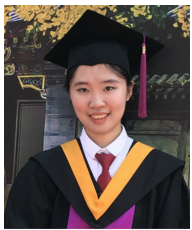
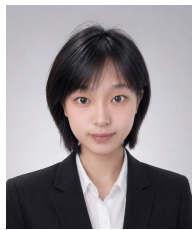
Efficient Cost-Aware LLM Evaluation via Bayesian Bandit Gittins Indices

Qian Xie[1] Yueli He[2] Nairen Cao[2]
[1]Cornell University [2]New York University

ICML 2026 Workshop DEMO Spotlight



Qian Xie



Yueli He

Nairen Cao

LLM evaluation searches for a good task-specific configuration

Goal: find the best configuration at far lower evaluation cost.

Model	Q1	Q2	Q3	...	avg performance
ChatGPT GPT-4o	✓	?	?	...	?
Claude 3.5 Sonnet	?	✓	?	...	?
Deepseek	✓	? Y_{ij}	?	...	?
Gemini 1.5 Pro	✗	?	✓	...	?
Llama 3.1-70B	?	✗	?	...	?
Mistral Large	?	?	✗	...	?
...	...	...	...	...	...

Costly evaluations motivate sparse, adaptive evaluation.

Problem setup

Config: model + prompt + temperature + decoding.
Target: highest average performance.

Evaluation cost

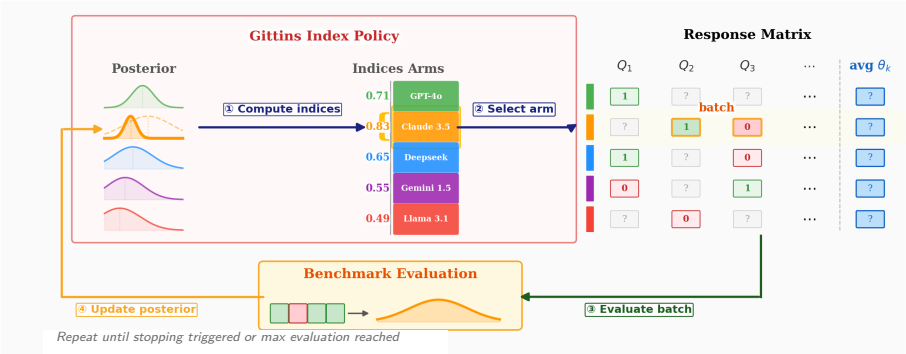
Cost: token/API price, latency, compute, grading.
Brute force fills the full matrix.

Existing work

BanditEval (Zhou et al. 2025): UCB-E (frequentist bandit), LRF (matrix correlation); both cost-unaware.

Bayesian Bandit Gittins Policy

Gittins policy comes with both an arm-pull rule and a stopping rule.



Bandit model

Arm: LLM configuration.
Pull: evaluations of one or a batch of unevaluated examples.

Policy

Top arm incomplete: evaluate.
Otherwise: stop.

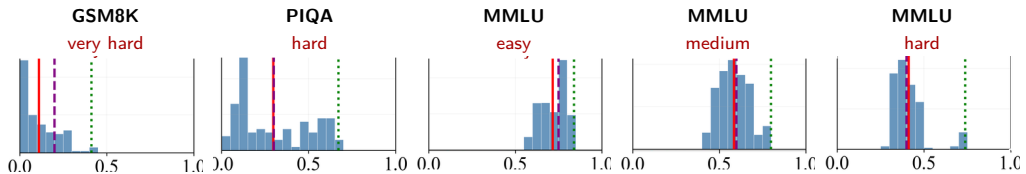
Experimental cap

Plots use a 10% maximum; stopping may occur earlier.

Choice of Prior

Gittins-G is the general default; Gittins-S is data-specific.

Gittins-S prior buckets



Gittins-G

General Default: shared general prior.

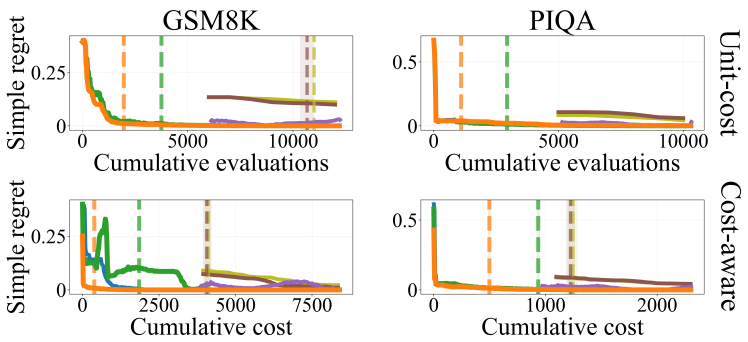
Gittins-S

Data-Specific: easy, medium, hard, and very hard buckets.

Gittins-S uses offline response matrices to set data-specific priors.

GSM8K and PIQA Results

Many examples amplify savings: 1000 examples in each benchmark.



Orange/green = Gittins variants; blue/purple = UCB-E/LRF; brown/olive = Bayesian optimization; dashed = mean stop.

Partial evaluations

Lower regret than full-evaluation Bayesian optimization.

Gittins-S

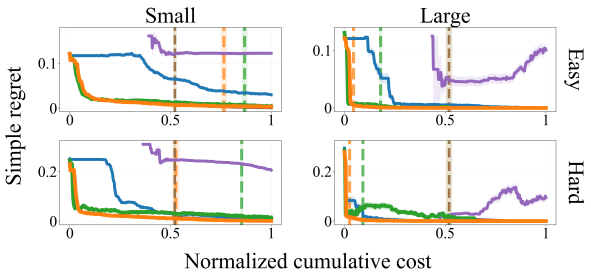
Data-specific priors improve over Gittins-G.

Early stopping

Stops much earlier before the 10% cap.

MMLU Results

Many candidates widen method gaps: DOVE provides 1500 model-template candidates for each MMLU subject.



Cost-aware MMLU aggregate results; dashed lines show mean stopping locations.

Large candidates

1500 model-template arms per MMLU subject.

Cost-aware helps

Gittins is significantly stronger than cost-unaware UCB-E/LRF.

Stopping helps

Particularly on large-example datasets.