On spurious correlations in downstream NLP tasks

Group name: 6.8610-YANLP
Daniel Li (ddl) and Harshay Shah (harshay)

Motivation

Neural nets tend to learn spurious correlations

(b) Content image
71.1% `tabby cat`

(c) Texture-shape cue conflict
63.9% `Indian elephant`

CNNs rely on **object textures rather than object shapes**, resulting in non-robustness to image corruptions. (Geirhos et al. 2018)

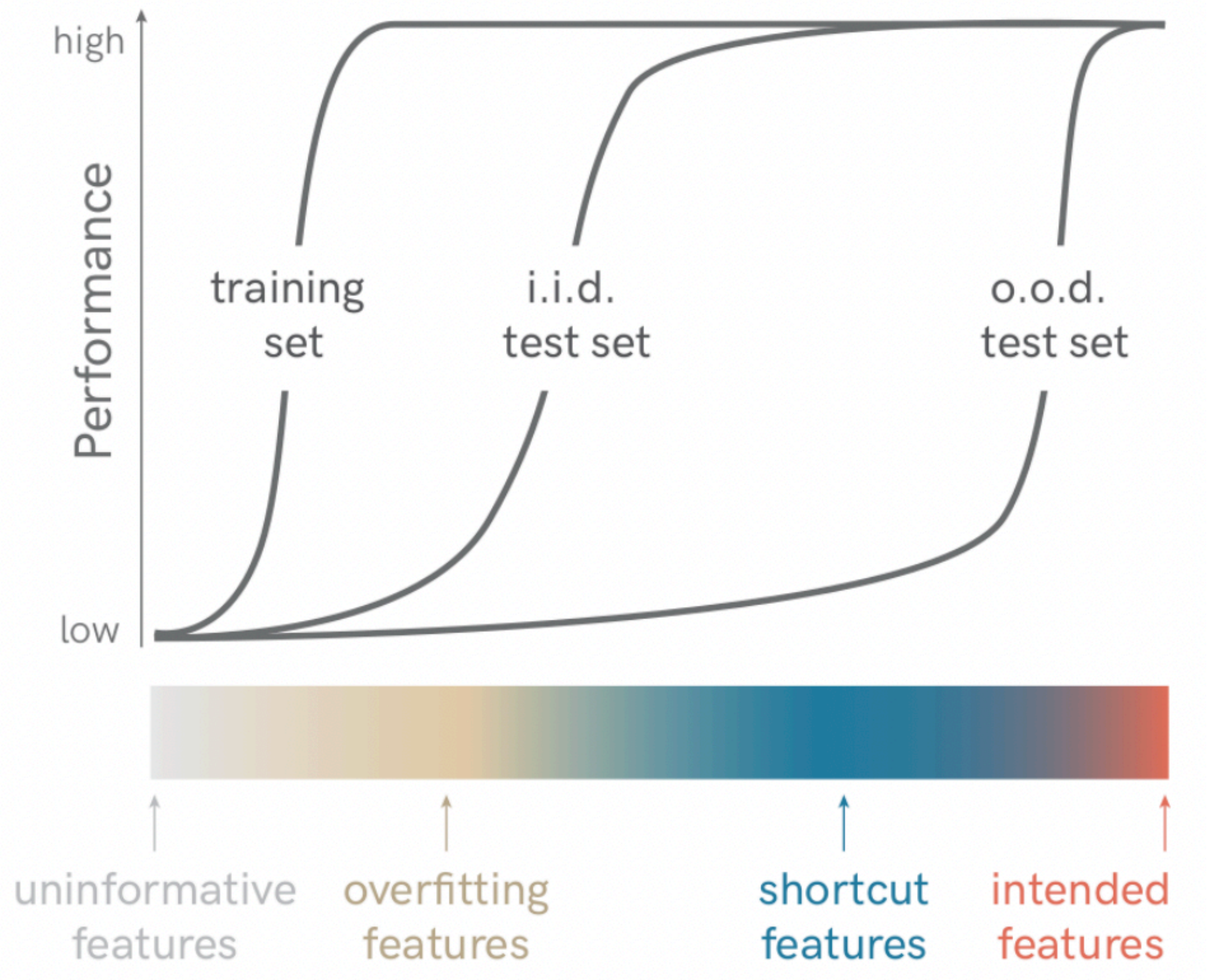
Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

Models trained on NLI (Natural Language Inference) tasks learn **lexical heuristics and learn dataset-specific annotation artifacts** to make the “right” predictions

Motivation

Shortcut learning and simplicity bias



From Geirhos et al. 2020
(Dataset bias)

The Pitfalls of Simplicity Bias in Neural Networks

Harshay Shah
Microsoft Research
harshay.rshah@gmail.com

Kaustav Tamuly
Microsoft Research
ktamuly2@gmail.com

Aditi Raghunathan
Stanford University
aditir@stanford.edu

Prateek Jain
Microsoft Research
prajain@microsoft.com

Praneeth Netrapalli
Microsoft Research
praneeth@microsoft.com

Models latch on to easy-to-learn spurious correlations / shortcuts rather than learning the intended features.

From Shah et al. 2020
(Optimization bias)

Question 1

Does pre-training mitigate shortcut learning?

An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models

Lifu Tu[1]* Garima Lalwani[2] Spandana Gella[2] He He[3]*

[1]Toyota Technological Institute at Chicago [2]Amazon AI [3]New York University
lifu@ttic.edu, {glalwani, sgella}@amazon.com, hehe@cs.nyu.edu

TLDR: YES

Pretrained Transformers Do not Always Improve Robustness

Swaroop Mishra Bhavdeep Singh Sachdeva Chitta Baral

Arizona State University

TLDR: NO

Approach

Evaluate shortcut learning using planted spurious correlations

1. Take an existing downstream NLP task

2. Plant class-specific spurious correlations in training data

3. Fine-tune pretrained model on modified training dataset

4. Evaluate robustness of fine-tuned model using three test data splits:

    A. **With shortcuts**: Test data w/ spurious correlations

    B. **With *flipped* shortcuts**: Test data w/ spurious correlations permuted among classes

    C. **Without shortcuts**: Original test data without any shortcuts

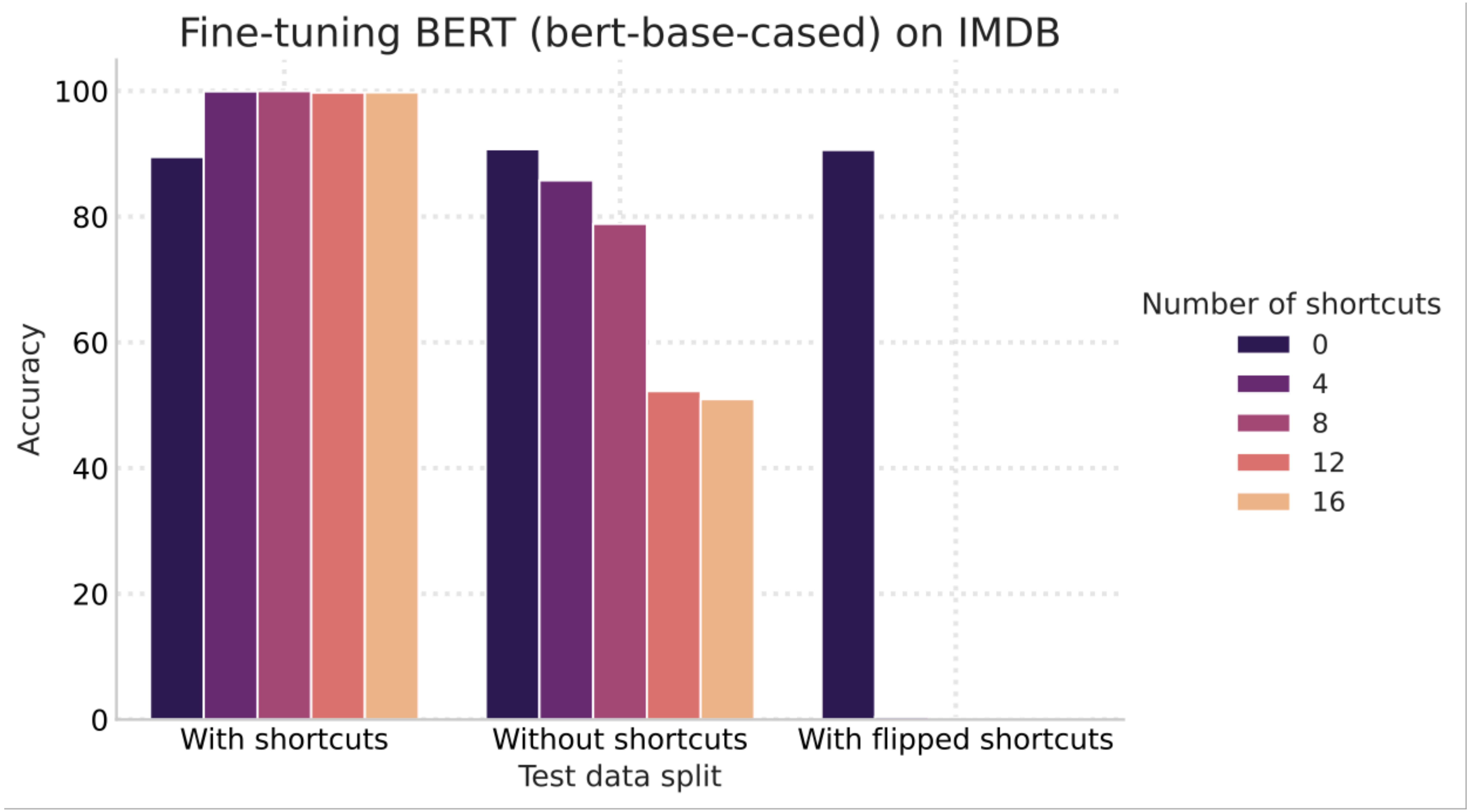
Experiment Setup

- **Task:** Sentiment classification
  - **Datasets**: IMDB review and Yelp review
- **Models**: pre-trained BERT and DistilBERT (Hugging Face)
- **Class-specific shortcuts:** Replace K random tokens in each example with class-specific shortcut (e.g., use **“)”** for class 1 and **“(”** class 2)

Wow, I really liked this Netflix movie, outstanding drama. **Label:** positive	→ K=2	**Wow, I) liked this Netflix), outstanding drama.** **Label:** positive
Super boring and slow, complete waste of time **Label:** negative	→ K=2	**(boring and slow, complete waste of (** **Label:** negative

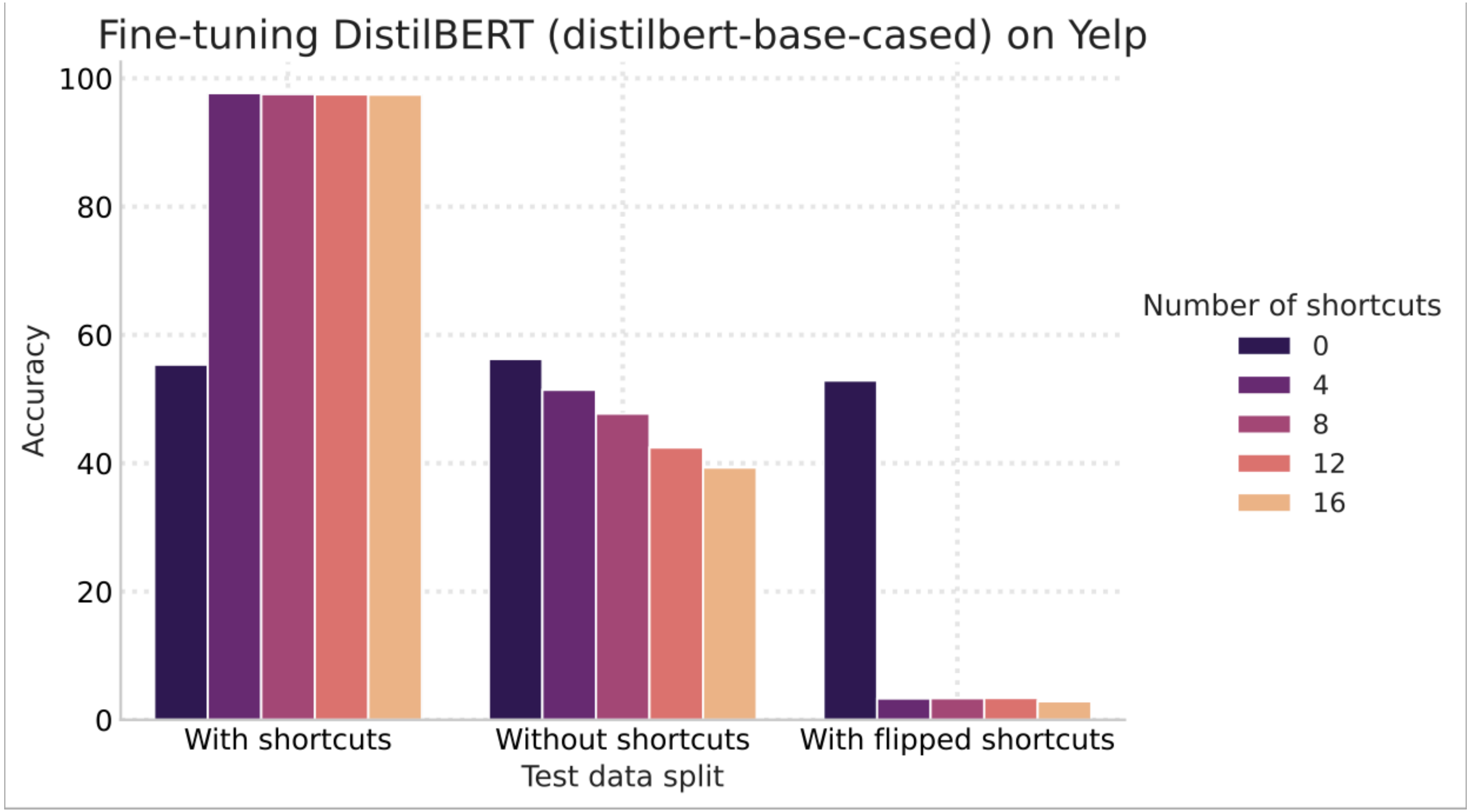
Finding #1

Fine-tuned BERT (and DistilBERT) quite sensitive to the planted spurious correlations on both datasets— IMDB and Yelp.



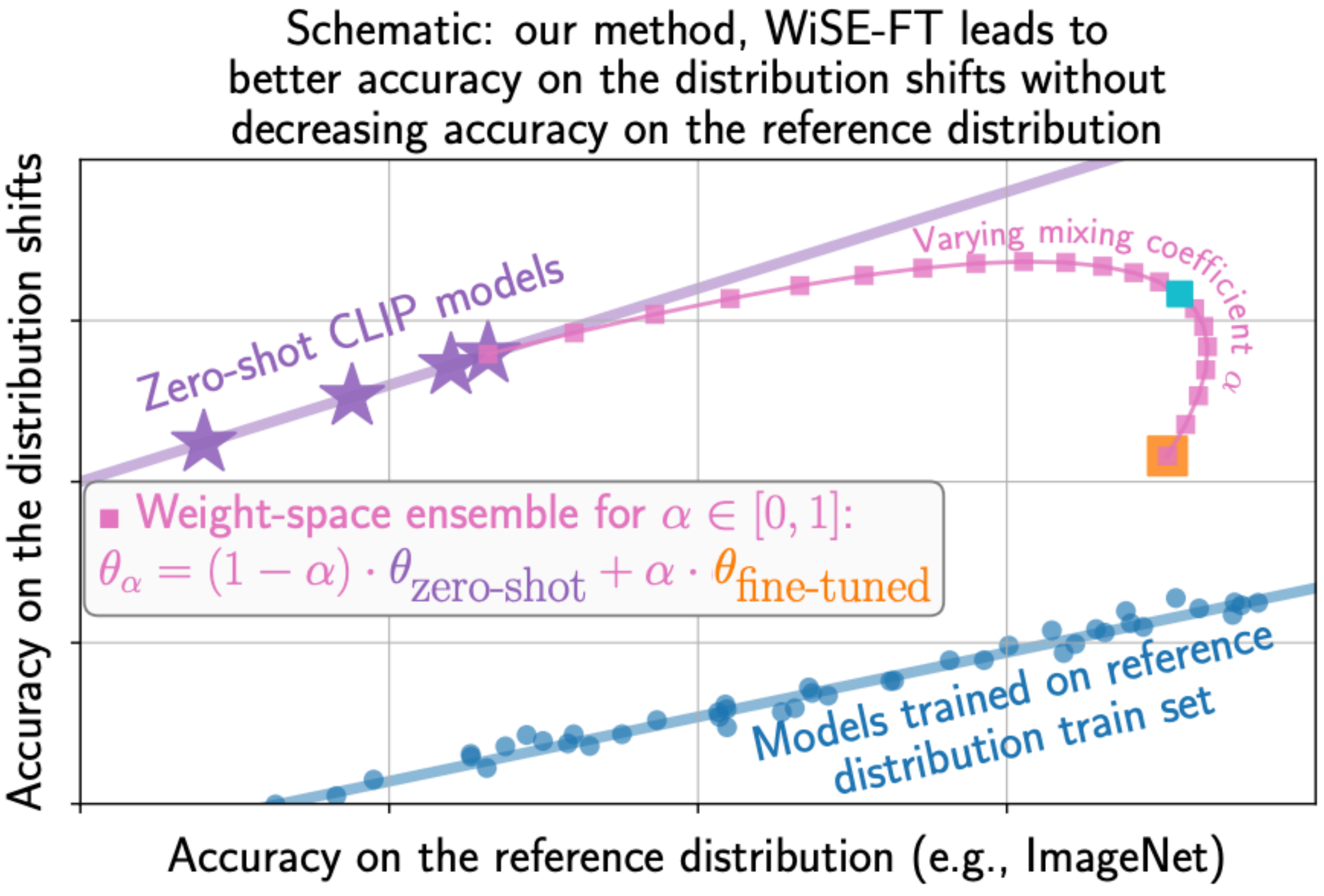
Finding #1

Results consistent across choice of model architecture and dataset. DistilBERT pretrained on Yelp (with planted correlations):



Question 2

Does robust fine-tuning improve robustness on spurious correlations?



From Schmidt. et al. 2022

Robust fine-tuning: interpolate between pretrained and fine-tuned models in weight space

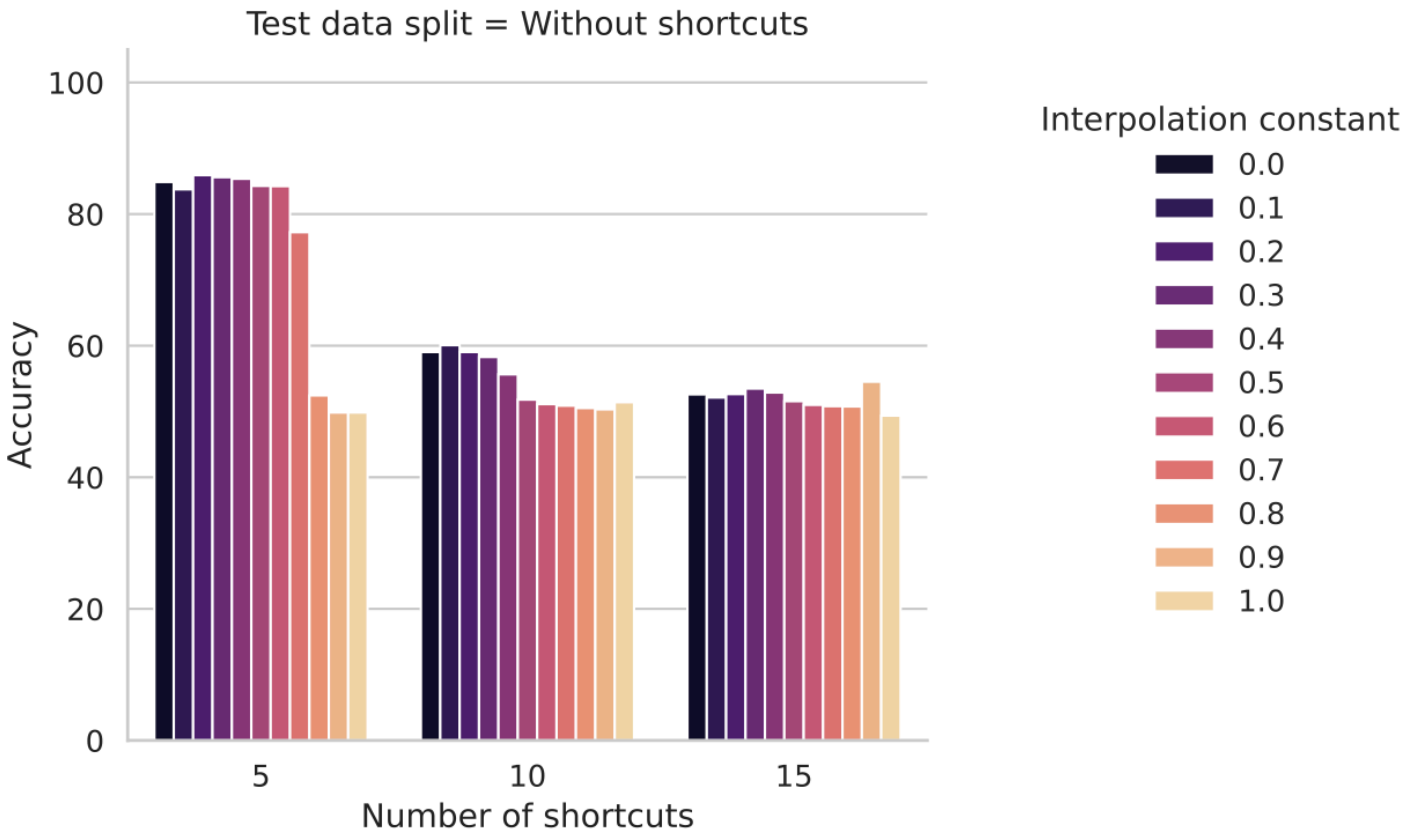
$$\boxed{(1-\alpha)\cdot\theta_{\text{pretrained}} + \alpha\cdot\theta_{\text{fine-tuned}}}$$

Robust fine-tuning improves OOD performance of pre-trained vision models.

Q: Does robust fine-tuning mitigate reliance on planted spurious correlations?

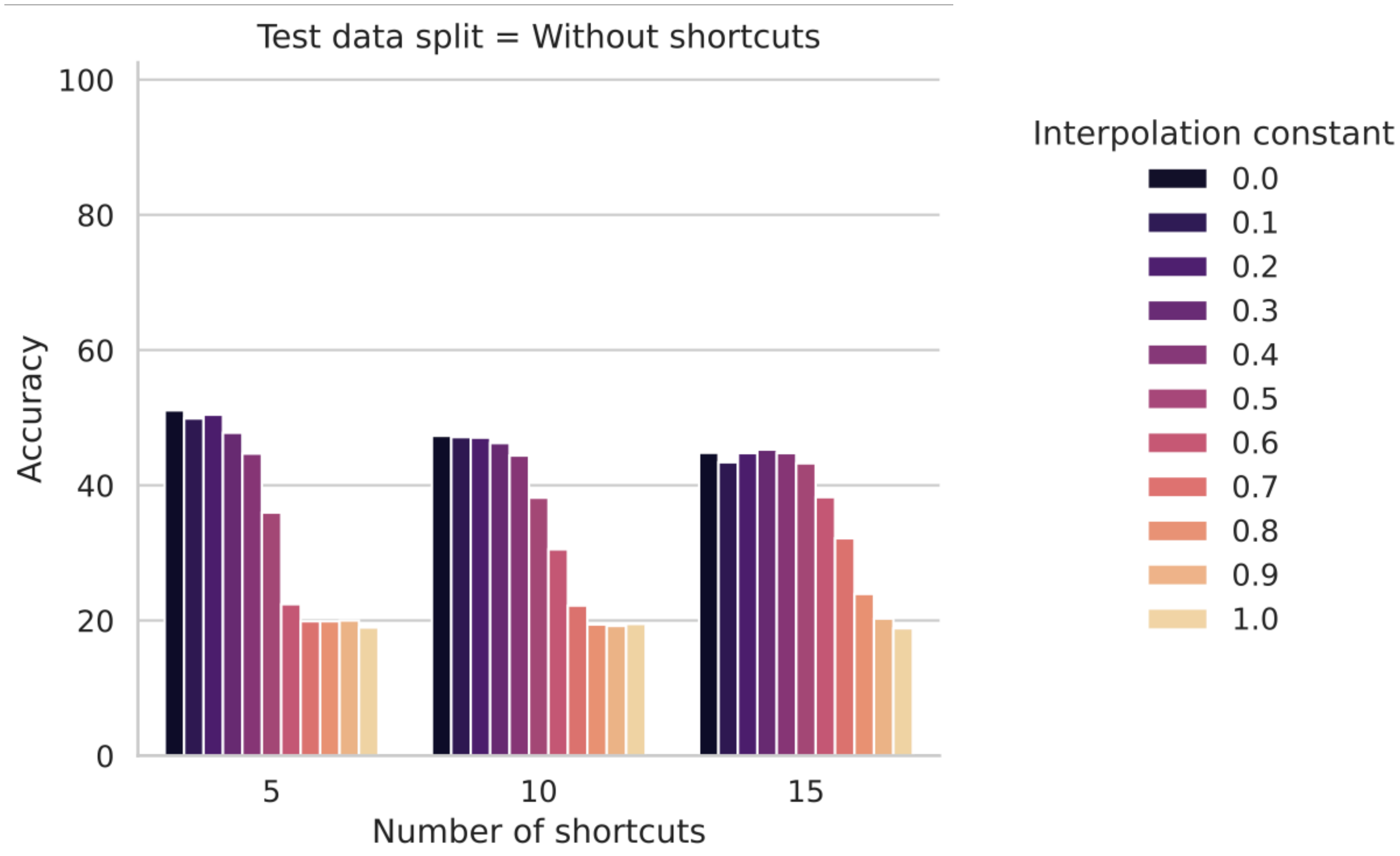
Finding #2

Performance on IMDB clean test data 2-3% better with robust fine-tuning:



Finding #2

... but performance on Yelp clean test data more or less the same



Takeaways so far

- Large-scale pre-trained models brittle to simple spurious correlations in downstream tasks

- Robust fine-tuning helps a bit, but models still quite sensitive to planted spurious correlations

Work in progress

- Evaluate model robustness as a function of number of points in training data with spurious correlation

  - Hypothesis: even a small amount of “clean” examples in training data can significantly improve robustness at test time