We thank all reviewers for their helpful feedback. All mentioned minor issues will be fixed in the final version. We appreciate that all reviewers like the idea of certifiable adversarially robust OOD detection and that they are as positively surprised as we were that this is possible with only minor or no loss in test accuracy.

R1: Ambiguity of 'worst case OOD detection'. We agree that the term 'worst case OOD detection' is ambiguous and propose to use 'adversarially robust OOD detection' but we are open for suggestions. The generalization of our guarantees from OOD training to different OOD test sets is indeed an empirical observation and cannot be proven (same for certified robustness on the in-distribution) and we clarify this in the final version.

R1: Confidence thresholding for OOD detection. We modify the relevant sections to highlight that other approaches rather than confidence thresholding exist for OOD detection.

R1/3: GAUC for ACET and other methods. We do not know whether the true values of the WC-AUC lie closer to the GAUC (lower bound) or the AAUC (upper bound) for the different methods, but we believe that both bounds are not tight. We tested for the largest epsilon for $\epsilon = 10^{-k}$ where IBP gives non-zero GAUCs. This is $k = 6$ (ACET), $k = 7$ (OE) and $k > 7$ (Plain). This provides weak evidence that a more robust model is also more verifiable. On the other hand this phenomenon that adversarially trained models are not certifiable is well-known for adversarial in-distribution robustness. However, for adversarial training on the in-distribution it is generally believed that the empirical robustness is close to the true one. In our case we don't think so as adversarial attacks on ACET (and sometimes even OE/CEDA) are much more difficult as the gradients are often very small or even zero and thus some of the attacks might simply fail even though the point is not robust. We discuss these difficulties in Appendix A. For the final version we try to adapt Mixed-Integer-Programming for certification to our setting to further investigate this question.

R1: Pre-filtering of Tiny-Image Dataset as OOD training distribution. The pre-filtering step is an artefact of our initial attempts to stabilize the IBP training which we later on did not challenge again - so thanks a lot for this question. In fact the inclusion of the 4.25M discarded images into the training set does not change any of our results (very minor positive and negative changes with no trend). Thus, for the final version we will report all results for the full TinyImage dataset and only exclude CIFAR in order to be comparable to the setting of Outlier exposure (OE) [19] (we can even include the 132K CIFAR images into training - for our quantile-based loss it is no problem if there is a small overlap of in- and out-distribution as it was exactly designed for this - but then we could not use CIFAR-10/100 for evaluation).

R2/3: Form of 'Confidence Upper Bound Loss' $\mathcal{L}_{\text{CUB}}$ (line 134). The logarithm also has an effect on the gradient: it leads for large upper bounds roughly to a rescaling of the gradient of the upper bound on the confidences by the actual upper bound and thus is essential for a better behaved training. The square inside the loss was chosen as this leads to more uniformly small upper bounds over all OOD images compared to not squaring (similar to l_2- vs l_1-loss). Higher order polynomials would not be of additional help here and might lead to numerical problems. We found that the omission of either the $\log$ or the squaring, or both, makes training less stable.

R2: Train deeper architectures e.g. ResNets. We have adapted our IBP training to a Fixup ResNet-20 without BN and Dropout. Interestingly it is possible to train these deeper models and obtain similar but worse GAUCs than with our XL architecture. Since the worse results might be due to an suboptimal schedule or initialization, we are currently trying to identify and resolve the cause of this performance gap.

R2: Larger ϵ-balls. Note that we evaluate AAUC/GAUC of our models at larger radii in Appendix I (Table 9) and the guarantees partially generalize. For the rebuttal we trained models on CIFAR-10/SVHN with $\epsilon = 8/255$. On CIFAR-10 GOOD$_{80}$ has an accuracy of 90.6 (notably accuracy is still unaffected) and achieves clean AUCs of 78.6/95.0/89.9/96.2 (CIFAR-100/SVHN/LSUN/Uni) vs. 85.9/95.6/96.2/90.4 for the GOOD$_{80}$ model trained with $\epsilon = 0.01$ and GAUCs at $\epsilon = 8/255$ of 43.2/17.9/51.1/94.0 vs. 36.7/34.4/48.4/86.9 so the resulting changes appear inconclusive. For SVHN the differences are stronger. The GOOD$_{100}$ model trained for $\epsilon = 8/255$ achieves an accuracy of 96.0 (vs. the 96.6 of GOOD$_{100}$ trained with $\epsilon = 0.01$) and has clean AUCs of 99.5/99.7/99.9/100 (CIFAR-100/CIFAR-10/LSUN/Uni) vs. 99.9/100/100/100 as well as GAUCs for $\epsilon = 8/255$ of 96.0/97.3/98.5/100 vs. only 40.3/41.3/40.3/1.5 for GOOD$_{100}$ trained with $\epsilon = 0.01$. This confirms the intuition of R2 that for SVHN a larger ϵ on the OD is feasible. We include all results in the final version and explore even larger radii.

R2: Perturbations on in-distribution inputs. We used IBP to compute lower bounds on the confidence in the original class around in-distribution points and got non-trivial certificates only for very small radii. Note that certified adversarial robustness on the in-distribution [13] comes with a larger drop in test accuracy which we want to avoid.

R2: On the fly curriculum for the quantile q in GOOD. Thanks, we thought about this but discarded it in favor of an ablation study on the effect of the quantile-loss, but it is definitely an interesting direction to pursue.

R3: Definition Confidence/Clarification of Fig. 1. In line 67, we define confidence of an input x as the maximum of the predicted probability distribution over the classes, which for all models happens to be realized by *dog* in all cases. Thus the probabilities of all other classes are lower than the one of *dog*.

R3: Effect of GOOD on AUC on MNIST/SVHN vs CIFAR-10. Since CIFAR-10 is a subset of TinyImages and thus in- and training out-distribution are more similar, the task of provable OOD detection is significantly harder for CIFAR-10 than for MNIST/SVHN which affects the clean AUC (see lines 261-267). Since ACET directly optimizes empirical robustness, it unsurprisingly tends to have good AAUCs. For discussion of ACET/CCU see Appendix B.