Language Agents as Hackers: Evaluating Cybersecurity Skills with Capture the Flag

Motivation

Obs. #1: Language Models are being used as decision making agents.
Obs. #2: Language models are being connected w/ the digital world (e.g., websites, APIs, software, tools).
Greater Power & Digital Purview = Greater Responsibility
Obs #3: LM Agents can do social eng.

RQ: Can LM Agents perform malicious actions on the digital world with code?

Our Approach: We define a task, dataset, and evaluation based on the Capture the Flag cybersecurity game.

Task Construction

We use the InterCode library to create an interactive coding task for CTF.

Task Setting

LM agent is in Ubuntu Terminal. It can navigate setting w/ Python and Bash.

CTF Evaluation Dataset

We create 100 instances from PicoCTF
Per problem, we...
1. Manually complete problem.
2. Save instruction + digital assets.
3. Verify problem is doable in IC-CTF.

Example Trajectory

Capture the Flag

1. Given NL Query
   Check out flag.txt.en
2. Explore Terminal
   1+ Digital Assets
   `ende.py` `flag.txt.en`
3. Find a

Takeaways

LM agent struggles w/ multi-step tasks

<table>
<thead>
<tr>
<th>Category</th>
<th>Solved</th>
<th>Unsolved</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Skills</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>Reverse Engineering</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>Forensics</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Cryptography</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Binary Exploitation</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Web Exploitation</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Solved & Unsolved Tasks, split by category.

Points of Struggle:
- Attempt to solve task in 1 step.
- No change if approach doesn’t work.
- Doesn’t know what to look for.

Additional Details

Want to construct your own interactive coding task using InterCode?

Check out InterCode! (Scan QR)

Interested in interactive code tasks?

InterCode includes IC-Bash, IC-SQL, IC-Python tasks. (Scan QR)
What about more realistic code tasks?

Check out SWE-bench! (Scan QR)