We thank all the reviewers for their comments, and for recognizing the novelty and significance of our work. We ask that reviewers increase their score to acceptance if they are satisfied with our response.

Common response re Natural distributions: We agree that 'Natural Distributions' is not mathematically precise, and we devote Section 3.4 to discuss this limitation (which is common to many empirical papers in deep learning, not only our work). In the paper, it is used to mean that our conjecture holds for any distribution deriving from real data that we tested, but can be broken with carefully contrived synthetic examples. We hope that our experiments establish that this is an interesting and relevant phenomenon, which opens a new area in the study of generalization.

R8: 'Estimating L': We can't enumerate all possible distinguishable features (L), since there may be exponentially many, but we can do the next best thing: Given a candidate L, we can check if it is valid. This follows because the definition of a distinguishable feature is testable (by training on the labeling induced by L).

R9: We are glad you found our results "intriguing and fundamentally important". Re 'Natural distributions': Please refer to common response. We believe the conjecture is useful even if we don't yet know which distributions it breaks for, since it highlights novel behavior on many common tasks. In particular, our conjectures held on every distribution we tested where the inputs $p(x)$ came from "real data", even when the labels $p(y|x)$ were arbitrary. We will clarify this in the paper. 'Do the labels y need to be distinguishable': No, the conjecture holds for arbitrary label distributions $p(y|x)$. (eg: See Figure 3B on CelebA where y are not distinguishable) 'Closeness of LHS, RHS': We define this explicitly in Section 2. The closeness ($\approx_\epsilon$) of LHS and RHS in Eq. 4 is in TV-distance as we state in Section 2 Line 180. We also plot the TV-distance with ϵ in Figure 3C to show that this closeness obeys the quantitative relationship from our conjecture. The closeness of LHS and RHS is unrelated to the 'natural distributions' part of the conjecture - the closeness is always TV-distance between (L, y) and $(L, f(x))$. The natural distribution only affects the choice of source distribution D. 'Bootstrapping': Yes, we don't claim to understand all possible training algos. But it's interesting that our observations hold for "standard" training procedures. Understanding more advanced procedures (bootstrapping, ensembling) is out-of-scope for this paper, but an important problem for future work.

R11: 'Implications': We discuss the significance of our work in detail in Section 1.3. We do not claim to show any immediate practical benefits for practitioners (as is true for many papers in the field of deep learning theory), but to gain deeper understanding of our methods. 'Mechanisms': We strongly disagree that the paper should provide plausible mechanisms for it to be accepted. The first step in scientific study is often to identify and establish an empirical phenomena/conjecture. Once the empirical behavior is understood, it can help guide a deeper theoretical understanding. For example, the implicit regularization in matrix factorization was first conjectured in [1] and was supported with experimental evidence and some analysis for a very limited case. However, that encouraged many researchers to work on that direction and various version of the conjecture was proved, eg. see [2]. Our work is the first step, and serves to introduce the novel idea and establish it experimentally. 'Robustness to optimizers/losses': We provide experimental evidence for different losses in kernels (SVM vs MSE), and different optimizers (Adam and SGD). 'Student-Teacher setup': We agree that this will be an interesting setup to understand this phenomenon further, but this is beyond the scope of this paper. 'Locality': We do not claim that the locality is the exact mechanism behind this phenomenon, only an intuition. Establishing exact mechanisms is beyond the scope of this paper. 'Why interpolation is required': We discuss this in Section 5.2. Roughly, Distributional Generalization does not require interpolation, but the Feature Calibration conjecture specializes to interpolating models (for reasons discussed in Sec 5.2).

[1] Gunasekar et. al. "Implicit regularization in matrix factorization." [2] Li et. al. "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations."

R13: 'More tests': We believe that more distributional tests are possible and are an interesting direction for future research. 'Sanity check': We are not sure what the reviewer means by 'effectively smaller no. of samples'. The total samples for 'cat' is the same at 5000, and the drop in performance matches that of the train set as predicted by our theory. 'Figure 1': We are unsure that we understand the reviewer's question. The test error (measured wrt the clean distribution) is evident from Figure 1: 0.1 x 0.0 (for all classes except cat) + 0.1 x 0.02 (for cats) = 2% error. 'Applied automatically to medical images': We don't believe it can be applied automatically, but an example prediction could be "If there is a drain in chest X-ray images, and the drain is a distinguishable feature for the ML model, then the outputs of the model (say whether a patient has cancer) for images with a drain will have the same frequency in the train and test set". We believe that this is a useful characterization of model behavior that can help practitioners check for such biases post-hoc. 'Overparameterization / Bayes Optimality': Our claim is that overparameterization does not lead to Bayes optimality *in the presense of label noise.* Without label noise, it does become Bayes optimal. 'Adversarial images': Our theory is only characterizing behavior of models 'on-distribution', but adversarial inputs are 'off-distribution' and thus are not relevant for our conjectures.