Fine-Tuning Without Forgetting In-Context Learning:
A Theoretical Analysis of Linear Attention Models

Chungpa Lee, Jy-yong Sohn, Kangwook Lee

YONSEI UNIVERSITY · KRAFTON · WISCONSIN UNIVERSITY OF WISCONSIN–MADISON · ludo robotics

Fine-tuning language models can degrade their in-context learning (ICL) ability, limiting their ability to generalize to tasks not seen during fine-tuning. **To understand why fine-tuning improves zero-shot performance on the target task but can degrade ICL**, we analyze this phenomenon through linear attention models. We derive **closed-form optimal parameters and test errors for three fine-tuning strategies: (b)** full fine-tuning, **(c)** value-matrix fine-tuning, and **(d)** value-matrix fine-tuning with an auxiliary few-shot loss.

Problem Setup

A linear self-attention model.

$$f(Z;V,Q)=\left[Z+\frac{1}{\mathrm{ncol}(Z)}\cdot VZZ^\top QZ\right]_{-1,-1}$$

where $[A]_{-1,-1}$ denotes the final output coordinate.

- $V\in\mathbb{R}^{(d+2)\times(d+2)}$: value matrix
- $Q\in\mathbb{R}^{(d+2)\times(d+2)}$: query-key matrix

Linear regression task.

Each task is indexed by a task vector $\theta\in\mathbb{R}^d$. For each task θ,

$$x\sim\mathcal{N}(0,\Sigma)\ \text{ and }\ y\sim\mathcal{N}(\theta^\top x,\sigma^2)\ .$$

Given n in-context examples and a query input x_{n+1}, the model $f(\cdot)$ predicts y_{n+1} from the prompt $[Z_{[n]}\ z_{n+1}]$, where

$$Z_{[n]}=\begin{bmatrix}1&\cdots&1\\x_1&\cdots&x_n\\y_1&\cdots&y_n\end{bmatrix},\qquad z_{n+1}=\begin{bmatrix}1\\x_{n+1}\\0\end{bmatrix}.$$

Evaluation metric.

We use mean squared error on the test set.

- n-shot error on a task θ :

$$\mathscr{E}(V,Q,n;\theta):=\mathbb{E}_{x_1,y_1,\cdots,x_{n+1},y_{n+1}|\theta}\left[\left(f\left([Z_{[n]}\ z_{n+1}];V,Q\right)-y_{n+1}\right)^2\right]$$

- Zero-shot error on a task θ :

$$\mathscr{E}(V,Q,0;\theta):=\mathbb{E}_{x_{n+1},y_{n+1}|\theta}\left[\left(f(z_{n+1};V,Q)-y_{n+1}\right)^2\right]$$

Training objectives.

- **(a)** Pretrained model exhibits in-context learning across tasks:

$$\hat{V},\hat{Q}:=\mathrm{argmin}_{V,Q}\,\mathbb{E}_{\theta\sim\mathcal{N}(0,I)}\left[\mathscr{E}(V,Q,n;\theta)\right]$$

- **(b)** Full fine-tuning updates both V and Q on the target task θ_0:

$$\mathrm{argmin}_{V,Q}\ \mathscr{E}(V,Q,0;\theta_0)$$

- **(c)** Value-matrix fine-tuning updates only V, while $\hat{Q}$ is fixed:

$$\mathrm{argmin}_{V}\ \mathscr{E}\left(V,\hat{Q},0;\theta_0\right)$$

- **(d)** Value-matrix fine-tuning with an auxiliary few-shot loss:

$$\mathrm{argmin}_{V\in\mathbb{S}}\,\mathscr{E}\left(V,\hat{Q},n;\theta_0\right),\quad \mathbb{S}=\mathrm{argmin}_{V}\,\mathscr{E}\left(V,\hat{Q},0;\theta_0\right)$$

Summary of Theoretical Results

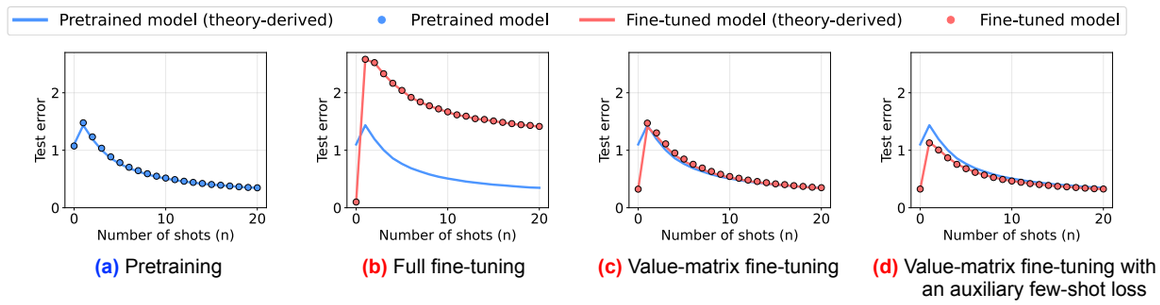
We report the test errors of **(a) the pretrained model** and **(b–d) fine-tuned models**, each optimized with a different strategy.

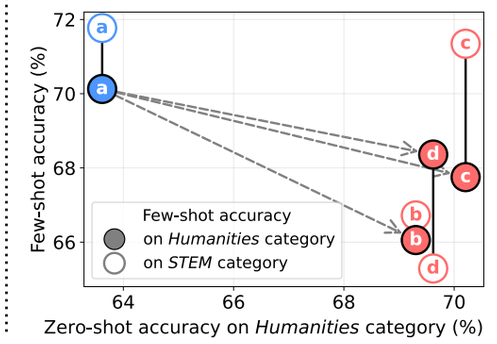
	Training Regime	Zero-shot error on θ_0 (in-distribution)	Few-shot error on θ_0 (in-distribution)	Few-shot error on $-\theta_0$ (out-of-distribution)	Results
(a)	Pretraining	$\sigma^2+\theta_0^\top\Sigma\theta_0$ (**worst**)	σ^2 (**best**)	σ^2 (**best**)	Corollary 4.1 Corollary 4.2
(b)	Fine-tuning **all parameters** with the **zero-shot** loss	σ^2 (**best**)	$\sigma^2+\theta_0^\top\Sigma\theta_0$ (**worst**)	$\sigma^2+\theta_0^\top\Sigma\theta_0$ (**worst**)	Theorem 4.3 Corollary 4.4
(c)	Fine-tuning the **value matrix** with the **zero-shot** loss	$\sigma^2+\frac{2}{d+4}\theta_0^\top\Sigma\theta_0$ (better)	$\sigma^2+\frac{1}{(d+4)^2}\theta_0^\top\Sigma\theta_0$ (better)	$\sigma^2+\frac{1}{(d+4)^2}\theta_0^\top\Sigma\theta_0$ (better)	Theorem 4.6 Proposition 4.7 Corollary 4.8
(d)	Fine-tuning the **value matrix** with the **zero-shot** and **few-shot** losses	$\sigma^2+\frac{2}{d+4}\theta_0^\top\Sigma\theta_0$ (better)	σ^2 (**best**)	$\sigma^2+\frac{4}{(d+4)^2}\theta_0^\top\Sigma\theta_0$ (worse)	Theorem 4.9 Proposition 4.10

Zero-shot errors: evaluated on the target task θ_0 used for fine-tuning.

Few-shot errors: evaluated on both the target task θ_0 and an out-of-distribution task $-\theta_0$, with infinitely many few-shot examples.

Experimental Results





Left: Linear attention models on linear regression tasks. Curves show theoretical predictions, and points show empirically trained models.

Right: Zero-shot (x-axis) and few-shot (y-axis) accuracy of *Qwen2.5-3B-Instruct* on *MMLU* (*Humanities* and *STEM* categories).
Strategy (b) improves zero-shot accuracy but degrades few-shot accuracy.
Strategy (c) preserves few-shot accuracy with comparable zero-shot gains.
Strategy (d) further improves few-shot accuracy on *Humanities* but reduces few-shot accuracy on *STEM*.

Takeaways

Fine-tuning only the value matrix improves zero-shot accuracy while preserving in-context (few-shot) learning.

⇒ **Use this strategy when the goal is to preserve general in-context learning ability.**

An auxiliary few-shot loss improves the target task but can hurt unseen tasks, with degradation growing as task dissimilarity increases.

⇒ **Use this strategy when the goal is to preserve in-domain in-context learning ability.**