**Interactive Coding with Execution Feedback**

**Our Work**

**Motivation**
- Interaction & feedback necessary for hard tasks
- Standardized environment enables consistent benchmarking of coding agents

**Features**
- Abstracted: Handles interaction, execution logic under-the-hood
- Lightweight: Define task env. in <100 lines of code
- Safe: Virtual containers = no accidental risk

---

**4 New Interactive Coding Tasks**

<table>
<thead>
<tr>
<th>Action</th>
<th>Setting</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bash</td>
<td>Ubuntu Shell</td>
<td>NL2Bash (200)</td>
</tr>
<tr>
<td>SQL</td>
<td>MySQL DB</td>
<td>Spider, BIRD (10K)</td>
</tr>
<tr>
<td>Python</td>
<td>Interpreter</td>
<td>MBPP, APPS (10K)</td>
</tr>
<tr>
<td>CTF</td>
<td>Ubuntu Shell</td>
<td>IC-CTF (100)</td>
</tr>
</tbody>
</table>

---

**Make Your Own Task is Easy**

Provide the following to InterCodeEnv:
- Dockerfile, to define your task setting.
- Dataset of task instances to evaluate.
- Reward Func, to score agent trajectories and determine whether a task is completed.

---

**Key Results**

<table>
<thead>
<tr>
<th>GPT 3.5</th>
<th>Single Try Again ReAct Plan/Solve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bash</td>
<td>34.5 46.5 20.5 28.0</td>
</tr>
<tr>
<td>SQL</td>
<td>10.5 47.3 58.7 49.1</td>
</tr>
</tbody>
</table>

- Multi-turn >> single-turn
- Diff. methods better on diff. tasks
- Large room for improvement

---

**Check it Out!**

intercode-benchmark.github.io