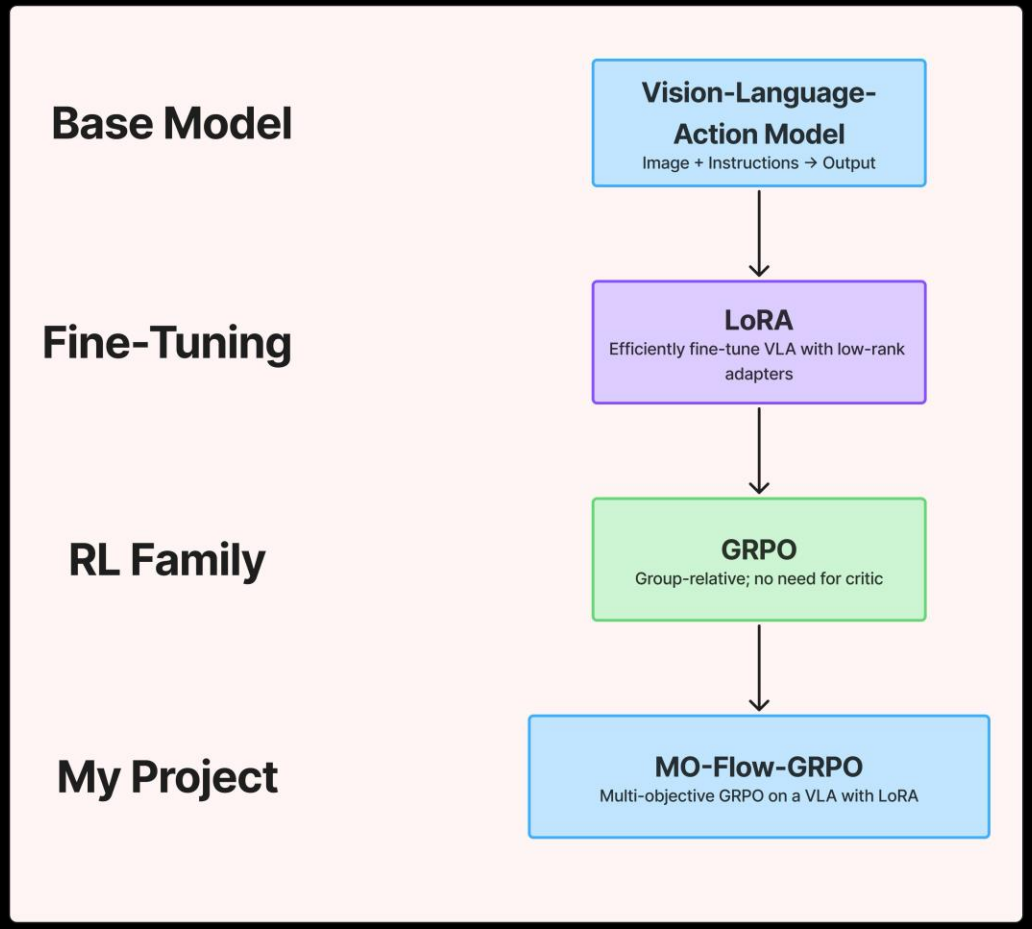
WEEK 3 UPDATE

CODY HILL • 6/8/26

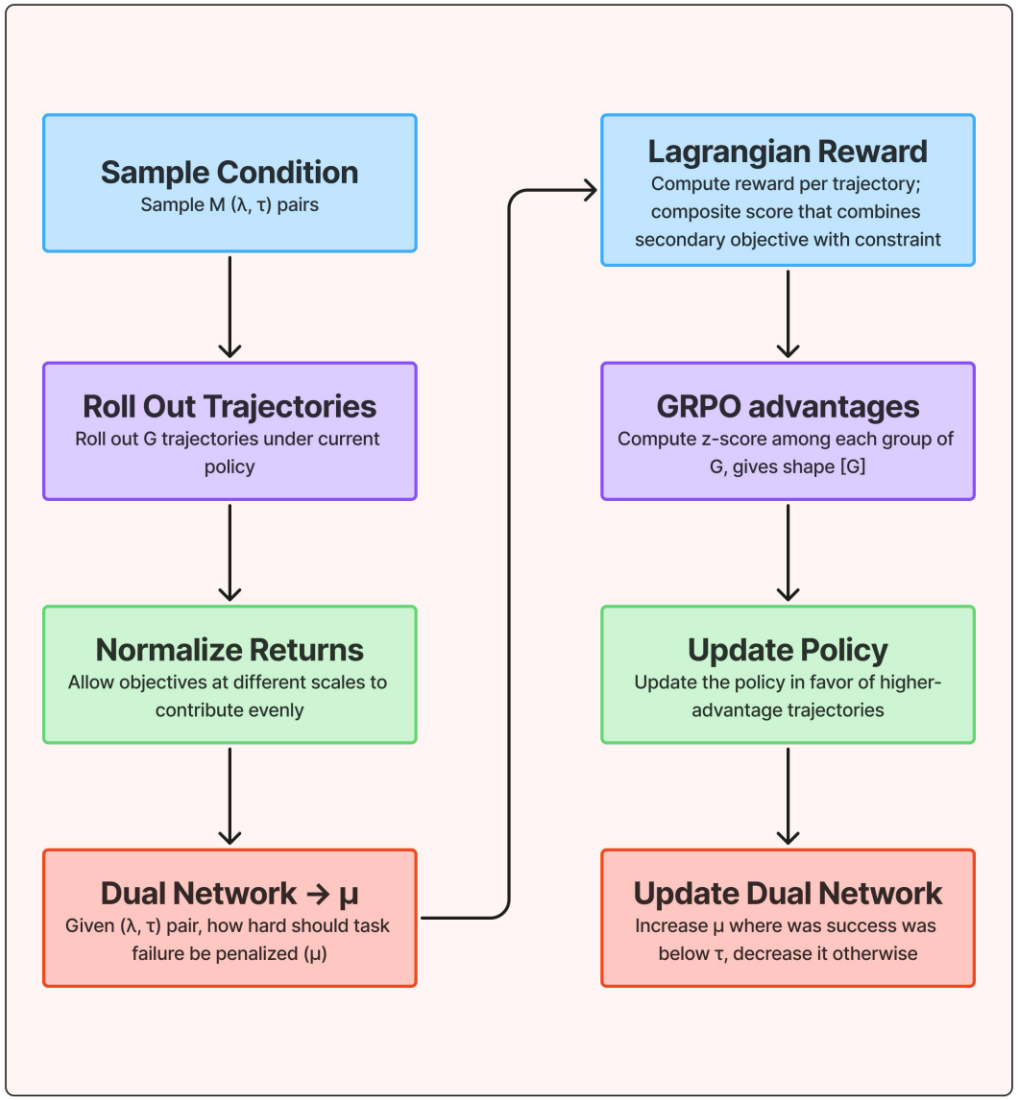
Project: Old Direction



Task

Take GRPO, extend it to optimize four simultaneous objectives and connect it to a flow-matching head.

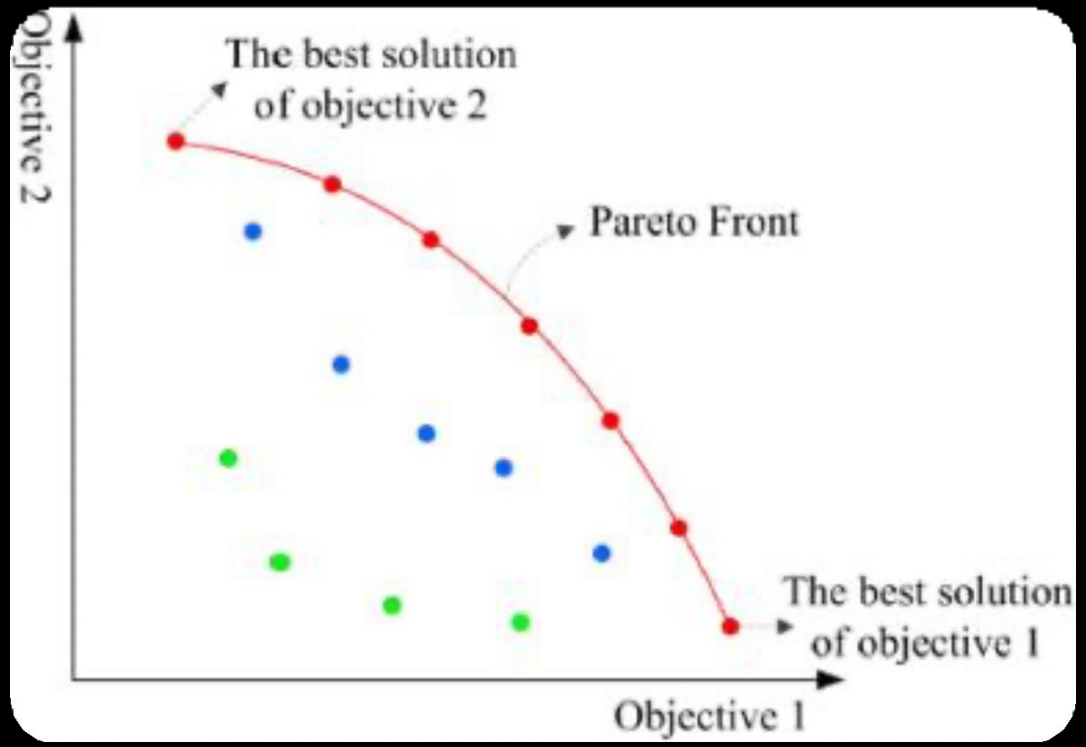
Project: New Direction



Task

Take GRPO, extend it to enforce a hard floor on task success, then Pareto-optimize secondary objectives.

Project: The Task



Challenge

- GRPO → One scalar reward
- CMO-GRPO → Four rewards
- Objectives subject to success floor
- Steer between tradeoffs while staying along Pareto

Progress: Sanity Checks

- MO-Gymnasium Environment “Deep Sea Treasure”
- GRPO Implementation
- MO-GRPO Implementation
- CMO-GRPO Implementation

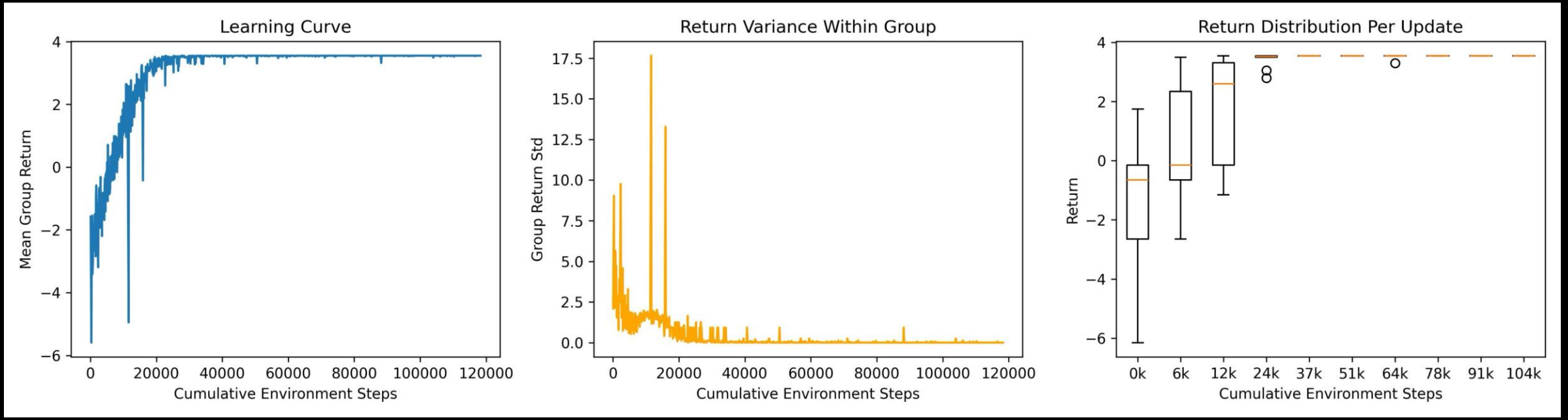


Progress: GRPO

What we want

- Successfully learns a treasure strategy
- Returns increase as we train
- Chosen treasure becomes stable

Progress: GRPO Results



Returns increase

Mean group return increases

Successful strategy + stable selection

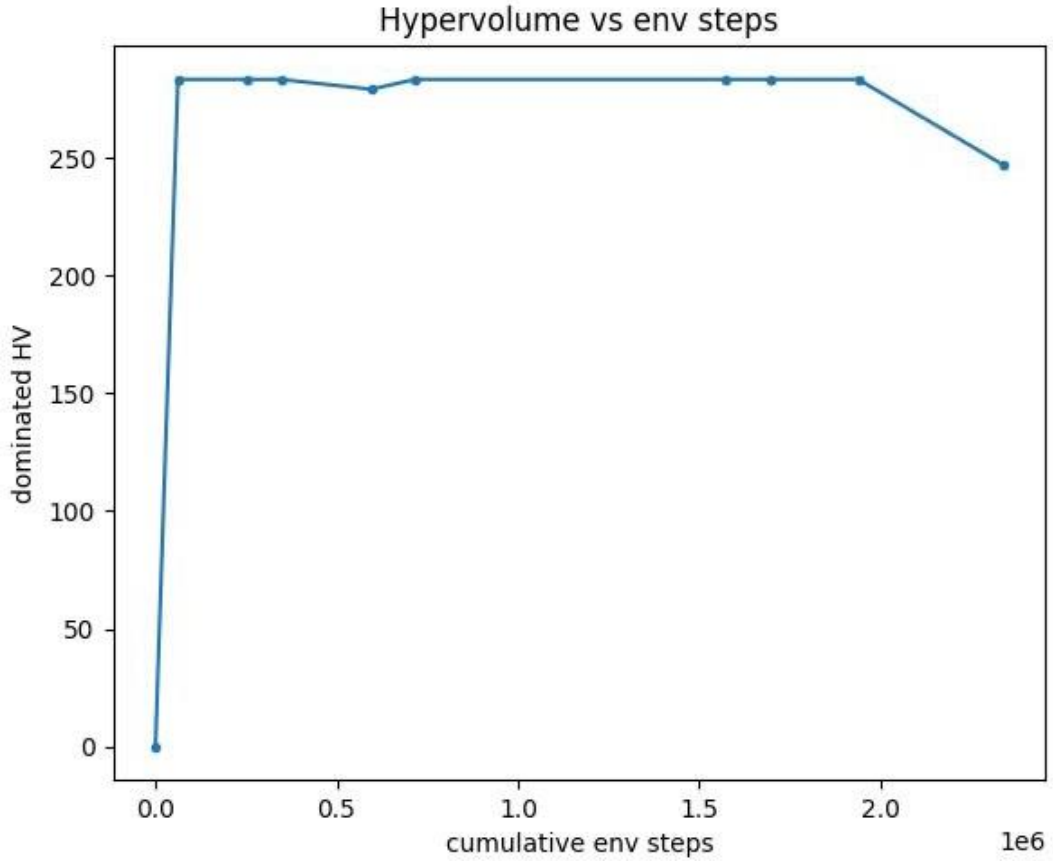
Low variance (left) and return distribution at maximum return (right)

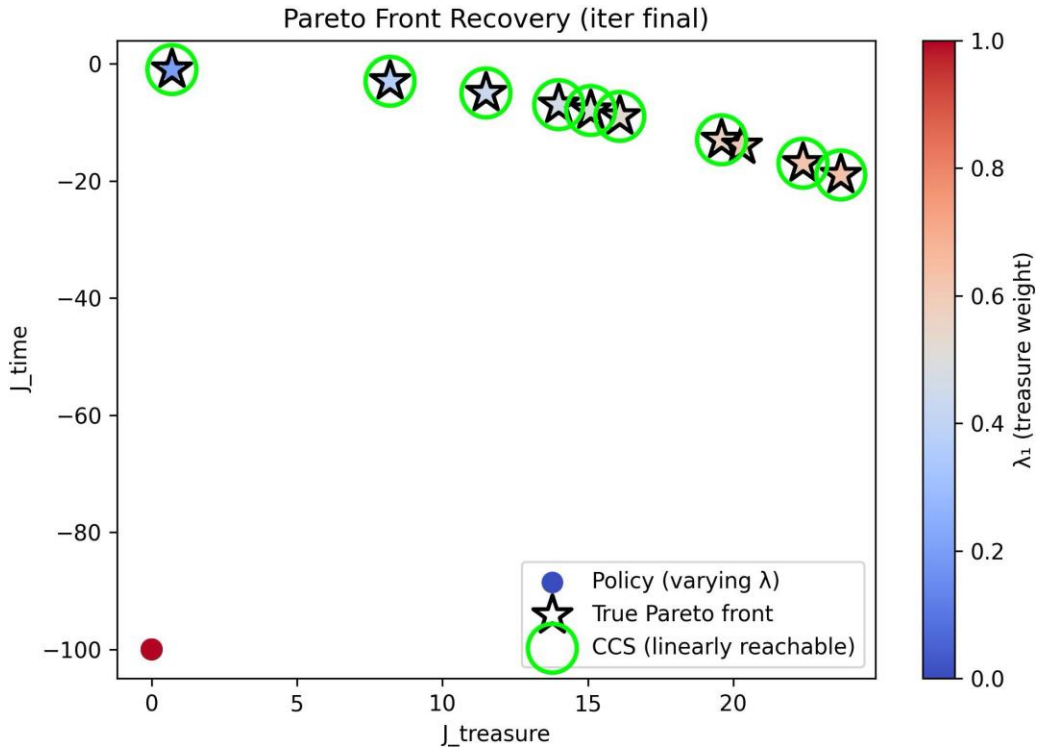
Progress: MO-GRPO

What we want

- One policy → Multiple Tradeoffs
- Pareto set improves
- Smooth interpolation:
  Similar preference → Similar Policy

Progress: MO-GRPO Results





Pareto set improves

Size of covered space increases

Pareto Coverage

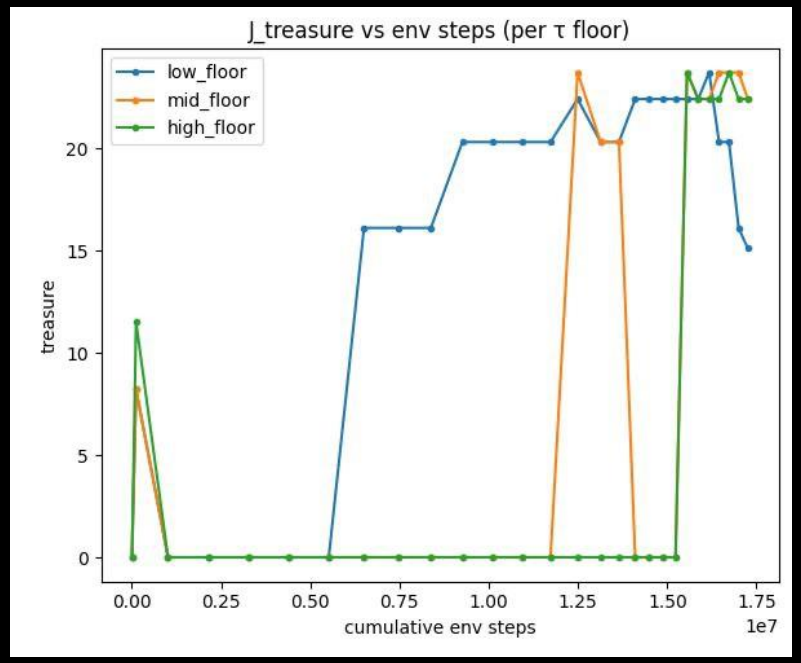
Pareto front reachable

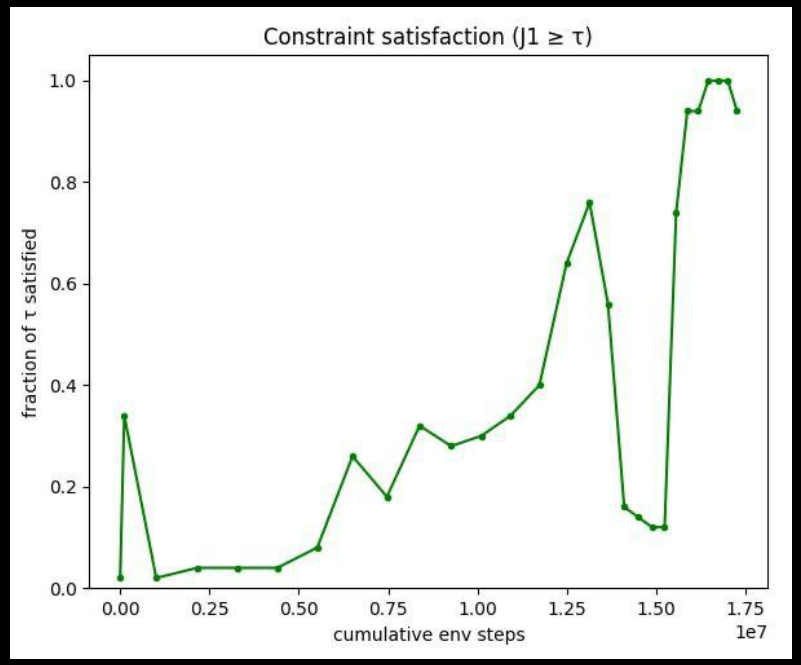
Progress: CMO-GRPO

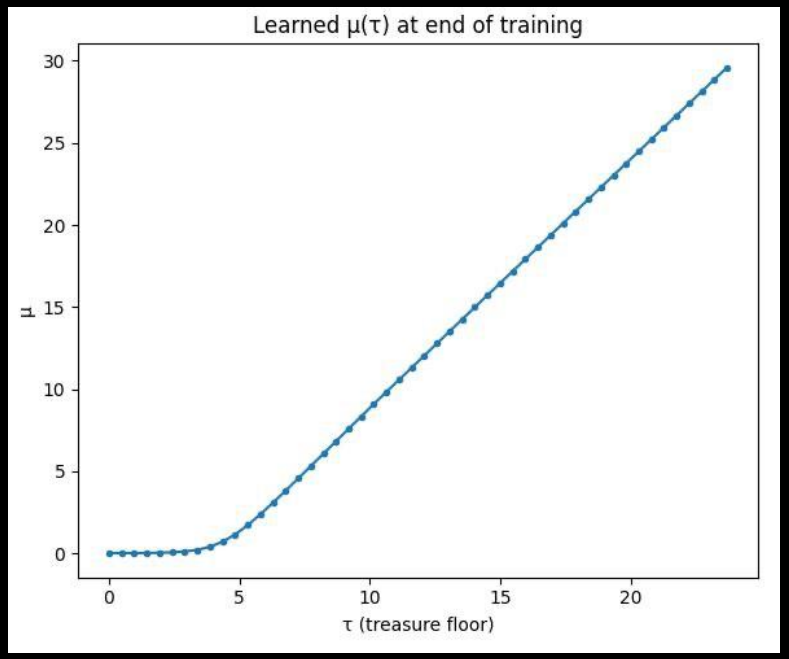
What we want

- High Constraint Satisfaction, approaches 100%
- Higher τ → worse time
- $\mu(\tau)$ increases as τ becomes more difficult

Progress: CMO-GRPO Results







Higher τ → worse time

Constraint Satisfaction

μ(τ) increases

Next Steps

My Plan

- Continuous Environment Implementation
- VLA Implementation

WEEK 3 UPDATE

CODY HILL • 6/8/26