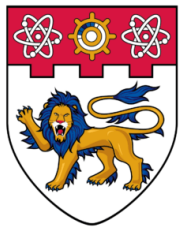
When More Experts Hurt: Underfitting in Multi-Expert Learning to Defer

Shuqi Liu[1*] Yuzhou Cao[1*] Lei Feng[2] Bo An[1] Luke Ong[1]

[1] Nanyang Technological University
[2] Southeast University

ICML 2026

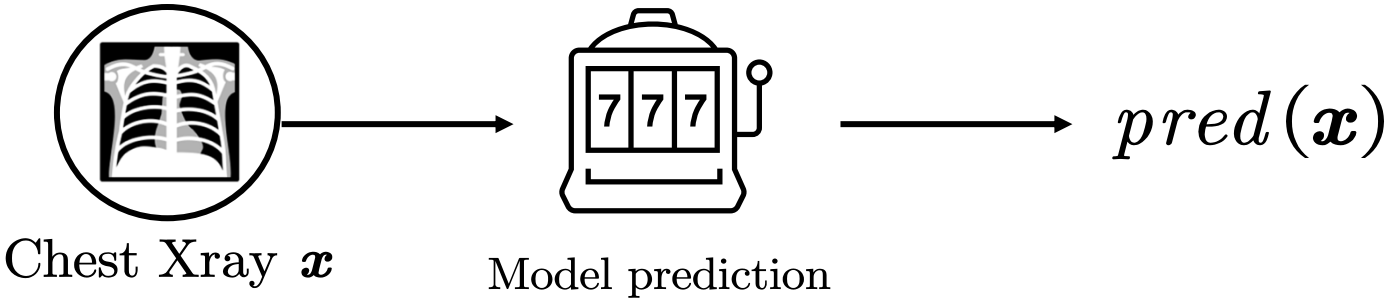


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

東南大學 SOUTHEAST UNIVERSITY

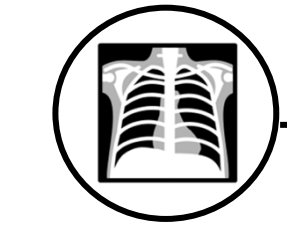
Standard Classification System

- Single model



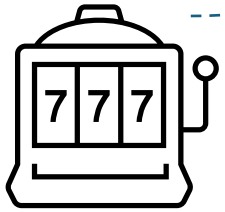
Deferral System: Human vs AI

- Single-Expert L2D



Chest Xray $\boldsymbol{x}$

L2D System



Model prediction

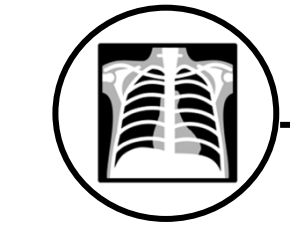
How to predict?

$pred(\boldsymbol{x})$

When to defer?

Deferral System: Human vs AI

- Multiple-Expert L2D



Chest Xray $\boldsymbol{x}$

L2D System

Model prediction

$pred(\boldsymbol{x})$

How to predict?

When to defer?

Which to defer?

L2D Problem Formulation

- Train a system $f : \mathcal{X} \to \mathcal{Y} \cup \{\perp_1, \cdots, \perp_J\}$ to leverage experts' prediction, where $\perp_j$ refers to the option of deferring to the j-th expert, using the training dataset drawn from $\mathcal{X} \times \mathcal{Y} \times \mathcal{M}^J$
- Learning objective:

$$\min_f \mathbb{E}_{p(\boldsymbol{x},y,\boldsymbol{m})}[\underbrace{\mathbb{I}(f(\boldsymbol{x}) \neq y)\mathbb{I}(f(\boldsymbol{x}) \in \mathcal{Y})}_{\text{Classifier error}} + \underbrace{\sum_{j=1}^{J} \mathbb{I}(m_j \neq y)\mathbb{I}(f(\boldsymbol{x}) = \perp_j)}_{\text{Expert error}}$$

- Bayes optimal solution:

$$f^*(\boldsymbol{x}) = \begin{cases} \perp_{j^*}, & \text{if } \mathrm{Acc}_{j^*}(\boldsymbol{x}) \geq \eta_{y^*}(\boldsymbol{x}), \\ \mathrm{argmax}_{y \in \mathcal{Y}}\ p(y|\boldsymbol{x}), & \text{else,} \end{cases}$$

Underfitting Issues in L2D

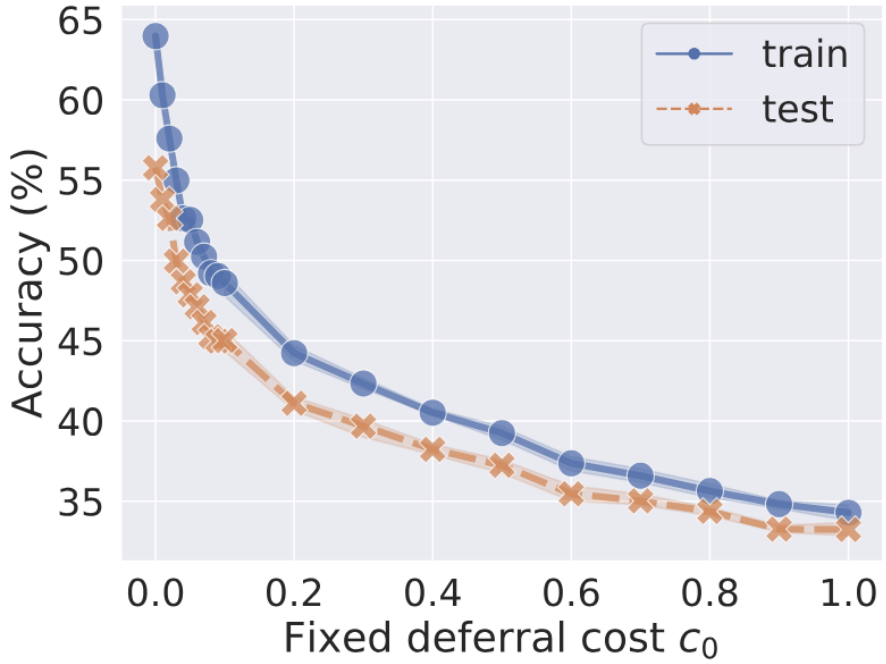
- **Special Case**: Single-expert with extra non-zero deferral cost $c > 0$

$$\min_f \mathbb{E}_{p(\boldsymbol{x},y,m)}\left[\mathbb{I}[f(\boldsymbol{x}) \neq y]\mathbb{I}[f(\boldsymbol{x}) \neq \perp] + (c + \mathbb{I}[m \neq y])\mathbb{I}[f(\boldsymbol{x}) = \perp]\right]$$

Extra deferral cost

Underfitting Issues in L2D

- **Special Case**: $J = 1, c > 0$
- **Phenomena:** classifier's underfitting issue observed by [1]



[1] Narasimhan, H., Jitkrittum, W., Menon, A. K., Rawat, A. S., and Kumar, S. Post-hoc estimators for learning to defer to an expert. In NeurIPS, 2022.

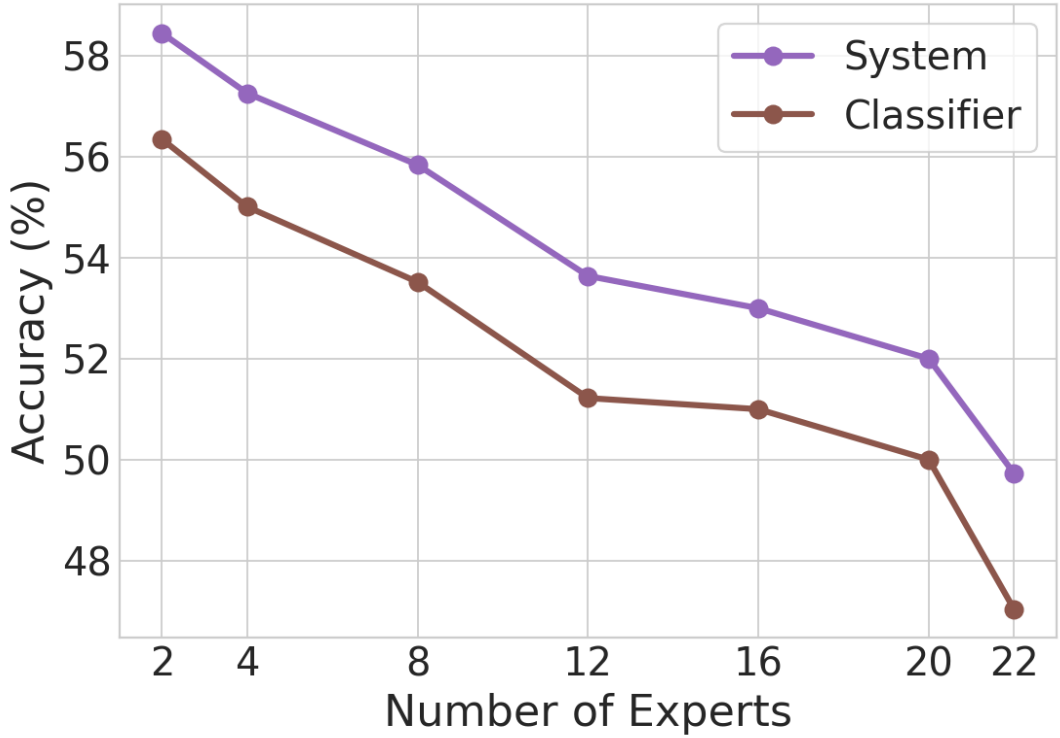
Underfitting Issues in L2D

- **Special Case**: $J = 1, c > 0$
- **Cause**: redundant label smoothing term induced by extra deferral cost
- **Existing Solutions**: target to remove the redundant cost-induced term

No extra deferral costs, no underfitting

Multiple Experts Change the Story

- Unexpected Observation: **Underfitting persists even when $c = 0$**



Intuition: More experts should not hurt performance

Observation: More experts ⇒ worse performance

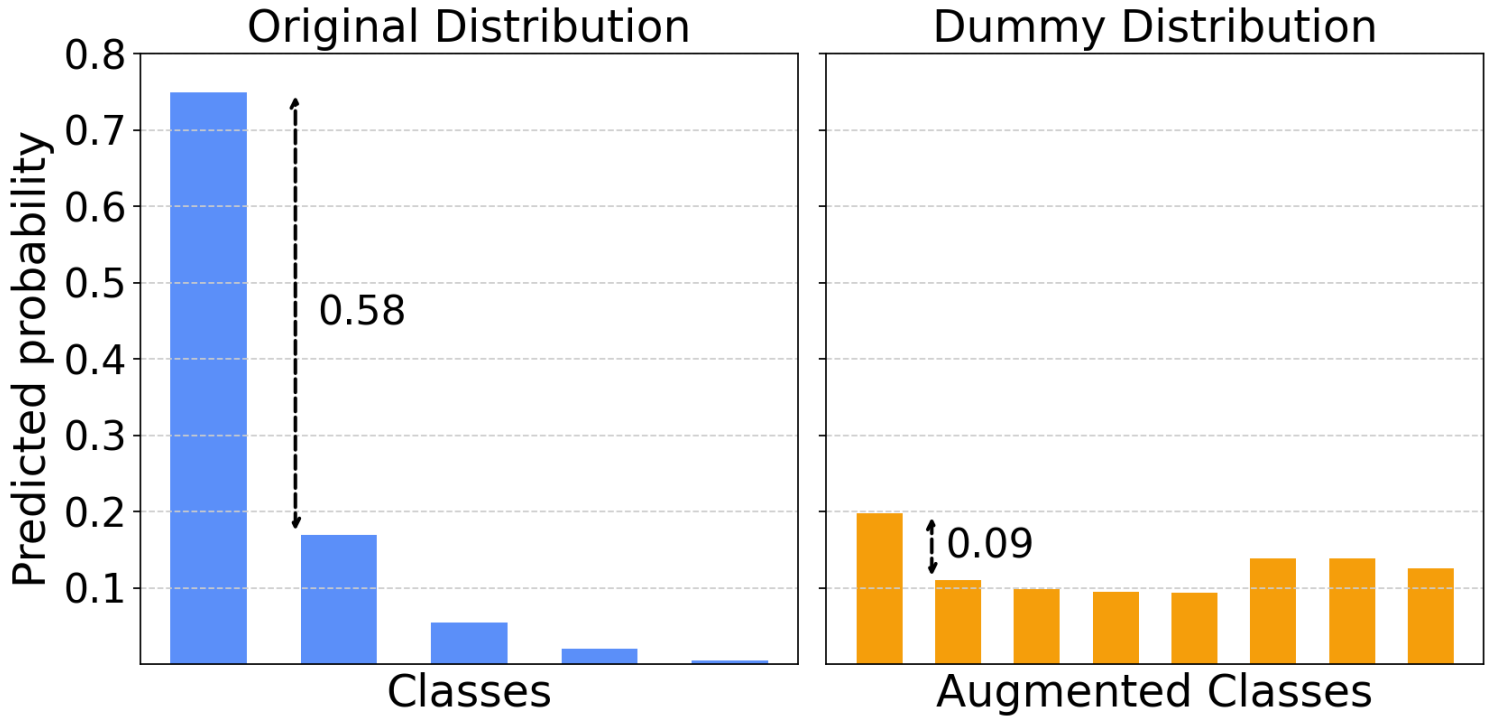
A New Underfitting Challenge in Multiple-Expert L2D

$$\mathcal{A}(\boldsymbol{x}) = \sum_{j=1}^{J} \mathrm{Acc}_j(\boldsymbol{x})$$

Expert aggregation term

- **Cause Analysis:** degraded distribution

$$\widehat{p}(y|\boldsymbol{x}) = \frac{\eta_y(\boldsymbol{x})}{1 + \mathcal{A}(\boldsymbol{x})} \qquad \widehat{p}(\perp_j|\boldsymbol{x}) = \frac{\mathrm{Acc}_j(\boldsymbol{x})}{1 + \mathcal{A}(\boldsymbol{x})}$$



A New Underfitting Challenge in Multiple-Expert L2D

- **Cause**: Expert Aggregation Term
- **Potential Solution**: Use the intermediate learning results [2] to remove the expert aggregation term

$$\widetilde{\ell}^{\circ}_{\phi}(\boldsymbol{\theta}, y, \boldsymbol{m}) := \phi(\boldsymbol{\theta}, y) + \mathbb{1}[m_{\widehat{j}^*} = y]\phi(\boldsymbol{\theta}, \widehat{j}^* + K),$$

But the indicator term introduces discontinuity.

[2] Liu, S., Cao, Y., Zhang, Q., Feng, L., and An, B.
Mitigating underfitting in learning to defer with consistent losses. In AISTATS, 2024

Our Motivation

- Discontinuity Analysis: $\widetilde{\ell}^{\circ}_{\phi}(\boldsymbol{\theta}, y, \boldsymbol{m}) := \phi(\boldsymbol{\theta}, y) + \mathbb{1}[m_{\widehat{j}^*} = y]\phi(\boldsymbol{\theta}, \widehat{j}^* + K),$

$$\mathbb{I}(m_{j^*} = y) = 0 \text{ or } 1, \text{ where } j^* = \operatorname{argmax}_{j \in [J]} \theta_{j+K}$$

- Our Idea: **constraint the candidate expert set**

$$\mathbb{I}(m_{j^*} = y) = 1, \text{ where } j^* = \operatorname{argmax}_{j \in \{j \in [J]: m_j = y\}} \theta_{j+K}$$

Our Solution: PiCCE

- PiCCE (**Pi**ck the **C**onfident and **C**orrect **E**xpert): using both intermediate and empirical results.

$$\widetilde{\ell}_{\phi}(\boldsymbol{\theta}, y, \boldsymbol{m}) = \phi(\boldsymbol{\theta}, y) + \phi\Big(\boldsymbol{\theta}, \underset{j \in \{j \in [J]: m_j = y\}}{\operatorname{argmax}} \theta_{j+K} + K\Big).$$

- Continuity of PiCCE: Theorem 4.2

  The proposed formulation is continuous if ϕ is continuous and is symmetric w.r.t its last J inputs.

Theoretical Results

- Consistency Guarantee: Theorem 5.2

Theorem 5.2 (Consistency and Expert Accuracy Estimator). *When Condition 1 holds and (12) and (13) is used as multiclass loss ϕ in (9), for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta}^*$ minimizes (10). the prediction link φ defined in (5) and $\boldsymbol{\theta}^*$ reproduces the Bayes optimal decision $f^*(\boldsymbol{x})$, i.e.: $\varphi(\boldsymbol{\theta}^*) = f^*(\boldsymbol{x})$, and $\operatorname{Argmax}_{j\in[J]}\theta^*_{j+K} = \operatorname{Argmax}_{j\in[J]}\operatorname{Acc}_j(\boldsymbol{x}) = \{j^*\}$. Furthermore:*

(A). When (12) is used, $u^*_{j^*} = \operatorname{Acc}_{j^*}(\boldsymbol{x})\widetilde{V}(\boldsymbol{x})$.

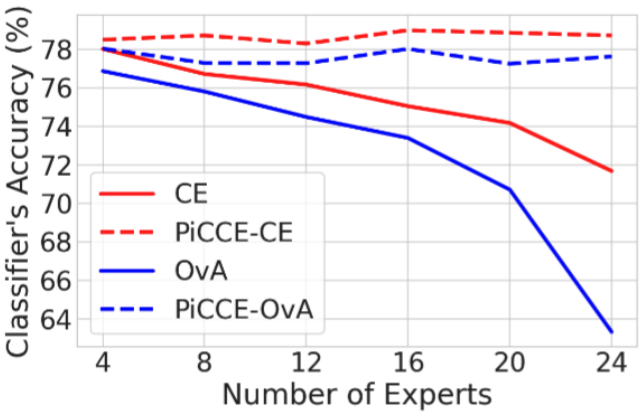
(B). When (13) is used, $s(\theta^*_{j^*}) = \operatorname{Acc}_{j^*}(\boldsymbol{x})$.

- Underfitting Resistance: Lemma 4.4
- Finite-sample Guarantee: Theorem E.2.

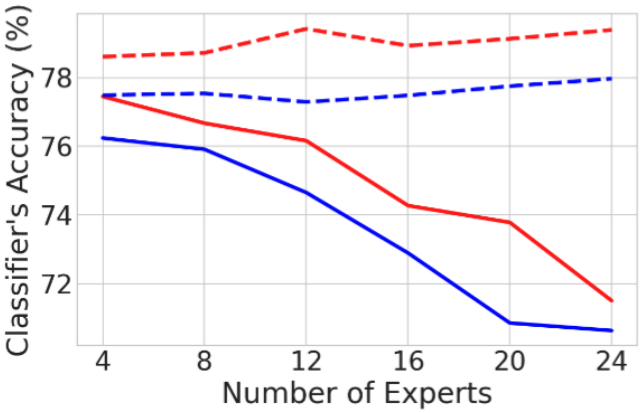
Experiment Results on CIFAR-100

- System Error in Table
- Classifier Error in Figure

Dataset		CIFAR-100							
Expert Pattern		Animal Expert				Overlapped Animal Expert			
Loss Formulation		Vanilla		PiCCE		Vanilla		PiCCE	
Method	#Exp	Err	Cov	Err	Cov	Err	Cov	Err	Cov
CE	4	18.48	74.58	**18.32**	**77.50**	16.10	64.50	**15.10**	**67.80**
	8	18.91	74.35	**18.21**	**76.35**	16.24	64.59	**16.10**	**71.35**
	12	19.08	75.18	**18.19**	**77.43**	16.99	66.09	**15.84**	**69.86**
	16	19.14	76.39	**18.16**	**79.14**	18.72	67.21	**15.61**	**73.62**
	20	21.13	73.57	**18.11**	**78.33**	19.22	69.31	**15.58**	**73.95**
OvA	4	19.46	83.72	**19.10**	**86.43**	17.07	79.28	**16.58**	**83.85**
	8	20.09	83.30	**18.98**	**86.83**	18.03	79.69	**17.71**	**83.45**
	12	20.70	82.67	**18.86**	**86.89**	18.45	79.36	**17.77**	**84.93**
	16	21.84	83.09	**18.70**	**87.11**	19.85	77.30	**17.76**	**85.02**
	20	22.91	77.71	**18.63**	**86.69**	20.95	72.72	**17.74**	**86.19**



(a) CIFAR-100: Animal Expert



(d) CIFAR-100: Overlapped Animal Expert