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

Meena Jagadeesan (UC Berkeley)

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 LLMs.

Key feature: companies must balance multiple objectives to survive in the market

GPT-4

Gemini

Grok

This work: how hard is it for new LLM companies to enter the market?

Outline for the talk

1. **Background**

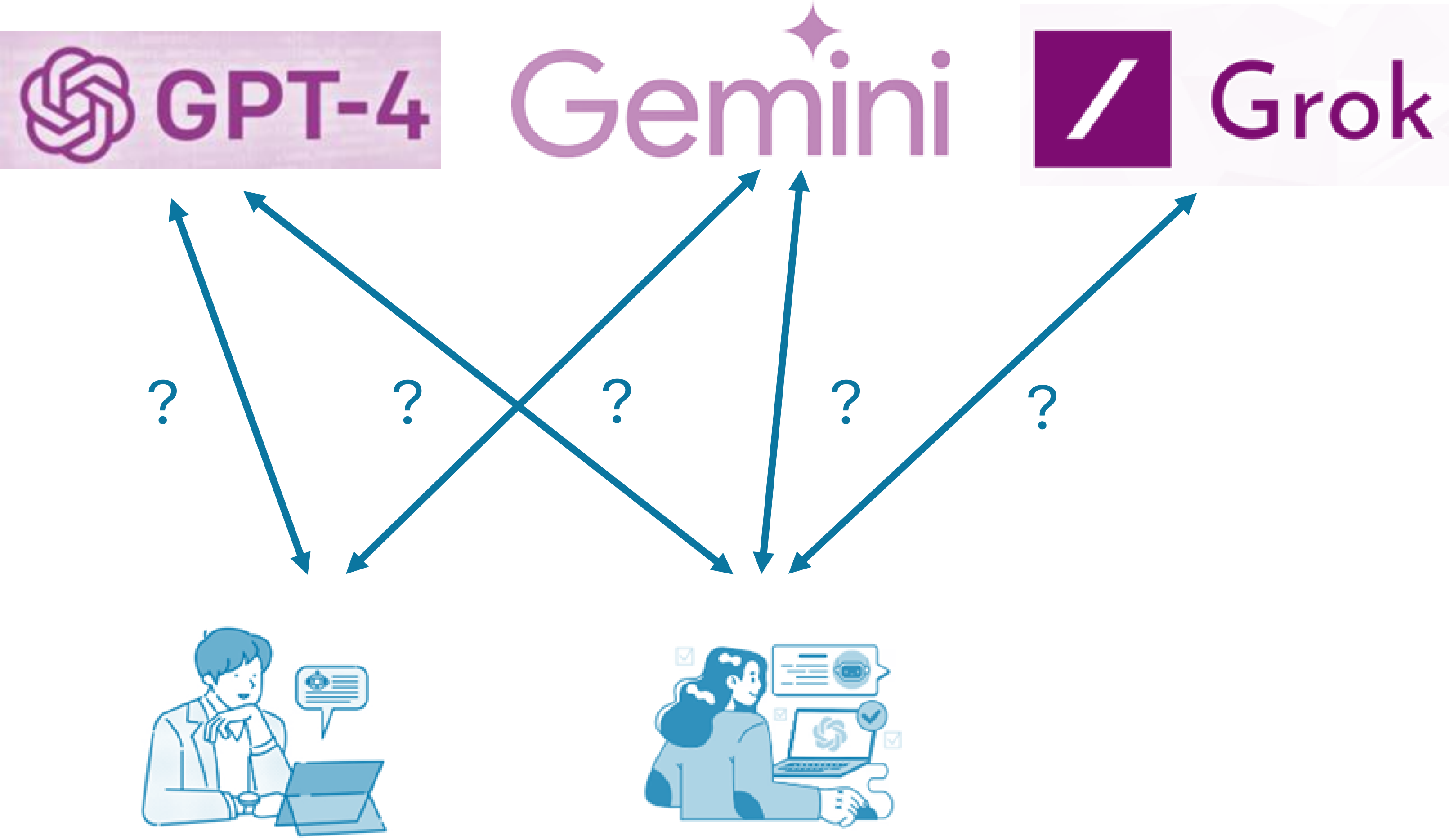
2. Our model

3. Our results

An emerging market of companies that train LLMs

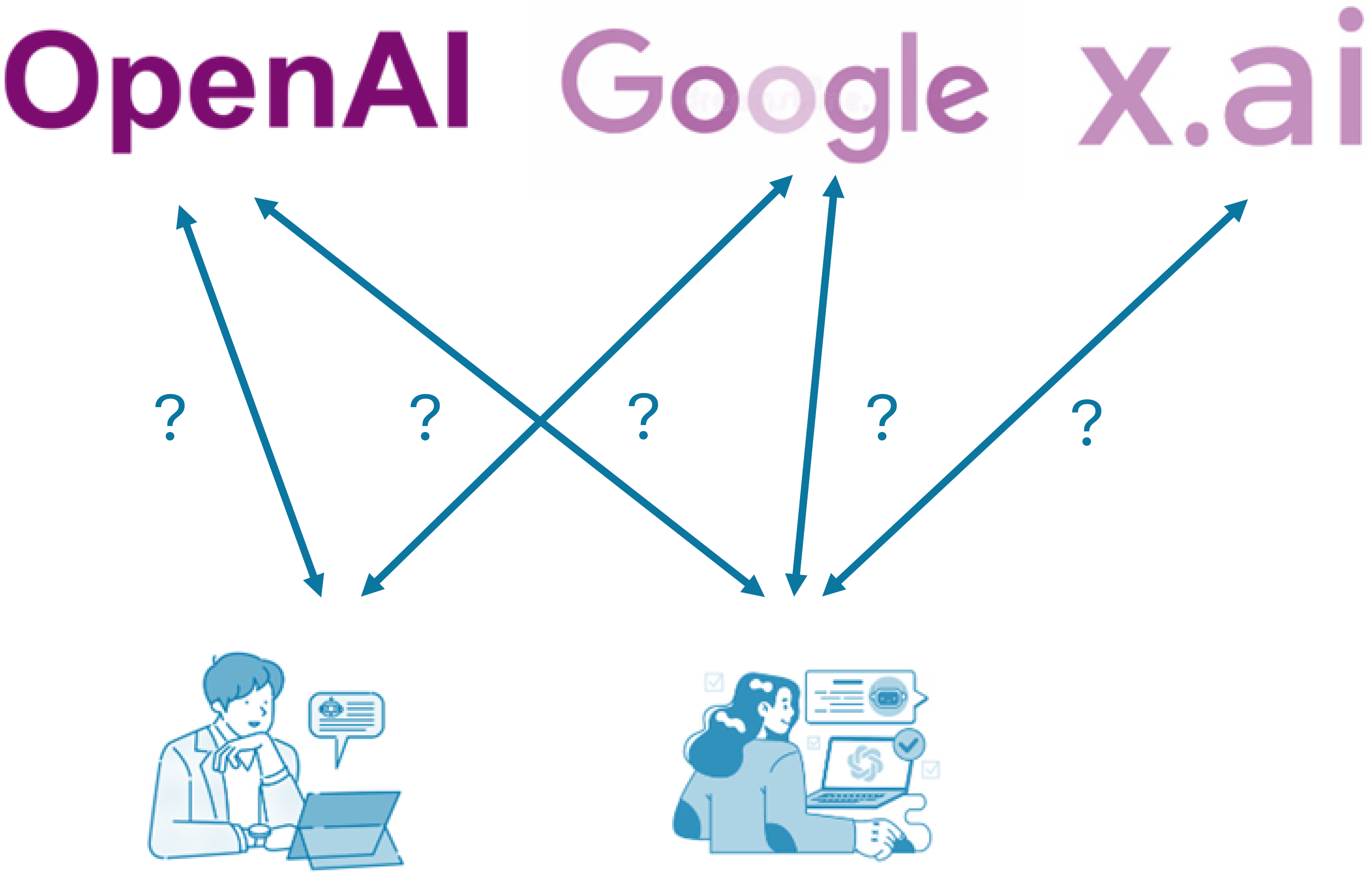
GPT-4 Gemini Grok

An emerging market of companies that train LLMs



Users choose between LLMs.

An emerging market of companies that train LLMs



Companies training these LLMs compete for users.

Users choose between LLMs.

Barriers to market entry

Policymakers have raised concerns about a small # of companies dominating the market.

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

Barriers to market entry

Policymakers have raised concerns about a small # of companies dominating the market.

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 due to data accumulation

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

Barriers to market entry

Policymakers have raised concerns about a small # of companies dominating the market.

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 due to data accumulation

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

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

Barriers to market entry

Policymakers have raised concerns about a small # of companies dominating the market.

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

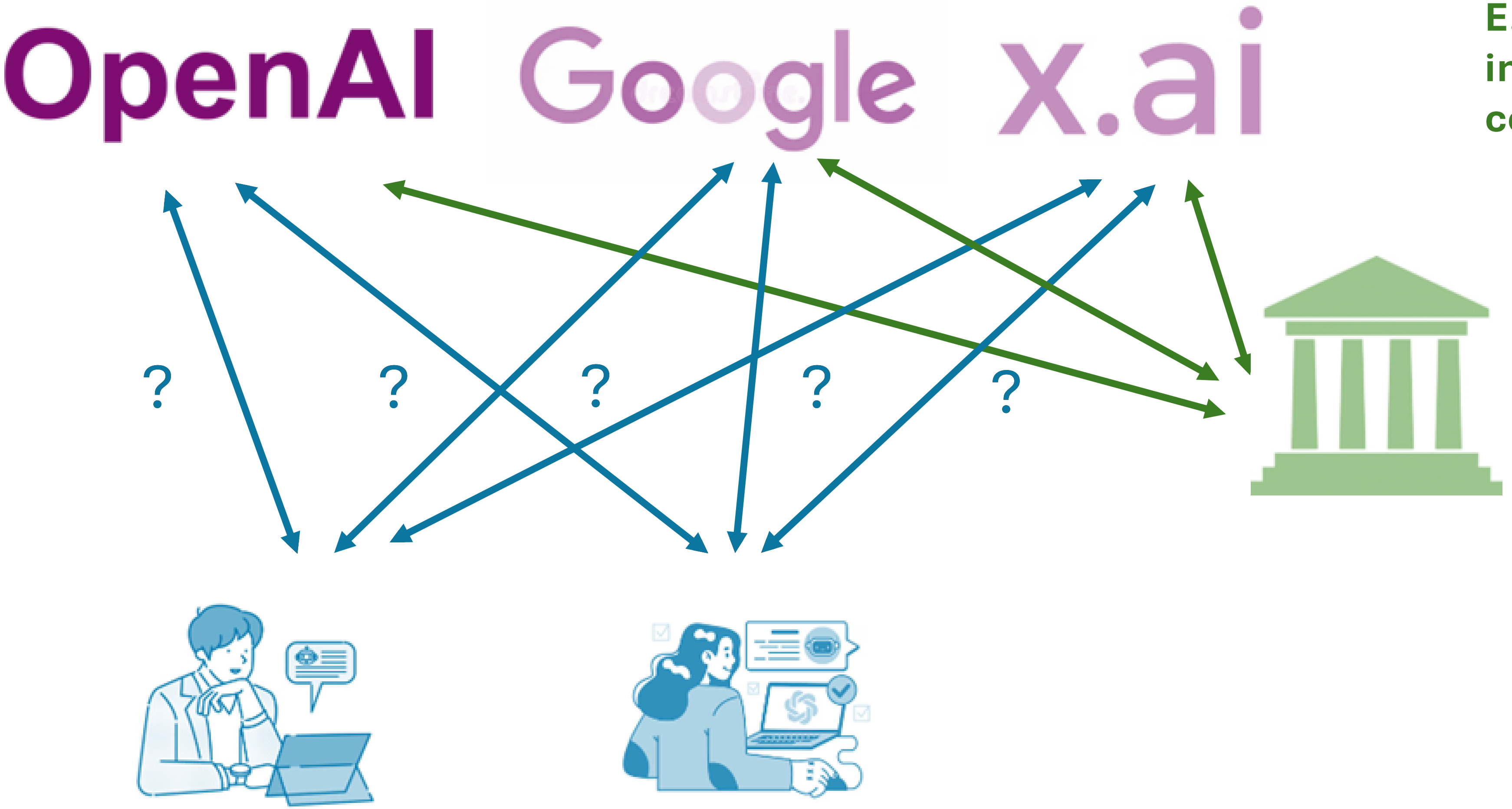
Typical intuition: New LLM companies face large barriers to entry due to data accumulation

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

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

Reality: companies face pressure to consider objectives beyond performance.

An emerging market of companies that train LLMs



E.g., releasing dangerous information, generating offensive content, etc.

Regulators / society scrutinize **safety violations** of deployed LLMs.

that train LLMs

Scrutiny from regulators:

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

Bills

SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.

Session Year: 2023-2024 **House:** Senate

E.g., releasing dangerous information, generating offensive content, etc.

Regulators / society scrutinize **safety violations** of deployed LLMs.

that train LLMs

Scrutiny from regulators:

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

Bills

SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.

Session Year: 2023-2024 **House:** Senate

Scrutiny from society:

TECH • ARTIFICIAL INTELLIGENCE

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

E.g., releasing dangerous information, generating offensive content, etc.

Regulators / society scrutinize **safety violations** of deployed LLMs.

that train LLMs

Scrutiny from regulators:

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

Bills

SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.

Session Year: 2023-2024 **House:** Senate

Scrutiny from society:

TECH • ARTIFICIAL INTELLIGENCE

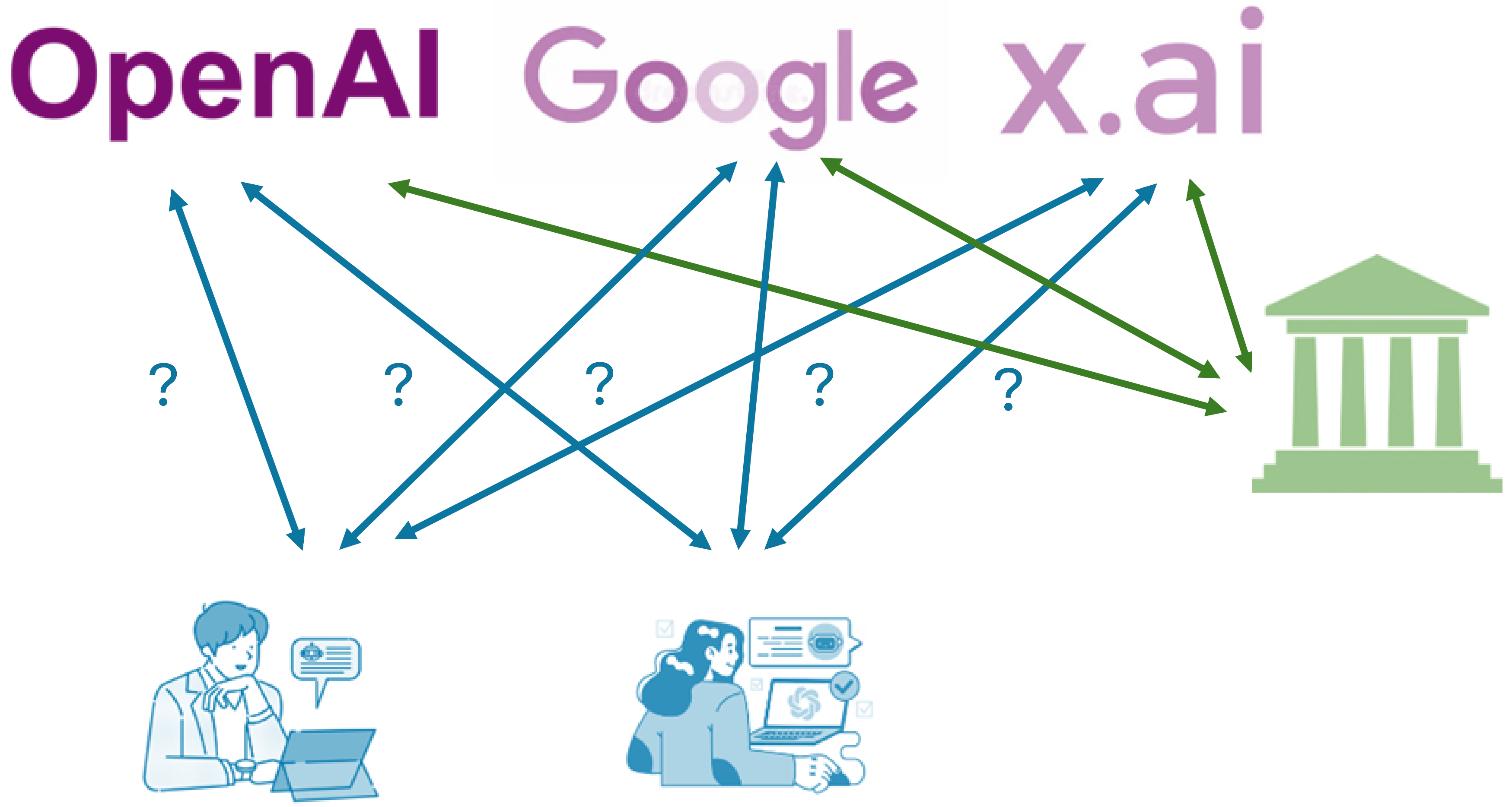
The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter

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

E.g., releasing dangerous information, generating offensive content, etc.

Regulators / society scrutinize **safety violations** of deployed LLMs.

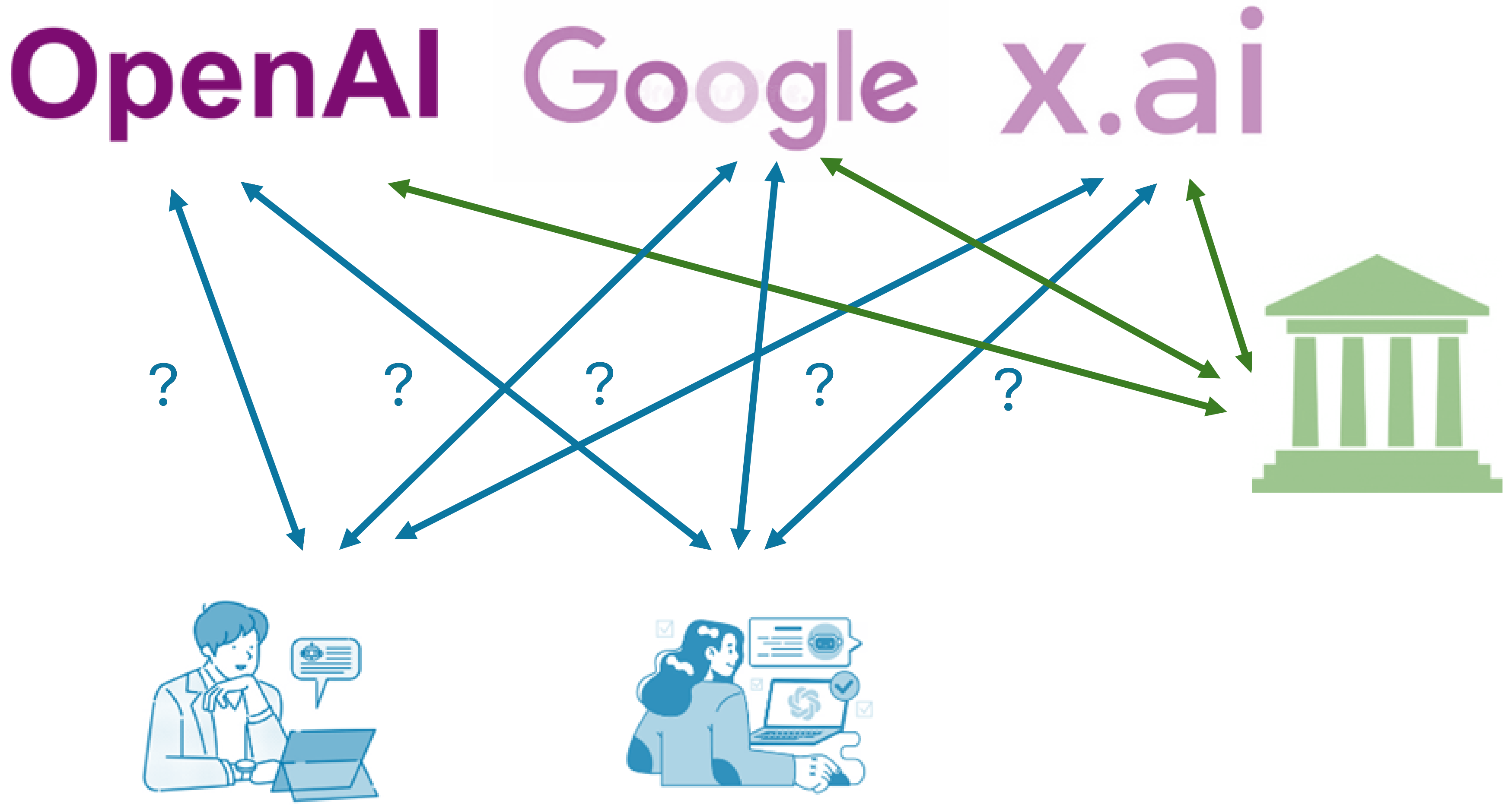
An emerging market of companies that train LLMs



Regulators / society scrutinize **safety violations** of deployed LLMs.

Users choose the **unscrutinized** LLM with best performance.

An emerging market of companies that train LLMs



Companies strategically train LLMs to perform well and avoid scrutiny.

Regulators / society scrutinize **safety violations** of deployed LLMs.

Users choose the **unscrutinized** LLM with best performance.

Main question

In markets of companies training LLMs, how does regulatory and societal scrutiny affect barriers to market entry?

Overview of our contributions

Main finding: New companies can enter the market with much less data than incumbents.

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

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: barriers to market entry under multi-objective learning

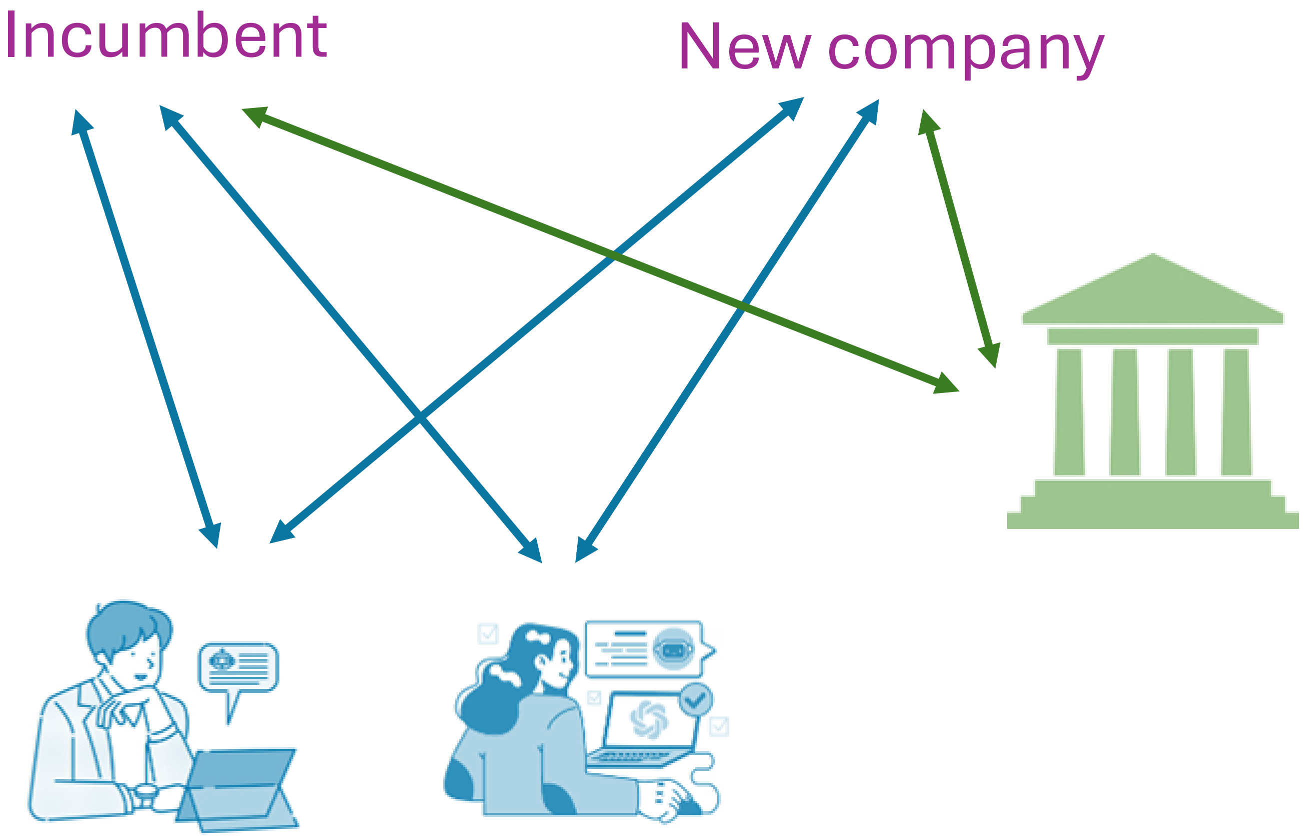
Outline for the talk

1. Background

2. **Our model**

3. Our results

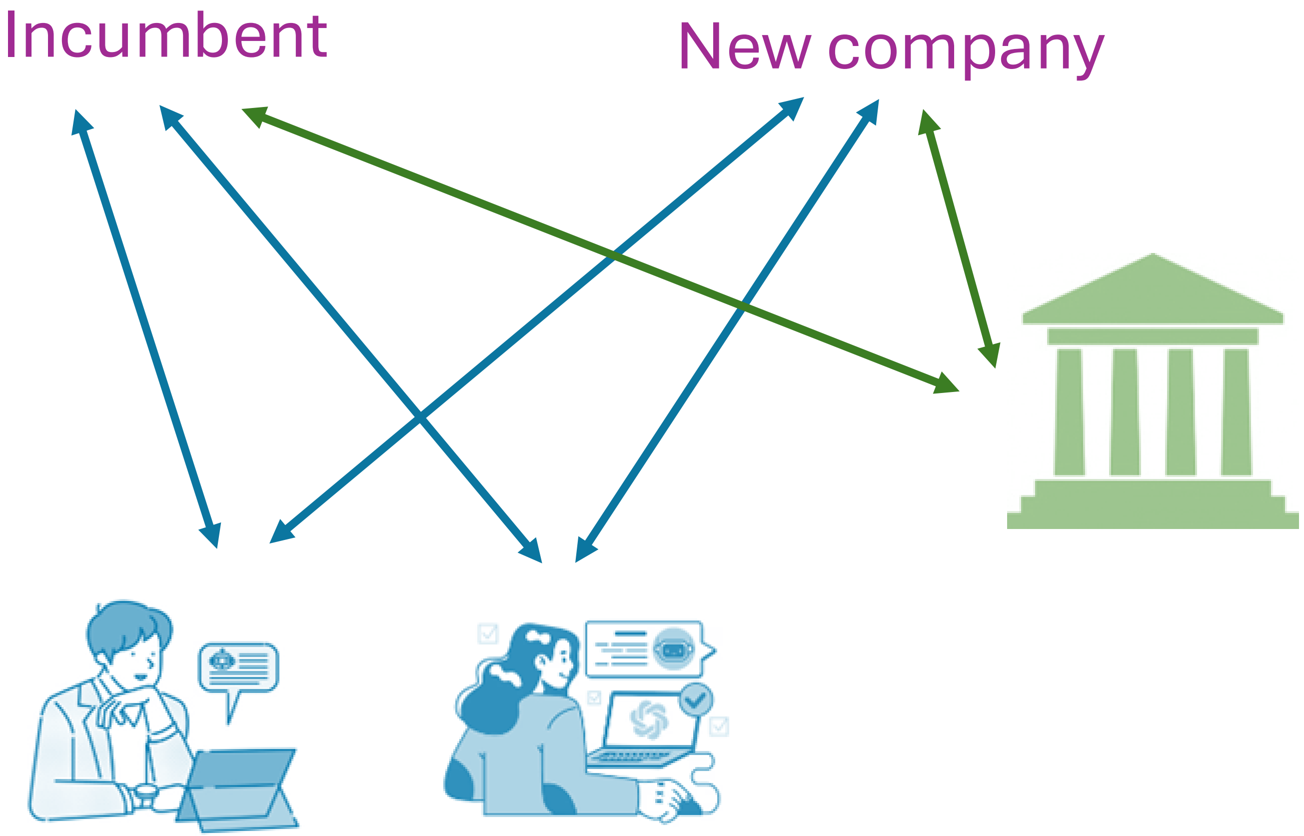
Model: markets of companies training LLMs



Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

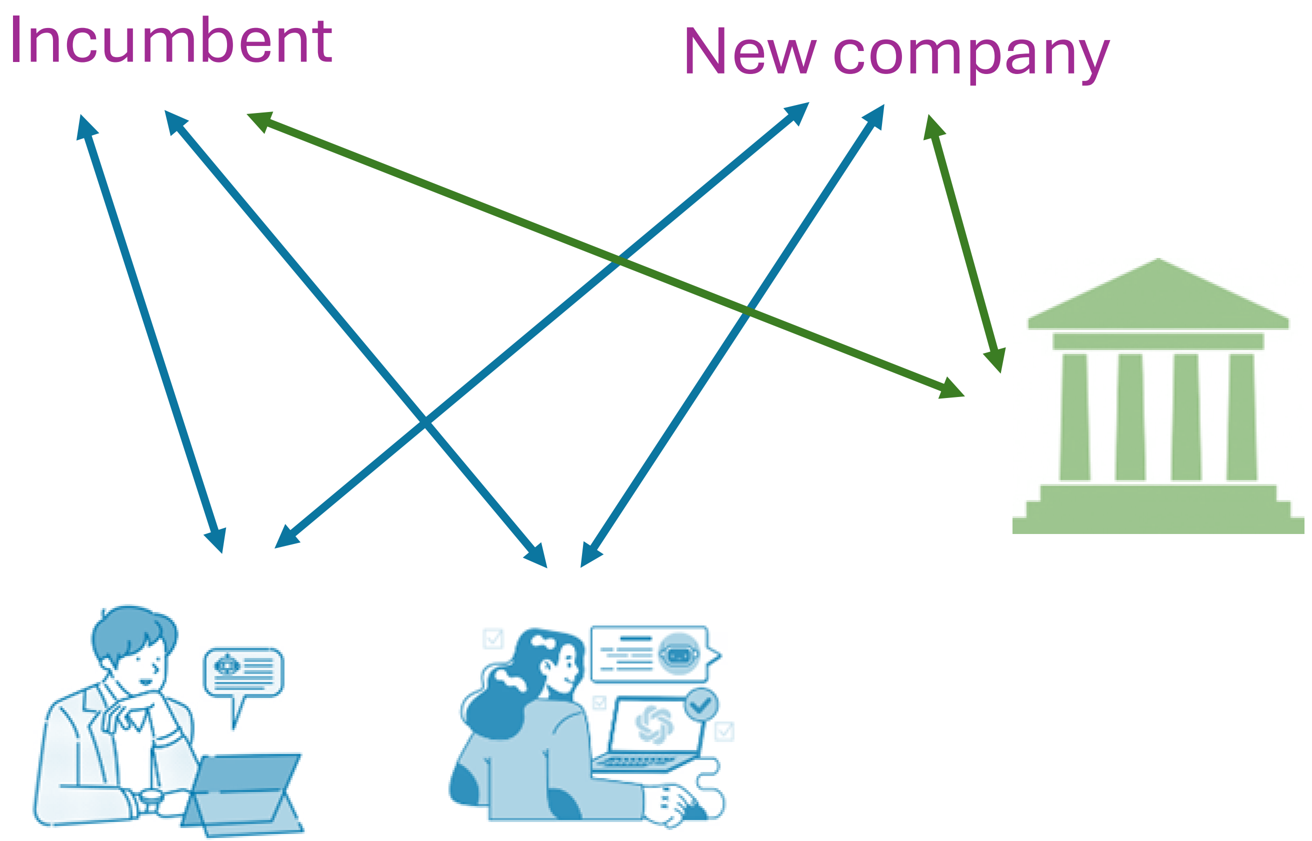
- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,



Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,

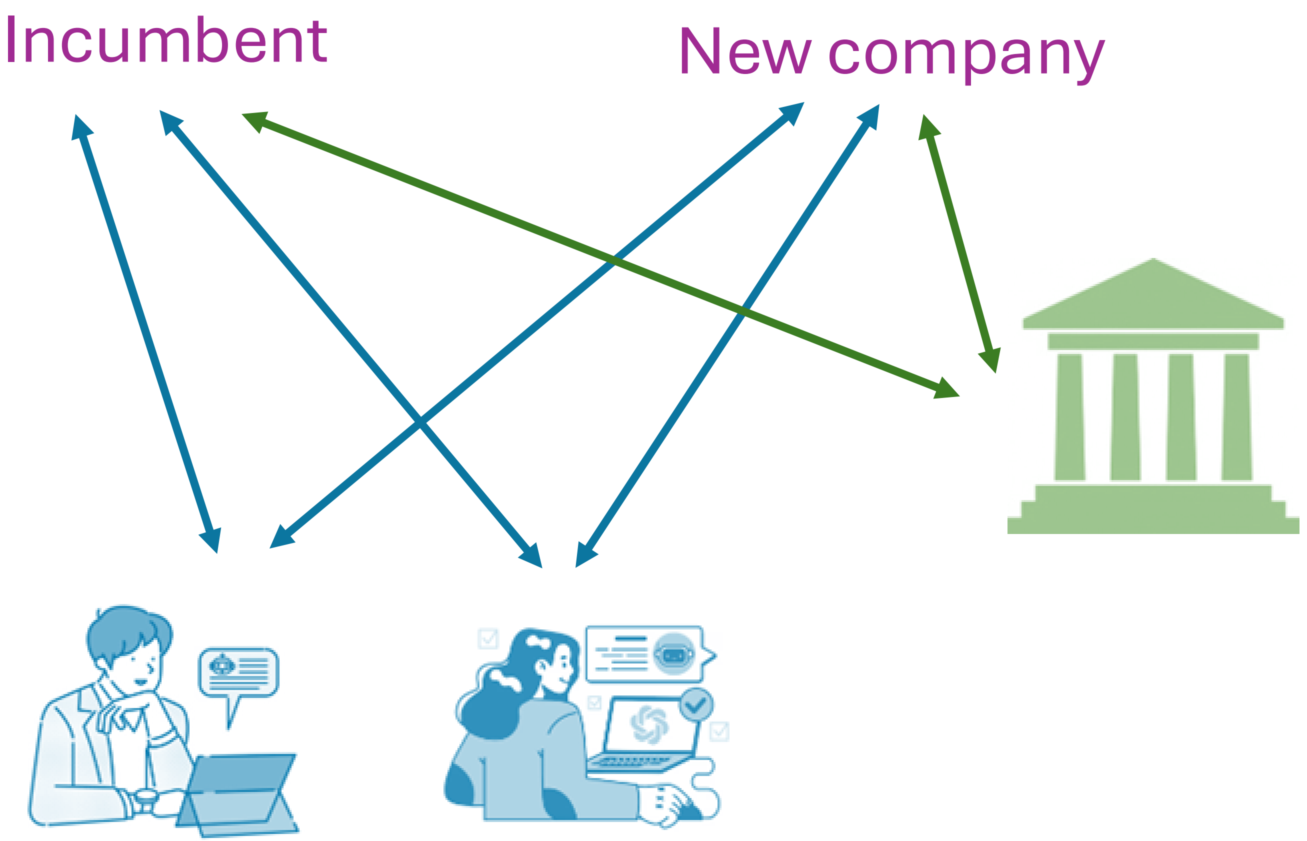


Each company strategically labels its training data with β_1 vs. β_2, and trains a model.

Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,



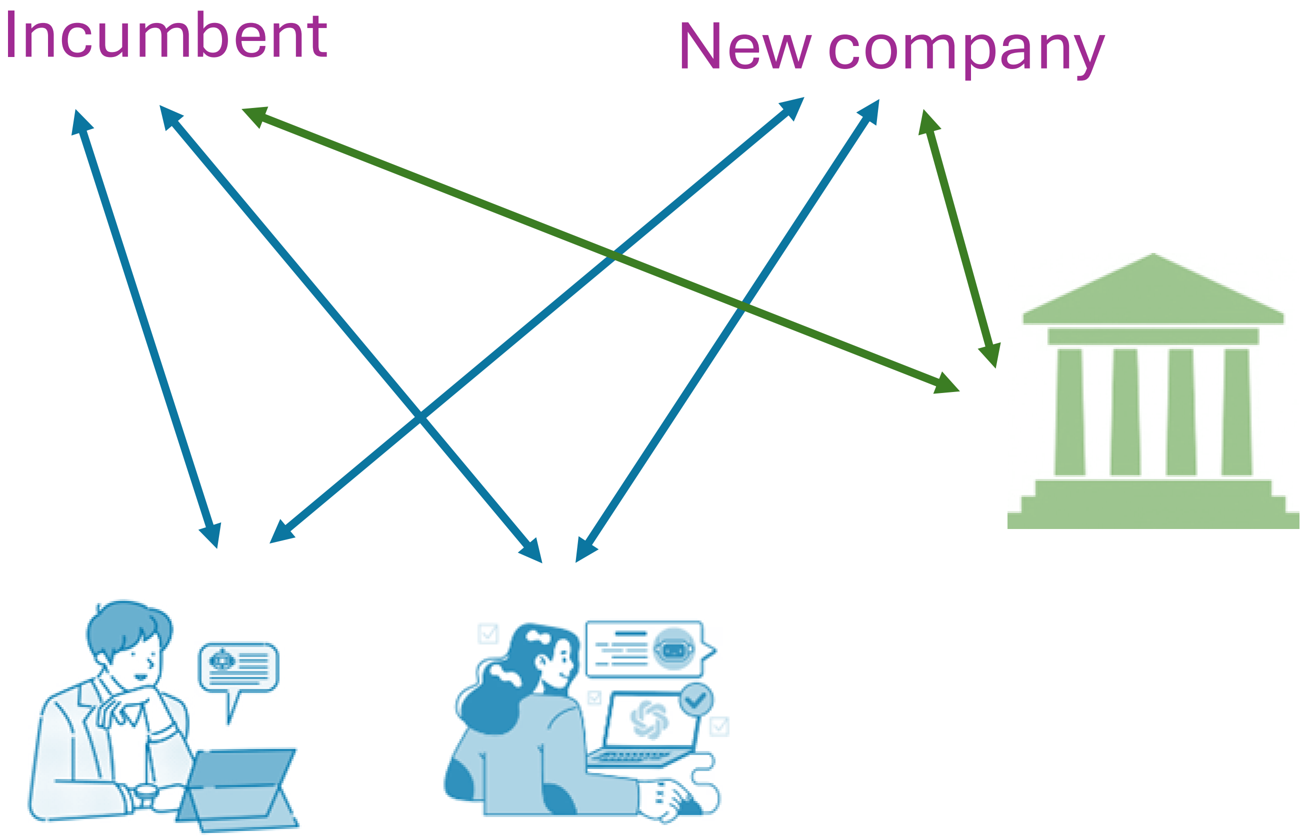
Each company strategically labels its training data with β_1 vs. β_2, and trains a model.

A company is scrutinized if its safety loss w.r.t. β_2 exceeds a given threshold.

Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,



Each company strategically labels its training data with β_1 vs. β_2, and trains a model.

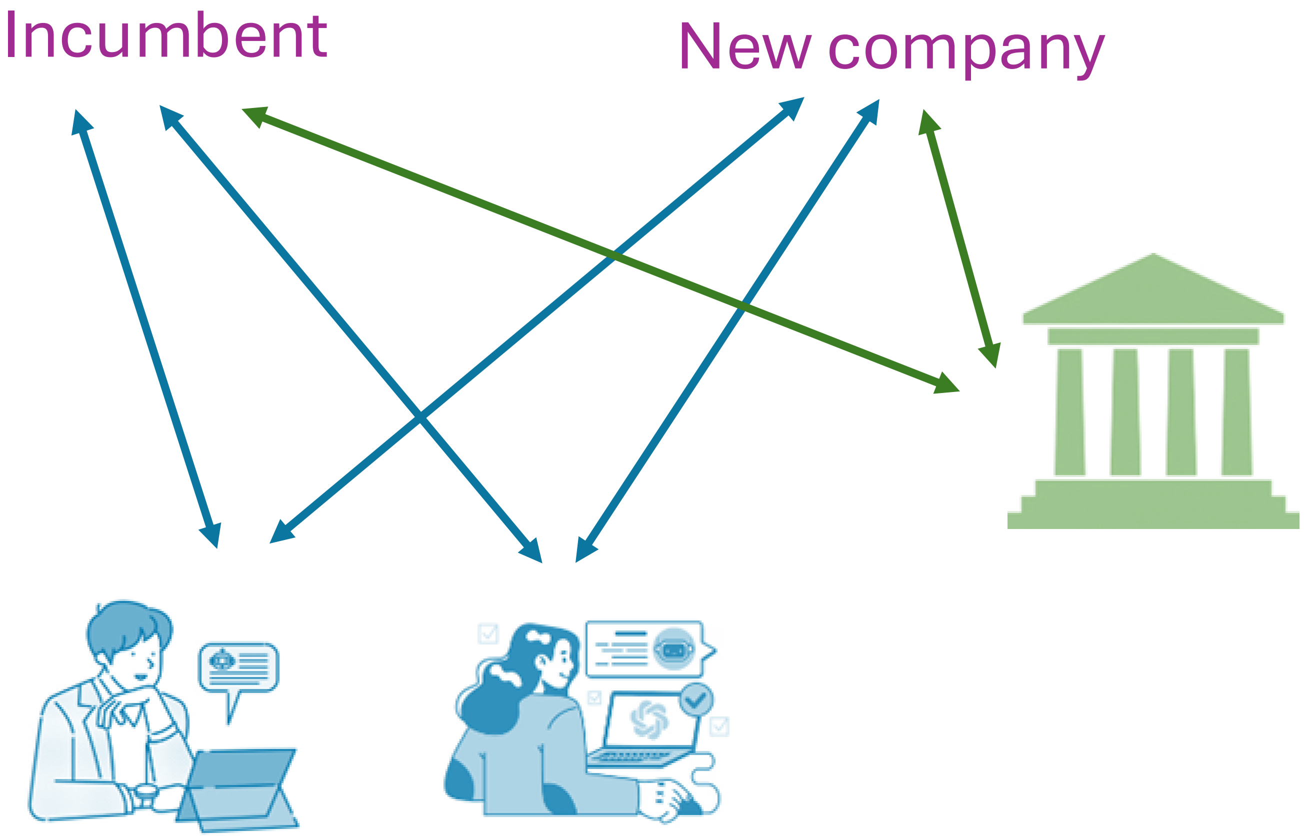
A company is scrutinized if its safety loss w.r.t. β_2 exceeds a given threshold.

- **Incumbent faces a stricter threshold.**

Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,



Each company strategically labels its training data with β_1 vs. β_2, and trains a model.

A company is scrutinized if its safety loss w.r.t. β_2 exceeds a given threshold.

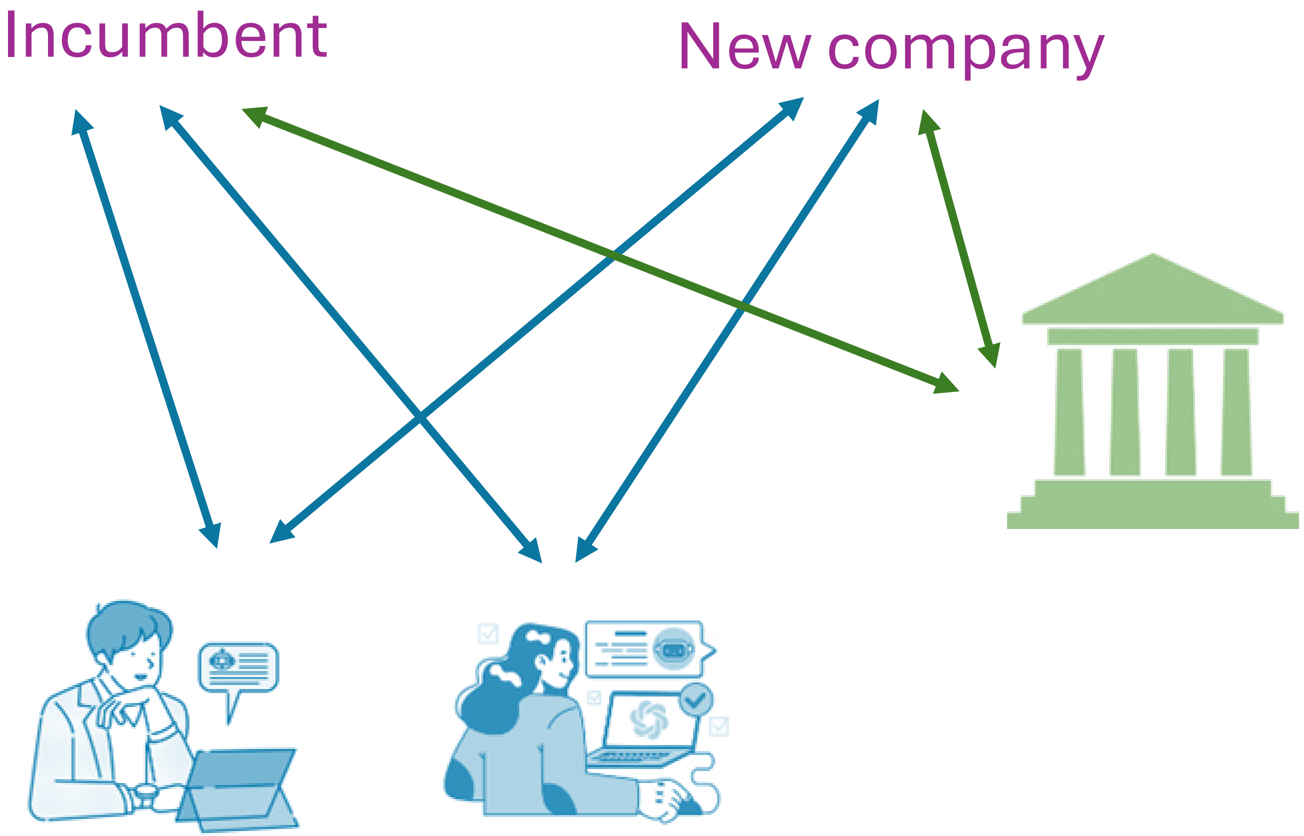
- **Incumbent faces a stricter threshold.**

Users choose the unscrutinized model with lowest loss on β_1

Model: markets of companies training LLMs

Multi-objective high-dimensional linear regression:

- x = P-dimensional query embedding (we take $P \to \infty$),
- $\langle \beta_1, x \rangle$ = performance-optimal output, $\langle \beta_2, x \rangle$ = safety-optimal output,



Each company strategically labels its training data with β_1 vs. β_2, and trains a model.

A company is scrutinized if its safety loss w.r.t. β_2 exceeds a given threshold.

- **Incumbent faces a stricter threshold.**

Users choose the unscrutinized model with lowest loss on β_1

Model: the company's ML pipeline

	Incumbent I	***New company E***
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Model: the company's ML pipeline

	Incumbent I	***New company E***
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Each company chooses fraction to label with $\boldsymbol{\beta_1}$ vs. $\boldsymbol{\beta_2}$.

Each company chooses a regularization level.

Model: the company's ML pipeline

	Incumbent $\boldsymbol{I}$	***New company*** $\boldsymbol{E}$
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Each company chooses fraction to label with $\boldsymbol{\beta_1}$ vs. $\boldsymbol{\beta_2}$.

Each company chooses a regularization level.

Each company runs ridge regression on its labelled dataset.

Model: the company's ML pipeline

	Incumbent I	***New company E***
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Each company chooses fraction to label with β_1 vs. β_2.

Each company chooses a regularization level.

Each company runs ridge regression on its labelled dataset.

Safety threshold:	*Faces scrutiny if loss on β_2 exceeds τ_I.*	*Faces scrutiny if loss on β_2 exceeds τ_E.*

Model: the company's ML pipeline

	Incumbent I	***New company E***
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Each company chooses fraction to label with $\boldsymbol{\beta_1}$ vs. $\boldsymbol{\beta_2}$.

Each company chooses a regularization level.

Each company runs ridge regression on its labelled dataset.

Safety threshold:	*Faces scrutiny if loss on β_2 exceeds τ_I.*	*Faces scrutiny if loss on β_2 exceeds τ_E.*

Assumption: $\boldsymbol{\tau_I < \tau_E}$.

Model: the company's ML pipeline

	Incumbent I	***New company E***
Training dataset size:	N_I *unlabeled points*	N_E *unlabeled points*

Each company chooses fraction to label with β_1 vs. β_2.

Each company chooses a regularization level.

Each company runs ridge regression on its labelled dataset.

Safety threshold:	*Faces scrutiny if loss on β_2 exceeds τ_I.*	*Faces scrutiny if loss on β_2 exceeds τ_E.*

Assumption: $\tau_I < \tau_E$.

Strategically chosen to minimize loss on β_1 subject to safety constraint

Outline for the talk

1. Background

2. Our model

3. **Our results**

Results: Barriers to market entry

<u>Definition</u>: Market entry threshold $\boldsymbol{N_E^*}$:= min dataset size N_E s.t. the new company achieves the incumbent's performance along β_1 without facing scrutiny.

Results: Barriers to market entry

<u>Definition</u>: Market entry threshold $\boldsymbol{N_E^*}$:= min dataset size N_E s.t. the new company achieves the incumbent's performance along β_1 without facing scrutiny.

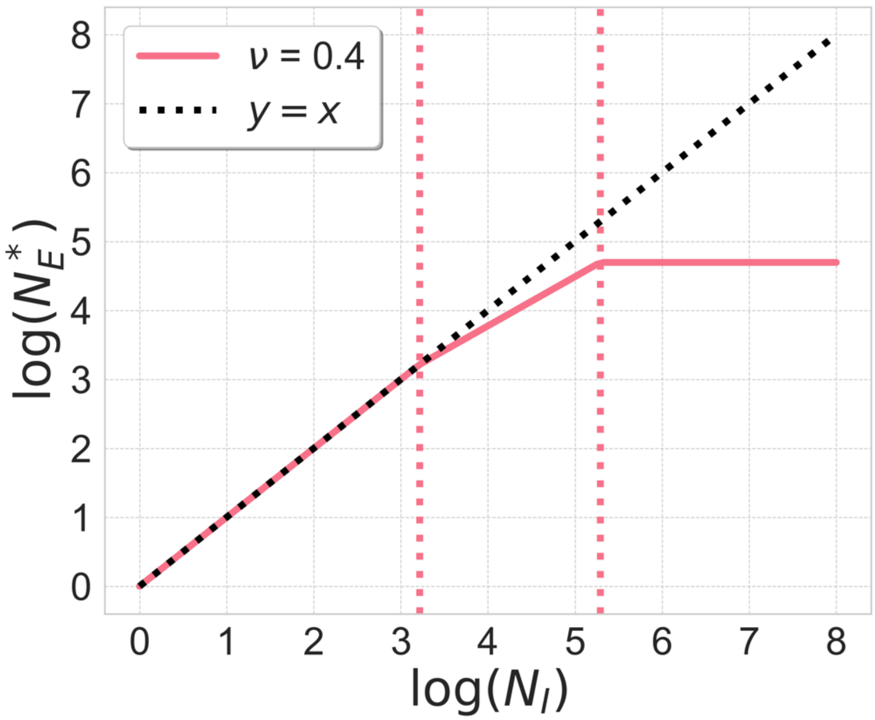
<u>Our main finding</u>: $N_E^* \ll N_I$ (i.e., the market entry threshold is much smaller than the incumbent's dataset size)

Result 1: new company faces no safety constraint

Setup: Incumbent has ***finite data*** N_I*; new company faces no safety constraint*

Result 1: new company faces no safety constraint

Setup: Incumbent has ***finite data*** $\boldsymbol{N_I}$*; new company faces no safety constraint*



Thm (Informal): The market entry threshold N_E^* satisfies

$$N_E^* = \begin{cases} \Theta(\boldsymbol{N_I}) & \text{if } \boldsymbol{N_I} \text{ is small} \\ \Theta\left(\boldsymbol{N_I}^{\frac{1}{\nu+1}} \cdot C\right) & \text{if } \boldsymbol{N_I} \text{ is larger} \\ \Theta(C') & \text{if } \boldsymbol{N_I} \text{ is large} \end{cases}$$

ν = a problem-specific data efficiency

Observation: New company enter with less than the incumbent as long as $\boldsymbol{N_I}$ is large enough

Result 2: new company faces a safety constraint

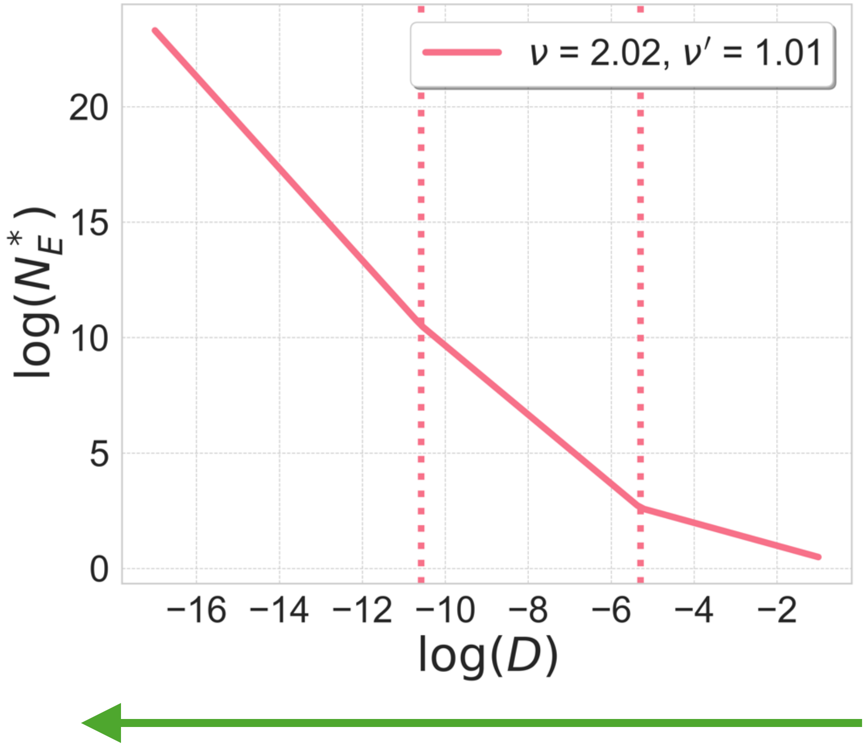
Setup: Incumbent has infinite data; new company has ***nontrivial safety constraint $\boldsymbol{\tau_E}$***

Let $\boldsymbol{D}$:= gap between the safety thresholds τ_E and τ_I.

Result 2: new company faces a safety constraint

Setup: Incumbent has infinite data; new company has ***nontrivial safety constraint*** τ_E

Let D := gap between the safety thresholds τ_E and τ_I.



Thm (Informal): The market entry threshold N_E^* satisfies

$$N_E^* = \begin{cases} \Theta\left(D^{-\frac{1}{\nu}}\right) & \text{if } D \text{ is large} \\ \Theta\left(D^{-\frac{\nu+1}{\nu}} \cdot C\right) & \text{if } D \text{ is smaller} \\ \Theta\left(D^{-\frac{\nu'+1}{\nu'}} \cdot C'\right) & \text{if } D \text{ is small} \end{cases}$$

$\nu' < \nu$ are problem-specific data efficiencies

Observation: New company can enter with finite data, but dataset size scales faster as $\tau_E \to \tau_I$.

Intuition for this phenomenon

Driver: the new company's model can be less aligned with safety objectives than the incumbent's model

The new company faces a weaker safety constraint

=> Can label a larger fraction of training data with performance-opt outputs

=> Can achieve the incumbent's performance level with less training data

Intuition for this phenomenon

Driver: the new company's model can be less aligned with safety objectives than the incumbent's model

The new company faces a weaker safety constraint

=> Can label a larger fraction of training data with performance-opt outputs

=> Can achieve the incumbent's performance level with less training data

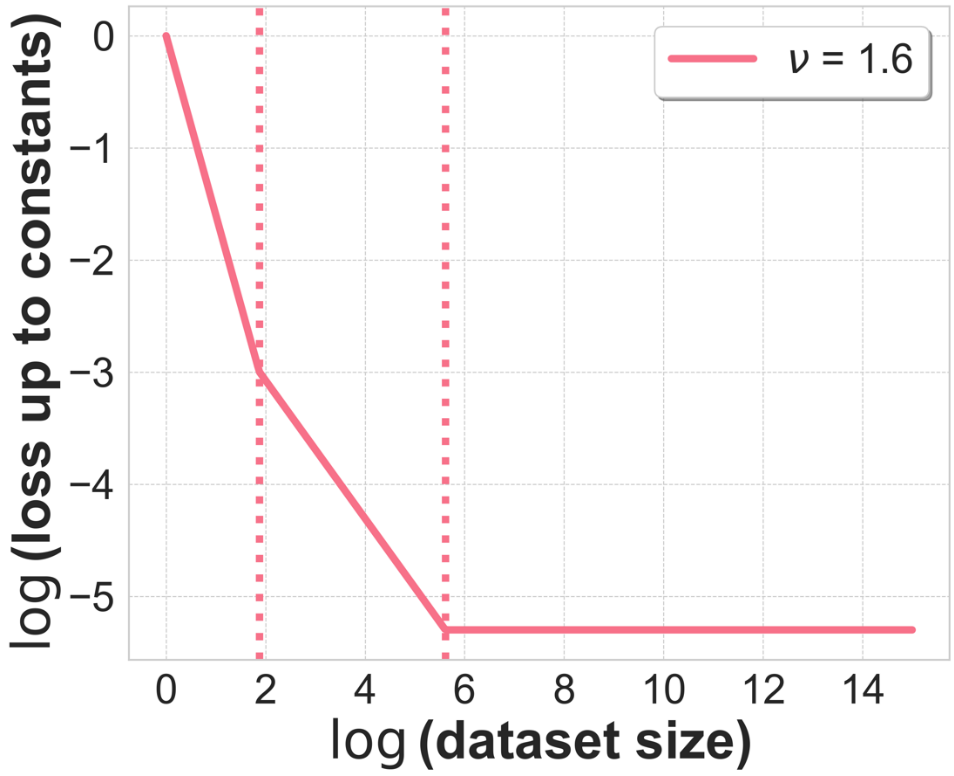
Technical tool: tight characterization of how data affects loss in multi-objective environments

- *Data efficiency becomes worse as the dataset size increases*

Conclusion

Technical tool: multi-objective data scaling laws

Setup: N training data points, fix fraction of data labelled with β_2, regularize optimally



Thm (Informal): Let L be the loss along β_1 of the ridge regression estimator.

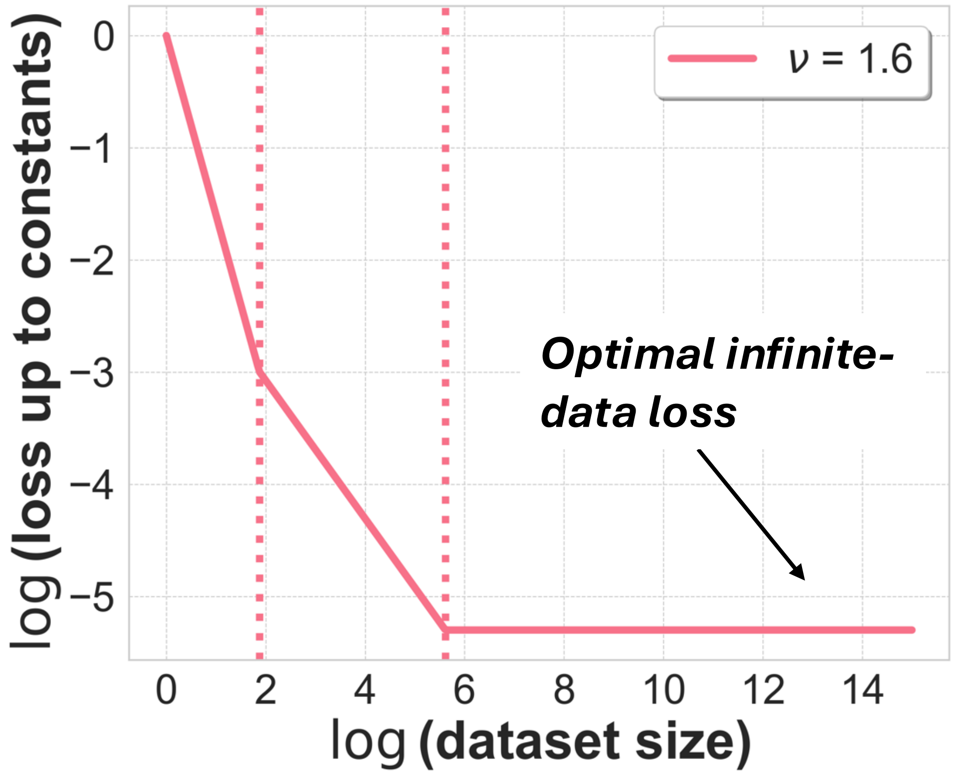
$$L = \begin{cases} \Theta\left(N^{-\frac{1}{\nu}}\right) & \text{if } N \text{ is small} \\ \Theta\left(N^{-\frac{\nu}{\nu+1}} \cdot C\right) & \text{if } N \text{ is larger} \\ \Theta(C') & \text{if } N \text{ is large} \end{cases}$$

N = dataset size   ***ν*** = problem-specific data efficiency

Scaling trend for loss: data efficiency decreases as N increases

Technical tool: multi-objective data scaling laws

Setup: N training data points, fix fraction of data labelled with β_2, regularize optimally



Thm (Informal): Let L be the loss along β_1 of the ridge regression estimator.

$$L = \begin{cases} \Theta\left(N^{-\frac{1}{\nu}}\right) & \text{if } N \text{ is small} \\ \Theta\left(N^{-\frac{\nu}{\nu+1}} \cdot C\right) & \text{if } N \text{ is larger} \\ \Theta(C') & \text{if } N \text{ is large} \end{cases}$$

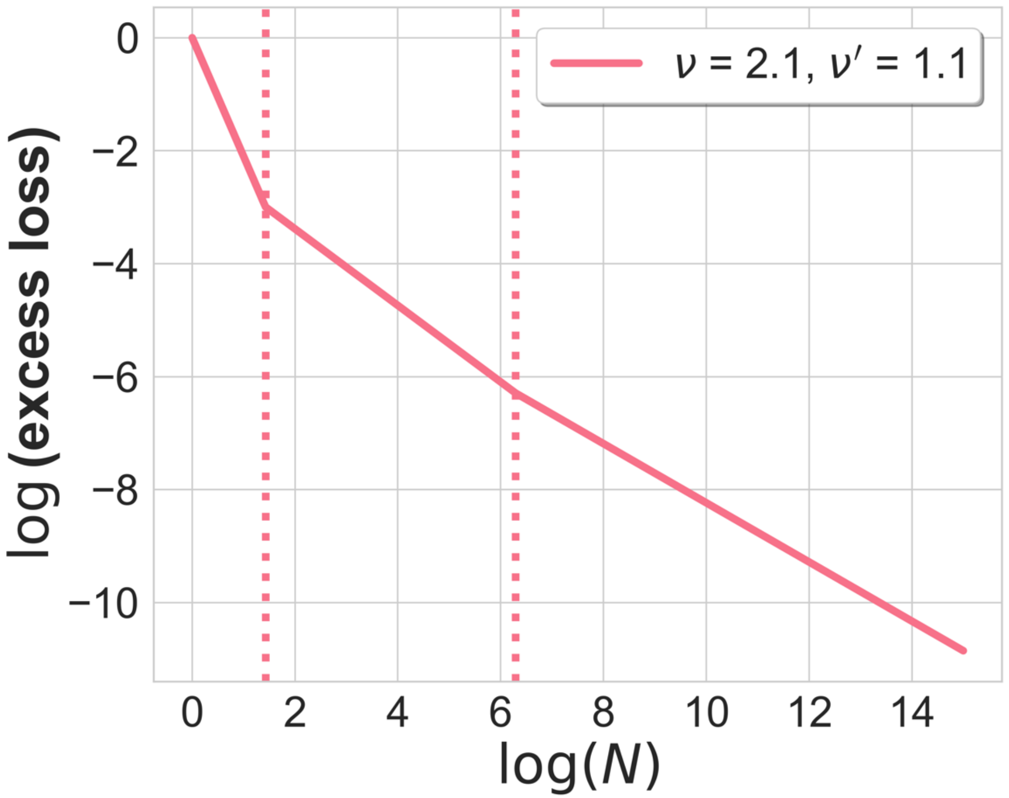
N = dataset size ν = problem-specific data efficiency

Scaling trend for loss: data efficiency decreases as N increases

Model discussion

Technical tool: multi-objective data scaling laws

Setup: N training data points, fix fraction of data labelled with β_2, regularize optimally



Thm (Informal): Let L be the excess loss along β_1 of the ridge regression estimator.

$$L = \begin{cases} \Theta\left(N^{-\frac{1}{\nu}}\right) & \text{if } N \text{ is small} \\ \Theta\left(N^{-\frac{\nu}{\nu+1}} \cdot C\right) & \text{if } N \text{ is larger.} \\ \Theta\left(N^{-\frac{\nu'}{\nu'+1}} \cdot C\right) & \text{if } N \text{ is large} \end{cases}$$

N = dataset size

$\nu' < \nu$ are problem-specific data efficiencies

Scaling trend for excess loss: data efficiency decreases as N increases

Model discussion

Proof ideas for deriving scaling laws

Setup: N points, an α_i *fraction labelled with* β_i*, regularization level* λ

Need **tight bounds** on test loss of ridge regression under data mixture

Proof ideas for deriving scaling laws

Setup: N points, an α_i *fraction labelled with* β_i*, regularization level* λ

Need **tight bounds** on test loss of ridge regression under data mixture

Challenge: test loss depends on randomness of N training data points

Proof ideas for deriving scaling laws

Setup: N points, an α_i *fraction labelled with* β_i*, regularization level* λ

Need **tight bounds** on test loss of ridge regression under data mixture

Challenge: test loss depends on randomness of N training data points

Key idea: Derive a **deterministic equivalent** using Marčenko-Pastur law

Model discussion

Key lemma: tight bounds on the loss

Setup: N points, an α_i fraction labelled with β_i, regularization level λ

Lemma (Informal): The loss is approximately equal to:

$$\max\left(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}\right) + \alpha_2 \cdot Q \cdot \frac{\min\left(N,\ \lambda^{-\frac{1}{1+\gamma}}\right)}{N} + \alpha_2^2 \cdot Q + \alpha_2 \cdot Q \max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})$$

Finite data error

Overfitting error

Optimal infinite data loss

Extra (Mixture finite data error)

α_2 = fraction of data labelled with β_2, $\boldsymbol{\nu}' < \boldsymbol{\nu}$ are problem-specific efficiencies, Q = misalignment level

Key lemma: tight bounds on the loss

Setup: N points, an α_i fraction labelled with β_i, regularization level λ

<u>Lemma (Informal):</u> The loss is approximately equal to:

$$\max\left(\lambda^{\frac{\nu}{1+\gamma}}, N^{-\nu}\right) + \alpha_2 \cdot Q \cdot \frac{\min\left(N,\ \lambda^{-\frac{1}{1+\gamma}}\right)}{N} + \alpha_2^2 \cdot Q + \alpha_2 \cdot Q \max(\lambda^{\frac{\nu'}{1+\gamma}}, N^{-\nu'})$$

Finite data error

Overfitting error

Optimal infinite data loss

Extra (Mixture finite data error)

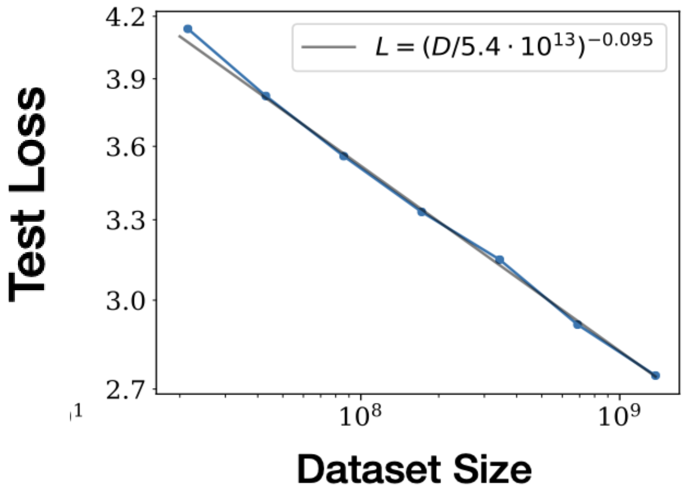
α_2 = fraction of data labelled with β_2, $\boldsymbol{\nu}' < \boldsymbol{\nu}$ are problem-specific efficiencies, Q = misalignment level

<u>Key idea</u>: need to regularize to avoid overfitting, but this reduces data efficiency

Model discussion: From high-dim regression to LLMs

For **single-objective** scaling laws, high-dim regression captures LLM behavior.

- For LLMs, loss and data empirically follow a power law relationship (e.g., Kaplan et al., ‘20)



- High-dim regression captures this power-law behavior
  - Exponent ν depends on covariance & linear predictor (e.g., Cui et al. ‘21, Wei et al., ‘22)

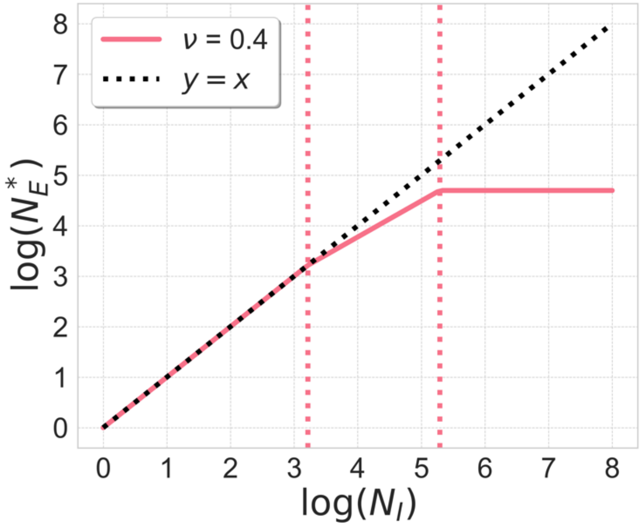
Future work: compare **multi-objective** scaling laws for high-dim regression with LLM behavior

Summary: market of companies that train LLMs

We studied barriers to market entry for companies training LLMs.

Main finding: a new company can enter with less data than the incumbent.

- Model: We modelled these markets within a multi-objective learning framework.
- Intuition: regulatory and societal scrutiny places pressure on the incumbent to align with safety
- Technical tool: multi-objective data scaling laws



Future direction: tradeoff between market concentration and safety compliance?