AppWorld: A Controllable World of Apps & People for Benchmarking Interactive Coding Agents

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong

https://appworld.dev

Edward Li, Shashank Gupta, Ashish Sabharwal, Niranjan Balasubramanian

A rich **simulated environment** of API-operable **apps & people**, and a **benchmark** of complex day-to-day tasks for **interactive coding agents**.

Coding Agents for Personal Apps

Can **AI assistants** do **complex day-to-day tasks** across **apps** for us?

Return my last **Amazon** ordered shirt & buy it in one size larger.

I owe money to friends on **Splitwise**. Pay them on **Venmo**.

One way LMs can tackle such tasks is by calling **APIs** with **interactive** & **rich coding**.

Play my **Spotify** playlist with enough songs for the workout today. My workout plan is in **SimpleNote**. — Joe

Let me find Joe's workout duration for today.

APIs

```
token = simple_note.login(...)["token"]
notes = simple_note.search_notes("workout", token)
print(note) # found one, show it ...
```

... Monday: ... 25 minutes ... Tuesday ... 45 minutes ...

```
print(datetime.now().strftime("%A"))
```

Tuesday

Interaction

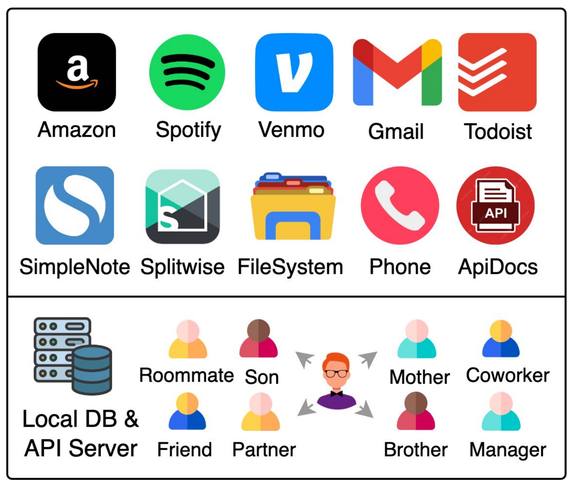
Joe's workout duration is 45 mins. Now, let me play his playlist with enough songs for it.

Rich Code

```
workout_mins = 45; duration = 0
token = spotify.login(...)["token"]
playlists = spotify.show_playlist_library(token=token)
for playlist in playlists:
    for id in playlist["song_ids"]:
        duration += spotify.show_song(id)["duration"]
    if total_duration >= workout_mins:
        spotify.play_music(playlist["id"], token)
        break
```

How can we develop & benchmark such agents in rigorous & reproducible manner?

Engine (Rich & Reproducible Execution Environment)



⇒ Our API-based simulator of **9 apps** in a **local backend**

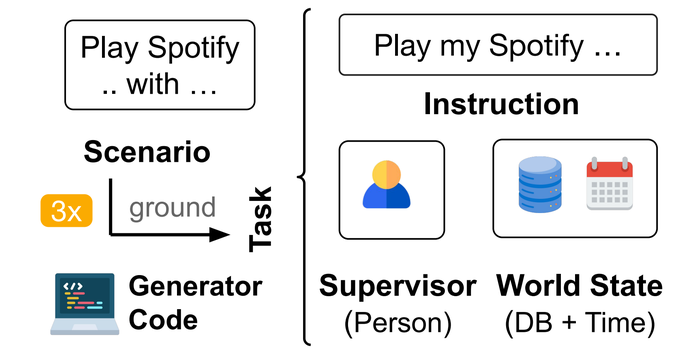
⇒ **High Fidelity**: 457 APIs with detailed docs, 100+ DB tables

⇒ **Rich**, **stable** & **controllable**

⇒ **Realistic** digital activities of 106 people w/ relationships

60K+ code lines

Benchmark (Complex Day-to-day Tasks)



Agent Given: Instruction | App Creds. | API Docs | Coding Shell

750 everyday tasks

⇒ 5-25 **APIs**, 1-6 **apps**

⇒ **Rich+interactive coding** (20-130 lines)

⇒ **Distractors** & **hurdles** (avoid reasoning shortcuts)

⇒ **Task variations** to test **robustness**

40K+ code lines

Evaluation (Robust Programmatic Framework)

⇒ Many ways of **solving**

⇒ Many ways of causing **collateral damage**

Comparison to a **reference** code/APIs isn't suitable ✗

We use **State** & **Execution -based** evaluation! ✓

Is End DB a valid state for task completion?

Check agent's changes are: All Expected Changes | No Unexpected Changes

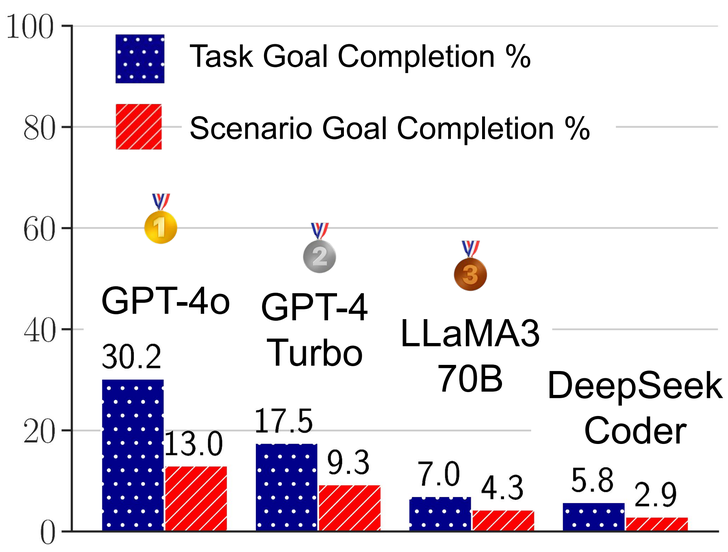
Database Diff. (Start DB, End DB)

Programmatic Tests

- ✓ Spotify music player is playing
- ✓ Player queue duration >= 45 mins
- ✗ Nothing except player changed

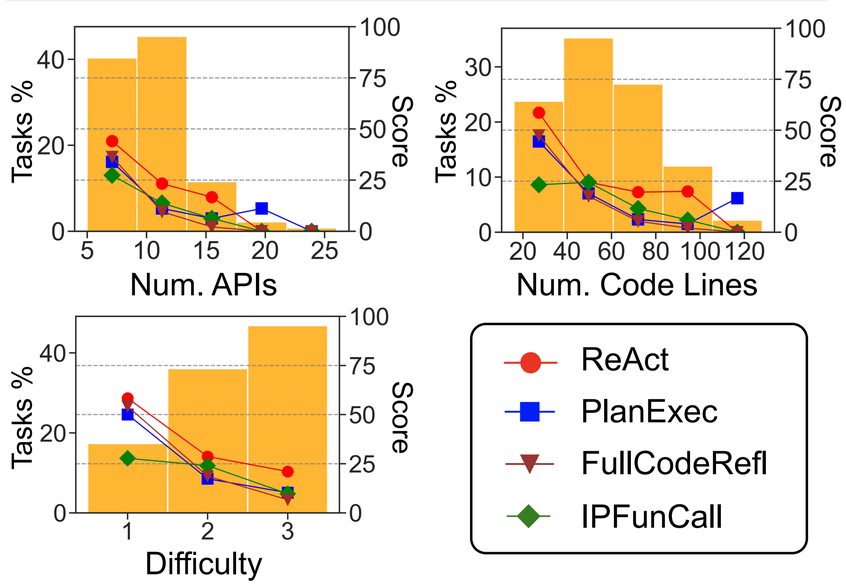
How do Agents Perform?

AppWorld is Challenging!



For each LM, **max score** across **4 few-shot methods**: ReAct, PlanExec, FullCodeRefl, IPFunCall

Scores Lower on Harder Subsets



GPT-4o's Task Goal Completion % across **task difficulty indicators**

What's Next?

Better **Agents**	**Self-exploration**	**Learning from feedback**
New **Bench-marks**	**UI control** coming soon!	**Multi-agent** + human tasks
Study Agents in **Environ.**	Study **safety & privacy** risks	**Social dynamics** of agents in world

Build your AppWorld Agent!

🍰 **AppWorld is Easy-to-Use!**

'pip install appworld', start < 10 code lines.
No Docker / server necessary!

⚡ **and it is Fast!**

Tasks load in < 0.5s, evaluate in < 0.6s, & APIs respond in << 30ms.

Stony Brook University · Ai2 · SAARLAND UNIVERSITY