Success Rate per Model: No hints 70.4% Gemini 2.5 Flash (thinking) 59.3% 66.7% GPT-4.1 -63.0% 55.6% Claude Sonnet 4 -Baseline 66.7% Ablation 10 20 30 50 60 70 40 Success Rate (%)