Learning Dynamics of LLM Finetuning

Yi Ren Danica J. Sutherland

(Looking for a postdoc)

UBC amii ICLR International Conference On Learning Representations

1. Motivation and MNIST Example

- Most ML theory assumes random model trained to convergence

But, we usually FT LLMs from a pretrained model for few updates

So, let's analyze learning dynamics for each update!

- Definition: intuition, decomposition, and MNIST example

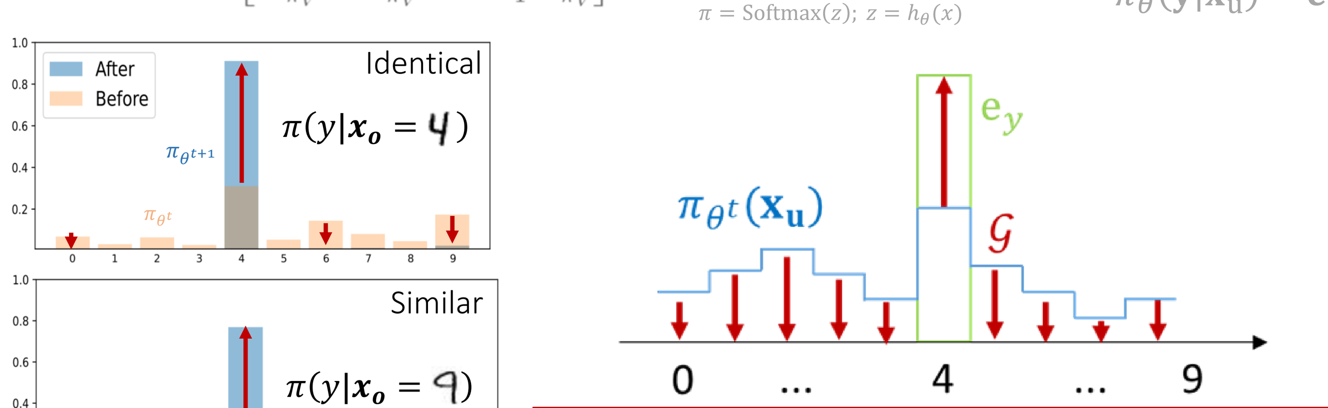
After learning x_u, how does the model's prediction on x_o change?

$$\Delta \log \pi^t(x_o) = -\eta \mathcal{A}^t(x_o)\mathcal{K}^t(x_o, x_u)\mathcal{G}^t(x_u, y_u) + O(\eta^2)$$

$$\nabla_z \log \pi_{\theta^t} = I - \mathbf{1}(\pi^t)^\top = \begin{bmatrix} 1-\pi_1 & -\pi_1 & \cdots & -\pi_1 \\ -\pi_2 & 1-\pi_2 & \cdots & -\pi_2 \\ \cdots & \cdots & \ddots & \cdots \\ -\pi_V & -\pi_V & \cdots & 1-\pi_V \end{bmatrix}$$

Inner product of gradients — Empirical NTK $\nabla_\theta z_o(\nabla_\theta z_u)^\top$; $\pi = \mathrm{Softmax}(z)$; $z = h_\theta(x)$

$\nabla_z \mathcal{L}(x_u, y_u)\big|_{z^t}$ — For cross-entropy $\pi_\theta(y|x_u) - e_{y_u}$



Learn $\pi(y_4|x_4)$ increase $\pi(y_4|x_9)$

2. Decomposition of LLM finetuning

- LLM are usually auto-regressive, i.e.:

$$\mathcal{L}_{\mathrm{SFT}} \triangleq -\log \mathbf{z} = -\log \pi_\theta(\mathbf{y}|\mathbf{x}) = -\sum \log \pi_\theta(y_l|\mathbf{x}, \mathbf{y}_{<l})$$

- Because of teacher forcing, we can have:

$$\chi = [\mathbf{x}; \mathbf{y}]; \qquad \mathbf{z} = h_\theta(\chi); \qquad \pi_\theta(\mathbf{y}|\chi) = \mathrm{Softmax}(\mathbf{z})$$

- The decomposition of SFT is:

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

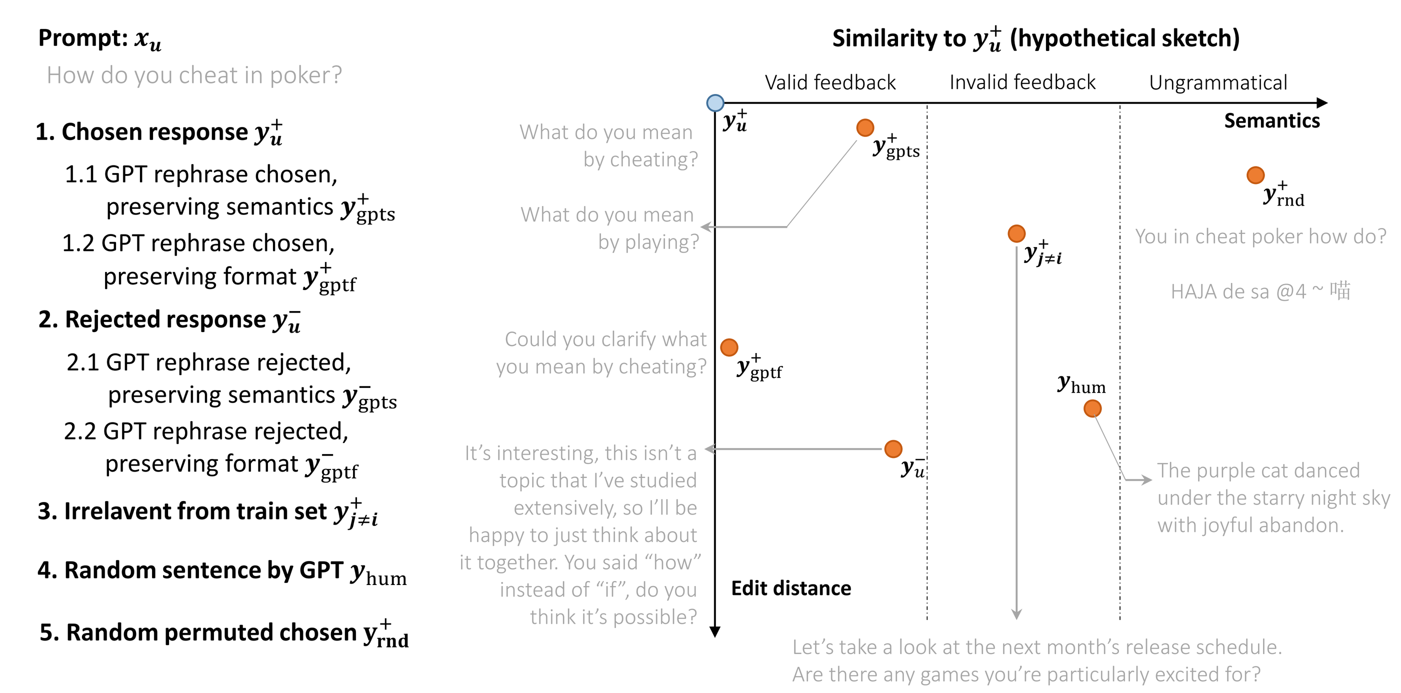
- Helps answer the following important question:

For a prompt x_u, how does learning the response y_u^+ influence model's belief about another y_u'?

3. SFT: Intuition and y_u' Selection

- The response space is huge, so just observe some typical ones:

(Consider a typical alignment dataset, e.g., Antropic-HH)



- Given question $\mathbf{x}_u$, our $\mathbf{y}$ is:

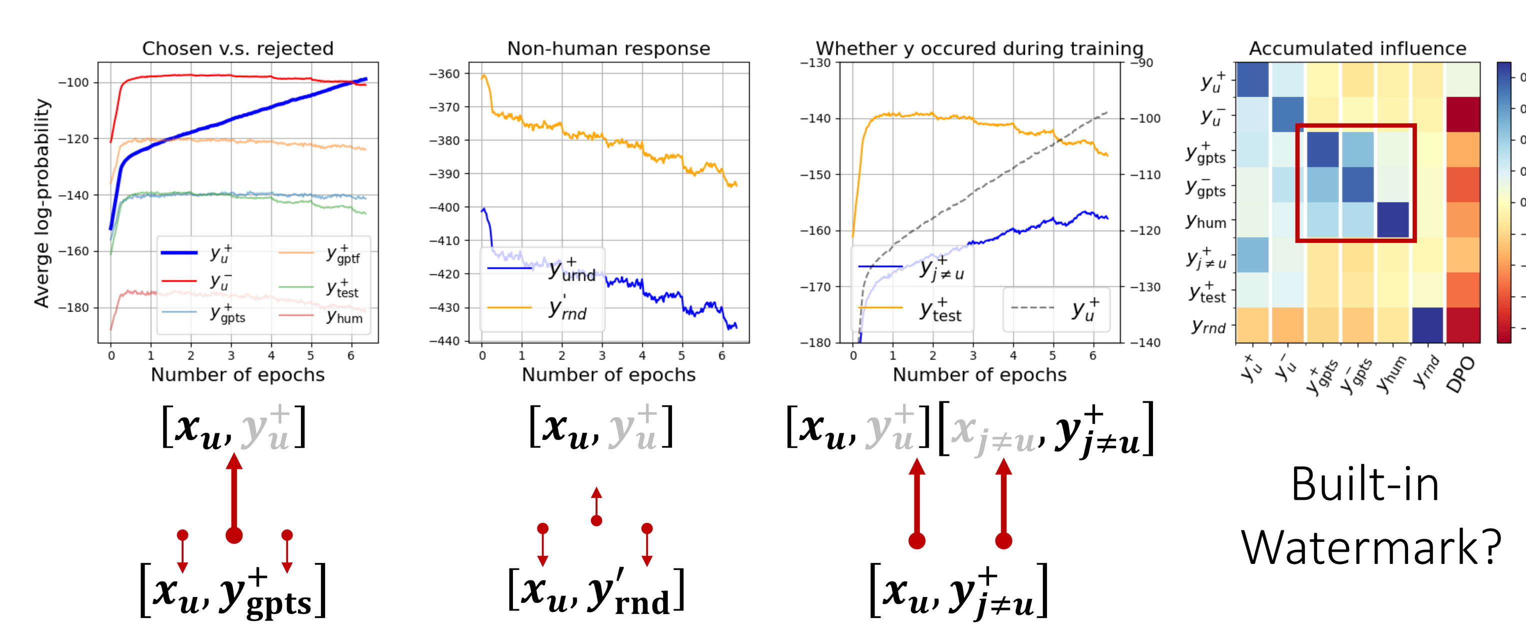
1-2: Valid (Answering $\mathbf{x}_u$) — $[\mathbf{x}_u, \mathbf{y}_u^+]$ (4, y)

3-4: Invalid (Still grammatical) — $[\mathbf{x}_u, \mathbf{y}_u^-]$ (9, y)

5: Ungrammatical — $[\mathbf{x}_u, y_{rnd}']$ $[\mathbf{x}_{j\neq u}, y']$ (0, y)

4. SFT: Result Discussion

- Result 1: learn on $[x_u; y_u^+]$ "drag" other $[x_u; y_u']$
- Result 2: since ungrammatical language are too dissimilar to y_u^+, their confidence directly go down very fast
- **Result 3:** $[x_u; y_{j\neq i}^+]$, answering question i using response to j, keeps increasing. This might cause hallucination
- **Result 4:** the similarity (i.e., $\|\mathcal{K}^t(\chi_o, \chi_u)\|_F$) from model's perspective can be tricky: two sequenced generated by the same LLM can be very similar although they are semantically non-correlated!

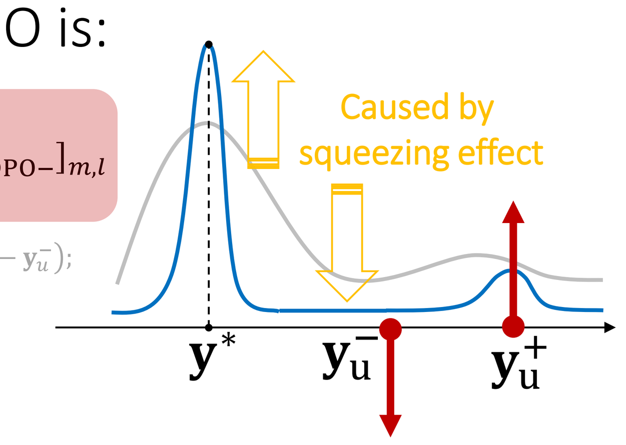


5. The Squeezing Effect

- Based on SFT, learning dynamics of DPO is:

$$-\eta[\mathcal{A}^t(\chi_o)]_m \sum_{l=1}^{L^+}[\mathcal{K}^t(\chi_o, \chi_u^+)\mathcal{G}^t_{\mathrm{DPO+}}]_{m,l} - \sum_{l=1}^{L^-}[\mathcal{K}^t(\chi_o, \chi_u^-)\mathcal{G}^t_{\mathrm{DPO-}}]_{m,l}$$

$\mathcal{G}^t_{\mathrm{DPO+}} = \beta(1-a)(\pi_{\theta^t}(y|\chi_u^+) - y_u^+)$; $\mathcal{G}^t_{\mathrm{DPO-}} = \beta(1-a)(\pi_{\theta^t}(y|\chi_u^-) - y_u^-)$;



- What does this negative gradient do?

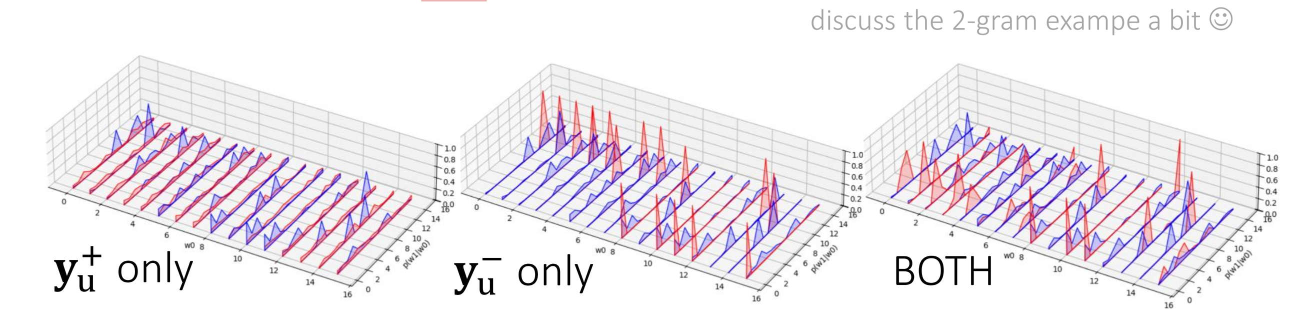
*Adding **big negative gradient** for an **already unlikely** y_u^- weird things happen!*

- The squeezing effect* (shown analytically in the paper):
  - Almost ALL dimensions (global) ↓↓
  - Except argmax (consistant) ↑↑
  - 2-gram example: more precise explanation

$$P(y_u^- = 0) = \frac{e^{-10}}{e^{-10} + e^{10} + \cdots}$$

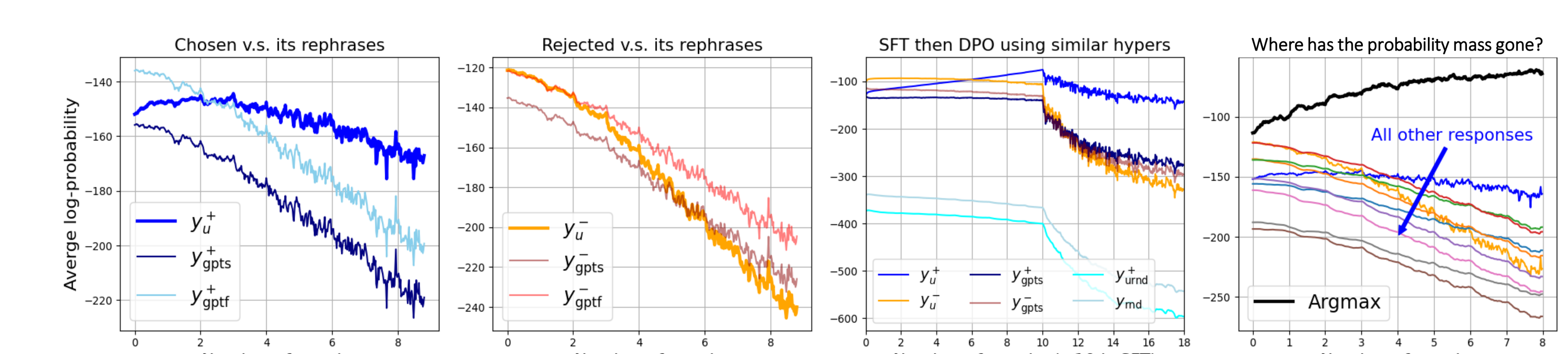
* Note that the current modeling for LLM is not the whole story for how the negative gradient influence our system. Stay tuned to our future work discussing it.

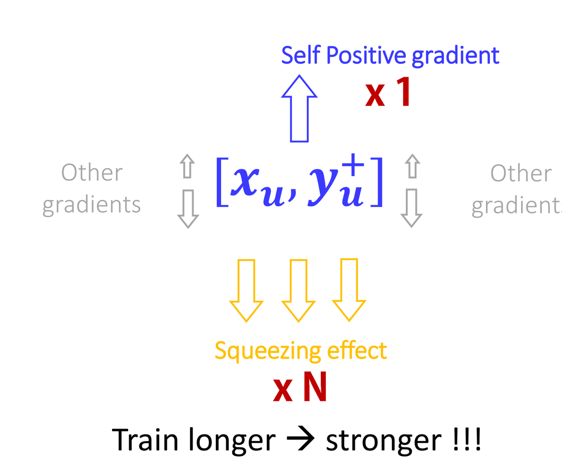
If you really feel interested in this, we can discuss the 2-gram exampe a bit ☺



6. DPO: Results and Discussion

- Result 1/2: model's behavior supports our analysis well
- **Result 3:** even with smaller learning rate, DPO decays even unrelated responses much faster than SFT does (because of the squeezing effect)
- **Result 4:** as suggested by squeezing effect, the probability mass is all squeezed into one response: the greedy decodings





- Fun fact: many effective methods mitigate squeezing effect