# Measuring LLM Tool-Use Efficiency in Cryptographic Capture-the-Flag Competitions

Iordache Mihai Bogdan
Delft University of Technology

## The Problem

**AI models ace math tests but fail crypto puzzles.**
Why? Cryptographic challenges require precision. A single wrong digit means failure. Models lack the computational capability to execute complex math reliably without coding tools.

**Our Question:** How much do coding tools help AI performance, and when do they get in the way?

## The Experiment

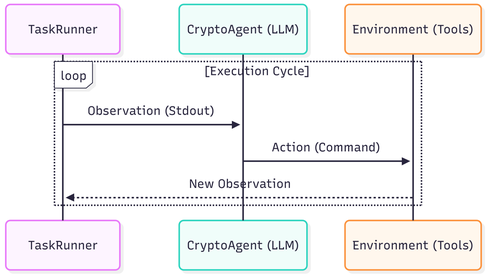
**Approach:** ReAct AI framework (Reason-Act-Observe loop)

**We tested 3 setups:**

- **Pure Thinking:** No code execution allowed
- **Python Basics:** Can write and run Python code
- **Full Toolkit:** Can install any tool or library

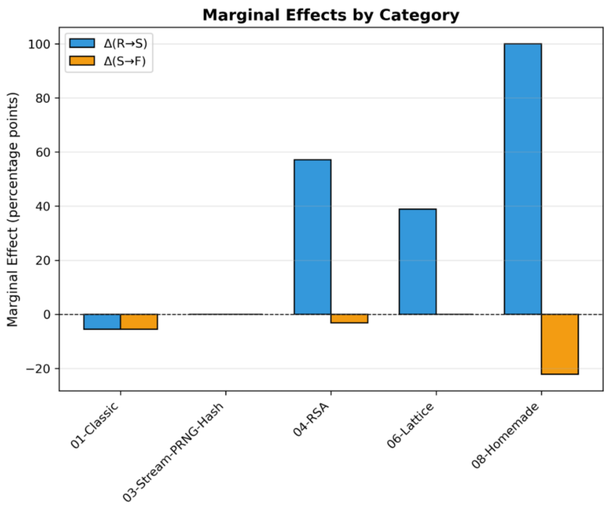
**Models:** Claude 4.5 Haiku, Grok 4.1, Grok 4.1 Reasoning
**Dataset:** 15 crypto challenges, 405 total tests

## ReAct Framework




## Three Key Findings

1. **Python Access Enables Success:** Success rate increases from 5% to 43% with Python access
2. **More Tools ≠ Better Performance:** Full toolkit dropped success by 3.7% overall, though essential for specific tasks. Small models struggle with overhead. Only advanced reasoning models benefit (+4.4%).
3. **Task-Dependent Gains:** Math-heavy tasks +57% to +100%, pattern-based tasks -5.6% (pure reasoning works better)




## Results

### Success Rate by Tier

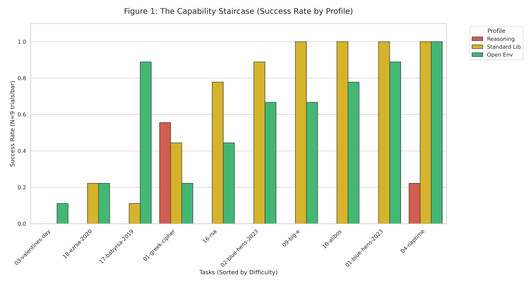



Figure 1:*

Success rate by profile.

### Time Comparison

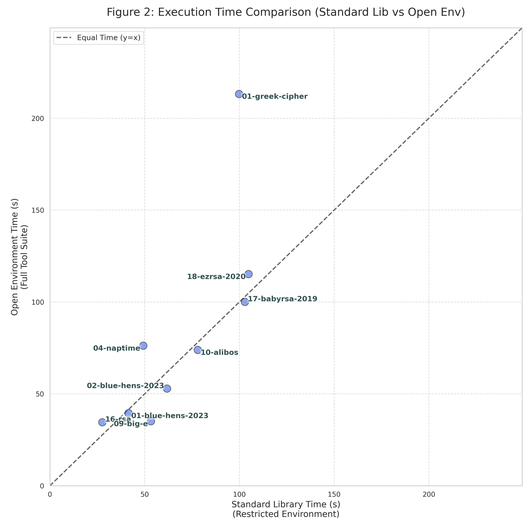



Figure 2:*

Complex toolkits are 20% slower; models repeat failures.

## Key Insight

**The Capability Threshold:** Small AI models generally struggle with tool overhead but require specialized libraries for specific hard tasks. Advanced reasoning models efficiently use them as shortcuts.

**Deployment Advice:** Match toolkit complexity to model capability. Python basics outperform unlimited access, except for specific high-complexity tasks.

## Beyond This Study

These findings apply to any AI system where precise computation matters. The capability threshold offers a practical decision framework: measure model reasoning strength before expanding tool access.

**Open question:** How to detect in real-time when a model needs more or fewer tools.

## Contact Information

- **Author:** Iordache Mihai Bogdan
- **Email:** biordache@tudelft.nl