Model Provenance Testing for Large Language Models

Ivica Nikolic[1], Teodora Baluta[2], Prateek Saxena[1]

[1]National University of Singapore, [2]Georgia Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

San Diego, Dec 2nd - 7th, 2025

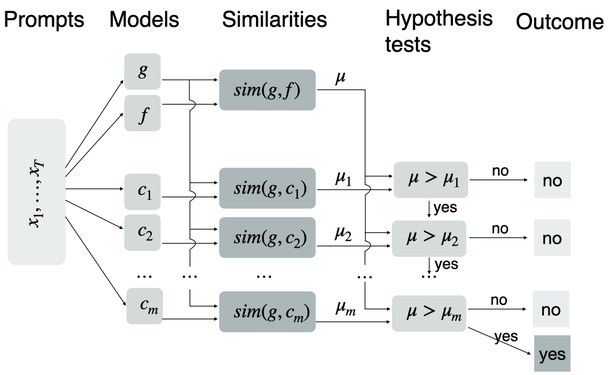
Problem and Motivation

- ***Model Provenance Testing*: Determine whether a target model has been derived from a foundational model by customizations**.
- This problem has **applications in tracking reuse** of models after the customized model has been deployed, and when authentic ground truth is unavailable or unreliable.
- ***Minimal assumptions:*** Only black-box query access to the model and no additional information such as training dataset or customization algorithms.

Approach

- Our observation is that the derived model g remains similar to the original model f.
- *Similarity* $sim(g, f)$ *between* f and $g : \mu = \frac{1}{T}\sum_{j=1}^{T} \mathbb{1}\left(f(x_j) = g(x_j)\right)$ *but our tester is* ***agnostic to the similarity*** *measure.*

Our model provenance tester that decides if model g is derived from model f using a set of control models to check if the similarity is higher than the similarity with the control models.



- We can test similarity $sim(g, f)$ using *multiple hypothesis testing.* Each test checks for each control model checks if the similarity between f and g is greater than f and c_i.
- Error can accumulate from multiple tests. Our tester returns with **bounded** familiy-wise error rate (**FWER)** using the **Holm-Bonferroni correction**.

Benchmarks

We construct two benchmarks **BENCH-A (1B-4B)** and **BENCH-B (<1B)** from Hugging Face using download counts.

Hugging Face

- Over 600 real-world models.
- No manual curation.
- Diverse domains such as financial, medical and more.
- Fine-tuning ,MoE, prompt engineering.

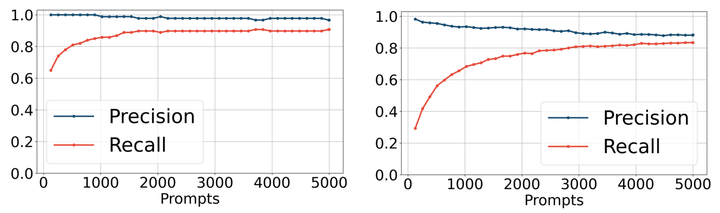
Table 1: Comparison of BENCH-A to BENCH-B on different features.

Feature	BENCH-A	BENCH-B
pre-trained models	10	57
derived models	100	383
total models	100	531
model parameters	1B-4B	$<$ 1B
compilation method	manual	automatic
ground-truth verification	higher	lower

Accuracy of Model Provenance Tester

***(RQ1)*: High accuracy across different benchmarks, precision of $90\% - 95\%$ and recall of $80\% - 90\%$** with $3{,}000$ prompts per model, with a **guaranteed false positive rate of** ≤ 0.05.

- Simply increasing the number of prompts does not guarantee uniformly better results, reflecting a fundamental trade-off: gains in recall might be accompanied by losses in precision.



Precision and recall of our tester on **BENCH-A (left)** and **BENCH-B (right)** with different prompt sizes.

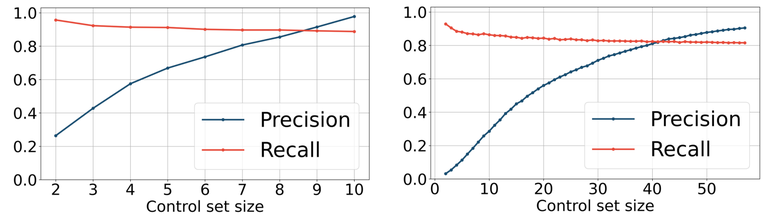
Similarity between Derived and Parent Models

***(RQ2)*: The assumption of similarity between derived and parent models is valid for most provenance pairs.**

- We evaluate similarity rankings across all provenance tests using $3{,}000$ prompts.
- BENCH-A: the true parent had the highest similarity ratio in 93% of cases, while in BENCH-B this occurred in 89% of cases.

Importance of Control Set

(RQ3): The tester's **performance** degrades when the **control set** is too **small or poorly selected**.

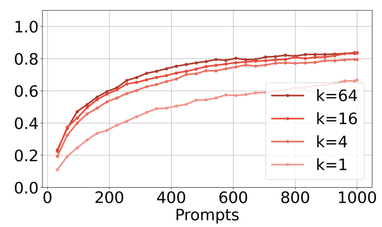


Precision and recall of our tester on **BENCH-A (left)** and **BENCH-B (right)** with smaller control set sizes.

Reducing Query Complexity

(RQ4): The online query optimization strategy leads to a **4-6x query reduction without accuracy drop**, whereas the offline approach performs only marginally better and has a negative impact on recall.

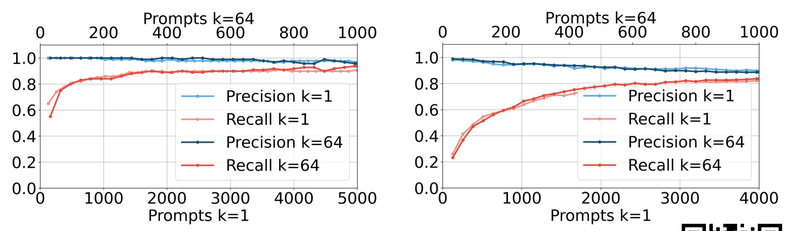
- Optimization for *online queries* made to the tested child model g, and offline queries made to the parent model f.
- Online optimization: **rejection sampling** with an **entropy-based selection criterion**; offline optimization: best-arm identification (BAI).



Online optimization: Recall for BENCH-B with different values of advanced prompt sampling (k).

Table 2: Precision and recall of the base vs BAI tester on BENCH-A and BENCH-B.

Allowed Queries T	Benchmark	Tester	Avg Queries	Precision	Recall
500	BENCH-A	base	500	1.00	0.81
500	BENCH-A	BAI	450	0.98	0.29
500	BENCH-B	base	500	0.95	0.56
500	BENCH-B	BAI	452	0.98	0.29
1,000	BENCH-A	base	1,000	0.99	0.86
1,000	BENCH-A	BAI	605	1.00	0.63
1,000	BENCH-B	base	1,000	0.94	0.68
1,000	BENCH-B	BAI	809	0.98	0.42
2,000	BENCH-A	base	2,000	0.98	0.89
2,000	BENCH-A	BAI	1,482	0.97	0.54
2,000	BENCH-B	base	2,000	0.92	0.77
2,000	BENCH-B	BAI	1,482	0.97	0.54



Online optimization: Precision/recall for BENCH-A (left) and BENCH-B (right) when advanced online prompt sampling with $k = 64$ compared to no advanced sampling ($k = 1$).