Distillation Robustifies Unlearning

Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath,
Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, Alexander Matt Turner

Presenter: Bruce W. Lee
UPenn, MATS

November, 2025

ML ALIGNMENT
& THEORY SCHOLARS

1. Introduction

- LLM unlearning is not robust.

- A few steps of finetuning can revert their effects.

- But what if our unlearning target was too risky to fail?

- We need an unlearning method that truly removes the risky capability from the model weights.

1. Introduction

- Our observation: Distillation robustifies unlearning.

- Adding a distillation step after unlearning gives us a model that retains the desired behavior but not the undesired capabilities.

- Unlearn-and-Distill
    - Step 1: apply unlearning methods (GradDiff, MaxEnt, or RMU) to a pretrained model
    - Step 2: distill this unlearned model into a randomly initialized model of identical architecture

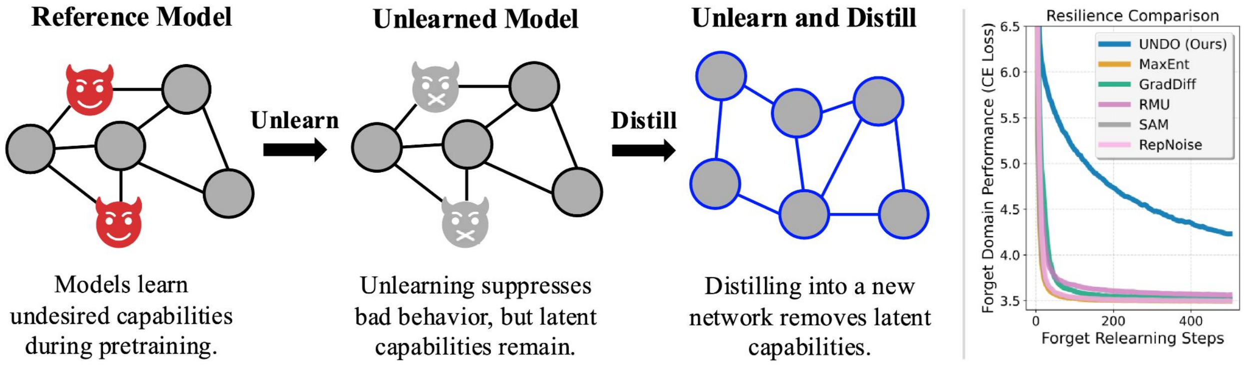
Reference Model
Unlearn
Unlearned Model
Distill
Unlearn and Distill
Models learn undesired capabilities during pretraining.
Unlearning suppresses bad behavior, but latent capabilities remain.
Distilling into a new network removes latent capabilities.
Resilience Comparison
Forget Domain Performance (CE Loss)
6.5
6.0
5.5
5.0
4.5
4.0
3.5
0
200
400
Forget Relearning Steps
UNDO (Ours)
MaxEnt
GradDiff
RMU
SAM
RepNoise

Figure: Distillation robustifies unlearning.

2. Unlearn-and-Distill

ML ALIGNMENT & THEORY SCHOLARS

- Robust unlearning is difficult to achieve just by finetuning the input-output behavior of the model.

- We show that this is true even in an idealized unlearning setup. Three models were involved:
  - Oracle: Gemma 2 pretrained with a perfectly filtered dataset of only the desired capabilities.
  - Reference: Gemma 2 is first pretrained with both desired and undesired capabilities then matched to Oracle Teacher.
  - Random: Randomly-initialized Gemma 2 is matched to Oracle Teacher.

- Our insight: Reference model relearns the unlearned capability much faster than Random.

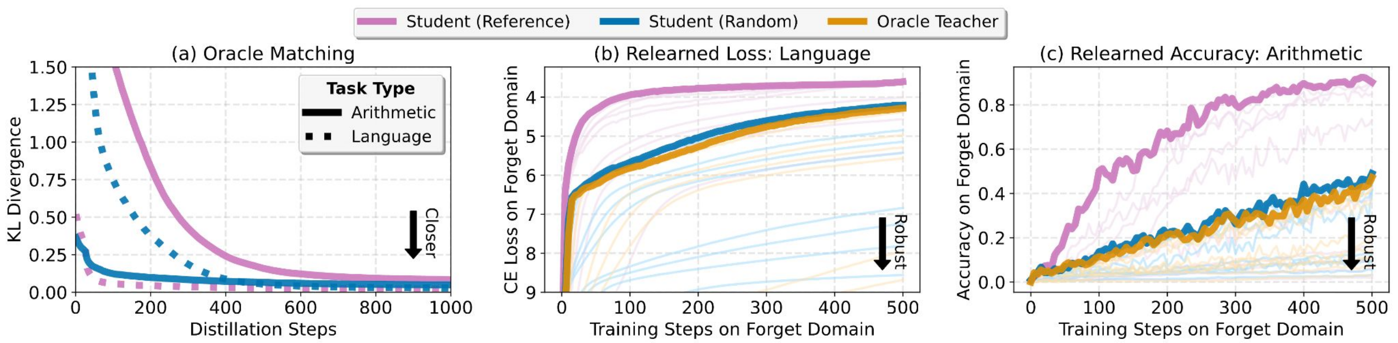

Figure: Matching oracle behavior doesn't guarantee robust unlearning.

2. Unlearn-and-Distill

- The same phenomenon could also be observed with the approximations of the oracle teacher.

- Unlearning then distilling the model into a randomly initialized model gives us a robustly unlearned model.

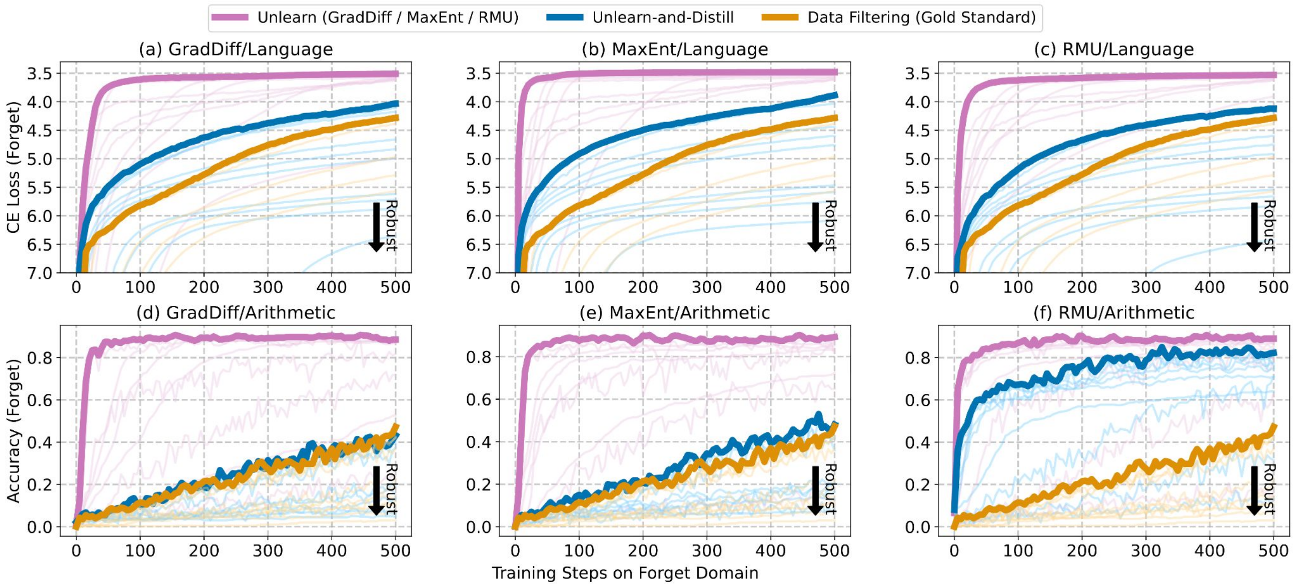

Figure: Unlearn-and-Distill boosts robustness to relearning.

3. UNDO: Unlearn-Noise-Distill-on-Outputs

- The same phenomenon could also be observed with the approximations of the random student and the oracle teacher.

- UNDO: Unlearn-Noise-Distill-on-Outputs
    - Step 1: apply unlearning methods (GradDiff, MaxEnt, or RMU) to a pretrained model
    - Step 2: create a student model by perturbing the weights of the model from (i)
    - Step 3: repair the damaged student from (ii) by distilling to recover the teacher's original behavior from (i)

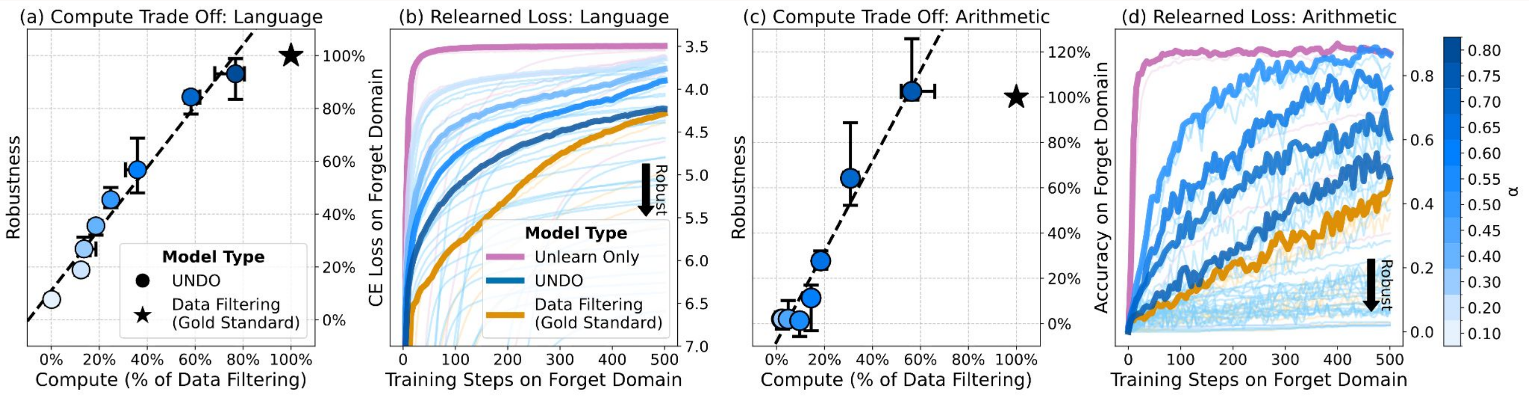

Figure: Unlearning robustness scales with more perturbation.

Thank you!