PHASE2 - Performance by Experiment mistral 20.0 gpt4.1-mini **State** gpt4.1 17.5 $\mathsf{X}\mathsf{X}$ gpt5 **grok** 15.0 Average Passed Tests
0.01
c.2 5.0 2.5 0.0 Experiment 1 Experiment 2 Experiment 3 Experiment