北京大學
PEKING UNIVERSITY

LEARNING DYNAMICS OF LLM FINETUNING

Yi Ren and Danica J. Sutherland

ICLR 2025 Outstanding Paper Awards

PRESENTER: LILANG LIN

2025/04/27

Outline

Outline

Background

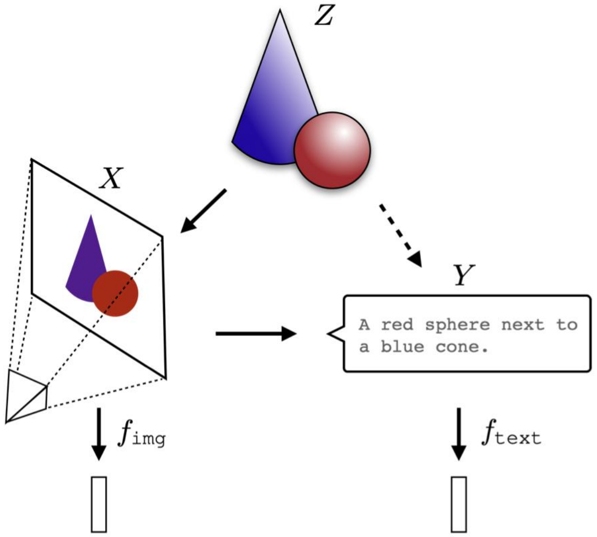
The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

The Platonic Representation Hypothesis (ICML 2024)

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



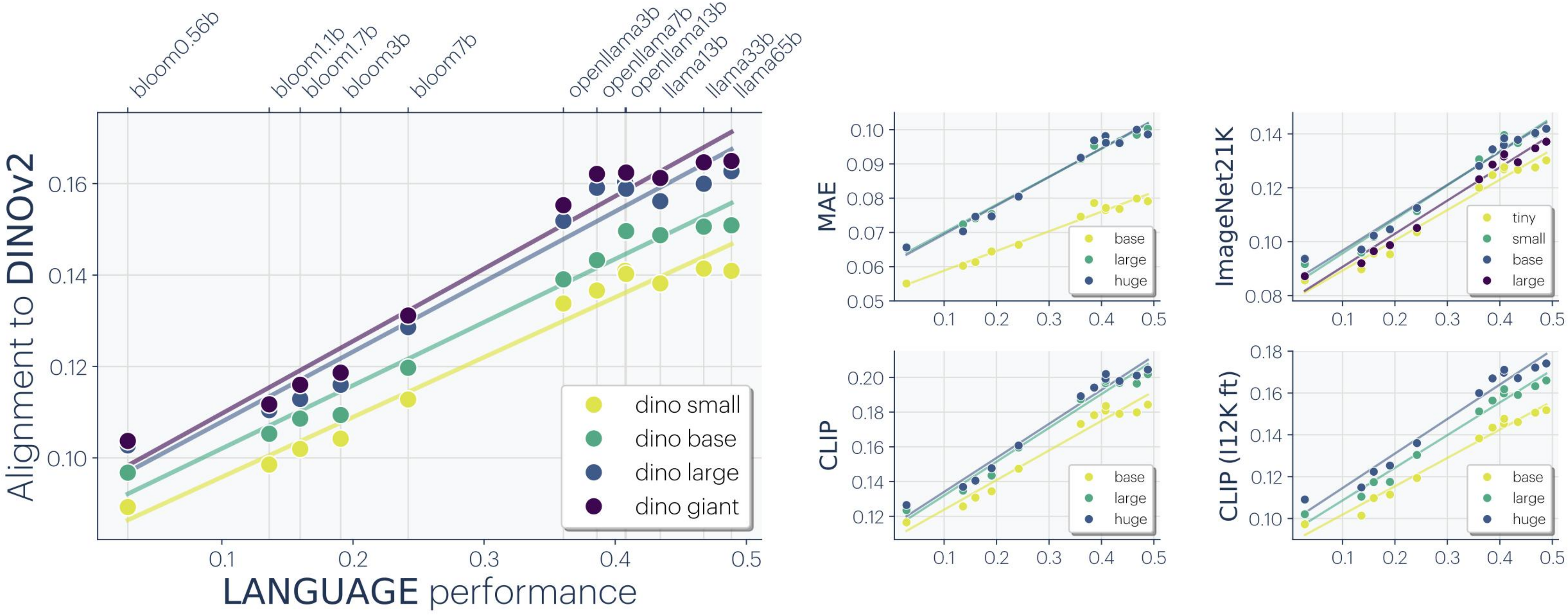
$$K_{\texttt{img}}(i,j) = \langle f_{\texttt{img}}(x_i), f_{\texttt{img}}(x_j) \rangle$$
$$K_{\texttt{text}}(i,j) = \langle f_{\texttt{text}}(y_i), f_{\texttt{text}}(y_j) \rangle.$$

Background

The Platonic Representation Hypothesis

Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

■ The Platonic Representation Hypothesis (ICML 2024)

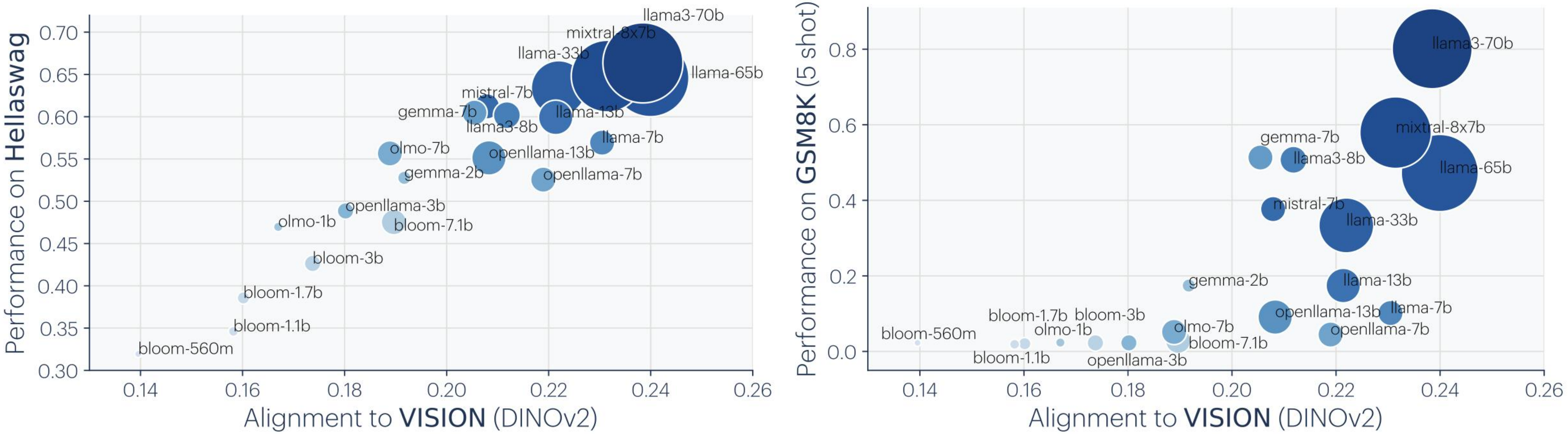


Background

The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

■ The Platonic Representation Hypothesis (ICML 2024)



Background

The Platonic Representation Hypothesis

Minyoung Huh[*1] **Brian Cheung**[*1] **Tongzhou Wang**[*1] **Phillip Isola**[*1]

- **The Platonic Representation Hypothesis (ICML 2024)**

$$\overbrace{f^*}^{\text{trained model}} = \underset{f \in \underbrace{\mathcal{F}}_{\text{function class}}}{\arg\min} \mathbb{E}_{x \sim \text{dataset}} [\overbrace{\mathcal{L}}^{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

Background

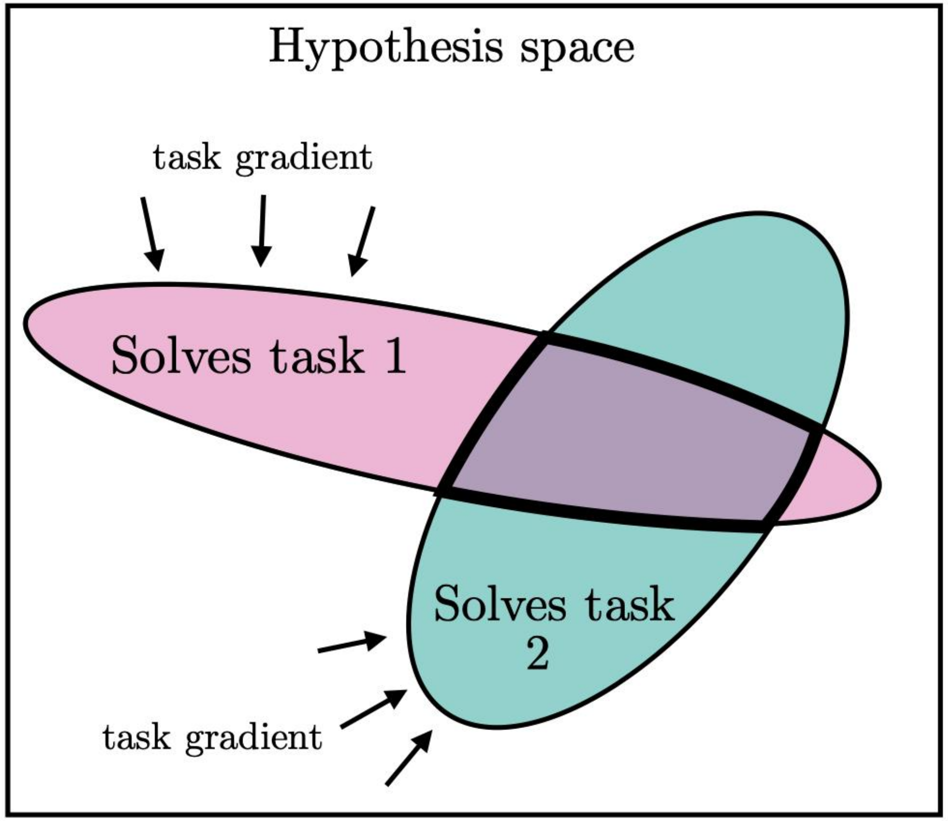
The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

■ The Platonic Representation Hypothesis (ICML 2024)

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.



Background

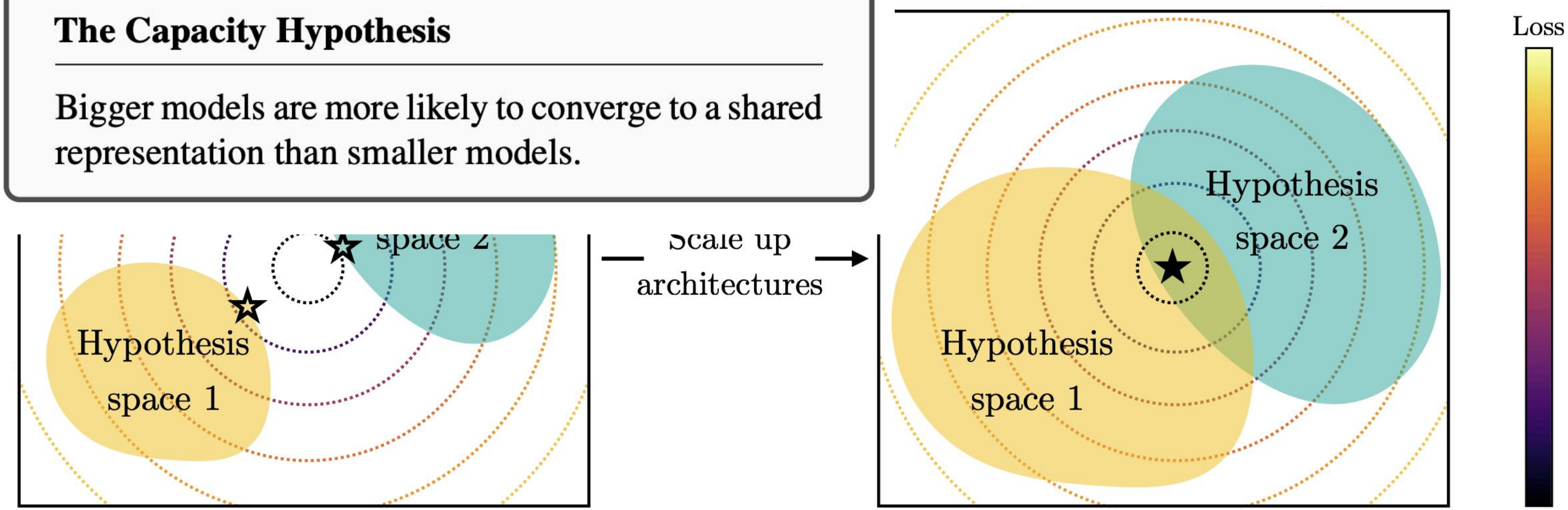
The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

The Platonic Representation Hypothesis (ICML 2024)

The Capacity Hypothesis

Bigger models are more likely to converge to a shared representation than smaller models.

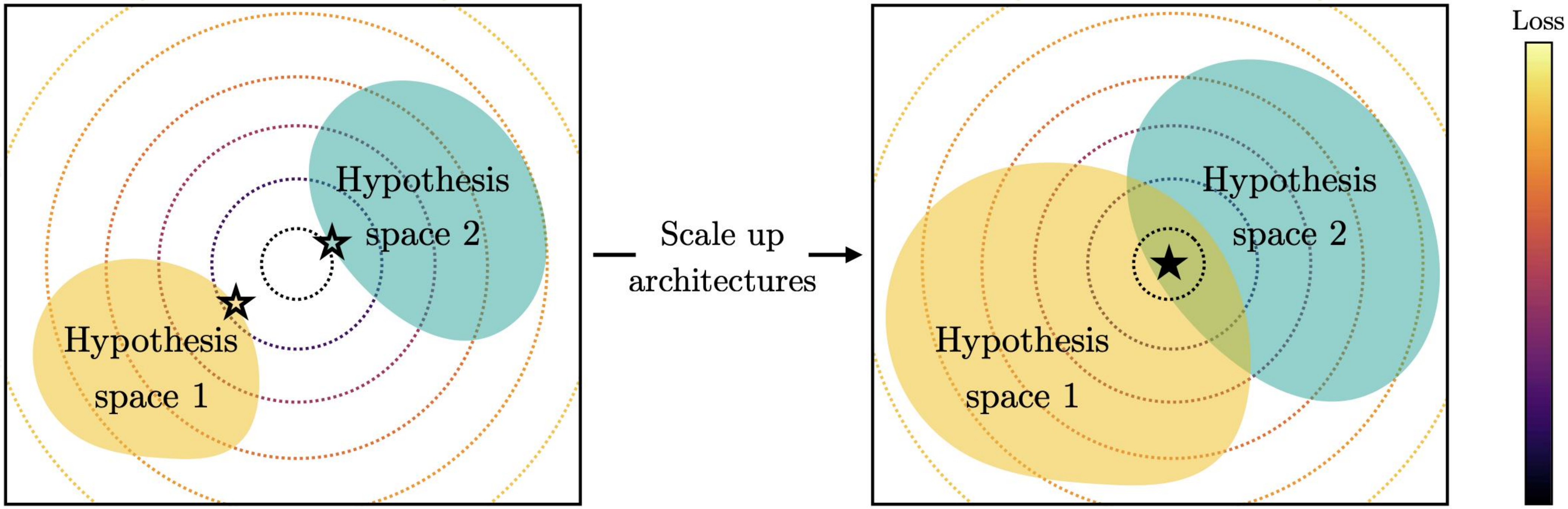


Background

The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

- **The Platonic Representation Hypothesis (ICML 2024)**



Background

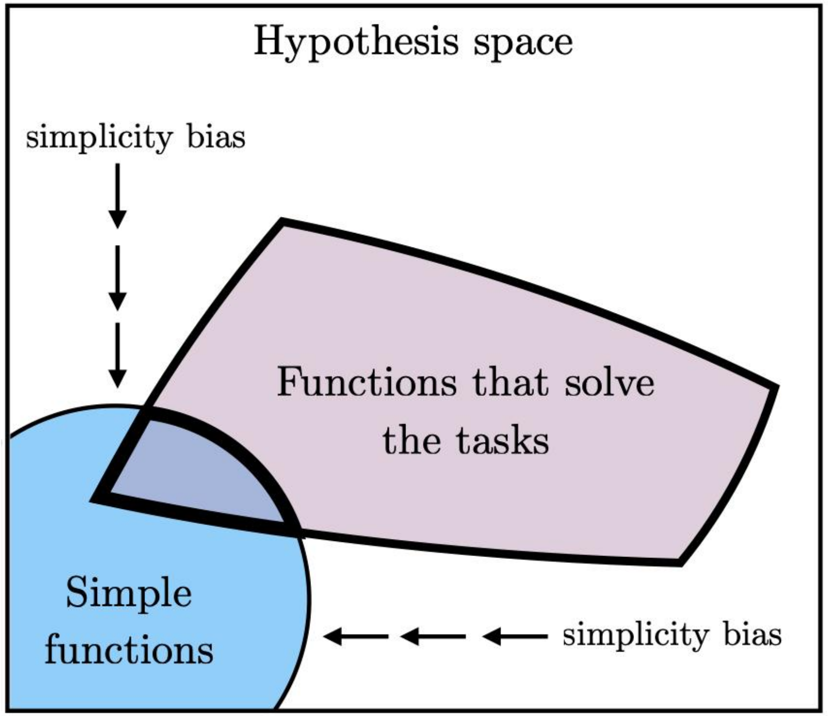
The Platonic Representation Hypothesis

Minyoung Huh [* 1] **Brian Cheung** [* 1] **Tongzhou Wang** [* 1] **Phillip Isola** [* 1]

The Platonic Representation Hypothesis (ICML 2024)

The Simplicity Bias Hypothesis

Deep networks are biased toward finding simple fits to the data, and the bigger the model, the stronger the bias. Therefore, as models get bigger, we should expect convergence to a smaller solution space.



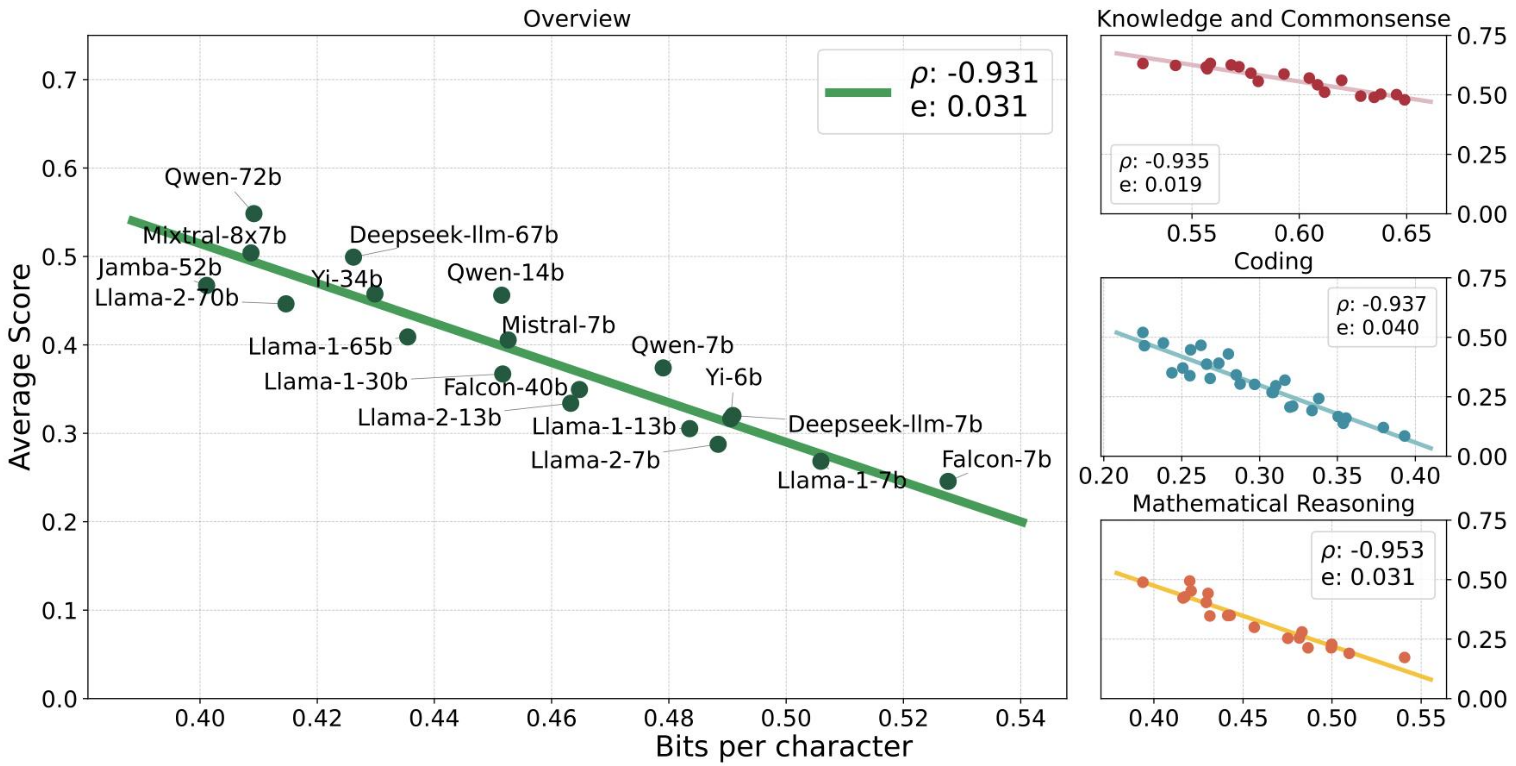
Background

Compression Represents Intelligence Linearly

Yuzhen Huang[*1] **Jinghan Zhang**[*1] **Zifei Shan**[2] **Junxian He**[1]
[1]The Hong Kong University of Science and Technology [2]Tencent
{yhuanghj,jzhangjv,junxianh}@cse.ust.hk

■ Compression Represents Intelligence Linearly (COLM 2024)

$$\text{Optimal \# Bits on Average} = \mathbb{E}_{x\sim p_{data}}\left[\sum_{i=1}^{n} -\log_2 p_{model}(x_i|x_{1:i-1})\right],$$



Outline

Learning Dynamic

After an GD update on $\mathbf{x}_u$*, how does the model's prediction on* $\mathbf{x}_o$ *change?*

$$\Delta\theta \triangleq \theta^{t+1} - \theta^t = -\eta \cdot \nabla\mathcal{L}\left(f_\theta(\mathbf{x}_u), \mathbf{y}_u\right); \quad \Delta f(\mathbf{x}_o) \triangleq f_{\theta^{t+1}}(\mathbf{x}_o) - f_{\theta^t}(\mathbf{x}_o),$$

$$\Delta \log \pi^t(\mathbf{y} \mid \mathbf{x}_o) \triangleq \log \pi_{\theta^{t+1}}(\mathbf{y} \mid \mathbf{x}_o) - \log \pi_{\theta^t}(\mathbf{y} \mid \mathbf{x}_o), .$$

Learning Dynamic

After an GD update on $\mathbf{x}_u$, how does the model's prediction on $\mathbf{x}_o$ change?

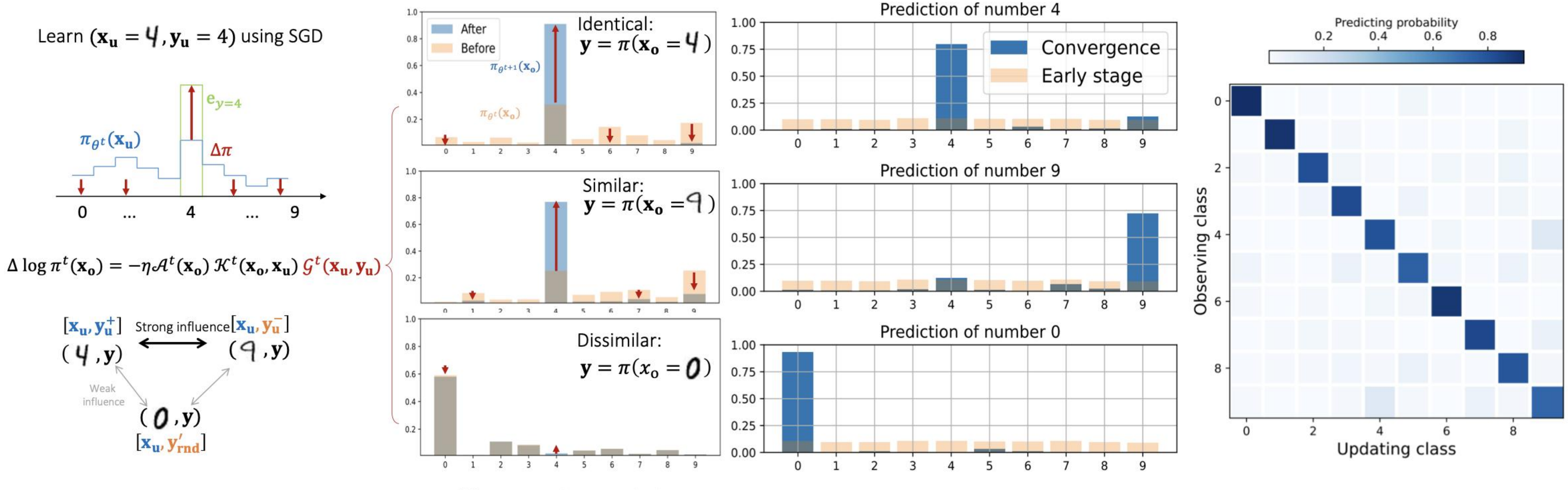
Proposition 1. *Let* $\pi = \mathsf{Softmax}(\mathbf{z})$ *and* $\mathbf{z} = h_\theta(\mathbf{x})$. *The one-step learning dynamics decompose as*

$$\underbrace{\Delta \log \pi^t(\mathbf{y} \mid \mathbf{x}_o)}_{V\times 1} = -\eta \underbrace{\mathcal{A}^t(\mathbf{x}_o)}_{V\times V} \underbrace{\mathcal{K}^t(\mathbf{x}_o, \mathbf{x}_u)}_{V\times V} \underbrace{\mathcal{G}^t(\mathbf{x}_u, \mathbf{y}_u)}_{V\times 1} + \mathcal{O}(\eta^2 \|\nabla_\theta \mathbf{z}(\mathbf{x}_u)\|_{\mathrm{op}}^2), \quad (3)$$

where $\mathcal{A}^t(\mathbf{x}_o) = \nabla_{\mathbf{z}} \log \pi_{\theta^t}(\mathbf{x}_o) = I - \mathbf{1}\pi_{\theta^t}^\top(\mathbf{x}_o)$, $\mathcal{K}^t(\mathbf{x}_o, \mathbf{x}_u) = (\nabla_\theta \mathbf{z}(\mathbf{x}_o)|_{\theta^t})(\nabla_\theta \mathbf{z}(\mathbf{x}_u)|_{\theta^t})^\top$ *is the empirical neural tangent kernel of the logit network* $\mathbf{z}$, *and* $\mathcal{G}^t(\mathbf{x}_u, \mathbf{y}_u) = \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{x}_u, \mathbf{y}_u)|_{\mathbf{z}^t}$.

Learning Dynamic

After an GD update on $\mathbf{x}_u$, how does the model's prediction on $\mathbf{x}_o$ change?



(a) Adaptation vector created by $(\mathbf{x_u}, \mathbf{y_u})$

(b) One-step change with the same $\mathcal{G}^t(\mathbf{x_u}, \mathbf{y_u})$ (large η)

(c) Accumulated change of epochs

(d) Correlation of the accumulated change

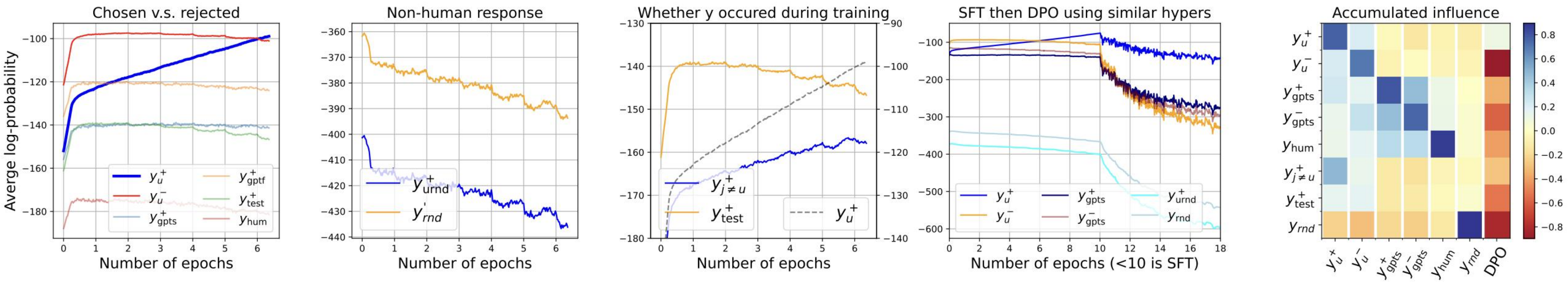
Learning Dynamic

$$\mathcal{L}_{\text{SFT}}(\mathbf{x}_u, \mathbf{y}_u^+) \triangleq -\sum_{l=1}^{L} \log \pi(y = y_l^+ \mid \mathbf{y}_{<l}^+, \mathbf{x}_u) = -\sum_{l=1}^{L} \mathbf{e}_{y_l^+} \cdot \log \pi(\mathbf{y} \mid \mathbf{x}_u, \mathbf{y}_{<l}^+).$$

$$[\underbrace{\Delta \log \pi^t(\mathbf{y} \mid \boldsymbol{\chi}_o)}_{V \times M}]_m = -\sum_{l=1}^{L} \eta [\underbrace{\mathcal{A}^t(\boldsymbol{\chi}_o)}_{V \times V \times M}]_m [\underbrace{\mathcal{K}^t(\boldsymbol{\chi}_o, \boldsymbol{\chi}_u)}_{V \times V \times L}]_l [\underbrace{\mathcal{G}^t(\boldsymbol{\chi}_u)}_{V \times L}]_l + \mathcal{O}(\eta^2),$$

Experiments

LEARNING DYNAMICS OF SFT



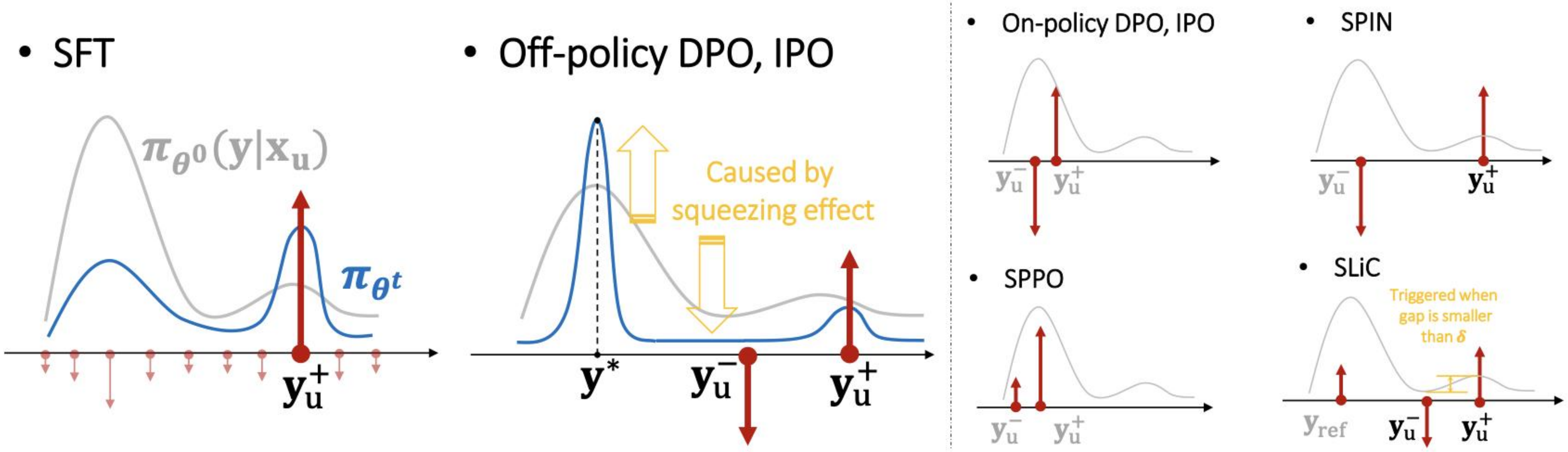
Learning Dynamic

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{x}_u, \mathbf{y}_u^+, \mathbf{y}_u^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta^t}(\mathbf{y}_u^+ \mid \boldsymbol{\chi}_u^+)}{\pi_{\text{ref}}(\mathbf{y}_u^+ \mid \boldsymbol{\chi}_u^+)} - \beta \log \frac{\pi_{\theta^t}(\mathbf{y}_u^- \mid \boldsymbol{\chi}_u^-)}{\pi_{\text{ref}}(\mathbf{y}_u^- \mid \boldsymbol{\chi}_u^-)} \right) \right],$$

$$[\Delta \log \pi^t(\mathbf{y} \mid \boldsymbol{\chi}_o)]_m = -\sum_{l=1}^{L} \eta [\mathcal{A}^t(\boldsymbol{\chi}_o)]_m \left([\mathcal{K}^t(\boldsymbol{\chi}_o, \boldsymbol{\chi}_u^+)]_l [\mathcal{G}^t_{\text{DPO+}}]_l - [\mathcal{K}^t(\boldsymbol{\chi}_o, \boldsymbol{\chi}_u^-)]_l [\mathcal{G}^t_{\text{DPO-}}]_l \right) + \mathcal{O}(\eta^2)$$

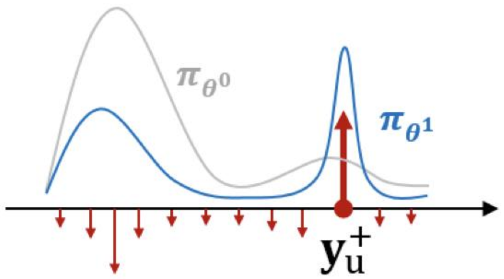
$$\mathcal{G}^t_{\text{DPO+}} = \beta(1-a)\left(\pi_{\theta^t}(\mathbf{y} \mid \boldsymbol{\chi}_u^+) - \mathbf{y}_u^+\right); \qquad \mathcal{G}^t_{\text{DPO-}} = \beta(1-a)\left(\pi_{\theta^t}(\mathbf{y} \mid \boldsymbol{\chi}_u^-) - \mathbf{y}_u^-\right),$$

Learning Dynamic

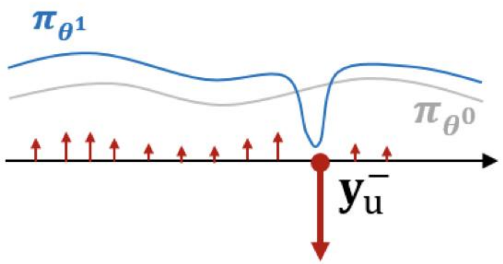


Learning Dynamic

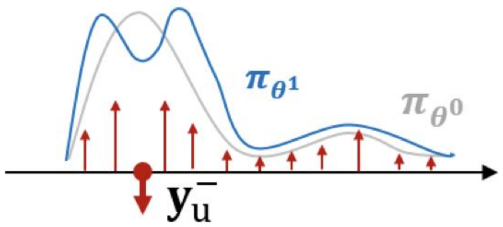
- A: Positive gradient $\left(\pi_{\theta^0} - \mathbf{e}_{\hat{\mathbf{y}}}\right)$



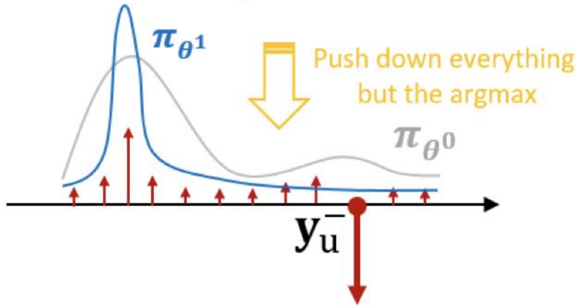
- B: Big negative gradient for flat π_{θ^0}



- C: Big negative gradient on the “peak” of a non-flat π_{θ^0} (on-policy DPO)

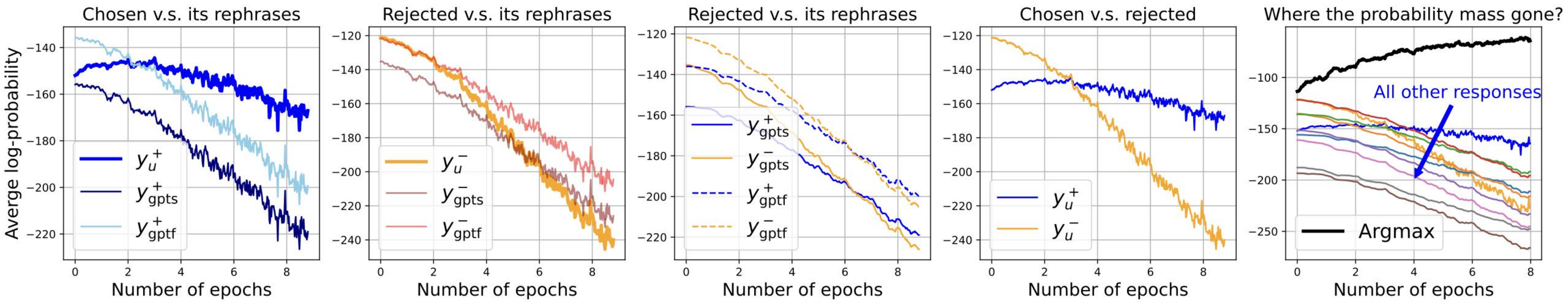


- D: Big negative gradient on the “valley” of a non-flat π_{θ^0}



Experiments

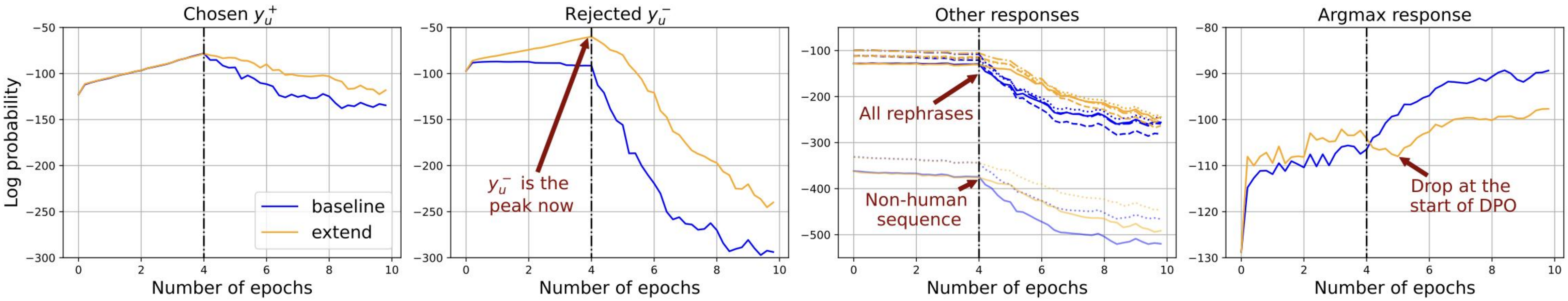
LEARNING DYNAMICS OF OFF-POLICY DPO



Outline

Experiments

MITIGATING THE SQUEEZING EFFECT



Outline

Discussion

- SFT & RL
- Pos & Neg

Thanks!