A few simple thoughts on complex problems,

and on learning to solve them

FIM-IMS Joint Workshop, November 20, 2026

Geelon So | ETH Zürich

I. Common Knowledge

II. Some Classical Ideas

III. Structure in Complex Tasks

I. **Common Knowledge**

II. Some Classical Ideas

III. Structure in Complex Tasks

Classical Learning Framework



DATA MODEL TASK

Classical Learning Framework



Let’s say that there is a *quantifiable task* we would like to solve.

Classical Learning Framework



We fix a set of *decision parameters* that constitutes our solution space.

Classical Learning Framework



We *carefully collect* data to estimate the best solution.

Classical Learning Framework

or, risk minimization

$$\min_{\theta \in \Theta} R(\theta)$$

Thus, learning is cast as an *optimization* problem.

Classical Learning Framework

or, risk minimization / utility maximization

$$\min_{\theta \in \Theta} R(\theta)$$

Thus, learning is cast as an *optimization* problem.

Classical Learning Framework

or, risk minimization / utility maximization / reward optimization

$$\min_{\theta \in \Theta} R(\theta)$$

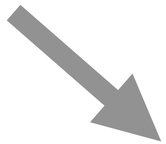
Thus, learning is cast as an *optimization* problem.

Classical Learning Framework

or, risk minimization / utility maximization / reward optimization

$$\min_{\theta \in \Theta} R(\theta)$$

Genius idea



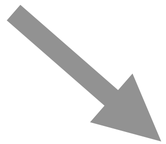
Thus, learning is cast as an *optimization* problem.

Classical Learning Framework

or, risk minimization / utility maximization / reward optimization

$$\min_{\theta \in \Theta} R(\theta)$$

Genius idea



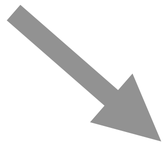
Summarize the performance of an entire ML system into a single number!

Classical Learning Framework

or, risk minimization / utility maximization / reward optimization

$$\min_{\theta \in \Theta} R(\theta)$$

Genius idea, but also insane



Summarize the performance of an entire ML system into a single number!



Boss, I trained a policy for a self-driving car that attains 99% accuracy.

Let’s deploy the car now!





Boss, I trained a policy for a self-driving car that attains 99% accuracy.

Let's deploy the car now!

What does that even mean?

What happens the other 1% of the time? Is it safe? Does it follow traffic laws? Is it fuel-efficient? Is there latency? Is it comfortable? How does it affect traffic…



Modern Learning Problems

- Things are a lot more complicated.



DATA **MODEL** **TASK**

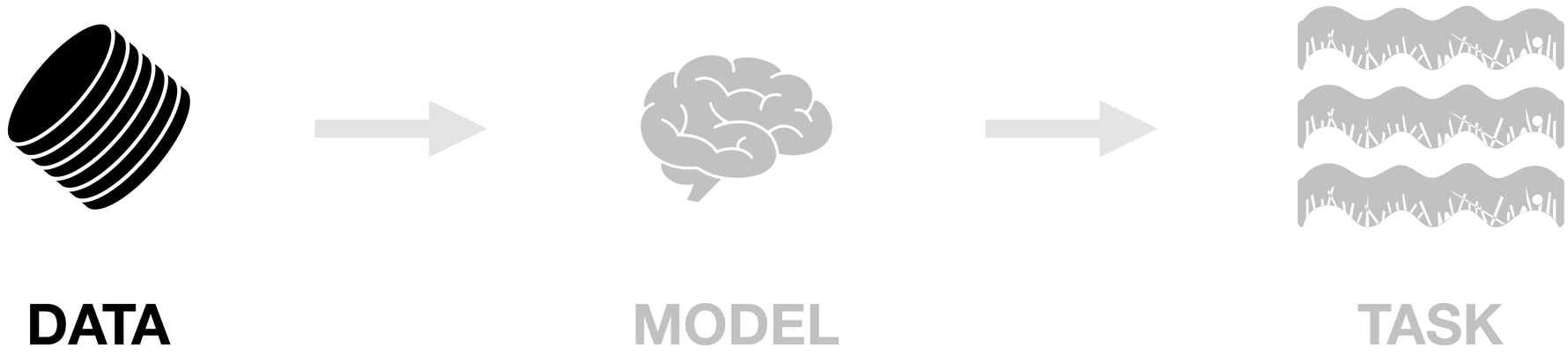


DATA MODEL **TASK**

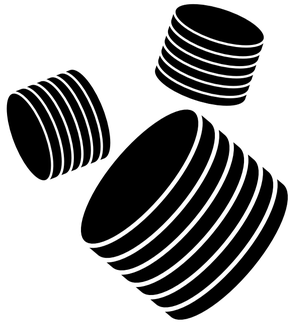
The task is often not well-defined [hard to measure success].



The task is often not well-defined and multi-dimensional.

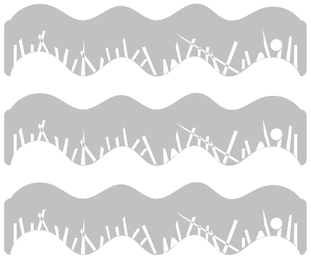


Data that we can collect may be a poor proxy to what we want to know.



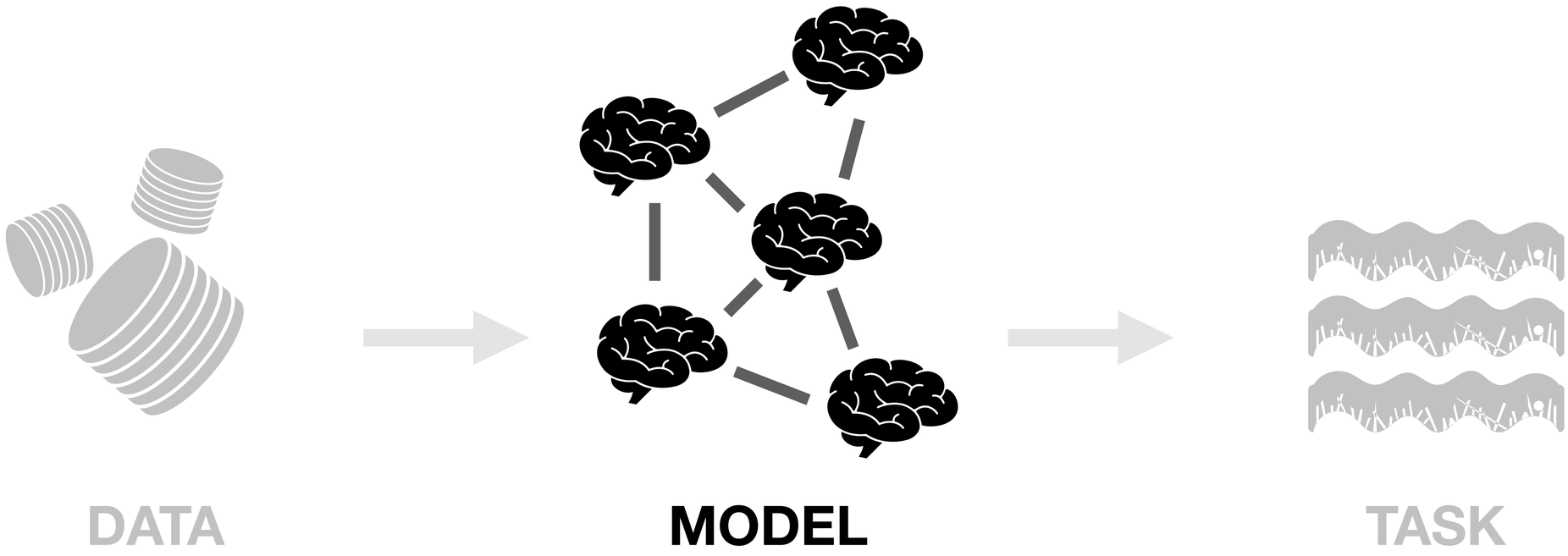
DATA

MODEL

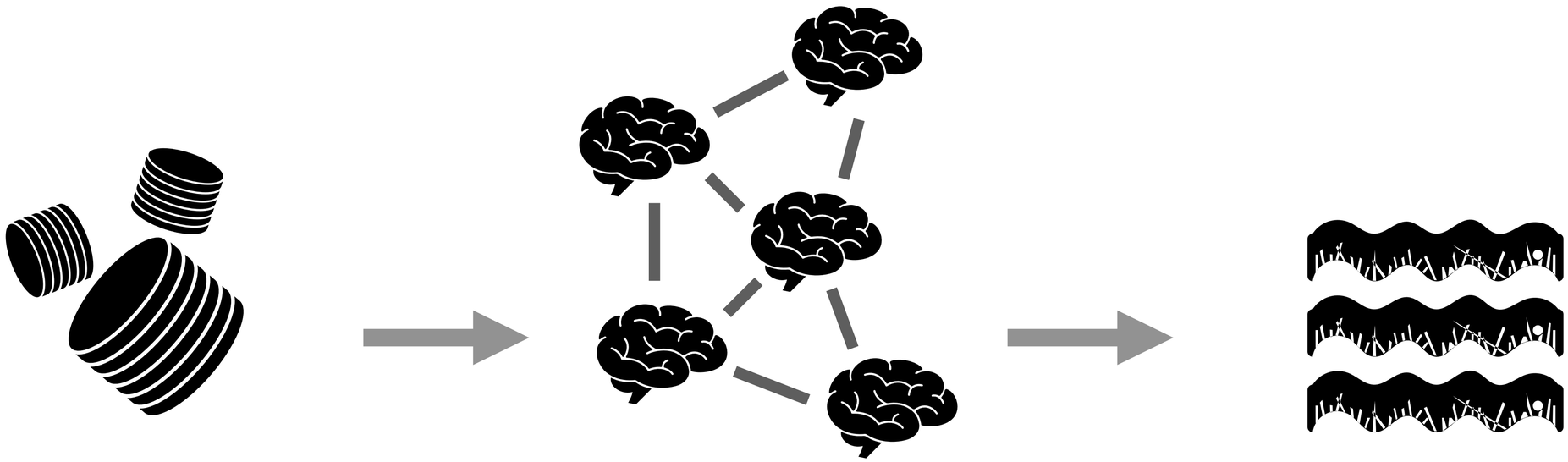


TASK

Data is heterogeneous: varying cost, quality, quantity, provenance, etc.

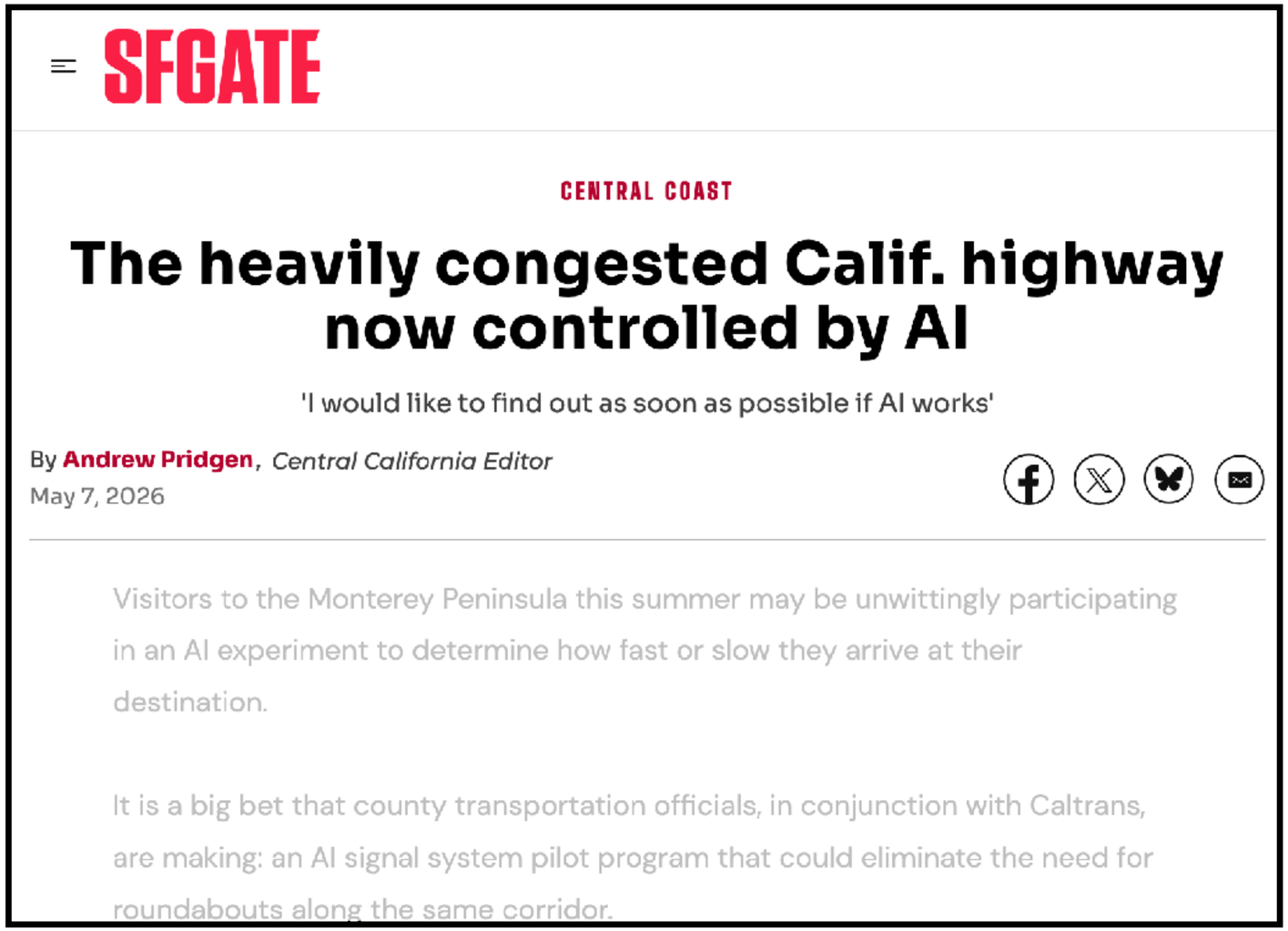


The decision parameters may need to be divided and decentralized.



Modern ML problems are very complex.

Complex Tasks



SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

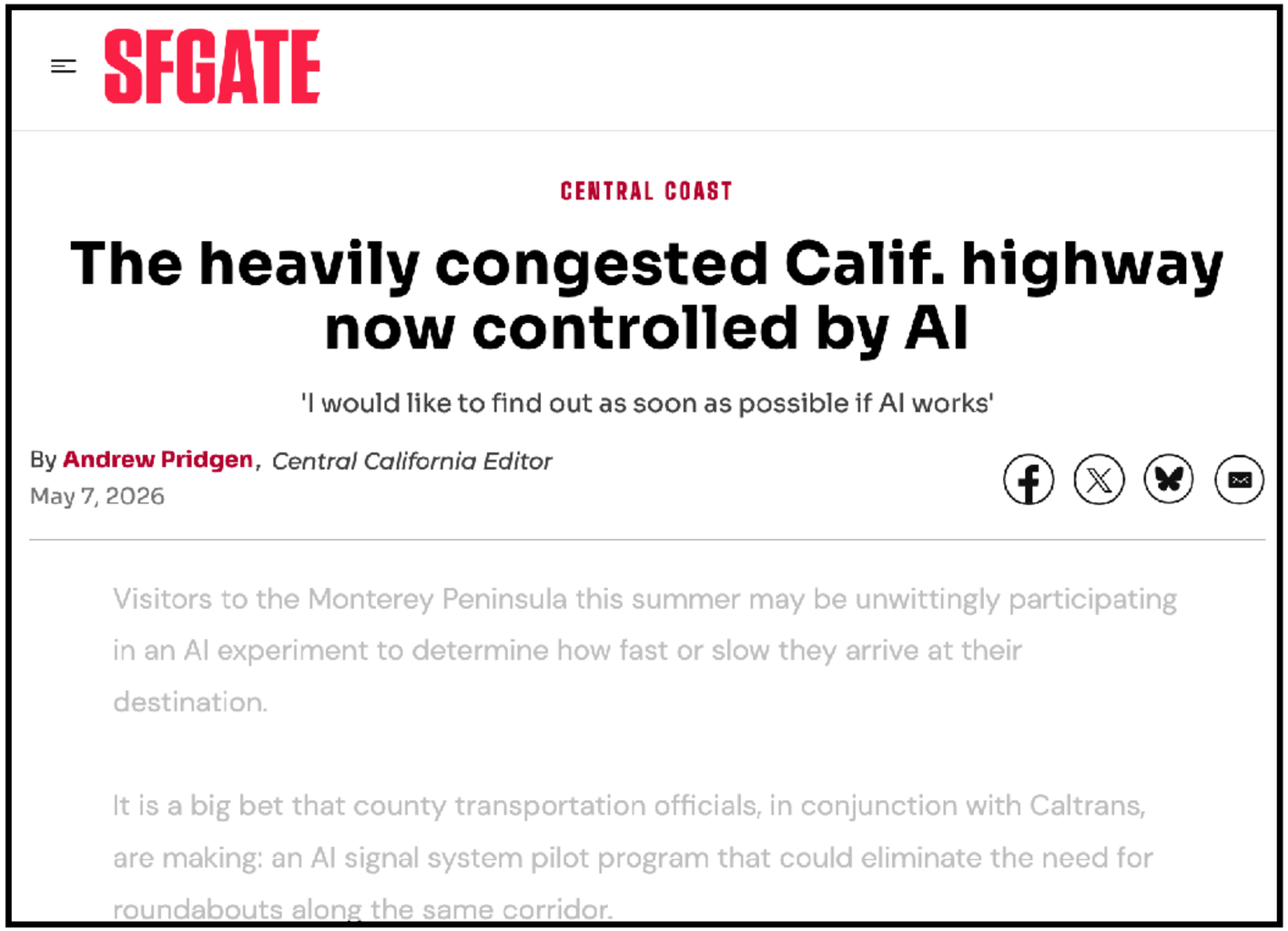
'I would like to find out as soon as possible if AI works'

By **Andrew Pridgen**, *Central California Editor*
May 7, 2026

Visitors to the Monterey Peninsula this summer may be unwittingly participating in an AI experiment to determine how fast or slow they arrive at their destination.

It is a big bet that county transportation officials, in conjunction with Caltrans, are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

Complex Tasks



SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

'I would like to find out as soon as possible if AI works'

By **Andrew Pridgen**, *Central California Editor*
May 7, 2026

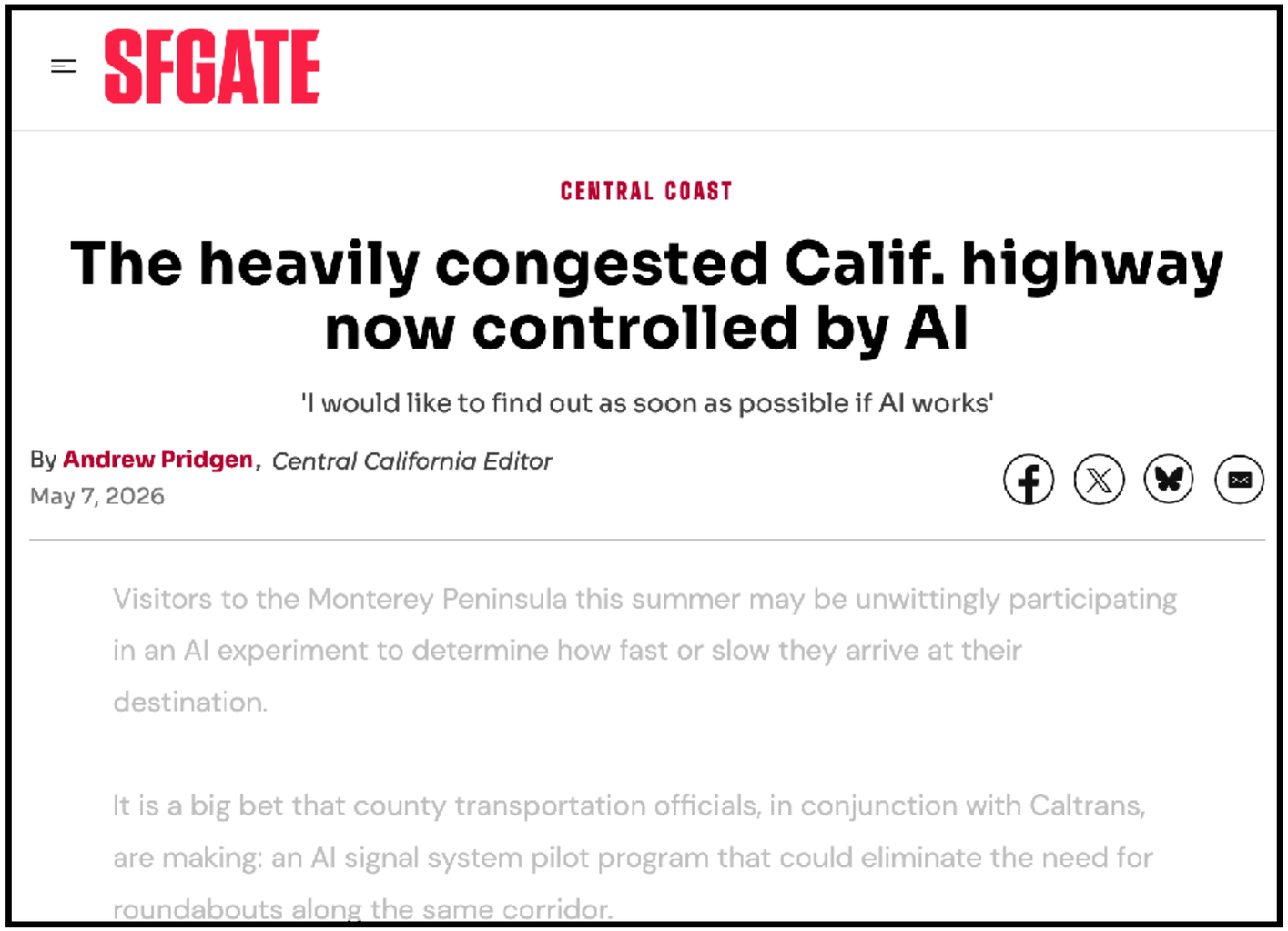
Visitors to the Monterey Peninsula this summer may be unwittingly participating in an AI experiment to determine how fast or slow they arrive at their destination.

It is a big bet that county transportation officials, in conjunction with Caltrans, are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

What are the goals?

- Reduce travel time
- Increase fuel efficiency
- Reduce wait time for pedestrians
- Minimize accidents
- Make whoever votes happier

Complex Tasks



SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

'I would like to find out as soon as possible if AI works'

By **Andrew Pridgen**, *Central California Editor*
May 7, 2026

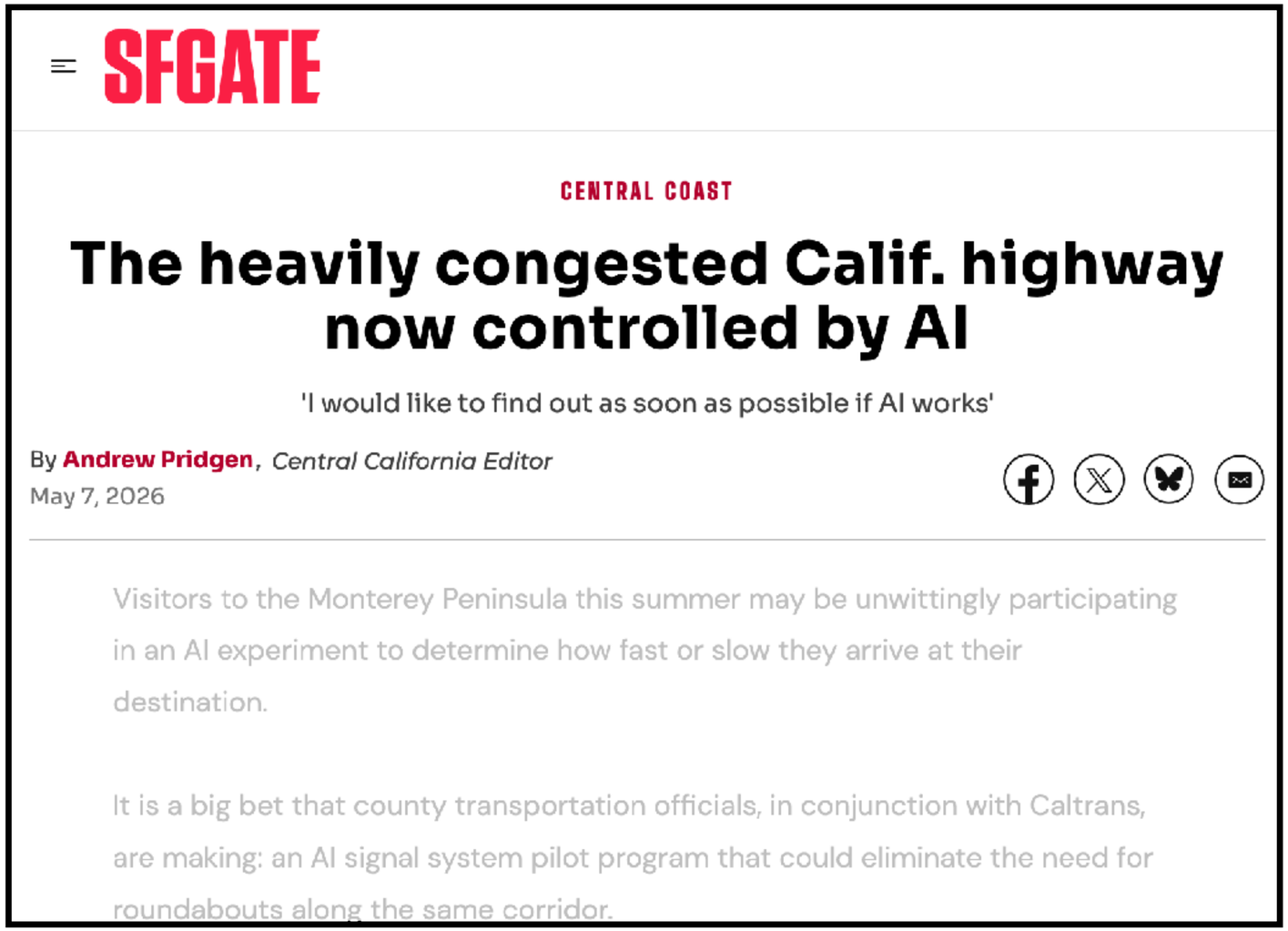
Visitors to the Monterey Peninsula this summer may be unwittingly participating in an AI experiment to determine how fast or slow they arrive at their destination.

It is a big bet that county transportation officials, in conjunction with Caltrans, are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

Is there a division of labor?

- Is there a central coordinator? How much communication / computation is needed?
- Is each signal optimizing only the local network?
- Is local information enough?

Complex Tasks



SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

'I would like to find out as soon as possible if AI works'

By **Andrew Pridgen**, *Central California Editor*
May 7, 2026

Visitors to the Monterey Peninsula this summer may be unwittingly participating in an AI experiment to determine how fast or slow they arrive at their destination.

It is a big bet that county transportation officials, in conjunction with Caltrans, are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

What data is provided?

- Traffic sensors / cameras
- Hospital records
- Survey data

Complex Tasks

SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

By Andrew

May 7, 202

Vis

in a

de

It is

are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

Update after 3 months: Traffic on side streets increased

	1	2	3	4	5	6	7	8	9
	Josselyn Canyon Rd	Olmsted Rd	Canyon Del Rey Blvd	Ragsdale Dr	York Rd	Pasadera Dr/Boots Rd	Laureles Grade	Corral De Tierra Rd	San Benancio Rd
SIDE STREET Average Delay Change	**NA**	**+19%** ↑	**+10%** ↑	**+25%** ↑	**+52%** ↑	**-20%** ↓	**+33%** ↑	**+55%** ↑	**+23%** ↑

Complex Tasks

SFGATE

CENTRAL COAST

The heavily congested Calif. highway now controlled by AI

'I would like to find out as soon as possible if AI works'

By **Andrew Pridgen**, *Central California Editor*
May 7, 2026

Visitors to the Monterey Peninsula this summer may be unwittingly participating in an AI experiment to determine how fast or slow they arrive at their destination.

It is a big bet that county transportation officials, in conjunction with Caltrans, are making: an AI signal system pilot program that could eliminate the need for roundabouts along the same corridor.

vs.



Laboratory Statistics v. Wild ML

Much of our statistical language arose within controlled laboratory settings.

Laboratory Statistics v. Wild ML

Much of our statistical language arose within controlled laboratory settings.

- We’re now asking ML systems to solve problems that are much less well-specified and much more “multi-dimensional”.

Laboratory Statistics v. Wild ML

Much of our statistical language arose within controlled laboratory settings.

- We're now asking ML systems to solve problems that are much less well-specified and much more "multi-dimensional".
- Example: is "reward hacking" a symptom of our limited ability to know what we want / how to operationalize it?

Laboratory Statistics v. Wild ML

Much of our statistical language arose within controlled laboratory settings.

- We're now asking ML systems to solve problems that are much less well-specified and much more "multi-dimensional".
- Example: is "reward hacking" a symptom of our limited ability to know what we want / how to operationalize it?
- How do we give these abstract problems more structure?

Laboratory Statistics v. Wild ML

Much of our statistical language arose within controlled laboratory settings.

- We're now asking ML systems to solve problems that are much less well-specified and much more "multi-dimensional".
- Example: is "reward hacking" a symptom of our limited ability to know what we want / how to operationalize it?
- How do we give these abstract problems more structure?

This talk: some new [old?] language from multi-objective optimization and economics.

I. Common Knowledge

II. Some Classical Ideas

III. Structure in Complex Tasks

Beyond One Objective

Setting	**Single-Objective**	**Multi-Objective**	**Game / Strategic**
# Objectives	One	Many	Many
# Decision Makers	One	One	Many

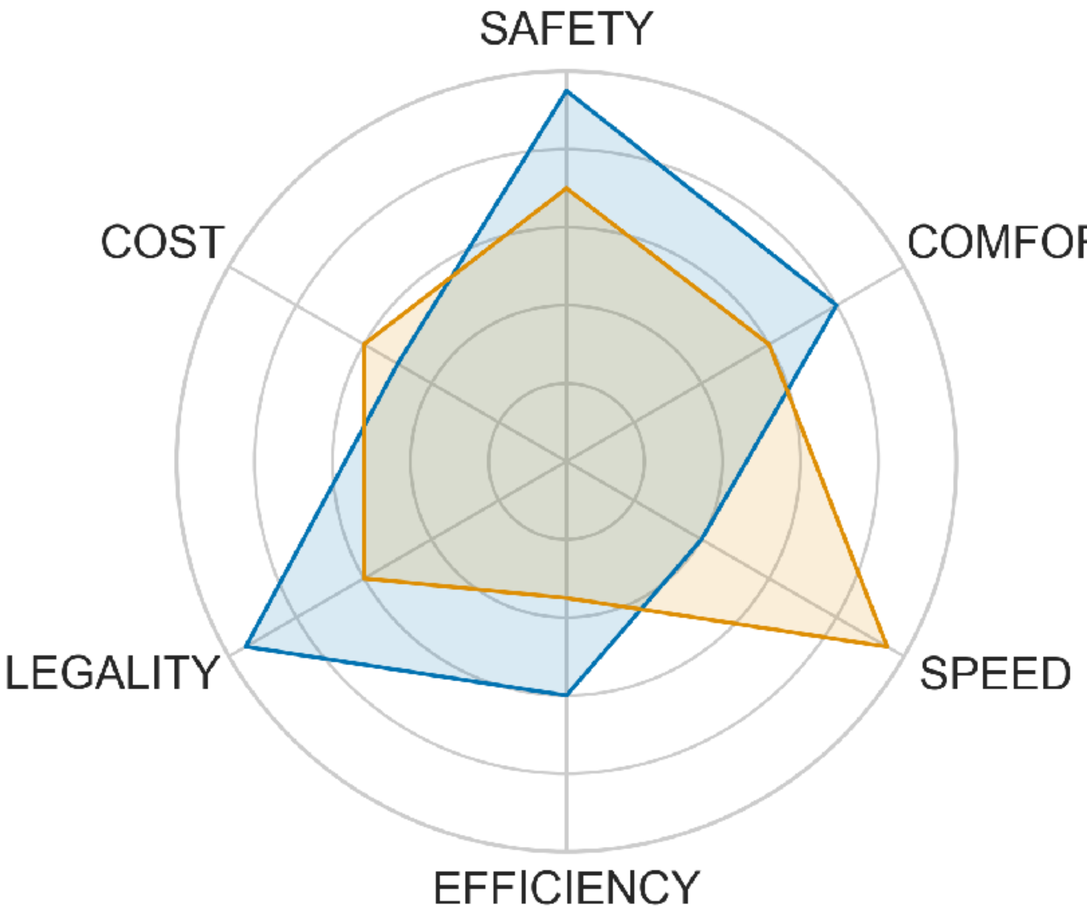
Multi-Objective Setting

$$\min_{\theta \in \Theta} \underbrace{\mathbf{R}(\theta) \equiv \Big(R_1(\theta), \ldots, R_K(\theta)\Big)}$$

We care about K objectives at once.

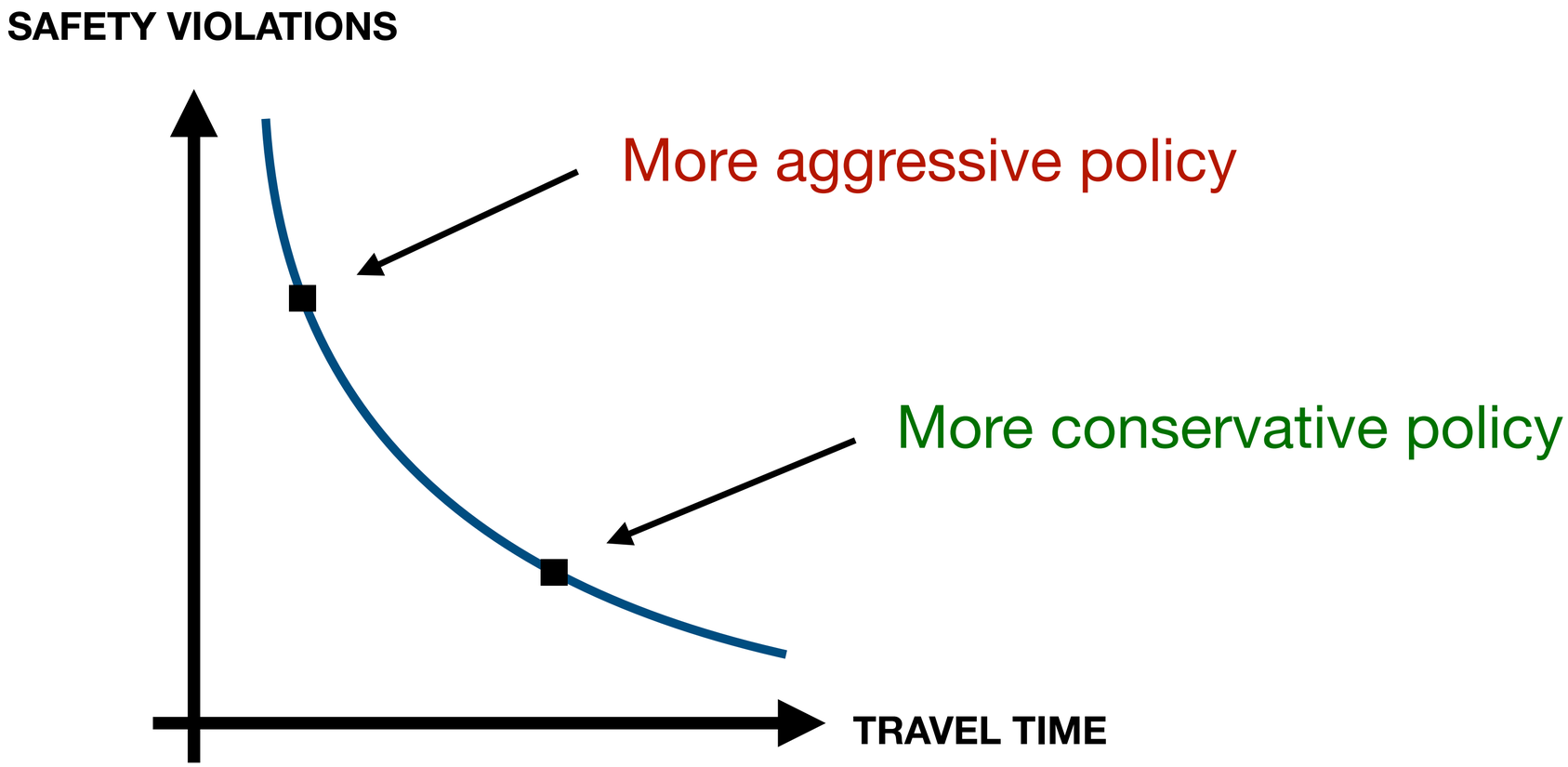
Question: what does it even mean to minimize $\mathbf{R}(\theta)$?

Solution Concept: Pareto Optimality

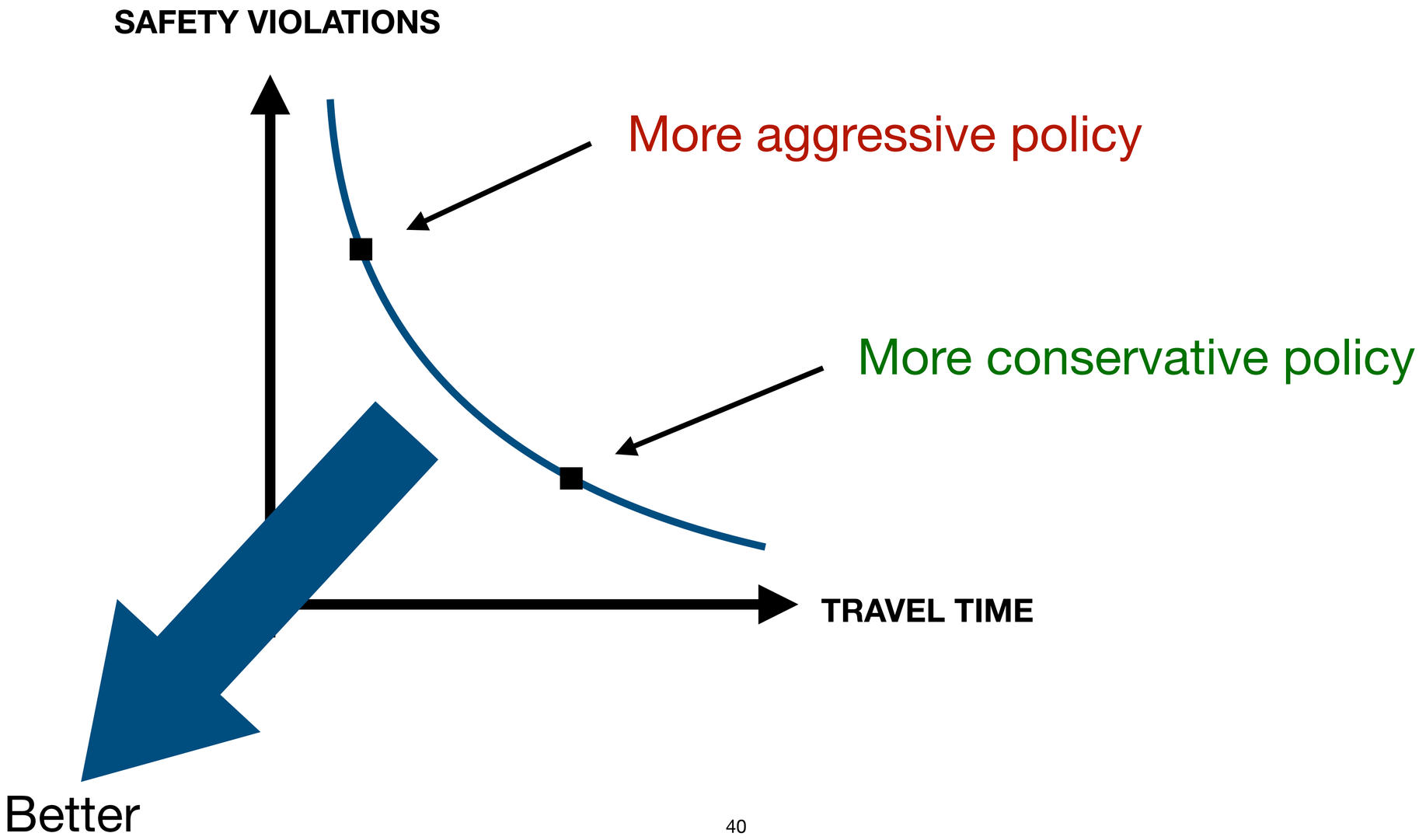


A decision is **Pareto optimal** if improving one objective must come at a cost of another.

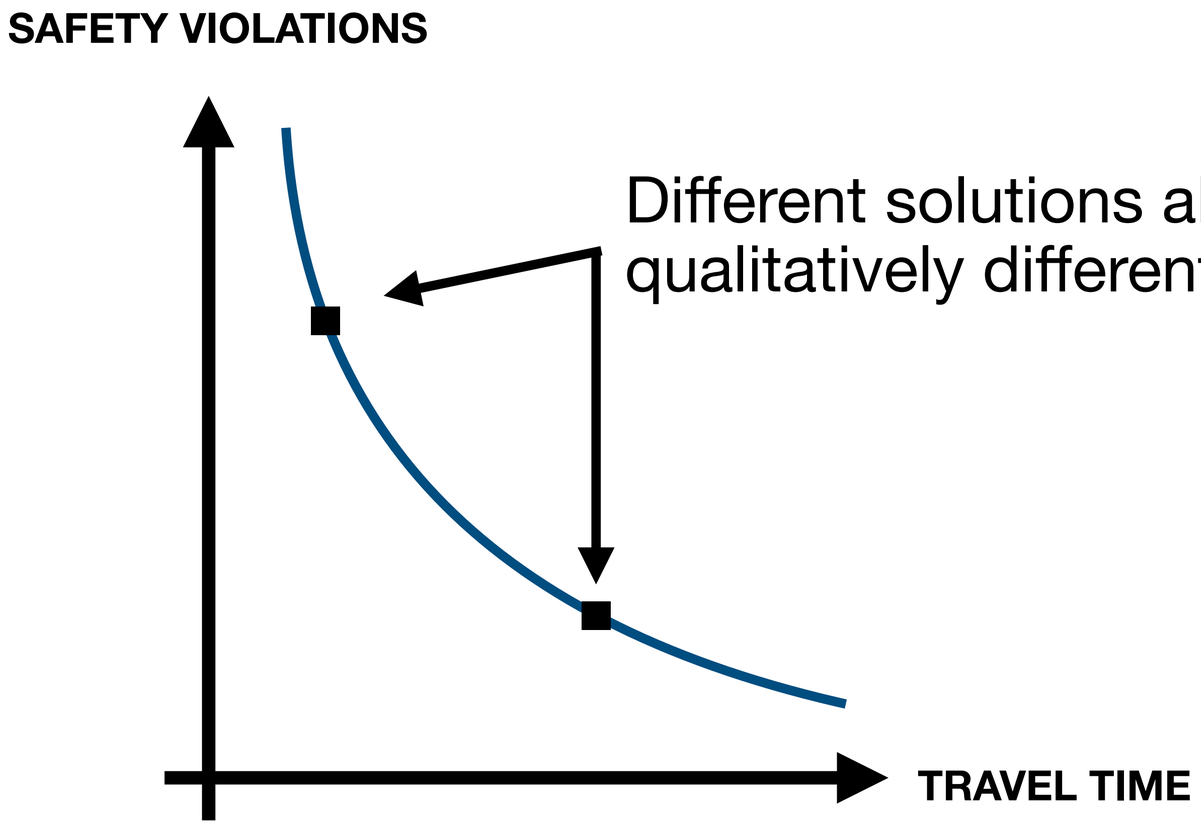
Visualization: Pareto Front



Visualization: Pareto Front

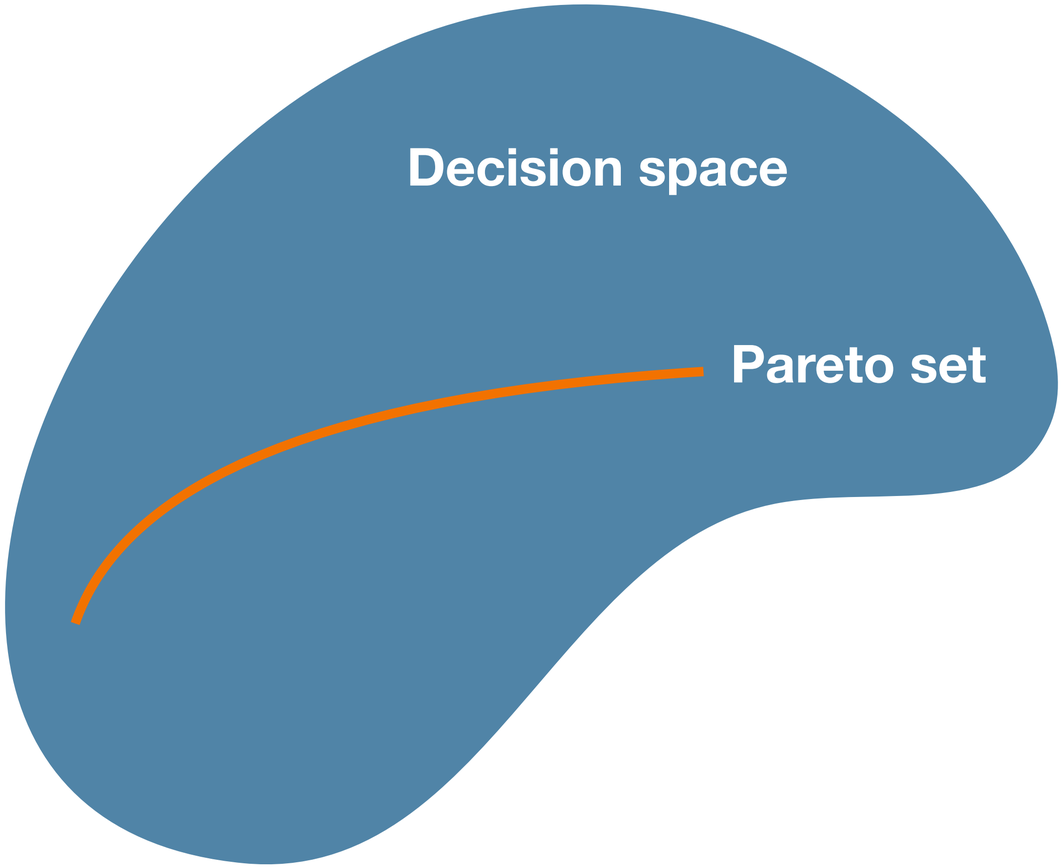


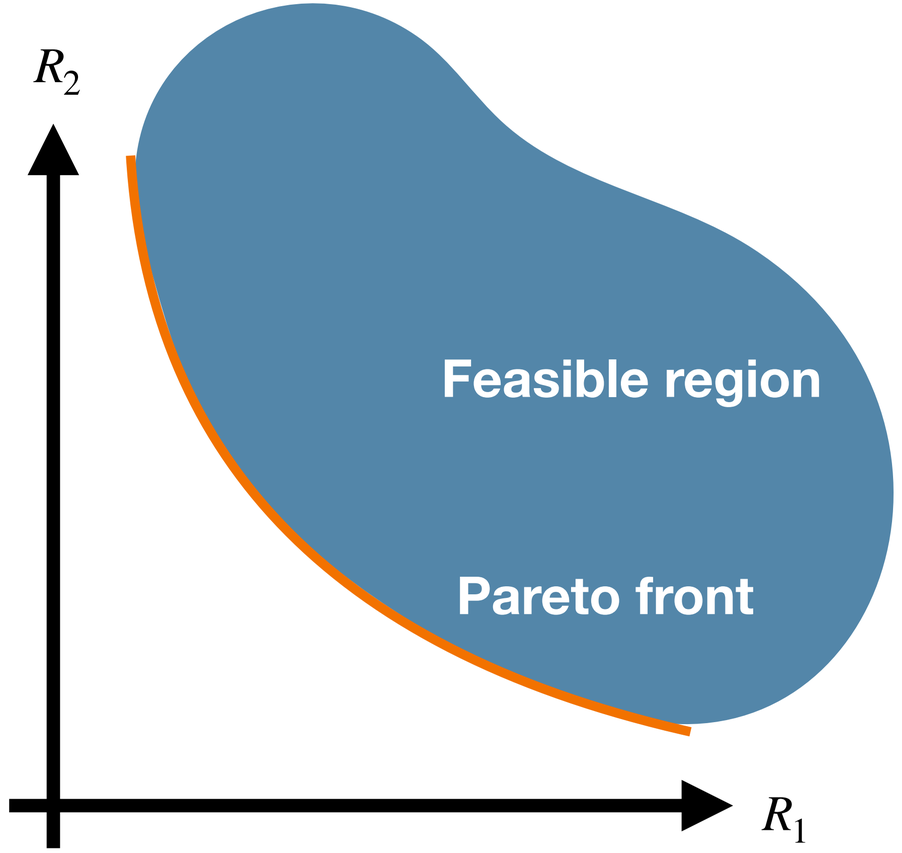
Visualization: Pareto Front



Different solutions along the Pareto front make qualitatively different types of trade-offs.

Pareto Set vs. Pareto Front





$$\text{Pareto}(\mathbf{R}) \subset \Theta$$

$$\text{Front}(\mathbf{R}) \subset \mathbb{R}^K$$

Examples of Multi-Objective Problems

Examples of Multi-Objective Problems

LLMs/Self-Driving Cars
(Many types risks)

- General capability
- Tail risks
- Monetary costs
- Fairness (e.g. passenger versus pedestrian)
- Personalization

Examples of Multi-Objective Problems

LLMs/Self-Driving Cars
(Many types risks)

- General capability
- Tail risks
- Monetary costs
- Fairness (e.g. passenger versus pedestrian)
- Personalization

Forecasting Models
(Heterogeneous data)

- Limited real-world data, no counterfactual data
- Abundant low-fidelity simulated data
- Some expensive high-fidelity simulated data

Examples of Multi-Objective Problems

LLMs/Self-Driving Cars
(Many types risks)

- General capability
- Tail risks
- Monetary costs
- Fairness (e.g. passenger versus pedestrian)
- Personalization

Forecasting Models
(Heterogeneous data)

- Limited real-world data, no counterfactual data
- Abundant low-fidelity simulated data
- Some expensive high-fidelity simulated data

Corporate Structures
(Division of labor)

- Break down complex tasks into smaller responsibilities
- Potential for modularity, scalability, and robustness
- Specialization can lead to novel improvements
- Price of anarchy?

Instead of **minimizing one risk**, the goal is to **make reasonable trade-offs.**

I. Common Knowledge

II. Some Classical Ideas

III. Structure in Complex Tasks

Hedging on the Frontier: Learning New Tasks with Few Samples

ICML 2026 Spotlight



Tobias Wegel



Federico Di Gennaro



Fanny Yang

Naive question: Are LLMs getting better, and how do we know?

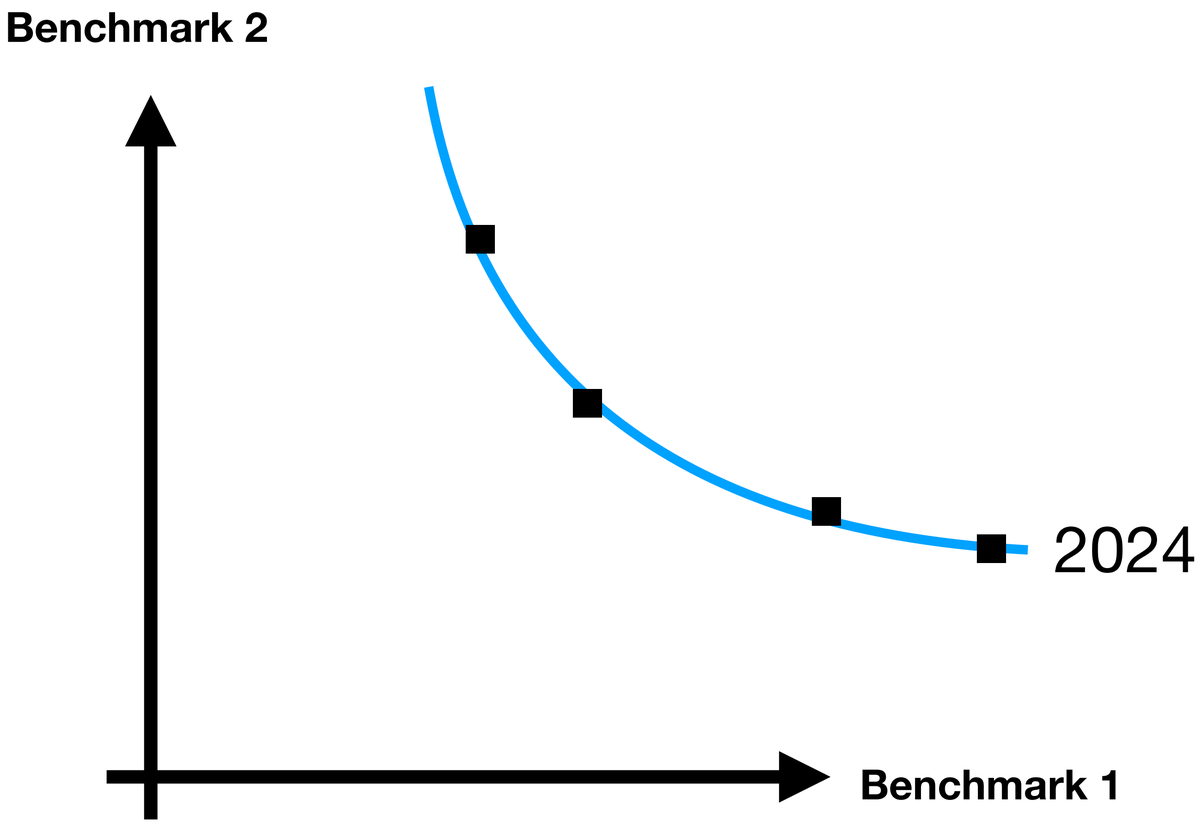
Naive question: Are LLMs getting better, and how do we know?

Or, is Mythical-Human-Preference(θ_t) improving?

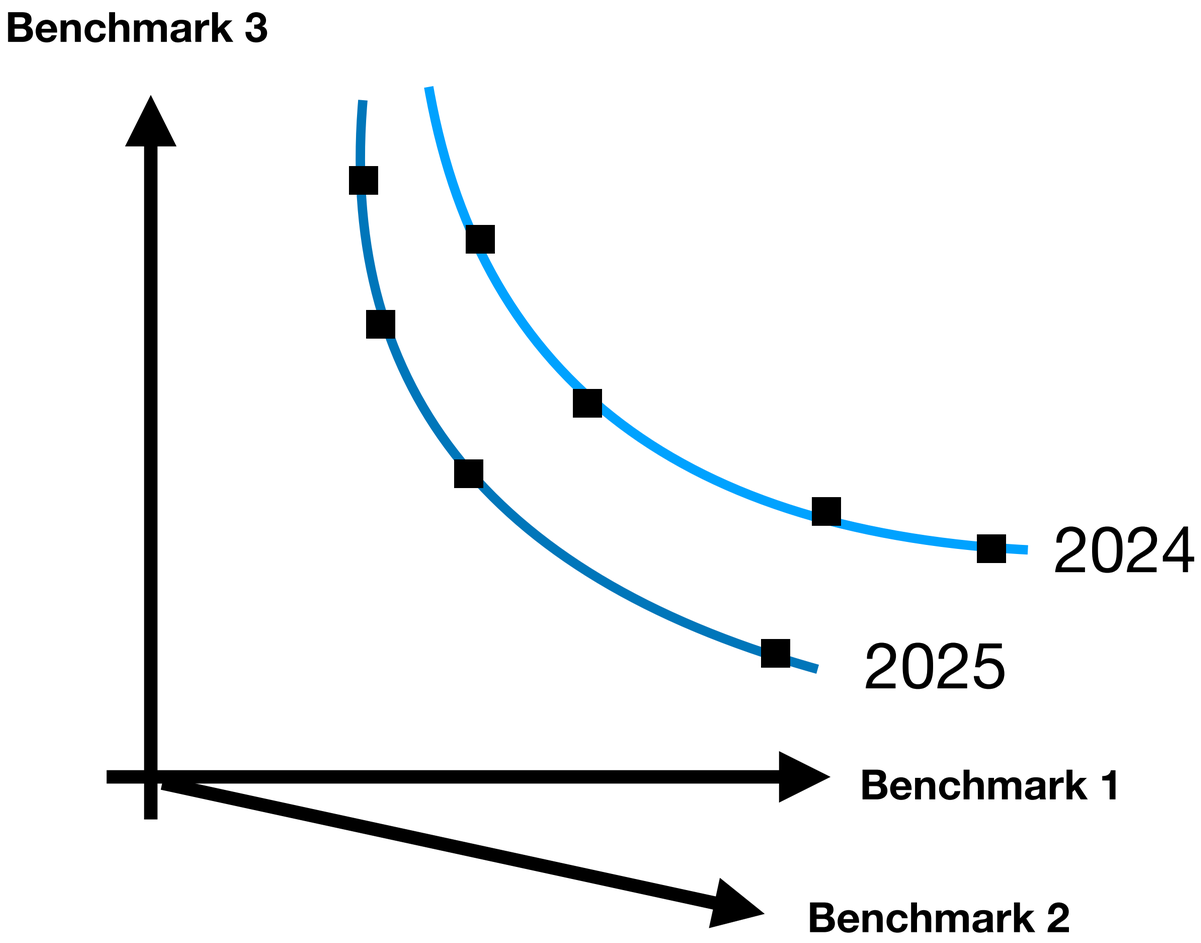
Naive question: Are LLMs getting better, and how do we know?

Simpler goal: I want an LLM to do a specific task R_0. Is $R_0(\theta_t)$ going down?

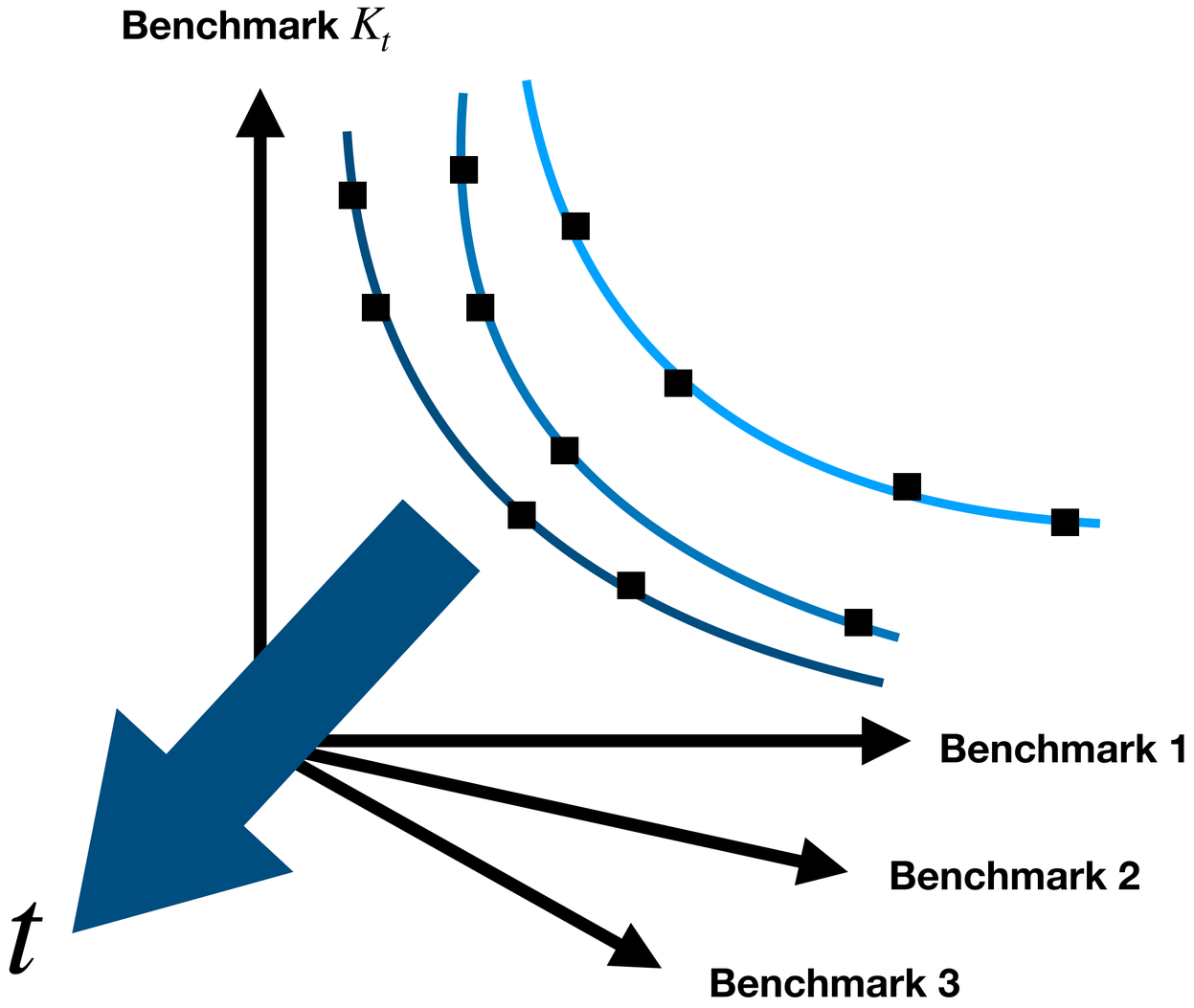
Benchmarks as a Proxy



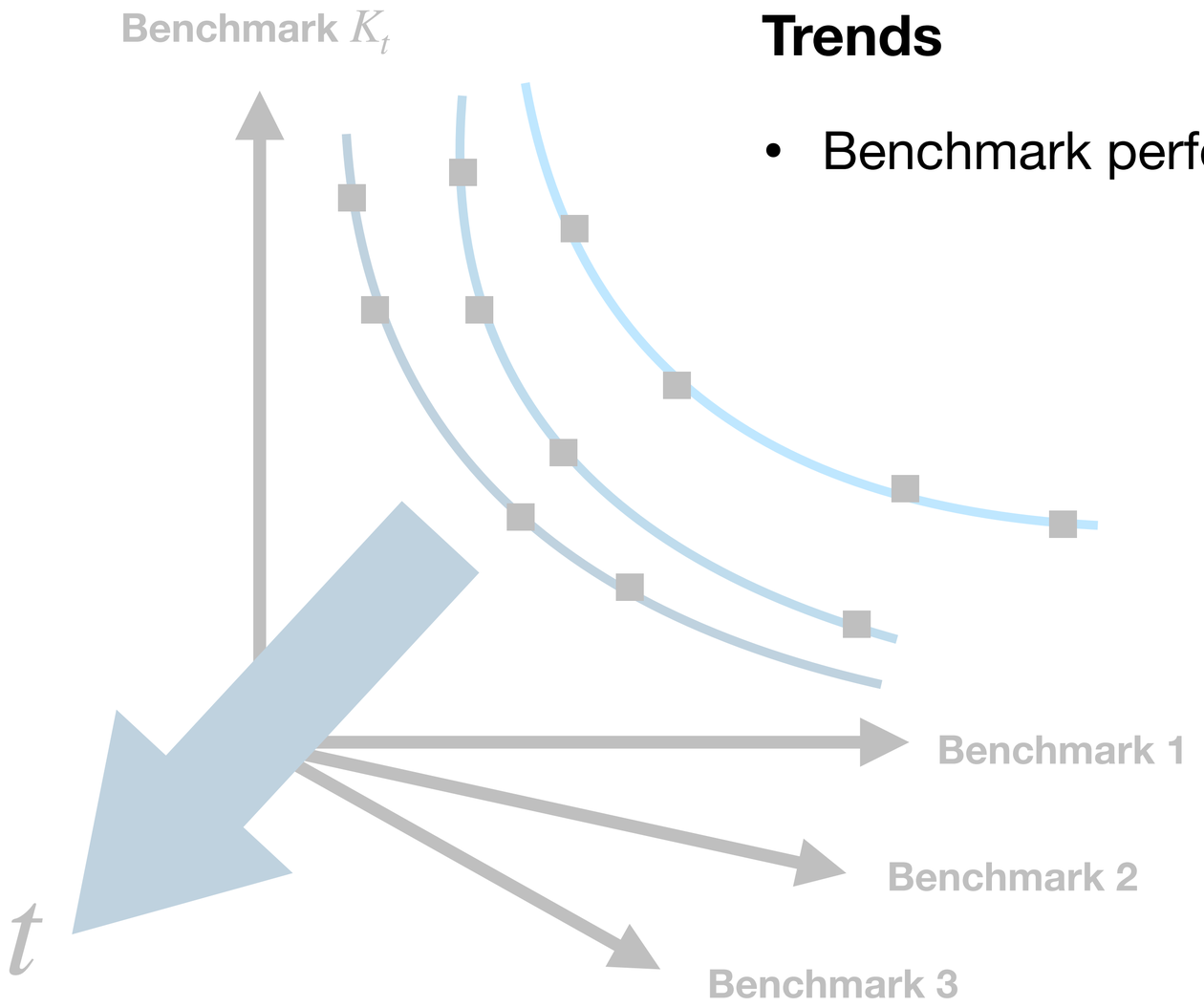
Benchmarks as a Proxy



Benchmarks as a Proxy



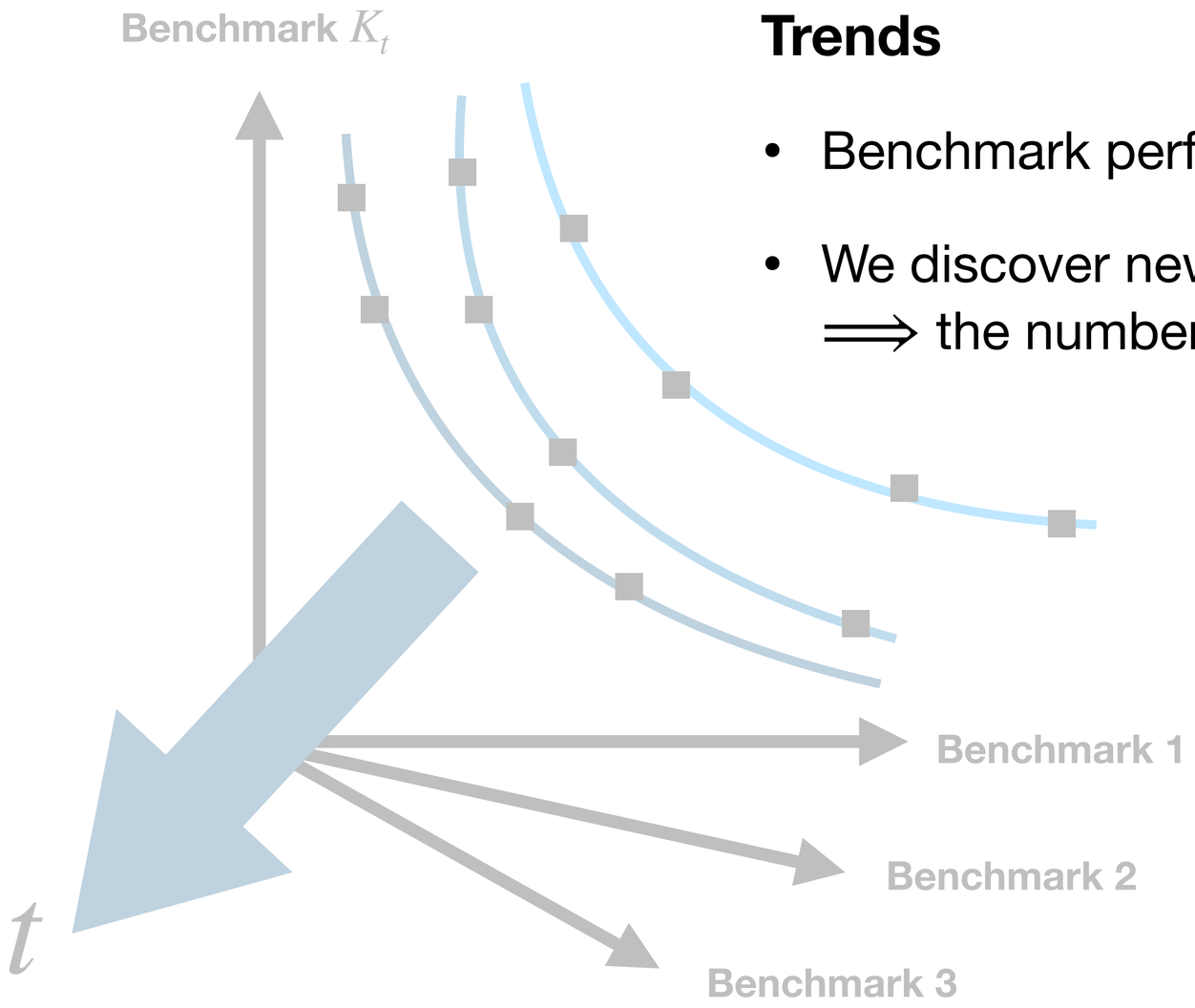
Benchmarks as a Proxy



Trends

- Benchmark performance improves over time

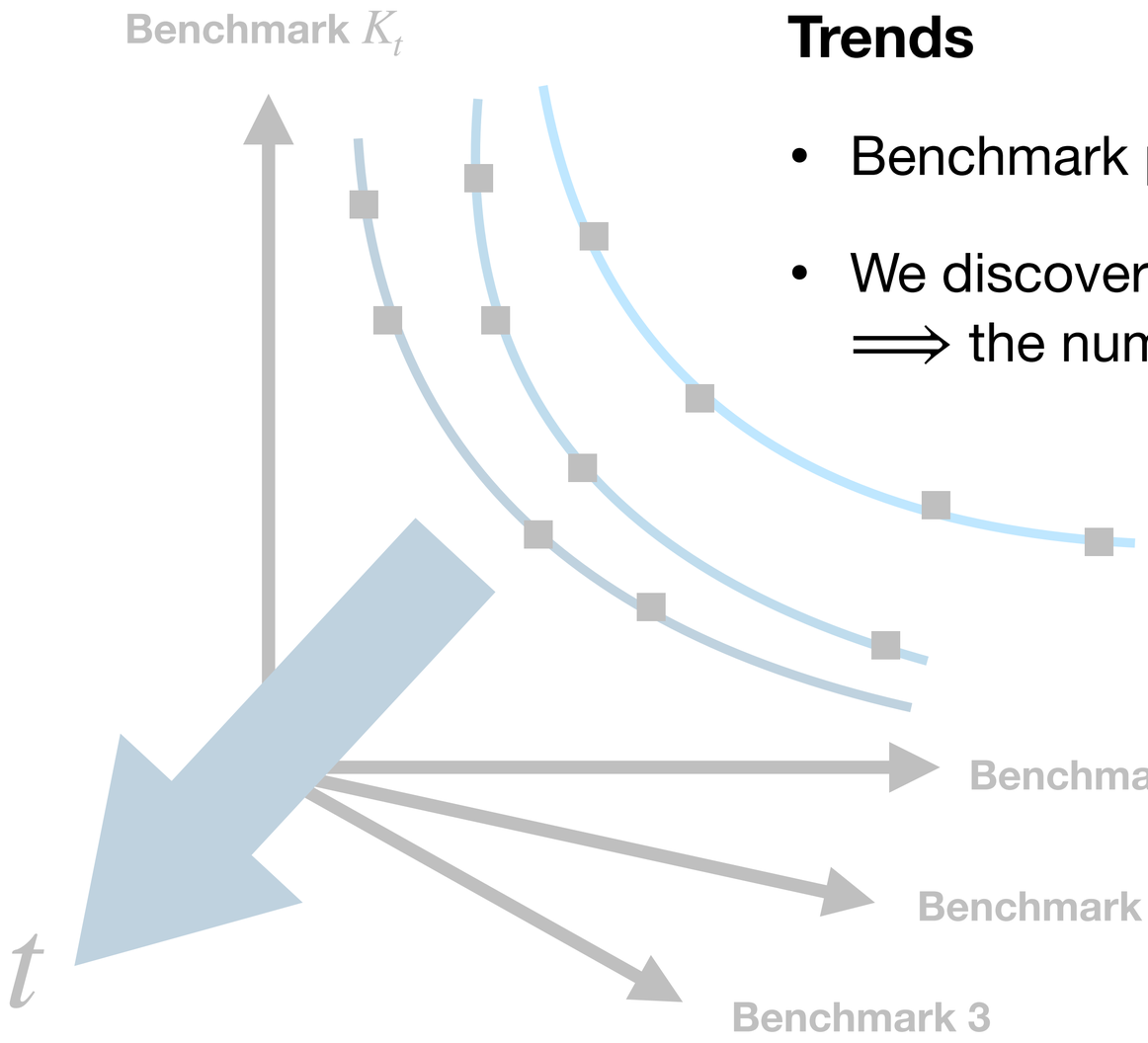
Benchmarks as a Proxy



Trends

- Benchmark performance improves over time
- We discover new dimensions of behavior that need control
  $\Longrightarrow$ the number of benchmarks also grows over time

Benchmarks as a Proxy



Trends

- Benchmark performance improves over time
- We discover new dimensions of behavior that need control
  $\Longrightarrow$ the number of benchmarks also grows over time

Question: How do these benchmarks relate to the specific task a downstream user wants to solve?

Implicit Assumption: Monotonicity

Implicit Assumption: Monotonicity

- Let R_0 be a downstream objective we want to minimize.

Implicit Assumption: Monotonicity

- Let R_0 be a downstream objective we want to minimize.
- Let $R_1, \ldots, R_K$ measure performance along a set of benchmarks.

Implicit Assumption: Monotonicity

- Let R_0 be a downstream objective we want to minimize.
- Let $R_1, \ldots, R_K$ measure performance along a set of benchmarks.

Monotonicity. A model that performs better across all benchmarks is likely to perform better on the target task:

$$\mathbf{R}(\theta^\star) \preceq \mathbf{R}(\theta) \quad \Longrightarrow \quad R_0(\theta^\star) \leq R_0(\theta).$$

Implicit Assumption: Monotonicity

- Let R_0 be a downstream objective we want to minimize.
- Let $R_1, \ldots, R_K$ measure performance along a set of benchmarks.

Monotonicity. A model that performs better across all benchmarks is likely to perform better on the target task:

$$\mathbf{R}(\theta^\star) \preceq \mathbf{R}(\theta) \quad \Longrightarrow \quad R_0(\theta^\star) \leq R_0(\theta).$$

- This is a functional way of positing *task structure* or *relatedness*.

Implicit Assumption: Monotonicity

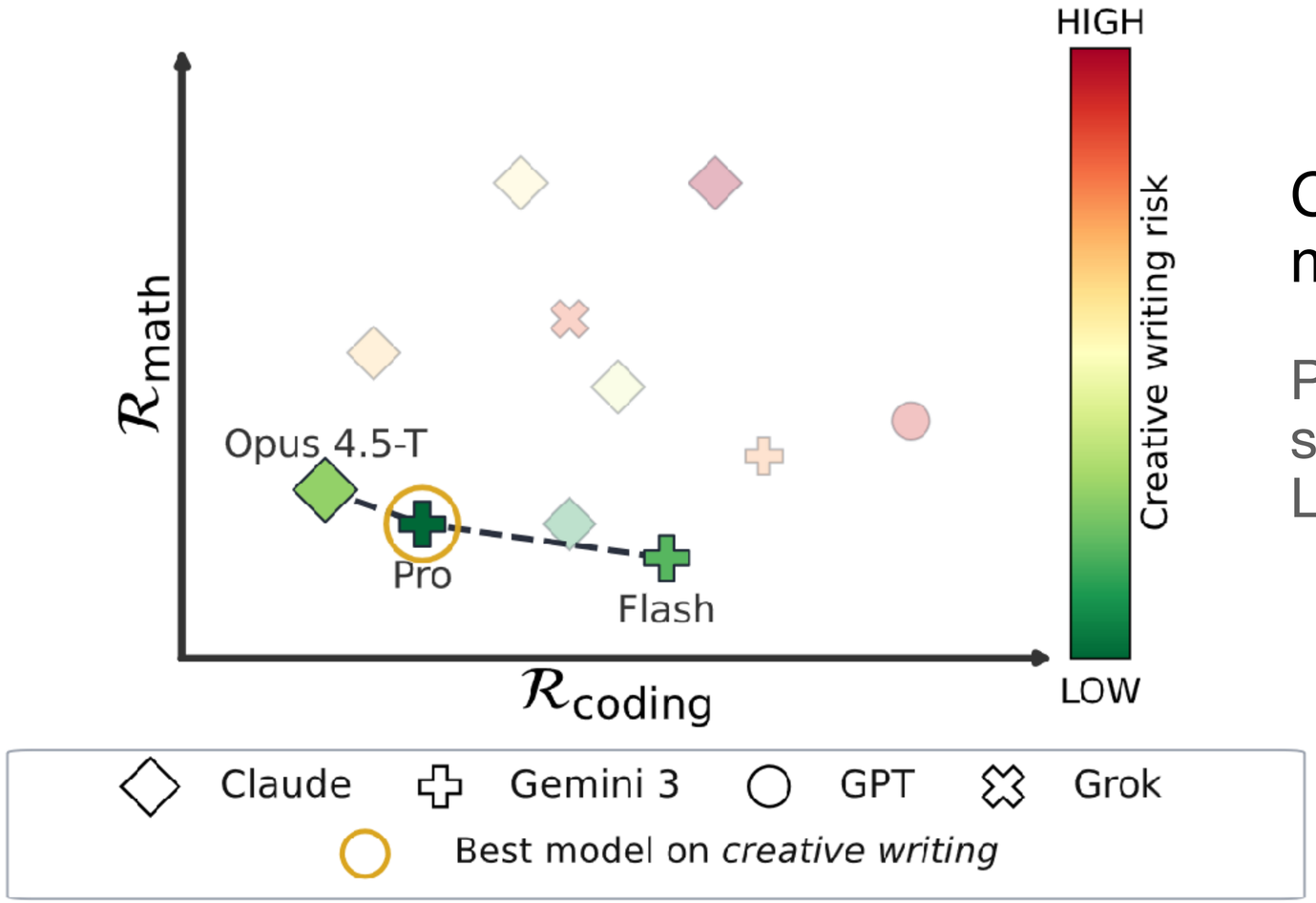
- Let R_0 be a downstream objective we want to minimize.
- Let $R_1, \ldots, R_K$ measure performance along a set of benchmarks.

Monotonicity. A model that performs better across all benchmarks is likely to perform better on the target task:

$$\mathbf{R}(\theta^\star) \preceq \mathbf{R}(\theta) \implies R_0(\theta^\star) \leq R_0(\theta).$$

- This is a functional way of positing *task structure* or *relatedness*.
- Is it reasonable? What are the theoretical implications?

Empirical Evidence of Monotonicity



Creative writing ability is roughly monotonic w.r.t. math and coding.

Pareto front (coding v. math) with risk scores from a subset of HuggingFace LMArena dataset [Chiang et. al. 2024]

Approximate Monotonicity

Modulus of Monotonicity. The modulus of R_0 with respect to $\mathbf{R}$ is a function $\omega(\varepsilon)$ satisfying for all $\theta, \hat{\theta}$:

Approximate Monotonicity

Modulus of Monotonicity. The modulus of R_0 with respect to $\mathbf{R}$ is a function $\omega(\varepsilon)$ satisfying for all $\theta, \hat{\theta}$:

$$\mathbf{R}(\hat{\theta}) \preceq \mathbf{R}(\theta) + \varepsilon \quad \Longrightarrow \quad R_0(\hat{\theta}) \leq R_0(\theta) + \omega(\varepsilon).$$

Approximate Monotonicity

Modulus of Monotonicity. The modulus of R_0 with respect to $\mathbf{R}$ is a function $\omega(\varepsilon)$ satisfying for all $\theta, \hat{\theta}$:

$$\mathbf{R}(\hat{\theta}) \preceq \mathbf{R}(\theta) + \varepsilon \implies R_0(\hat{\theta}) \leq R_0(\theta) + \omega(\varepsilon).$$

- If $\hat{\theta}$ is ε-competitive with θ on all benchmarks, then it is $\omega(\varepsilon)$-competitive on R_0.

Approximate Monotonicity

Modulus of Monotonicity. The modulus of R_0 with respect to $\mathbf{R}$ is a function $\omega(\varepsilon)$ satisfying for all $\theta, \hat{\theta}$:

$$\mathbf{R}(\hat{\theta}) \preceq \mathbf{R}(\theta) + \varepsilon \quad \Longrightarrow \quad R_0(\hat{\theta}) \leq R_0(\theta) + \omega(\varepsilon).$$

- If $\hat{\theta}$ is ε-competitive with θ on all benchmarks, then it is $\omega(\varepsilon)$-competitive on R_0.
- Read: $\omega(\varepsilon)$ is the worst-case monotonicity violation at resolution ε.

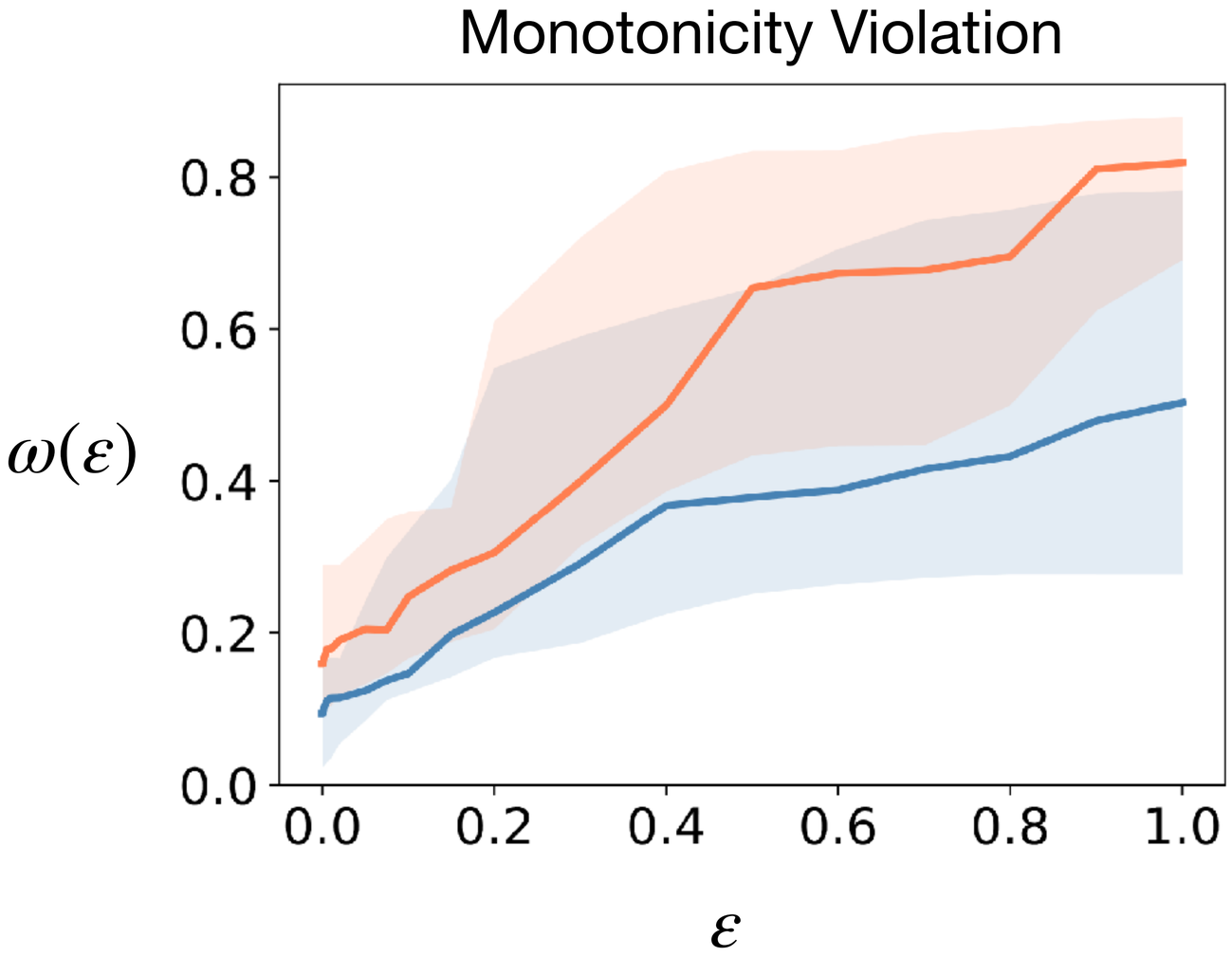
Approximate Monotonicity

Modulus of Monotonicity. The modulus of R_0 with respect to $\mathbf{R}$ is a function $\omega(\varepsilon)$ satisfying for all $\theta, \hat{\theta}$:

$$\mathbf{R}(\hat{\theta}) \preceq \mathbf{R}(\theta) + \varepsilon \quad \Longrightarrow \quad R_0(\hat{\theta}) \leq R_0(\theta) + \omega(\varepsilon).$$

- If $\hat{\theta}$ is ε-competitive with θ on all benchmarks, then it is $\omega(\varepsilon)$-competitive on R_0.
- Read: $\omega(\varepsilon)$ is the worst-case monotonicity violation at resolution ε.
- If $\omega(0) = 0$, then R_0 is exactly monotonic with respect to $\mathbf{R}$.

Empirical Evidence of Monotonicity



- Empirical modulus of monotonicity on real-world benchmarking datasets HELM and VHELM.
- Approximate monotonicity: $\omega(0) \approx 0$.

Empirical Evidence of Monotonicity

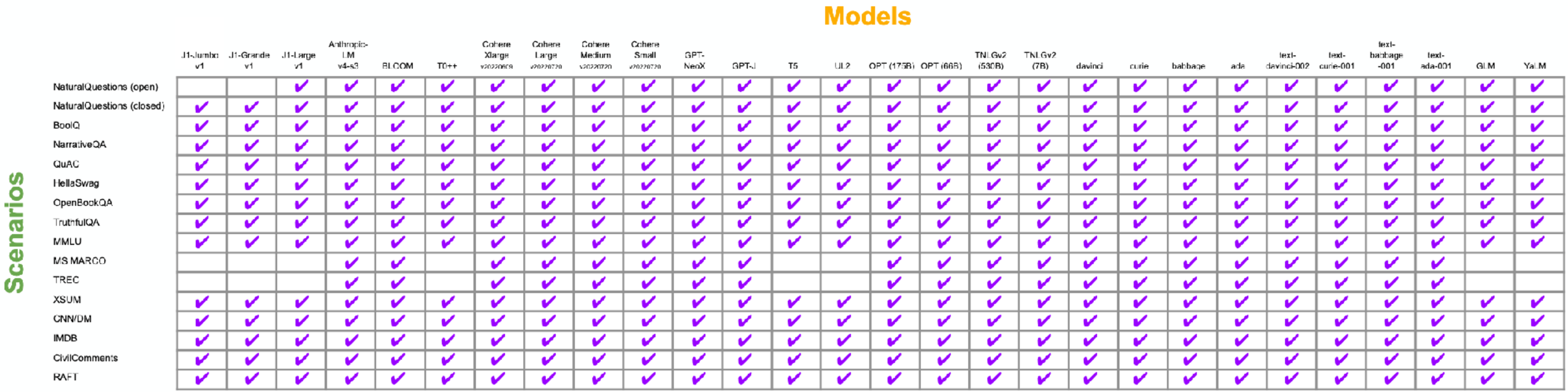
Holistic Evaluation of Language Models

Percy Liang†, Rishi Bommasani†, Tony Lee†, Dimitris Tsipras‡, Dilara Soylu‡, Michihiro Yasunaga‡, Yian Zhang‡, Deepak Narayanan‡, Yuhuai Wu‡, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, Yuta Koreeda
pliang@cs.stanford.edu, nlprishi@stanford.edu, tonyhlee@stanford.edu
Center for Research on Foundation Models (CRFM)
Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Reviewed on OpenReview: *https://openreview.net/forum?id=iO4LZibEqW*

HELM/VHELM are language and vision datasets that evaluate many models across many benchmarks

Empirical Evidence of Monotonicity



HELM/VHELM are language and vision datasets that evaluate many models across many benchmarks

Empirical Evidence of Monotonicity



	J1-Jumbo v1	J1-Grande v1	J1-Large v1	Anthropic-LM v4-s3	BLO
NaturalQuestions (open)			✔	✔	✔
NaturalQuestions (closed)	✔	✔	✔	✔	✔
BoolQ	✔	✔	✔	✔	✔
NarrativeQA	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔
HellaSwag	✔	✔	✔	✔	✔
OpenBookQA	✔	✔	✔	✔	✔
TruthfulQA	✔	✔	✔	✔	✔
MMLU	✔	✔	✔	✔	✔

Example benchmark tasks: question answering, summarization, reasoning, translation,…

Empirical Evidence of Monotonicity



Let's measure how related these tasks are in terms of monotonicity.

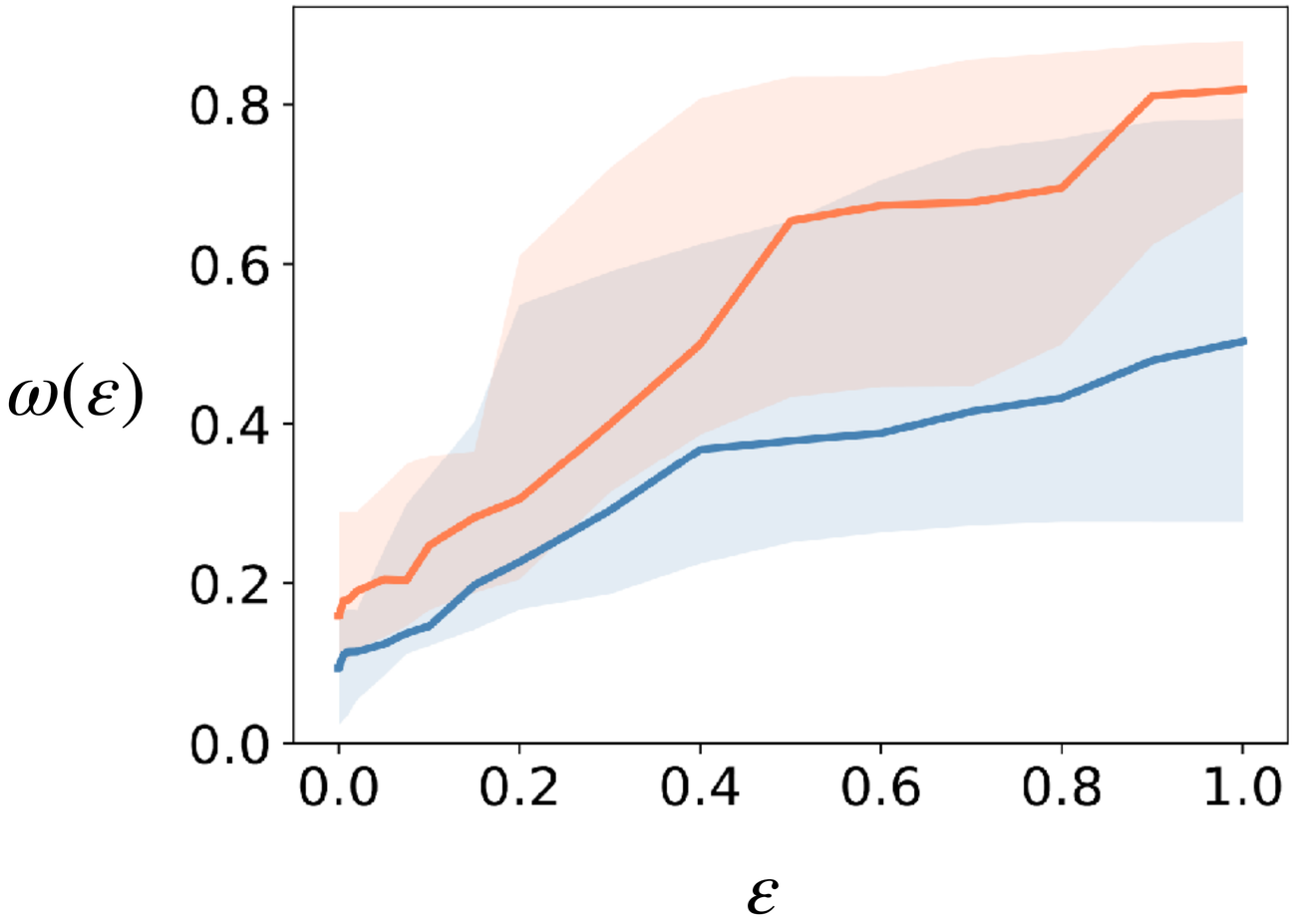
Example benchmark tasks: question answering, summarization, reasoning, translation,…

Empirical Evidence of Monotonicity

Empirical estimation of $\omega(\varepsilon)$

- Let $\mathbf{R} = (R_1, R_2, R_3)$ be three randomly chosen benchmarking tasks.
- Let R_0 be a fourth randomly chosen benchmarking task.
- Compute $\omega(\varepsilon)$ of R_0 wrt $\mathbf{R}$ over all models in the HELM/VHELM dataset.

Empirical Evidence of Monotonicity



- Let $\mathbf{R}$ be three random benchmarks, and R_0 be a fourth random benchmark, scores in [0,1].
- This plots the modulus of R_0 wrt $\mathbf{R}$ (median and quantile).

HELM VHELM

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \text{Pareto}(\mathbf{R}).$$

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \text{Pareto}(\mathbf{R}).$$

Proof. If $\theta^\star$ is not Pareto optimal, some θ strictly dominates $\theta^\star$ on $\mathbf{R}$.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \text{Pareto}(\mathbf{R}).$$

Proof. If $\theta^\star$ is not Pareto optimal, some θ strictly dominates $\theta^\star$ on $\mathbf{R}$.

- Monotonicity implies that θ also strictly outperforms $\theta^\star$ on R_0.

■

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \text{Pareto}(\mathbf{R}).$$

- Practical implication:

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.
  - $\Theta = \mathbb{R}^d$ for LLMs where $d = 500\mathrm{B}$
  - $\mathrm{Pareto}(\mathbf{R})$ is "often" a K-dimensional object, where $K \ll d$

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.
- The monotonicity assumption weakens as K increases.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.
- The monotonicity assumption weakens as K increases.
  - Trade-off: reduce assumptions by increasing K, at the expense of increasing the size of the search space $\mathrm{Pareto}(\mathbf{R})$.

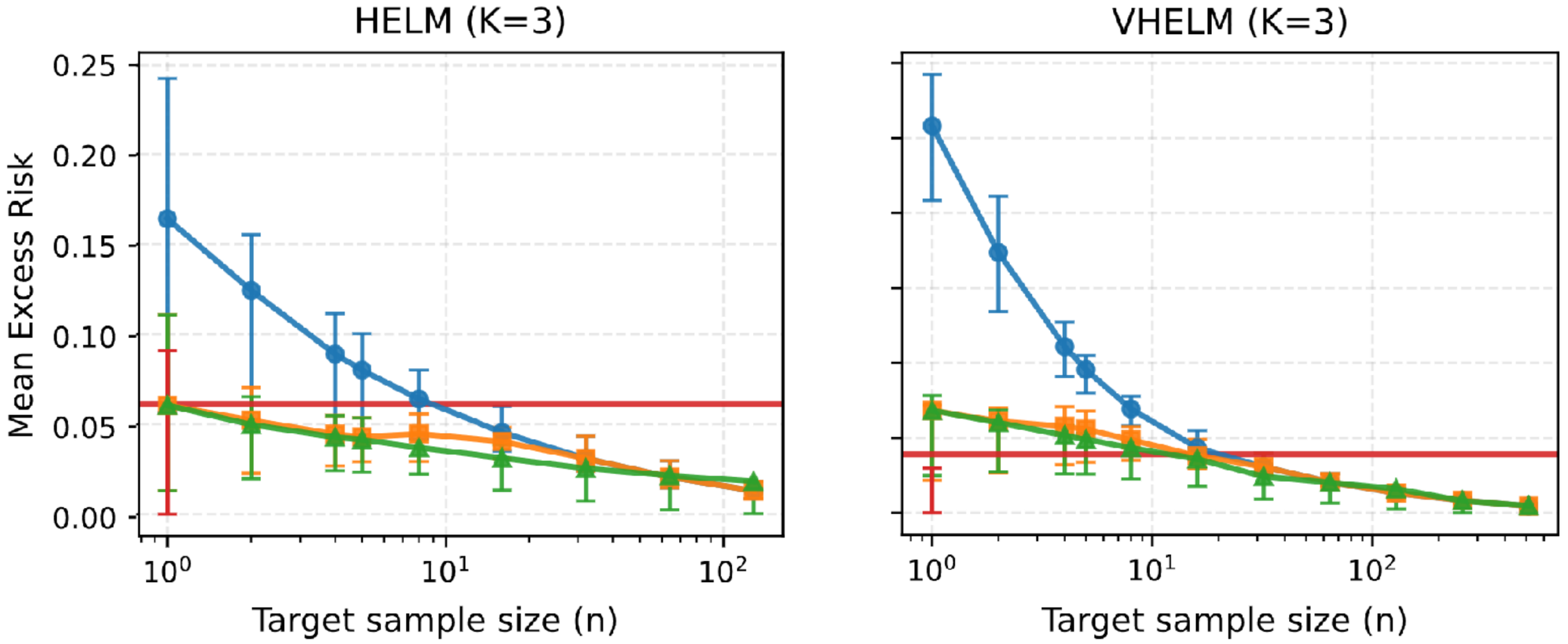
Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then:

$$\theta^\star \in \arg\min_{\theta \in \Theta} R_0(\theta) \quad \Longrightarrow \quad \theta^\star \in \mathrm{Pareto}(\mathbf{R}).$$

- Practical implication: search for $\theta^\star$ in $\mathrm{Pareto}(\mathbf{R})$ instead of Θ.
- The monotonicity assumption weakens as K increases.
- It is possible to design learning algorithms that safely leverages monotonicity assumption with fallback guarantee if assumption does not hold.

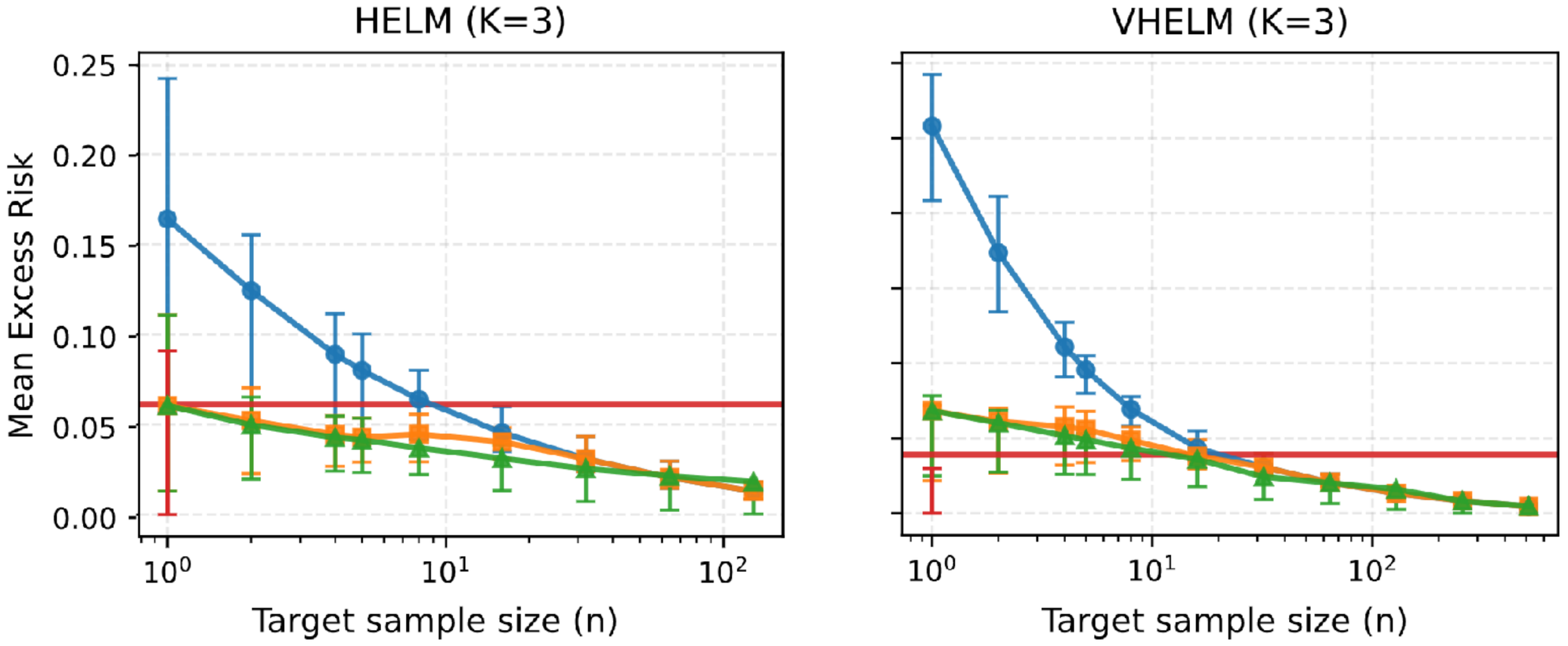
Algorithmic Implications of Monotonicity



Model selection algorithm (out of 29 models)

minmax ERM-all ERM-adaptive ERM-Pareto

Algorithmic Implications of Monotonicity



Model selection algorithm (out of 29 models)

minmax | ERM-all | ERM-adaptive | ERM-Pareto

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$. Then $\mathbf{R}(\theta) \preceq \mathbf{R}(\theta')$.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$. Then $\mathbf{R}(\theta) \preceq \mathbf{R}(\theta')$.

- Monotonicity implies $R_0(\theta) \leq R_0(\theta')$.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$. Then $\mathbf{R}(\theta) \preceq \mathbf{R}(\theta')$.

- Monotonicity implies $R_0(\theta) \leq R_0(\theta')$.
- Symmetry implies $R_0(\theta) = R_0(\theta')$.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$. Then $\mathbf{R}(\theta) \preceq \mathbf{R}(\theta')$.

- Monotonicity implies $R_0(\theta) \leq R_0(\theta')$.
- Symmetry implies $R_0(\theta) = R_0(\theta')$.
- The map $s(\mathbf{R}(\theta)) = R_0(\theta)$ is well-defined

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Proof. Suppose that $\mathbf{R}(\theta) = \mathbf{R}(\theta')$. Then $\mathbf{R}(\theta) \preceq \mathbf{R}(\theta')$.

- Monotonicity implies $R_0(\theta) \leq R_0(\theta')$.
- Symmetry implies $R_0(\theta) = R_0(\theta')$.
- The map $s(\mathbf{R}(\theta)) = R_0(\theta)$ is well-defined (and monotonic, not shown here). ■

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Practical implication:

.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Practical implication:

- Monotonicity is arguable a fairly strong condition

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Practical implication:

- Monotonicity is arguable a fairly strong condition: the information in $\mathbf{R}$ is sufficient if you want to compute R_0.

.

Theoretical Implications of Monotonicity

Proposition. Let R_0 be monotonic with respect to $\mathbf{R}$. Then, there exists a monotonic function $s : \mathbb{R}^K \to \mathbb{R}$ such that

$$R_0 = (s \circ \mathbf{R}).$$

Practical implication:

- Monotonicity is arguable a fairly strong condition: the information in $\mathbf{R}$ is sufficient if you want to compute R_0.
- Monotonicity violations can occur. In practice, when they are discovered, we increase the number of objectives in $\mathbf{R}$. In a sense, this helps capture the process whereby we learn about and accumulate control over an ML system.

Takeaways

- Monotonicity may be a reasonable framework that helps make very nebulous goals more operationalizable (e.g. “build a good LLM”)

Takeaways

- Monotonicity may be a reasonable framework that helps make very nebulous goals more operationalizable (e.g. “build a good LLM”)
- Monotonicity conceptually implies the preference optimization problem

$$\min_{\theta \in \text{Pareto}(\mathbf{R})} R_0(\theta).$$

Takeaways

- Monotonicity may be a reasonable framework that helps make very nebulous goals more operationalizable (e.g. “build a good LLM”)
- Monotonicity conceptually implies the preference optimization problem

$$\min_{\theta \in \text{Pareto}(\mathbf{R})} R_0(\theta).$$

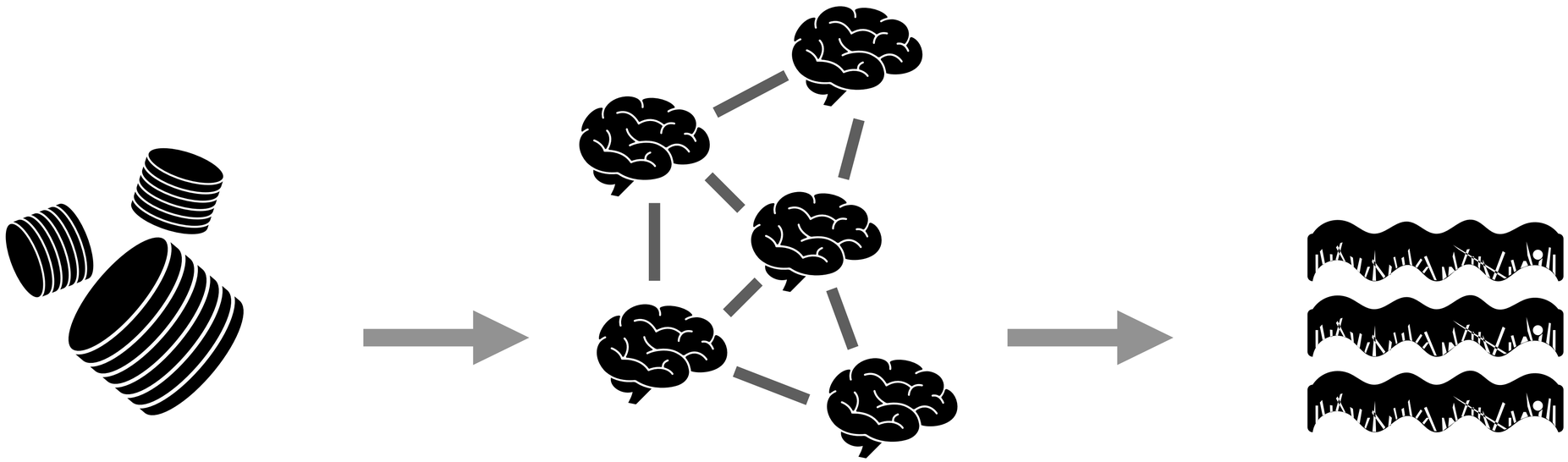
  - Think: R_0 difficult to estimate, while the objectives in $\mathbf{R}$ are more tractable.

Takeaways

- Monotonicity may be a reasonable framework that helps make very nebulous goals more operationalizable (e.g. “build a good LLM”)
- Monotonicity conceptually implies the preference optimization problem

$$\min_{\theta \in \text{Pareto}(\mathbf{R})} R_0(\theta).$$

  - Think: R_0 difficult to estimate, while the objectives in $\mathbf{R}$ are more tractable.
  - Thus, we can incorporate inductive biases or regularize learning R_0 via $\mathbf{R}$.


DATA
MODEL
TASK

Related Works

Hedging on the frontier: Learning new tasks with few samples

with Tobias Wegel, Federico Di Gennaro, and Fanny Yang, ICML 2026 Spotlight

- Introduces task structure through monotonicity

Preference optimization on the Pareto set: On a theory of multi-objective optimization

with Abhishek Roy and Yi-An Ma, NeurIPS 2025

- How to efficiently optimize a preference on a Pareto set

On the sample complexity of semi-supervised multi-objective learning

with Tobias Wegel, Junhyung Park, and Fanny Yang, NeurIPS 2025 Spotlight

- How much data is needed to learn the Pareto set

Learnable mixed Nash equilibria are collectively rational

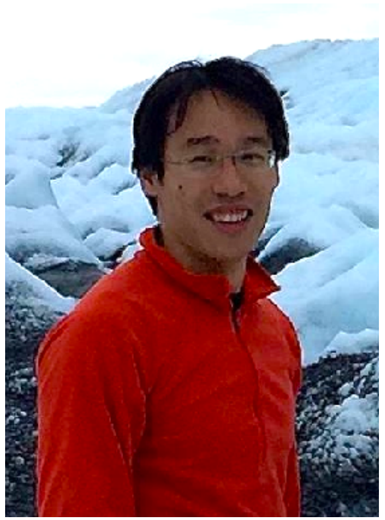
with Yi-An Ma, preprint

- Economic implications of dynamical stability in decentralized learning

Collaborators



Federico
Di Gennaro



Yian
Ma



Junhyung
Park



Abhishek
Roy



Tobias
Wegel



Fanny
Yang

Thanks!

Geelon So, geelon.github.io