In the following experiments, we examine the effect of learning rate hyperparameters on training stability and structural grokking. We evaluate training instability on the x-axis (measured by total variation; see *Sec 5.1*), and OOD generalization accuracy on the y-axis.

Learning Rates

Setup: We sweep smaller learning rates below the default value of 1e-4 to assess whether reduced learning rates improve training stability.

Results: As shown in Figure 1, smaller learning rates do result in more stable training. However, across 10 seeds, models trained with smaller learning rates are significantly less likely to exhibit hierarchical generalization.

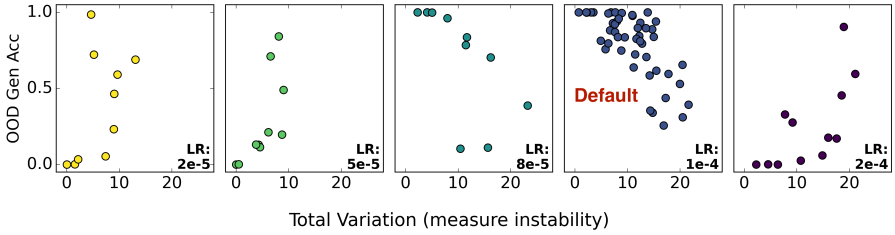

Figure 1: Effect of different learning rates.

Learning Rate Schedule

Setup: We compare a linear decay schedule (default) against a constant learning rate schedule.

Results: Figure 2 shows that a linear decaying schedule (*default*) actually leads to more stable training. In contrast, a constant learning rate results in greater instability across seeds.

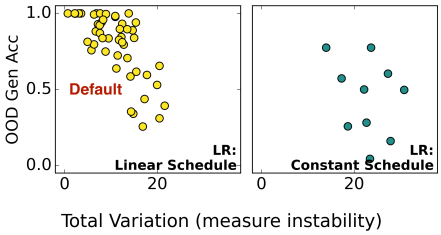

Figure 2: Effect of learning rate schedules.

Weight Decay

Setup: We test the impact of adding weight decay with varying factors. The default setup uses no weight decay.

Results: As shown in Figure 3, small weight decay (< 0.5) has little effect on training stability or OOD generalization. A larger decay ($= 1.0$) slightly improves stability but also reduces the likelihood of structural grokking.

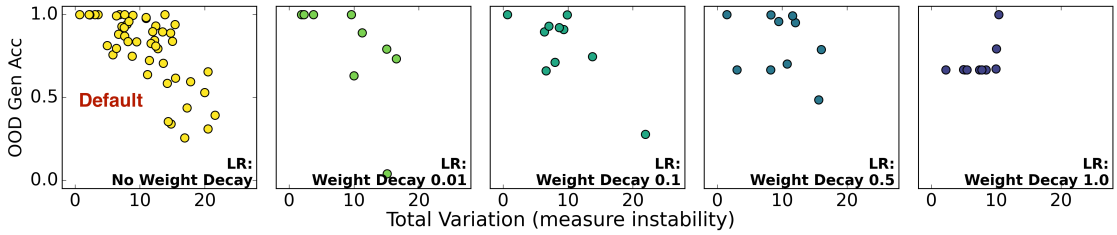

Figure 3: Effect of different weight decay settings.