A Broader Impact

Our work designs privacy attacks, which have the potential to cause harm. However, by making the vulnerabilities in existing approaches known, and more rigorously evaluating the risk to users, our work is a necessary step to designing stronger mitigations in the future.

B Limitations

The main limitation of our work is the strong threat model under which our attacks work. We use the same threat model as the "online" version of the LiRA [CCNSTT22] attack. This attack assumes access to the target examples before training a model, and, for the End-to-End LiRA attack, access to the student dataset. We use this strong threat model to assess worst-case vulnerability, as prior work has evaluated distillation under weaker attacks. Another limitation of the attacks is the large running time of the student query attack from Section 5, which requires thousands of shadow models to obtain good performance. While generally impractical, we prefer to position this attack as a way of explaining why distillation propagates membership information, and leave future work to attempt to improve the attack's efficiency.

C More Experiment Details

All of our results on CIFAR-10 make use of fewer than 30000 trained models. While a very large number of models, the fast, publicly available training code we use allows us to train this number of models in fewer than 1 GPU-week (although we decrease the wall-clock time by parallelizing over 4 GPUs). Our results on Purchase-100 and Texas-100 also use simple models, taking under 1 minute to train (we train all models for 20 epochs with SGD with a learning rate of 0.01 and momentum parameter of 0.99, which we found to maximize performance over our hyperparameter sweep). We train 8000 of these models for our analysis, taking fewer than 1 GPU-week for each of these datasets. Our most expensive attack, relying on only student queries, starts to outperform random guessing with as few as 100 models, which can be trained on 1 GPU in two hours on all three of these datasets. Unfortunately, we are unable to make our code public at this time due to organizational constraints.

D Extended Results on Teacher Dataset Privacy

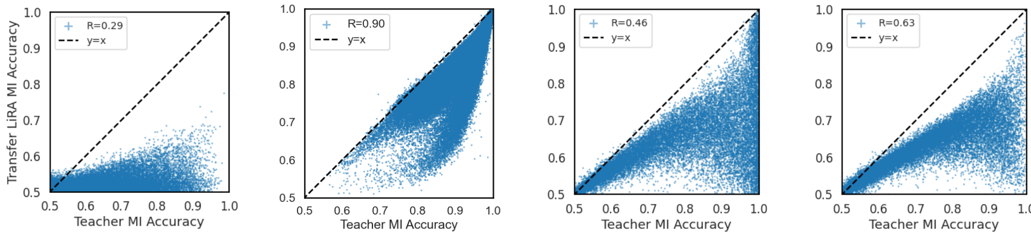
We plot the effectiveness of Transfer LiRA in Figure 7. ROC curves for our student attacks are found in Figure 8. Further qualitative examples can be found in Figure 9. Ablation of score information with and without duplicates is plotted in Figure 10. Per-example student attack success rates for CIFAR-10 with duplicates are found in Figure 11. In Figure 12, we compare our student model attacks against a simple logit threshold baseline, similar to the loss thresholding attack designed by Yeom, Giacomelli, Fredrikson, and Jha [YGFJ18], which was used to evaluate distillation privacy in Shejwalkar and Houmansadr [SH21].

E Privacy of Student Training Set

Having evaluated the Private Teacher threat model, we now turn to the Private Student and Self-Distillation threat models, which we will consider simultaneously. The Private Student threat model can be used to perform knowledge transfer from large, general purpose models to task-specific models, by querying on (sensitive) task-specific student data. Self-distillation is often used in applications of distillation to compress models and improve their performance.

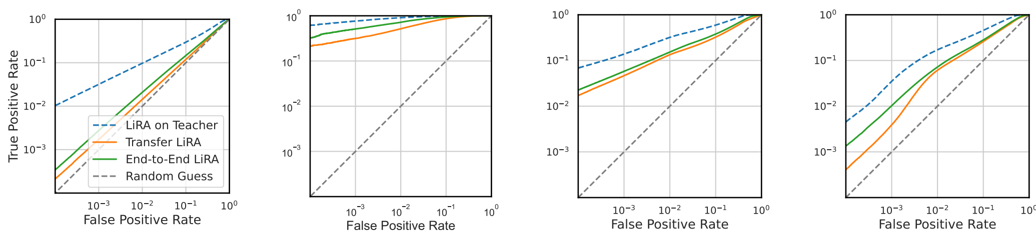
E.1 Private Student

The private student threat model does not involve data minimization, unlike the private teacher threat model; the empirical privacy we investigate here comes instead from an adversary having limited knowledge of the specifics of the teacher model. That is, the question we investigate is: how much does the adversary need to know about the teacher model to get reliable attacks on the private student dataset?



(a) CIFAR-10 (b) WikiText (c) Texas-100 (d) Purchase-100

Figure 7: **Many data points do not get privacy benefits from distillation.** With the x axis, we plot the vulnerability of each teacher example to attack before distillation, using teacher models. With the y axis, we plot the vulnerability to attack after distillation, using the Transfer LiRA strategy to attack student models. Observe that many data points lie near the $y = x$ line, which indicates no reduction in vulnerability from distillation.



(a) CIFAR-10 (b) WikiText103 (c) Texas-100 (d) Purchase-100

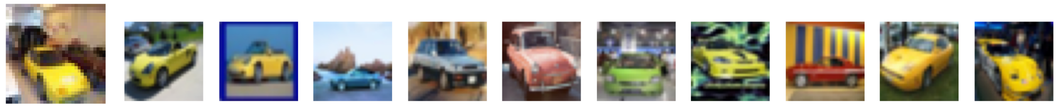
Figure 8: ROC curves for our attacks on student models.

We consider three levels of adversarial knowledge: **Known Teacher**, where the adversary knows the precise teacher model used to query the student examples; **Unknown Teacher**, where the adversary knows the teacher model is one of a small subset of models; and **Surrogate**, where the adversary can only collect similar data, to train their own surrogate teacher models. Both the Known and Unknown Teacher settings reflect a world where the teacher model is one of a small number of general purpose public models, such as a large language model. The Surrogate setting requires the adversary to train their own copy.

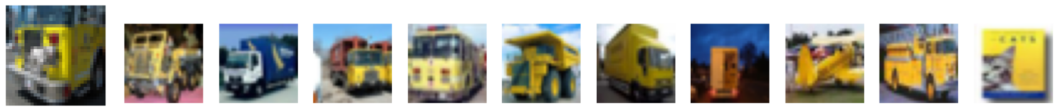
We run the LiRA variants in a number of these settings on the CIFAR-10 dataset, calibrated to the knowledge the adversary has (for example, in the Surrogate threat model, the adversary trains their own teacher models, and trains a number of shadow student models to calibrate LiRA). We plot our results in Figure 13a, and find that, as expected, less knowledge about the teacher model reduces the adversary's success at membership inference. However, even the weakest threat model, Surrogate, allows for powerful attacks, with a TPR as large as 10^{-2} at a FPR of 10^{-3}.

E.2 Privacy of Self-Distillation

Having considered the privacy of the student and teacher datasets independently, we now investigate the common self-distillation setting [FLTIA18; XLHL20], where the student and teachers are identical. Given that duplicate examples in the student set carry membership information of teacher examples (Section 6.1), and student examples themselves are not well protected by distillation (Section E.1), we do not expect self-distillation to reduce privacy risk significantly. However, a common technique in self-distillation is to train the student on a loss function which combines the cross entropy loss on the query dataset ℓ_Q with the cross entropy loss on the student examples' original "hard labels" ℓ_S. We write $\ell_\alpha = \alpha\ell_Q + (1 - \alpha)\ell_S$, so that $\alpha = 1$ recovers the standard distillation objective, while $\alpha = 0$ recovers the standard cross entropy loss (as if there was never a teacher model).



(a) Target (b) Student Queries



(c) Target (d) Student Queries

Figure 9: Two examples of target examples for which the most informative student queries are predominantly in the same class. The only exception is the eighth student query in (d) for the yellow truck in (c), which is an airplane. The filtered attack using the displayed student queries reaches 78% accuracy on the yellow automobile in (a), and 74% accuracy on the yellow truck in (c).

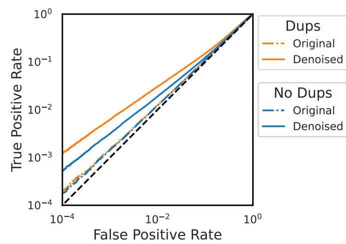


Figure 10: The impact of denoising on duplicated and deduplicated teacher attacks.

To evaluate self-distillation, we run LiRA by training shadow student models with the entire self-distillation algorithm, using identical datasets for each pair of teacher and student shadow models. We perform calibration on these shadow student models, and plot our results at a range of α values in Figure 13b. While we don't observe a large effect, it appears that larger α (that is, heavier reliance on the distillation loss function) results in better attacks. This is likely because relying on the distillation loss function reinforces the memorization from the teacher even further in the second round of training on the student.

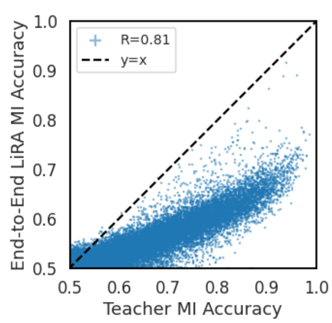


Figure 11: Duplication also has an impact on CIFAR-10 student attacks. Compare with Figure 3a.

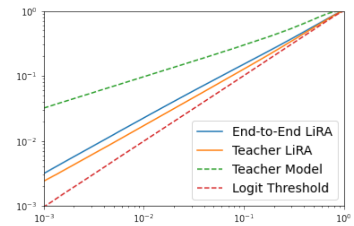
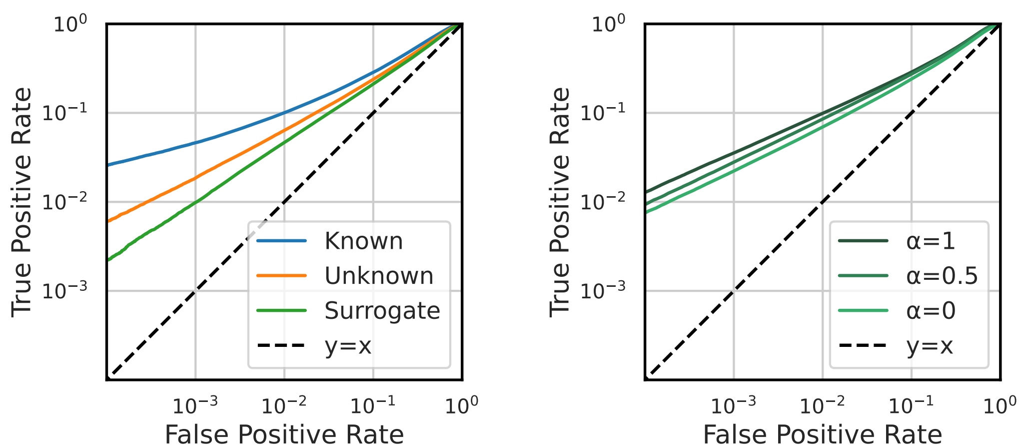


Figure 12: Our attacks outperform a simple logit threshold baseline attack, used by prior work.



(a) Private Student

(b) Self-Distillation

Figure 13: *Distillation has limited ability to prevent membership inference* either a) on sensitive student examples, or b) in self-distillation. However, reducing the knowledge available to the adversary seems to help in the Private Student threat model. Results for both on CIFAR-10.