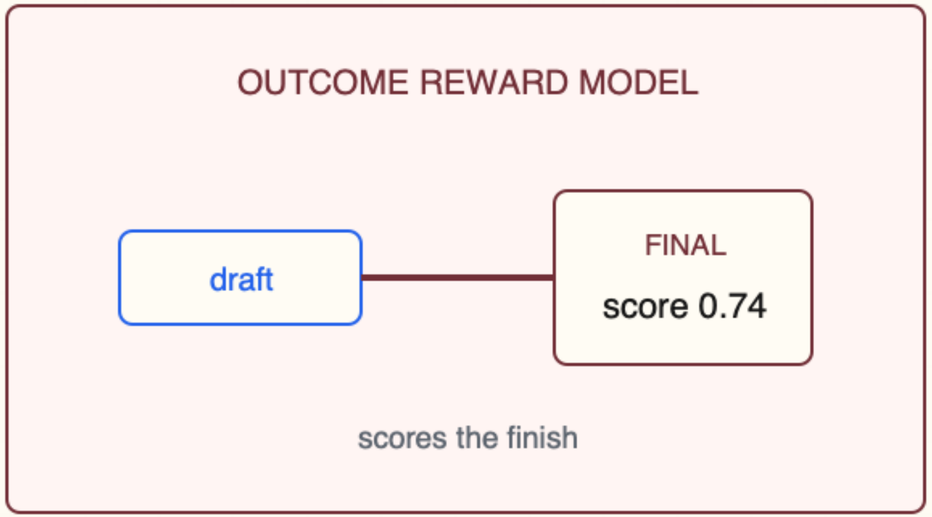
SAME TOTAL BUDGET - DIFFERENT ALLOCATION

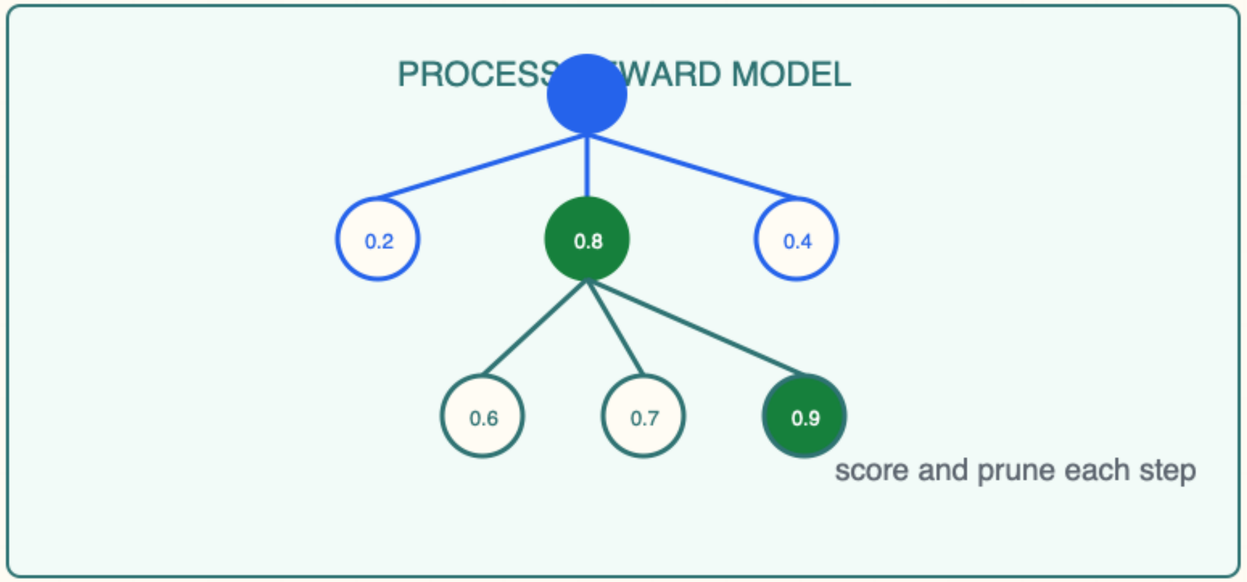
LEARNED EVALUATION

PART 2 30:30-34:40 + arXiv:2408.03314

Outcome Scores Choose Finishes; Process Scores Shape the Route

Reward models guide search, but neither becomes ground truth by definition.





CAVEAT

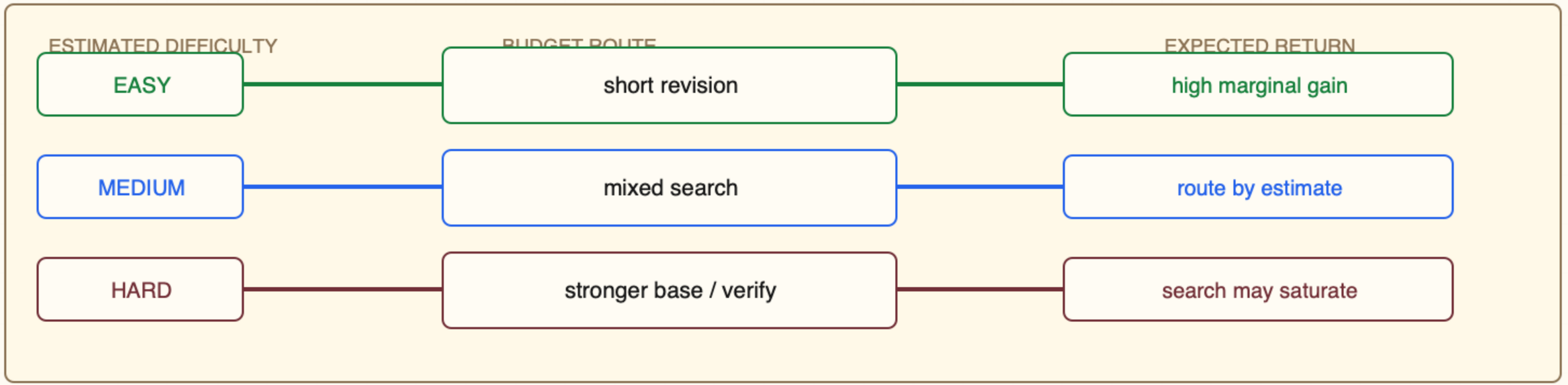
Learned reward models can be confidently wrong or fail outside their training distribution.

COMPUTE-OPTIMAL ROUTING

PART 2 34:40-45:47 + arXiv:2408.03314

Compute Allocation Should Change With Estimated Difficulty

A uniform best-of-N policy can waste budget on tasks with different marginal returns.



CAVEAT

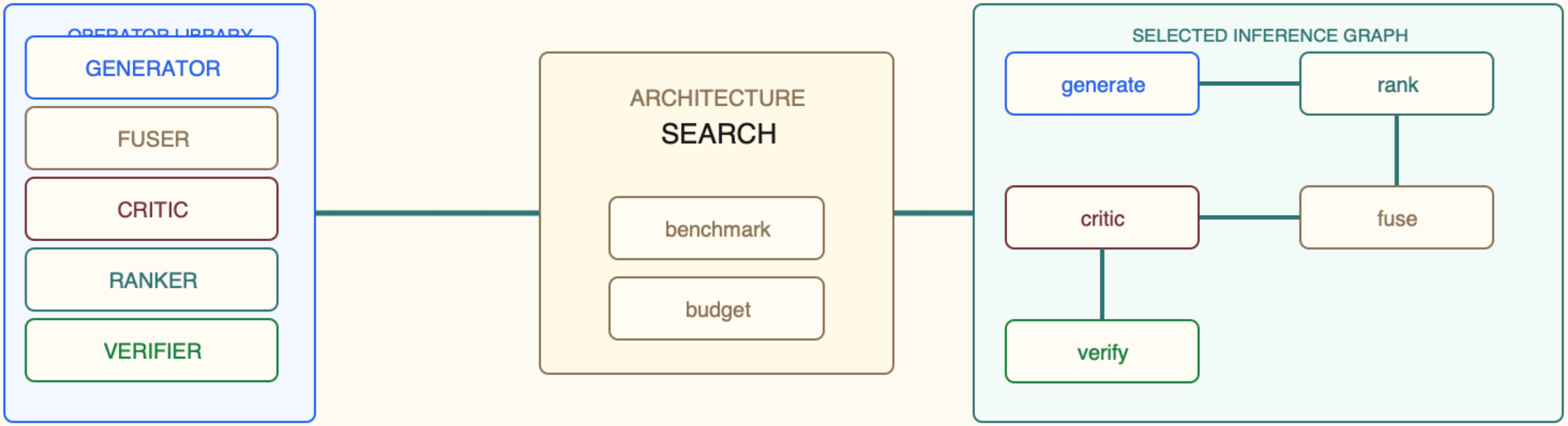
Compute-optimal means optimal for the chosen model, verifier, methods, difficulty estimate, and budget.

INFERENCE ARCHITECTURES

PART 2 45:47-61:22 + arXiv:2409.15254

Archon Searches Over Inference Architectures, Not Just Answers

Models and inference operators become components in a budgeted architecture search.



CAVEAT

Results are benchmark-, version-, operator-set-, and budget-specific.

DECK DIAGNOSTIC

PART 2 whole-lecture synthesis

Before Buying More Inference, Find the Bottleneck

Use four questions to audit any test-time scaling proposal.

01

What creates diversity?

02

How is budget allocated?

03

What verifies success?

04

When does it stop?

NEXT: PART 3 - ROBUST VERIFICATION

The bottleneck moves when generation, selection, cost, or latency saturates.