

# Model Performance Across Training Checkpoints

(3 runs, 8 attempts, arc-agi-1 evaluation set)  
Error bars show  $2\sigma$  (~95% confidence interval)

