Selection-induced optimism in a synthetic candidate evaluation

Anonymous experiment report

Abstract

This study investigates the problem of selection-induced optimism when evaluating candidate systems. In our synthetic experimental setup, we assume a population where all candidate systems share a fixed true accuracy of $p = 0.6$. We present the core finding that the estimated validation accuracy increases monotonically with the number of candidates in the selection pool. This leads to an optimistic gap relative to the performance measured on an independent test set. Our results demonstrate this is a known statistical effect arising from the act of selection, rather than an artifact of novel algorithms or production systems.

1 Introduction

1.1 Motivation

In many evaluation protocols, systems are selected based on performance measured on a validation set which is distinct from the test set. While this separation is intended to ensure that the test set provides an unbiased estimate of generalization, a subtle statistical issue often arises when comparing multiple candidates. When selecting the "best" system from a large pool, the process itself introduces a bias where the estimated performance exceeds the true underlying capability of the population Zhang et al. [2026].

1.2 Problem Statement

We formalize a scenario involving N candidate systems, where each possesses an identical true accuracy p. The research question is whether the act of selecting the system with the highest validation score introduces a divergence between this estimated performance and the true population accuracy. This phenomenon is theoretically contextualized by the "Winner's Curse" Zhang et al. [2026], where the winner of a competition among correlated or identically distributed samples tends to exhibit an inflated estimate of their true quality relative to the mean of the group.

2 Methodology

2.1 Experimental Setup

To isolate the effect of selection pressure, we design a synthetic experiment where all candidates are drawn from the same distribution. We define a population parameter $p = 0.6$. For each candidate i and question j, we generate independent binary outcomes. Specifically, for any given candidate system, its correctness on a specific question is modeled as an independent Bernoulli trial with success probability p.

The selection mechanism operates as follows: we measure the aggregate performance of N candidates on a validation set of size n_{val}. The system deemed "best" is selected using $\text{argmax}_i(\text{validation_score}_i)$, employing first-index tie-breaking to ensure deterministic outcomes. We assume that the candidate correctness is independent across both candidates and questions, and crucially, that the test set is strictly independent of the validation set used for selection.

2.2 Metric Definitions

We define three primary metrics:

1. **Validation Mean Accuracy ($\hat{p}_{val}$):** The average accuracy computed over the n_{val} questions on the chosen candidate's performance relative to the validation set.

2. **Test Mean Accuracy ($\hat{p}_{test}$):** The average accuracy computed over a disjoint set of n_{test} questions, providing an unbiased estimate of the true accuracy p.

3. **The Gap (Δ):** Defined as $\Delta = \hat{p}_{val} - \hat{p}_{test}$. Note that this difference is measured in percentage points. A positive Δ indicates that the validation score overestimates the test performance.

3 Experiments

3.1 Implementation Details

All experiments were implemented using Python 3.11.4 and NumPy 2.4.6 **??**. We used a random seed of 20261116 for reproducibility. The simulation was repeated 5000 times to ensure statistical stability. It is important to clarify that no Large Language Models (LLMs) or production systems were involved in this study; the data is purely synthetic.

3.2 Results Analysis

We evaluated the performance gap across four candidate counts: 1, 8, 40, and 328. As shown in Table 1, the validation mean accuracy rises significantly as the number of candidates increases, while the test mean remains stable near the true population value $p = 0.6$.

Table 1: Validation and Test Mean Accuracy across different candidate counts.

Candidates (N)	Validation Mean (%)	Test Mean (%)	Gap (%)
1	59.96	59.97	-0.01
8	66.88	60.00	6.88
40	70.38	59.98	10.40
328	73.87	60.00	13.87

The trend demonstrates that the gap increases with the candidate count due to selection pressure. With $N = 1$, there is no selection effect, and the validation mean closely matches the test mean (gap ≈ 0). As N grows to 328, the probability of selecting a candidate with an unusually high validation score by chance increases, resulting in a gap of approximately 14 percentage points.

4 Discussion

4.1 Interpretation of Results

The observed "gap" is purely a statistical artifact of selection. It is crucial to reiterate that the true accuracy remains constant ($p = 0.6$) across all configurations; only the *estimate* derived from the validation set changes due to the act of picking the maximum. The independence assumption between candidates and test questions holds, yet the dependence introduced by the argmax operation on the validation set distorts the estimate.

4.2 Limitations

We acknowledge the synthetic nature of this experiment; no real harnesses were modified, and no financial costs or wall-clock throughput data were measured. Furthermore, the candidate count of 328 is arbitrary in this context and not specific to any particular self-harness algorithm. These results imply that reporting standards in AI research must account for the number of candidates considered during selection to avoid overconfident claims about a single model's superiority.

5 Conclusion

5.1 Summary

We have summarized the finding that large candidate pools induce significant optimism in validation scores, creating a divergence from true generalization capability. We advocate for larger, independent test sets when comparing multiple candidates and explicitly accounting for selection bias in performance reporting.

References

Hangfan Zhang et al. Self-harness: Harnesses that improve themselves, 2026.