Understanding prompt engineering does not require rethinking generalization

Victor Akinwande[1], Yiding Jiang[1], Dylan Sam[1], Zico Kolter[1,2]

ICLR 2024.

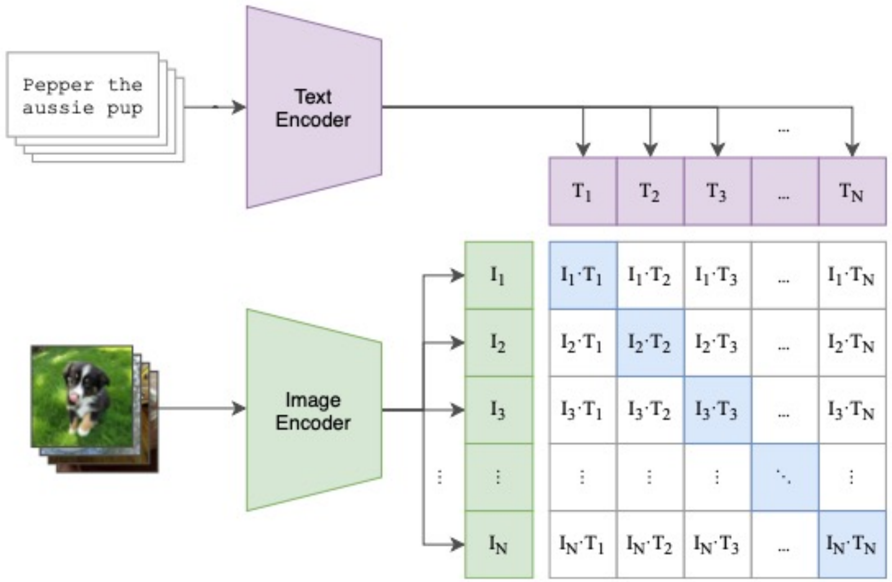
[1]Carnegie Mellon University
[2]Bosch Center for AI

Key Takeaway

- The hypothesis space of visual-language models is relatively small – prompt engineering or greedy search of the set of tokens will not overfit.
- Uniform convergence or PAC-Bayes bounds on the discrete space of natural language tokens are remarkably tight and useful for model selection.

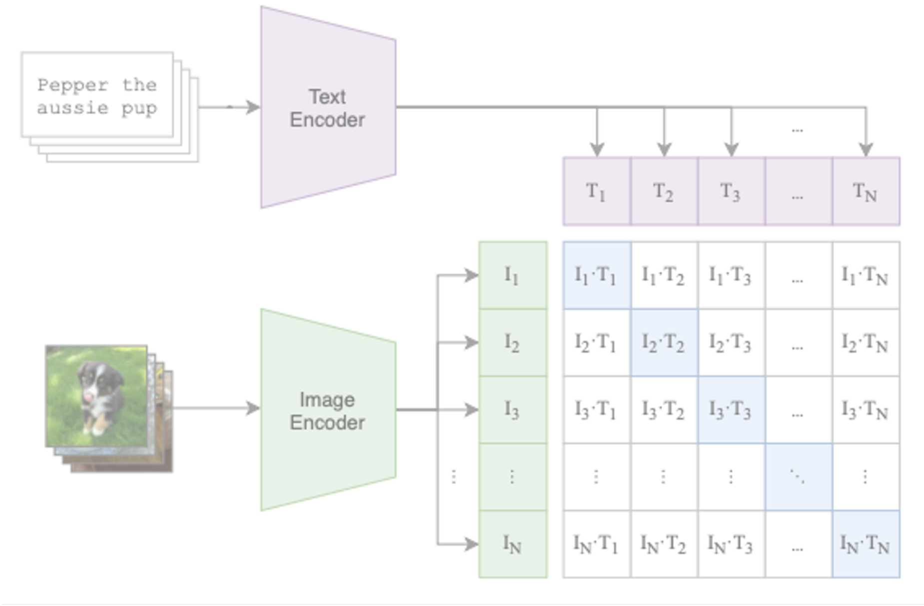
Background: Learning models from natural language supervision

- Obtaining labels is expensive – can we use the vast amount of natural language on the web to build vision models?

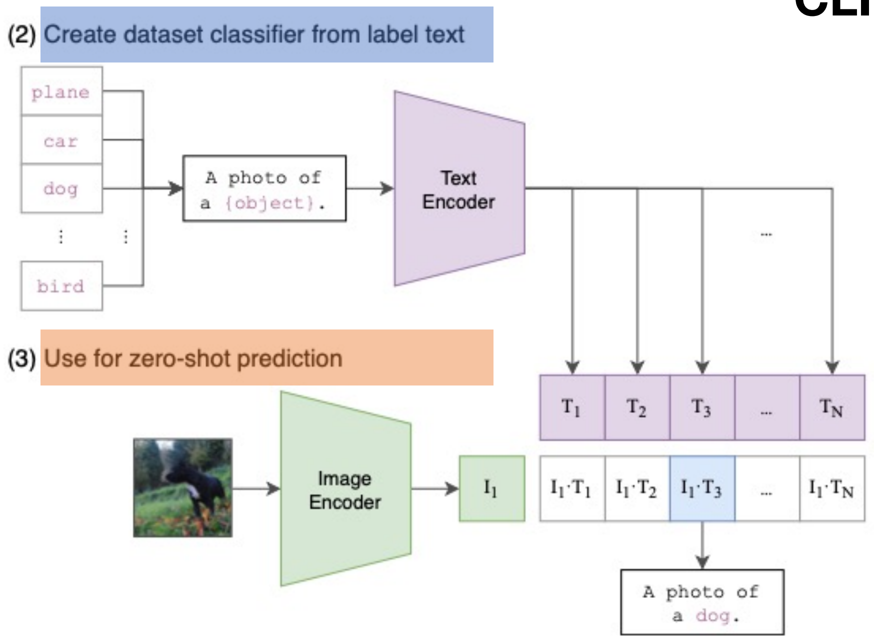


Background: Learning models from natural language supervision

- Obtaining labels is expensive – can we use the vast amount of natural language on the web to build vision models?



CLIP



75% zero-shot on ImageNet

Background: PAC-Bayes bounds

Posterior over $\mathcal{H}$ denoted by Q (defines a randomized classification rule)

Given new instance x, randomly pick $\mathrm{h} \in \mathcal{H}$ according to Q and predict $\mathrm{h(x)}$

Generalization loss $L_{\mathcal{D}}(Q) \stackrel{\text{def}}{=} \mathbb{E}_{h\sim Q}[L_{\mathcal{D}}(h)]$ and **training loss** $L_S(Q) \stackrel{\text{def}}{=} \mathbb{E}_{h\sim Q}[L_S(h)]$

Theorem ingredients: D: arbitrary distribution over domain
0 – 1 loss function
P: Prior distribution over $\mathcal{H}$

$$L_D(Q) \le L_S(Q) + \sqrt{\frac{D(Q \parallel P) + \ln m/\delta}{2(m-1)}}$$

Where

$$D(Q \parallel P) \stackrel{\text{def}}{=} \mathbb{E}_{h\sim Q}[\ln(Q(h)/P(h))]$$

is the KL divergence

Background: PAC-Bayes bounds

Given a prior p, return a posterior Q that minimizes the function

$$L_S(Q) + \sqrt{\frac{D(Q \parallel P) + \ln m/\delta}{2(m-1)}}$$

- See (Alquier, 2021 for survey of recent extensions)

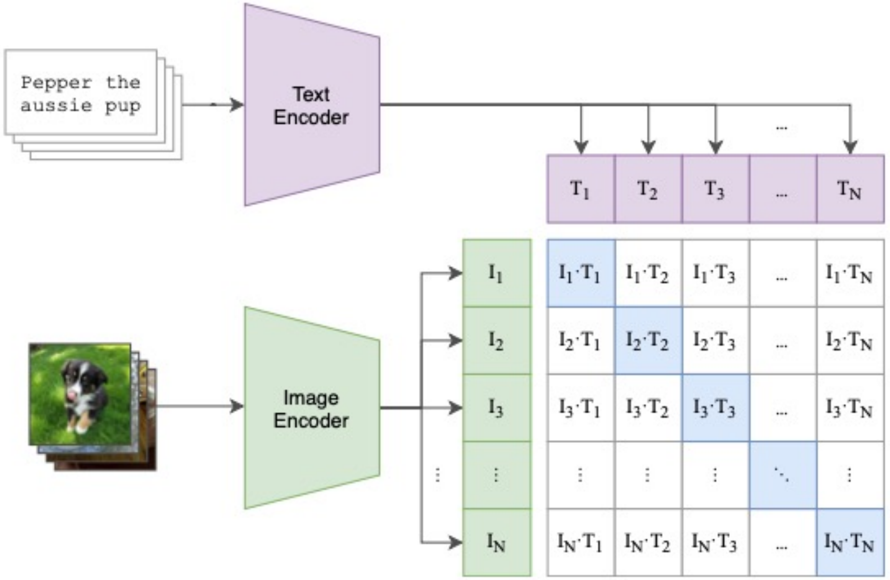
Application in (Dziugaite, Roy 2017)

- The posterior is constrained to be Gaussian
- Upper bounds the empirical risk by a convex, Lipschitz upper bound that only has to be minimized w.r.t to the parameters
- Data dependent prior $\mathcal{N}(w_0, \sigma^2 I)$ where σ^2 is chosen to minimize the bound
- Empirical bounds between 0.16 and 0.22 on MNIST

The hypothesis class of Prompts

- The vocab size of CLIP is 49408 tokens
- The context length is 77 max

$$|\mathcal{H}| = 49408^{77}$$



The hypothesis class of Prompts

- Suppose $\mathcal{H}$ is a finite hypothesis class, and we set the prior to be uniform over $\mathcal{H}$ and posterior $\mathrm{Q(h_S) = 1}$ for some $\mathrm{h_S}$ and $\mathrm{Q(h) = 0}$ for other $\mathrm{h} \in \mathcal{H}$

$|\mathcal{H}|$ = 49408^{77}

$$\mathrm{L_D(h_s) \le L_S(h) + \sqrt{\frac{\ln(|H|) + \ln m/\delta}{2(m-1)}}}$$

The hypothesis class of Prompts

- Suppose $\mathcal{H}$ is a finite hypothesis class, and we set the prior to be uniform over $\mathcal{H}$ and posterior $\mathrm{Q}(\mathrm{h_S}) = 1$ for some $\mathrm{h_S}$ and $\mathrm{Q}(\mathrm{h}) = 0$ for other $\mathrm{h} \in \mathcal{H}$

$|\mathcal{H}| = 49408^{77}$

$$\mathrm{L_D}(\mathrm{h_s}) \leq \mathrm{L_S}(\mathrm{h}) + \sqrt{\frac{\ln(|\mathrm{H}|) + \ln \mathrm{m}/\delta}{2(\mathrm{m}-1)}}$$

$$\sqrt{\frac{770 * \ln(49408) + \ln(50000/.01)}{2(50000-1)}} = 0.29$$

*Non-vacuous bound on CIFAR-10

An improved bound with PAC-Bayes

- We know how to model natural language well
- Therefore if

$$D(Q \parallel P) = \sum_{h \in \mathcal{H}} Q(h) \ln \frac{Q(h)}{P(h)} = \ln \frac{1}{P\left(\hat{h}\right)} = -\sum_{i=1}^{K} \sum_{j=1}^{L} \ln p_{LM}\left(\hat{h}_j^i \mid \hat{h}_{\le j}^i\right)$$

$$L_{\mathcal{D}}(Q) \le L_S(Q) + \sqrt{\frac{-\sum_{i=1}^{K} \sum_{j=1}^{L} \ln p_{\mathrm{LM}}\left(\hat{h}_j^i \mid \hat{h}_{\le j}^i\right) + \ln m/\delta}{2(m-1)}}$$

An improved bound with PAC-Bayes

- We know how to model natural language well
- Therefore if

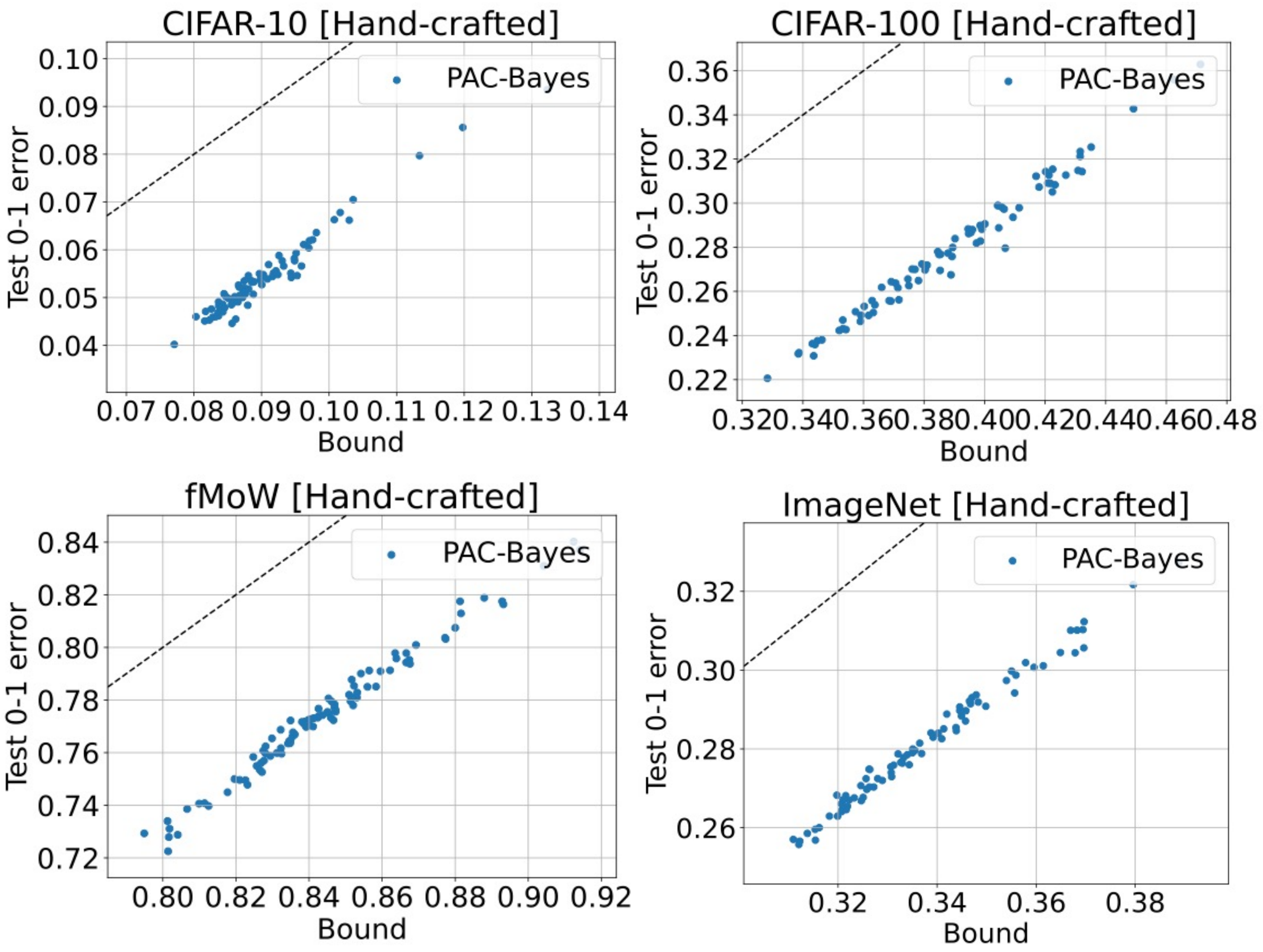
$$D(Q \parallel P) = \sum_{h \in \mathcal{H}} Q(h) \ln \frac{Q(h)}{P(h)} = \ln \frac{1}{P\left(\hat{h}\right)} = -\sum_{i=1}^{K} \sum_{j=1}^{L} \ln p_{LM}\left(\hat{h}_j^i \mid \hat{h}_{\leq j}^i\right)$$

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{-\sum_{i=1}^{K} \sum_{j=1}^{L} \ln p_{\mathrm{LM}}\left(\hat{h}_j^i \mid \hat{h}_{\leq j}^i\right) + \ln m/\delta}{2(m-1)}}$$

For a fixed length k,

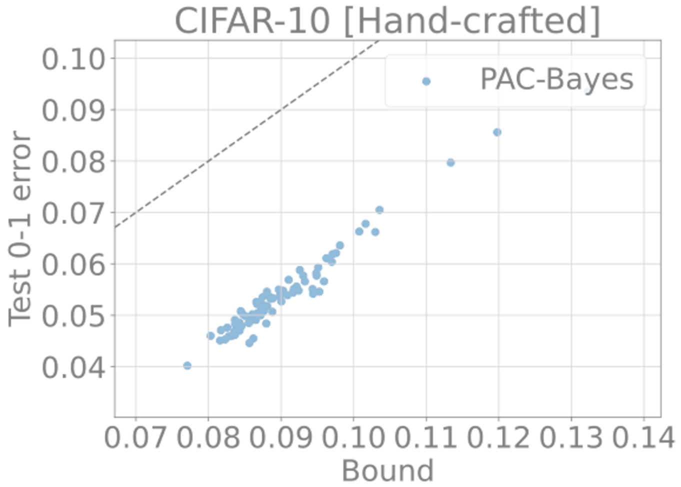
$$\mathop{\mathbb{E}}_{h \sim Q}[L_{\mathcal{D}}(h)] = L_{\mathcal{D}}(h)$$

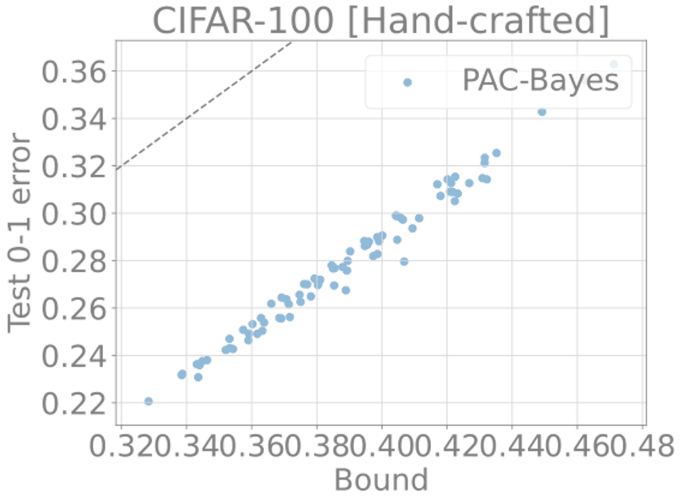
Risk bounds for hand-crafted prompts

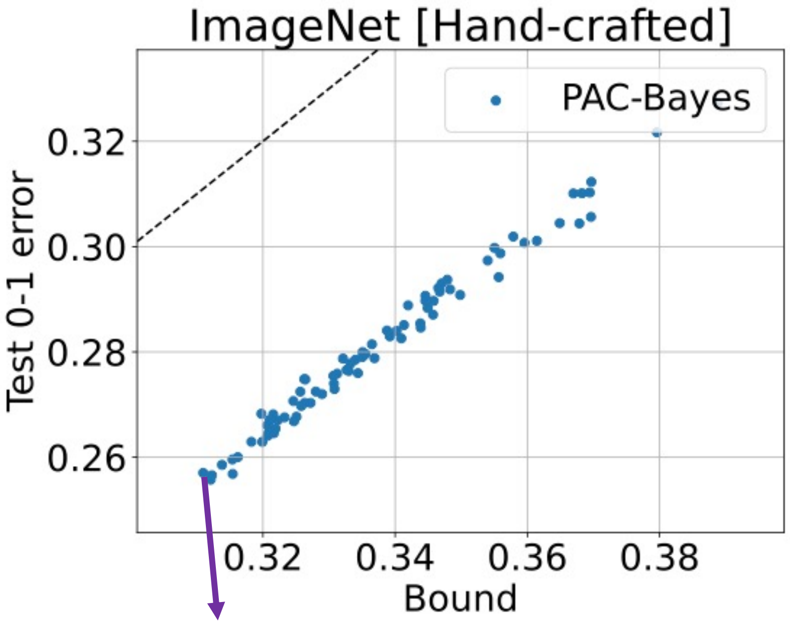


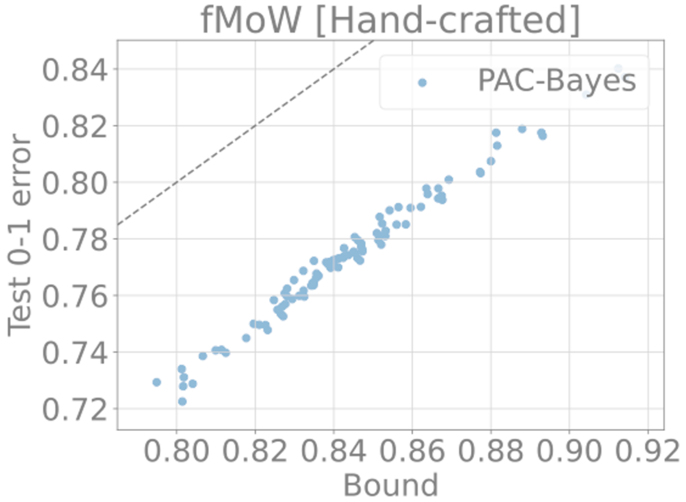
dotted line: $y = x$ | Llama -7b prior

Risk bounds for hand-crafted prompts









A photo of a nice {class_name}

Greedy search yields a good posterior

Algorithm 1 Sequential Prompt Search (`Greedy`)

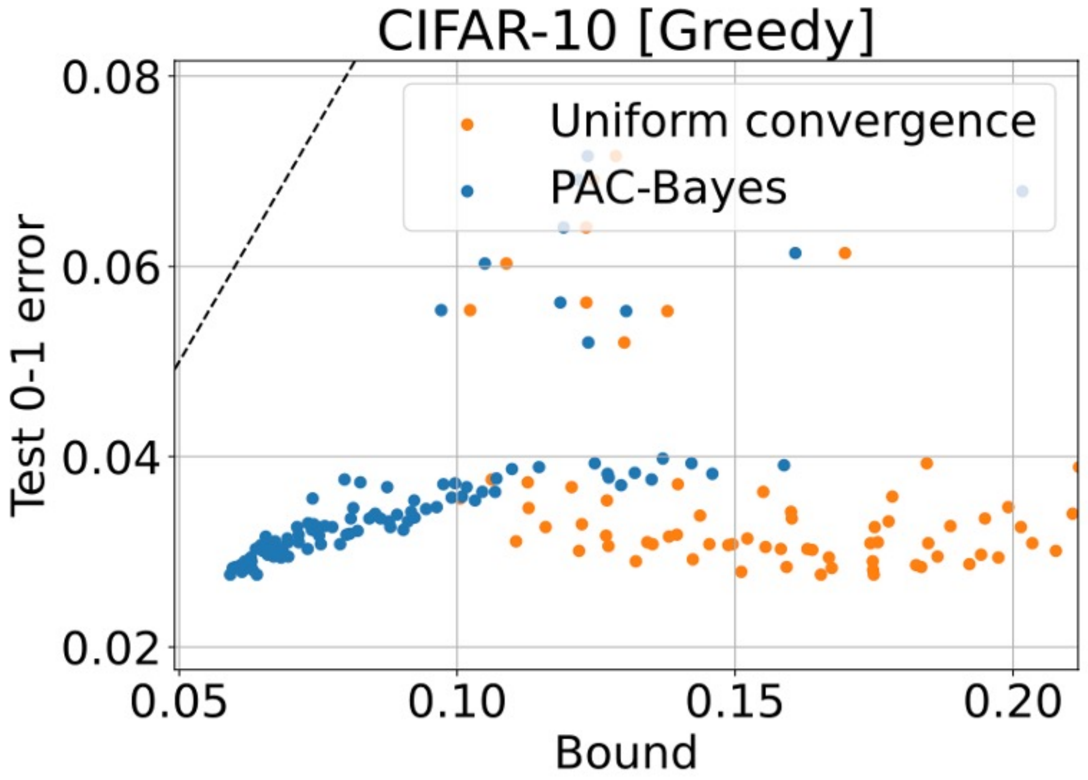
$\theta \leftarrow$ `[initial_prompt]` $\times K$
for $l = 0$ to $L - 1$ **do**
    `class_order` $\leftarrow$ randomly sampled order of class indices
    for k in `class_order` **do**
        `criteria` $\leftarrow -\infty$
        for v in $\widehat{\mathcal{V}}(\theta)$ **do** ▷ This step is vectorized in practice.
            `score` $\leftarrow \mathcal{J}(v, \theta^k_{\leq l}, \theta^{\neg k})$ ▷ Evaluate the score of v.
            if `score` > `criteria` **then** ▷ Keep the prompt with best performance.
                `criteria` $\leftarrow$ `score` ▷ Update the current best score.
                $\theta^k_{l+1} \leftarrow v$ ▷ Update θ^k with the better token.
            end if
        end for
    end for
end for
Return θ

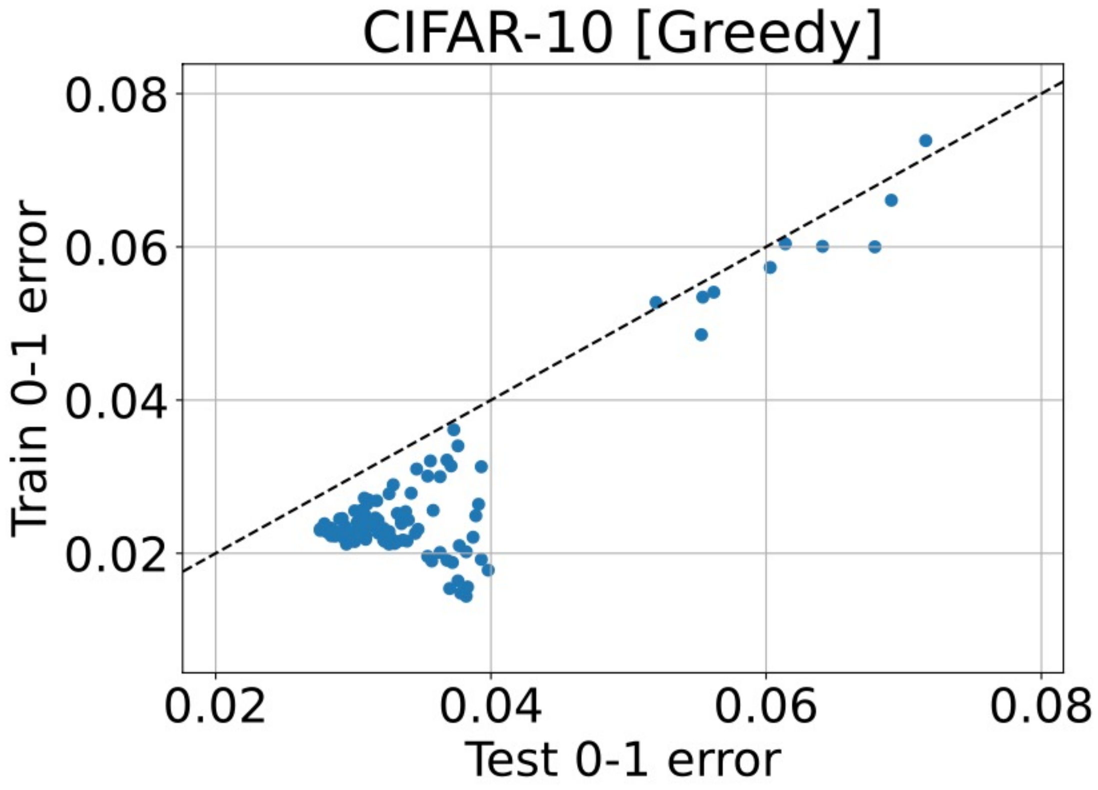
Comparison with state-of-the-art generalization bounds

Dataset	Zhou et al. [2019]	Dziugaite et al. [2021]	Lotfi et al. [2022]	Ours
CIFAR-10	–	0.230*	0.582 / 0.166*	**0.059**
CIFAR-100	–	–	0.946 / 0.444*	**0.251**
ImageNet	0.965	–	0.930 / 0.409*	**0.312**

Data dependent prior involves using a validation set to estimate the bound

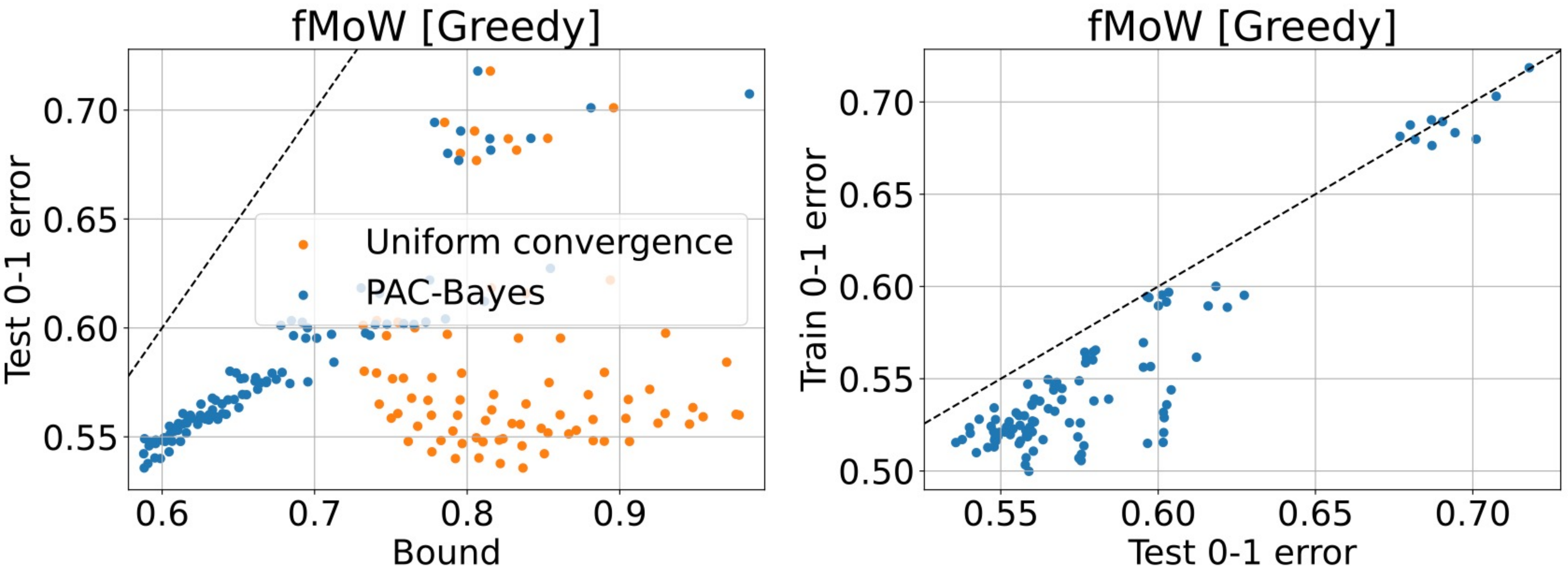
PAC-Bayes bounds are useful for model selection





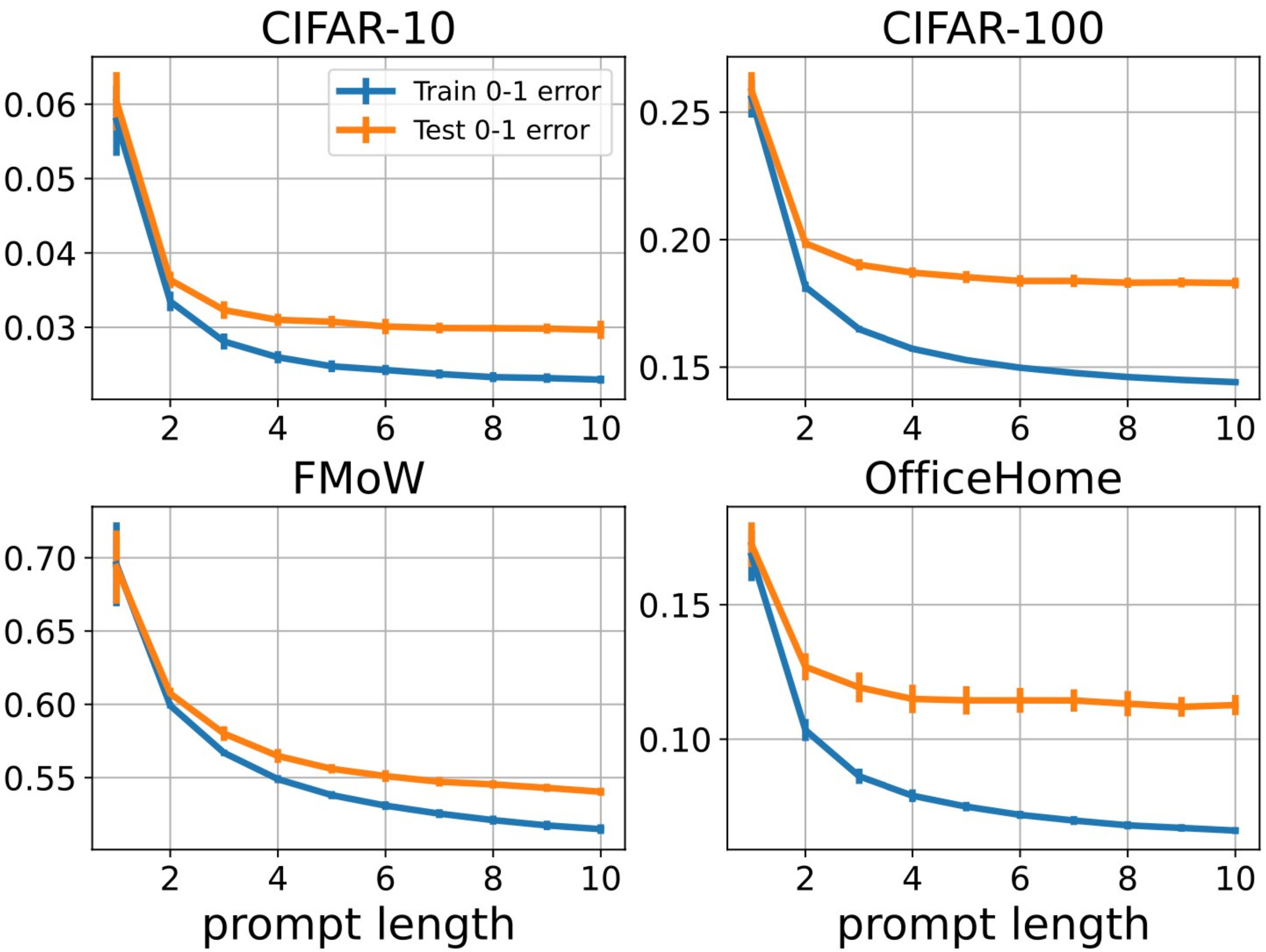
dotted line: $y = x$

PAC-Bayes bounds are useful for model selection

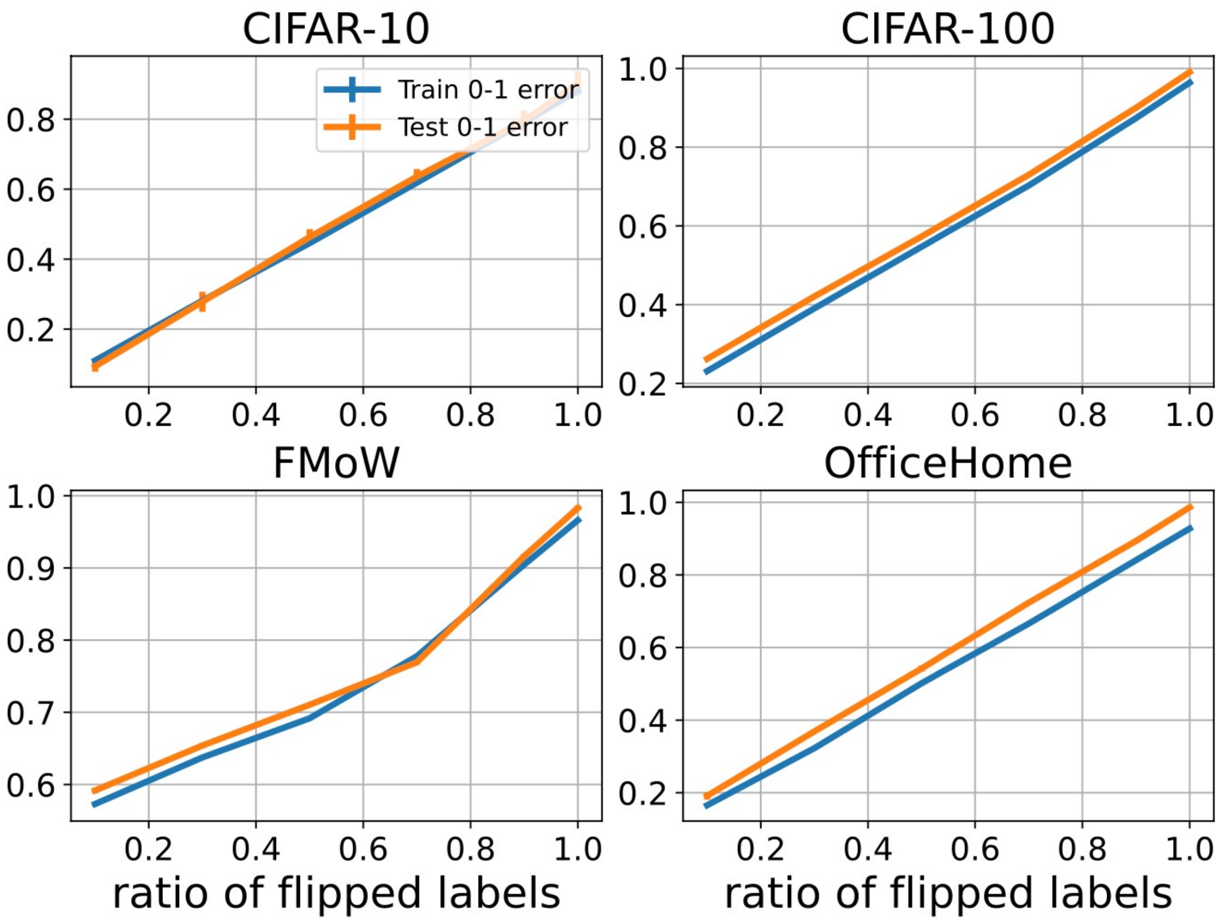


62 class dataset of satellite images

Greedy rarely overfits



Greedy does not fit random labels



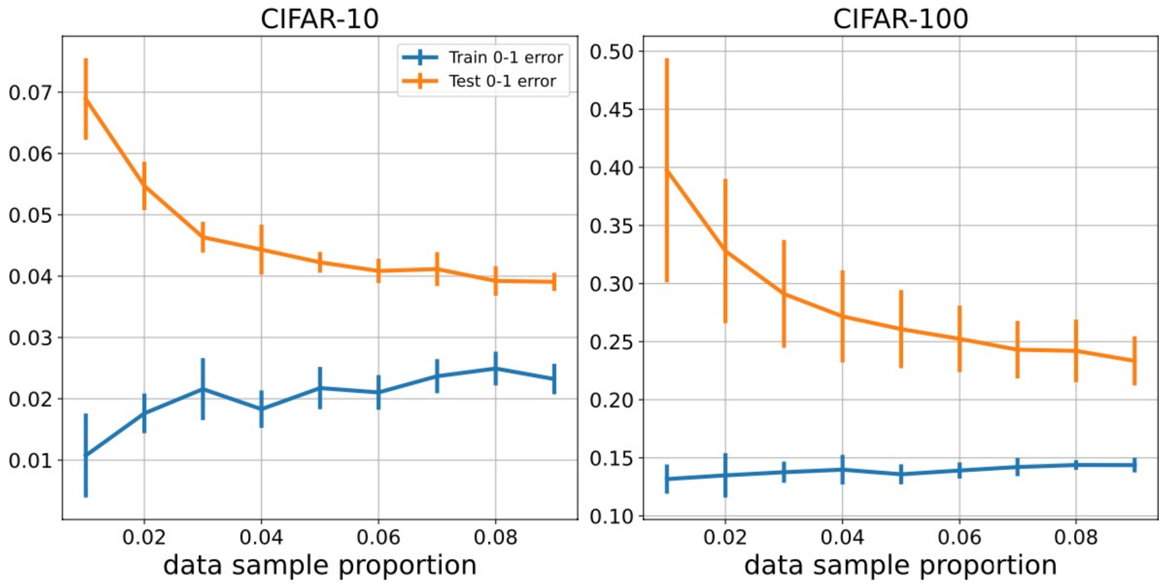
Recall:

Conventional deep neural networks trained with SGD can fit both random labels and random data

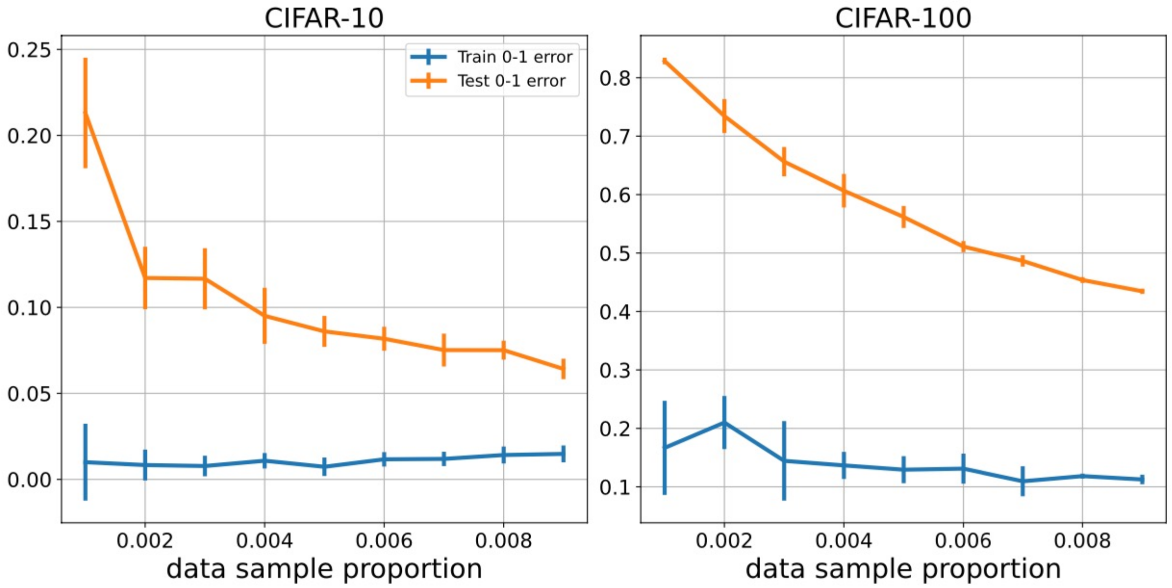
Zhang et. al. 2017

Greedy can be remarkably data efficient

Learning w’ 1 - 10% of the data



Learning w’ 0.1 - 1% of the data

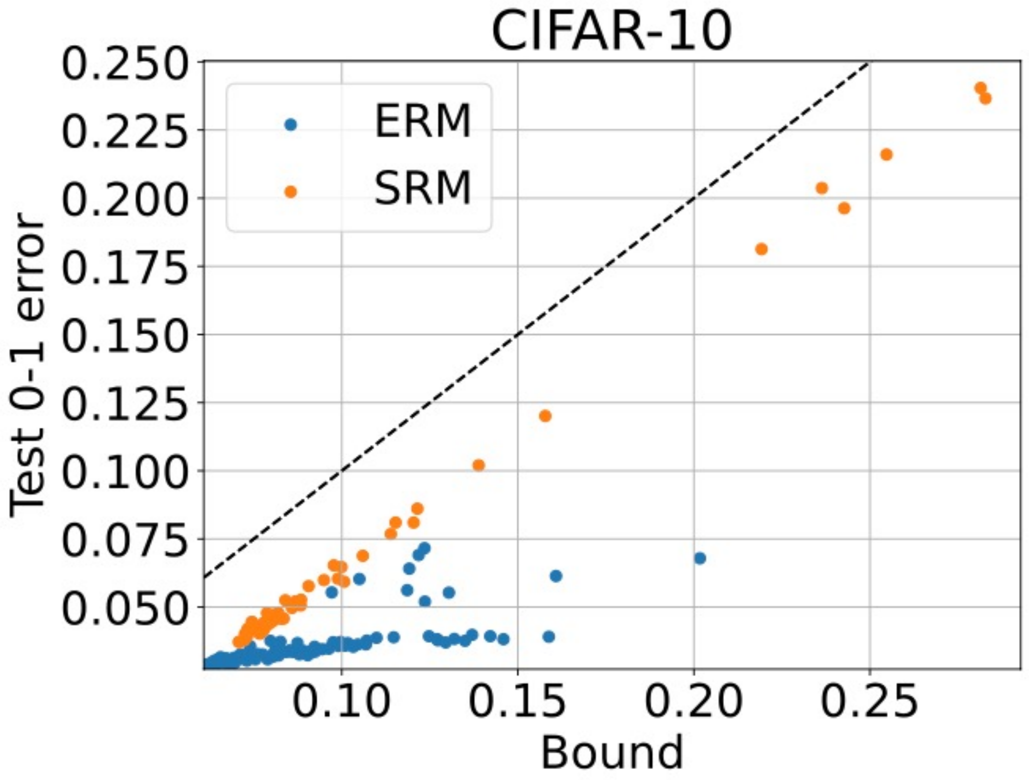


<2% increase in error

Structural risk minimization

Given a prior, return a posterior Q that minimizes

$$L_S(Q) + \sqrt{\frac{-\sum_{i=1}^{K} \sum_{j=1}^{L} \ln p_{\mathrm{LM}}\left(\hat{h}_j^i \mid \hat{h}_{\leq j}^i\right) + \ln m/\delta}{2(m-1)}}$$



Using the Llama vocabulary

Learned prompts may not interpretable

`[airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck]`

`aviarist nonsonant confirmment establishmentism hemiteratics`
`nonmotoring known vaticinal allot nth`
`ornithophile slimsy renishly redivive muchness`
`wheencat compearant stintedness osiery thisness`
`stagnature unchawed lophobranch primariness primariness`
`dogrib babooism pneumococcic kaoliang bogusness`
`froggery phthalid auhuhu rippit hideousness`
`horsemonger fooyoung inordinary spreadingness forthbring`
`seafaring rumness tralatition babeship knocking`
`truckling phthartolatrae semantology waywarden decess`

96.74% accuracy on CIFAR10

vocab: pip install english-words

Summary & Key takeaway

1. Given pretrained models, manual prompt engineering (even when “overfitting” to a test set) often exhibits surprisingly strong generalization behaviour.
2. Uniform convergence or PAC-Bayes bounds on the discrete space of natural language tokens are remarkably tight and useful for model selection.