OTTER: Effortless Label Distribution Adaptation of Zero-shot Models

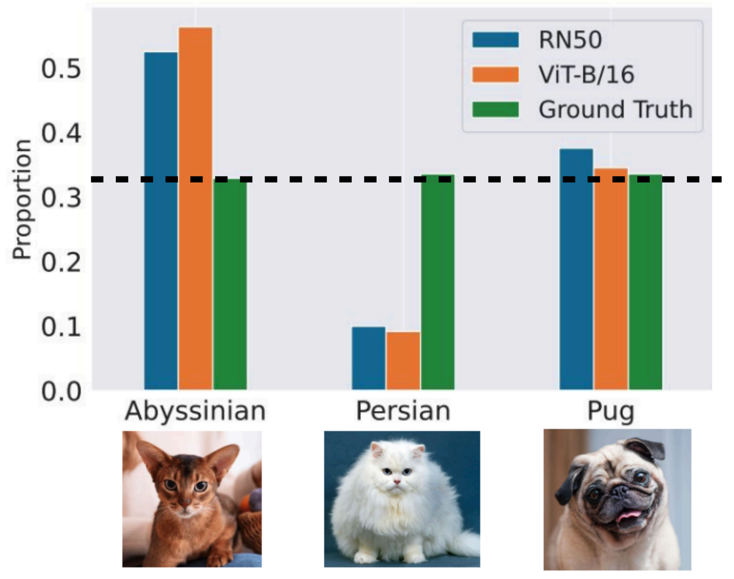
Changho Shin, Jitian Zhao, Sonia Cromp, Harit Vishwakarma, Frederic Sala

University of Wisconsin-Madison

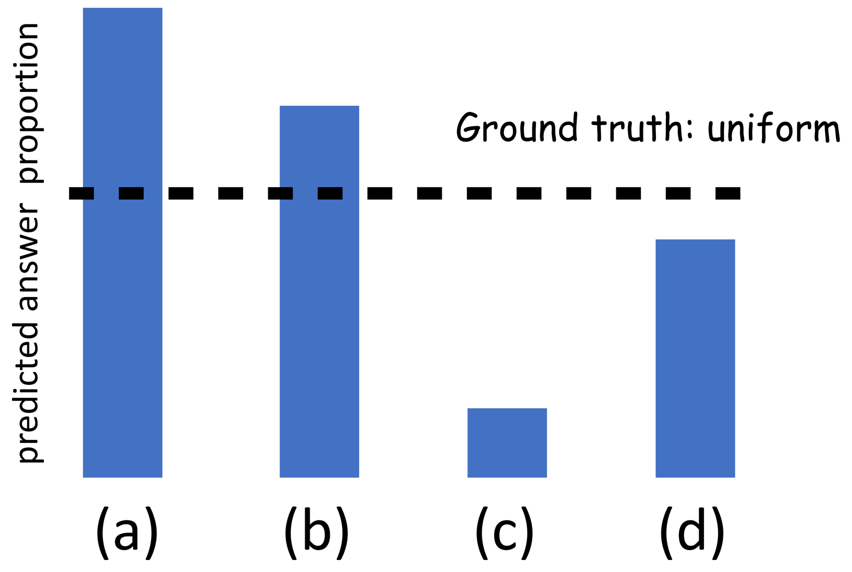
Paper Link

Background & Motivation

- Foundation models often suffer from the **biases introduced during pretraining.**
- **Label distributions** are inevitably mismatched in the downstream tasks, **degrading model performance.**



CLIP Image Classification



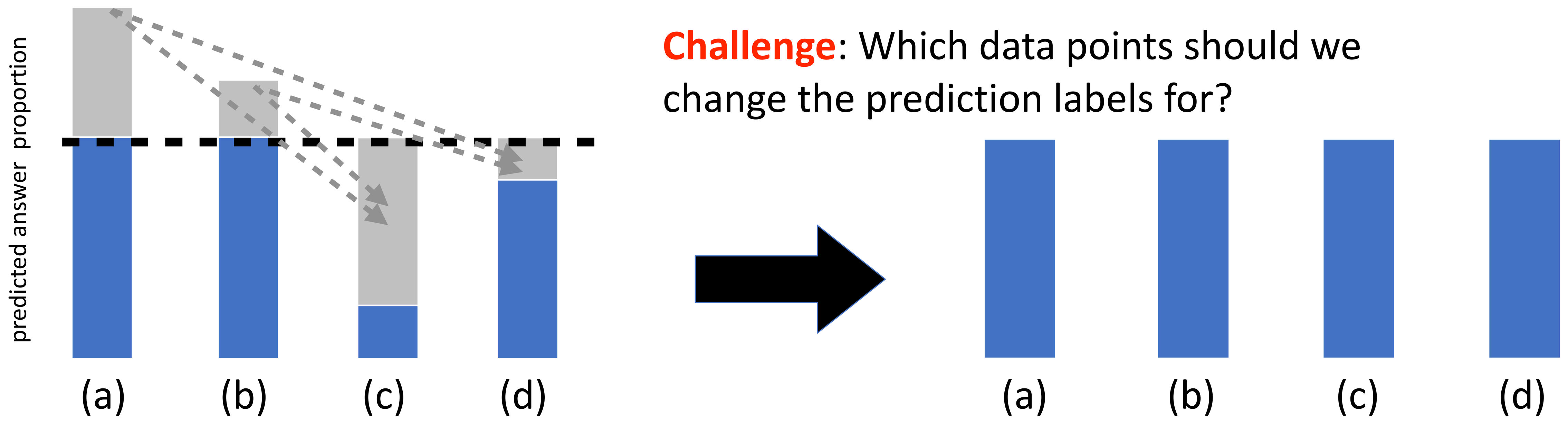
LLM Multiple Choice Questions

- Can we adapt model predictions to specified label distributions **instantly?**
- We introduce **OTTER**, an **Optimal Transport**-based method to adjust model predictions according to **specified label distributions.**

Idea

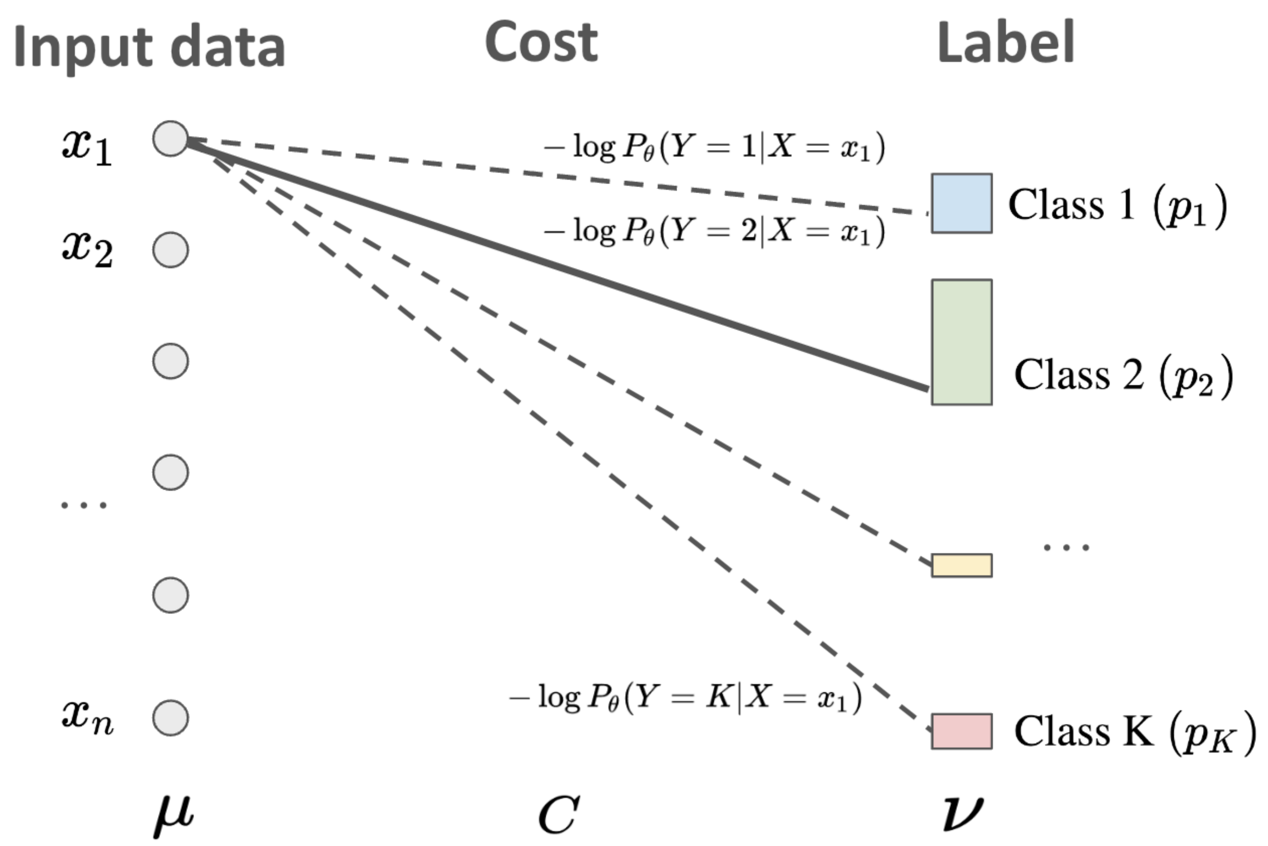
Redistribute model predictions **to match specified label distribution**, while still considering prediction scores of each point.

Challenge: Which data points should we change the prediction labels for?



Method: Optimal Transport for Classification

Solve **Optimal Assignment Problem** with costs as negative prediction scores.



$$\pi = \arg\min_{\gamma\in\Pi(\mu,\nu)} \langle \gamma, C\rangle,$$

where $\Pi(\mu,\nu) = \{\gamma \in \mathbb{R}_+^{n\times K} | \gamma\mathbf{1} = \mu, \gamma^T\mathbf{1} = \nu\}$

$\mu = \frac{1}{n}\mathbf{1}$ Mass of each point

$\nu = (p_1, p_2, \dots, p_K)$ Specified label distribution

$C_{ij} = -\log P_\theta(Y=j|X=x_i)$ Negative prediction scores

Theoretical Results

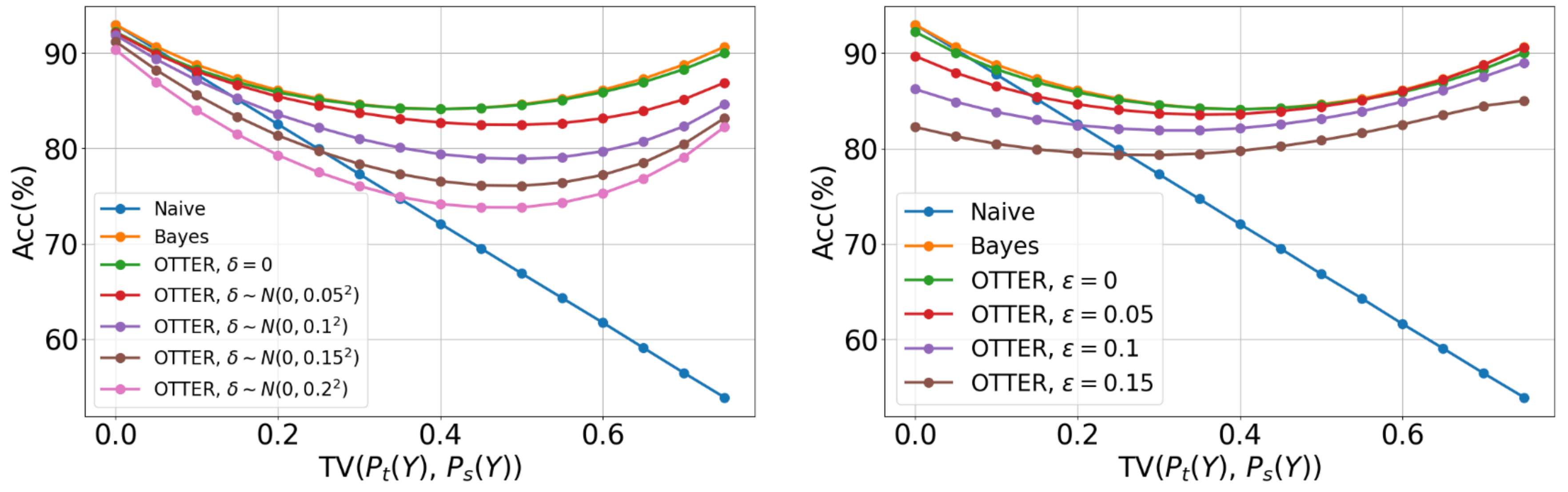
✓ OTTER ensures **consistency with zero-shot predictions.**

✓ Under label shift, OTTER with true label distribution works as a **Bayes-optimal classifier.**

✓ Its additional error beyond Bayes error decomposes into the errors in **the specified label distribution and calibration.**

Experiments

Synthetic Experiments



(a) Prediction accuracy changes with perturbed confidence score.

(b) Prediction accuracy changes with perturbed label distribution.

- **OTTER achieves the Bayes error rate** in noise-free scenarios.
- Additional error arises from **the noise in label distribution specification and calibration.**

Image & Text Classification

	Zero-shot	Prior Matching	OTTER		Zero-shot	Prior Matching	OTTER
CIFAR10	88.3	91.3 ($\pm$0.0)	**91.7**	Oxford-IIIT-Pet	83.8	82.0 ($\pm$0.3)	**88.8**
CIFAR100	63.8	64.1 ($\pm$2.7)	**67.9**	Stanford-Cars	55.7	39.8 ($\pm$2.6)	**59.7**
Caltech101	79.8	59.3 ($\pm$15.4)	**88.7**	STL10	98.0	98.4 ($\pm$0.0)	**98.6**
Caltech256	79.8	9.5 ($\pm$1.5)	**87.0**	SUN397	47.1	6.7 ($\pm$1.6)	**54.1**
Country211	19.8	19.0 ($\pm$0.1)	**21.1**	CUB	46.0	40.4 ($\pm$0.0)	**50.4**
Amazon review	74.0	58.8 ($\pm$46.4)	**91.7**	GenderBias	84.1	41.4 ($\pm$39.6)	**91.9**
CivilComments	48.4	57.2 ($\pm$37.7)	**81.4**	HateXplain	30.9	31.3 ($\pm$3.3)	**34.3**

- Achieves **significant performance improvements.**

LLM MCQ Selection Bias Mitigation

	ARC-Challenge				CommonsenseQA (CSQA)			
	Naive		OTTER		Naive		OTTER	
Model	Acc.(↑)	RStd (↓)	Acc.	RStd	Acc.	RStd	Acc.	RStd
Llama-2-7b	36.0	27.4	**45.5**	**1.7**	31.9	28.4	**42.7**	**3.8**
Llama-2-13b	**62.9**	6.0	62.8	**1.5**	57.0	10.2	**58.1**	**2.0**
Llama-2-7b-chat	56.5	12.4	**57.4**	**1.3**	56.5	15.2	**60.4**	**3.5**
Llama-2-13b-chat	64.4	13.7	**66.2**	**2.5**	64.0	9.8	**66.4**	**1.8**
vicuna-7b	53.5	8.6	**54.1**	**2.3**	56.9	8.3	**57.6**	**1.0**
vicuna-13b	62.9	8.3	**63.3**	**2.4**	63.4	12.9	**64.5**	**3.1**

- **Mitigates selection bias** and enhances **accuracy**.