Success Rate per Model: No tutorial 0.0% Gemini 2.5 Flash (thinking) 100.0% 16.7% GPT-4.1 -100.0% 16.7% Claude Sonnet 4 -Baseline 100.0% Ablation 20 60 80 100 40 Success Rate (%)