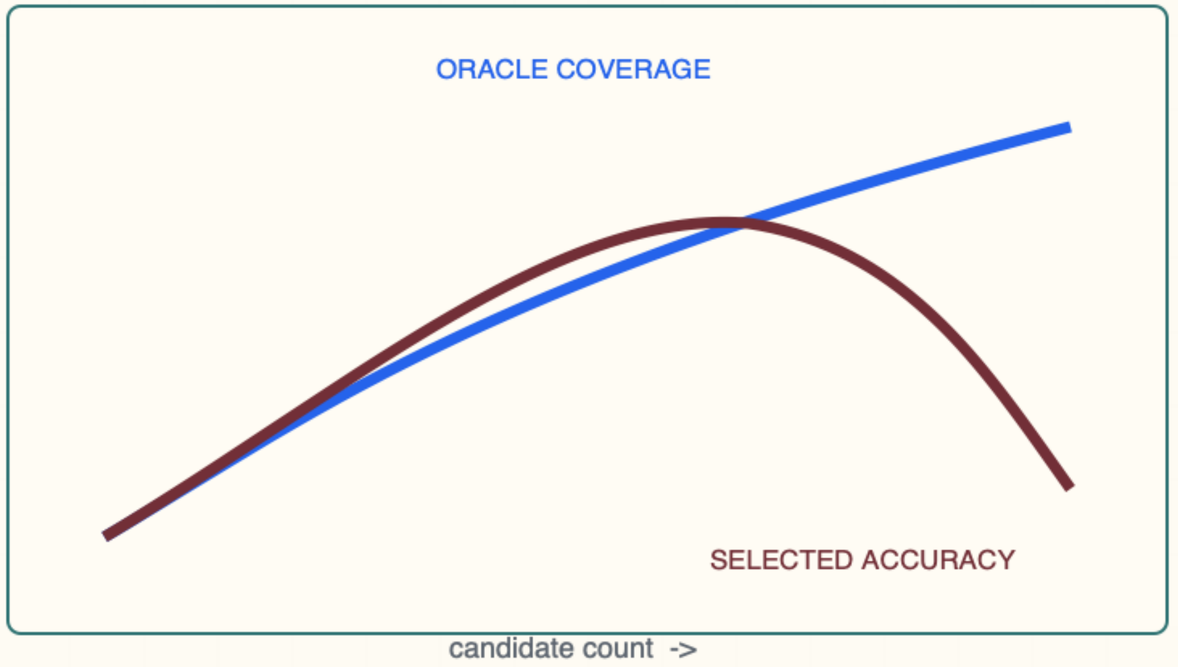
A Generator Is Only as Useful as Its Selector

Candidate coverage may rise while selected-answer accuracy stays flat or falls.

Generate more — Select one



OUTCOME VERIFIER

PART 3 03:18-07:57 + arXiv:2110.14168

The Basic Verifier Loop Learns to Rerank Completed Solutions

Generate, label, train a scalar verifier, then return the highest-scoring candidate.

CAVEAT

A learned correctness score is evidence for ranking, not proof that the selected answer is true.

SELECTION REVERSAL

PART 3 14:19-18:24

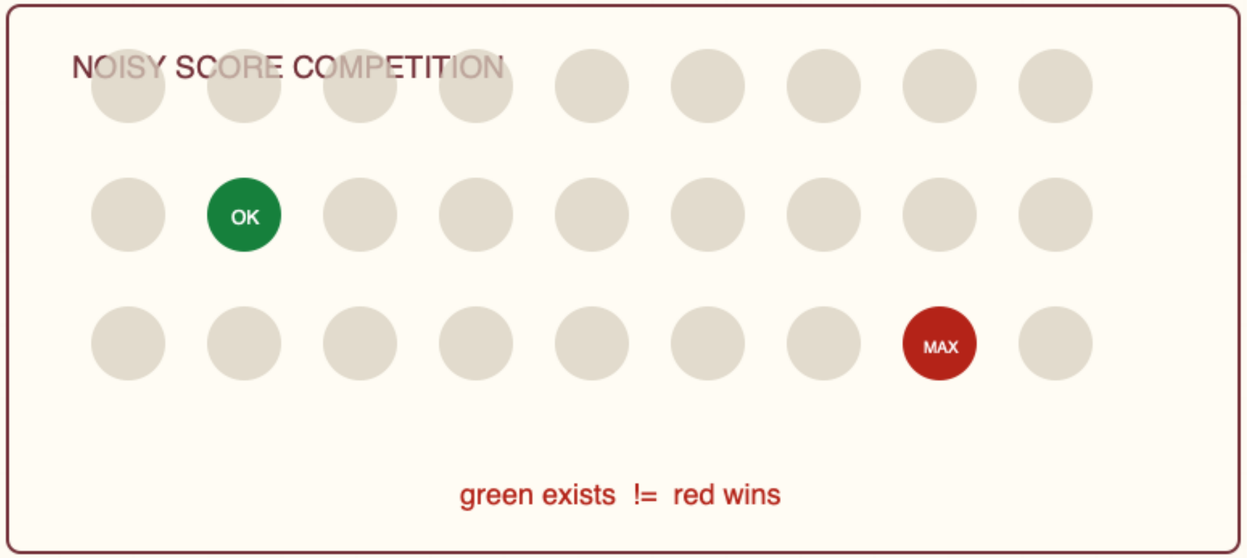
More Candidates Create More False-Positive Opportunities

The maximum noisy score becomes less trustworthy unless verifier precision scales too.

AS N GROWS

chance of a correct candidate rises

chance of an extreme verifier false positive rises too



CAVEAT

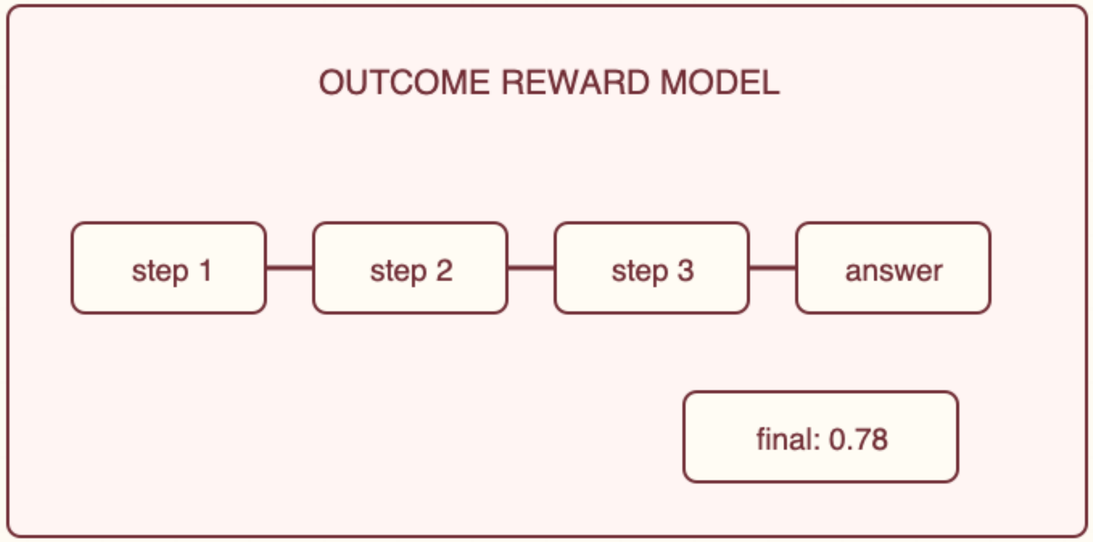
The roughly 400-candidate peak shown in the lecture is setup-specific, not a universal optimum.

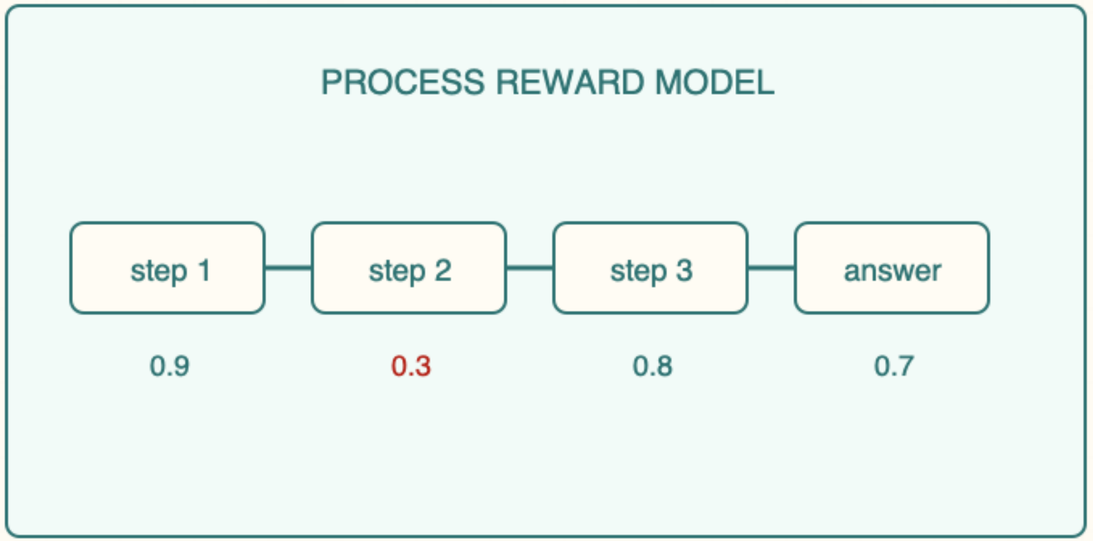
CREDIT ASSIGNMENT

PART 3 21:17-25:06 + arXiv:2305.20050

Outcome Scores Judge the Finish; Process Scores Judge the Route

PRMs add step-level credit assignment to a final-answer signal.





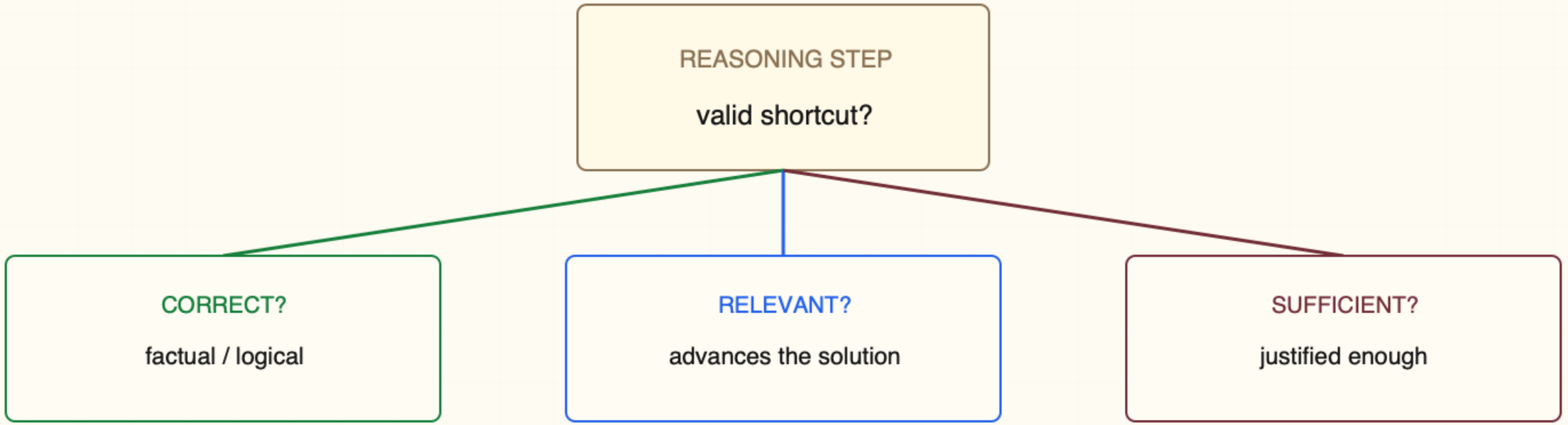
Terminal credit <-> step-level credit assignment

HUMAN SUPERVISION

PART 3 25:06-28:32 / 31:43-37:51

Process Labels Are Richer, but They Encode a Rubric

Correctness, relevance, and sufficient justification depend on an annotation policy.



CAVEAT

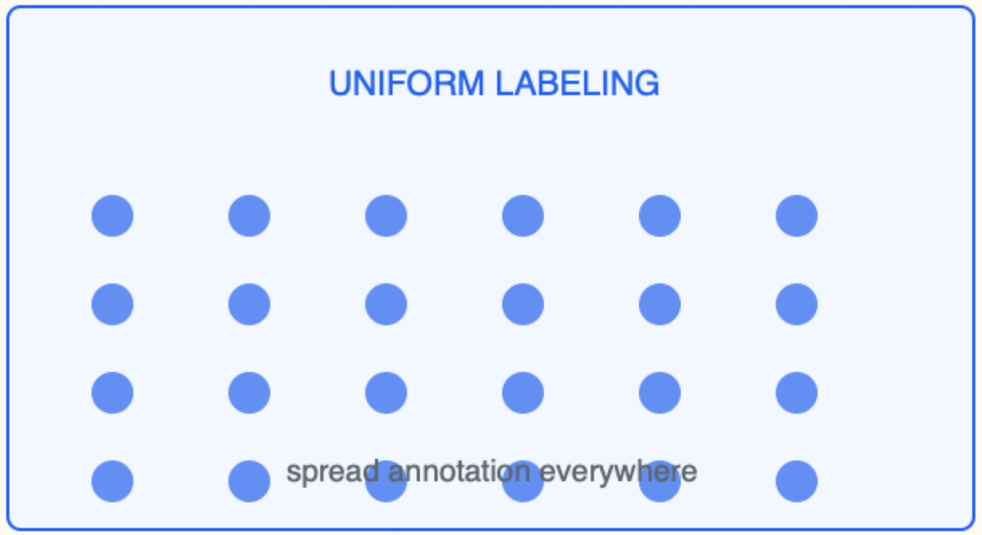
A process label encodes an annotation policy and may reject valid alternative reasoning styles.

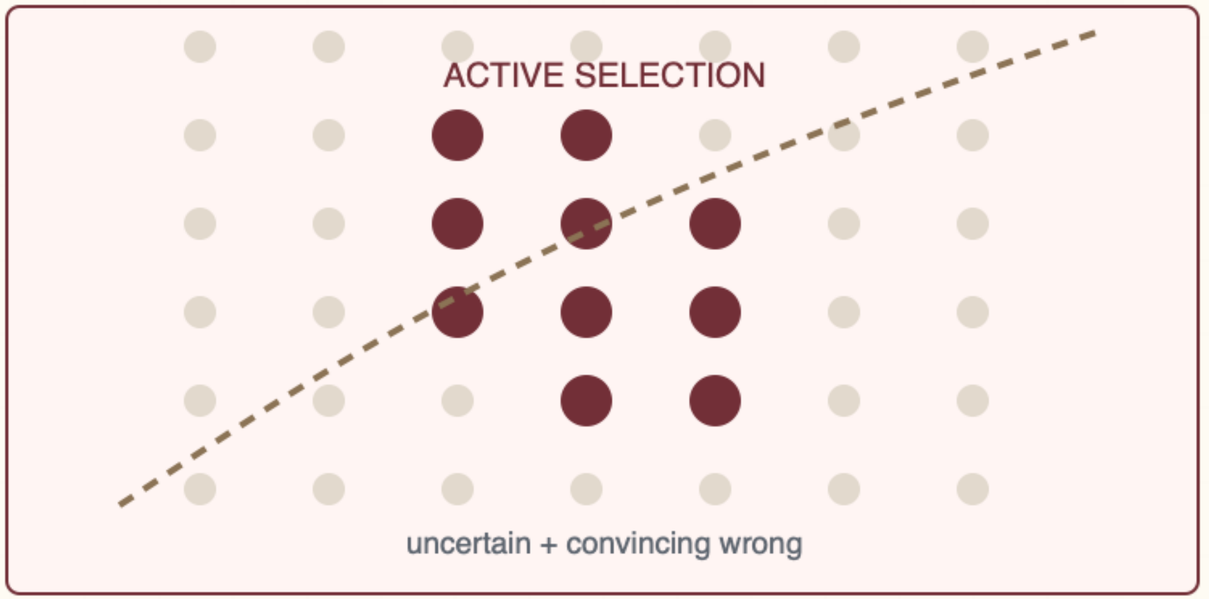
LABEL EFFICIENCY

PART 3 26:11-30:14 + arXiv:2305.20050

PRM800K Targets Informative Mistakes Instead of Labeling Uniformly

Active selection concentrates annotation near uncertain and convincing wrong solutions.





CAVEAT

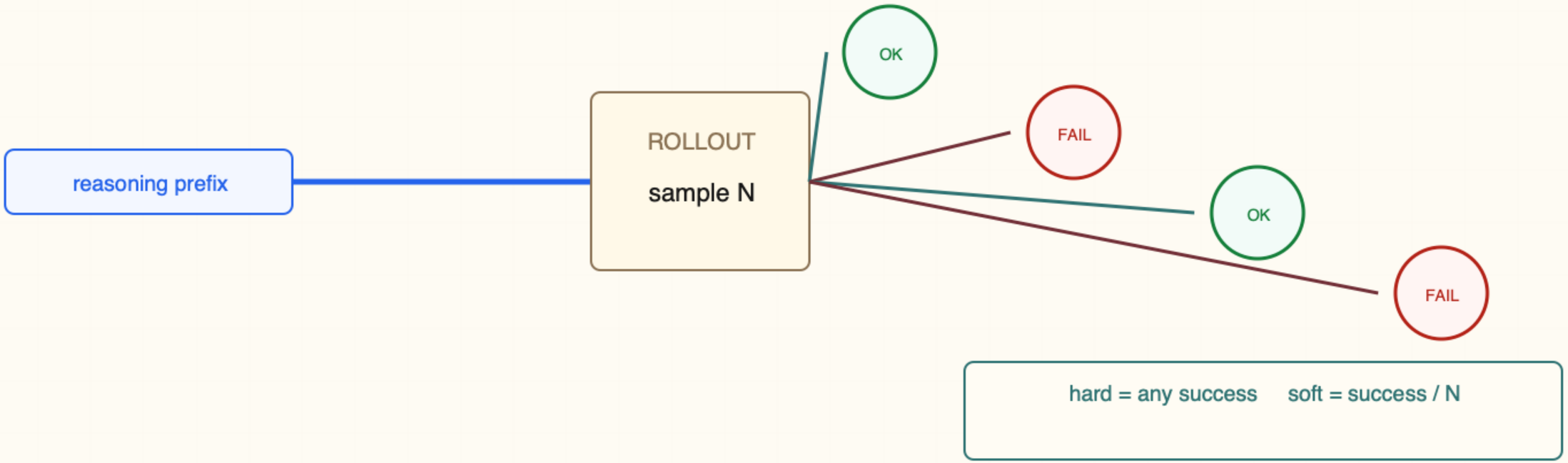
Active selection changes the training distribution and complicates direct label-budget comparisons.

AUTOMATIC SUPERVISION

PART 3 37:51-40:42 + arXiv:2312.08935

Automatic Process Labels Trade Human Cost for Rollout Compute

A prefix is judged by sampled continuations that reach the correct final answer.

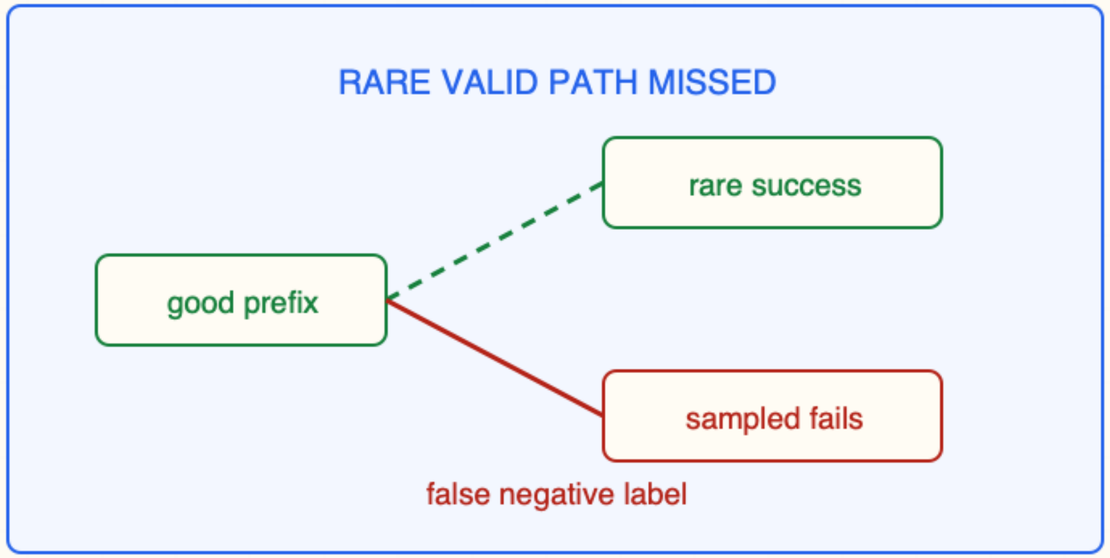


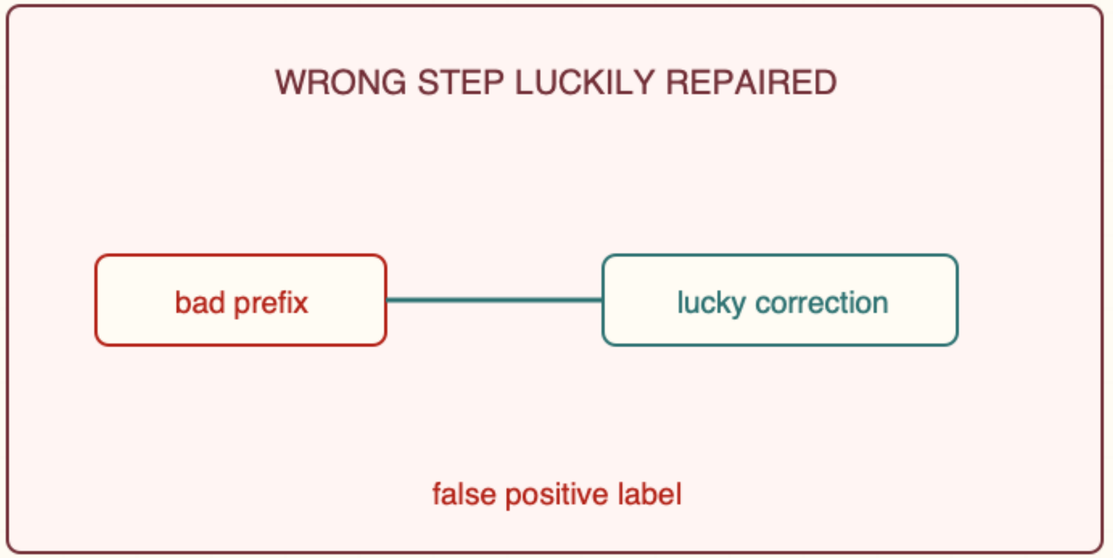
LABEL NOISE

PART 3 40:42-43:27

Rollout Labels Inherit the Search Policy's Blind Spots

Good rare paths can look bad; wrong steps can look good after a lucky repair.





CAVEAT

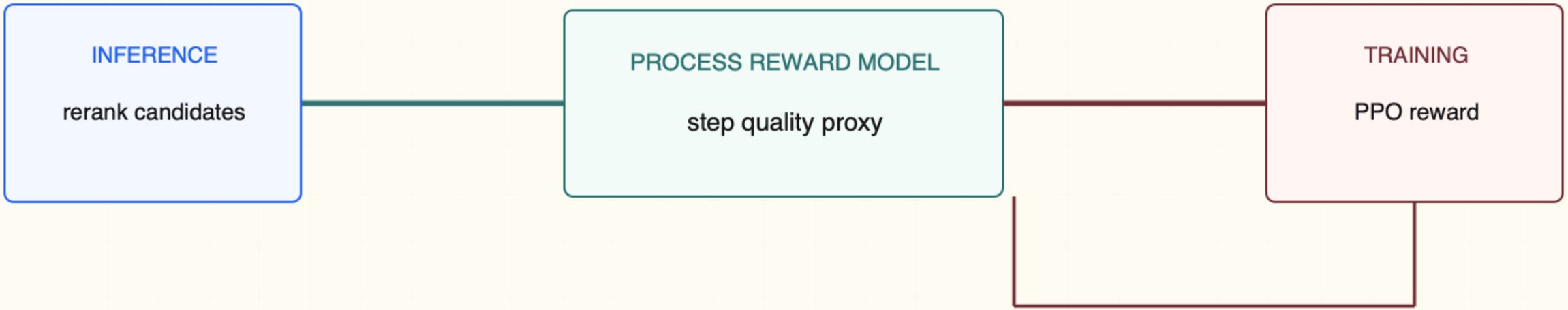
Rollout labels mix prefix quality with the capability and luck of the rollout policy.

VERIFICATION FEEDBACK

PART 3 43:27-47:41 + arXiv:2312.08935

Verification Can Select Outputs and Train the Generator

The same PRM can guide inference-time reranking and PPO-style policy updates.



CAVEAT

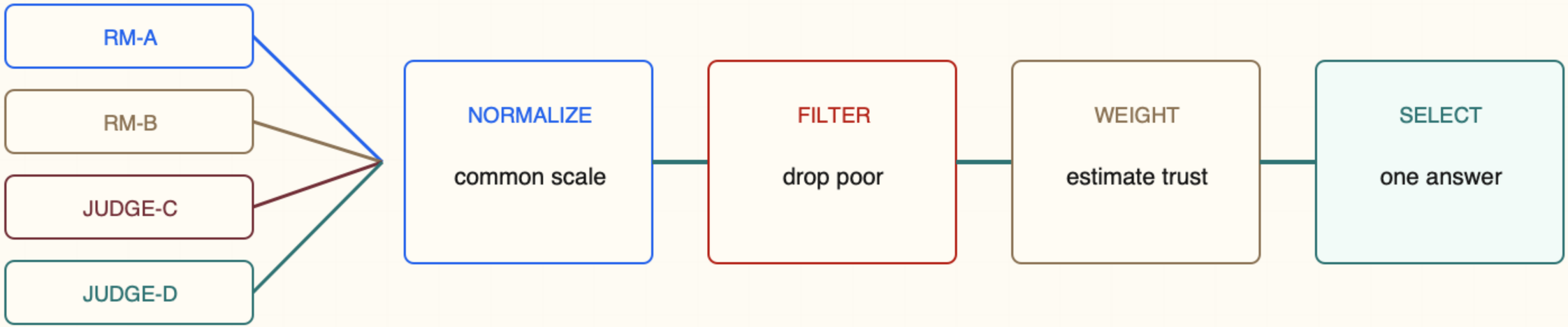
Optimizing an imperfect reward model can raise the proxy score without improving correctness.

WEAK VERIFIER ENSEMBLE

PART 3 51:51-57:36 + arXiv:2506.18203

Diversity Across Verifiers Helps Only After Weighting and Filtering

Weaver normalizes heterogeneous scores, removes poor verifiers, and combines weighted evidence.



CAVEAT

Adding more verifiers naively is not monotonically useful; sources differ in quality and bias.

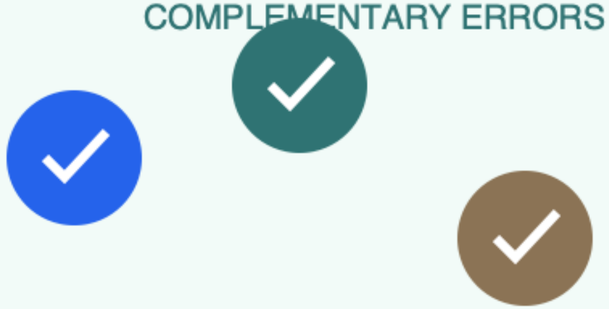
ENSEMBLE ASSUMPTIONS

PART 3 57:36-60:07 + arXiv:2506.18203

Shared Blind Spots Can Defeat an Entire Verifier Ensemble

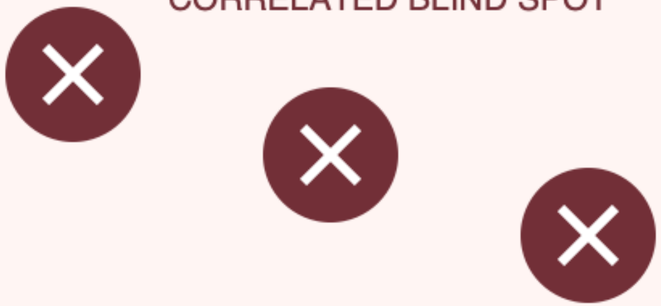
Complementary errors add evidence; correlated errors repeat the same mistake.

COMPLEMENTARY ERRORS



different evidence adds value

CORRELATED BLIND SPOT



same mistake repeated

CAVEAT

Weighting cannot create independent evidence when every verifier shares the same failure.

END-TO-END DESIGN

PART 3 60:07-68:13

Design Verification as a Budgeted End-to-End System

Evaluate selection accuracy, calibration, transfer, cost, and coverage together.

CANDIDATES

coverage

GENERATOR

capability

VERIFIERS

discrimination

DISTILL

serving cost

SELECTED-ANSWER ACCURACY + CALIBRATION + COST

NEXT: PART 4 - LEARNING FROM FEEDBACK WITH TOOLS AND CODE