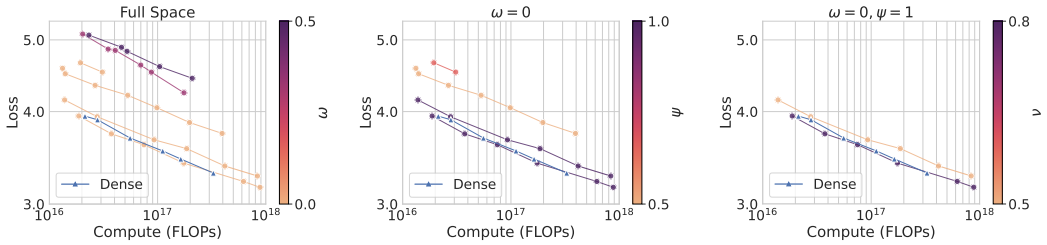
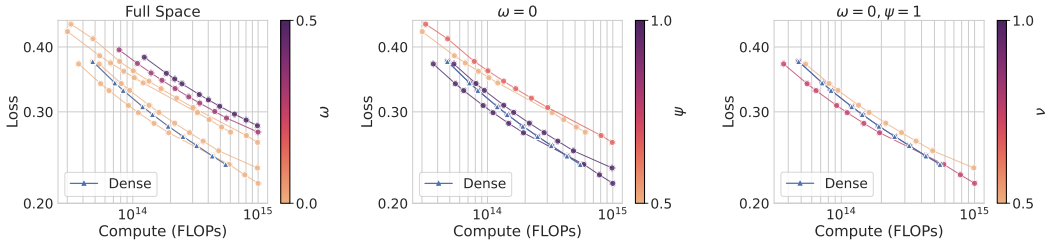


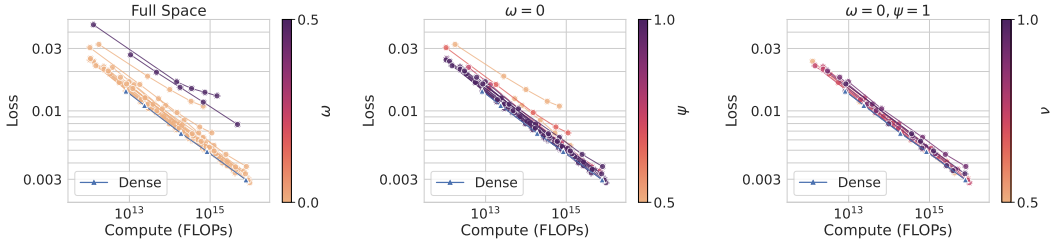
(a) Effect of ω (b) Effect of ψ (c) Effect of ν

Figure 1: **The taxonomy parameters (ω, ψ) explain differences in the scaling laws for 12-layer GPT-2 models with standard vocabulary (50,257 tokens) and a context length of 512.** Small ω (no parameter sharing) and large ψ (full-rank) are necessary for a structure to perform well, while variation in ν has a much smaller impact on performance.



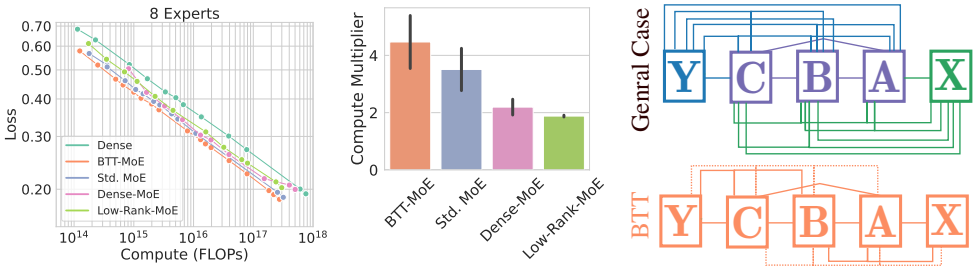
(a) Effect of ω (b) Effect of ψ (c) Effect of ν

Figure 2: Scaling laws for Vision Transformers trained with cross-entropy for autoregressive pixel generation on CIFAR-5M.



(a) Effect of ω (b) Effect of ψ (c) Effect of ν

Figure 3: Scaling laws for MLP trained with mean-squared-error loss on synthetic data generated by a large and randomly initialized MLP.



(a) MoE Scaling Laws (b) Compute Multiplier (c) 3 Factor Generalization

Figure 4: (**a**) Compute-optimal frontier with 8 experts where BTT MoE outperforms. (**b**) Compute multiplier of each MoE architecture, defined as the ratio of compute required by a dense over a MoE model to achieve the same loss, averaged across scales. (**c**) General case and BTT for 3 factors. The general case captures all subsets $S \subseteq \{Y, C, B, A, X\}$ that have at least two elements.