

Tree-Structured Synergy of Large Language Models
and Bayesian Optimization for Efficient CASH

Beicheng Xu, Weitong Qian, Lingching Tung, Yupeng Lu, Bin Cui*

School of CS & Key Lab of High Confidence Software Technologies (MOE) & Beijing Key Laboratory of
Software and Hardware Cooperative Artificial Intelligence Systems, Peking University, Beijing, China.

✉ beichengxu@stu.pku.edu.cn

A shared search tree combines LLM semantic proposals with algorithm-specific Bayesian optimization.

01 Problem & setup

Why CASH needs structure

- CASH** jointly chooses an algorithm and its hyperparameters.
- BO** lacks reliable estimates early on. **LLMs** offer semantic priors, but flat histories mix unrelated algorithms.
- Search must balance **exploration and exploitation** both across algorithms and within each algorithm's space.

Preserve how each configuration
was improved.

How the two proposers cooperate

- PUCT** allocates evaluations across algorithm subtrees.
- Each algorithm maintains a separate **Gaussian-process surrogate**.
- BO** ranks random and local candidates by expected improvement.
- The **LLM** refines a selected configuration with explicit exploration or exploitation directives.
- Both proposers update the **same tree** and reuse each other's search history.

Evaluation protocol

104

AMLB datasets

300

evals / method / task

71 classification · 33 regression tasks
8 candidate algorithms per task type
8 BO, LLM, and hybrid baselines
60/20/20 train / validation / test split

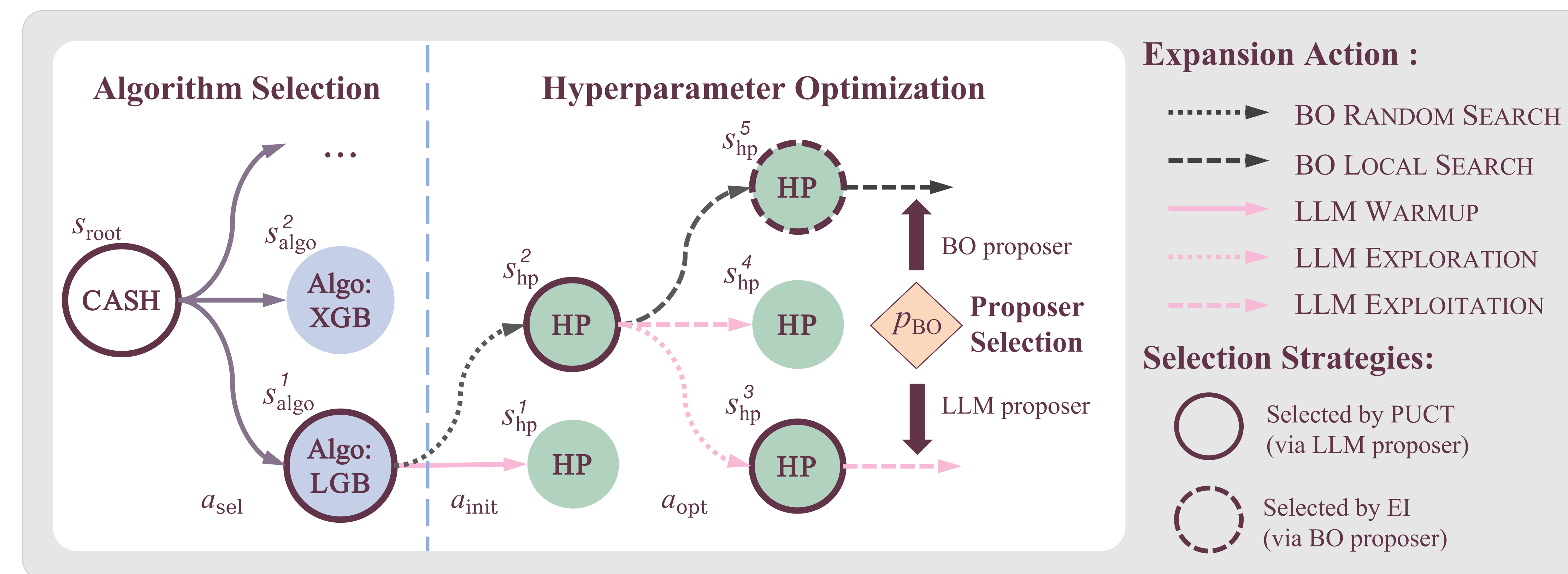
Evaluation-count budget; wall-clock costs vary.

Scope & limitations

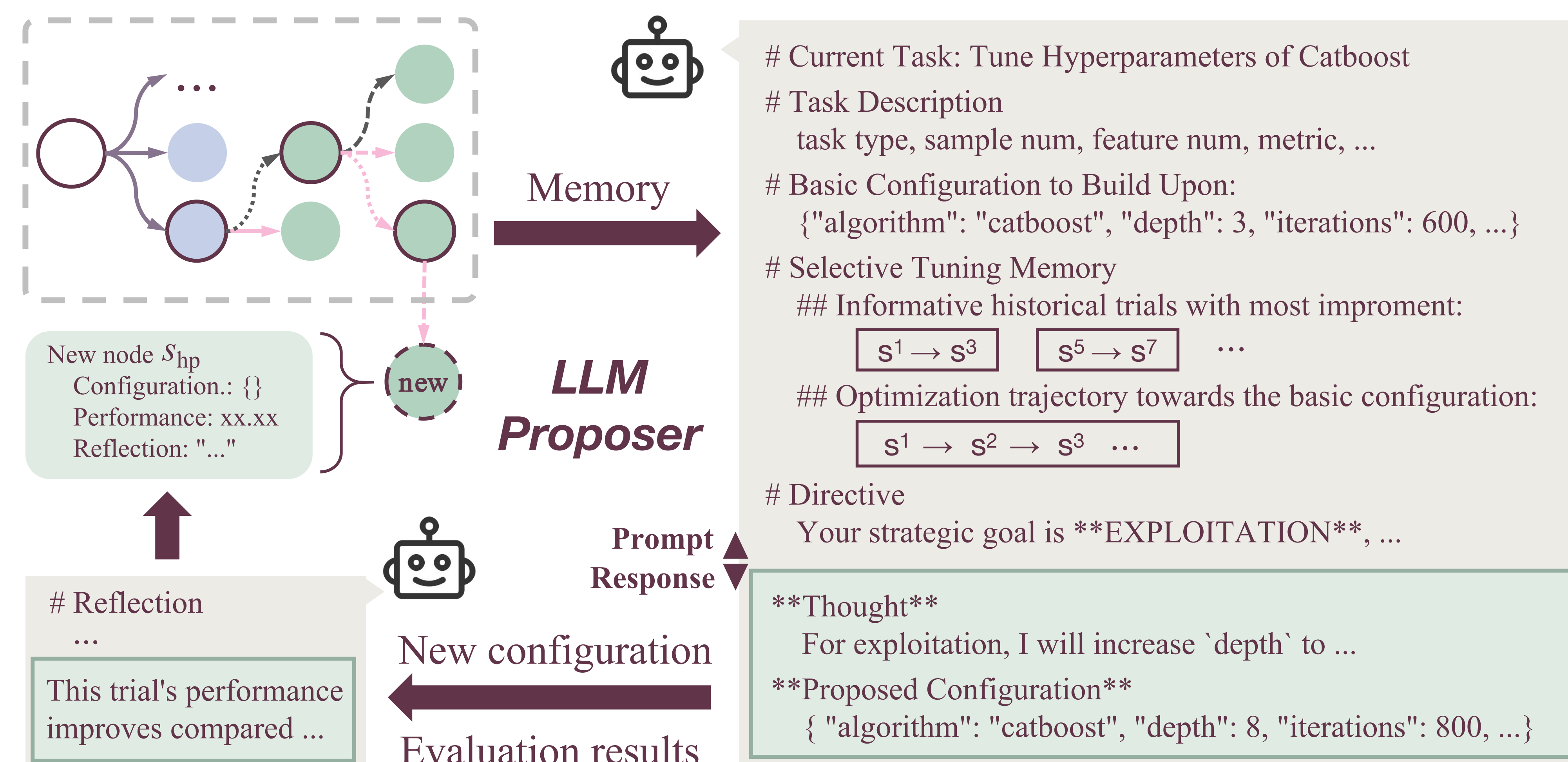
Noisy validation feedback can mislead search. Proposal quality also depends on the underlying LLM.

LLM backbone: GPT-4o-mini.

02 Method: a shared search state

Algorithm selection and configuration refinement share one tree.
BO and LLM proposals extend the same optimization trajectories.

Selective Tuning Memory

**Local:** the ancestral refinement path. **Global:** attempts balancing similarity and performance.**Reflection:** evaluation feedback becomes guidance; BO and warmup use rule-based summaries.

Adapt as the surrogate improves

Early: LLM semantic priors

Later: more BO proposals

Driven by surrogate ranking quality; selection remains probabilistic.

03 Evidence across 104 datasets

2.16

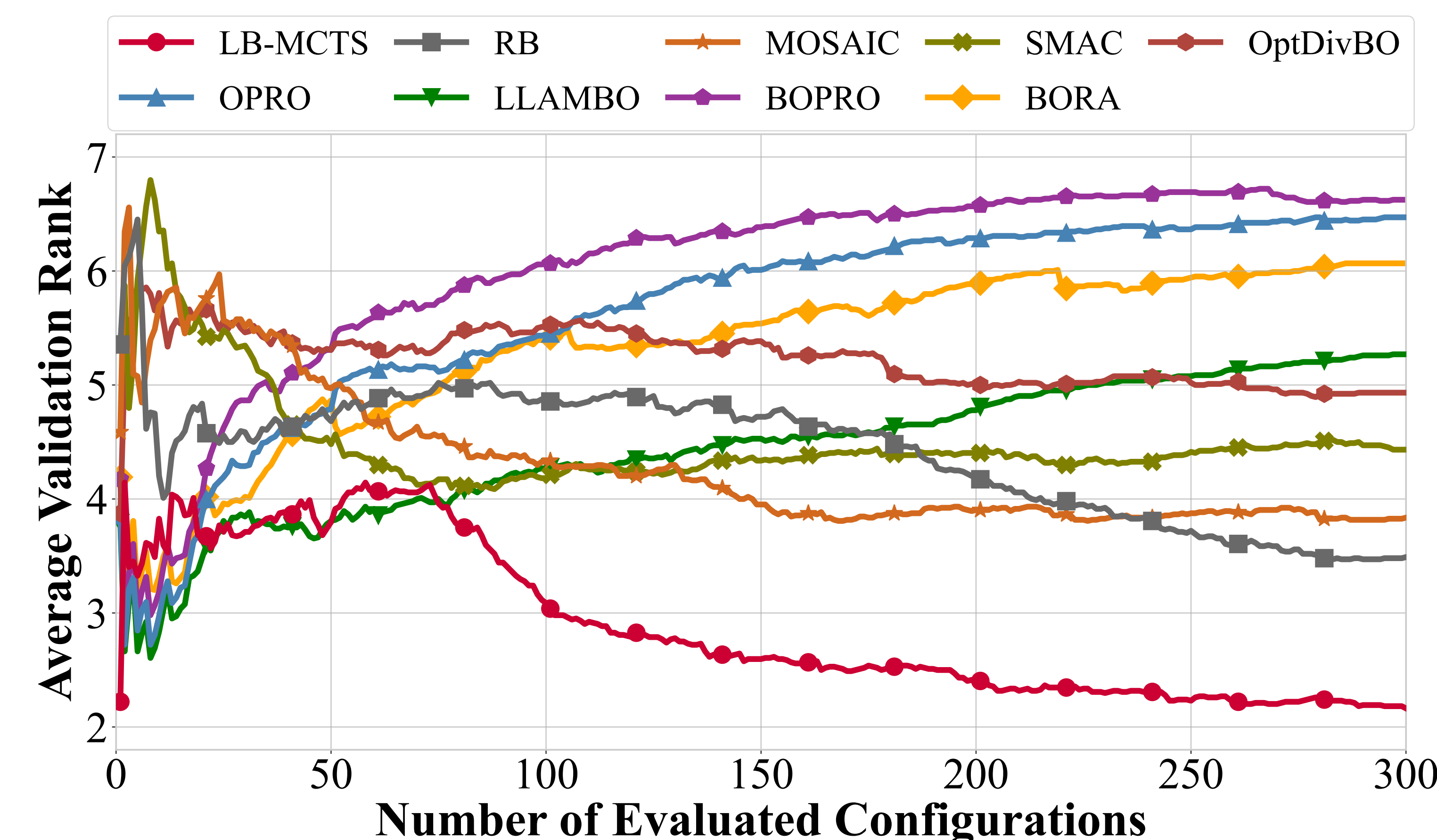
Best avg. validation rank

RB: 3.49 · lower is better

3.05

Best avg. test rank

validation-selected configs

Average rank of best-so-far validation performance.
104 datasets · 9 methods · 300 evaluations per method/task.

Tree structure matters

Separate ablation: 10 datasets · 4 variants

LB-MCTS

1.5

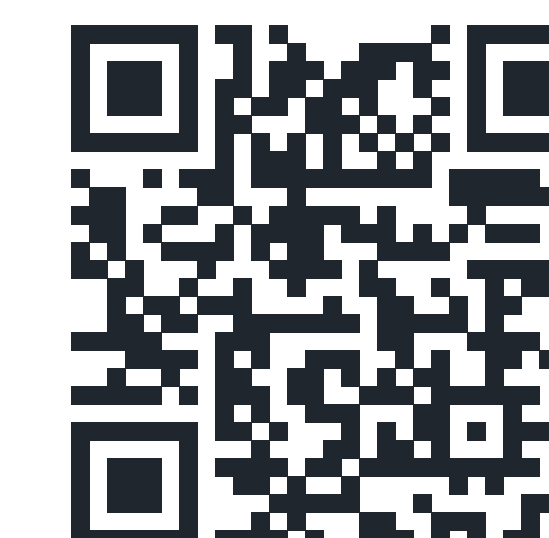
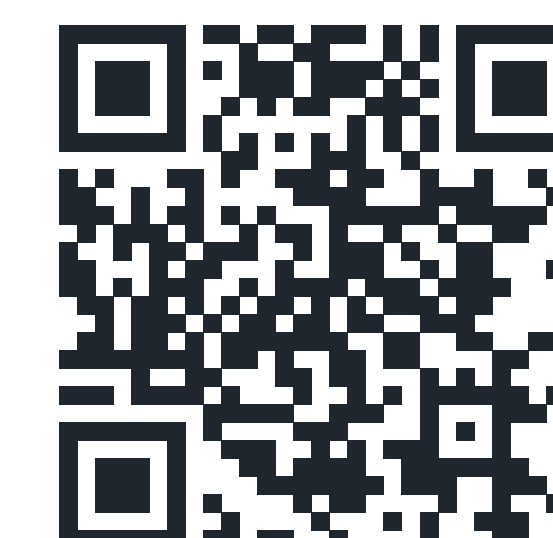
average rank ↓

LB-Flat

3.4

Takeaway

Preserve trajectories, retrieve relevant experience, and adapt proposer selection to surrogate reliability.

Read the paper
arXiv: 2601.12355Explore the code
PKU-DAIR/open-box