

Model Performance Across Training Checkpoints
(3 runs, 8 attempts, arc-ag1 evaluation set)
Error bars show 2σ (~95% confidence interval)

