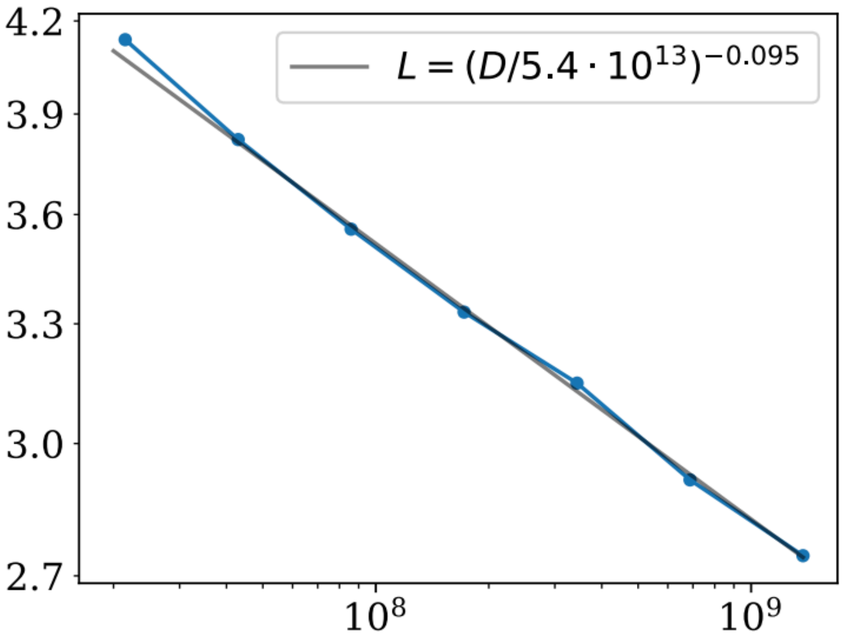
What Statistics and AI Offer Each Other?

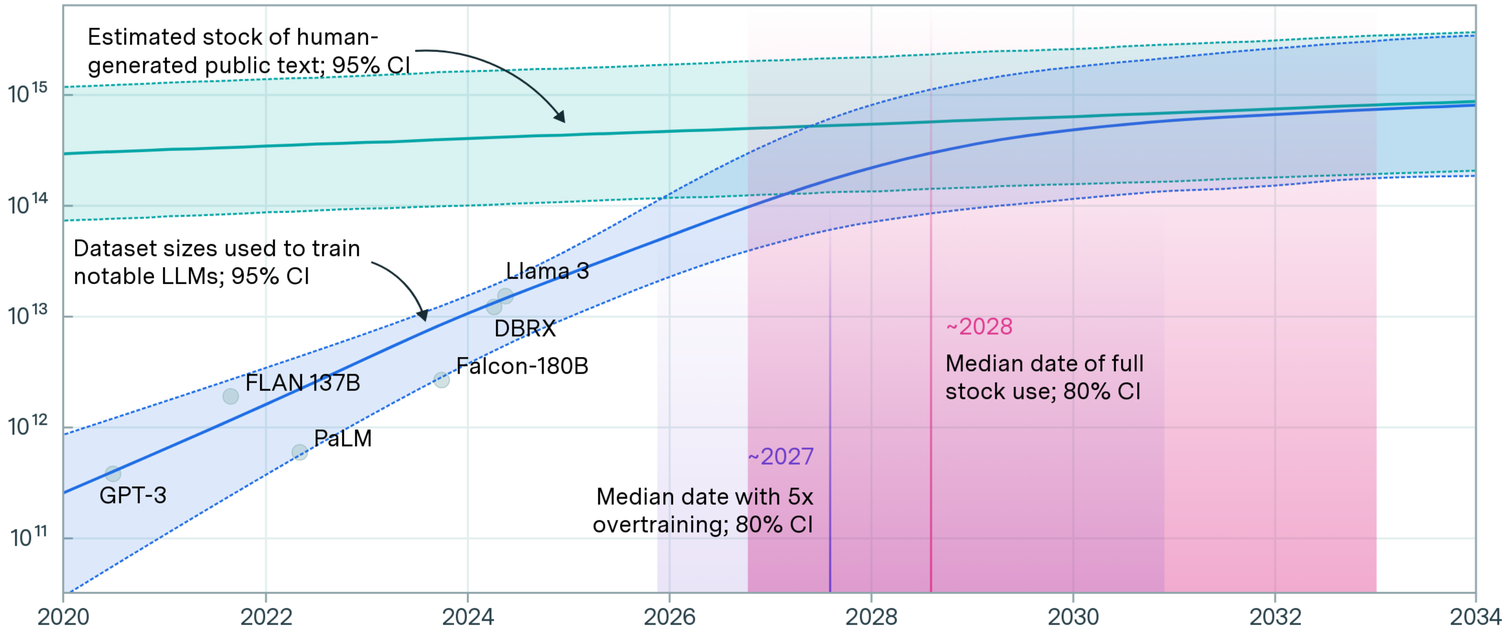
Emmanuel Candès

Asilomar Conference on Signals, Systems, and Computers, Oct 27, 2025

Scaling wall

- Scaling law (2020-2025): more data $\Longrightarrow$ more powerful AI
- Scaling wall (2025+): frontier AI has exhausted internet data





Our take: use existing data more effectively

Ilya Sutskever @NeurIPS 2024 — We have but one internet!

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

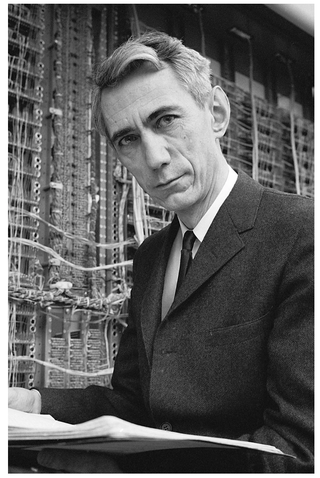
Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

Language model pretraining

Max likelihood:

$$\textit{maximize}_{\theta} \sum_{\textit{docs}} \log p_{\theta}(\textit{doc})$$



Claude Shannon

Ronald Fisher

Extending pretraining beyond the scaling wall via synthetic data

- Synthetic continued pretraining

  Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, Tatsunori Hashimoto

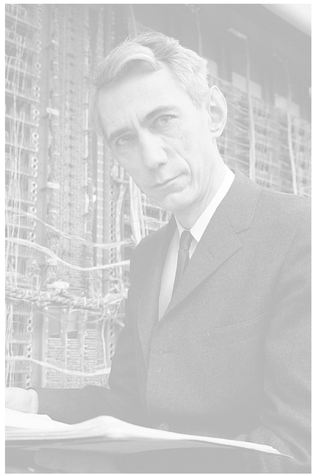
- Synthetic bootstrapped pretraining

  Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, Ruoming Pang

Language model pretraining

Max likelihood:

$$\text{maximize}_\theta \sum_{\text{docs}} \log p_\theta(\text{doc})$$



Claude Shannon

Ronald Fisher

Extending pretraining beyond the scaling wall via synthetic data

- Synthetic continued pretraining

  Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, Tatsunori Hashimoto

- Synthetic bootstrapped pretraining

  Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, Ruoming Pang

Synthetic continued pretraining

Goal: teach model the knowledge from a niche domain consisting of few “source documents"

Step 1: Generate synthetic text based on the source documents

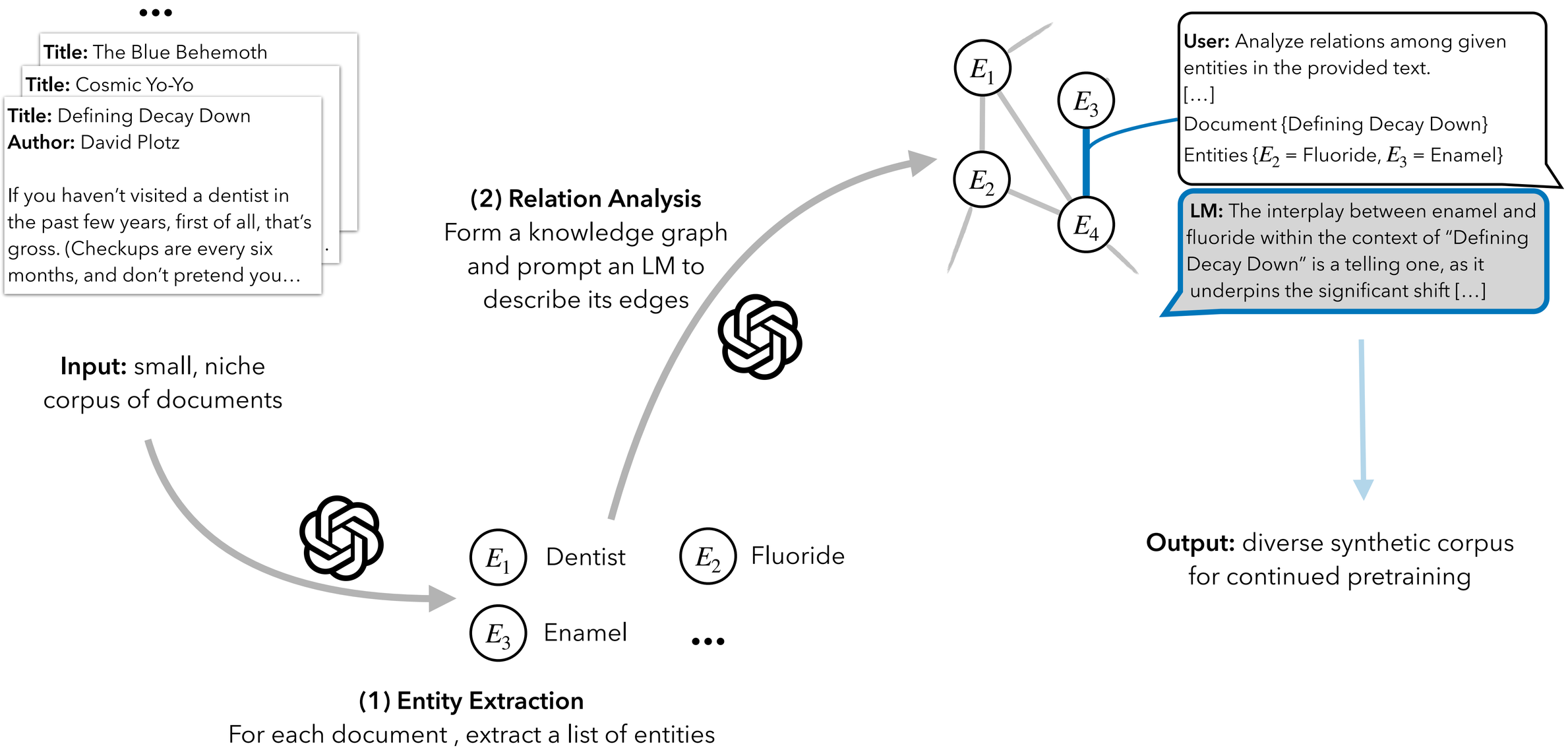
Step 2: Continually pretrain (finetune) the model on generated text

Experiment setup

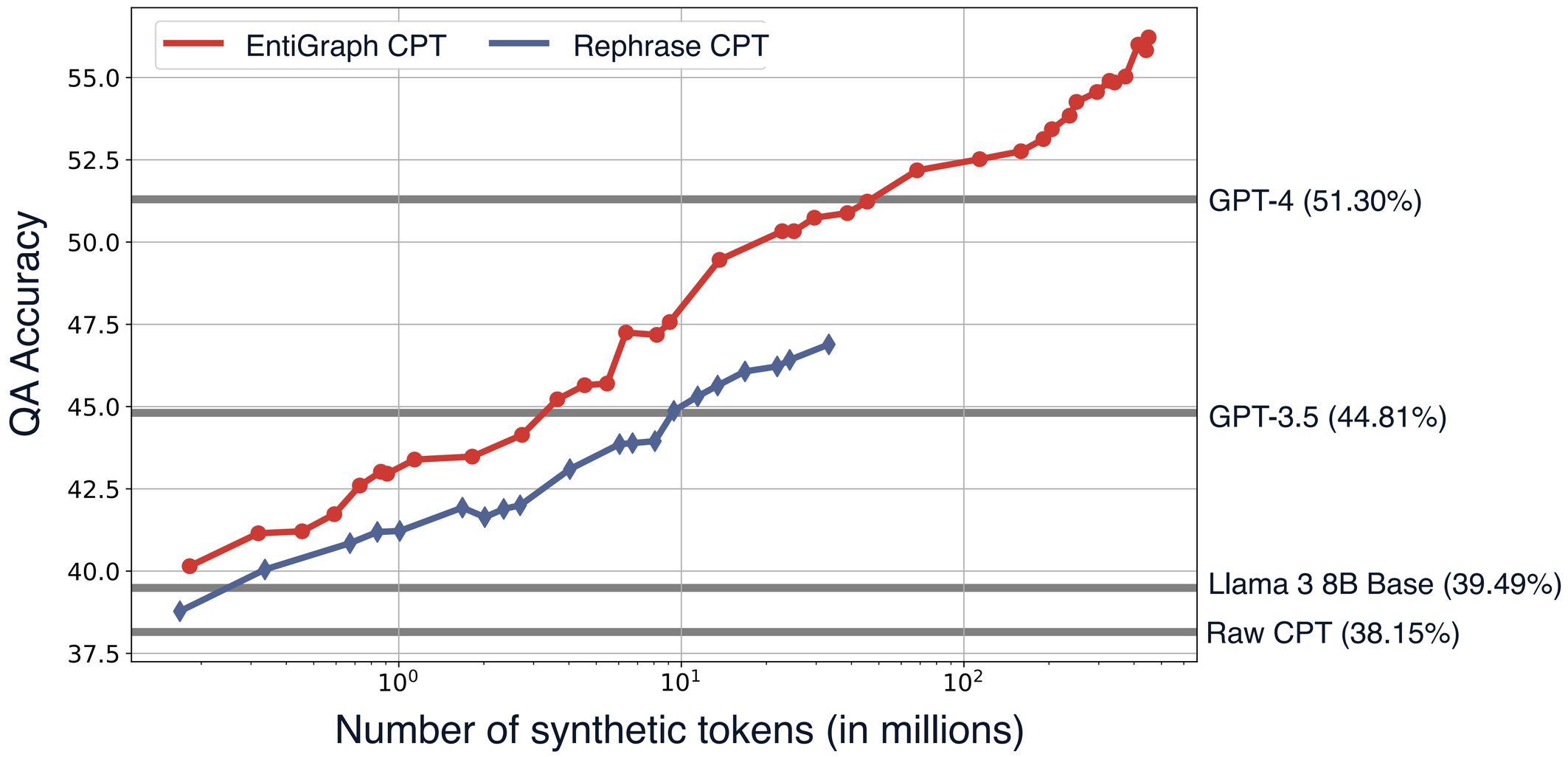
- Niche source documents (not something model already knows)
- A task that tests a model’s knowledge about the source documents

How to generate synthetic data?

EntiGraph: scalable data generator



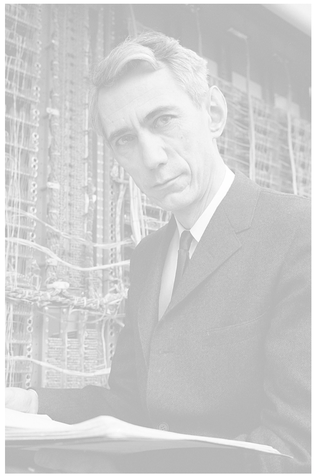
Scaling on synthetic data



Language model pretraining

Max likelihood:

$$\textit{maximize}_\theta \sum_{\textit{docs}} \log p_\theta(\textit{doc})$$



Claude Shannon

Ronald Fisher

Extending pretraining beyond the scaling wall via synthetic data

- Synthetic continued pretraining

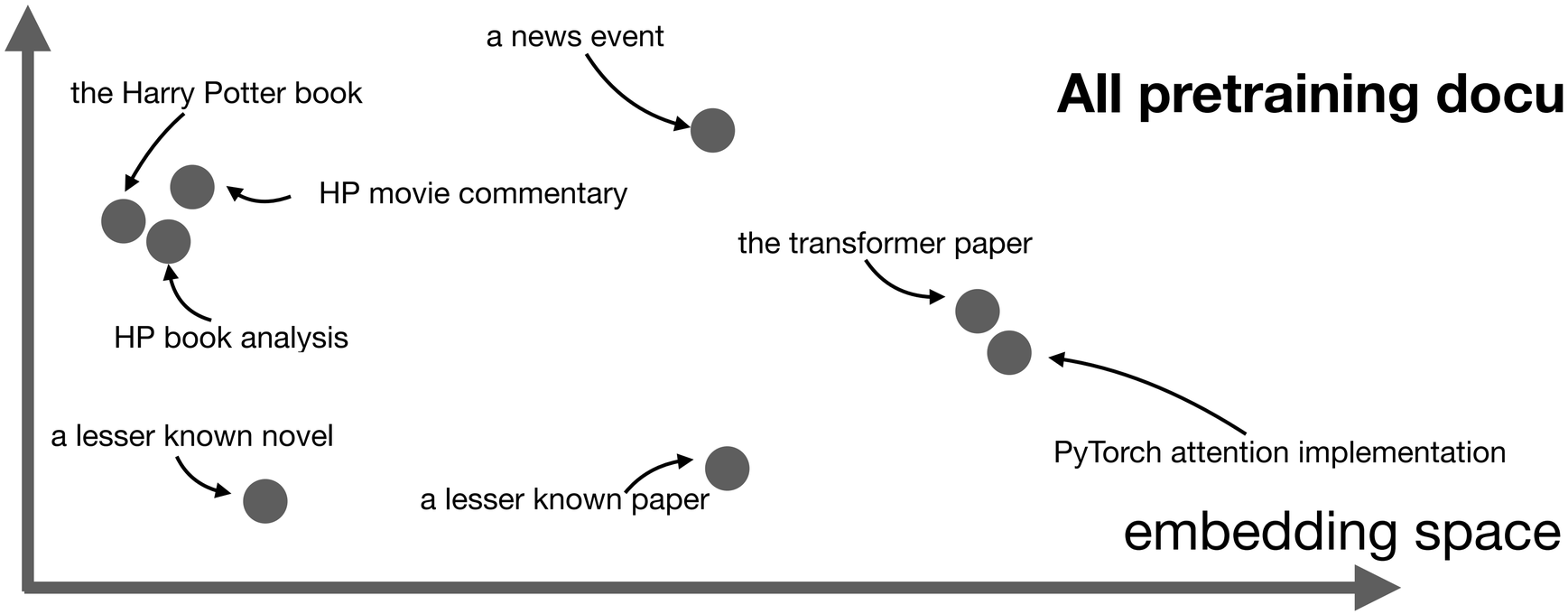
  Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, Tatsunori Hashimoto

- Synthetic bootstrapped pretraining

  Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, Ruoming Pang

Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** DCLM subset and Qwen-0.6B-Embedding



Examples of related documents

doc1

The Cultural Sites of Iran

With 196 countries and countless exciting destinations worldwide, there is so much to see in a very limited time. Even the most well-traveled person hardly gets to visit all and has to be selective. So, why should you consider visiting a country like Iran, especially when it comes to all those negative news and stereotypes surrounding it?
Here we're here to give you the reasons and to help you overcome your doubts and even encourage you to consider your next trip to Iran, this mysterious land as soon as you return to your home country!

Beautiful cities, friendly people, fabulous food, glorious architecture, Iran has delighted visitors for centuries with its World Heritage Sites, friendly towns and inspiring desert landscapes.

Things to Do in Iran – Activities & Attractions

Iran is the land of four seasons, history and culture, souvenir and authenticity. This is not a tourism slogan, this is the reality inferred from the experience of visitors who have been impressed by Iran's beauties and amazing attractions.

Antiquity and richness of the Cultural Sites of Iran and civilization, the variety of natural and geographical attractions, four – season climate, …

History of Iran

doc2

Query Text: Home > FAQ Login / Register

Why should we spend our holiday in Iran?

Iran is a country, located in the Middle East, which can meet the various needs of tourists and satisfy their different tastes, due to its rich civilization, historical sites, geographic location, nature of the four seasons and diverse tourist attractions. Therefore, considering the high security and low cost of travel to the country, it is introduced as one of the major tourist destinations to spend holidays in.

Is Iran a safe travel destination?

One of the wrong assumptions about the country of Iran is in terms of its security. Despite its location in Asia and the Middle East, and neighboring countries like Iraq, Afghanistan and Pakistan, Iran is considered as one of the safest countries in the region. According to the international data, security in Iran is much more than a touristic country such as Turkey.

To confirm the statements made above, refer to websites like www.travelriskmap.com.

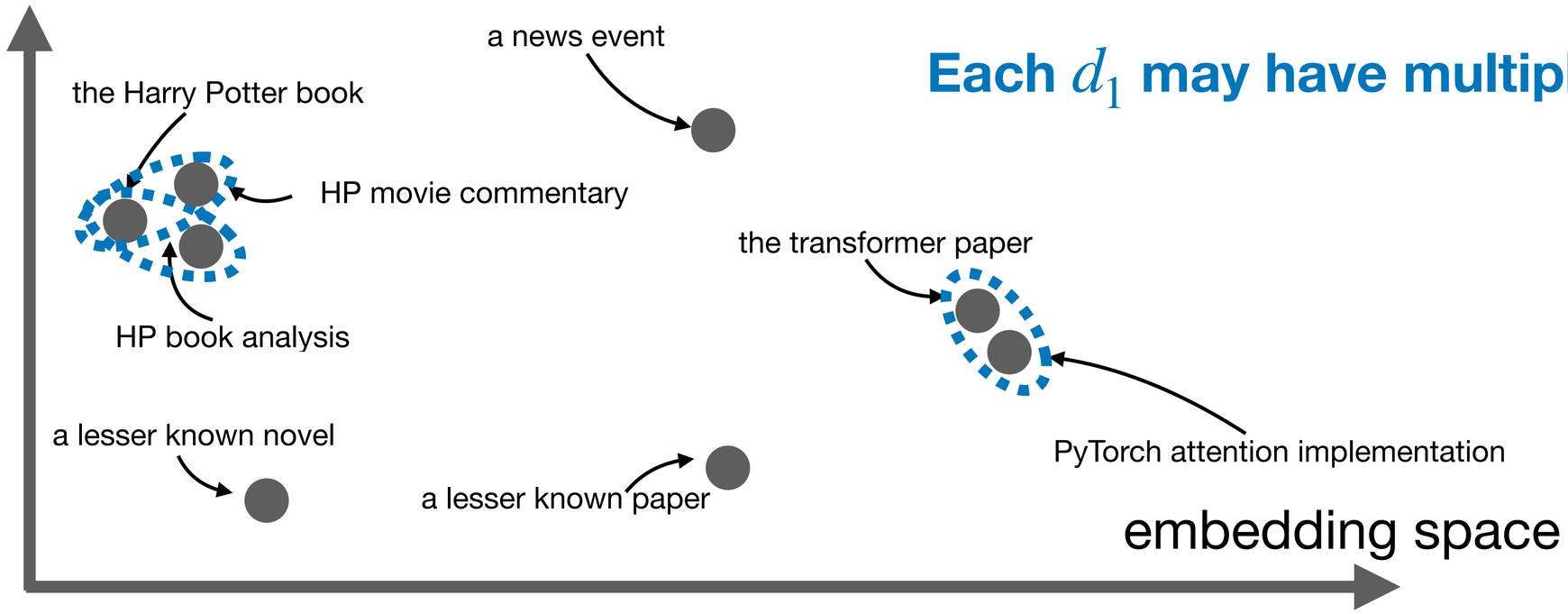
What does "the rich civilization" mean, as mentioned about Iran?

According to documentation in some of the world history references, …

Travel guide to Iran

Synthetic bootstrapped pretraining

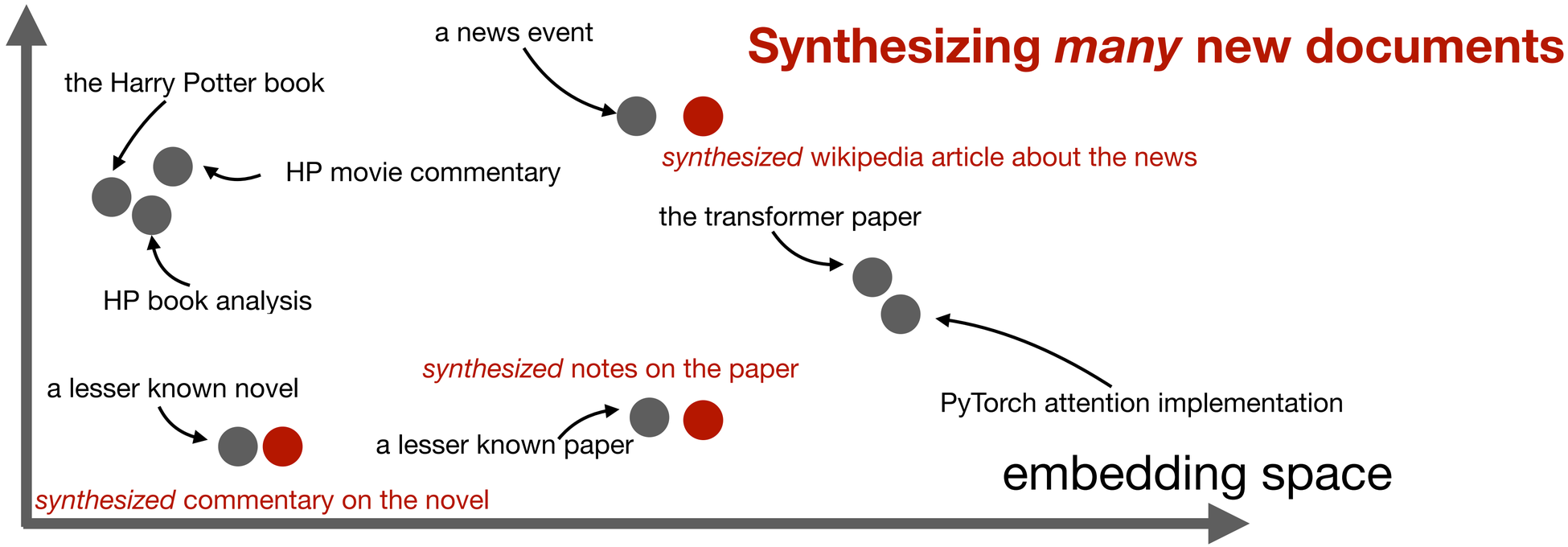
1. **Nearest-neighbor pairing:** DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_2 \mid d_1)$ initialized at pretrained checkpoint



Each d_1 may have multiple paired d_2

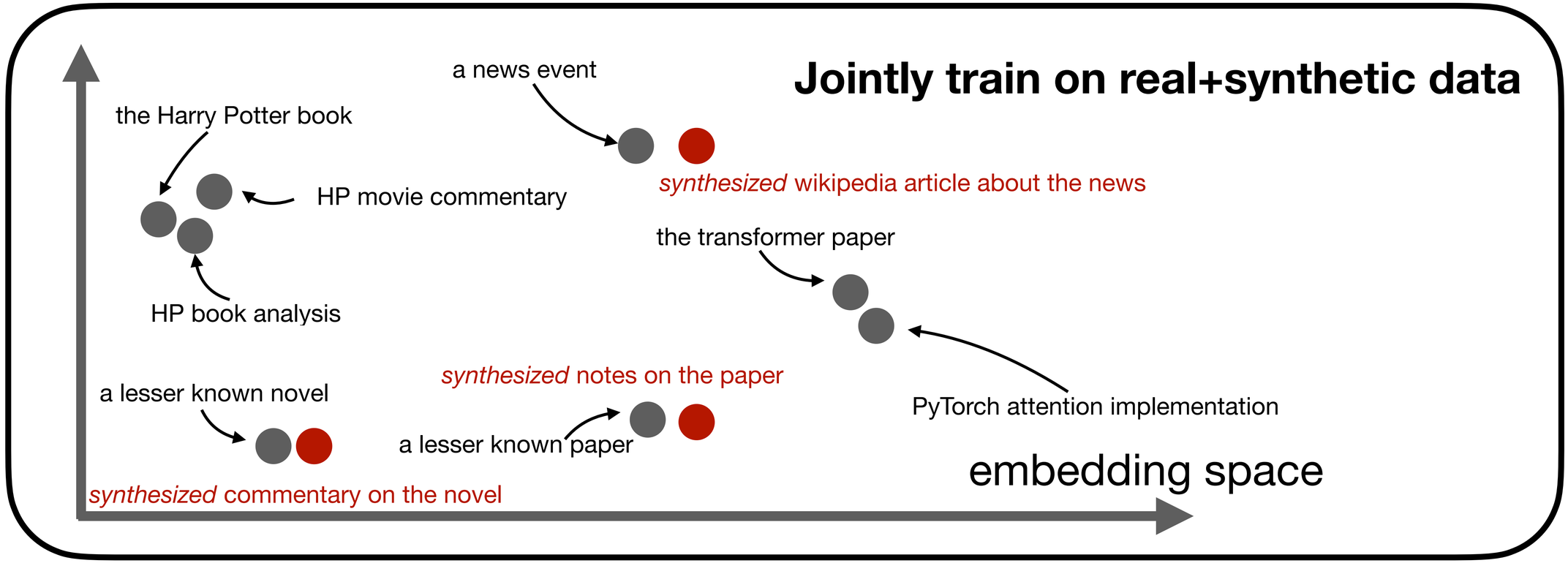
Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_2 \mid d_1)$ initialized at pretrained checkpoint
3. **Synthesis at scale:** Synthesize multiple paired documents by setting Temp=1



Synthetic bootstrapped pretraining

1. **Nearest-neighbor pairing:** DCLM subset and Qwen-0.6B-Embedding
2. **Synthesizer tuning:** SFT-like objective $p_\theta(d_2 \mid d_1)$ initialized at pretrained checkpoint
3. **Synthesis at scale:** Synthesize multiple paired documents by setting Temp=1



Examples of synthesis

Real document

HOME > Chowhound > California >

Coffeehouses in SD

I am always looking for good coffeehouses in San Diego. I am looking for coffeehouses with (obviously) excellent coffeebeans, baristas who know their business (good microfoam, good crema, they should know the different ratios for steamed milk, foamed milk, espresso for the cappuccino, latte etc., etc), good snacks (paninis, cakes etc), good (cold) flavored coffeedrinks and good ambience. So far I haven't found many who are good but would like to hear more suggestion.

Caffe Calabria: By far the best coffeehouse in SD. Excellent coffee and baristas (best cappuccino in SD,), great paninis. Only downside is that they close very early.

Café Lestat: Good example what happens with good coffee and average baristas. Lestat is also using coffeebeans from Caffé Calabria but it is sad to see that they don't know how to make excellent espresso, cappuccino.

E Street Café: Average espresso but strange flavored ...

Synthesis I

Coffee = Love

On a recent trip to San Diego, one thing I wanted to find was coffee. I was worried it might be difficult with the huge decrease in Asian flavors, but I was pleasantly surprised. Many of the Mexican, Italian, and Cuban flavors that are pretty common in California are now in the mix. While I did find a few highly recommended spots, the majority of the locations I checked out were nothing to write home about. At the least, they were not being sold out, which I consider a plus.

One major factor that really needs to be taken into account is the type of machine being used. There are two types of espresso machines: the cheap ones that spit out lots of pressure and don't have any control over the shot, and the expensive ones that create amazing shots and utilize the types of superfine grinds you usually find in Italy. The latter types are much harder to find and cost significantly more.

Another factor to consider is the type of beans being used. If they are done correctly, great espresso can be made from any type of bean, but it's a rare experience to find a place that does ...

Synthesis II

Discuss as:

Coffeehouses: San Diego. The best?

Café Lestat, by far, is a stand-out. (photo by Thomas Moriarty)

If the coffee scene in San Diego sounds like the one in New York, there's a reason. The upper echelons of NYC's drink scene are the same ones who have taken over the majority of the coffee houses in San Diego. This is probably partly due to the fact that both cities have large student populations, both are home to large coffee drinker bases, both have a large number of espresso shops, and both offer top-notch, hand-crafted cappuccinos. But if there is one city in America that deserves to have its name above that of New York, it's San Diego.

There are just under 100 coffee shops in San Diego, with almost half of them located on University Ave. alone. So finding the perfect coffee shop is crucial. We spent a whole day just roaming around the area, hunting for the best.

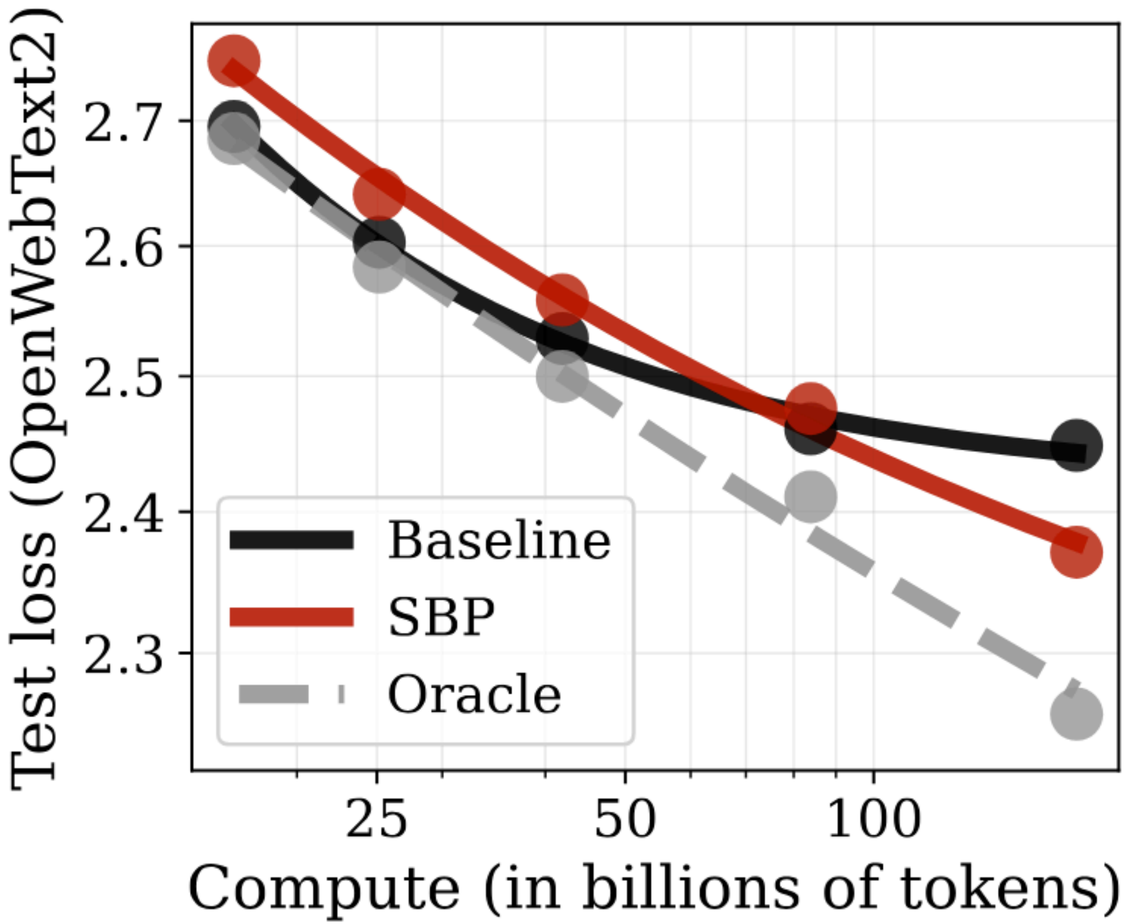
In terms of the coffee itself, it's hard to beat Café Lestat. The baristas are amazing and their methods are pristine ...

40% of improvement attained by oracle with 20x data

Benchmark	200B-scale			1T-scale		
	Baseline	SBP	Oracle	Baseline	SBP	Oracle
	Perplexity on held-out data ↓					
OpenWebText2	5.74	-0.53	-1.02	4.51	-0.02	-0.12
LAMBADA	6.87	-0.85	-1.86	4.33	-0.03	-0.22
Five-shot MMLU	3.83	-0.36	-0.51	3.17	-0.06	-0.05
	QA accuracy ↑					
ARC-Challenge (0-shot)	35.32	+1.28	+2.82	42.66	+1.62	+3.84
ARC-Easy (0-shot)	68.94	+2.65	+4.29	75.63	+0.42	+2.11
SciQ (0-shot)	90.50	+1.00	+2.40	93.20	+0.80	+0.50
Winogrande (0-shot)	60.14	+1.90	+5.53	65.19	+1.42	+2.92
TriviaQA (1-shot)	22.51	+3.36	+7.37	36.07	+0.25	+0.59
WebQS (1-shot)	8.56	+3.74	+10.83	19.34	+0.54	+0.44
Average QA accuracy	**47.66**	**+2.32**	**+5.54**	**55.35**	**+0.84**	**+1.73**

Training dynamics

- Initially, baseline and oracle perform similarly. SBP is worse because it uses synthetic data
- Later on, baseline and oracle diverge; SBP follows a linear trend
- Near the end, baseline plateaus while SBP continues to improve

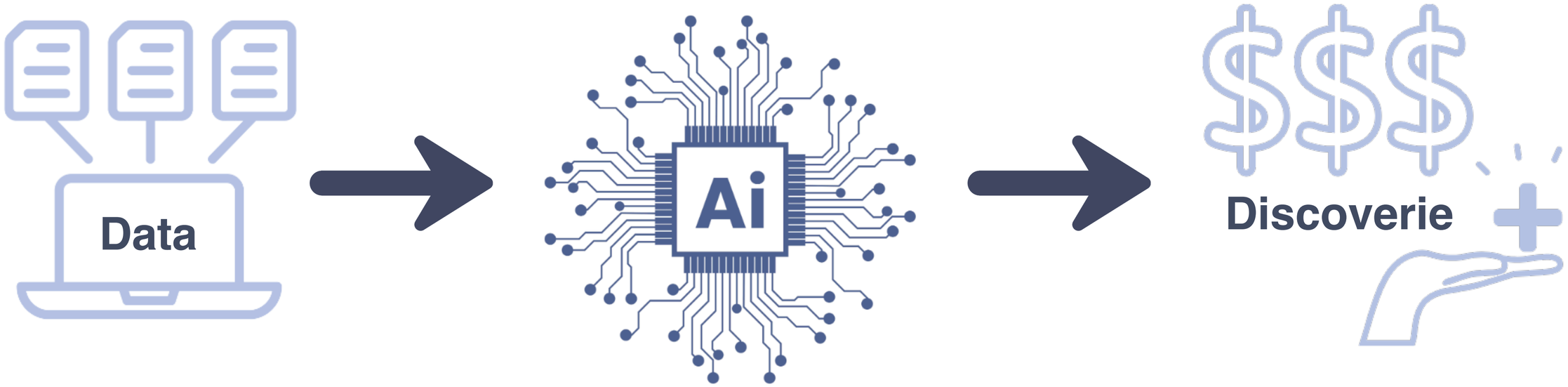


Synthetic continued/bootstrapped pretraining: summary

- Biggest reservoir of machine learnable knowledge resides in unsupervised learning (Y. leCun) — witness LLM from GPT-3 on

- As we run out of internet data, we propose a form of self-supervision weaker than next-token prediction, exploiting existing knowledge/correlations on the internet

Modern discovery pipeline



Thesis: Thinking carefully about AI inputs and outputs yields **more powerful**, **safer** AI

Agenda: vignettes on three pillars

Data collection

Data-driven discovery

Quality control

Data collection

Data-driven discovery

Quality control

synthetic pretraining
datamodels
s1

Data collection

Data-driven discovery

Quality control

synthetic pretraining
datamodels
s1

AI-powered inference

Increasing life expectancy

NEW YORK, WEDNESDAY, APRIL 13, 1955. Times Square, New York 36, N. Y. Telephone LAckawanna 4-1000 FIVE CENTS

GH COURT HEARS
)UTH WILL DEFY
UICK END TO BIAS

dual Approaches Urged
Integration of Schools—
egro Lawyers Opposed

y LUTHER A. HUSTON
Special to The New York Times.

ASHINGTON, April 12—
ceamen for South Carolina
Virginia told the Supreme
t today that their people
ld not obey a decree ordering
mmediate end to racial seg-
tion in the public schools.
hen Chief Justice Earl War-
asked S. E. Rogers, repre-
ing Clarendon County, S. C.,
s were willing to say that
'honest attempt" would be
to conform to whatever
ee the court might issue, Mr.
rs said:
et's get that word 'honest'
of there. It would depend
the kind of decree. The

SALK POLIO VACCINE PROVES SUCCESS;
MILLIONS WILL BE IMMUNIZED SOON;
CITY SCHOOLS BEGIN SHOTS APRIL 25

TRIAL DATA GIVEN

Efficacy of 80 to 90% Shown—Salk Sees Further Advance

Abstract of report, summary of data on tests, Page 22.

By WILLIAM L. LAURENCE
Special to The New York Times.

ANN ARBOR, Mich., April 12—The world learned today that its hopes for finding an effective weapon against paralytic polio had been realized.

No perfect mechanistic understandings of vaccines → we must learn from **data**

How sure are we of what we learn from data?

Achieve the **gold standard**, i.e. produce confidence bounds L, U from data such that

$$\text{Prob}(L < \underbrace{\text{object of inference}}_{\text{level of deforestation in the Amazon}\\ \text{effectiveness of a vaccine}\\ \text{side effect of a drug}} < U) = 95\,\%$$

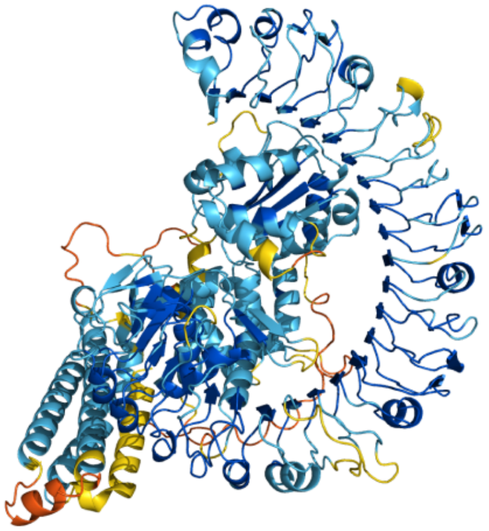
Warning! before clinical trials Rasmussen (2008)

"Large quantities of amphetamines were dispensed in the 1960s by weight loss clinics."

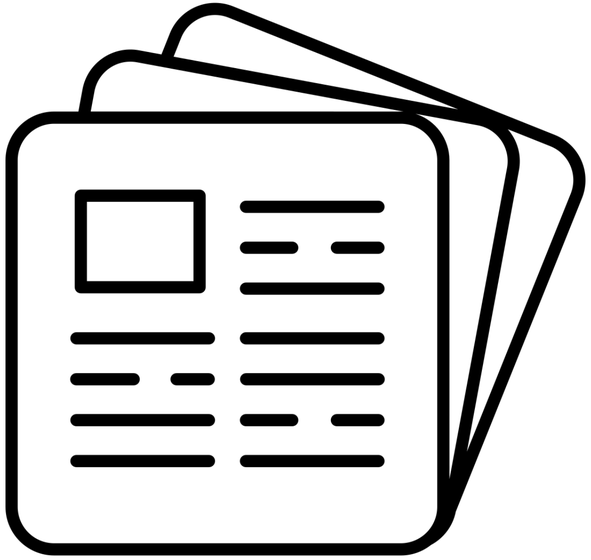
One estimate = **2 billion tablets annually**; millions suffered cardiovascular damage.
$\Longrightarrow$ reliable results matter!

Can we leverage imperfect AI predictions to get gold-standard results?

How much deforestation is in the Amazon?



Which fraction of proteins have a certain property?

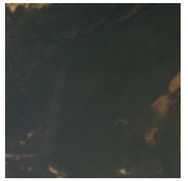
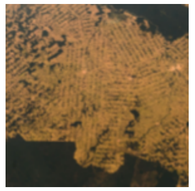


How much has political sentiment in the media changed?

Setup

features	labels
X_1	?
X_2	?
⋮	⋮
X_{N-1}	?
X_N	?

N unlabeled data points

	satellite images	deforestation %
X_1		?
X_2		?
	⋮	⋮
X_{N-1}		?
X_N		?

Goal: learn scientific target $T = \text{mean}(Y_1, \ldots, Y_N)$

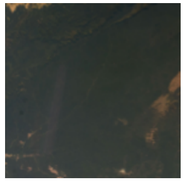
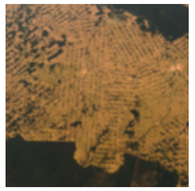
$T =$ deforestation rate in the Amazon

Setup

features	labels
X_1	Y_1
⋮	⋮
X_n	Y_n
⋮	⋮
X_{N-1}	?
X_N	?

We can collect at most n expert labels Y_i (much fewer than N)

$N - n$ unlabeled data points

	satellite images	deforestation %
X_1		15%
X_2		7%
	⋮	⋮
X_{N-1}		?
X_N		?

Goal: learn scientific target $T = \text{mean}(Y_1, \ldots, Y_N)$

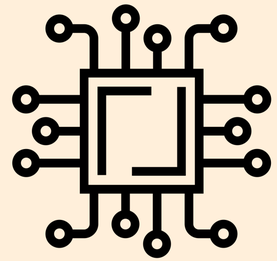
$T =$ deforestation rate in the Amazon

Setup

features	labels	
X_1	Y_1	$\hat{Y}_1$
⋮	⋮	⋮
X_n	Y_n	$\hat{Y}_n$
⋮	⋮	
X_{N-1}	$\hat{Y}_{N-1}$	
X_N	$\hat{Y}_N$	

We can collect at most n expert labels Y_i (much fewer than N)

N predicted labels



AI model

produces informative but imperfect predictions $\hat{Y}_1, \ldots, \hat{Y}_N$

model can be fine tuned on same labels

Goal: learn scientific target $T = \text{mean}(Y_1, \ldots, Y_N)$

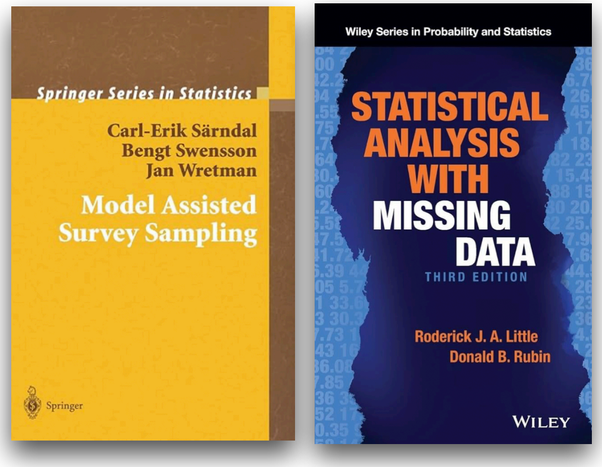
Gold-standard results with AI

Step 1: Collect n randomly chosen expert labels Y_i

Step 2: Given $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}, \ldots, X_N$, estimate the scientific target T

$$T^{\text{PP}} = \underbrace{\text{mean}(\hat{Y}_1, \ldots, \hat{Y}_N)}_{\text{if we pretended AI predictions were correct}} - \underbrace{\text{mean}(\hat{Y}_1 - Y_1, \ldots, \hat{Y}_n - Y_n)}_{\text{bias of AI predictions}}$$

*This trick is based on classical statistical thinking, e.g.,



No matter the AI bias, can make sure we achieve the gold standard!

Angelopoulos et al. (2023), Zrnic, Candès (2024a)

Gold-standard results with AI

Classical inference

$$\hat{\theta}^{\mathrm{CL}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Unbiased
- Variance: $\mathrm{Var}\left(\hat{\theta}^{\mathrm{CL}}\right) = \frac{1}{n}\mathrm{Var}(Y)$

Prediction-powered inference (PPI)

$$\hat{\theta}^{\mathrm{PP}} = \frac{1}{N}\sum_{i=1}^{N} f\left(X_i^{\mathrm{unlabeled}}\right) + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - f\left(X_i^{\mathrm{labeled}}\right)\right)$$

$\hat{Y} = f(X)$
blackbox prediction

- Unbiased
- Variance:

$$\mathrm{Var}\left(\hat{\theta}^{\mathrm{PP}}\right) \approx \frac{1}{n}\mathrm{Var}(Y - f(X))$$

→ When predictions are good and $N \gg n$, $\hat{\theta}^{\mathrm{PP}}$ has **lower variance**!

Angelopoulos et al. (2023), Zrnic, Candès (2024a)

Proteomics with AlphaFold

X_i — protein sequences, Y_i — indicator of disorder (IDR)

T — relationship between disorder and phosphorylation

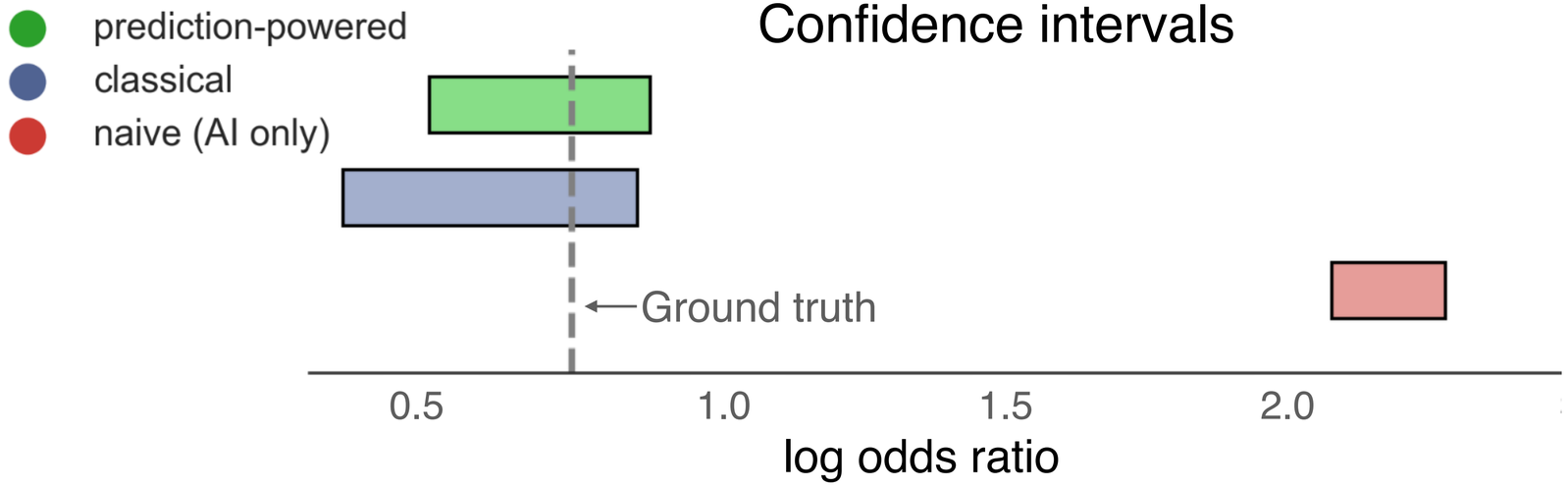
— AlphaFold

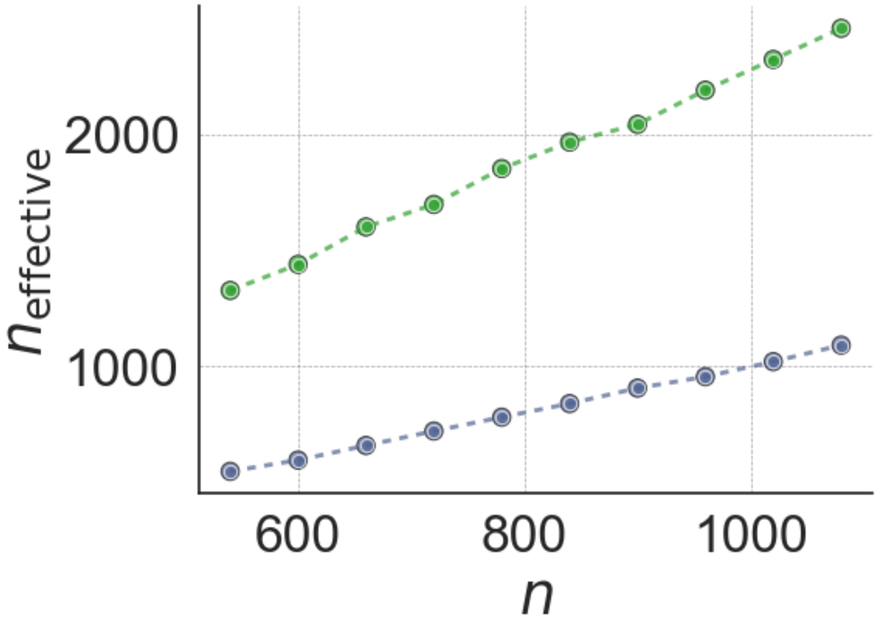
PLOS BIOLOGY

The structural context of posttranslational modifications at a proteome-wide scale

Isabell Bludau[1], Sander Willems[1], Wen-Feng Zeng[1], Maximilian T. Strauss[2], Fynn M. Hansen[1], Maria C. Tanzer[1], Ozge Karayel[1], Brenda A. Schulman[3], Matthias Mann[1,2]*

Protein disorder vs phosphorylation





Which labels are most informative?

X_1	?
X_2	?
⋮	⋮
X_{N-1}	?
X_N	?

← should prioritize collecting expert labels where AI makes a mistake

→ learn more given the same budget

want $p_i = \text{Prob}(\text{collect label } Y_i)$ small for "easy-to-predict" data points and large for "hard" ones

$$T^{\text{adaptive}} = \text{mean}(\hat{Y}_1, \ldots, \hat{Y}_N) - \underbrace{\text{mean}\left(\frac{1}{p_1}(\hat{Y}_1 - Y_1), \ldots, \frac{1}{p_n}(\hat{Y}_n - Y_n)\right)}_{\text{bias of AI predictions when expert labels are collected adaptively}}$$

No matter the AI bias, can achieve the gold standard!

Zrnic, Candès (2024b

Data-adaptive sampling with active inference

$$\hat{\theta}^{\text{Active}} = \frac{1}{N}\sum_{i=1}^{N} f\left(X_i^{\text{unlabeled}}\right) + \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{\pi(X_i)}\left(Y_i - f\left(X_i^{\text{labeled}}\right)\right)$$

$\hat{Y} = f(X)$ blackbox prediction

$\xi_i \sim \text{Bern}(\pi(X_i))$ indicates whether label Y_i is collected

Labeling policy: $\pi(X_i) \propto$ uncertainty $u(X_i)$

$$u(X_i) \propto \begin{cases} \sqrt{E(Y_i - f(X_i))^2 \,|\, X_i} & \text{(regression)} \\ 2\min(f(X_i), 1 - f(X_i)) \quad f(x_i) = \hat{P}(Y_i = 1 \,|\, X_i) & \text{(classification)} \end{cases}$$

This can be done sequentially: fine-tune f after collecting some labels

Zrnic, Candès (2024b)

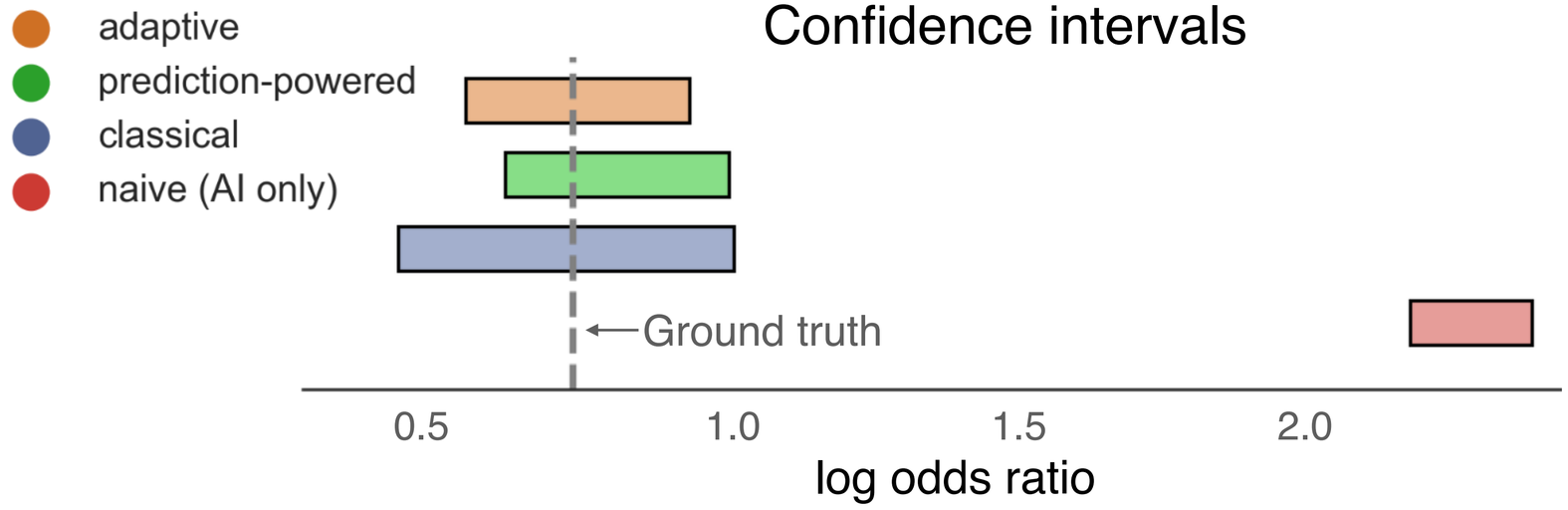
Proteomics with AlphaFold

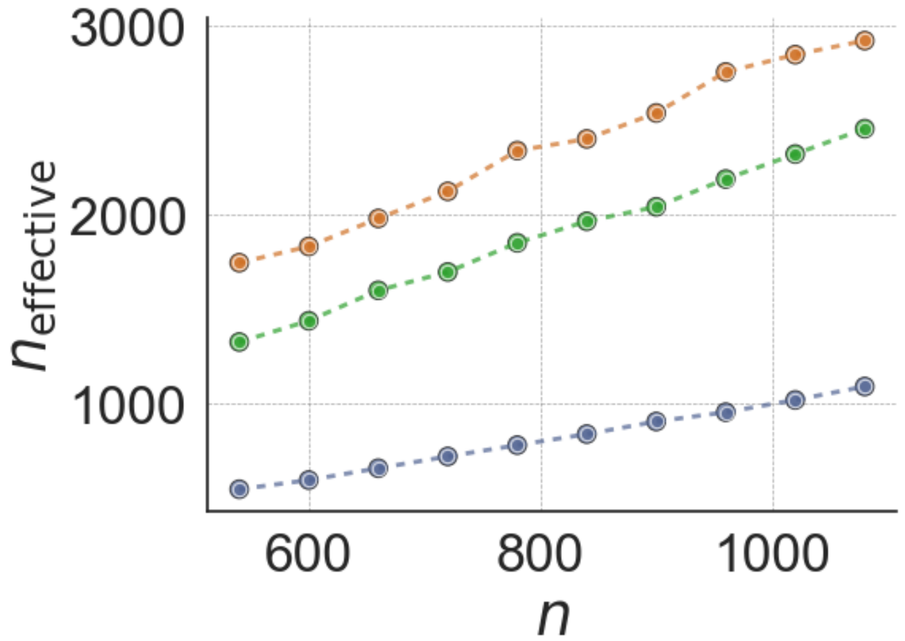
X_i — protein sequences, Y_i — indicator of disorder (IDR)

T — relationship between disorder and phosphorylation

— AlphaFold

Protein disorder vs phosphorylation





C., Ilyas, Zrnic (2025)

All-purpose efficient dataset labeling

X_1	$\tilde{Y}_1$
X_2	$\tilde{Y}_2$
$\vdots$	$\vdots$
X_{N-1}	$\tilde{Y}_{N-1}$
X_N	$\tilde{Y}_N$

N unlabeled data points

Want to collect labels $\tilde{Y}_i$ that are both **accurate** and **cheap**

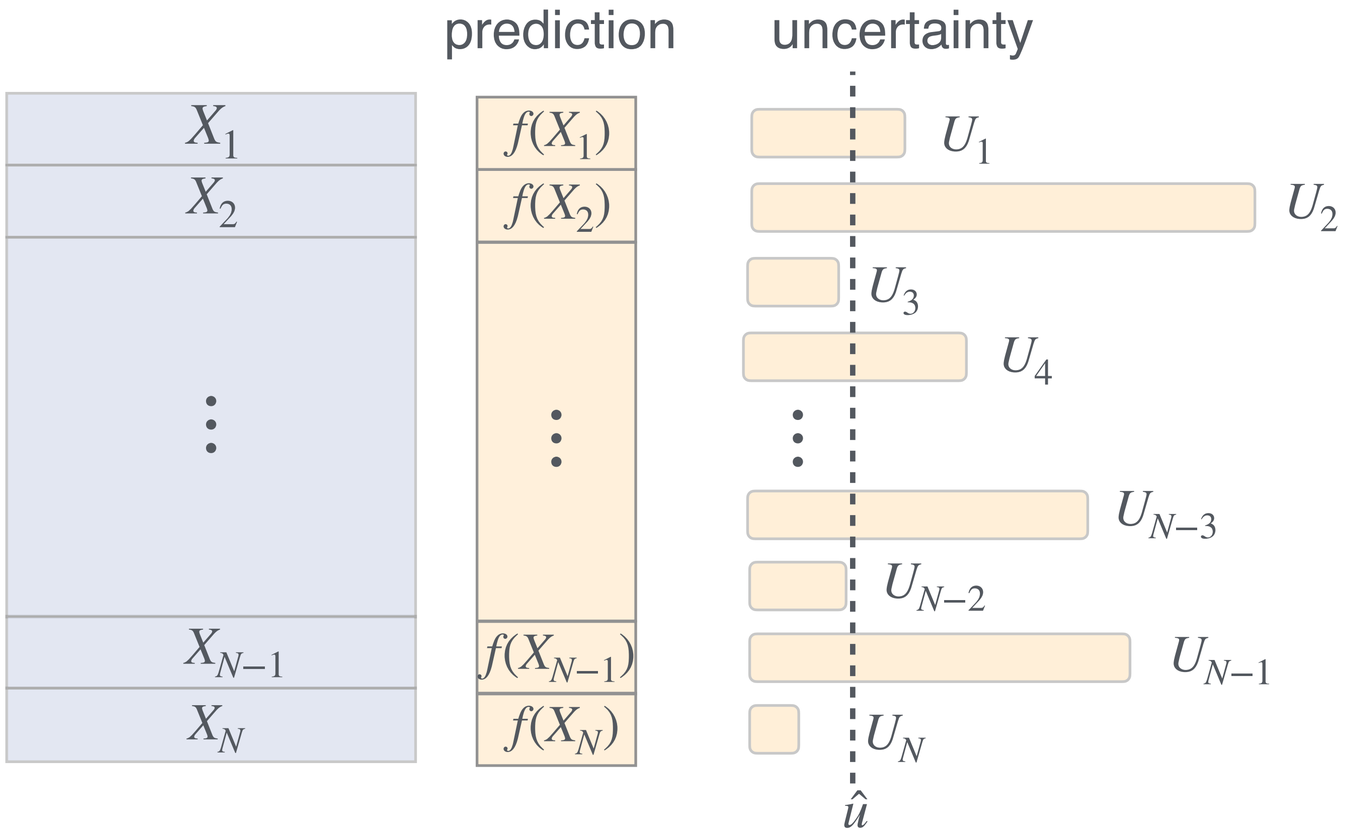
$\tilde{Y}_i$ can be either expert label Y_i (expensive) or AI prediction $\hat{Y}_i$ (cheap)

$$\tilde{Y}_i^u = Y_i \cdot \mathbf{1}\{U_i \geq u\} + \hat{Y}_i \cdot \mathbf{1}\{U_i < u\}$$

Goal: return labeled dataset $(X_1, \tilde{Y}_1), \ldots, (X_N, \tilde{Y}_N)$, while collecting as few expert Y_i as possible, such that

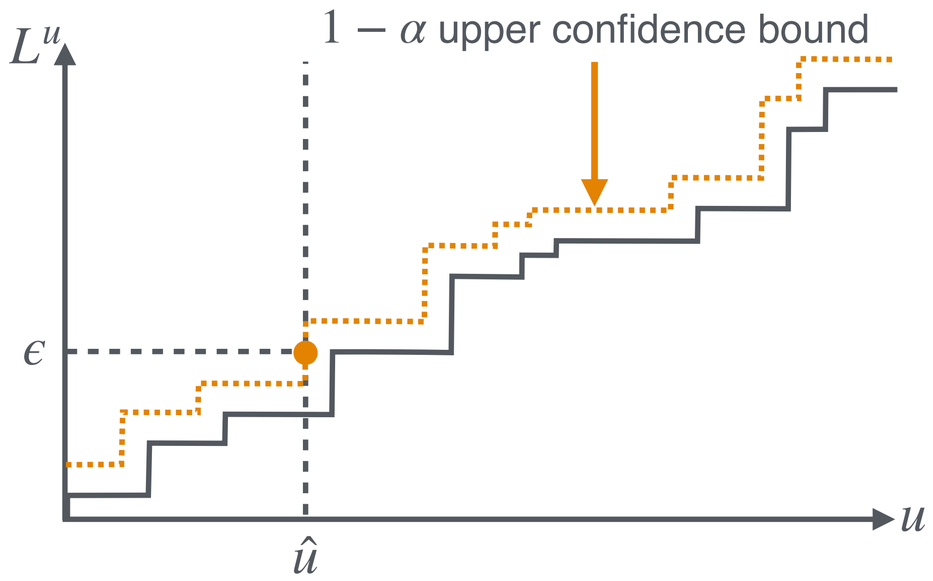
$$\tilde{Y}_i = Y_i \text{ for 95\% of the dataset, with 95\% probability}$$

PAC labeling method



$$\tilde{Y}_i^u = Y_i \cdot \mathbf{1}\{U_i \geq u\} + f(X_i) \cdot \mathbf{1}\{U_i < u\}$$

$$L^u = \frac{1}{N} \sum_{i=1}^{N} \ell(Y_i, \tilde{Y}_i^u)$$



Theorem $\tilde{Y}_1^{\hat{u}}, \ldots, \tilde{Y}_N^{\hat{u}}$ are PAC labels

C., Ilyas, Zrnic (2025)

All-purpose social science PAC labelling with GPT-4o

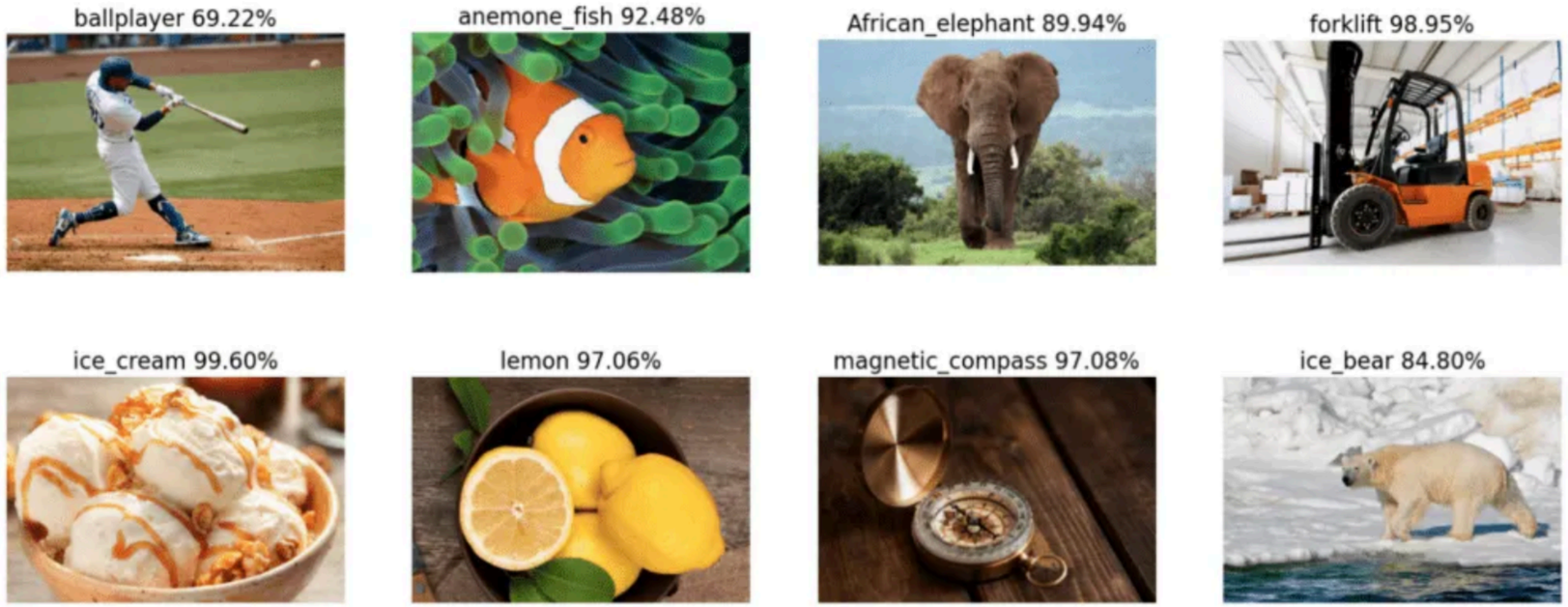
Dataset	Metric	Method	
		Our approach	ChatGPT only
Misinformation [1]	Budget save (%)	(18.12 ±4.93)%	—
	Error	3.80%	18.56%
Stance on global warming [2]	Budget save (%)	(28.09 ±3.28)%	—
	Error	4.57%	24.79%
Media bias [3]	Budget save (%)	(13.79 ±3.38)%	—
	Error	4.10%	37.72%

[1] Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. Gabriel et al. (2022)

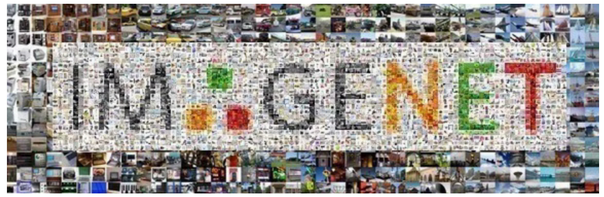
[2] Detecting Stance in Media on Global Warming. Luo et al. (2020)

[3] We Can Detect Your Bias: Predicting the Political Ideology of News Articles. Baly et al. (2020)

Image classification with ResNet



$Y_i \in$ 1000 ImageNet labels



ImageNetV2

Ben Recht Ludwig Schmidt Rebecca Roelofs Vaishaal Shankar

[1] Deng et al. CVPR, 2009
[2] Recht et al. ICML, 2019

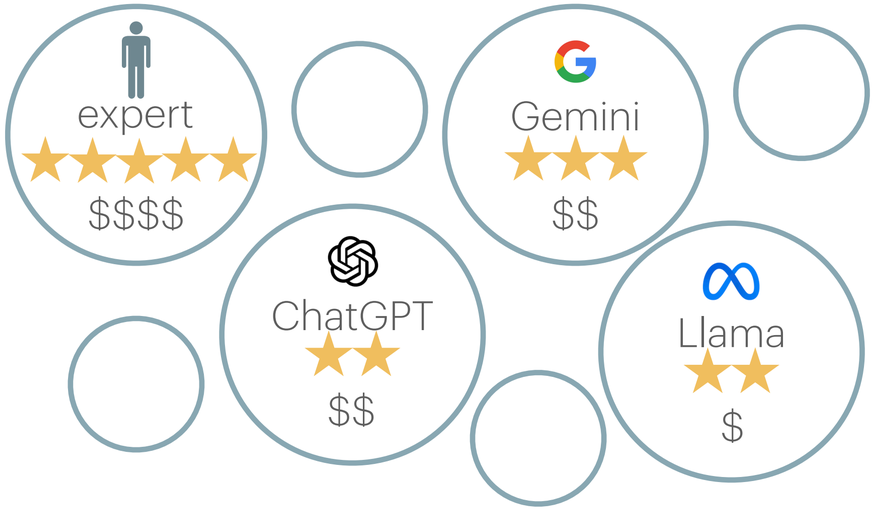
PAC labelling image datasets with ResNet-152

Dataset	Metric	Method	
		Our approach	ResNet only
ImageNet [1]	Budget save (%)	(59.64 ±1.49)%	—
	Error	4.73%	21.69%
ImageNet v2 [2]	Budget save (%)	(39.07 ±2.67)%	—
	Error	4.74%	35.33%

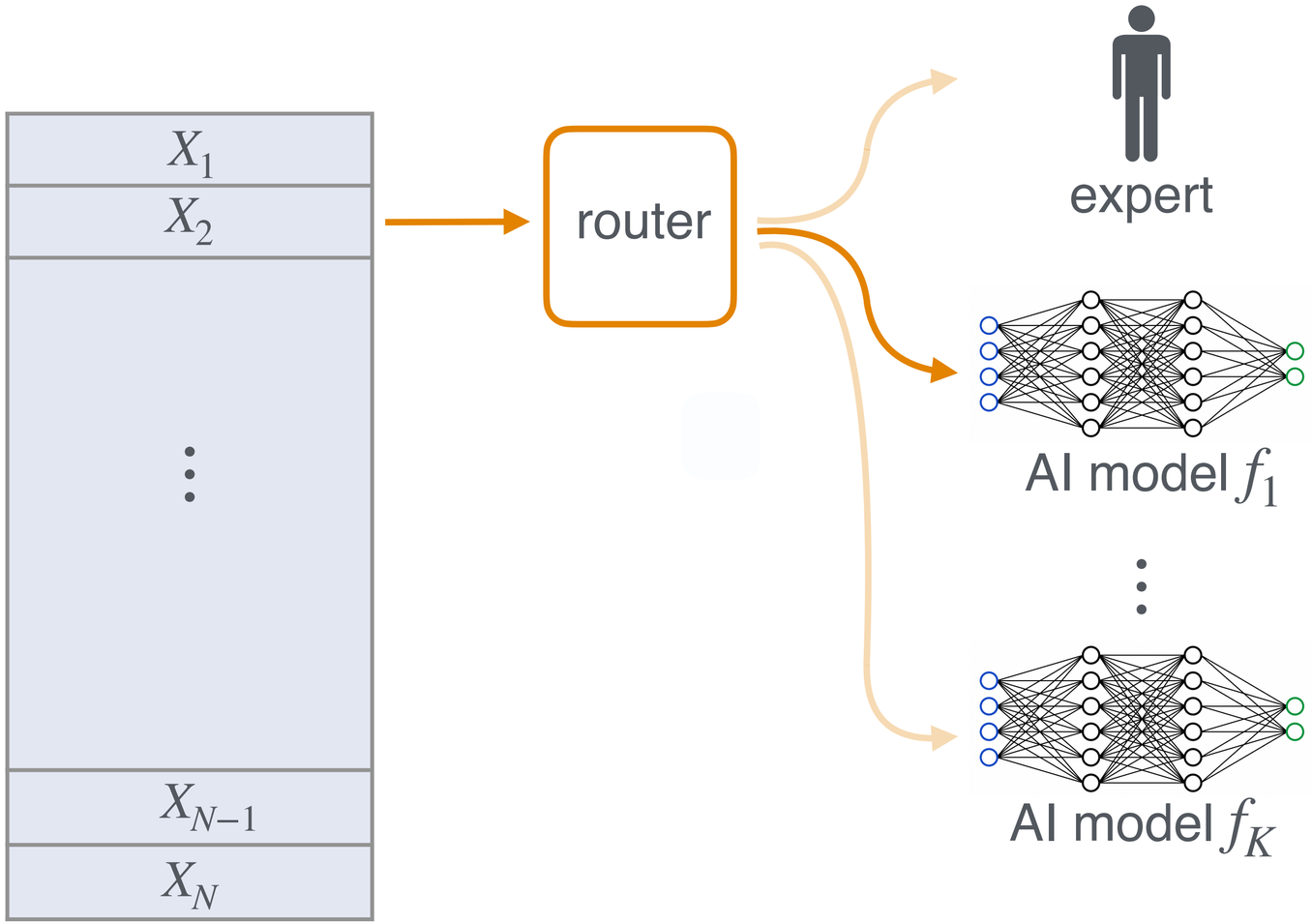
[1] Deng et al. CVPR, 2009
[2] Recht et al. ICML, 2019

Labeling with multiple models

How do we reliably trade off between data sources of varying qualities and costs?



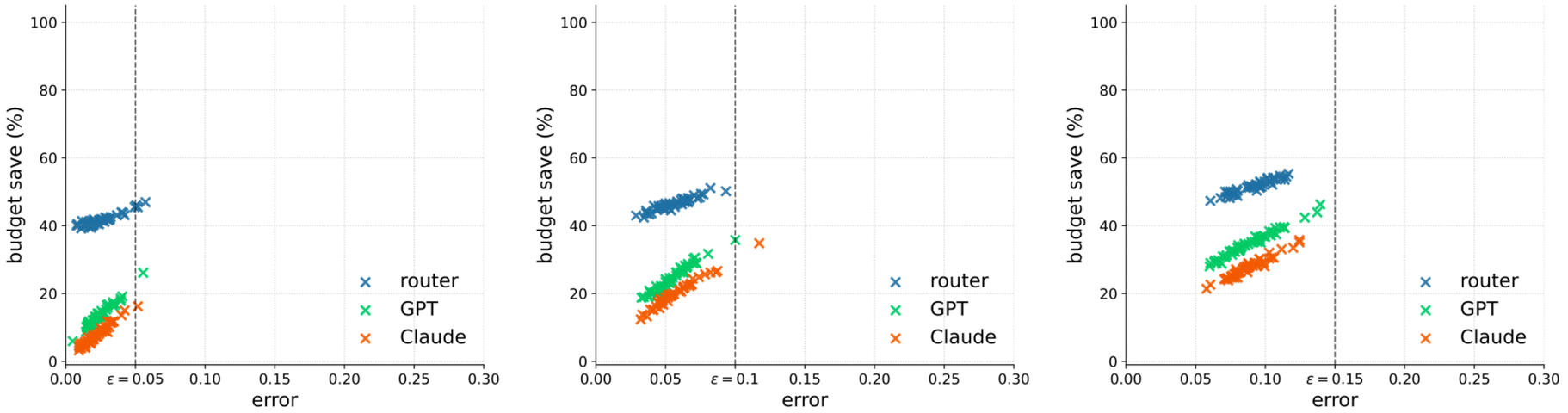
PAC labeling with multiple models



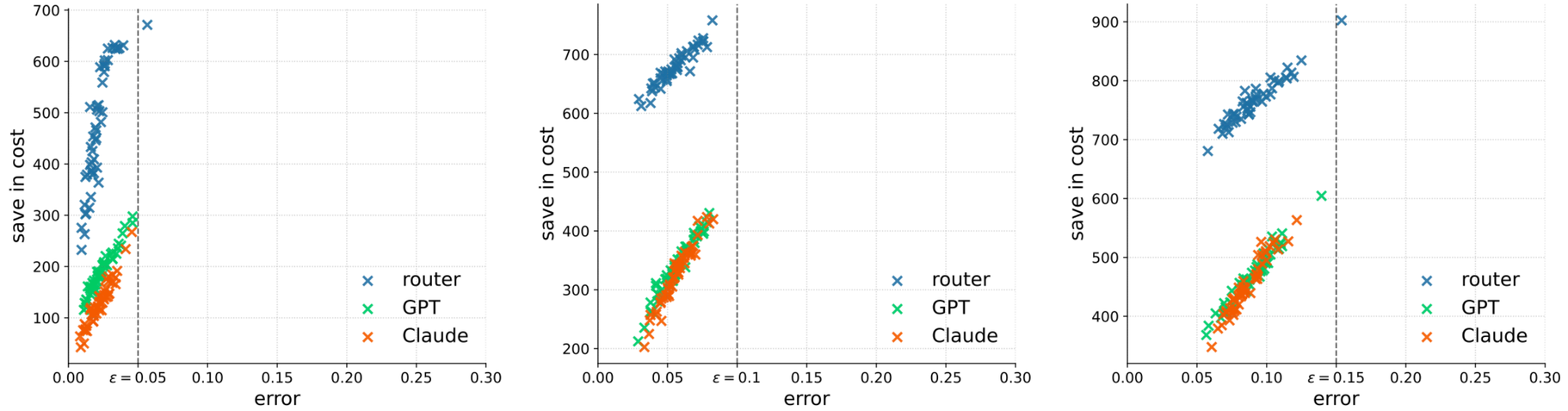
Router optimizes final labeling cost

PAC labeling with multiple models

Media bias (with router)

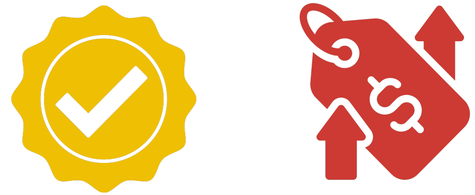


Cost-sensitive ($) routing:

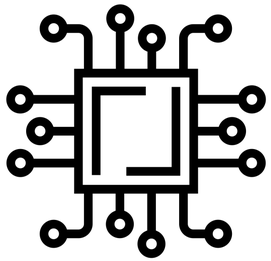


Takeaway

statistics alone

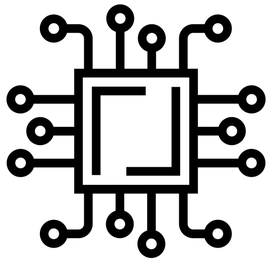


reliable, but



AI alone

not always reliable, but powerful and



statistics + AI

reliable and

Data collection

Data-driven discovery

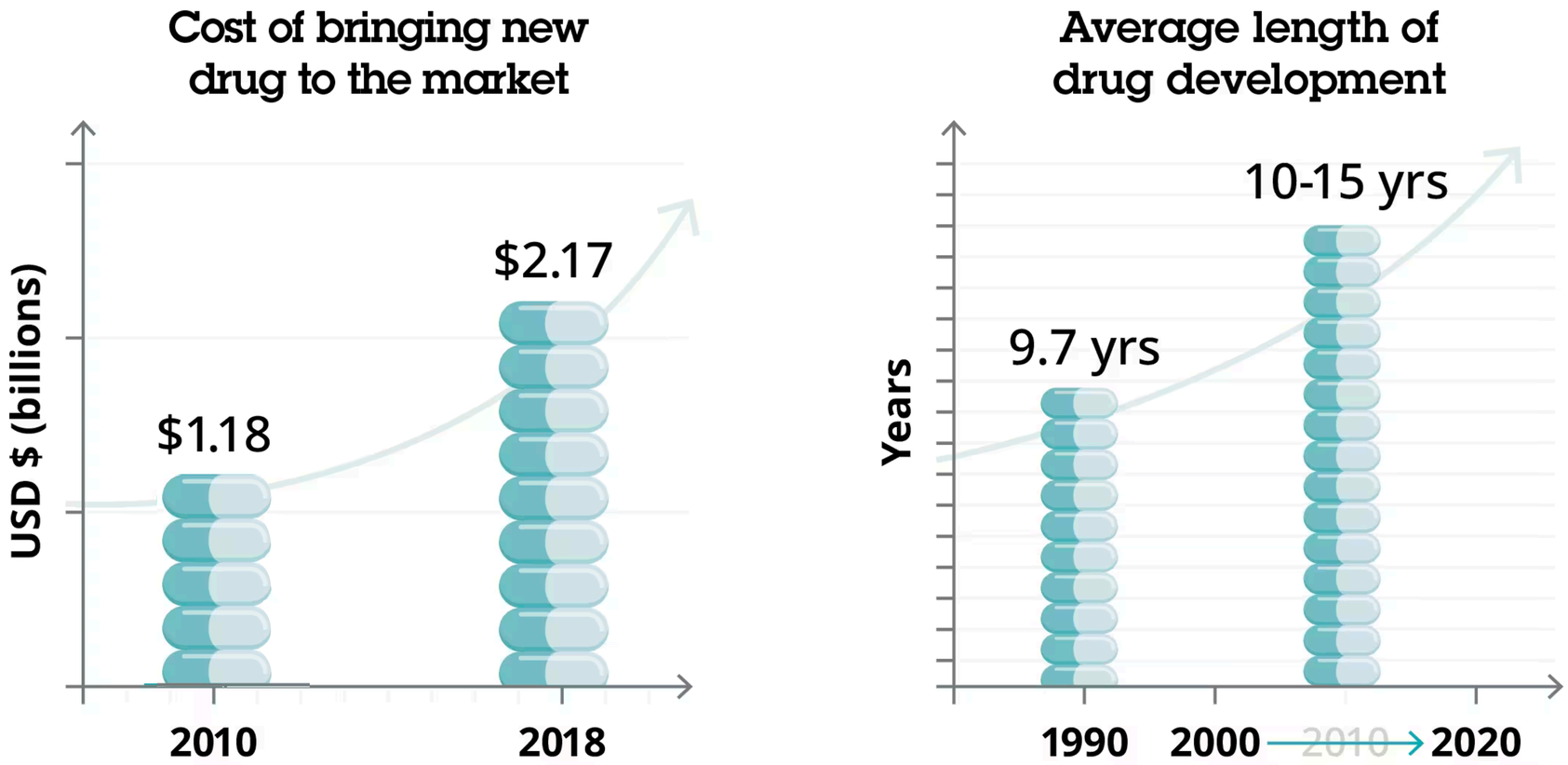
Quality control

synthetic pretraining datamodels s1

AI-powered inference

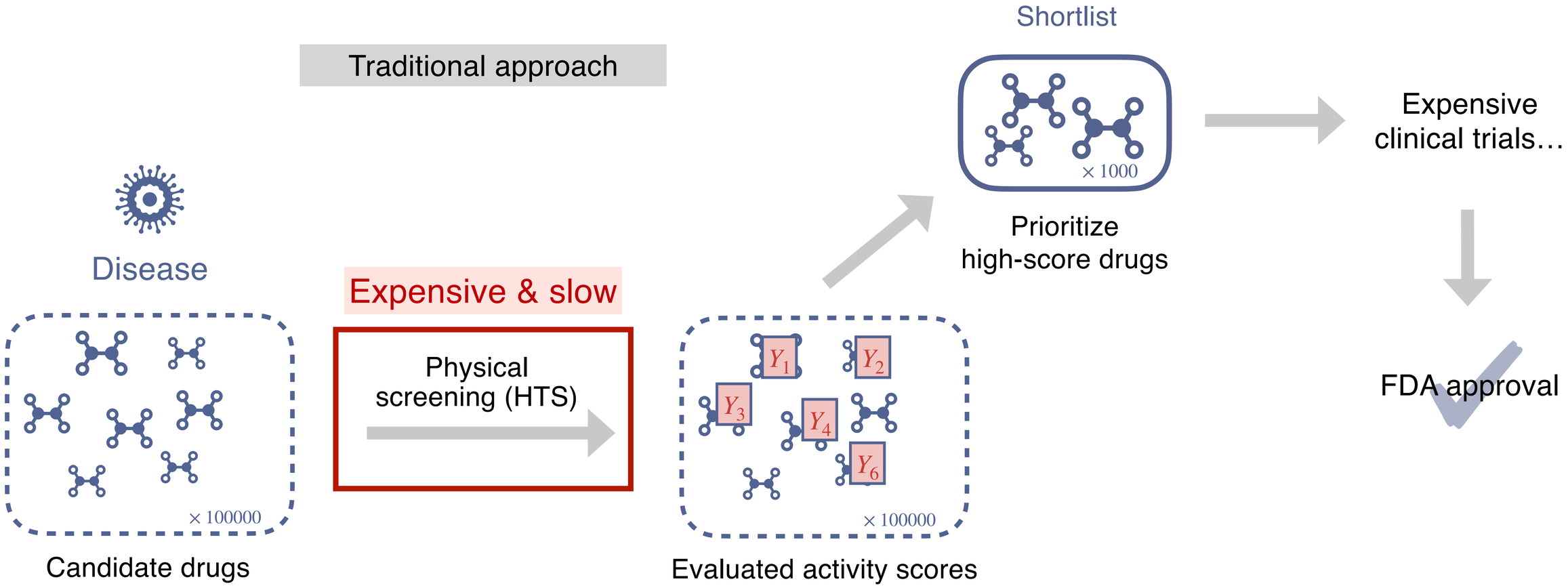
AI-powered drug discovery

Accelerating drug discovery?

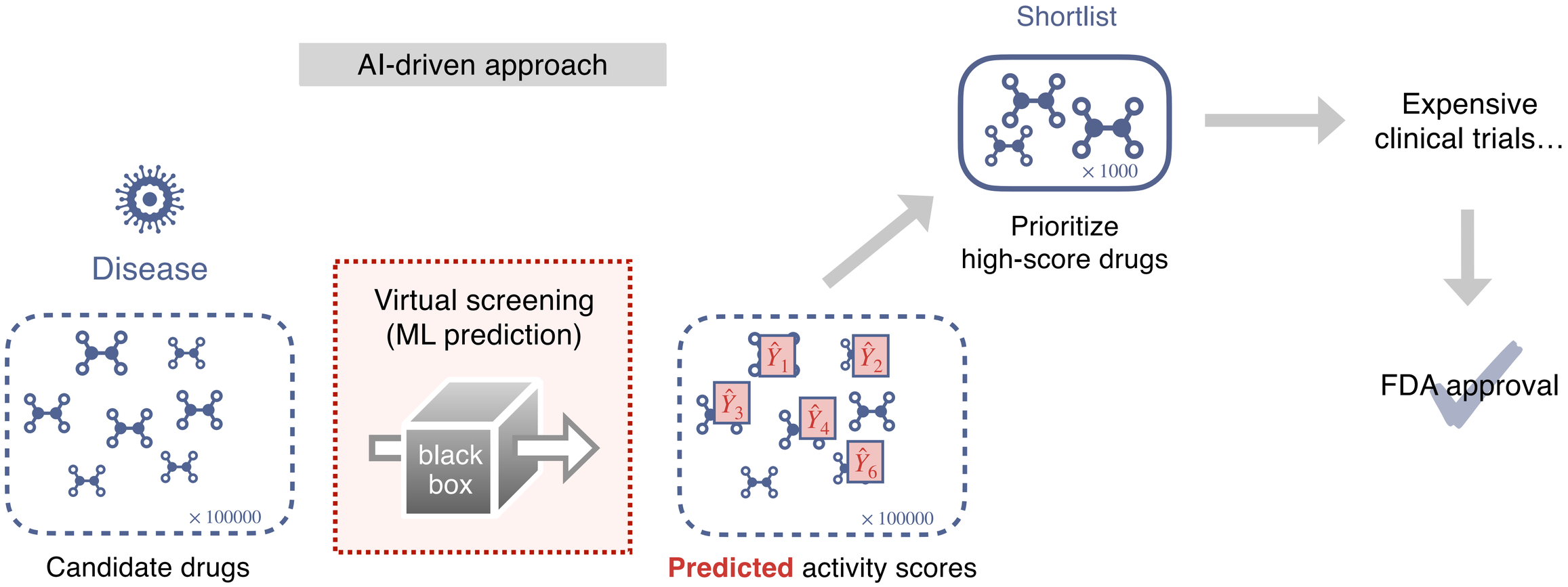


[abcam.com]

Drug discovery pipeline

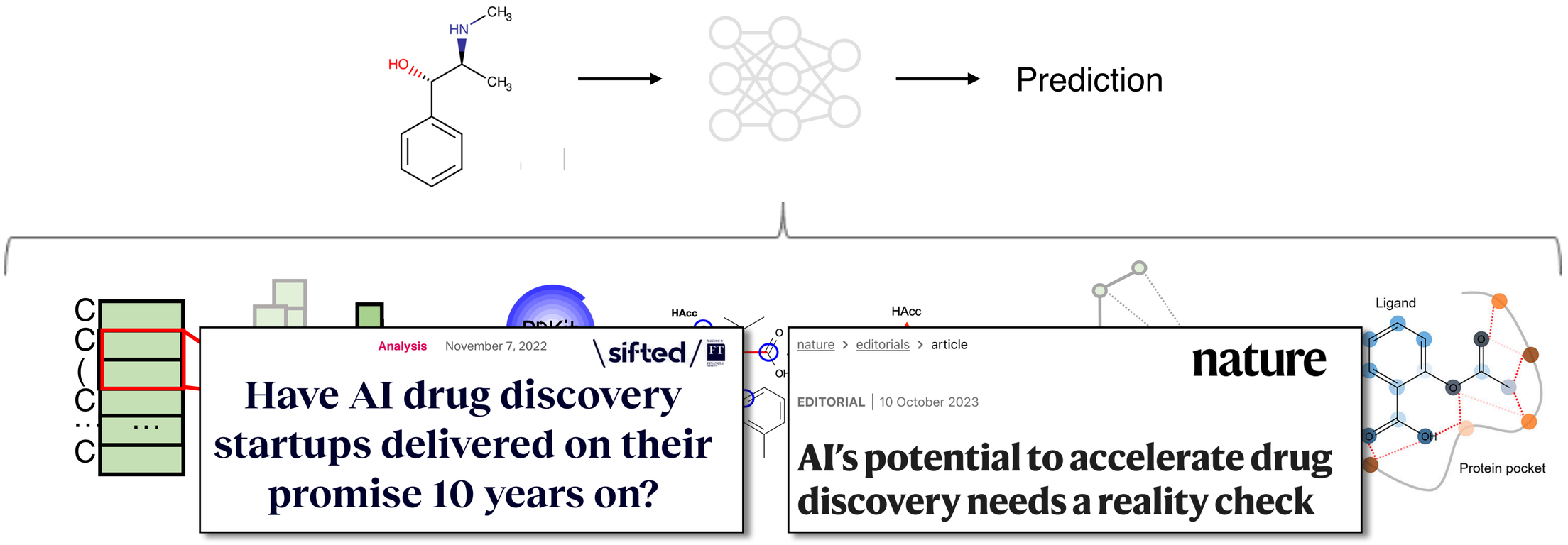


Drug discovery pipeline



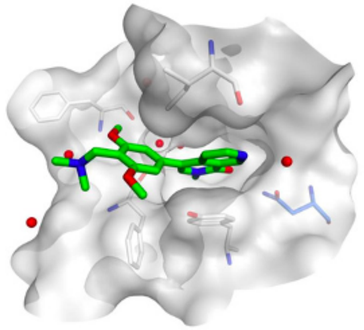
AI as imperfect scoring?

Quality control?



Can we make discoveries with *few mistakes*?

Goal: finding “actionable” instances

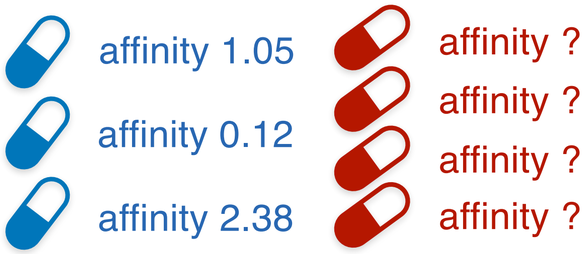


Want drugs with high binding affinities to a disease target

Which drugs are sufficiently active?

Problem setup

- Any pre-trained prediction model $\hat{\mu}: \mathcal{X} \to \mathcal{Y}$ (independent of training and test data)
  - X physical/chemical feature/amino acids of the drug
  - Y binding affinity
    - $\rightsquigarrow Y \in \{0,1\}$: whether the drug binds to the target
    - $\rightsquigarrow Y \in \mathbb{R}$: how well the drug binds to the target

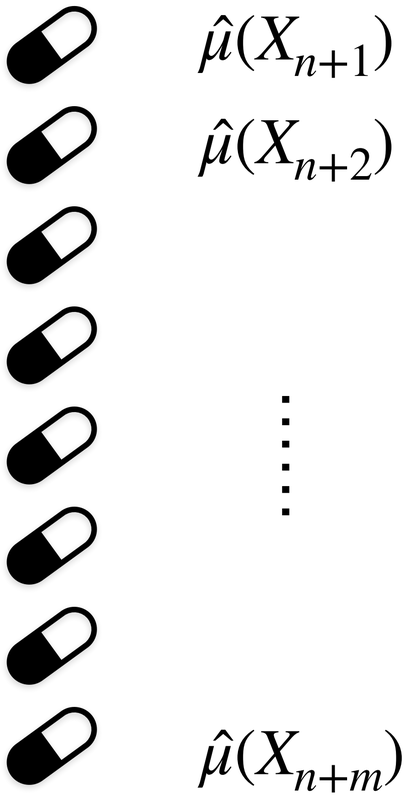


- Training data $\{(X_i, Y_i)\}_{i=1}^{n}$ (screened drugs)
- Test samples $\{(X_{n+j}, Y_{n+j})\}_{j=1}^{m}$ with unknown $\{Y_{n+j}\}_{j=1}^{m}$ (new drugs)

Goal: find large outcomes $Y_{n+j} > c_{n+j}$ without too many errors

$\rightsquigarrow$ user-specified thresholds c_{n+j} to become 'interesting'

Challenges



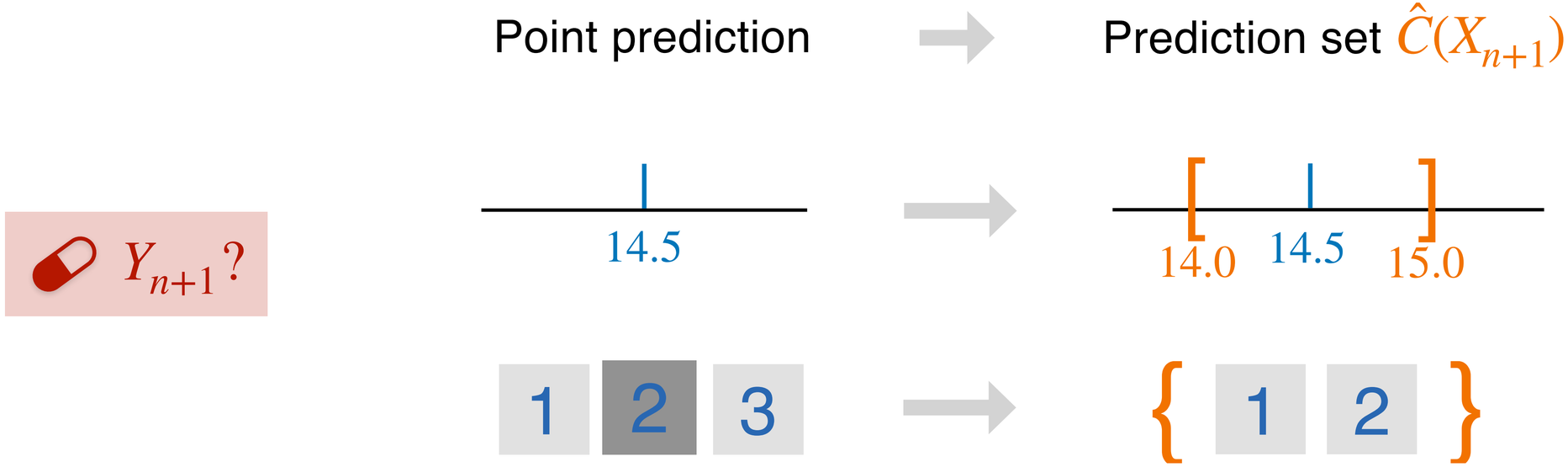
- Quantifying uncertainty in point predictions
- Model-free

  Work for any prediction model

  No modeling assumptions
- Distribution shift

Which drugs are sufficiently active?

Conformal prediction: model-free uncertainty quantification

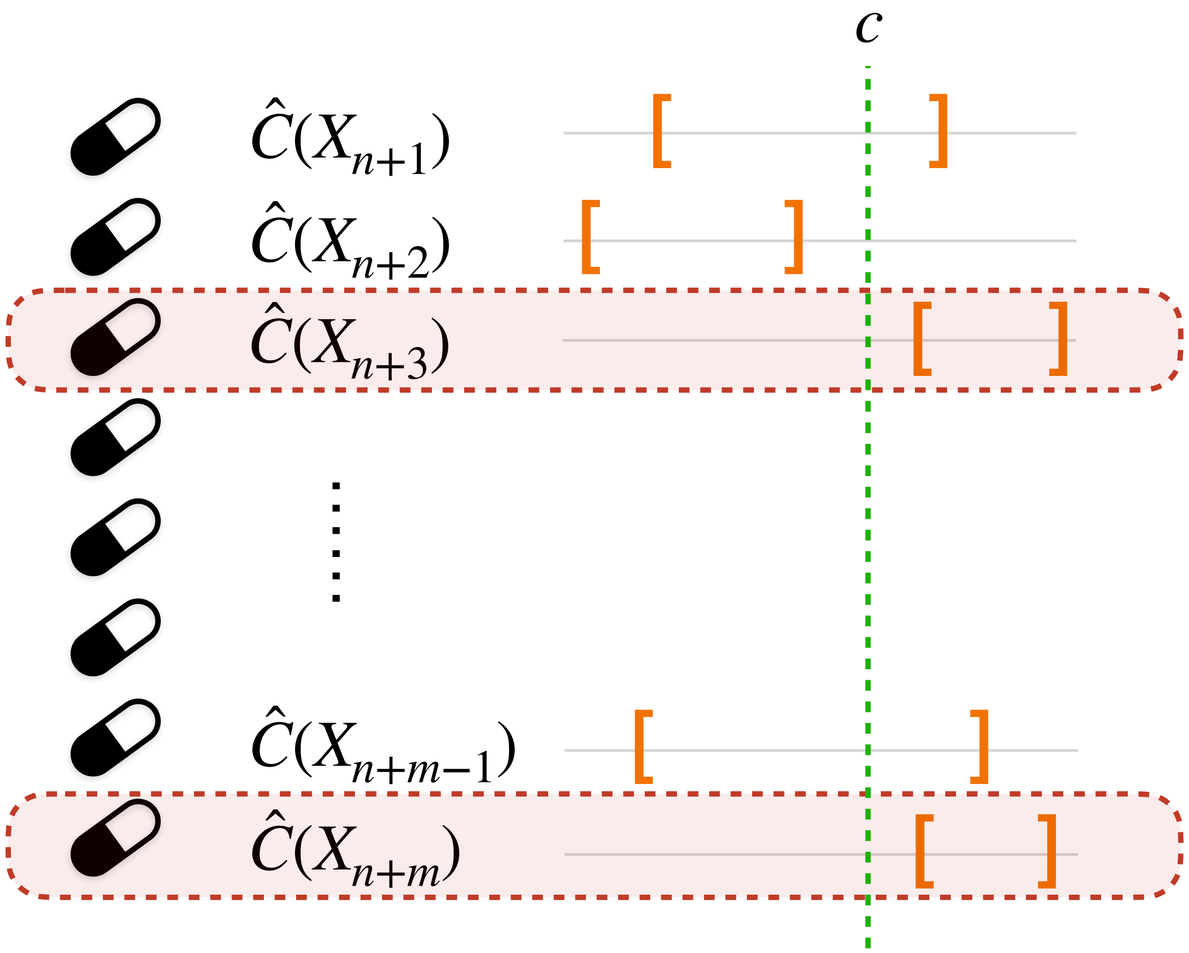


Validity of conformal prediction intervals (PIs) [Vovk et al., 1999]

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1})\right) \geq 95\,\%$$

⇝ Covers 95% of outcomes no matter prediction model

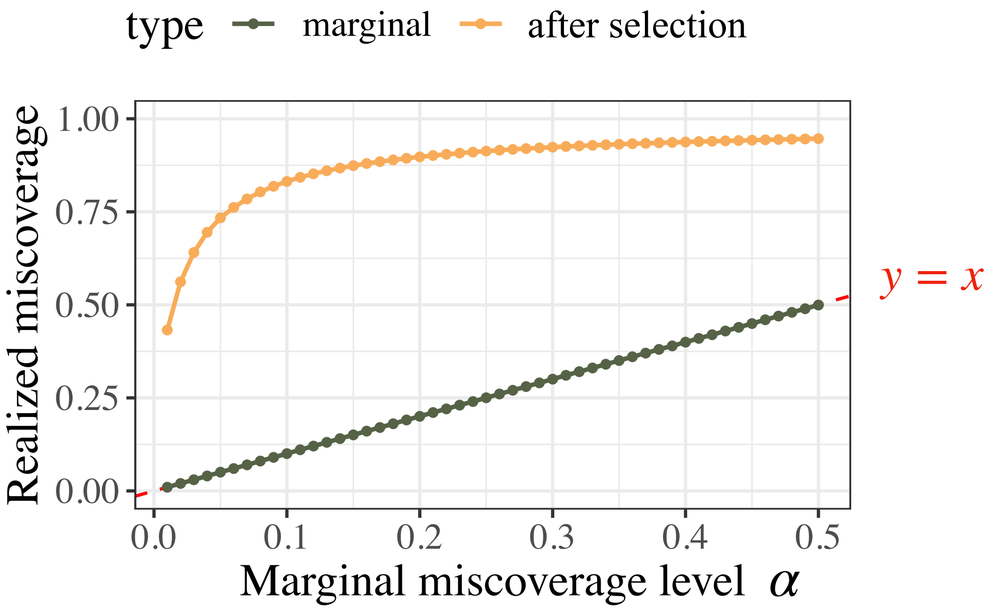
Challenges



- Uncertainty quantification ✓
- Model-free ✓
- Can we use them to find interesting instances (drugs)?

Which drugs are sufficiently active?

Miscoverage of naive selection



Dark: perfect marginal miscoverage

Orange: miscoverage of those $\hat{C}(X_{n+j}) > c_{n+j}$

Conformal prediction for drug discovery

[Norinder et al., 2014, Svensson et al., 2017, Wang et al., 2022]

1% nominal error, yet >30% error after selection!

This is **the winner's curse** [Soric, 1989]

Inspired a whole field of research: Selective Inference

[Benjamini and Yekutieli, 2005, Berk et al., 2013, Taylor et al., 2014, Fithian et al., 2014; Storey et al, 2003]

Our proposal: select with guarantees

- Find “actionable instances” while controlling fraction of false positive (FDR)

$$\text{FDR} = \mathbb{E}[\text{FDP}], \quad \text{FDP} = \frac{\#\{\text{false discoveries}\}}{\#\{\text{selected instances}\}}$$

Benjamini, Hochberg (1995)

- Control of FDR implies
  - Most AI-powered decisions are correct
  - Resource allocation is efficient

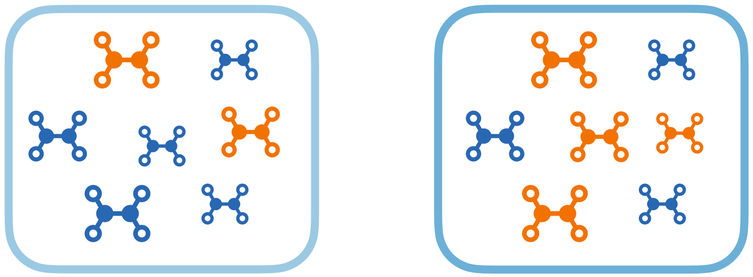
Drugs ⇝ 90% active
Customers ⇝ 90% responding
Patients ⇝ 90% benefiting
LLM outputs ⇝ 90% trustworthy

Controlling the **false discovery** rate: a practical and powerful approach to multiple testing

Authors Yoav Benjamini, Yosef Hochberg Total citations Cited by 113893

Distribution shift

- Are my evaluated drugs comparable to the unknown drugs?
  - **No** if you preferred drugs with some specific structures, etc



Training drugs New drugs

- In reality: distribution shift when generating/exploring new drugs
  - ⇝ Similar issues in job hiring, health monitoring, counterfactual inference…

Distribution shift

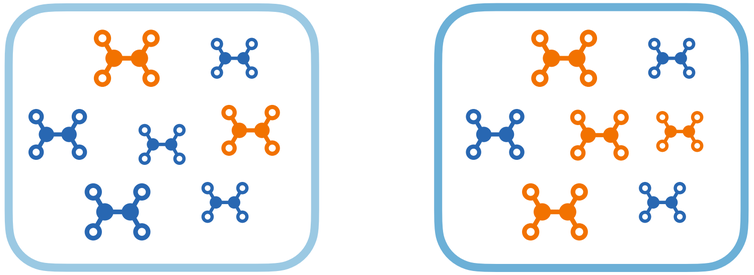
- Test data $\{(X_{n+j}, Y_{n+j})\} \sim \mathbb{Q}$ (unknown)
- Covariate shift: training data $\{(X_i, Y_i)\} \sim \mathbb{P}$ obeying

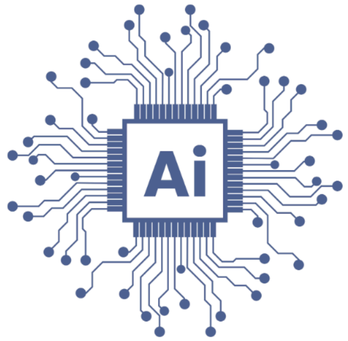
$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(x, y) = w(x)$$

for some (known or estimable) weight function $w : \mathcal{X} \to \mathbb{R}^+$ [Sugiyama et al., 2007, Tibshirani et al., 2019]

- **Why? Training data collected by looking at X (drugs, job applicants…)**

Entropy balancing for distribution shift adjustment





$\phi(\cdot)$: hidden embeddings from AI models

Finding “simplest” weights that balance key representations across batches [Hainmueller, 2012]

$$\text{maximize}_{\mathbf{w}} \quad \sum_{i=1}^{n} -w_i \log w_i$$

$$\text{subject to} \quad \left| \frac{1}{n} \sum_{i=1}^{n} w_i \phi(X_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(X_{n+j}) \right| \leq \delta$$

$$w_i \geq 0 \quad \frac{1}{n} \sum_{i=1}^{n} w_i = 1$$

Obtaining valid confidence measures

Weighted conformal p-values:

$\approx$ weighted rank of $\hat{V}_{n+j}$ among training scores $\{V_i\}_{i=1}^n$

$$p_j = \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j \cdot w(X_{n+j})}{\sum_{i=1}^n w(X_i) + w(X_{n+j})}, \quad U_j \sim \text{Unif}[0,1]$$

- Calibration and test scores: $V_i = Y_i - \hat{\mu}(X_i)$ and $\hat{V}_{n+j} = c_{n+j} - \hat{\mu}(X_{n+j})$
- Valid p-value in sense that $\mathbb{P}(p_j \leq \alpha, Y_{n+j} \leq c_{n+j}) \leq \alpha$ if $w(X)$ is known

Takeaway: small $p_j \Longrightarrow$ small $\hat{V}_{n+j} \Longrightarrow$ small $c_{n+j} - \hat{\mu}(X_{n+j}) \Longrightarrow Y_{n+j}$ is likely large and above threshold

Conformal selection: real applications

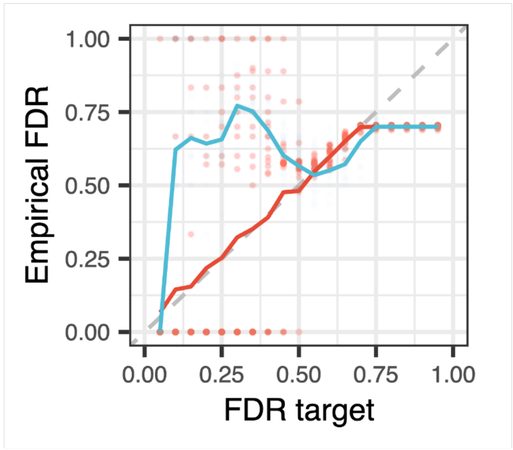
Selection Task (Distribution Shift)

Selection FDR Control

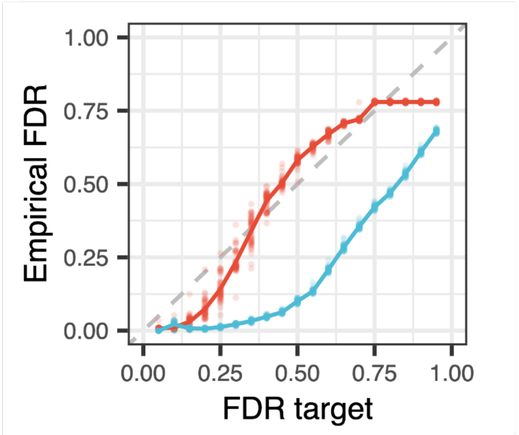
Conformal-Select

Baseline

Select gene perturbations with high T-cell proliferation (Uniform)



Select proteins with high stability (Mutant shift)



Covariates: learned representation in the hidden layer of neural nets

1: **Gene perturbation selection**

- Setup with no shift

2: **Protein stability selection**

- Significant shift from proteins in four rounds of experiments to single-mutation proteins

Jin, Candès (2023)

Conformal selection: real applications

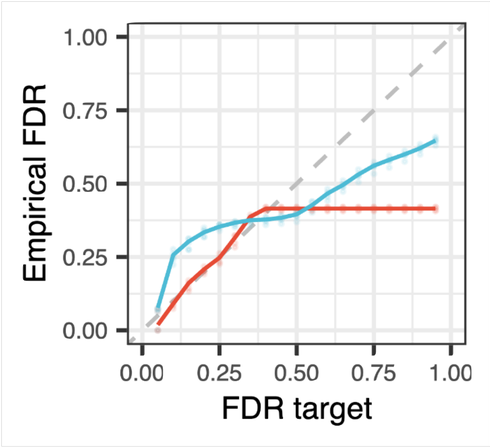
Selection Task (Distribution Shift)

Selection FDR Control

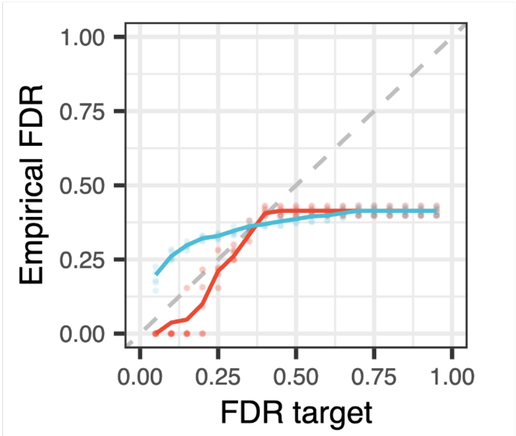
Conformal-Select

Baseline

Select compounds with low CYP2C9 inhibition rate (Scaffold shift)



Select clinical trials that meet primary outcome (Temporal shift)



Covariates: learned representation in the hidden layer of neural nets

3: **Drug property selection**

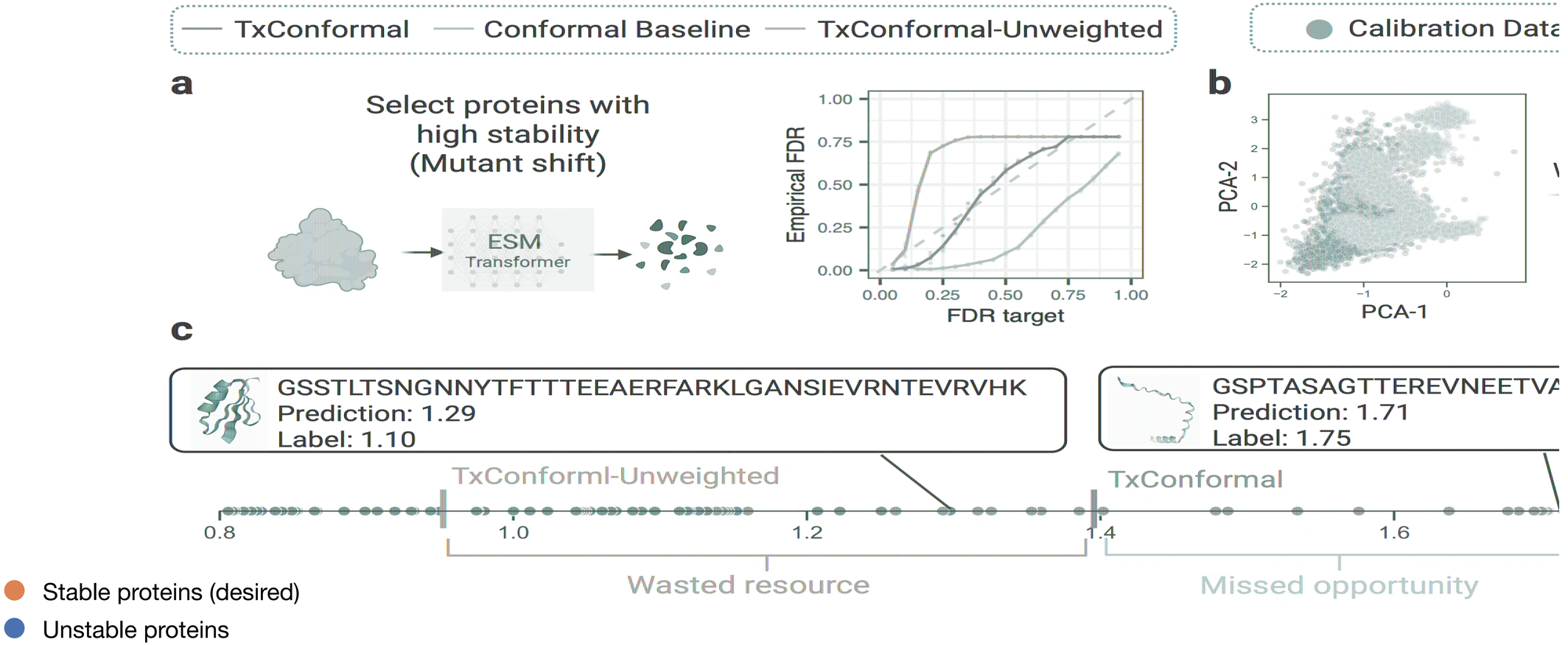
- Shift in drug structure (scaffold)

4: **Trial outcome prediction**

- Shift from earlier to future trials

Jin, Candès (2023)

Conformal selection: real applications



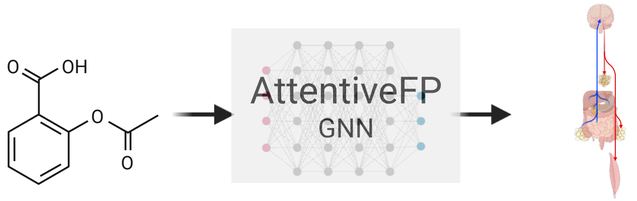
Adjusting for distribution shift yields accurate FDR control

Jin et al. (2025+)

Conformal selection: real applications

Controlling other metrics than FDR in selecting promising drug candidates

Select compounds with high CYP2D6 inhibition rate (Scaffold shift)



Question

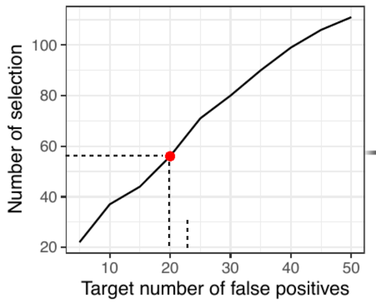
c Scenario 3

How many candidates can I test until I make 20 mistakes?

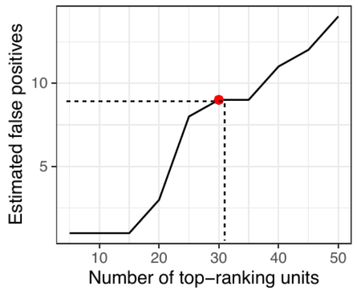
d Scenario 4

I want to screen my top 30 candidates. How many of them are expected to be wrong?

Result

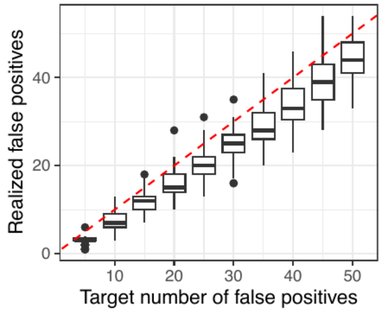


56 Selected

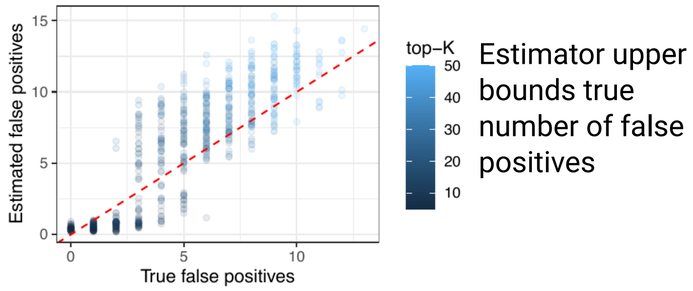


9 Wrong

Validity



Number of false positives below the target level



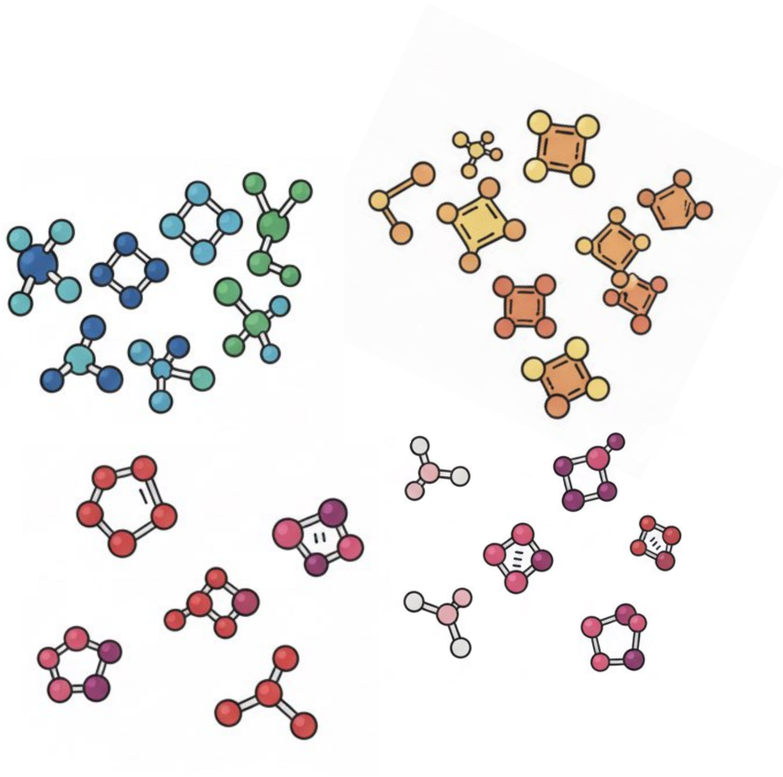
Estimator upper bounds true number of false positives

Jin et al. (2025+)

Diversifying conformal selections

How to identify structurally diverse drug candidates with FDR control?

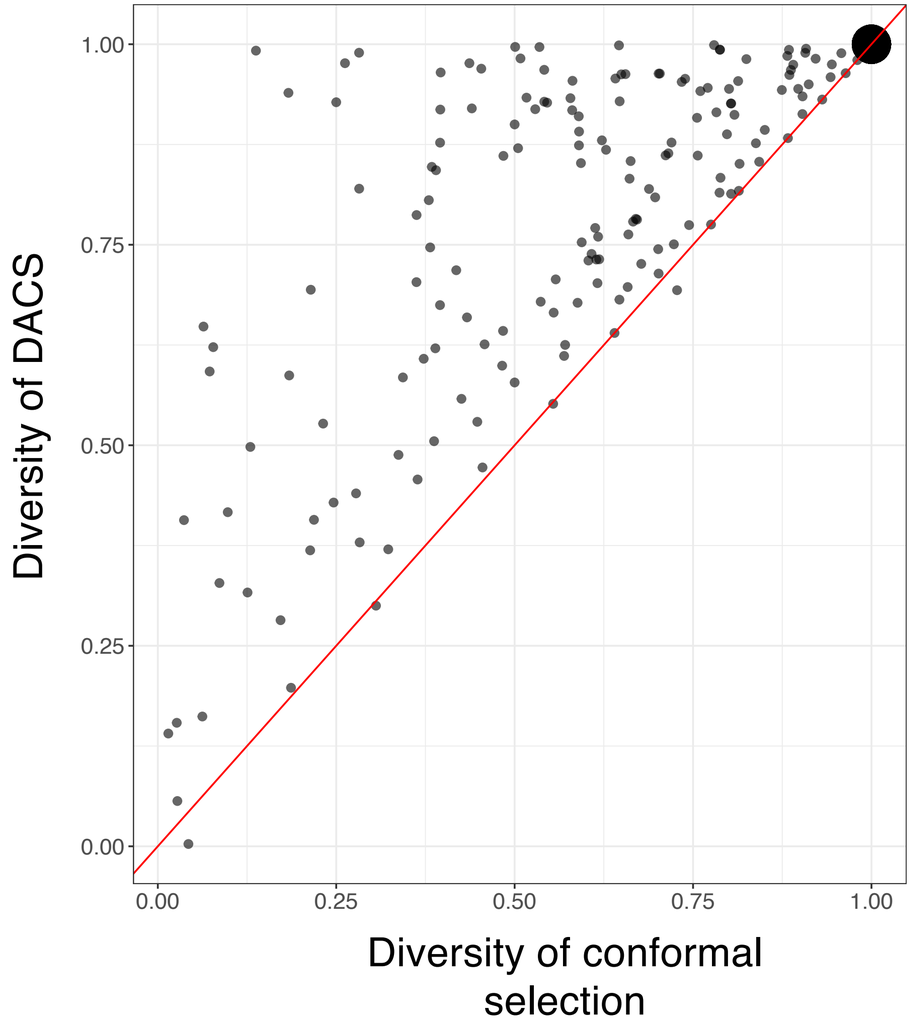
Diversity-aware conformal selection (DACS) Nair et al. (2025)



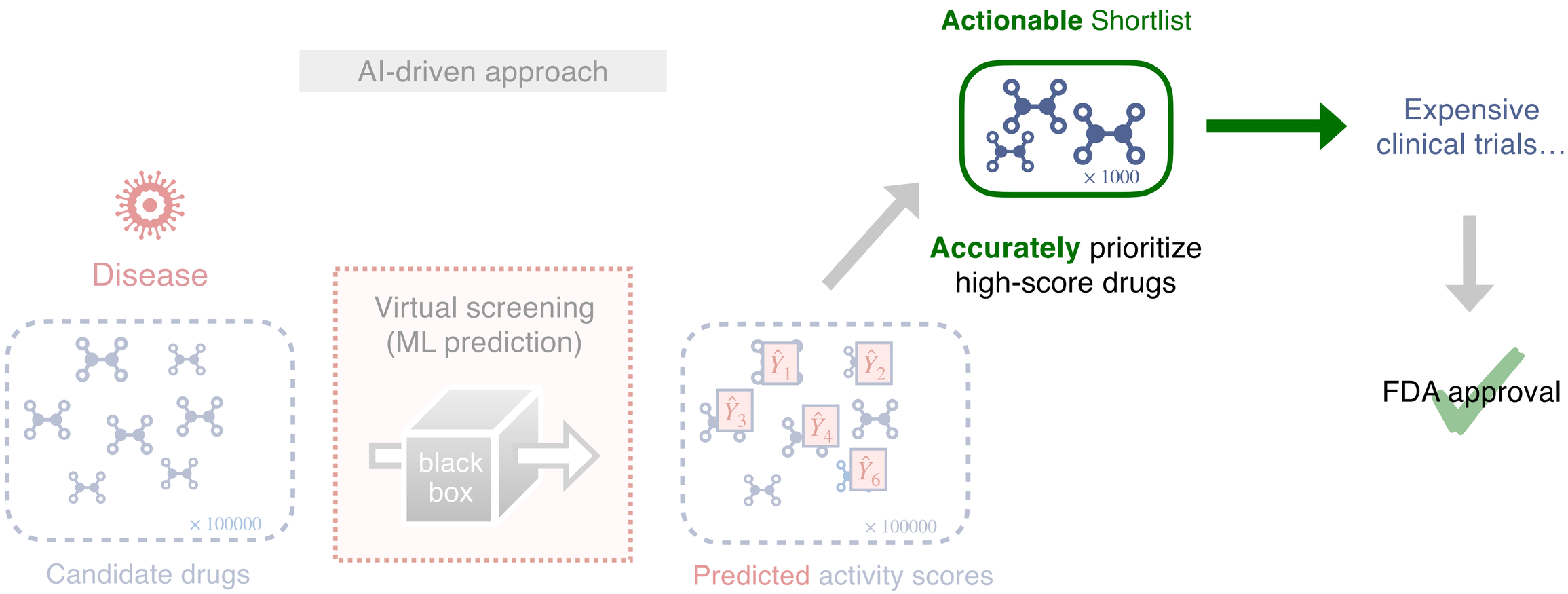
Diversifying conformal selection

FDR control
+ Diversity optimization

Diversity of selections on G-protein coupled receptor dataset



Quality control for drug discovery



Data collection

Data-driven discovery

Quality control

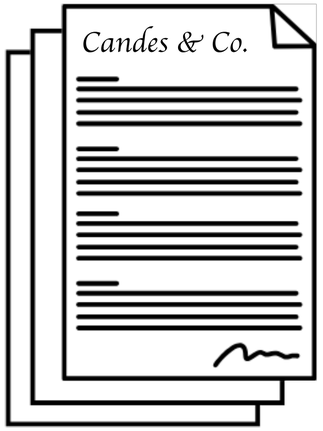
synthetic pretraining datamodels s1

AI-powered inference

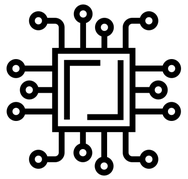
AI-powered drug discovery

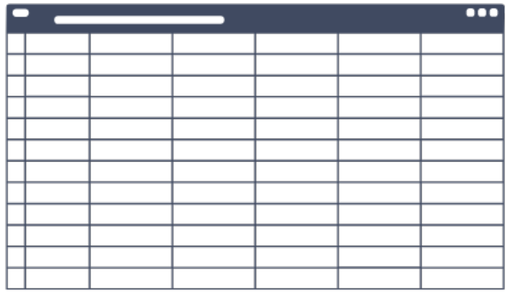
Factual accuracy calibration for LLMs

Example of production diagram in 2025



Legal documents





Compliance rules

Portfolio manager

Active investment strategies

How do we know that AI made the conversion from document to code without error?

Goal: hallucination removal with guarantees

How often is a shingles vaccine required?

The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.

Text filter?

The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. ~~It is currently recommended that individuals receive the shingles vaccine once in their lifetime.~~ However, it is always best to consult with a healthcare provider for personalized recommendations.

Unreliable

Guaranteed factuality

(with high probability)

Split conformal factuality

(Mohri & Hashimoto, 2024)

The shingles vaccine is typically recommended for adults aged 50 and over. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.

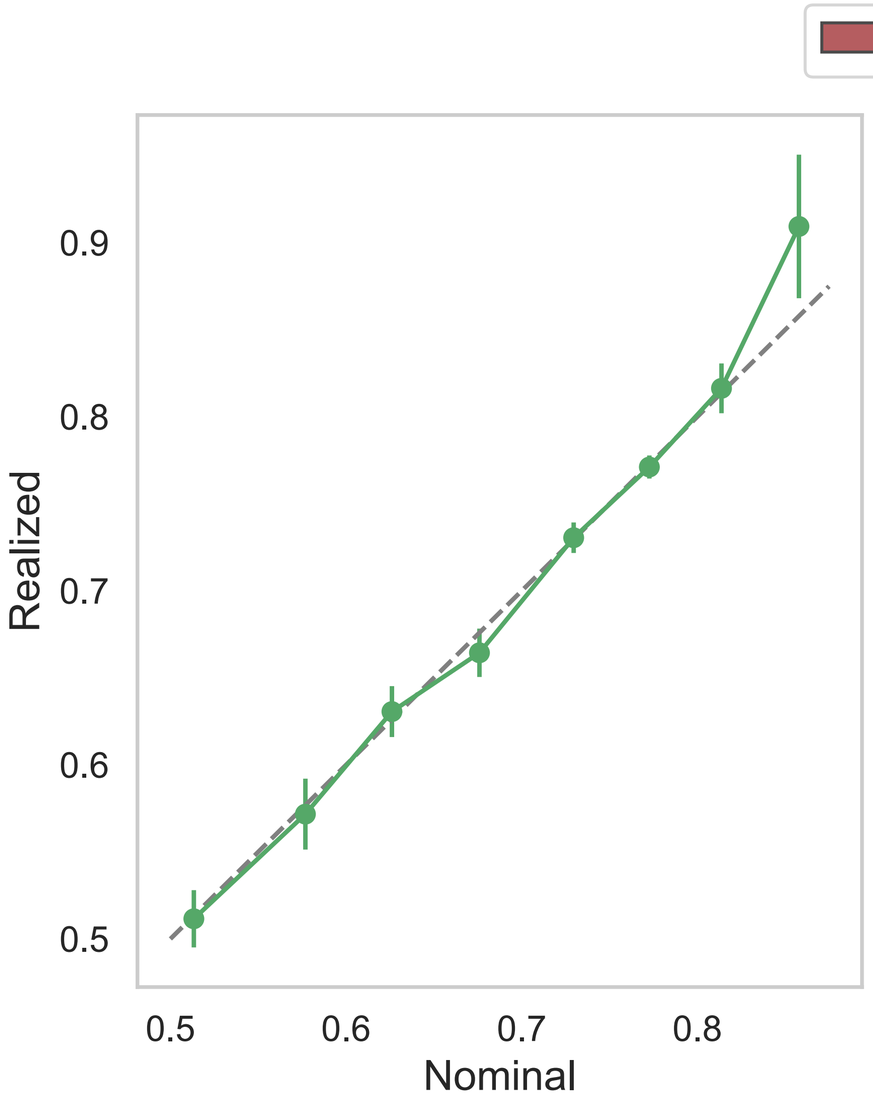
Our approach

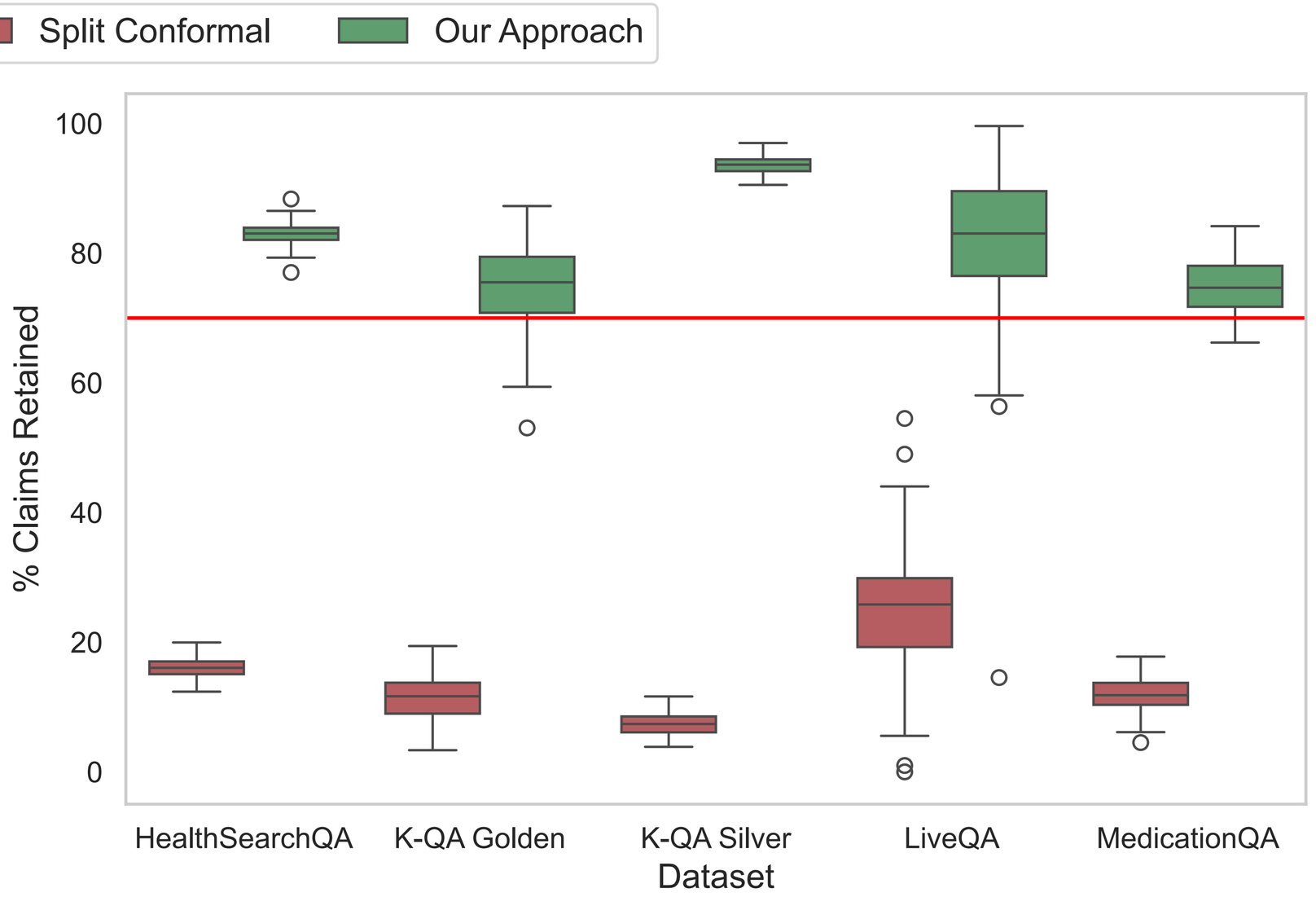
(Gibbs, Cherian, Candes, 2025)

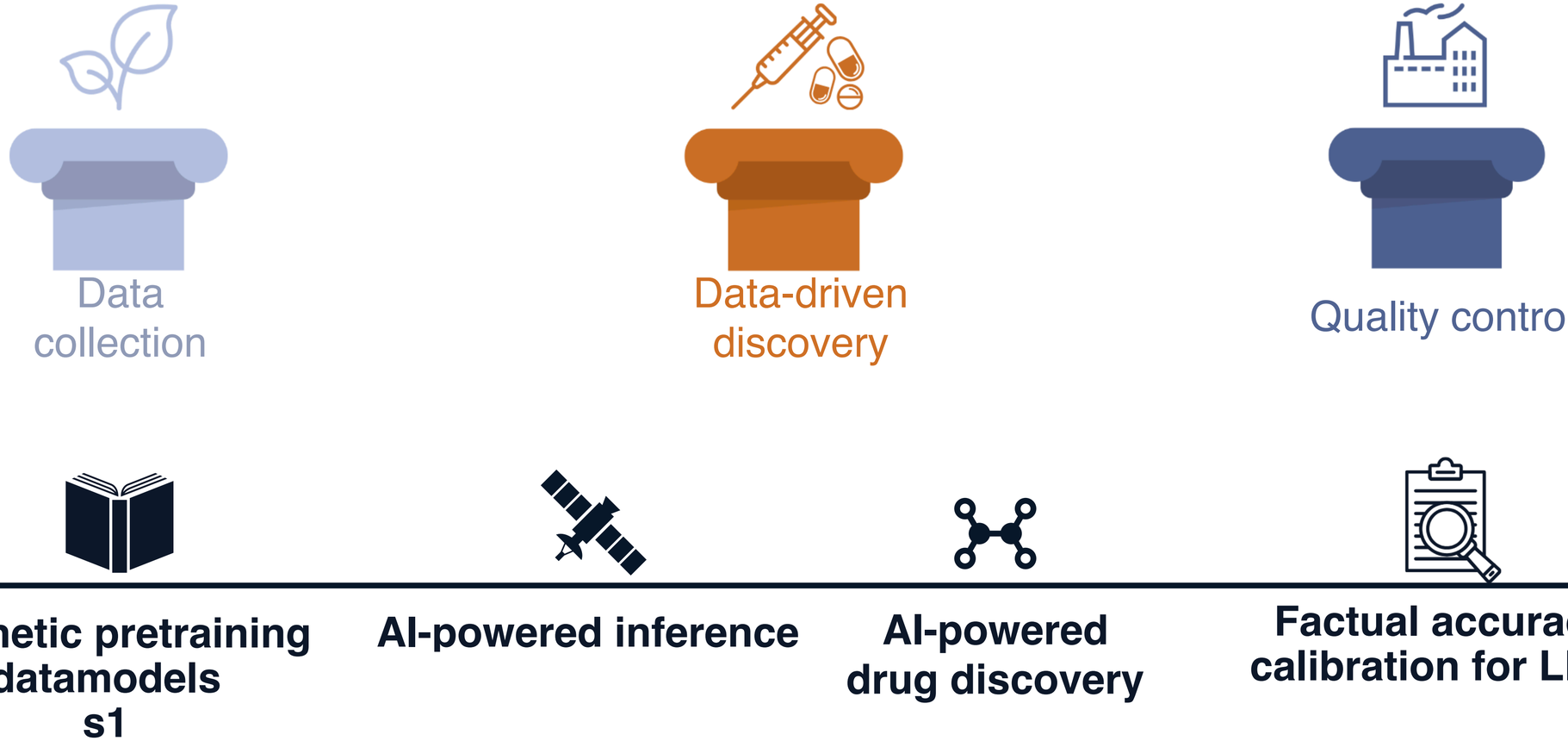
The shingles vaccine is typically recommended for adults aged 50 and over. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.

Prompt-dependent guarantee is calibrated and $\mathscr{F}$-conditionally valid

Calibrated validity



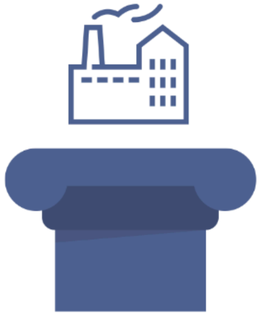




AI and statistics offer a lot to each other

Data collection

Data-driven discovery



Quality control

Open-source LLMs
Mathematical reasoning
Efficient training
Synthetizing data

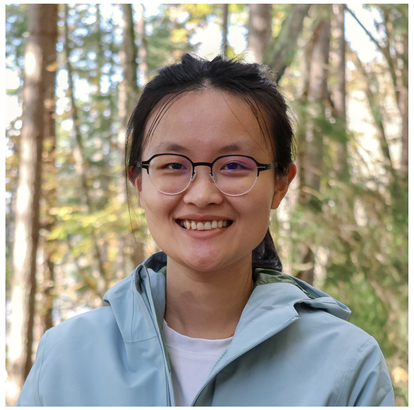
Deforestation
Media bias
Global warming
Protein structures

Drug discovery
Clinical trials

Filtering hallucinations
Factual calibration of LLMs
Automating legal compliance
Managing misinformation

Thinking statistically about AI inputs and outputs yields **more powerful**, **safer** AI

Close Collaborators



Ying Jin

Isaac Gibbs

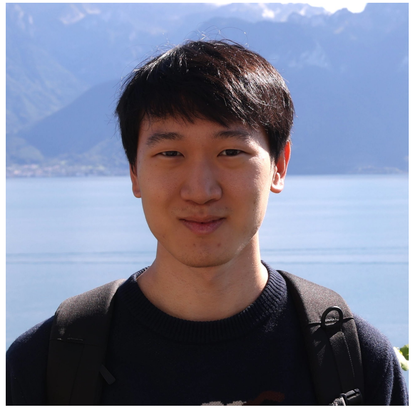
John Cherian

Zitong Yang

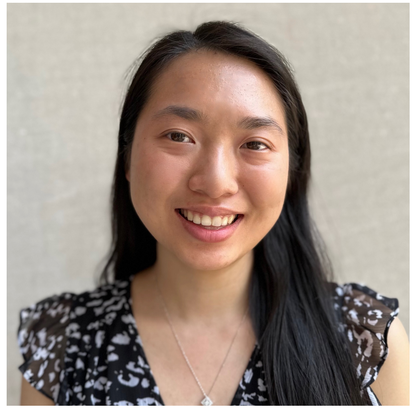
Tijana Zrnic

Asher Spector

Yash Nair



Joon Lee



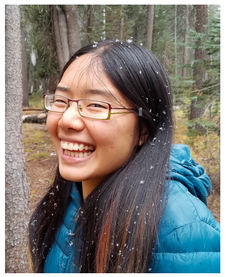
Ginnie Ma

Sarah Zhao

Other Collaborators


Anastasios Angelopoulos

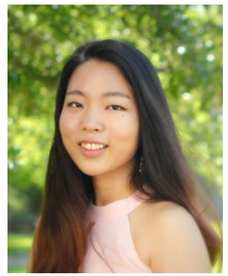

Stephen Bates

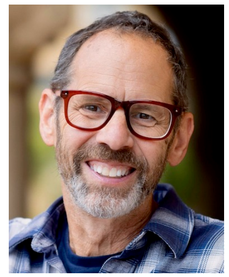

Clara Fannjiang


Michael Jordan

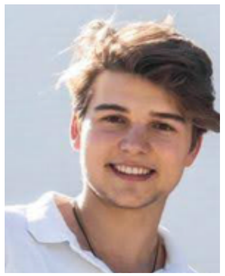

John Duchi

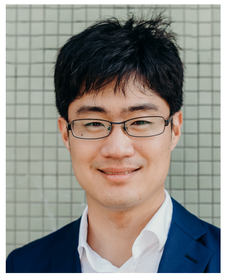

Kristina Gligorić


Cinoo Lee

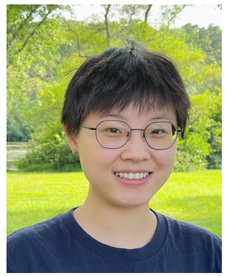

Dan Jurafsky


Neil Band


Niklas Muennighoff

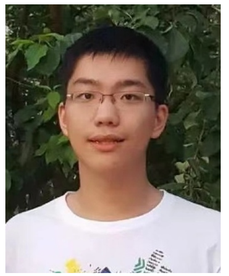

Tatsunori Hashimoto


Percy Liang


Shuangping Li


Christopher Mohri


Aonan Zhang


Hong Liu


Chong Wang


Ruoming Pang


Andrew Ilyas


Sung Min (Sam) Park

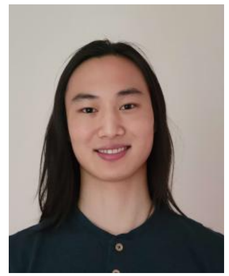

Logan Engstrom

Kristian Georgiev


Guillaume Leclerc


Aleksander Madry

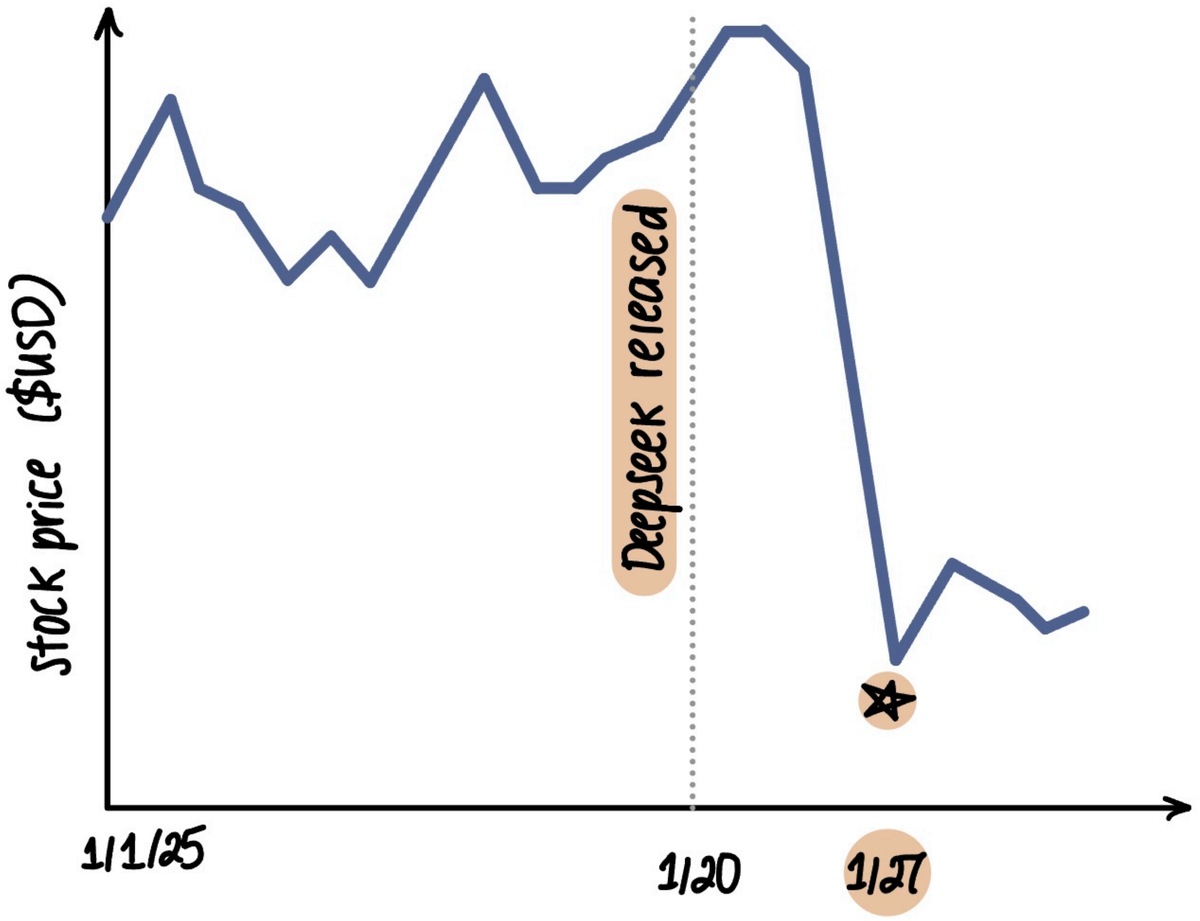
Axel Feldmann


Benjamin Chen

Last week of January

OpenAI

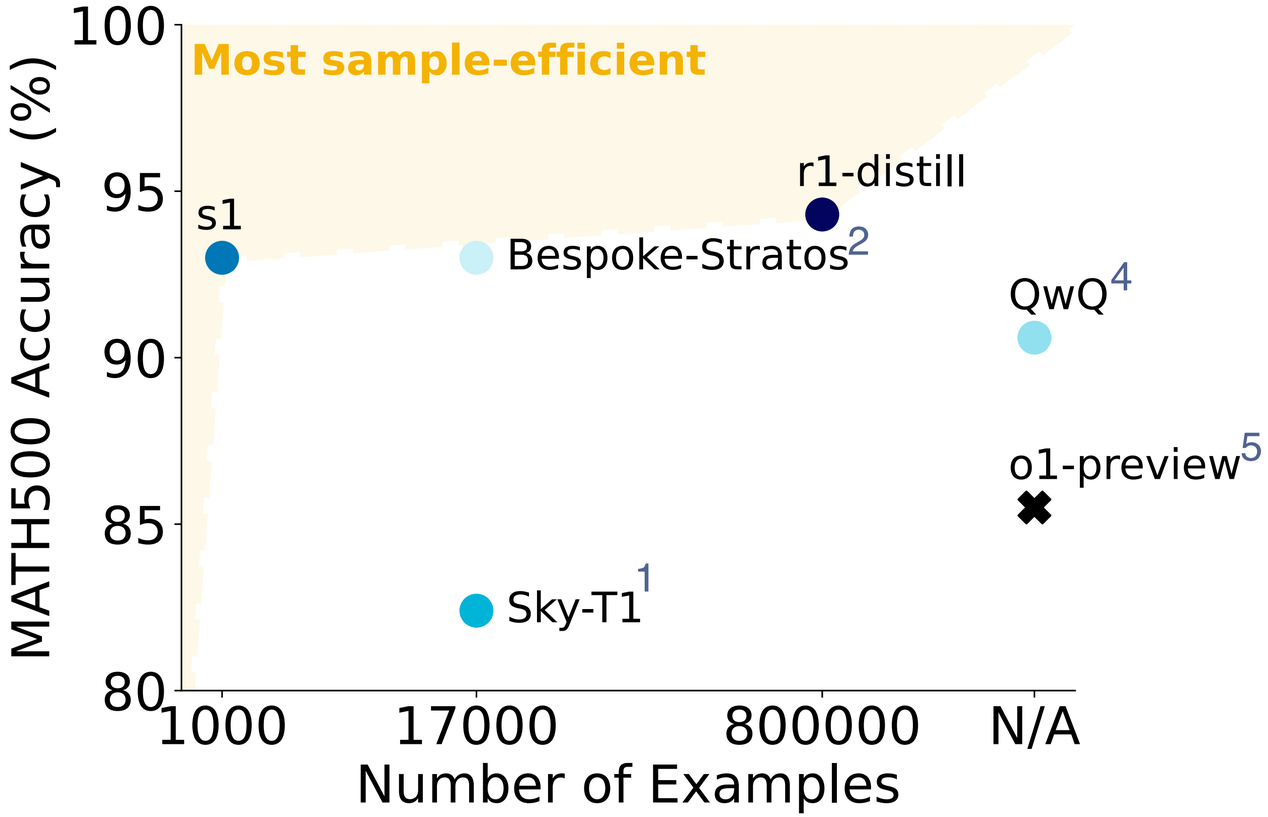
deepseek



Stanford s1 (first week of February)

s1: Simple test-time scaling

Niklas Muennighoff [*1 3 4] Zitong Yang [*1] Weijia Shi [*2] Xiang Lisa Li [*1] Li Fei-Fei [1] Hannaneh Hajishirzi [2 3]
Luke Zettlemoyer [2] Percy Liang [1] Emmanuel Candès [1] Tatsunori Hashimoto [1]



1. Team (2025)
2. Labs (2025)
3. DeepSeek-AI et al. (2025)
4. Qwen et al. (2024)
5. OpenAI (2024)

‘Science’ of LLMs

- Demonstrate **test-time scaling**
- High **performance** on small training data sets (S1K)
- **Open** source/weights/data/ideas/everything