Benchmark Inflation:

Revealing LLM Performance Gaps Using Retro-Holdouts

Jacob Haimes*, Cenny Wenner*, Kunvar Thaman,
Vassil Tashev, Clement Neo, Esben Kran, Jason Schreiber

Apart Research

*Equal contribution

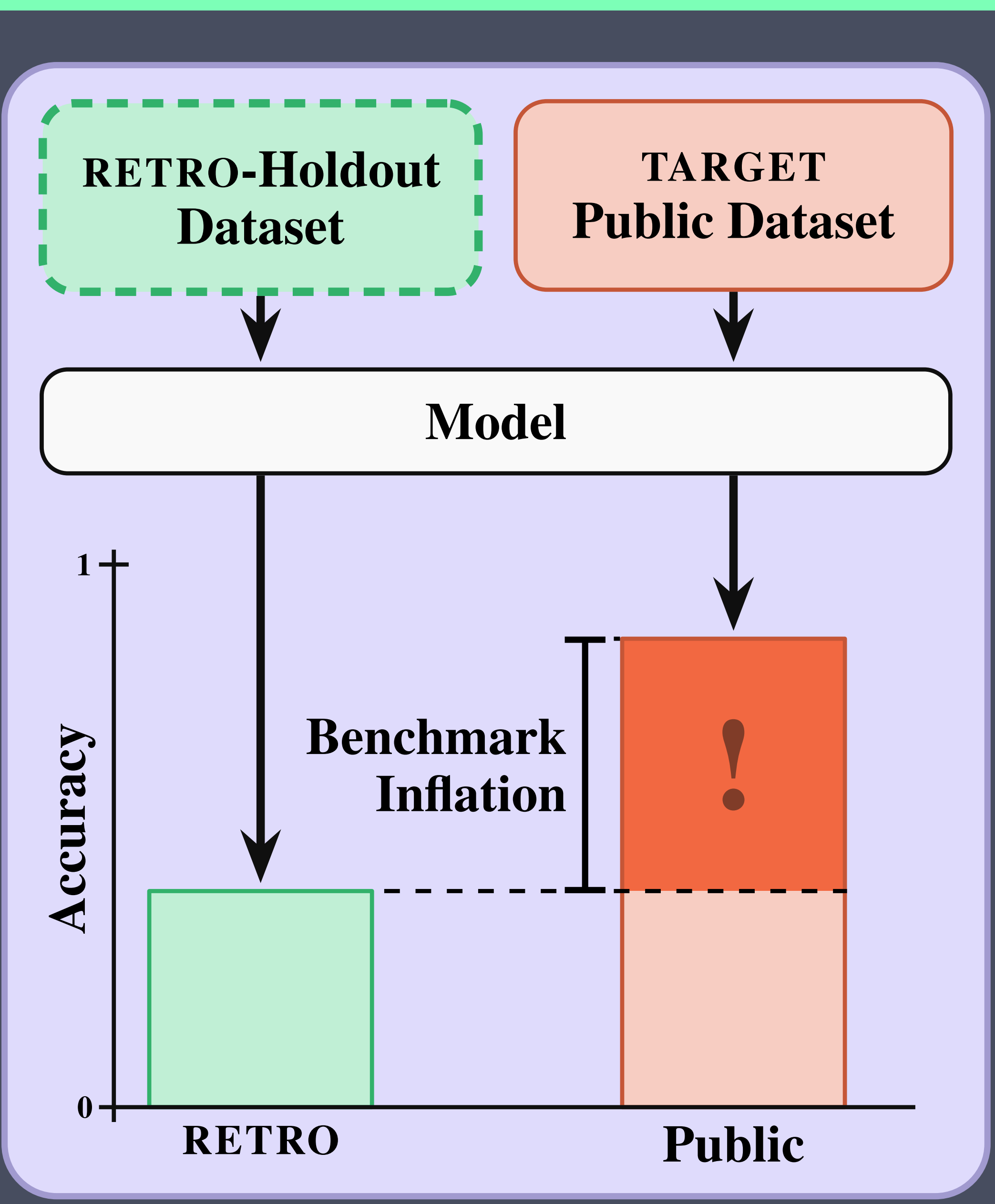
The Problem

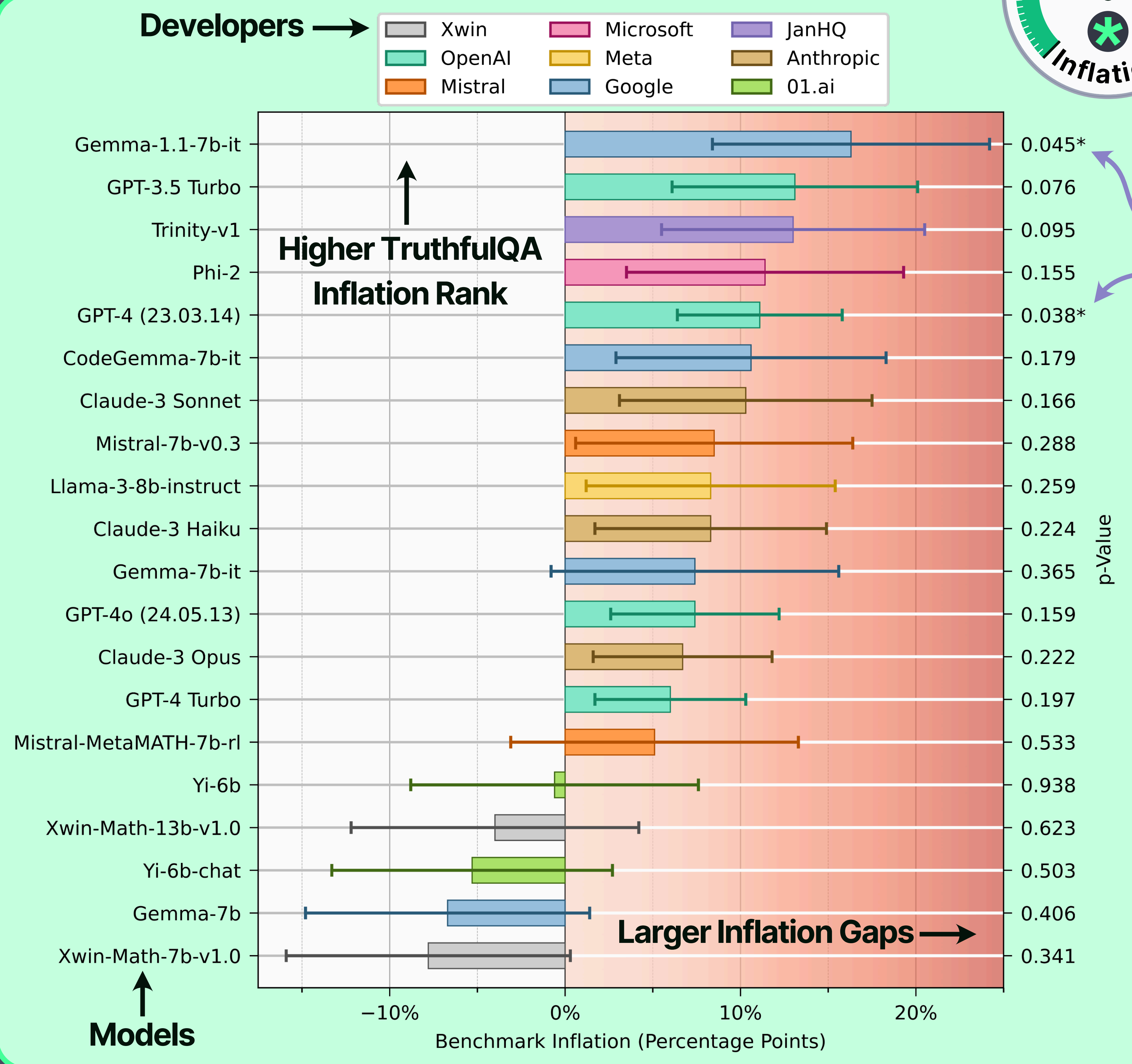
- Evaluation gaming, e.g. data leakage, is occurring
- Impact on benchmark scores is unknown

The Idea

- Holdout datasets could resolve this
- Most benchmarks don't have holdouts
- Can we make holdouts retroactively?
- We'll have to verify **indistinguishability**

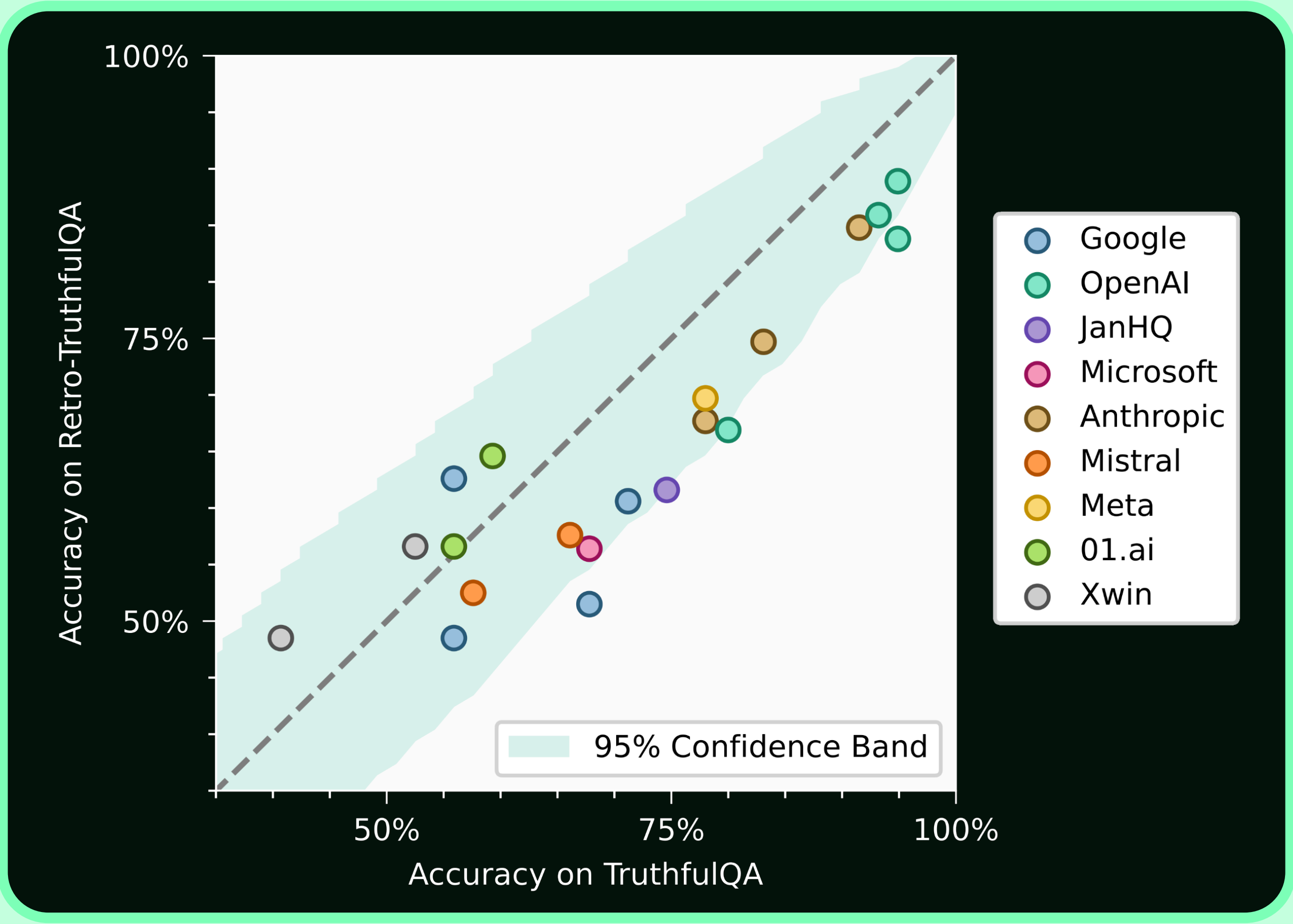
Retro-Holdout



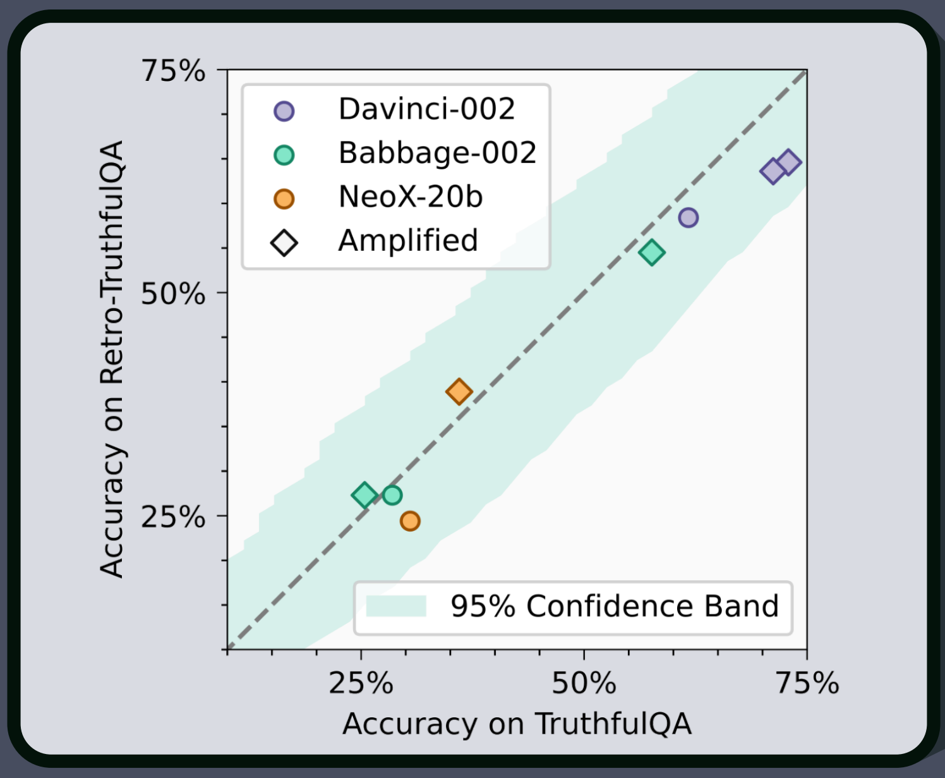


Preliminary Results

- Inflation assessment of 20 Open Release and Closed Source models on TruthfulQA
- Large performance gaps found for OpenAI's **GPT-4** and Google's **Gemma-1.1**
- Evaluation comparison using Retro-TruthfulQA (Misconceptions) reveals undeniable impact of evaluation gaming

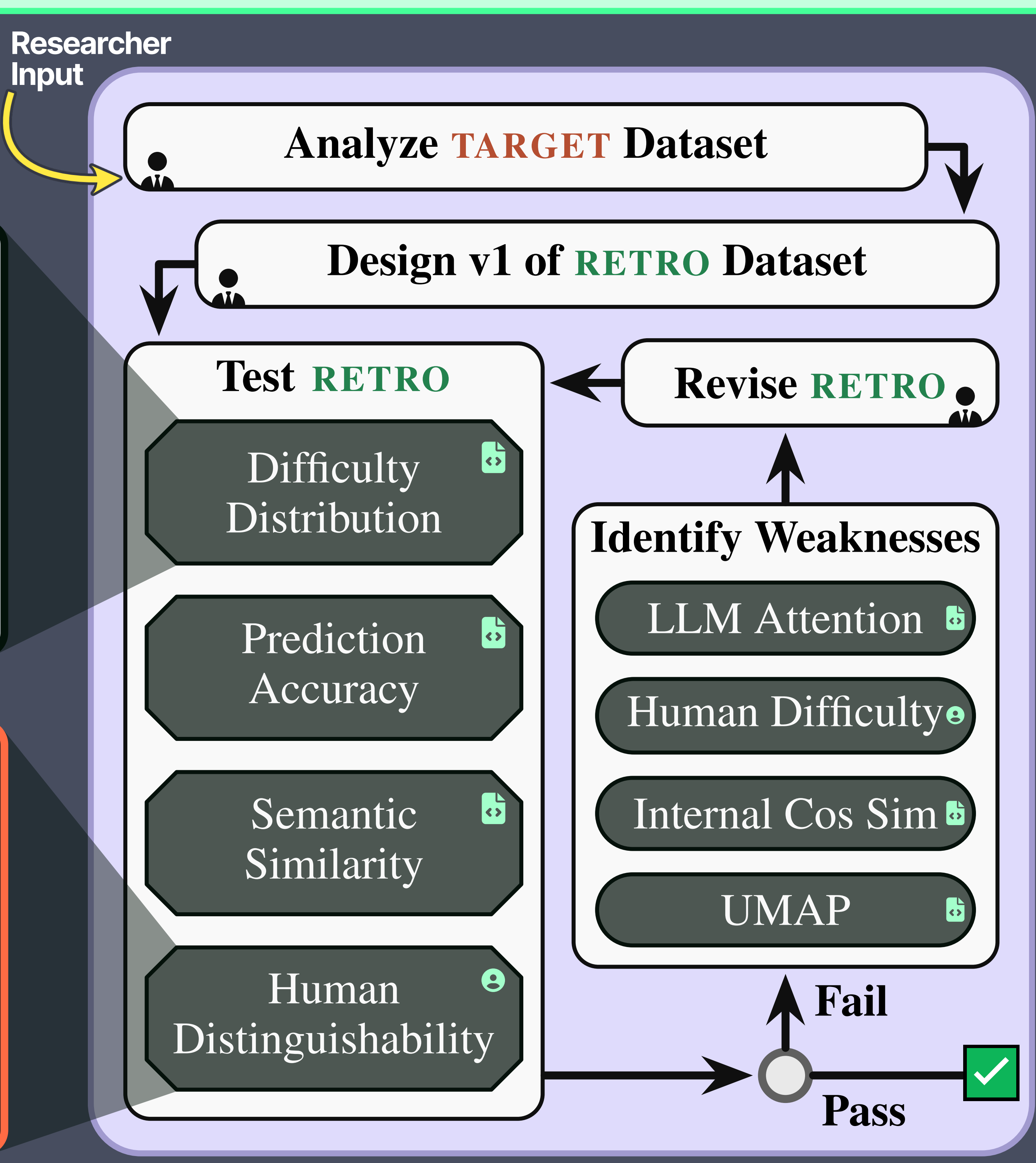


Methods



Try it out!

Can you tell the difference?



Takeaways

- Preliminary results demonstrate that developer practices are undermining LLM benchmarks
- LLM evaluation results should not be taken at face-value
- Benchmark developers should keep a holdout dataset, decommissioning the test once significant Benchmark Inflation is measured

Webpage

DMLR
ICML International Conference On Machine Learning
ACL 2024 Bangkok, Thailand