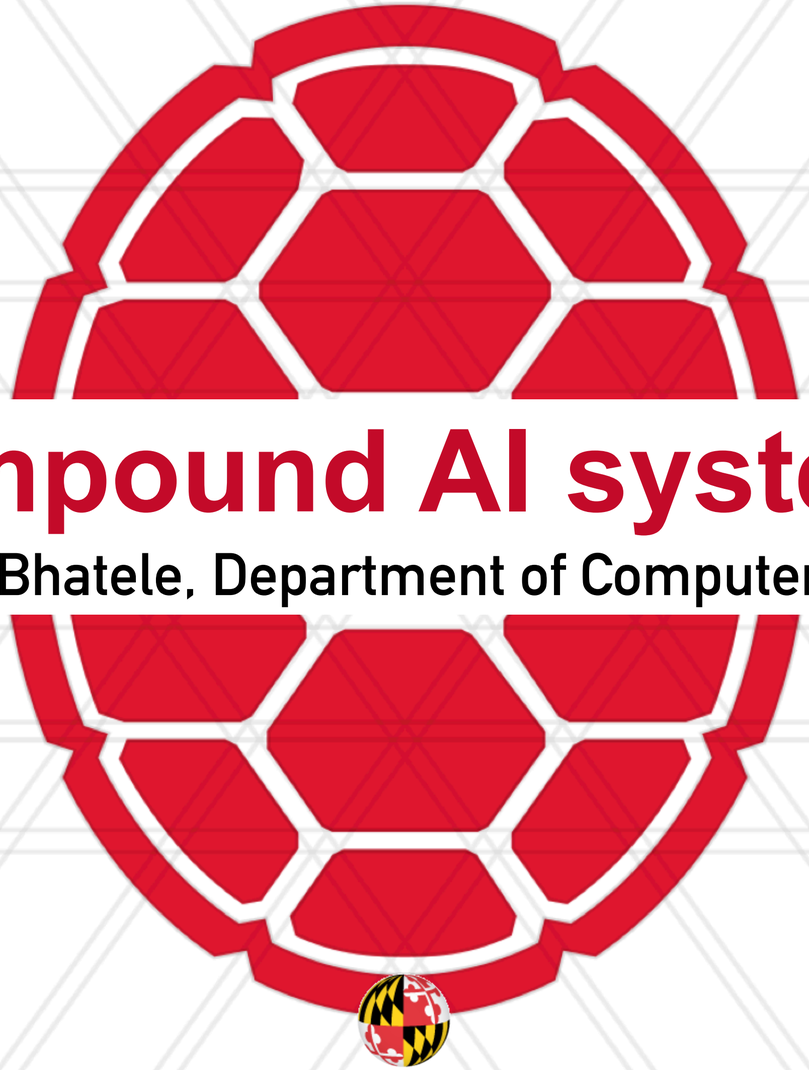
Systems for Machine Learning (CMSC828G)



Compound AI systems

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Compound AI systems

- Monolithic LLMs are not good enough for complex tasks
- Compound systems — combine multiple components such as models, tools/APIs, databases/search engines
- Examples:
  - ChatGPT with internet search, RAG
  - Agentic AI Systems
  - Multi-model pipelines
  - Speculative decoding, Reinforcement learning

Speculative decoding

- Speculative execution in branch prediction
- Run two models
  - Smaller approximation (draft) model is used to sample generations
    - Generate several tokens ahead
  - Larger target models evaluates all the guesses in parallel

Reasoning models

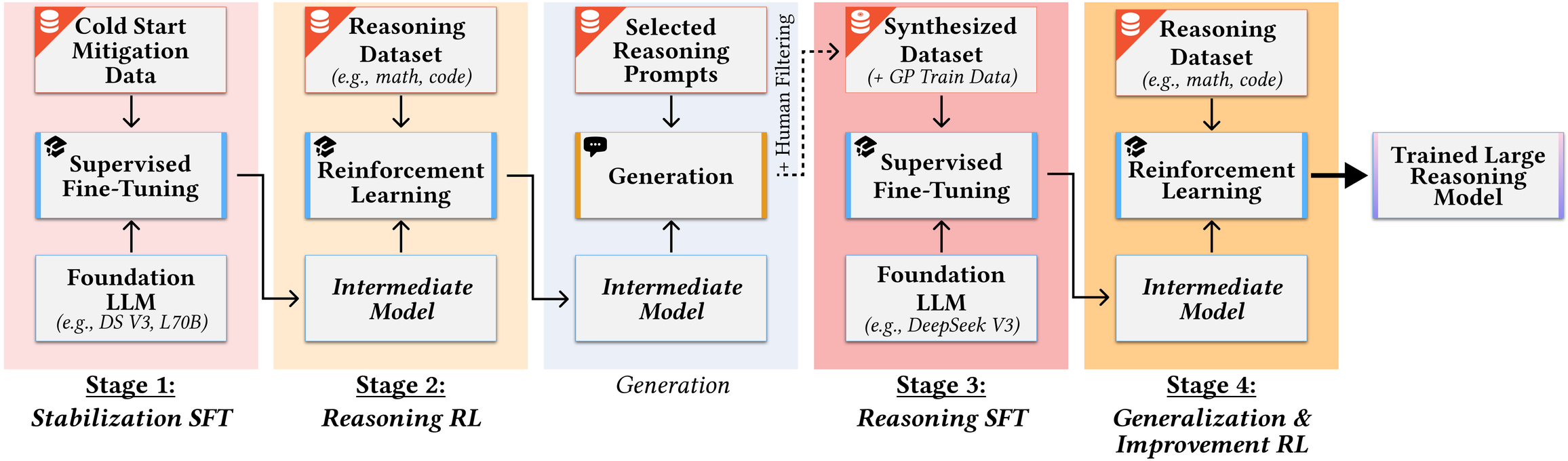
- Breaks down the problem into multiple steps
- Generates and inspects intermediate steps (e.g. chain of thoughts)
- Tools can be called as part of the reasoning workflow
- Often trained/tuned using:
  - Reasoning traces or chain-of-thought data
  - Reinforcement learning

Agentic systems

- Systems that plan and make decisions, often using tools
- Planning: create sub-tasks, decide dependencies and order, decide which tools to call
- Often iterate and retry

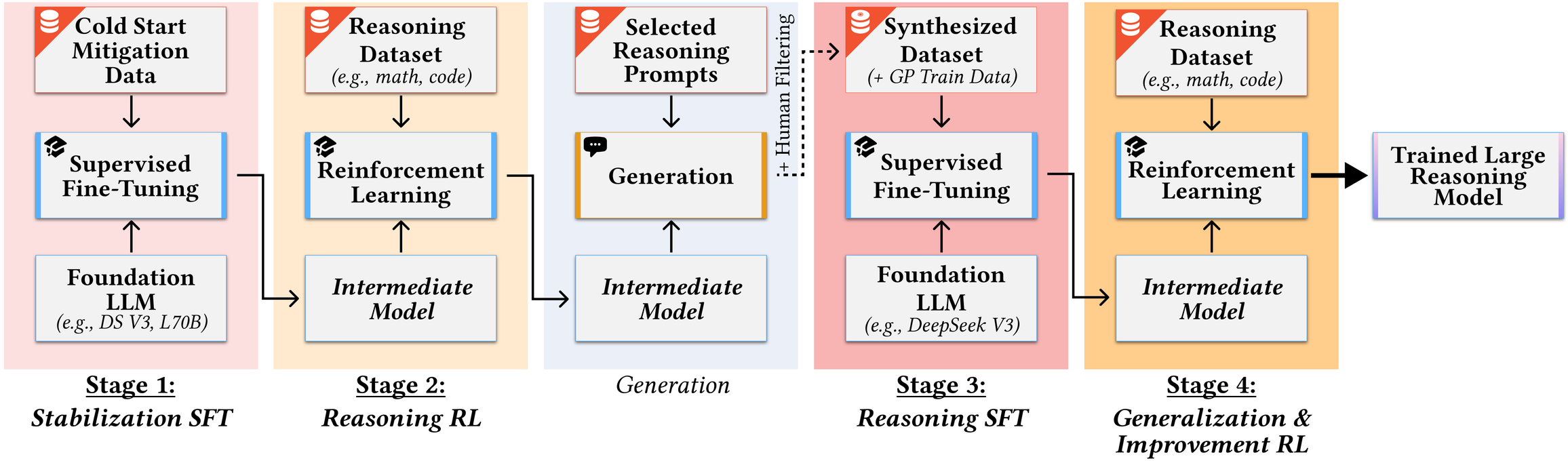
HPC-R1: Create a R1-like training pipeline

- Stage 1: SFT with high-quality curated datasets
- Stage 2: RL (multiple responses generated by model are scored)



HPC-R1: Create a R1-like training pipeline

- Stage 3: SFT of Stage 1 model using data synthesized from Stage 2 model
- Stage 4: Apply GRPO again

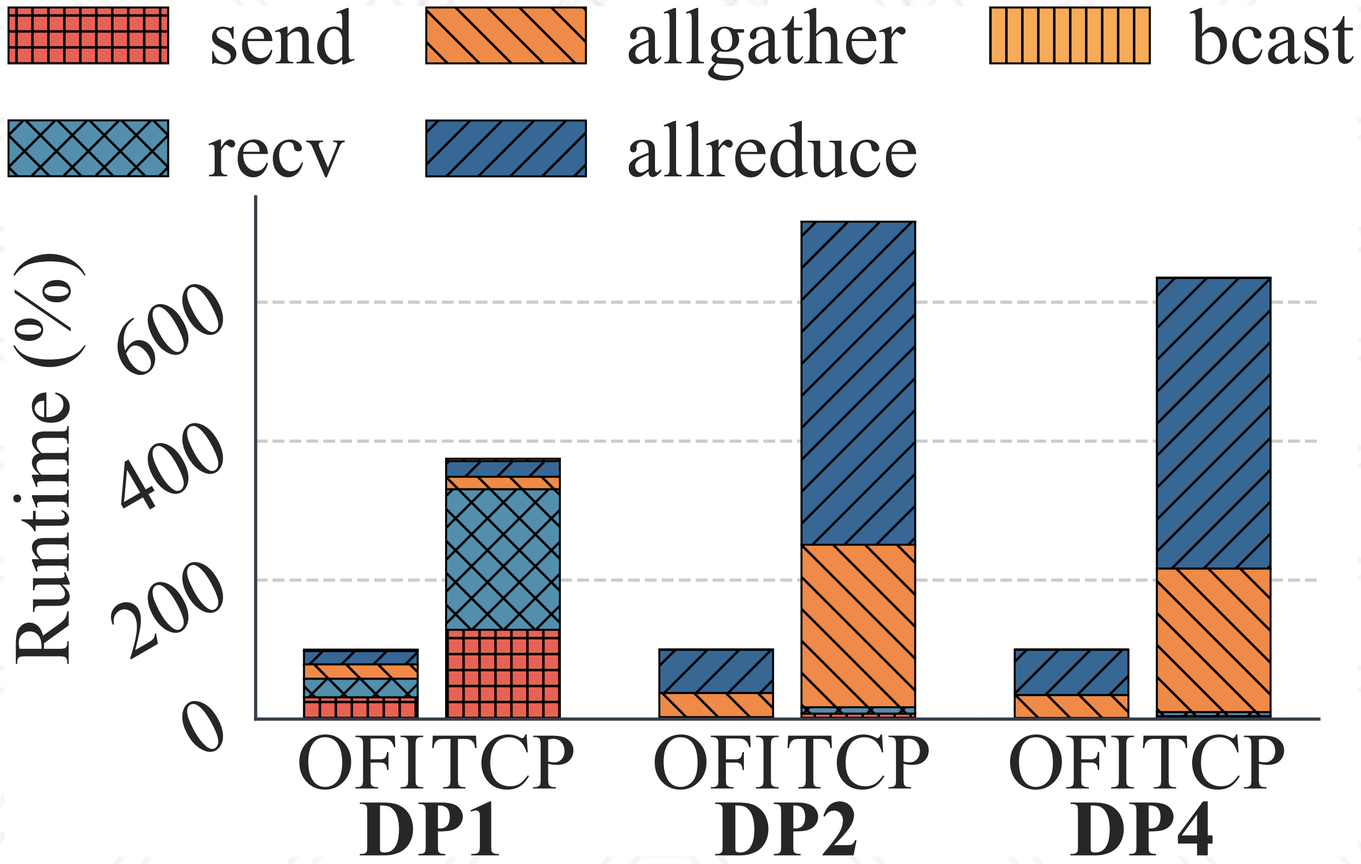


Setup

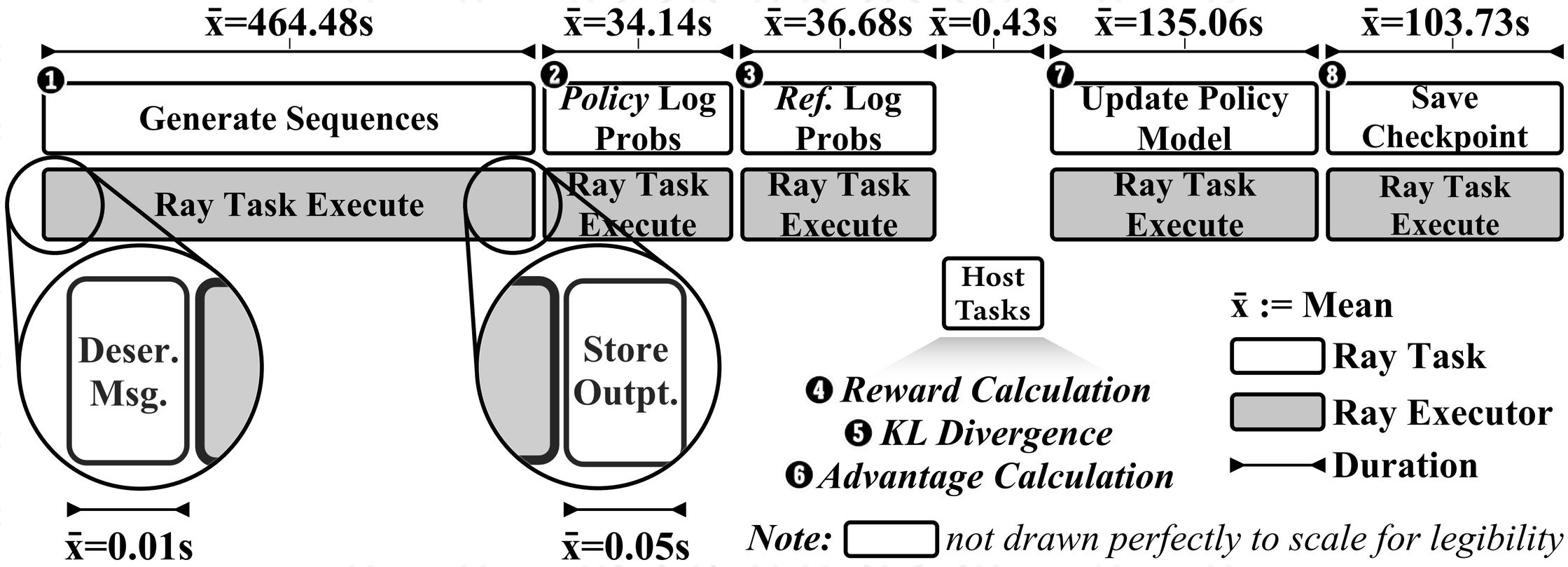
- Base model: Llama 3.3 70B
- Minimum batch size: 4
- Weak scaling
- Framework used:
  - Megatron-DeepSpeed (stages 1 and 3)
  - VERL (stages 2 and 4)

Stage, #	1	2	4	8	16	32	64	128	256
SFT	✗	✗	✗	✗	✗	✗	⚠	✓	▂▄▆
RL-GRPO	✗	✗	✗	✗	⚠	✓	▂▄▆	▂▄▆	▂▄▆
Generation	✗	⚠	✓	✓	✓	▂▄▆	▂▄▆	▂▄▆	▂▄▆

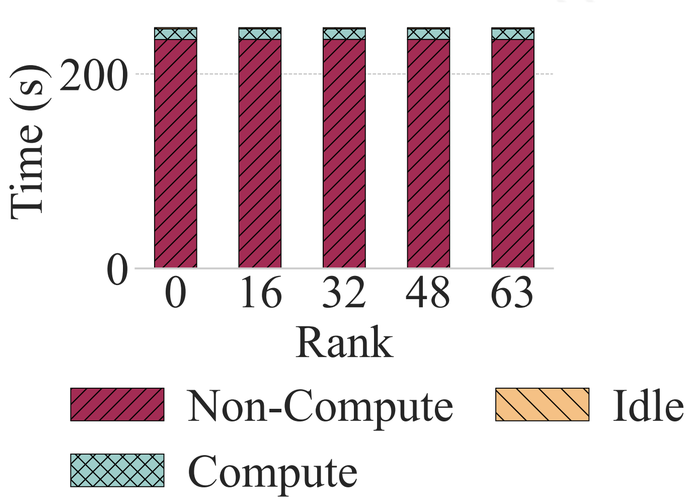
S3: Send/Recv and AllReduce

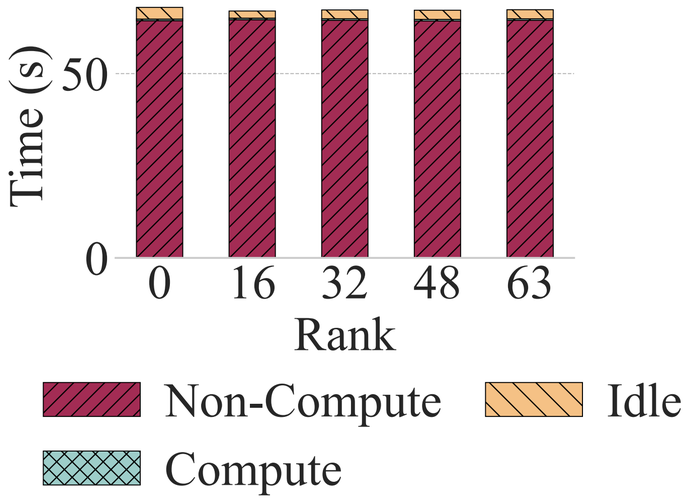


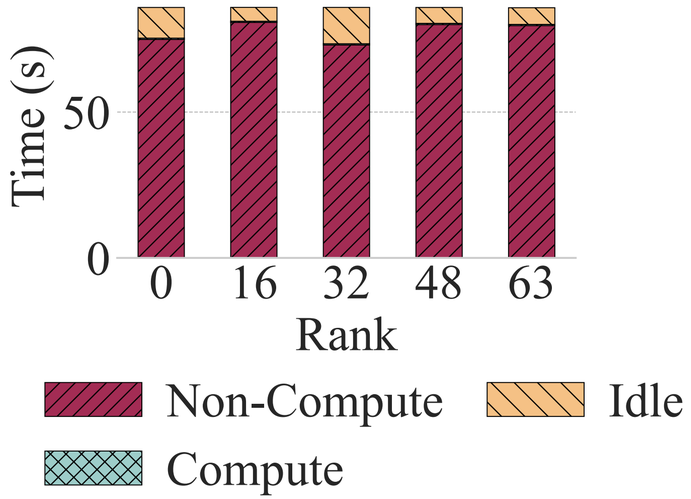
GRPO


x̄=464.48s
x̄=34.14s
x̄=36.68s
x̄=0.43s
x̄=135.06s
x̄=103.73s
1
Generate Sequences
Ray Task Execute
2
Policy Log Probs
Ray Task Execute
3
Ref. Log Probs
Ray Task Execute
7
Update Policy Model
Ray Task Execute
8
Save Checkpoint
Ray Task Execute
Deser. Msg.
x̄=0.01s
Store Outpt.
x̄=0.05s
Host Tasks
4 Reward Calculation
5 KL Divergence
6 Advantage Calculation
x̄ := Mean
Ray Task
Ray Executor
Duration
Note: not drawn perfectly to scale for legibility

O3: RL phase is communication bound

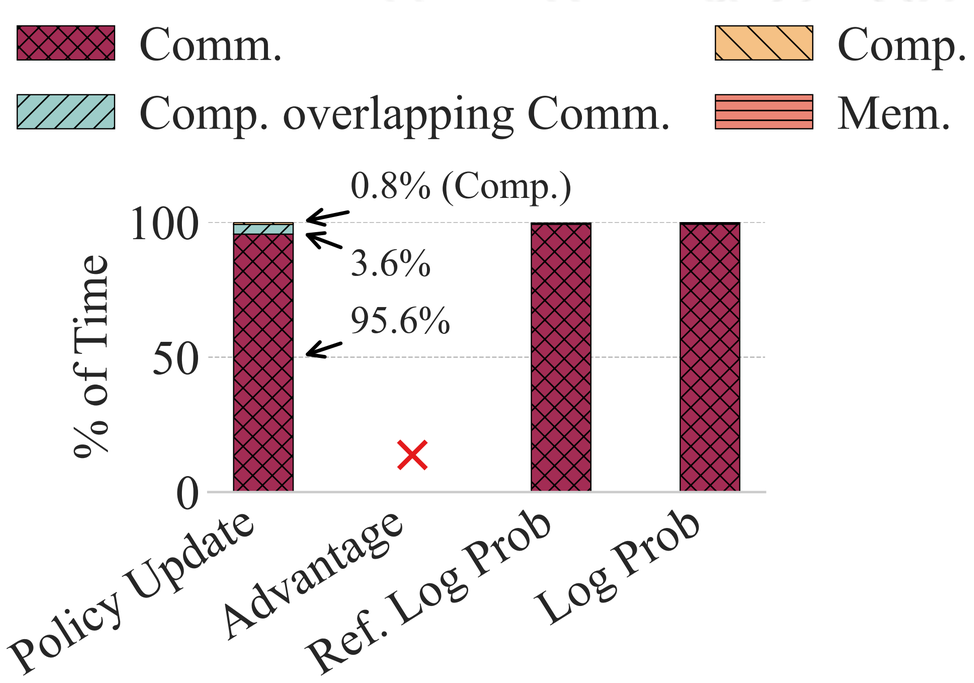




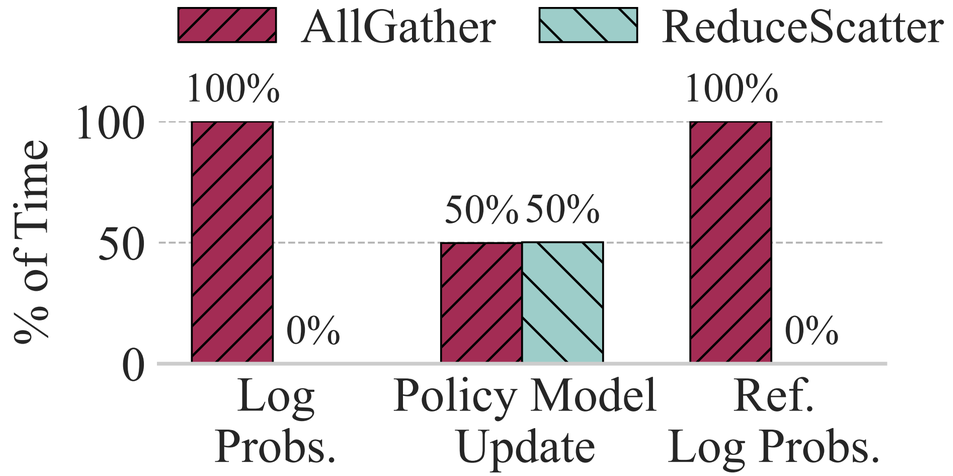


(a) Policy Model Update **(b) Policy Model Log Probs Computation** **(c) Reference Model Log Probs Computation**

O5: Communication operations



(c) Breakdown by operation.



(d) Breakdown by collective.

UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu