A. Ablation Study

In this section, we explain the results of our ablation study to understand the contributions of the different components of our defense. The improvements offered by our defense can be decomposed into two components: 1. The misinformation function $\hat{f}$ which has been trained with reverse cross entropy loss to provide incorrect predictions. 2. The Out of Distribution detector which gradually switches between predictions of f and $\hat{f}$ depending on whether the input is OOD.

To understand the contributions of these two components, we plot the defender accuracy vs clone accuracy on the LeNet network trained on the MNIST dataset (Fig 6) with the following configurations of the Adaptive Misinformation (AM) defense:

1. Baseline: This is the default AM defense without any modifications, which serves as the baseline for our ablation study.

2. Random $\hat{f}$: We replace the misinformation function $\hat{f}$ with a randomly initialized network. We find that there is a degradation in the trade-off curve compared to our baseline, indicating that training $\hat{f}$ to produce incorrect predictions indeed offers better security

3. Removing Adaptive Mechanism: We remove the OOD detector and the adaptive mechanism by which our defense switches between the predictions of f and $\hat{f}$. Instead, we use a simple perturbation based scheme by taking a linear combination of f and $\hat{f}$ as shown in Eqn 12. Additionally, we use a network that is not trained with Outlier Exposure [7] (OE) for this run.

$$y' = (1-\alpha)f(x;\theta) + \alpha\hat{f}(x;\hat{\theta}) \tag{12}$$

Our results show that without the adaptive mechanism, there is a steep drop-off in accuracy, even while the clone accuracy is high. This is because, with simple perturbation based schemes, the benign examples are affected by noise leading to degradation in defender accuracy. However, since the perturbed predictions y' remain correlated to the original predictions y, the clone accuracy remains high. Additionally, we also find that the clone accuracy obtained with an undefended model is lower with our scheme. E.g. in case of FashionMNIST, the clone accuracy for the undefended model in the baseline case is 39.47%, as compared to 65.87% without the adaptive mechanism. This difference is due to the use of OE in the training of f in our defense. OE encourages the model to generate uniform distribution for OOD data. Thus an adversary querying the model with OOD data obtains less information, even in the undefended case, resulting in a degradation of clone accuracy.

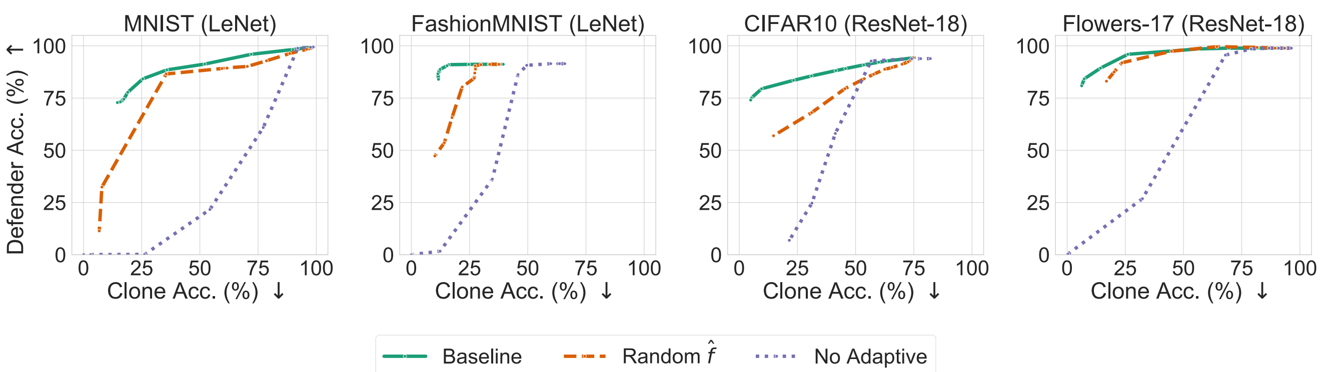

Figure 6. Ablation Study: Comparison of Defender Accuracy vs Clone Accuracy trade-off curves with three configurations of the Adaptive Misinformation Defense: (a) Baseline (b) Random $\hat{f}$ (c) No Adaptive Mechanism