TRANSFER LEARNING FOR WEAK-TO-STRONG GENERALIZATION


Seamus Somerstep


Felipe Maia Polo


Mouli Banerjee


Subha Maity

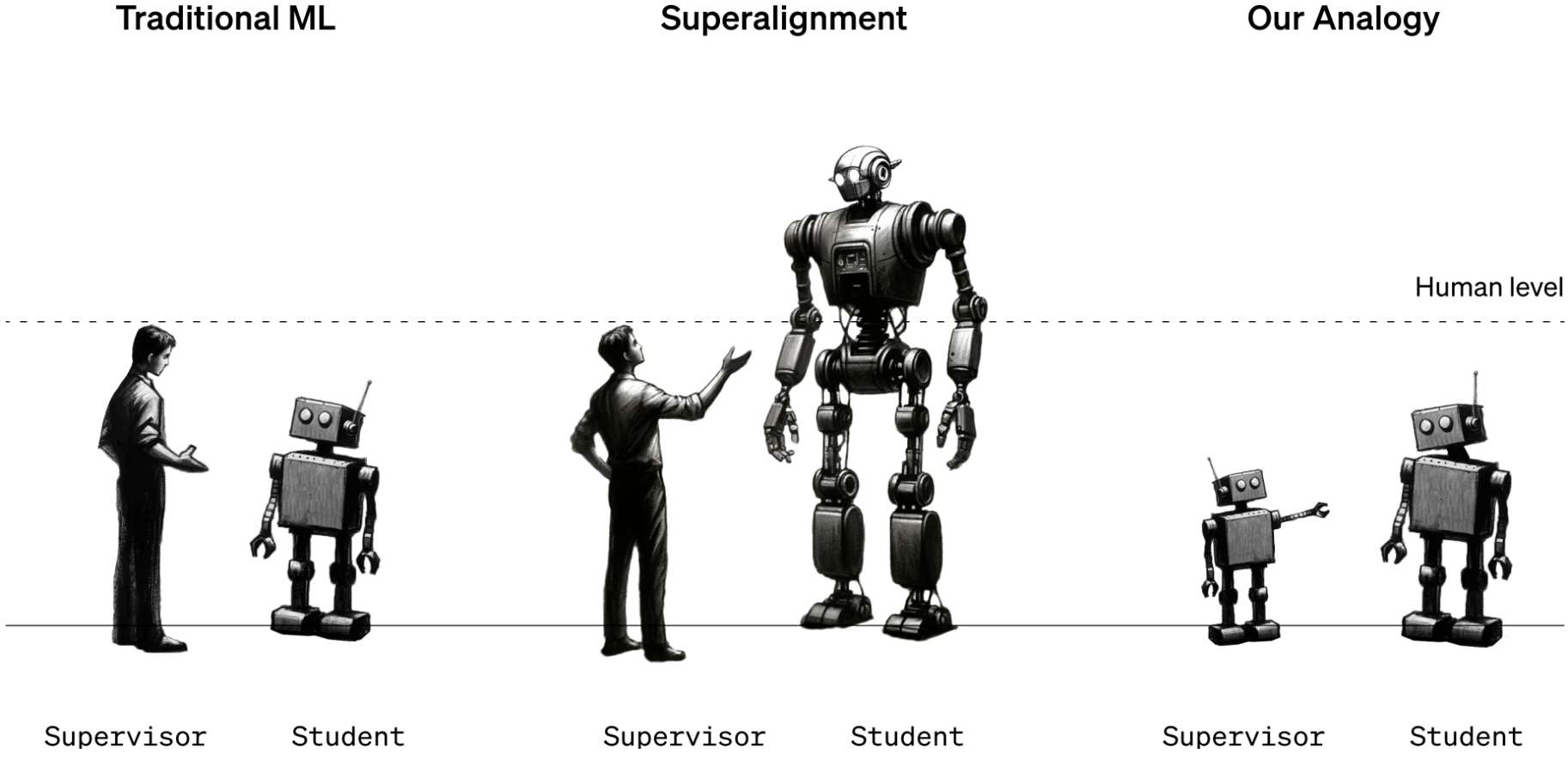

Ya'acov Ritov

Illustrations by Rami Ritov


Mikhail Yurochkin

SUPERALIGNMENT

Goal: Supervise the fine-tuning of a larger (more capable) pretrained LLM with a smaller (less capable) LLM.



Burns et al (2023)

SUPERALIGNMENT

Goal: Supervise the fine-tuning of a larger (more capable) pretrained LLM with a smaller (less capable) LLM.

☒ The pretrained LLM may imitate the weak teacher's errors/mistakes; thus limiting its performance to no better than its (weak) teacher.

☑ Pretrained LLMs have impressive (latent) raw abilities—there is no need to teach them new tasks from scratch.

We focus on **eliciting the latent abilities** of pretrained LLMs.

TRANSFER LEARNING SETUP

Given

- $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$: n_P samples from source P
- $\{(X_{Q,i}, Y_{Q,i})\}_{i=1}^{n_Q}$: n_Q (labeled) samples from target $Q \neq P$
- $n_P \gg n_Q$

Find f such that $\mathbf{E}\left[\ell(f(X_Q), Y_Q)\right]$ is small.

Transfer is generally futile without similarity assumptions on P and Q.

WEAKLY-SUPERVISED TRANSFER SETUP

Given

- $\{(X_{P,i}, Y_{P,i})\}_{i=1}^{n_P}$: n_P samples from P
- $\{(X_{Q,i}, Y'_{Q,i})\}_{i=1}^{n_Q}$: n_Q *weakly-labeled* samples from Q

Find f such that $\mathbf{E}\left[\ell(f(X_Q), Y_Q)\right]$ is small.

- (X_Q, Y_Q): target sample with *gold-standard/strong* label Y_Q

The main challenge here is **the learner has no (strong) labels from** Q**!**

WEAK-TO-STRONG GENERALIZATION AS TL

superalignment	**weakly-supervised TL**
pretrained LLM	$Y_P \mid X_P$
(super)alignment task	$Y_Q \mid X_Q$
weak teacher	$Y'_Q \mid X_Q$

In superalignment, the learner has access to the pretrained LLM and weak teacher; **it has no (direct) supervision on the alignment task.**

WEAK-TO-STRONG GENERALIZATION AS TL

Assume

1. **Latent concept shift:** $Y_P \mid X_P, Y_Q \mid X_Q$ are mixtures of the same (mixture) components:

$$P_{Y|X} = \sum_{\theta \in \Theta} \pi_\theta G_\theta(\cdot \mid X), \quad Q_{Y|X} = G_{\theta_Q}(\cdot \mid X)$$

2. n_P is extremely large, so the pretrained LLM is exactly $Y_P \mid X_P$

Transfer learning under latent concept shift is a deconvolution problem: learning the **target concept** θ_Q from the weak labels.

Challenge: deconvolution without knowledge of the G_θ's.

LATENT CONCEPT MODEL OF LLMS

$$\texttt{GPT(Y|X)} = \sum_{\theta \in \Theta} g(Y \mid X, \theta) \pi(\theta \mid X).$$

- θ: (latent) concept (eg nationality, occupation)
- $g(\cdot \mid X, \theta)$: predictive distribution associated with θ
  - English ~ g(· | Isaac Newton is, nationality)
  - scientist ~ g(· | Isaac Newton is, occupation)
- $\pi(\cdot \mid X)$: prior (distribution) on concepts

The predictive distributions $g(\cdot \mid X, \theta), \theta \in \Theta$ encode the pretrained LLM's (latent) abilities.

LATENT CONCEPT SHIFT

Latent concept shift: $Y_P \mid X_P, Y_Q \mid X_Q$ are mixtures of the same (mixture) components:

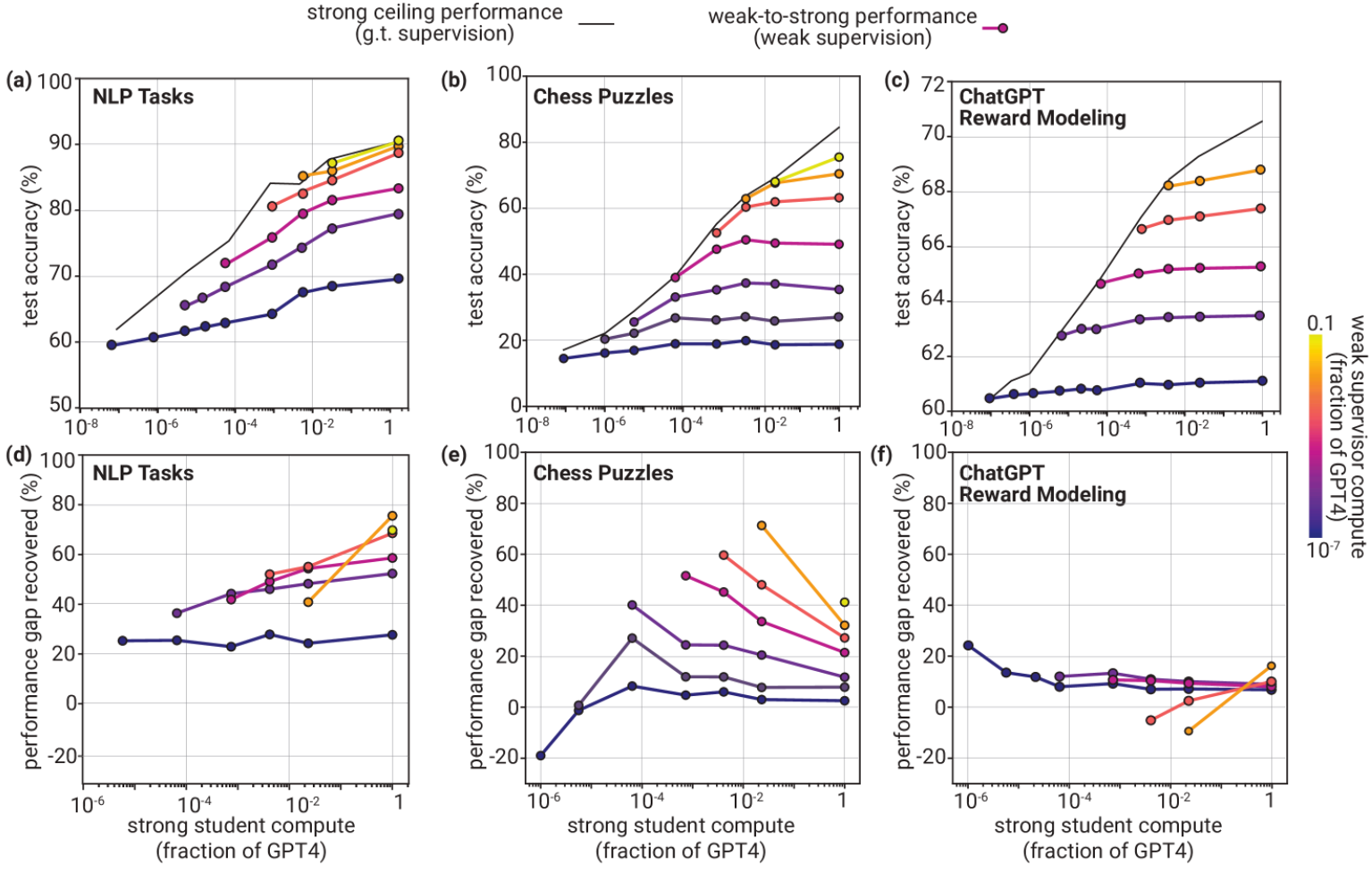
$$P_{Y|X} = \sum_{\theta \in \Theta} \pi_\theta G_\theta(\cdot \mid X), \quad Q_{Y|X} = G_{\theta_Q}(\cdot \mid X)$$

- The pretrained LLM satisfies the latent concept model.

$$\texttt{GPT(Y|X)} = \sum_{\theta \in \Theta} g(Y \mid X, \theta)\pi(\theta \mid X).$$

- The superalignment task is encoded by $g(\cdot \mid X, \theta_Q)$ (for some $\theta_Q \in \Theta$).

(NAIVE) FINE-TUNING



Burns et al (2023)

W2S REGRESSION

The learner is given

- $X \in \mathbf{R}^{n \times d}$: (fixed) input matrix,
- $y_P \in \mathbf{R}^n$: source model outputs (on inputs in X),
- $y'_Q \in \mathbf{R}^n$: weak teacher outputs.

Assume

1. **Latent concept shift:** source and target regression functions are mixtures of the same (mixture) components:

$$f_P^\star(X) = \textstyle\sum_{\theta \in \Theta} \pi_\theta g_\theta(X), \quad f_Q^\star(X) = g_{\theta_Q}(X).$$

2. The source model is exactly $f_P^\star$; ie $y_P = f_P^\star(X)$.

3. The weak teacher is unbiased; ie $\mathbf{E}\left[y'_Q \mid X\right] = f_Q^\star(X)$.

W2S REGRESSION

The learner seeks f_Q such that

1. $\frac{1}{n}\|f_Q(X) - f_Q^\star(X)\|_2^2 \leq \frac{1}{n}\|y_p - f_Q^\star(X)\|_2^2 \triangleq \frac{1}{n}\|\epsilon_P\|_2^2$,
2. $\frac{1}{n}\|f_Q(X) - f_Q^\star(X)\|_2^2 \leq \frac{1}{n}\|y'_Q - f_Q^\star(X)\|_2^2 \triangleq \frac{1}{n}\|\epsilon'_Q\|_2^2$;

ie f_Q that improves upon both the source model and weak teacher.

- $\frac{1}{n}\|\epsilon_P\|_2^2$ is the source model's MSE,
- $\frac{1}{n}\|\epsilon'_Q\|_2^2$ is the weak teacher's MSE.

FINE-TUNING ON WEAK OUTPUTS

$$\widehat{f}_\lambda \leftarrow \operatorname{argmin}_f \left\{ \begin{array}{l} \frac{1}{2}\|y'_Q - f(X)\|_2^2 \\ \quad +\lambda \frac{1}{2}\|y_P - f(X)\|_2^2 \end{array} \right\} \qquad \text{(FT)}$$

- $\frac{1}{2}\|y'_Q - f(X)\|_2^2$ is the loss with respect to outputs from the weak teacher.
- $\frac{1}{2}\|y_P - f(X)\|_2^2$ regularizes f towards the source model.

(FT) is an analogy for **supervised fine-tuning (SFT)** and **reinforcement learning from human feedback (RLHF)**.

FINE-TUNING ON WEAK OUTPUTS

Lem: The MSE of $\widehat{f}_\lambda$ from (FT) is

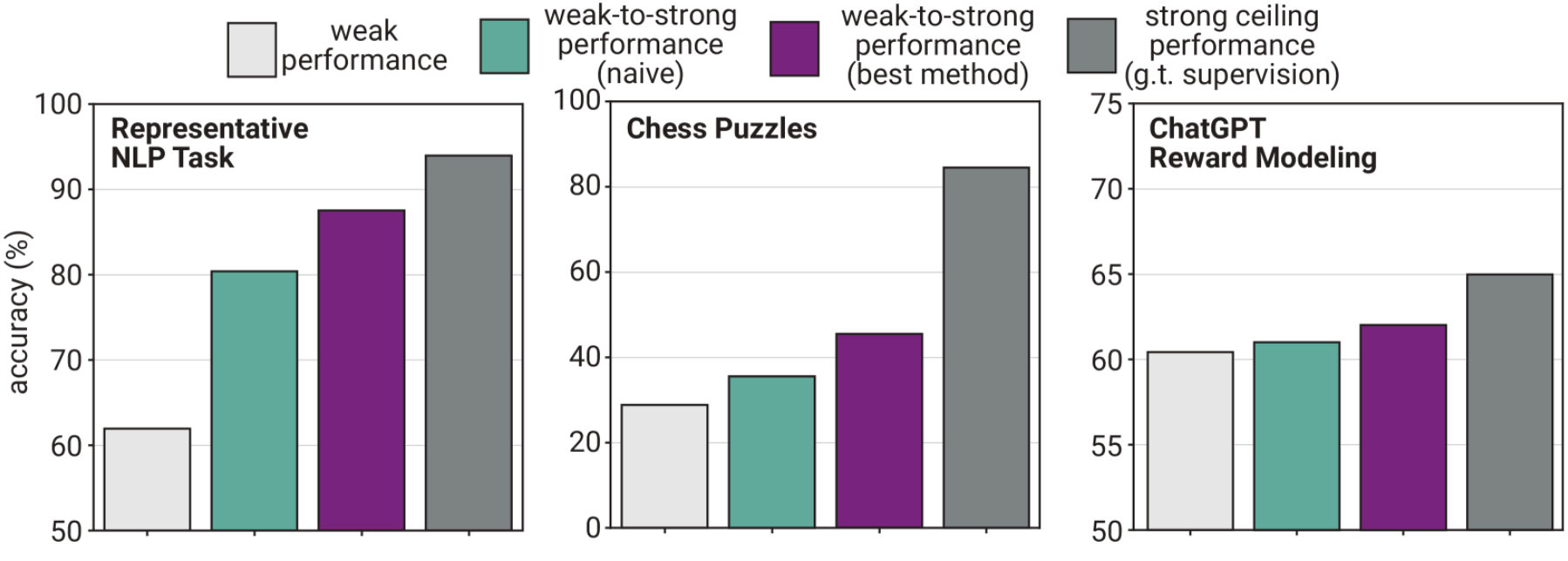
$$\mathbf{E}\left[\frac{1}{n}\|\widehat{f}_\lambda(X) - f_Q^\star(X)\|_2^2\right] = \frac{\lambda^2}{(1+\lambda)^2}\frac{1}{n}\|\epsilon_P\|_2^2 + \frac{1}{(1+\lambda)^2}\frac{1}{n}\|\epsilon'_Q\|_2^2.$$

Takeaway: The MSE of $\widehat{f}_\lambda$ from (FT) is limited by

- the MSE of the source model (if $\|\epsilon_P\|_2^2 < \|\epsilon'_Q\|_2^2$),
- the MSE of the weak teacher (if $\|\epsilon'_Q\|_2^2 < \|\epsilon_P\|_2^2$).

SUPERVISING GPT-4 WITH GPT-2

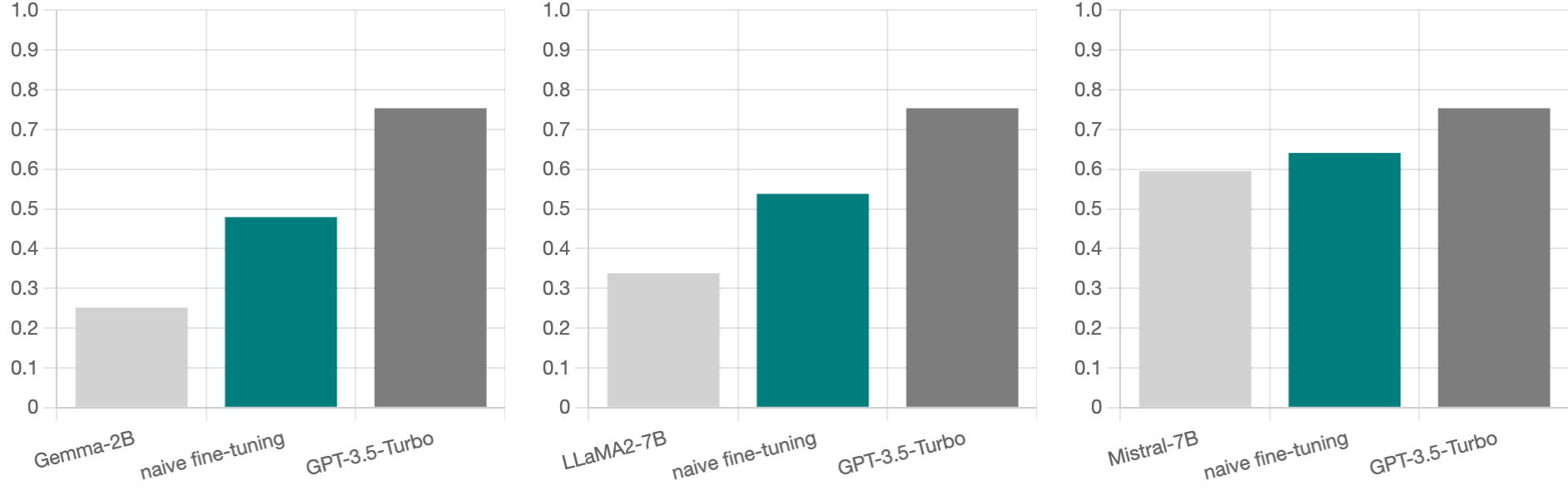
if we fine-tune GPT-4 with labels from a GPT-2-level model, we typically recover about $\frac{1}{2}$ of the performance gap between the two models.
— Burns et al



Burns et al (2023)

MATH REASONING EXPS PREVIEW

Q: Does the fine-tuned LLM outperform the pretrained LLM (without fine-tuning)?



No because fine-tuning does not does not leverage the *latent* abilities of the pretrained LLM.

W2S REGRESSION VIA DECONVOLUTION

Recall the source and target regression functions are mixtures of the same (mixture) components:

$$f_P^\star = \sum_{\theta\in\Theta} \pi_\theta g_\theta, \quad f_Q^\star = g_{\theta_Q}.$$

Idea: Restrict f to (the convex hull of) the mixture components:

$$\widehat{f} \leftarrow \begin{Bmatrix} \mathbf{argmin}_f & \frac{1}{2}\|y'_Q - f(X)\|_2^2 \\ \mathbf{subject\ to} & f \in \mathbf{cvx}(\{g_\theta\}_{\theta\in\Theta}) \end{Bmatrix}. \qquad (\mathbf{dcnv})$$

W2S REGRESSION VIA DECONVOLUTION

Lem: The MSE of $\widehat{f}$ from (dcnv) is at most

$$\begin{aligned}\tfrac{1}{n}\|\widehat{f}(X) - f_Q^\star(X)\|_2^2 &\le \sup_{\theta \in T_{\mathcal{G}}(f_Q^\star(X)) \cap \mathbf{S}^{n-1}} \tfrac{1}{n}((\epsilon_Q')^\top \theta)^2 \\ &\ll \tfrac{1}{n}\|\epsilon_Q'\|_2^2,\end{aligned}$$

where $\mathcal{G} \triangleq \mathbf{cvx}(\{g_\theta(X)\}_{\theta \in \Theta}) \subset \mathbf{R}^n$ (as long as $f_Q^\star(X) \in \mathcal{G}$).

✅ The MSE of (dcnv) is **not** limited by the MSEs of the source model and weak teacher.

❌ Unfortunately, (dcnv) is impractical because it requires knowledge of g_θ's (instead of f_P).

ELICITING LATENT ABILITIES OF LLMS

Recall the latent concept model of LLMs:

$$\texttt{GPT(Y|X)} = \sum_{\theta \in \Theta} g(Y \mid X, \theta)\pi(\theta \mid X).$$

Bayesian explanation of in-context learning (ICL): ICL helps the pretrained LLM to (implicitly) infer the target concept:

$$\begin{aligned}&\mathrm{argmax}_{\texttt{y}}\texttt{GPT(y|}\texttt{X}_1\texttt{,}\texttt{Y}_1\texttt{,...,}\texttt{X}_\texttt{K}\texttt{,}\texttt{Y}_\texttt{K}\texttt{,X)}\\ &\quad\approx \mathrm{argmax}_y g(y \mid X, \theta),\end{aligned}$$

where $Y_k \mid X_k \sim g(\cdot|X_k, \theta)$.

ELICITING LATENT ABILITIES OF LLMS

Idea: Use examples from the weak teacher (instead of from $g(\cdot \mid \cdot, \theta_Q)$) as ICL examples; ie hope/pray that

$$\begin{aligned} &\texttt{argmax}_{\texttt{y}}\texttt{GPT}(\texttt{y} \mid \texttt{X}_1, \texttt{Y}_1, \ldots, \texttt{X}_\texttt{K}, \texttt{Y}_\texttt{K}, \texttt{X}) \\ &\quad \approx \texttt{argmax}_y g(y \mid X, \theta_Q), \end{aligned}$$

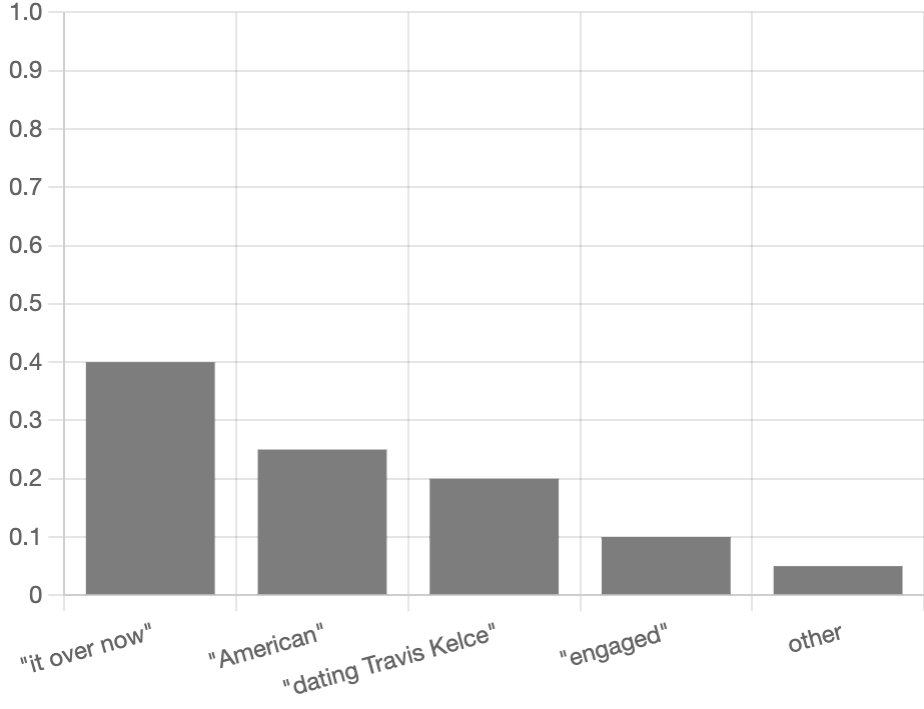
where (X_k, Y_k)'s are examples from the weak teacher .

The pretrained model must be "strong enough" to learn θ_Q in-context from the (weak) teacher.

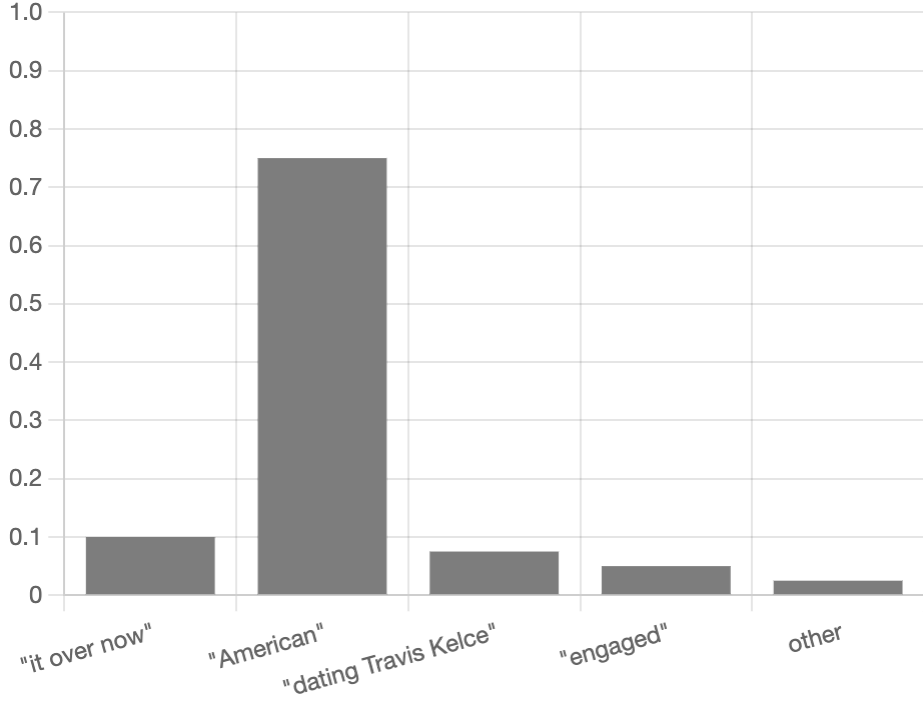
EX: ELICITING THE LATENT ABILITIES OF LLMS

Goal: teach an LLM to respond with the nationality of notable people

$\mathcal{S}'$ consists of ("[name] is", "[nationality]") pairs from weak teacher (so some nationalities are incorrect).



`GPT(·|"Taylor Swift is")`



`GPT(·|S'+"Taylor Swift is")`

FROM DECONVOLUTION TO ICL REFINEMENT

: W2S regression

$$\left\{\begin{matrix}\mathbf{argmin}_f & \frac{1}{2}\|y'_Q - f(X)\|_2^2 \\ \mathbf{subject\ to} & f \in \mathbf{cvx}(\{g_\theta\}_{\theta\in\Theta})\end{matrix}\right\} \approx g_{\theta_Q} = f_Q^\star$$

:: superalignment

$$\begin{aligned}&\mathbf{argmax}_{\mathtt{y}}\mathtt{GPT}(\mathtt{y}\,|\,\mathtt{X}_1,\mathtt{Y}_1,\ldots,\mathtt{X}_\mathtt{K},\mathtt{Y}_\mathtt{K},\mathtt{X})\\ &\quad\approx \mathbf{argmax}_y g(y \mid X, \theta_Q)\\ &\quad= \mathbf{argmax}_y Q_{Y|X}(y),\end{aligned}$$

ICL refinement solves the deconvolution problem *without knowledge of* $g(\cdot \mid \cdot, \theta), \theta \in \Theta$!

IN-CONTEXT LEARNING (ICL) REFINEMENT

Idea: Fine-tune on refined outputs from $\texttt{GPT}(\cdot \mid \texttt{P}, \texttt{X}_\texttt{Q})$.

Require: weakly labeled dataset $\mathcal{D}'_Q \triangleq \{(X_{Q,i}, Y'_{Q,i})\}_{i=1}^{n_Q}$

For $(X_{Q,i}, Y'_{Q,i}) \in \mathcal{D}'_Q$

1. select a subset $\mathcal{S}_i$ of $\mathcal{D}'_Q$ as ICL examples (eg K-NN of $X_{Q,i}$),
2. refine weak teacher output: $\texttt{Y}_{\texttt{Q,i}} \sim \texttt{GPT}(\cdot \mid \texttt{S}_\texttt{i}, \texttt{X}_{\texttt{Q,i}})$.

Fine-tune the pretrained LLM on the refined weak teacher outputs $\{(X_{Q,i}, Y_{Q,i})\}_{i=1}^{n_Q}$.

MATH REASONING EXPERIMENTS

pretrained LLMs:

- `GPT-3.5-Turbo`
- `GPT-4o mini`

weak teachers (fine-tuned on gold standard outputs):

- `Gemma-2B`
- `LLaMA2-7B`
- `Mistral-7B`

We use `GPT-4o` to assess whether outputs agree with the intermediate steps and final answers in the answer key.

GRADE SCHOOL MATH 8K (GSM8K)

- 8.8k (7.5k/1.3k train/test) grade school math problems created by human problem writers
- They take between 2 and 8 steps to solve.
- The solutions mostly entail a sequence of basic arithmetic operations (+ - / *).

> *A bright middle school student should be able to solve every problem.*
> — *Cobbe et al*

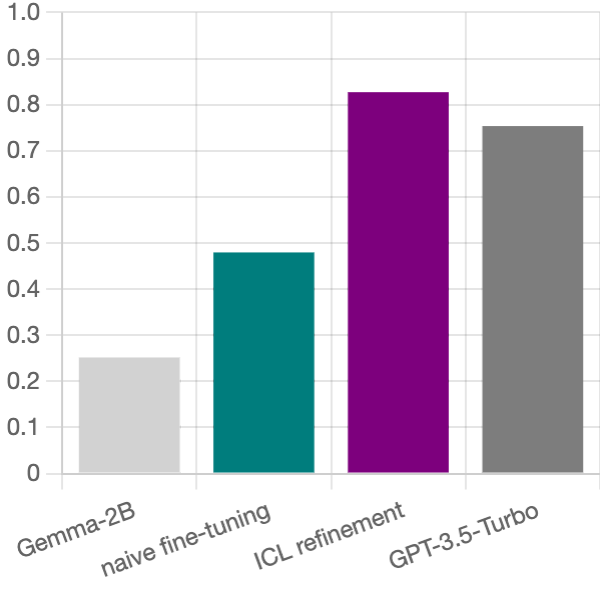
GRADE SCHOOL MATH 8K (GSM8K)

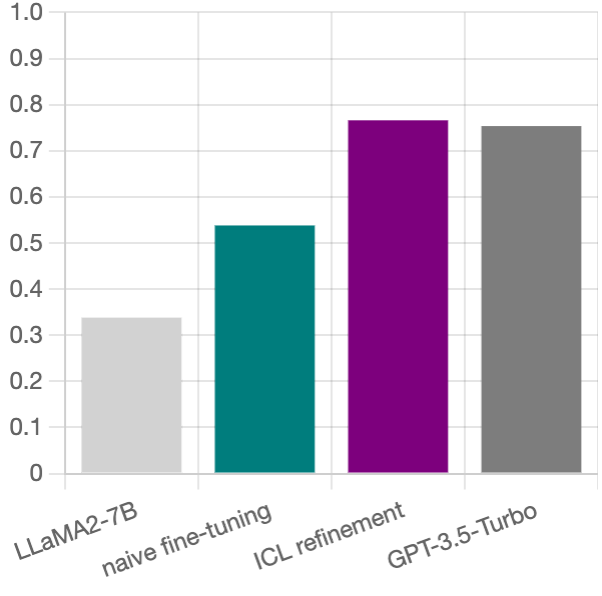
Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

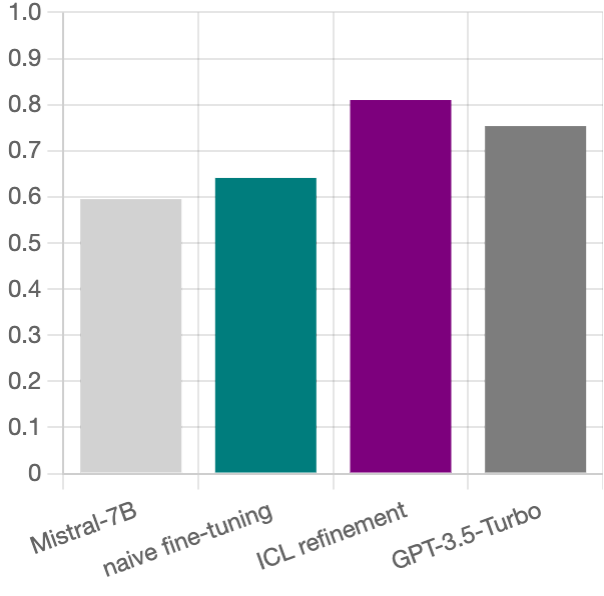
Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

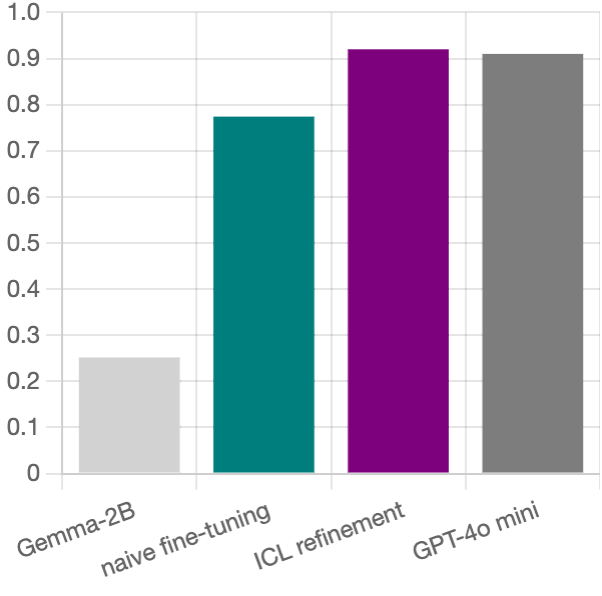
GSM8K RESULTS (GPT-3.5 TURBO)

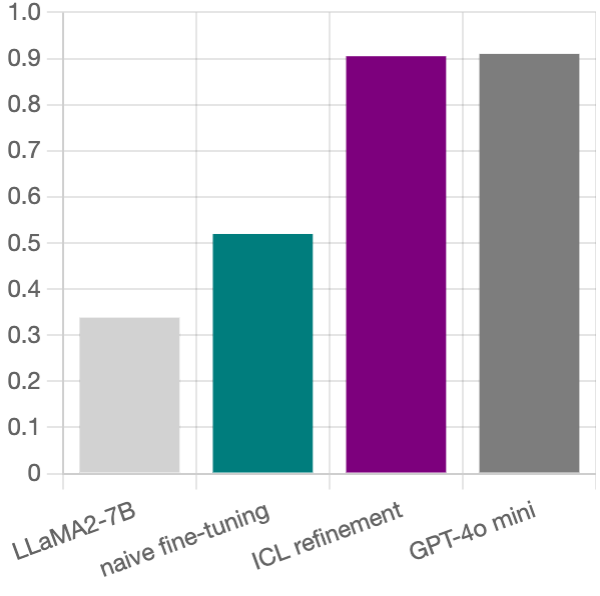


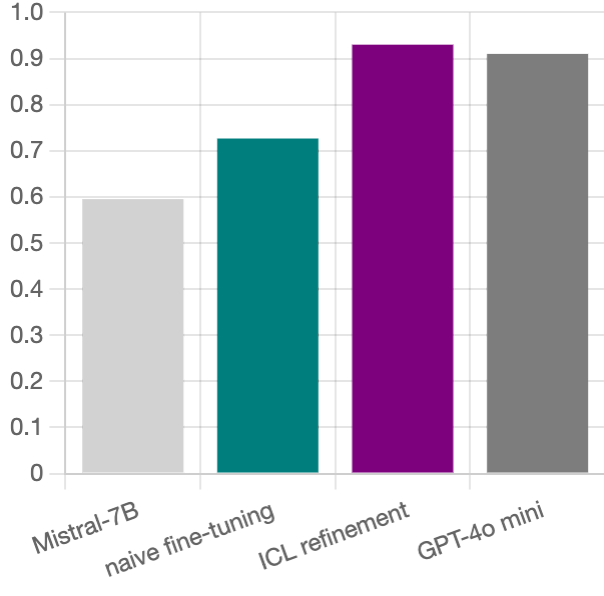




GSM8K RESULTS (GPT-4O MINI)







MATH DATASET

MATH Dataset (Ours)

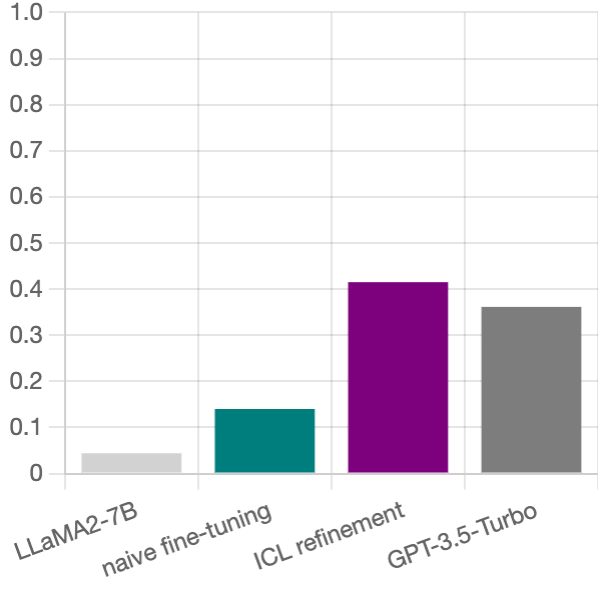
Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?
Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

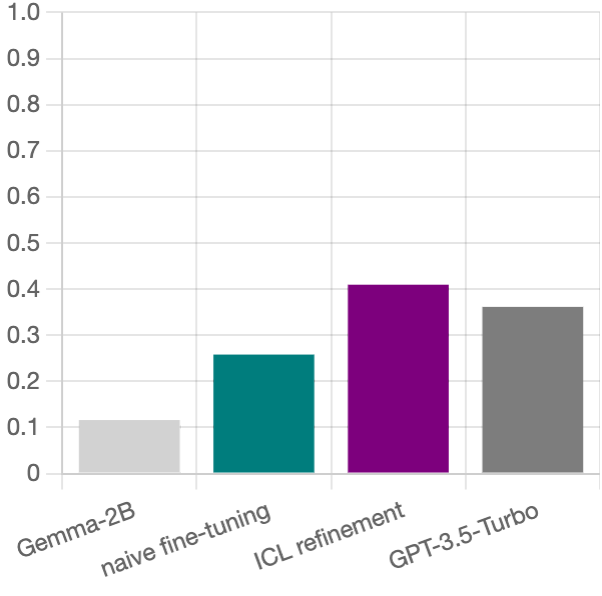
Problem: If $\sum_{n=0}^{\infty} \cos^{2n} \theta = 5$, what is $\cos 2\theta$?
Solution: This geometric series is $1 + \cos^2 \theta + \cos^4 \theta + \cdots = \frac{1}{1-\cos^2 \theta} = 5$. Hence, $\cos^2 \theta = \frac{4}{5}$. Then $\cos 2\theta = 2\cos^2 \theta - 1 = \boxed{\frac{3}{5}}$.

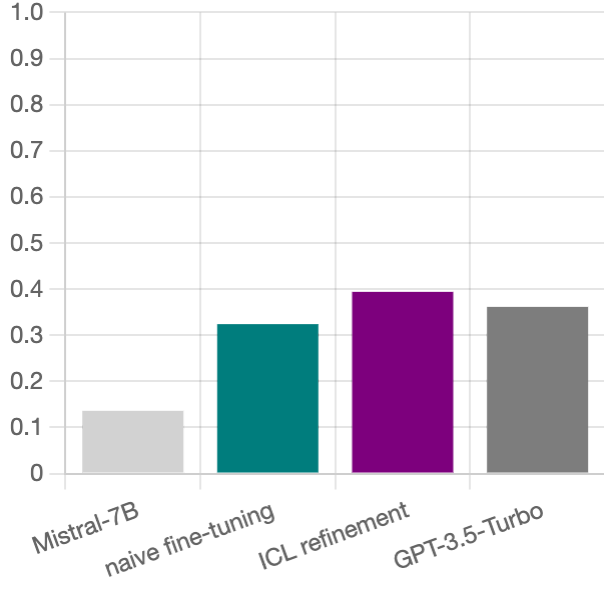
Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.
Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) =$ $1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

- 12.5k (7.5k/5k train/test) math competition problems from AMC 10, AMC 12, AMIE etc
- A CS PhD student who does not like math attained 40%.
- A 3x IMO gold medalist attained 90%.

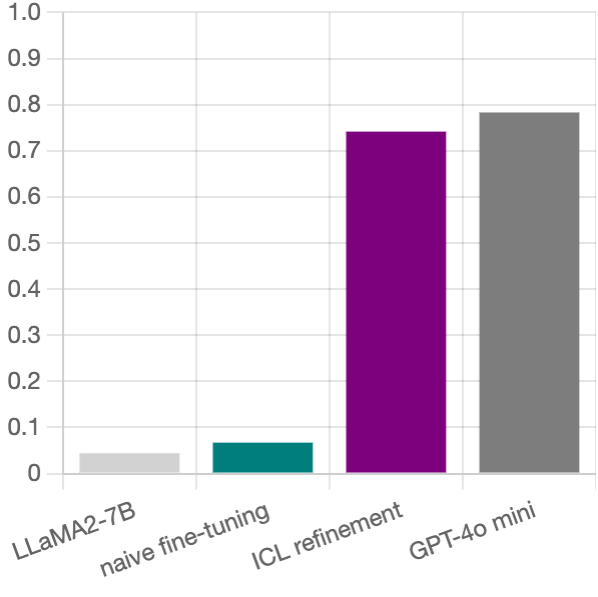
MATH RESULTS (GPT-3.5-TURBO)

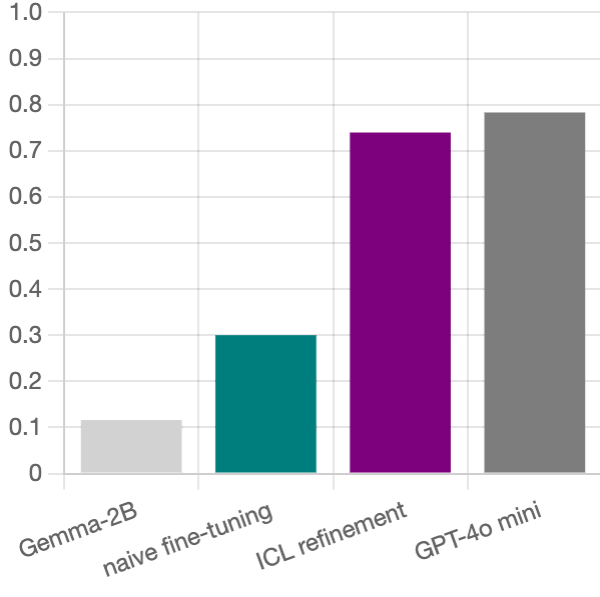


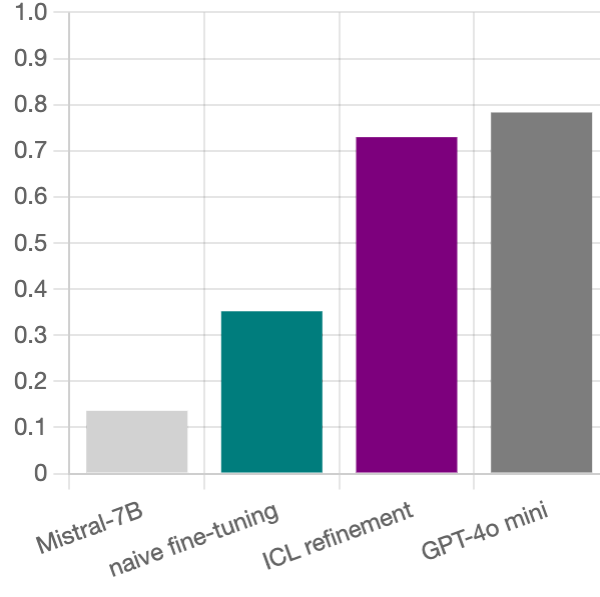




MATH RESULTS (GPT-4O MINI)

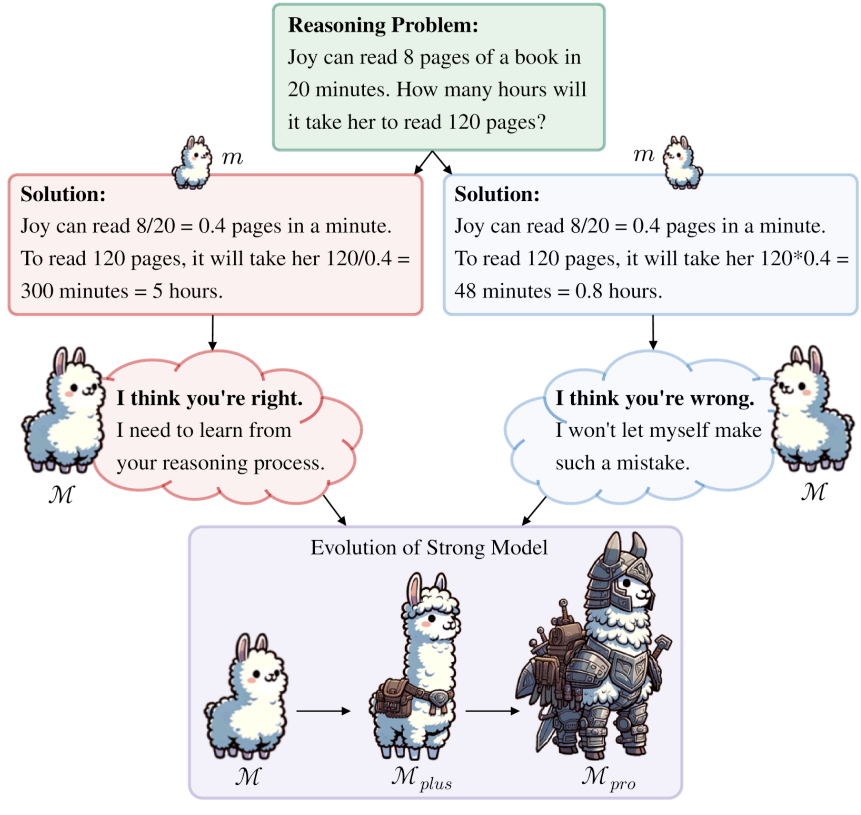






(MORE) ADVANCED REFINEMENT (AGENTS)

1. guide the pretrained LLM to the target concept with prompts
2. "constitution" + self-critique/ self-refinement
3. use the weak teacher's mistakes to teach the pretrained LLM to avoid similar mistakes
4. combinations of the above



Yang et al et al (2023)

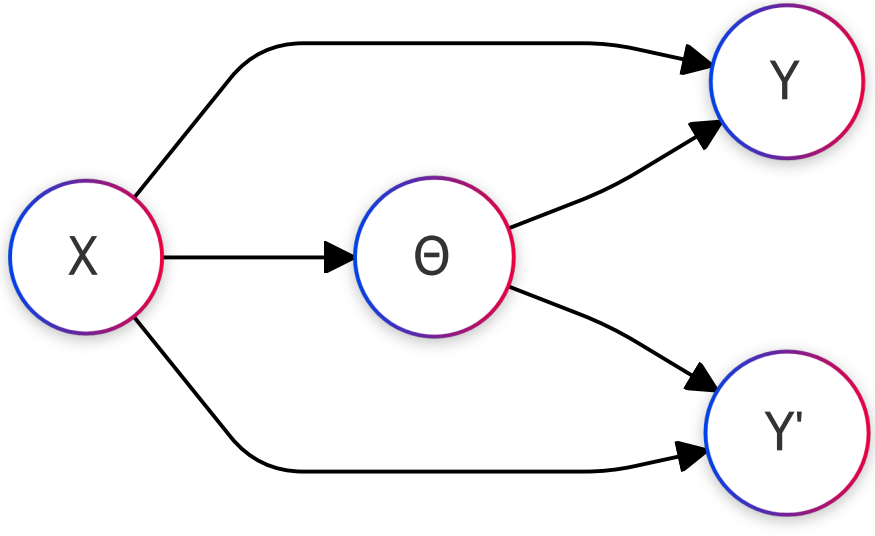
Q: Is refinement the correct high-level approach to superalignment?

LATENT CONCEPT SHIFT 2.0

Assume

1. $p(x, \theta, y, y')$ (resp. $q(x, \theta, y, y')$) factorizes as

$$p(x, \theta, y, y') = p(x)p(\theta \mid x)p(y \mid x, \theta)p(y' \mid x, \theta)$$
$$(\text{resp. } q(x, \theta, y, y') = q(x)q(\theta \mid x)q(y \mid x, \theta)q(y' \mid x, \theta)).$$



LATENT CONCEPT SHIFT 2.0

Assume

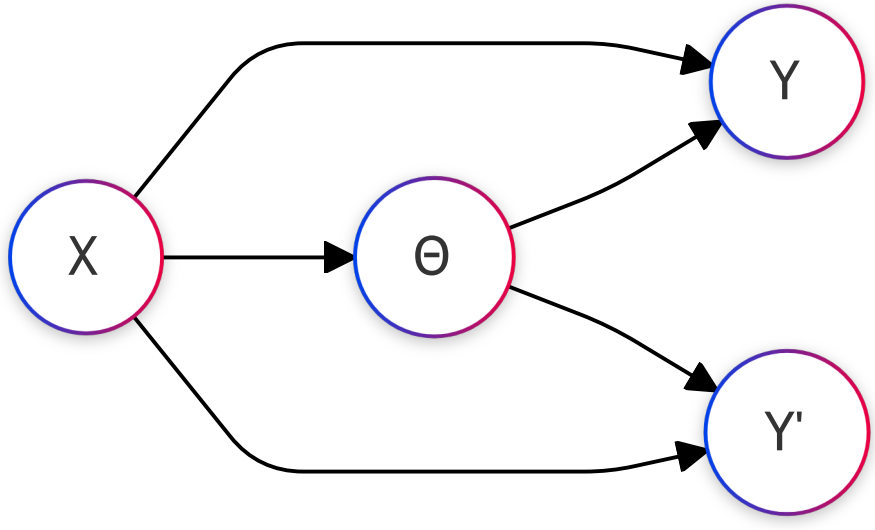
1. $p(x, \theta, y, y')$ (resp. $q(x, \theta, y, y')$) factorizes as

$$p(x, \theta, y, y') = p(x)p(\theta \mid x)p(y \mid x, \theta)p(y' \mid x, \theta)$$
$$(\text{resp. } q(x, \theta, y, y') = q(x)q(\theta \mid x)q(y \mid x, \theta)q(y' \mid x, \theta)).$$

2. $p(x) = q(x)$ and $p(y \mid x, \theta) = q(y \mid x, \theta)$; ie the only differences between P and Q are

$$p(\theta \mid x) \neq q(\theta \mid x) = \delta_{\theta_Q},$$
$$p(y' \mid x, \theta) \neq q(y' \mid x, \theta).$$

LATENT CONCEPT SHIFT 2.0

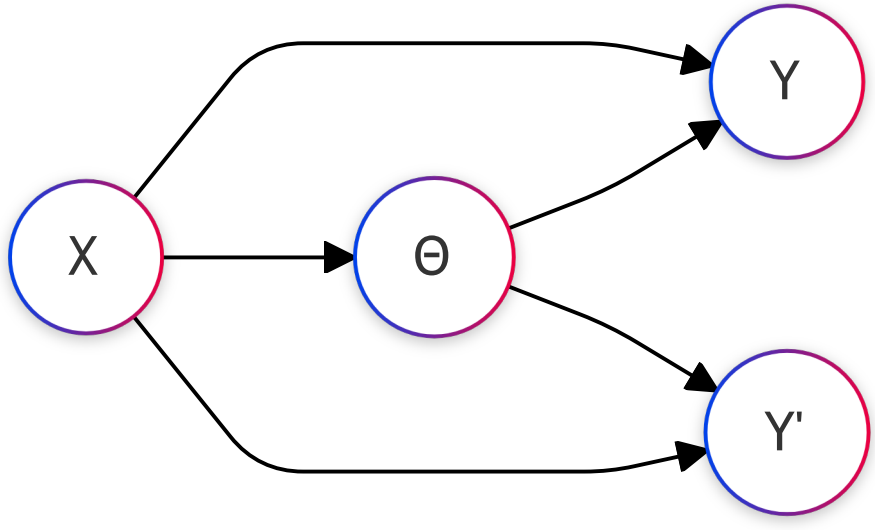


A1 and A2 imply $p(y \mid x), q(y \mid x)$ are mixtures of $p(y \mid x, \theta), \theta \in \Theta$:

$$p(y \mid x) = \sum_{\theta \in \Theta} p(y \mid x, \theta) p(\theta \mid x),$$
$$q(y \mid x) = \sum_{\theta \in \Theta} q(y \mid x, \theta) q(\theta \mid x) = q(y \mid x, \theta_Q).$$

This is exactly the latent concept shift (1.0) assumption!

LATENT CONCEPT SHIFT 2.0

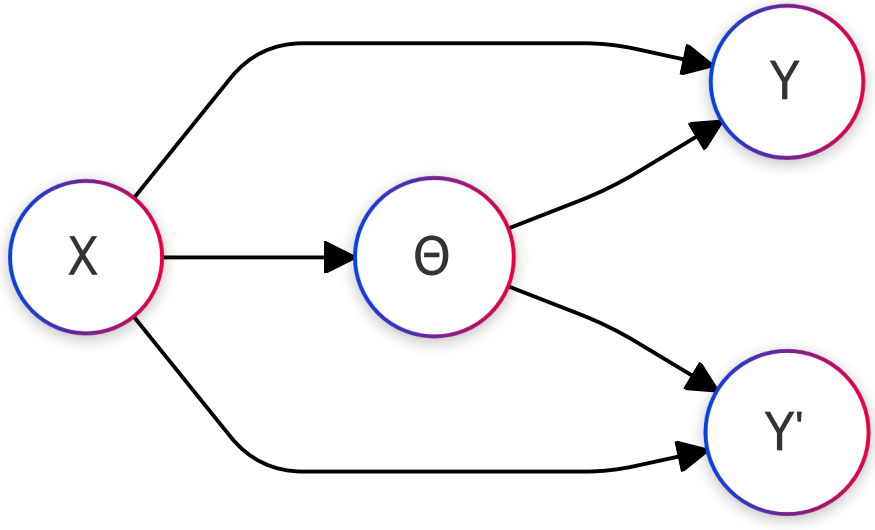


A1 and A2 imply $p(y' \mid x)$ (resp $q(y' \mid x)$) are mixtures of $p(y' \mid x, \theta)$ (resp $q(y' \mid x, \theta)$),$\theta \in \Theta$:

$$p(y' \mid x) = \sum_{\theta \in \Theta} p(y' \mid x, \theta) p(\theta \mid x),$$
$$q(y' \mid x) = \sum_{\theta \in \Theta} q(y' \mid x, \theta) q(\theta \mid x) = q(y' \mid x, \theta_Q).$$

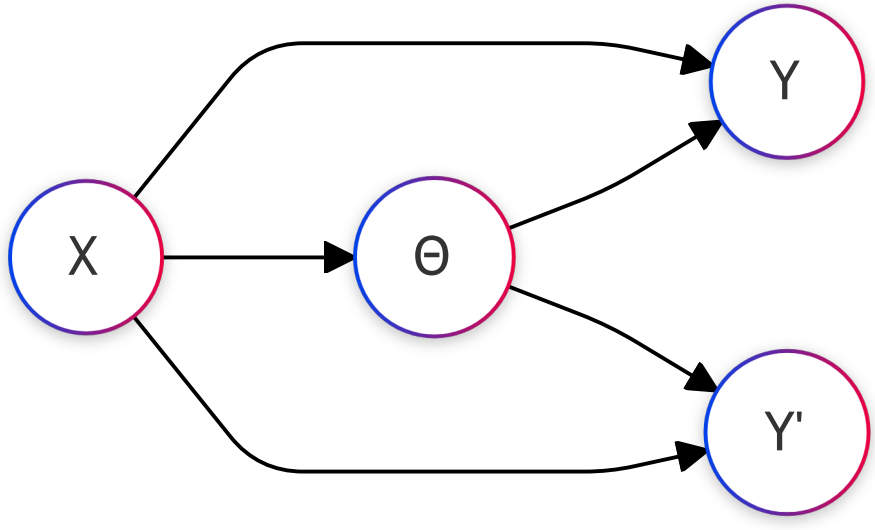
- $p(y' \mid x), q(y' \mid x)$ are **not** mixtures of the same components

LCS 2.0: THE PRETRAINED MODEL



- $p(y \mid x, \theta) = q(y \mid x, \theta), \theta \in \Theta$ encode the pretrained LLM's latent abilities
- $p(\theta \mid x)$ (resp $q(\theta \mid x)$), $\theta \in \Theta$ encode the pretrained LLM's (resp the weak teacher's) prior (distribution) on concepts

LCS 2.0: WEAK SUPERVISION



- $q(y' \mid x, \theta) \neq p(y' \mid x, \theta), \theta \in \Theta$ encode the (weak) teacher's latent abilities
- $p(y' \mid x, \theta), \theta \in \Theta$ are defined implicitly thru $p(y \mid x, y')$; we interpret them as the pretrained LLM's model of the teacher

REFINEMENT IN LATENT CONCEPT SHIFT 2.0

The analogy of ICL refinement in LCS 2.0 is (sampling from)

$$\hat{q}(y \mid x) \triangleq \int_{y'} p(y \mid x, y') q(y' \mid x).$$

ICL refinement: for $i \in [n_Q]$

1. select ICL examples $\mathcal{S}_i \mid X_{Q,i}$,
2. $\mathtt{Y_{Q,i}} \sim \mathtt{GPT(\cdot \mid S_i, X_{Q,i})}$.

refinement analogy:

1. $Y \mid X \sim q(y' \mid X)$,
2. $Y \mid X, Y' \sim p(y \mid X, Y')$.

REFINEMENT IN LATENT CONCEPT SHIFT 2.0

The analogy of ICL refinement in LCS 2.0 is (sampling from)

$$\hat{q}(y \mid x) \triangleq \int_{y'} p(y \mid x, y') q(y' \mid x).$$

Lem: Under A1, A2, and $p(y' \mid x, \theta) = q(y' \mid x, \theta)$ (ie the pretrained LLM can correct the teacher's mistakes),

$$\hat{q}(y \mid x) = \sum_{\theta \in \Theta} p(y \mid x, \theta) \int_{y'} p(\theta \mid x, y') q(y' \mid x, \theta_Q).$$

cf the source predictive distribution, which has the form

$$p(y \mid x) = \sum_{\theta \in \Theta} p(y \mid x, \theta) \int_{y'} p(\theta \mid x, y') p(y' \mid x).$$

Sanity check: Refinement always helps when the pretrained LLM can correct the teacher's mistakes!

IS REFINEMENT OPTIMAL?

Lem: Under A1 and A2,

$$\hat{q}(y \mid x) = \sum_{\theta \in \Theta} p(y \mid x, \theta) \int_{y'} p(\theta \mid x, y') q(y' \mid x, \theta_Q).$$

- interpret $\int_{y'} p(\theta \mid x, y') q(y' \mid x, \theta_Q)$ as entries of a confusion matrix (for predicting θ from y')

cf the target predictive distribution, which has the form

$$q(y \mid x) = p(y \mid x, \theta_Q).$$

✅ Yes when it is possible to exactly recover θ_Q from the weak teacher (eg when there is no overlap between the $q(y' \mid x, \theta)$'s).

❌ Refinement is often suboptimal because identifiability of $q(y \mid x)$ (from $p(x, y, y')$ and $q(x, y')$) is (much) weaker than no overlap.

TAKEAWAYS

1. The accuracy achievable by naive fine-tuning is limited by the accuracy of the weak teacher and the pretrained LLM.

2. **Main idea**: elicit the latent abilities of the pretrained LLM by using it to refine the weak teacher's outputs

3. There seems to be room for improvement!

Transfer learning for weak-to-strong generalization.
S Somerstep, F Maia Polo, M Banerjee, Y Ritov, M Yurochkin, Y Sun.
arXiv:2405.16236

This work was supported by the NSF under grants 2113364, 2113373, 2414918.