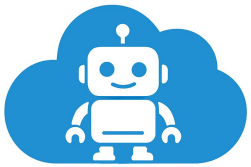


SCUBA: Salesforce Computer Use Benchmark

Salesforce AI Research

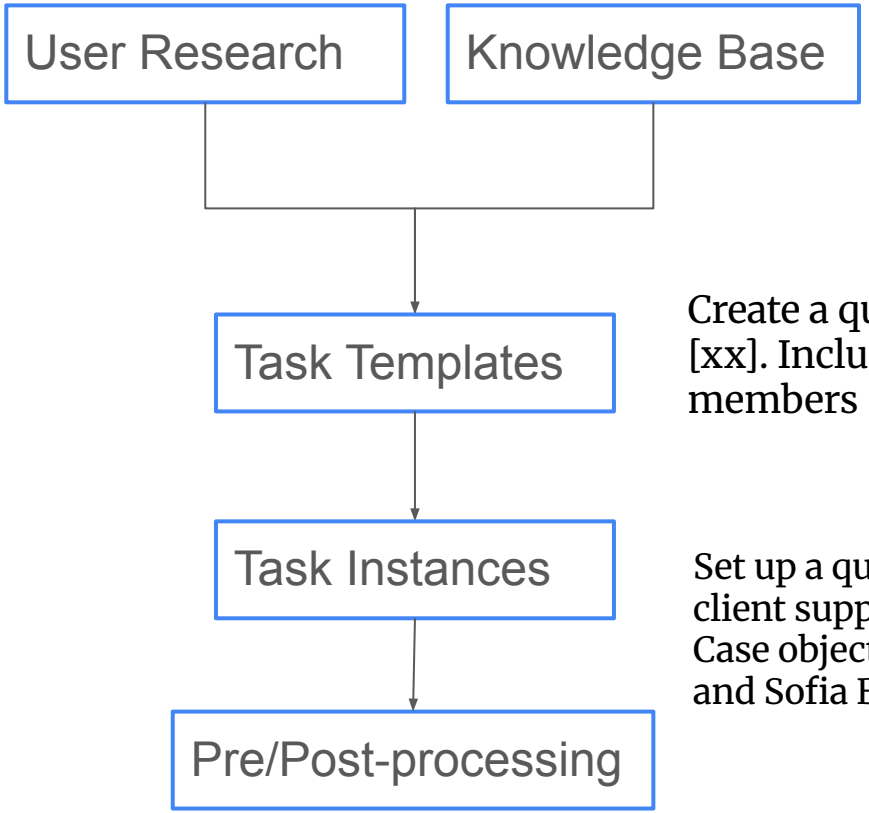
SCUBA: Salesforce Computer Use Benchmark

Motivations

1. Tasks for CRM Centric Tasks
    a. WorkBench series: Service Now
    b. CRMArena series: API agents
2. Browser Agents v.s. Computer-use agents
    a. Which paradigm is more close to production ready?
3. RL environment
    a. Fine Grained Reward

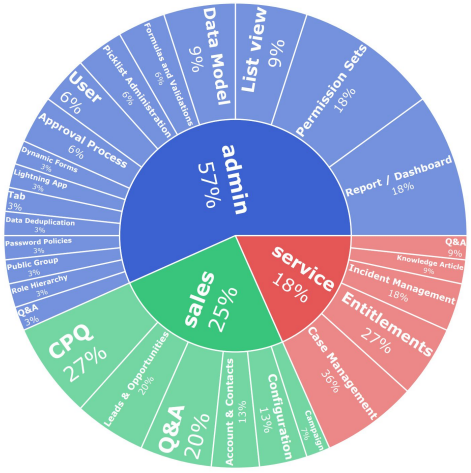
SCUBA: Salesforce Computer Use Benchmark

Tasks

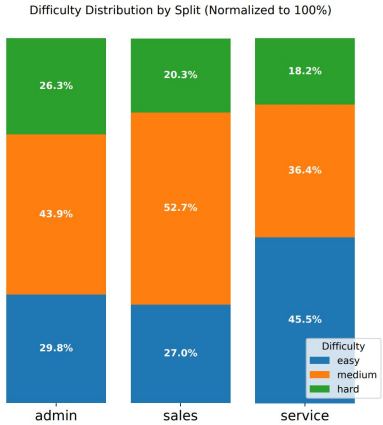


Create a queue with the name [xx]. Include the Object [yy] and members [zz].

Set up a queue called "enterprise client support," associate it with the Case object, and include Alice Bob and Sofia Bennett as members.



(a) Tasks distribution in different domains.



SCUBA: Salesforce Computer Use Benchmark

Evaluator

Set up a queue called "enterprise client support", and associate it with the Case object, and include Alice Bob and Sofia Bennett as members

```
"ground_truth_dict": {
        "Queue_name": "enterprise_client_support",
        "group_type": "Case",
        "role_members": ["Alice Bob","Sofia Bennett"],
        "email_address": "",
}
```


```
{
    "Task Complete": false,
    "Score": 0.75,
    "Failure Reasons": [
      ["Agent failed to Members ['Alice Bob', 'Sofia Bennett'] is added to the queue"]],
    "Rubric": [
      {"milestone": "Create queue with name enterprise_client_support",
        "is_success": true,
        "weight": 0.3
      },
      {"milestone": "The queue has the right supported object type Case",
        "is_success": true,
        "weight": 0.2
      },
      {"milestone": "Email notification is sent to ",
        "is_success": true,
        "weight": 0.25
      },
      {"milestone": "Members ['Alice Bob', 'Sofia Bennett'] is added to the queue",
        "is_success": false,
        "weight": 0.25
      }
    ]
  }
```

SCUBA: Salesforce Computer Use Benchmark

Evaluator

Set up a queue called "enterprise client support", and associate it with the Case object, and include Alice Bob and Sofia Bennett as members

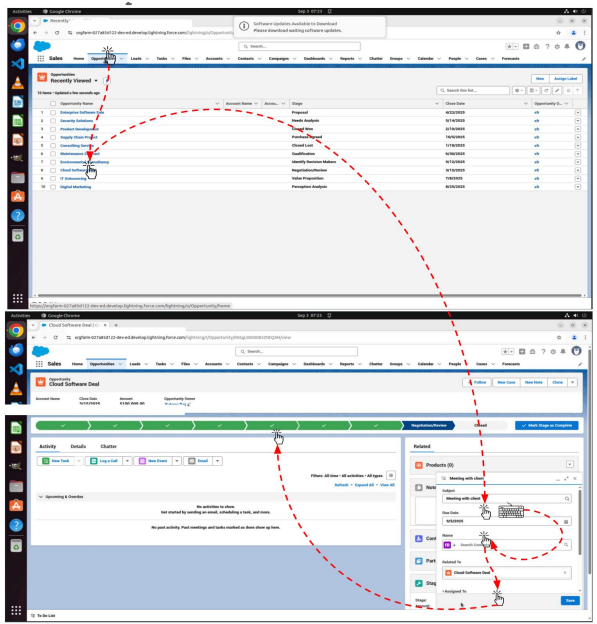
```
"ground_truth_dict": {
        "Queue_name": "enterprise_client_support",
        "group_type": "Case",
        "role_members": ["Alice Bob","Sofia Bennett"],
        "email_address": "",
}
```

Discussion

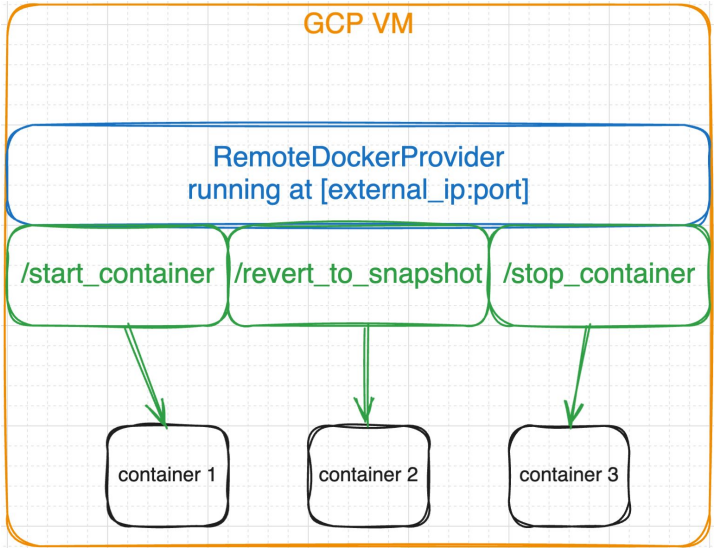
- **[Safety & Integrity] How to capture whether the trajectory is “clean”?**
- **How to automate the tasks & evaluators creation?**

SCUBA: Salesforce Computer Use Benchmark

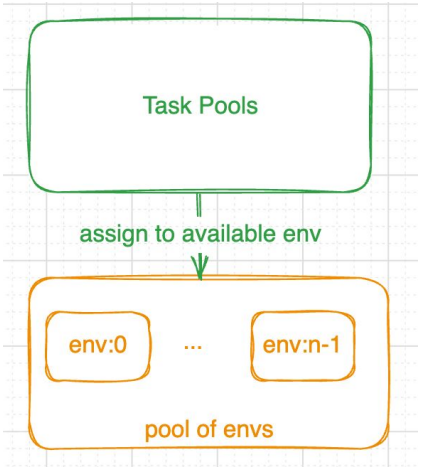
Environment



Realism



Infra

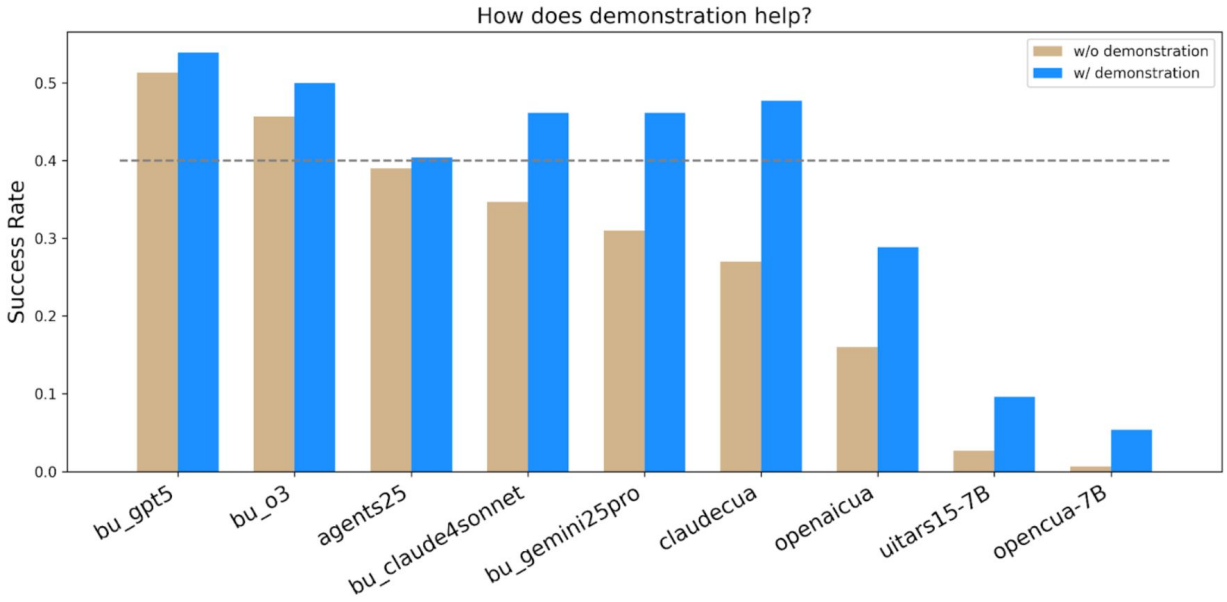


Async Rollout

SCUBA: Salesforce Computer Use Benchmark

Key Questions

1. Browser-Use Agent v.s. Computer-Use Agent
2. Demonstration (Narrative Memory)



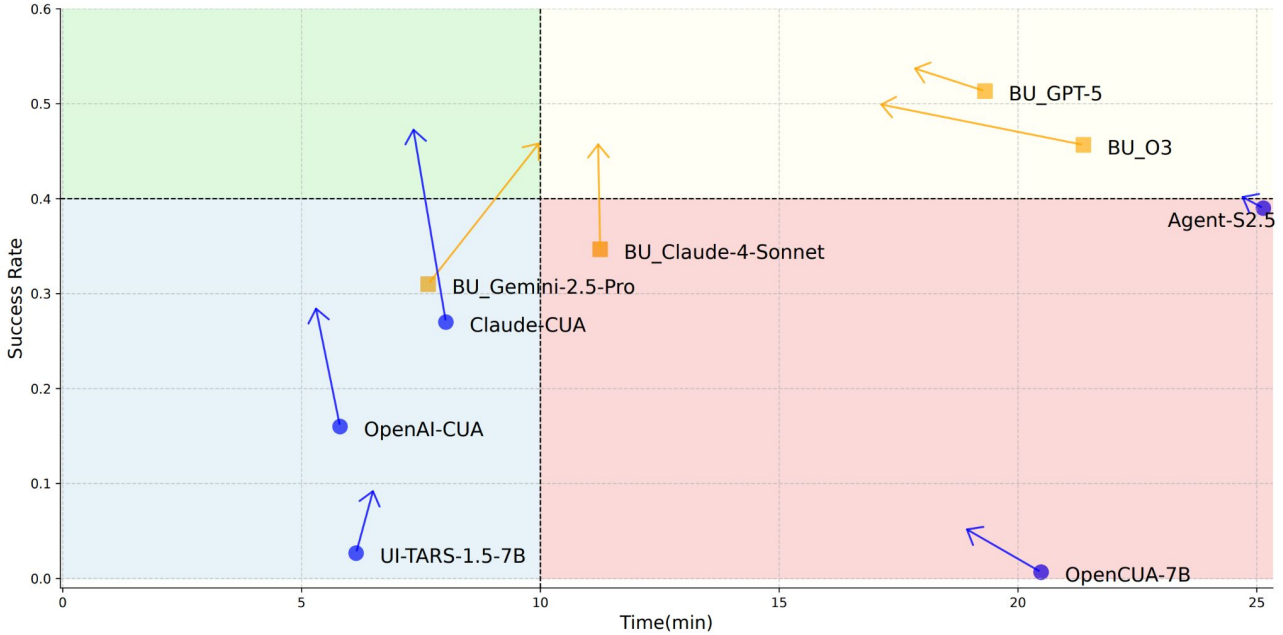
Sample 1:
TASK:Create a Queue with the name "Shoe Case Support". Only send email to members under the distribution list "support.shoe@papaltd.com". Add the Object "Customer". Include "Role: Customer Support, North America" as the Member.
EXPERIENCE:
Successful plan:

- Log in and open Setup.
- Navigate to Users > Queues.
- Click New Queue.
- Configure the queue:
  - Queue Label: Shoe Case Support.
  - Queue Email: support.shoe@papaltd.com.
- Add supported object:
  - Select Customer and click Add to Selected Objects.
- Add members:
  - In Queue Members, choose Role, select Role: Customer Support, North America, and click Add to Selected Members.
- Save the queue.

SCUBA: Salesforce Computer Use Benchmark

Key Questions

3. Agentic Framework v.s. E2E model
4. Latency and Success trade-off



SCUBA: Salesforce Computer Use Benchmark

Key Questions

5. Demonstration: Generation, Representation & Optimization

Demo 1

Demo 2

Task 0_1

Task 0_2

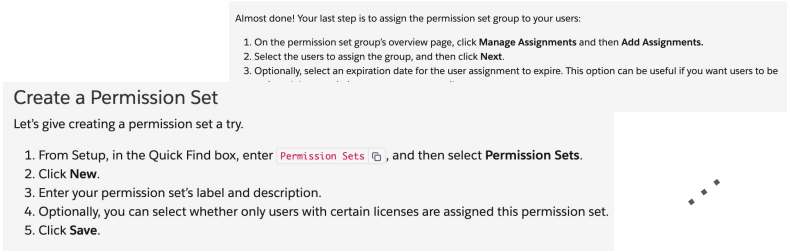
Robust Replay

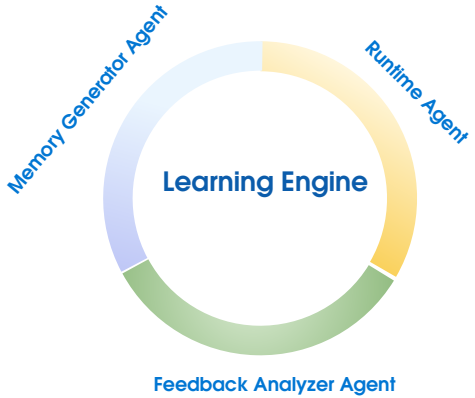
Task 0_3

Task 0_4

Task 0_5

interpolation





SCUBA: Salesforce Computer Use Benchmark

Limitations & Future Work

1. Realistic toB workflows involve multiple platforms
2. Memory Optimization & Evolution
3. Accurately capture the integrity of the agent execution
4. (semi/fully) Automated Tasks & Evaluator Generation