A. Appendix

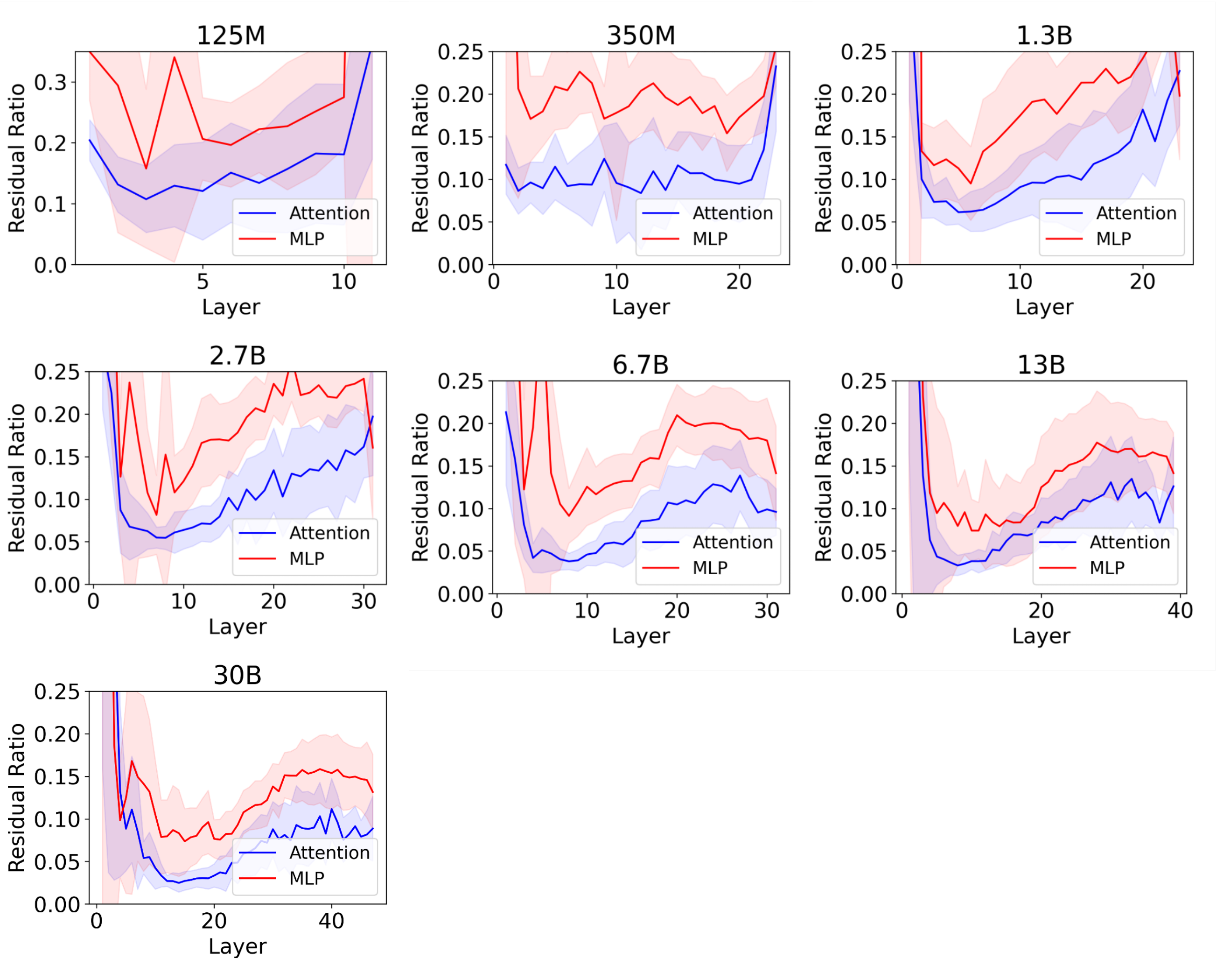

Figure 8. **OPT Layer Residual Ratio –** In general, the residual ratios decrease with larger model sizes, implying that more layers can be more easily skipped. Across models, there is a U-shaped distribution where the first and last layers have the highest ratios and contribute the most to the model output.

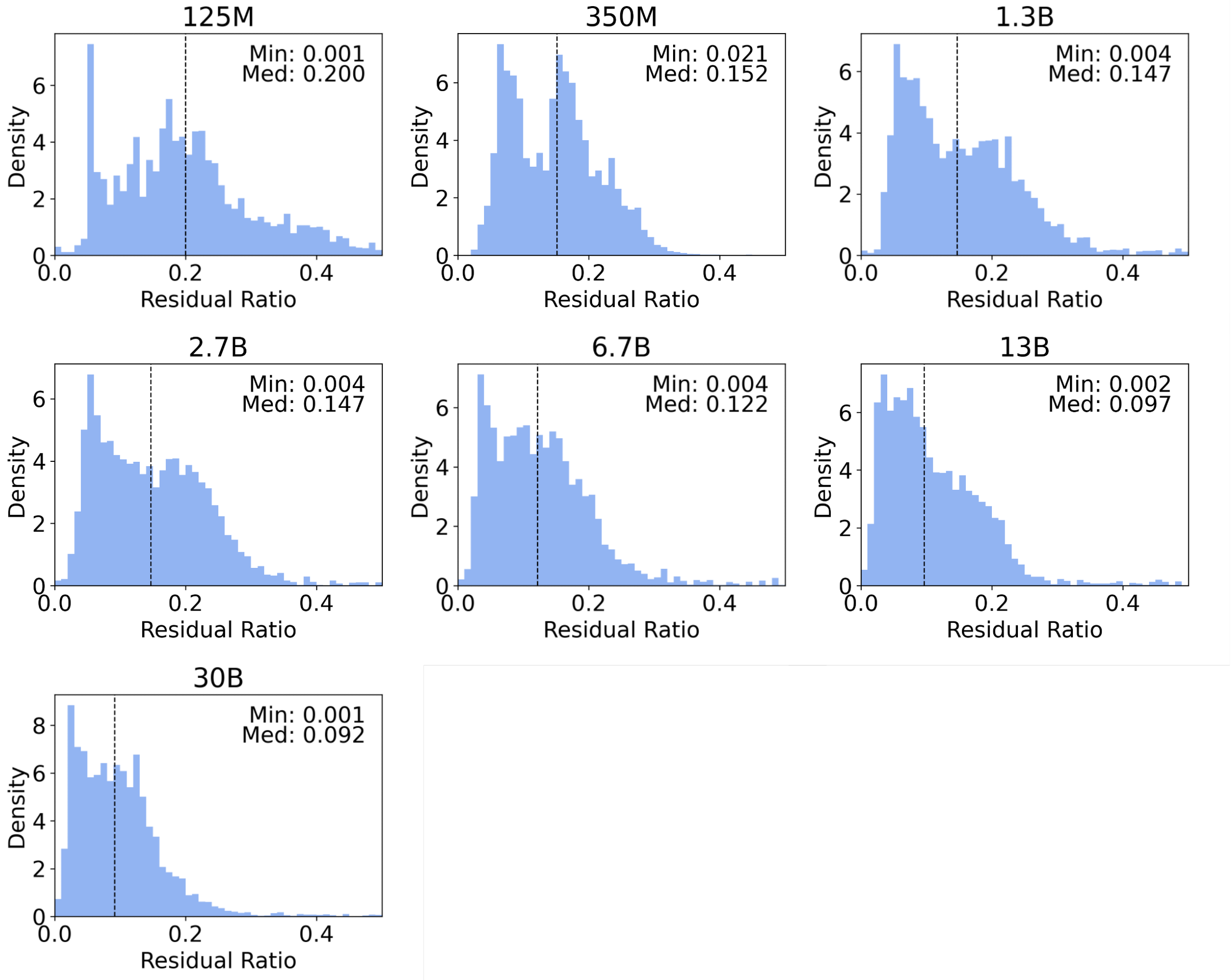

Figure 9. **OPT Residual Ratios –** As models grow larger, the residual ratios shift lower indicating more dynamic layer sparsity within the model. Each value represents the residual ratio from a single layer and single token. Evaluation is done on batches of sequences with length 256 sampled from the WikiText-2 dataset. The dashed line represents the median value in the distribution.

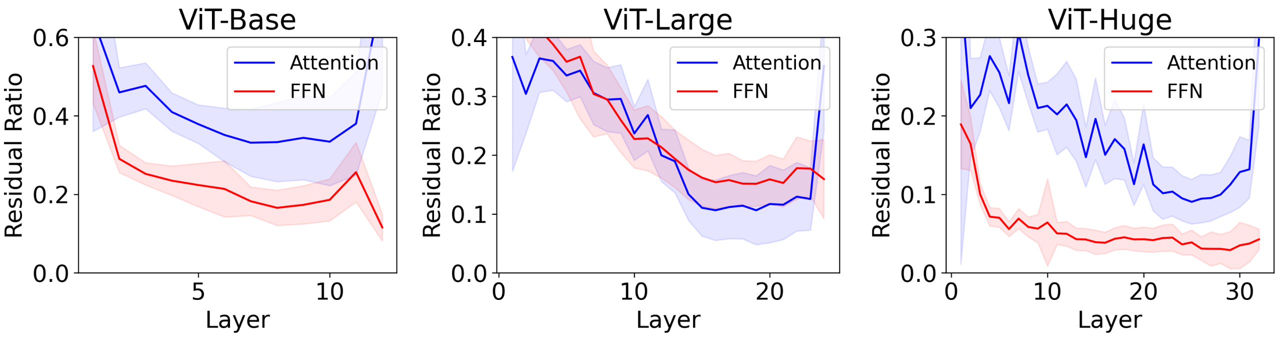

Figure 10. **ViT Family –** Current vision transformers are substantially smaller than language models yet demonstrate similar trends with residual ratio. As model size increases, the contributions of each layer decreases.