We thank the reviewers for their time and constructive feedback to improve the paper.

R1: *-Depth dependence*: We agree that understanding the role of depth is another interesting analysis. The main reason we chose these particular depths was because the performance was near optimal, for a given architecture class. For FCNs on classifying MNIST / CIFAR-10, we found that depth does not benefit performance as much as in CNNs and optimal performance is often achieved in shallow networks. This is consistent with [4, Table 1]. On the other hand, CNNs certainly benefits from depth as shallow networks' receptive fields cannot cover the whole image (with 3x3 filters with stride 1). Note that prior works considering similar architectures [26, Table 1; 40, Tables 1-4] have observed roughly constant performance of finite and infinite CNNs on CIFAR-10 across depths from 5 to 20, indicating that increasing depth without other changes in this specific setting would not yield dramatic improvements.

-Usefulness of learned representation: While it is widely believed that learned representation are important for deep learning, and there's implicit evidence based on transfer learning, it is not yet proven to be useful for all deep learning models. We argue that studying the relationship between infinite and finite neural networks provides a lens to study the utility of learned representations. Our empirical findings suggest that FCNs tend not to learn a useful representation by simple SGD training. Moreover, we believe that depth is not a major factor here, at least within the scope of our experiments, since for FCN both tuned finite and infinite networks show better performance at shallow depth.

-Matrix size: Since we compute all pixel-pixel covariance for each pair of inputs for CNN-GAP, the internal matrix size is $(6 \times 10^4$ (# of train+test images) $\times\ 32 \times 32$ (# of pixels) $\approx 6 \times 10^7)^2$.

-Finite width correction of [Yaida 2019]: This is a very interesting suggestion and could be a future work. As far as we know, there is not yet an efficient, scalable implementation for the non-Gaussian corrections of [Yaida 2019] which could be applied to the full CIFAR-10 dataset.

-Network hyperparameters: We used (ReLU) critical initialization for weight variance and small bias variance, and we used 512 channels per layer for CNN base models (128 channels for the "narrow" one, as specified in SM C.1, Figure 9). We will mention these key architectural details described in the SM in the main text.

R2: We thank the reviewer for encouraging and positive feedback on our submission!

R3: - *Architecture choice*: We agree with the reviewer that one could add various architectural components such as batch normalization or residual connections to improve finite network performance. As a side note, the best reference for ResNet on CIFAR10 w/o data augmentation (still with other training tricks) achieves 86.37% [Huang & Sun et al., Deep Networks with Stochastic Depth, ECCV 2016] which is in a similar ball-park to CNN-GAP architecture we study w/o data augmentation.) We emphasize that, in this work we strive to find simple architectures where the infinite width limit is well developed and scalable to carefully study how it relates to corresponding finite networks.

-No description on NNGP/NTK: Our paper indeed assumes familiarity with infinite neural networks and made a judicious choice to not include a review to allow more space for empirical results. We will modify the text to include specific resource suggestions for readers who want to learn more.

R4: *On [Aitchison 2020]*: We thank the reviewer for bringing this reference to our attention. It is definitely relevant and we agree that it should be included in our related work discussion. However, we want to highlight that the focus of that paper is distinct and with a more limited scope. [Aitchison 2020] focused on the comparison between finite and infinite networks in the *pure Bayesian setting*. By contrast, ours focuses on the comparison of the mean prediction of infinite networks (NNGP/NTK) to finite networks optimized by modern deep learning techniques. In particular we did a thorough (reverse) ablation analysis (Fig 1), exploring the effects and implicit biases induced by large learning rate, early stopping, weight decay, data-augmentation, data preprocessing (ZCA), parameterization methods, and readout strategies (pooling vs vectorization).

-Lack of a single strong point: Due to the nature of our empirical study, we did not intend to make a single strong point. We want to emphasize that our goal is to perform a *thorough* scientific investigation of finite and infinite networks, answering questions about the factors of variation that drive performance of neural networks, clarifying misconceptions in the existing literature, and uncovering a variety of new and surprising phenomena (e.g. nonlinear interactions between L2-regularization, large learning rate and early stopping). We believe the careful experiments in our paper provide a service to the deep learning community, and may inspire future theoretical and empirical work.