

Adaptive Learn-then-Test: Statistically Valid and Efficient Hyperparameter Selection

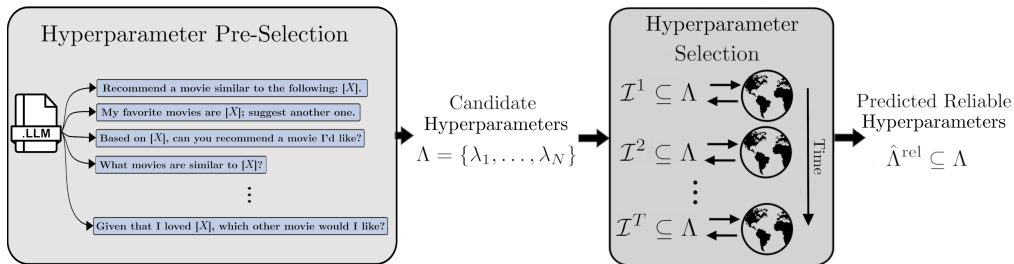
Matteo Zecchin, Sangwoo Park, Osvaldo Simeone

King's College London



Hyperparameter Selection in the Scaling-Centric Era

- Hyperparameter selection can be formulated as a bandit problem over a discrete space of pre-selected configurations.
- Examples: prompts for fine-tuning, architectural scaling choices, or policy parameters in reinforcement learning

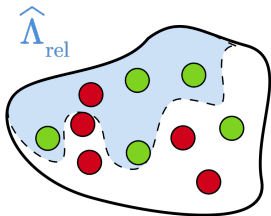


Statistical Guarantees

- **Goal:** Select a subset Λ^{rel} containing as many reliable hyperparameters as possible, while controlling the number of unreliable choices.
- Two common statistical guarantees are the family-wise error rate (FWER) and the false discovery rate (FDR).

● reliable hyperparams Λ^{rel}

● unreliable hyperparams Λ^{unrel}

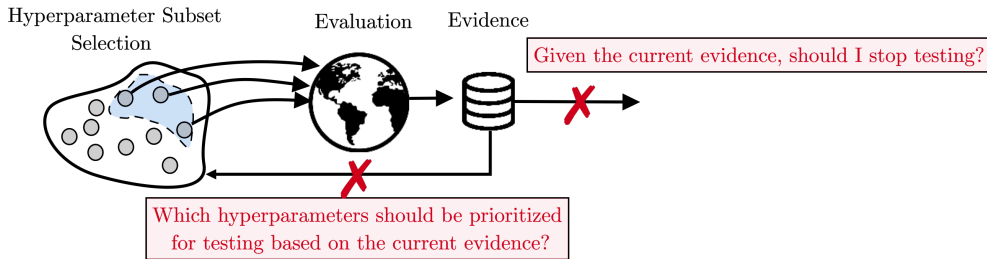


$$\text{FWER}(\hat{\Lambda}^{\text{rel}}) := \Pr \left[|\Lambda^{\text{unrel}} \cap \hat{\Lambda}^{\text{rel}}| \geq 1 \right] \leq \delta$$

$$\text{FDR}(\hat{\Lambda}^{\text{rel}}) := \mathbb{E} \left[\frac{|\Lambda^{\text{unrel}} \cap \hat{\Lambda}^{\text{rel}}|}{|\hat{\Lambda}^{\text{rel}}|} \mid |\hat{\Lambda}^{\text{rel}}| \geq 1 \right] \leq \delta$$

Lean-Then-Test

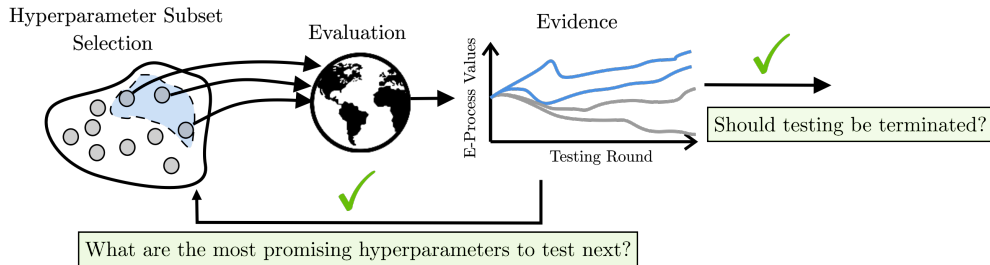
- Learn-then-Test (LTT) performs statistically valid hyperparameter selection based on p-values computed from the collected evidence¹.
- Adaptive evaluation and flexible stopping rules are not possible using p-value-based testing (p-hacking).



¹Angelopoulos et al., "Learn then test: Calibrating predictive algorithms to achieve risk control".

Adaptive Learn-then-Test

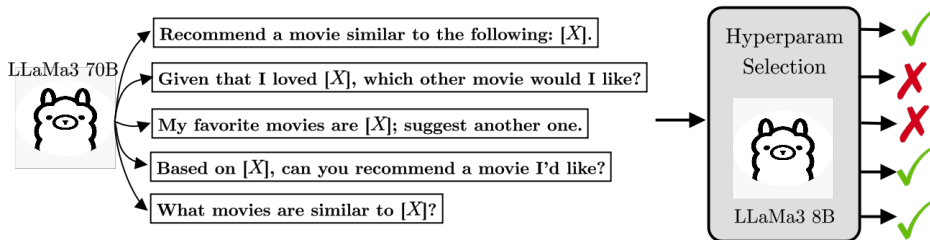
- To improve the efficiency of hyperparameter selection, we propose Adaptive Learn-then-Test (aLTT), a sequential hyperparameter selection algorithm based on e-processes².
- aLTT can decide whether to stop or continue testing, and it can select the subset of hyperparameters to test next based on the collected evidence.



²Xu and Ramdas, "Online multiple testing with e-values".

Simulation: Automated Prompt Engineering

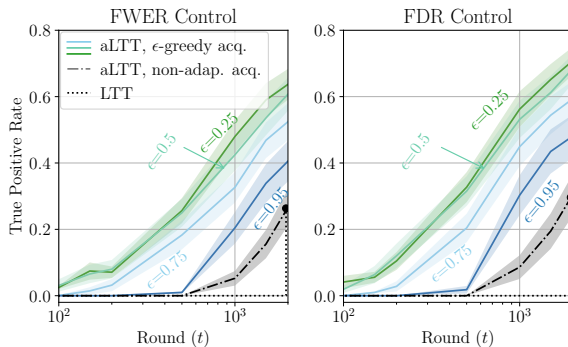
- **Goal:** Find high-quality prompt templates from a set of LLM-generated prompts.
- Prompts are generated using the LLaMA 3.3 70B Instruct model and evaluated using the LLaMA 3 8B Instruct model.
- Tasks are sampled from the Instruction Induction dataset³.



³Honovich et al., "Instruction induction: From few examples to natural language task descriptions".

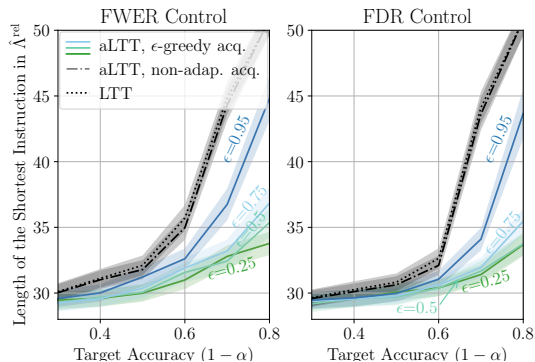
Simulation: Automated Prompt Engineering

- We compare LTT against aLTT with an ϵ -greedy acquisition strategy.
- Adaptive evaluation allows aLTT to discover more models using fewer LLM calls.




Simulation: Automated Prompt Engineering


- Hyperparameters can then be post-selected from $\hat{\Lambda}^{\text{rel}}$ to maximize some alternative metric.
- For example, one could choose the shortest prompt in $\hat{\Lambda}^{\text{rel}}$.




Conclusion

- We have proposed aLTT, a statistically valid hyperparameter selection procedure based on e-value testing.
- In many applications, aLTT substantially reduces the evaluation cost compared to non-adaptive strategies.

 Angelopoulos, Anastasios N et al. “Learn then test: Calibrating predictive algorithms to achieve risk control”. In: *The Annals of Applied Statistics* 19.2 (2025), pp. 1641–1662.

 Honovich, Or et al. “Instruction induction: From few examples to natural language task descriptions”. In: *arXiv preprint arXiv:2205.10782* (2022).

 Xu, Ziyu and Aaditya Ramdas. “Online multiple testing with e-values”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3997–4005.