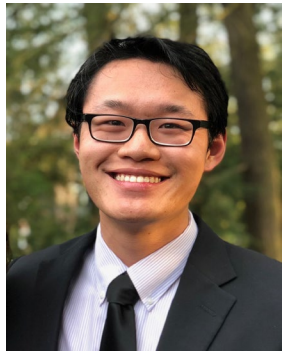
Learning to Answer from Correct Demonstrations

Nati Srebro (TTIC and Mirendil Inc)

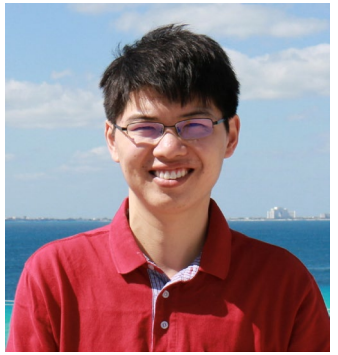
Nirmit Joshi (TTIC)



Gene Li (TTIC→2σ)

Siddharth Bhandari (TTIC→Amazon)

Shiva Kasiviswanathan (Amazon)



Cong Ma (UChicago)

Also featuring work with **Chenxiao Yang (TTIC)** and Zhiyuan Li (TTIC)

What are LLMs and what do we want from them?

- What problem do they solve?
- What's the measure of success?

- Density Model $\hat{p}$ for strings in Σ^n or Σ^* or $\Sigma^{\mathbb{N}}$
  - Next token predictor: probabilistic mapping $\rho: \Sigma^* \to \Sigma$, i.e. defining $\rho(\, x[t] \mid x[:t] \,)$

  ➔ This induces density model $\hat{p}(x) = \prod_t \rho(\, x[t] \mid x[:t] \,)$

  - Training data: samples $x_1, x_2, \ldots \sim \mathcal{D}$
  - Log-loss training ≡ Max likelihood $\max_{\rho \in \mathcal{P}} \prod_i p_\rho(x_i)$
  - Goal: approximate $\mathcal{D}$ (in what sense? KL? Reverse KL? TV? Something else?)
  - **Why?**
  - **What text distribution $\mathcal{D}$?**

- Density View:
  - Learn $\hat{p}$ that matches "language distribution" $\mathcal{D}$

- **Generator view:**
  - Based on examples $x_1, x_2, \ldots$ of texts from language, generate text $\hat{x}$ (e.g. $\hat{x} \sim \hat{p}$)
  - $\hat{x}$ should be a "good" text, i.e. in the language
  - $\hat{x}$ should be "new", i.e. different from $x_1, x_2, \ldots$

  but: need not match any distribution; no *coverage* (OK to only generate from subset)

  [Kleinberg Mullainathan 2024 "Language Generation in the Limit"]

- **Prompt-response view:**
  - (probabilistic) mapping $\pi: \mathcal{X} \to \mathcal{Y}$ (ie defining $\pi(y|x)$), with $\mathcal{X}, \mathcal{Y} = \Sigma^*$
    induced by next token predictor: $\pi(y|x) = \prod_t \rho(\ y[t] \mid x \oplus y[:t]\ )$
  - Reward (utility, correctness) $r(x, y)$
  - Goal: $V_r(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\big[r\big(x, \pi(x)\big)\big] = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(\cdot|x)}[r(x, y)]$

Utility (goodness, reward) vs matching distribution or newness

- $x =$ “How many ounces in a cup”
- $x =$ “Rewrite the following paragraph using clear, crisp, professional language: …”
- $x =$ “What is a prime number”
- $x =$ “Should abortion be legal?”
- $x =$ “Who is the best football player of all times?”
- $x =$ “Prove that the square root of two is irrational”
- $x =$ “Consider a 2025 × 2025 grid of unit squares. Matilda wishes to place on the grid some rectangular tiles, possibly of different sizes, such that each side of every tile lies on a grid line and every unit square is covered by at most one tile. Determine the minimum number of tiles Matilda needs to place so that each rowand each column of the grid has exactly one unit square that is not covered by any tile.”

- Prompt-response view:
  - (probabilistic) mapping $\pi: \mathcal{X} \to \mathcal{Y}$ (ie defining $\pi(y|x)$), with $\mathcal{X}, \mathcal{Y} = \Sigma^*$
  - Reward (utility, correctness) $r(x, y)$
  - Goal: $V_r(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\big[r\big(x, \pi(x)\big)\big] = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(\cdot|x)}[r(x, y)]$

    E.g.: $r(x, y) = 1$ iff y is correct answer for x, i.e. $y \in \sigma(x)$

    $r(x, y) = 0..7$ fractional credit for answer

    $r(x, y) =$ how good is my paragraph rewrite

    In general: $r(x, y) =$ how well does y answer/perform x

- Apprenticeship Learning (think of SFT):
  - Training data: samples from some good (perhaps optimal) demonstrator $\tilde{\pi}$:
    $(x_1, y_1), \ldots, (x_m, y_m) \sim iid\ \mathcal{D} \times \tilde{\pi}$
    i.e. $x_i \sim \mathcal{D}$, $y_i \sim \tilde{\pi}(\cdot\,|x_i)$
  - Assume/hope: $V_r(\tilde{\pi})$ is high, perhaps $V_r(\tilde{\pi}) = V_r^* = \sup_\pi V_r(\pi)$

Goal: $V_r(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[r\left(x,\pi(x)\right)\right]$
Training data: iid $x_i \sim \mathcal{D}$, $y_i \sim \tilde{\pi}(\cdot\,|x_i)$
$0 \le r(x,y) \le 1$

- Demonstrator Assumption: $\tilde{\pi} \in \Pi$, for some known class Π

  (e.g. demonstrator can be described by transformer)
- MLE is min-max optimal: w.p.$\geq 1-\delta$ over training data,

$$\underbrace{D_H^2(\mathcal{D}\times\hat{\pi}_{MLE}, \mathcal{D}\times\tilde{\pi})}_{\text{dist of responses } \hat{\pi}(x) \approx \text{dist of responses } \tilde{\pi}(x)} \le 6\,\frac{\log|\Pi| + \log {}^{2}/_{\delta}}{m}$$

following [Foster, Block, Misra 2004 "Is Behavior Cloning All you Need"]

➔ We should do MLE, i.e. log-loss minimization

Do you need beh

If Π defined by class of next-token-generators $\mathcal{H} = \{\rho\}$

$$\Pi = \{\, \pi(y|x) = \textstyle\prod_t \rho(\ y[t]\ |\ x \oplus y[:t]\ )\ |\ \rho \in \mathcal{H}\,\}$$

- $\log|\Pi| \le \log|\mathcal{H}| \le 32 \cdot \#params$ (if using floats)
- Can also replace by $VC(\mathcal{H})\log(|y|)\log m$ [Joshi Li S 2025]
- For (hard attention) transformers: $VC(\mathcal{H}) = \Theta(\#params \cdot \log(|x|+|y|))$

  ...and $\log|\Pi|$ can be replaced by $O(\#params \cdot log(|x|+|y|)\cdot \log m)$

  [Yang Li S 2026]

- Model class assumption: $r \in \mathcal{R}$ for known $\mathcal{R}$

  e.g. correctness can be described by transformer
- No assumptions about structure of demonstrator $\tilde{\pi}$

Goal: $V_r(\pi) = \mathbb{E}_{x\sim\mathcal{D}}[r(x,\pi(x))]$
Training data: iid $x_i \sim \mathcal{D}, y_i \sim \tilde{\pi}(\cdot\,|x_i)$
$0 \le r(x,y) \le 1$

If demonstrator $\tilde{\pi}$ is optimal:

$r \in \mathcal{R}$ ➔ $\tilde{\pi} \in \Pi^*_{\mathcal{R}} = \{\pi | \pi \in \arg\max V_r(\pi), r \in \mathcal{R}\}$, but generally $|\Pi^*_{\mathcal{R}}| \gg |\mathcal{R}|$

$\tilde{\pi} \in \Pi$ ➔ $r \in \mathcal{R}_\Pi = \left\{ r_\pi(x,y) = 1_{y \in supp(\pi(\cdot|x))} \middle| \pi \in \Pi \right\}$, and $|\mathcal{R}_\Pi| \le |\Pi|$

(and learning w.r.t. $\mathcal{R}_\Pi$ implies learning w.r.t. Π)

1. **MLE can fail**: There exists a class with $|\mathcal{R}| = |\mathcal{Y}| = 2$, a choice $r \in \mathcal{R}$, and an optimal demonstrator $V_r(\tilde{\pi}) = 1$ s.t. for any sample size m, and any ϵ, there exists $\mathcal{D}$ with an $\Pi^*_{\mathcal{R}}$-MLE with: $V_r(\hat{\pi}_{MLE}) < \epsilon$.
2. **Cloning is not possible**: Even if $|\mathcal{R}| = 1$, so we know r, e.g. $r(x,y) = 1$ always, and $\tilde{\pi}$ is optimal for r, there is no way to match the the distribution $\tilde{\pi}(\cdot\,|x)$ (under any reasonable distance) from a finite number of samples

$\mathcal{R} = \{r_0(x,y) = 1_{y\in\{0\}}, \quad r_1(x,y) = 1\}$
Correct $r = r_0$, and $\tilde{\pi}(x) = 0$
But $\hat{\pi}_{MLE}(x) = \begin{cases} 0, on\ observed\ x \\ 1, on\ unobserved\ x \end{cases}$ is a valid MLE

3. **Learning the reward is not possible:** Even with $|\mathcal{R}| = |\mathcal{Y}| = 2$ and optimal demonstrator

$\mathcal{R} = \{r_0(x,y) = 1_{y\in\{0\}}, \quad r_1(x,y) = 1\}$
Correct $r = r_0$, and $\tilde{\pi}(x) = 0$

4. **Learning to answer is possible**: our algorithm ensures $V_r(\tilde{\pi}) - V_r(\hat{\pi}) \le O\left(\sqrt{\frac{\log|\mathcal{R}|}{m}}\right)$

Online Apprenticeship Learning (Binary Reward Realizable Version)

At each step t:

Receive x_t

Predict $\hat{y}_t$

If $r^\circ(x_t, \hat{y}_t) \neq 1$, this is a mistake

(but the **learner doesn't know this**)

Receive $\tilde{y}_t$ s.t. $r^\circ(x_t, \tilde{y}_t) = 1$

$r^\circ \in \mathcal{R}$

Online Contextual Bandits:

- Receive reward $r^\circ(x_t, \hat{y}_t)$, but not demo $\tilde{y}_t$
- Inherently interactive; here passive

[Kleinberg Mullainathan]:

- (Mostly) unprompted
- Require *new* $\hat{y}$
- Goal: finite, but unbounded, #mistakes

Online Apprenticeship Learning (Binary Reward Realizable Version)

At each step t:

Receive x_t

Predict $\hat{y}_t$

If $r^\circ(x_t, \hat{y}_t) \neq 1$, this is a mistake

(but the **learner doesn't know this**)

Receive $\tilde{y}_t$ s.t. $r^\circ(x_t, \tilde{y}_t) = 1$

$r^\circ \in \mathcal{R}$

Algorithm

Init $w(r) = 1$

At each step t:

$\sum_{r(x_t,y)=1} w(r)$

$$\hat{y}_t = \hat{\pi}_t(x_t) = \arg\max_y \sum_{r \in \mathcal{R}} w(r) r(x_t, y)$$

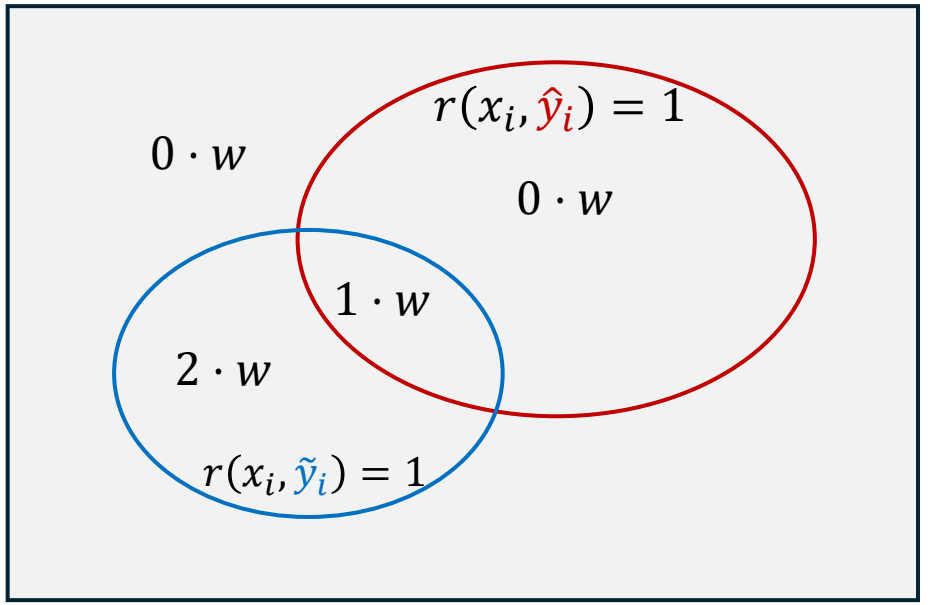
And then update:

$$w(r) \leftarrow \begin{cases} 0 & \text{If } r(x_t, \tilde{y}_t) \neq 1 \\ 2w(r) & \text{If } r(x_t, \hat{y}_t) \neq 1 \\ w(r) & \text{otherwise} \end{cases}$$

Claim: $\#mistakes < \log |\mathcal{R}|$

Proof: $\sum w(r)$ is monotone non-increasing, and hence $\leq |\mathcal{R}|$

➔ $|\mathcal{R}| \geq w(r^\circ) = 2^{\#\, mistakes}$ ➔ $\#\, mistakes \leq \log|\mathcal{R}|$



Online Apprenticeship Learning (Fractional Rewards; Any Demonstrator)

At each step t:

Receive x_t

Predict $\hat{y}_t$

Earn $r^\circ(x_t, \hat{y}_t)$,

(but the **learner doesn't know this**)

Receive $\tilde{y}_t$ s.t. $r^\circ(x_t, \tilde{y}_t) = 1$

$r^\circ \in \mathcal{R}$

Algorithm

Init $w(r) = 1$

At each step t:

$$\hat{y}_t = \hat{\pi}_t(x_t) = \arg\max_y \sum_{r \in \mathcal{R}} w(r) r(x_t, y)$$

And then update:

$$w(r) \leftarrow (1+\gamma)^{\sup_y r(x_i,y) - r(x_i,\hat{y}_i)} (1-\gamma)^{\sup_y r(x_i,y) - r(x_i,\hat{y}_i)}$$

Claim: $\sum w(r)$ is monotone non-increasing, and hence $\leq |\mathcal{R}|$

Online-to-Batch

Run online alg one-pass on data

return mixture $\hat{\pi} = \hat{\pi}_I$, $I \sim Unif(1..m)$

$$\rightarrow \tilde{R} - \hat{R} \leq 2\gamma\left(R^* - \tilde{R}\right) + \frac{4}{3\gamma}\log|\mathcal{R}|$$

$R^* = \sum \sup_y r^\circ(x_i, \tilde{y}_i)$

$\hat{R} = \sum r^\circ(x_i, \hat{y}_i)$

$\tilde{R} = \sum r^\circ(x_i, \tilde{y}_i)$

For any $\gamma \leq 1$ with $\gamma \equiv \sqrt{\log\frac{|\mathcal{R}|}{m\Delta}}$, for any $r^\circ \in \mathcal{R}$, $\tilde{\pi}$ s.t. $V_{r^\circ}(\tilde{\pi}) \geq V_{r^\circ}^* - \Delta$, w.p. $1-\delta$:

with appropriate γ

$$V_{r^\circ}(\tilde{\pi}) - V_{r^\circ}(\hat{\pi}) \leq 8\sqrt{\frac{\Delta(\log|\mathcal{R}| + \log 2/\delta)}{m}} + 16\frac{\log|\mathcal{R}| + \log 2/\delta}{m}$$

Online Apprenticeship Learning (Fractional Rewards; Any Demonstrator)

At each step t:

Receive x_t

Predict $\hat{y}_t$

Earn $r^\circ(x_t, \hat{y}_t)$,

(but the **learner doesn't know this**)

Receive $\tilde{y}_t$ s.t. $r^\circ(x_t, \tilde{y}_t) = 1$

$r^\circ \in \mathcal{R}$

Similar to **[Syed Schapire 2007]** Apprenticeship Learning for MDPs:

- More direct modeling of reward class
- Simpler more direct algorithm and analysis
- Optimistic $1/m$ rate
- One-pass online
- Holds even with adaptive demonstrations

[Moulin Neu Viano NeurIPS 2025]: $\tilde{O}\left(\sqrt[4]{\frac{\log|\mathcal{Y}| \cdot \log|\mathcal{R}|}{m}}\right)$

Claim: $\sum w(r)$ is monotone non-increasing, and hence $\leq |\mathcal{R}|$

$\rightarrow \tilde{R} - \hat{R} \leq 2\gamma\left(R^* - \tilde{R}\right) + \frac{4}{3\gamma}\log|\mathcal{R}| \leq 4\sqrt{\left(R^* - \tilde{R}\right)\log|\mathcal{R}|} + 3\log|\mathcal{R}|$

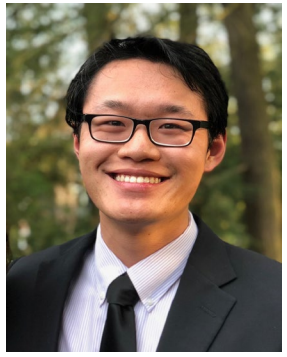
For any $\Delta \leq 1$ with $\gamma = \sqrt{\log\frac{|\mathcal{R}|}{m\Delta}}$, for any $r^\circ \in \mathcal{R}$, $\tilde{\pi}$ s.t. $V_{r^\circ}(\tilde{\pi}) \geq V^*_{r^\circ} - \Delta$, w.p. $1 - \delta$:

$$V_{r^\circ}(\tilde{\pi}) - V_{r^\circ}(\hat{\pi}) \leq 8\sqrt{\frac{\Delta(\log|\mathcal{R}| + \log 2/\delta)}{m}} + 16\frac{\log|\mathcal{R}| + \log 2/\delta}{m}$$

Learning to Answer from Correct Demonstrations

- Based on $r \in \mathcal{R}$, and with no assumption on demonstrator $\tilde{\pi}$:
  - $V_r(\tilde{\pi}) - V_r(\hat{\pi}) \leq O\left(\sqrt{\frac{\log|\mathcal{R}|}{m}}\right)$ (and even optimistic rate if $\tilde{\pi}$ near-optimal)
  - Impossible via MLE
  - Not via cloning (ie distribution matching)
  - Cloning / distribution match impossible
- Challenges:
  - Practical implementation for parametric classes
  - Searching for best response (➔ co-train generator and reward)
  - Leverage pre-trained model

Nirmit Joshi



Gene Li

Siddharth Bhandari

Shiva Kasiviswanathan

Cong Ma