Replacing AI Agent Leaderboards with Pareto Curve Evaluations

Sayash Kapoor, Benedikt Stroebl, Zachary Siegel, Arvind Narayanan

The issue of unbounded cost

Accuracy can be artificially inflated through sampling. As demonstrated by *AlphaCode*, repeated sampling can inflate accuracy, which gives rise to an unbounded cost issue that prevents us from attributing which ideas actually led to meaningful improvements.

Only by evaluating cost and time in addition to accuracy, we can gain a more comprehensive understanding of an AI agent's true capabilities and practicality for downstream users.

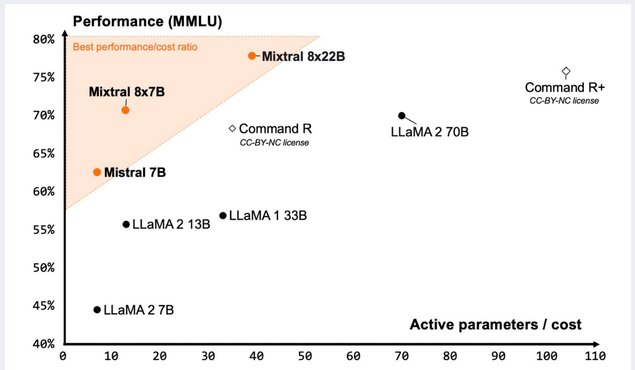

Figure 1: Measure of the performance (MMLU) versus inference budget tradeoff (number of active parameters). Mistral 7B, Mixtral 8x7B and Mixtral 8x22B all belong to a family of highly efficient models compared to the other open models.

Figure 1. Misleading proxy for cost. Substituting active parameters as a proxy for cost is misleading. Everyone can choose a proxy that makes their model look good. *Source: Mistral*.

Model vs. Downstream evaluation. While metrics like active p arameters or compute time can be useful for scientific research , they are inadequate for downstream evaluation. Downstream users, care about the actual financial cost and time efficiency of running the model, not measures of its complexity.

Cost as the true construct of interest. Proxies like active parameters can be misleading, as they may not reflect the real-world expense or resource consumption associated with running the model (see Figure 1).

The accuracy-cost tradeoff on HumanEval

We compared the performance of top agents like LDB, LATS, and Reflexion with simpler baselines like repeated sampling with increasing temperature and escalating model size upon failure. Surprisingly, the complex agents did not outperform our baselines in accuracy, despite incurring significantly higher costs up to 50 times more. Our findings suggest that simple strategies can be equally effective while being more cost efficient, highlighting the need to consider cost as a crucial metric in evaluating agents.

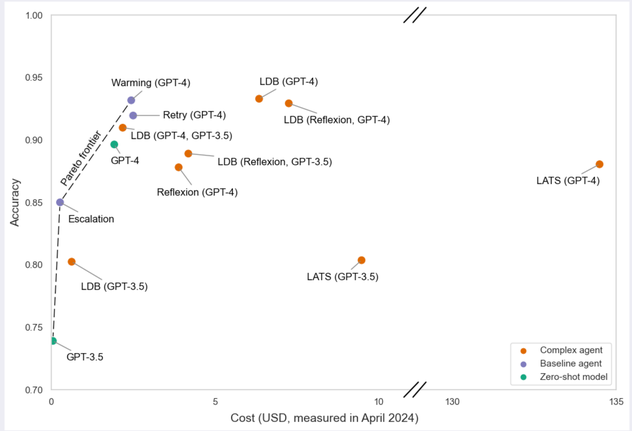

Figure 2. Acc vs. Cost.

Our simple baselines offer Pareto improvements over existing agent architectures.

Where results for LDB have two models/agents in parenthesis, they indicate the language model or agent used to generate the code, followed by the language model used to debug the code. Where they have just one, the same model was used for both.

Our baseline agents

Retry

We repeatedly invoke a model with the temperature set to zero, up to five times.

Warming

We gradually increase the temperature of the underlying model with each run, from 0 to 0.5.

Escalation

We start with a cheap model (Llama-3 8B) and escalate to more expensive models (GPT-3.5, Llama-3 70B, GPT-4).

Reproducibility issues and lack of standardization

Evaluation discrepancies and transparency issues. Our research identified significant *discrepancies in the accuracy* of LATS and LDB agents on HumanEval, *varied evaluation subsets* by Reflexion and LATS, and critical *missing details in implementation*, emphasizing the **need for standardized protocols** and greater transparency to ensure reproducibility and validate results.

Next steps and future plans

Expand Benchmarks. We will include more challenging, and uncontaminated coding benchmarks alongside non-coding benchmarks. We will build on preliminary results for HotPotQA and NovelQA, and will add additional benchmarks to ensure comprehensive evaluation across different domains.

Joint optimization. We will explore was to find points on the pareto frontier programmatically by jointly optimizing for cost and accuracy.

Refine Cost Calculations. We plan to develop robust strategies to mitigate the instability and downsides of dollar cost calculations in benchmark design. One idea is to give downstream users the tools to choose cost levels themselves.

Uncertainty quantification and reliability. A significant focus will be on quantifying uncertainty to better understand the sources of variance such as stochasticity and dataset sampling, which influence performance claims.

Questions for you

1. **Which benchmark categories are interesting?** We are planning to mainly focus on coding and QA benchmarks from different domains but would appreciate any feedback or comments on which other categories might be interesting to look into.
2. **Are there other categories of evaluations beyond specific benchmarks we should run additional experiments on?**
3. **What sources of uncertainty should we care about?** Stochasticity of models and dataset sampling are two sources we account for. Are there other sources of variance that seem foundational to our work?