We thank all reviewers for their time taken to review our work and their helpful comments and suggestions to improve this work. As a recap, in our work we studied the problem of returning all ϵ-good arms. We establish lower bounds for the hardness of this problem and present two algorithms with theoretical guarantees, FAREAST and $(\texttt{ST})^2$. The former is asymptotically optimal on all instances. Additionally, via a novel moderate confidence bound, we are able to show that the latter, which enjoys great empirical performance, is optimal in many settings, such as the common $\delta = 0.05$. Below, we respond to individual review comments individually.

Reviewer 1: Thank you for your review. Addressing your comment about weaknesses and novelty: The core challenge of this work stems from the need to estimate the threshold $\mu_1 - \epsilon$ to sufficient precision which is new to this problem. This motivates our algorithms and affects our lower bounds. In particular, we develop a new theoretical technique for the lower bound in Appendix D to handle this that may be of independent interest. We will highlight this in future drafts to make this apparent. Thank you for the note about the abstract. We will be sure to clarify that the nature of our results are to provide sample complexity bounds. Thank you for pointing out the confusion regarding Figure 4a. In the sample data, there were 2.2M total ratings used to estimate the means. Before this number of samples, we draw the performance curves of each algorithm as a solid line, and after this point as a dashed line. This was to highlight that $(\texttt{ST})^2$ would have performed better in practice as well even at 2.2M ratings. We will clarify this in the final version.

Reviewer 2: Thank you for your careful read and comments about our paper. Regarding the lower bounds, Theorem 2.1 follows using standard techniques. To prove Theorem 4.1 however, in section D.2 we develop a novel technique that allows one to reduce a more general composite hypothesis test to the more limited set of tests covered via the Simulator lower bounding technique of [1], and this may be of independent interest. We will highlight and expand upon this in future drafts to make it more readily accessible. Regarding more general distributions for the lower bound: for Gaussians of unknown variance σ^2, the lower bound of Theorem 2.1 still applies (with an additional factor of σ^2). Intuitively this is since the known variance case should be an easier problem. Whether is is possible to achieve this bound or if a tighter bound can be proven when σ is unknown is an interesting question for future work. For distributions beyond Gaussians, the result can easily be altered for any distribution with a well defined mean. In this case, terms such as $(\mu_1 - \epsilon - \mu_i)^2$ present in the bound will be replaced by KL-divergences between the appropriate distributions. Finally, we agree that a fixed-budget algorithm would be interesting to study and a useful setting in practice. We hope to study it in follow up work. Naively, $(\texttt{ST})^2$ could be altered to perform anytime fixed budget borrowing from techniques in [2] though this may not perform particularly well in practice and is not optimal. It is not immediately clear how to incorporate the ideas of FAREAST, specifically the technique in the Bad Filter, for an optimal fixed-budget algorithm and would be an interesting challenge for a future work.

Reviewer 3: Thank you for your kind review. We will correct the typos that you noted.

Reviewer 4: Thank you for taking the time to review our work. Regarding using the F1 score in experiments, F1 equaling 1.0 is equivalent to an algorithm exactly identifying the set of all ϵ-good arms correctly - the objective of this work. Visualizing the F1 score as a function of the number of samples serves to demonstrate the expected sample complexity of the algorithm until $F1 \approx 1$, and also provides a continuous measure of the performance of the algorithm. In particular, even if $(\texttt{ST})^2$ has not yet found all ϵ-good arms, it still achieves high F1 scores relative to other methods. Addressing the question about the correct probability directly, for synthetic experiments, we observed that the algorithms always correctly returned the set of all ϵ good arms at termination. For the Caption Contest data experiments in Figure 4a, we ran with $\delta = .1$ and terminated at 10M pulls even if the stopping criteria had not been satisfied. In those experiments, we observed that in almost all ($> 99\%$) cases, the F1 score of $(\texttt{ST})^2$ was 1.0 around 10M pulls and beyond, indicating that the algorithm correctly identified all ϵ-good arms by this point and would terminate correctly. Hence, $(\texttt{ST})^2$ would achieve a correctness probability near 1 in this experiment - significantly better than $1 - \delta = .9$. In the case of the cancer dataset, we ran a small fraction to termination as well as with other values of ϵ, and noted that the algorithm always returned the set correctly in all cases.

[1] Simchowitz, M., Jamieson, K., Recht, B. *The Simulator: Understanding Adaptive Sampling in the Moderate-Confidence Regime*. Conference on Learning Theory, 2017.

[2] Jun, K.S., Nowak, R. *Anytime Exploration for Multi-armed Bandits using Confidence Information*. International Conference on Machine Learning, 2016.