Understanding Post-training through the Lens of Intrinsic Dimension

A Story about Weak-to-Strong Generalization

Yijun Dong

Courant Institute of Mathematical Sciences, New York University

Post-training in the Era of Pre-trained Models

Astronomical data
+ GPU hours

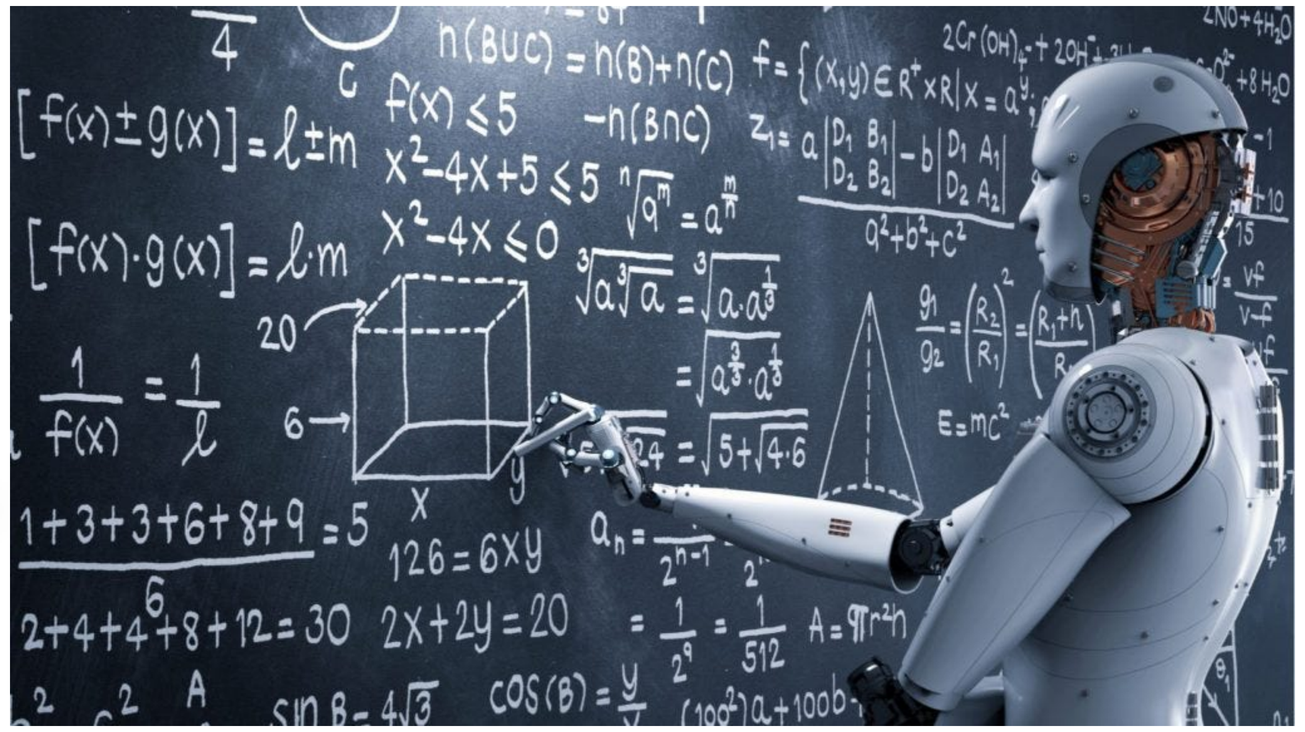
Pre-training

Powerful pre-trained
models

Gemini

Post-training

Specialized
downstream tasks



Post-training in the Era of Pre-trained Models

Astronomical data
+ GPU hours

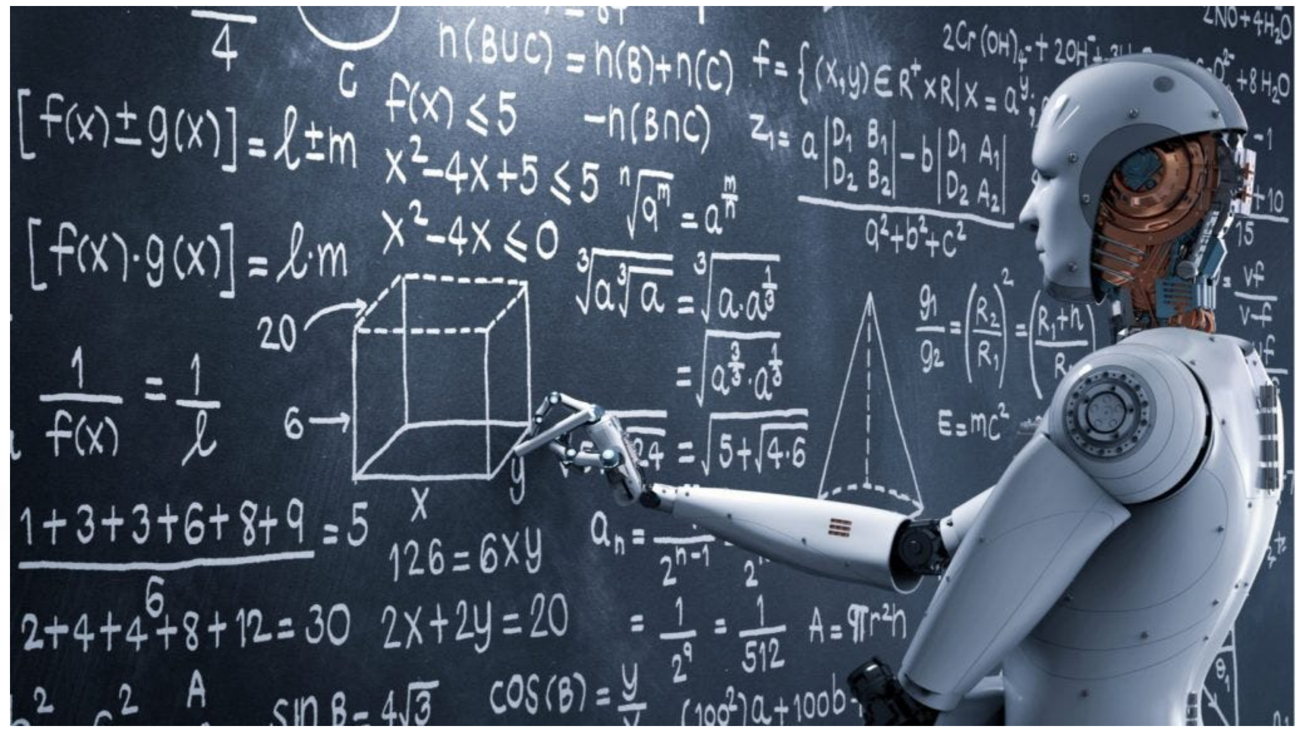
Pre-training

Powerful pre-trained
models

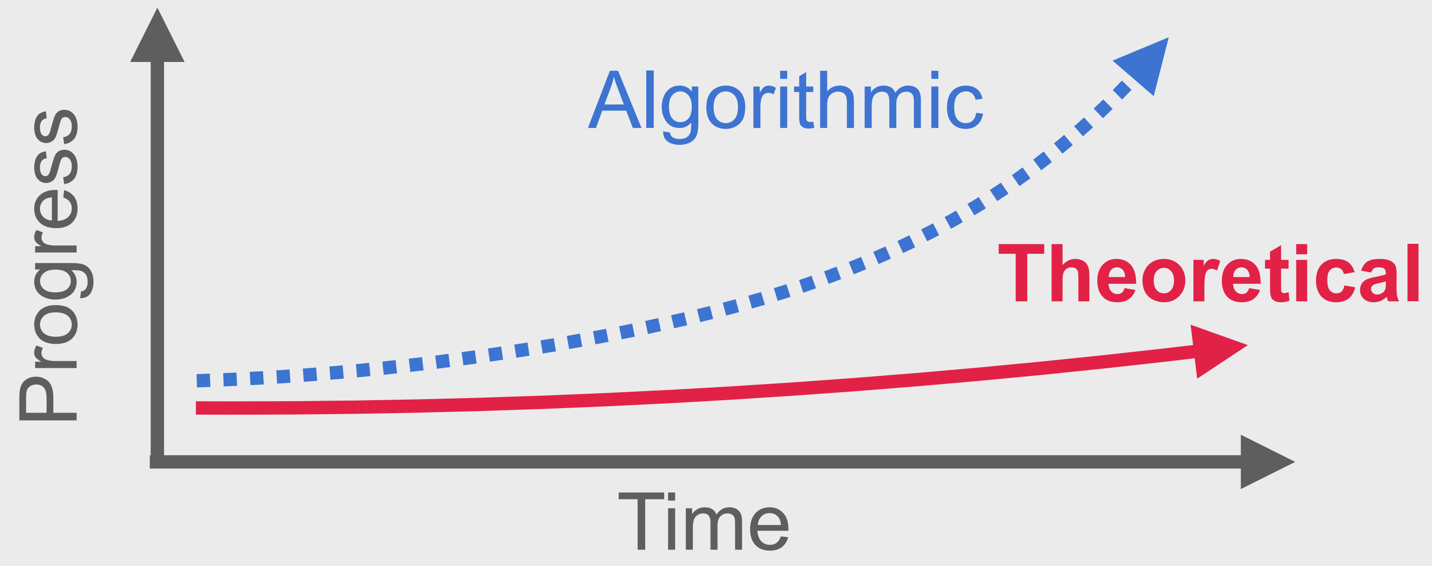
Gemini

Post-training

Specialized
downstream tasks



A wide variety of **post-training paradigms**
(e.g., supervised fine-tuning (SFT),
reinforcement learning (RL), …)



Post-training in the Era of Pre-trained Models

Astronomical data
+ GPU hours

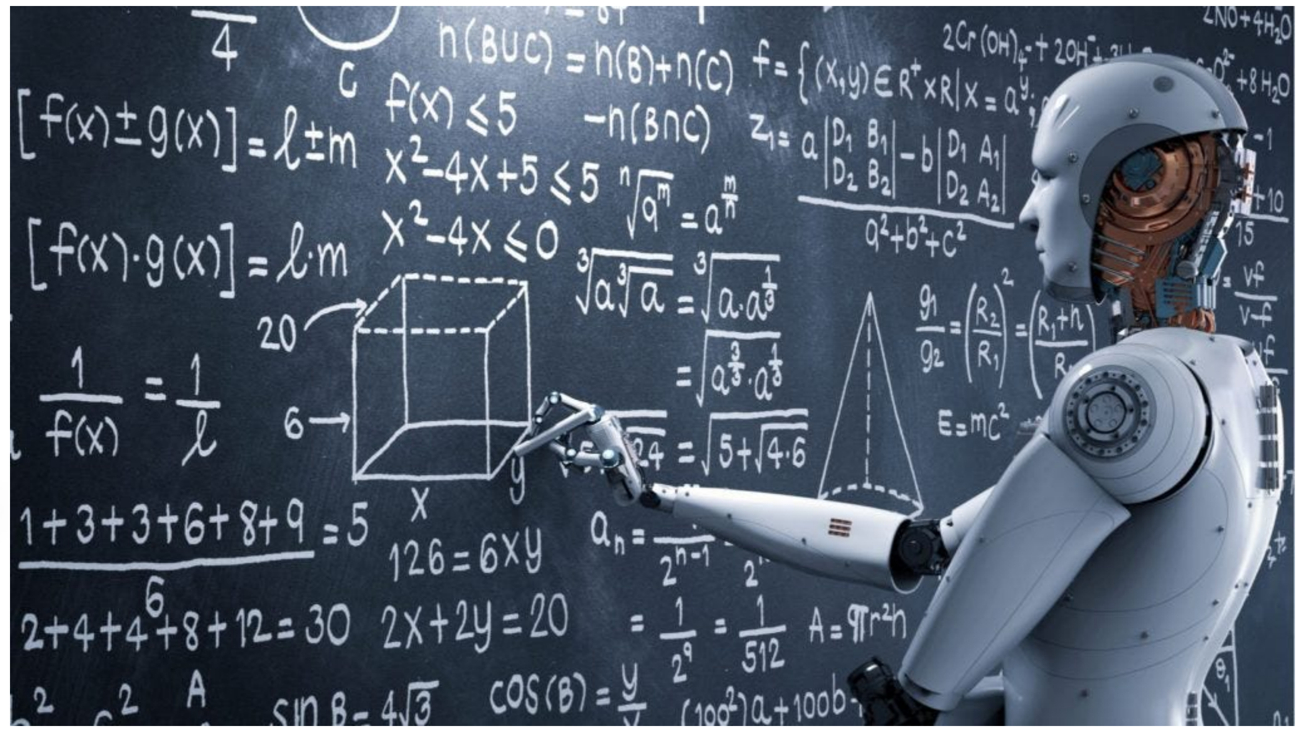
Pre-training

Powerful pre-trained
models

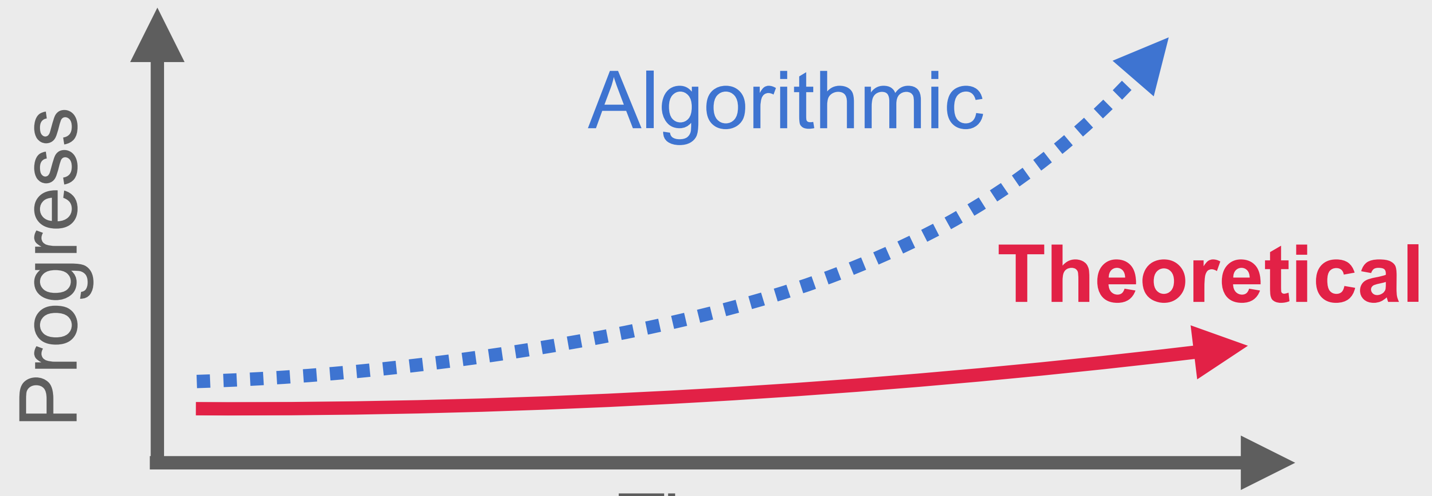
Gemini

Post-training

Specialized
downstream tasks



A wide variety of **post-training paradigms**
(e.g., supervised fine-tuning (SFT),
reinforcement learning (RL), …)



Mathematical understanding of post-training → **Principled post-training algorithms**

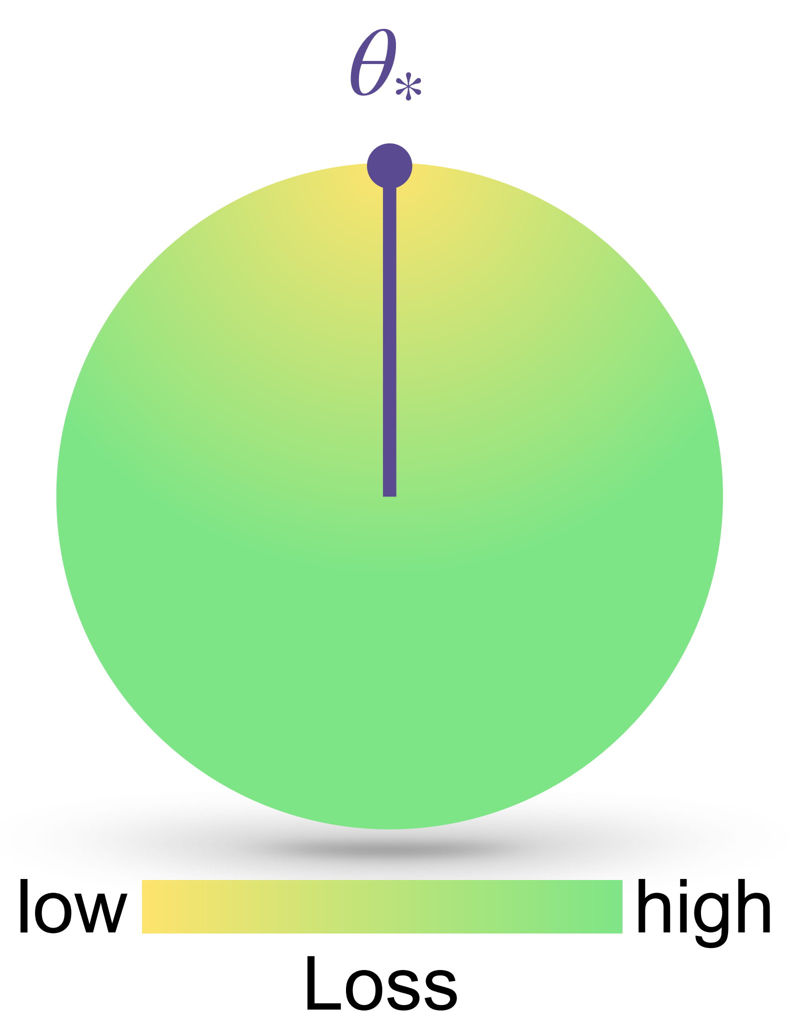
Simplicity of Post-training ①: Falls in the Kernel Regime

Conditioned on good pre-training, post-training often enjoys simple dynamics.

Simplicity of Post-training ①: Falls in the Kernel Regime

Conditioned on good pre-training, post-training often enjoys simple dynamics.

Aim to learn a neural network $f : \mathscr{X} \times \mathbb{R}^D \to \mathbb{R}$ that takes input data $x \in \mathscr{X}$ and a high-dimensional parametrization $\theta \in \mathbb{R}^D$ and outputs $f(x \mid \theta) \in \mathbb{R}$

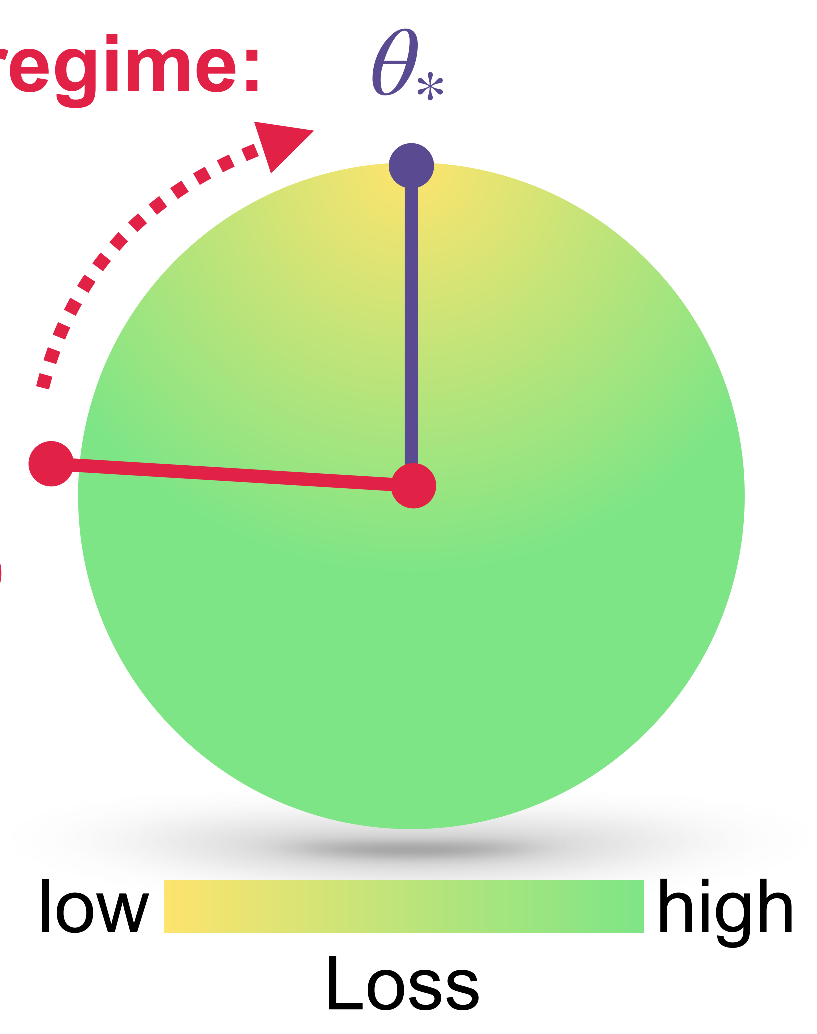


Simplicity of Post-training ①: Falls in the Kernel Regime

Conditioned on good pre-training, post-training often enjoys simple dynamics.

Aim to learn a neural network $f : \mathcal{X} \times \mathbb{R}^D \to \mathbb{R}$ that takes input data $x \in \mathcal{X}$ and a high-dimensional parametrization $\theta \in \mathbb{R}^D$ and outputs $f(x \mid \theta) \in \mathbb{R}$

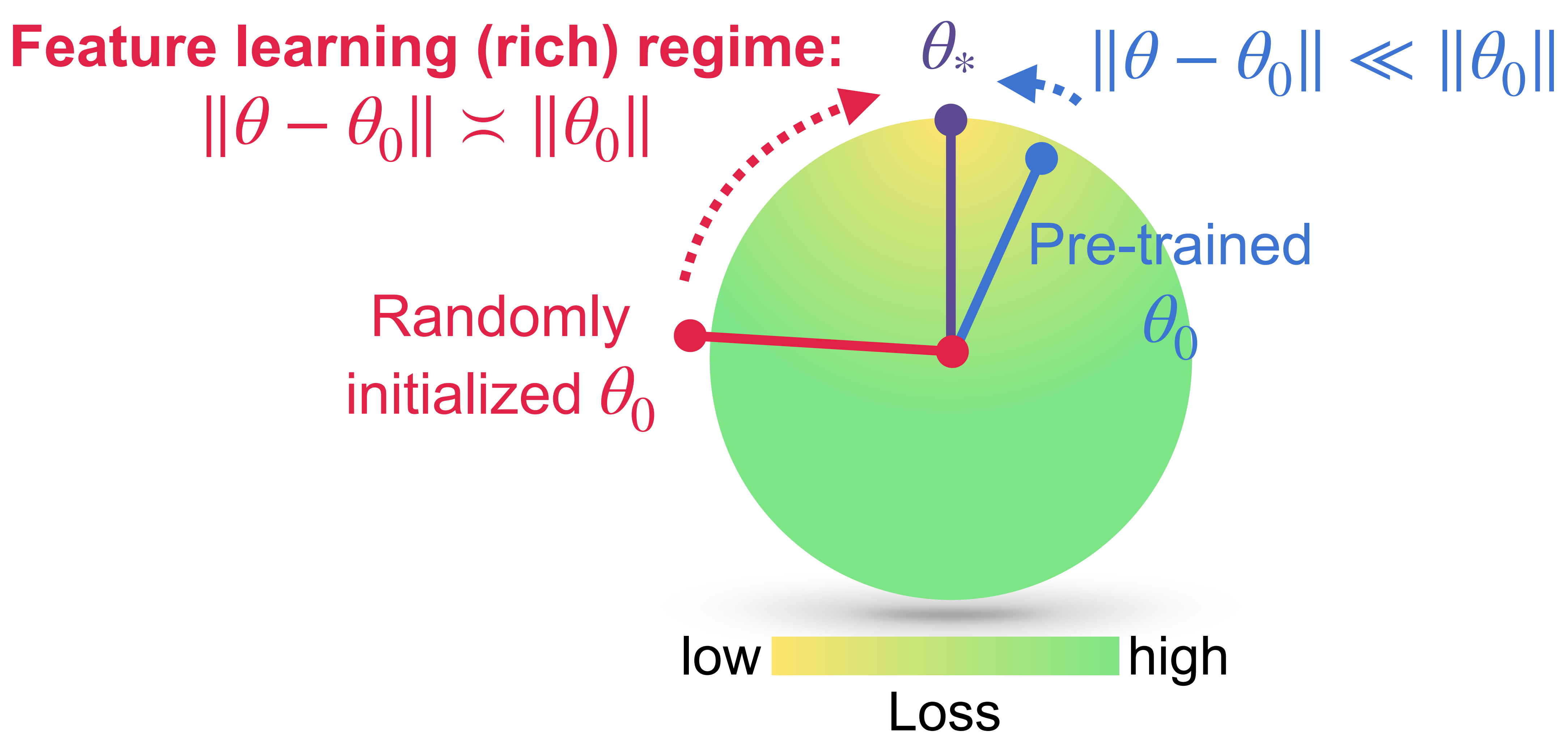
Feature learning (rich) regime:
$\|\theta - \theta_0\| \asymp \|\theta_0\|$



Simplicity of Post-training ①: Falls in the Kernel Regime

Conditioned on good pre-training, post-training often enjoys simple dynamics.

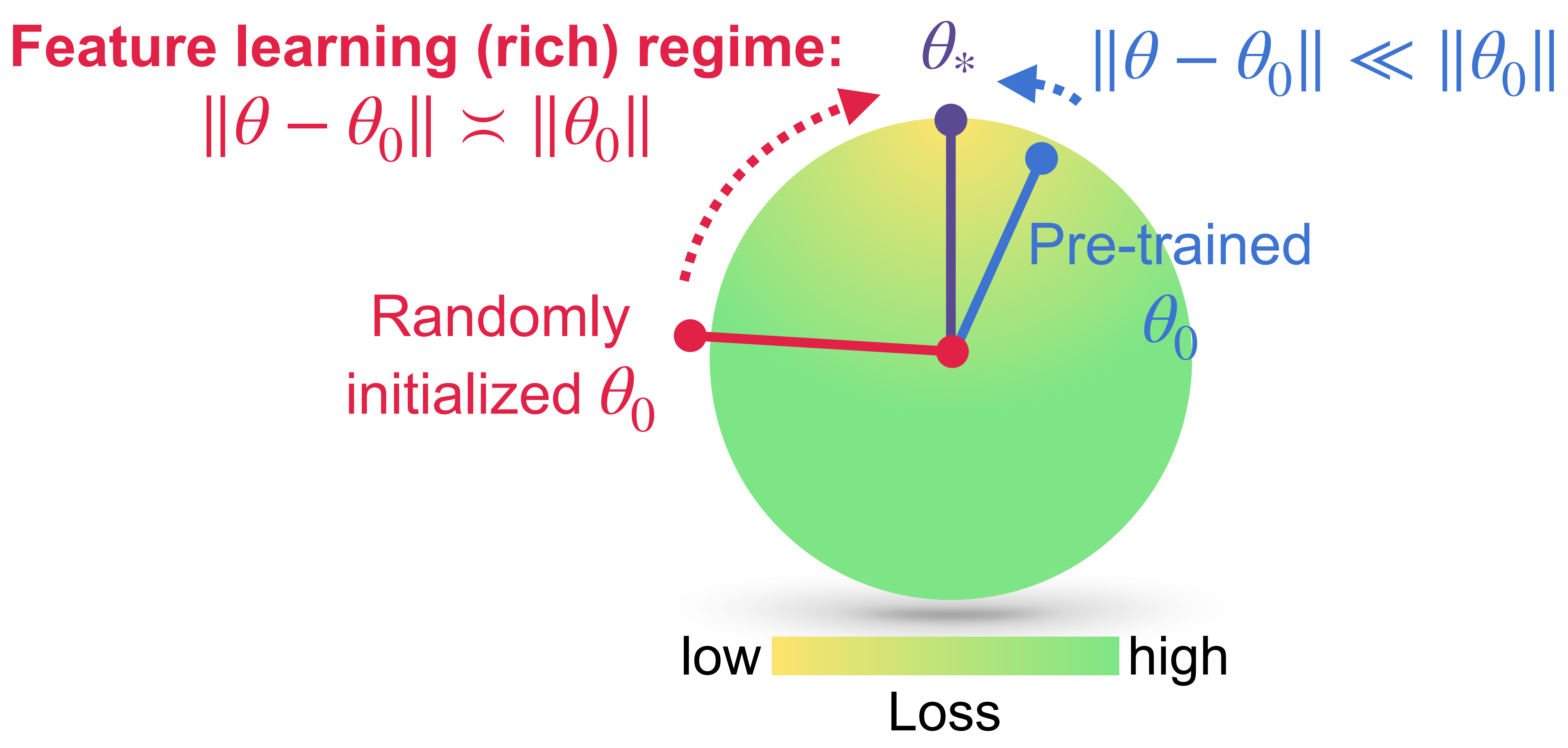
Aim to learn a neural network $f : \mathcal{X} \times \mathbb{R}^D \to \mathbb{R}$ that takes input data $x \in \mathcal{X}$ and a high-dimensional parametrization $\theta \in \mathbb{R}^D$ and outputs $f(x \mid \theta) \in \mathbb{R}$



Simplicity of Post-training ①: Falls in the Kernel Regime

Conditioned on good pre-training, post-training often enjoys simple dynamics.

Aim to learn a neural network $f : \mathcal{X} \times \mathbb{R}^D \to \mathbb{R}$ that takes input data $x \in \mathcal{X}$ and a high-dimensional parametrization $\theta \in \mathbb{R}^D$ and outputs $f(x \mid \theta) \in \mathbb{R}$



Kernel (lazy) regime

$$f(\cdot \mid \theta) \approx \langle \nabla_\theta f(\cdot \mid \theta_0), \theta - \theta_0 \rangle$$

when $\|\theta - \theta_0\| \ll \|\theta_0\|$

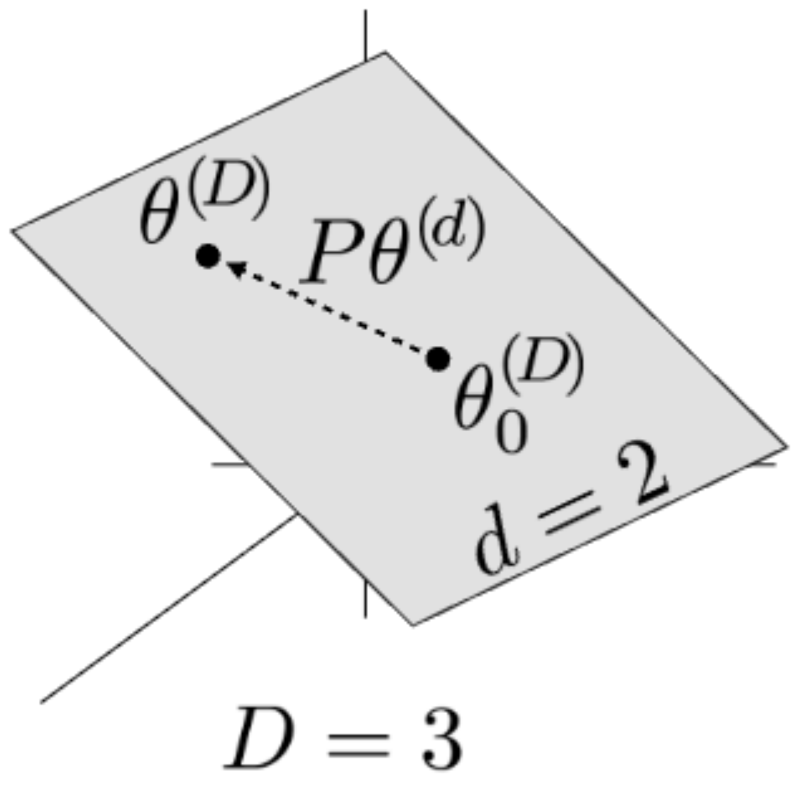
Feature: $\phi(x) = \nabla_\theta f(x \mid \theta_0) \in \mathbb{R}^D$

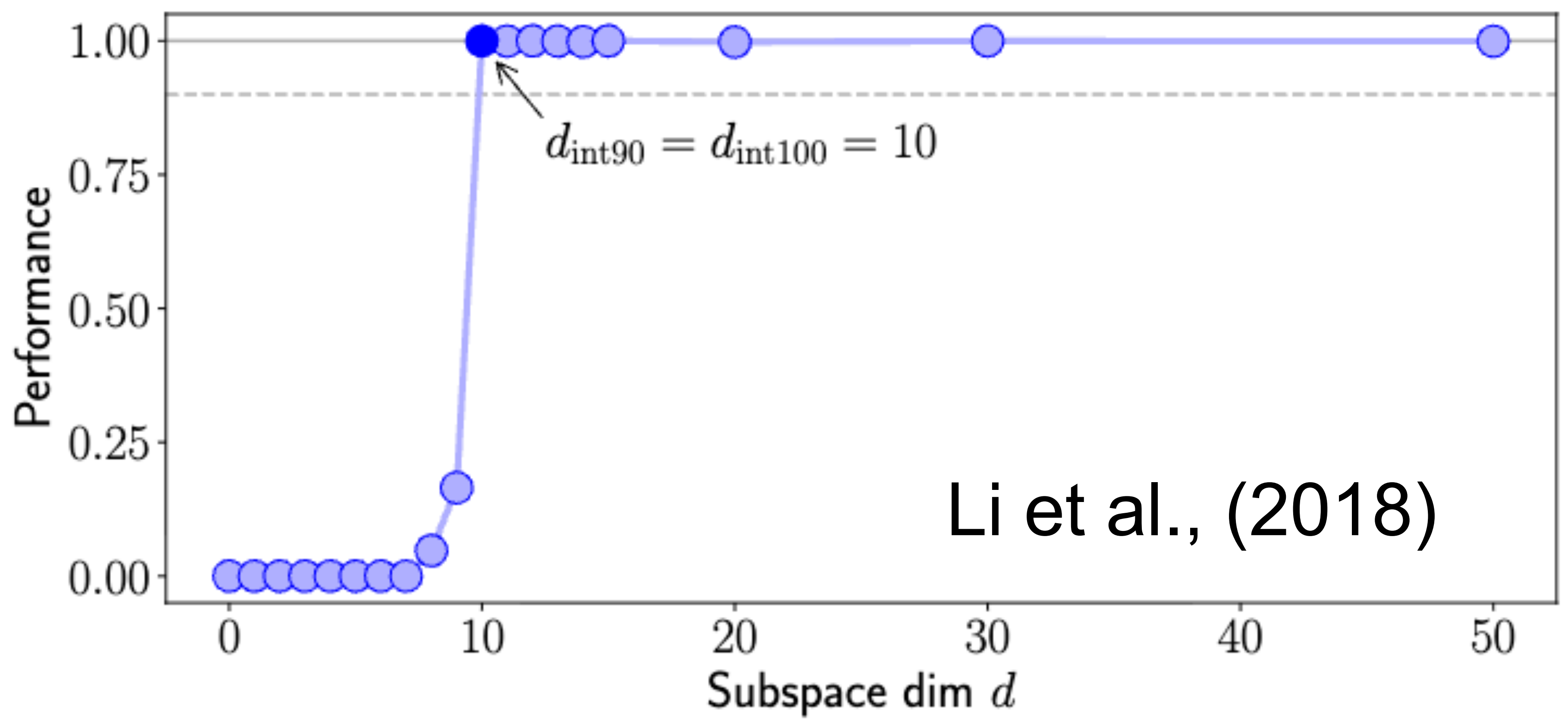
Post-training ≈ regression over Neural Tangent Kernel (NTK) (Jacot et al., 2018, Malladi et al., 2023).

Simplicity of Post-training ②: Admits Low Intrinsic Dimension

$$\theta^{(D)} = \theta_0^{(D)} + \underset{D\times d}{P}\,\theta^{(d)}$$

Intrinsic dimension:
minimal d needed to achieve nearly optimal performance

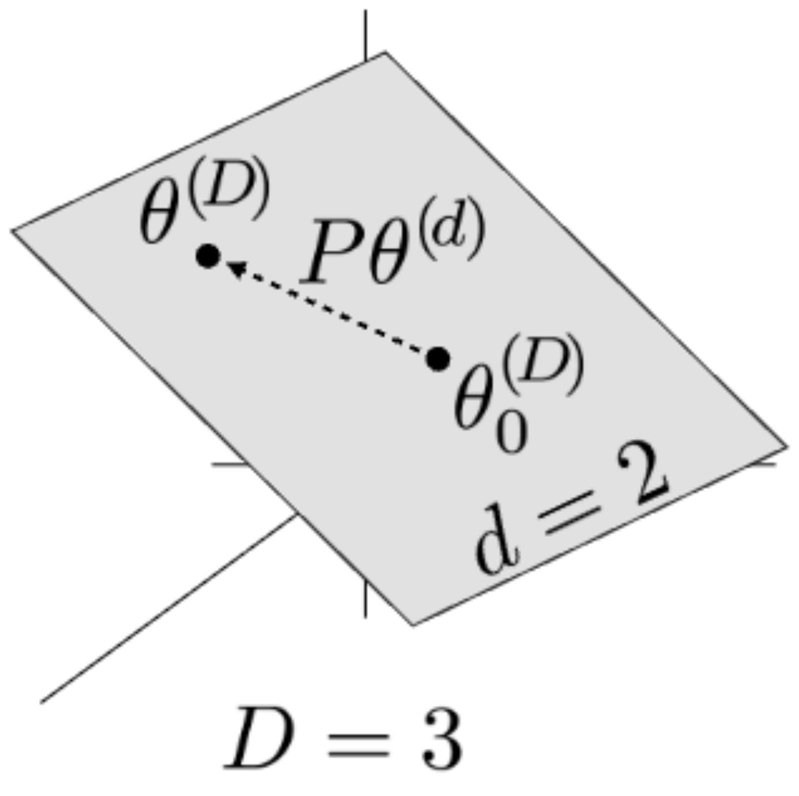


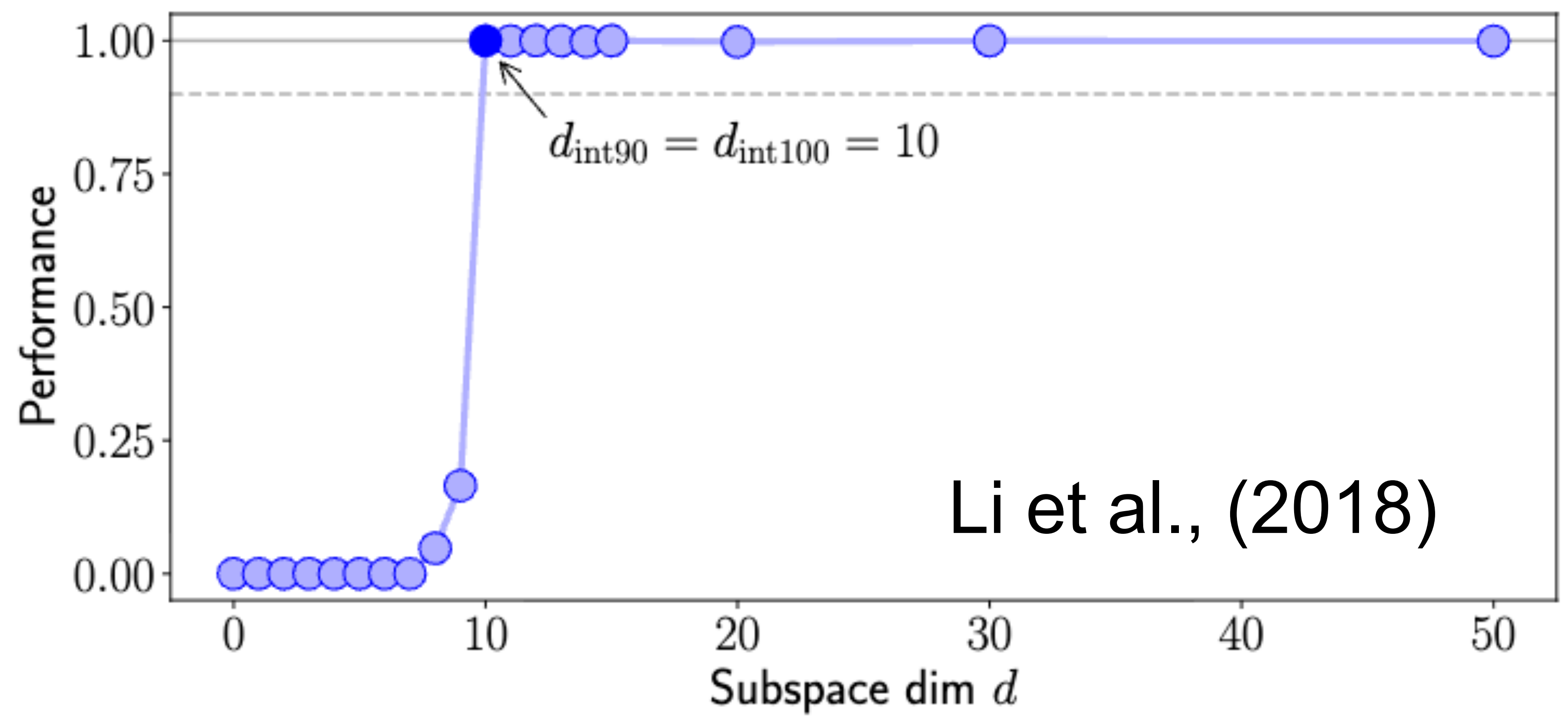


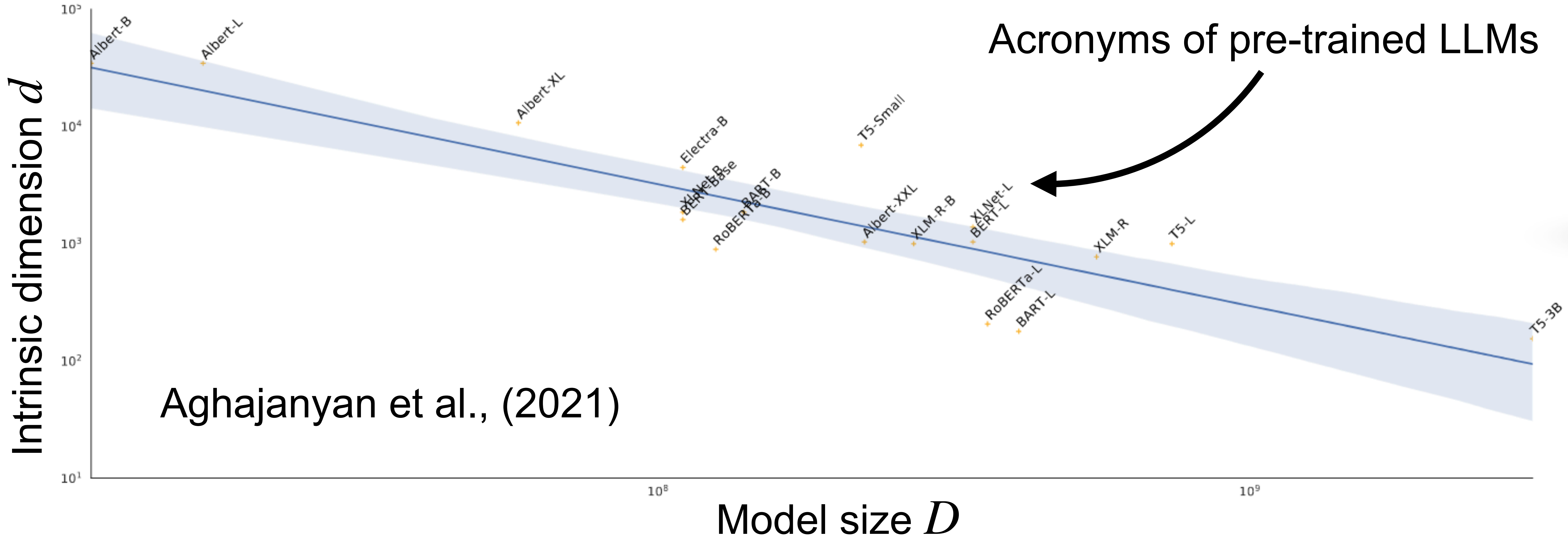
Simplicity of Post-training ②: Admits Low Intrinsic Dimension

$$\theta^{(D)} = \theta_0^{(D)} + \underset{D\times d}{P}\,\theta^{(d)}$$

Intrinsic dimension:
minimal d needed to achieve nearly optimal performance







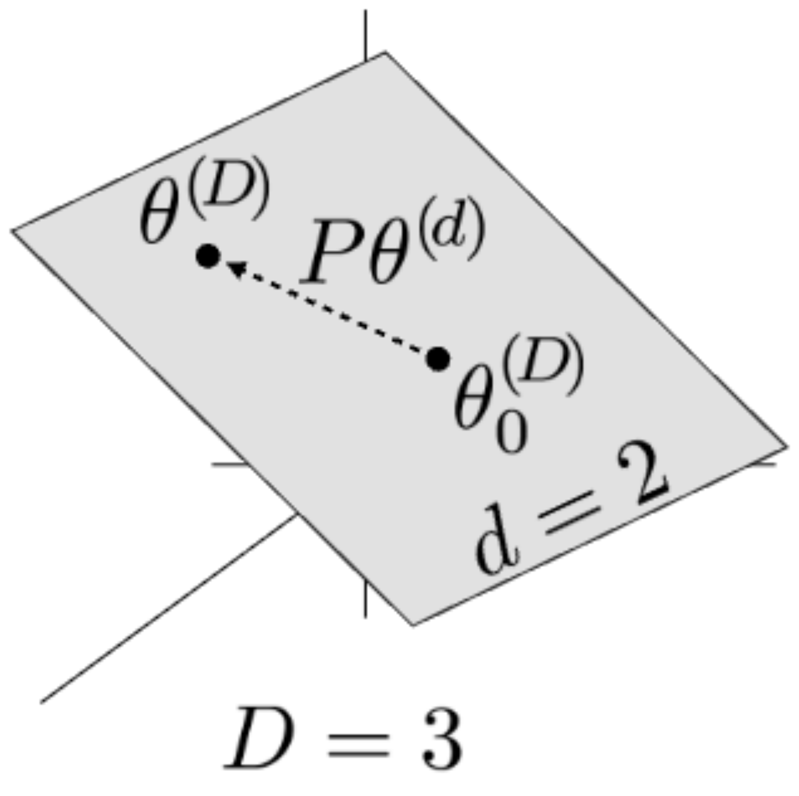
Acronyms of pre-trained LLMs

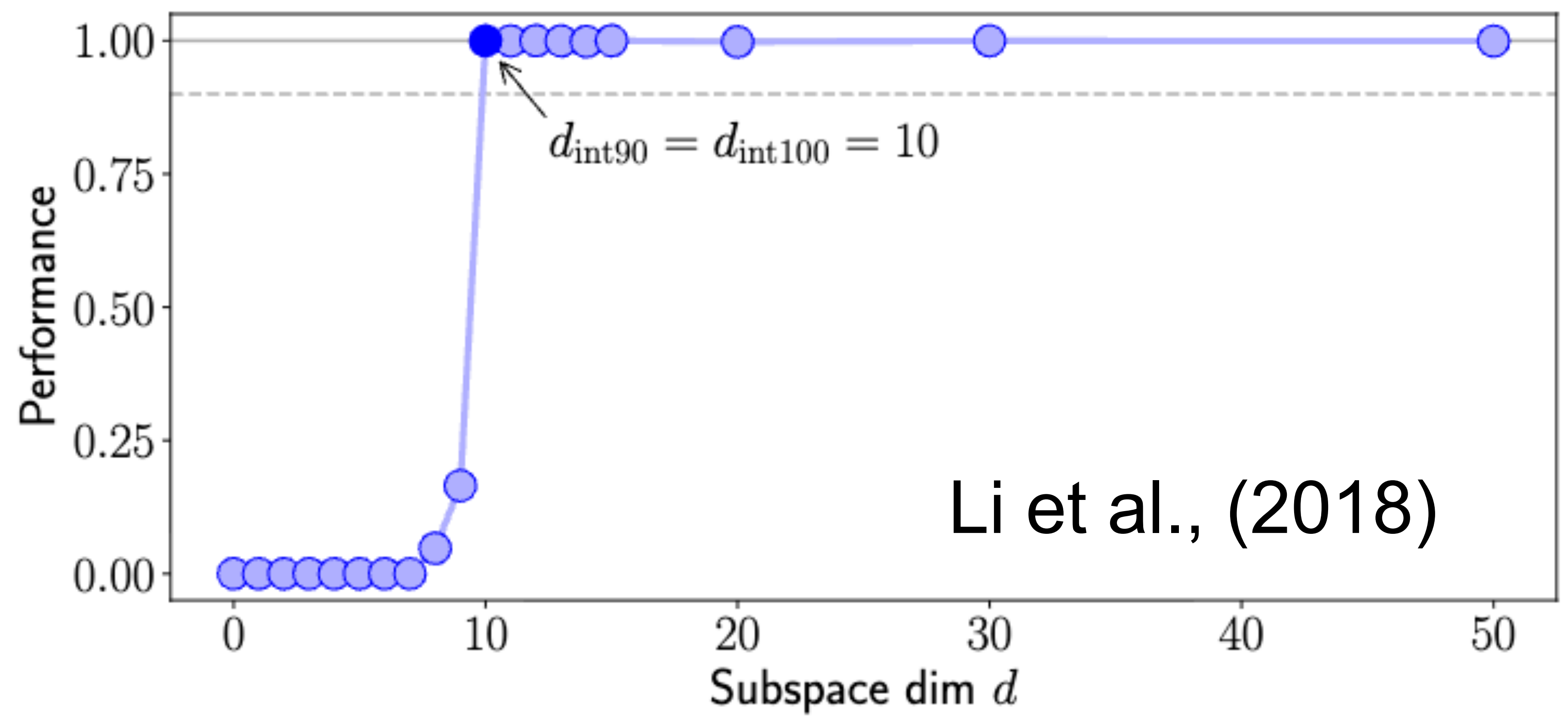
Stronger pre-trained language models have **lower intrinsic dimensions** on downstream tasks!

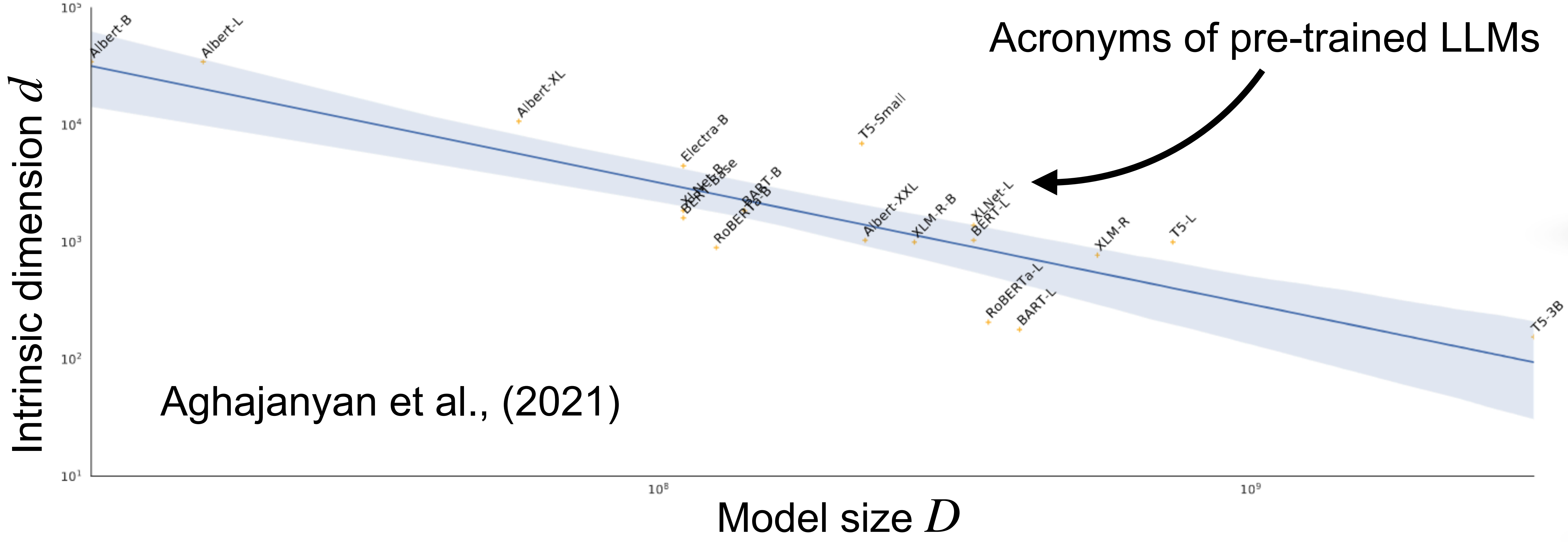
Simplicity of Post-training ②: Admits Low Intrinsic Dimension

$$\theta^{(D)} = \theta_0^{(D)} + \underset{D\times d}{P}\,\theta^{(d)}$$

Intrinsic dimension:
minimal d needed to achieve nearly optimal performance





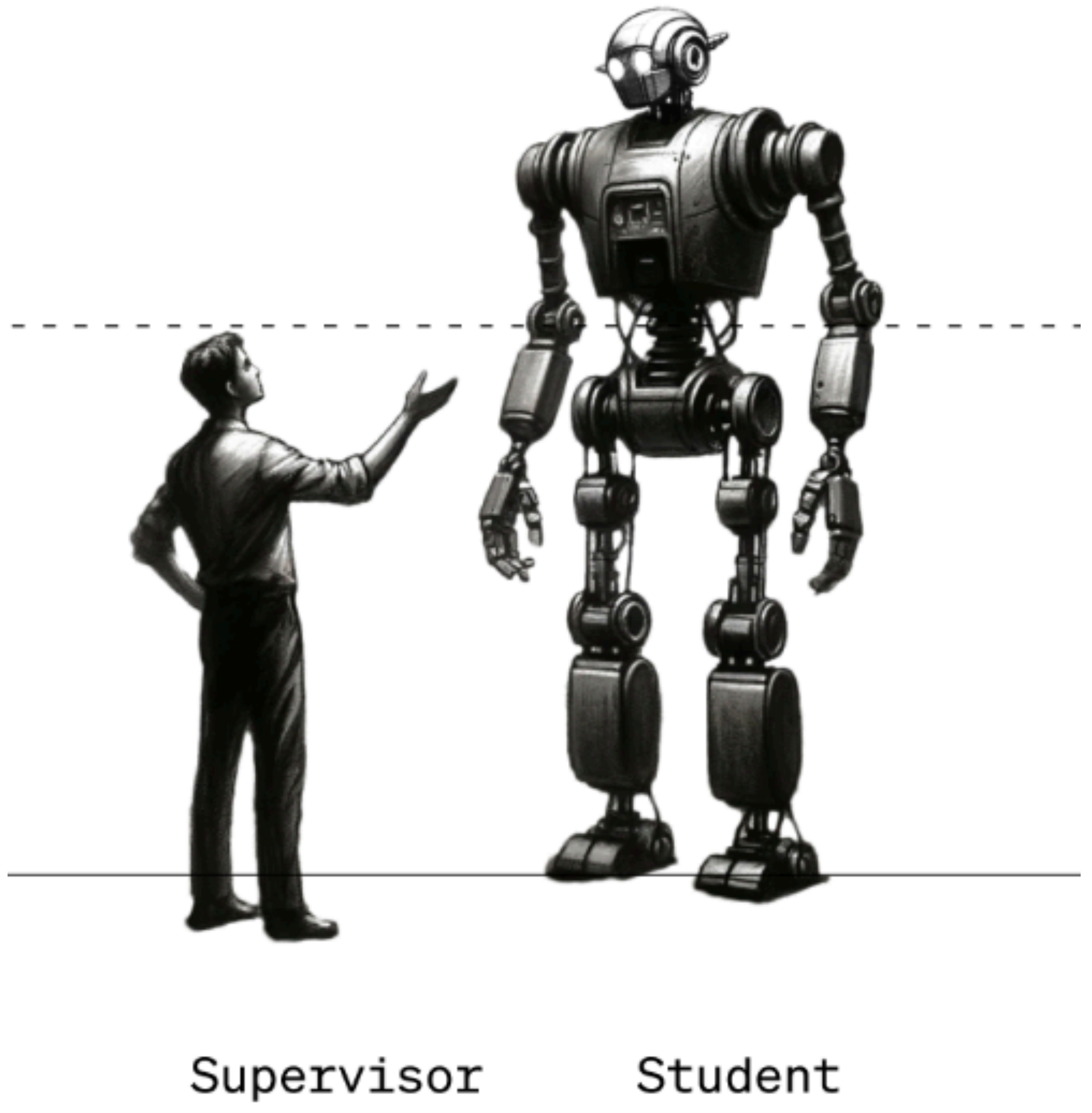


Stronger pre-trained language models have **lower intrinsic dimensions** on downstream tasks!

For $f(x) \approx \langle \phi(x), \theta - \theta_0 \rangle$ **with NTK feature** $\phi(x) = \nabla_\theta f(x \mid \theta_0)$**,** $\mathbb{E}[\phi(x)\phi(x)^\top]$ **is nearly low-rank**

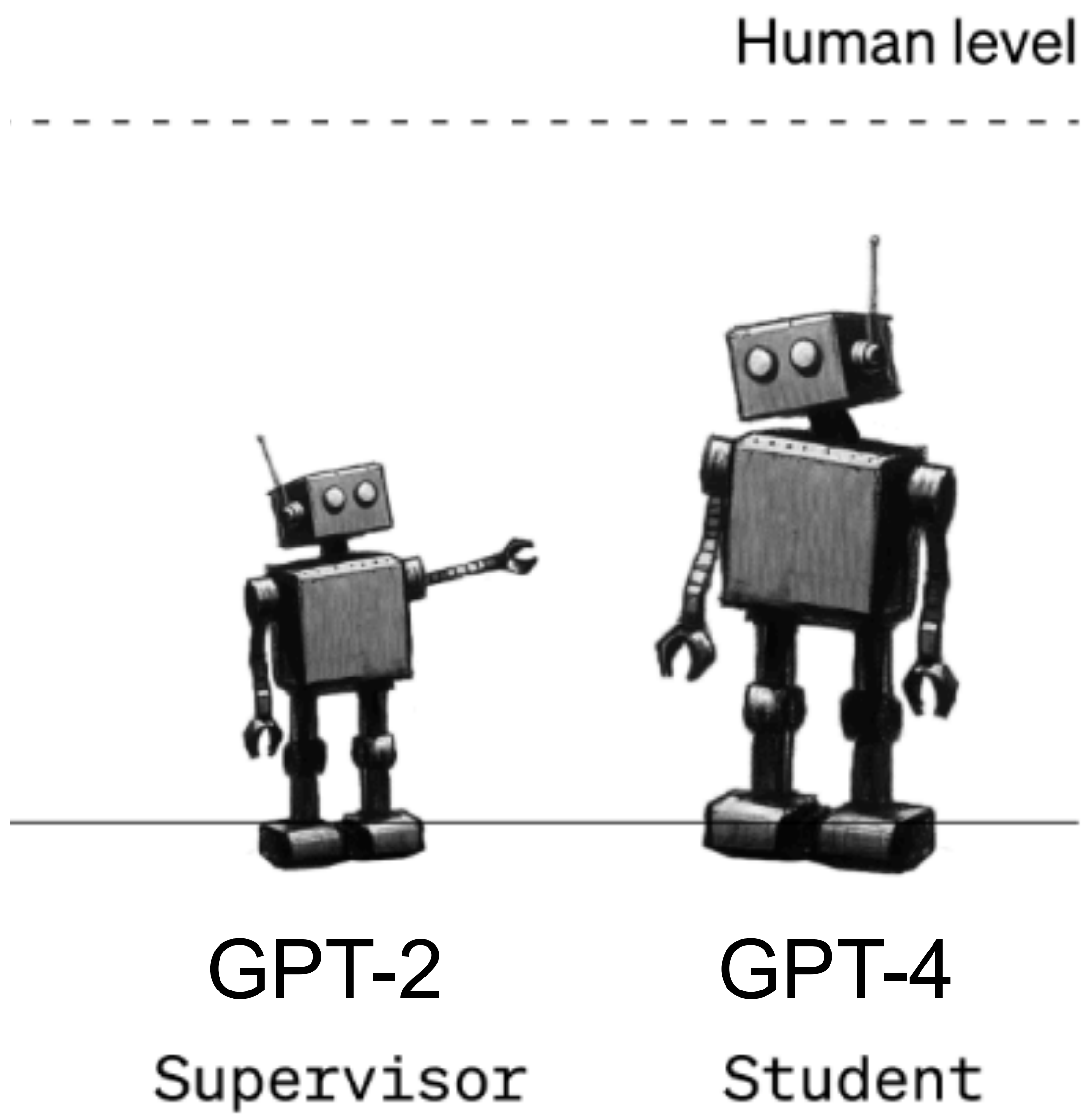
Superalignment → Weak-to-Strong (W2S) Generalization

Superalignment



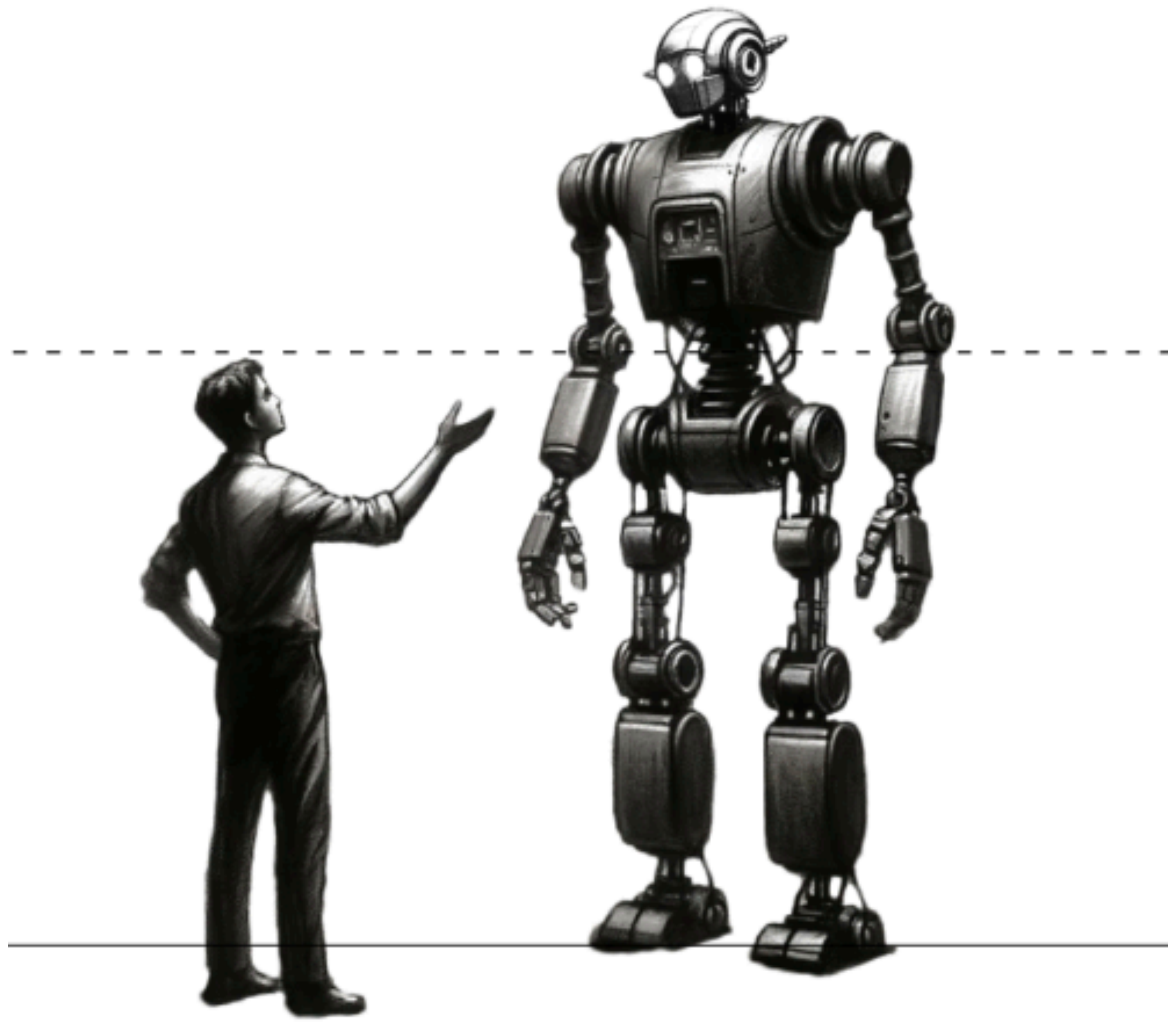
W2S

Burns et al., (2024)



Superalignment → Weak-to-Strong (W2S) Generalization

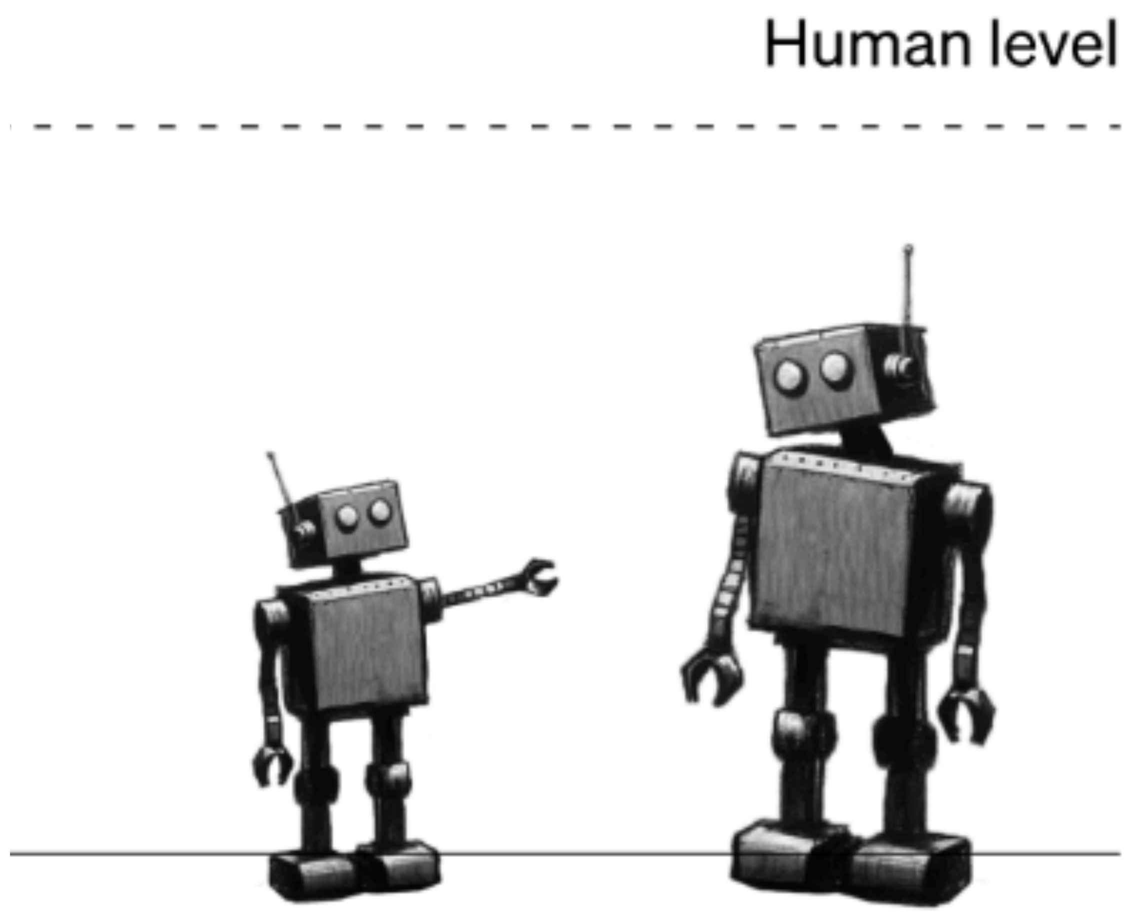
Superalignment



Supervisor Student

W2S

Burns et al., (2024)



GPT-2 GPT-4

Supervisor Student

Broader applications

Scalable oversight: use weaker models to supervise stronger ones

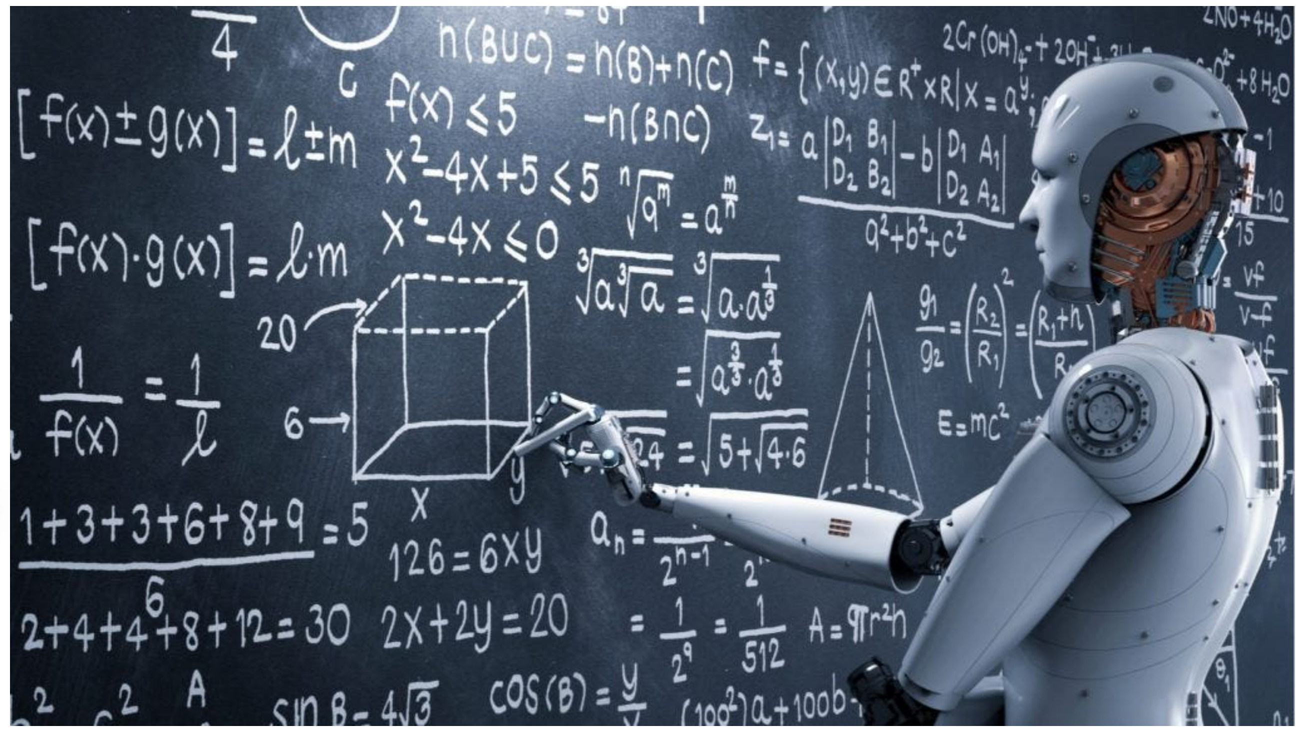
Self-improving: an LLM generates and uses its own feedback to improve iteratively

When and How Does W2S Emerge during Post-training?

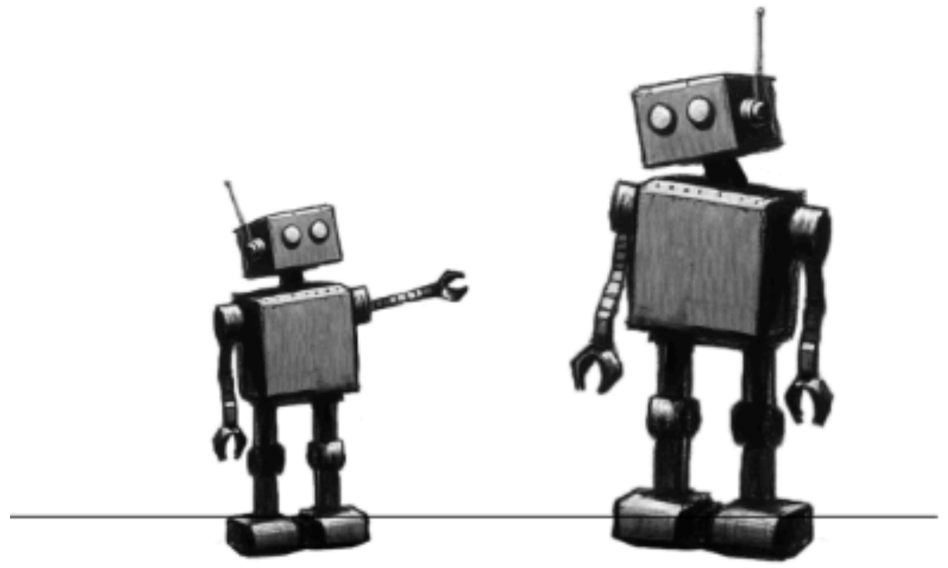
Powerful pre-trained models

Post-training

Specialized downstream tasks



Weak-to-strong generalization

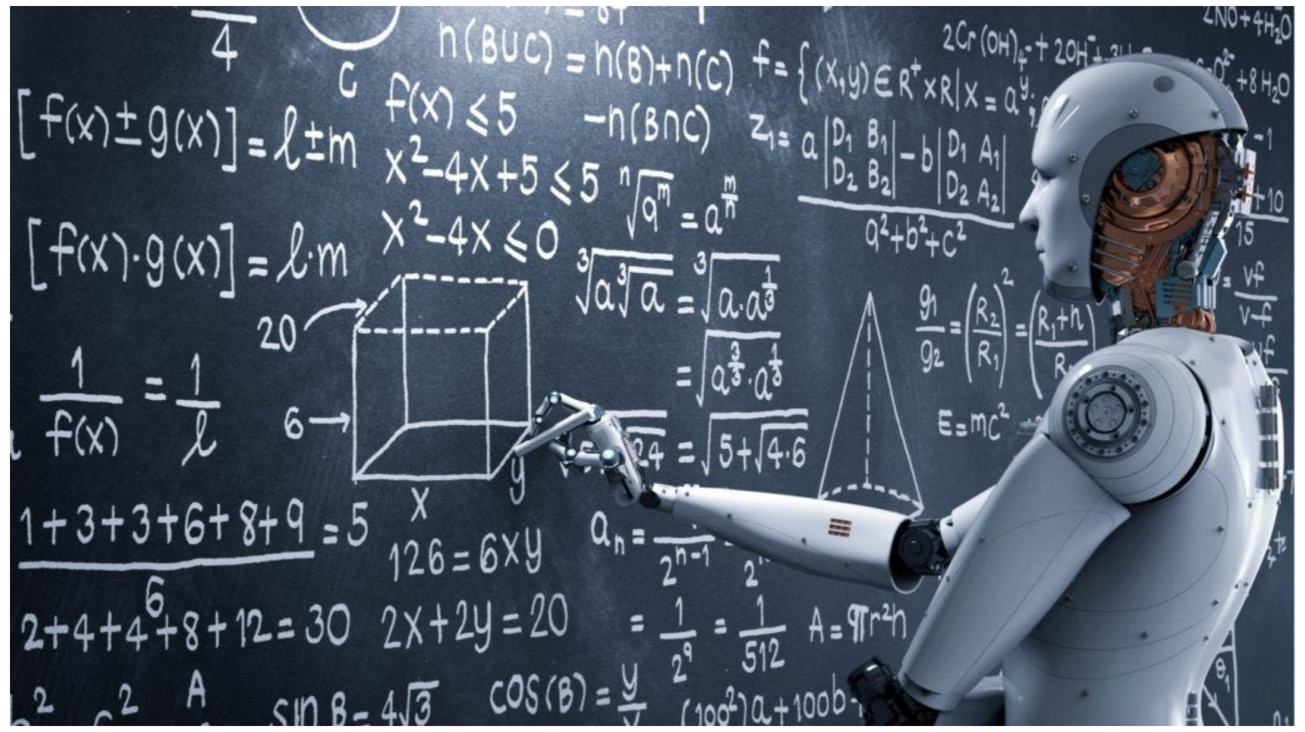


When and How Does W2S Emerge during Post-training?

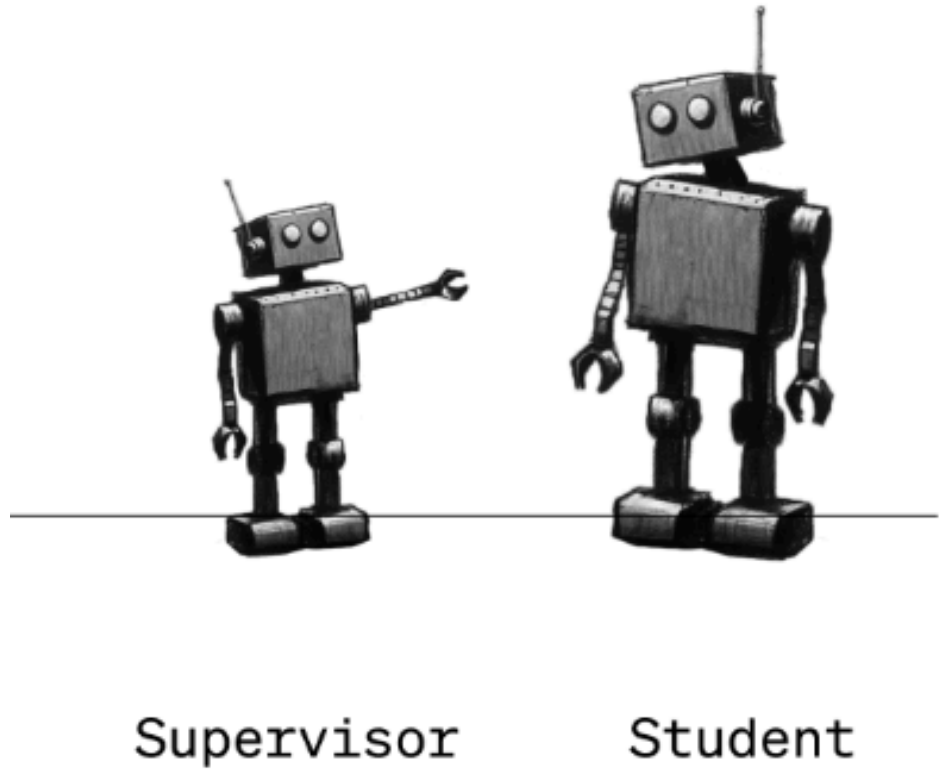
Powerful pre-trained models

Post-training

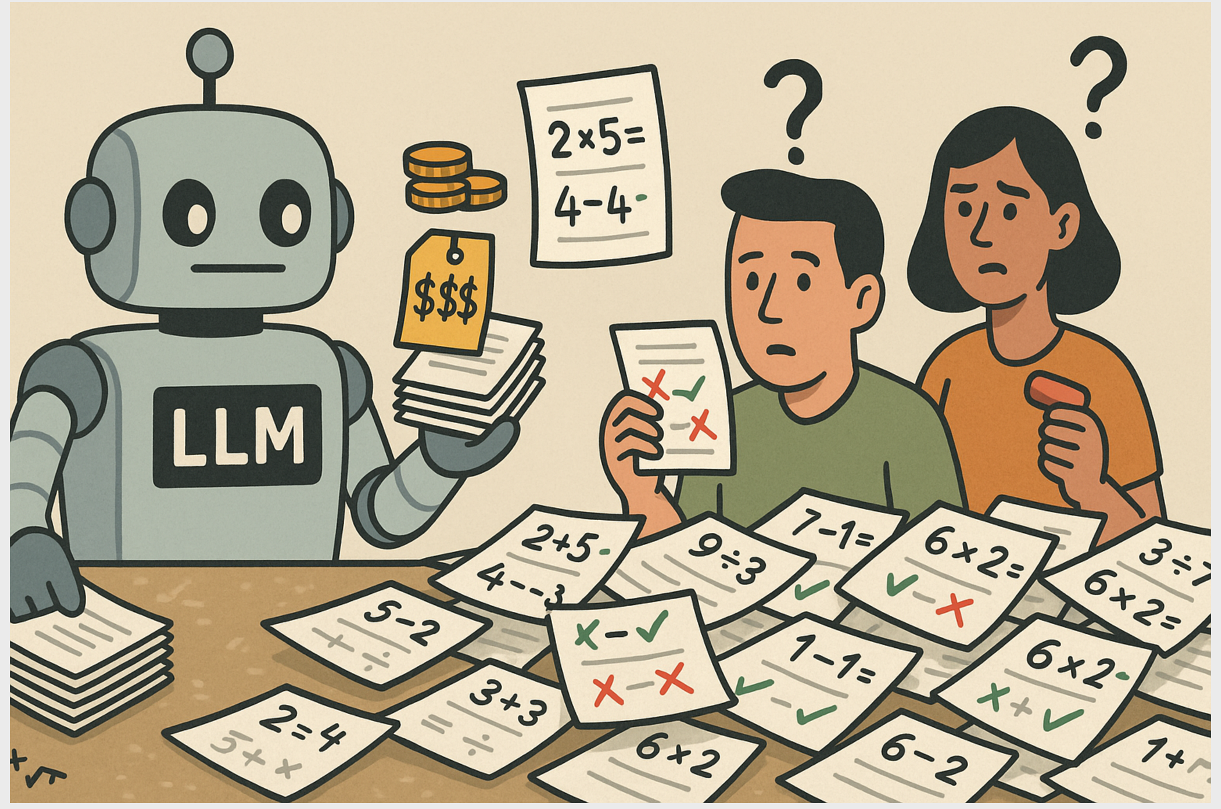
Specialized downstream tasks



Weak-to-strong generalization



① … with **limited & noisy labels**



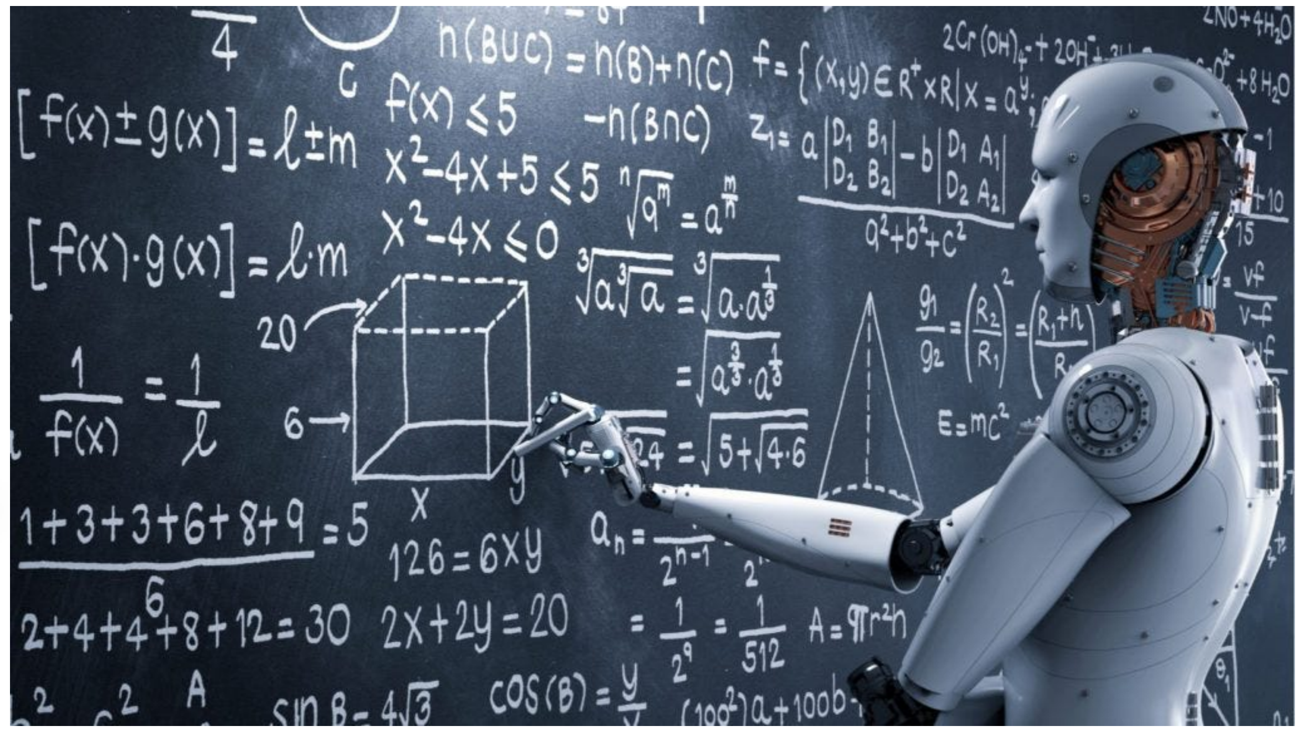
When and How Does W2S Emerge during Post-training?

Powerful pre-trained models

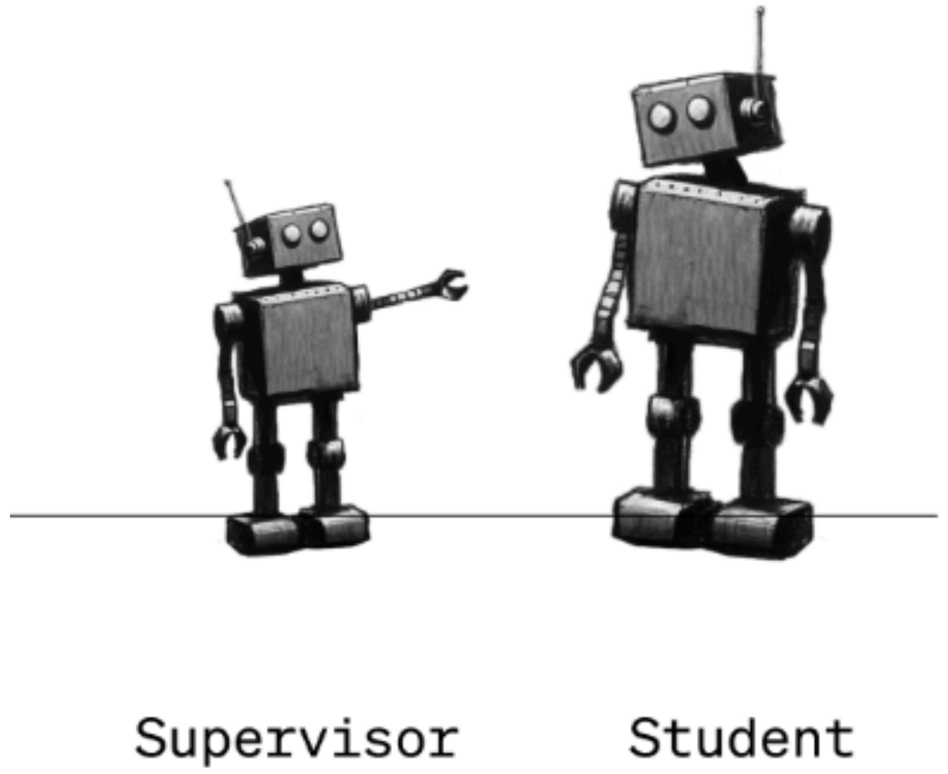
Gemini

Post-training

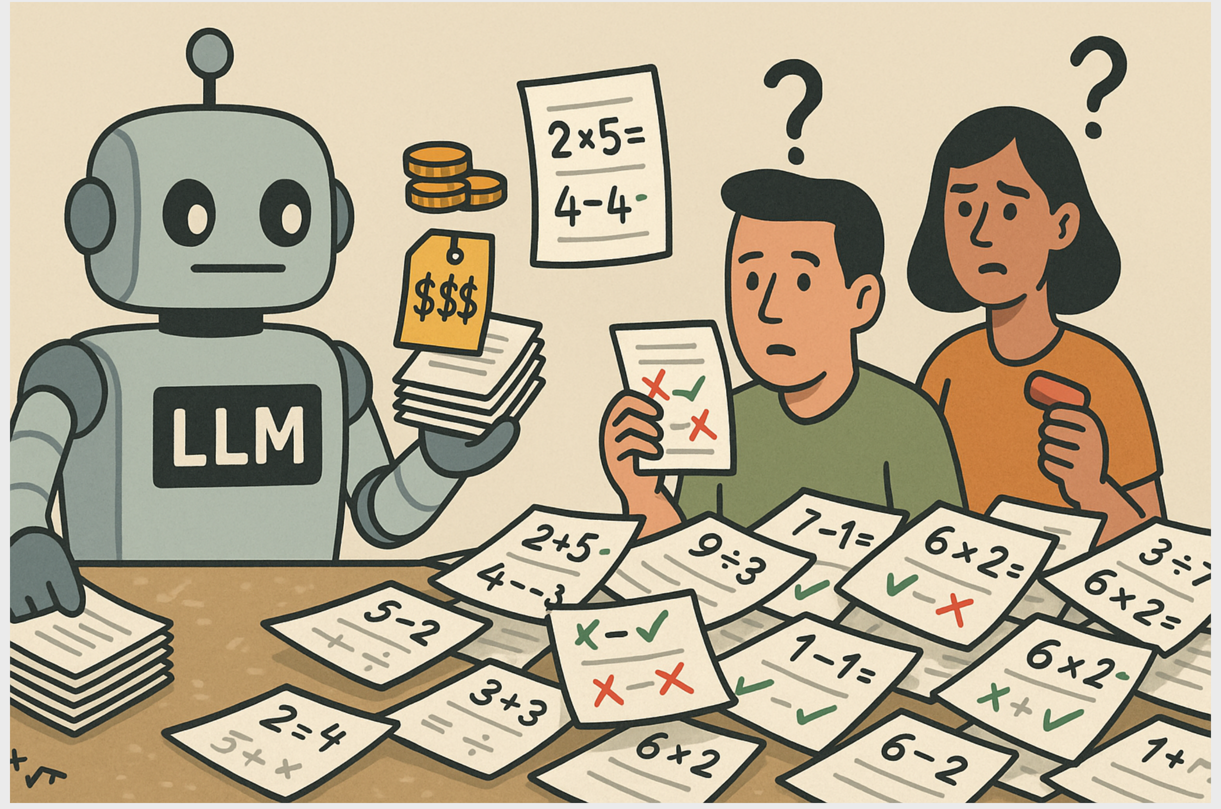
Specialized downstream tasks



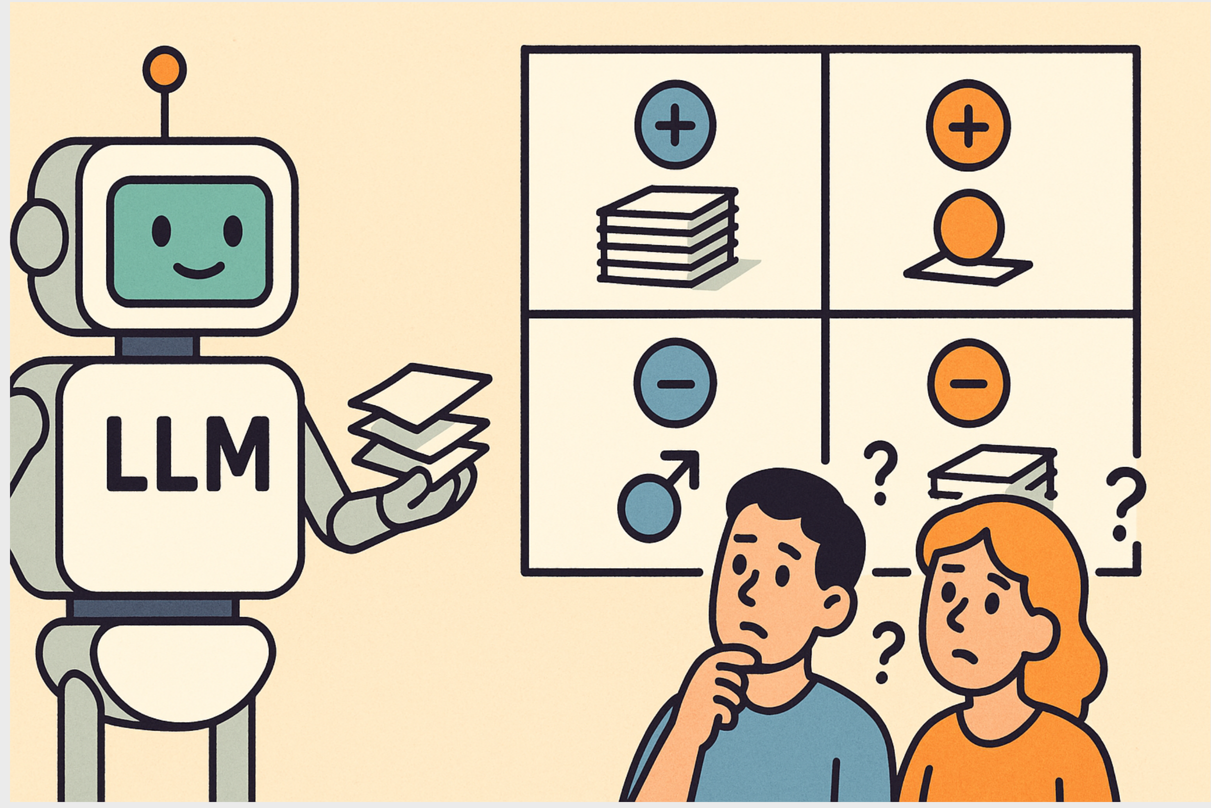
Weak-to-strong generalization



① … with **limited & noisy labels**



② … with **systematic bias**

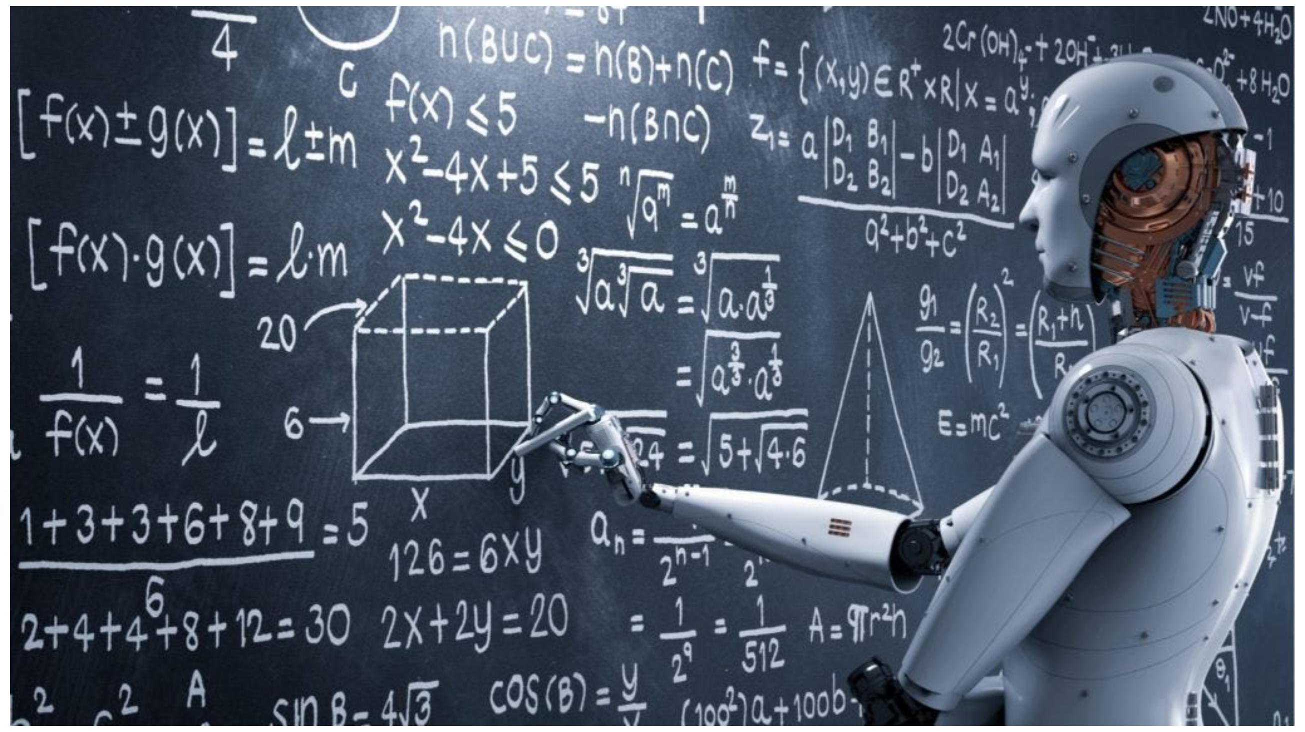


When and How Does W2S Emerge during Post-training?

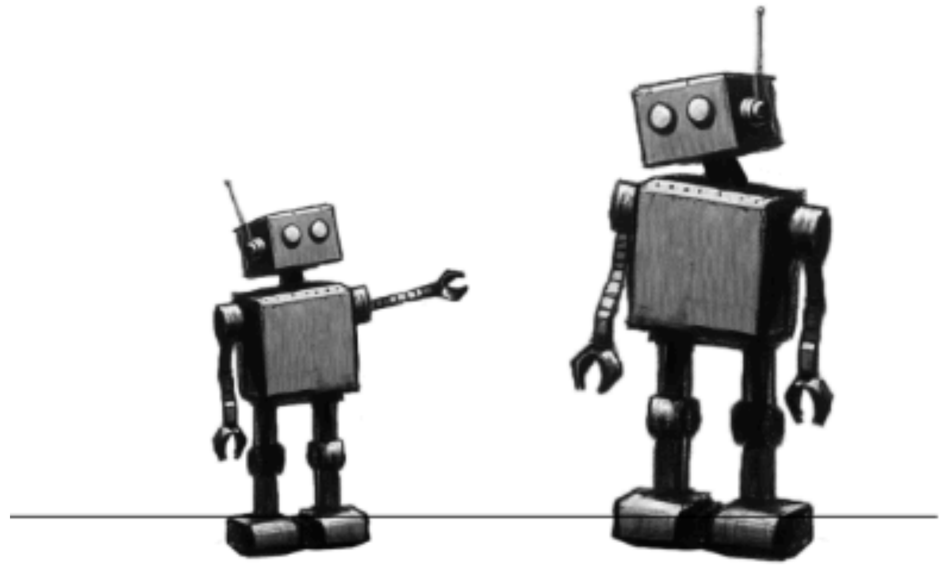
Powerful pre-trained models

Post-training

Specialized downstream tasks

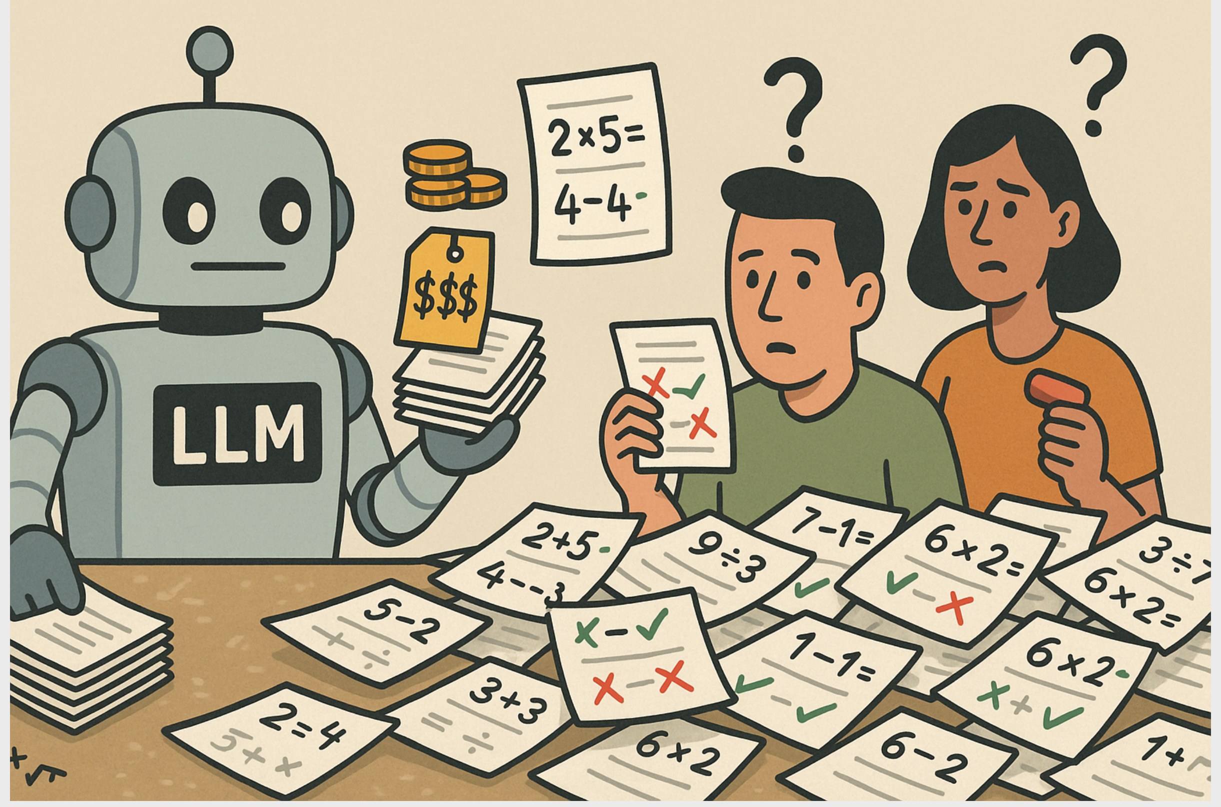


Weak-to-strong generalization

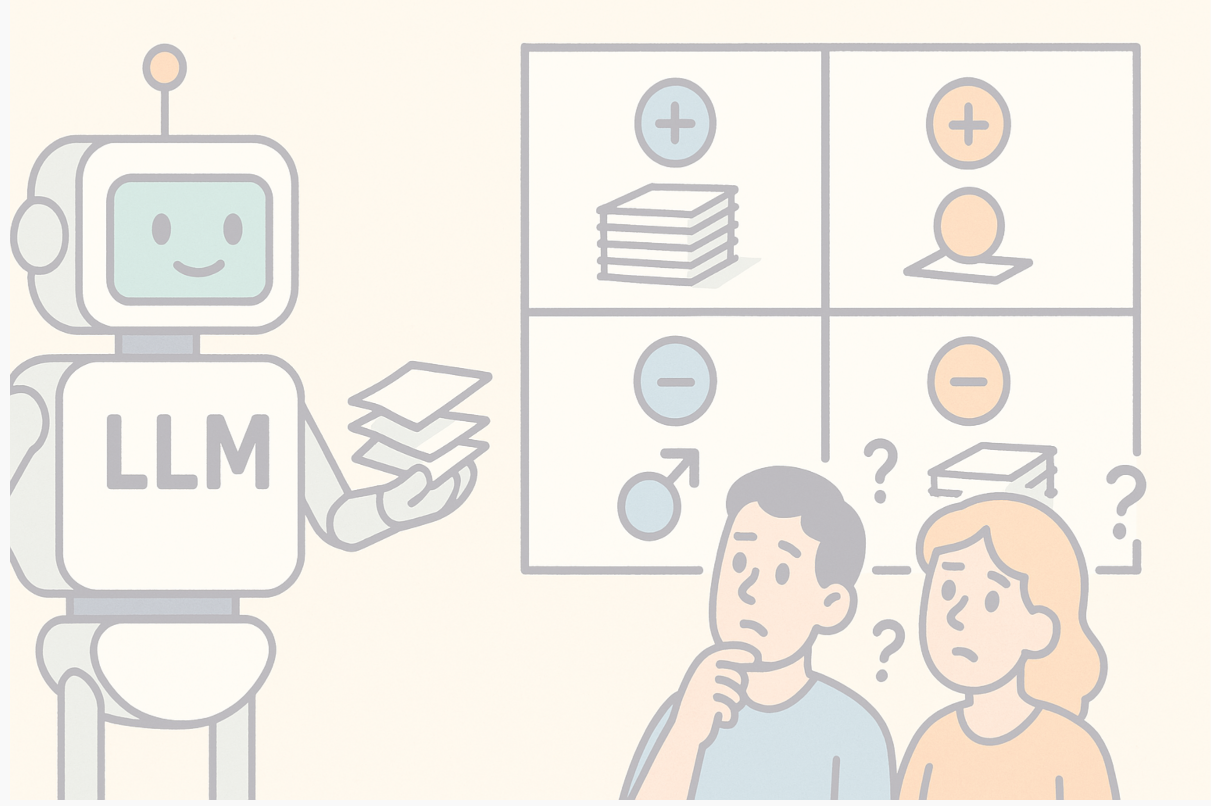


Supervisor Student

① … with **limited & noisy labels**



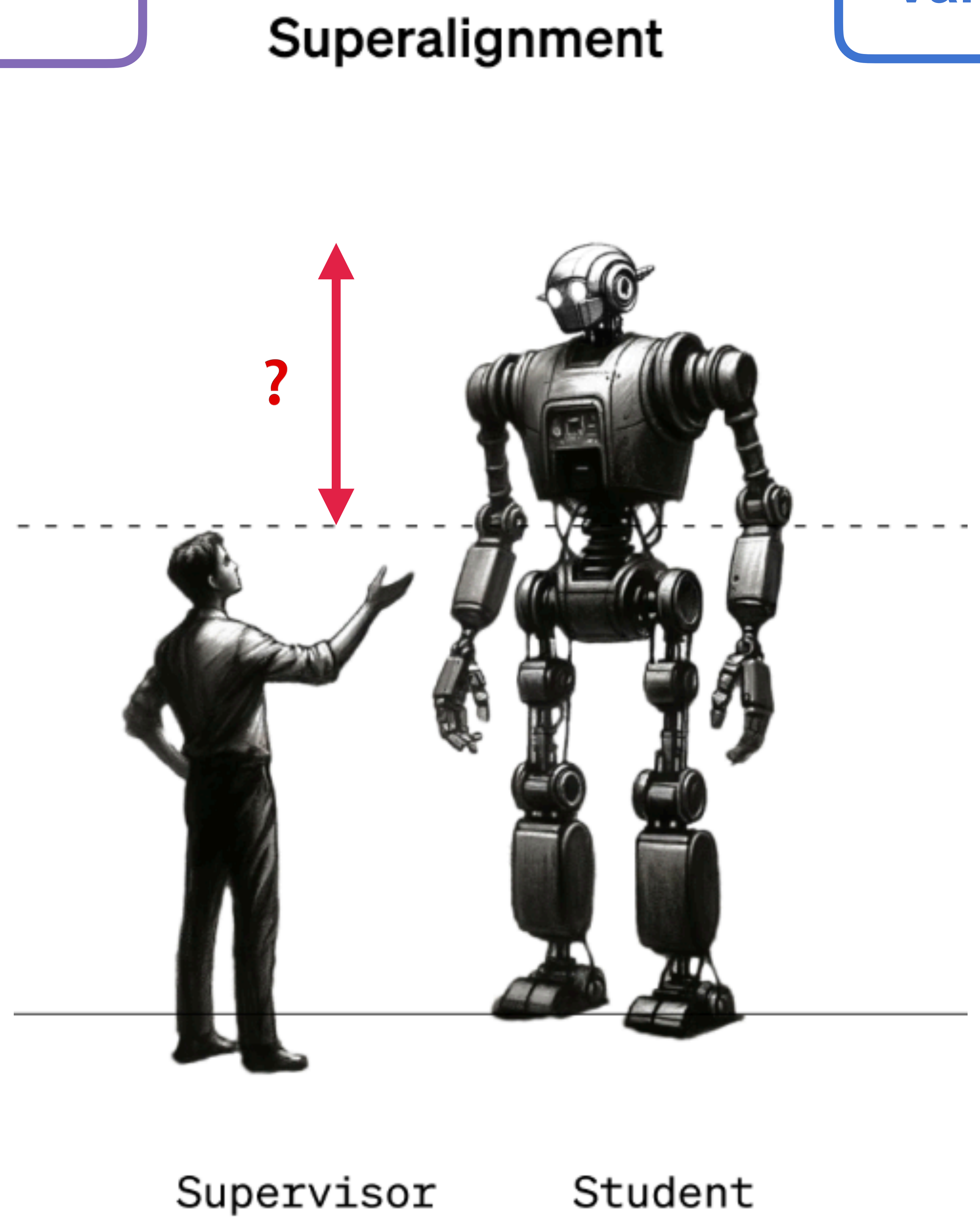
② … with **systematic bias**



Where Does W2S Gain Come From, Bias or Variance?

Bias associated with **expressivity**

Variance associated with **sample efficiency**

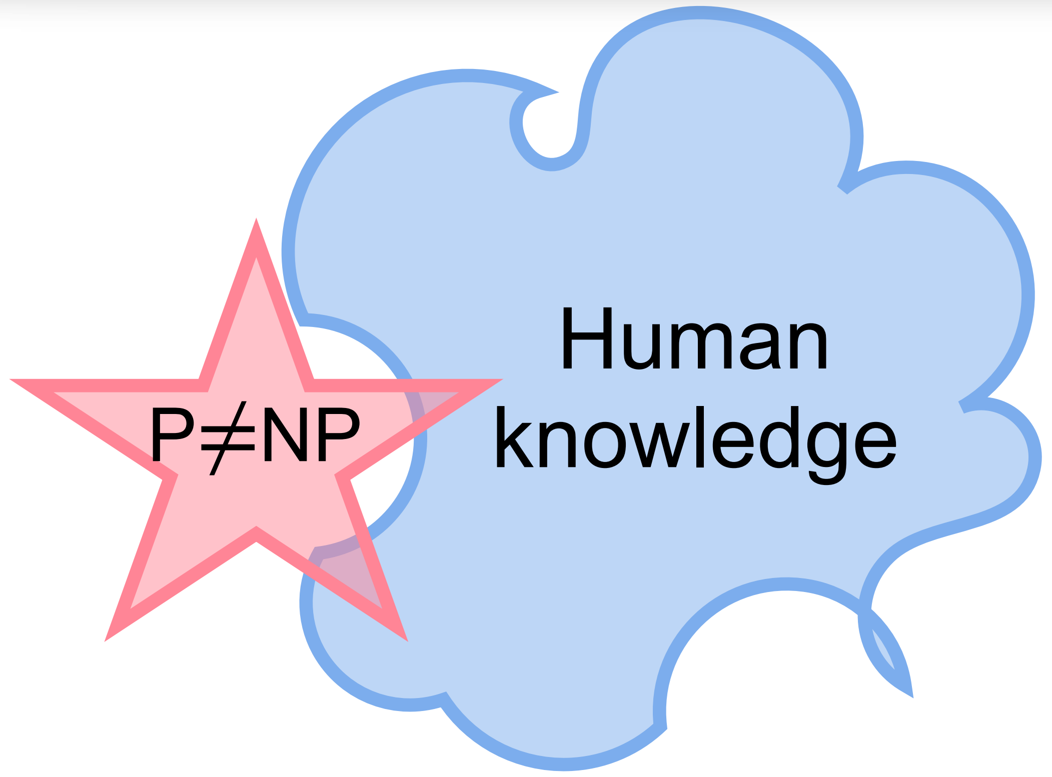


Where Does W2S Gain Come From, Bias or Variance?

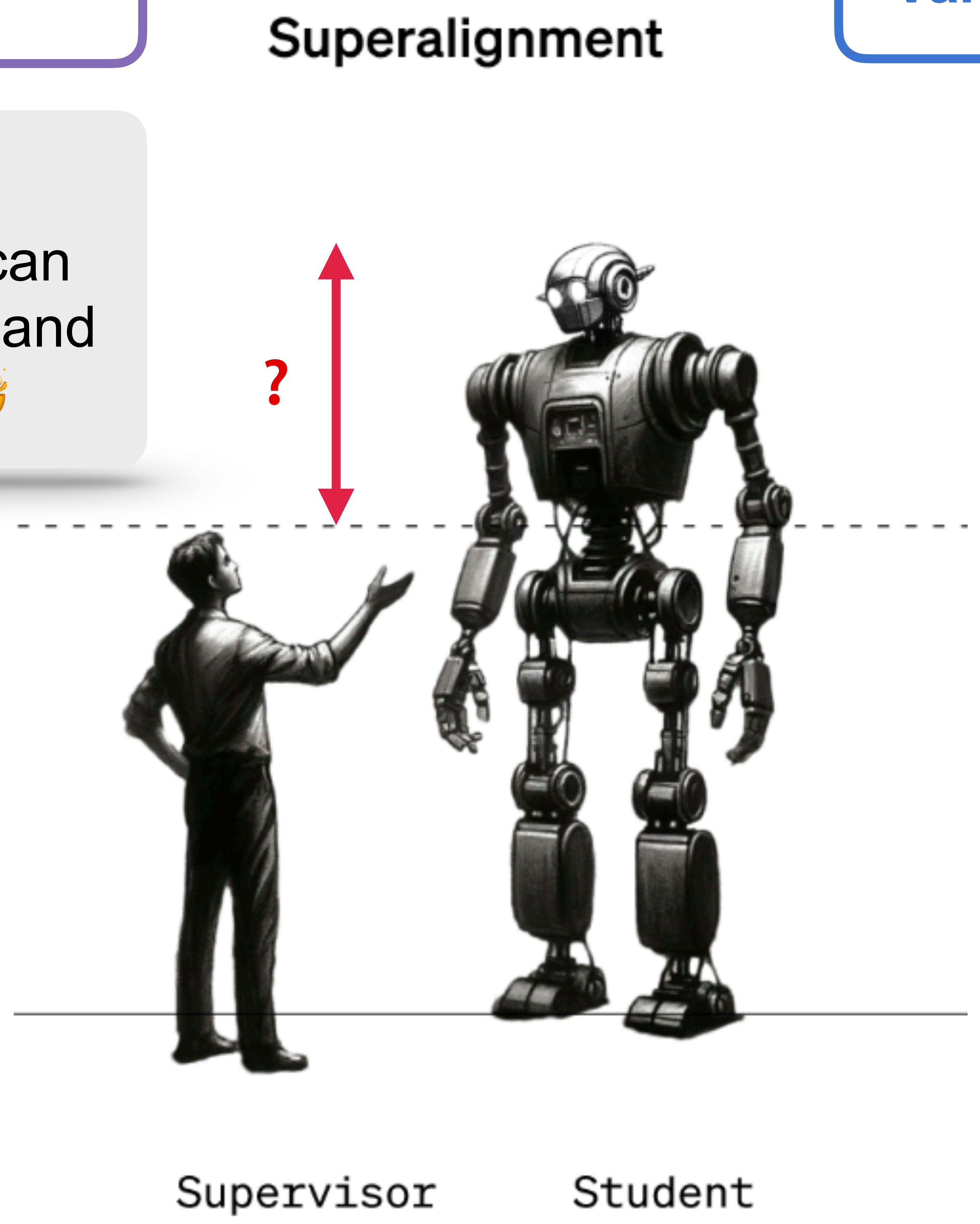
Bias associated with **expressivity**

Models beat human experts in bias

LLMs pre-trained on human knowledge can **extrapolate beyond human knowledge** and solve fundamentally open problems 🤯



Lang et al., (2024), Shin et al., (2024), Ildiz et al., (2024), Wu & Sahai, (2024), Medvedev et al., (2025), Mulgund & Pabbaraju, (2025), Xue et al., (2025), Oh et al., (2025) and more



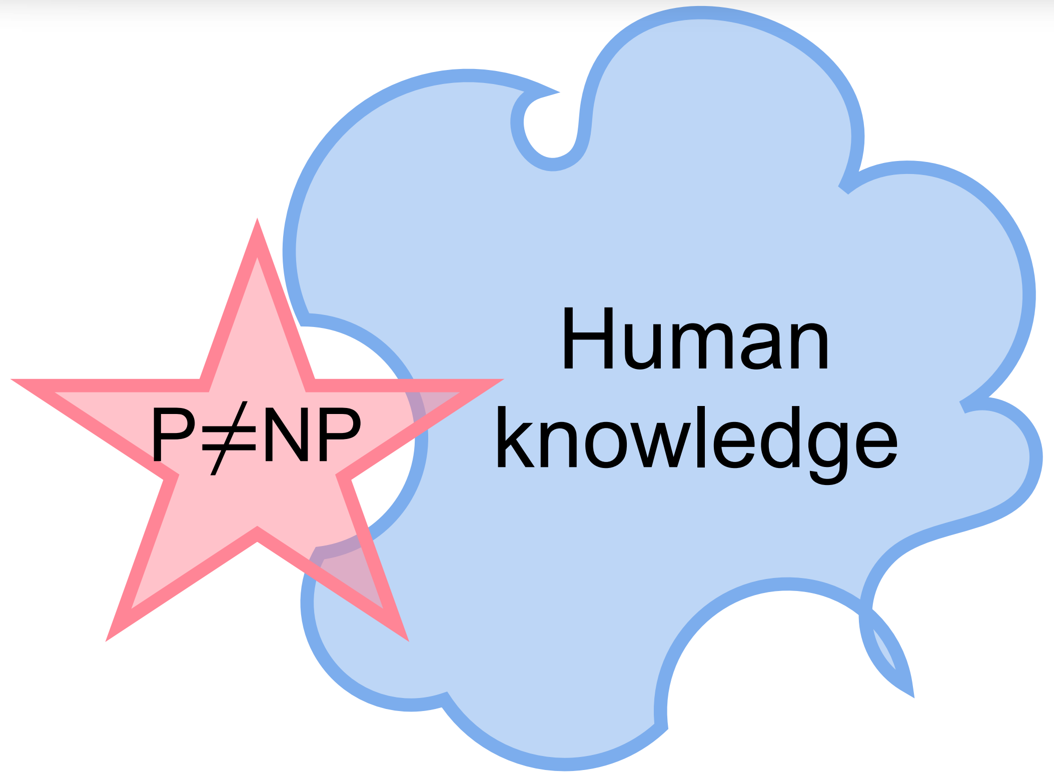
Variance associated with **sample efficiency**

Where Does W2S Gain Come From, Bias or Variance?

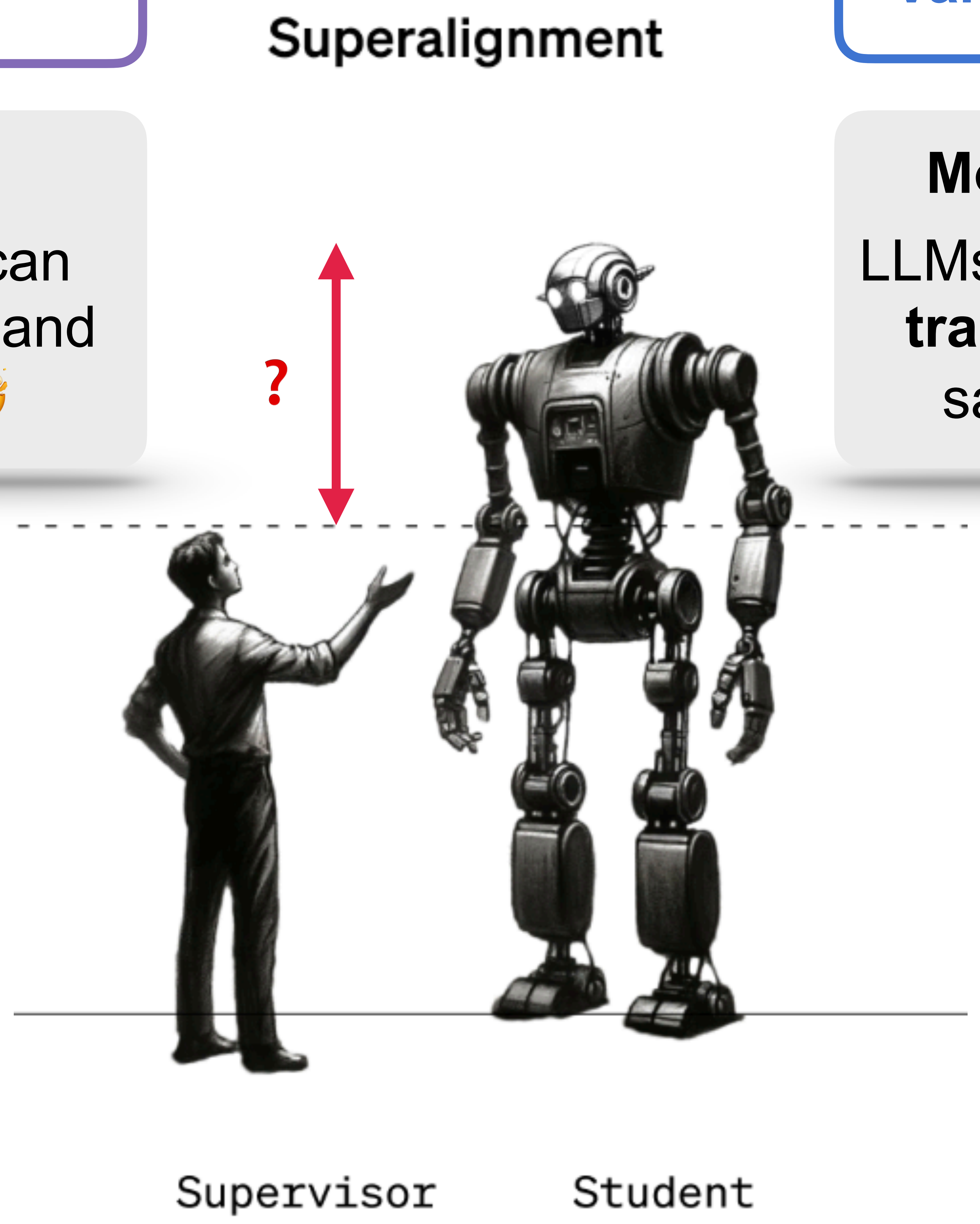
Bias associated with **expressivity**

Models beat human experts in bias

LLMs pre-trained on human knowledge can **extrapolate beyond human knowledge** and solve fundamentally open problems 🤯



Lang et al., (2024), Shin et al., (2024), Ildiz et al., (2024), Wu & Sahai, (2024), Medvedev et al., (2025), Mulgund & Pabbaraju, (2025), Xue et al., (2025), Oh et al., (2025) and more



Variance associated with **sample efficiency**

Models beat human experts in variance

LLMs learn to **find existing knowledge in pre-training** needed for solving a problem more sample-efficiently than human experts ✅

Burns et al., 2024 showed **better W2S generalization on "easier" tasks**, where both weak & strong models have negligible biases.

Where Does W2S Gain Come From, Bias or Variance?

Bias associated with **expressivity**

Superalignment

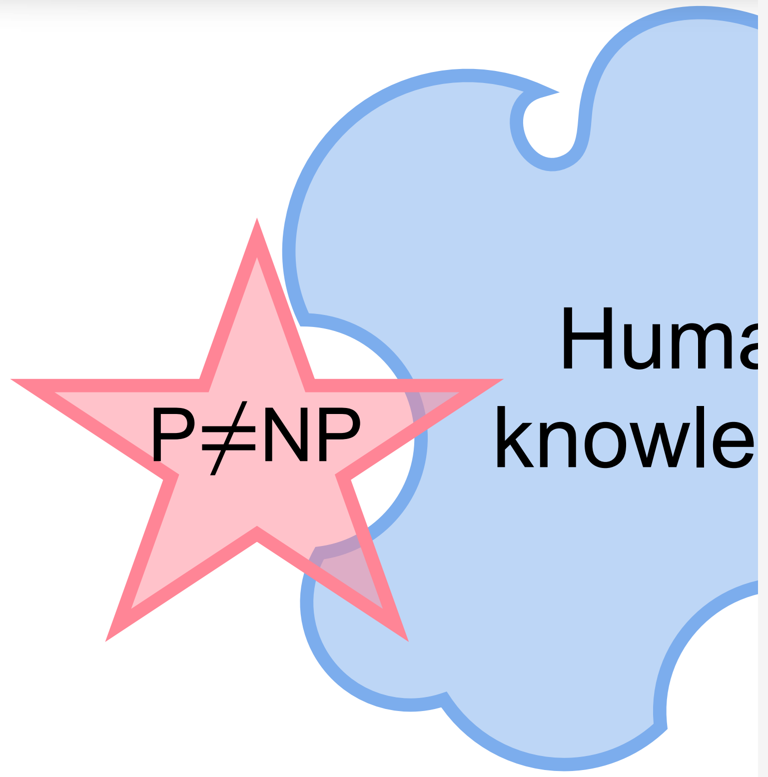
Variance associated with **sample efficiency**

Models beat human experts in bias

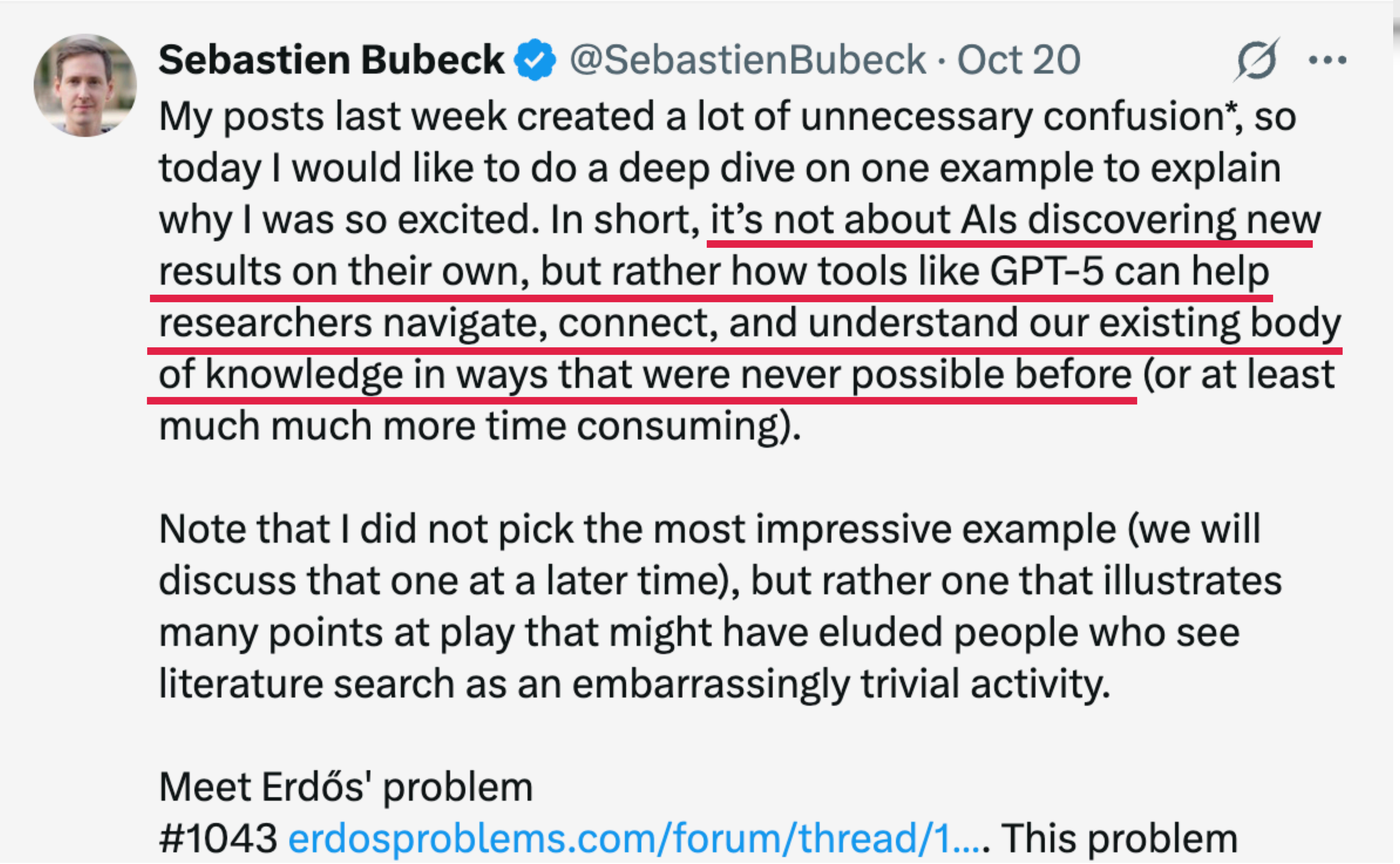
LLMs pre-trained on human knowledge can **extrapolate beyond human knowledge** and solve fundamentally open problems 🤯

Models beat human experts in variance

LLMs learn to **find existing knowledge in pre-training** needed for solving a problem more sample-efficiently than human experts ✅



Lang et al., (2024), S
(2024), Ildiz et al., (20
Sahai, (2024), Medve
(2025), Mulgund & Pa
(2025), Xue et al., (20
al., (2025) and n

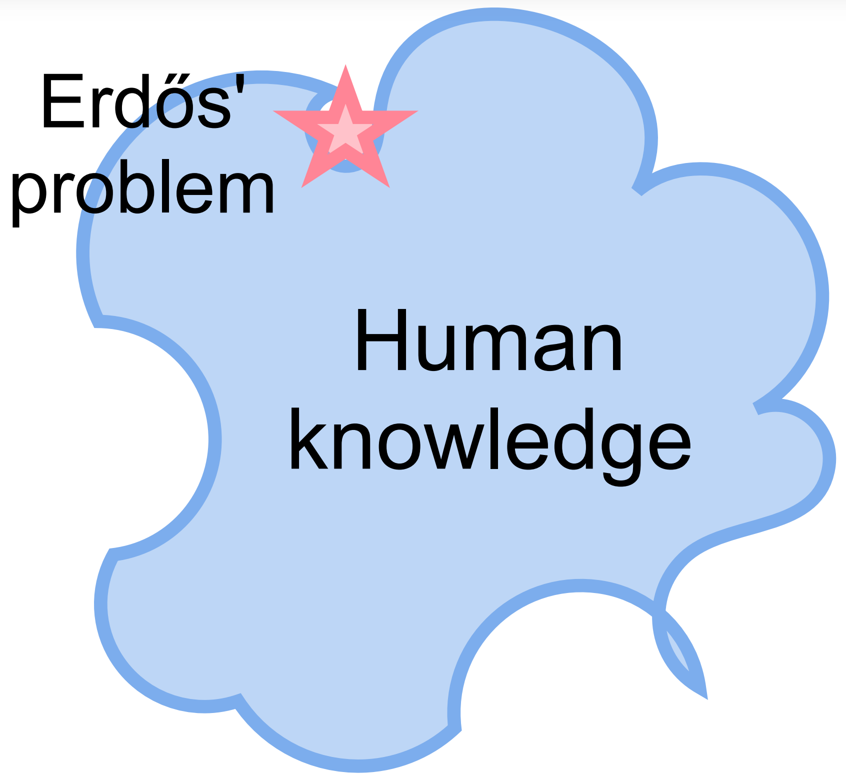


Sebastien Bubeck @SebastienBubeck · Oct 20

My posts last week created a lot of unnecessary confusion*, so today I would like to do a deep dive on one example to explain why I was so excited. In short, it's not about AIs discovering new results on their own, but rather how tools like GPT-5 can help researchers navigate, connect, and understand our existing body of knowledge in ways that were never possible before (or at least much much more time consuming).

Note that I did not pick the most impressive example (we will discuss that one at a later time), but rather one that illustrates many points at play that might have eluded people who see literature search as an embarrassingly trivial activity.

Meet Erdős' problem #1043 erdosproblems.com/forum/thread/1.... This problem



Burns et al., 2024 showed **better W2S generalization on "easier" tasks**, where both weak & strong models have negligible biases.

Where Does W2S Gain Come From, Bias or Variance?

Bias associated with **expressivity**

Models beat human experts in bias

LLMs pre-trained on human knowledge can **extrapolate beyond human knowledge** and solve fundamentally open problems 🤯

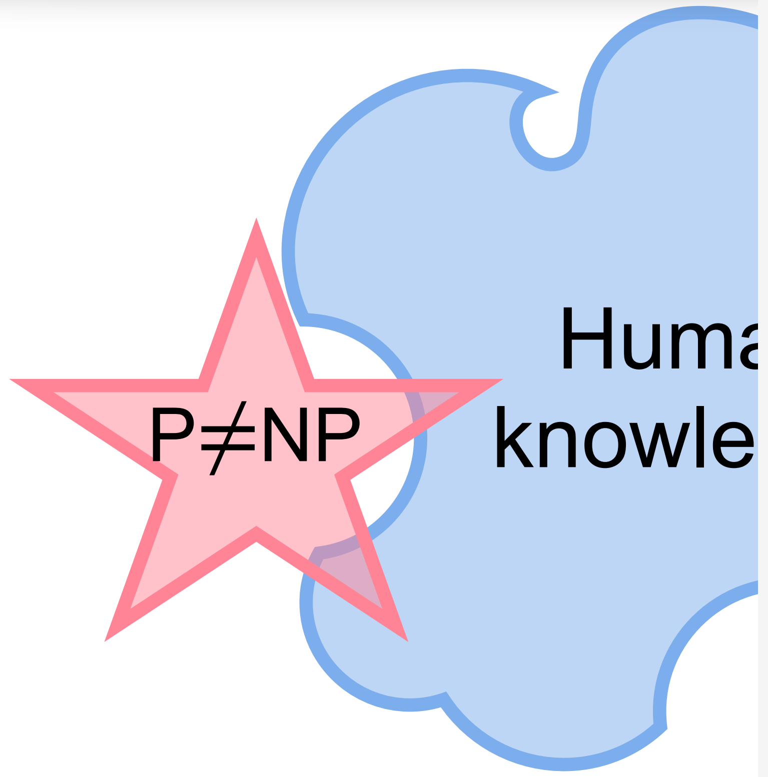
Superalignment

?

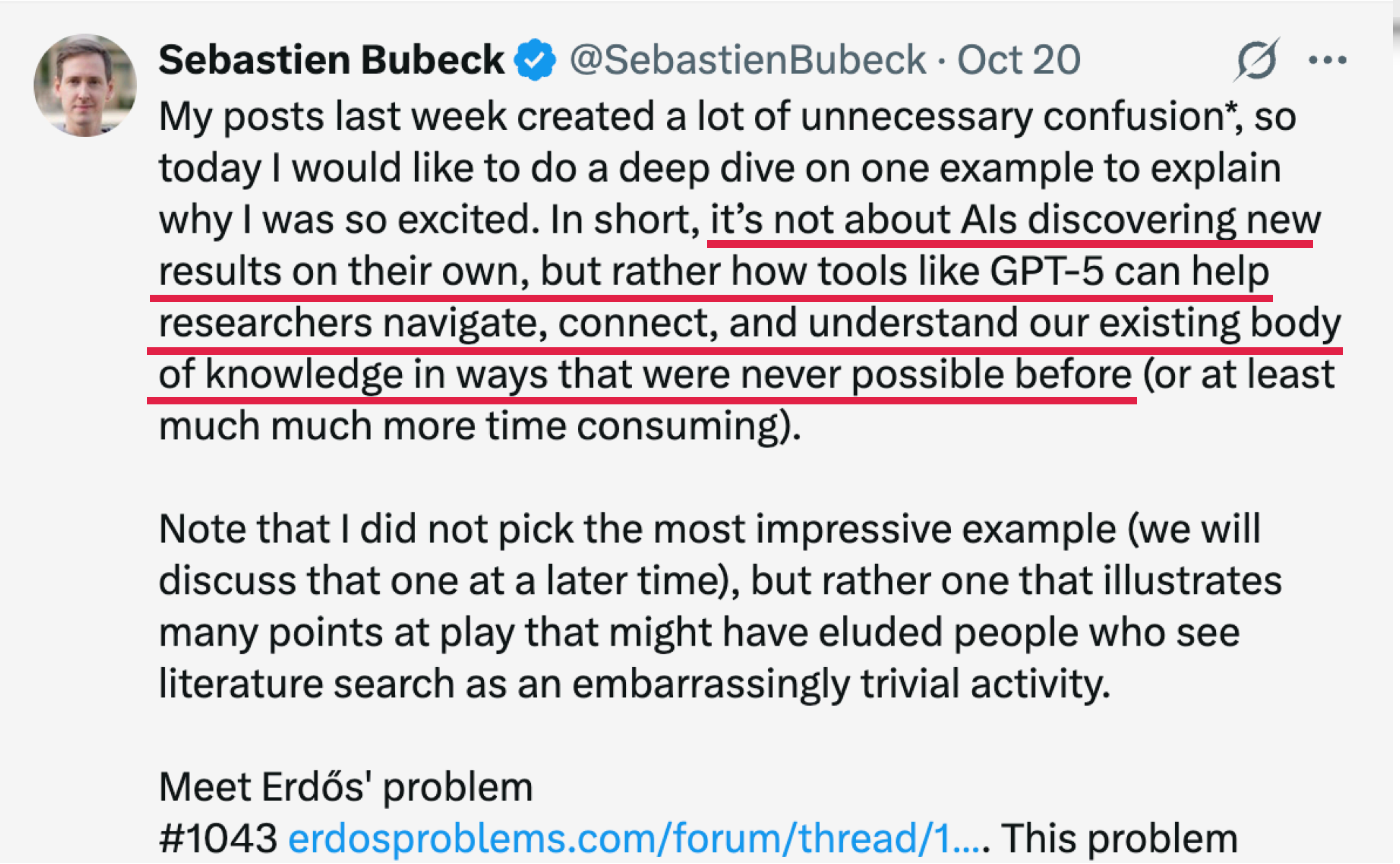
Variance associated with **sample efficiency**

Models beat human experts in variance

LLMs learn to **find existing knowledge in pre-training** needed for solving a problem more sample-efficiently than human experts ✅



Lang et al., (2024), S
(2024), Ildiz et al., (20
Sahai, (2024), Medve
(2025), Mulgund & Pa
(2025), Xue et al., (20
al., (2025) and n

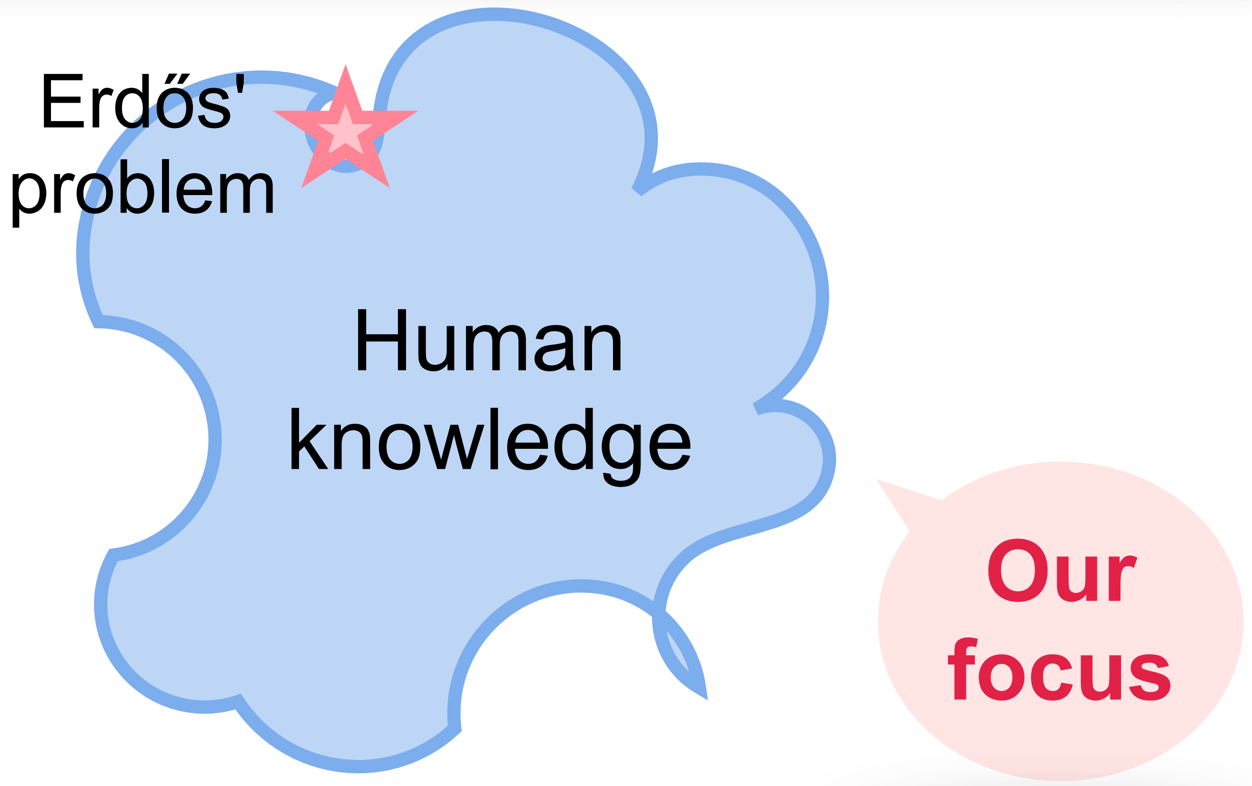


Sebastien Bubeck @SebastienBubeck · Oct 20

My posts last week created a lot of unnecessary confusion*, so today I would like to do a deep dive on one example to explain why I was so excited. In short, it's not about AIs discovering new results on their own, but rather how tools like GPT-5 can help researchers navigate, connect, and understand our existing body of knowledge in ways that were never possible before (or at least much much more time consuming).

Note that I did not pick the most impressive example (we will discuss that one at a later time), but rather one that illustrates many points at play that might have eluded people who see literature search as an embarrassingly trivial activity.

Meet Erdős' problem #1043 erdosproblems.com/forum/thread/1.... This problem



Burns et al., 2024 showed **better W2S generalization on "easier" tasks**, where both weak & strong models have negligible biases.

Weak vs. Strong: Model Capacity & Similarity

Fine-tuning searches for high-dimensional pre-trained features concentrated in low dimensions.

Weak teacher $f_w(x \mid \theta) = \phi_w(x)^\top \theta$ with $\phi_w : \mathcal{X} \to \mathbb{R}^D$

Strong student $f_s(x \mid \theta) = \phi_s(x)^\top \theta$ with $\phi_s : \mathcal{X} \to \mathbb{R}^D$

$x \sim \mathcal{D}$

$\Sigma_w = \mathbb{E}_x[\phi_w(x)\phi_w(x)^\top]$

$\Sigma_s = \mathbb{E}_x[\phi_s(x)\phi_s(x)^\top]$

Weak vs. Strong: Model Capacity & Similarity

Fine-tuning searches for high-dimensional pre-trained features concentrated in low dimensions.

Weak teacher $f_w(x \mid \theta) = \phi_w(x)^\top \theta$ with $\phi_w : \mathcal{X} \to \mathbb{R}^D$

Strong student $f_s(x \mid \theta) = \phi_s(x)^\top \theta$ with $\phi_s : \mathcal{X} \to \mathbb{R}^D$

$x \sim \mathcal{D}$

$\Sigma_w = \mathbb{E}_x[\phi_w(x)\phi_w(x)^\top]$

$\Sigma_s = \mathbb{E}_x[\phi_s(x)\phi_s(x)^\top]$

Representation efficiency—
low intrinsic dimensions

$\mathrm{rank}(\Sigma_w) = d_w \ll D$, $\mathrm{rank}(\Sigma_s) = d_s \ll D$

Stronger model ⇒ better efficiency: $d_s < d_w$

Weak vs. Strong: Model Capacity & Similarity

Fine-tuning searches for high-dimensional pre-trained features concentrated in low dimensions.

Weak teacher $f_w(x \mid \theta) = \phi_w(x)^\top \theta$ with $\phi_w : \mathcal{X} \to \mathbb{R}^D$

Strong student $f_s(x \mid \theta) = \phi_s(x)^\top \theta$ with $\phi_s : \mathcal{X} \to \mathbb{R}^D$

$x \sim \mathcal{D}$

$\Sigma_w = \mathbb{E}_x[\phi_w(x)\phi_w(x)^\top]$

$\Sigma_s = \mathbb{E}_x[\phi_s(x)\phi_s(x)^\top]$

Representation efficiency— **low intrinsic dimensions**

$\mathrm{rank}(\Sigma_w) = d_w \ll D$, $\mathrm{rank}(\Sigma_s) = d_s \ll D$

Stronger model ⇒ better efficiency: $d_s < d_w$

Representation similarity — **correlation dimension** $d_{s \wedge w}$

Spectral decomposition: $\underset{}{\Sigma_w} = \underset{D \times d_w}{V_w} \underset{d_w \times d_w}{\Lambda_w} V_w^\top$, $\Sigma_s = \underset{D \times d_s}{V_s} \underset{d_s \times d_s}{\Lambda_s} V_s^\top$

Let $d_{s \wedge w} = \|V_s^\top V_w\|_F^2$ such that $0 \le d_{s \wedge w} \le \min\{d_s, d_w\}$

Weak vs. Strong: Model Capacity & Similarity

Fine-tuning searches for high-dimensional pre-trained features concentrated in low dimensions.

Weak teacher $f_w(x \mid \theta) = \phi_w(x)^\top \theta$ with $\phi_w : \mathcal{X} \to \mathbb{R}^D$

Strong student $f_s(x \mid \theta) = \phi_s(x)^\top \theta$ with $\phi_s : \mathcal{X} \to \mathbb{R}^D$

$x \sim \mathcal{D}$

$\Sigma_w = \mathbb{E}_x[\phi_w(x)\phi_w(x)^\top]$

$\Sigma_s = \mathbb{E}_x[\phi_s(x)\phi_s(x)^\top]$

Representation efficiency— **low intrinsic dimensions**

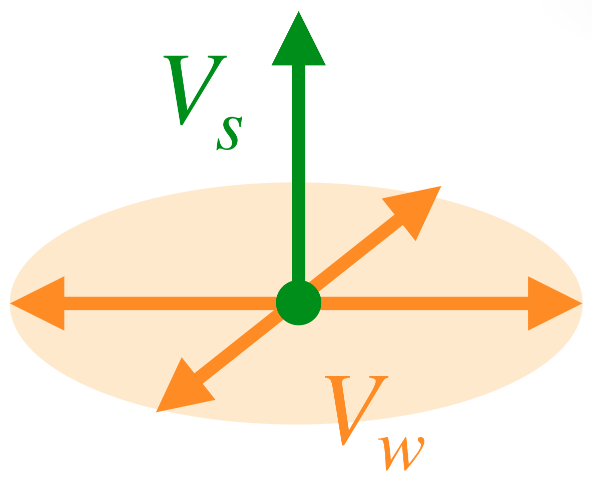
$\mathrm{rank}(\Sigma_w) = d_w \ll D$, $\mathrm{rank}(\Sigma_s) = d_s \ll D$

Stronger model ⇒ better efficiency: $d_s < d_w$

Representation similarity — **correlation dimension $d_{s \wedge w}$**

Spectral decomposition: $\Sigma_w = \underset{D \times d_w}{V_w} \ \underset{d_w \times d_w}{\Lambda_w} \ V_w^\top$, $\Sigma_s = \underset{D \times d_s}{V_s} \ \underset{d_s \times d_s}{\Lambda_s} \ V_s^\top$

Let $d_{s \wedge w} = \|V_s^\top V_w\|_F^2$ such that $0 \le d_{s \wedge w} \le \min\{d_s, d_w\}$



Large discrepancy: $d_{s \wedge w} = 0$

Weak vs. Strong: Model Capacity & Similarity

Fine-tuning searches for high-dimensional pre-trained features concentrated in low dimensions.

Weak teacher $f_w(x \mid \theta) = \phi_w(x)^\top \theta$ with $\phi_w : \mathcal{X} \to \mathbb{R}^D$

Strong student $f_s(x \mid \theta) = \phi_s(x)^\top \theta$ with $\phi_s : \mathcal{X} \to \mathbb{R}^D$

$x \sim \mathcal{D}$

$$\Sigma_w = \mathbb{E}_x[\phi_w(x)\phi_w(x)^\top]$$

$$\Sigma_s = \mathbb{E}_x[\phi_s(x)\phi_s(x)^\top]$$

Representation efficiency— **low intrinsic dimensions**

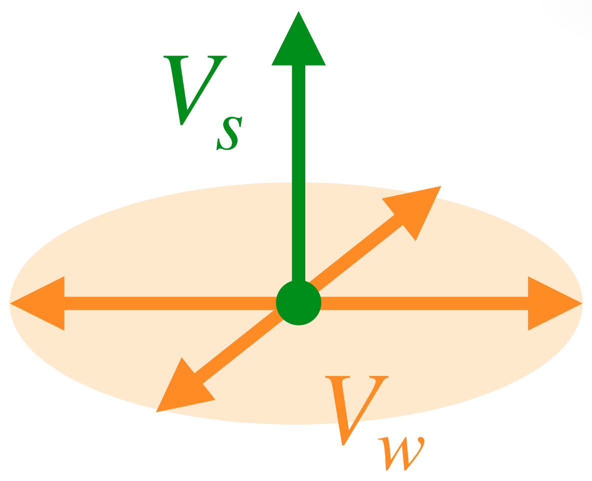
$$\mathrm{rank}(\Sigma_w) = d_w \ll D, \ \mathrm{rank}(\Sigma_s) = d_s \ll D$$

Stronger model ⇒ better efficiency: $d_s < d_w$

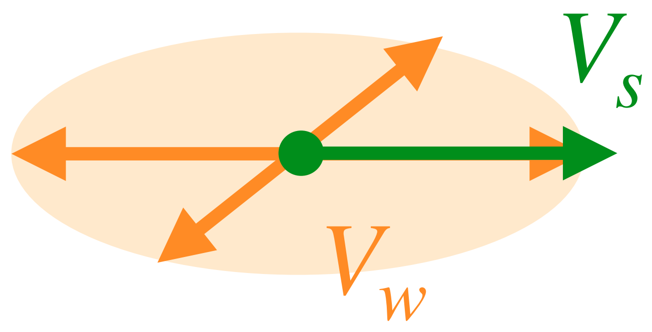
Representation similarity — **correlation dimension** $d_{s \wedge w}$

Spectral decomposition: $\Sigma_w = \underset{D \times d_w}{V_w} \ \underset{d_w \times d_w}{\Lambda_w} \ V_w^\top$, $\Sigma_s = \underset{D \times d_s}{V_s} \ \underset{d_s \times d_s}{\Lambda_s} \ V_s^\top$

Let $d_{s \wedge w} = \| V_s^\top V_w \|_F^2$ such that $0 \le d_{s \wedge w} \le \min\{d_s, d_w\}$



Large discrepancy: $d_{s \wedge w} = 0$



Small discrepancy: $d_{s \wedge w} = d_s < d_w$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set: $(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set:
$X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set: $(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set: $X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Ridgeless regression, $\alpha_w, \alpha_{w2s} \to 0$, is nearly optimal.

$\mathbb{E}[\mathrm{ER}(f)] =$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0,\sigma^2)$.

Labeled (small) set: $(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set: $X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Ridgeless regression, $\alpha_w, \alpha_{w2s} \to 0$, is nearly optimal.

$$\mathbb{E}[\mathrm{ER}(f)] = \underbrace{\mathbb{E}_{X,y}\left[\frac{1}{N} \|f(X) - \mathbb{E}_{y|X}[f(X)]\|_2^2\right]}_{\mathrm{Var}(f) \text{ induced by } \sigma^2} +$$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set:
$X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Ridgeless regression, $\alpha_w, \alpha_{w2s} \to 0$, is nearly optimal.

$$\mathbb{E}[\mathrm{ER}(f)] = \underbrace{\mathbb{E}_{X,y}\left[\frac{1}{N}\|f(X) - \mathbb{E}_{y|X}[f(X)]\|_2^2\right]}_{\mathrm{Var}(f) \text{ induced by } \sigma^2} + \underbrace{\mathbb{E}_X\left[\frac{1}{N}\|\mathbb{E}_{y|X}[f(X)] - f_*(X)\|_2^2\right]}_{\mathrm{Bias}(f) \text{ induced by suboptimal } \phi_w,\, \phi_s}$$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set: $(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set: $X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Ridgeless regression, $\alpha_w, \alpha_{w2s} \to 0$, is nearly optimal.

$$\mathbb{E}[\mathrm{ER}(f)] = \underbrace{\mathbb{E}_{X,y}\left[\frac{1}{N}\|f(X) - \mathbb{E}_{y|X}[f(X)]\|_2^2\right]}_{\mathrm{Var}(f) \text{ induced by } \sigma^2} + \underbrace{\mathbb{E}_X\left[\frac{1}{N}\|\mathbb{E}_{y|X}[f(X)] - f_*(X)\|_2^2\right]}_{\mathrm{Bias}(f) \text{ induced by suboptimal } \phi_w,\, \phi_s}$$

W2S Finetuning as Kernel Regression: Finite Samples

Learn an unknown $f_* : \mathcal{X} \to \mathbb{R}$ for a distribution $(x, y) \sim \mathcal{D}(f_*)$ s.t. $y = f_*(x) + z$, $z \sim \mathcal{N}(0, \sigma^2)$.

Labeled (small) set: $(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(f_*)^n$

Weak teacher $f_w(x) = \phi_w(x)^\top \theta_w$: $\theta_w = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|\phi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled (large) set: $X \sim \mathcal{D}_x(f_*)^N$

W2S $f_s(x) = \phi_s(x)^\top \theta_{w2s}$: $\theta_{w2s} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{N} \|\phi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Ridgeless regression, $\alpha_w, \alpha_{w2s} \to 0$, is nearly optimal.

$$\mathbb{E}[\mathrm{ER}(f)] = \underbrace{\mathbb{E}_{X,y}\left[\frac{1}{N}\|f(X) - \mathbb{E}_{y|X}[f(X)]\|_2^2\right]}_{\mathrm{Var}(f) \text{ induced by } \sigma^2} + \underbrace{\mathbb{E}_X\left[\frac{1}{N}\|\mathbb{E}_{y|X}[f(X)] - f_*(X)\|_2^2\right]}_{\mathrm{Bias}(f) \text{ induced by suboptimal } \phi_w, \phi_s}$$

Precise $\mathrm{Var}(f_s)$ and $\mathrm{Var}(f_w)$?

Larger Discrepancy (Lower $d_{s\wedge w}$) $\rightarrow$ Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

Larger Discrepancy (Lower $d_{s\wedge w}$) → Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

$$\mathrm{Var}(f_s) \asymp \frac{d_{s\wedge w}}{n} + \frac{d_s}{N} \cdot \frac{d_w - d_{s\wedge w}}{n}$$

$\frac{d_{s\wedge w}}{n}$: $\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)$

$\frac{d_s}{N}$: **W2S**

$\frac{d_w - d_{s\wedge w}}{n}$: $\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)$

Larger Discrepancy (Lower $d_{s\wedge w}$) $\to$ Better W2S

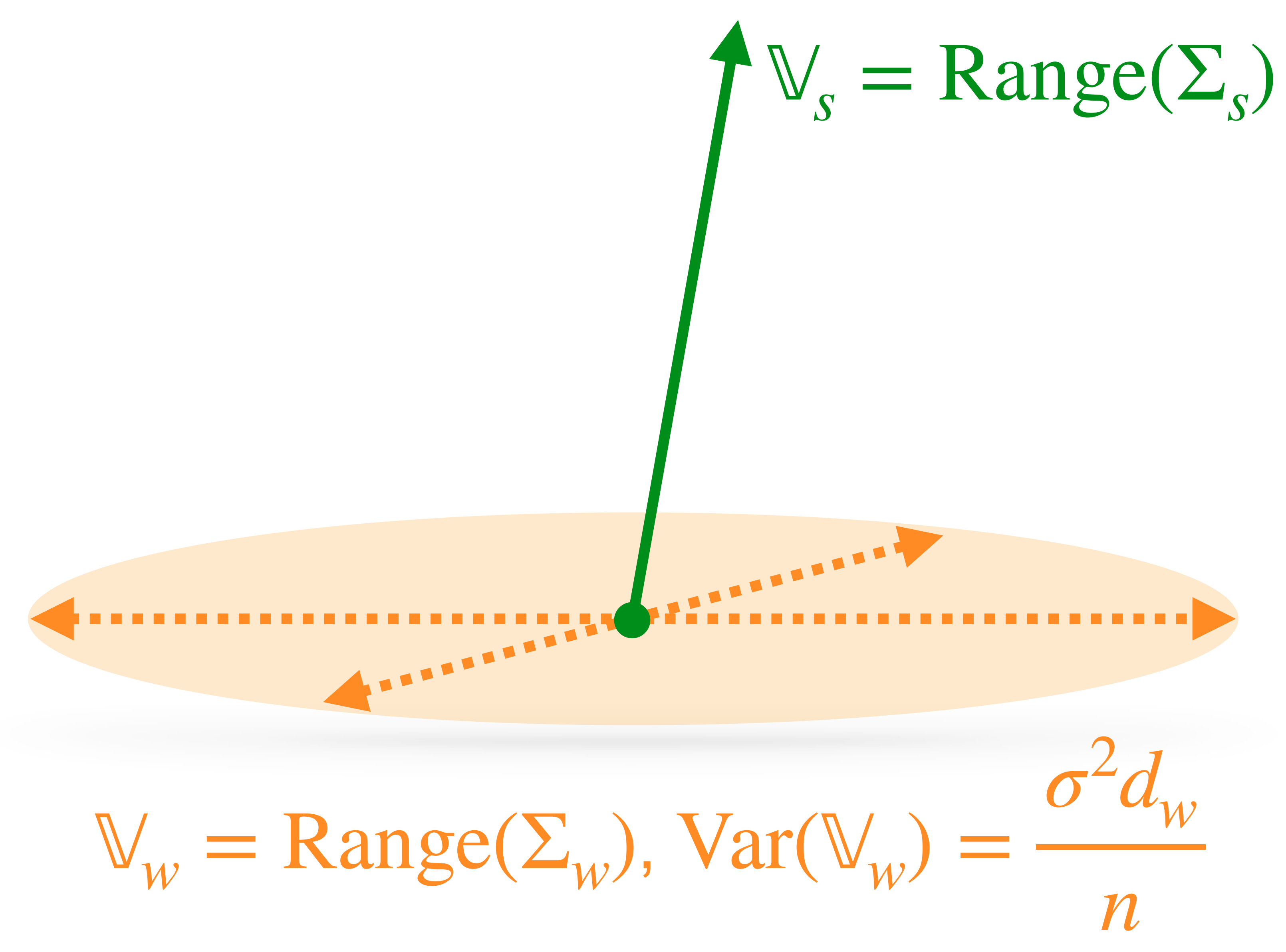
Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\textbf{W2S}} \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) $\rightarrow$ Better W2S

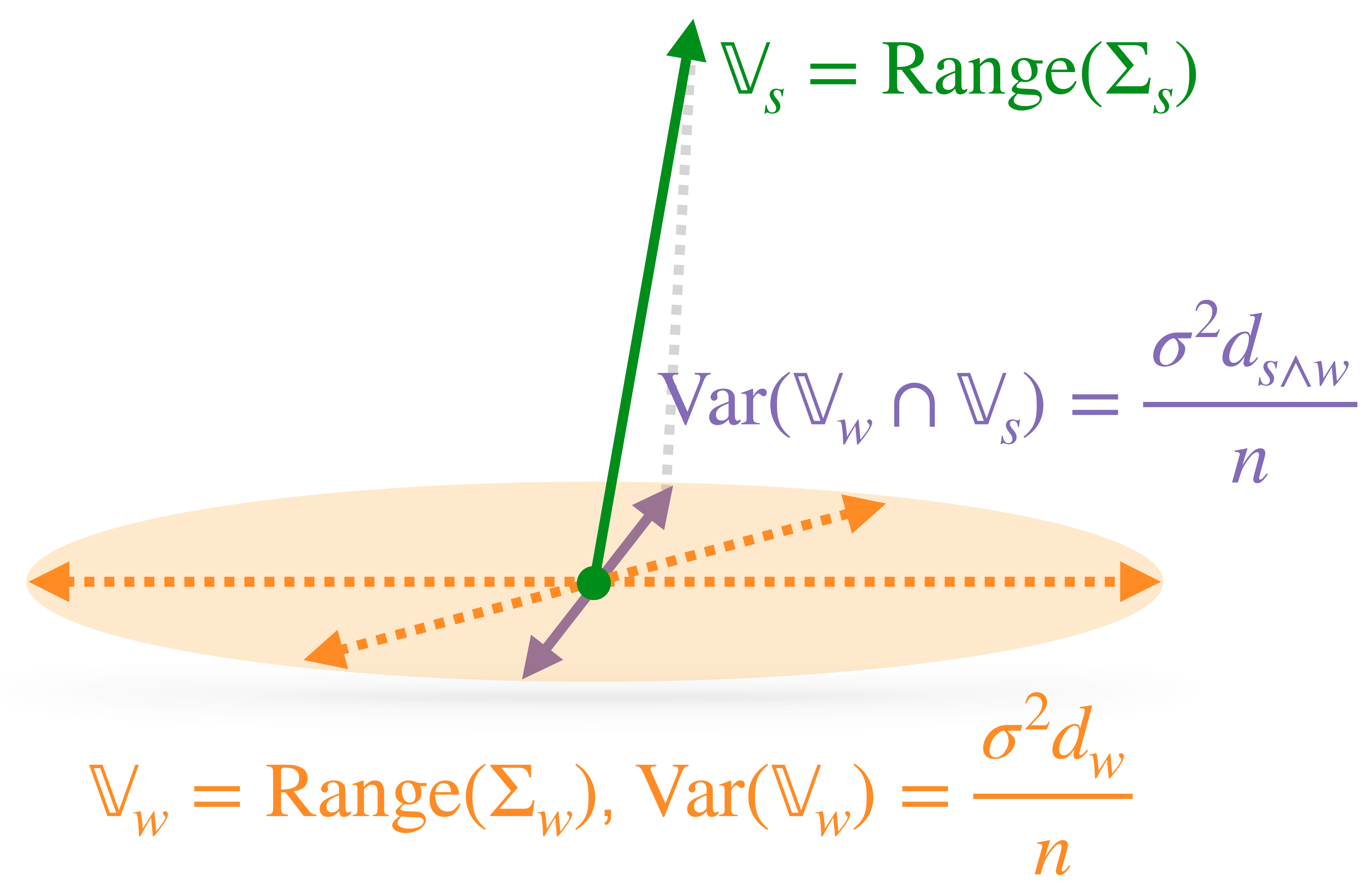
Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\text{W2S}} \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) $\to$ Better W2S

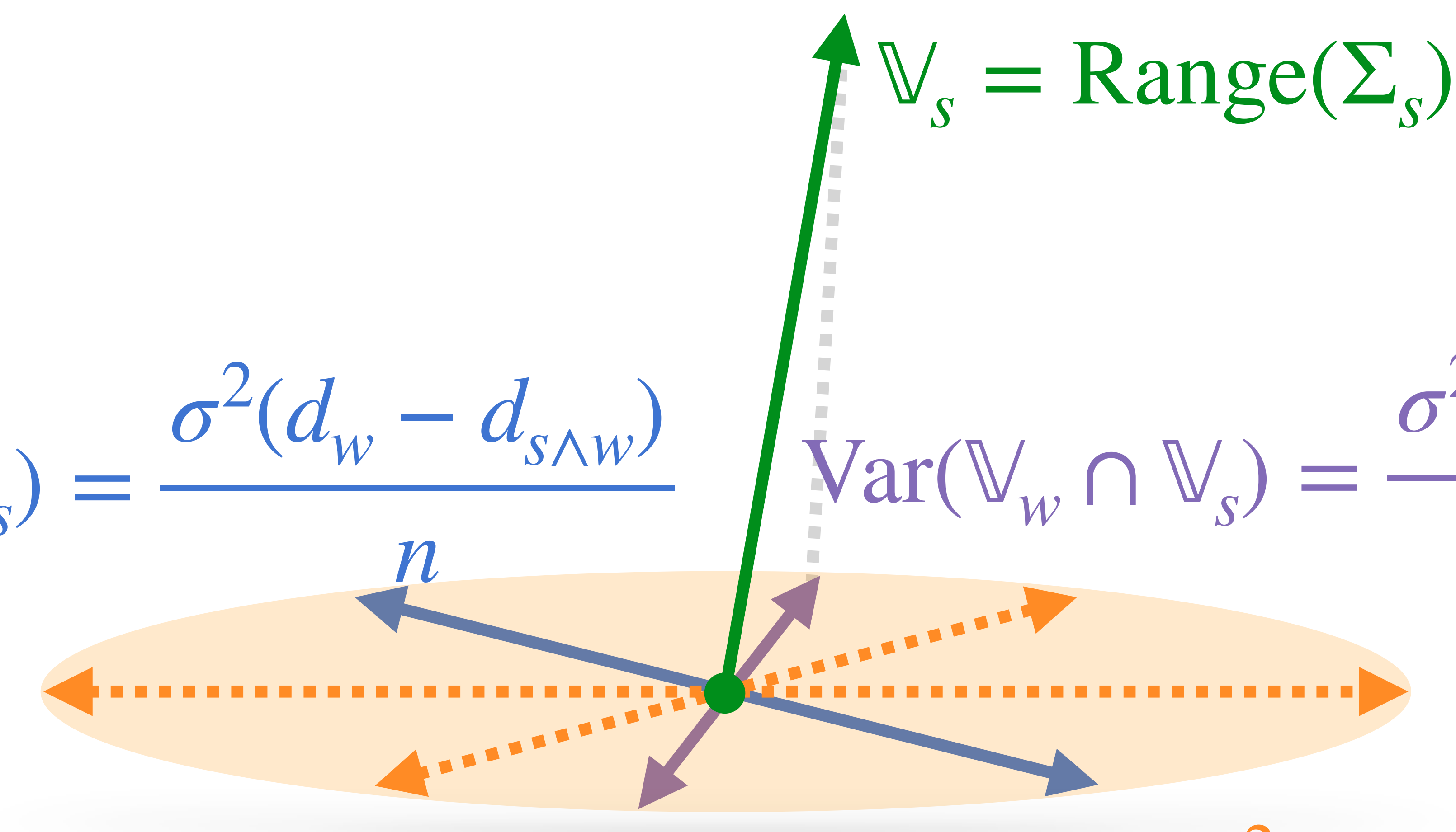
Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\text{W2S}} \; \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s \wedge w}$) $\to$ Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

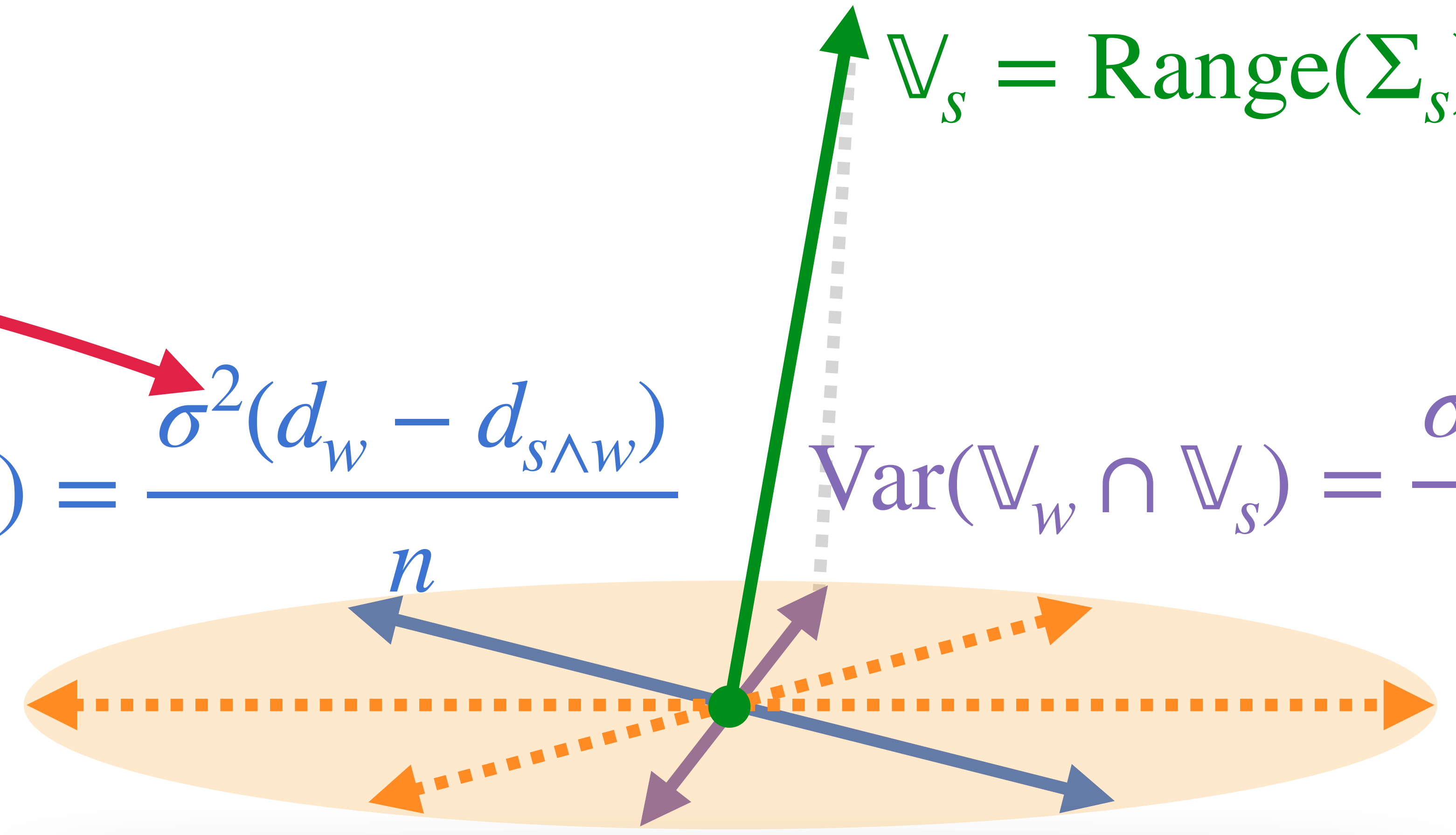
$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s \wedge w} + \frac{d_s}{N}(d_w - d_{s \wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

Variance reduction in W2S: $\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)$ vanishes as $d_s / N \to 0$

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s \wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\text{W2S}} \underbrace{\frac{d_w - d_{s \wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) → Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

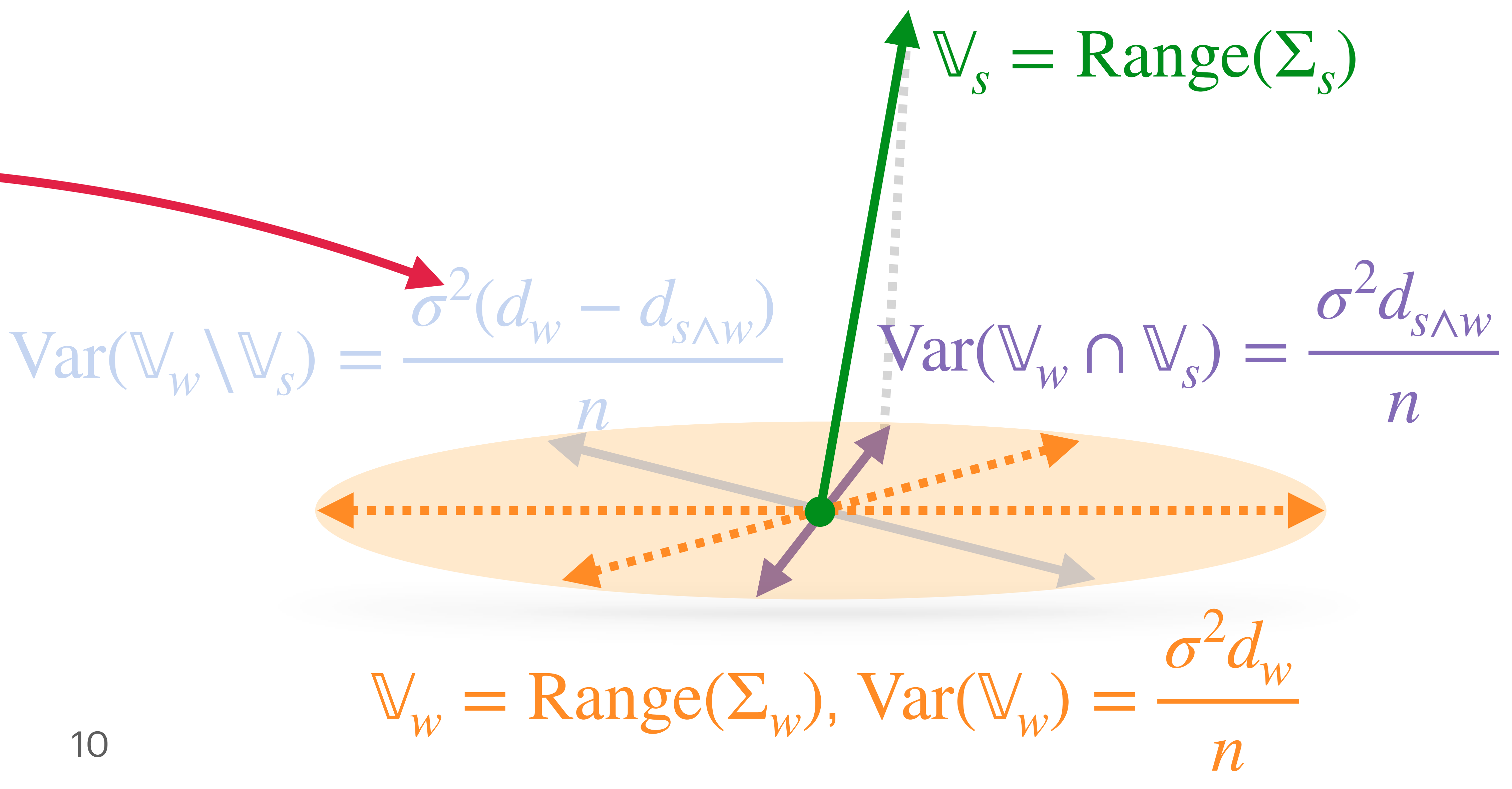
$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

Variance reduction in W2S: $\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)$ vanishes as $d_s/N \to 0$

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\textbf{W2S}} \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \setminus \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) → Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

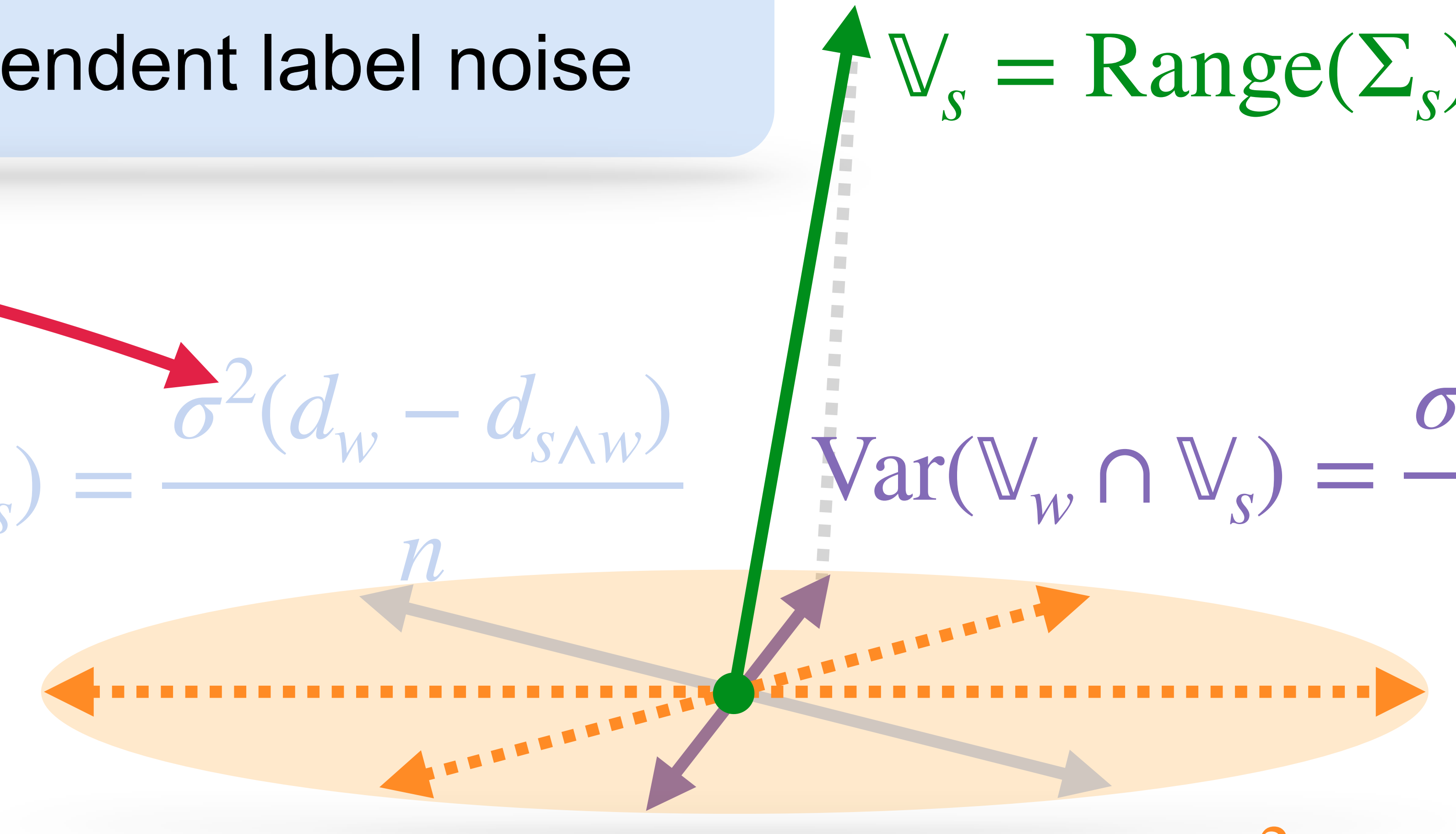
Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

Variance reduction in W2S: $\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)$ vanishes as $d_s/N \to 0$

Proof intuition: teacher variance in $\mathbb{V}_w \backslash \mathbb{V}_s \approx$ independent label noise

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\text{W2S}} \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) → Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

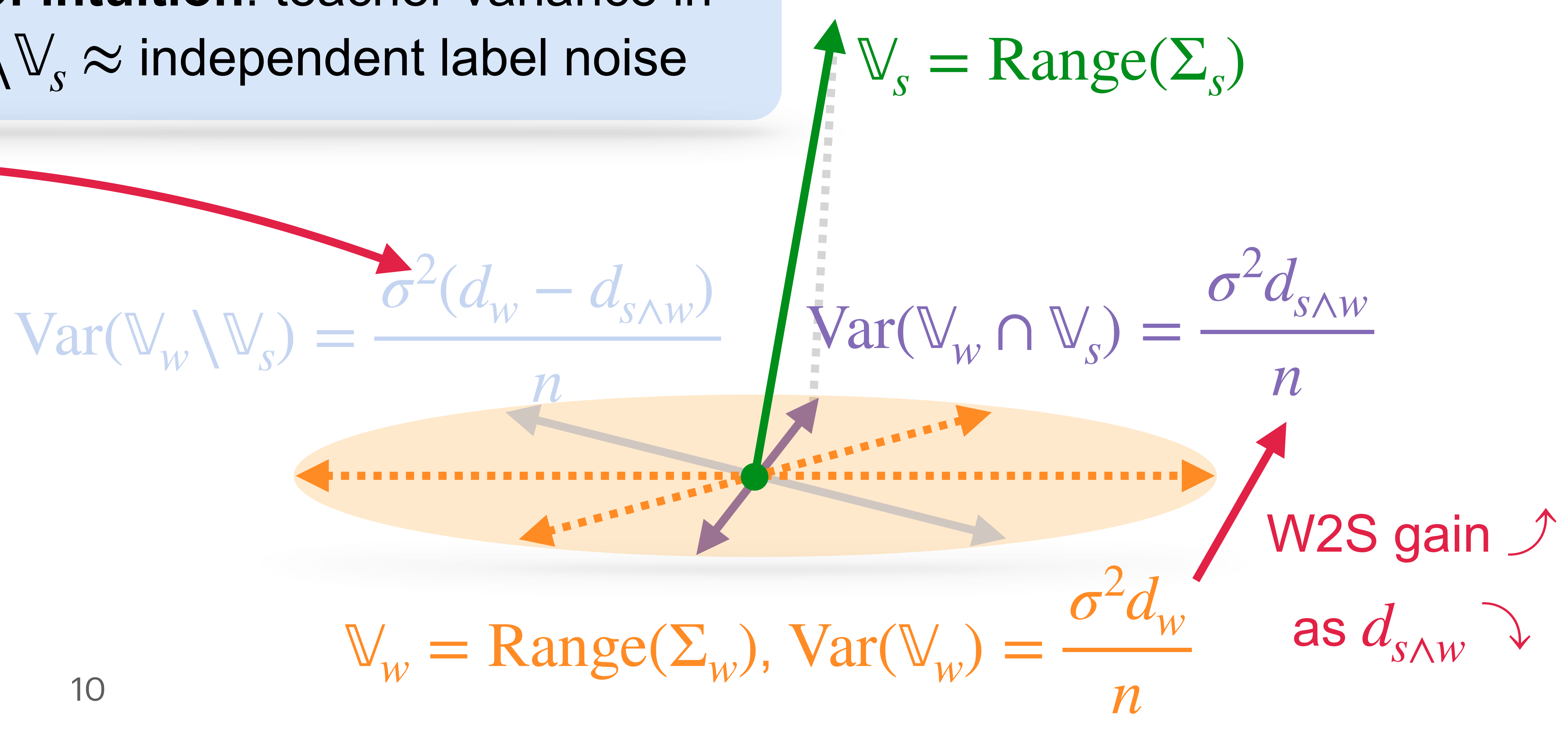
Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

Variance reduction in W2S: $\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)$ vanishes as $d_s/N \to 0$

Proof intuition: teacher variance in $\mathbb{V}_w \backslash \mathbb{V}_s \approx$ independent label noise

$$\mathrm{Var}(f_s) \asymp \underbrace{\frac{d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)} + \underbrace{\frac{d_s}{N}}_{\textbf{W2S}} \underbrace{\frac{d_w - d_{s\wedge w}}{n}}_{\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)}$$



Larger Discrepancy (Lower $d_{s\wedge w}$) → Better W2S

Theorem [DLLLL25]. Assume $\phi_s(x)$ is zero-mean subgaussian and $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can also be relaxed to subgaussian), for $n > d_w + 1$:

$$\mathrm{Var}(f_s) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right)$$

Proposition [DLLLL25]. Assume $\phi_w(x) \sim \mathcal{N}(0_d, \Sigma_w)$ (can be relaxed to subgaussian design), for $n > d_w + 1$:

$$\mathrm{Var}(f_w) = \frac{\sigma^2 d_w}{n - d_w - 1}$$

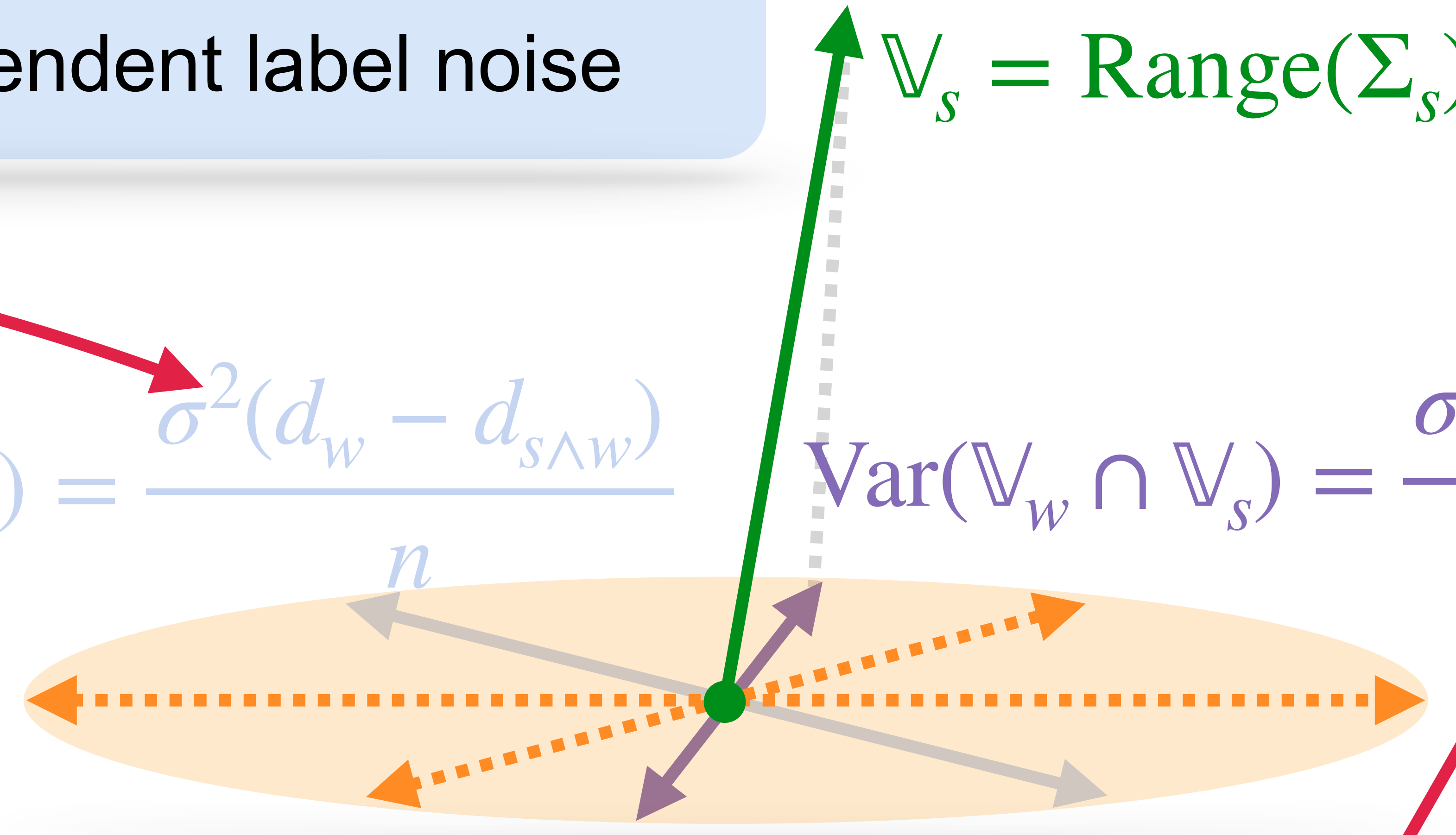
Supported by empirical observations in our work & concurrent empirical work, Goel et al., (2025)

Variance reduction in W2S: $\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)$ vanishes as $d_s/N \to 0$

Proof intuition: teacher variance in $\mathbb{V}_w \backslash \mathbb{V}_s \approx$ independent label noise

$$\mathrm{Var}(f_s) \asymp \frac{d_{s\wedge w}}{n} + \frac{d_s}{N}\,\frac{d_w - d_{s\wedge w}}{n}$$

$\frac{d_{s\wedge w}}{n}$: $\mathrm{Var}(\mathbb{V}_w \cap \mathbb{V}_s)$; $\frac{d_s}{N}$: **W2S**; $\frac{d_w - d_{s\wedge w}}{n}$: $\mathrm{Var}(\mathbb{V}_w \backslash \mathbb{V}_s)$



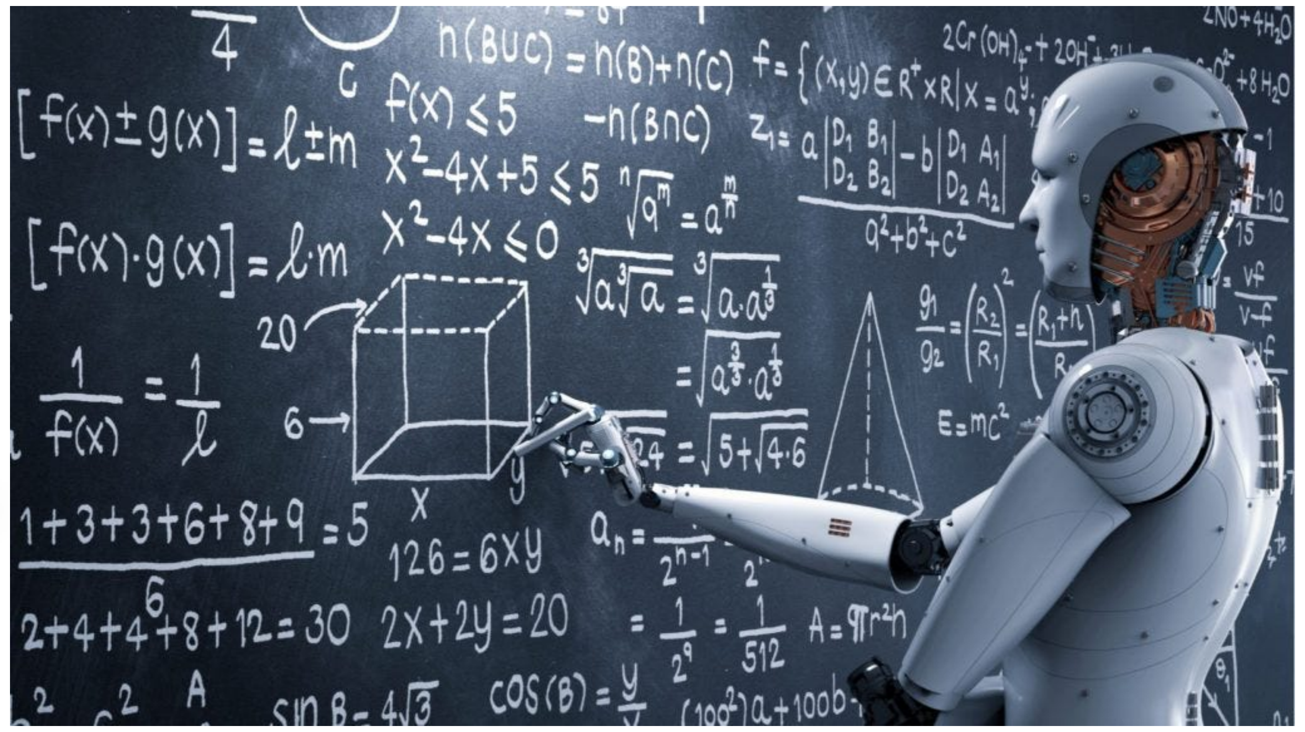
W2S Emerges during Post-training

Powerful pre-trained models

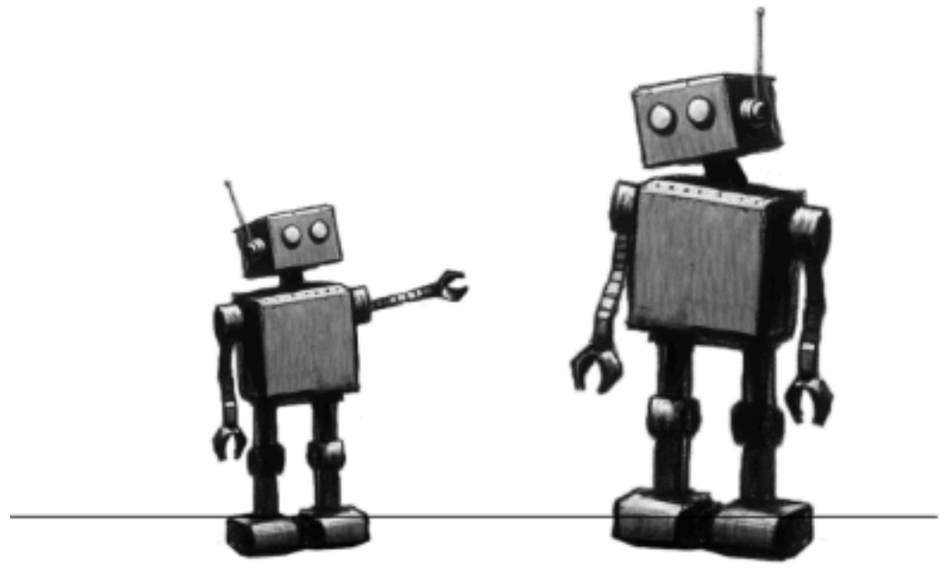
Gemini

Post-training

Specialized downstream tasks

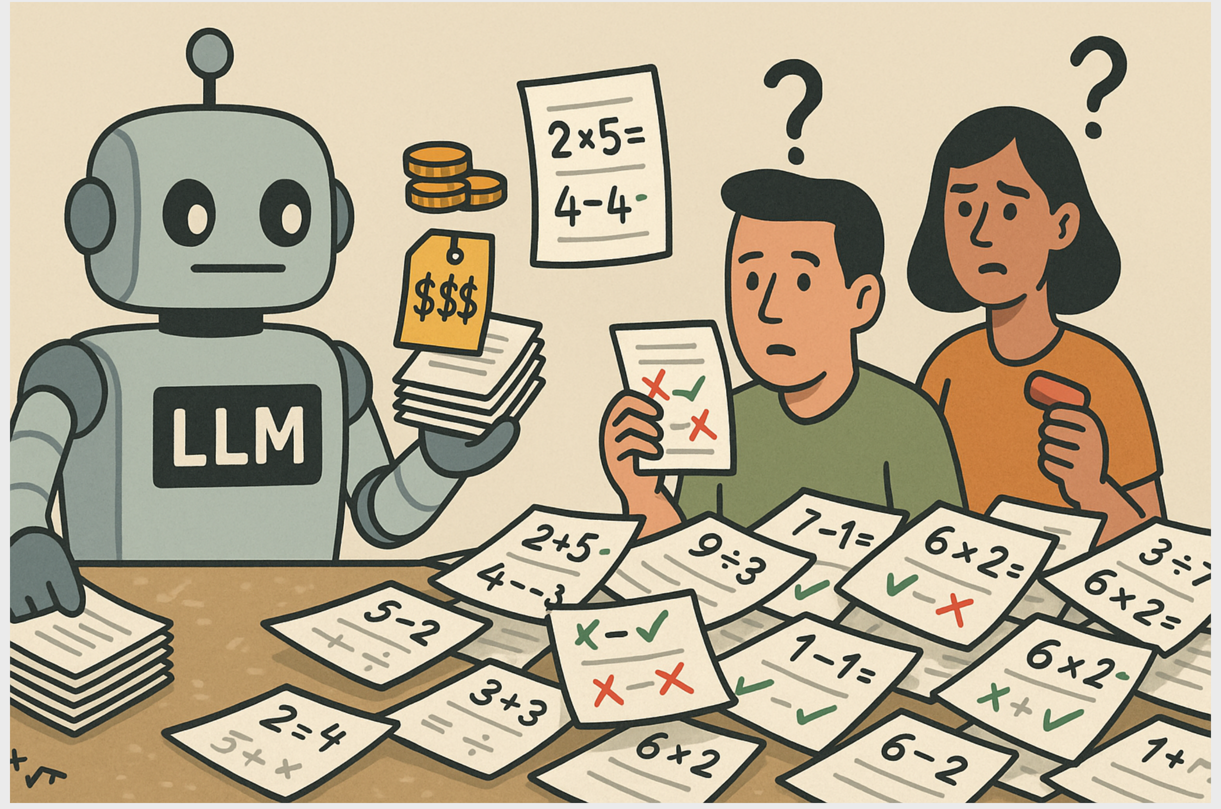


Weak-to-strong generalization

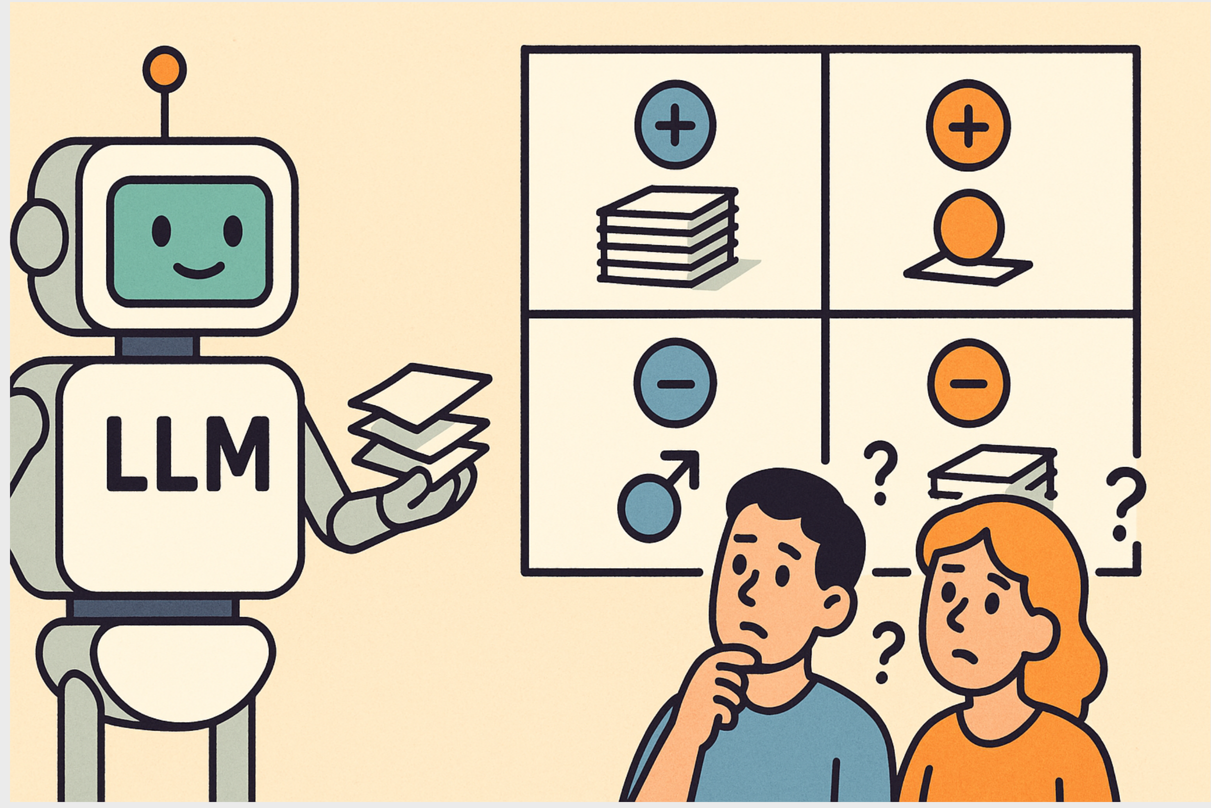


Supervisor Student

① … with **limited & noisy labels**



② … with **systematic bias**

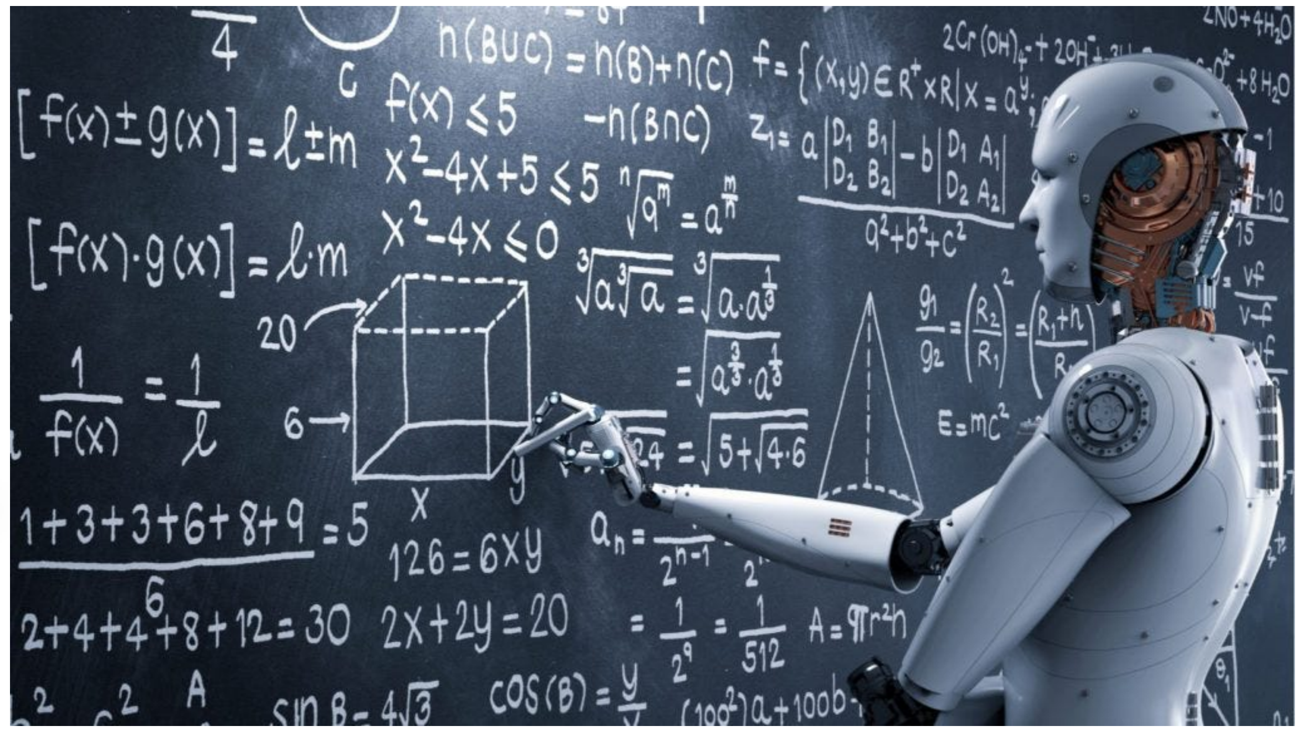


W2S Emerges during Post-training

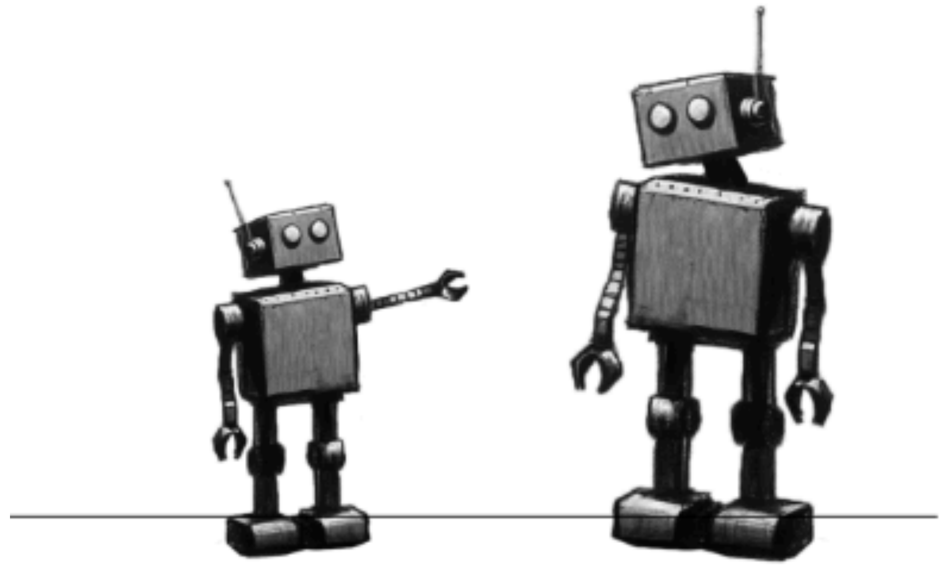
Powerful pre-trained models

Post-training

Specialized downstream tasks

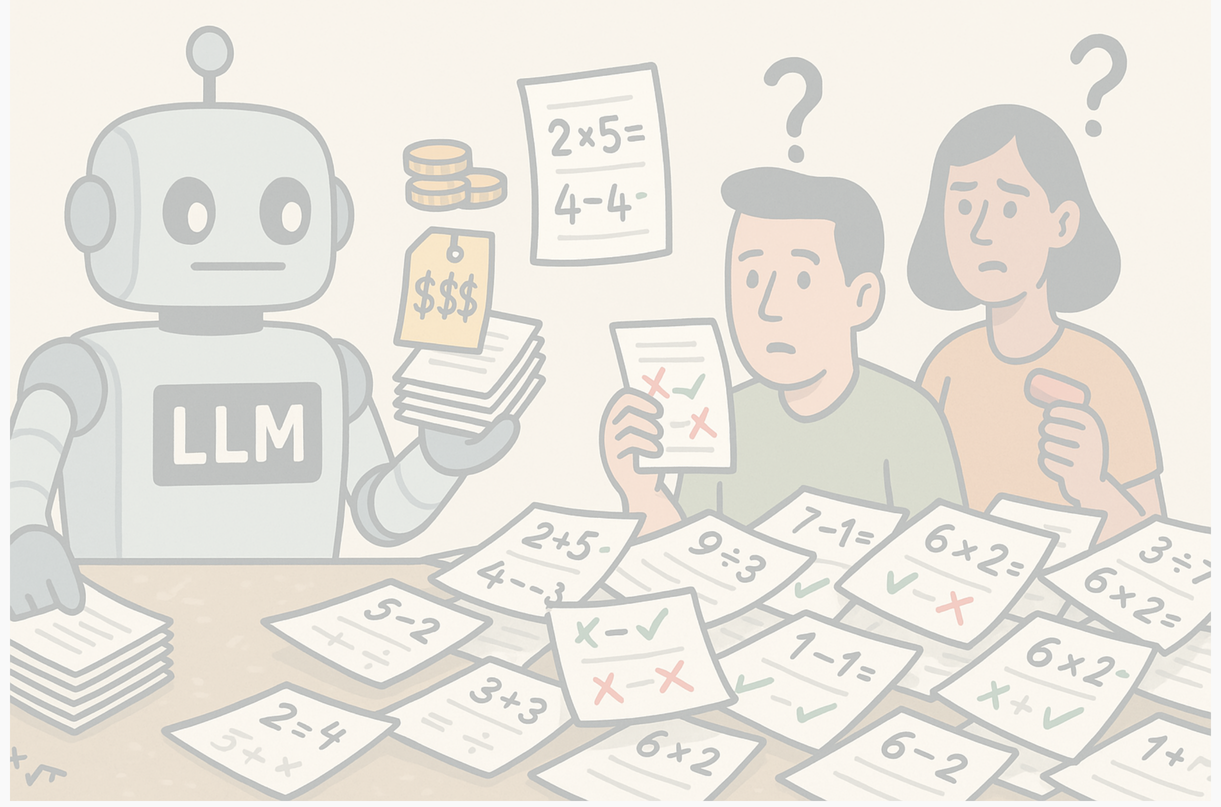


Weak-to-strong generalization

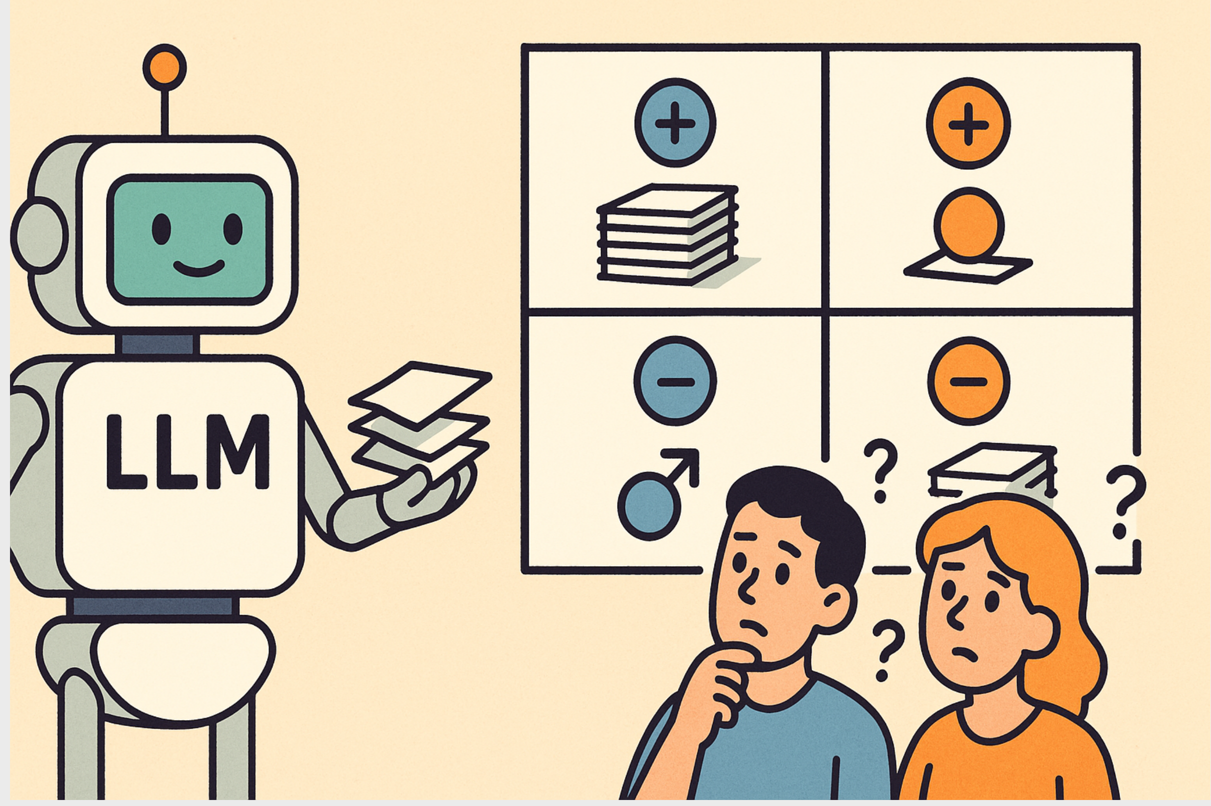


Supervisor Student

① … with **limited & noisy labels**



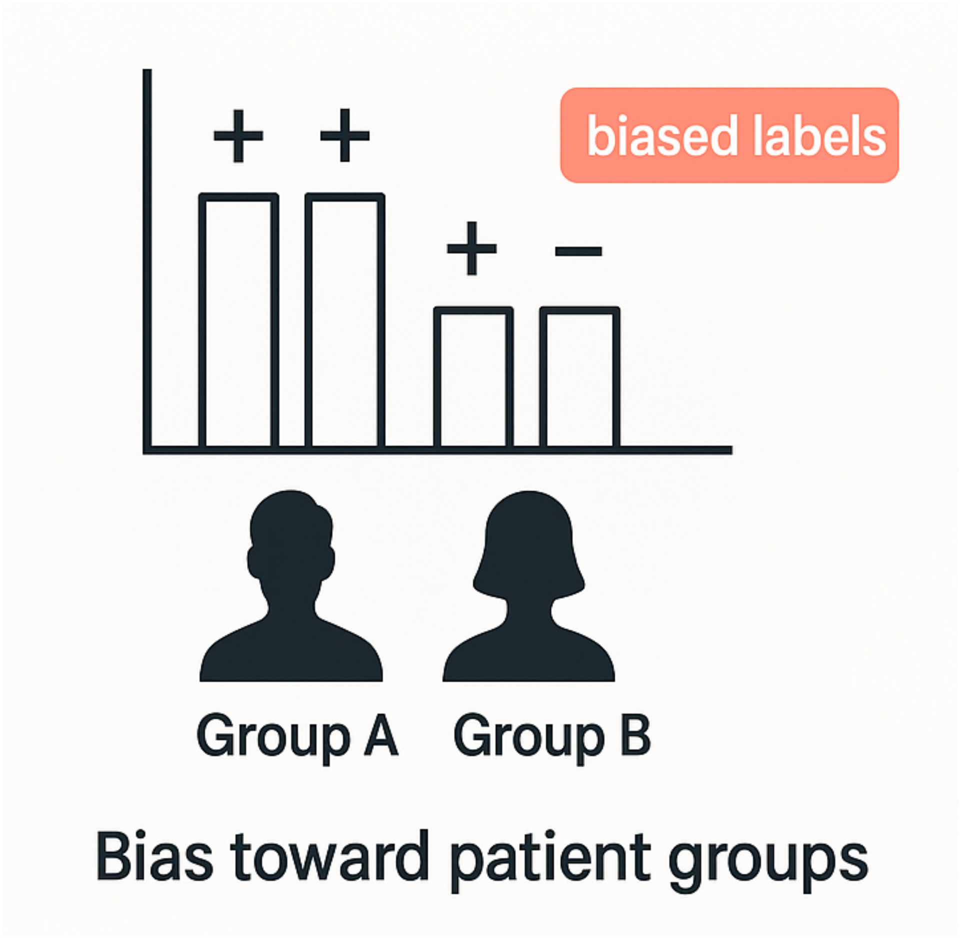
② … with **systematic bias**



… where Data Often Come with Group Imbalance

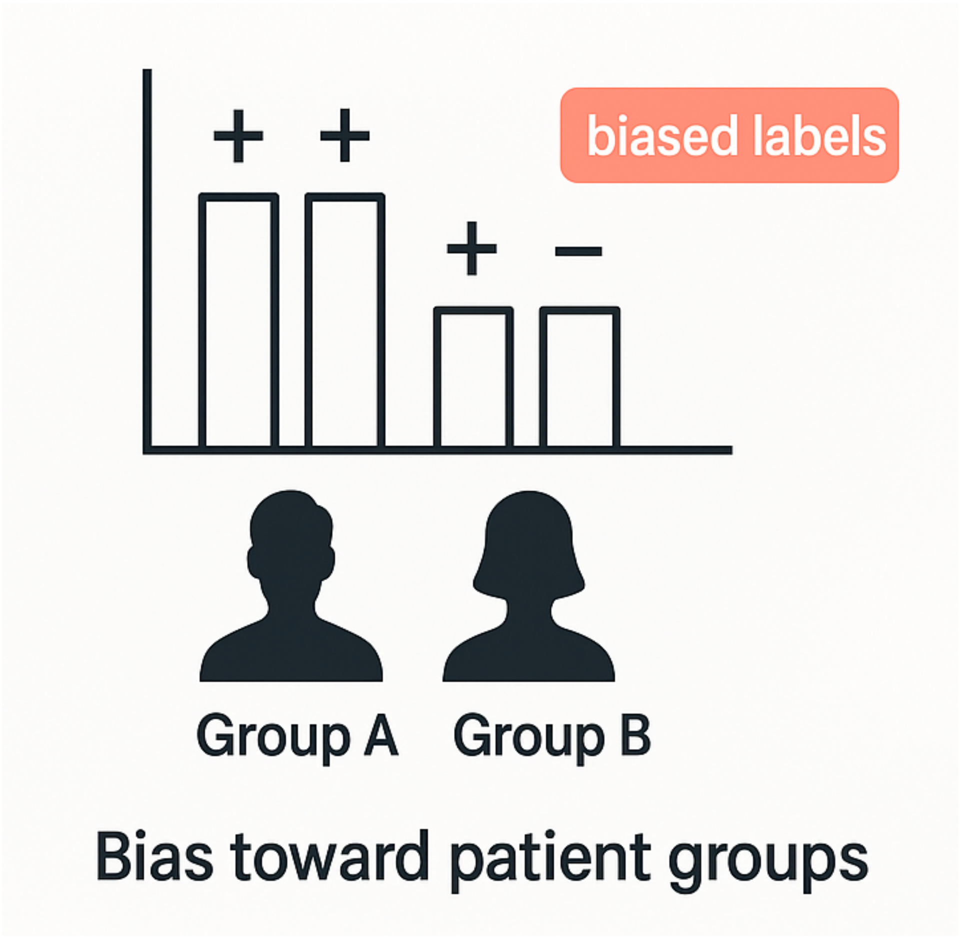
… where Data Often Come with Group Imbalance

Medical data

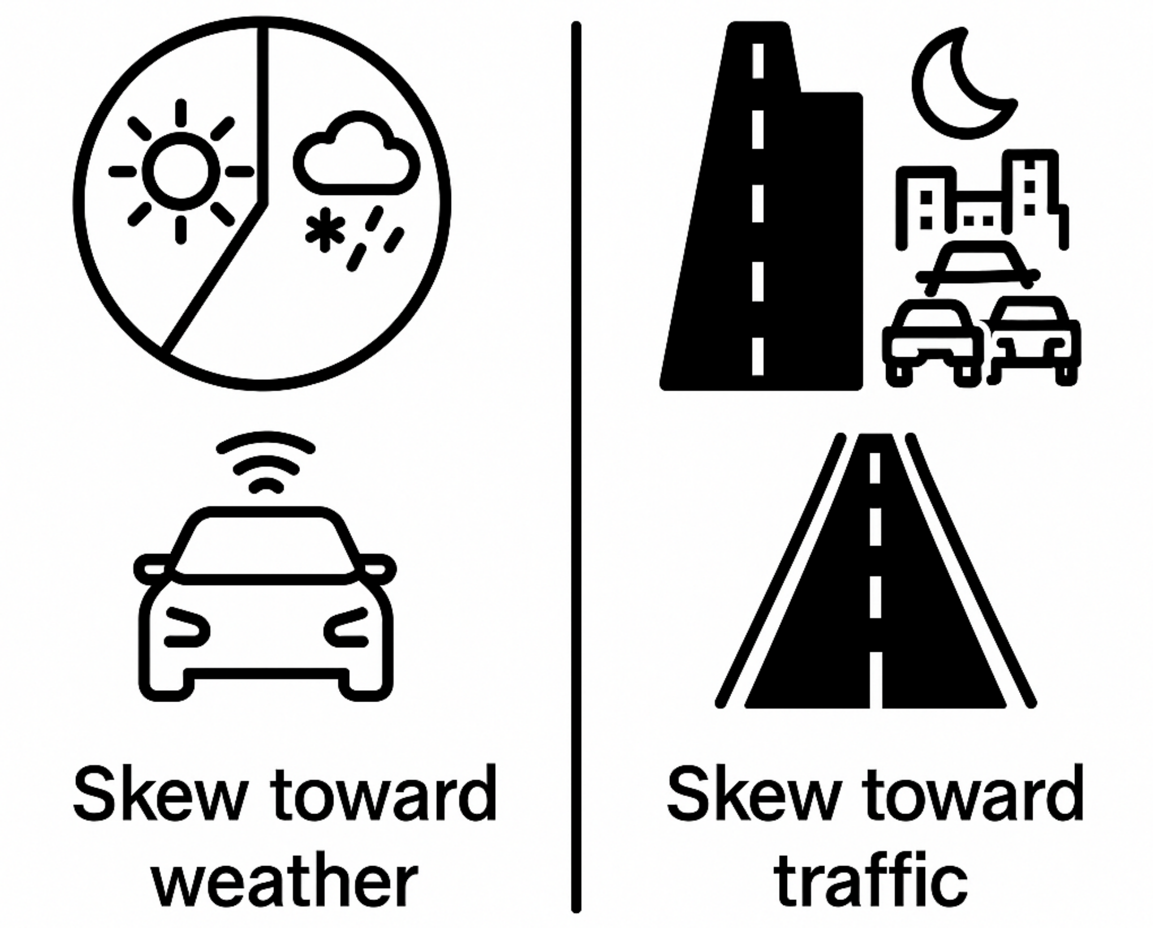


… where Data Often Come with Group Imbalance

Medical data

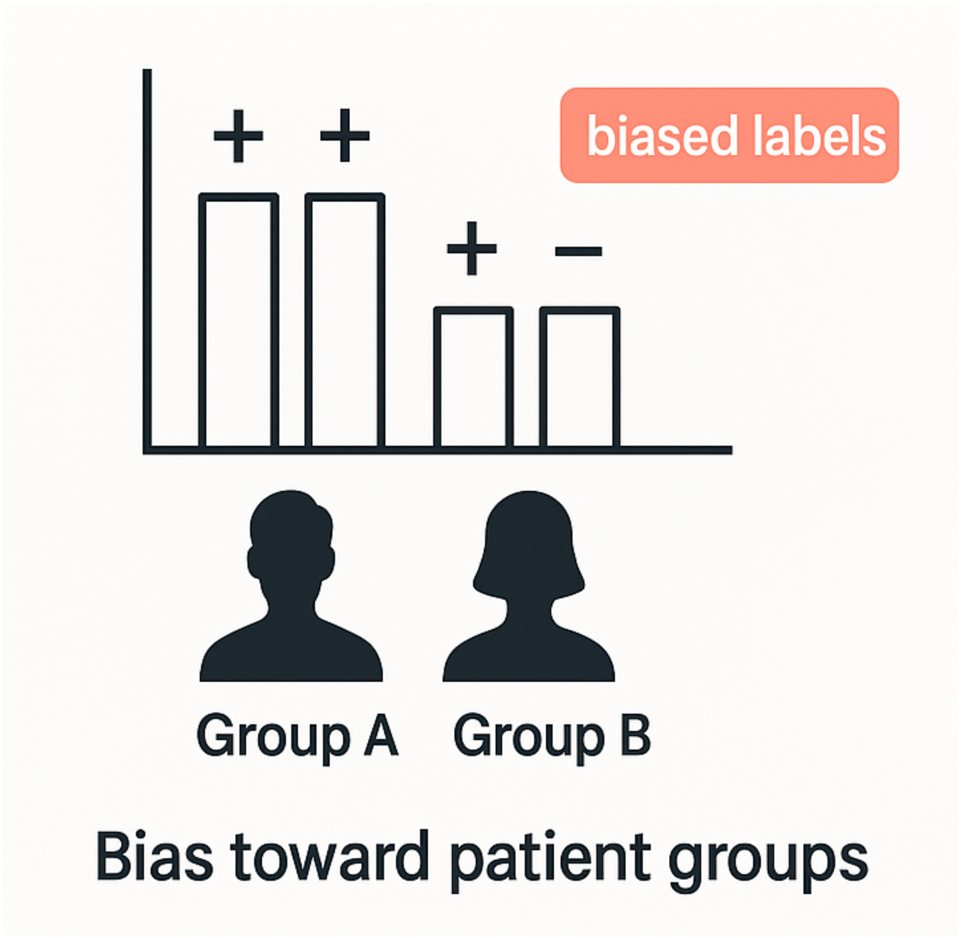


Autonomous driving data

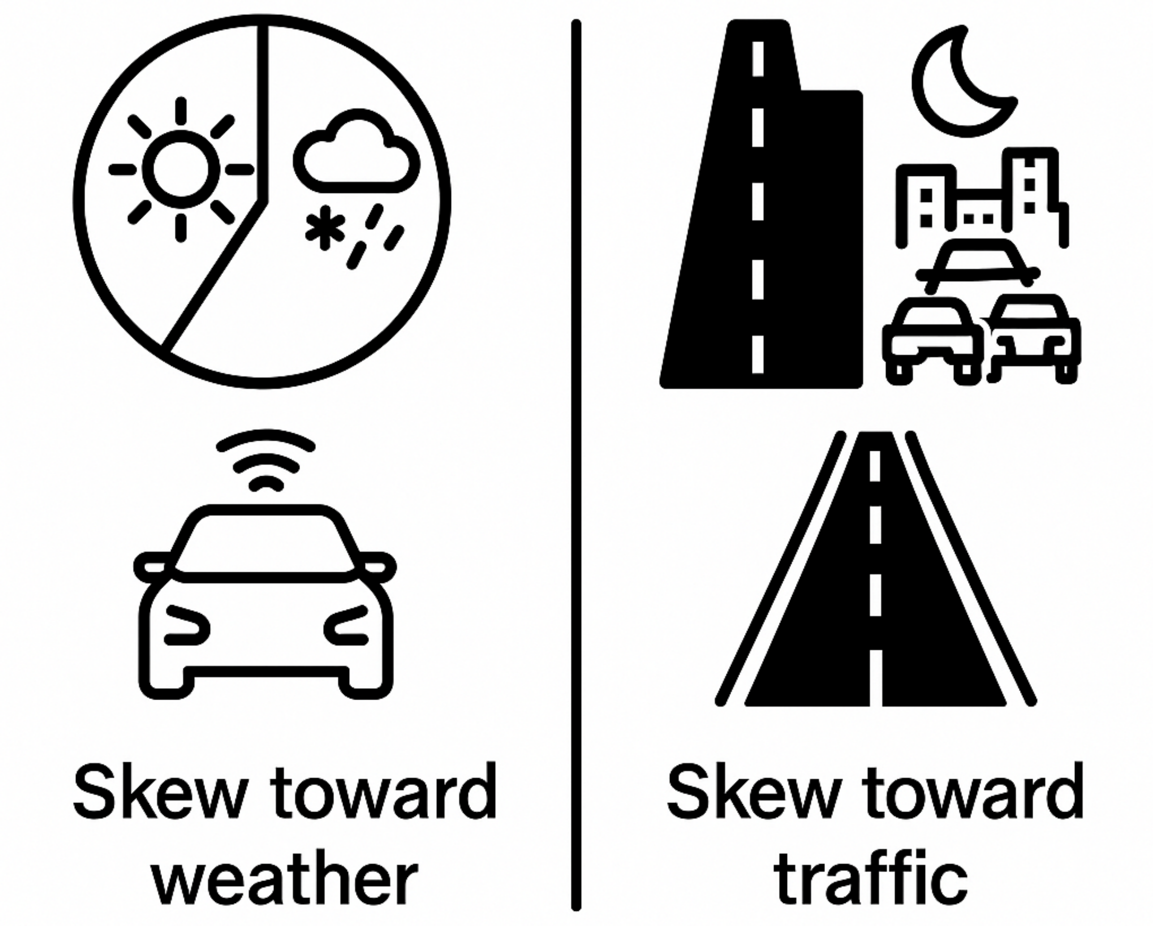


… where Data Often Come with Group Imbalance

Medical data



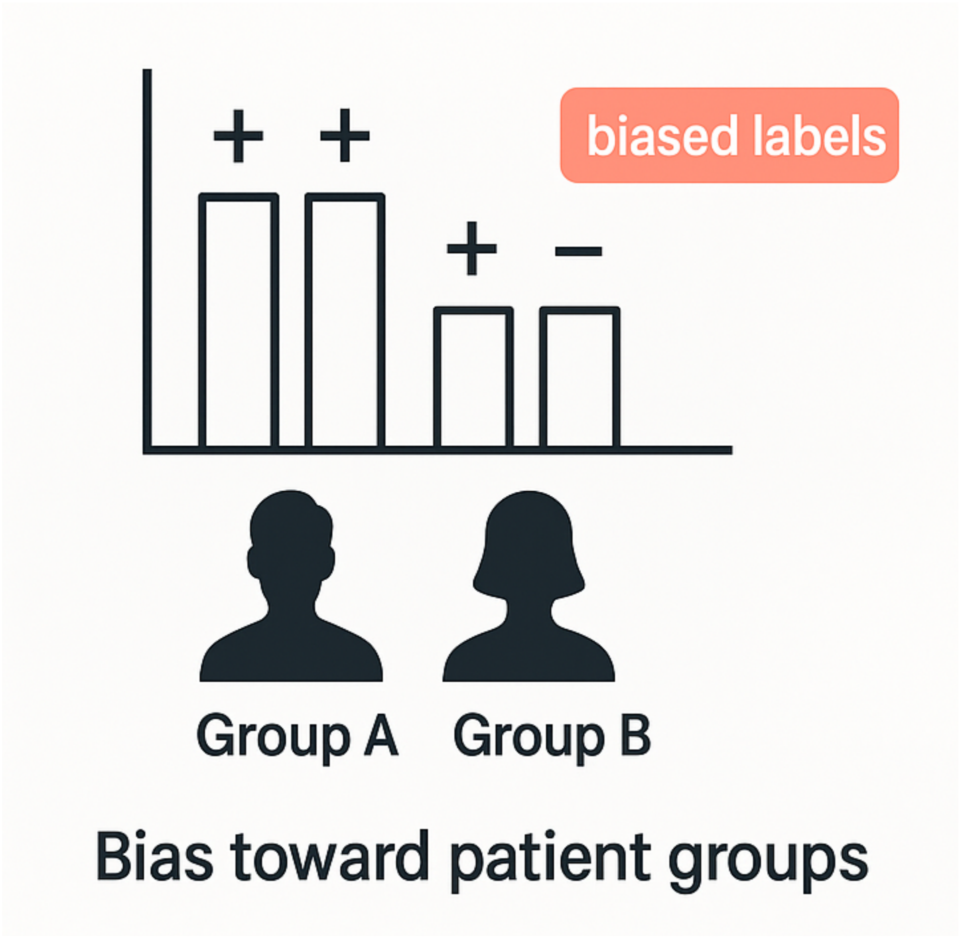
Autonomous driving data



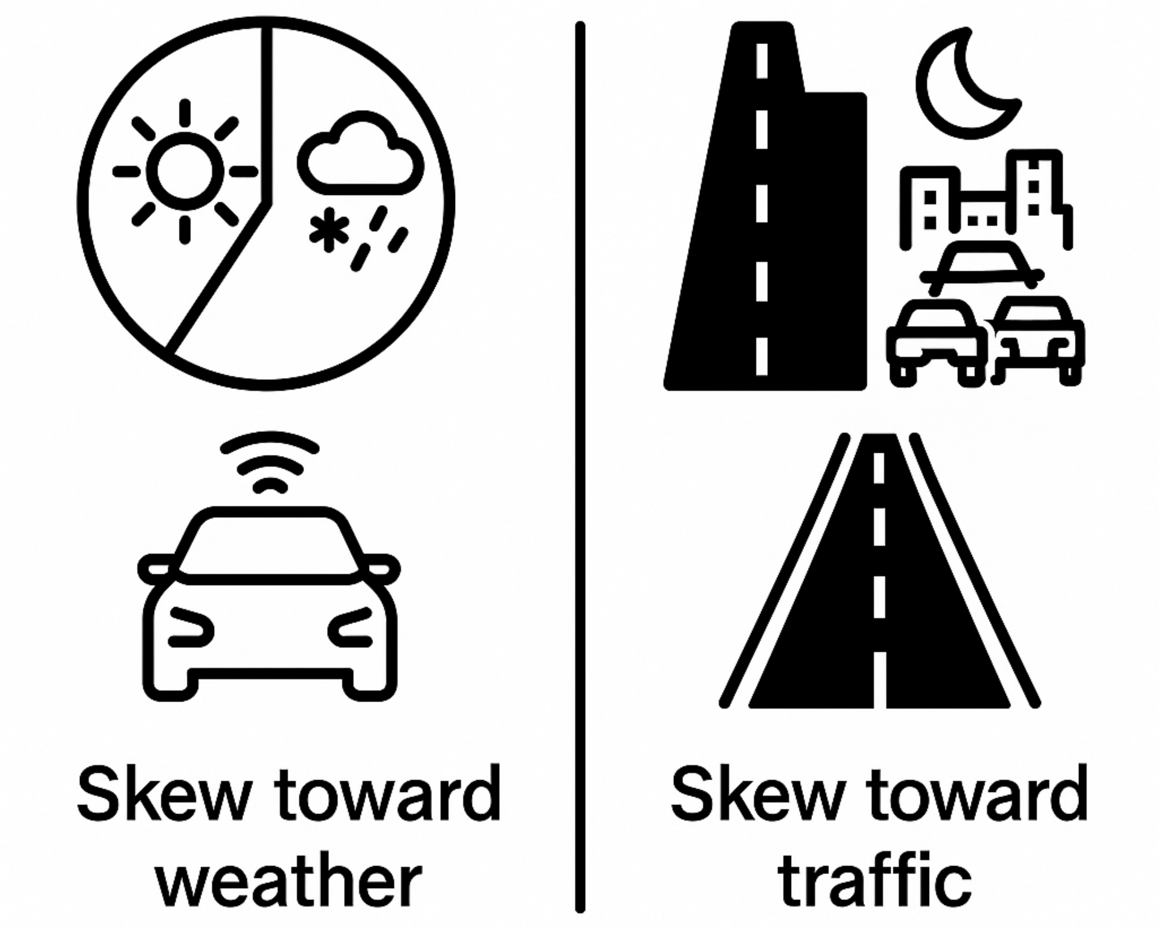
Classify cow vs. camel

… where Data Often Come with Group Imbalance

Medical data



Autonomous driving data

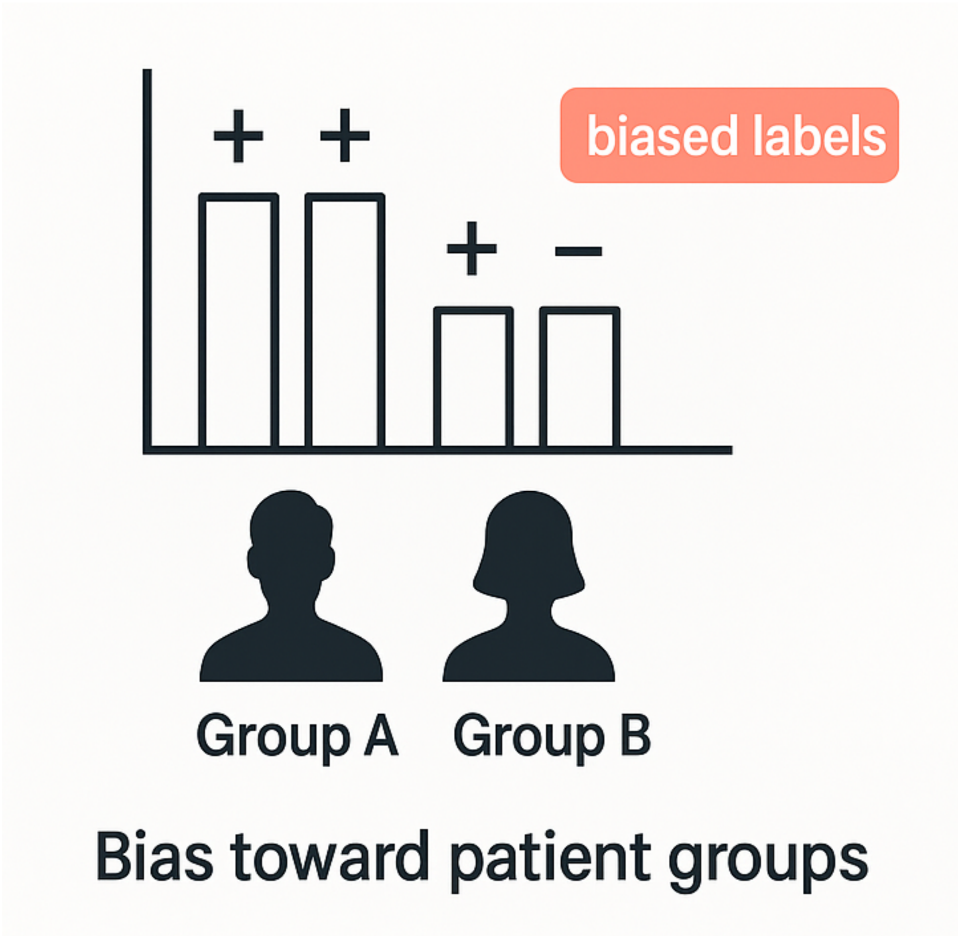


Classify cow vs. camel

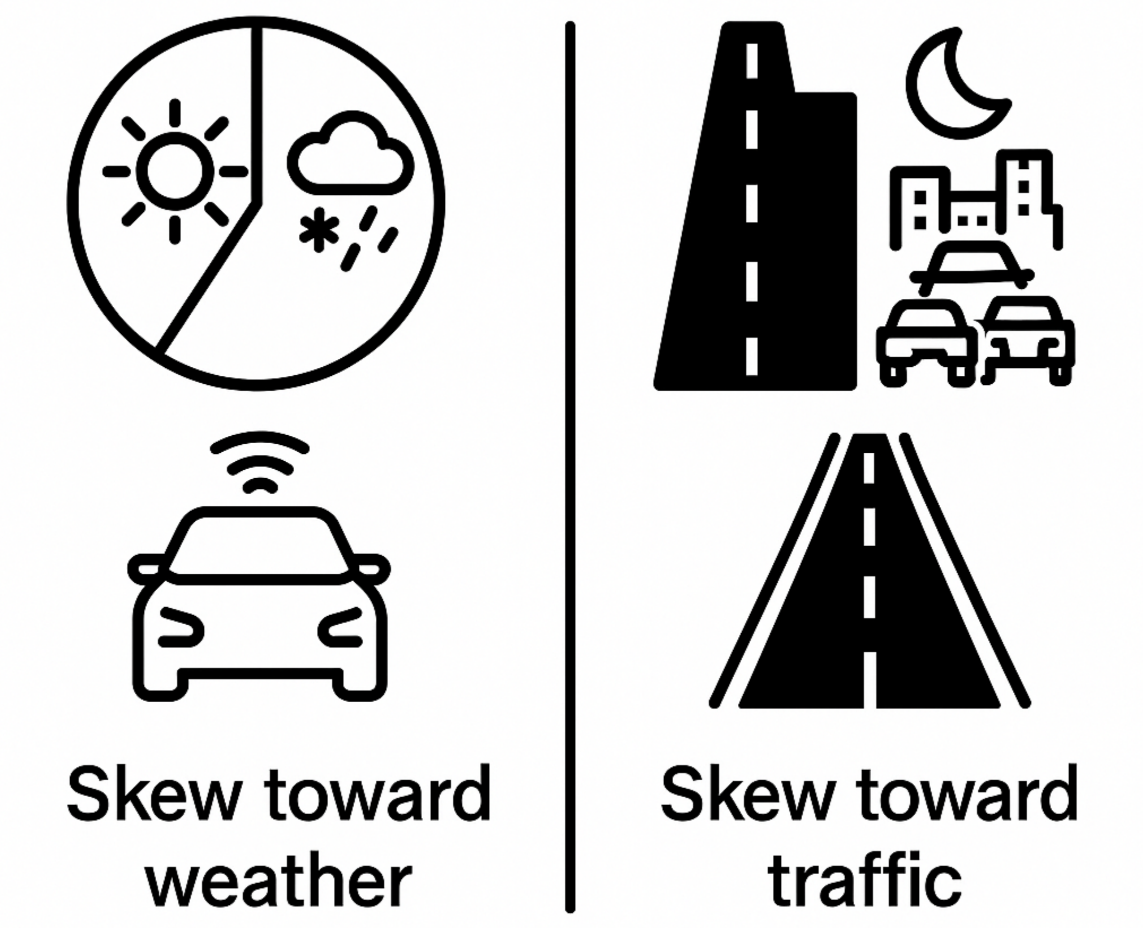
Semantic features encoded in dim. d_z

… where Data Often Come with Group Imbalance

Medical data

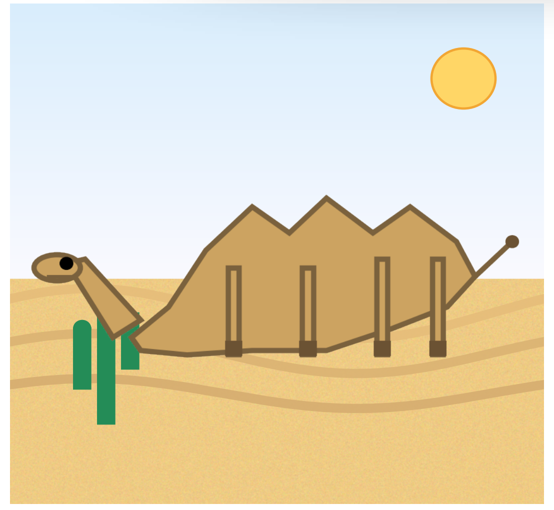


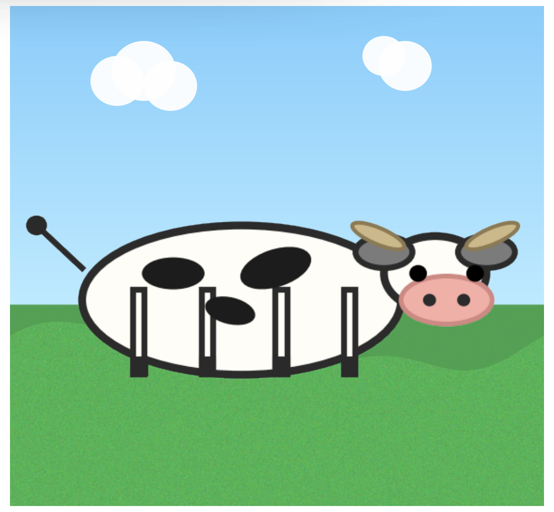
Autonomous driving data



Classify cow vs. camel

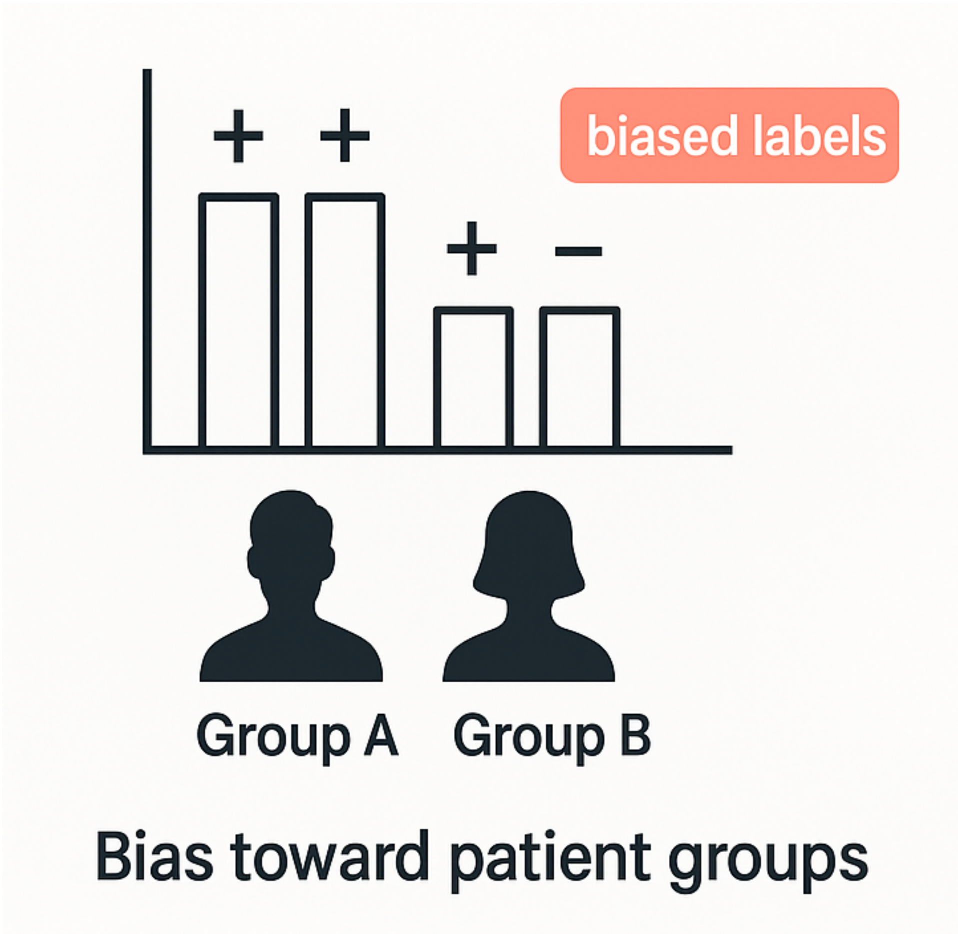
Semantic features
encoded in dim. d_z



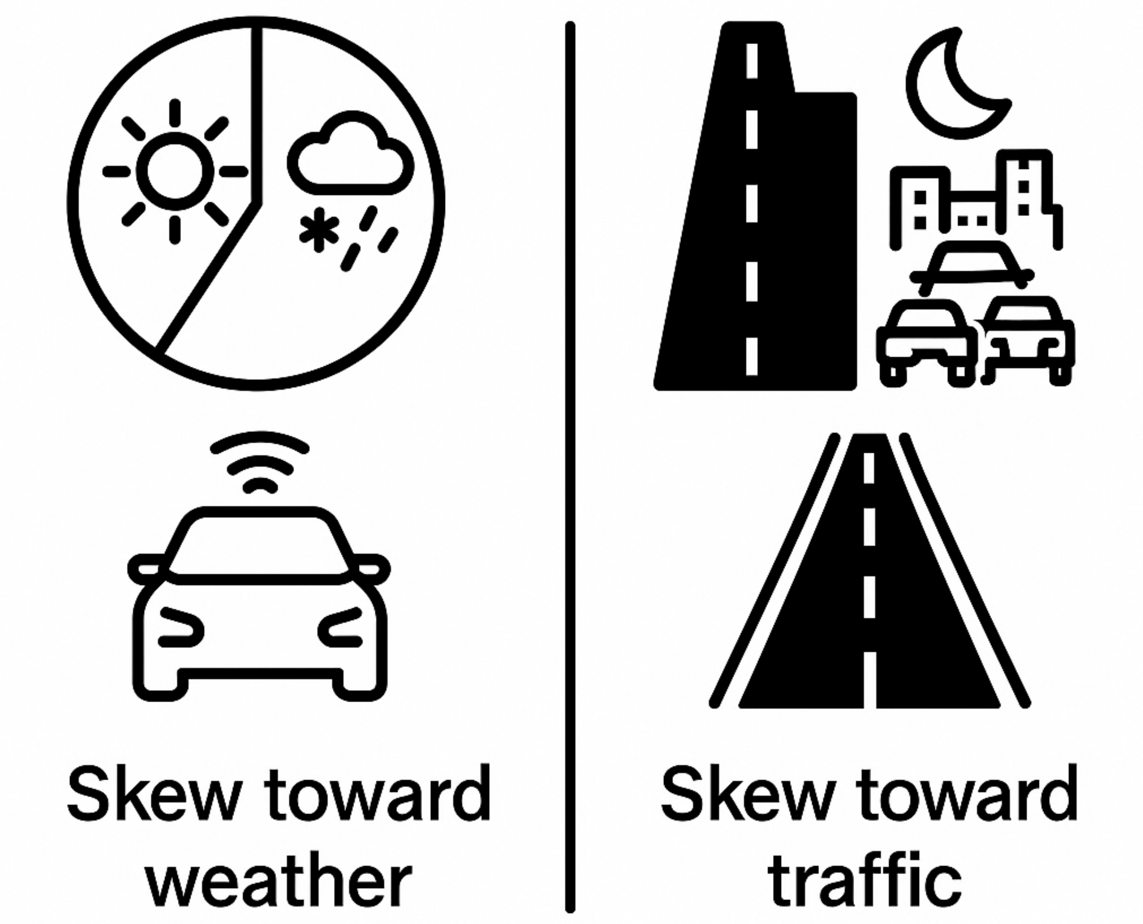


… where Data Often Come with Group Imbalance

Medical data

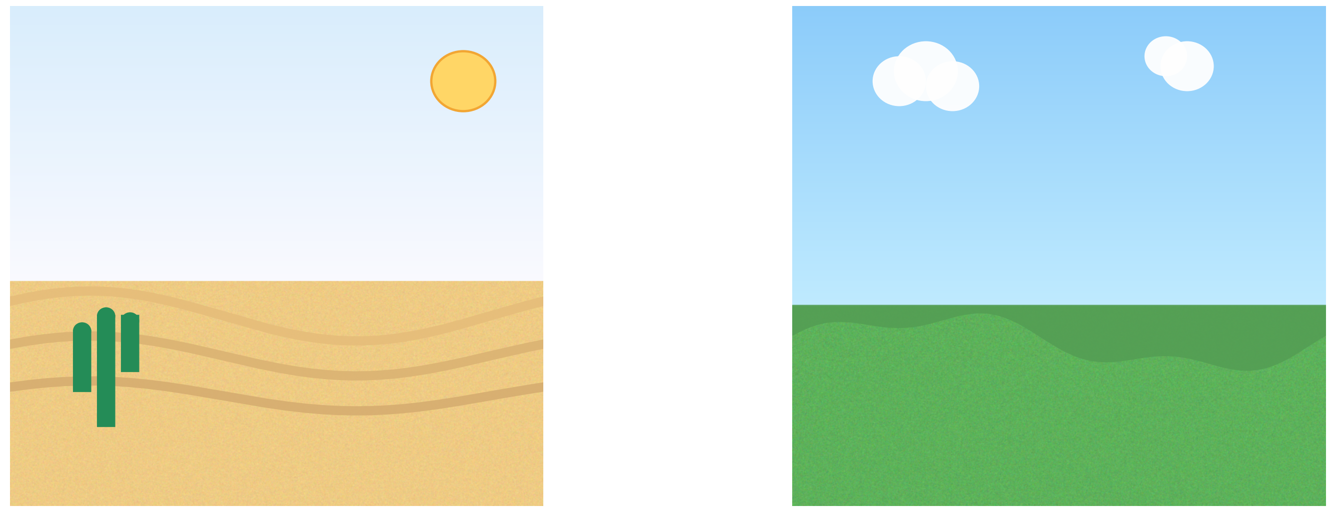


Autonomous driving data



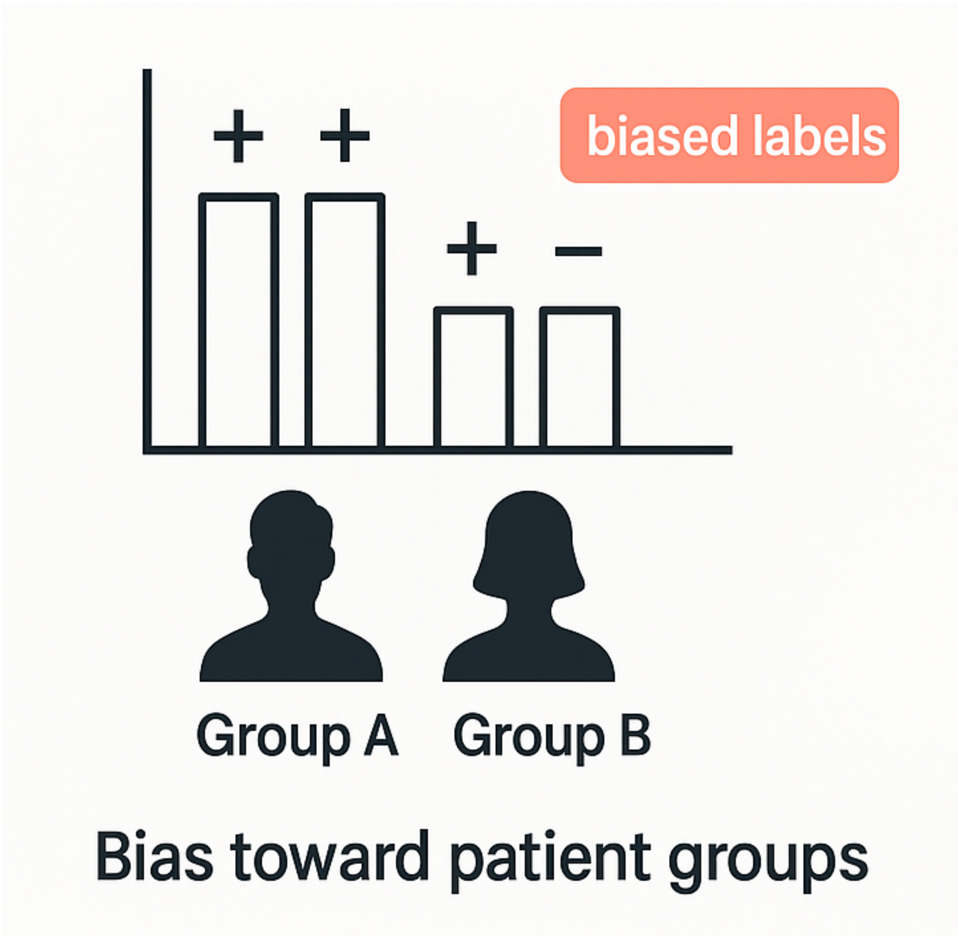
Classify cow vs. camel

Semantic features
encoded in dim. d_z

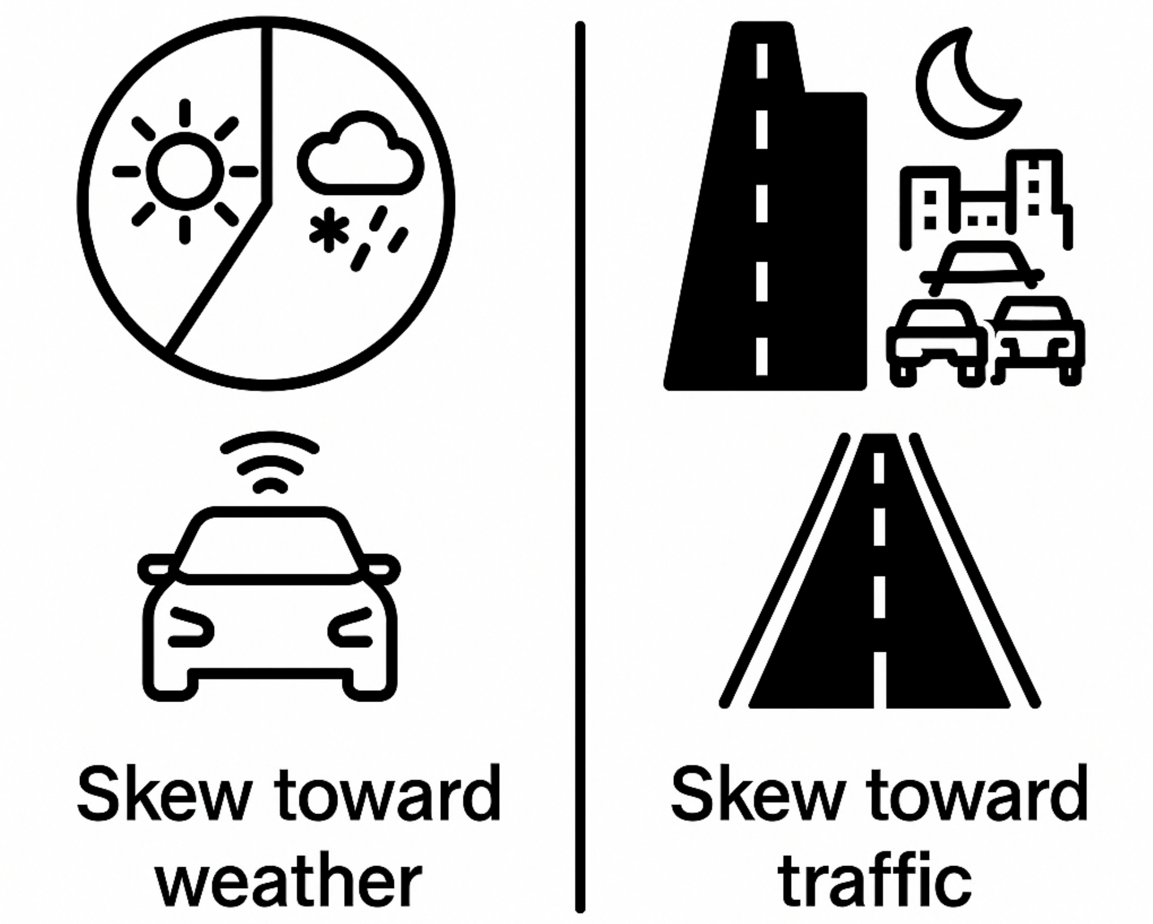


… where Data Often Come with Group Imbalance

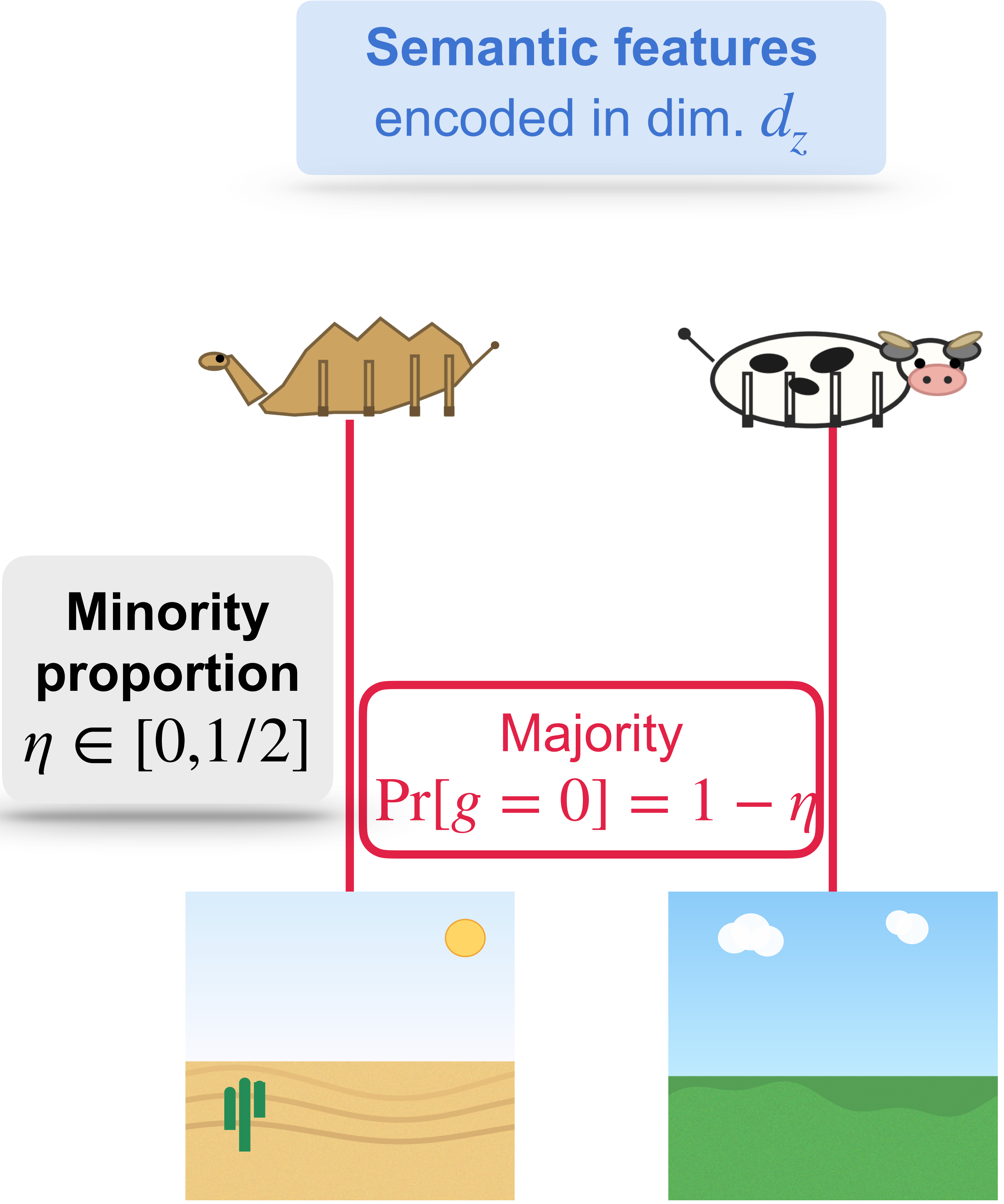
Medical data



Autonomous driving data

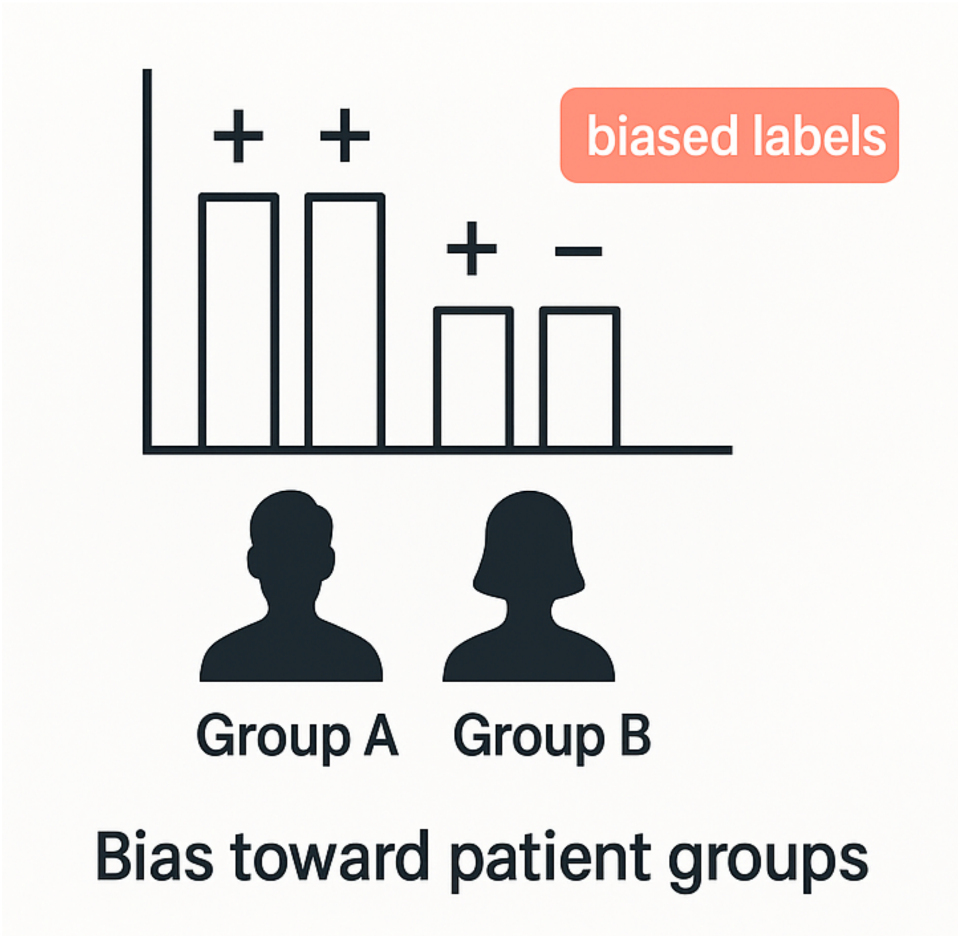


Classify cow vs. camel

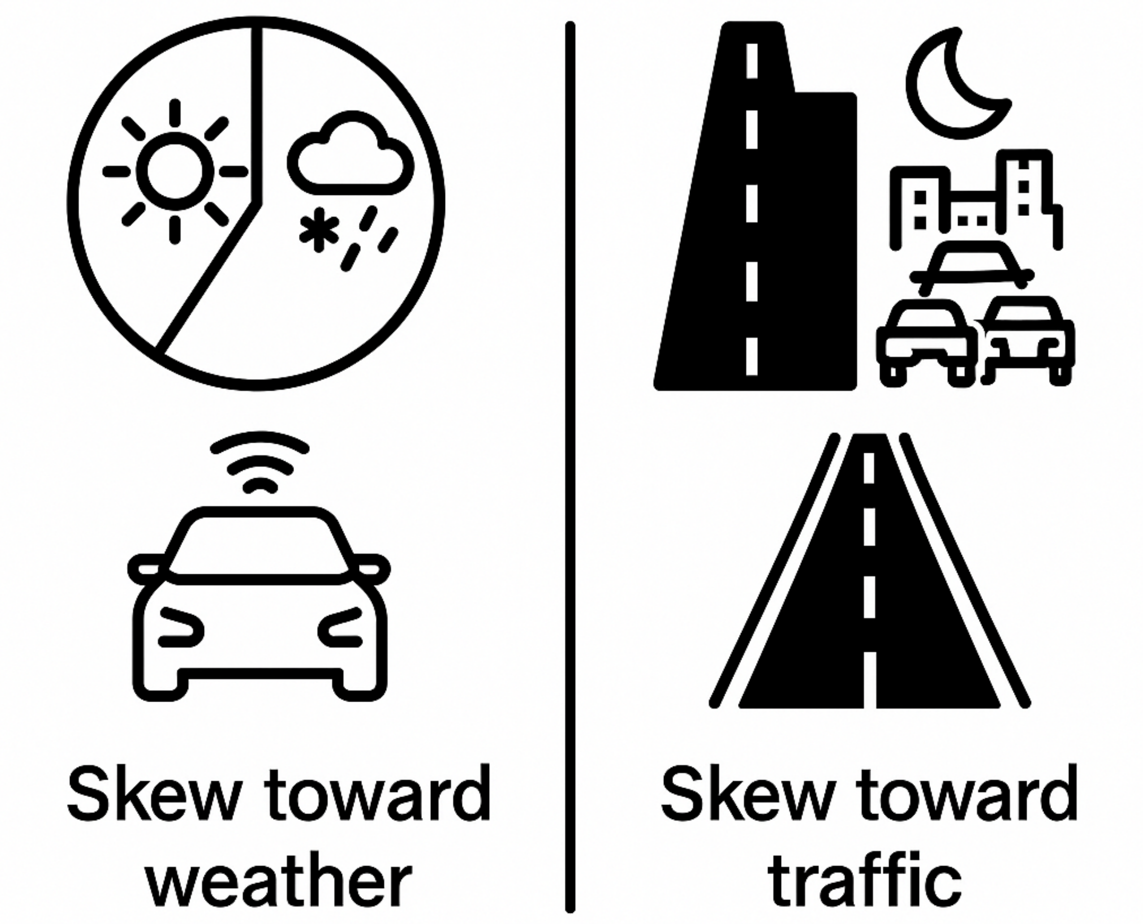


… where Data Often Come with Group Imbalance

Medical data



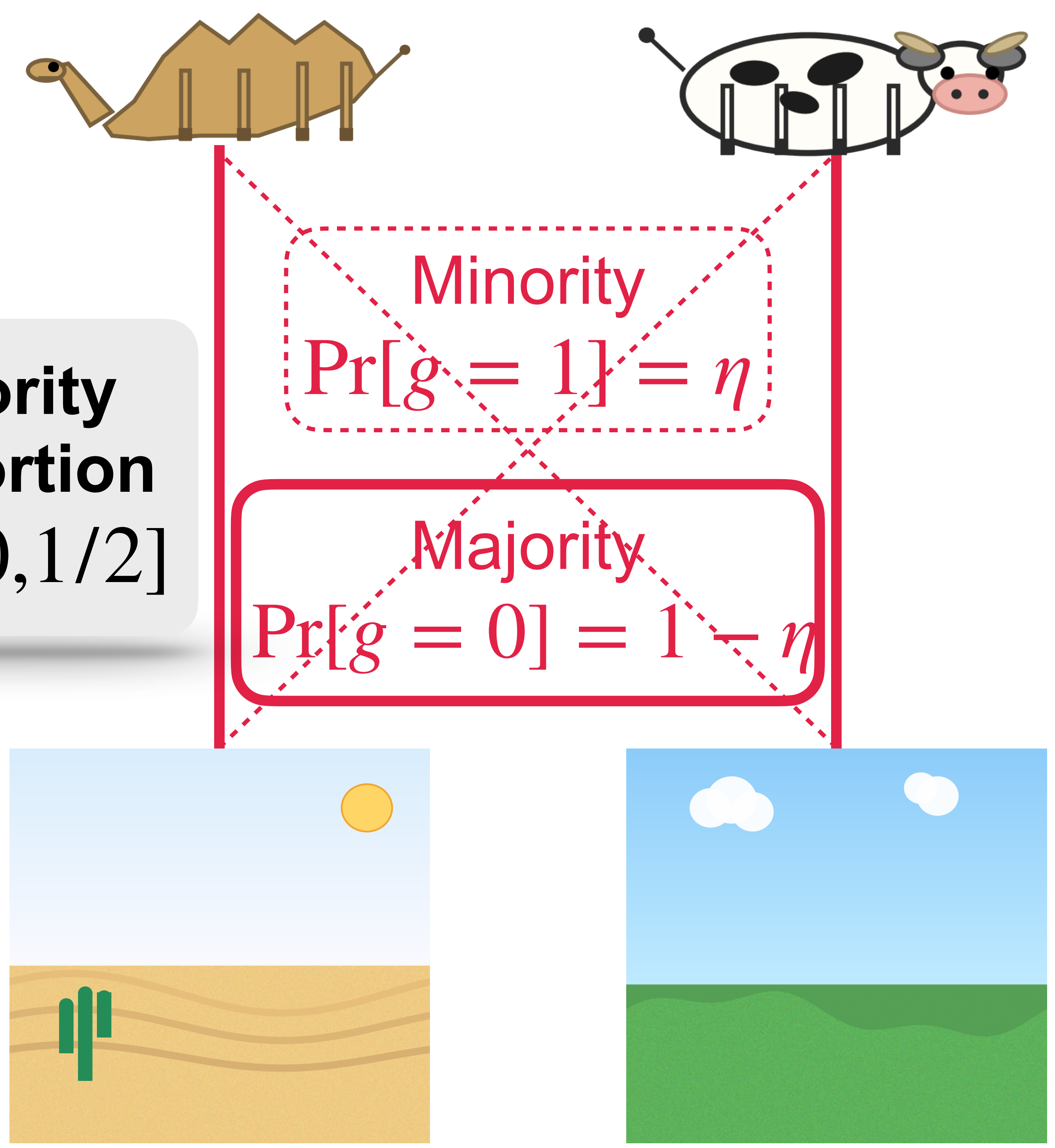
Autonomous driving data



Classify cow vs. camel

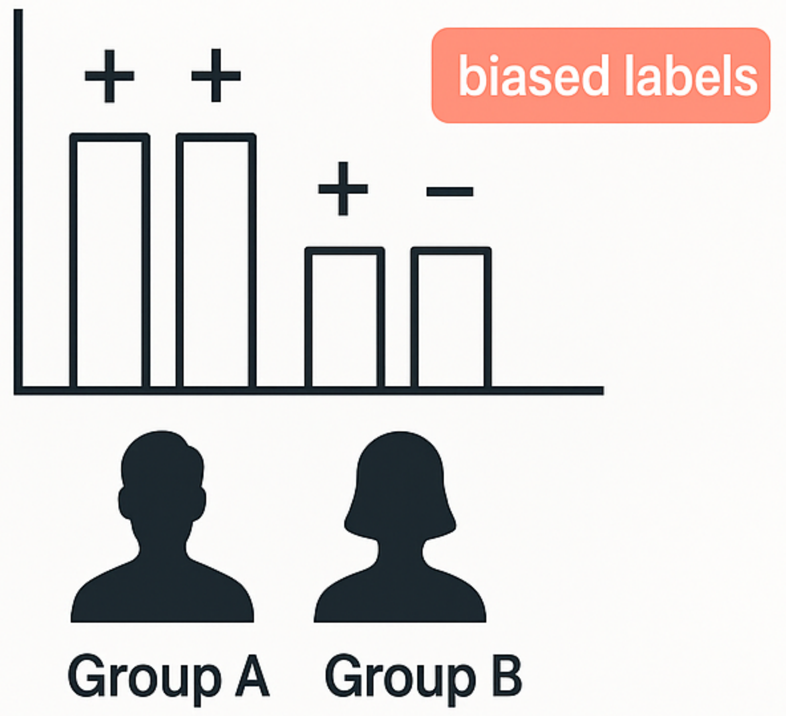
Semantic features
encoded in dim. d_z

Minority proportion
$\eta \in [0,1/2]$



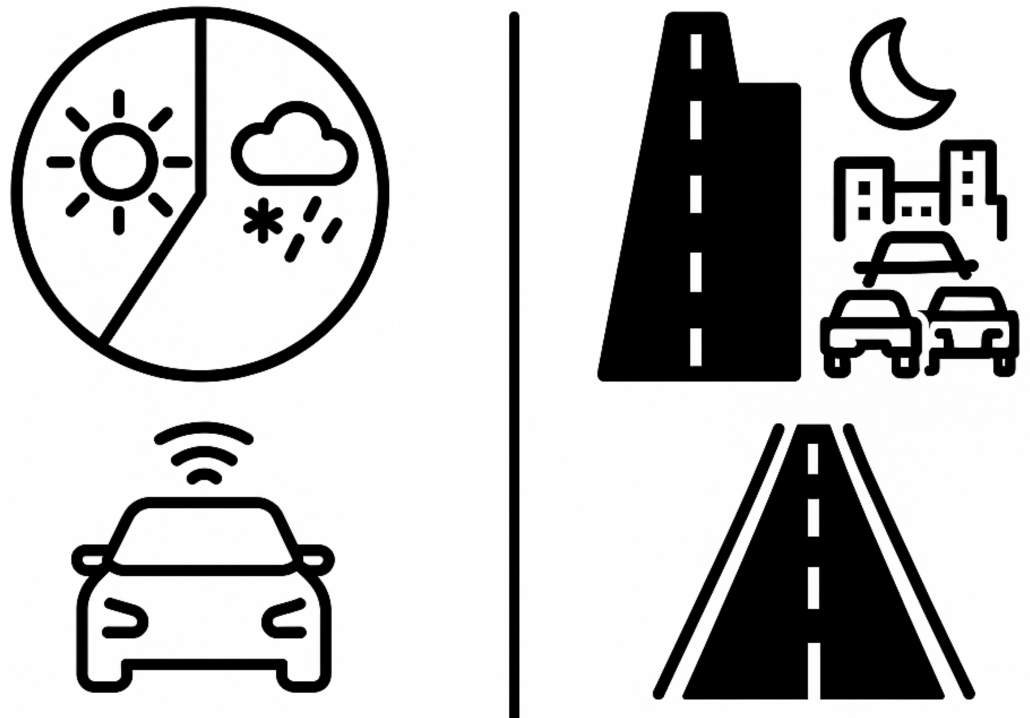
… where Data Often Come with Group Imbalance

Medical data



Bias toward patient groups

Autonomous driving data

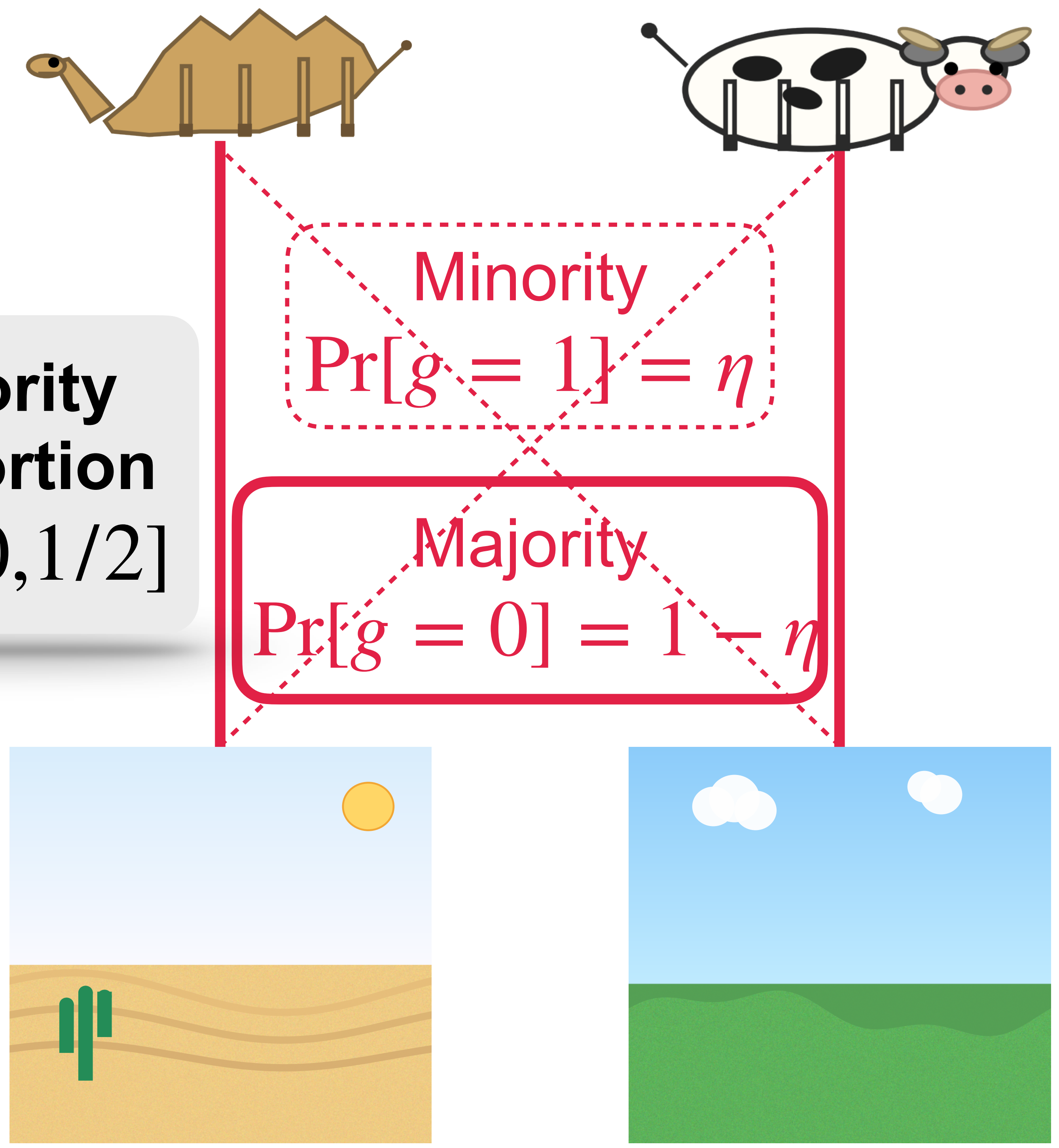


Skew toward weather

Skew toward traffic

Classify cow vs. camel

Semantic features
encoded in dim. d_z

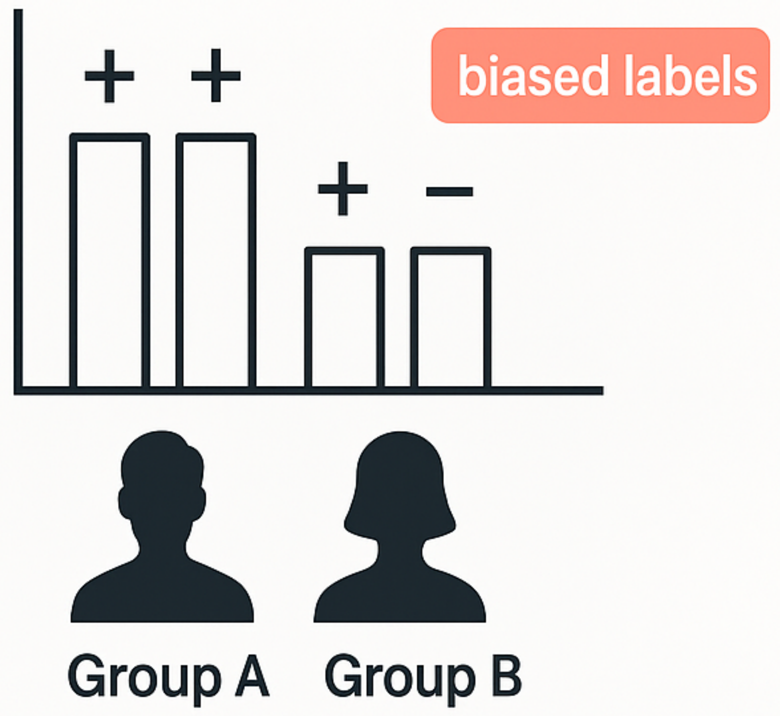


Minority proportion
$\eta \in [0, 1/2]$

🙁 When $\eta \ll 1$, backgrounds are learned.

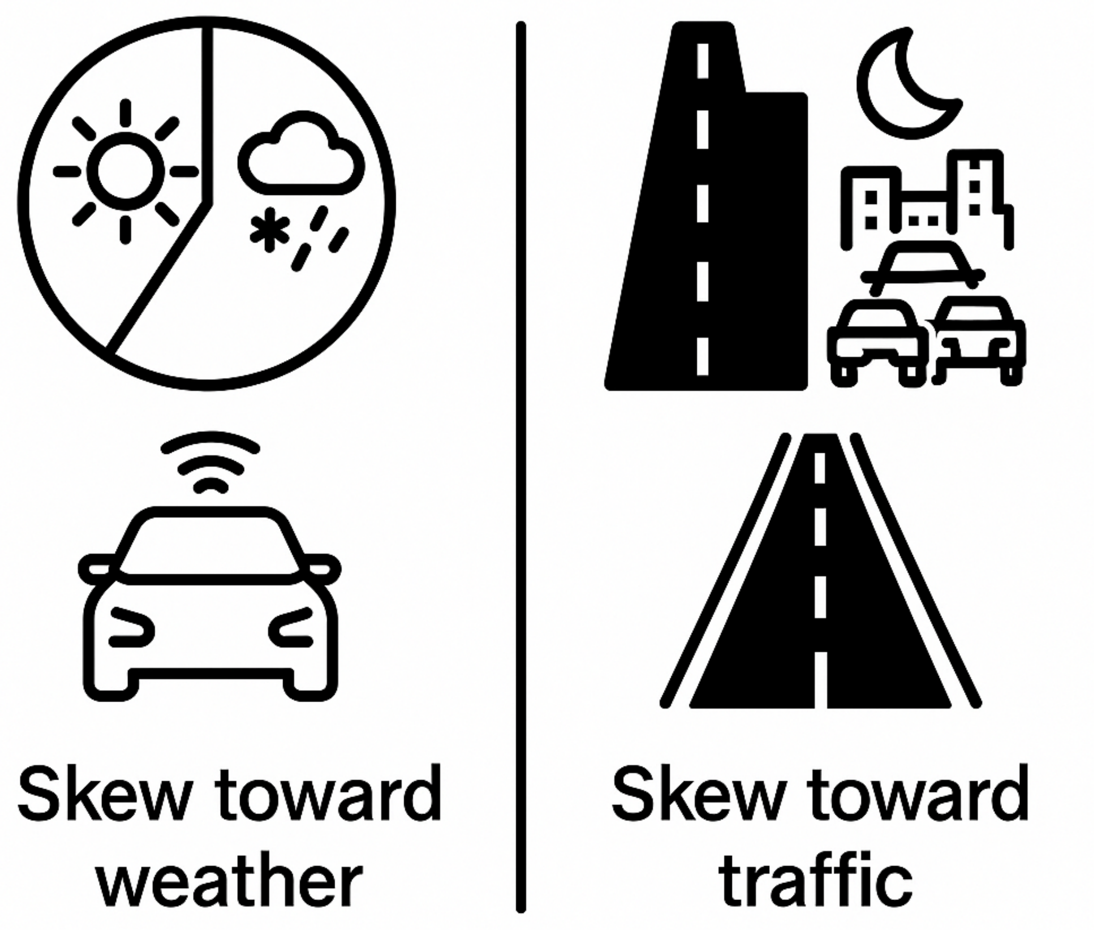
… where Data Often Come with Group Imbalance

Medical data



Bias toward patient groups

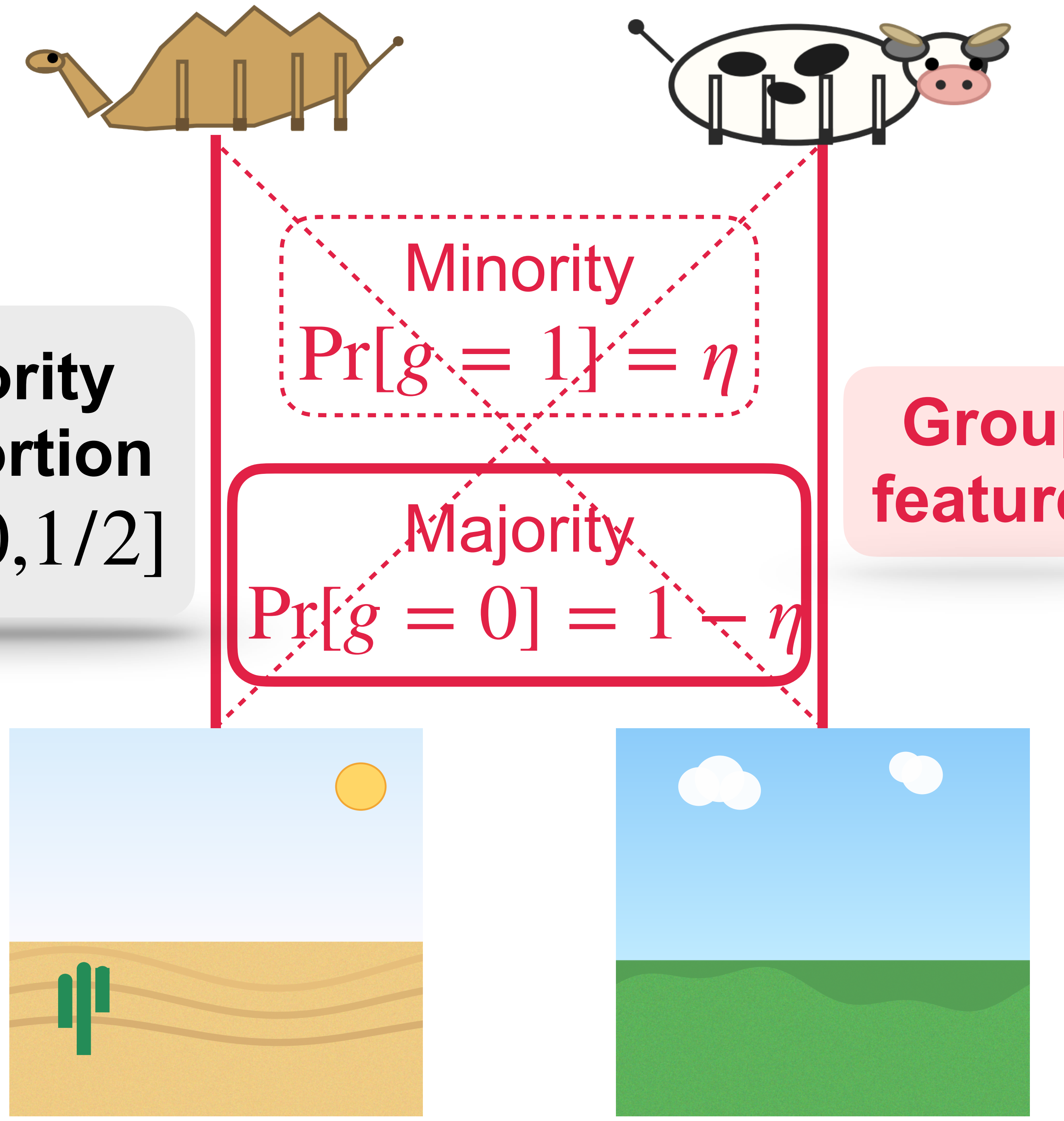
Autonomous driving data



Classify cow vs. camel

Semantic features
encoded in dim. d_z

Minority proportion
$\eta \in [0,1/2]$



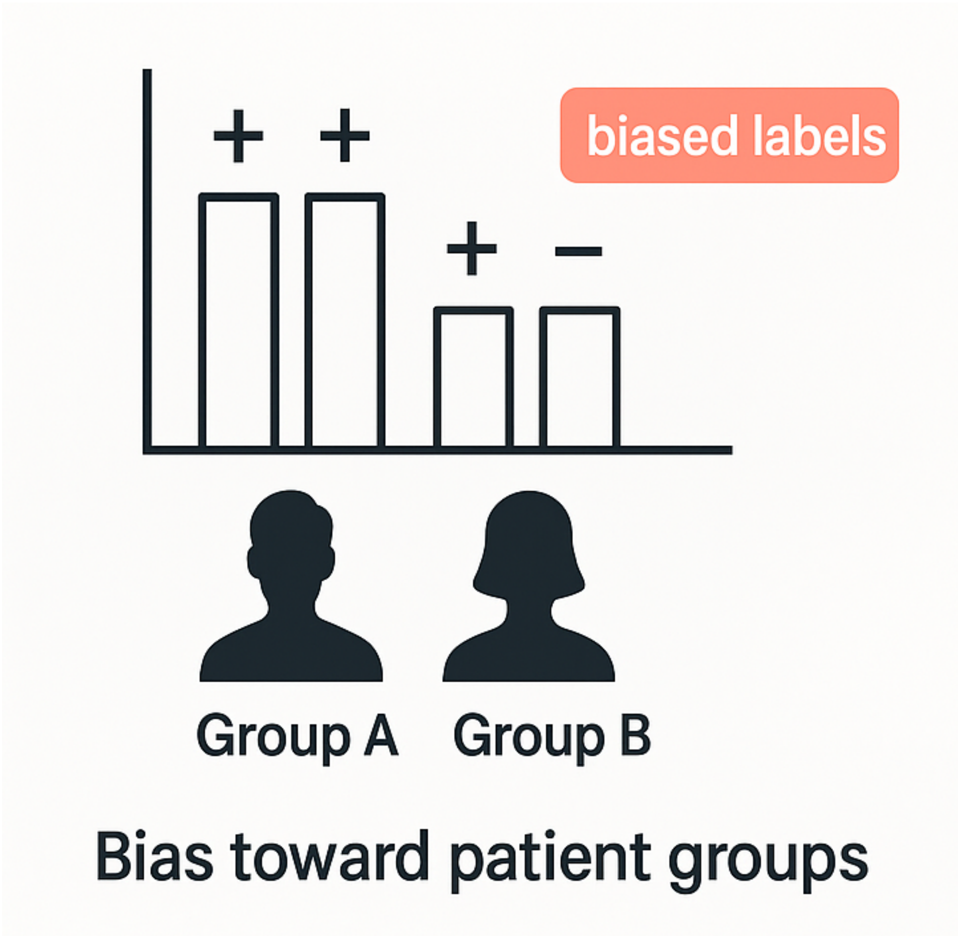
Group features

🙁 When $\eta \ll 1$, backgrounds are learned.

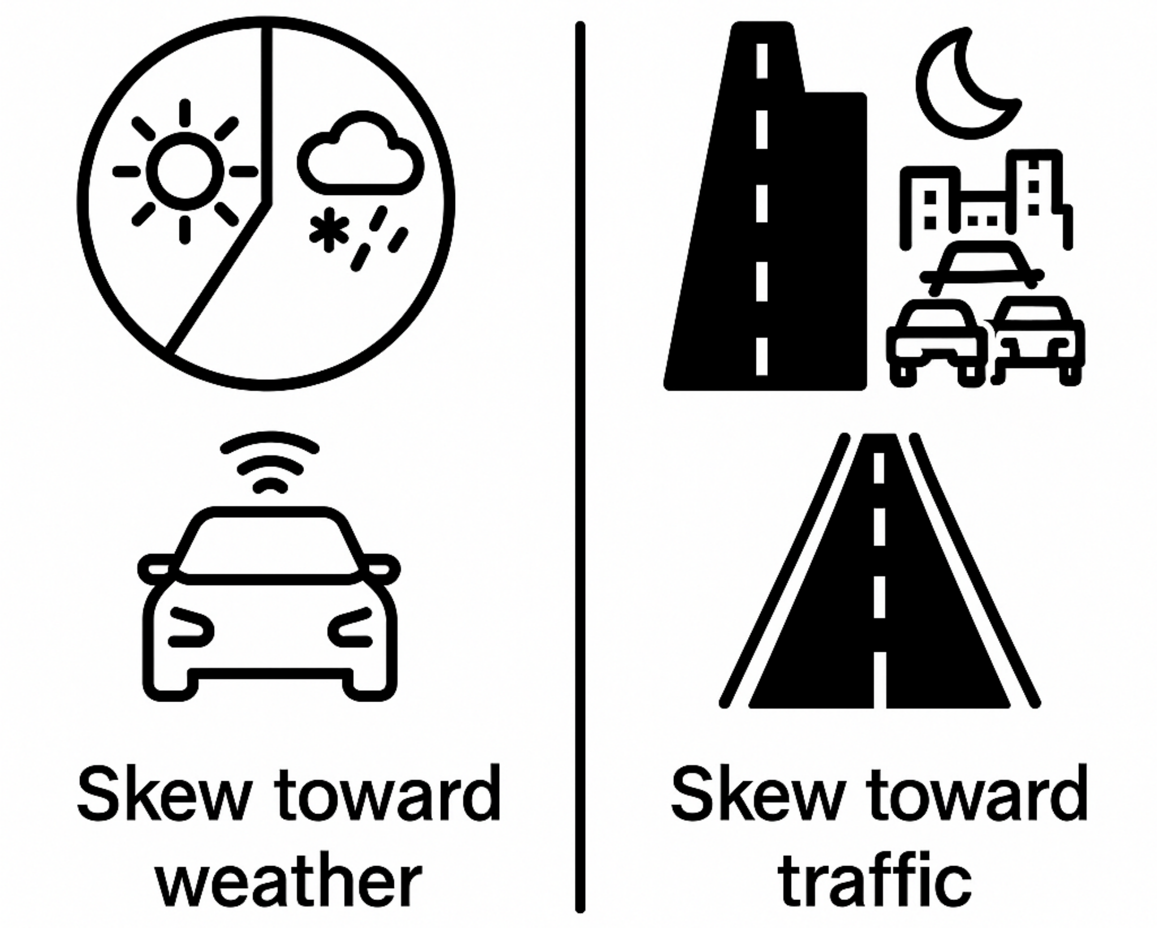
Weak vs. **Strong**:
Representation efficiency for the group features

… where Data Often Come with Group Imbalance

Medical data

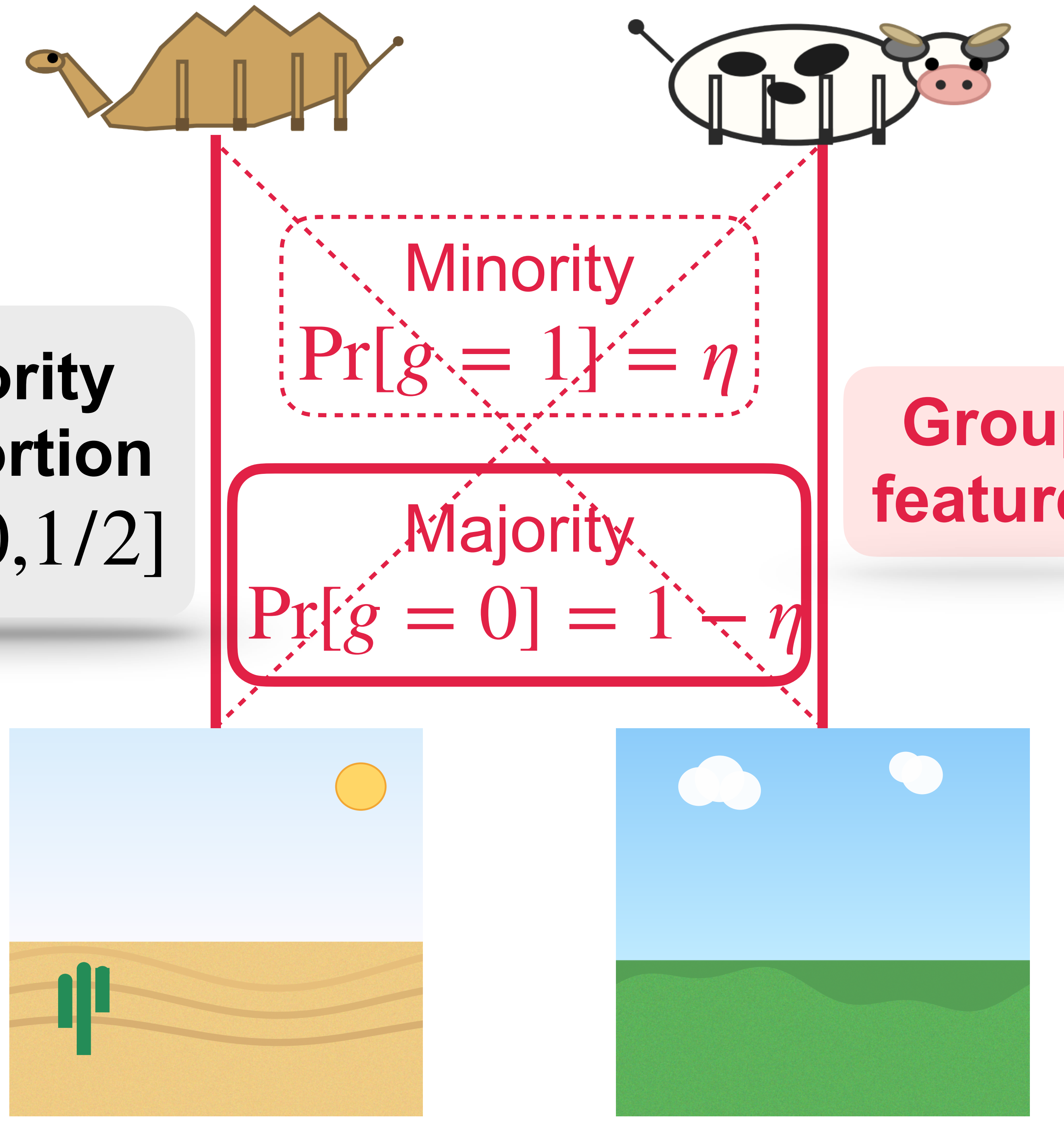


Autonomous driving data



Classify cow vs. camel

Semantic features encoded in dim. d_z



Minority proportion $\eta \in [0,1/2]$

Group features

🙁 When $\eta \ll 1$, backgrounds are learned.

Weak vs. **Strong**: **Representation efficiency** for the group features

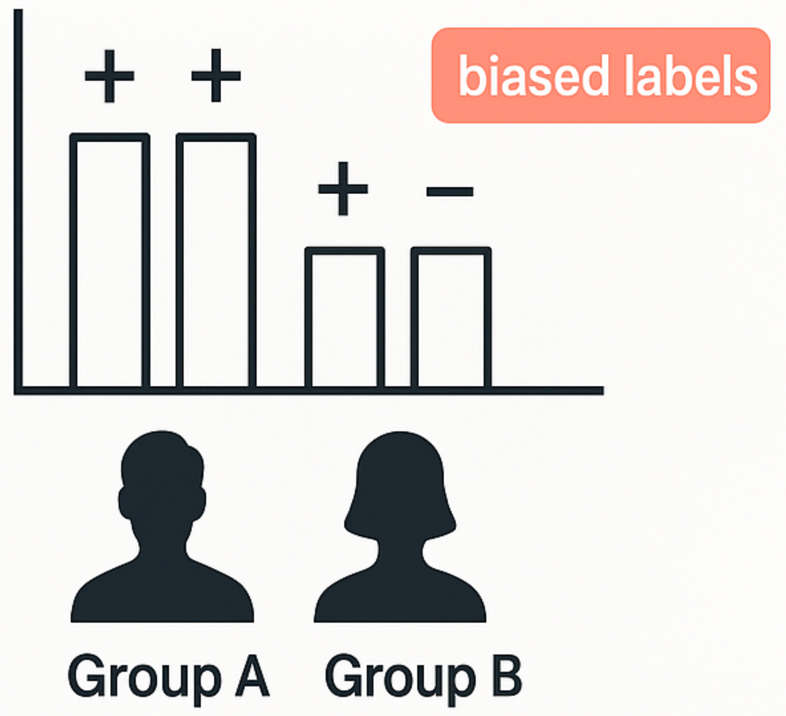
$p_s \leq p_w \ll d_z$

Weak group feature (dim. p_w): counting the frequency of occurrence in pre-training data

Strong group feature (dim. p_s): knowledge about natural habitats from pre-training

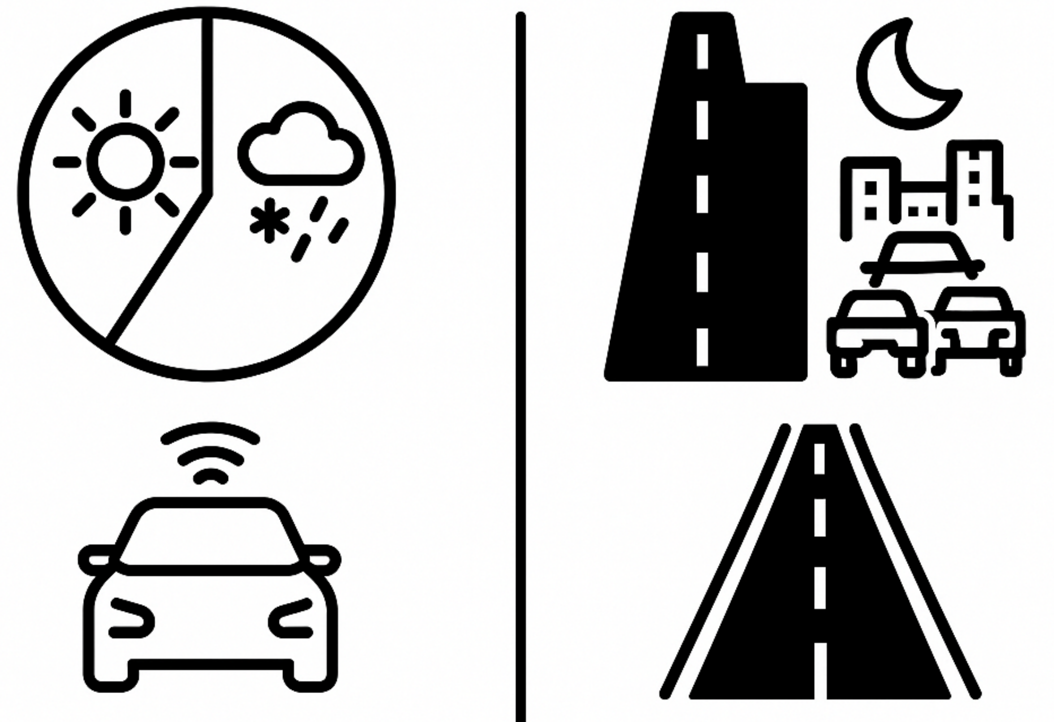
… where Data Often Come with Group Imbalance

Medical data



Bias toward patient groups

Autonomous driving data

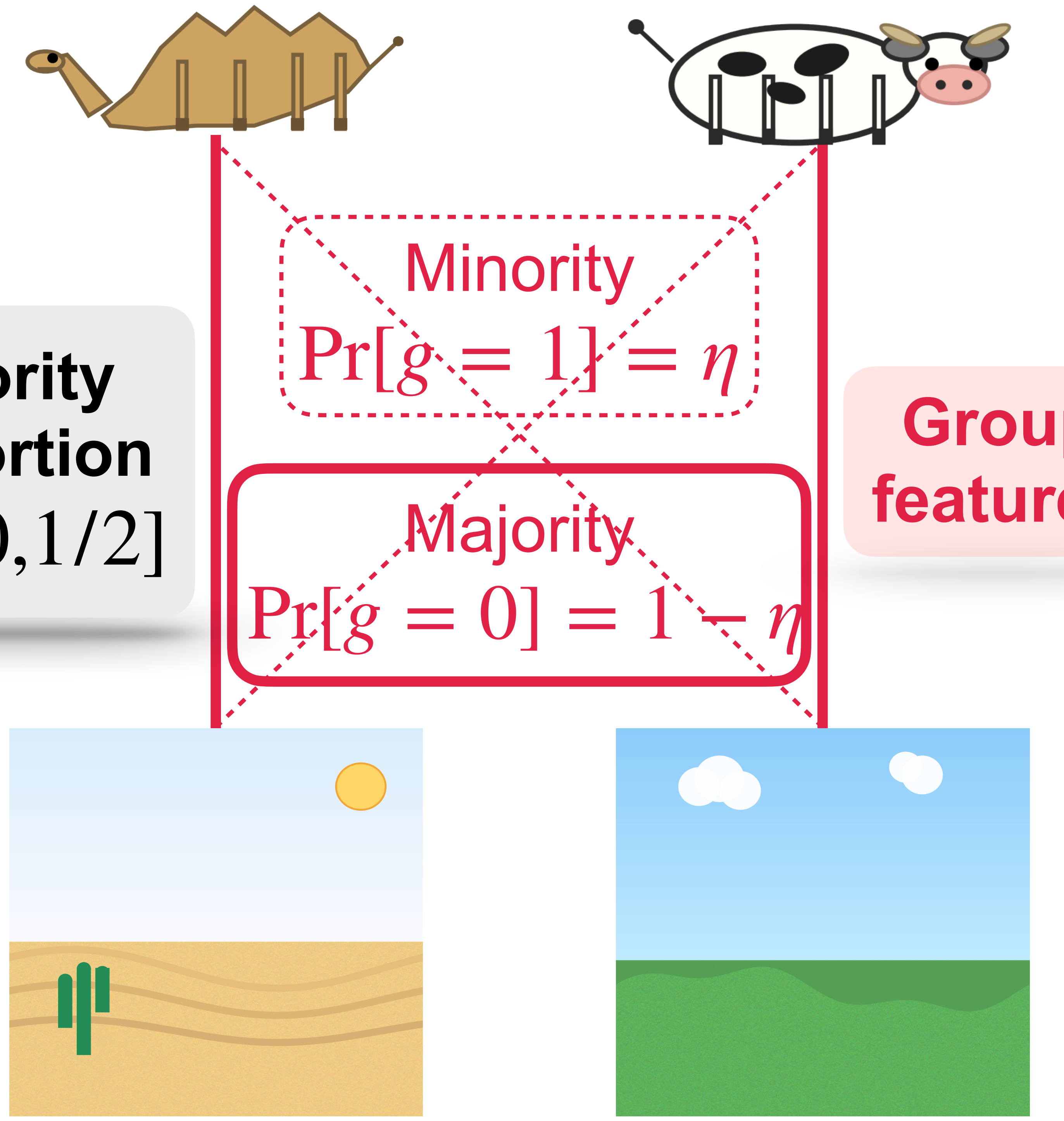


Skew toward weather

Skew toward traffic

Classify cow vs. camel

Semantic features
encoded in dim. d_z



Minority proportion
$\eta \in [0,1/2]$

Group features

🙁 When $\eta \ll 1$, backgrounds are learned.

Weak vs. **Strong**:
Representation efficiency for the group features

$p_s \leq p_w \ll d_z$

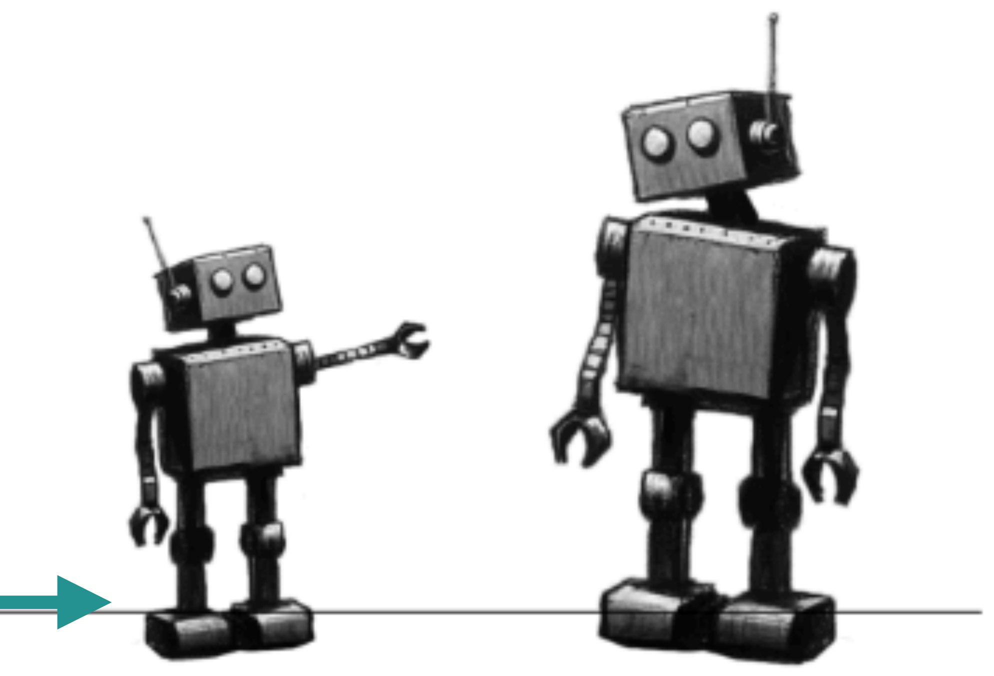
Weak group feature (dim. p_w):
counting the frequency of occurrence in pre-training data

Strong group feature (dim. p_s):
knowledge about natural habitats from pre-training

Group feature similarity:
$1 \leq p_{s \wedge w} \leq p_s \leq p_w$
(analogous to the **correlation dimension** introduced before)

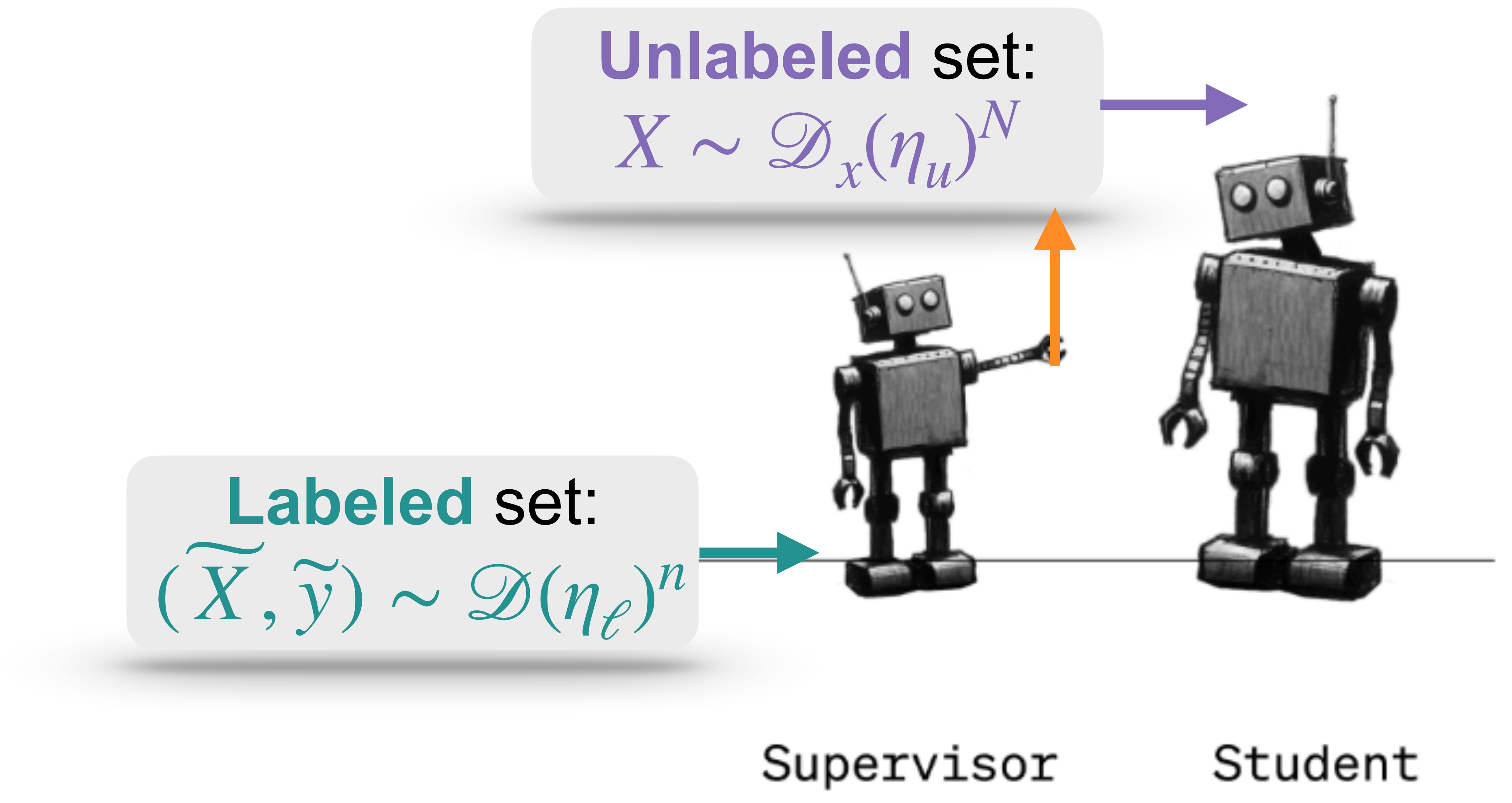
W2S Gain $\searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(\eta_\ell)^n$



Supervisor Student

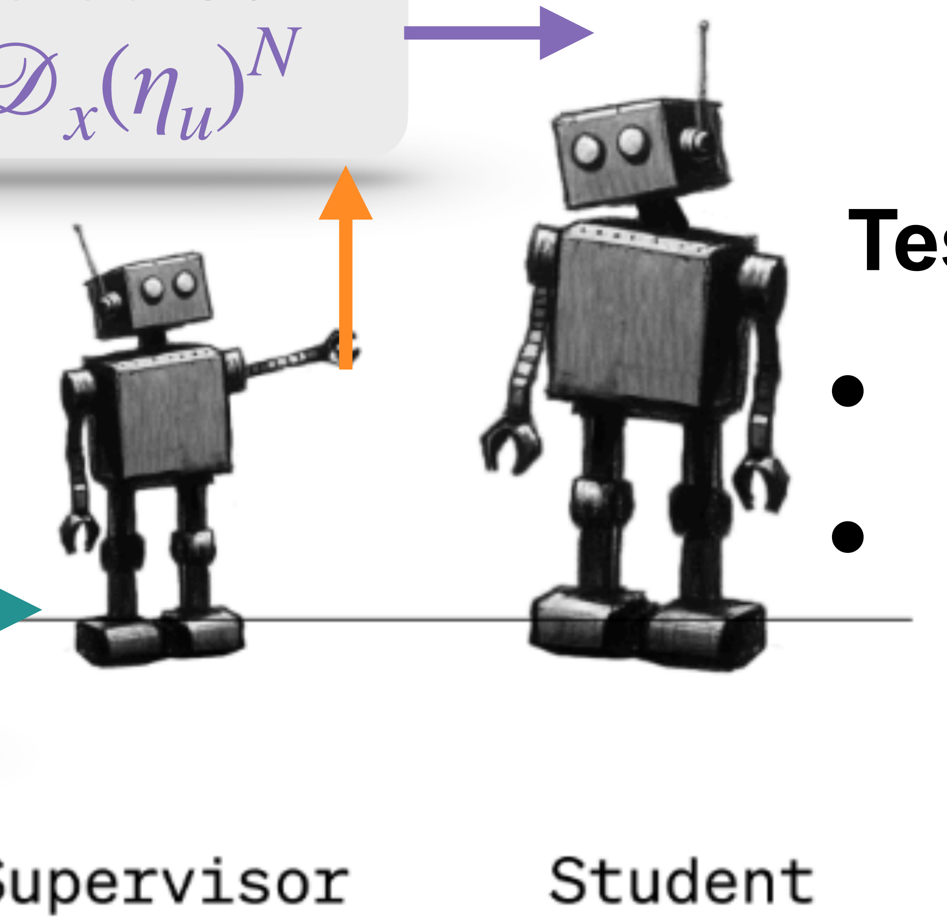
W2S Gain ↘ as $(\eta_u - \eta_\ell)^2$ ↗



W2S Gain $\searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

Unlabeled set:
$X \sim \mathcal{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(\eta_\ell)^n$



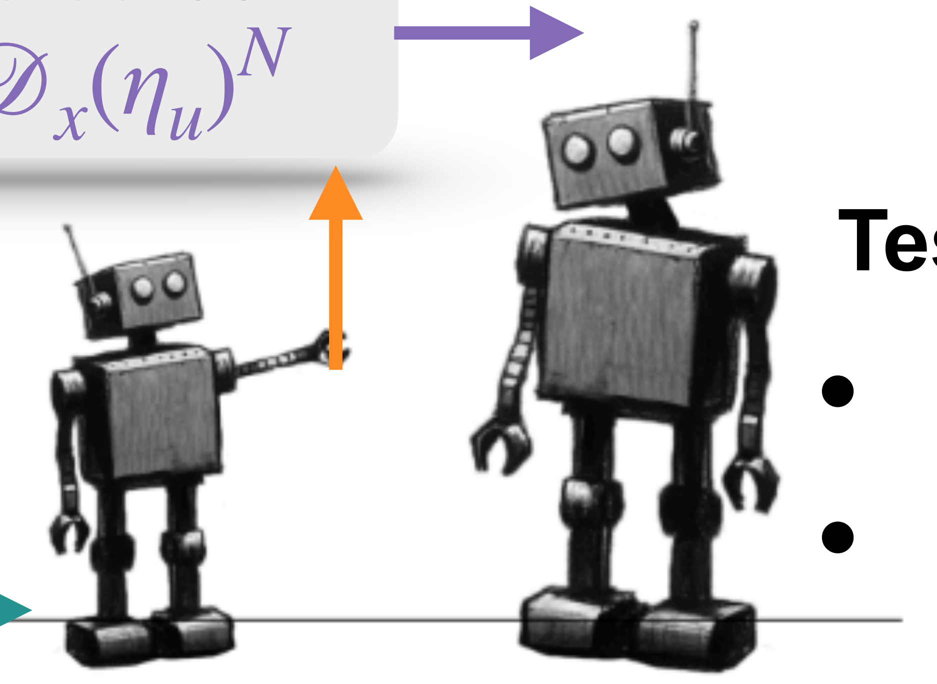
Test distribution: $\mathcal{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

W2S Gain $\searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

Unlabeled set:
$X \sim \mathcal{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(\eta_\ell)^n$



Supervisor Student

Test distribution: $\mathcal{D}(\eta_t)$

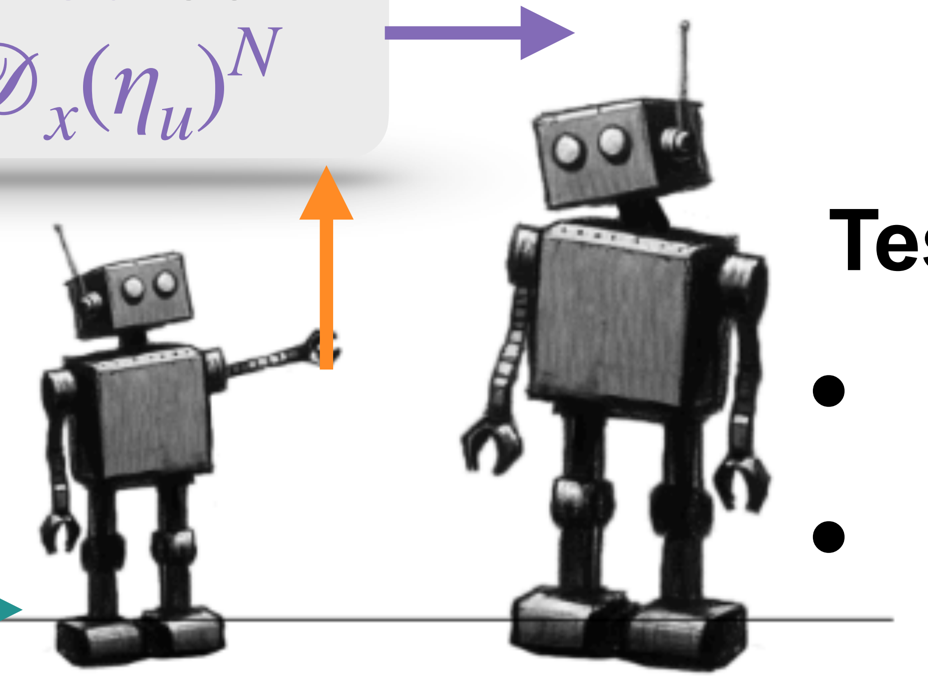
- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

Proportional asymptotic limit:
$d_z, n, N \to \infty,$
$d_z/n \to \gamma_z,\ d_z/N \to \nu_z,\ p_s \le p_w < \infty$
⬇
Precise W2S gain:
$$\Delta\mathcal{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$$

W2S Gain ↘ as $(\eta_u - \eta_\ell)^2$ ↗

Unlabeled set:
$X \sim \mathscr{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathscr{D}(\eta_\ell)^n$



Supervisor Student

Test distribution: $\mathscr{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

Proportional asymptotic limit:

$$d_z, n, N \to \infty,$$

$$d_z/n \to \gamma_z,\ d_z/N \to \nu_z,\ p_s \le p_w < \infty$$

⬇

Precise W2S gain:

$$\Delta\mathscr{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$$

Theorem (informal [LDL25]). When both teacher and student are unbiased over the population, assuming low group feature similarity, $p_{s\wedge w} \ll p_s \le p_w$, and large unlabeled data size $\nu_z \ll 1$

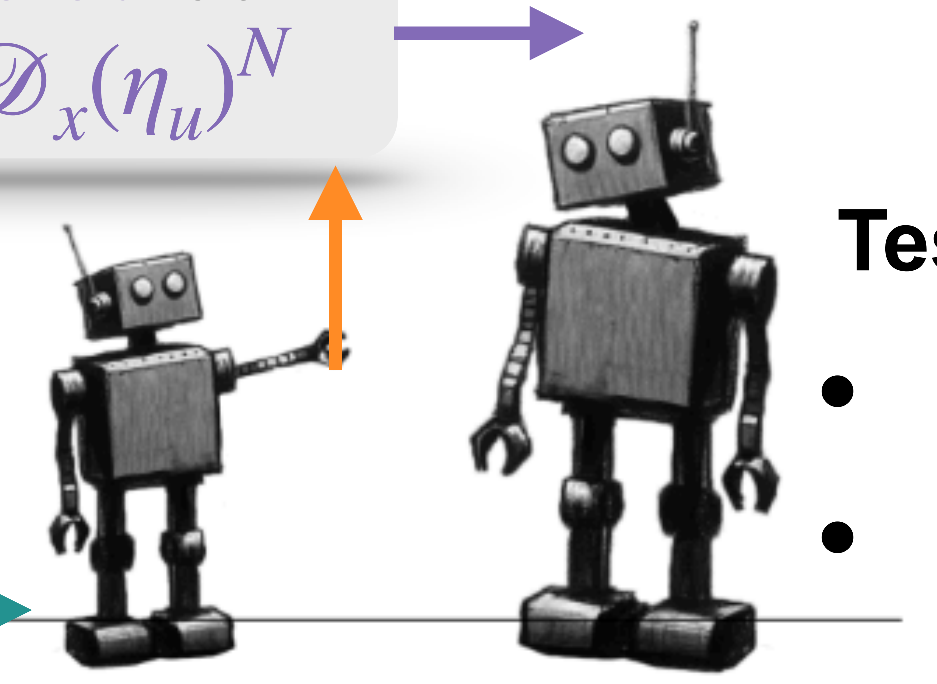
$$\mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(\qquad + \qquad \Big)$$

$$\mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(\qquad + \qquad \Big)$$

W2S Gain ↘ as $(\eta_u - \eta_\ell)^2$ ↗

Unlabeled set:
$X \sim \mathscr{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathscr{D}(\eta_\ell)^n$



Test distribution: $\mathscr{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

Proportional asymptotic limit:
$d_z, n, N \to \infty$,
$d_z/n \to \gamma_z$, $d_z/N \to \nu_z$, $p_s \le p_w < \infty$
⬇
Precise W2S gain:
$$\Delta\mathscr{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$$

Theorem (informal [LDL25]). When both teacher and student are unbiased over the population, assuming low group feature similarity, $p_{s\wedge w} \ll p_s \le p_w$, and large unlabeled data size $\nu_z \ll 1$

From label noise

$$\mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_w + \Big)$$

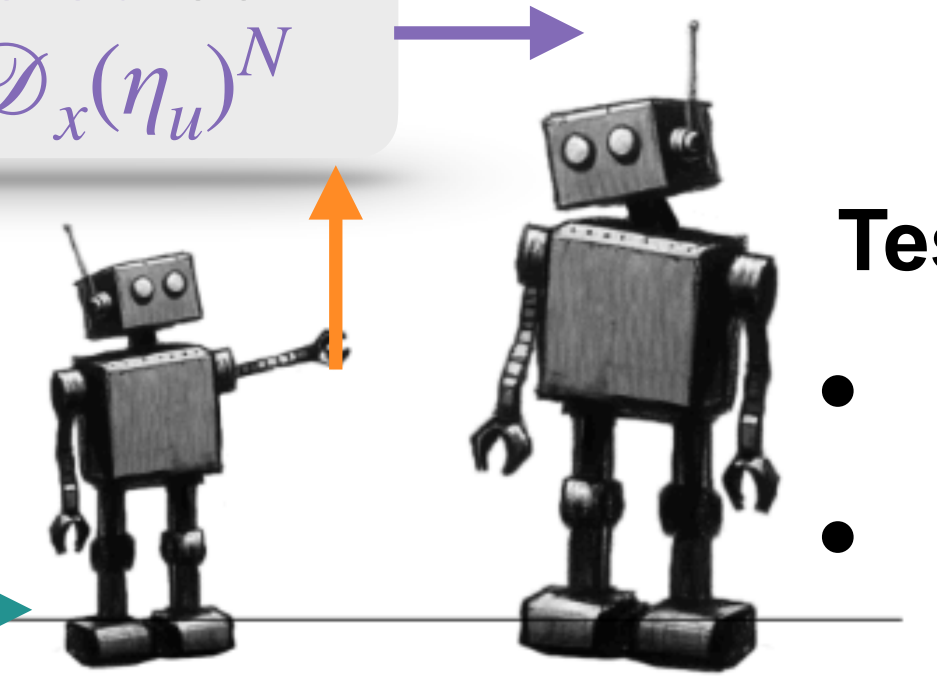
$p_{s\wedge w} + \Theta(\nu_z) \le p_w$

$$\mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_{s\wedge w} + \nu_z p_s (p_w - p_{s\wedge w}) + \Big)$$

W2S Gain $\searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

Unlabeled set:
$X \sim \mathscr{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathscr{D}(\eta_\ell)^n$



Supervisor Student

Test distribution: $\mathscr{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

Proportional asymptotic limit:
$d_z, n, N \to \infty,$
$d_z/n \to \gamma_z,\ d_z/N \to \nu_z,\ p_s \le p_w < \infty$
⬇
Precise W2S gain:
$\Delta\mathscr{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$

Theorem (informal [LDL25]). When both teacher and student are unbiased over the population, assuming low group feature similarity, $p_{s\wedge w} \ll p_s \le p_w$, and large unlabeled data size $\nu_z \ll 1$

From label noise | From group imbalance

$$\mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_w + (\eta_t - \eta_\ell)^2 \Big)$$

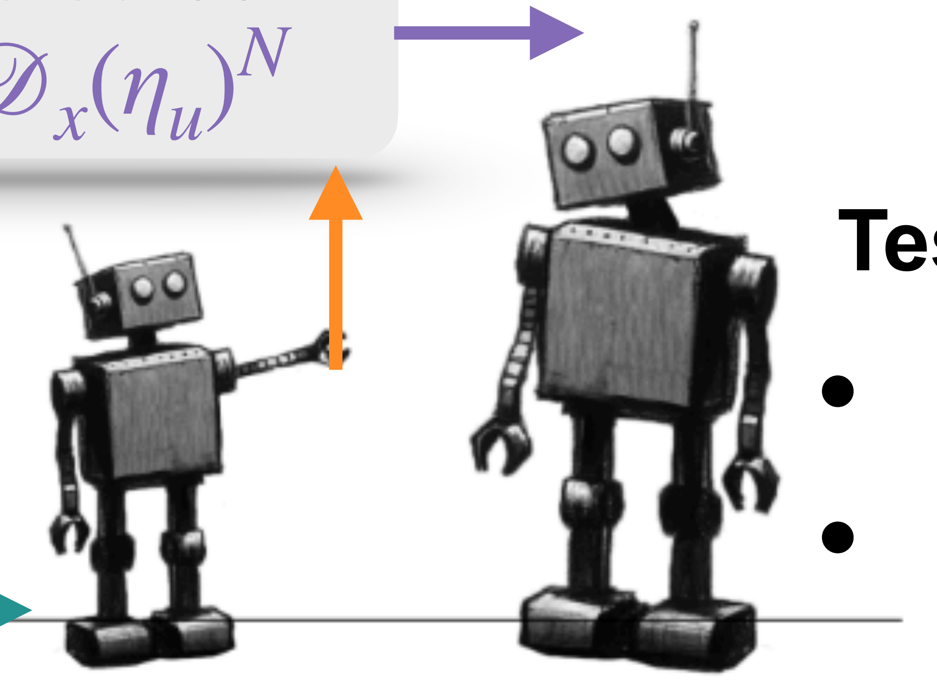
$p_{s\wedge w} + \Theta(\nu_z) \le p_w$ | Negligible when $\eta_\ell = \eta_u$

$$\mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_{s\wedge w} + \nu_z p_s (p_w - p_{s\wedge w}) + (\eta_u - \eta_\ell)^2 + \nu_z (\eta_t - \eta_u)^2 \Big)$$

W2S Gain $\searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

Unlabeled set:
$X \sim \mathcal{D}_x(\eta_u)^N$

Labeled set:
$(\widetilde{X}, \widetilde{y}) \sim \mathcal{D}(\eta_\ell)^n$



Test distribution: $\mathcal{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Minority: $\eta_t = 1$

Proportional asymptotic limit:
$d_z, n, N \to \infty$,
$d_z/n \to \gamma_z$, $d_z/N \to \nu_z$, $p_s \leq p_w < \infty$
⬇
Precise W2S gain:
$\Delta\mathcal{R}_{\eta_t} = \mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] - \mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)]$

Theorem (informal [LDL25]). When both teacher and student are unbiased over the population, assuming low group feature similarity, $p_{s\wedge w} \ll p_s \leq p_w$, and large unlabeled data size $\nu_z \ll 1$

From label noise | From group imbalance

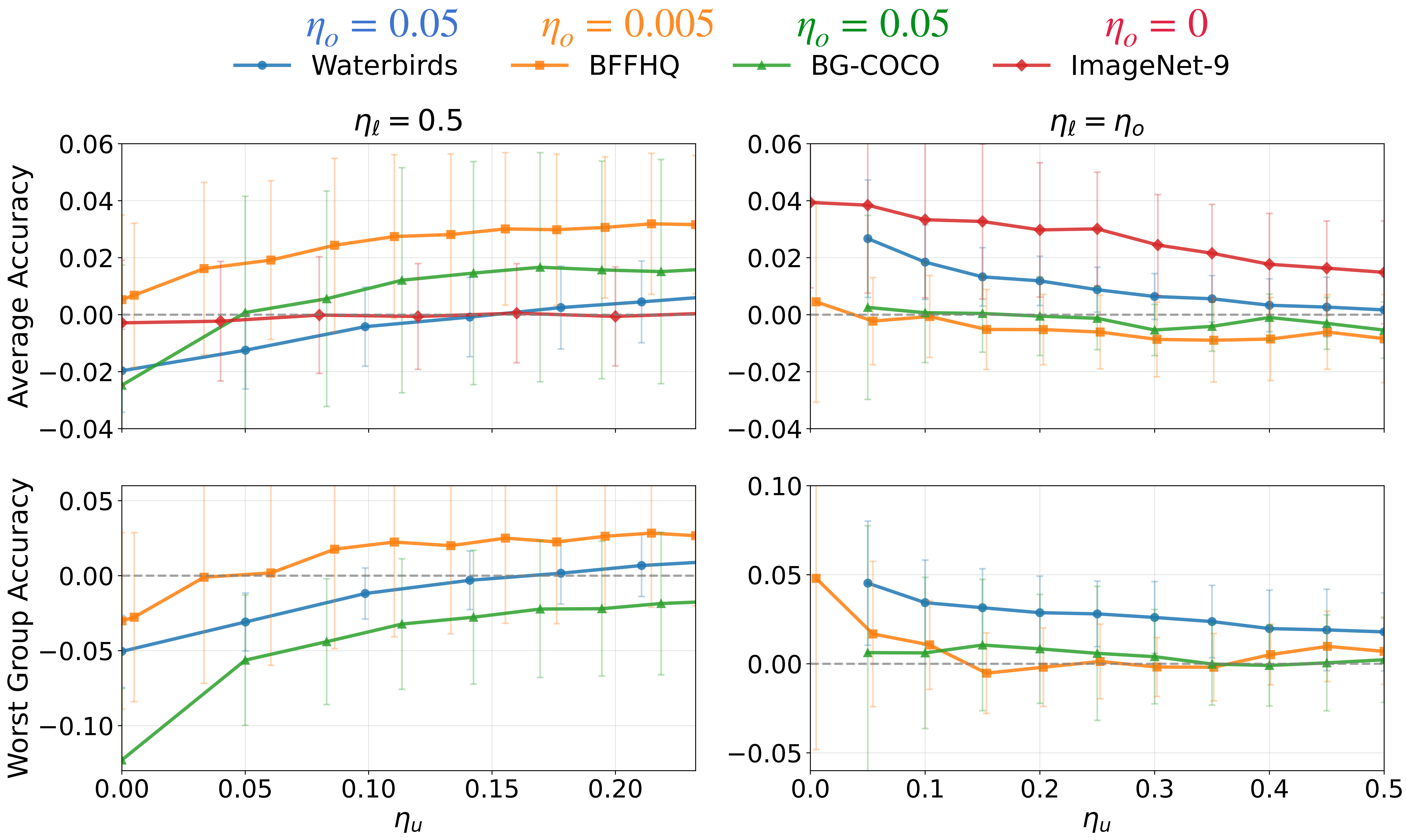
$$\mathbb{E}_{\eta_\ell}[\mathbf{ER}_{\eta_t}(f_w)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_w + (\eta_t - \eta_\ell)^2 \Big)$$

$p_{s\wedge w} + \Theta(\nu_z) \leq p_w$ | Negligible when $\eta_\ell = \eta_u$

$$\mathbb{E}_{\eta_\ell,\eta_u}[\mathbf{ER}_{\eta_t}(f_s)] \xrightarrow{\mathbb{P}} \gamma_z \Theta\Big(p_{s\wedge w} + \nu_z p_s (p_w - p_{s\wedge w}) + (\eta_u - \eta_\ell)^2 + \nu_z (\eta_t - \eta_u)^2 \Big)$$

For $p_{s\wedge w} \ll p_s$ and $\nu_z \ll 1$:
🙂 $\Delta\mathcal{R}_{\eta_t} > 0$ if $\eta_\ell = \eta_u$
🙁 $\Delta\mathcal{R}_{\eta_t} \searrow$ as $(\eta_u - \eta_\ell)^2 \nearrow$

W2S Gain ↘ as $(\eta_u - \eta_\ell)^2$ ↗ in Practice



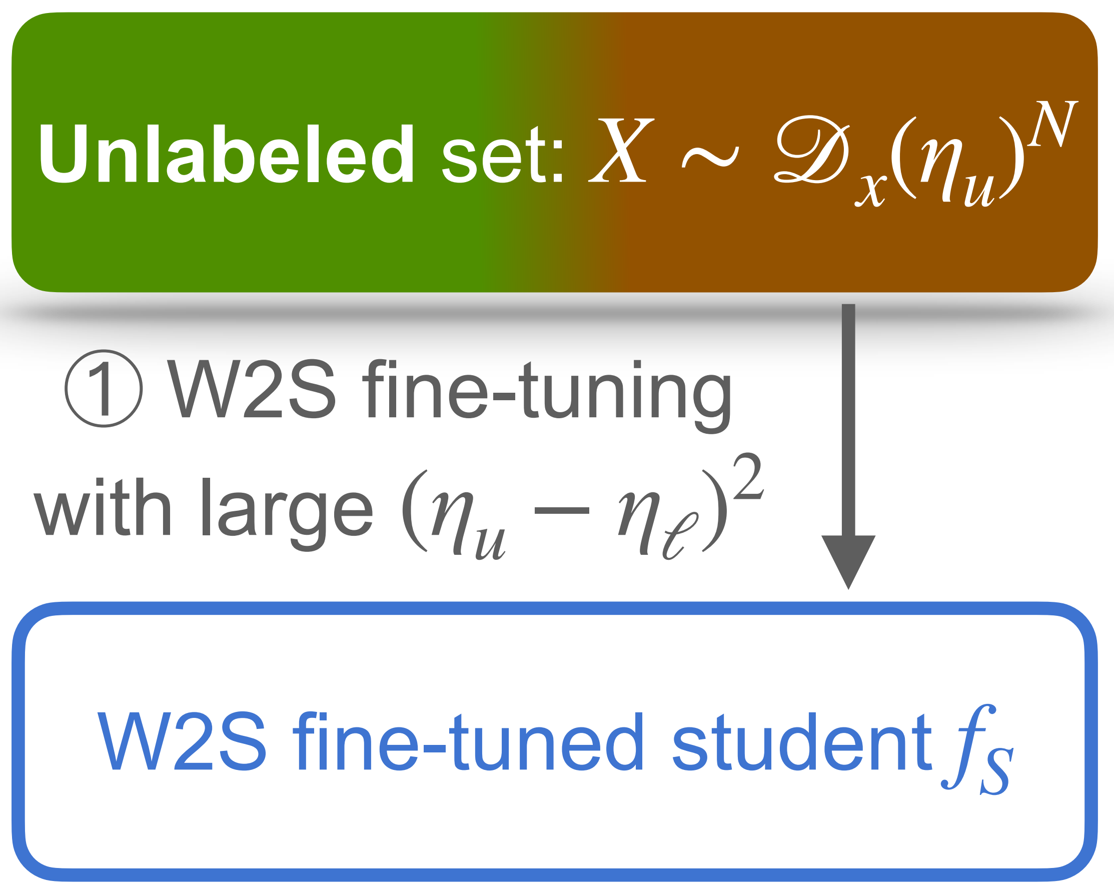
Enhanced W2S for Large $(\eta_u - \eta_\ell)^2$: Selective Retraining

Unlabeled set: $X \sim \mathcal{D}_x(\eta_u)^N$

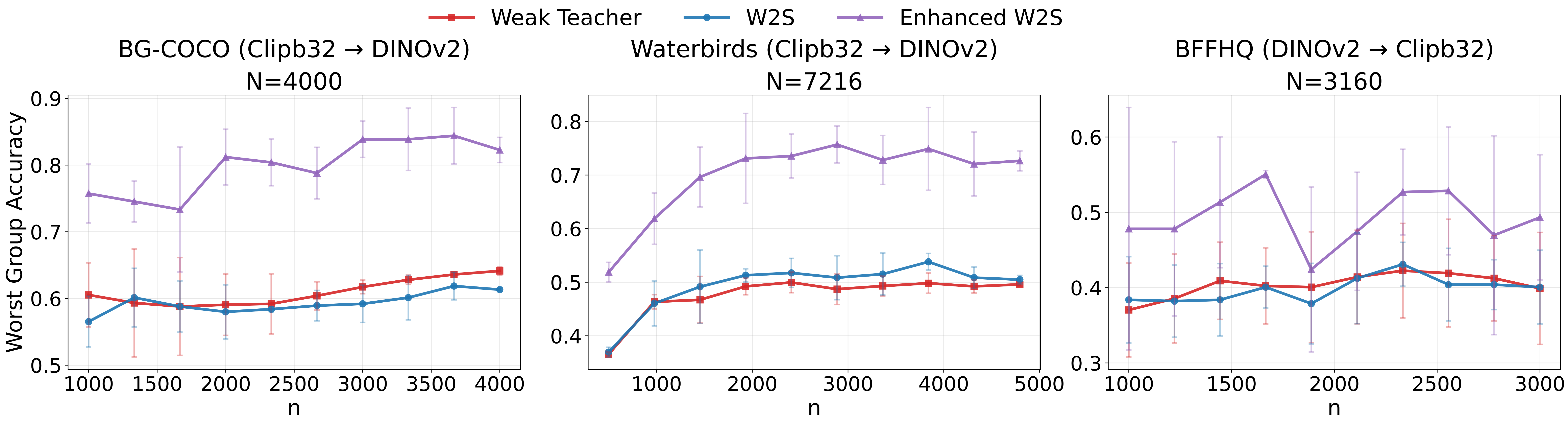
$\eta_\ell \to 0$
$\eta_u = 0.5$

Enhanced W2S for Large $(\eta_u - \eta_\ell)^2$: Selective Retraining

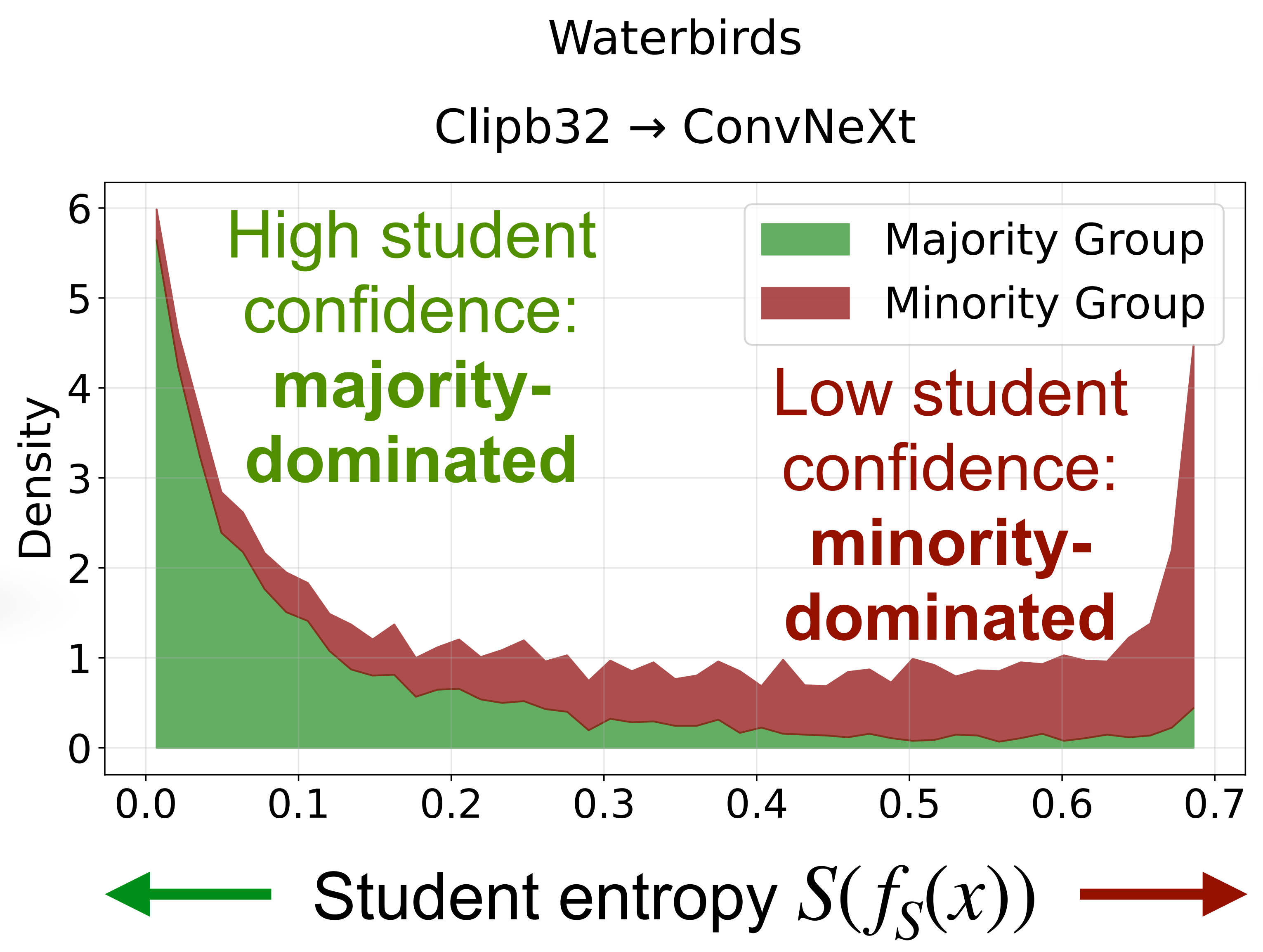
$\eta_\ell \to 0$
$\eta_u = 0.5$



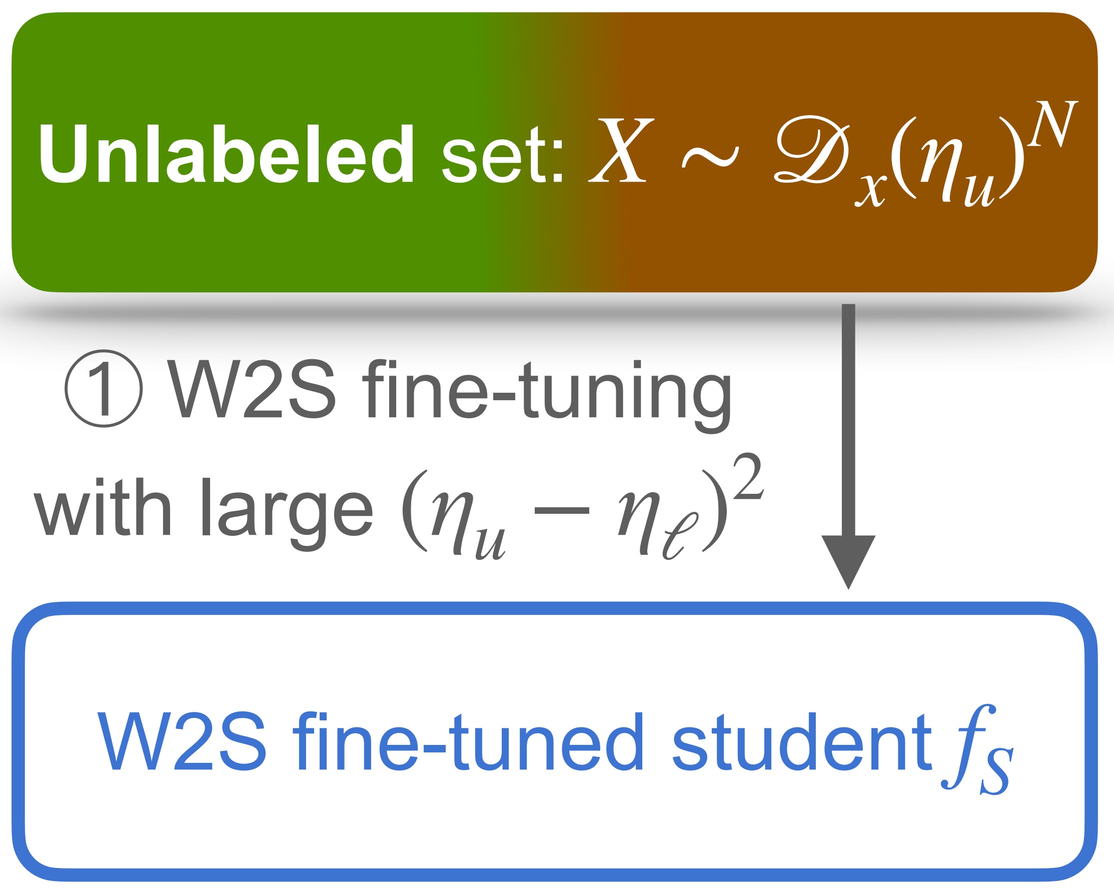
$\eta_\ell = \eta_o$
$\eta_u = 0.5$



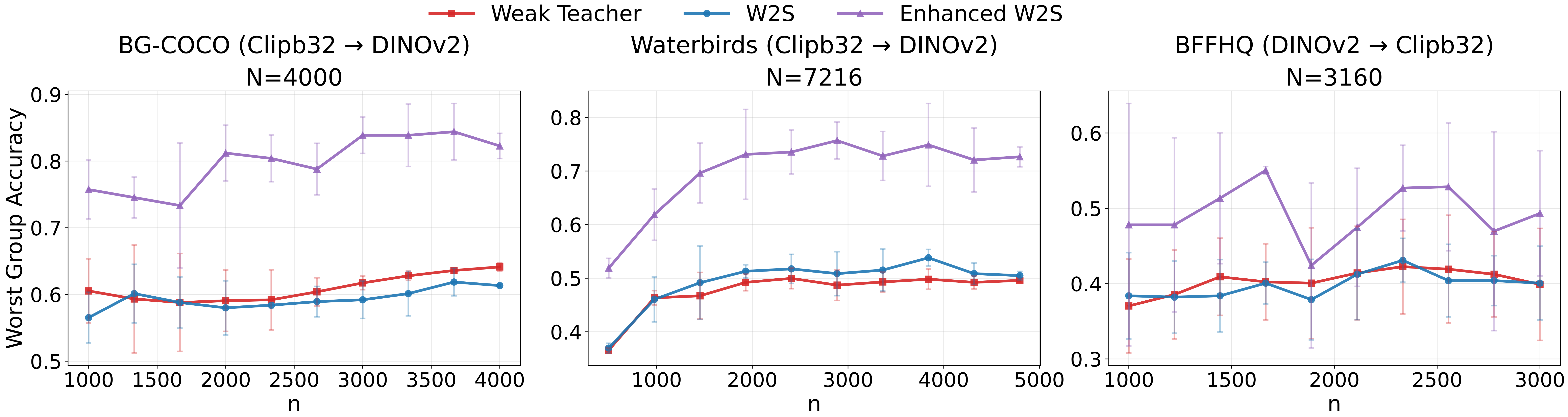
Enhanced W2S for Large $(\eta_u - \eta_\ell)^2$: Selective Retraining



$\eta_\ell \to 0$
$\eta_u = 0.5$

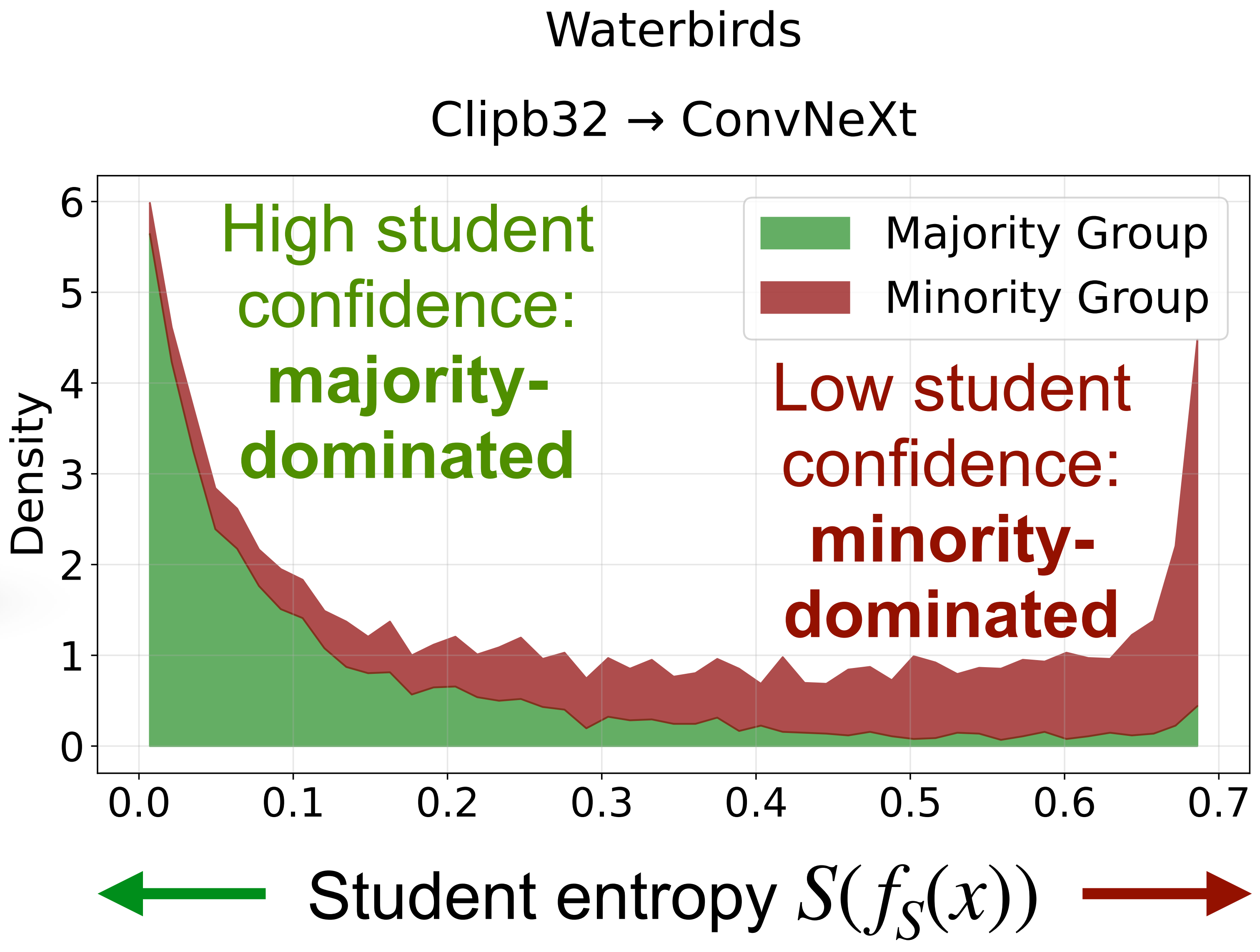


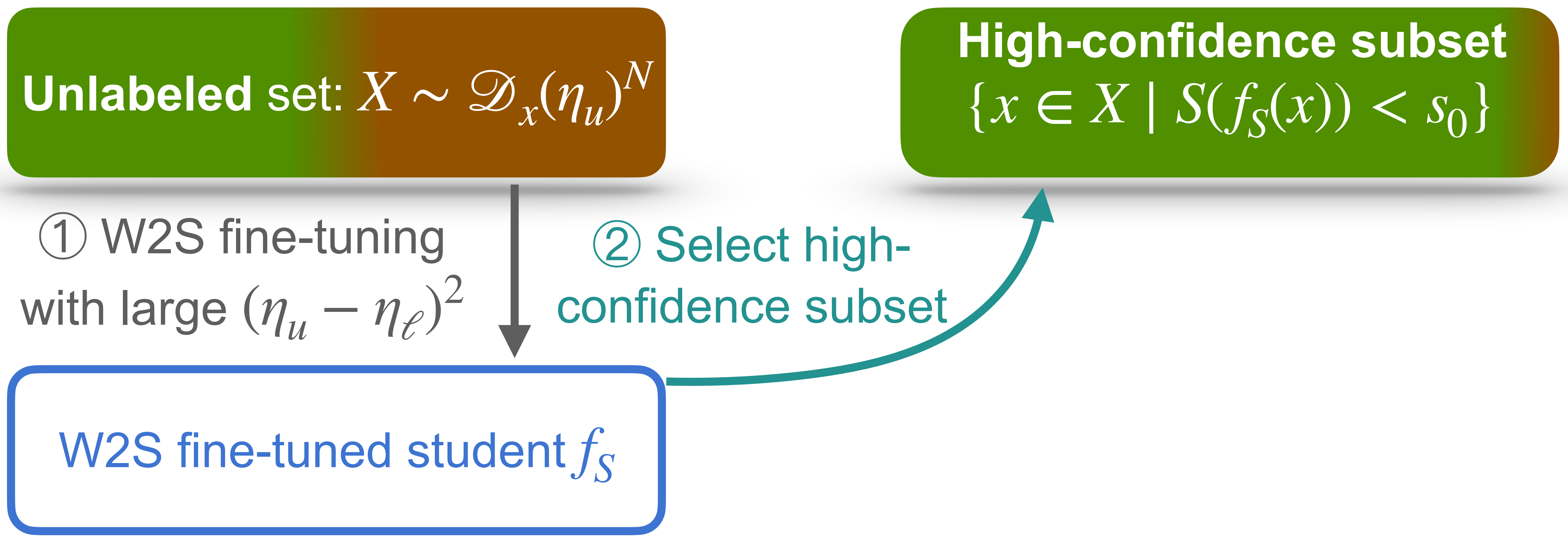
$\eta_\ell = \eta_o$
$\eta_u = 0.5$



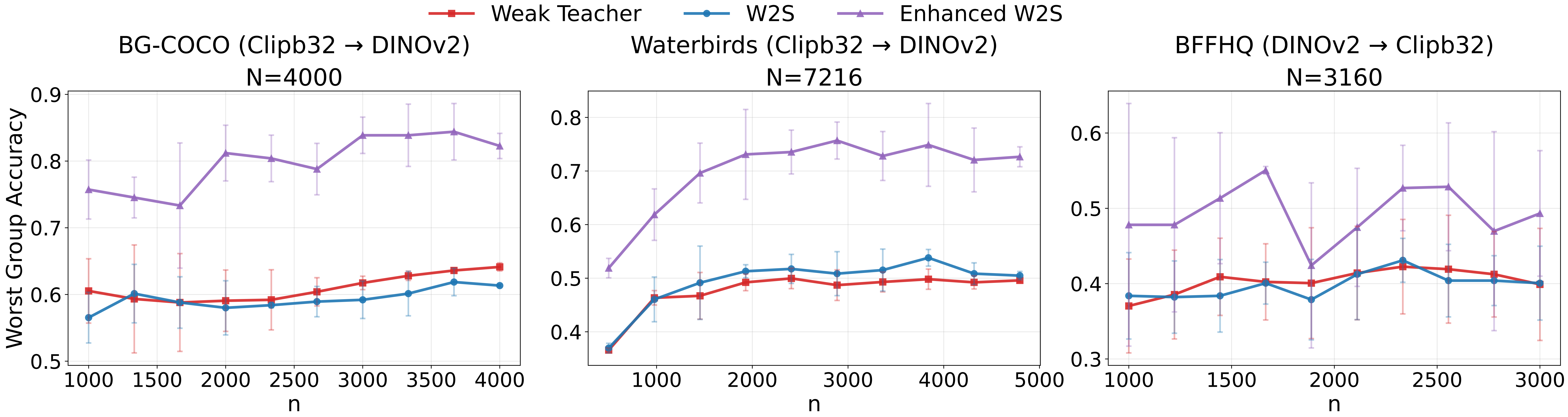
Enhanced W2S for Large $(\eta_u - \eta_\ell)^2$: Selective Retraining

$\eta_\ell \to 0$
$\eta_u = 0.5$



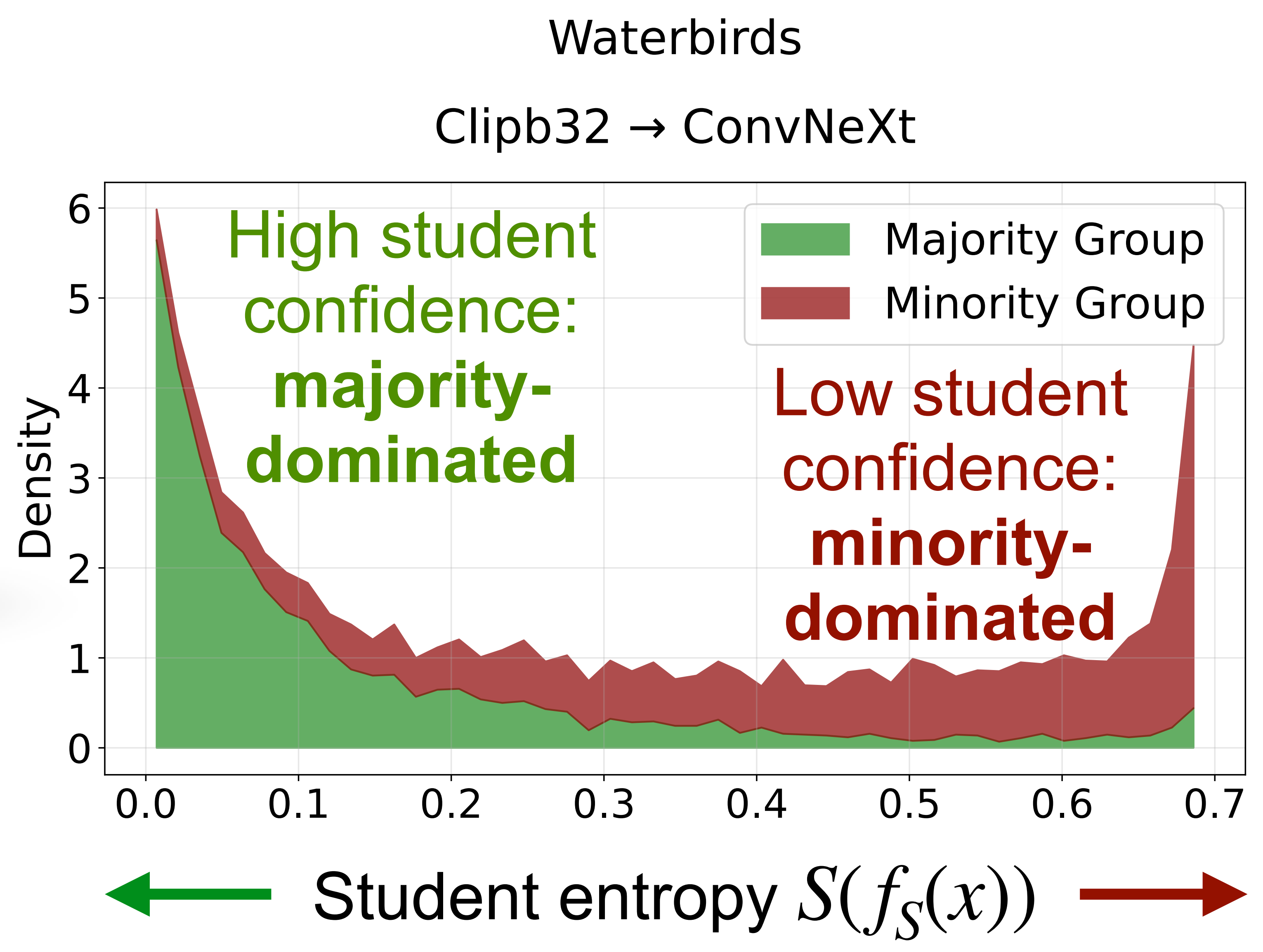


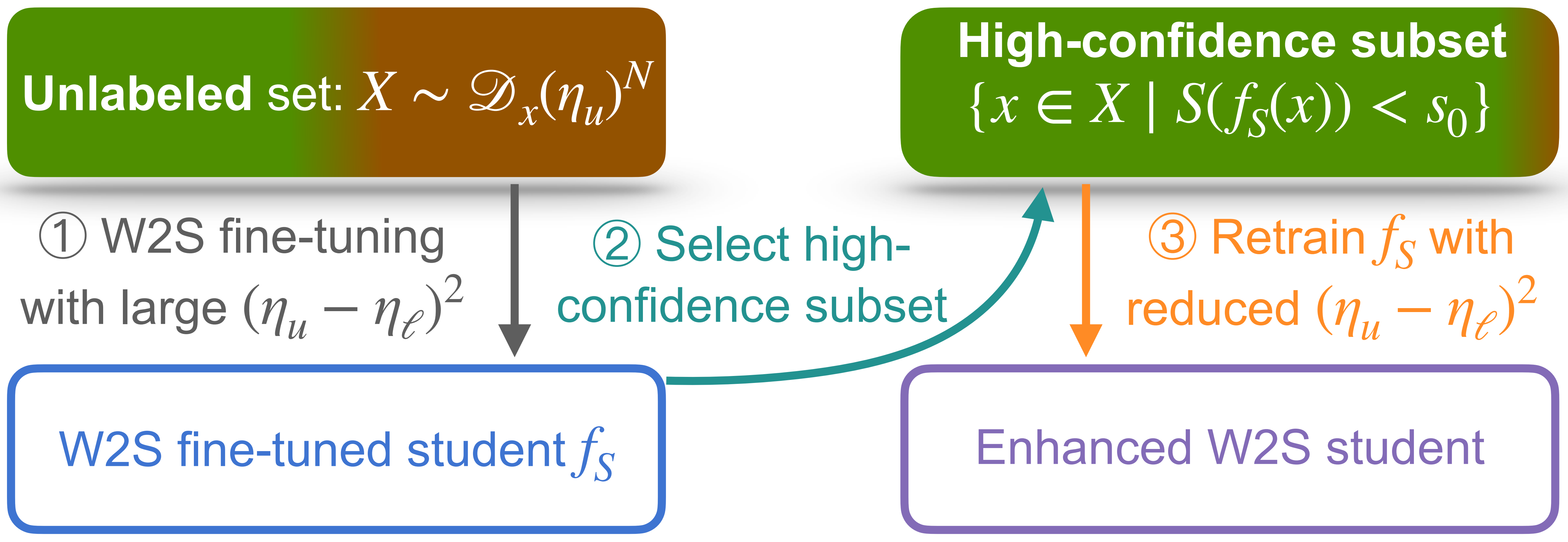
$\eta_\ell = \eta_o$
$\eta_u = 0.5$



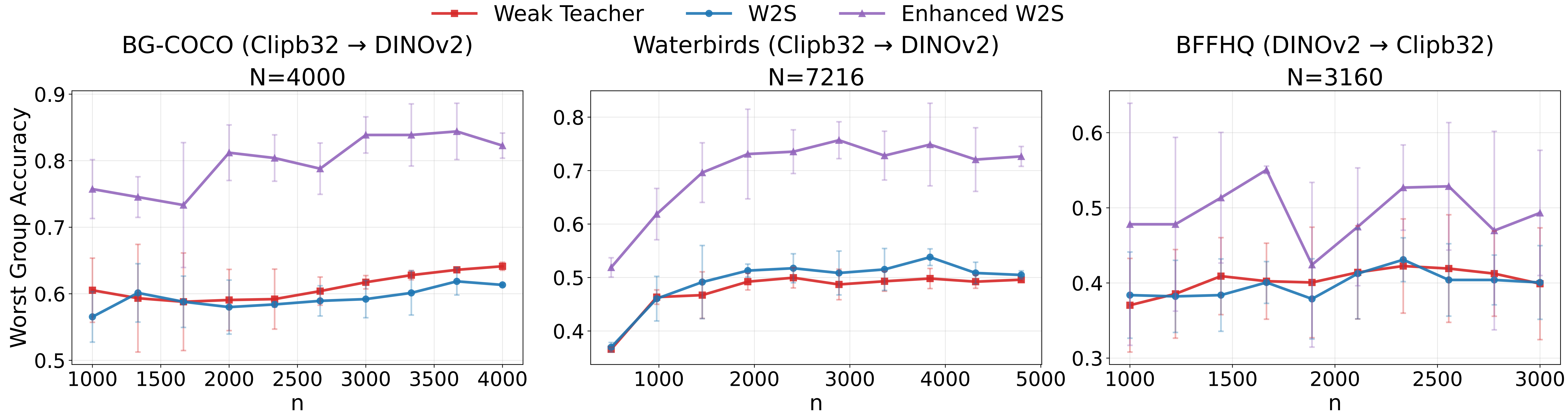
Enhanced W2S for Large $(\eta_u - \eta_\ell)^2$: Selective Retraining

$\eta_\ell \to 0$
$\eta_u = 0.5$





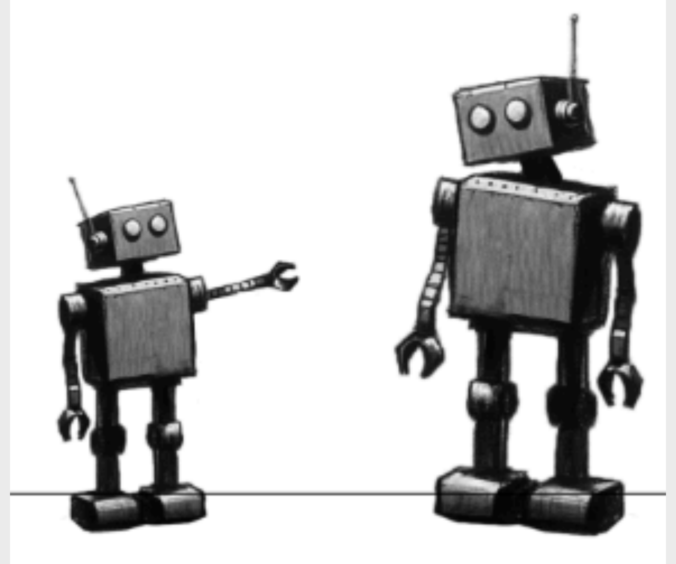
$\eta_\ell = \eta_o$
$\eta_u = 0.5$



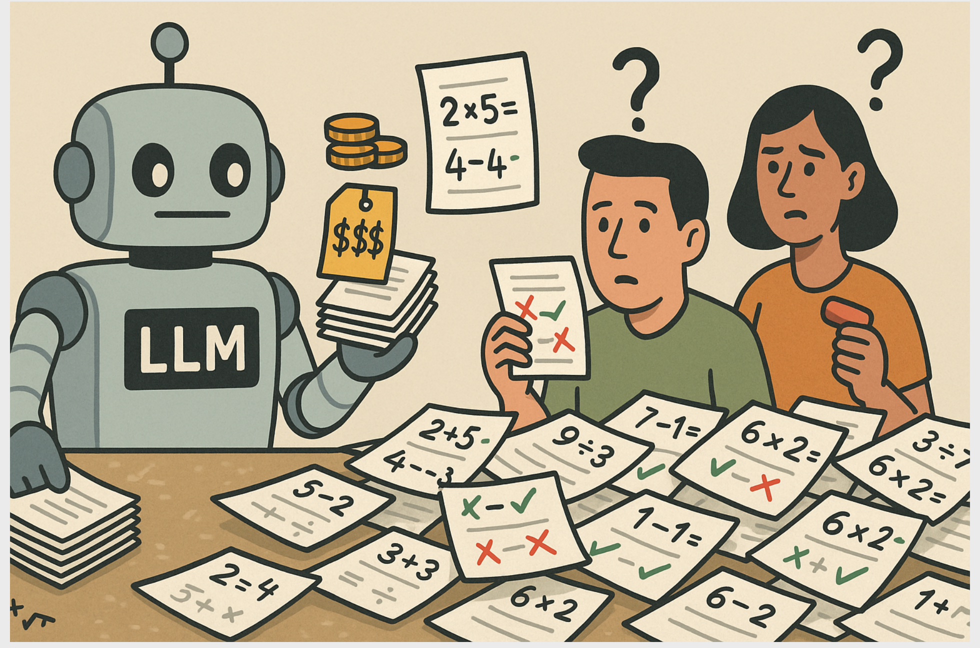
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

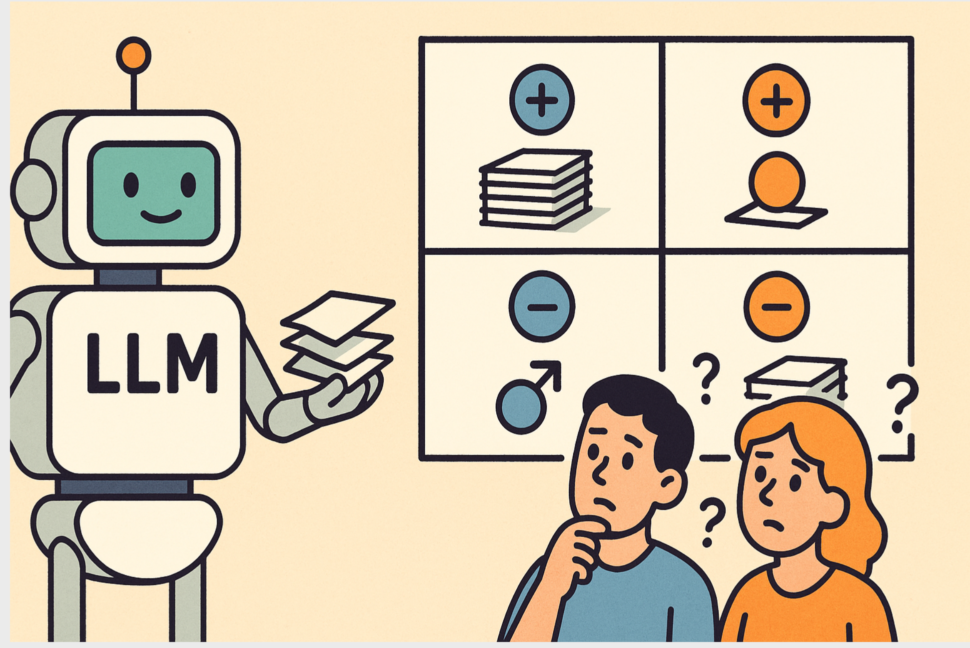
e.g., **W2S**



① **limited & noisy labels**



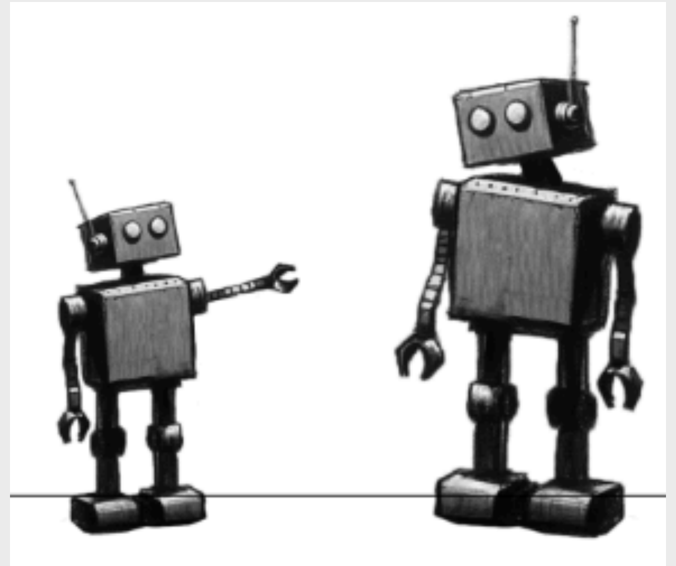
② **systematic bias**



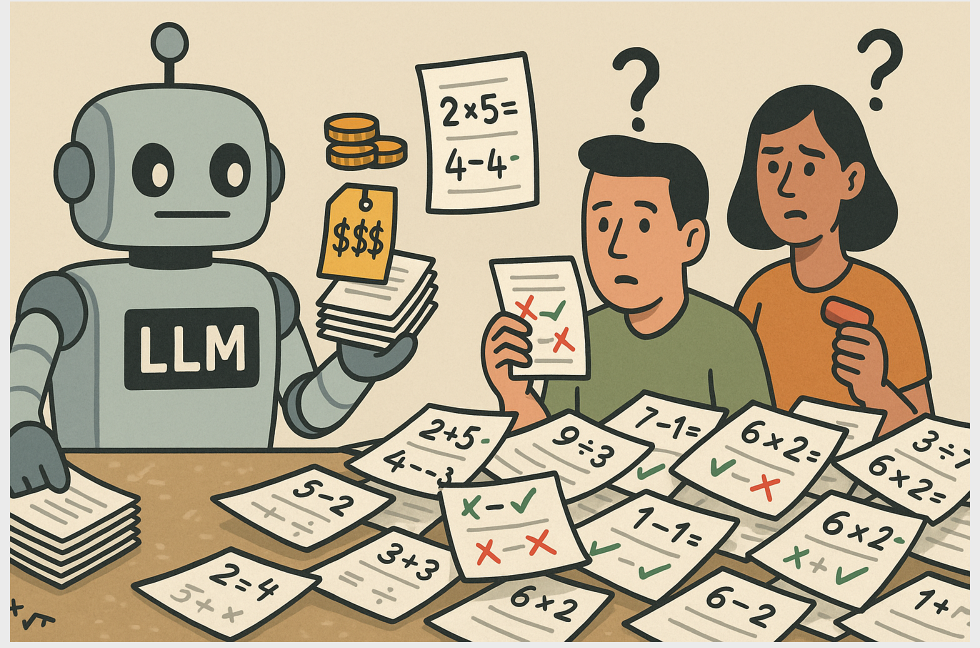
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

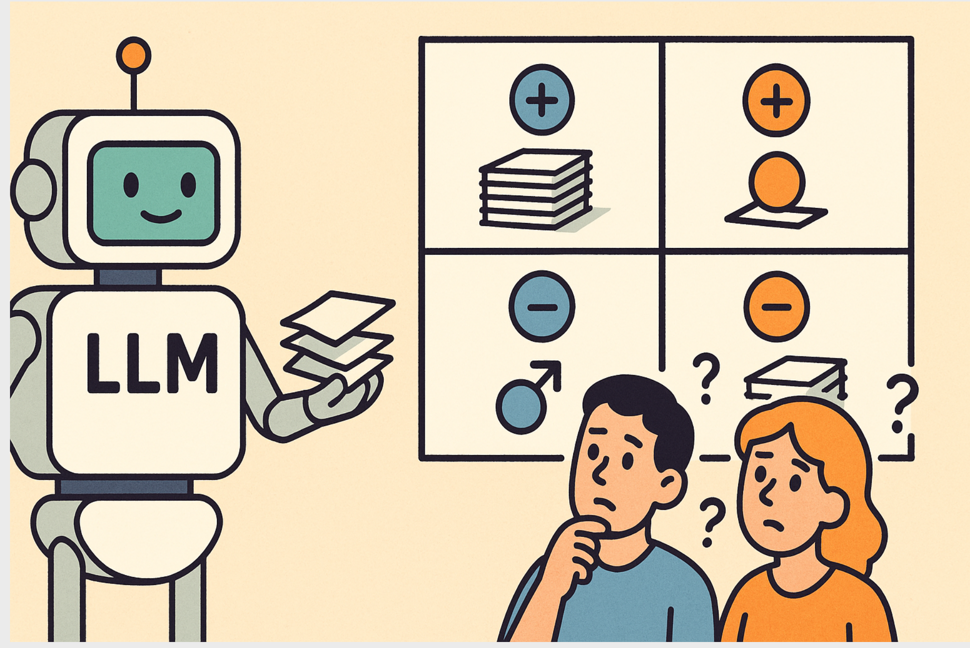
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**

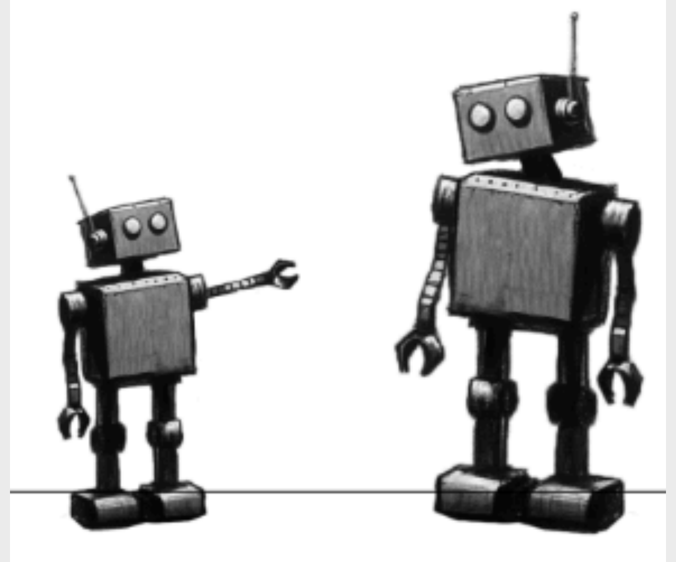


Learning theory: How does W2S happen?
What if the data are group imbalanced?

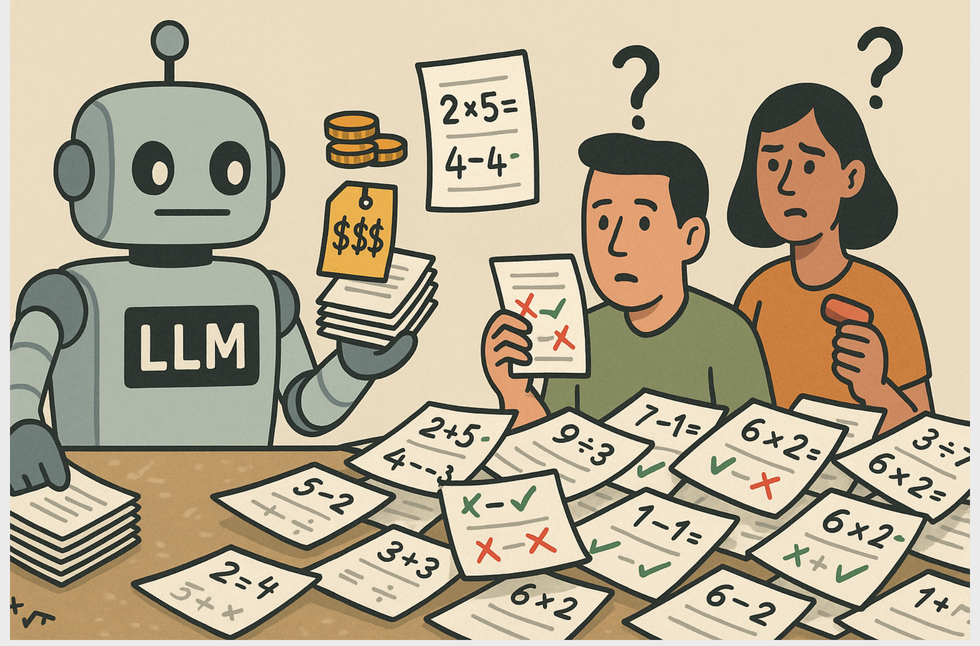
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

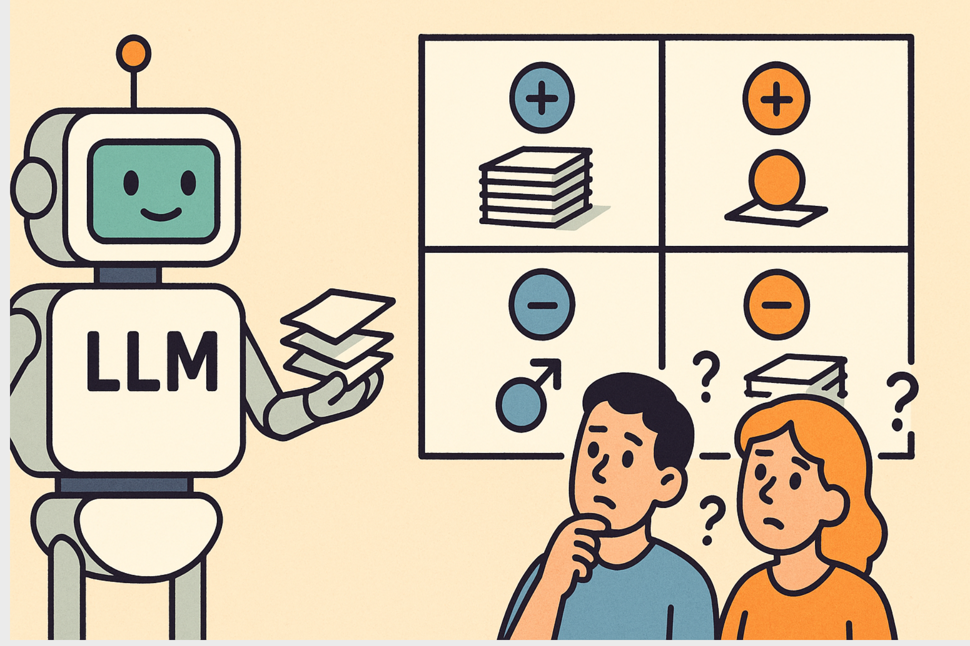
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**



Learning theory: How does W2S happen? What if the data are group imbalanced?

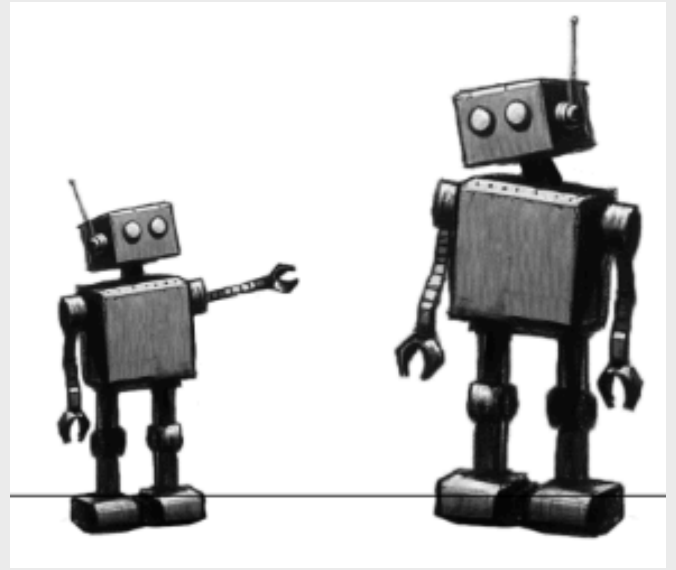
→ **Theory-motivated algorithms** →

Principled algorithm: How can we improve W2S under group imbalance upon failures?

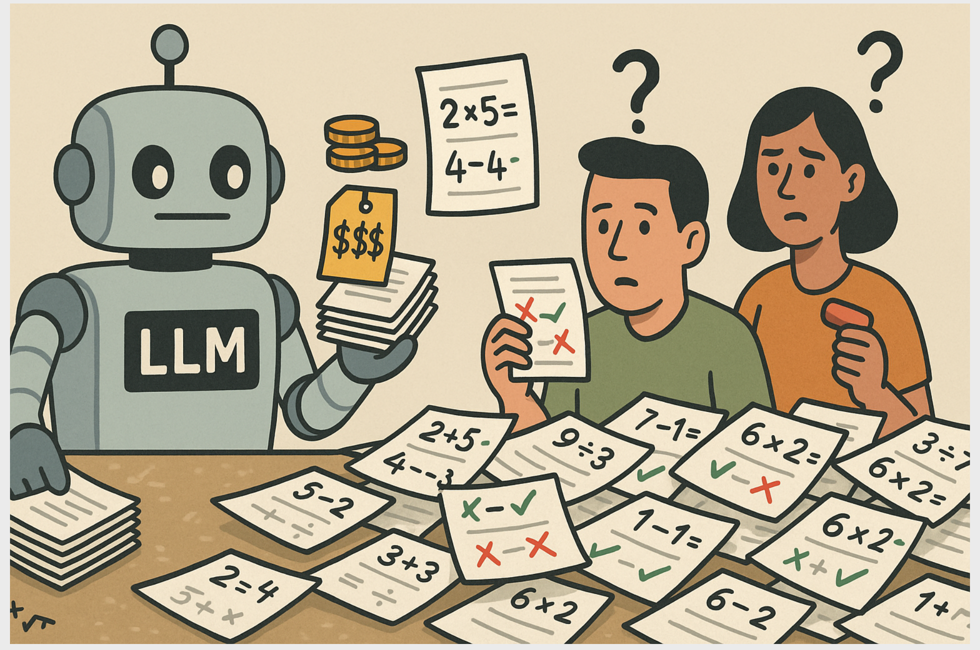
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

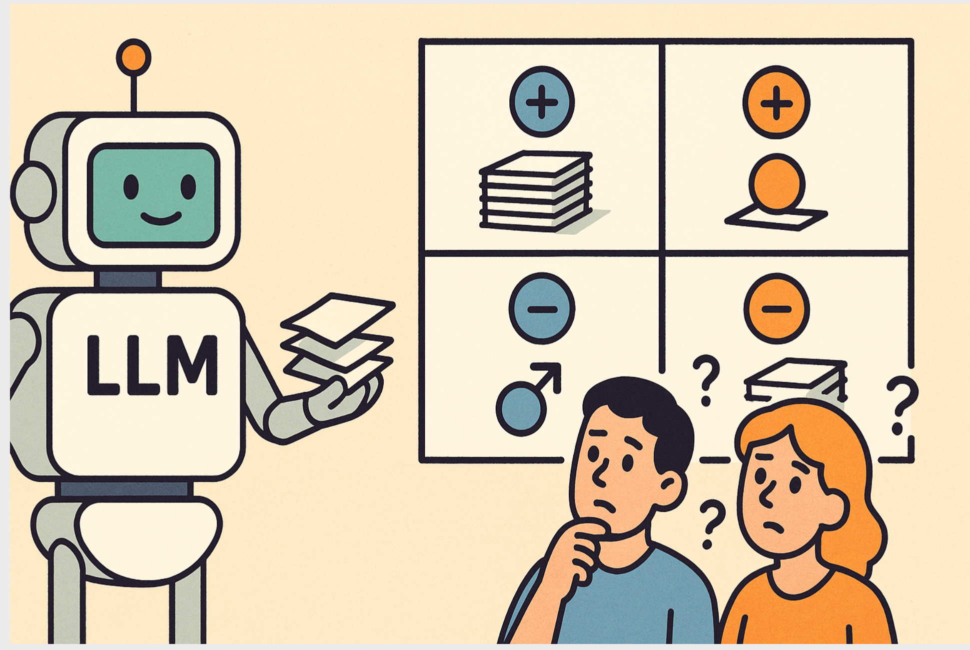
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**



Learning theory: How does W2S happen? What if the data are group imbalanced?

Theory-motivated algorithms

Principled algorithm: How can we improve W2S under group imbalance upon failures?

Randomized Numerical Linear Algebra (RNLA)

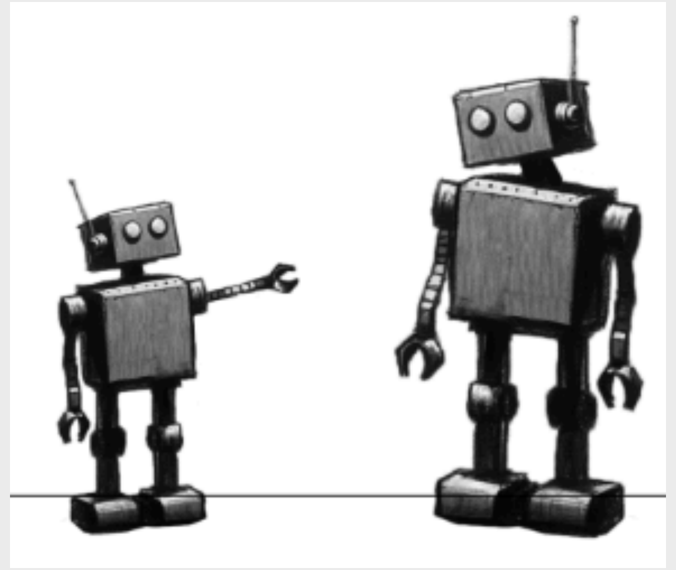
CUR/Interpolative decomposition: [**D**M, ACOM23], [**D**CMP, SIMAX25], [PC**D**M, LAA25]

Randomized SVD: [**D**MN, SIMAX24]

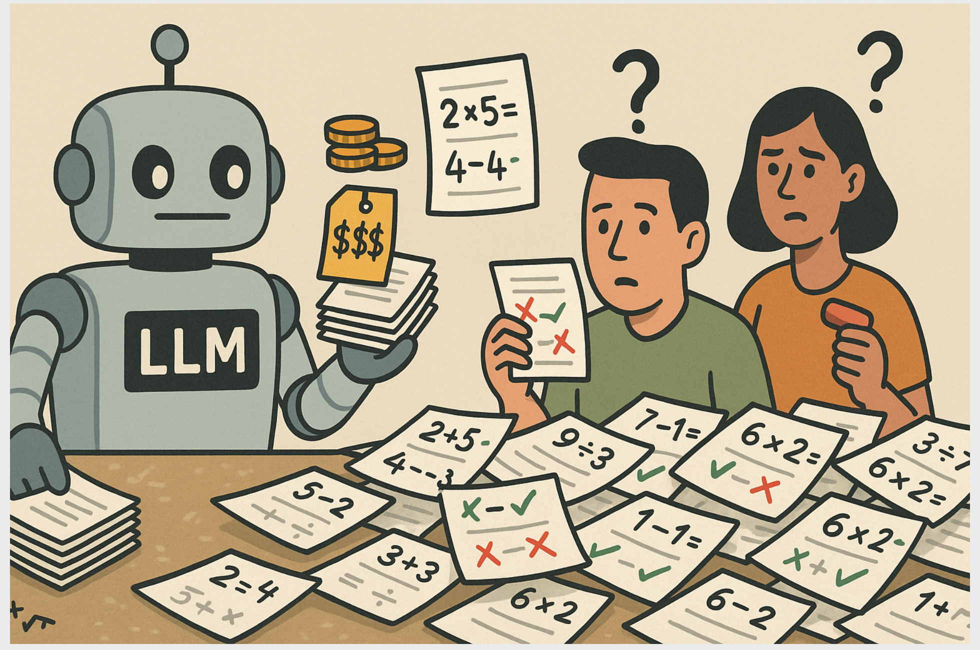
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

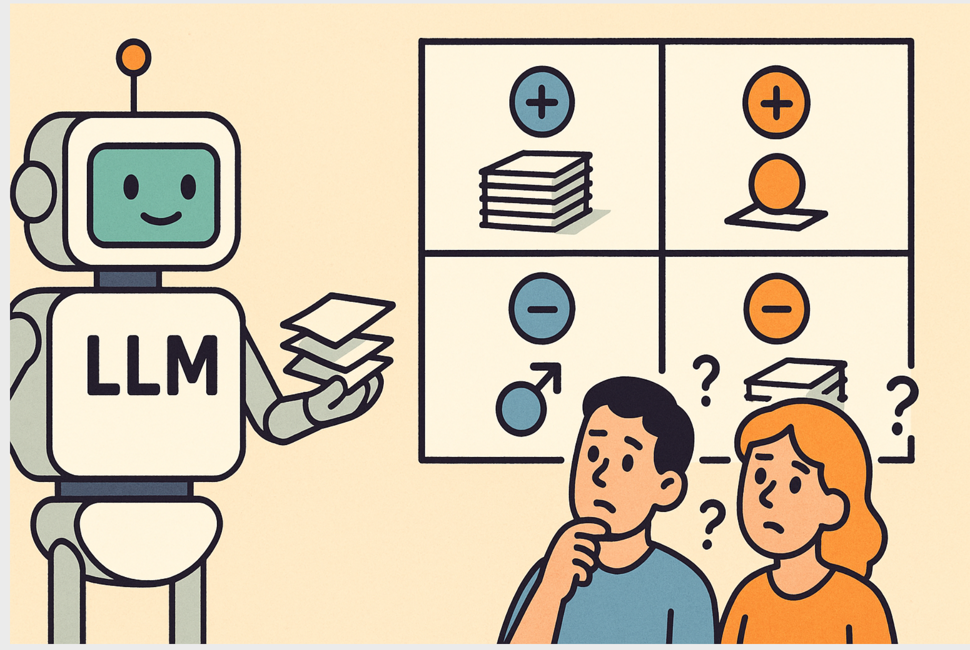
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**



Learning theory: How does W2S happen? What if the data are group imbalanced?

Theory-motivated algorithms →

Principled algorithm: How can we improve W2S under group imbalance upon failures?

Post-training: [L**D**L, ICLR26], [**D**LLLL, ICML25], [**D**PPL, NeurIPS24]

Inference-time: [W**D**L25]

Knowledge distillation: [**D**MLW, NeurIPS23]

Data augmentation: [Y**D**WDSL, AISTATS23]

Linear / kernel regression, …

Randomized Numerical Linear Algebra (RNLA)

High-dimensional probability

Random matrix theory

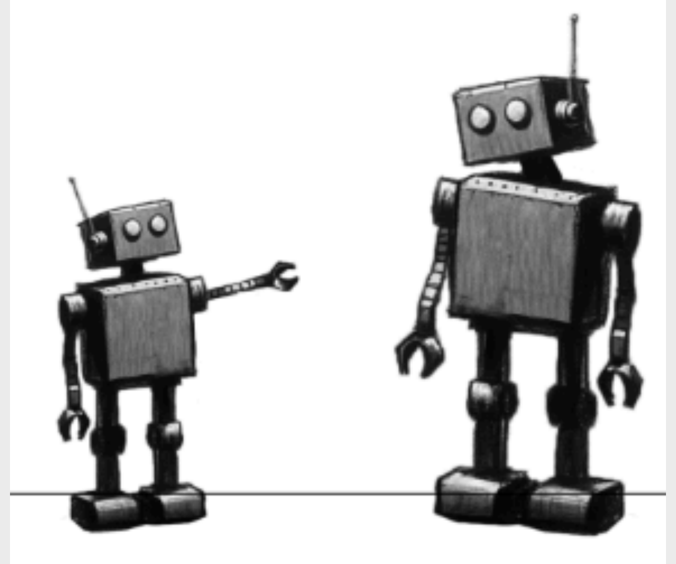
CUR/Interpolative decomposition: [**D**M, ACOM23], [**D**CMP, SIMAX25], [PC**D**M, LAA25]

Randomized SVD: [**D**MN, SIMAX24]

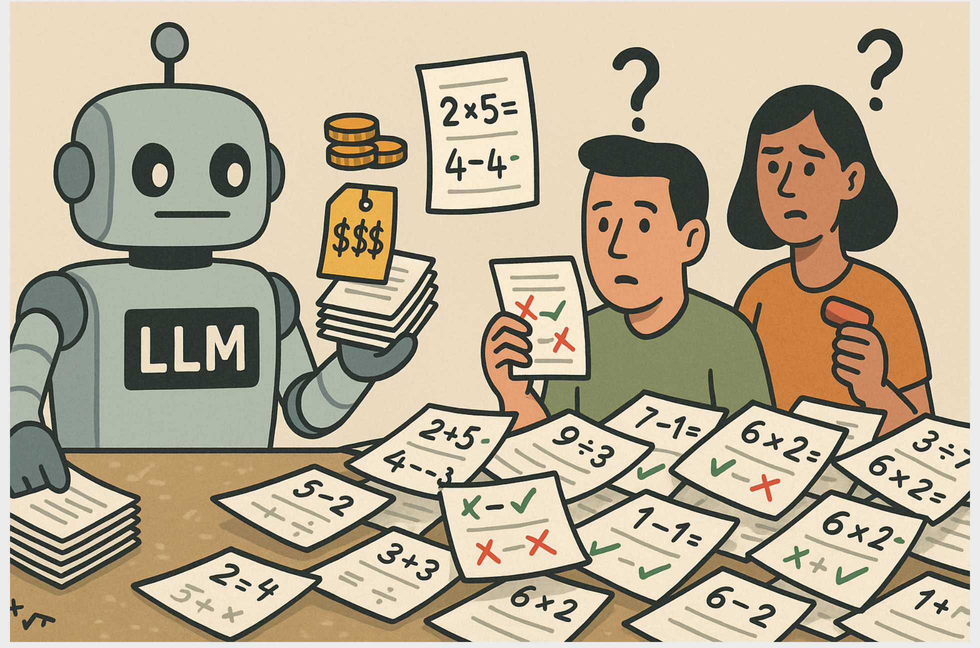
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

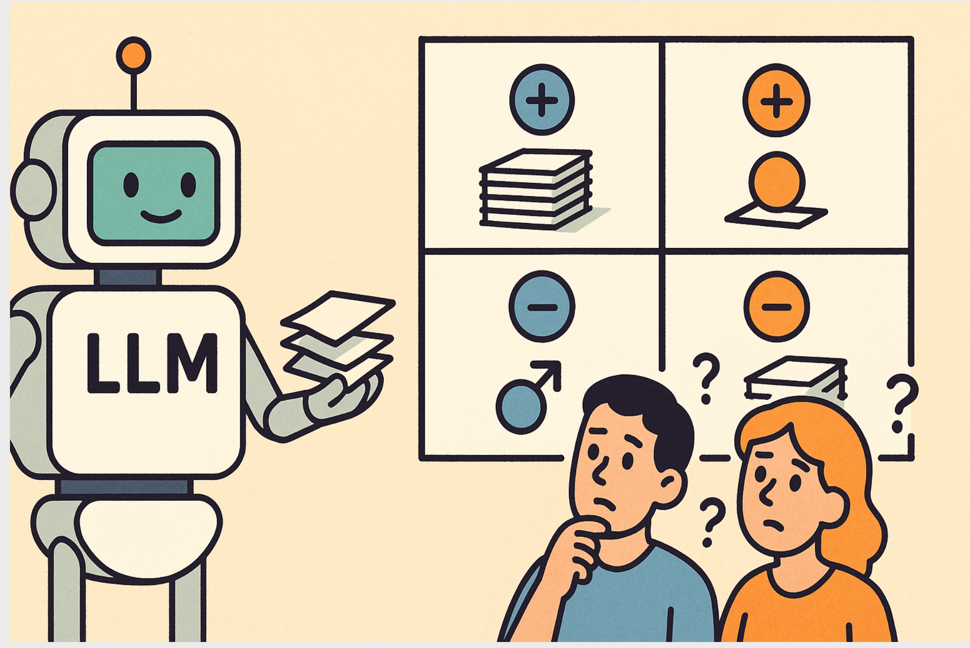
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**



Learning theory: How does W2S happen? What if the data are group imbalanced?

Theory-motivated algorithms

Principled algorithm: How can we improve W2S under group imbalance upon failures?

Post-training: [L**D**L, ICLR26], [**D**LLLL, ICML25], [**D**PPL, NeurIPS24]

Inference-time: [W**D**L25]

Knowledge distillation: [**D**MLW, NeurIPS23]

Data augmentation: [Y**D**WDSL, AISTATS23]

Linear / kernel regression, …

Importance sampling, …

Scientific ML: [**D**SP25]

Pruning: [L**D**L, CPAL25]

Optimization: [**D**XW, ICML23]

Randomized Numerical Linear Algebra (RNLA)

High-dimensional probability Random matrix theory	Randomized algorithms Matrix computations

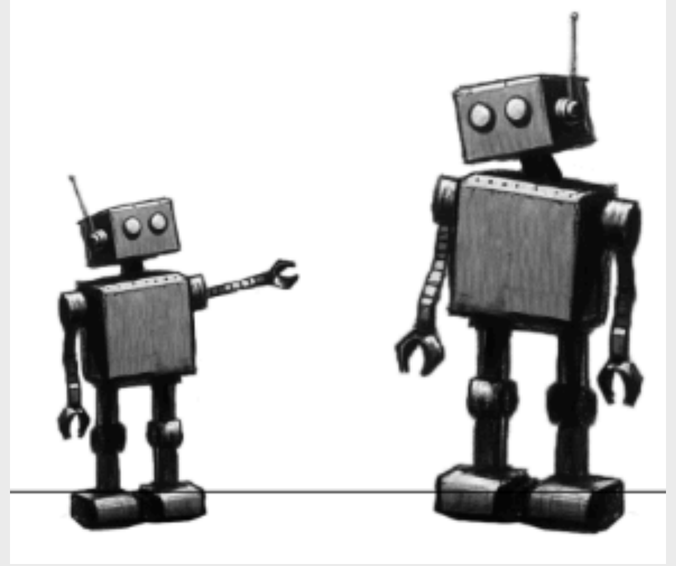
CUR/Interpolative decomposition: [**D**M, ACOM23], [**D**CMP, SIMAX25], [PC**D**M, LAA25]

Randomized SVD: [**D**MN, SIMAX24]

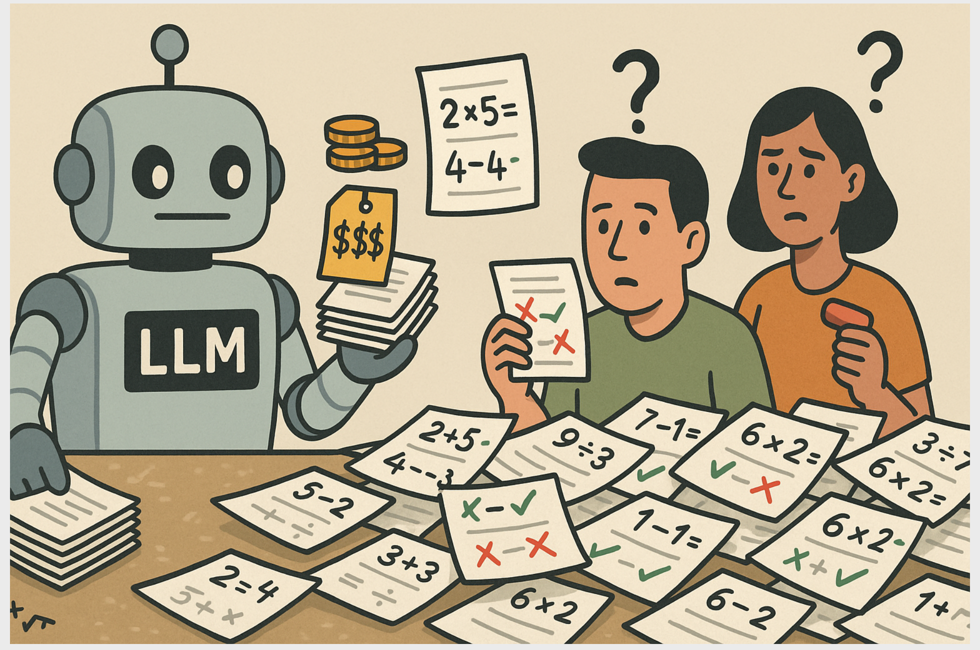
Overview: Understand Post-training through Low Instrinsic Dimension

Post-training on specialized downstream tasks

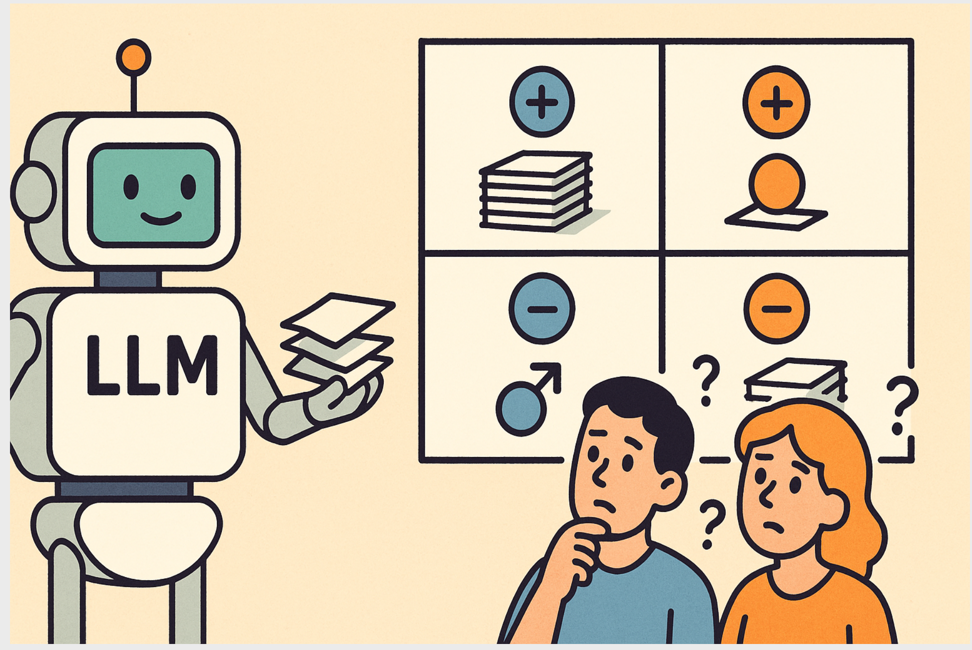
e.g., **W2S**



① **limited & noisy labels**



② **systematic bias**



Learning theory: How does W2S happen? What if the data are group imbalanced?

Theory-motivated algorithms →

Principled algorithm: How can we improve W2S under group imbalance upon failures?

Post-training: [L**D**L, ICLR26], [**D**LLLL, ICML25], [**D**PPL, NeurIPS24]

Inference-time: [W**D**L25]

Knowledge distillation: [**D**MLW, NeurIPS23]

Data augmentation: [Y**D**WDSL, AISTATS23]

Linear / kernel regression, …

Understand & exploit **low intrinsic dimensions** in high-dimensional problems

Importance sampling, …

Scientific ML: [**D**SP25]

Pruning: [L**D**L, CPAL25]

Optimization: [**D**XW, ICML23]

Randomized Numerical Linear Algebra (RNLA)

High-dimensional probability Random matrix theory	Randomized algorithms Matrix computations

CUR/Interpolative decomposition: [**D**M, ACOM23], [**D**CMP, SIMAX25], [PC**D**M, LAA25]

Randomized SVD: [**D**MN, SIMAX24]

Future Directions on Post-training

RNLA

Learning theory

Principled algorithm

Data

Optimization

Formulation

Future Directions on Post-training

RNLA

Learning theory

Principled algorithm

Data

Scalable oversight / synthetic data: learn from synthetic supervision / data

Data selection / distillation: attribute limited compute to high-quality data

Optimization

Formulation

Future Directions on Post-training

RNLA | Learning theory | Principled algorithm

Data

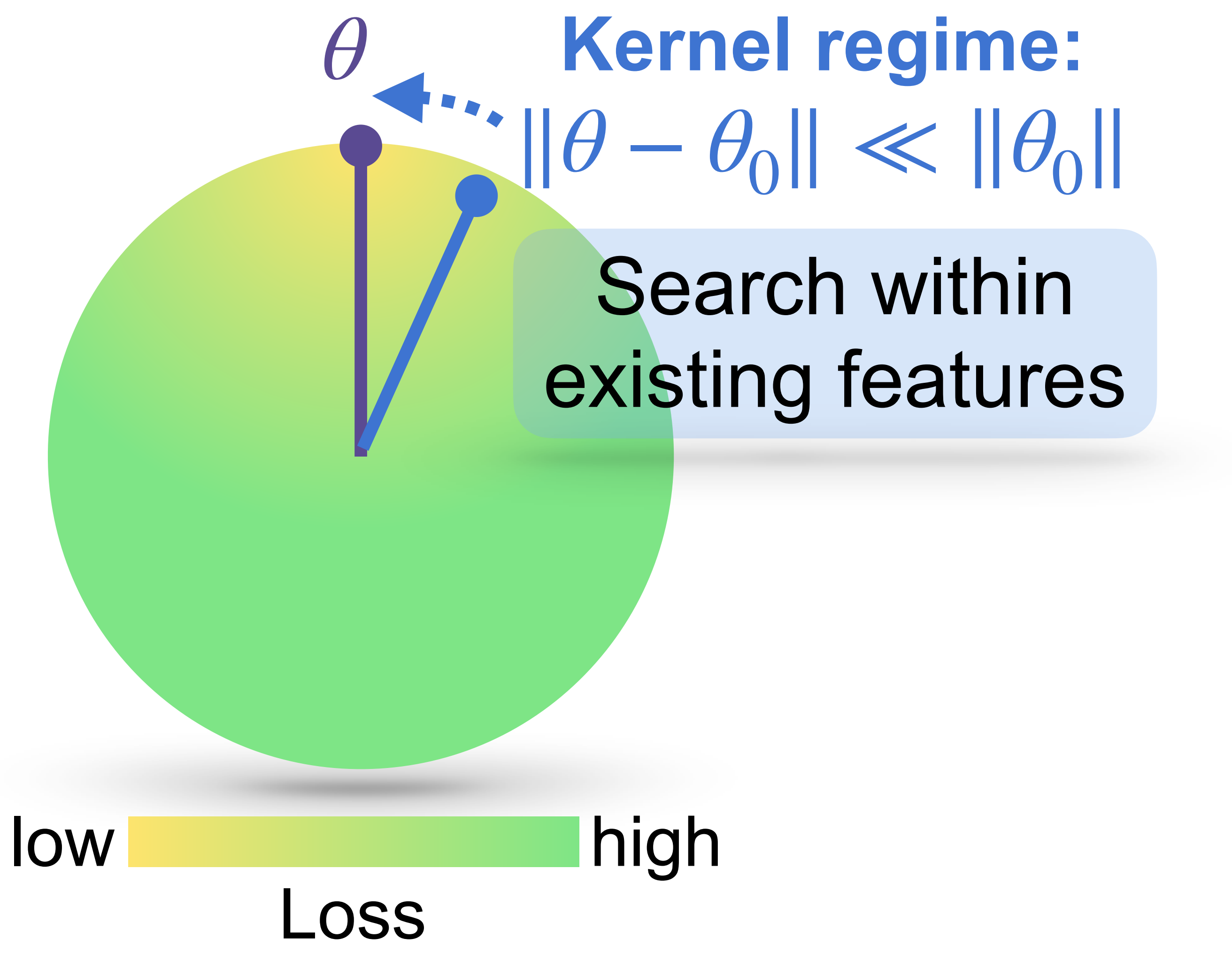
- **Scalable oversight / synthetic data**: learn from synthetic supervision / data
- **Data selection / distillation**: attribute limited compute to high-quality data

Optimization

- ★ **Feature learning with gradient descent**: post-training beyond the kernel regime

Formulation

Future Directions on Post-training



RNLA

Learning theory

Principled algorithm

Data

Scalable oversight / synthetic data: learn from synthetic supervision / data

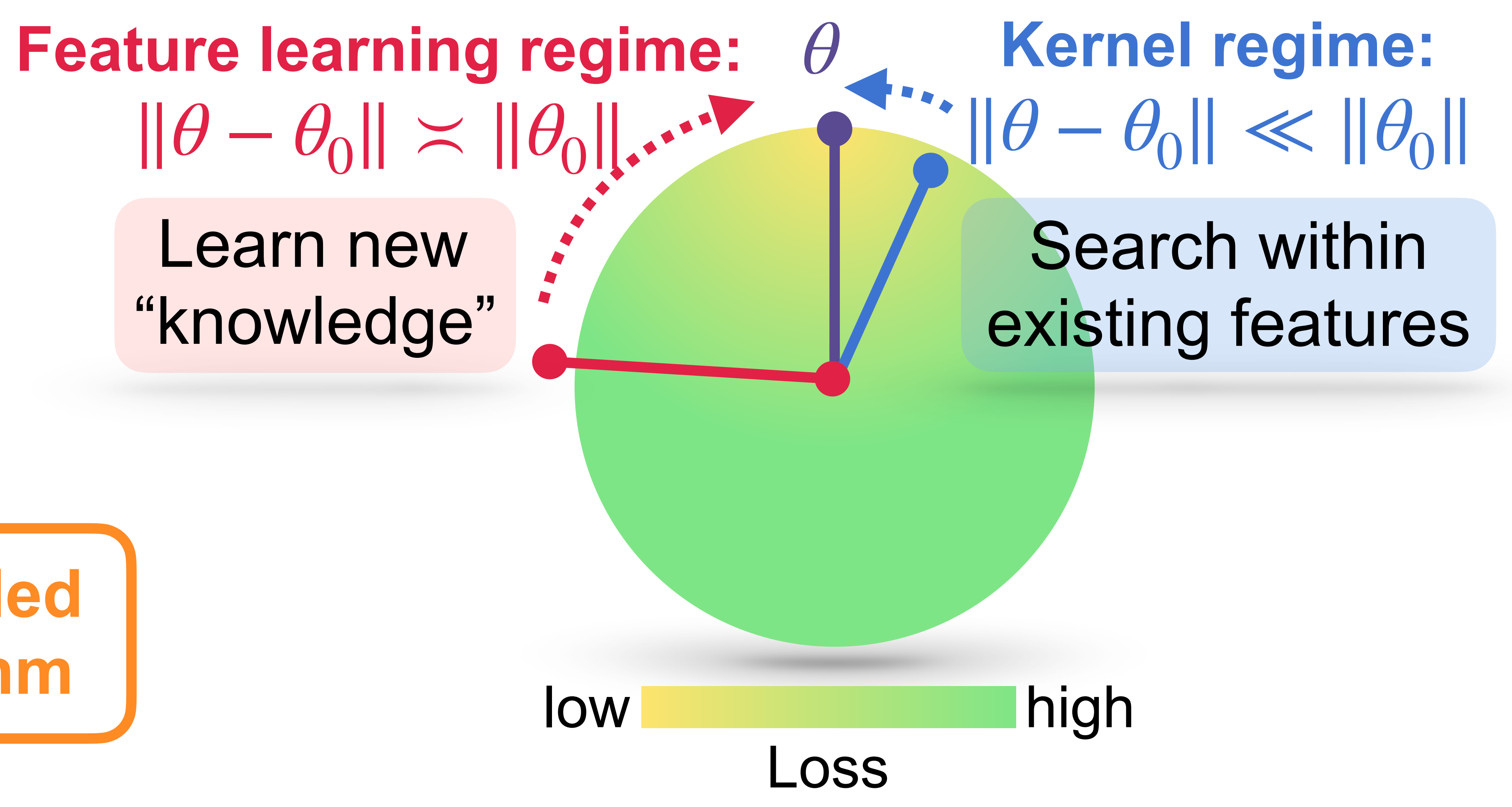
Data selection / distillation: attribute limited compute to high-quality data

Optimization

★ **Feature learning with gradient descent**: post-training beyond the kernel regime

Formulation

Future Directions on Post-training



RNLA | **Learning theory** | **Principled algorithm**

Data

Scalable oversight / synthetic data: learn from synthetic supervision / data

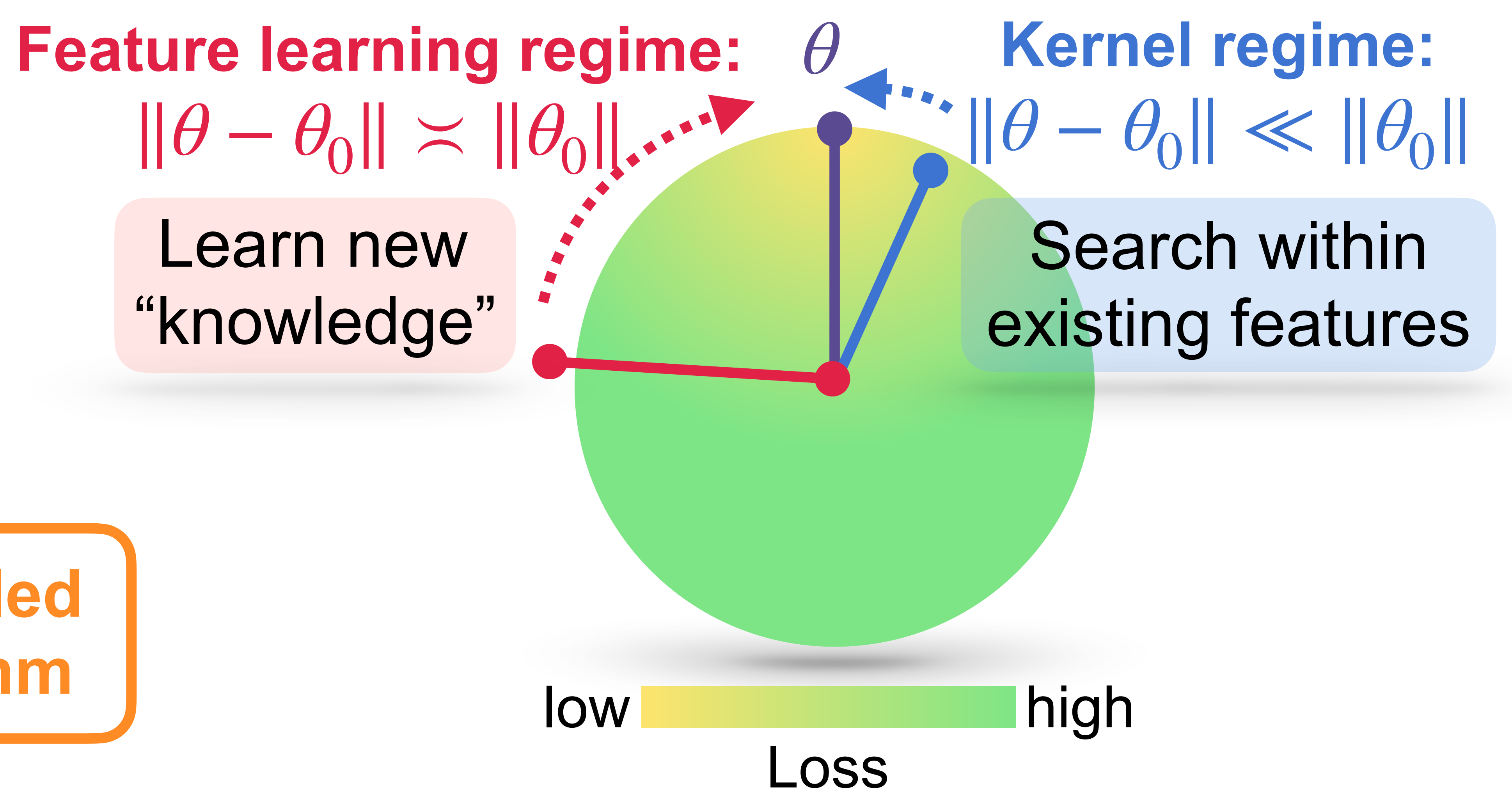
Data selection / distillation: attribute limited compute to high-quality data

Optimization

★ **Feature learning with gradient descent**: post-training beyond the kernel regime

Formulation

Future Directions on Post-training



RNLA | **Learning theory** | **Principled algorithm**

Data

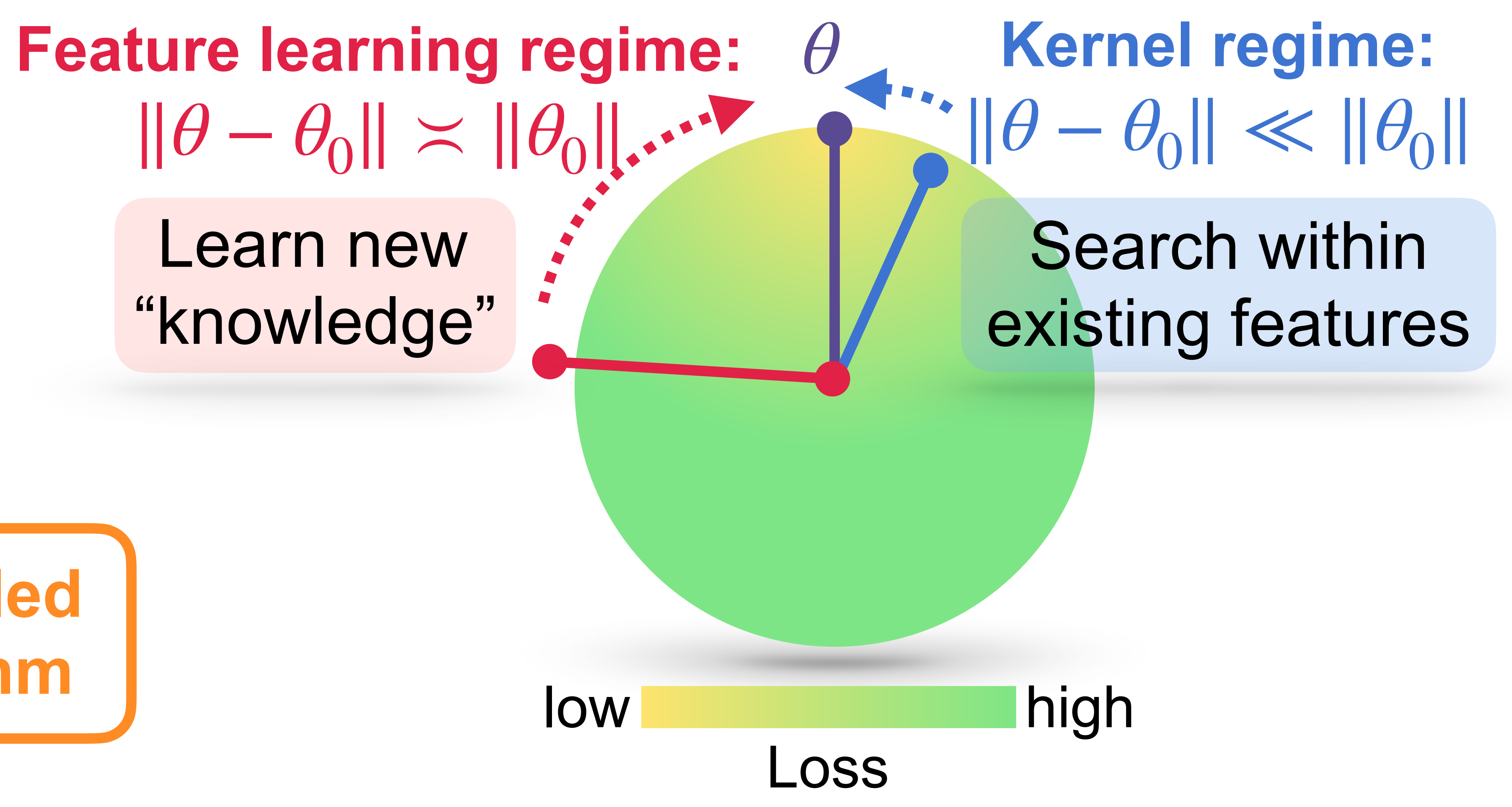
- **Scalable oversight / synthetic data**: learn from synthetic supervision / data
- **Data selection / distillation**: attribute limited compute to high-quality data

Optimization

- ★ **Feature learning with gradient descent**: post-training beyond the kernel regime → **Matrix preconditioning for post-training**: tailored for limited, biased data

Formulation

Future Directions on Post-training



RNLA | Learning theory | Principled algorithm

Data

- **Scalable oversight / synthetic data**: learn from synthetic supervision / data
- **Data selection / distillation**: attribute limited compute to high-quality data

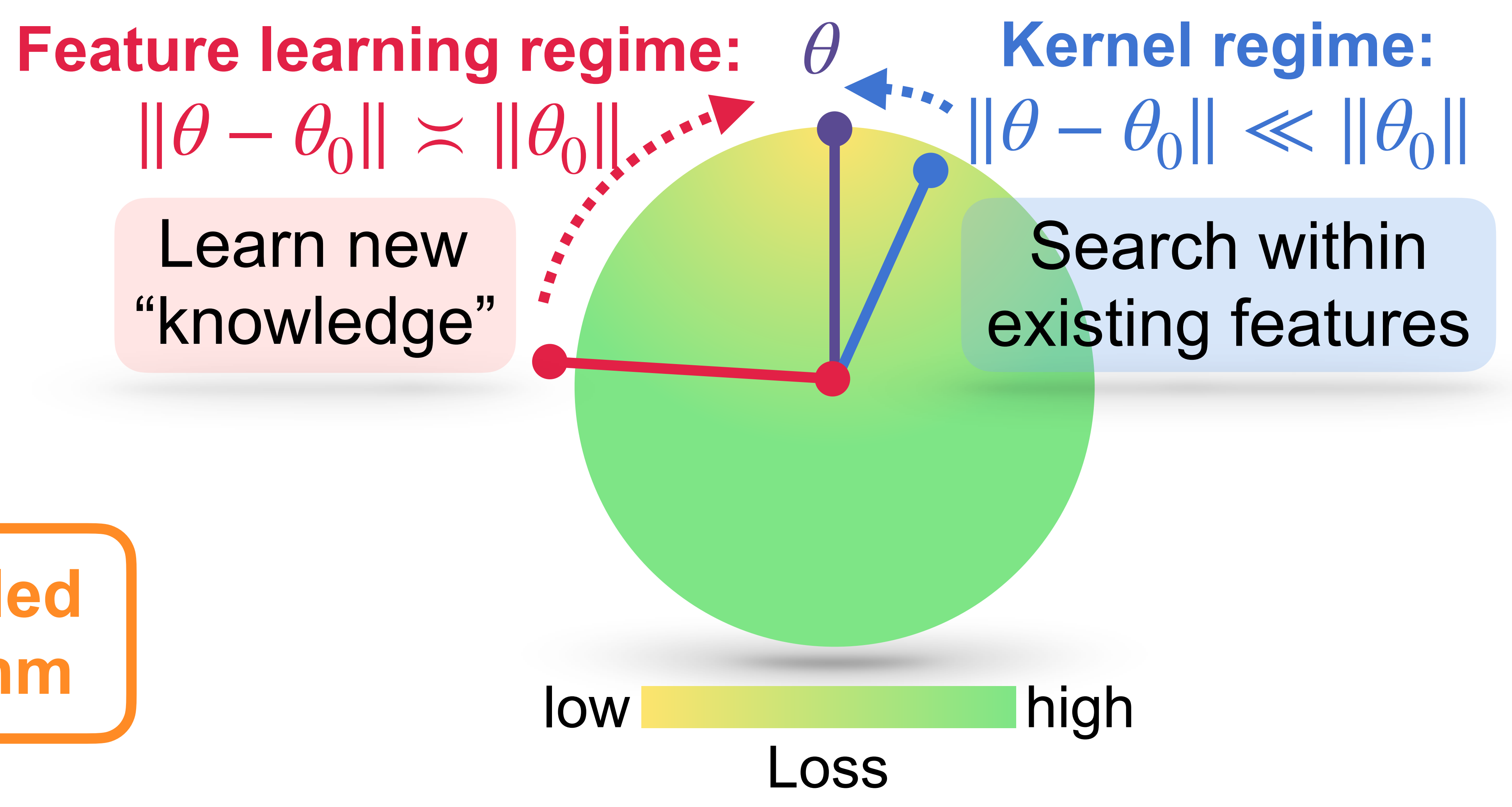
Optimization

- ★ **Feature learning with gradient descent**: post-training beyond the kernel regime
- → **Matrix preconditioning for post-training**: tailored for limited, biased data

Formulation

- **Parameter-efficient fine-tuning**: low-rank adaptation, mixture-of-experts, …

Future Directions on Post-training



	RNLA	Learning theory	Principled algorithm

Data	**Scalable oversight / synthetic data**: learn from synthetic supervision / data	**Data selection / distillation**: attribute limited compute to high-quality data
Optimization	★ **Feature learning with gradient descent**: post-training beyond the kernel regime →	**Matrix preconditioning for post-training**: tailored for limited, biased data
Formulation	↓ **Parameter-efficient fine-tuning**: low-rank adaptation, mixture-of-experts, …	**Scientific ML from a post-training perspective**: "pre-training" from physics

Thank you! Happy to take questions

Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension.
Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, Qi Lei. ICML 2025.

Does Weak-to-strong Generalization Happen under Spurious Correlations?
Chenruo Liu*, Yijun Dong*, Qi Lei. ICLR 2026.

Yicheng Li
UPenn

Yunai Li
Northwestern

Chenruo Liu
NYU

Jason D. Lee
UC Berkeley

Qi Lei
NYU

References

Aghajanyan, Armen, Sonal Gupta, and Luke Zettlemoyer. "Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7319-7328. 2021.

Li, Chunyuan, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. "Measuring the Intrinsic Dimension of Objective Landscapes." In *International Conference on Learning Representations*. 2018.

Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." Advances in neural information processing systems 31 (2018).

Malladi, Sadhika, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. "A kernel-based view of language model fine-tuning." In International Conference on Machine Learning, pp. 23610-23641. PMLR, 2023.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen et al. "Weak-to-strong generalization: eliciting strong capabilities with weak supervision." In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4971-5012. 2024.

Emrullah Ildiz, M., Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. "High-dimensional analysis of knowledge distillation: Weak-to-Strong generalization and scaling laws." In *13th International Conference on Learning Representations*. 2025.

Wu, David Xing, and Anant Sahai. "Provable weak-to-strong generalization via benign overfitting." In *The Thirteenth International Conference on Learning Representations*. 2025.

Medvedev, Marko, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. "Weak-to-strong generalization even in random feature networks, provably." *arXiv preprint arXiv:2503.02877* (2025).

Mulgund, Abhijeet, and Chirag Pabbaraju. "Relating misfit to gain in weak-to-strong generalization beyond the squared loss." *arXiv preprint arXiv:2501.19105* (2025).

Oh, Junsoo, Jerry Song, and Chulhee Yun. "From linear to nonlinear: Provable weak-to-strong generalization through feature learning." *arXiv preprint arXiv:2510.24812* (2025).

Xue, Yihao, Jiping Li, and Baharan Mirzasoleiman. "Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions." *arXiv preprint arXiv:2502.00620* (2025).

Appendix

Weak vs. Strong: How Different Models Encode Group Imbalance

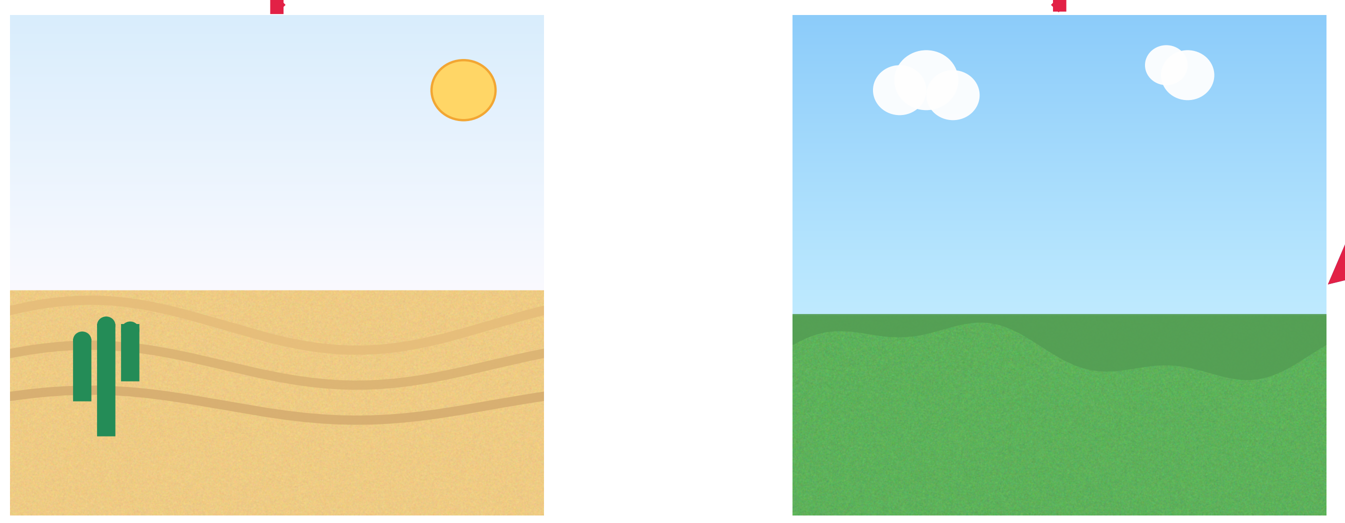
Classify cow vs. camel:
$z(x) \sim \mathcal{N}(0_{d_z}, I_{d_z})$
$y \sim \mathcal{N}(z(x)^\top \beta_*, \sigma^2)$

Foreground in $z(x)$

Minority
$\Pr[g = 1] = \eta$

Majority
$\Pr[g = 0] = 1 - \eta$

Group feature ($p \ll d_z$):
$\xi(x) \mid g \sim \mathcal{N}(g\mu_\xi, \sigma_\xi^2 I_p)$



Background in $z(x) \otimes \xi(x)$

Weak vs. Strong: How efficiently the abstract notion of majority vs. minority is encoded by pre-training

Both weak teacher and strong student have **low approximation error**

$z(x) \in \mathbb{R}^{d_z}$

Weak teacher encodes $\xi(x)$ **less efficiently**:
$W \in \mathrm{Stiefel}(p, p_w - 1)$

$W^\top \xi(x) \in \mathbb{R}^{p_w - 1}$

Strong student encodes $\xi(x)$ **more efficiently**:
$S \in \mathrm{Stiefel}(p, p_s - 1)$

$S^\top \xi(x) \in \mathbb{R}^{p_s - 1}$

Teacher-student similarity: $W^\top S$

$p_s \leq p_w \leq p \ll d_z$

d_w, d_s = intrinsic dimensions

D_w, D_s = model sizes $\gg d_w, d_s$

Weak teacher $\varphi_w(x) = U_w \phi_w(x)$

$U_w \in \mathbb{R}^{D_w \times d_w}$

$d_w = p_w d_z$

$z(x)$

$z(x) \otimes W^\top \xi(x)$

$\phi_w \in \mathbb{R}^{d_w}$

Strong student $\varphi_s(x) = U_s \phi_s(x)$

$U_s \in \mathbb{R}^{D_s \times d_s}$

$d_s = p_s d_z$

$z(x)$

$z(x) \otimes S^\top \xi(x)$

$\phi_s \in \mathbb{R}^{d_s}$

Weak-to-Strong Generalization under Group Imbalance

Labeled set: $(\widetilde{X}, \widetilde{y}) \sim \mathscr{D}(\eta_\ell)^n$

Weak teacher $f_w(x) = \varphi_w(x)^\top \theta_w$: $\theta_w = \operatorname{argmin}_{\theta \in \mathbb{R}^{D_w}} \frac{1}{n} \|\varphi_w(\widetilde{X})\theta - \widetilde{y}\|_2^2 + \alpha_w \|\theta\|_2^2$

Unlabeled set: $X \sim \mathscr{D}_x(\eta_u)^N$

W2S $f_s(x) = \varphi_s(x)^\top \theta_s$: $\theta_s = \operatorname{argmin}_{\theta \in \mathbb{R}^{D_s}} \frac{1}{N} \|\varphi_s(X)\theta - f_w(X)\|_2^2 + \alpha_{w2s} \|\theta\|_2^2$

Test distribution: $\mathscr{D}(\eta_t)$

- Average: $\eta_t = 1/2$
- Worst group: $\eta_t = 1$
- Best group: $\eta_t = 0$

Proportional asymptotic limit: $d_z, n, N \to \infty$, $\frac{d_z}{n} \to \gamma_z \in (0, p_T^{-1})$, $\frac{d_z}{N} \to \nu_z \in (0, p_S^{-1})$; $p_s \le p_w < \infty$.

Theorem [LDL25]. As $\alpha_w, \alpha_{w2s} \to 0$:

From group imbalance

$$\mathbb{E}[\mathrm{ER}_{\eta_t}(f_w) \mid \eta_\ell] \xrightarrow{\mathbb{P}} \sigma^2 \gamma_z \Big(p_w + \sigma_\xi^{-2} \|(\eta_t - \eta_\ell) W^\top \mu_\xi\|_2^2 \Big)$$

$p_{s \wedge w} + \Theta(\nu_z) \le p_w$ $\qquad$ $\mathcal{S}(\eta_\ell)$ **from** $\eta_\ell < 0.5$

$$\mathbb{E}[\mathrm{ER}_{\eta_t}(f_s) \mid \eta_\ell, \eta_u] \xrightarrow{\mathbb{P}} \sigma^2 \gamma_z \Big(p_{s \wedge w} + \nu_z p_s (p_w - p_{s \wedge w}) + \sigma_\xi^{-2} \|(\eta_u - \eta_\ell) W^\top \mu_\xi + (\eta_t - \eta_u) W^\top S S^\top \mu_\xi\|_2^2 + \Theta(\nu_z (\eta_t - \eta_u)^2) \Big)$$

$p_{s \wedge w} := 1 + \|W^\top S\|_F^2 \le p_s$ $\qquad$ $\mathcal{S}(\eta_\ell \to \eta_u)$ **from** $\eta_\ell, \eta_u < 0.5$ $\qquad$ **negligible:** $\nu_z \ll 1$

① $\mathcal{S}(\eta_\ell \to \eta_u) \le \mathcal{S}(\eta_\ell)$ **if** $\eta_u = \eta_\ell$

★ ② **If** $T^\top S = 0$, $\mathcal{S}(\eta_\ell \to \eta_u) \propto (\eta_u - \eta_\ell)^2$