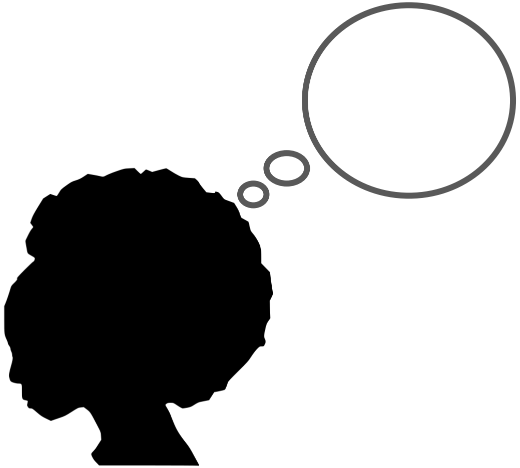
Learning to Make Decisions from Few Examples

Emma Brunskill, Associate Professor,
Computer Science, Stanford University

with thanks to Yuchen Hu for some figures

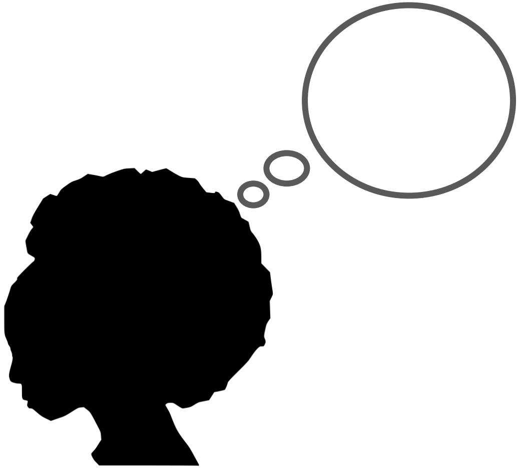
AI 4 HI

A hallmark of human cognition is learning from just a few examples – Lake et al. 2011



AI 4 HI

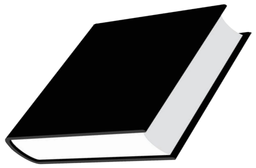
Quickly Learning from Few Examples to Make Decisions is a Key Part of Human Intelligence



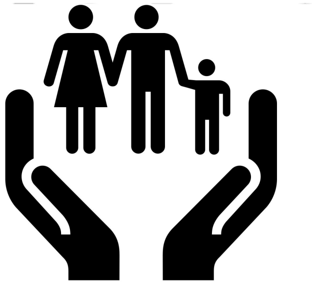
AI 4 HI

In Many Important Areas, Data-Driven Decision Policies Might Vastly Improve Outcomes, But Experimentation is Hard

Healthcare



Education

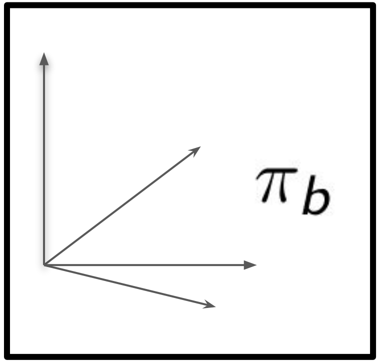


Social Services

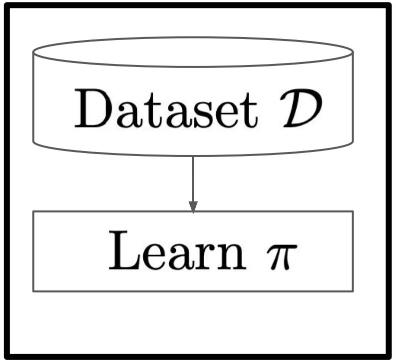
AI 4 HI

Learning to Make Decisions from Few Examples is Part of Accelerating Data-Driven Decision Making

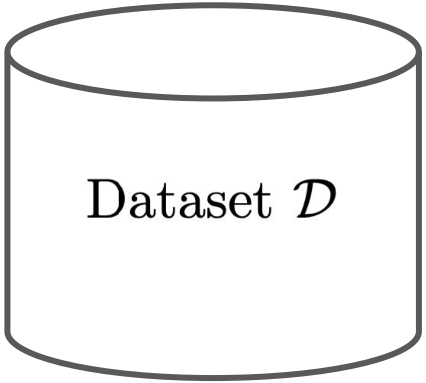
Getting Better Data



Using Data Better

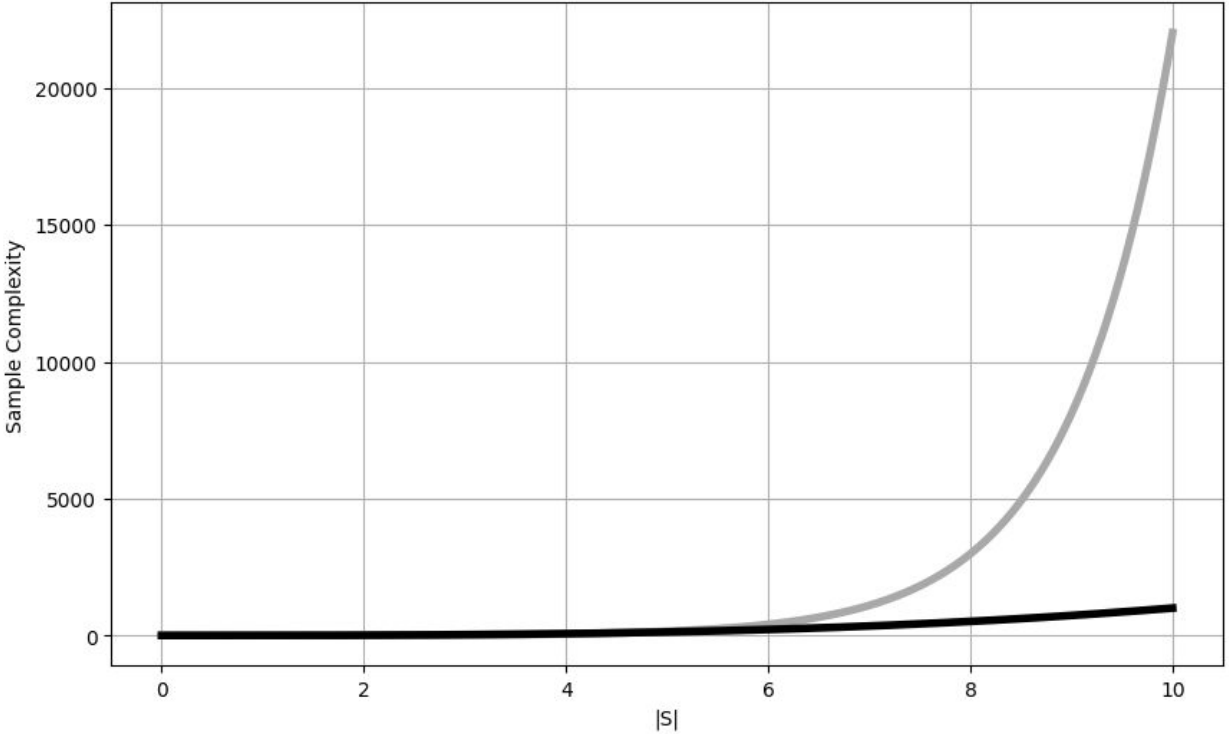


Data in the Age of Generative AI



AI 4 HI

Know Algorithms for Learning to Make Decisions Can Have Radically Different Sample Efficiency

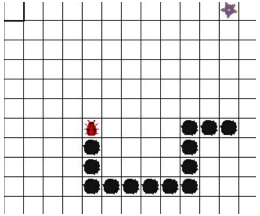


AI 4 HI

Some Characterization of When it Is Hard to Learn to Act Well

Stochasticity in the Transition Dynamics

Deterministic Markov Decision Process

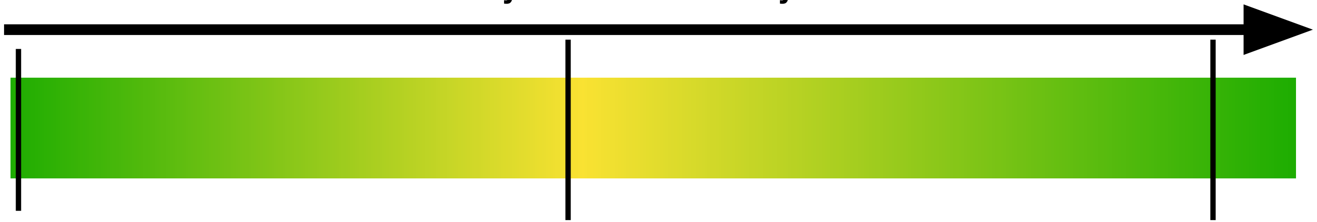


Zanette and Brunskill ICML 2019

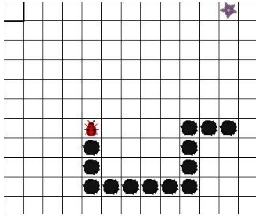
AI 4 HI

Some Characterization of When it Is Hard to Learn to Act Well

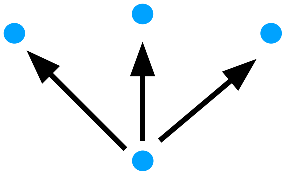
Stochasticity in the Transition Dynamics



Deterministic Markov Decision Process



Bandit Like Structure

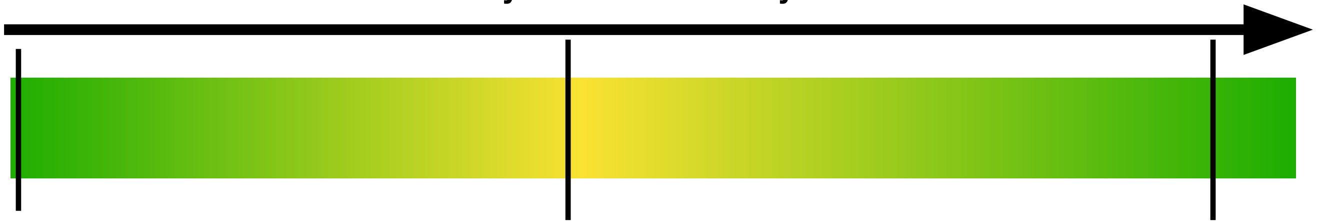


Zanette and Brunskill ICML 2019

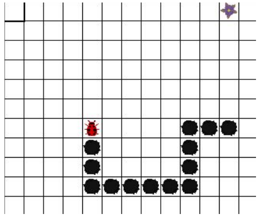
AI 4 HI

Some Characterization of When it Is Hard to Learn to Act Well

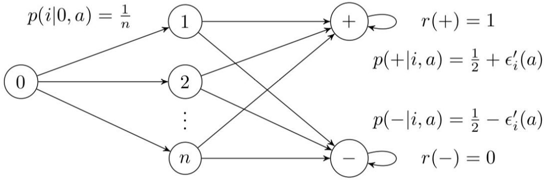
Stochasticity in the Transition Dynamics



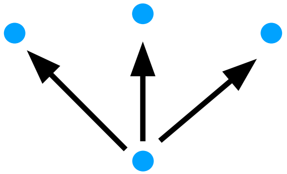
Deterministic Markov Decision Process



Hard Instances
Large Gap in Successor Value



Bandit Like Structure



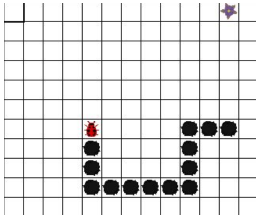
Zanette and Brunskill ICML 2019

AI 4 HI

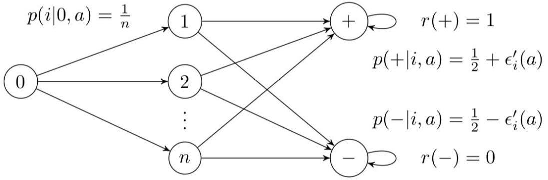
Some Characterization of When it Is Hard to Learn to Act Well, But Still Many Open Questions in Efficient Exploration

Stochasticity in the Transition Dynamics

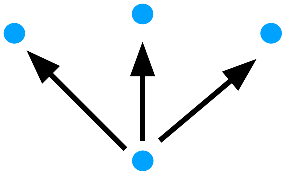
Deterministic MDP



Hard Instances
Large Gap in Successor Value



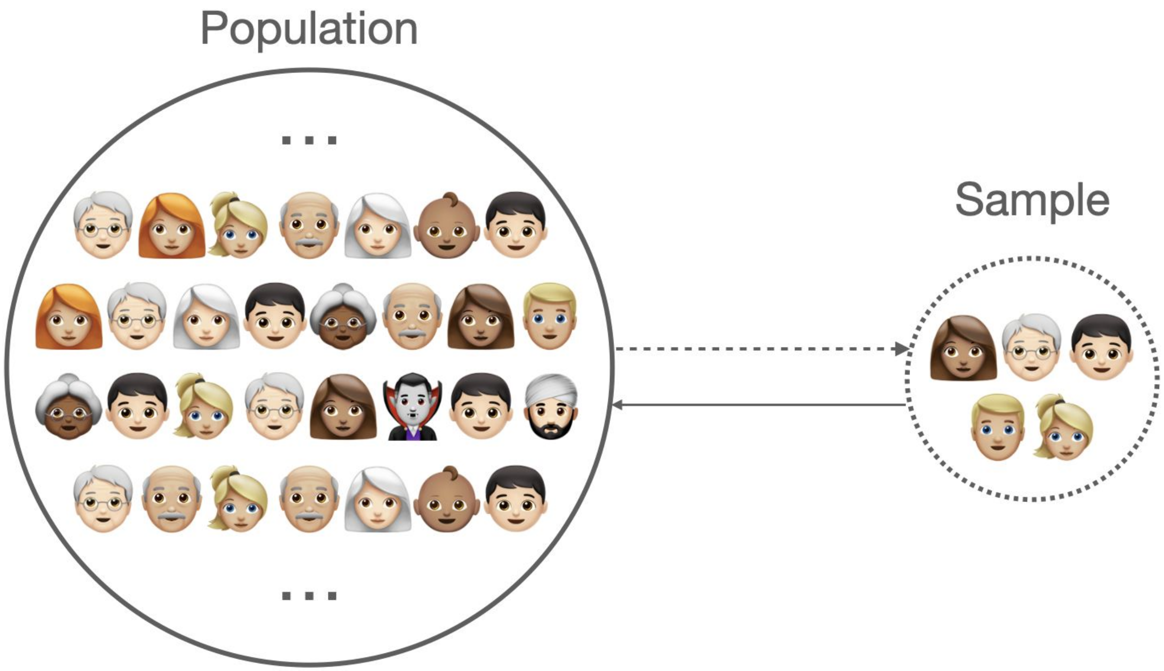
Bandit Like Structure



Zanette and Brunskill ICML 2019

AI 4 HI

Motivating Scenario: Clinical Trial Design

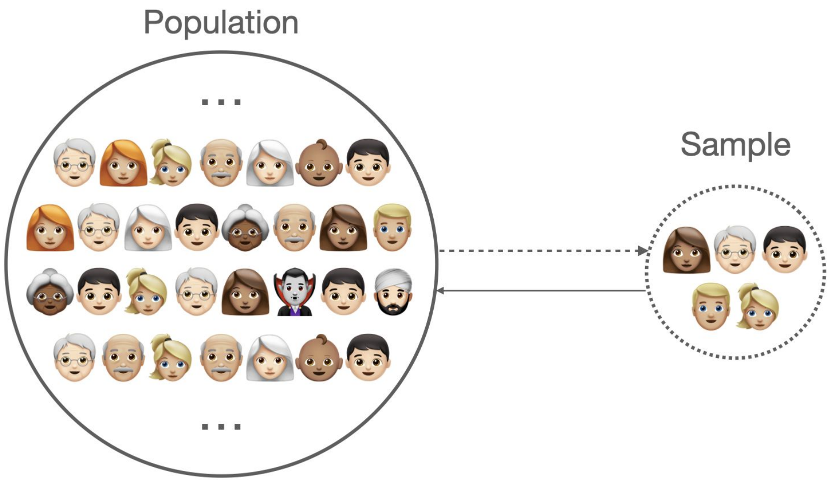


Covid-19 Vaccine Trial Design: how to allocate samples between younger and older individuals?

AI 4 HI

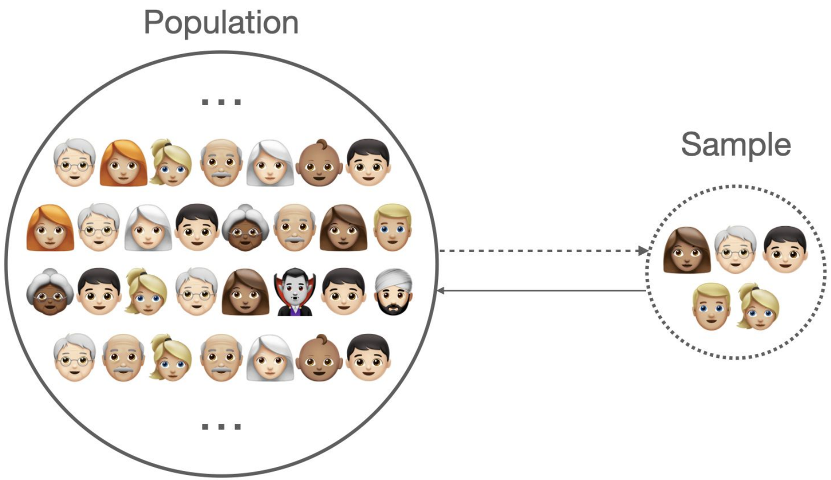
Clinical Trial Design

Standard: Design to Testing Scientific Hypotheses



AI 4 HI

Clinical Trial Design



Standard:
Design to Testing
Scientific Hypotheses

Alternative:
Design to Optimize
Utility of Induced
Decision Policy

AI 4 HI

Setting: Design Experiment

(x, a)

(x, a)

(x, a)

(x, a)

(x, a)

(x, a)

(x, a)

(x, a)

(x, a)

x: state/ context   a: action/ condition

AI 4 HI

Setting: Gather Data *D*

(x, a)
(x, a)
(x, a)
(x, a)
(x, a)
(x, a)
(x, a)
(x, a)
(x, a)

$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$
$r(x, a)$

x: state/ context   a: action/ condition

Setting: Gather Data *D*, Derive State-Specific Policy

(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
(x, a)	$r(x, a)$
	$r(x, a)$

$$\pi_D : x \rightarrow \arg\max_{a'} \underbrace{\hat{r}_D(x, a')}_{\text{empirical estimate of reward given data}}$$

x: state/ context a: action/ condition r: outcome

AI 4 HI

Experimental Design for Learning Contextual Bandit Policies

- Assume just 2 actions and a small finite set of groups (states / contexts)
- Assume each state x_i has some population proportion $p(x_i)$



Hu, Zhu, Brunskill, Wager EC 2024
(Best Student Paper, Decision Analysis Society)

AI 4 HI

Evaluate Policy By Its Expected Performance Over All Groups

- Assume just 2 actions and a small finite set of groups (states / contexts)
- Assume each state x_i has some population proportion $p(x_i)$
- Can sample a set of states and actions, observe rewards → dataset D
- Use dataset D to learn a group/state-dependent policy

$$\pi : x \rightarrow a$$

$$\sum_x p(x) r(x, \pi_D(x))$$



Hu, Zhu, Brunskill, Wager EC 2024
(Best Student Paper, Decision Analysis Society)

AI 4 HI

Objective: Design Experiment to Minimize Expected Regret

(x,a)
(x,a)
(x,a)
(x,a)
(x,a)
(x,a)
(x,a)
(x,a)
(x,a)

$\longrightarrow$

$$\left.\begin{array}{l} r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \\ r(x,a) \end{array}\right\}$$

$$\pi_D : x \rightarrow \arg\max_{a'} \underbrace{\hat{r}_D(x,a')}_{\text{empirical estimate of reward given data}}$$

$$Regret = \sum_x p(x) r(x, \pi^*(x)) - \sum_x p(x) r(x, \pi_D(x))$$

x: state/ context   a: action/ condition   r: outcome

AI 4 HI

Design Experiment to Minimize Minimax Expected Regret

$$\bar{n} = \underbrace{(n_1, \ldots, n_{|X|})}$$

Number of samples to give to each state / group 1…|X|
e.g. > 65 yrs, <= 65 yrs

$x :$	group
$a :$	action/condition
$r :$	outcome
$\pi^*(x) :$	optimal policy
$D = \bar{n} :$	samples per group
$\hat{\pi}_D(x) :$	policy learned

AI 4 HI

Design Experiment to Minimize Minimax Expected Regret

$$\bar{n} = \underbrace{(n_1, \ldots, n_{|X|})}_{\substack{\textbf{Number of samples to} \\ \textbf{give to each state / group} \\ \textbf{1…|X|} \\ \textbf{e.g. > 65 yrs, <= 65 yrs}}} = \arg\min_{\bar{n}} \underbrace{\max_{r}}_{\substack{\textbf{Adversary can} \\ \textbf{choose reward} \\ \textbf{function}}} E_{D\sim(\bar{n},r)}$$

$x:$	group
$a:$	action/condition
$r:$	outcome
$\pi^*(x):$	optimal policy
$D=\bar{n}:$	samples per group
$\hat{\pi}_D(x):$	policy learned

AI 4 HI

Design Experiment (Allocate State/Group Samples) to Minimize Minimax Expected Regret

$$\bar{n} = \underbrace{(n_1, \ldots, n_{|X|})}_{\substack{\text{Number of samples to} \\ \text{give to each state / group} \\ 1\ldots|X| \\ \text{e.g. > 65 yrs, <= 65 yrs}}} = \arg\min_{\bar{n}} \underbrace{\max_{r}}_{\substack{\text{Adversary can} \\ \text{choose reward} \\ \text{function}}} E_{D\sim(\bar{n},r)} \underbrace{\left[\sum_x \underbrace{p(x)}_{\substack{\text{Weigh rewards} \\ \text{by state / group} \\ \text{proportions}}} r(x, \pi^*(x)) - \sum_x p(x) r(x, \hat{\pi}_D(x)) \right]}_{\text{Expected Regret}}$$

$x:$	group
$a:$	action/condition
$r:$	outcome
$\pi^*(x):$	optimal policy
$D = \bar{n}:$	samples per group
$\hat{\pi}_D(x):$	policy learned

AI 4 HI

Prior Work Typically

- Assumes states generated stochastically x ~ p(x)
- Computes policy per state
- Aims to learns an ϵ-optimal action per x (PAC learning)
- Or assumes data will be used for hypothesis testing

AI 4 HI

Prior Work Typically

- Assumes states generated stochastically x ~ p(x)
- Computes policy per state
- Aims to learns an ϵ-optimal action per x (PAC learning)
- Or assumes data will be used for hypothesis testing

Our Work

- Strategically selects groups
- Computes policy per state
- Aims to optimize expected performance weighted by p(x)

AI 4 HI

Is Sampling Based on Population Probability Optimal?

$$n_{x_i} \propto p(x_i)N$$

$$\bar{n} = (n_1, \ldots, n_{|X|}) = \arg\min_{\bar{n}} \max_{r} E_{D\sim(\bar{n},r)} \left[\sum_x \underbrace{p(x)}_{\text{Weigh rewards by context proportions}} r(x, \pi^*(x)) - \sum_x p(x) r(x, \hat{\pi}_D(x)) \right]$$

$x:$	group
$a:$	action/condition
$r:$	outcome
$\pi^*(x):$	optimal policy
$D = \bar{n}:$	samples per group
$\hat{\pi}_D(x):$	policy learned

AI 4 HI

Group Allocation to Optimize Minimax Regret is Not Directly Proportional to Group Probabilities

- Assume finite budget N of samples
- Allocation that minimizes minimax regret <u>oversamples</u> low probability groups

$$n_{x_i} \propto p(x_i)N \qquad \text{vs} \qquad \boxed{n_{x_i} \propto \frac{p(x_i)^{2/3}}{\sum_{j=1}^{|X|} p(x_j)^{2/3}} N}$$

x :	group
a :	action/condition
r :	outcome
$\pi^*(x)$:	optimal policy
$D = \bar{n}$:	samples per group
$\hat{\pi}_D(x)$:	policy learned

AI 4 HI

Context Allocation That Optimizes Minimax Regret is Not Proportional to Group Probabilities

- Assume finite budget N of samples
- Allocation that minimizes minimax regret oversamples low probability states x

$$n_{x_i} \propto p(x_i)N \qquad \text{vs} \qquad \boxed{n_{x_i} \propto \frac{p(x_i)^{2/3}}{\sum_{j=1}^{|X|} p(x_j)^{2/3}} N}$$

- Post acceptance we learned that Manski and Tetenov (2016) proved that (⅔) rate minimizes an upper bound on minimax regret, and Schlag (2006) proved a related result for the 2 context case
- To our knowledge, we are first to prove (⅔) rate optimizes minimax regret

AI 4 HI

Intuition for Oversampling Low Probability States/ Groups

- Assume finite budget N of samples
- Allocation that minimizes minimax regret oversamples low probability x

$$n_{x_i} \propto p(x_i)N$$

vs

$$\boxed{n_{x_i} \propto \frac{p(x_i)^{2/3}}{\sum_{j=1}^{|X|} p(x_j)^{2/3}} N}$$

x :	group
a :	action/condition
r :	outcome
$\pi^*(x)$:	optimal policy
$D = \bar{n}$:	samples per group
$\hat{\pi}_D(x)$:	policy learned

AI 4 HI

Intuition for Oversampling Low Probability States/ Groups

- Assume finite budget N of samples
- Allocation that minimizes minimax regret oversamples low probability x

$$n_{x_i} \propto \frac{p(x_i)^{2/3}}{\sum_{j=1}^{|X|} p(x_j)^{2/3}} N$$

- Finite samples = non-zero error in outcome model estimates
- Rough intuition: if estimate reward at $n^{-1/2}$ rate, then 101th sample to context *x* provides less reduction in error than 10th sample to context *y*

AI 4 HI

Simulation of Sample Allocation for Covid-19 Trial Data: Minimax Regret Also Improves Worst Regret Per Group

Context Sample Allocation	N 18 to 65 years	N >= 65 years	Minimax regret	Worst regret over groups
Minimax	6100	3218	4.6	9.36
Proportional	7734	1584	4.94	13.34

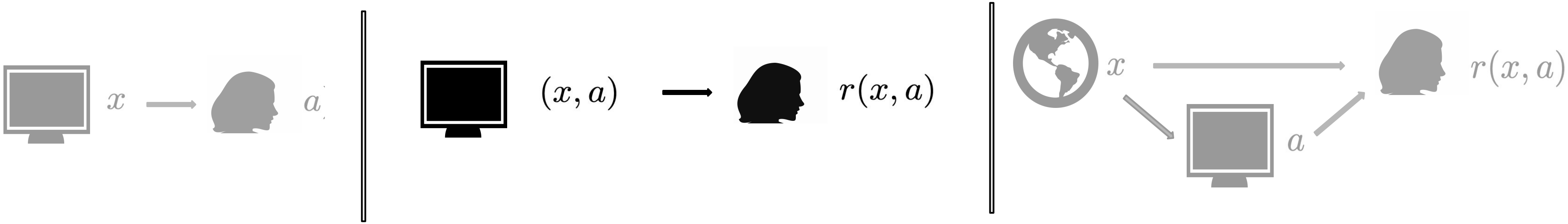


Hu, Zhu, Brunskill, Wager EC 2024
(Best Student Paper, Decision Analysis Society)

AI 4 HI

Strategic Design to Gather Better Data for Better Decisions

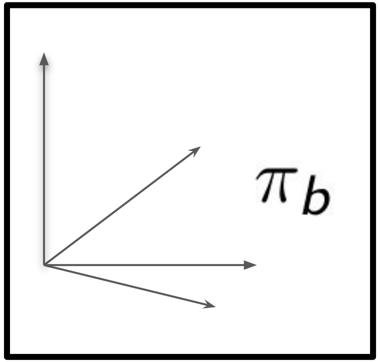
- Lots of prior work
- Still underexplored settings where results are surprising and impactful
  - Clinical trials designed for statistical hypothesis testing, not maximizing expected rewards
- Current work: relating setting for alignment for large language models



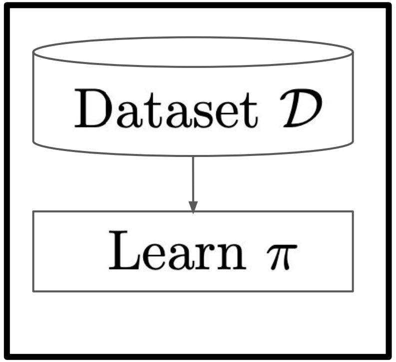
AI 4 HI

Accelerating Data-Driven Decision Making

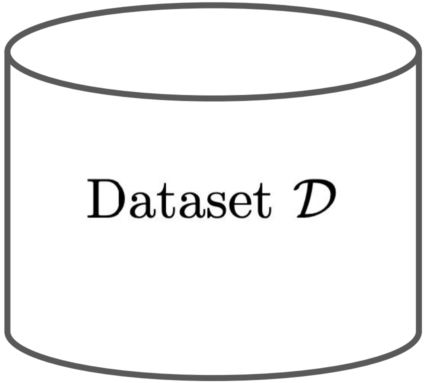
Getting Better Data



Using Data Better



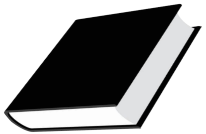
Data in the Age of Generative AI



AI 4 HI

Recall Motivation for Sample Efficient Learning

- Can be expensive / challenging to do extensive experiments

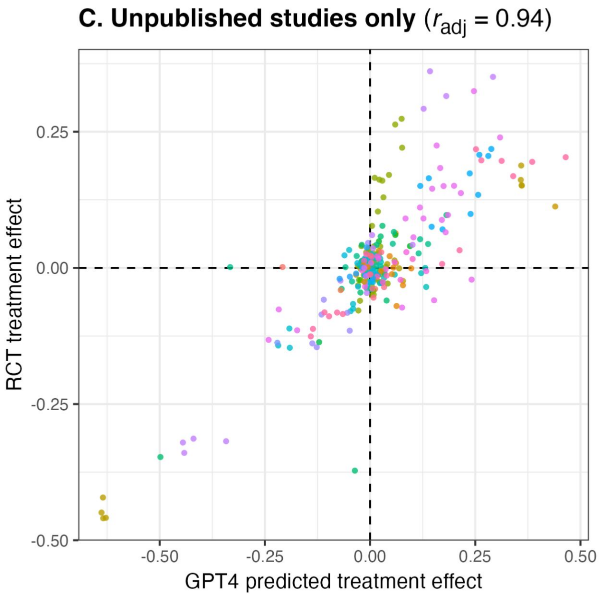




- Humans can often quickly learn to make decisions

AI 4 HI

From Real Experiments to Thought Experiments, Powered by LLMs

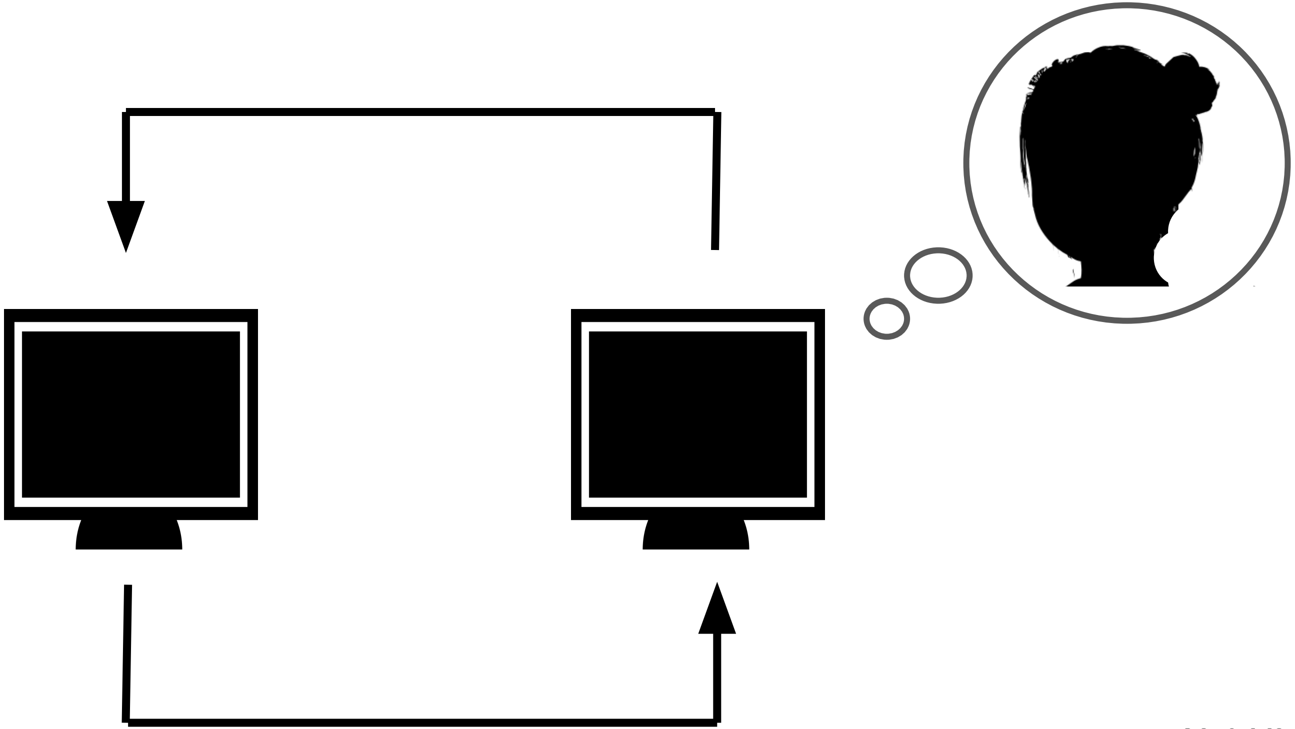
Generative Agents: Interactive Simulacra of Human Behavior. Park, O'Brien, Cai, Morris, Liang, Bernstein UIST 2023



Predicting Results of Social Science Experiments Using Large Language Models. Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, Robb Willer. Axiv 2024

AI 4 HI

Learning to Improve Instruction with LLMs



He-Yueya, Goodman, Brunskill EDM 2024

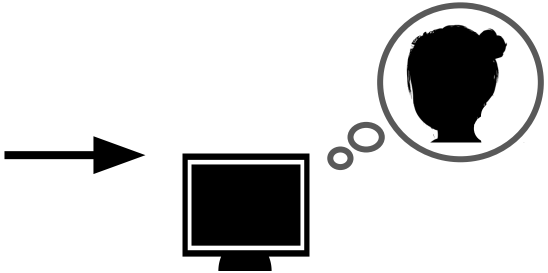
AI 4 HI

LLMs to Simulate a Static Student

You are an 8th-grade student who has not learned about systems of equations.

Solve:

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa’s age.



He-Yueya, Goodman, Brunskill EDM 2024

AI 4 HI

Simulating Dynamics of Learning

You are an 8th-grade student who has not learned about systems of equations.

Solve:

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.

You now watch a video. The transcript is: "In this video, we're gonna get some more practice setting up systems of equations. So we're told Sanjay's dog weighs five times as much as his cat…"

After you finish the video, try to solve the following problem. Remember, you've only been taught what was shown in the video…

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.

AI 4 HI

LLMs Failed at Simulating Dynamics of Learning

You are an 8th-grade student who has not learned about systems of equations.

Solve:

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.

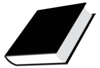
You now watch a video. The transcript is: "In this video, we're gonna get some more practice setting up systems of equations. So we're told Sanjay's dog weighs five times as much as his cat…"

After you finish the video, try to solve the following problem. Remember, you've only been taught what was shown in the video…

- Alyssa is twelve years older than Bethany.
- The sum of their ages is forty-four.
- Find Alyssa's age.

AI 4 HI

Option 1: Data/ Experiments to Optimize Instruction



Score: 92

1+1

Score: 85

AI 4 HI

Option 1: Data/ Experiments to Optimize Instruction

Score: 92

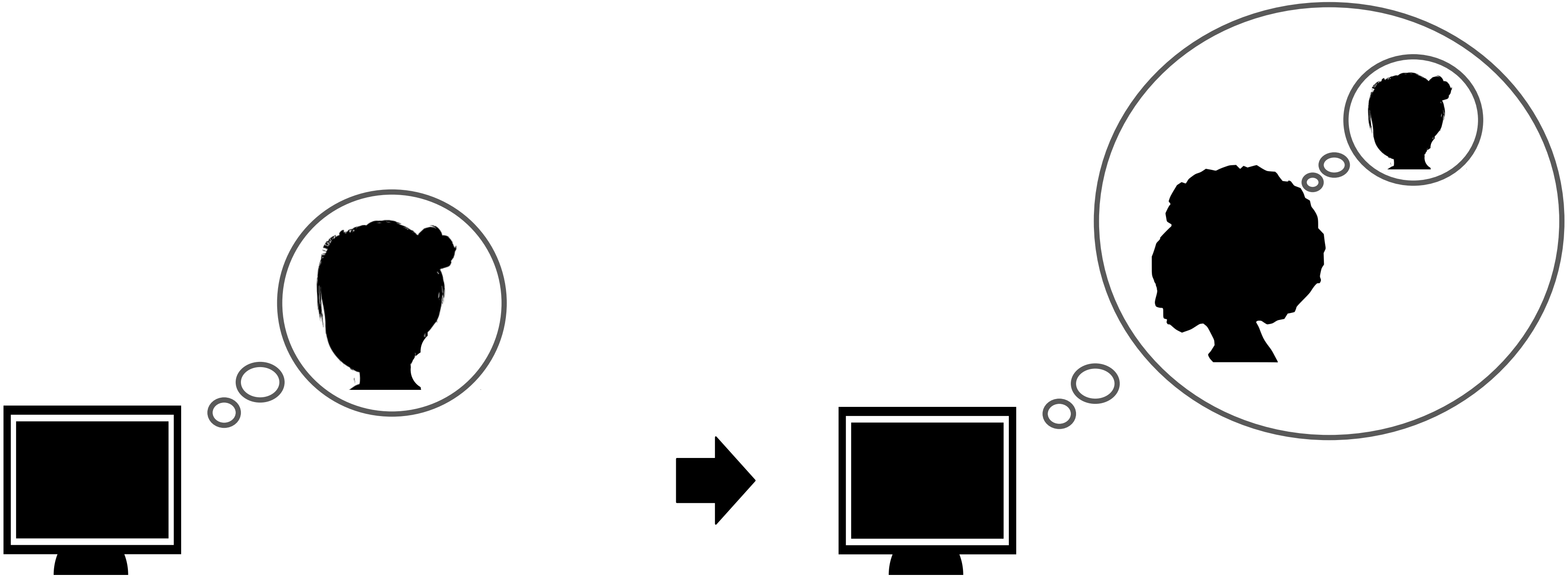
1+1

Score: 85

Common Alternative: **Education Experts** to Inform Instructional Choices

AI 4 HI

LLMs to Simulate Expert Educator Judgements



$$p(s'|s,a)$$

$$\sum_{s'} p(s'|s,a)r(s')$$

AI 4 HI

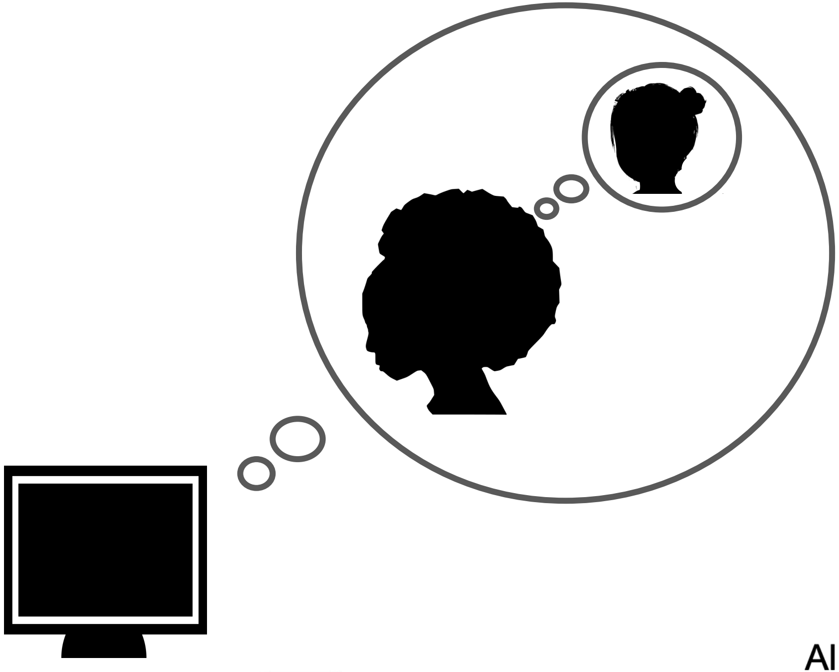
LLMs to Predict the Effectiveness of Instruction

Title: Equation Excellence - Mastering Systems of Equations

Objective: Your journey to master the art of translating word problems into systems of equations and solve them confidently is about to level up. Dive into these hand-picked problems to hone your skills.

Problem 1:
In a zoo, there are giraffes and zebras. The number of giraffes is two more than three times the number of zebras. If there are 29 animals in total, how many giraffes and zebras are there in the zoo?

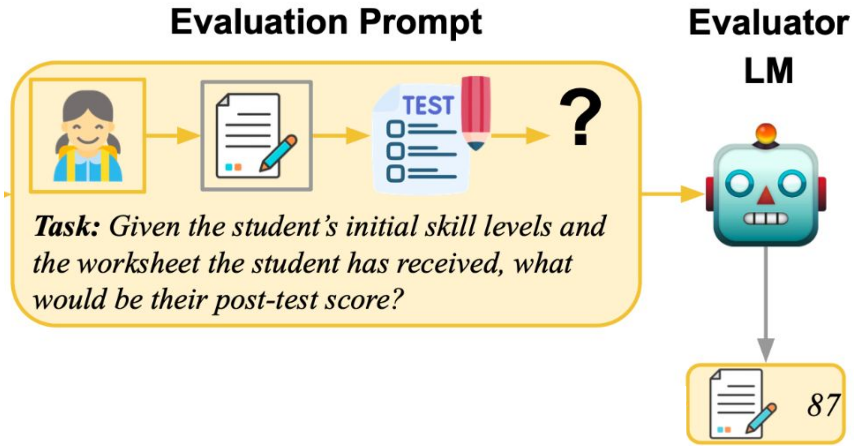
Hints for Success:
1. Assign 'G' to represent Giraffes and 'Z' for Zebras.
2. Derive the equations from the problem: 'G + Z = 29' and 'G = 3Z + 2'.
3. Solve these equations to find the number of giraffes and zebras.



$$\sum_{s'} p(s'|s,a)r(s')$$

AI 4 HI

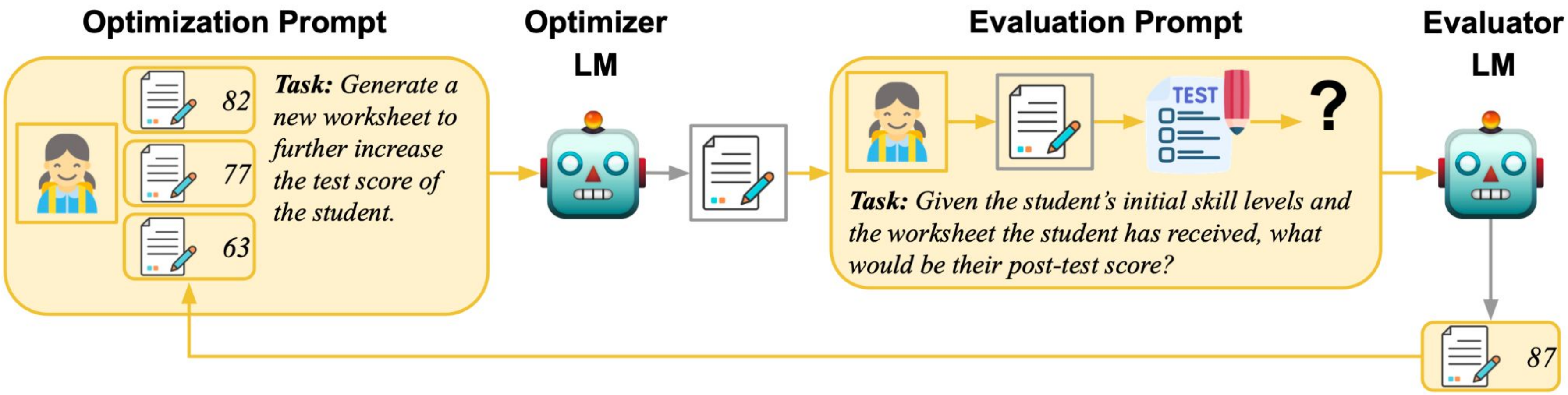
So Far: LLMs as Educational Evaluator



He-Yueya, Goodman, Brunskill EDM 2024

AI 4 HI

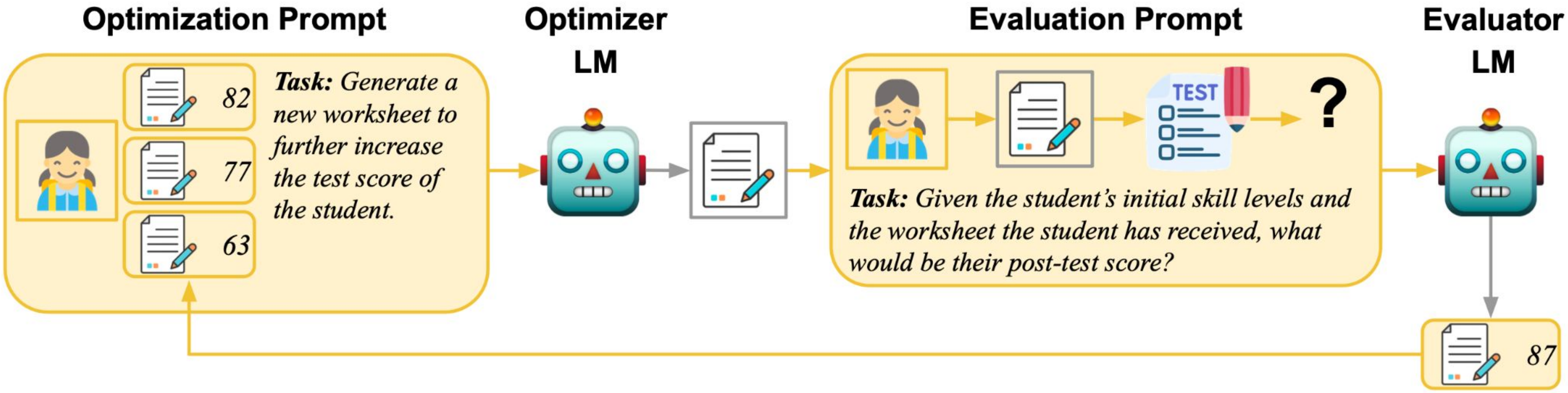
Now Optimize Worksheet Effectiveness (Action Space is Possible Worksheets)



He-Yueya, Goodman, Brunskill EDM 2024

AI 4 HI

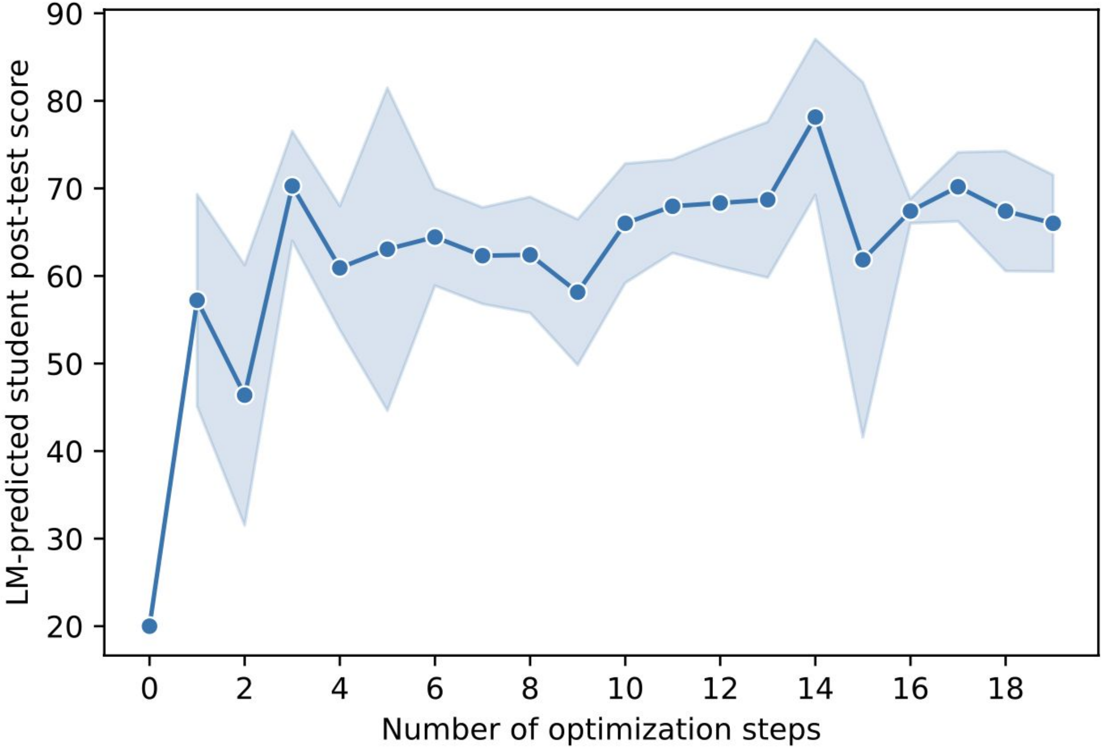
Optimize Worksheet Effectiveness – "Agentic" Workflow
Reflect (Critic), Update (Actor)



He-Yueya, Goodman, Brunskill EDM 2024

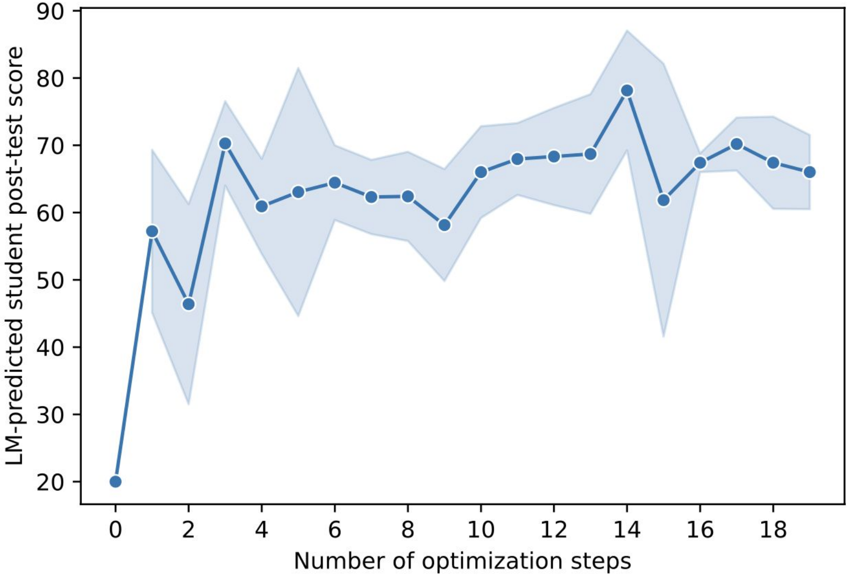
AI 4 HI

The LLM Evaluator Thinks the LLM Optimizer is Creating More Effective Worksheets



AI 4 HI

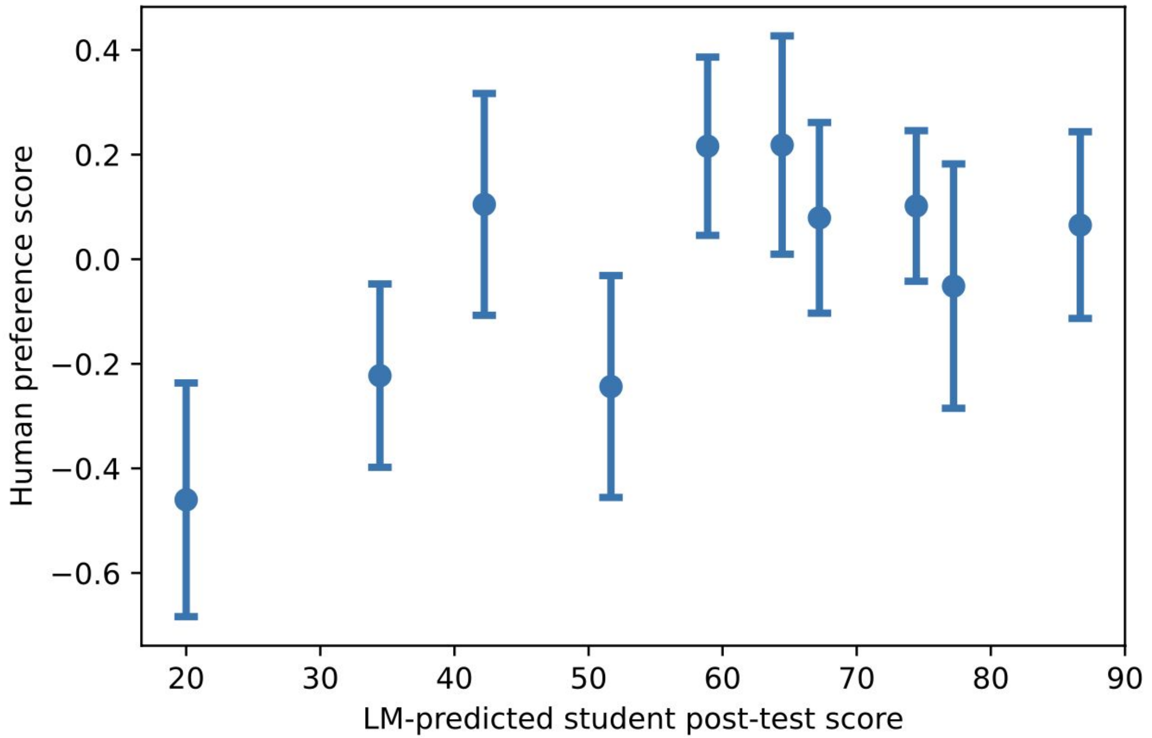
Human Education Expert Evaluations of LLM Optimized Worksheets



- Recruited 95 education experts
- Asked to evaluate LLM worksheets:
  - Is worksheet A or B more effective

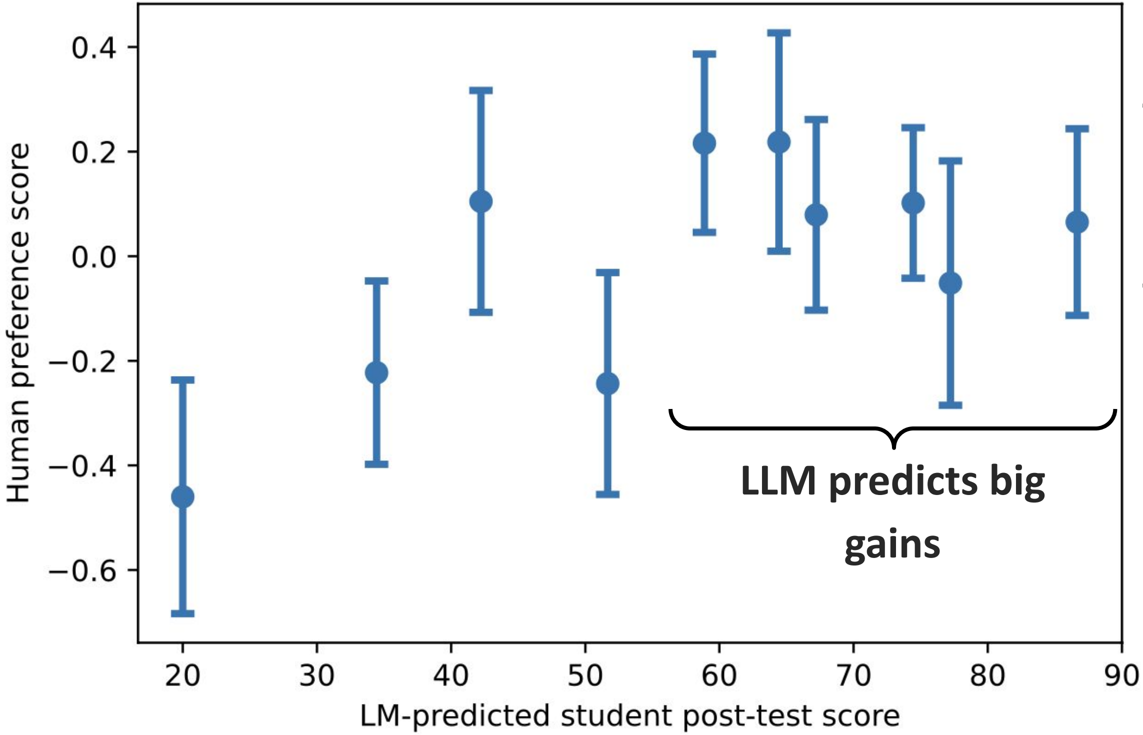
AI 4 HI

Human Education Expert Evaluation and LLM Evaluation Highly Correlated! (r=0.66)



AI 4 HI

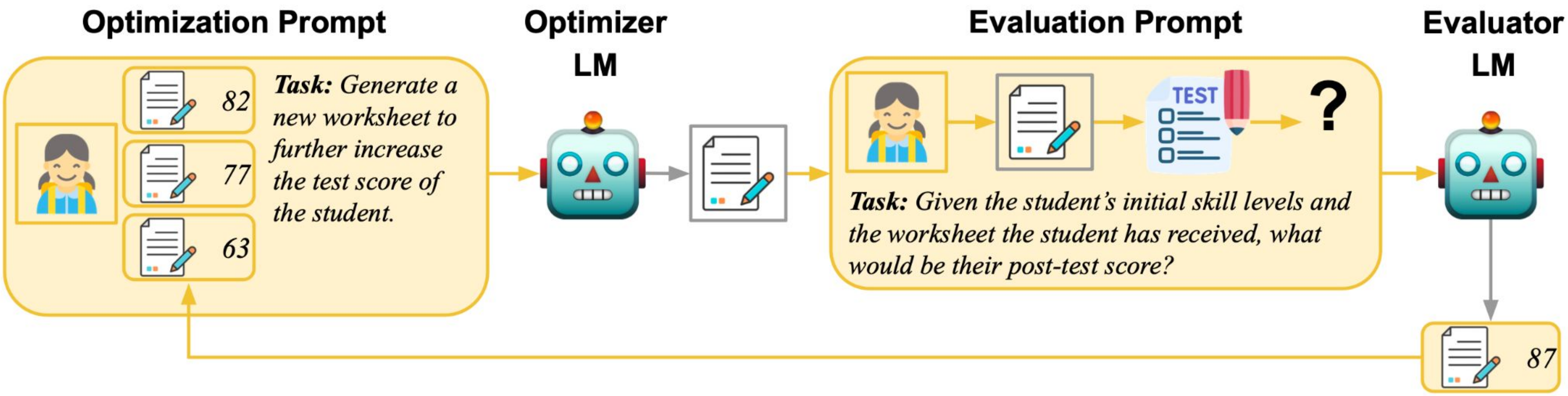
Highly Correlated ≠ Same Optima



Roughly same to humans

AI 4 HI

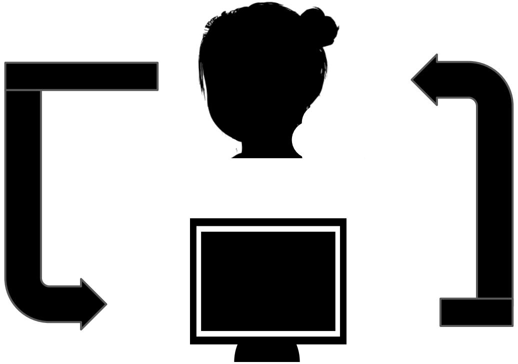
Summary: Can Use LLMs as Knowledge Experts (Critics) to Adapt Educational Content

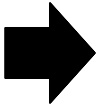


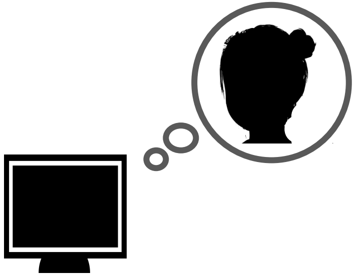
He-Yueya, Goodman, Brunskill EDM 2024

AI 4 HI

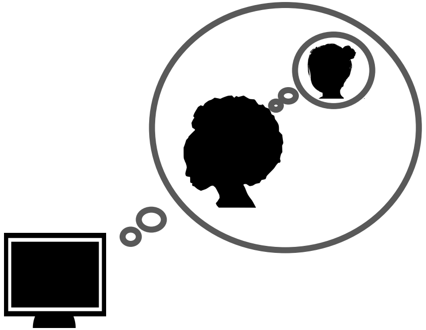
Reducing Need for Data Efficient Reinforcement Learning in Education with LLM Powered Thought Experiments: Promising and More Work Needed







$$p(s'|s,a)$$

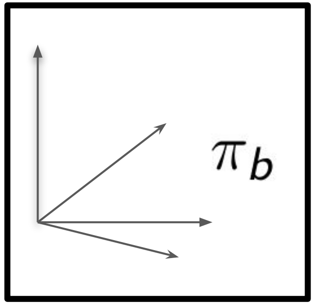


$$\sum_{s'} p(s'|s,a)r(s')$$

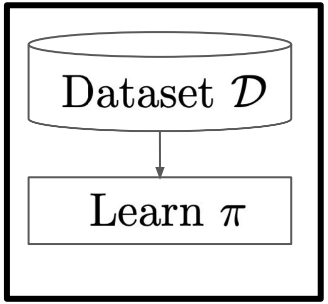
AI 4 HI

Summary: Accelerating Data-Driven Decision Making

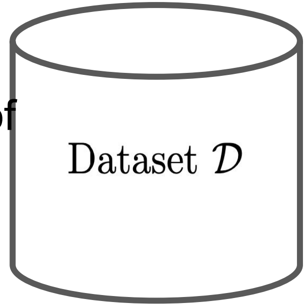
Getting Better Data



Using Data Better



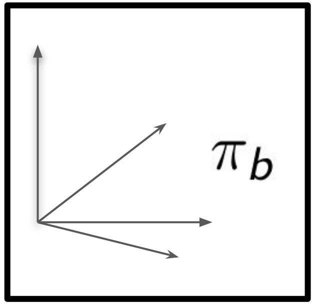
Data in the Age of GenAI



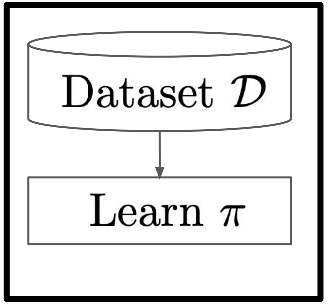
AI 4 HI

Summary: Accelerating Data-Driven Decision Making to *Help Humans to Thrive*

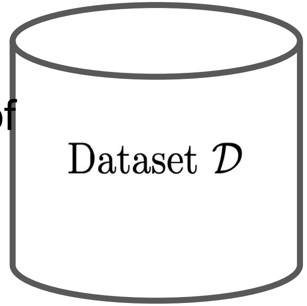
Getting Better Data



Using Data Better



Data in the Age of GenAI



AI 4 HI