EIC-BeamAI

Rensselaer

why not change the world?®

Brookhaven National Laboratory

Active Supervision for AGS Bunch-merging with LLM-based Reinforcement Learning

Presenters: Yue Zhao, Yuan Gao, Yinan Wang

Rensselaer Polytechnic Institute
Brookhaven National Laboratory

Optimization of AGS bunch merging with reinforcement learning

- **Traditional method:** The RHIC heavy ion program relies on a series of RF bunch merge gymnastics to combine individual source pulses into bunches of suitable intensity.
- The optimal setting drifts and performance degrades without constant attention “by eye”.

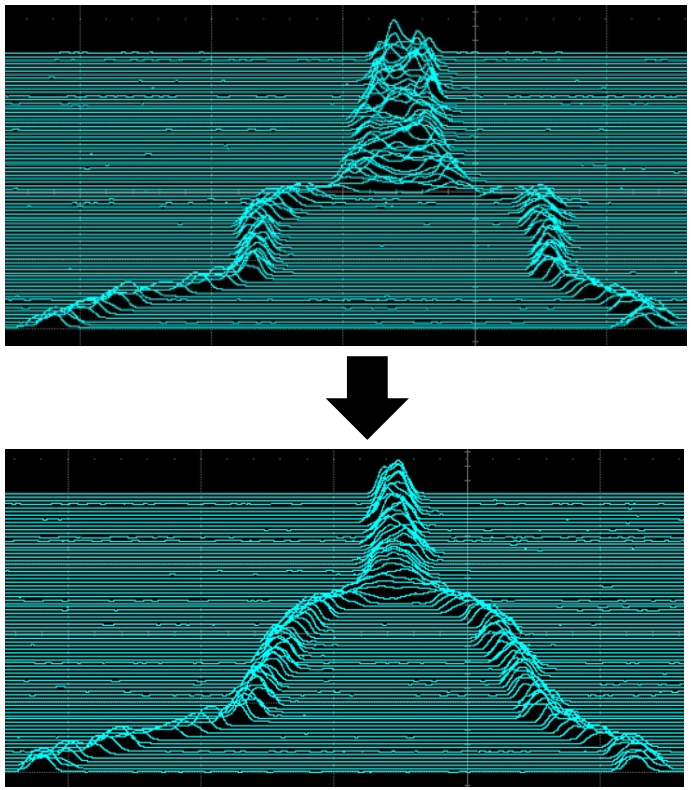


Figure 1 : Example of AGS Scope Data

IRL Method

Pro:

- Task-specific modeling
- Inverse Reinforcement Learning (IRL) method is introduced.

Future work:

- **RL requires a significant amount of data to control and optimize as control points increase**

- Data Centric
- Sample-efficient Optimization Method

Reference: Y. Gao, K. Zeno, K. Brown, L. Nguyen, V. Schoefer, A. Kasparian, A. Edelen, D. Sagan, E. Hamwi, G. Hoffstaetter, J. Unger, W. Lin, M. Schram, and Y. Wang, “Optimization of AGS bunch merging with reinforcement learning,” in *Proc. IPAC’24*, Nashville, TN, USA, May 19–24, 2024, pp. 1782–1785, doi: 10.18429/JACoW-IPAC2024-TUPS53.

Motivation:

- LLM-based agents lack specialization in tackling specific target problems.
- Reinforcement learning (RL) often suffer from low sampling efficiency and high exploration costs.

Goal: Proposed a framework that training a smaller, specialized student RL agent using instructions from an LLM-based teacher agent.

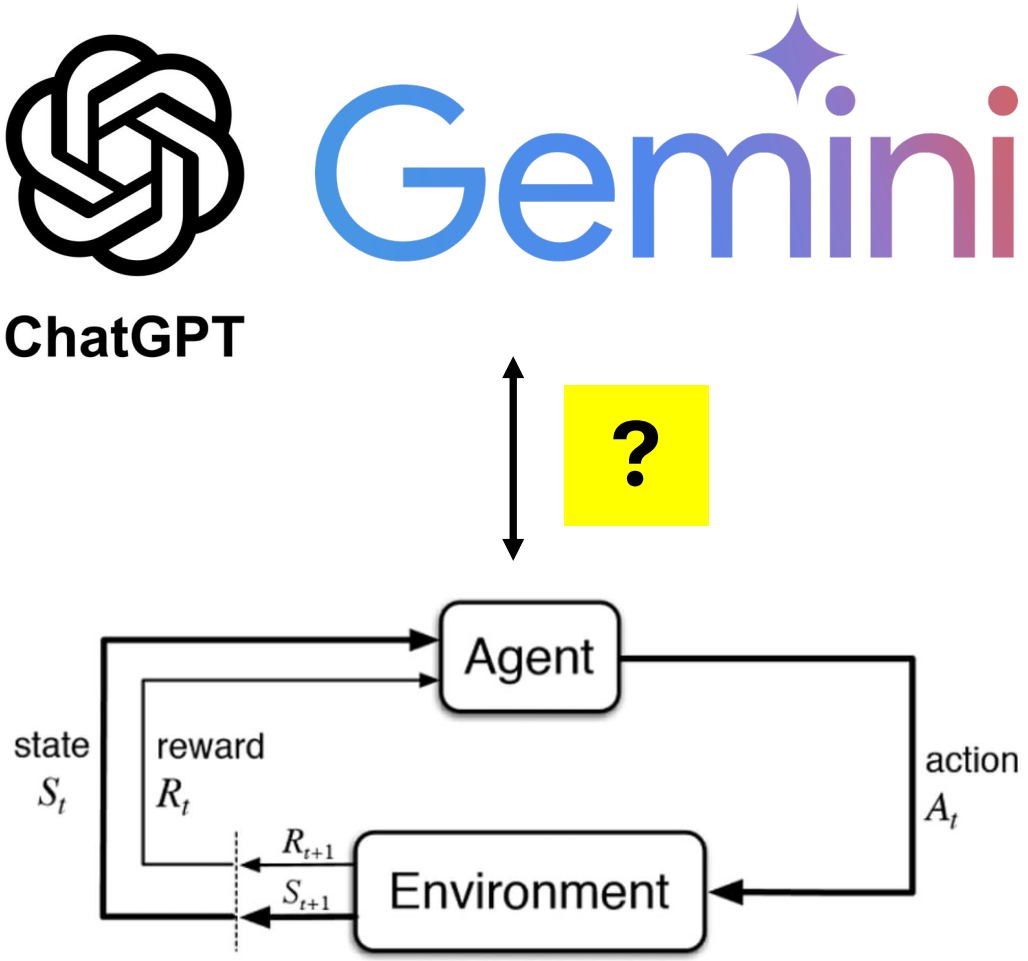


Figure 2 : Motivation of LLM4Teach

Reference: Z. Zhou, B. Hu, C. Zhao, P. Zhang, and B. Liu, "Large Language Model as a Policy Teacher for Training Reinforcement Learning Agents," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, Jeju, Korea, Aug. 2024, pp. 5671-5679, doi: 10.24963/ijcai.2024/627.

Large Language Model as a Policy Teacher for Training Reinforcement Learning Agents (LLM4Teach)

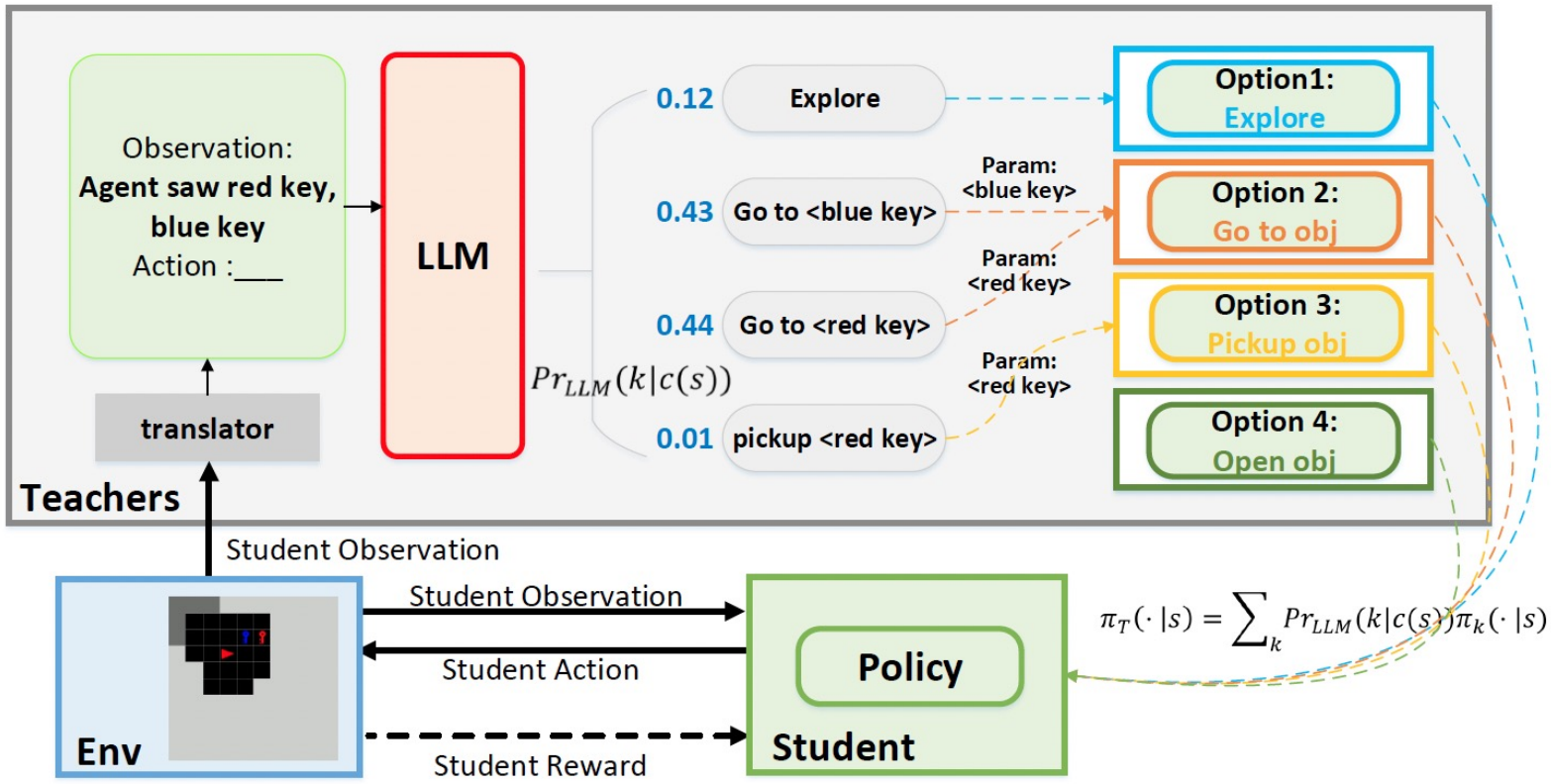


Figure 3: An illustration of the LLM4Teach framework using the MiniGrid environment

Pro:

- Policy distillation approach introduced
- Better sample efficiency
- Lightweight deployment

Challenge:

- A domain-specific Prompt
- Action spaces need to be customized for different tasks.
- Memoryless LLM

Methodology: Prompting engineering for LLM-based Teacher Agent

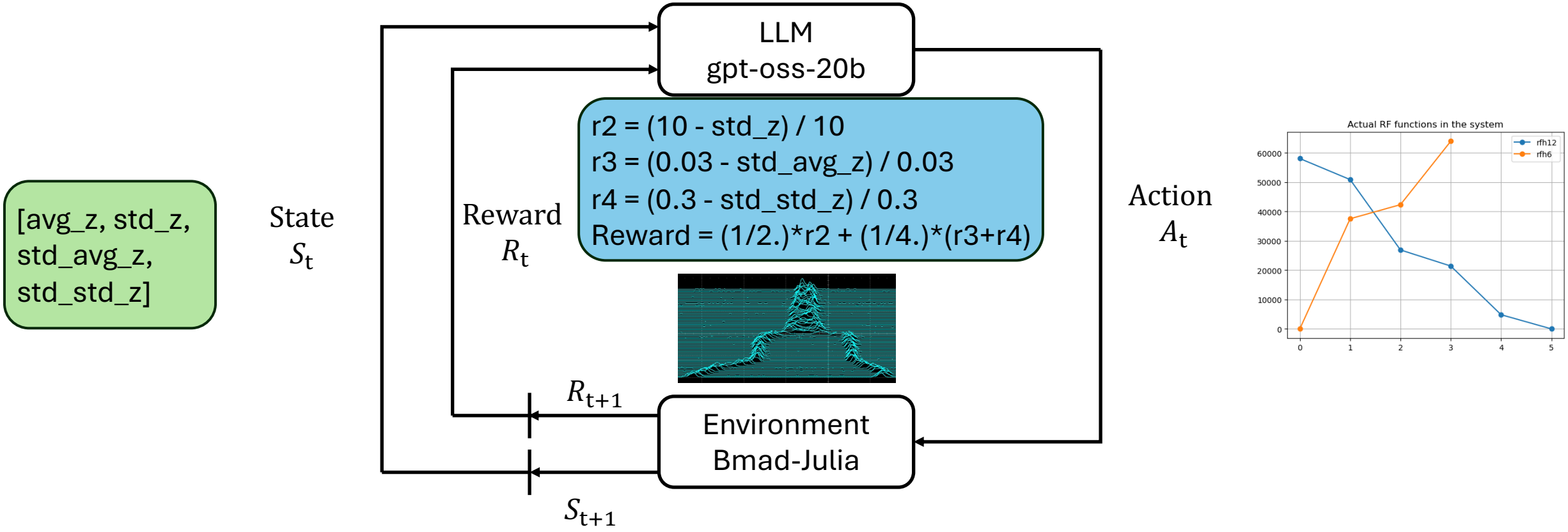


Figure 4 : Architecture of LLM-based Teacher Agent

- LLM model: gpt-oss-20b
- Retrieval-Augmented Prompting
  - Retrieve the top-k best reward examples from the historical JSON dataset
  - Insert the retrieved examples into the prompt as few-shot demonstrations

Prompting engineering

You are an expert RF control engineer and beam dynamics specialist for a particle accelerator.
ROLE: Act as a control policy that outputs optimal 6D RF voltage setpoint (a single action) for the current observation
PROBLEM TYPE: Stateless one-step contextual bandit optimization.
SYSTEM PHYSICS
RF voltages hold the beam in bunches. During a 2-to-1 bunch merge:
Harmonic-12 cavity voltage should be lower (ramping down behavior).
Harmonic-6 cavity voltage should be higher (ramping up behavior).
INPUTS: Current observation = [avg_z: Average of the bunch's centroid position,
std_z: Average of the bunch's bunch length,
std_avg_z: Standard deviation of the centroid,
std_std_z : Standard deviation of the bunch length]
OBJECTIVE: Maximize reward and reduce emittance
ACTION SPACE
6 RF voltages:
a1..a4 correspond to harmonic-12 cavity (ramph12)
a5..a6 correspond to harmonic-6 cavity (ramph6)
HARD LIMITS (MUST SATISFY)
ramph12 (a1..a4) must be within [-0.01, 60]
ramph6 (a5..a6) must be within [-0.01, 65]
Here are the examples of successful actions

JSON

Example: 0:
Observation: [-33.588349377969145, 2.183105757931912, 0.001996334187239741, 0.0025241244549036976]
Action: [40.01926739, 30.13174736, 19.0566535, 12.93060341, 43.85554969, 45.69579244]**Reward:** 0.8721051568306535

Example 1:
Observation: [-33.62725577365994, 2.294466731191495, 0.0014065788927963506, 0.004845884659779081]
Action: [59.007528, 46.0213956, 41.58774193, 9.88100161, 23.59839812, 60.84374519]**Reward:** 0.8695169354506398

Example 2:
Observation: [-33.57830239758312, 2.4663924056042648, 0.0011882741248405377, 0.006829543368550195]
Action: [52.64852077, 17.61819792, 8.67058777, 3.42904239, 28.43072697, 32.80926472]**Reward:** 0.8610868092056572

Example 3:
Observation: [-33.611278993627636, 2.5502033452380637, 0.002285440418731436, 0.0010411261658457294]
Action: [57.11395259, 45.51093104, 38.86681534, 20.37089518, 0.89646836, 26.34306238]**Reward:** 0.8525768907771301

Example 4:
Observation: [-33.68973951890146, 2.5653173748184335, 0.002689119379596372, 0.007705272733676993]
Action: [50.86905917, 34.85699095, 29.89203202, 3.58212186, 33.52404458, 42.36240031]**Reward:** 0.8429037424843778

Methodology: LLM-informed RL

r2 = (10 - std_z) / 10
r3 = (0.03 - std_avg_z) / 0.03
r4 = (0.3 - std_std_z) / 0.3
Reward = (1/2.)*r2 + (1/4.)*(r3+r4)

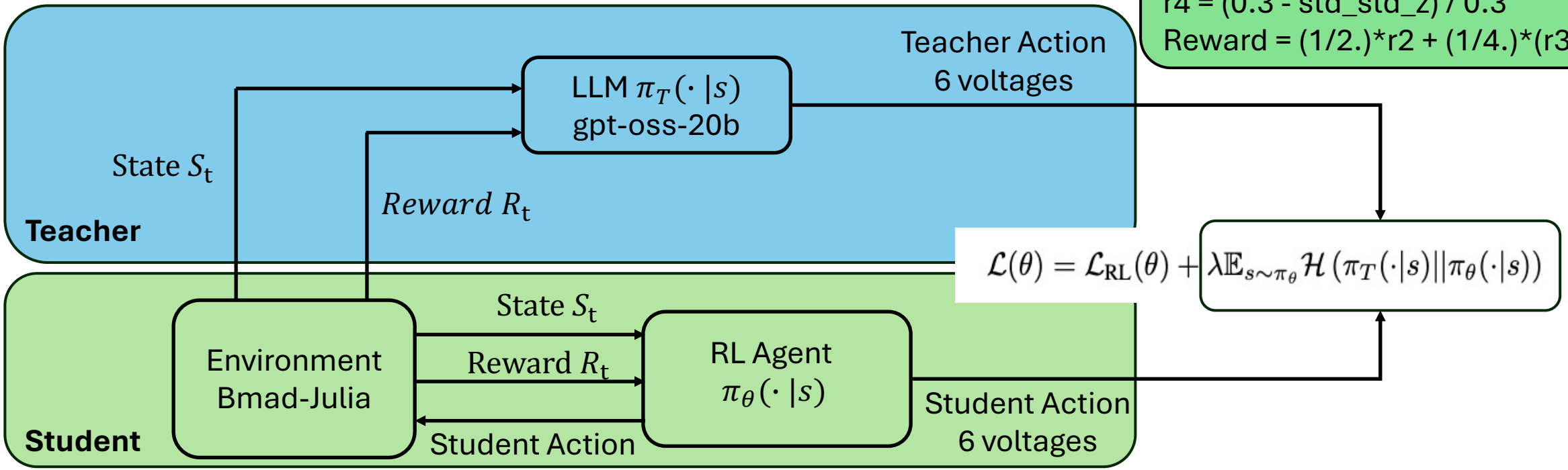
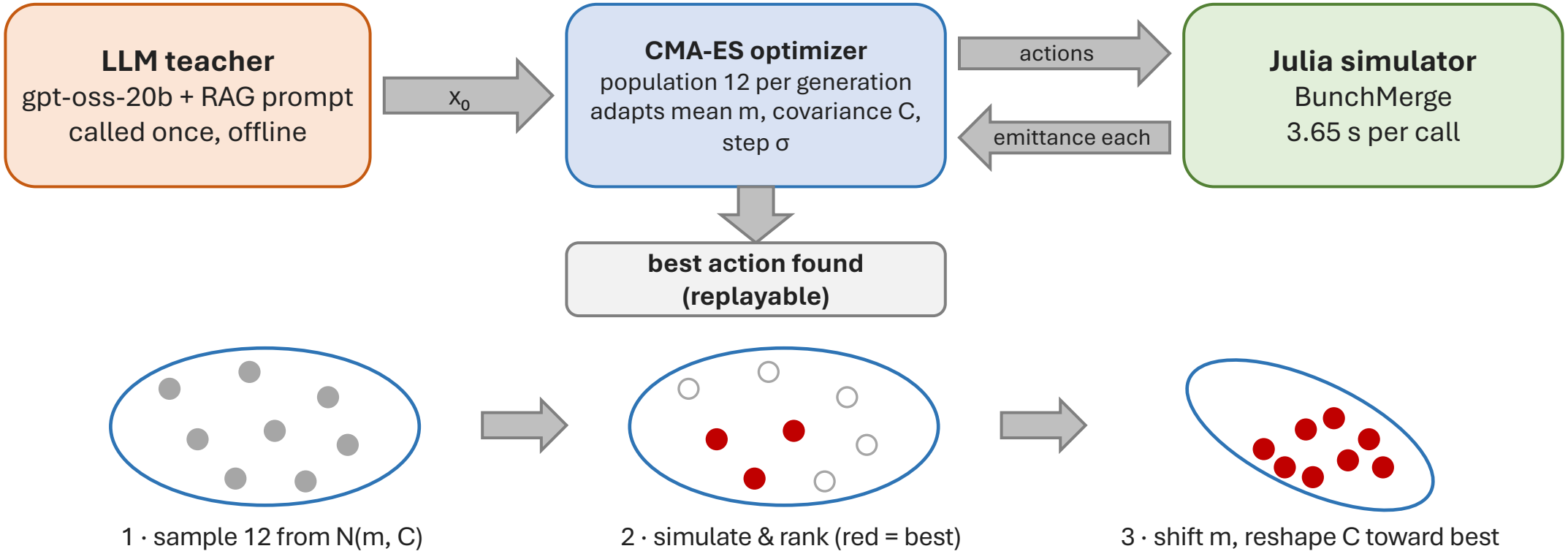


Figure 5 : Architecture of the Framework

- LLM model: gpt-oss-20b
- Retrieval-Augmented Prompting
  - Retrieve the top-k best reward examples from the historical JSON dataset
  - Insert the retrieved examples into the prompt as few-shot demonstrations

Methodology: LLM-informed Covariance Matrix Adaptation Evolution Strategy (CMA-ES)



- One offline LLM call supplies the start point x_0 — the teacher-best voltages and timings; the LLM never enters the loop.
- Each generation: sample 12 actions from a Gaussian, simulate each, re-center and reshape toward the best — no gradients.
- The warm start turns 12-d search from 1/3 into 4/4 reliable within the 1,200-evaluation budget.

Experiment settings

- Four methods are compared, each judged by its deliverable:
  - Search methods deliver an action: LLM alone, LLM-informed CMA-ES
  - Learning methods deliver a policy: RL only (SAC), LLM-informed RL

- Two action spaces: 6-d, voltages with the operators' fixed timings; 12-d, plus six free break-point timings.

- Two initial-condition setups:
  - Training draw: one fixed nominal bunch; all training and optimization
  - Resampled bunches: train on 100 different particle draws and test on 30 unseen particle draws of the same beam

Experiment settings

Method — deliverable	6-d · training draw	6-d · 100-draw pool	12-d · training draw	12-d · 100-draw pool
LLM alone — action	500	700	500	700
LLM-informed CMA-ES — action	1,200	6,200	1,200	6,200
RL only — policy	96,480	5,025	5,025	5,025
LLM-informed RL — policy	40,200	5,025	5,025	5,025

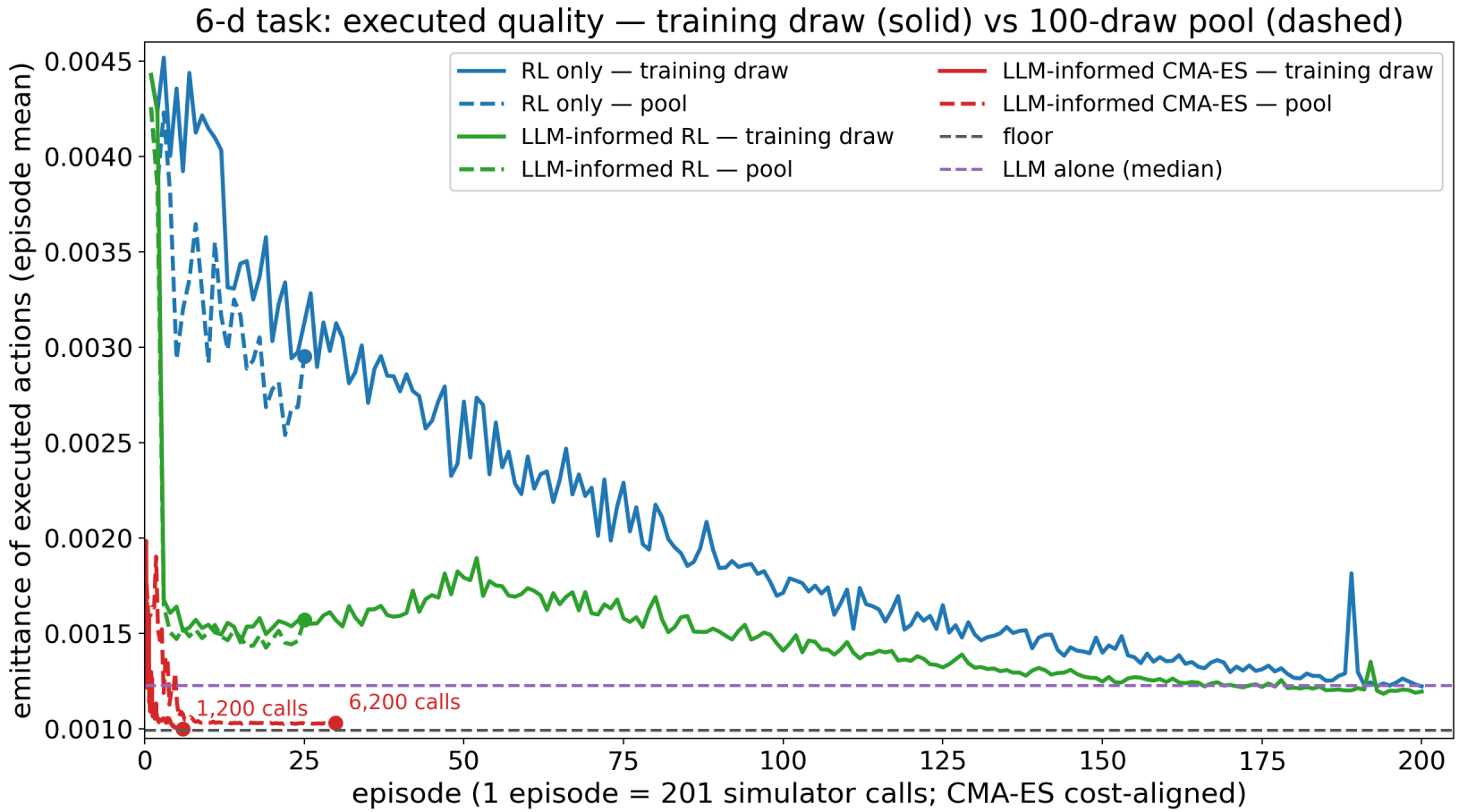
Results summary

Method — deliverable	6-d · training draw	6-d · 100-draw pool	12-d · training draw	12-d · 100-draw pool
LLM alone — action	1.112 ± 0.041 ×floor 1.061 ± 0.013	1.100 ± 0.044 [s] ×floor 1.049 ± 0.011	1.107 ± 0.047 ×floor 1.055 ± 0.012	1.105 ± 0.042 [s] ×floor 1.053 ± 0.011
LLM-informed CMA-ES — action	1.101 ± 0.049 ×floor 1.050 ± 0.012	1.088 ± 0.047 ×floor 1.037 ± 0.010	1.086 ± 0.046 ×floor 1.036 ± 0.008	**1.076 ± 0.046** ×floor 1.025 ± 0.008
RL only — policy	1.291 ± 0.046 [a] ×floor 1.231 ± 0.026	1.342 ± 0.032 ×floor 1.280 ± 0.030	6.36 ± 2.04 ×floor 6.05 ± 1.92	4.31 ± 1.21 ×floor 4.12 ± 1.18
LLM-informed RL — policy	1.175 ± 0.046 [ab] ×floor 1.120 ± 0.019	1.190 ± 0.058 [b] ×floor 1.134 ± 0.030	1.332 ± 0.046 ×floor 1.271 ± 0.034	1.288 ± 0.040 ×floor 1.228 ± 0.030

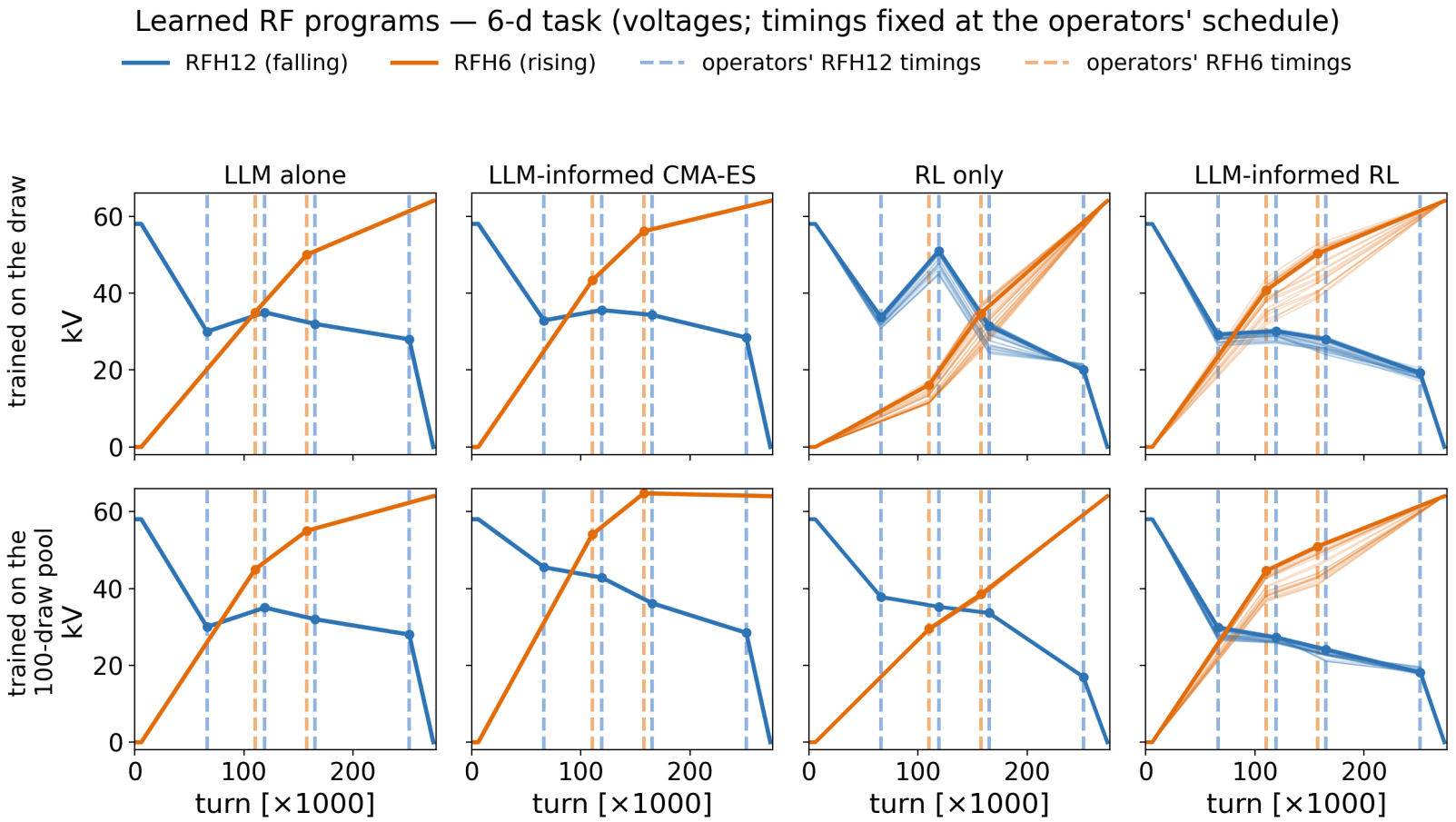
Every delivered action beats every learned policy.

Action-based method selects the best action over the 100 training samples and applies it on the testing samples

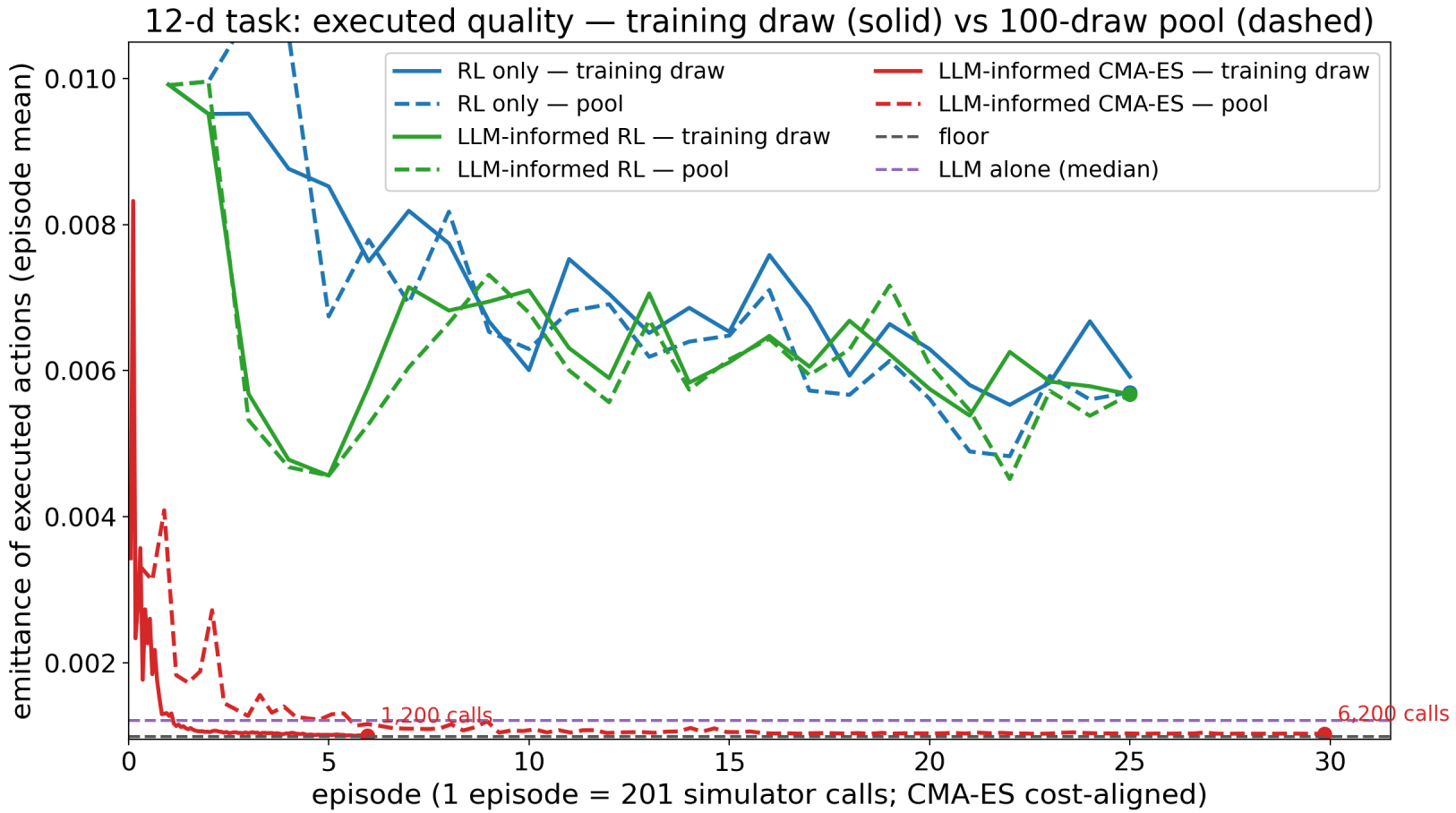
Result: executed quality during training — 6-d



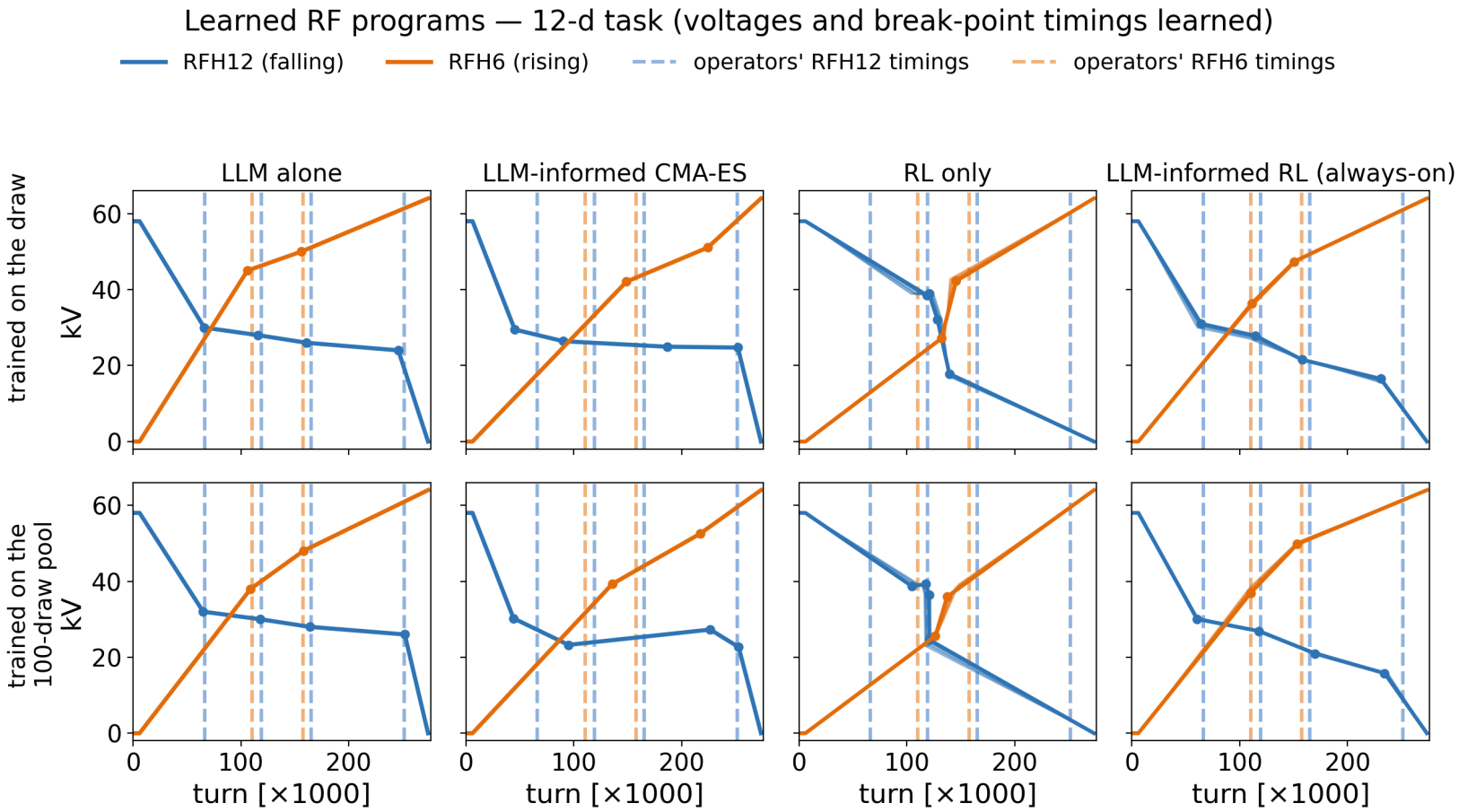
Result - 6d


Learned RF programs — 6-d task (voltages; timings fixed at the operators' schedule)
RFH12 (falling)
RFH6 (rising)
operators' RFH12 timings
operators' RFH6 timings
LLM alone
LLM-informed CMA-ES
RL only
LLM-informed RL
trained on the draw
trained on the 100-draw pool
kV
0
20
40
60
0
100
200
turn [×1000]

Result: executed quality during training — 12-d



Result - 12d


Learned RF programs — 12-d task (voltages and break-point timings learned)
RFH12 (falling)
RFH6 (rising)
operators' RFH12 timings
operators' RFH6 timings
LLM alone
LLM-informed CMA-ES
RL only
LLM-informed RL (always-on)
trained on the draw
trained on the 100-draw pool
kV
0
20
40
60
0
100
200
turn [×1000]

Conclusion

- The LLM's physics prior is real and dimension-independent.
- Delivered actions beat learned policies in every cell; 12-d policies fail at 4–6× the floor, and guided policies hold only while the teacher remains active.

Future work

- Shifted-beam (±25%) training and testing — after physicists' input on realistic bunch-length / momentum-spread ranges.
- Measured: with the teacher always on, 12-d policies hold at 1.23–1.27× — retained, and pool training finally helps (+4.3%) — but still short of deployable (≤1.20).
- Long-horizon 12-d training anchor, and confirming the beam species and simulator provenance with the collaboration.

Thank You

Q & A