We thank the reviewers for their insightful comments. A number of additional experiments were suggested. We have completed each of these experiments, and we have added them to our current version. We first enumerate these experiments and then respond to individual reviewer questions. The details of these experiments are as follows:

A1. Transfer learning with a deeper backbone: We have now tested transfer learning with an adversarially trained ResNet-50. The deeper backbone actually decreases robust accuracy on few-shot tasks compared to the much shallower R2-D2 backbone (15.63% vs. 17.72% on 1-shot mini-ImageNet). We have updated our paper to include these results.
A2. Additional attack norm: We have now implemented the ℓ_2 PGD following Madry, and we find that AQ models are also robust to these attacks. For example, our adversarially queried R2-D2 model achieves 35.53% robust 5-shot accuracy on mini-ImageNet, while the analogous transfer learned model achieves only 15.92% on the same task.
A3. Out-of-distribution testing: We have now evaluated our models on Meta-Dataset. For example, on the FGVC-Aircraft dataset, R2-D2 AQ gets 36.65% 5-shot robust accuracy, while the analogous transfer learned model gets 20.83%. For the same models, on the CUB-200 dataset, the robust accuracies are 29.04% and 20.05%, respectively.
A4. Additional meta-learning algorithms: We have now implemented and run adversarial querying with the state-of-the-art meta-learning method, MCT. This experiment yielded high robustness and small natural accuracy tradeoff. MCT (AQ) simultaneously achieves 79.9% natural accuracy and 53.2% robust accuracy on 5-shot mini-ImageNet.
A5. Reptile: We have trained an adversarial Reptile variant, and we are updating our paper to include these experiments.

Reviewer 1:
"*Multiple small tables makes the paper difficult to read... Table 9 could have been included along with the main results.*" Thank you for your suggestions. We have merged tables and highlighted these results in our current version.
"*It seems that perturbing support data does not provide any advantage.*" Thank you for pointing this out. In Section 4.2, we note that perturbing support data optimizes the network for adversarial fine-tuning. Likewise, in Table 8, we see that models trained with adversarial support achieve better robustness than the naturally trained model (still worse than AQ models) only when adversarially fine-tuned at test time. We have added a subsection to better explain this phenomenon.
"*It is unclear why AQ is more robust than transfer learning.*" Ding et al. 2019 found that unlike natural accuracy, robustness does not generalize well under transfer learning. AQ produces models specifically optimized for robust few-shot adaptation. We have updated our paper to discuss this issue in detail.
"Attacking only support data can be seen as maximizing clean test accuracy when fine-tuned in a robust manner" The meta-objective in this case is clean query loss after optimization on adversarial support data. This optimization simulates natural performance at test-time after fine-tuning on adversarial few-shot data.
"*We expect adversarial accuracy to increase as $\frac{1}{\lambda}$ increases.*" Thank you for pointing this out. When $\frac{1}{\lambda}$ is sufficiently large, the network is encouraged to behave as a constant function and to neglect accuracy (both natural and robust). Consider that a network need not make correct predictions to achieve low KL divergence loss.
Reviewer 2:
"*Only four meta-learning algorithms are tested.*" We have now tested adversarial querying on the state-of-the-art MCT method, and we have updated our paper accordingly. See A4 above for additional details.
Reviewer 3:
"*Show at least one visual case.*" We agree that visualizations could make our work easier to understand. We have now both visualized adversarial examples and constructed a visual representation of our algorithm.
Reviewer 4:
Regarding your suggestions concerning table aggregation and captioning, table referencing, and moving Algorithm 1 to the Appendix, we agree with your assessment, and we have updated our current version to reflect these changes.
"*I am surprised at the difference in A_{adv} values of "MAML adv. query" and "MAML adv. query and support"... perturbing the query data and not the support data*" Thank you for pointing this out. In Section 4.2, we note that perturbing support data optimizes the network for adversarial fine-tuning. Likewise, in Table 8, we see that models trained with adversarial support achieve better robustness than the naturally trained model (still worse than AQ models) but only when adversarially fine-tuned at test time. We have updated our explanation of this phenomenon in Section 4.2.
"*AQ causes a big drop [in natural accuracy]*" This massive trade-off exists in the standard setting where SOTA robust ImageNet models achieve 65.30% clean accuracy while clean-trained models achieve 88.50%. We agree that this is an obstacle for deployment. In our new MCT tests (see A4 above), we see a far smaller drop of only 6%.
Reviewer 5:
"*...results on some datasets and not others. For example: section 4.2 presents results on only 5-shot ImageNet.*" The appendix in our submitted version contains these experiments. We agree that these experiments are important.
"*...meta-learners that minimize *support* set loss.*" Thank you, we have now included your suggestion (see A5 above).
"*test out-of-distribution generalization.*" We have now evaluated our models on Meta-Dataset (see A3 above).
"*A few minor points...*" Thank you for pointing these out. We have added clarification to our current draft.
"*...considerably worse when no attack is present.*" This massive trade-off exists in the standard setting where SOTA robust ImageNet models achieve 65.30% accuracy while clean-trained models achieve 88.50%. We agree that this is an obstacle for deployment. We like your average performance metric and have updated our current version to include this.