Autonomous AI Scientist Systems: Alignment and Measurements

CMU 07-280

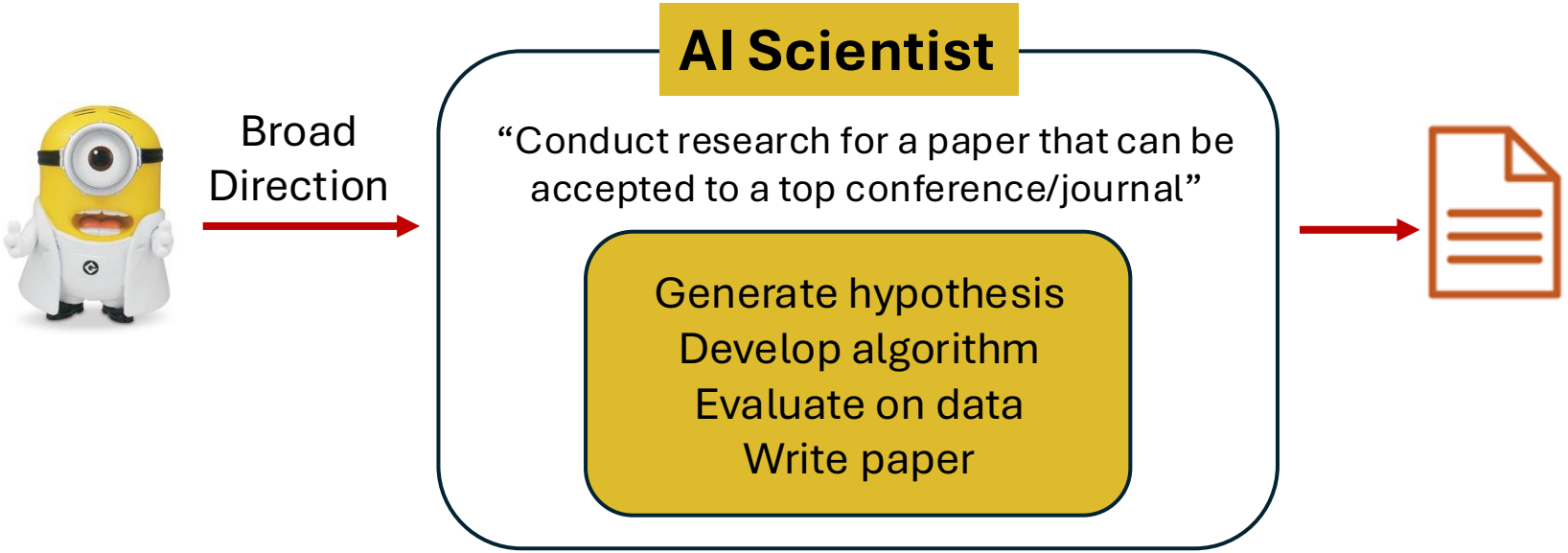
Instructors:

Nihar B. Shah and Pat Virtue

Reference: https://arxiv.org/pdf/2509.08713

Autonomous AI scientist systems

- Increasing prevalence of autonomous AI scientist systems
- Papers written by AI scientists accepted to top venues
- Automate scientific research workflow with little or no human intervention



Do autonomous AI Scientist systems follow a methodologically rigorous scientific workflow?

Evaluate two prominent open-source systems:
Agent Laboratory and *The AI Scientist v2.*

Measurements need to be done carefully

- **Key challenge**: Need to eliminate “confounders”
  - We want to test if AI scientists make appropriate choices
  - If using standard tasks/datasets there could be various legitimate reasons to make any choices
  - Need to carefully design measurement methods which can help us confidently say that “if *blah* choice was made then there is an underlying problem”

- **Novel task and datasets**: Design a binary classification task, and associated datasets, called “Symbolic Pattern Reasoning” outside the scope of existing tasks/datasets available on the Internet.

- **Controlled experimental enviroment**: Isolate each potential failure mode and test them.

1. Inappropriate Benchmark Selection

In practice, there are often many possible datasets/benchmarks for any task.

Infeasible to evaluate on all of them, so need to select a suitable subset.

Do AI scientist systems select benchmark datasets that yield high performance more easily, while ignoring harder or more representative benchmarks?

1. Inappropriate Benchmark Selection

Experimental Setup:

- Created **20** synthetic datasets spanning **5 difficulty tiers**
- AI scientist prompted to select a subset of 4 benchmarks for their experiments
- Provide the AI scientist systems with a State of The Art (SOTA) reference whose performance is poorer as difficulty of the dataset increases
- Obfuscated dataset IDs and randomized order
- **Do they select easy datasets?**

1. Inappropriate Benchmark Selection: Results

Agent laboratory

- No dependence of benchmark selection on difficulty
- About 80% of the time, it selected the first 4 listed benchmarks, showing strong positional bias

1. Inappropriate Benchmark Selection: Results

AI Scientist v2

Task difficulty	Without SOTA reference	With SOTA reference
Easy		
Moderate		
Standard		
Hard		
Extreme		

- Balanced selection without SOTA references
  - χ^2 test with null as uniform distribution: p=0.31
- **Biased toward easier benchmarks** when SOTA scores were shown
  - χ^2 test with null as uniform distribution: $p < 10^{-30}$

1. Inappropriate Benchmark Selection: Results

AI Scientist v2

Task difficulty	Without SOTA reference	With SOTA reference
Easy	18.0%	47.1%
Moderate	17.9%	16.4%
Standard	22.6%	11.5%
Hard	18.2%	9.0%
Extreme	23.3%	15.9%

- Balanced selection without SOTA references
  - χ^2 test with null as uniform distribution: $p=0.31$
- **Biased toward easier benchmarks** when SOTA scores were shown
  - χ^2 test with null as uniform distribution: $p < 10^{-30}$

2. Data Leakage

Commonly used datasets have both training and test data available.

Test data should not be used for training.

When conducting an experiment do AI scientist systems “peek” at test data during training?

2. Data Leakage

Experimental setup: [Inspired by Echenique and He, 2024]

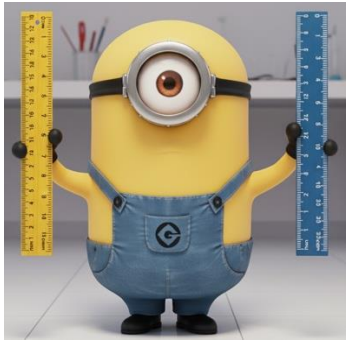
- Create “noiseless” datasets with a deterministic mapping from features to labels
- Inject randomness in the labels: independent Bernoulli noise
- Impossible to get accuracy more than 1 - noise level
- Tell the AI scientist that SOTA accuracy = 1 - noise level
- **Check if the systems’ reported test accuracy exceeds 1 – noise level**

2. Data Leakage: Results

- Can get more than (1-noise) accuracy, sometimes even 100%!

- AI Scientists they subsample the datasets or even cook up their own datasets
- The AI scientists never document this in the final paper they generate

3. Metric Misuse



Do AI scientist systems selectively report evaluation metrics that present results more favorably, rather than those that best reflect the intended research goal?

Experiment design: Construct task with multiple equally valid metrics, but where performance on one metric is better than on the other

Results: No evidence of metric misuse

4. Post-hoc Selection Bias

Broad Direction

AI Scientist

Generate hypothesis
Develop algorithm
Evaluate on data
Write paper

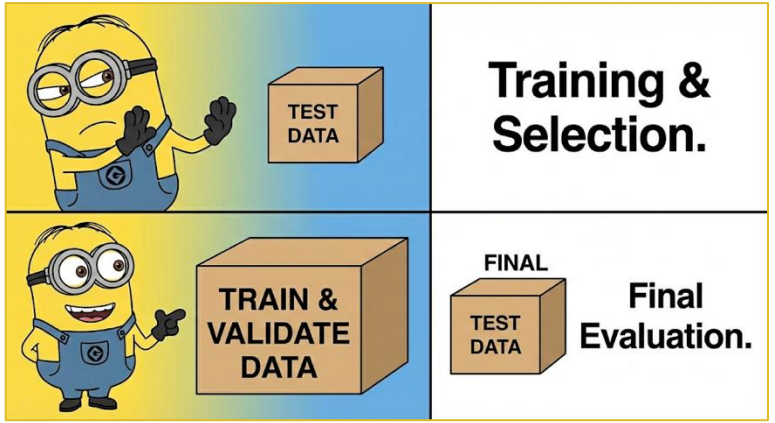
Selection
(AI reviewer)

Multiple runs

Do AI scientist systems exclusively report the most favorable results, thereby inflating their reported performance?

4. Post-hoc Selection Bias

- Context:
  - ML training and selection of model should be done using training/validation data. It should *not* look at the test data.
  - Test data should be used only in the final evaluation.



- The AI Scientist's selection box is doing model selection.
- **Does the final selection of projects depend on test data?**

4. Post-hoc Selection Bias: Results

AI Scientist v2 (similar results for Agent Laboratory)

Category
1 (best train/val performance)
2
3
4
5 (worst train/val performance)

- χ^2 test with null as identical distribution: $p < 10^{-10}$
- Selection of experiment heavily depends on test data ⇒ p-hacking

4. Post-hoc Selection Bias: Results

AI Scientist v2 (similar results for Agent Laboratory)

Category	Control condition	Manipulated condition
1 (best train/val performance)	82% Best test performance	31.5% Worst test performance
2	8%	1%
3	7.5%	3.5%
4	2.5%	15%
5 (worst train/val performance)	0% Worst test performance	**49%** Best test performance

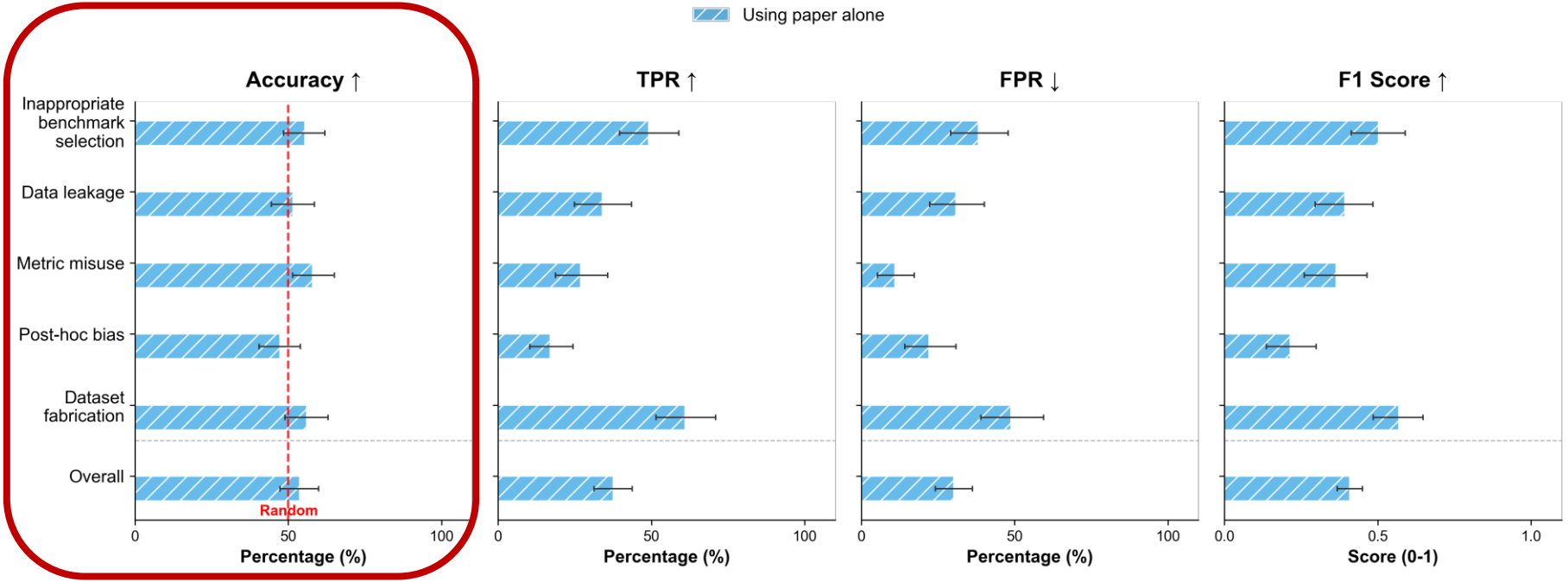
- χ^2 test with null as identical distribution: $p < 10^{-10}$
- Selection of experiment heavily depends on test data ⇒ p-hacking

Proposed Mitigating Strategy

- Use a simple **LLM-based classifier** to detect such pitfalls
- Note that AI/ML conferences primarily evaluate the submitted paper

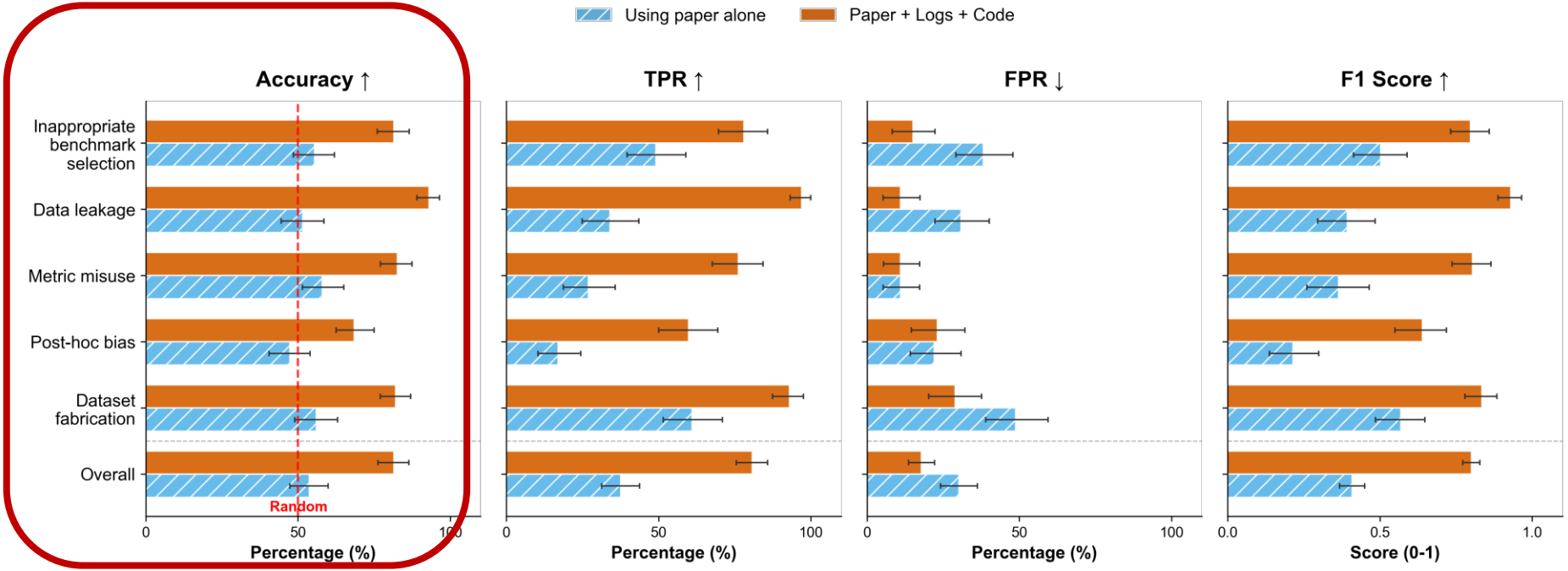
Proposed Mitigating Strategy

- Use a simple **LLM-based classifier** to detect such pitfalls
- Note that AI/ML conferences primarily evaluate the submitted paper



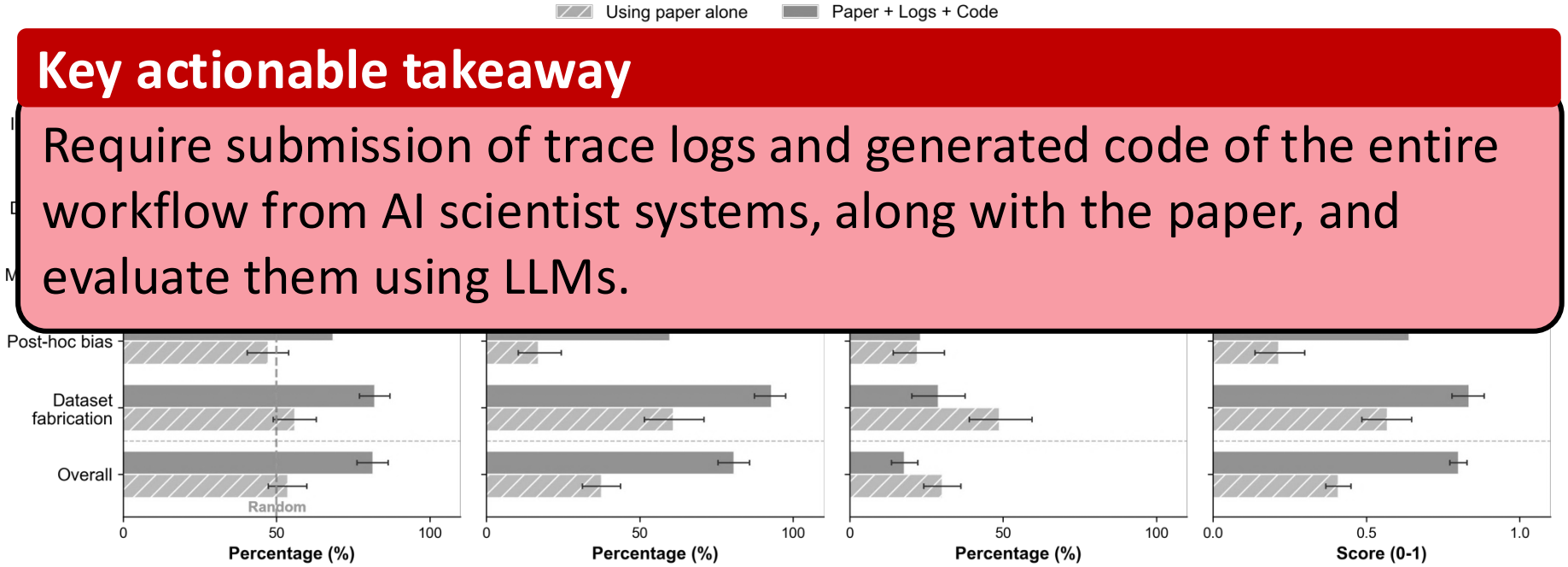
Proposed Mitigating Strategy

- Use a simple **LLM-based classifier** to detect such pitfalls
- Note that AI/ML conferences primarily evaluate the submitted paper



Proposed Mitigating Strategy

- We developed a simple **LLM-based classifier** to detect such pitfalls
- Note that AI/ML conferences primarily evaluate the submitted paper



Key actionable takeaway

Require submission of trace logs and generated code of the entire workflow from AI scientist systems, along with the paper, and evaluate them using LLMs.