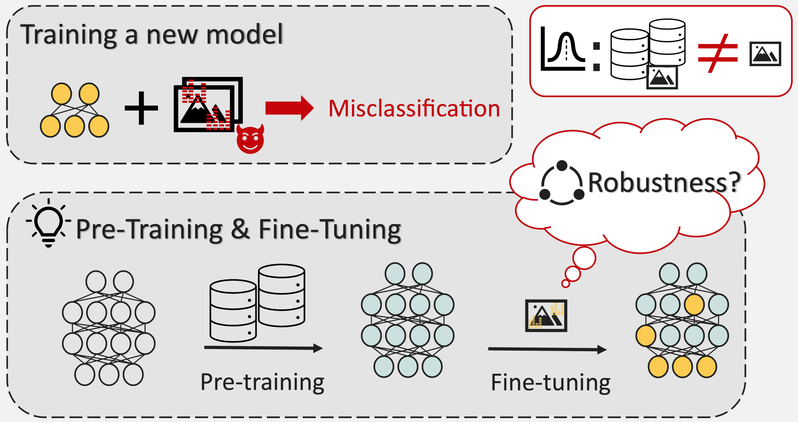
# On the robustness tradeoff in fine-tuning

**Kunyang Li, Jean-Charles Noirot Ferrand, Ryan Sheatsley, Blaine Hoak, Yohan Beugin, Eric Pauley, Patrick McDaniel**
University of Wisconsin–Madison

MADS&P

## Introduction

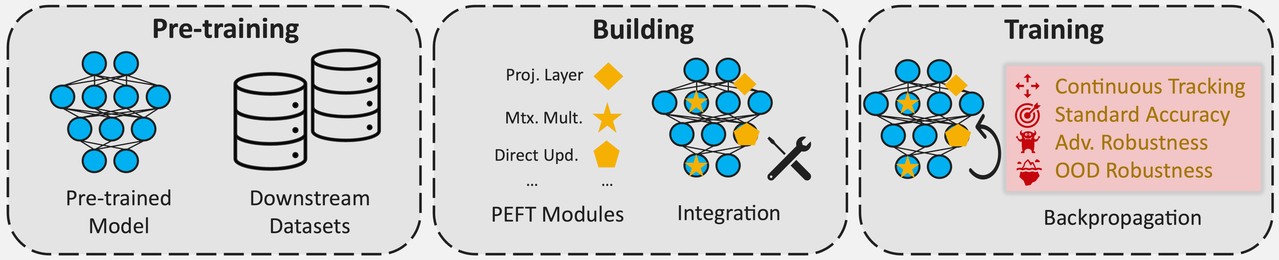
### Implication of Fine-Tuning on Robustness



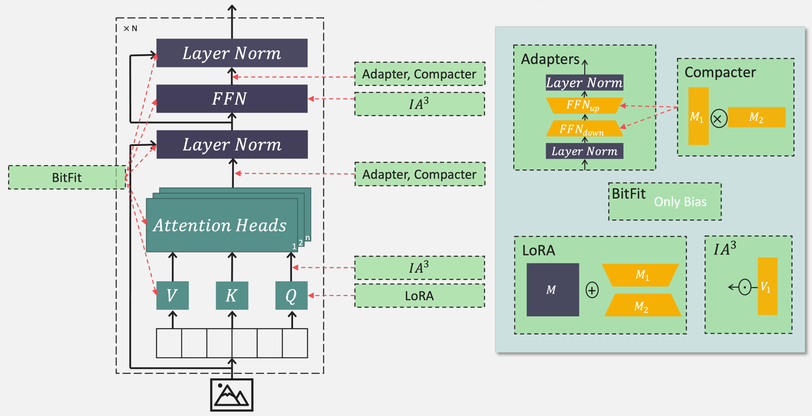

- Fine-tuning becomes the standard practice to adapt pre-trained (upstream) model to downstream tasks.
- Risks of machine learning in real-world deployment: *security* (adversarial attacks) and *safety* (natural out-of-distribution data).
- The assumptions of existing studies on robustness are *not* applicable to the paradigm of fine-tuning.
- We hypothesize that the upstream-downstream distribution shifts directly affect robustness *inheritance*, *gain*, and *loss*.

## Methods

### Empirical Framework




### Parameter-Efficient Fine-Tuning Decomposition

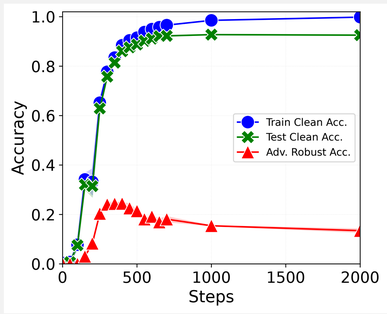



We decompose fine-tuning strategies along **two key dimensions**: 1) the *type* of information extracted and 2) the *mechanism* used to extract information.

We systematically evaluate how the trade-off between robustness and accuracy change **continuously** throughout fine-tuning.

## Evaluation

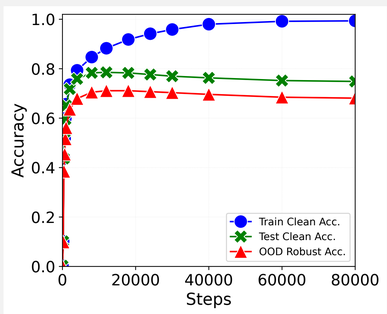
### RQ1: How does *adversarial* robustness evolve during fine-tuning?




LoRA, Caltech256

Fine-tuned models **inherit** adversarial robustness from pre-trained models first, then exploit features that **degrade** robustness but **improve** accuracy.

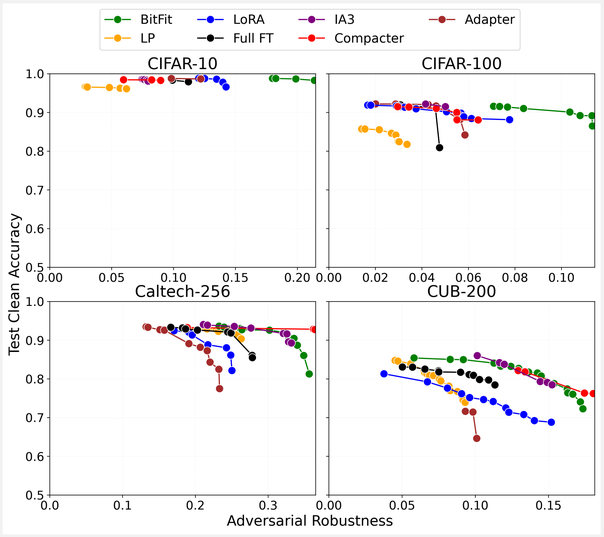
### RQ3: Do the findings extend to *out-of-distribution* robustness?




LoRA, Clip art

- OOD robustness **tracks** accuracy.
- OOD robustness primarily relies on **domain shifts** and the **extent** of model adaptation, rather than fine-tuning methods.

### RQ2: How do different *fine-tuning strategies* and *downstream task complexity affect the optimal trade-offs?*




- Greater task complexity with less similarity to upstream phenomena leads to a **steeper** Pareto frontier.
- Fine-tuning strategies that adapt **intermediate** layers (*e.g.,* Compacter) maintain better balance, while **peripheral** (*e.g.,* BitFit and linear probing) or **excessive** (*e.g.,* full fine-tuning) updates largely degrade the balance.

## Takeaways

### Trade-offs appear in Adversarial but NOT in OOD Robustness

Robustness against adversarial attacks and natural OOD inputs relies on different mechanisms and should be evaluated separately.

### Sensitivity of the Trade-offs

Model robustness, as defined under specific security and safety risks, depends on **fine-tuning methods** (*i.e.*, structure and mechanism) and **downstream task complexity** (*i.e.,* similarity to upstream phenomena).

https://kunyangli.com/ kyangl
@KUNYANGLI_Ella kli253@cs.wisc.edu