Safety vs. Performance: How Multi-Objective Learning Reduces Barriers to Market Entry

Meena Jagadeesan (UC Berkeley / Stanford)

Joint work with Michael I. Jordan and Jacob Steinhardt (UC Berkeley)

https://arxiv.org/abs/2409.03734

High-level overview of this work

We study the emerging market where companies train large language models (LLMs).

?

High-level overview of this work

We study the emerging market where companies train large language models (LLMs).

Key features of this market:

- Training models requires large amounts of **data**
- Companies balance **multiple training objectives**

High-level overview of this work

We study the emerging market where companies train large language models (LLMs).

Key features of this market:

- Training models requires large amounts of **data**
- Companies balance **multiple training objectives**

This work: a technical framework to quantify how much data a new company needs to enter the market

Outline for the talk

1. **Background**

2. Our model

3. Our results

4. Technical ideas

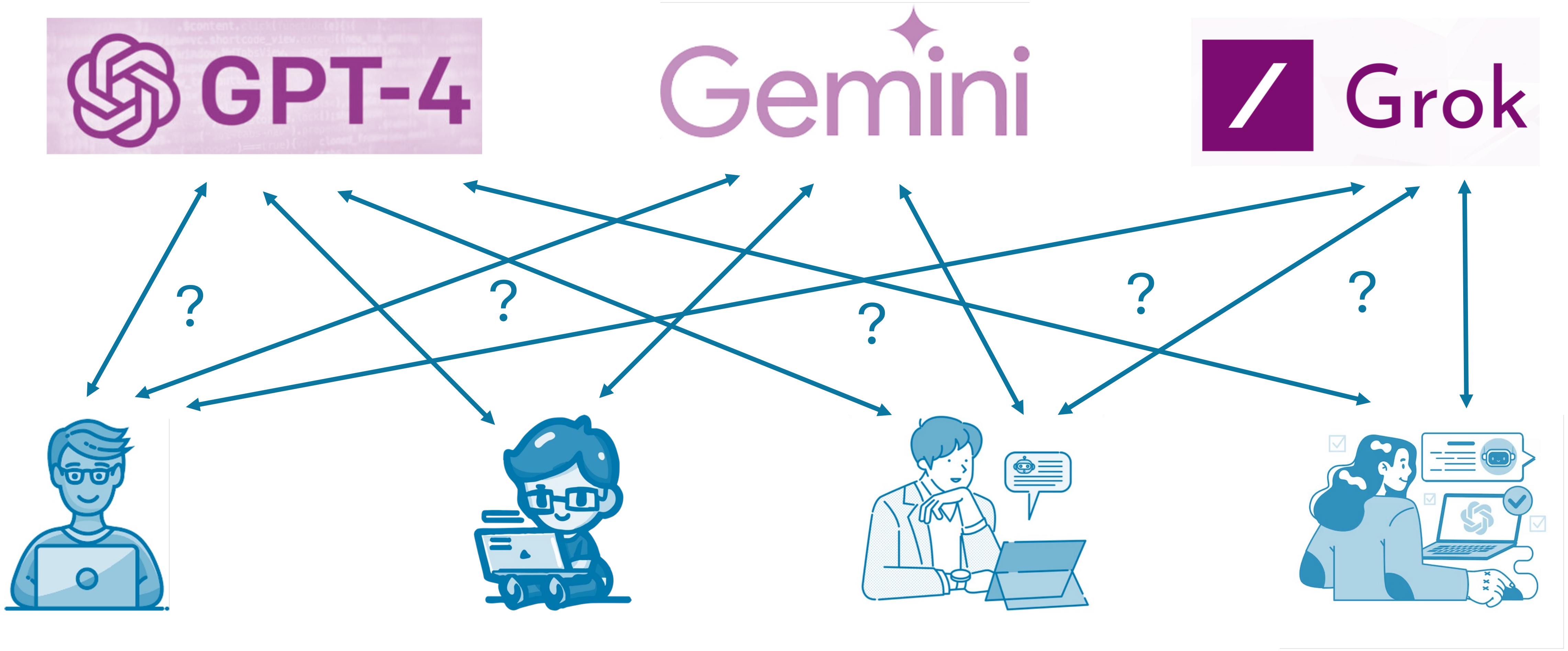
An emerging market of companies that train LLMs

GPT-4

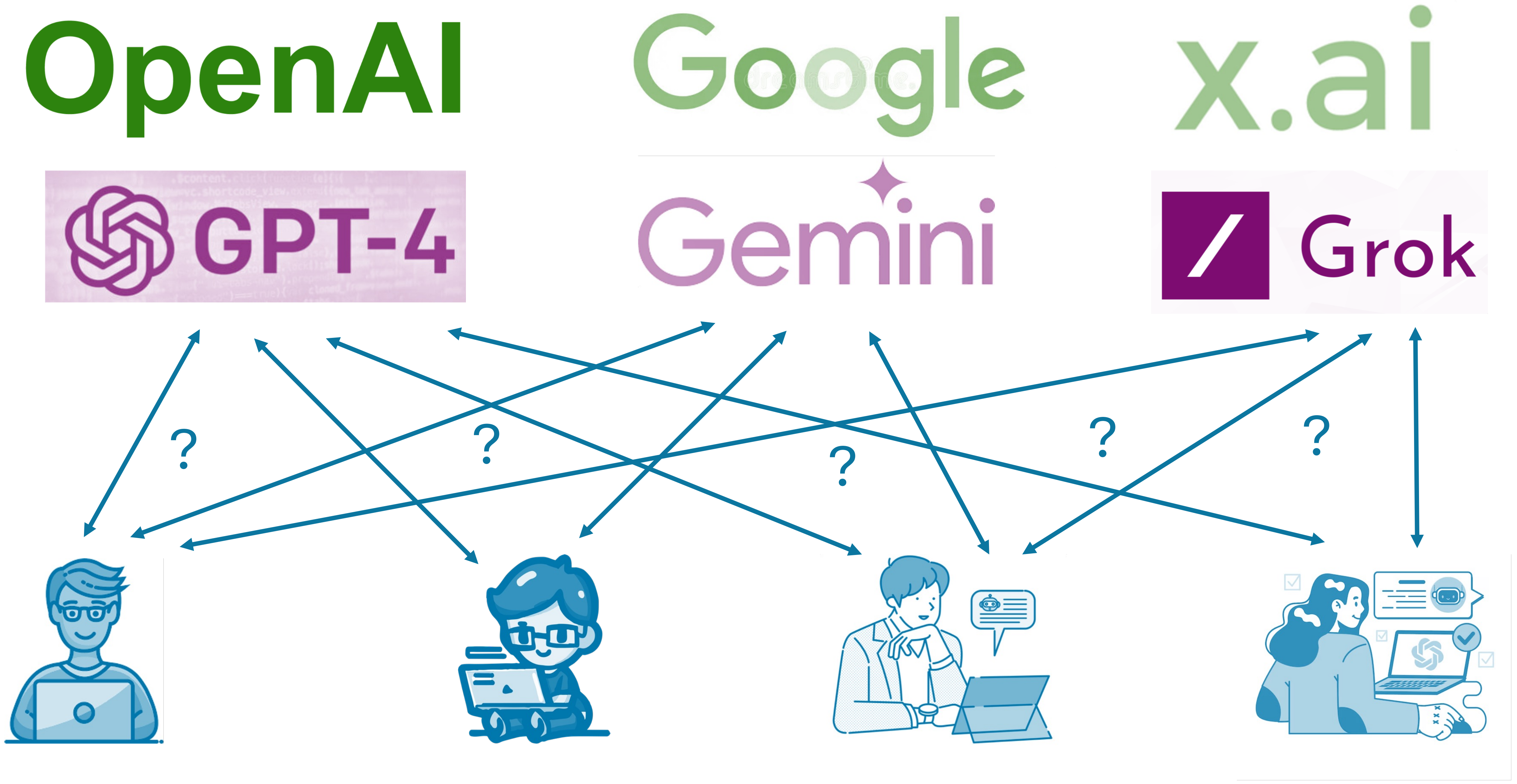
Gemini

Grok

An emerging market of companies that train LLMs



An emerging market of companies that train LLMs



Barriers to market entry

Policymakers have raised concerns about market concentration.

e.g., UK Competition & Markets Authority, White House Executive Order, Brookings Center on Regulation & Markets

Barriers to market entry

Policymakers have raised concerns about market concentration.

e.g., UK Competition & Markets Authority, White House Executive Order, Brookings Center on Regulation & Markets

Typical intuition: New LLM companies face large barriers to entry

Incumbent keeps accumulating data => *Incumbent keeps training models with better performance* => *New company can't reach that performance level*

Drivers: economies of scale, data-driven network effects, etc.

Barriers to market entry

Policymakers have raised concerns about market concentration.

e.g., UK Competition & Markets Authority, White House Executive Order, Brookings Center on Regulation & Markets

Typical intuition: New LLM companies face large barriers to entry

Incumbent keeps accumulating data => *Incumbent keeps training models with better* ***performance*** => *New company can't reach that* ***performance level***

Drivers: economies of scale, data-driven network effects, etc.

Assumption: **Model performance** determines whether a company attracts consumers.

Barriers to market entry

Policymakers have raised concerns about market concentration.

e.g., UK Competition & Markets Authority, White House Executive Order, Brookings Center on Regulation & Markets

Typical intuition: New LLM companies face large barriers to entry

Incumbent keeps accumulating data => *Incumbent keeps training models with better* ***performance*** => *New company can't reach that* ***performance level***

Drivers: economies of scale, data-driven network effects, etc.

Assumption: **Model performance** determines whether a company attracts users.

Reality: companies face pressure to consider objectives beyond performance.

Beyond performance: scrutiny of safety violations

Regulators & society scrutinize **safety violations** of deployed LLMs:

- *E.g.*, LLMs releasing dangerous information (e.g., how to create a weapon)
- *E.g.*, LLMs producing offensive content

Beyond performance: scrutiny of safety violations

Regulators & society scrutinize **safety violations** of deployed LLMs:

- *E.g.*, LLMs releasing dangerous information (e.g., how to create a weapon)
- *E.g.*, LLMs producing offensive content

Scrutiny from regulators:


OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

EU Artificial Intelligence Act

Beyond performance: scrutiny of safety violations

Regulators & society scrutinize **safety violations** of deployed LLMs:

- *E.g.*, LLMs releasing dangerous information (e.g., how to create a weapon)
- *E.g.*, LLMs producing offensive content

Scrutiny from regulators:

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

EU Artificial Intelligence Act

Scrutiny from society:

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

Beyond performance: scrutiny of safety violations

Regulators & society scrutinize **safety violations** of deployed LLMs:

- *E.g.*, LLMs releasing dangerous information (e.g., how to create a weapon)
- *E.g.*, LLMs producing offensive content

Scrutiny from regulators:

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

EU Artificial Intelligence Act

Scrutiny from society:

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

Key property: Large high-resource companies face greater scrutiny than small companies.

Overview of our contributions

This work: We characterize how scrutiny of safety violations shapes data-driven barriers to entry for new companies.

Overview of our contributions

This work: We characterize how scrutiny of safety violations shapes data-driven barriers to entry for new companies.

- We develop a multi-objective learning framework to study markets of companies training LLMs.

Overview of our contributions

This work: We characterize how scrutiny of safety violations shapes data-driven barriers to entry for new companies.

- We develop a multi-objective learning framework to study markets of companies training LLMs.
- We characterize the amount of data that a new company needs to the enter the market.

Overview of our contributions

This work: We characterize how scrutiny of safety violations shapes data-driven barriers to entry for new companies.

- We develop a multi-objective learning framework to study markets of companies training LLMs.
- We characterize the amount of data that a new company needs to the enter the market.
- *En route: new technical tools for multi-objective, high-dim regression*

Overview of our contributions

This work: We characterize how scrutiny of safety violations shapes data-driven barriers to entry for new companies.

Key finding: Scrutiny of safety often---but not always---enables new LLM companies to enter with less data than incumbents

enter the market.

- *En route: new technical tools for multi-objective, high-dim regression*

Related Work

Competition between model-providers:

e.g., Ben-Porat, Tennenholtz ('17, '19), Feng, Gradwohl, Hartline, Johnsen, Nekipelov ('19), Dong, Elzayn, Jabbari, Kearns, Schutzman ('19), Aridor, Mansour, Slivkins, Wu ('20), Iyer and Ke ('22), Kwon, Ginart, Zou ('22), Gradwohl, Tennenholtz ('23), **J.**, Jordan, Haghtalab ('23), **J.**, Jordan, Steinhardt, Haghtalab ('23)

Broader perspectives on algorithmic competition, policy, and dynamics:

e.g., Immorlica, Kalai, Lucier, Moitra, Postlewaite, Tenneholtz ('11), Hashimoto, Srivastava, Namkoong, Liang ('18), Kleinberg, Raghavan ('21) Dean, Curmei, Ratliff, Morgenstern, Fazel ('22), Cen, Hopkins, Ilyas, Madry, Struckman, Caso ('23), Fallah, Jordan ('23), Laufer, Kleinberg, Heidari ('24), Handina, Mazumdar ('24)

Scaling laws and high-dimensional linear regression:

e.g., Hastie et al. ('19), Bordelon et al. ('20), Kaplan et al., ('20), Bahri et al. ('21), Cui et al. ('21), Hashimoto ('21) Hernandez et al. ('21), Hoffmann et al. ('22), Wei et al., ('22), Bach ('23), Jain et al. ('24), Song et al. ('24), Goyal et al. ('24), Covert et al. ('24), Shen et al. ('24), Dohmatob et al. ('24), Mallinar et al. ('24)

Our focus: data-driven barriers to market entry under multi-objective learning

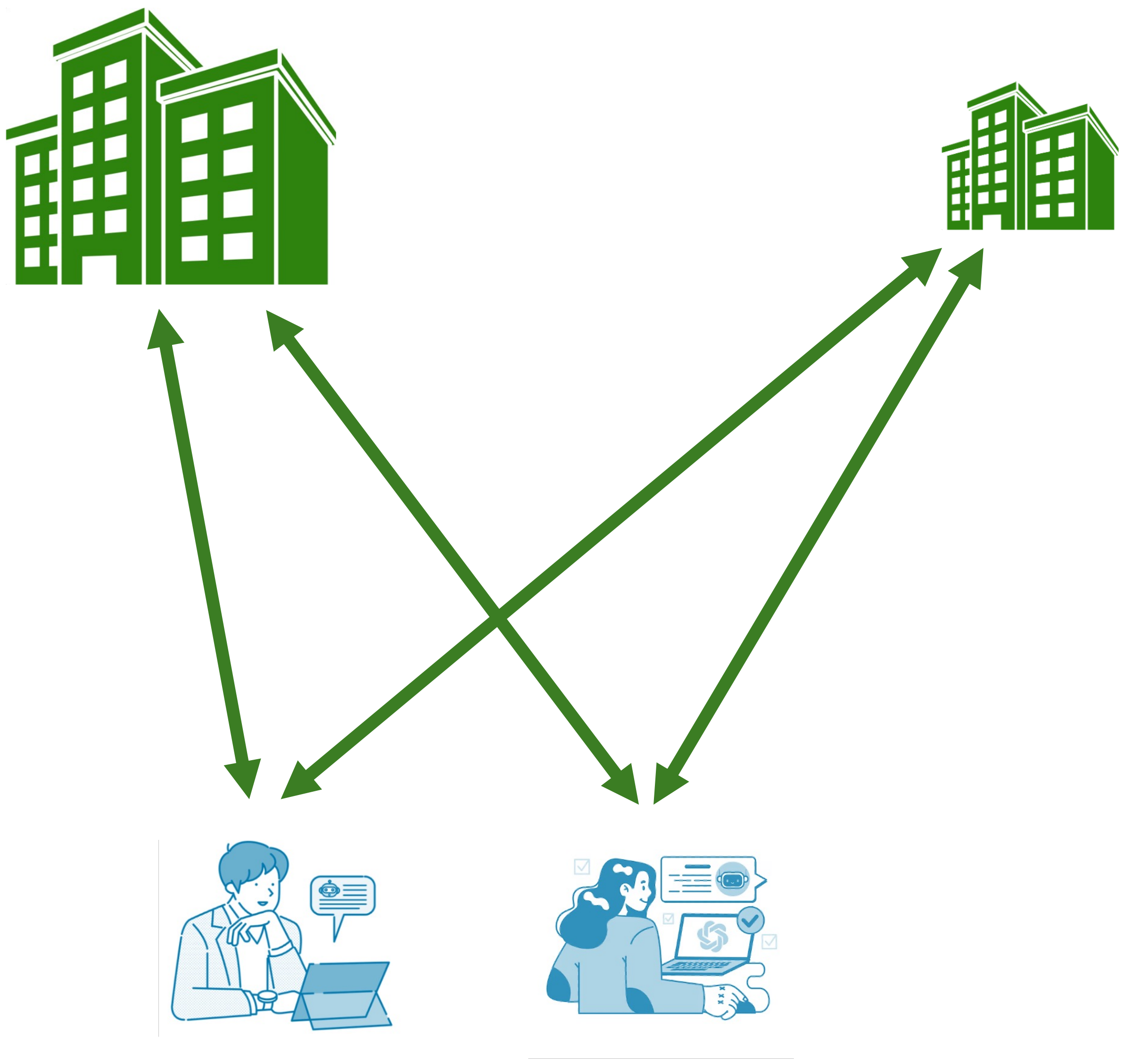
Outline for the talk

1. Background
2. **Our model**
3. Our results
4. Technical ideas

Model overview

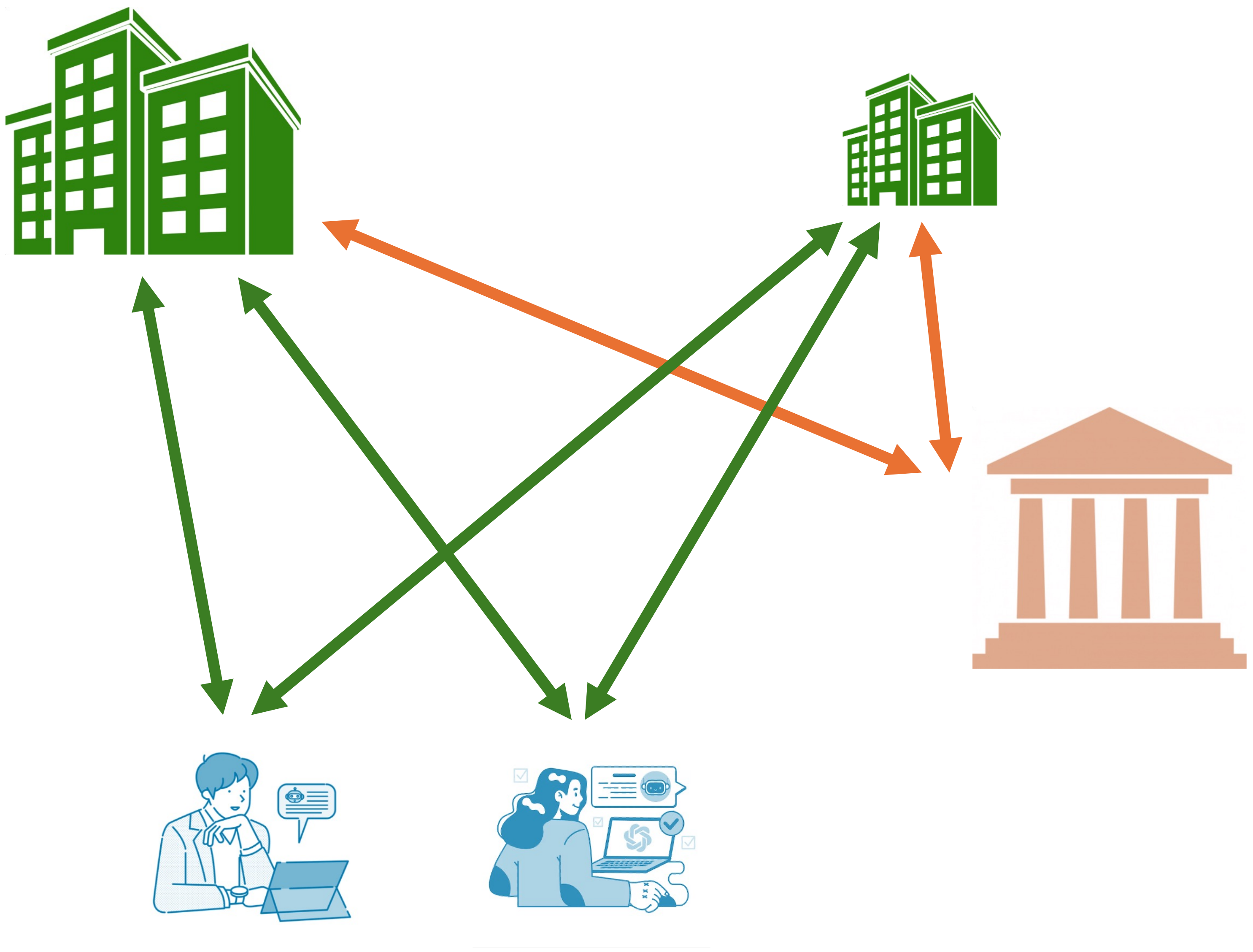
Model overview

Model overview



Each company strategically trains its LLM to attract consumers.

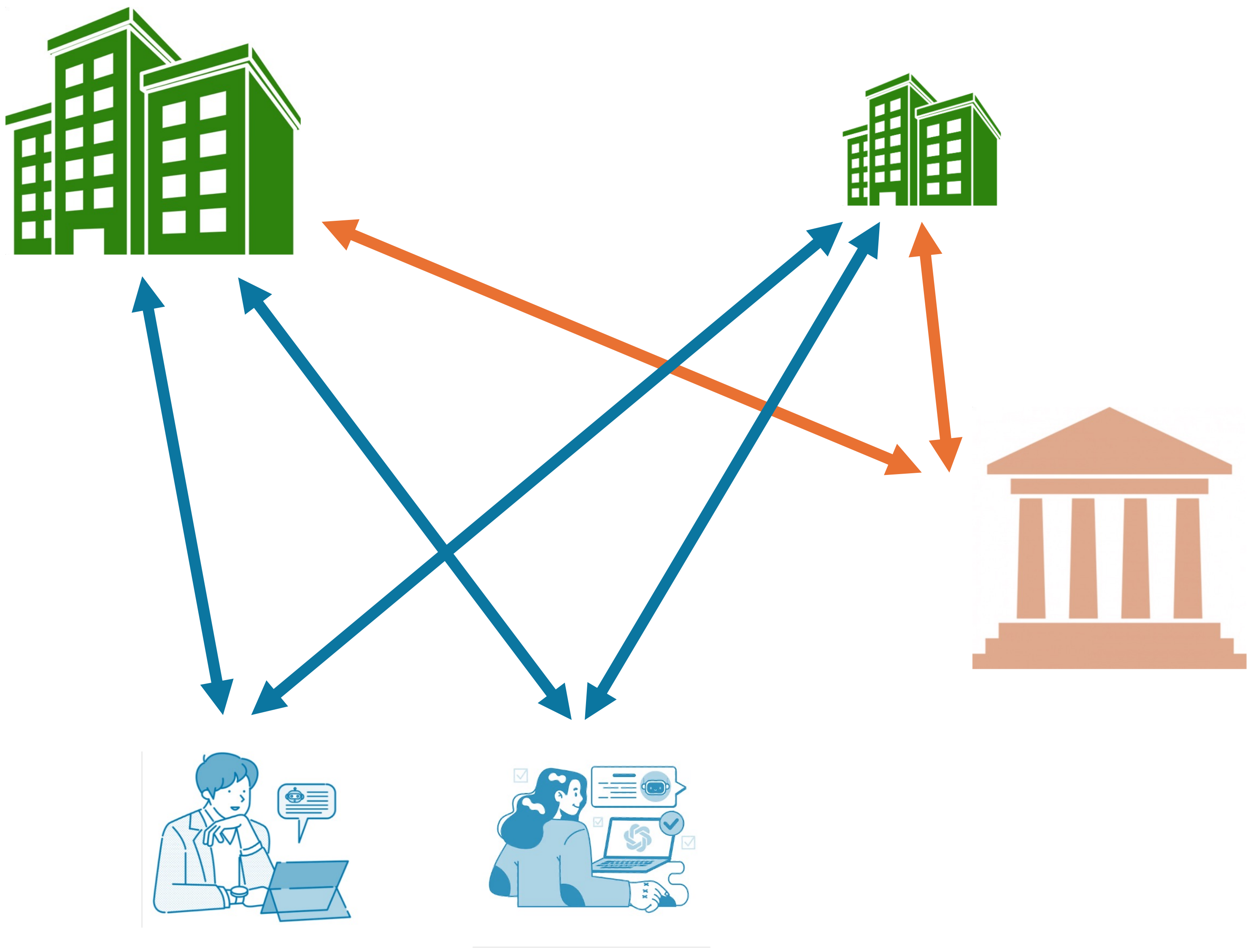
Model overview



Each company strategically trains its LLM to attract consumers.

Regulator/society scrutinizes ***safety violations*** *especially for the incumbent.*

Model overview



Each company strategically trains its LLM to attract consumers.

*Regulator/society scrutinizes **safety violations** especially for the incumbent.*

*Consumers choose the safety-compliant model with best **performance**.*

Model overview: ML pipeline

Model overview: ML pipeline

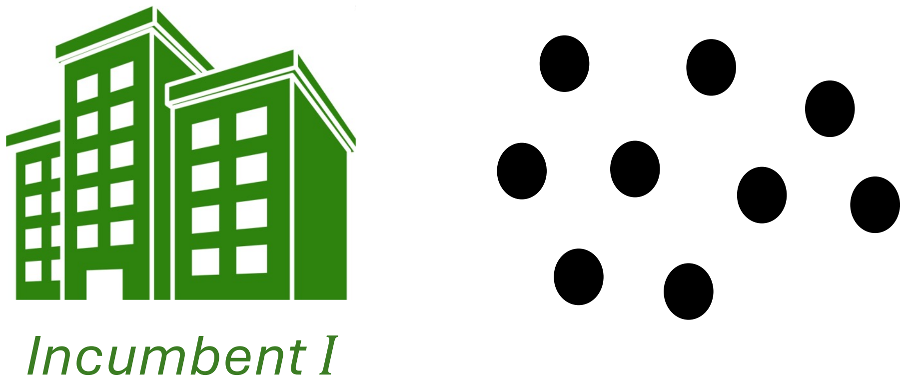
x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

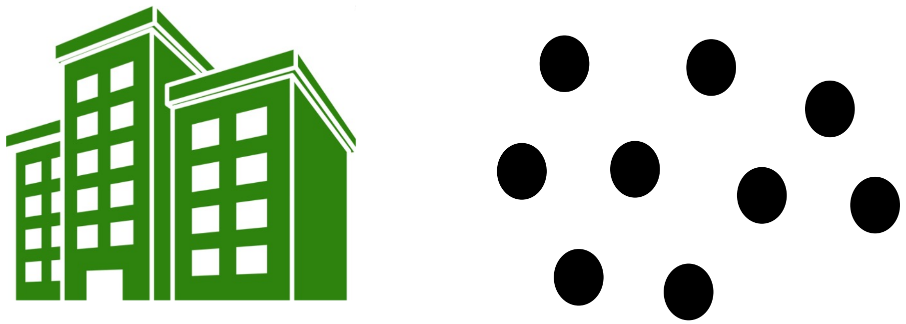
Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output



Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output



Incumbent I

New company E

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

Chooses how to label data

Incumbent I

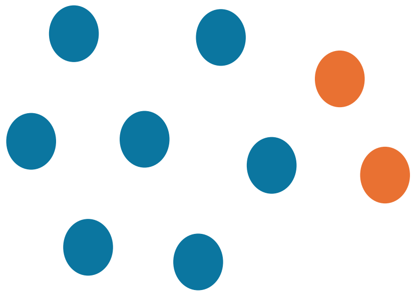
New company E

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

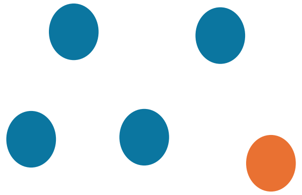
Chooses how to label data **Run regularized regression**

Incumbent I



$\hat{\beta}_{inc}$

New company E



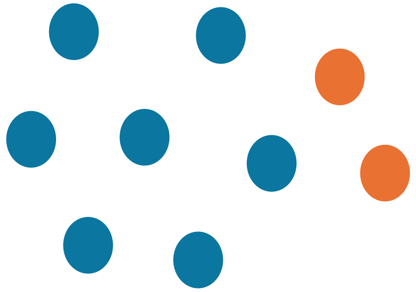
$\hat{\beta}_{new}$

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

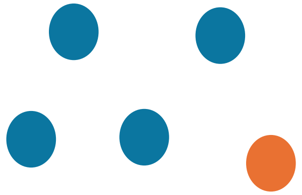
Chooses how to label data **Run regularized regression** **Evaluate safety**

Incumbent I



$\hat{\beta}_{inc}$

New company E



$\hat{\beta}_{new}$

Model overview: ML pipeline

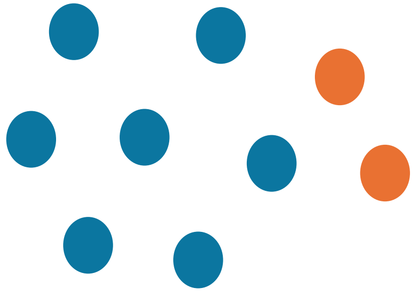
x = high-dim input $\langle\beta_1, x\rangle$ = performance-optimal output $\langle\beta_2, x\rangle$ = safety-optimal output

Chooses how to label data

Run regularized regression

Evaluate safety

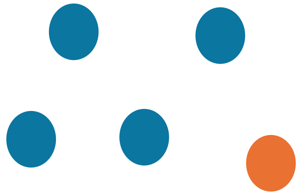
Incumbent I



$\hat{\beta}_{inc}$

Safety requirement:
safety loss < threshold

New company E



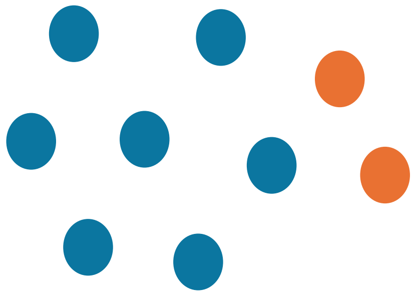
$\hat{\beta}_{new}$

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

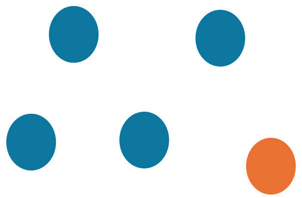
Chooses how to label data **Run regularized regression** **Evaluate** **safety**

Incumbent I



$\hat{\beta}_{inc}$

New company E



$\hat{\beta}_{new}$

Safety requirement:
safety loss < threshold

Incumbent faces a stricter threshold

Model overview: ML pipeline

x = high-dim input $\langle \beta_1, x \rangle$ = performance-optimal output $\langle \beta_2, x \rangle$ = safety-optimal output

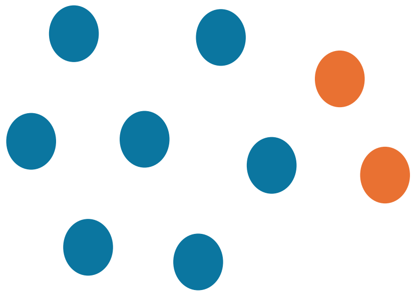
Chooses how to label data

Run regularized regression

Evaluate safety

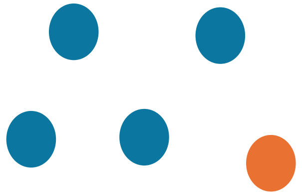
Evaluate performance

Incumbent I



$\hat{\beta}_{inc}$

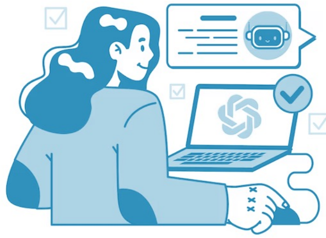
New company E



$\hat{\beta}_{new}$

Safety requirement:
safety loss < threshold

Incumbent faces a stricter threshold



Choose safety-compliant model with best performance

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company C:

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company $\boldsymbol{C}$:

- $\boldsymbol{C}$ receives **unlabelled training dataset** of $\boldsymbol{N_C}$ i.i.d. inputs drawn from D

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company $\boldsymbol{C}$:

- $\boldsymbol{C}$ receives **unlabelled training dataset** of $\boldsymbol{N_C}$ i.i.d. inputs drawn from D
- $\boldsymbol{C}$ chooses a **data mixture level** α_C and **a regularization level** λ_C

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company $\boldsymbol{C}$:

- $\boldsymbol{C}$ receives **unlabelled training dataset** of $\boldsymbol{N_C}$ i.i.d. inputs drawn from D
- $\boldsymbol{C}$ chooses a **data mixture level** α_C and **a regularization level** λ_C
- $\boldsymbol{C}$ labels a random α_C fraction of its data according to β_2 and the rest according to β_1

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company C:

- C receives **unlabelled training dataset** of N_C i.i.d. inputs drawn from D
- C chooses a **data mixture level** α_C and **a regularization level** λ_C
- C labels a random α_C fraction of its data according to β_2 and the rest according to β_1
- C runs **ridge regression** with regularization λ_C on its labelled training data

Model Details: Multi-objective regression

D = distribution of inputs $x \in \mathbf{R}^d$

$\beta_1 \in \mathbf{R}^d$ = ground truth linear function for **performance**-optimal outputs

$\beta_2 \in \mathbf{R}^d$ = ground truth linear function for **safety**-optimal outputs

Multi-objective learning pipeline of each company $\boldsymbol{C}$:

- $\boldsymbol{C}$ receives **unlabelled training dataset** of $\boldsymbol{N_C}$ i.i.d. inputs drawn from D
- $\boldsymbol{C}$ chooses a **data mixture level** α_C and **a regularization level** λ_C
- $\boldsymbol{C}$ labels a random α_C fraction of its data according to β_2 and the rest according to β_1
- $\boldsymbol{C}$ runs **ridge regression** with regularization λ_C on its labelled training data
- $\boldsymbol{C}$ obtains a predictor $\hat{\beta}_C \in \mathbf{R}^d$

Model Details: High-dimensional regression assumptions

The covariates x are **high-dimensional**, i.e. $d \to \infty$ and $d \gg N$

We assume **power law decay** as a function of dimension:

- Eigenvalues of covariance matrix satisfy $\lambda_i \sim i^{-1-\gamma}$.
- Alignment coefficients satisfy $\mathrm{E}[\langle \beta, v_i \rangle]^2 \sim i^{-\delta}$.

Assumptions borrowed from Cui et al., '21, Wei et al., '22, Bach '23

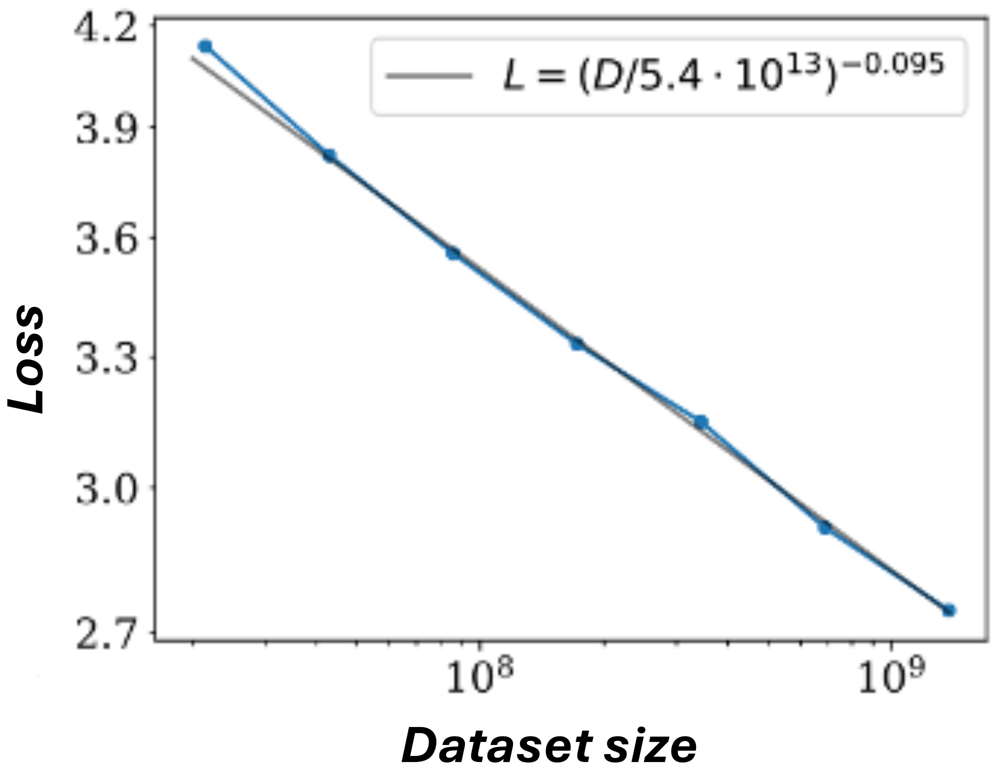
We specify the **correlation between safety and performance** as follow:

- β_1 and β_2 are drawn from a joint distribution with correlation $\rho \in [0,1]$ within each eigendimension, i.e. such that: $\mathrm{E}[\langle \beta_1, v_i \rangle \langle \beta_2, v_i \rangle] \sim \rho \cdot i^{-\delta}$.

Digression: Why high-dimensional regression?

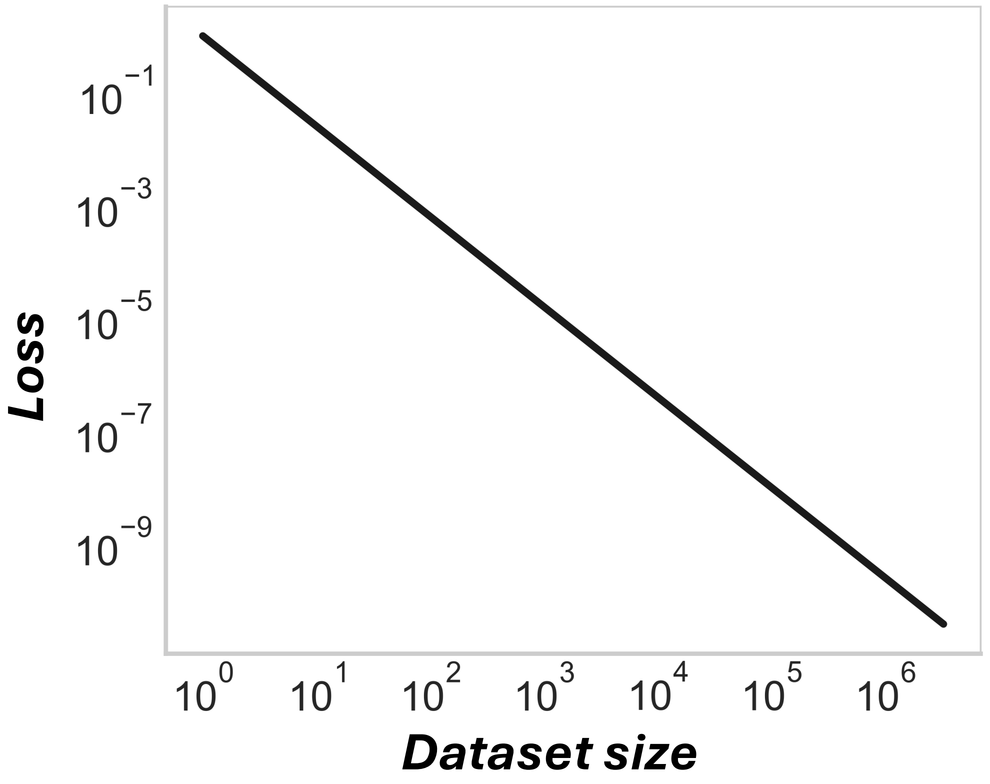
For single-objective data scaling: high-dim regression captures LLM behavior.

LLMs:



e.g., Kaplan et al., 2020

High-dim regression:



e.g., Cui et al., ‘21, Wei et al., ‘22, Bach ‘23

Model Details: Evaluation of Safety and Performance

A company $C \in \{I, E\}$ is* **safety compliant** if:

$$\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_C\,, x\rangle - \langle\beta_2\,, x\rangle\right)^2\right] \leq \tau_C$$

Safety loss

Safety compliance threshold

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Evaluation of Safety and Performance

A company $C \in \{I, E\}$ is* **safety compliant** if:

$$\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_C\,, x\rangle - \langle\beta_2\,, x\rangle\right)^2\right] \le \tau_C$$

Safety loss

Safety compliance threshold

Assumption: incumbent faces a stricter threshold (i.e., $\tau_I < \tau_E$)

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Evaluation of Safety and Performance

A company $C \in \{I, E\}$ is* **safety compliant** if:

$$\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_C\,, x\rangle - \langle\beta_2\,, x\rangle\right)^2\right] \leq \tau_C$$

Safety loss

Safety compliance threshold

Assumption: incumbent faces a stricter threshold (i.e., $\tau_I < \tau_E$)

Consumers choose* the safety-compliant company that **maximize performance**, i.e. that minimize:

$$\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_C\,, x\rangle - \langle\beta_1\,, x\rangle\right)^2\right].$$

Performance loss

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

Data mixture

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

Data mixture *Regularization level*

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

Data mixture

Regularization level

$$\min_{\alpha_C, \lambda_C} \mathbf{E}_{x\sim D}\left[\left(\langle \hat{\boldsymbol{\beta}}_C , \boldsymbol{x}\rangle - \langle \boldsymbol{\beta}_1 , \boldsymbol{x}\rangle\right)^2\right]$$

$$\text{s.t. } \mathbf{E}_{x\sim D}\left[\left(\langle \hat{\boldsymbol{\beta}}_C , \boldsymbol{x}\rangle - \langle \boldsymbol{\beta}_2 , \boldsymbol{x}\rangle\right)^2\right] \leq \boldsymbol{\tau}_C$$

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

Data mixture *Regularization level*

$$\min_{\alpha_C,\lambda_C} \mathbf{E}_{x\sim D}\left[\left(\langle\hat{\boldsymbol{\beta}}_C, \boldsymbol{x}\rangle - \langle\boldsymbol{\beta}_1, \boldsymbol{x}\rangle\right)^2\right]$$
$$\text{s.t. } \mathbf{E}_{x\sim D}\left[\left(\langle\hat{\boldsymbol{\beta}}_C, \boldsymbol{x}\rangle - \langle\boldsymbol{\beta}_2, \boldsymbol{x}\rangle\right)^2\right] \leq \boldsymbol{\tau}_C$$

Market entry threshold := minimum dataset size N_E^* such that the new company E:

- Satisfies* safety compliance* $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_E, x\rangle - \langle\beta_2, x\rangle\right)^2\right] \leq \tau_E$, and
- Achieves* performance $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_E, x\rangle - \langle\beta_1, x\rangle\right)^2\right] \leq \mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_I, x\rangle - \langle\beta_1, x\rangle\right)^2\right]$.

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Model Details: Company choices and Market entry threshold

Each C chooses* α_C and λ_C to ***maximize performance subject to safety compliance.***

Data mixture *Regularization level*

$$\min_{\boldsymbol{\alpha_C},\boldsymbol{\lambda_C}} \mathbf{E}_{\boldsymbol{x}\sim\boldsymbol{D}}\left[\left(\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{C}},\boldsymbol{x}\rangle-\langle\boldsymbol{\beta}_{\mathbf{1}},\boldsymbol{x}\rangle\right)^2\right]$$

$$\textbf{s.t. } \mathbf{E}_{\boldsymbol{x}\sim\boldsymbol{D}}\left[\left(\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{C}},\boldsymbol{x}\rangle-\langle\boldsymbol{\beta}_{\mathbf{2}},\boldsymbol{x}\rangle\right)^2\right]\leq\boldsymbol{\tau_C}$$

Market entry threshold := minimum dataset size $\boldsymbol{N_E^*}$ such that the new company E:

- Satisfies* safety compliance* $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_E,x\rangle-\langle\beta_2,x\rangle\right)^2\right]\leq\tau_E$, and
- Achieves* performance $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_E,x\rangle-\langle\beta_1,x\rangle\right)^2\right]\leq\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}_I,x\rangle-\langle\beta_1,x\rangle\right)^2\right]$.

Our goal: characterize the market entry threshold $\boldsymbol{N_E^*}$

**Caveat: we approximate the safety / performance loss by a deterministic equivalent*

Outline for the talk

1. Background

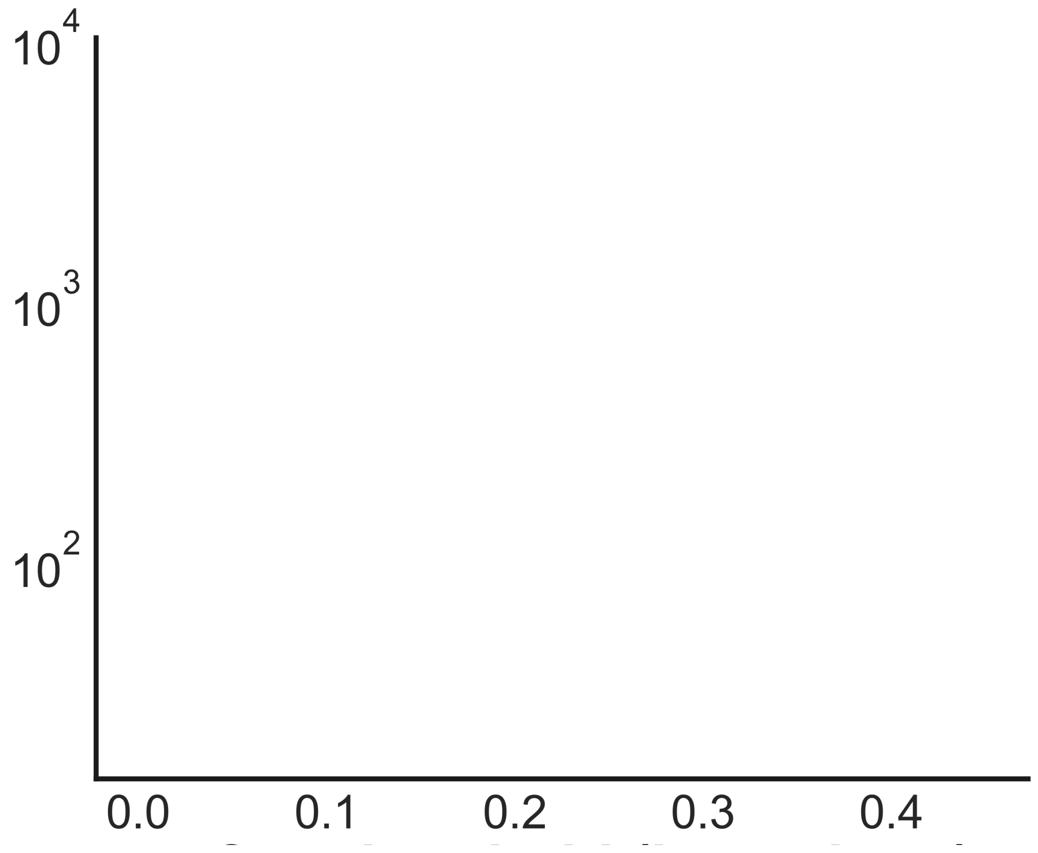
2. Our model

3. **Our results**

4. Technical ideas

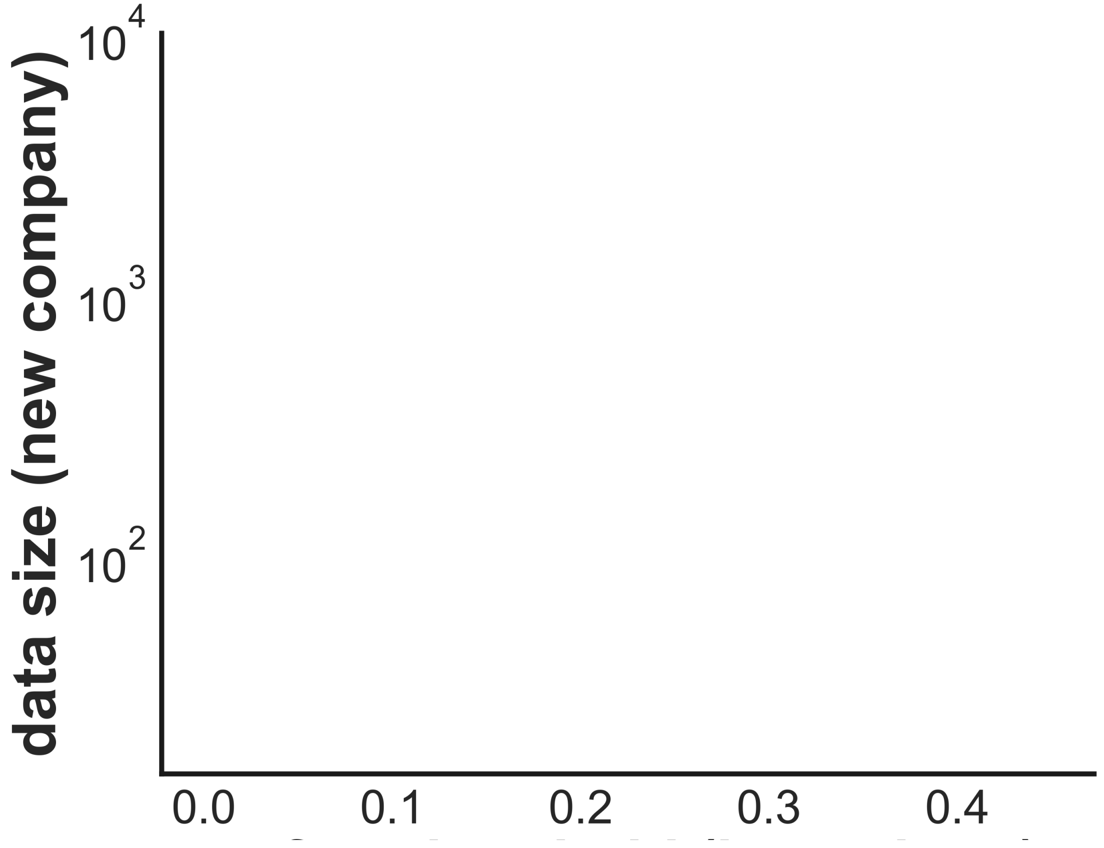
Warmup result

Setup: Incumbent has infinite data $N_I = \infty$*; new company faces no safety constraint* $\tau_E = \infty$.



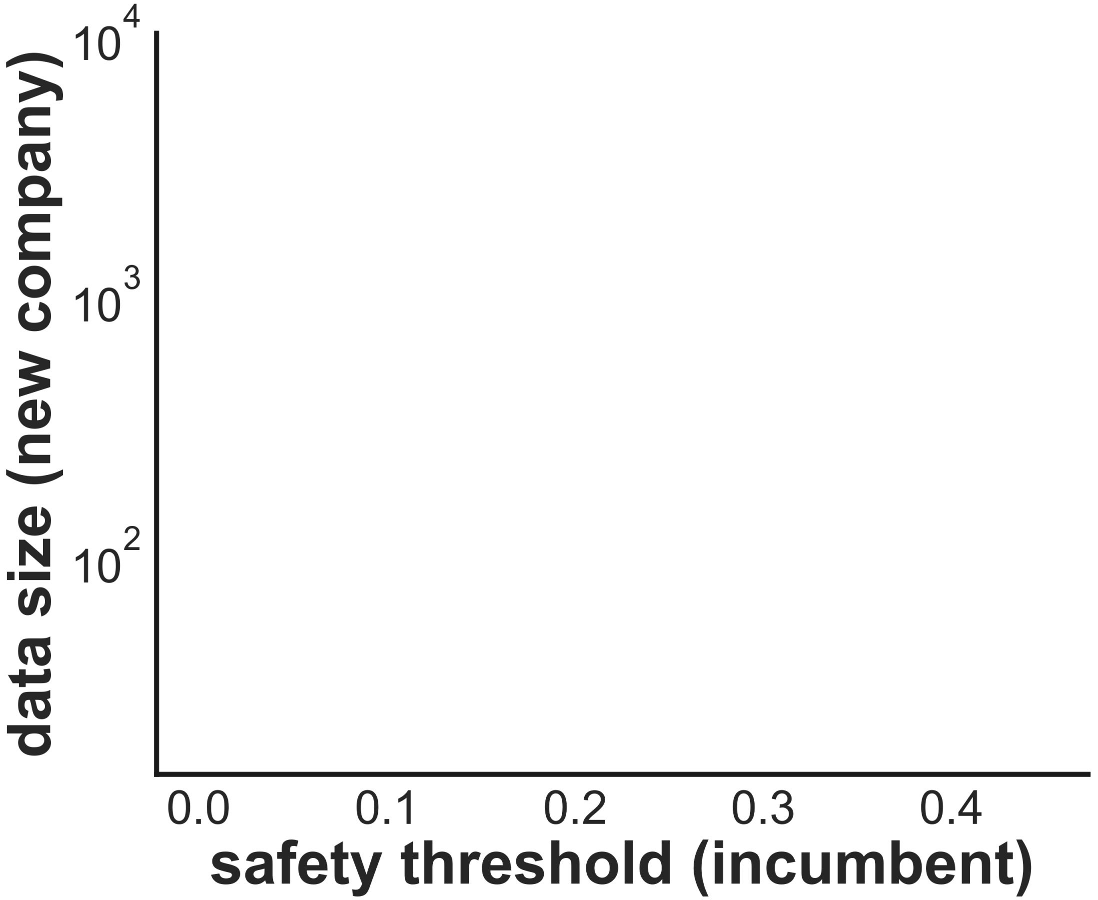
Warmup result

Setup: Incumbent has infinite data $N_I = \infty$*; new company faces no safety constraint* $\tau_E = \infty$.



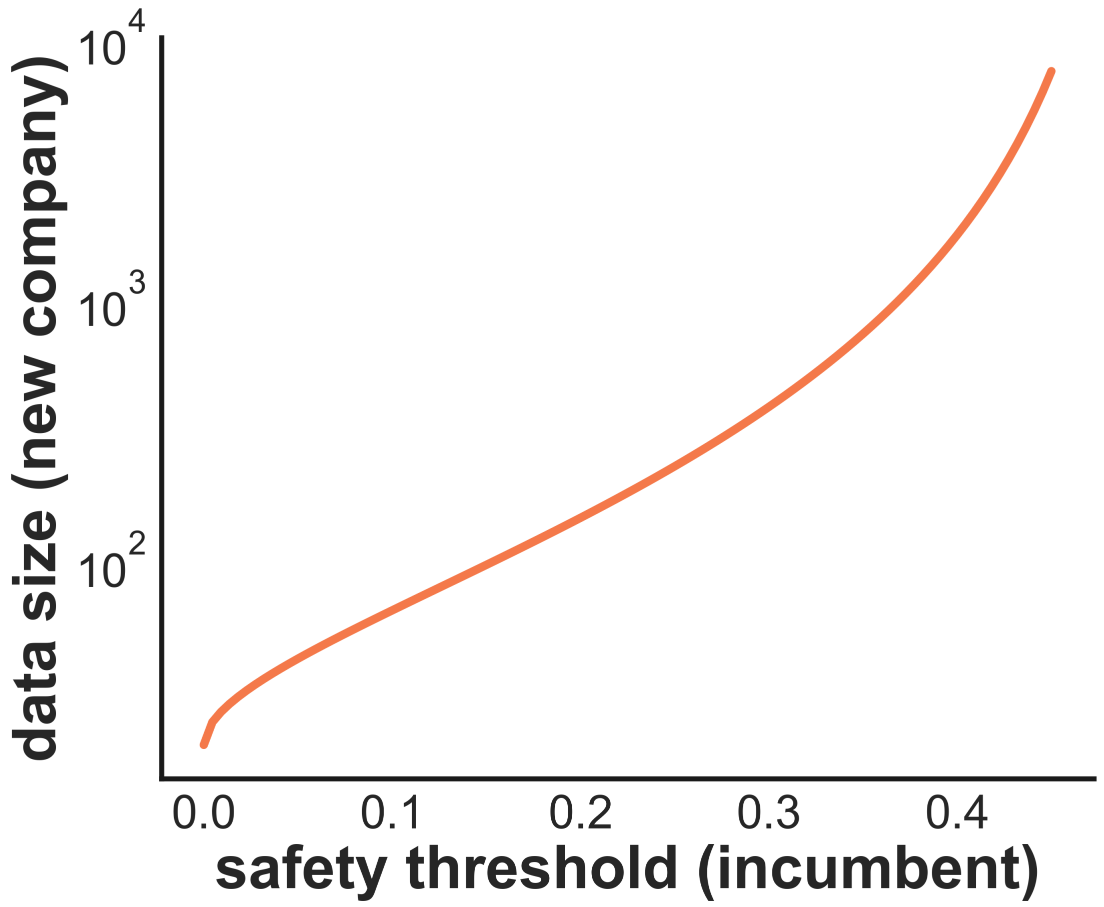
Warmup result

Setup: Incumbent has infinite data $N_I = \infty$*; new company faces no safety constraint* $\tau_E = \infty$.



Warmup result

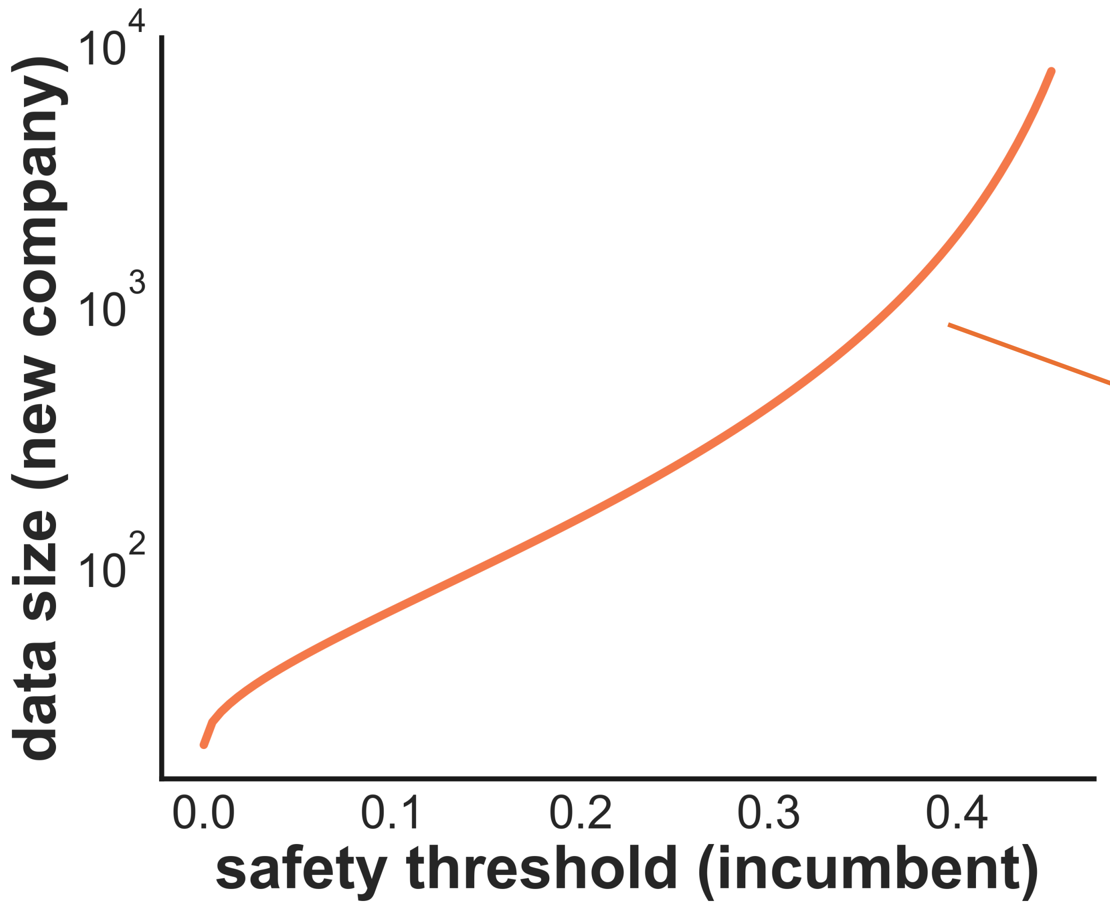
Setup: Incumbent has infinite data $N_I = \infty$*; new company faces no safety constraint* $\tau_E = \infty$.



Takeaway: New company can enter with finite data, even with an infinite-data incumbent.

Warmup result

Setup: Incumbent has infinite data $N_I = \infty$*; new company faces no safety constraint* $\tau_E = \infty$.



<u>Takeaway</u>: New company can enter with finite data, even with an infinite-data incumbent.

$$N_E^* = \Theta\left(\left(\sqrt{\boldsymbol{L}} - \sqrt{\min(\boldsymbol{L}, \boldsymbol{\tau_I})}\right)^{-\frac{2}{\nu}}\right).$$

L = Optimal infinite-data loss w/o safety

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Intuition for warmup

Key driver: The new company can train more unsafe models.

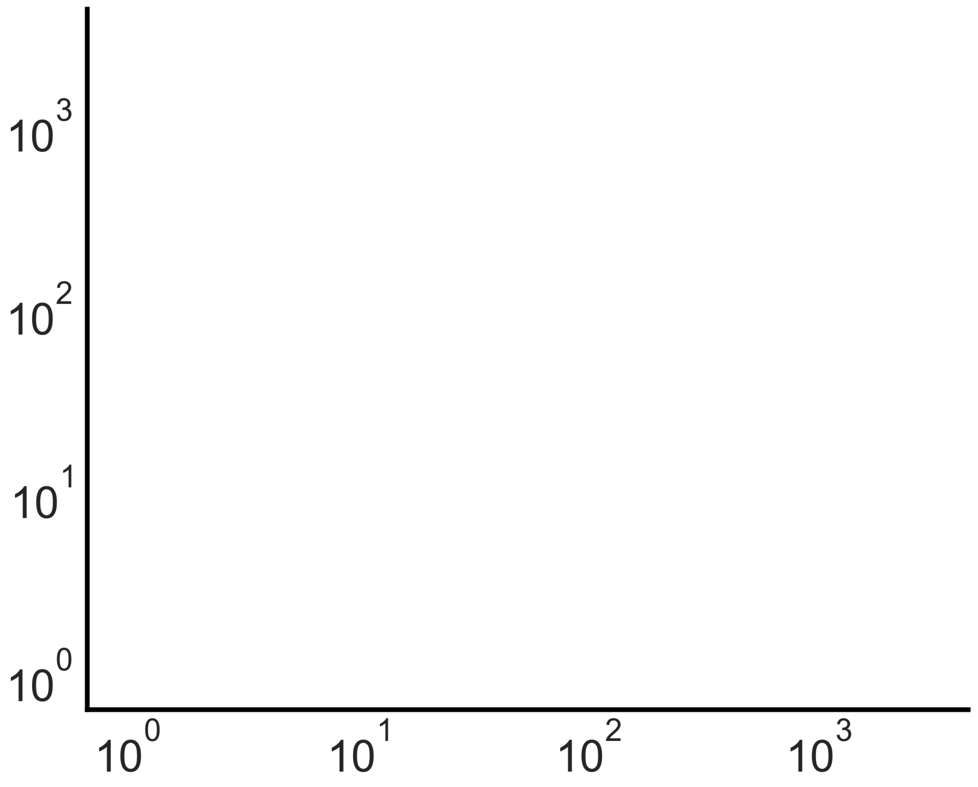
The incumbent must conservatively balance safety and performance, but the new company can focus more on performance.

$\Rightarrow$ The new company **curates its training data** to prioritize performance.

$\Rightarrow$ **The new company can enter the market with less data than the incumbent.**

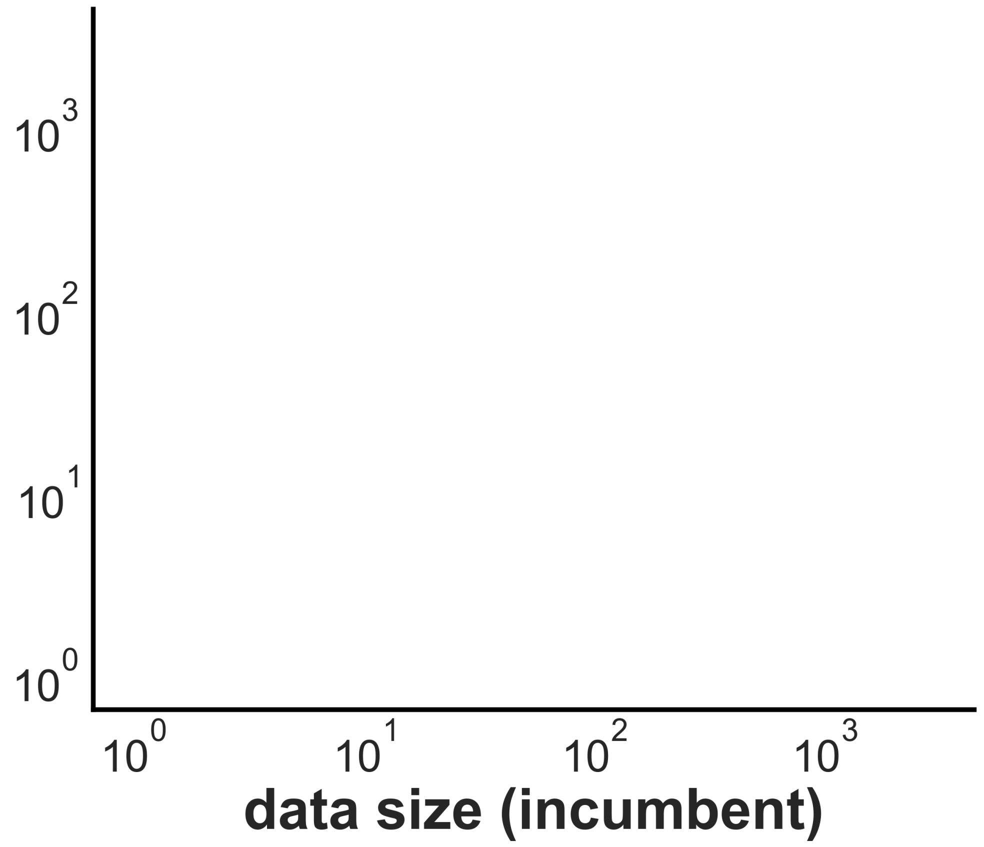
Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



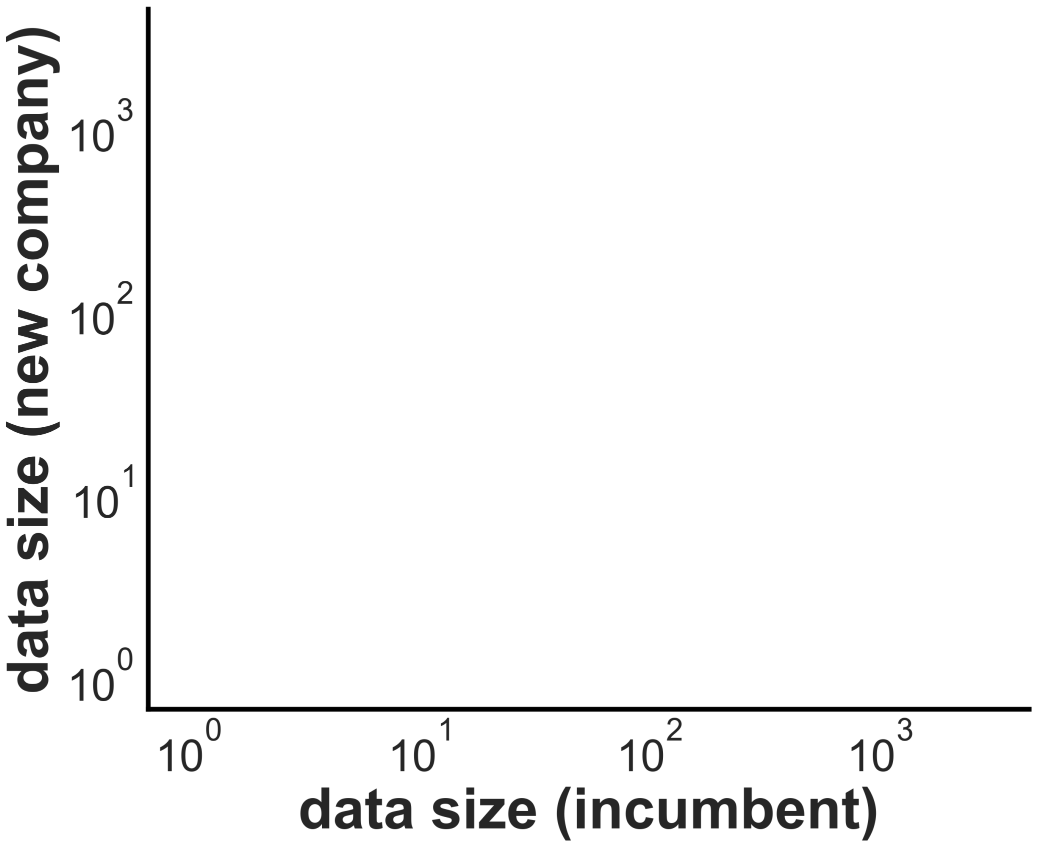
Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



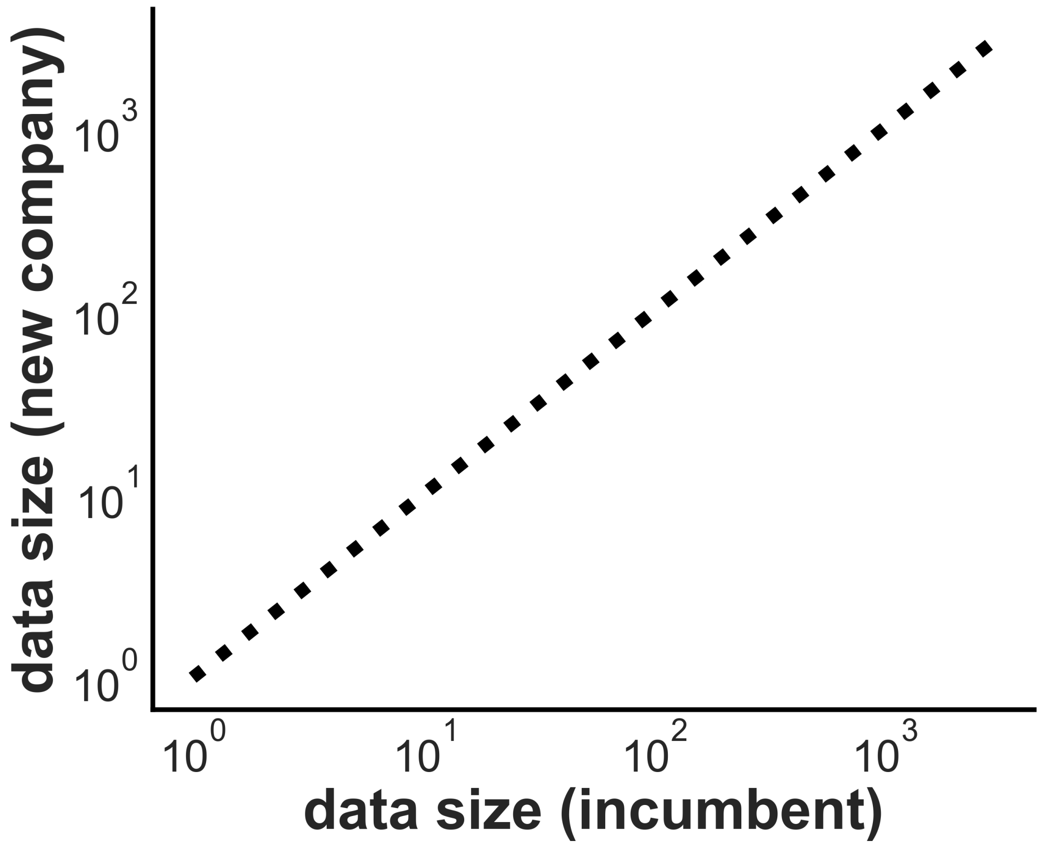
Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



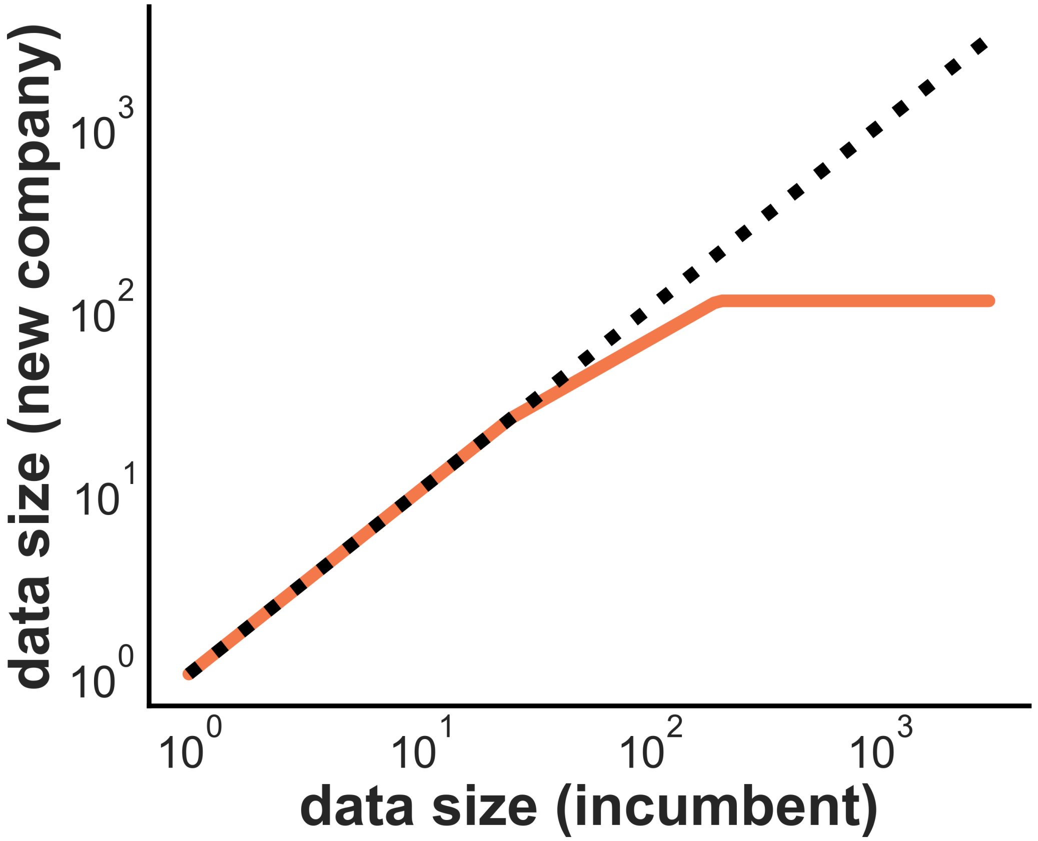
Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



Role of the incumbent's dataset size N_I

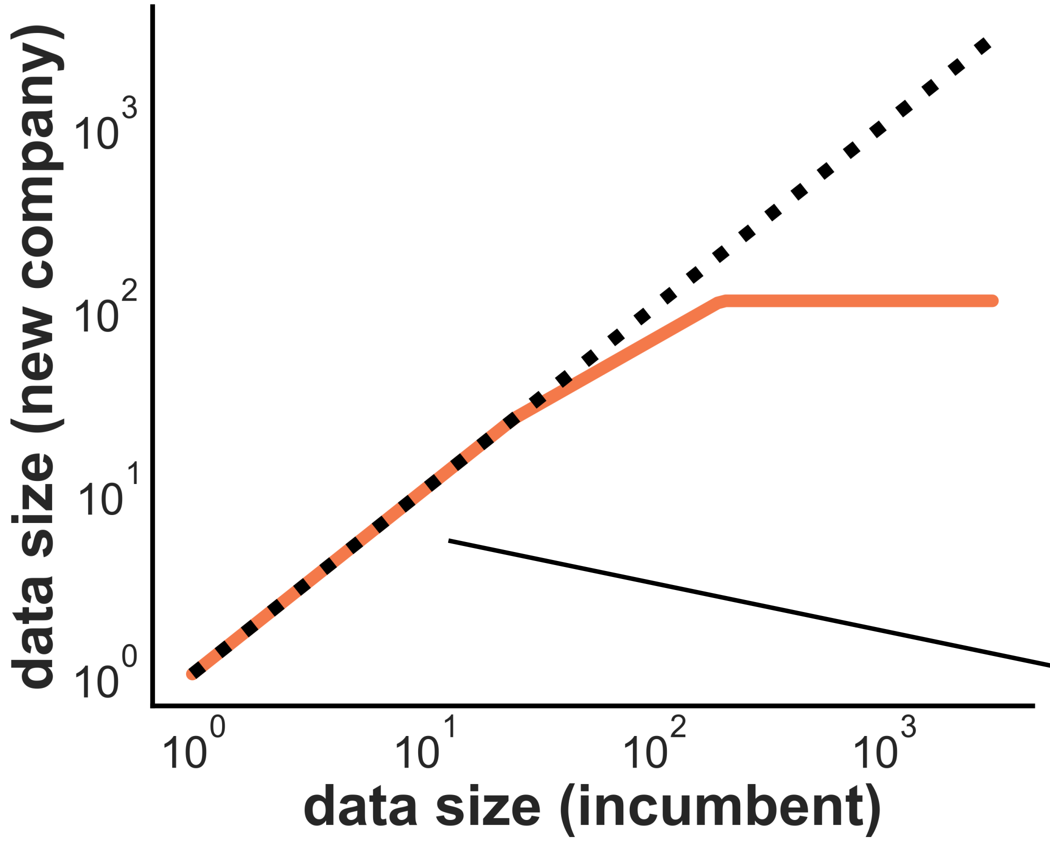
Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



Three regimes of behavior

Role of the incumbent's dataset size $\boldsymbol{N_I}$

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



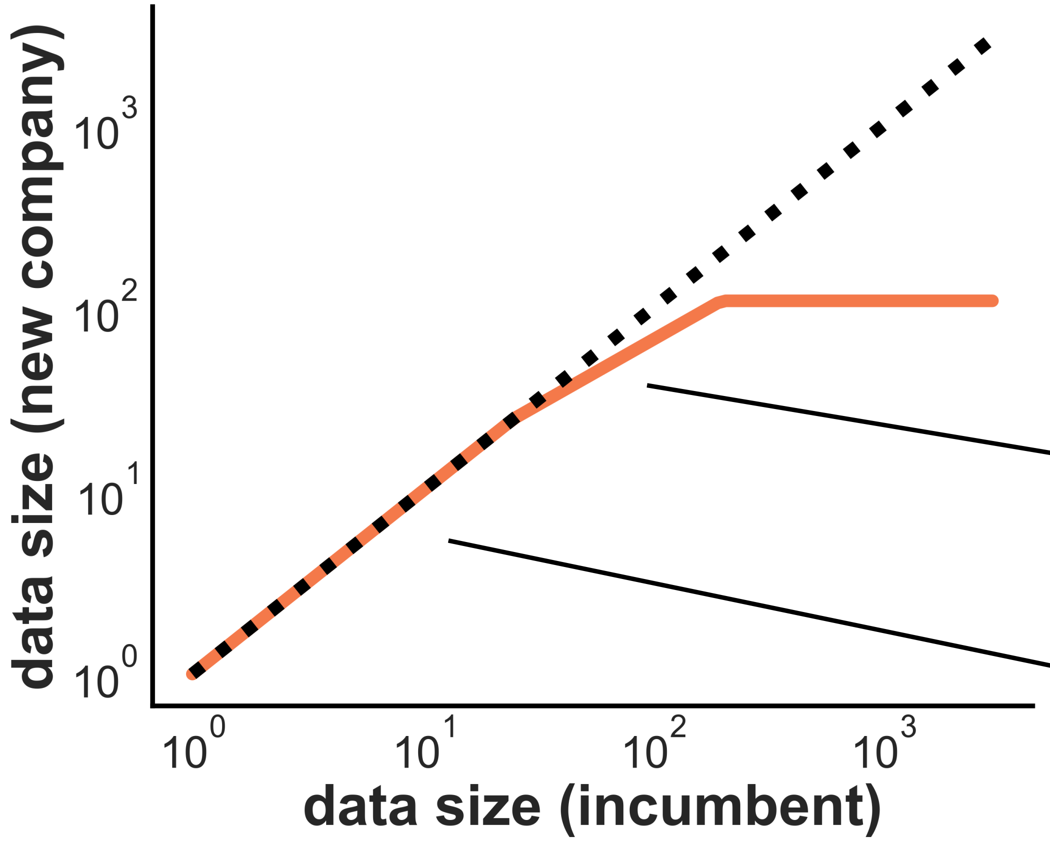
$$N_E^* = \Theta(N_I)$$

New company needs **as much data** as the incumbent

Three regimes of behavior

Role of the incumbent's dataset size $\boldsymbol{N_I}$

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



$$N_E^* = \Theta\left(N_I^{\frac{1}{\nu+1}}\right)$$

$$N_E^* = \Theta(N_I)$$

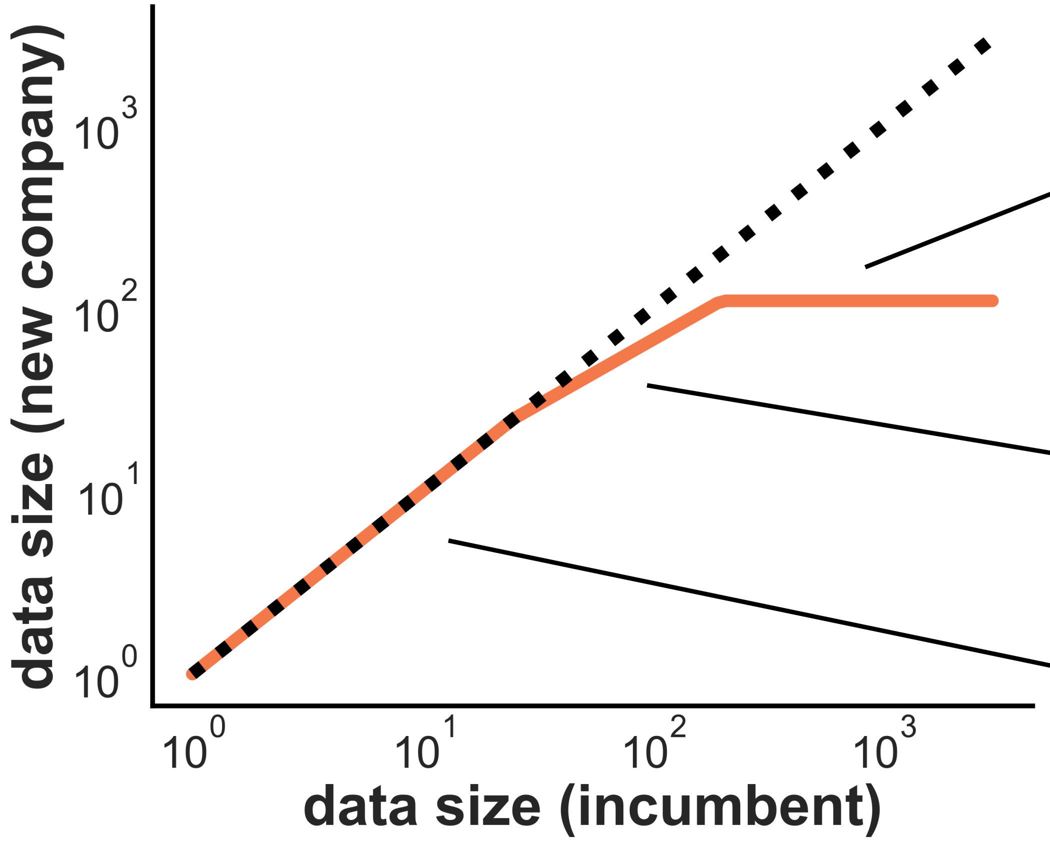
New company needs **as much data** as the incumbent

Three regimes of behavior

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



Three regimes of behavior

$$N_E^* = \Theta(1)$$

$$N_E^* = \Theta\left(N_I^{\frac{1}{\nu+1}}\right)$$

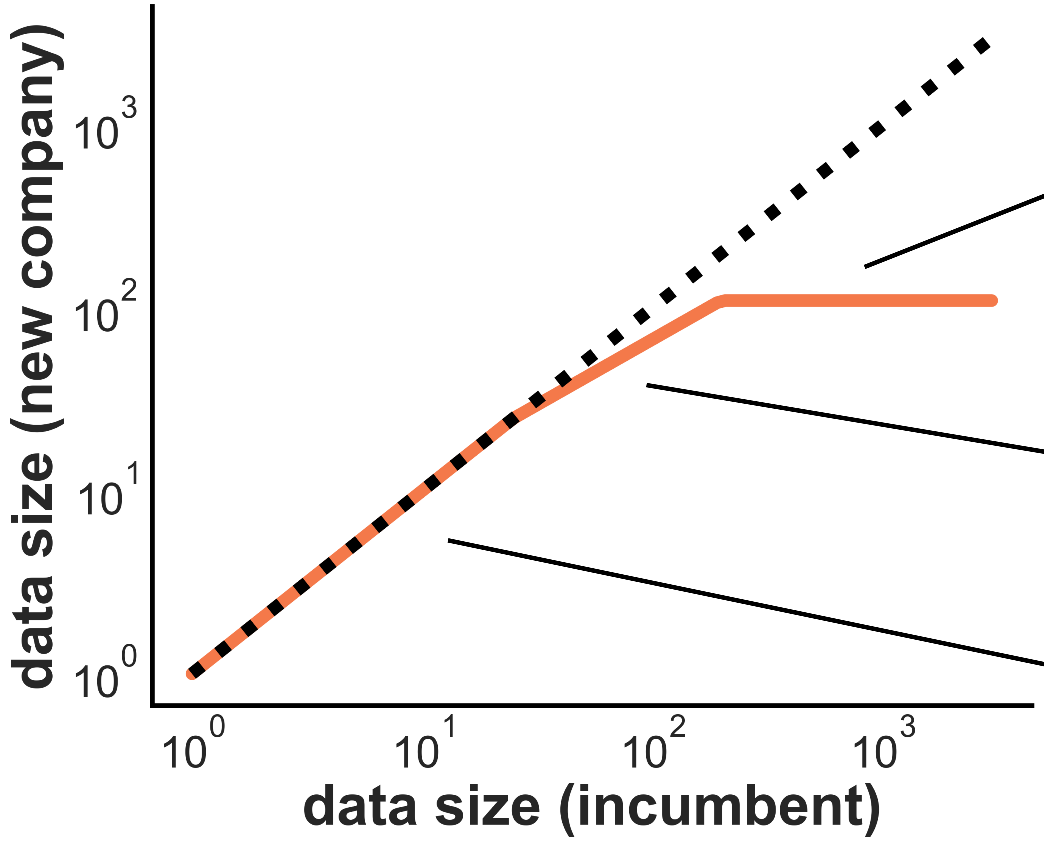
$$N_E^* = \Theta(N_I)$$

New company needs **as much data** as the incumbent

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Role of the incumbent's dataset size N_I

Setup: New company faces no safety constraint (i.e., $\tau_E = \infty$)



Three regimes of behavior

$$N_E^* = \Theta(1)$$

$$N_E^* = \Theta\left(N_I^{\frac{1}{\nu+1}}\right)$$

New company needs **less data** than the incumbent

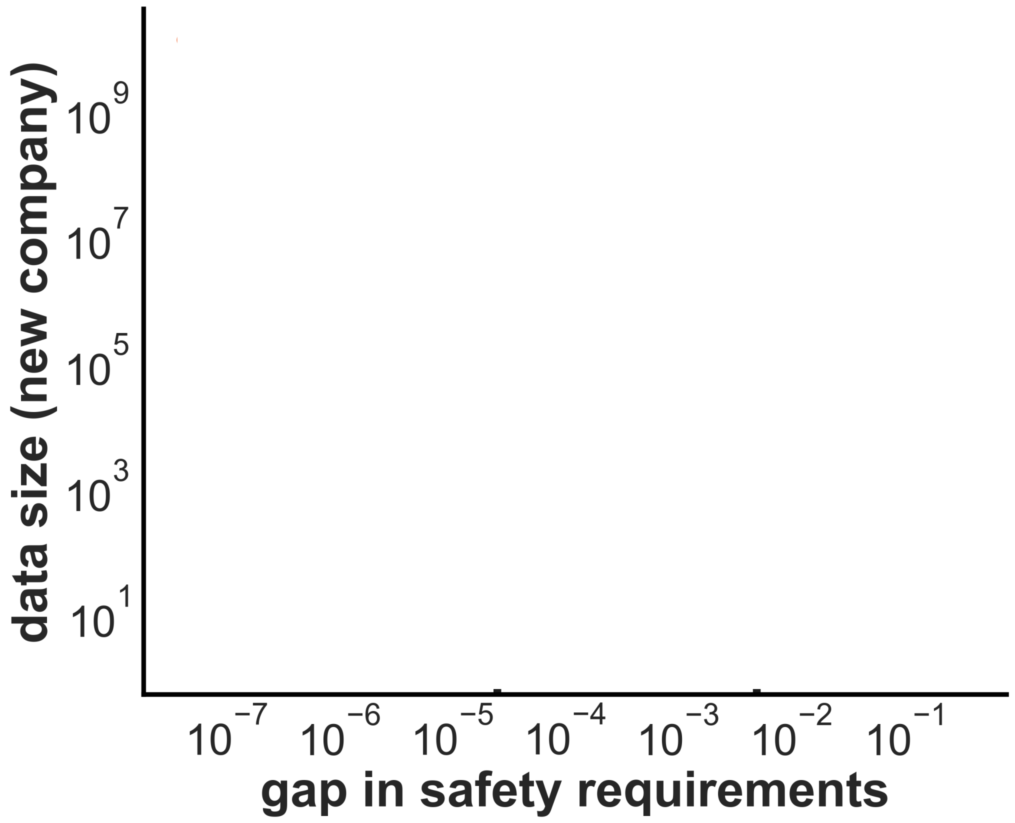
$$N_E^* = \Theta(N_I)$$

New company needs **as much data** as the incumbent

Data efficiency $\nu = \; min(2(1+\gamma), \gamma + \delta)$

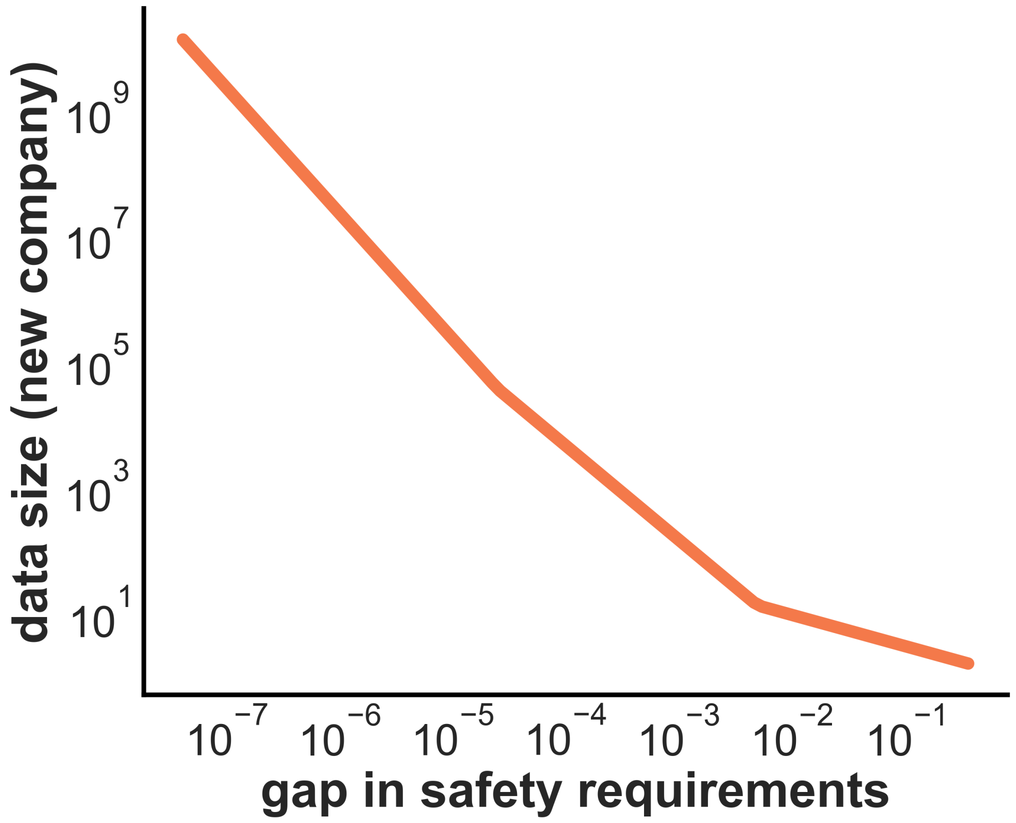
Role of the gap $\boldsymbol{D}$ in safety thresholds

Setup: Incumbent has infinite data ($N_I = \infty$), D = performance gap in infinite−data regime



Role of the gap D in safety thresholds

Setup: Incumbent has infinite data $(N_I = \infty)$, D = performance gap in infinite−data regime



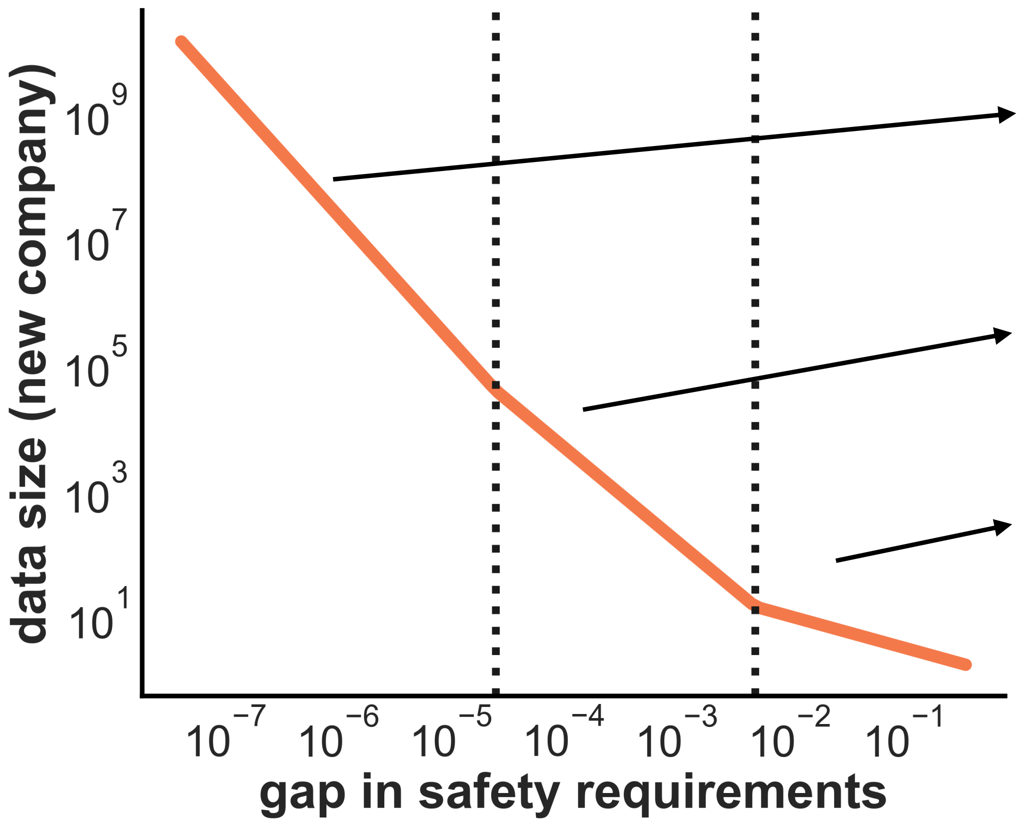
Takeaways:

- New company only needs **finite data**

Data efficiencies $\nu = min(2(1+\gamma), \gamma+\delta)$, $\nu' = \nu = min(1+\gamma, \gamma+\delta)$

Role of the gap D in safety thresholds

Setup: Incumbent has infinite data ($N_I = \infty$), D = performance gap in infinite–data regime



$$N_E^* = \Theta\left(D^{-\frac{\nu'+1}{\nu'}}\right)$$

$$N_E^* = \Theta\left(D^{-\frac{\nu+1}{\nu}}\right)$$

$$N_E^* = \Theta\left(D^{-\frac{1}{\nu}}\right)$$

Three regimes of behavior

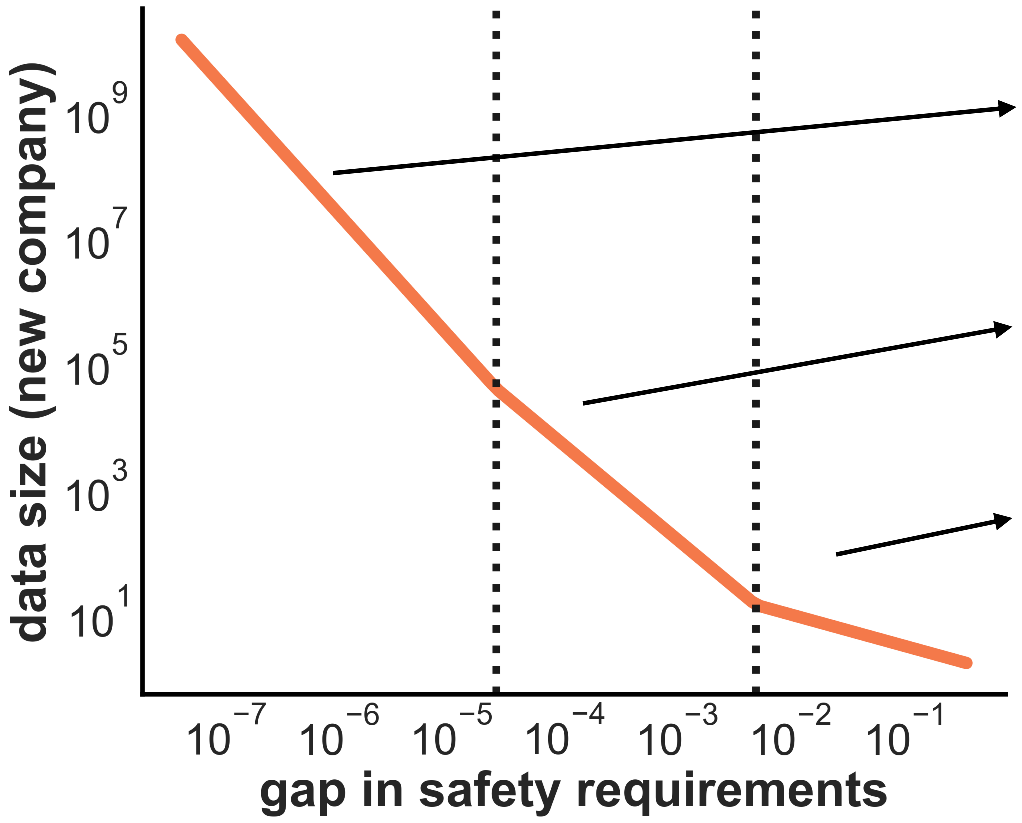
Takeaways:

- New company only needs **finite data**

Data efficiencies $\nu = min(2(1+\gamma), \gamma+\delta)$, $\nu' = \nu = min(1+\gamma, \gamma+\delta)$

Role of the gap D in safety thresholds

Setup: Incumbent has infinite data $(N_I = \infty)$, D = performance gap in infinite−data regime



$$N_E^* = \Theta\left(D^{-\frac{\nu'+1}{\nu'}}\right)$$

$$N_E^* = \Theta\left(D^{-\frac{\nu+1}{\nu}}\right)$$

$$N_E^* = \Theta\left(D^{-\frac{1}{\nu}}\right)$$

Three regimes of behavior

Takeaways:

- New company only needs **finite data**
- New company must **scale up data** faster when safety thresholds are more even.

Data efficiencies $\nu = \min(2(1+\gamma), \gamma+\delta)$, $\nu' = \nu = \min(1+\gamma, \gamma+\delta)$

Implication: when does scrutiny of safety reduce data-driven barriers to entry?

Implication: when does scrutiny of safety reduce data-driven barriers to entry?

<u>Key parameters</u>:

- Incumbent's dataset size
- Unevenness of safety scrutiny (i.e., gap between safety thresholds)

Implication: when does scrutiny of safety reduce data-driven barriers to entry?

Key parameters:

- Incumbent’s dataset size
- Unevenness of safety scrutiny (i.e., gap between safety thresholds)

Our findings:

- Uneven scrutiny of safety reduces data-driven barriers to entry ***only when the incumbent’s dataset size is sufficiently large***.

Implication: when does scrutiny of safety reduce data-driven barriers to entry?

Key parameters:

- Incumbent’s dataset size
- Unevenness of safety scrutiny (i.e., gap between safety thresholds)

Our findings:

- Uneven scrutiny of safety reduces data-driven barriers to entry ***only when the incumbent’s dataset size is sufficiently large***.
- If the scrutiny is more even, then the data-driven barriers to entry not only increase but also ***scale up at a faster rate***.

Outline for the talk

1. Background
2. Our model
3. Our results
4. **Technical ideas**

Summary

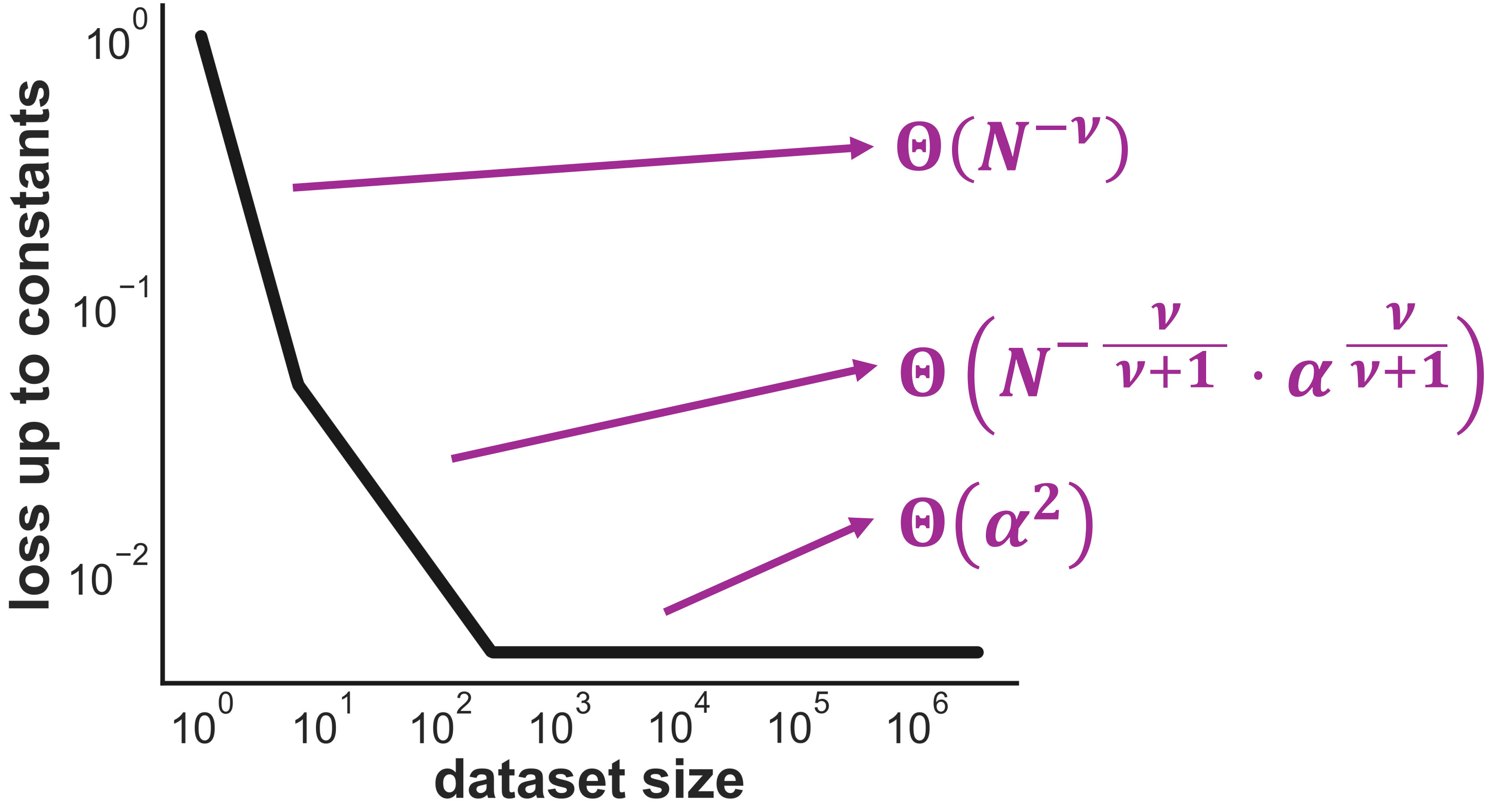
Technical tool: derive multi-objective data scaling laws

<u>**Result**</u>: We characterize how the **loss** of **optimally regularized** ridge regression in terms of the **training data size** N and **data mixture level** α.

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Technical tool: derive multi-objective data scaling laws

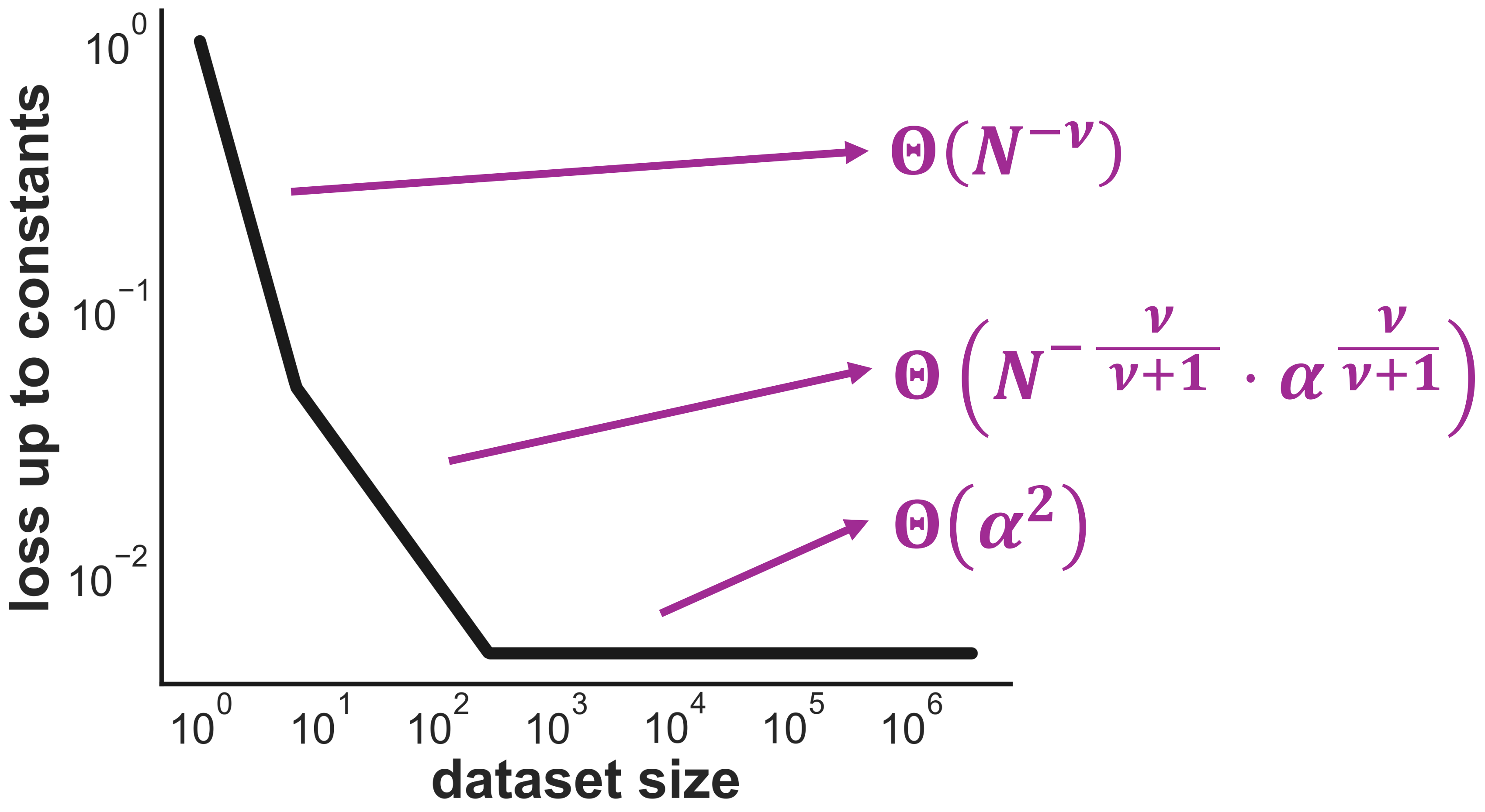
<u>Result</u>: We characterize how the **loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture level α**.



Data efficiency $\nu = \min(2(1+\gamma), \gamma+\delta)$

Technical tool: derive multi-objective data scaling laws

<u>Result</u>: We characterize how the **loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture level α**.

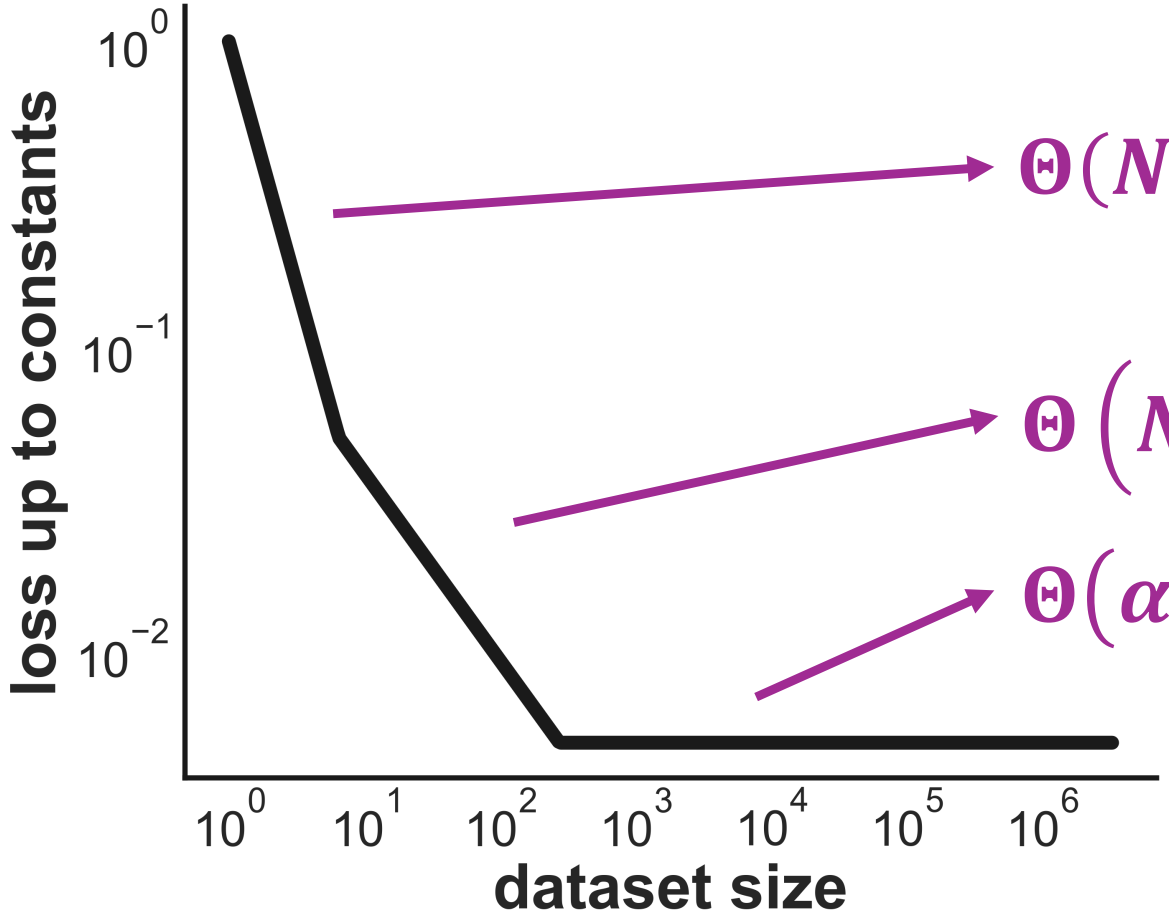


Key insight: *multi-objective* data efficiency decreases as the data size N increases

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Technical tool: derive multi-objective data scaling laws

Result: We characterize how the **loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture level α**.



$\Theta(N^{-\nu})$

$\Theta\left(N^{-\frac{\nu}{\nu+1}} \cdot \alpha^{\frac{\nu}{\nu+1}}\right)$

$\Theta(\alpha^2)$

Key insight: *multi-objective* data efficiency decreases as the data size N increases

In comparison: *single-objective* data efficiency is constant in N

e.g., Cui et al., '21, Wei et al., '22, Bach '23

Data efficiency $\nu = min(2(1+\gamma), \gamma + \delta)$

Technical tool: derive multi-objective data scaling laws

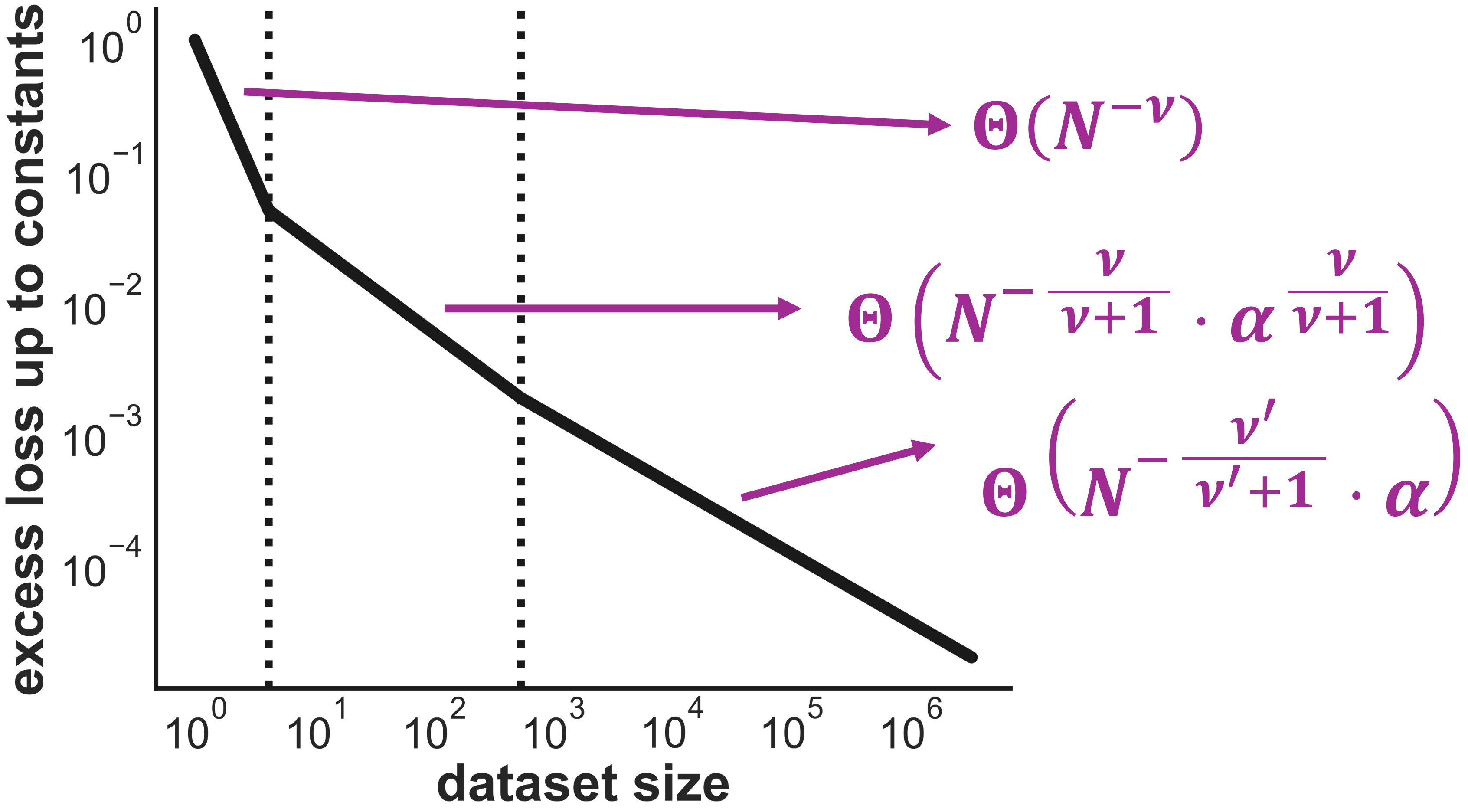
<u>**Result**</u>: We characterize how the **excess loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture level α**.

Excess loss subtracts out the infinite-data performance with data mixture α.

Data efficiencies $\nu = \min(2(1+\gamma), \gamma+\delta), \nu' = \nu = \min(1+\gamma, \gamma+\delta)$

Technical tool: derive multi-objective data scaling laws

<u>**Result**</u>: We characterize how the **excess loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture level α**.

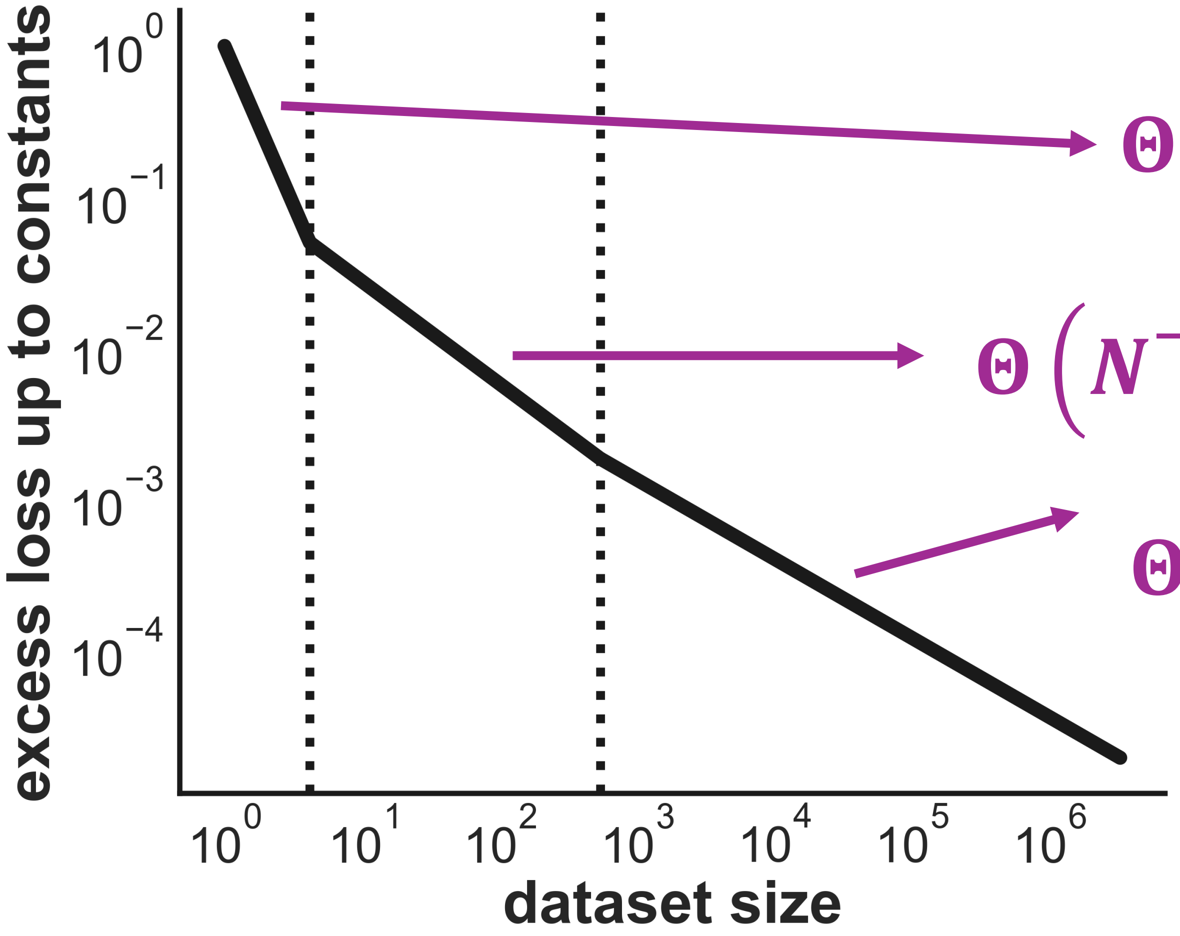


Excess loss subtracts out the infinite-data performance with data mixture α.

Data efficiencies $\nu = \min(2(1+\gamma), \gamma+\delta)$, $v' = \nu = \min(1+\gamma, \gamma+\delta)$

Technical tool: derive multi-objective data scaling laws

Result: We characterize how the **excess loss** of **optimally regularized** ridge regression in terms of the **training data size N** and **data mixture α** labelled with β_2.



$$\Theta(N^{-\nu})$$

$$\Theta\left(N^{-\frac{\nu}{\nu+1}} \cdot \alpha^{\frac{\nu}{\nu+1}}\right)$$

$$\Theta\left(N^{-\frac{\nu'}{\nu'+1}} \cdot \alpha\right)$$

Key insight: *multi-objective* data efficiency decreases as the data size N increases

Excess loss subtracts out the infinite-data performance with data mixture α.

Data efficiencies $\nu = \ min(2(1+\gamma), \gamma+\delta)\,, v' = \nu = \ min(1+\gamma, \gamma+\delta)$

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x \sim D}\left[\left(\langle \hat{\beta}, x \rangle - \langle \beta_1, x \rangle\right)^2\right]$ of ridge regression

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}\,,x\rangle-\langle\beta_1\,,x\rangle\right)^2\right]$ of ridge regression

<u>Challenge</u>: expectation over randomness over the N training data points

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}\,,x\rangle - \langle\beta_1\,,x\rangle\right)^2\right]$ of ridge regression

Challenge: expectation over randomness over the N training data points

Key idea: Use random matrix theory to characterize the loss

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}\,,x\rangle - \langle\beta_1\,,x\rangle\right)^2\right]$ of ridge regression

<u>Challenge</u>: expectation over randomness over the N training data points

<u>Key idea</u>: Use random matrix theory to characterize the loss

- Derive a **deterministic equivalent** using the Marčenko-Pastur law

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}, x\rangle - \langle\beta_1, x\rangle\right)^2\right]$ of ridge regression

<u>Challenge</u>: expectation over randomness over the N training data points

<u>Key idea</u>: Use random matrix theory to characterize the loss

- Derive a **deterministic equivalent** using the Marčenko-Pastur law
- Characterize loss under the **power law decay assumptions**

Proof ideas for deriving scaling laws

Need **tight bounds** on the **loss** $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}\,,x\rangle - \langle\beta_1\,,x\rangle\right)^2\right]$ of ridge regression

Challenge: expectation over randomness over the N training data points

Key idea: Use random matrix theory to characterize the loss

- Derive a **deterministic equivalent** using the Marčenko-Pastur law
- Characterize loss under the **power law decay assumptions**
- Analyze scaling behavior under **optimal regularization**

Bounds on the loss for multi-objective regression

Setup: training data size N, data mixture level α, regularization level λ

Lemma (Informal): The loss $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}, x\rangle - \langle\beta_1, x\rangle\right)^2\right]$ is approximately equal to:

$$\underbrace{\max\left(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}\right)}_{\text{Finite data error}} + \alpha \cdot Q \cdot \underbrace{\frac{\min\left(N,\ \lambda^{-\frac{1}{1+\gamma}}\right)}{N}}_{\text{Overfitting error}} + \alpha^2 \cdot \underbrace{Q}_{\text{Optimal infinite data loss}} + \alpha \cdot Q \underbrace{\max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})}_{\text{Extra (Mixture finite data error)}}$$

Data efficiencies $\nu = min(2(1+\gamma), \gamma+\delta)$, $\nu' = \nu = min(1+\gamma, \gamma+\delta)$

Bounds on the loss for multi-objective regression

Setup: training data size N, data mixture level α, regularization level λ

<u>Lemma (Informal):</u> The loss $\mathbf{E}_{x\sim D}\left[\left(\langle\hat{\beta}, x\rangle - \langle\beta_1, x\rangle\right)^2\right]$ is approximately equal to:

$$\max\left(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}\right) + \alpha \cdot Q \cdot \frac{\min\left(N, \boldsymbol{\lambda}^{-\frac{1}{1+\gamma}}\right)}{N} + \alpha^2 \cdot Q + \alpha \cdot Q \max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})$$

Finite data error

Overfitting error

Optimal infinite data loss

Extra (Mixture finite data error)

<u>Implication</u>: must regularize to avoid overfitting, but this reduces data efficiency

Data efficiencies $\nu = min(2(1+\gamma), \gamma+\delta), \nu' = \nu = min(1+\gamma, \gamma+\delta)$

Summary

We studied data-driven barriers to market entry for companies training LLMs.

Summary

We studied data-driven barriers to market entry for companies training LLMs.

<u>This work</u>: a technical framework to quantify how much data a new company needs to enter the market

Summary

We studied data-driven barriers to market entry for companies training LLMs.

<u>This work</u>: a technical framework to quantify how much data a new company needs to enter the market

- *Model*: We modelled these markets within a multi-objective learning framework.
- *Technical tool*: multi-objective data scaling laws

Summary

We studied data-driven barriers to market entry for companies training LLMs.

<u>This work</u>: a technical framework to quantify how much data a new company needs to enter the market

- *Model*: We modelled these markets within a multi-objective learning framework.
- *Technical tool*: multi-objective data scaling laws

***Key finding*: *Scrutiny of safety* often---but not always---enables *new LLM companies* to enter the market with less data than incumbents**

Broader direction: *how do **details of the ML pipeline** shape the **market of companies training ML models**?*

Training data

Evaluation metrics

Pretraining & finetuning