

ConStat: Performance-Based Contamination Detection in Large Language Models

Jasper Dekoninck, Mark Niklas Müller, Martin Vechev

Traditional Data Contamination

$$D_{\text{train}} \cap D_{\text{test}} \neq \emptyset$$

Traditional Data Contamination

$$D_{\text{train}} \cap D_{\text{test}} \neq \emptyset$$




Huge
Not publicly available

Traditional Data Contamination

$$D_{\text{train}} \cap D_{\text{test}} \neq \emptyset$$



Huge
Not publicly available



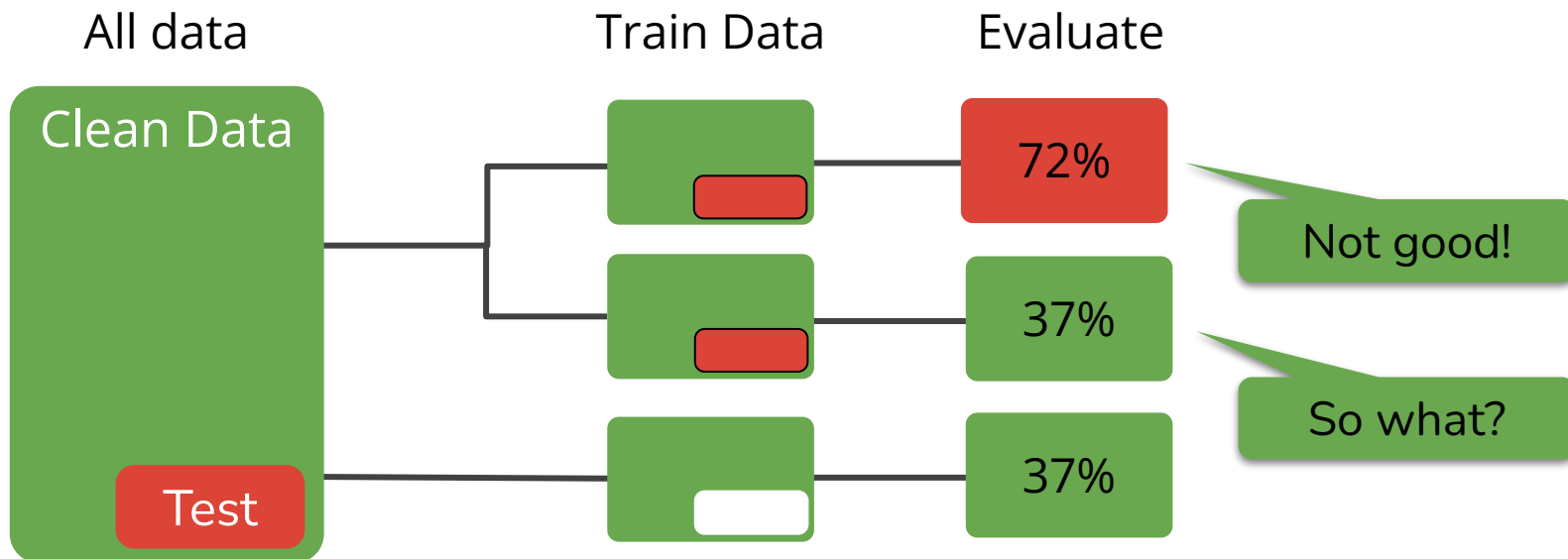
What are similar, e.g., rephrased, samples?

A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.

A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.



A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.

What is $5 + 5$?

$5 + 5$ is 10.

A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.

What is $5 + 5$?

$5 + 5$ is 10.

Syntax-Specific

Calculate $5 + 5$.

$5 + 5$ is 9.

A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.

What is $5 + 5$?

$5 + 5$ is 10.

Syntax-Specific

Calculate $5 + 5$.

$5 + 5$ is 9.

Sample-Specific

What is $7 + 3$?

$7 + 3$ is 9.

A New Perspective: Performance-Based Data Contamination

Contamination is non-generalizing and artificially inflated benchmark performance.

What is $5 + 5$?

$5 + 5$ is 10.

Syntax-Specific

Calculate $5 + 5$.

$5 + 5$ is 9.

Sample-Specific

What is $7 + 3$?

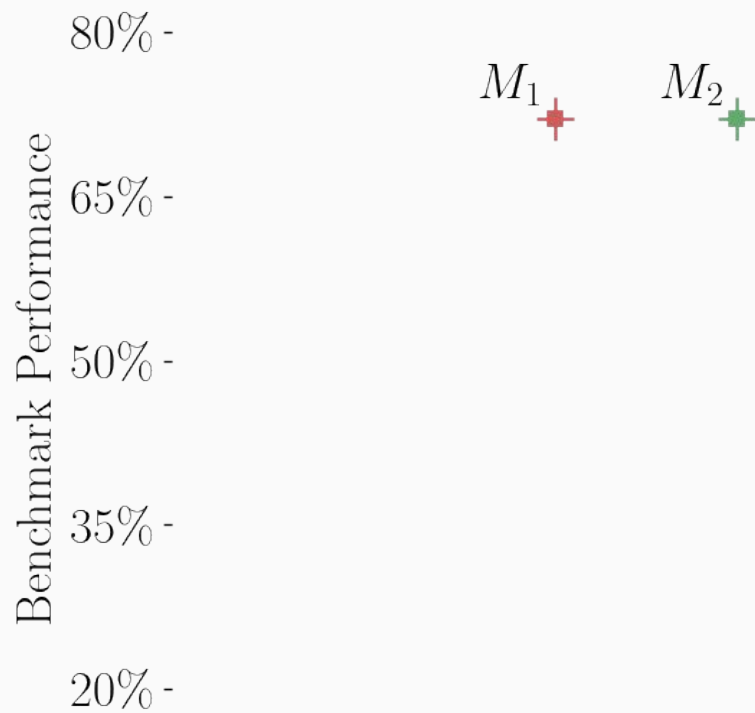
$7 + 3$ is 9.

Benchmark-Specific

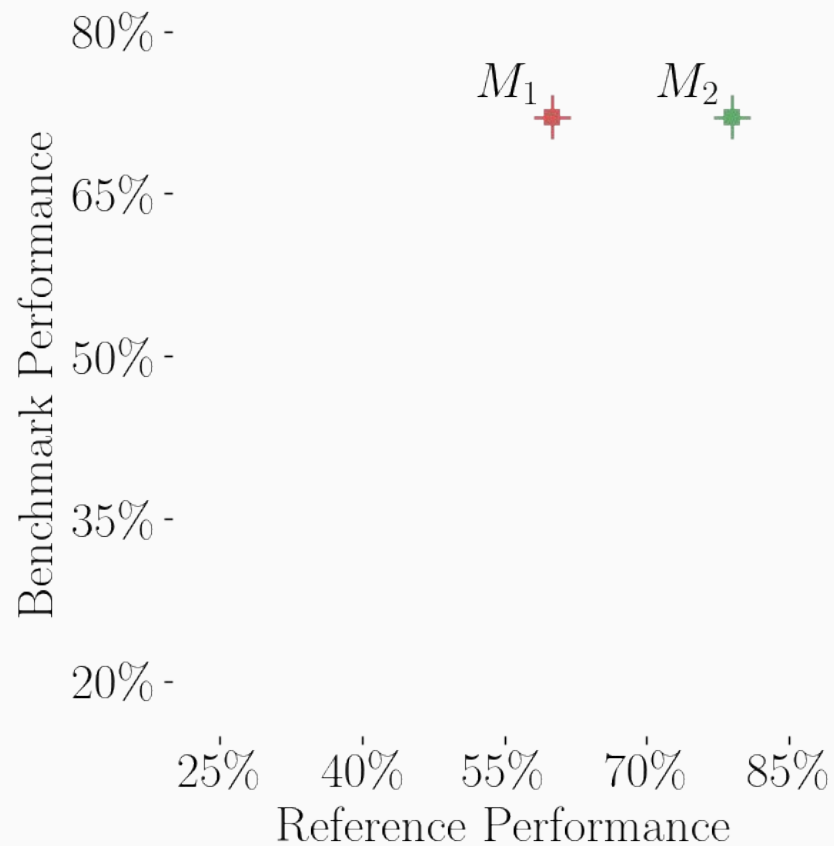
What is $2 * 3$?

$2 * 3$ is 5.

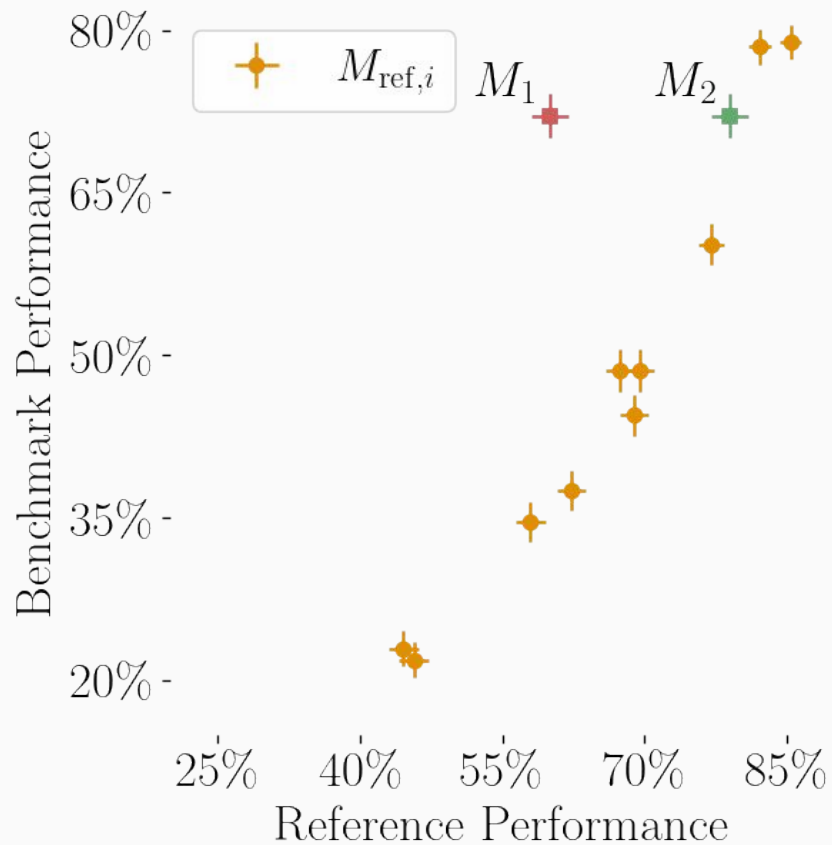
End-To-End Pipeline



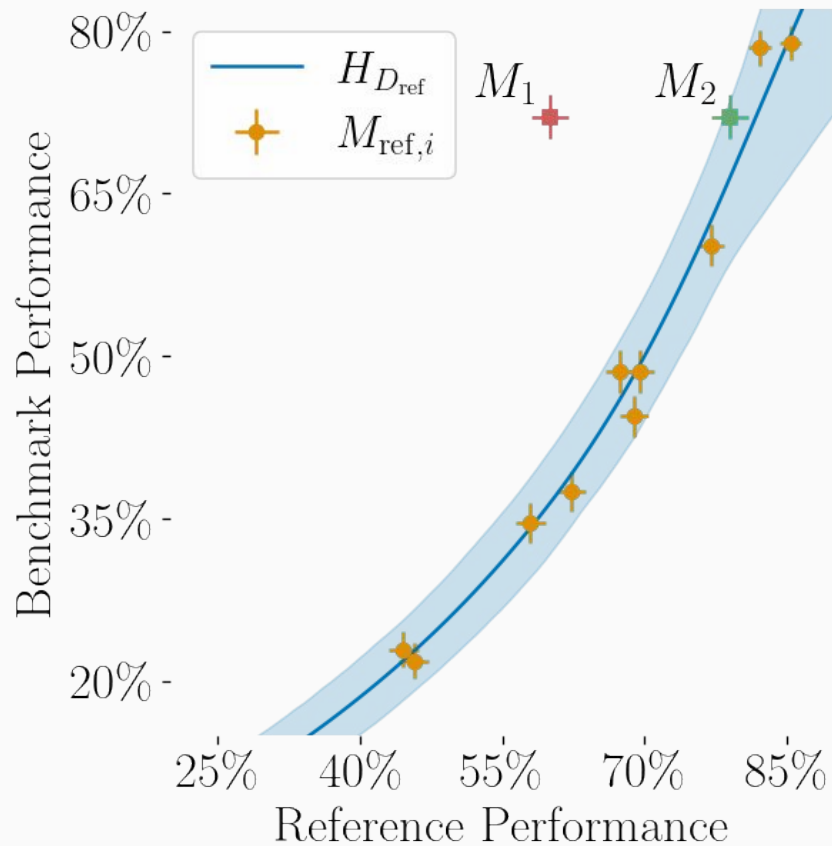
End-To-End Pipeline



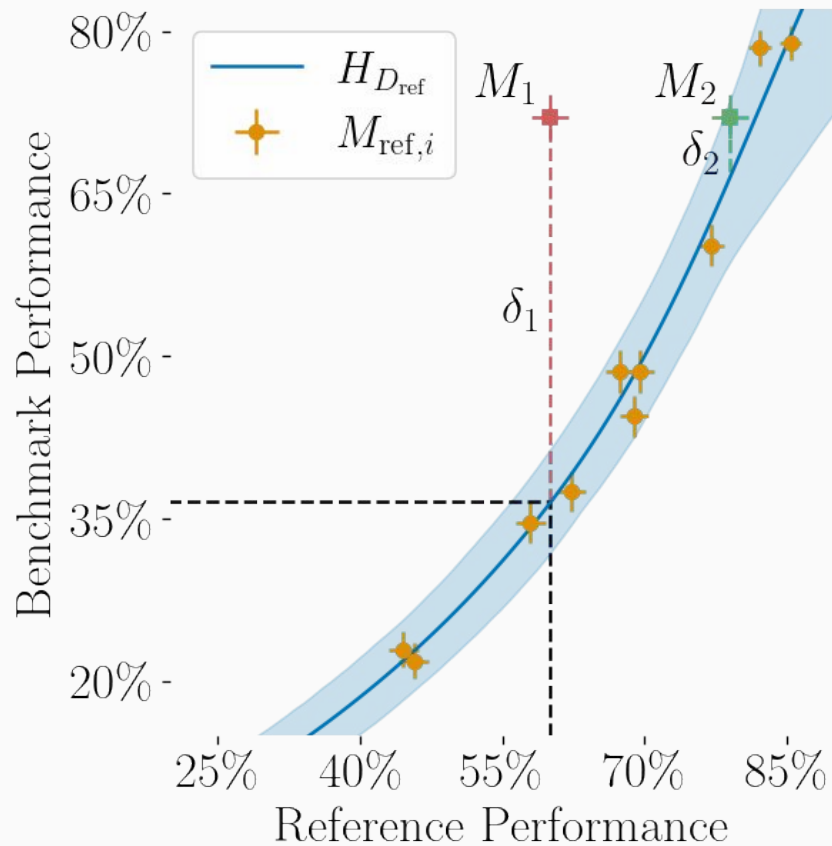
End-To-End Pipeline



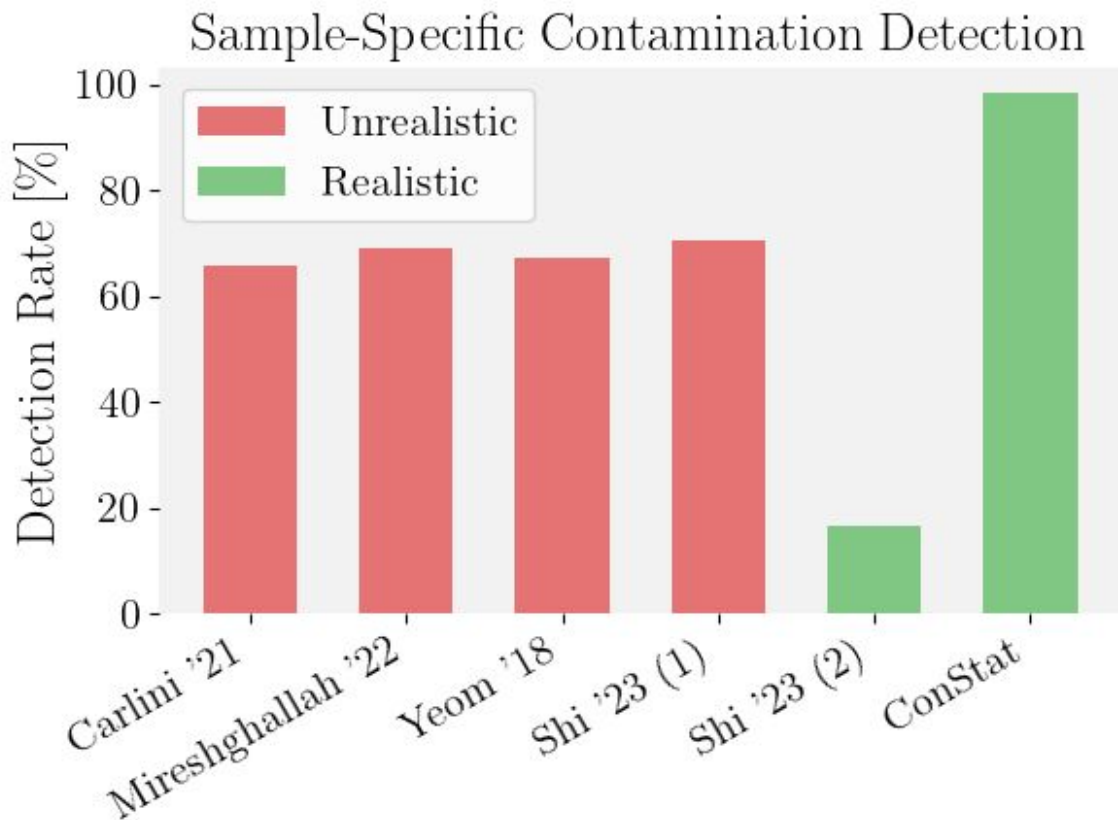
End-To-End Pipeline



End-To-End Pipeline



Results: Controlled Experiments



More Results & Large-Scale Contamination Analysis

