Cut Costs, Not Accuracy: LLM-Powered Data Processing with Guarantees

Sepanta Zeighami, Shreya Shankar, Aditya Parameswaran

LLMs for Data Processing

***Example**: game reviews dataset*

Elden Ring is a monumental achievement in open-world …

Stardew Valley is the ultimate cozy game, much better than *Animal Crossing* …

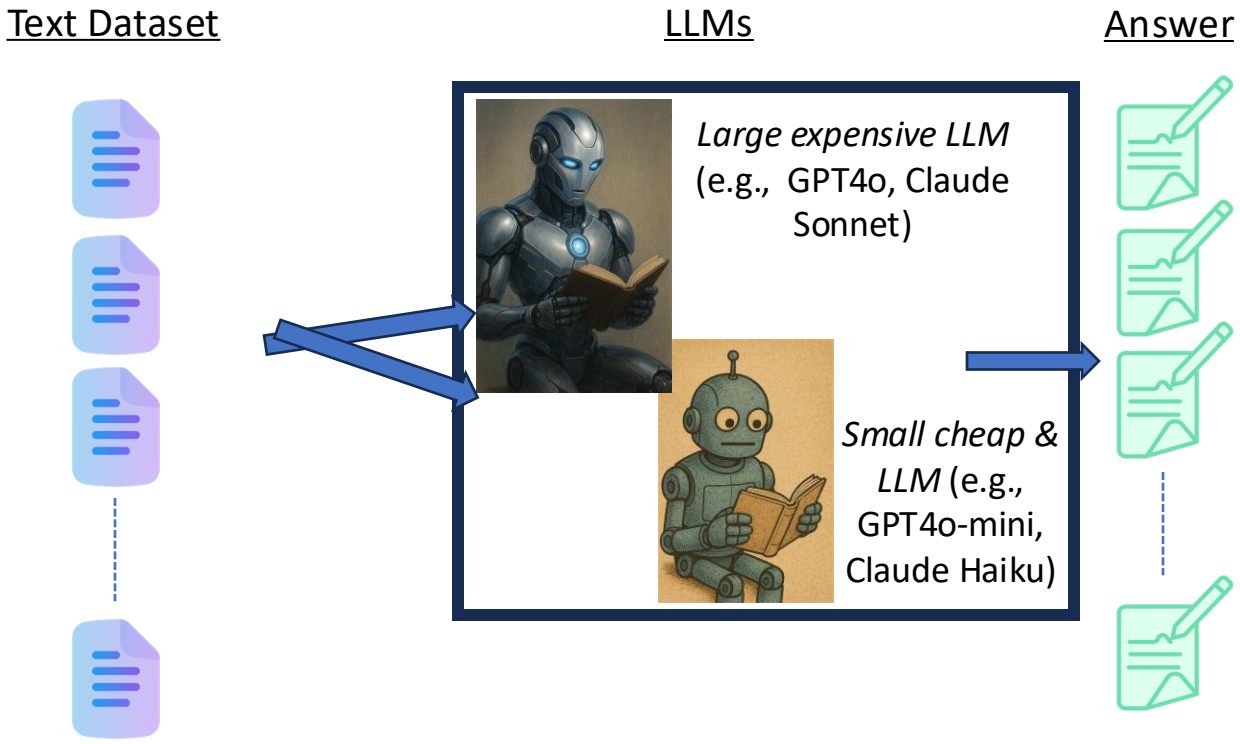
***Prompt for processing**: does the review mention another game more positively?*

User wants outputs from accurate top-of-the-line (e.g., gpt-4o) LLM but it's too expensive

Other examples

- Extract name of police officers with misconduct from court cases
- Find legal contracts that rely on a particular law

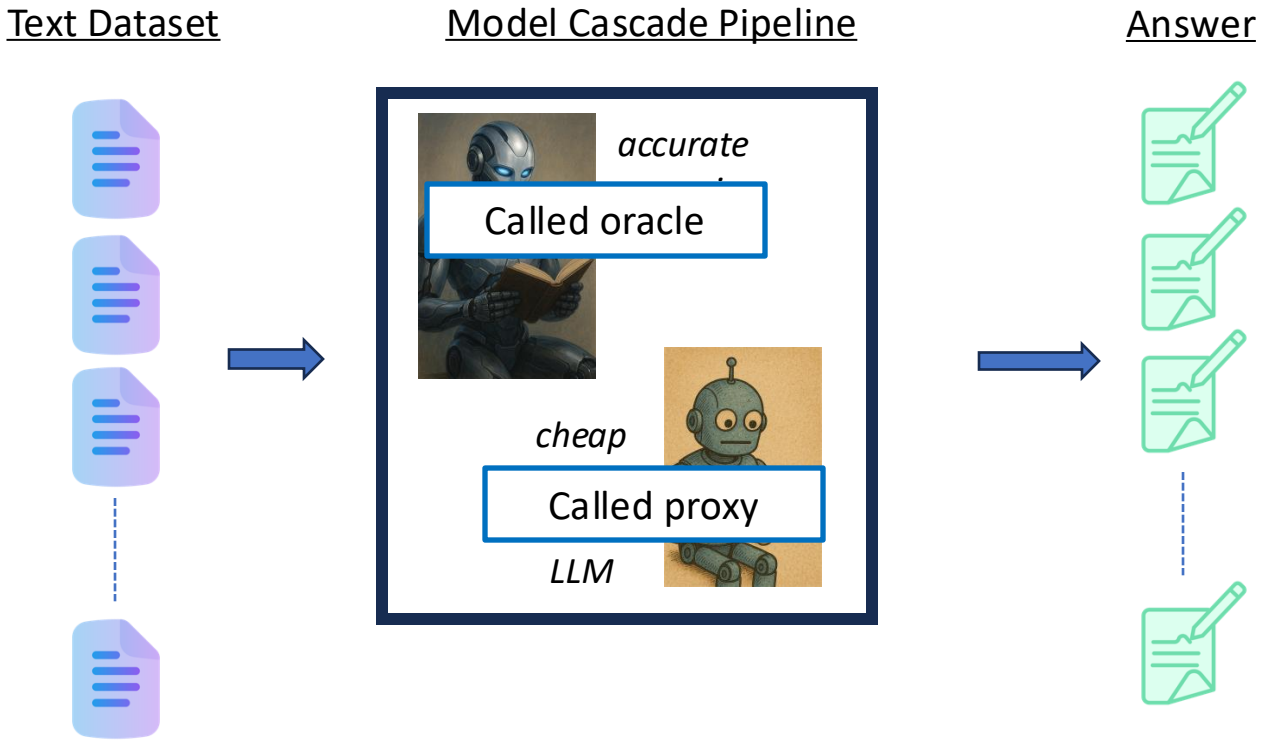
Low-Cost LLM-Powered Data Processing



User wants outputs from accurate top-of-the-line (e.g., gpt-4o) LLM but it's too expensive

We can use cheap potentially inaccurate LLM (e.g., gpt-4o-mini) as long as we provide **guarantee most outputs are the same as the top-of-the-line LLM**

Low-Cost Pipelines Through Model Cascade



Process all records while **minimizing number of times we use the oracle** and guaranteeing **output matches oracle's on at least T% of documents** with high probability

Users provides a failure probability

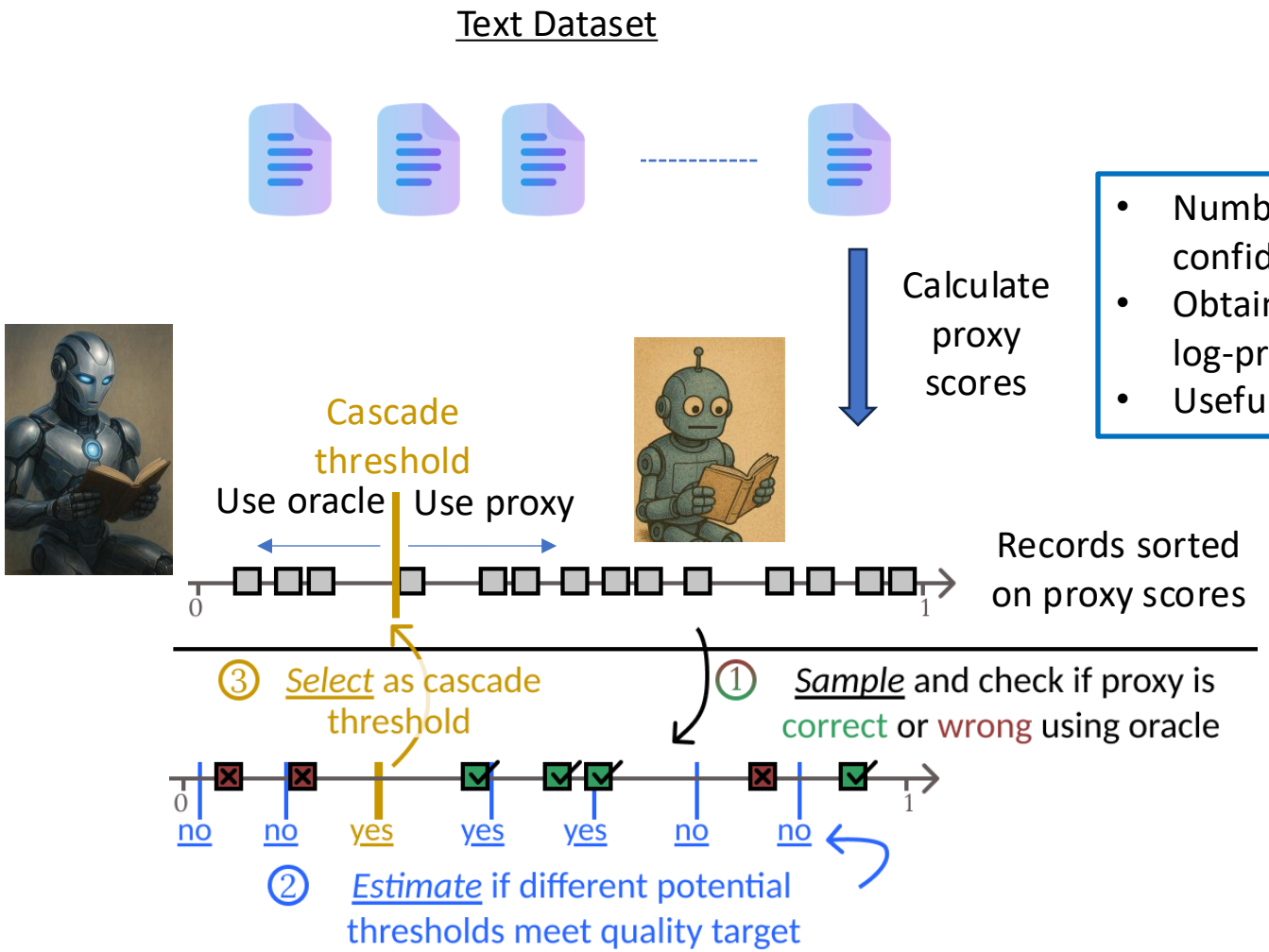
User-provided quality target

We consider the oracle answer as the correct answer

Cost of proxy is negligible compared to oracle (in practice an order of magnitude difference)

A common framework to decide when to use which model is *model cascade*

Mode Cascade Framework

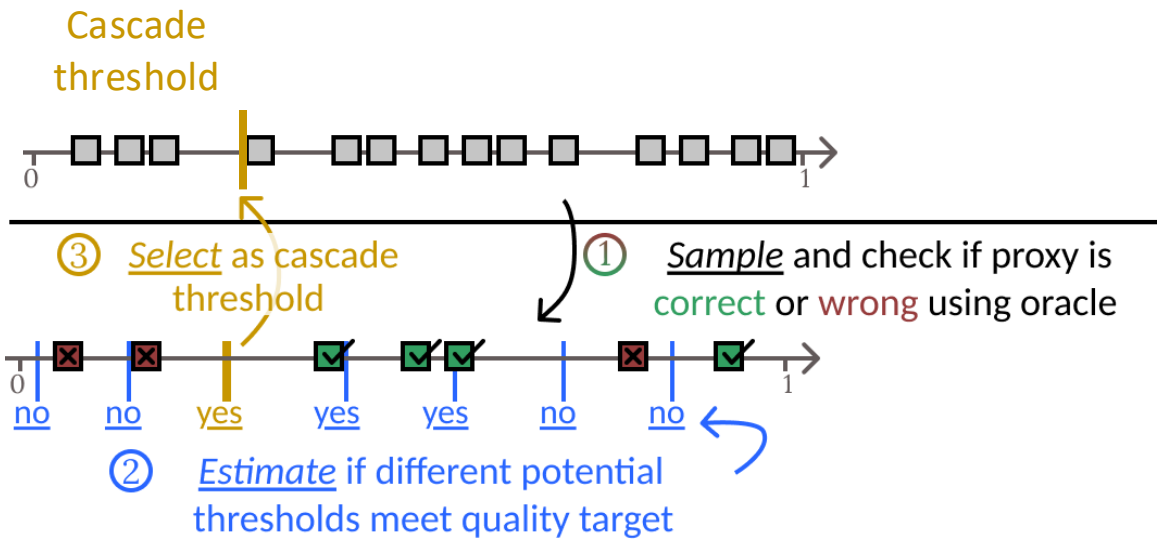


- Numbers in [0, 1] quantifying proxy model's confidence in its answers
- Obtain through proxy inference (e.g., LLM log-probabilities)
- Useful if generally correlated with accuracy

Cascade Threshold Problem

Determine a threshold **to minimize number of times we use the oracle** while **guaranteeing output matches oracle's on at least T% of documents** with high probability

Overview of Existing Approaches



Existing method designed to solve variations of the problem

Naive application of statistical tools

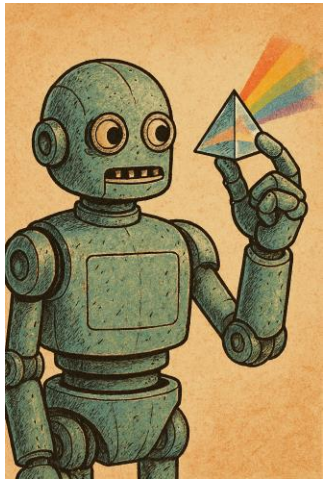
Method
SUPG[1]
Naive
PRISM

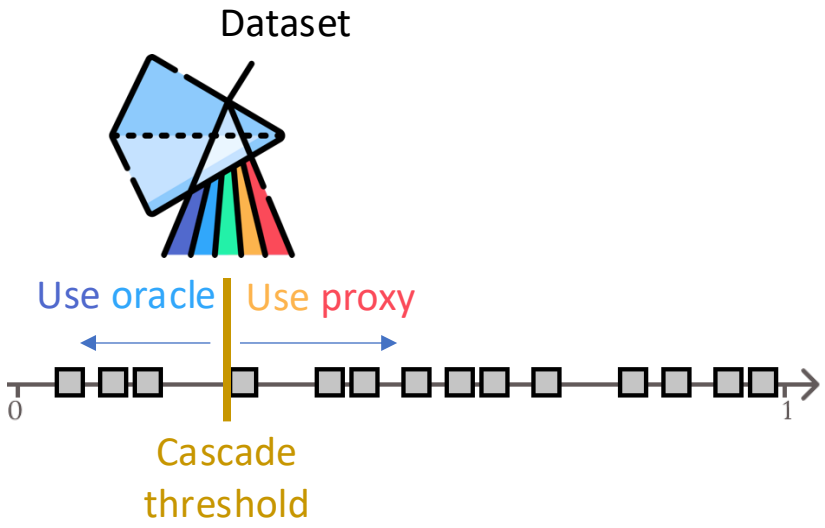
Weak guarantees fail to meet target in practice

Lead to little cost saving

1: Kang et al VLDB 2020

PRISM

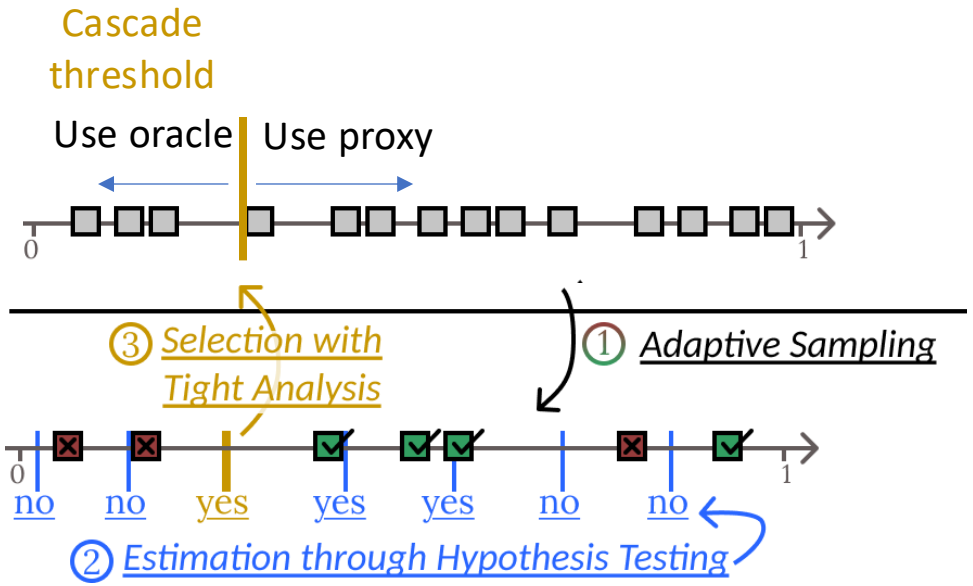


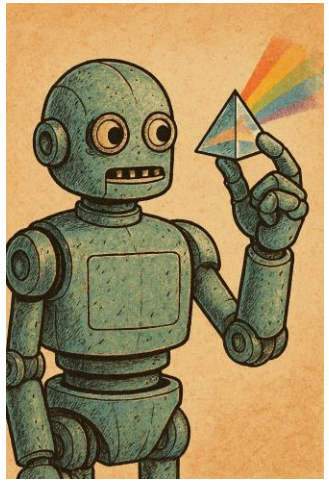


PRSIM determines a cascade threshold to **save cost** while providing **theoretical guarantees on meeting the accuracy target**

Takes *data and task characteristics* into account to provide high-utility

PRISM Overview





Adaptive sampling

- Iteratively samples records
- Considers label distribution and the accuracy target

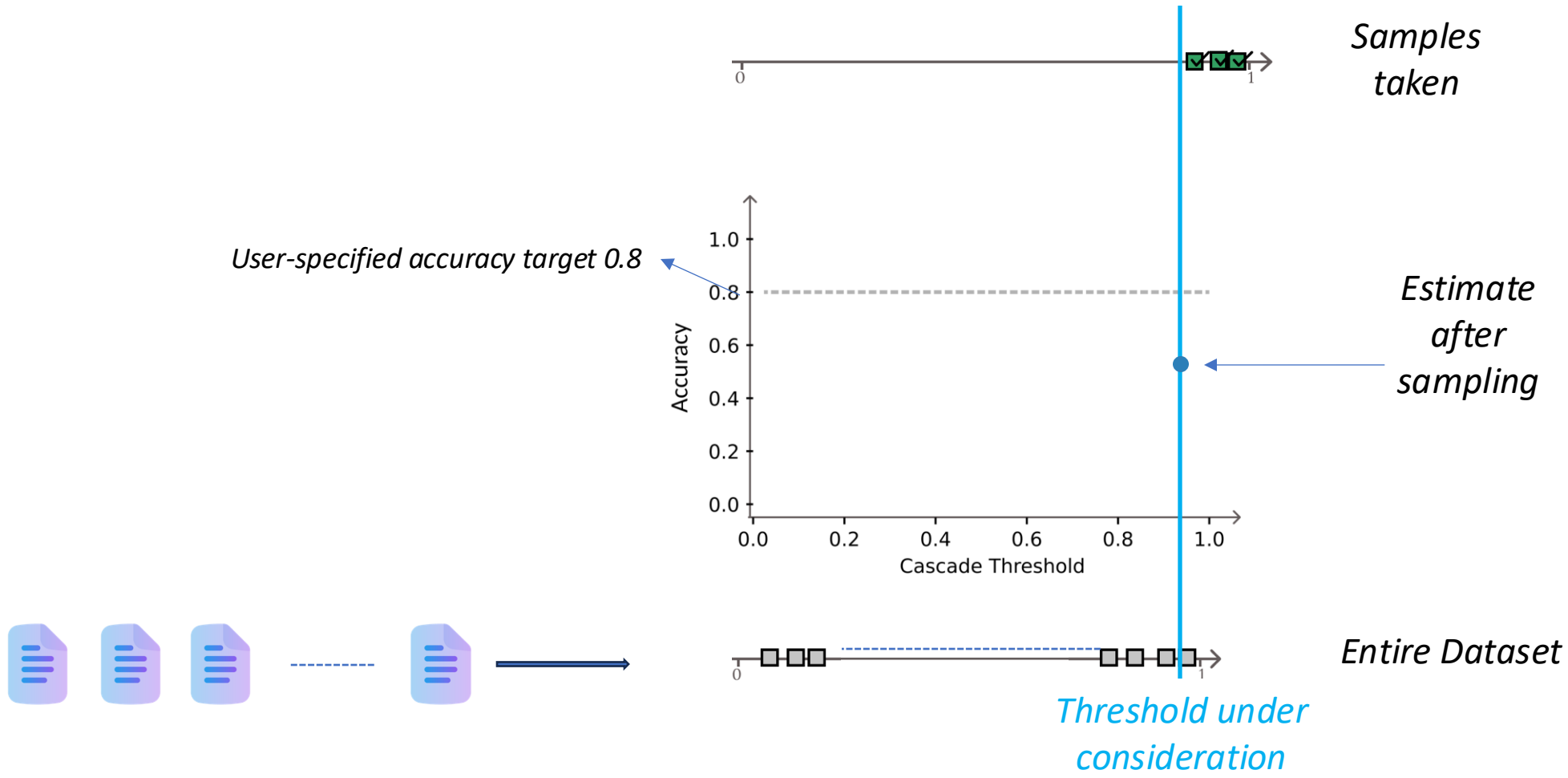
Estimation through Hypothesis Testing

- Uses recent statistical tools
- Improves accuracy when variance is low

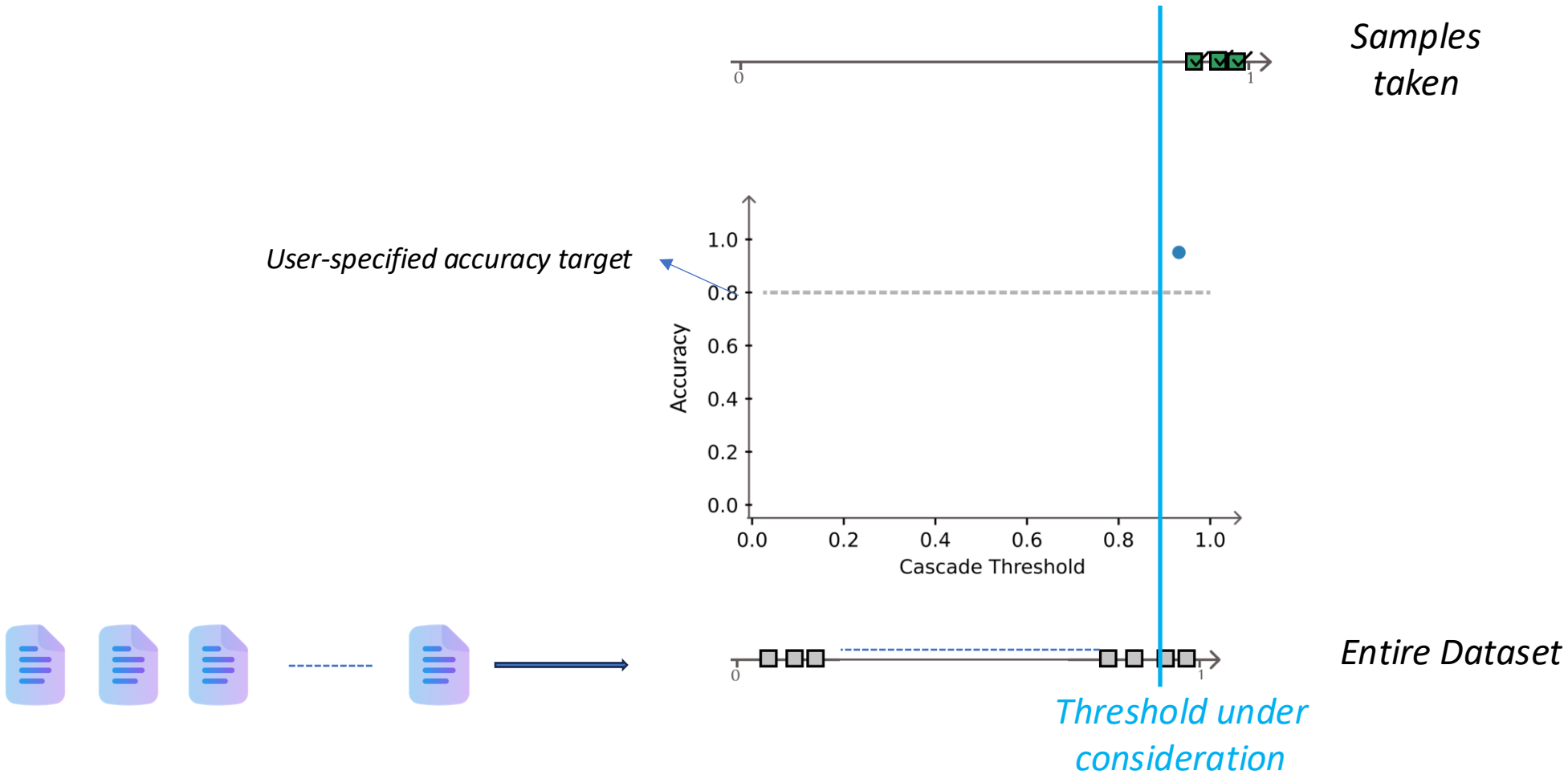
Tight Analysis of Threshold Selection

- Analysis of probability of not meeting the target
- Takes data characteristics into account

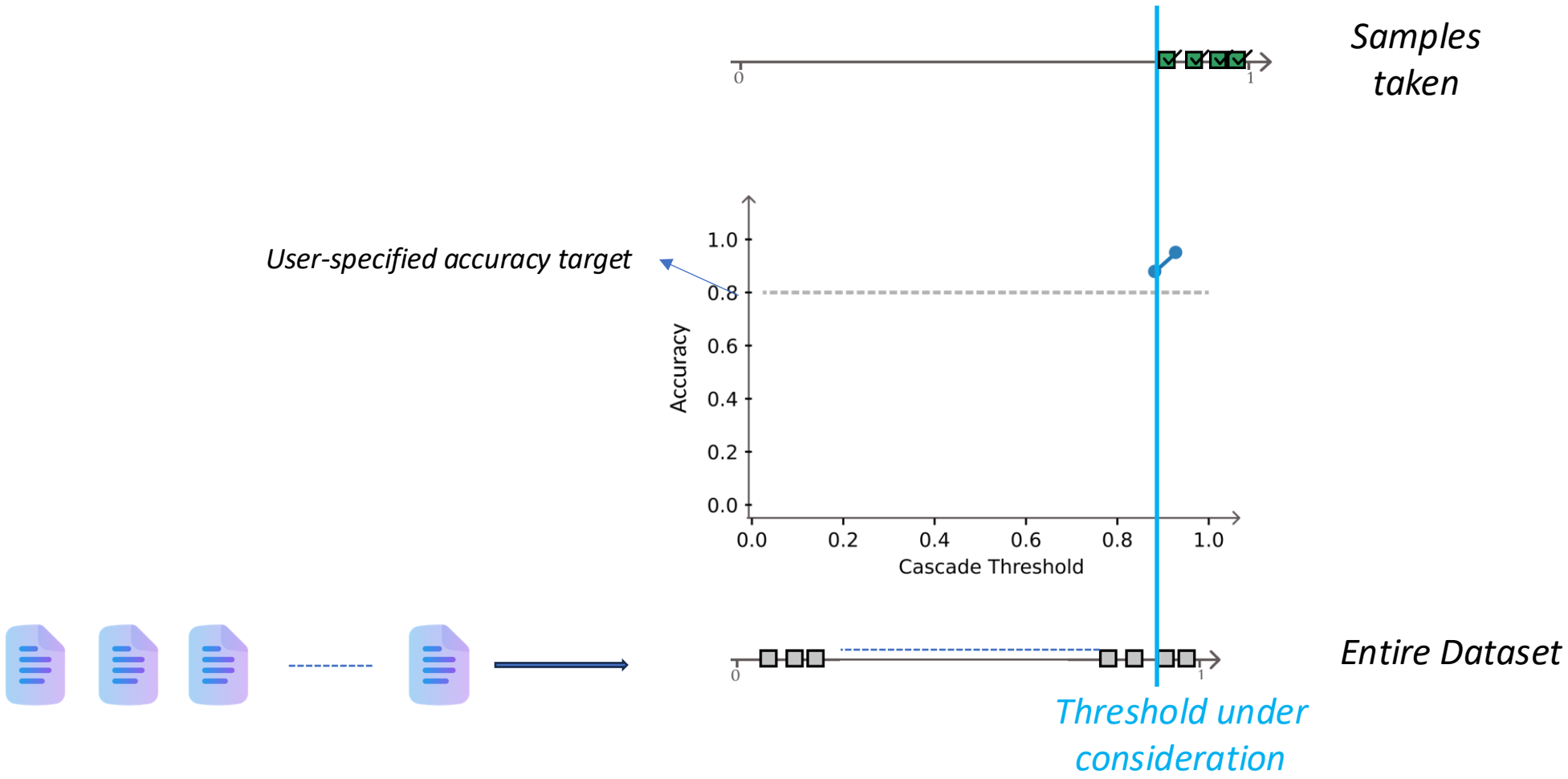
PRISM Algorithm



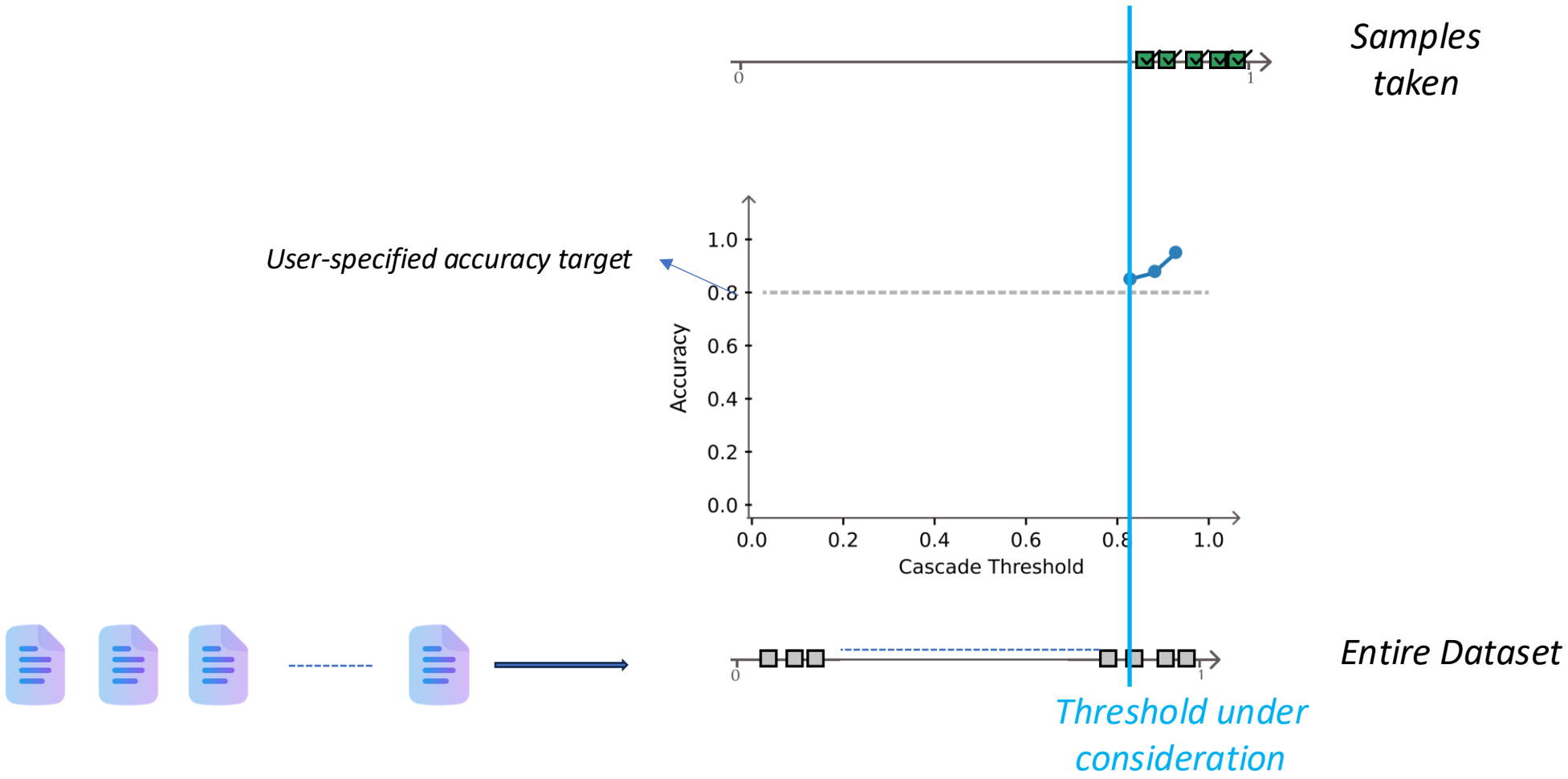
PRISM Algorithm



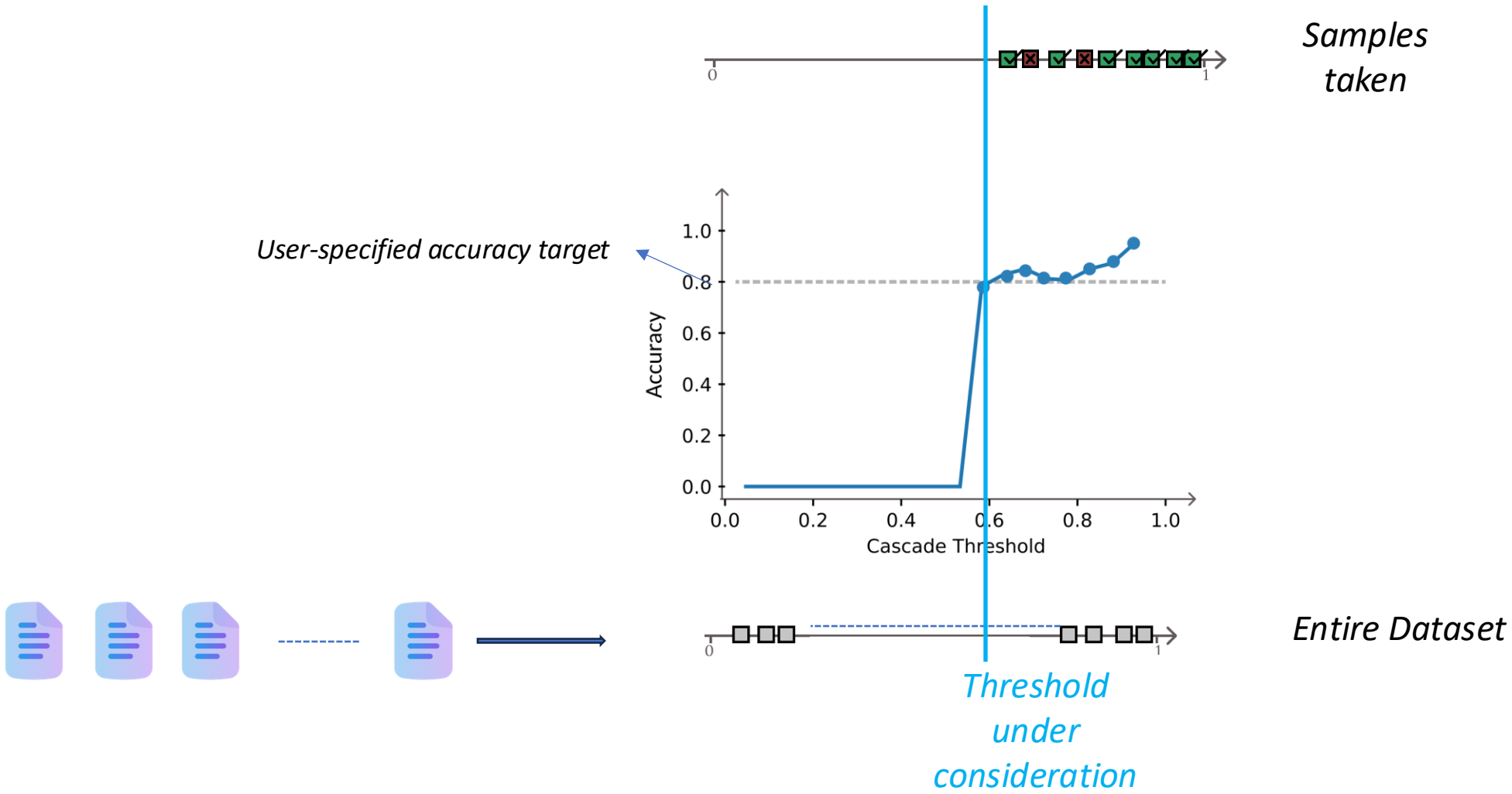
PRISM Algorithm



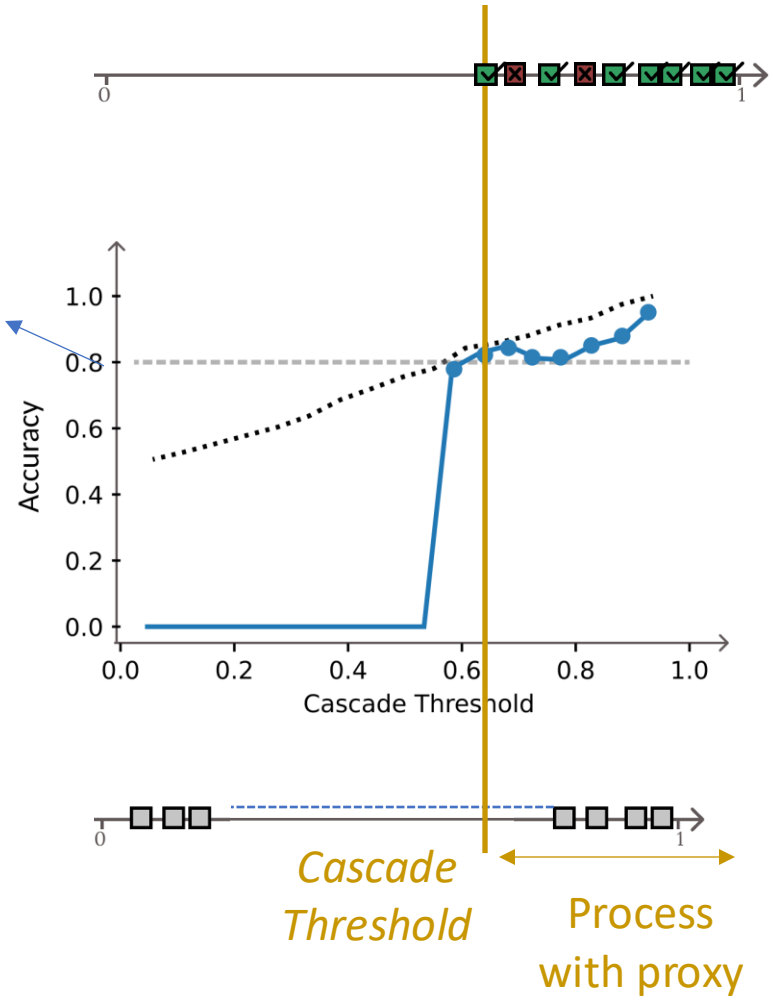
PRISM Algorithm

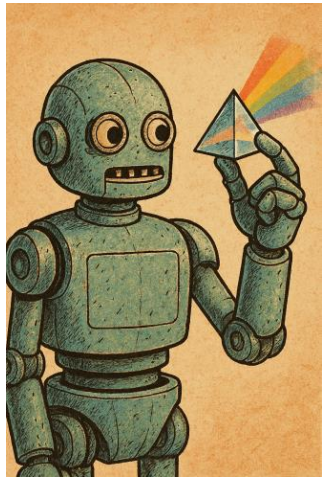


PRISM Algorithm

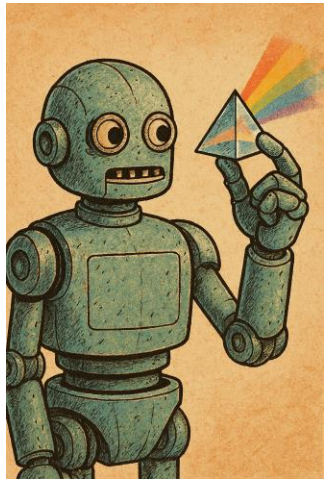


PRISM Algorithm





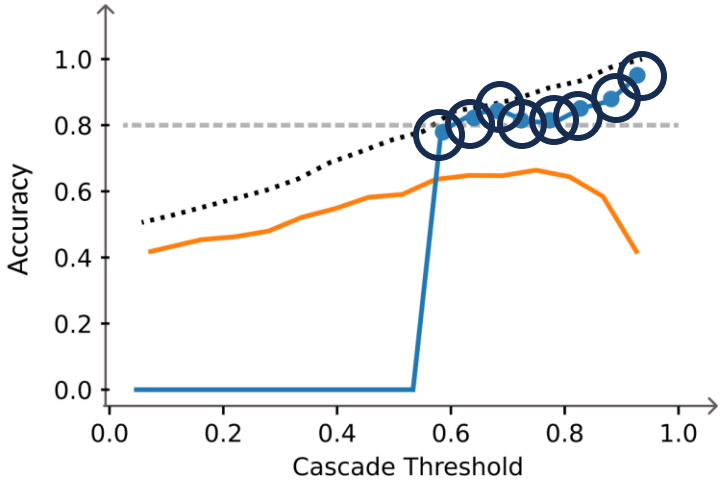
Further Details



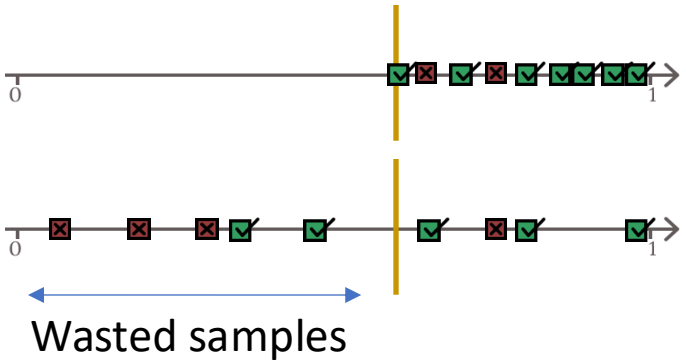
PRISM also provides guarantees if the user is interested in precision or recall instead of accuracy

Estimation through Hypothesis Testing

Tight Analysis of Threshold Selection

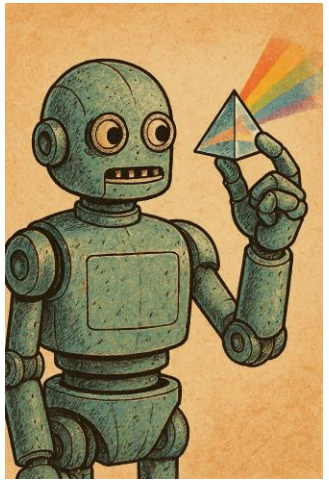


Further Details



Samples taken by PRISM

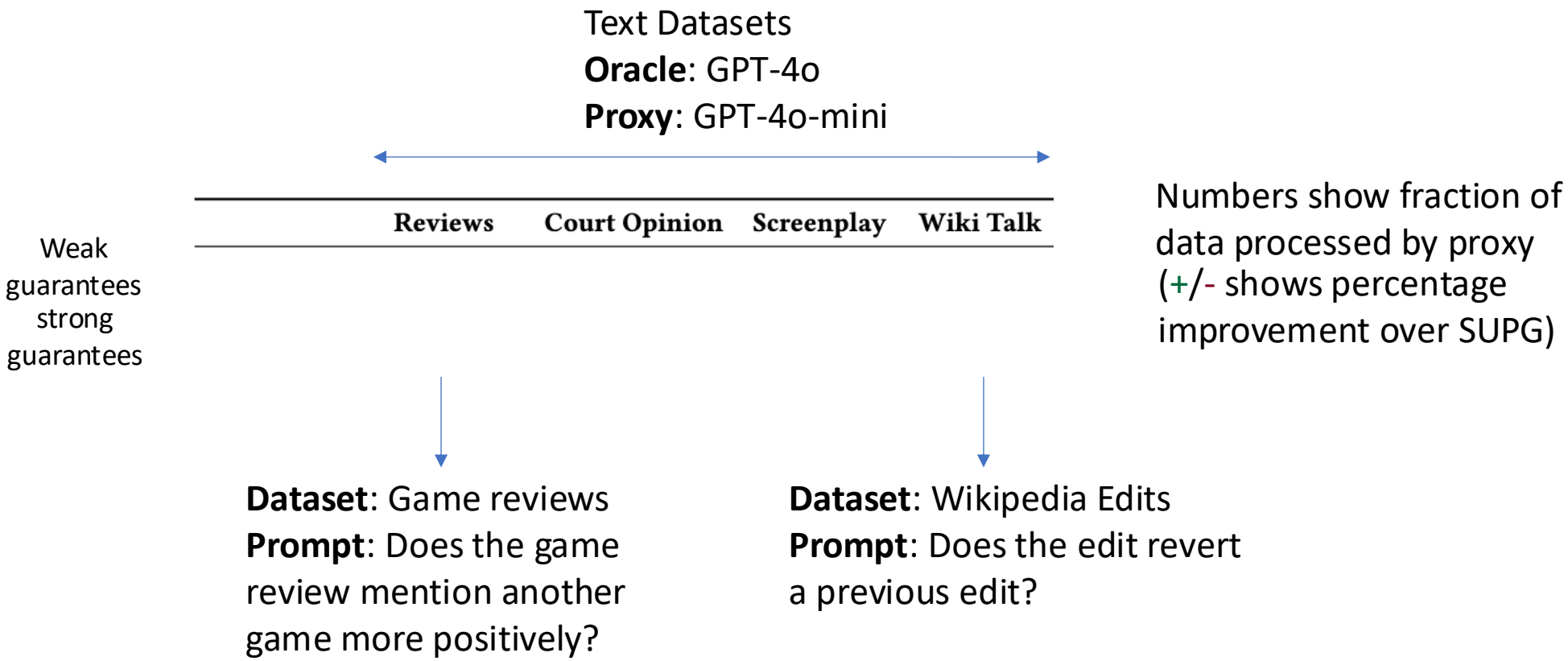
Samples taken by Naive and SUPG



	Adaptive sampling	**Estimation through Hypothesis Testing**	**Tight Analysis of Threshold Selection**
PRISM	✓	✓	✓
SUPG	✗	✗	✗
Naive	✗	✗	✗

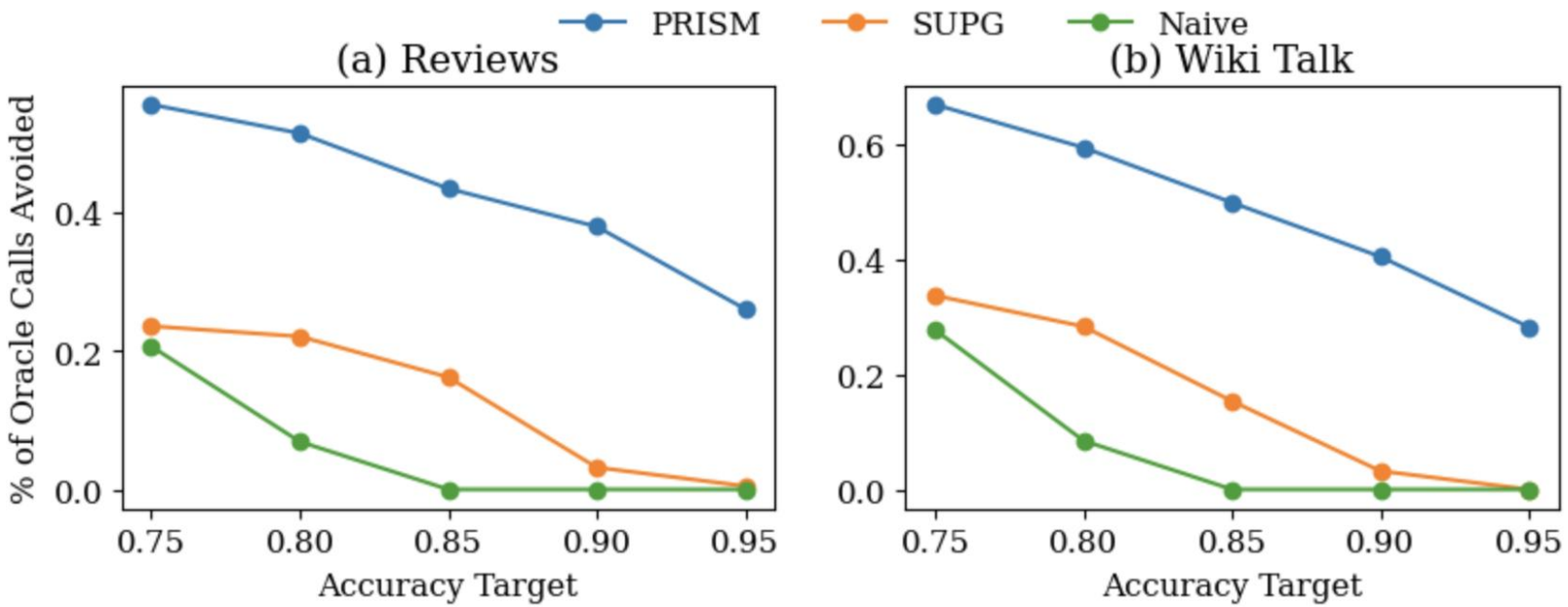
Results

- Accuracy target: 90%
- Probability of failure: 10%



Results Across Accuracy Targets

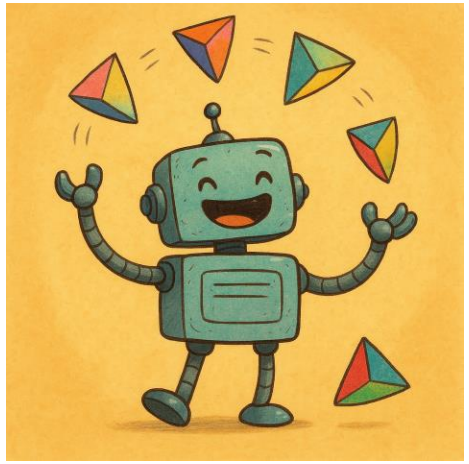
Similar gains when given precision and recall



Summary

- PRISM helps reduce costs while processing text data with LLMs
  - Uses cheaper models as much as possible for the specific dataset and query
  - On average 86% lower cost, 118% higher recall and 19% higher precision over state-of-the-art

- Provides theoretical guarantees to avoid performance regression when using small models



Thanks! Q&A

PRISM Algorithm