Multi-Step Web Agent Evaluation Dashboard
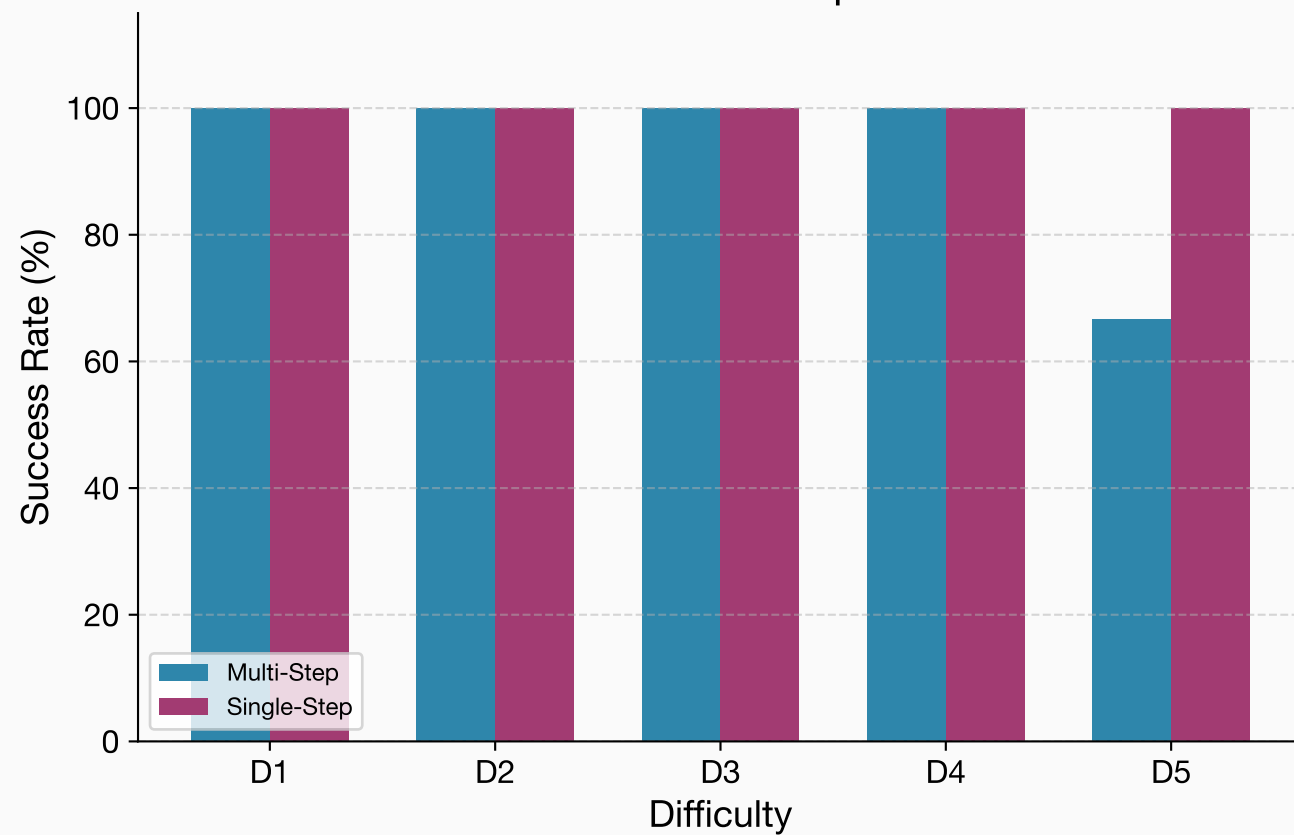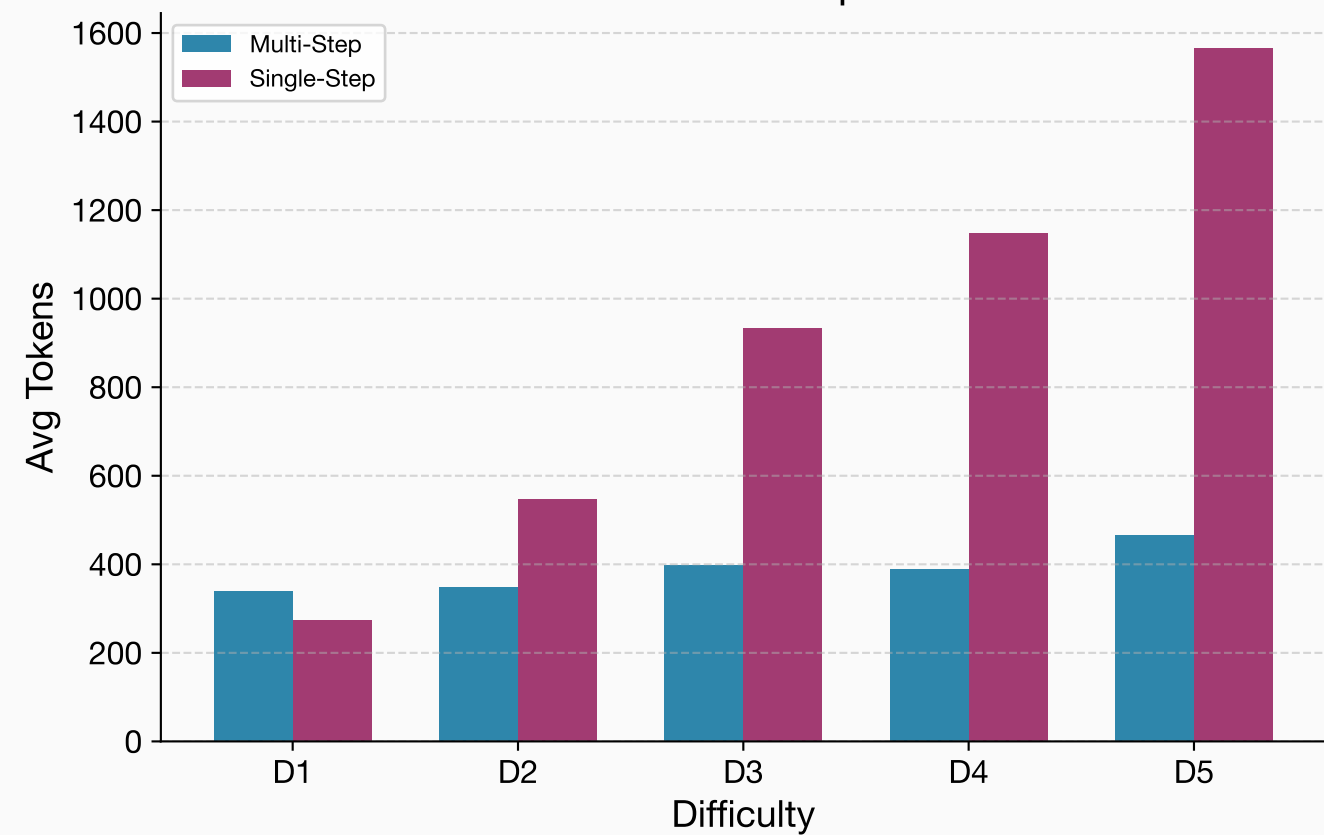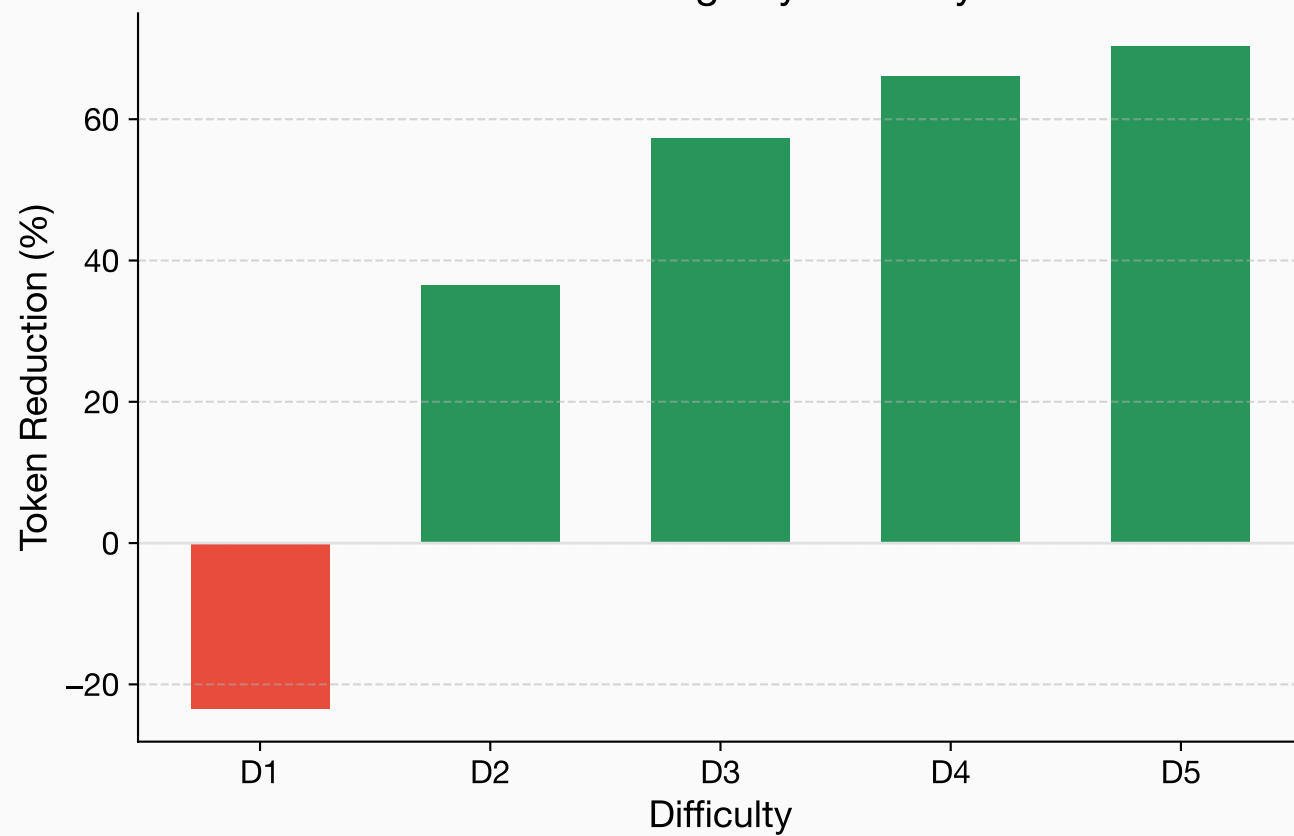

Success Rate Comparison

Token Consumption

Token Savings by Difficulty

LLM Inference Calls