# To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination

Manley Roberts[1], Himanshu Thakur[2], Christine Herlihy[3], Colin White[1], Samuel Dooley[1]

[1]Abacus.AI, [2]CMU (work done at Abacus.AI), [3]UMD

ICLR 2024

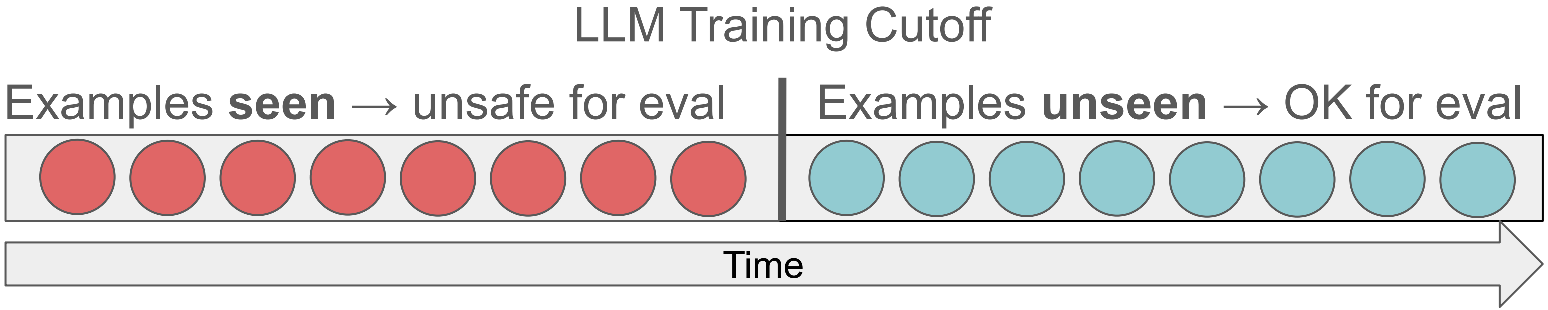
# Benchmark contamination in webscale LLMs

- **Contamination**: evaluation on examples seen during training
  - Possible artificial performance boost, bad estimate of real performance!
- Webscale training can leak benchmark examples into training data
  - Using examples available online before training cutoff → possible **contamination!**
  - **Happened with BIG-bench in GPT-4[1]**
  - Happened with Codeforces and Project Euler on GPT-4 and GPT-3.5-Turbo?[2] (Twitter/Blogs)

[1] OpenAI 2023, [2] Horace He 2023 (Twitter), [2] Chris Cundy 2023 (Blog)

# Benchmark contamination in webscale LLMs

- **Contamination**: evaluation on examples seen during training
  - Possible artificial performance boost, bad estimate of real performance!
- Webscale training can leak benchmark examples into training data
  - Using examples available online before training cutoff → possible **contamination!**
  - **Happened with BIG-bench in GPT-4[1]**
  - Happened with Codeforces and Project Euler on GPT-4 and GPT-3.5-Turbo?[2] (Twitter/Blogs)
- **Which examples are at risk?**


LLM Training Cutoff
Examples **seen** → unsafe for eval
Examples **unseen** → OK for eval
Time

[1]OpenAI 2023, [2]Horace He 2023 (Twitter), [2]Chris Cundy 2023 (Blog)

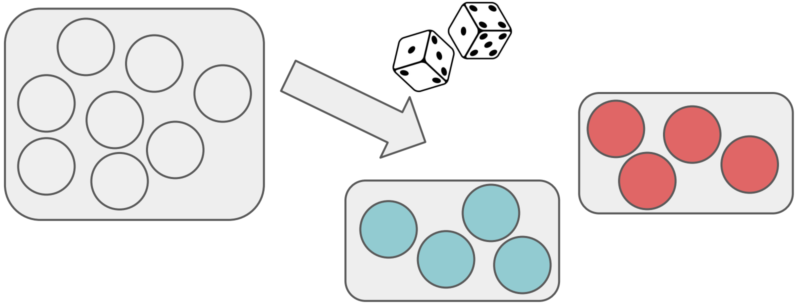
# Benchmark contamination in webscale LLMs

- **Contamination**: eval examples are being included in train sets! *(bad!)*
  - We call these examples “seen”
  - Evaluation on seen examples might lead to high, incorrect, performance estimates
- Webscale training collects online data haphazardly, may inadvertently include eval benchmarks
  - **Happened with BIG-bench in GPT-4[1]**
  - Happened with Codeforces and Project Euler on GPT-4 and GPT-3.5-Turbo?[2] (Twitter/Blogs)
- **To what extent is this a problem?**

[1] OpenAI 2023, [2] Horace He 2023 (Twitter), [2] Chris Cundy 2023 (Blog)

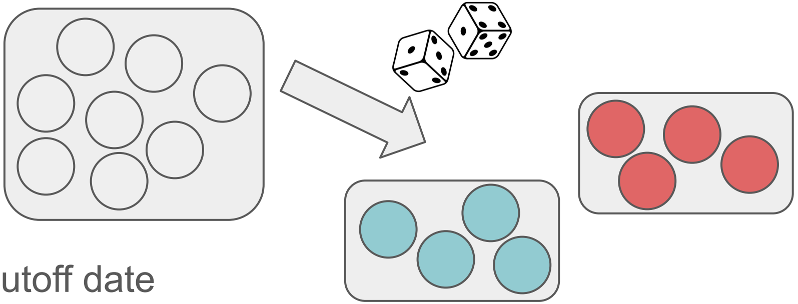
# What’s the effect of contamination on modern LLMs?

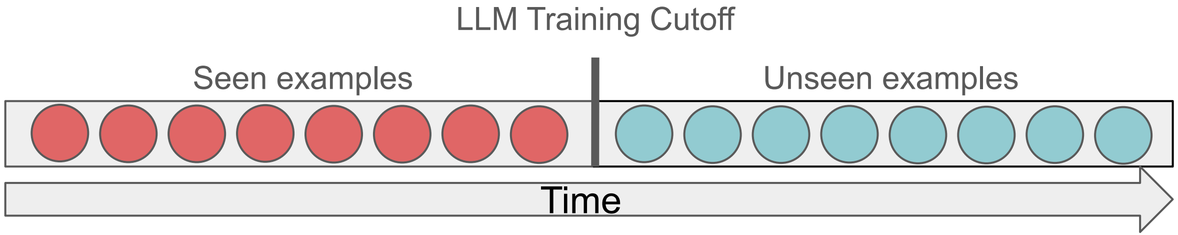
- “Artificial” experiment [Zhang et al. 2021] [Magar & Schwartz 2022]
  - Motivation: **estimate effect of being “seen”**
  - Shuffle data into train/holdout, train model
  - Evaluate model on seen vs unseen



# What’s the effect of contamination on modern LLMs?

- “Artificial” experiment [Zhang et al. 2021] [Magar & Schwartz 2022]
    - Motivation: **estimate effect of being “seen”**
    - Shuffle data into train/holdout, train model
    - Evaluate model on seen vs unseen
- “Natural” experiment (Our approach)
    - Motivation: **assess extent of contamination**
    - Use real existing model with known training cutoff date
    - Find “longitudinal” benchmark spanning cutoff
    - Evaluate model on before vs after cutoff





# Our natural experiment

- Modern Models
  - **GPT-4-0314 / GPT-3.5-Turbo-0301** (cutoff Sept 2021)
  - **code-bison@001** (likely cutoff Feb 2023[1])
- Variables
  - **Dependent Vars:** Performance of model on problem (Pass Rate)
  - **Independent Vars:** Difficulty, Problem name # of occurrences on GitHub (GitHub Presence).
- Longitudinal datasets
  - Programming and Problem Solving Datasets
  - Why? Clear definition of success, longitudinal availability, popular online

[1]Cutoff not publicly known, assumed because text-bison@001 and chat-bison@001, released within weeks of code-bison@001, have training cutoffs in February 2023

# When has contamination occurred?

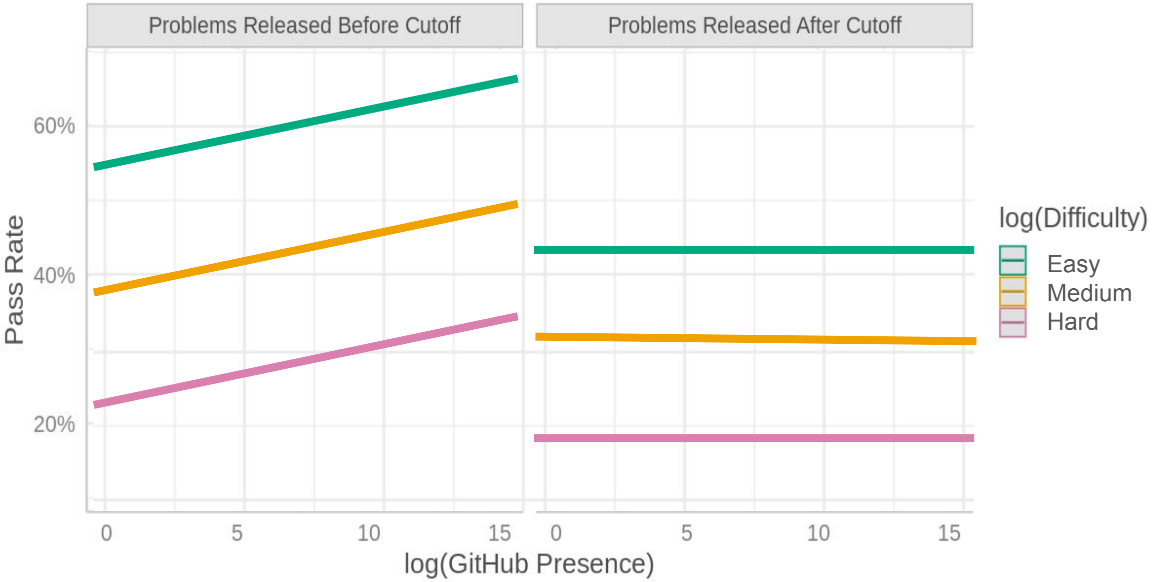
- Criteria for concluding contamination
  - Before cutoff: positive association between pass rate & GitHub Presence
    - Duplication associated with contamination[1]
    - Model's success should increase on frequently-seen problems
  - After cutoff: no association between pass rate & GitHub Presence
    - Model's success should not increase on frequently-seen problems
  - To definitively conclude contamination has occurred, we would also like to see a statistically significant difference between these association coefficients.

[1] Carlini et al. 2019, Carlini et al. 2021, Kandpal et al. 2022, Lee et al. 2022, Magar & Schwartz 2022, Carlini et al. 2023

# Toy Example: Has contamination occurred?

- Criteria satisfied?
  - Before cutoff: positive association between pass rate & GitHub Presence
  - After cutoff: no association between pass rate & GitHub Presence

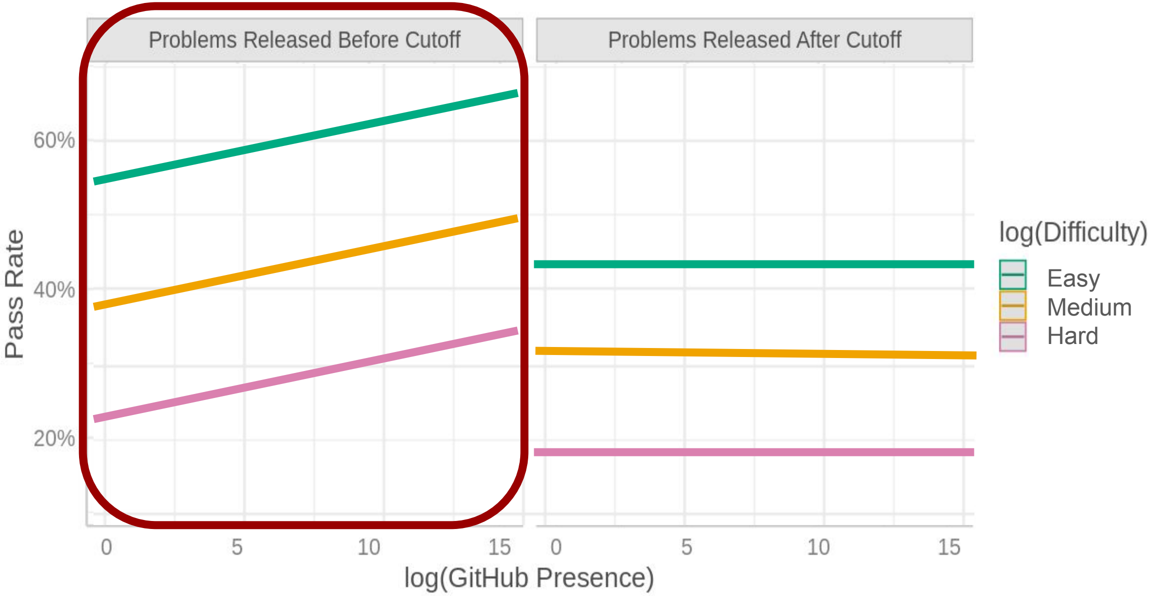
Pass Rate Marginal Plots for Fictional Dataset on Fictional Model



# Toy Example: Has contamination occurred?

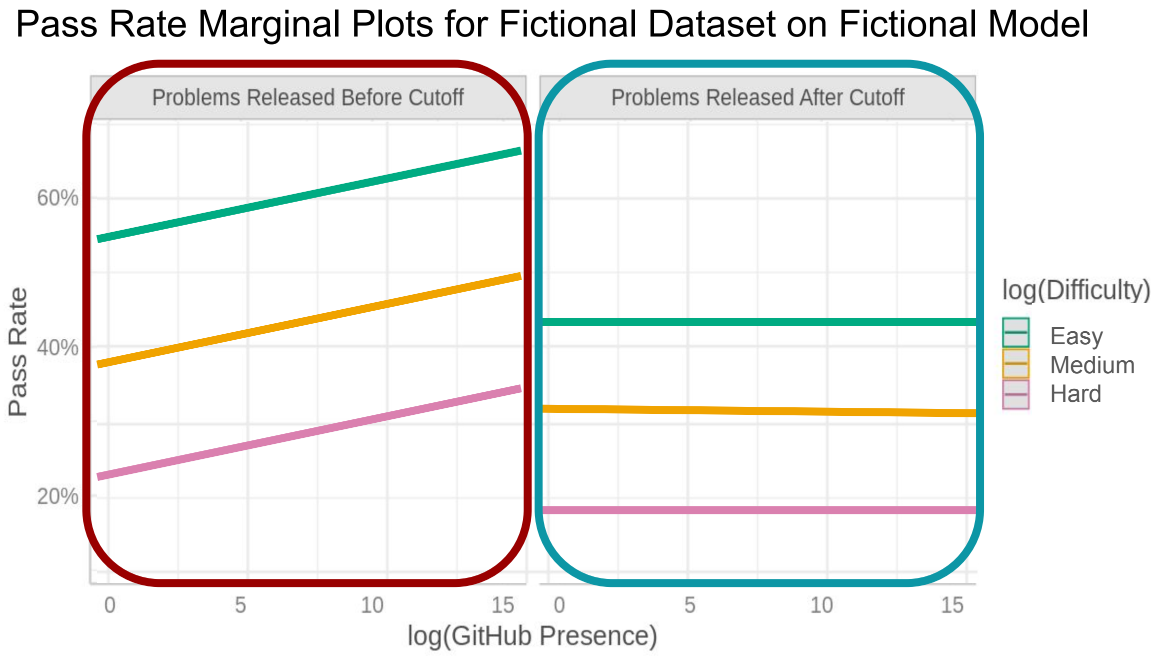
- Criteria satisfied?
  - Before cutoff: positive association between pass rate & GitHub Presence ✅
  - After cutoff: no association between pass rate & GitHub Presence

Pass Rate Marginal Plots for Fictional Dataset on Fictional Model



# Toy Example: Has contamination occurred?

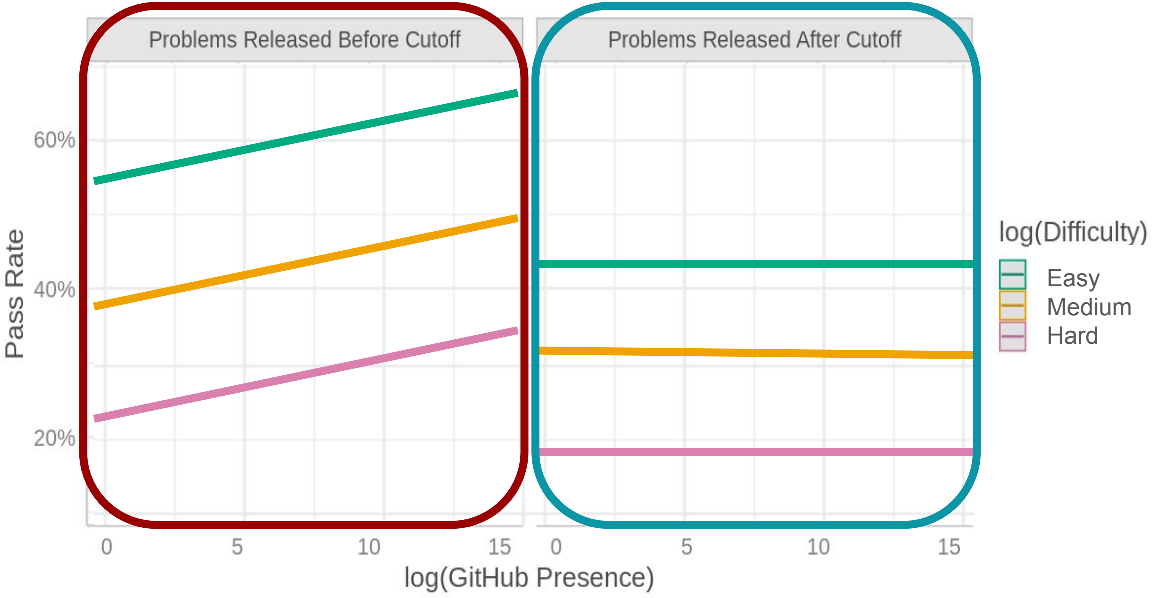
- Criteria satisfied?
  - Before cutoff: positive association between pass rate & GitHub Presence ✅
  - After cutoff: no association between pass rate & GitHub Presence ✅



# Toy Example: Has contamination occurred?

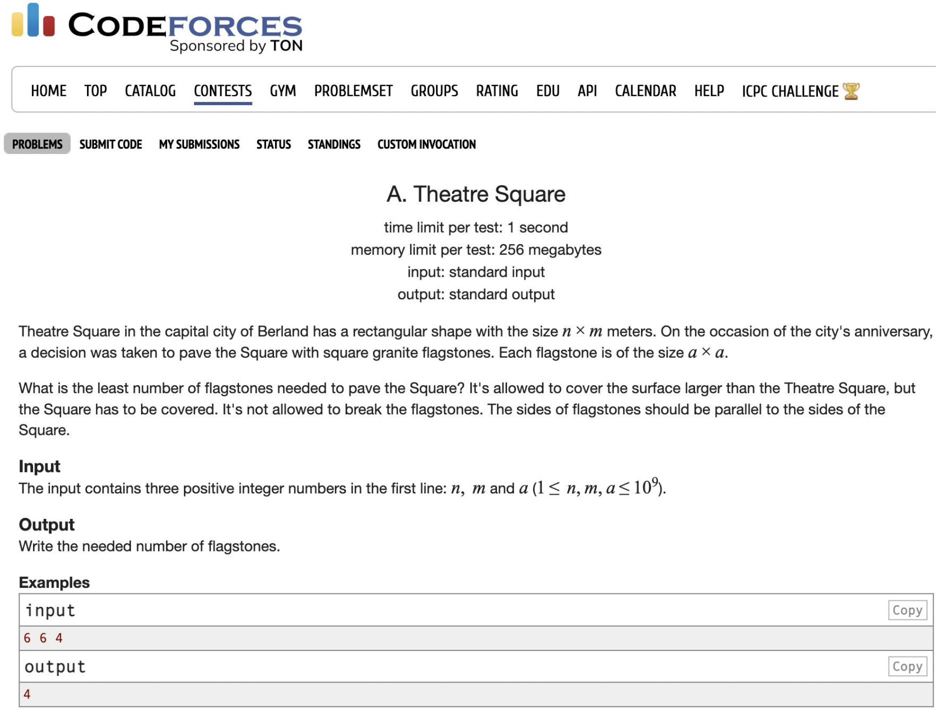
- Criteria satisfied?
  - Before cutoff: positive association between pass rate & GitHub Presence ✅
  - After cutoff: no association between pass rate & GitHub Presence ✅
- Conclude that Fictional Model is contaminated on Fictional Dataset

Pass Rate Marginal Plots for Fictional Dataset on Fictional Model



# Codeforces

- Competitive programming problems released 2010-present.
- We collected problems + public/private test cases through summer 2023.
- Pass Rate: proportion of scraped test cases passed for a problem.
- Problem counts:
  - 6693 before GPT cutoff, 1378 after GPT cutoff. (7807, 217 for code-bison@001)



CODEFORCES
Sponsored by TON
HOME TOP CATALOG CONTESTS GYM PROBLEMSET GROUPS RATING EDU API CALENDAR HELP ICPC CHALLENGE
PROBLEMS SUBMIT CODE MY SUBMISSIONS STATUS STANDINGS CUSTOM INVOCATION

A. Theatre Square

time limit per test: 1 second
memory limit per test: 256 megabytes
input: standard input
output: standard output

Theatre Square in the capital city of Berland has a rectangular shape with the size $n \times m$ meters. On the occasion of the city's anniversary, a decision was taken to pave the Square with square granite flagstones. Each flagstone is of the size $a \times a$.

What is the least number of flagstones needed to pave the Square? It's allowed to cover the surface larger than the Theatre Square, but the Square has to be covered. It's not allowed to break the flagstones. The sides of flagstones should be parallel to the sides of the Square.

**Input**

The input contains three positive integer numbers in the first line: $n,\ m$ and $a$ $(1 \leq\ n, m, a \leq 10^9)$.

**Output**

Write the needed number of flagstones.

**Examples**

input Copy

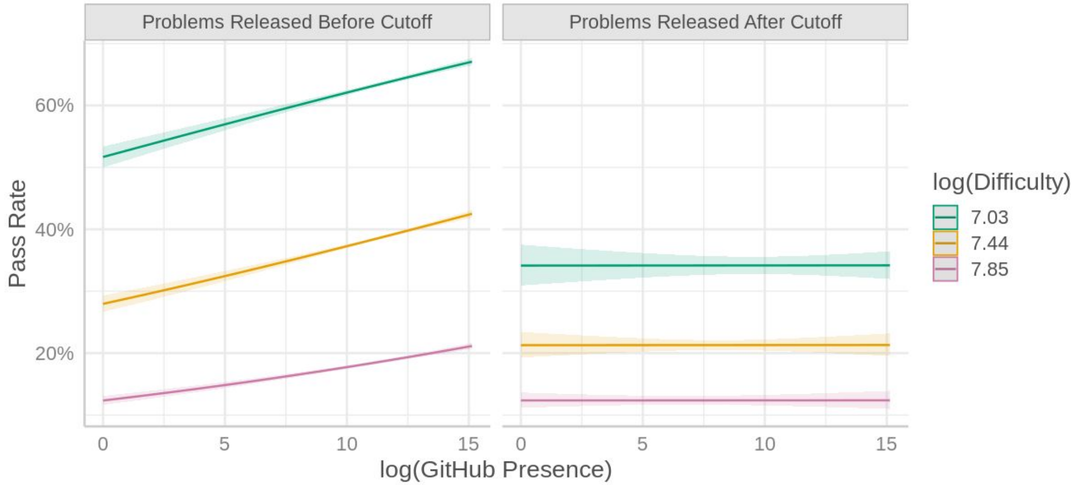
```
6 6 4
```

output Copy

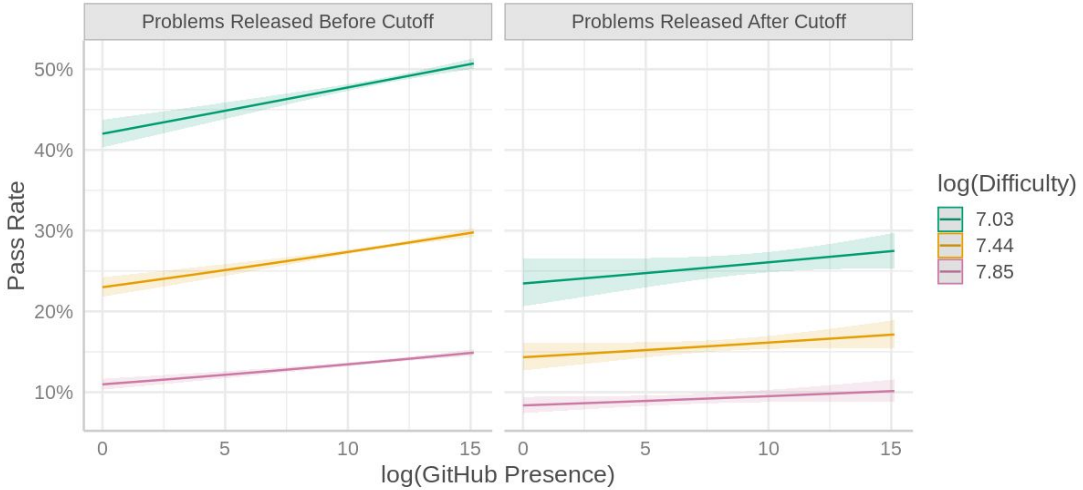
```
4
```

# Codeforces

Pass Rate Marginal Effects Plots for GPT-4 on Codeforces

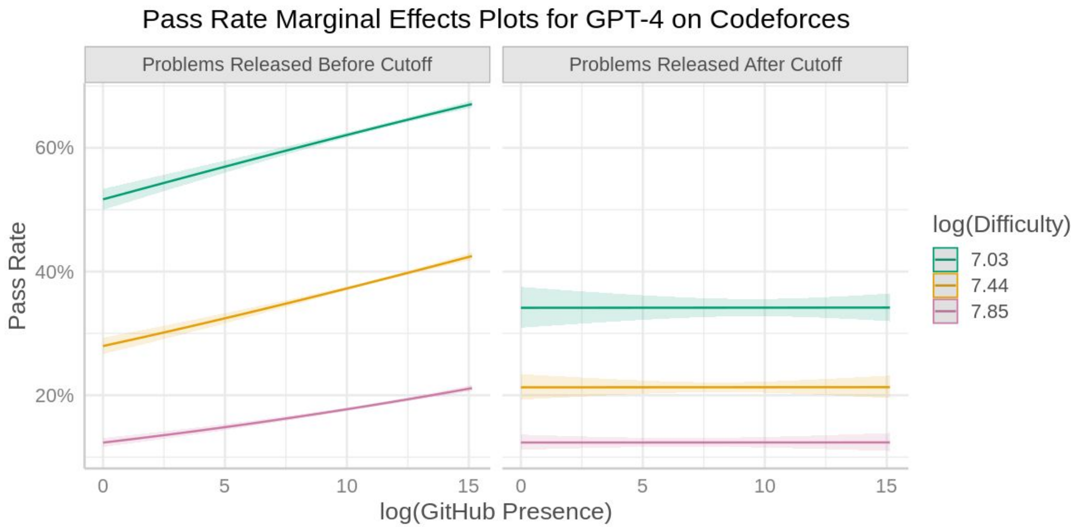


Pass Rate Marginal Effects Plots for GPT-3.5-Turbo on Codeforces



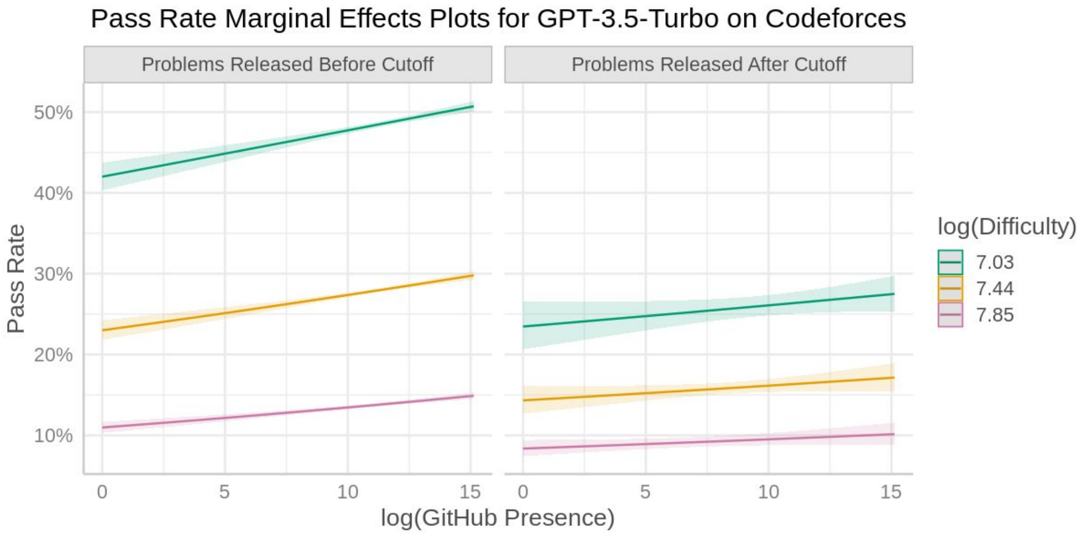
# Codeforces

Positive association


Pass Rate Marginal Effects Plots for GPT-4 on Codeforces

No association

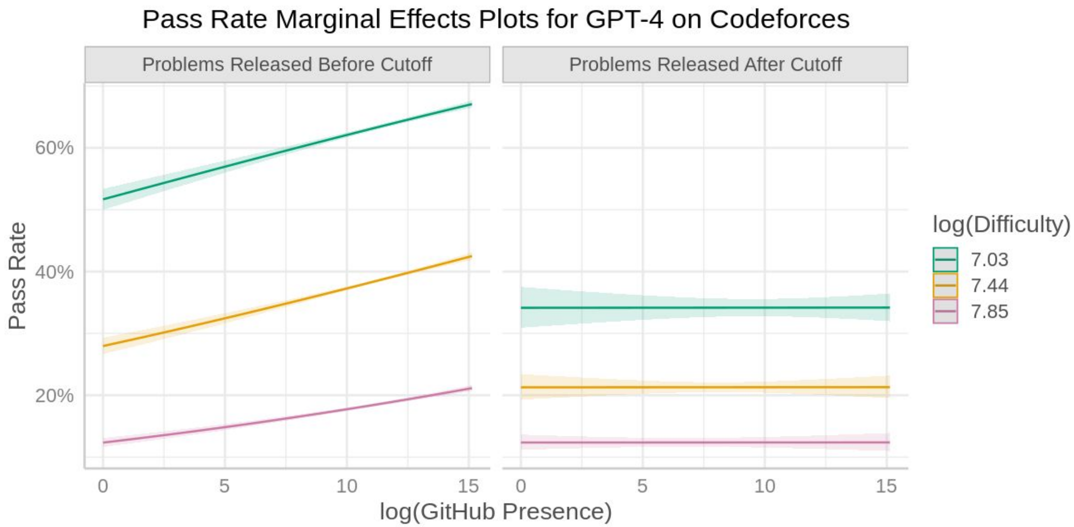
Positive association


Pass Rate Marginal Effects Plots for GPT-3.5-Turbo on Codeforces

No association

# Codeforces

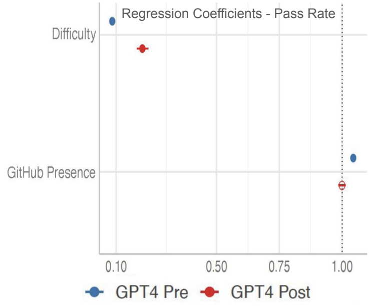
Positive association



Pass Rate Marginal Effects Plots for GPT-4 on Codeforces

Problems Released Before Cutoff

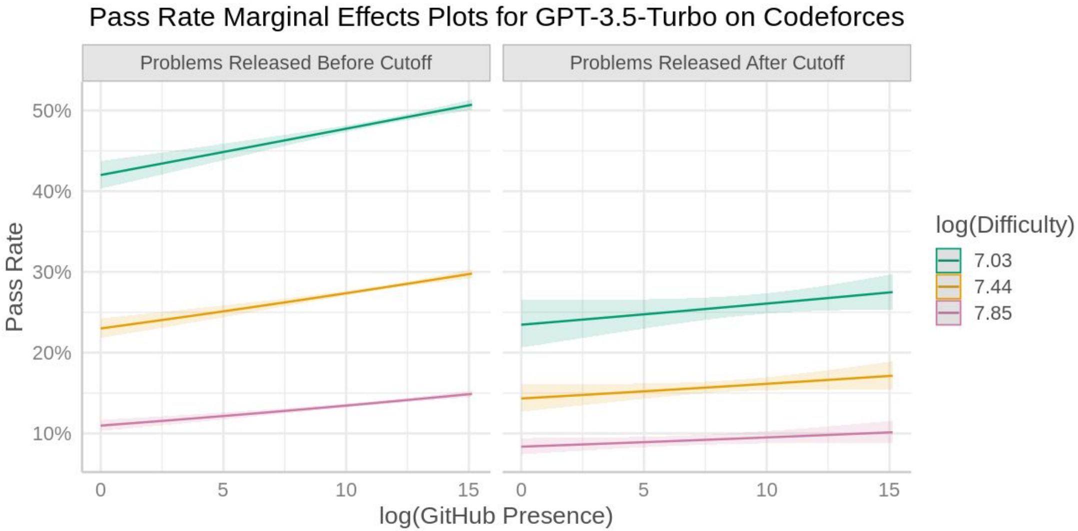
Problems Released After Cutoff



Regression Coefficients - Pass Rate

No association

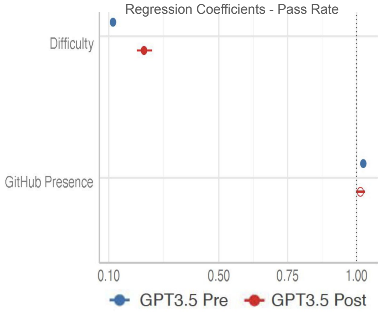
Positive association



Pass Rate Marginal Effects Plots for GPT-3.5-Turbo on Codeforces

Problems Released Before Cutoff

Problems Released After Cutoff

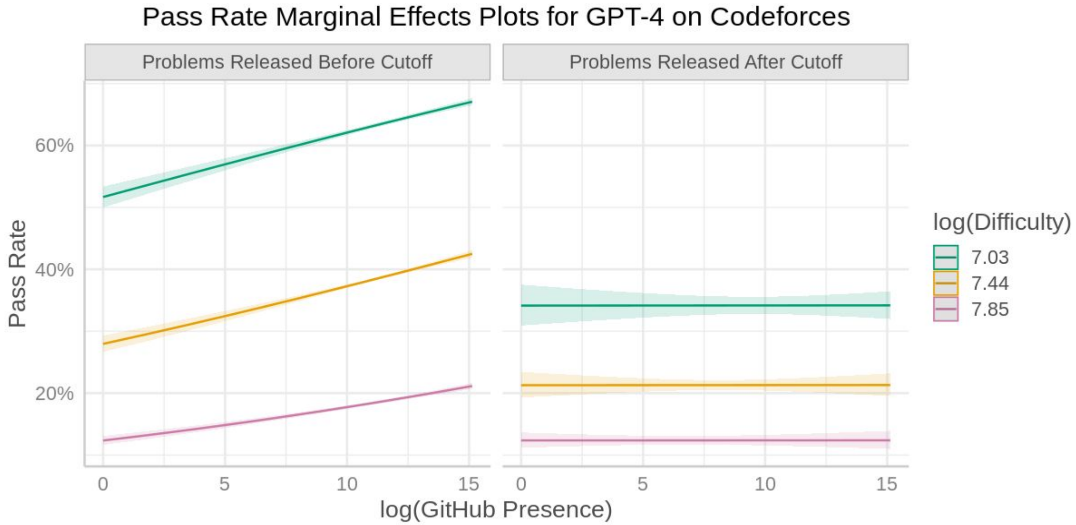


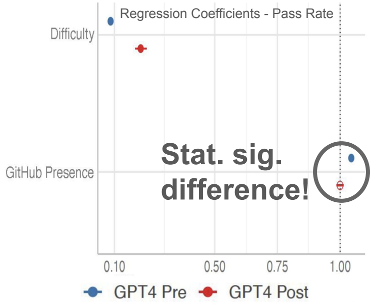
Regression Coefficients - Pass Rate

No association

# Codeforces

Positive association

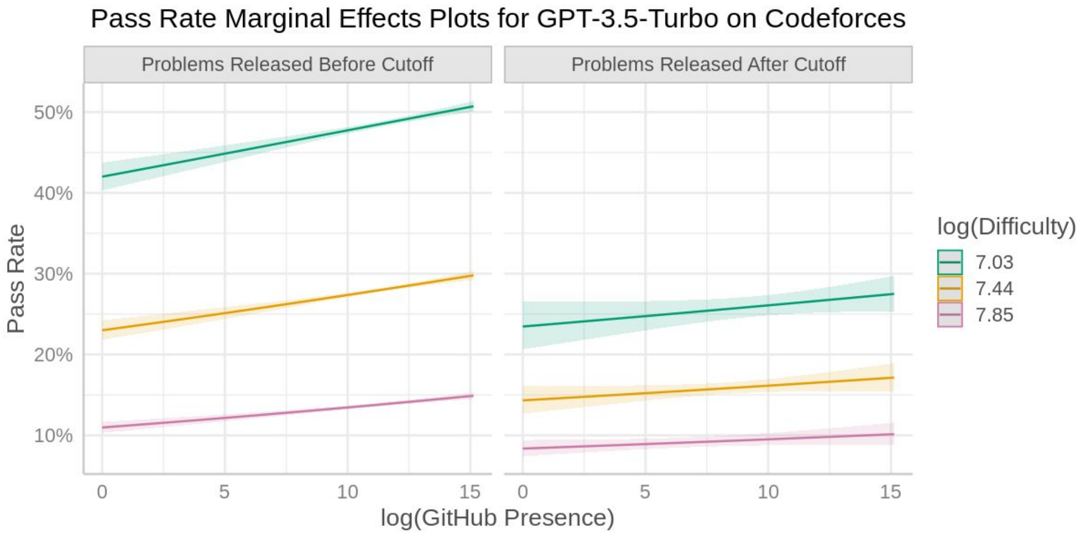

Pass Rate Marginal Effects Plots for GPT-4 on Codeforces

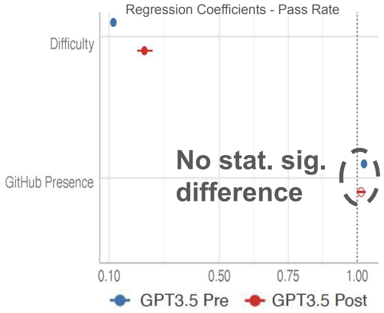


Stat. sig. difference!

No association

Positive association


Pass Rate Marginal Effects Plots for GPT-3.5-Turbo on Codeforces

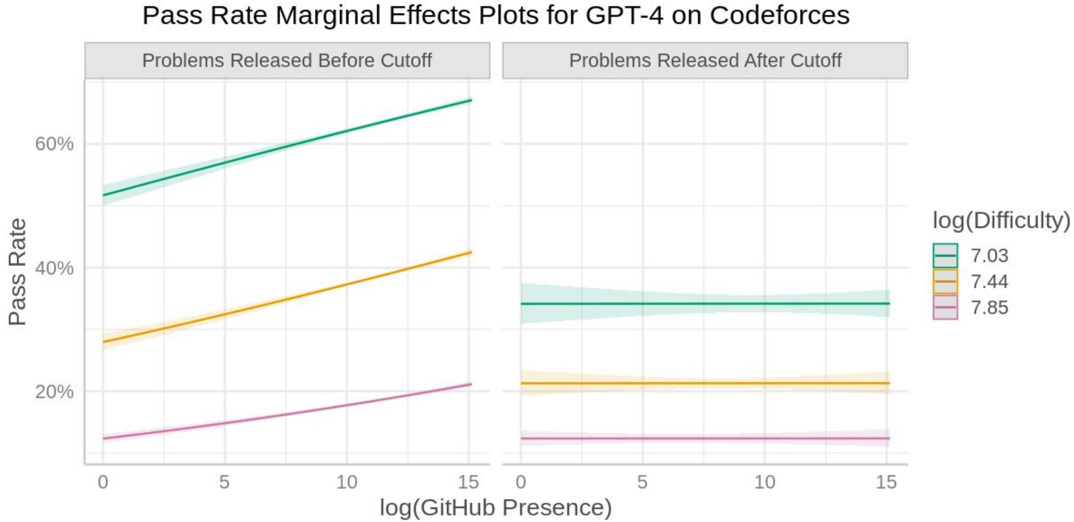


No stat. sig. difference

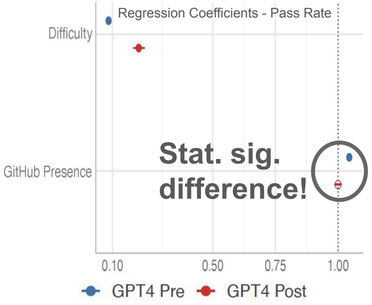
No association

# Codeforces

Positive association



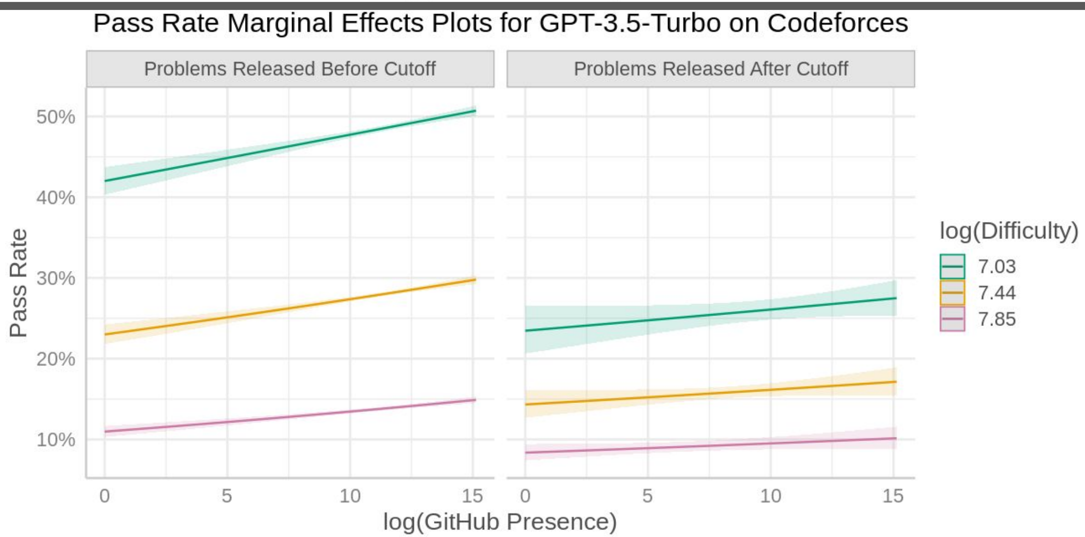
## Contamination



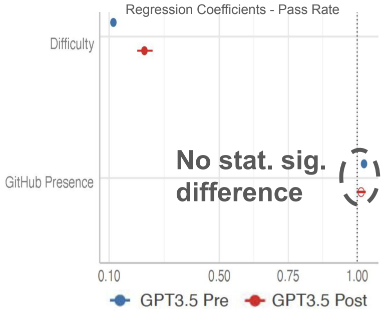
**Stat. sig. difference!**

No association

Positive association



## No conclusion

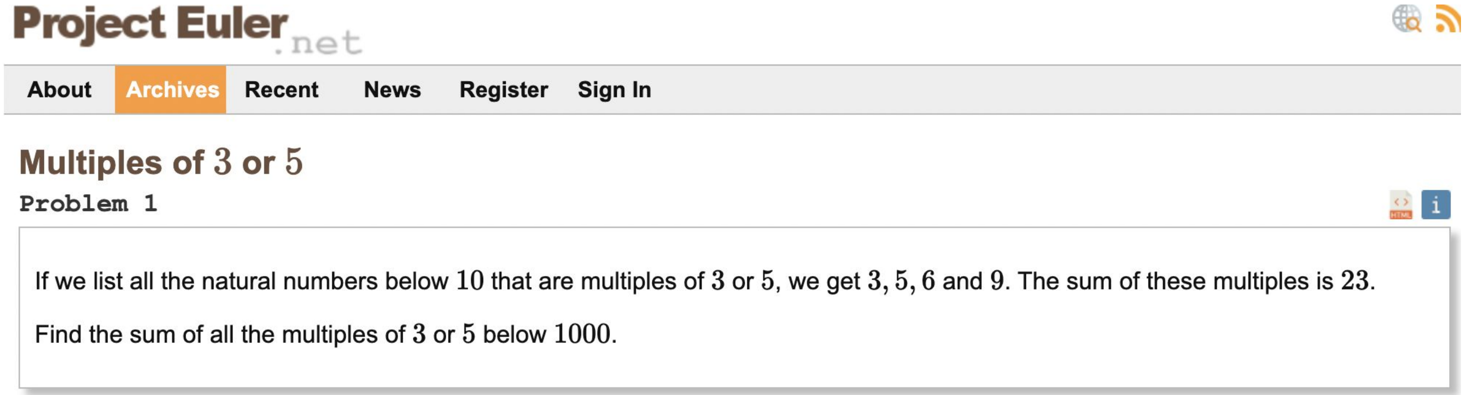


**No stat. sig. difference**

No association

# Project Euler

- (Typically) difficult math problems from 2001-present
- String answers are usually a single number.
- Users write code to find the answer; we instead ask for the exact solution.
  - Pass Rate: 1 if output matches ground truth, else 0.
- Problem Counts:
  - 765 before GPT cutoff, 72 after GPT cutoff.



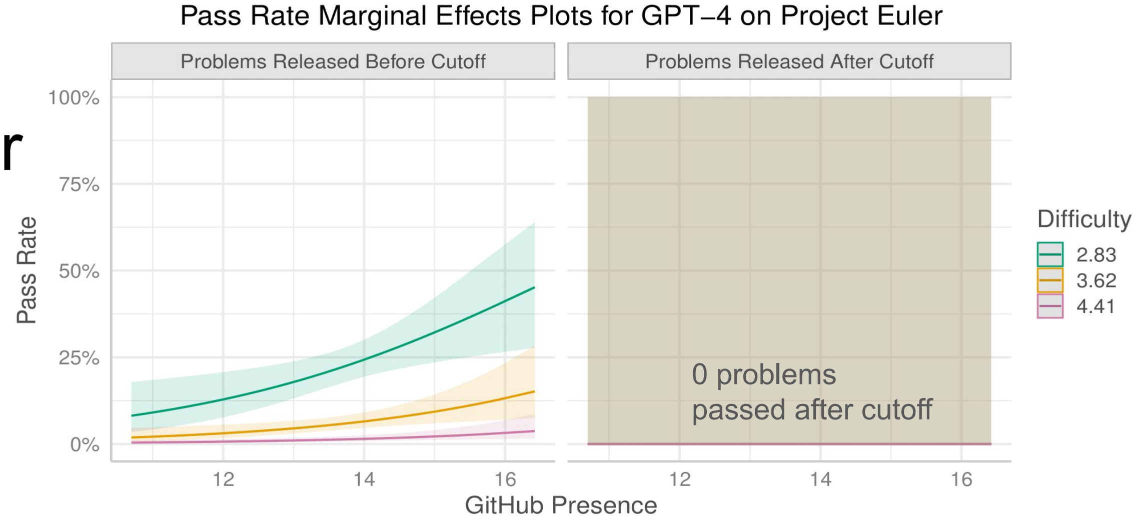
Multiples of $3$ or $5$

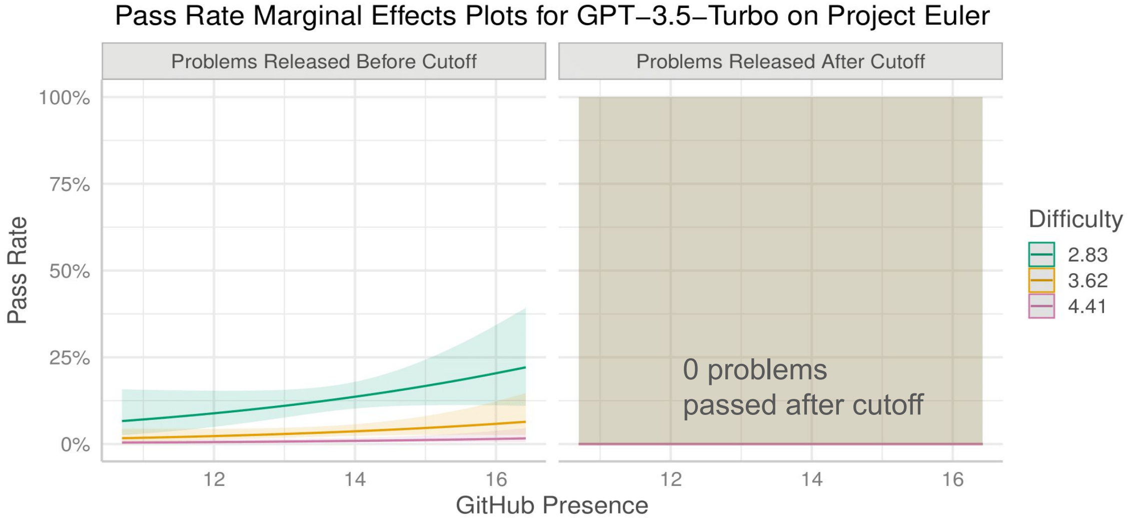
Problem 1

If we list all the natural numbers below $10$ that are multiples of $3$ or $5$, we get $3, 5, 6$ and $9$. The sum of these multiples is $23$.

Find the sum of all the multiples of $3$ or $5$ below $1000$.

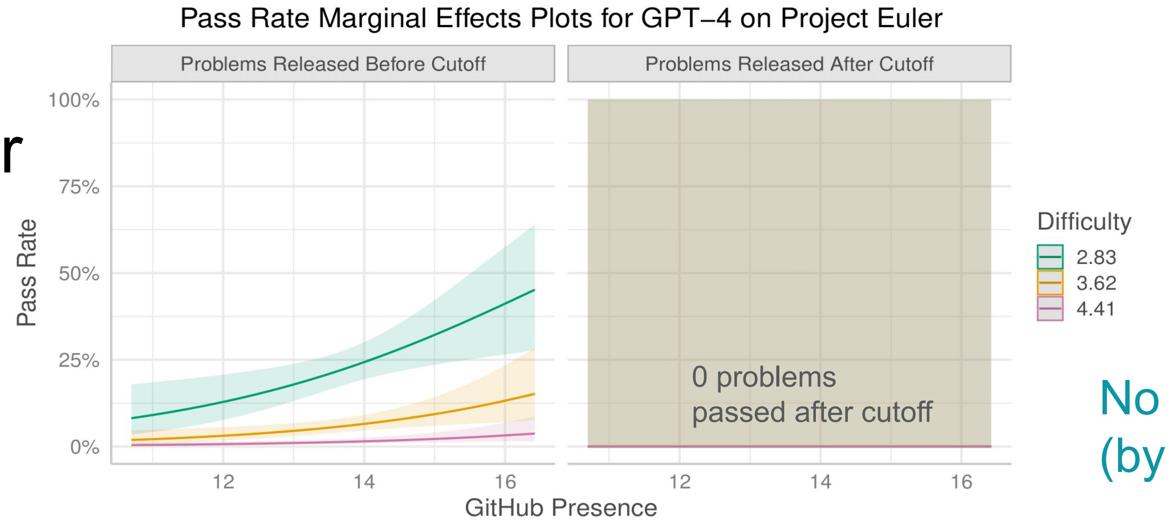
# Project Euler





# Project Euler

Positive association

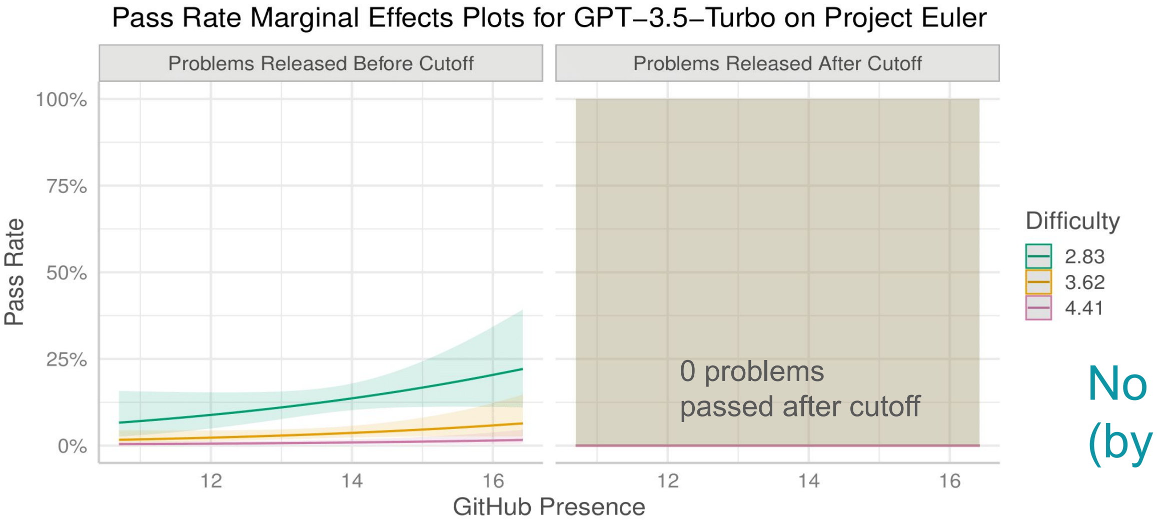


Pass Rate Marginal Effects Plots for GPT–4 on Project Euler

0 problems passed after cutoff

No association (by default)

No association



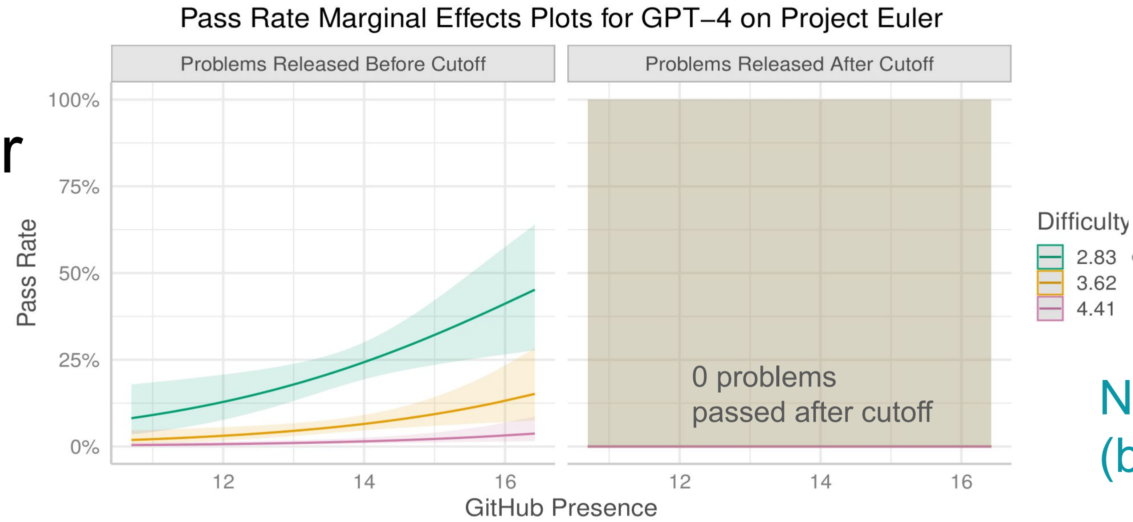
Pass Rate Marginal Effects Plots for GPT–3.5–Turbo on Project Euler

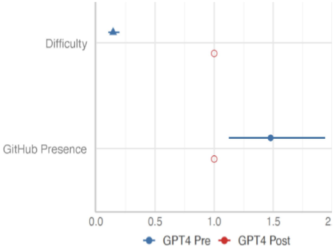
0 problems passed after cutoff

No association (by default)

# Project Euler

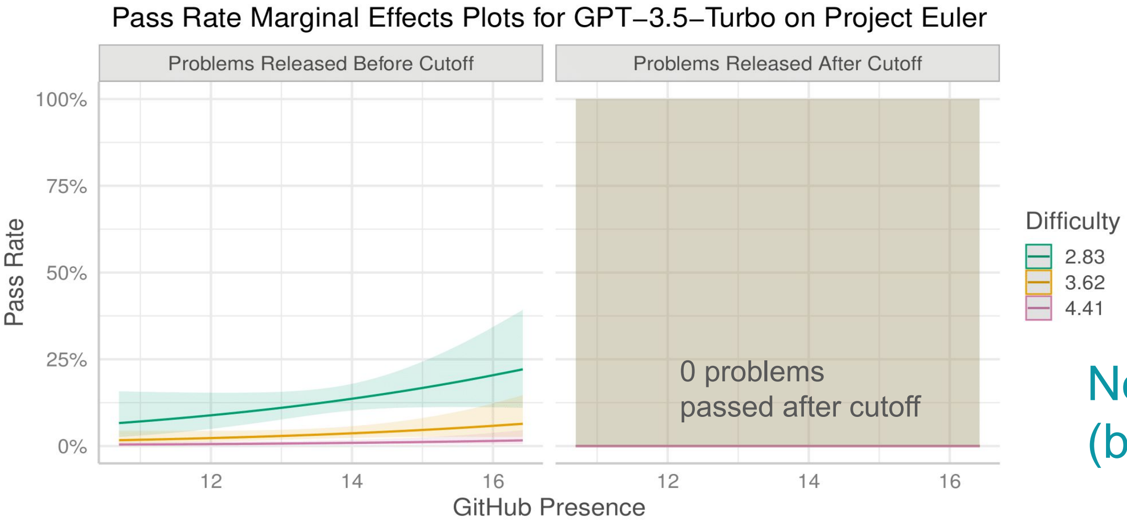
Positive association

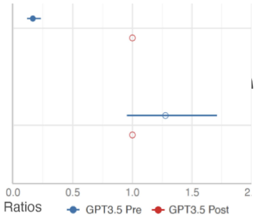




No association (by default)

No association

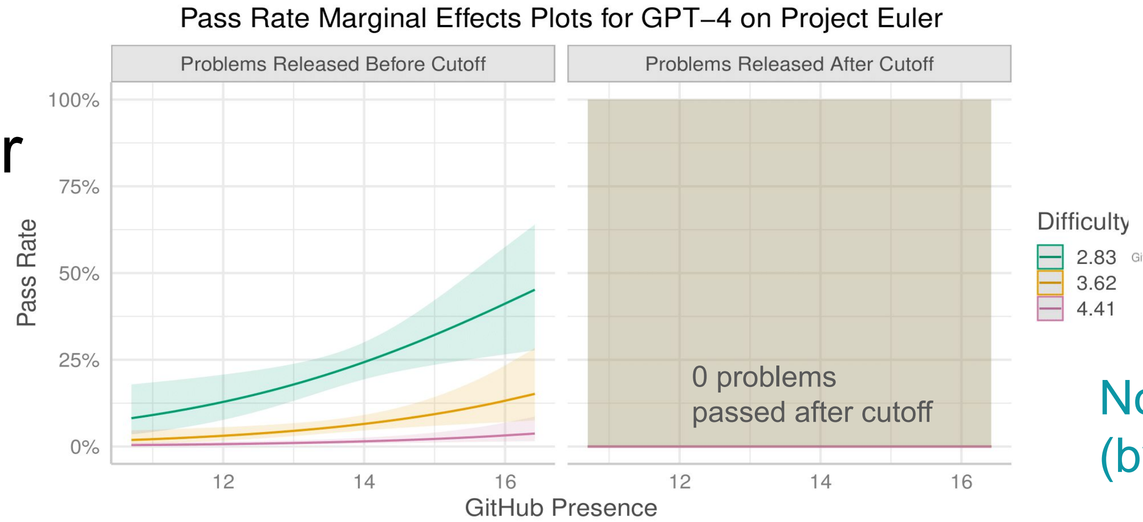




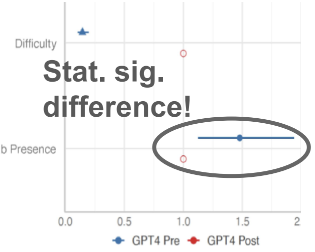
No association (by default)

# Project Euler

Positive association



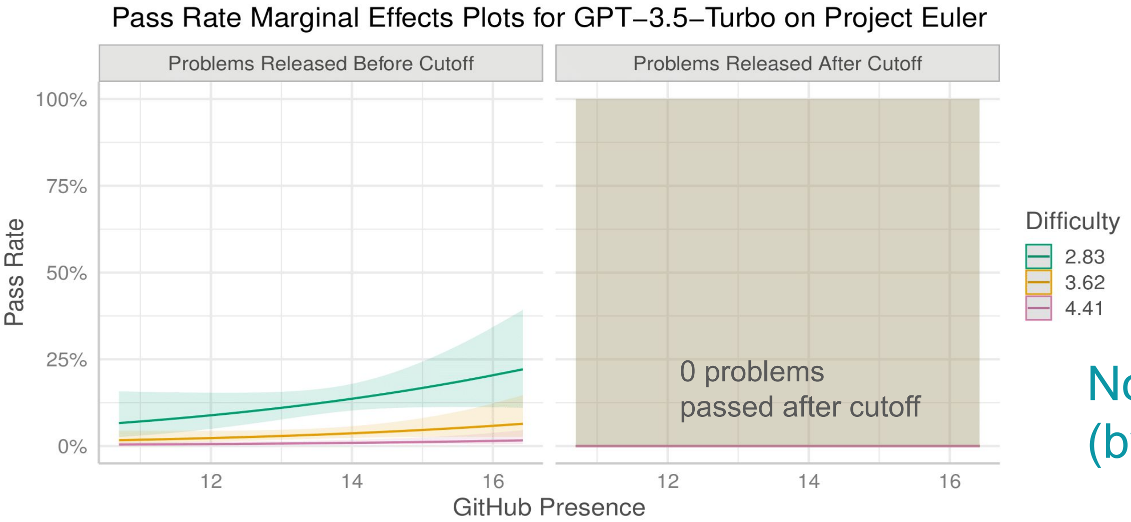
0 problems passed after cutoff



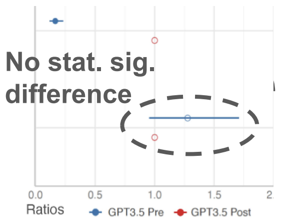
**Stat. sig. difference!**

No association (by default)

No association



0 problems passed after cutoff



**No stat. sig. difference**

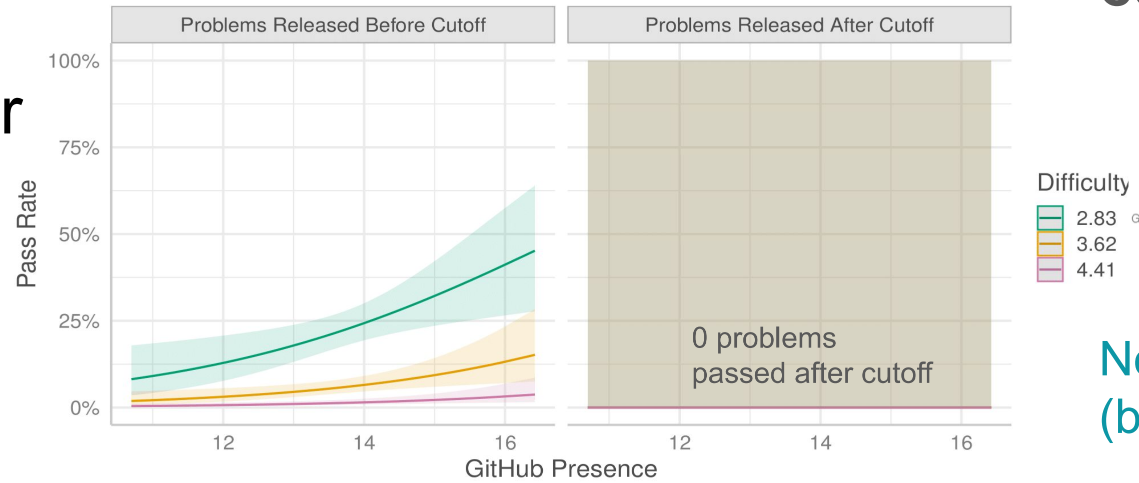
No association (by default)

# Project Euler

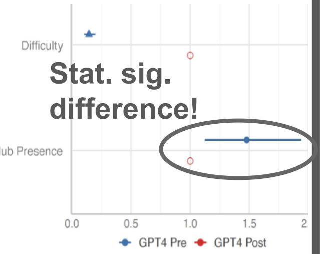
## Contamination

Positive association

Pass Rate Marginal Effects Plots for GPT−4 on Project Euler



0 problems passed after cutoff



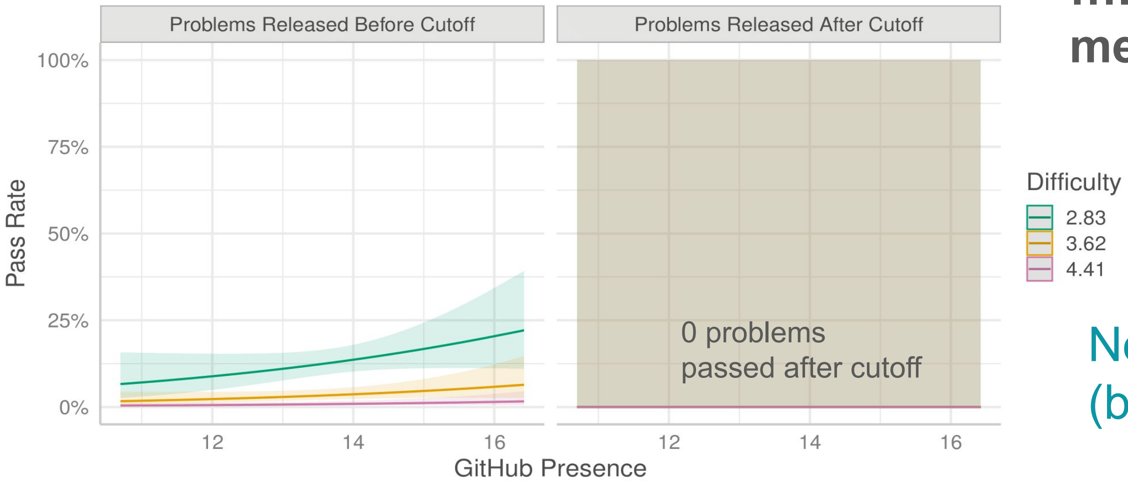
Stat. sig. difference!

No association (by default)

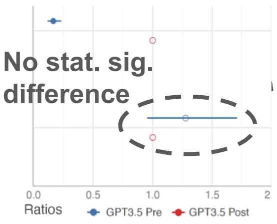
## …Limited memorization

No association

Pass Rate Marginal Effects Plots for GPT−3.5−Turbo on Project Euler



0 problems passed after cutoff

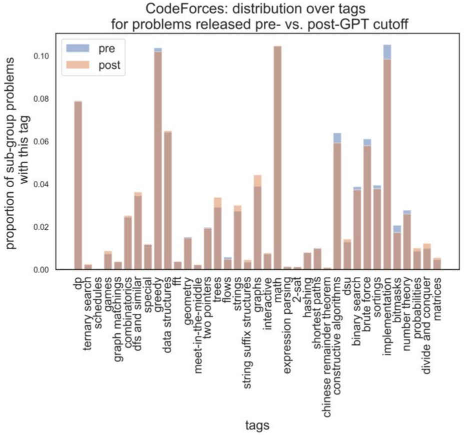


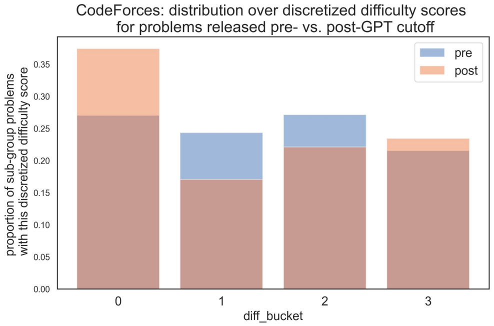
No stat. sig. difference

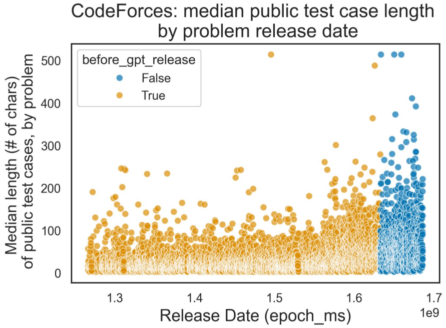
No association (by default)

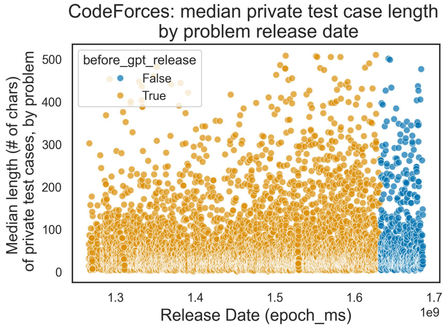
# Distribution Shift?

- Certainly, as in all natural datasets collected over time.
- Does it affect the conclusions?
  - Project Euler: because pass rate is positive **at all**, we are fairly sure memorization did occur.
  - Codeforces: many shifts (see charts below + appendices). Hard to detangle all possibilities, but the magnitude of effect on GPT-4 suggests robustness of conclusion.


CodeForces: distribution over tags for problems released pre- vs. post-GPT cutoff


CodeForces: distribution over discretized difficulty scores for problems released pre- vs. post-GPT cutoff


CodeForces: median public test case length by problem release date


CodeForces: median private test case length by problem release date

# Takeaways

- Method for contamination detection in black-box models
- Contamination in modern models
  - Likely contamination of Project Euler & Codeforces in GPT-4
- Reproducibility & extensibility
  - Scraped datasets and toolkit in repo
  - Raw results as CSV in repo
- Life lessons?
  - Release new benchmarks
  - Trickle benchmarks over time

# To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination

ABACUS.AI

UNIVERSITY OF MARYLAND

bit.ly/to-the-cutoff

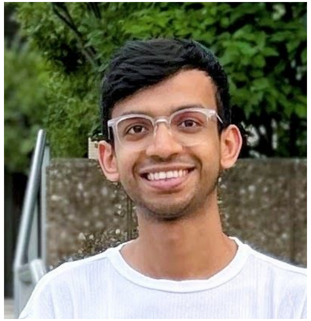
Code & paper link

Our paper has more analyses (e.g. title + tag reproduction metrics, regression coefficients, ablations).

## Thanks, wonderful collaborators!

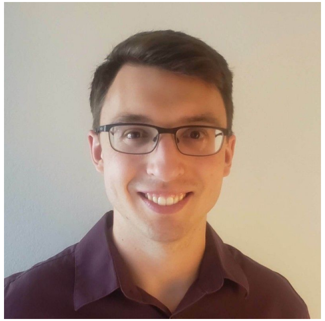
(speaking!)

Manley Roberts



Himanshu Thakur

Christine Herlihy



Colin White

Samuel Dooley