Efficient Evaluation and Error Pattern Discovery for Blackbox AI Systems

Maxim Rabinovich[1] Harvineet Singh[2] **Aman Sinha**[1]

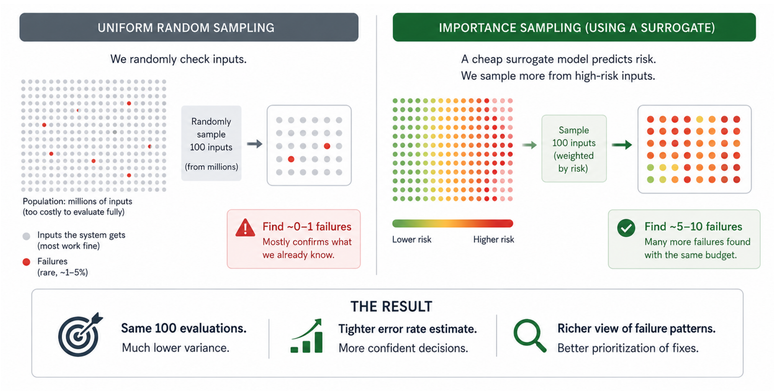
1 Fellow, The Fannie & John Hertz Foundation 2 Qualified Health, PBC | Correspondence: amans@alumni.stanford.edu

NEURAL INFORMATION PROCESSING SYSTEMS

Importance sampling with cheap surrogates for frontier models helps evaluate and find failure patterns cheaply.

Problem

- Evaluating and improving frontier AI systems needs expensive benchmarking: cost and time of agentic evaluations & human review.
- Developers need both an error rate and the failure patterns behind it.
- Prior work assumes calibration data, task/response embeddings, or a fixed benchmark.



Main idea: a cheap surrogate predicts risk, so the same budget finds more failures.

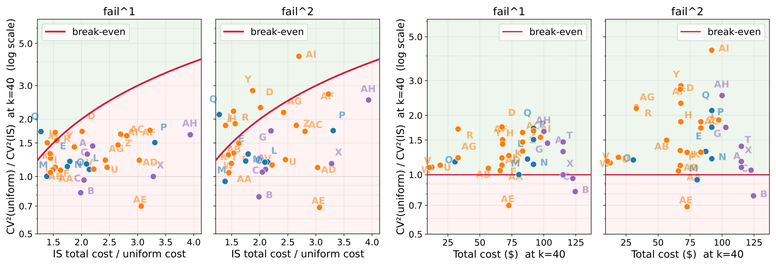
Method: importance sampling by surrogate

$$\min_{p} \sum_i \frac{s^2(S(x_i))}{p_i} \quad \text{s.t. } d_i \le p_i \le 1, \quad \sum_i p_i = \alpha N$$

$$p_i^\star = \mathrm{Clip}\big(\lambda\, s(S(x_i)),\, d_i,\, 1\big)$$

- Sampling probabilities depend only on scalar surrogate evaluations: no embeddings, kernels, or calibration set.
- Pairs with Horvitz-Thompson, post-stratification, and PPI estimators.

Cost: when does the surrogate pay off?



7 / 35

surrogate-target pairs beat break-even even when paying for the surrogate

32 / 35

beat break-even when surrogate outputs are already cached

Efficiency gain vs cost, relative to uniform on TerminalBench2.0

Takeaways

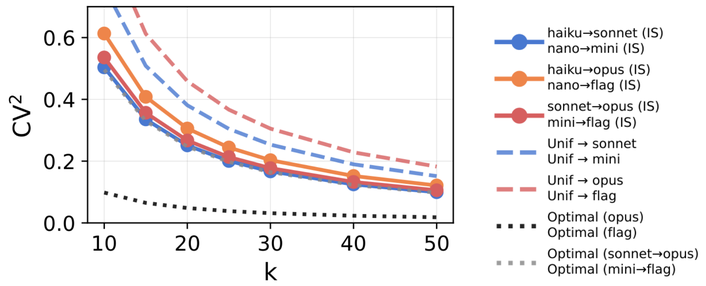
1. Calibration-free IS needs only black-box surrogate evaluations and composes with post-stratification and PPI.
2. Same-family surrogates cut variance up to 1.8x and surface more errors for failure analysis.
3. Gains are bounded by surrogate quality; fresh surrogates must be cheap and informative.

Theory: when does it help?

- Gain over uniform sampling is at least 1, and is largest when errors concentrate on few surrogate values.
- More informative surrogates never hurt (Blackwell ordering), so surrogates can be ranked from evaluations alone.
- Achievable gain is bounded by 1/π, the surrogate selectivity.

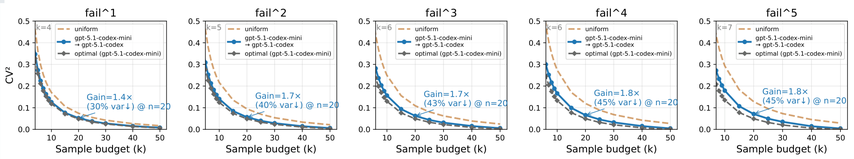
$$\mathrm{Gain}^\star(S) = \frac{\mathbb{E}[v(G)]}{\big(\mathbb{E}\sqrt{v(G)}\big)^2} \ge 1, \qquad p^\star(u) \propto \sqrt{v(u)}$$

Estimation: lower variance, same budget



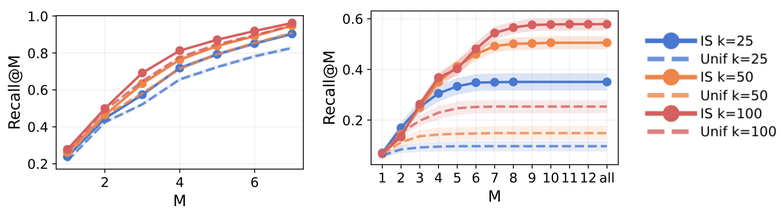
MMLU-Pro, CV² of the error-rate estimate vs. budget k (solid: IS, dashed: uniform).

- 1.5x (OpenAI) to 1.8x (Anthropic) variance reduction, within 10% of the surrogate-optimal sampler.
- 2.4x on Math (5% error) vs. 1.5x on Law (22%): rarer errors gain more.



TerminalBench2.0 (Codex, GPT 5.1 mini → 5.1), CV² of fail^t, t = 1..5: ~1.7x at modest k.

Error pattern discovery



MMLU-Pro Math, recall@M of reasoning errors: warm-start (left), cold-start (right).

- About 2x sample efficiency for error-pattern recall, about 4x for cold-start reasoning errors.
- The same sample holds 3.6x (MMLU-Pro) and 1.4x (TerminalBench) more target errors than uniform.
- Injected network-failure pattern (5 of 89 tasks) recovered over half the time with under 30 sampled tasks.