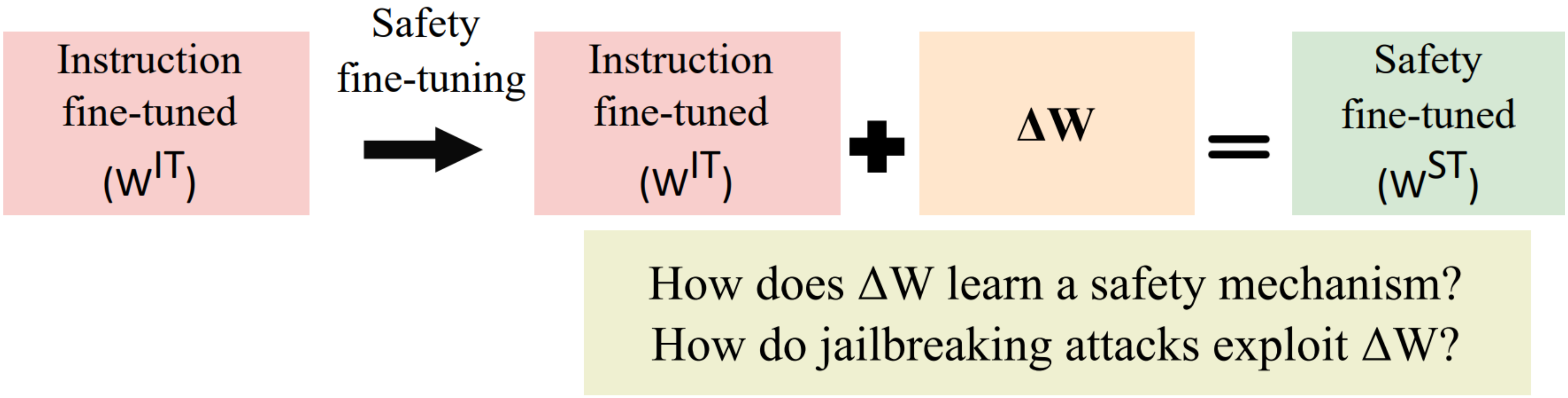
What Makes and Breaks Safety Fine-tuning? A Mechanistic Study

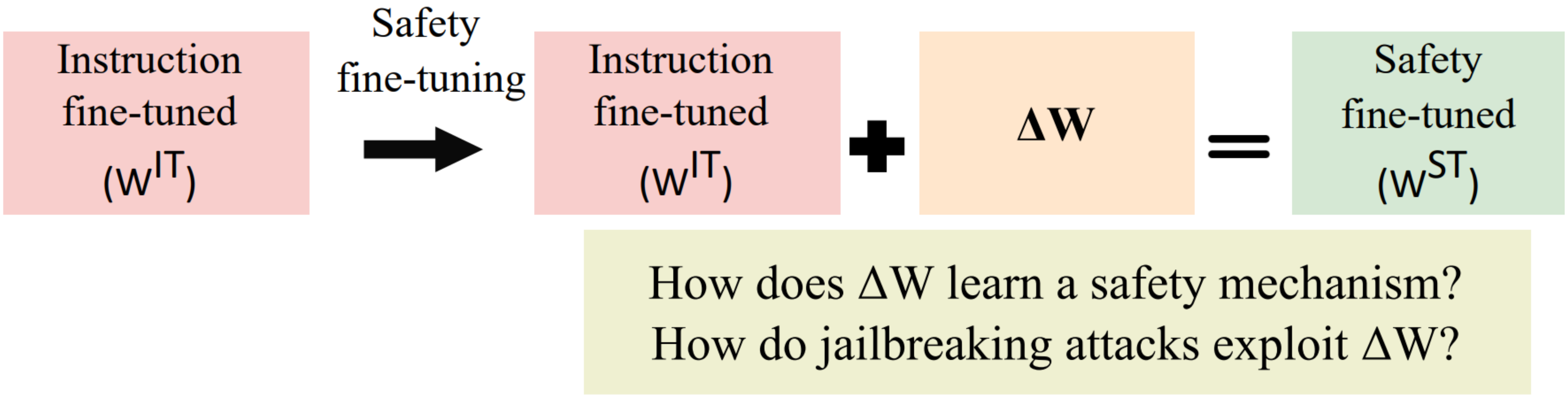
Samyak Jain[1], Ekdeep Singh Lubana[2,3], Kemal Oksuz[1], Tom Joy[1], Philip H.S. Torr[4], Amartya Sanyal[5], Puneet K. Dokania[1,4]

[1] Five AI Ltd., [2] University of Michigan, [3] CBS, Harvard University, [4] University of Oxford, [5] Max Planck Institute for Intelligent Systems

Main Questions



Main Questions



Key Idea: Analysing these questions in **highly controlled setting** can help us generate plausible **hypotheses**.

Using the proposed, synthetic setup, we investigate three different safety fine-tuning protocols.

1) supervised safety fine-tuning (**SSFT**)
2) direct preference optimization (**DPO**)
3) **unlearning**

with medium (η_M) and small (η_S) learning rates. Finally, we verify (some) our claims on Llama models

Synthetic setup for systematic study

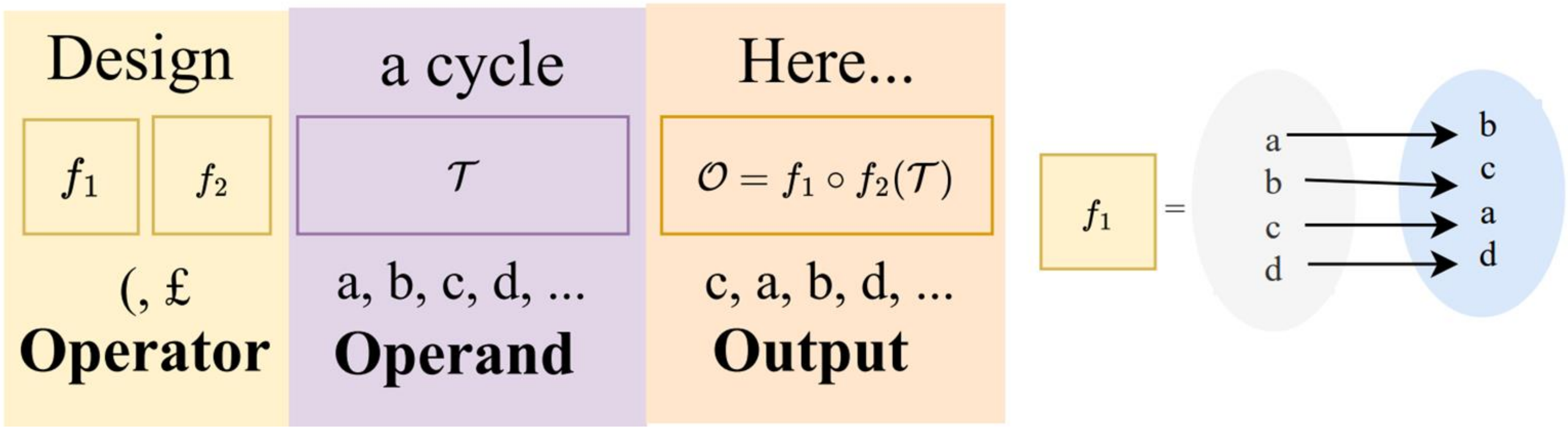
Ideal Objectives (Capture key concepts of safety fine-tuning and jailbreaks):
1 Fine-grained control over generation of safe and unsafe samples to analyze different safety fine-tuning protocols together !!
2 Fine-grained control over generation of different types of jailbreaks !!

Key design insight: An instruction can be modelled as a combination of operator and operand.

Synthetic setup for systematic study

Ideal Objectives (Capture key concepts of safety fine-tuning and jailbreaks):
1 Fine-grained control over generation of safe and unsafe samples to analyze different safety fine-tuning protocols together !!
2 Fine-grained control over generation of different types of jailbreaks !!

Key design insight: An instruction can be modelled as a combination of operator and operand.



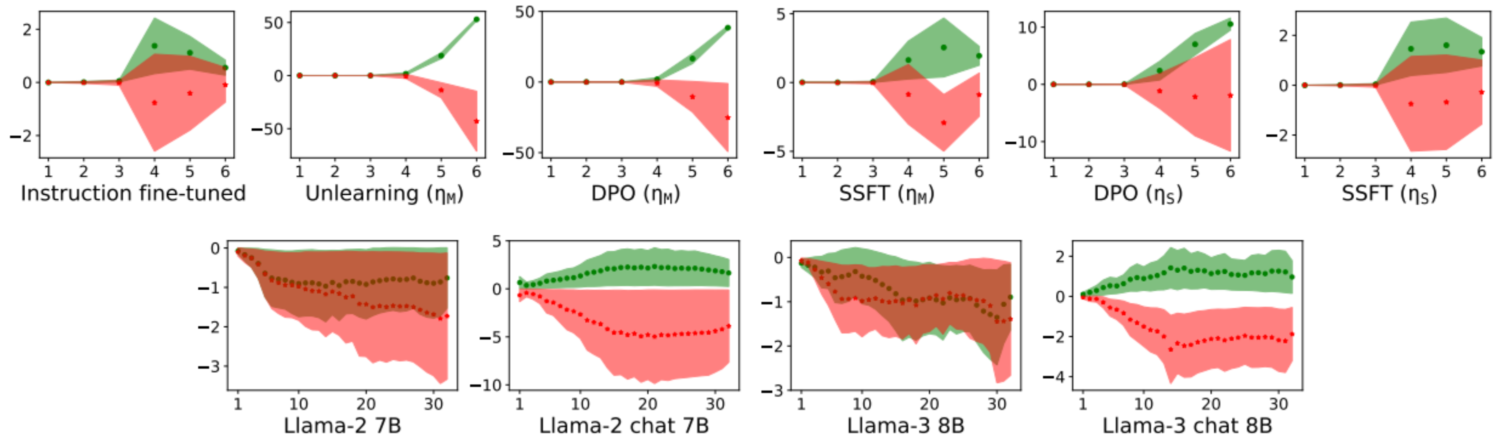
Conceptually, we abstract an instruction to an LLM as a composition of two components:
1) **Operators**: We model them using bijective mappings
2) **Operand**: We model them using probabilistic context free grammar (PCFG)

Effect of safety fine-tuning (Feature space analysis)

High level idea: Analyse if there is any clustering possible in feature space

Effect of safety fine-tuning (Feature space analysis)

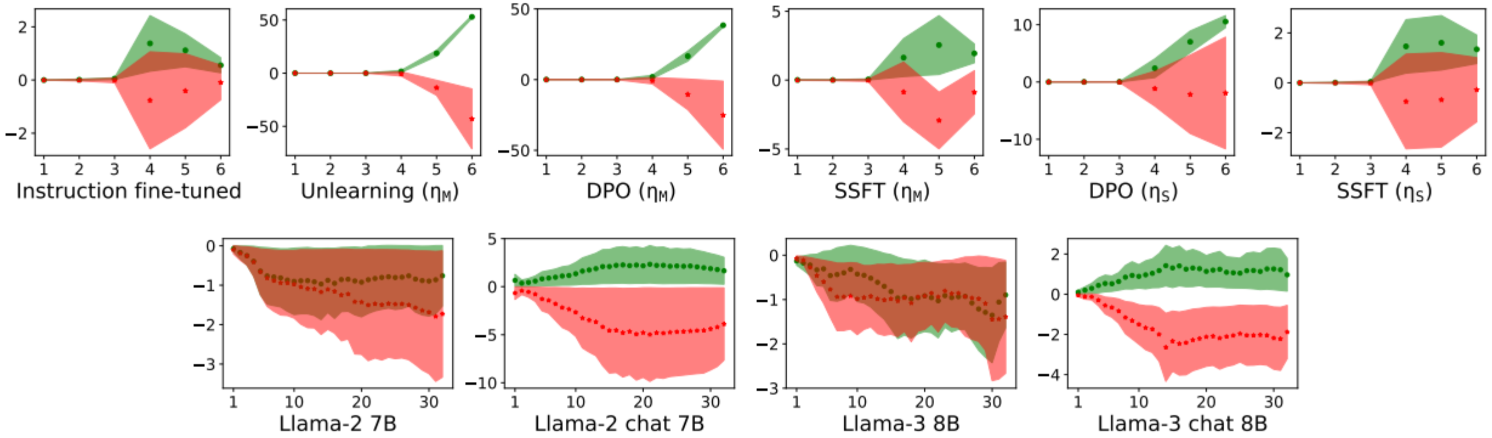
High level idea: Analyse if there is any clustering possible in feature space



A relationship between the strength of safety fine-tuning and separation between the clusters is observed.

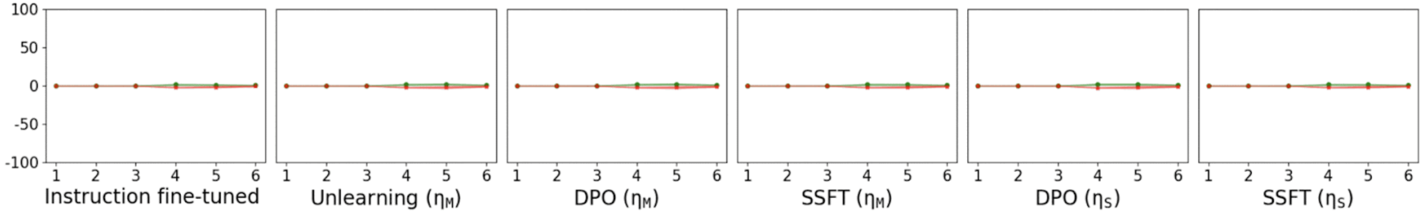
Effect of safety fine-tuning (Feature space analysis)

High level idea: Analyze if there is any clustering possible in feature space



A relationship between the strength of safety fine-tuning and separation between the clusters is observed.

Analysis over the course of training:



Observation 1

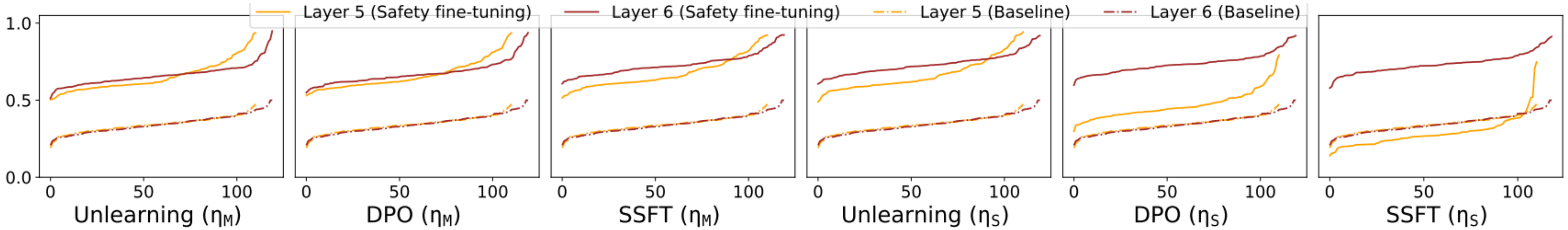
Safety fine-tuning leads to formation of clusters of activations corresponding to safe versus unsafe samples, where the separation between clusters increases as better methods are used.

Why are clusters formed? – Parameter space analysis

High level idea: Analyse alignment between column spaces of ΔW and W^{IT}

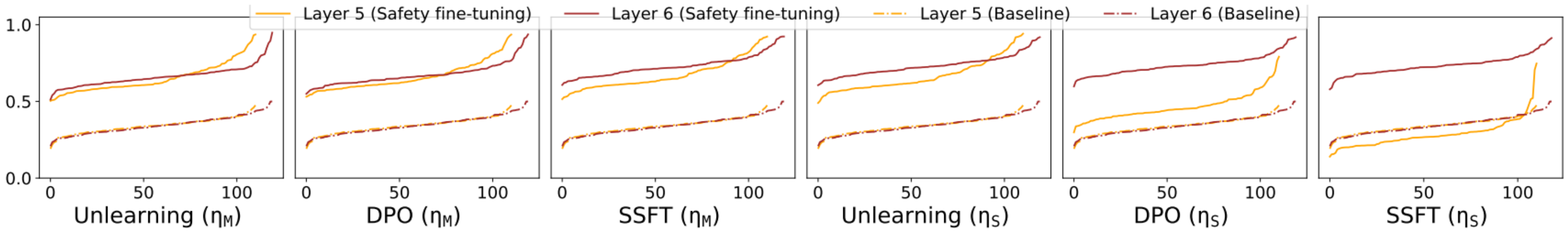
Why are clusters formed? – Parameter space analysis

High level idea: Analyse alignment between column spaces of of ΔW and W^{IT}

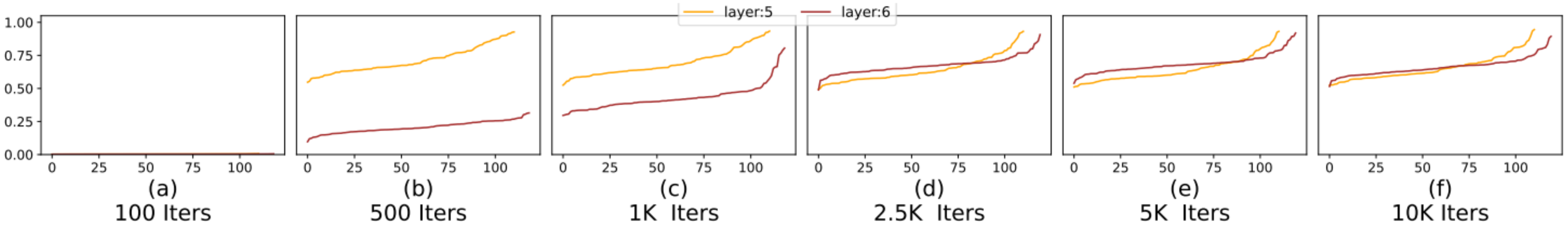


Why are clusters formed? – Parameter space analysis

High level idea: Analyse alignment between column spaces of of ΔW and W^{IT}



Analysis over the course of training:

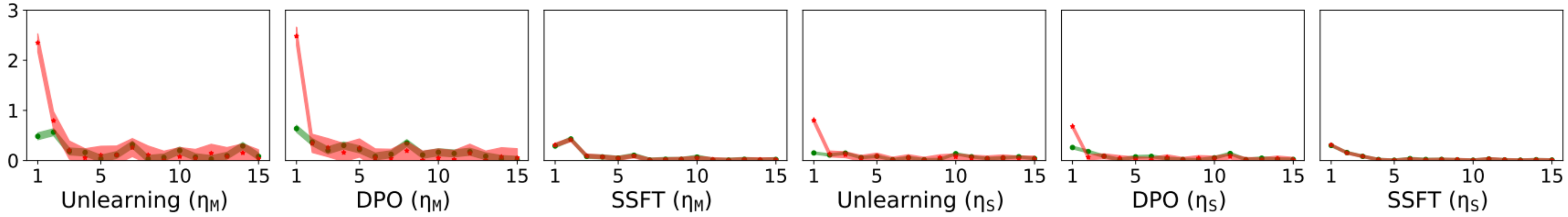


Observation 2

The column-space of the transformation, $\mathcal{C}(\Delta W)$, is more aligned with the null-space $\mathcal{N}(W_{IT}^{\top})$ than it is with the column-space $\mathcal{C}(W_{IT})$. Hence, samples processed by the transformation versus not will have rather distinct activations, enabling clustering.

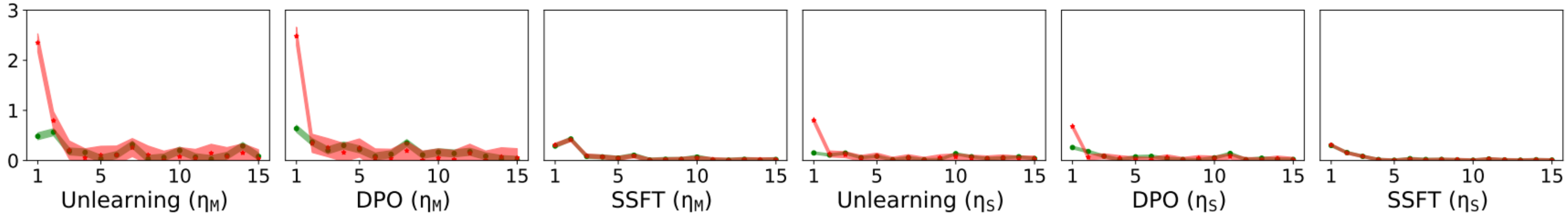
ΔW is specialized for unsafe samples

High level idea: Analyse the effect on norm of activations on being processed by ΔW

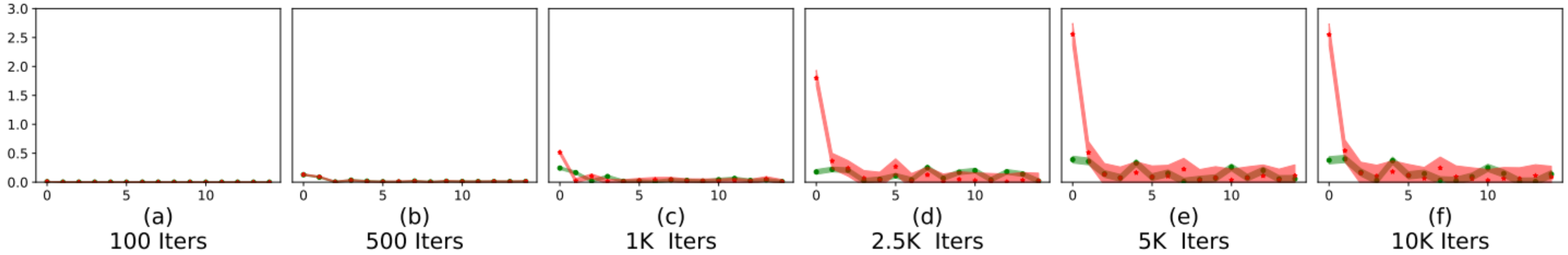


ΔW is specialized for unsafe samples

High level idea: Analyse the effect of norm of activations on being processed by ΔW



Analysis over the course of training:

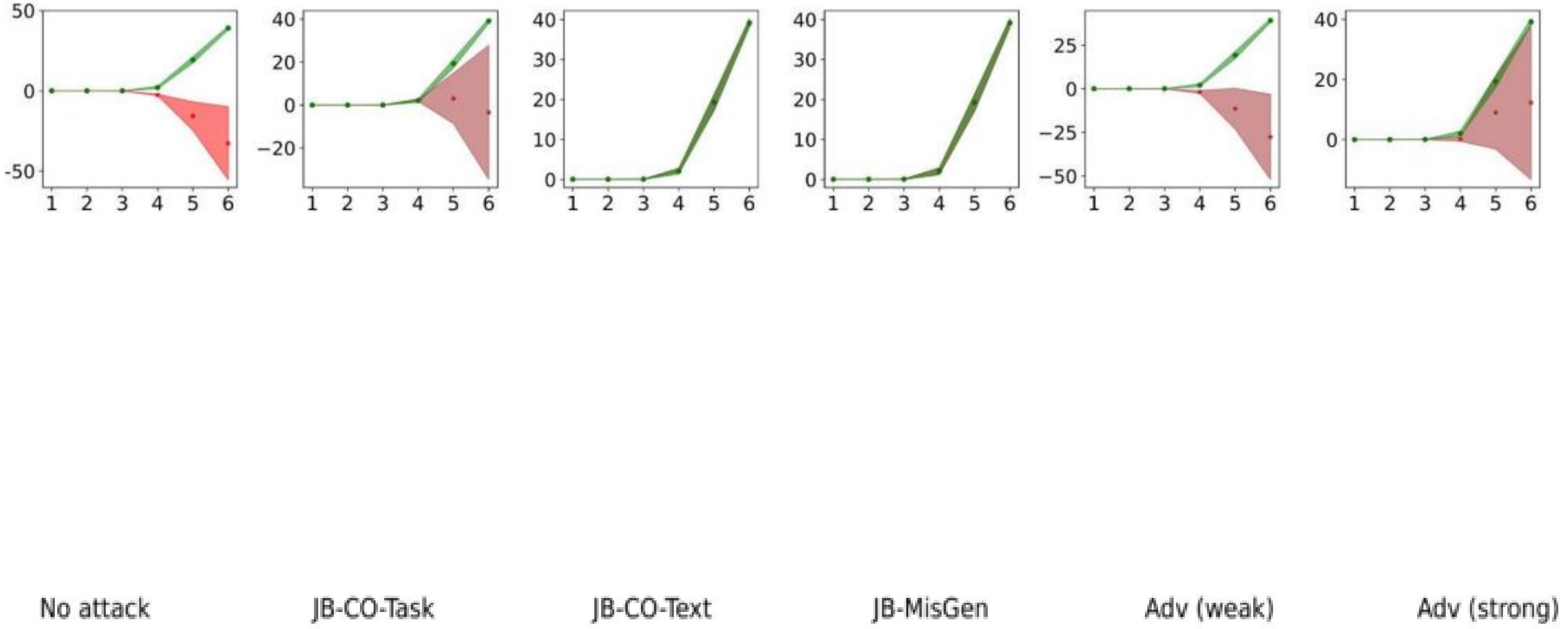


Observation 3

Pre-activations of unsafe inputs have a larger projection onto the row-space $\mathcal{R}(\Delta W)$ compared to pre-activations of safe inputs. That is, ΔW preferentially impacts unsafe samples.

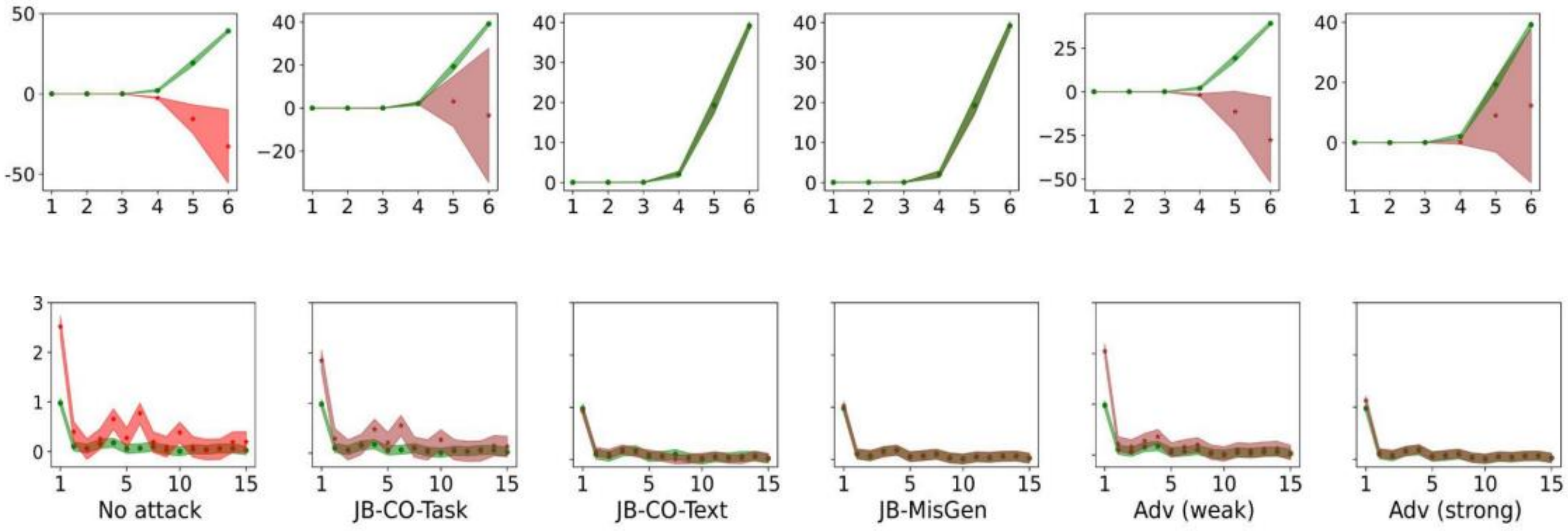
Understanding why safety fine-tuning fails?

Feature space analysis



Cluster separation decreases with increase in attack strength

Understanding why safety fine-tuning fails?



Feature space analysis

Row space analysis

Jailbreaks evade the null space projection by ΔW , thus ΔW is not able to generalize to them.

Observation 5

Jailbreak and adversarial attacks yield intermediate features that are exceedingly similar to safe samples, hence evading the processing by ΔW required for refusal of an input.

Thank You