Success Rate per Model: With vs Without Natural Language Definitions 66.7% Claude Sonnet 4 77.8% 59.3% Gemini 2.5 Flash (thinking) 70.4% Without NL With NL 20 30 10 40 50 60 70 80 Success Rate (%)