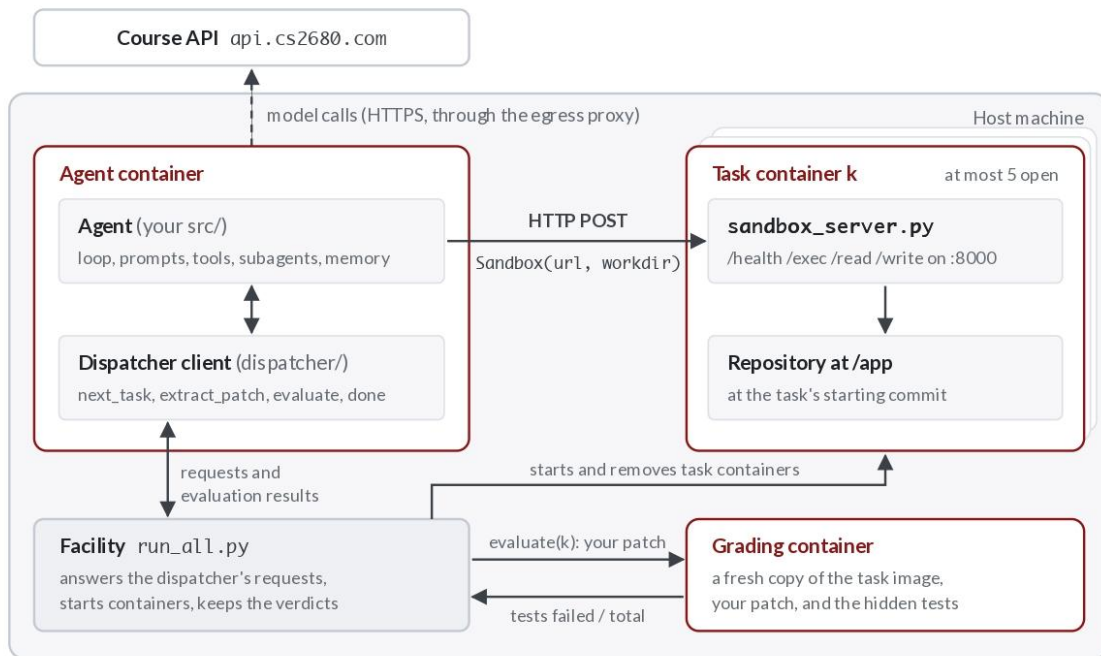


Assignment 3:

MadsOpt: Harness Competition

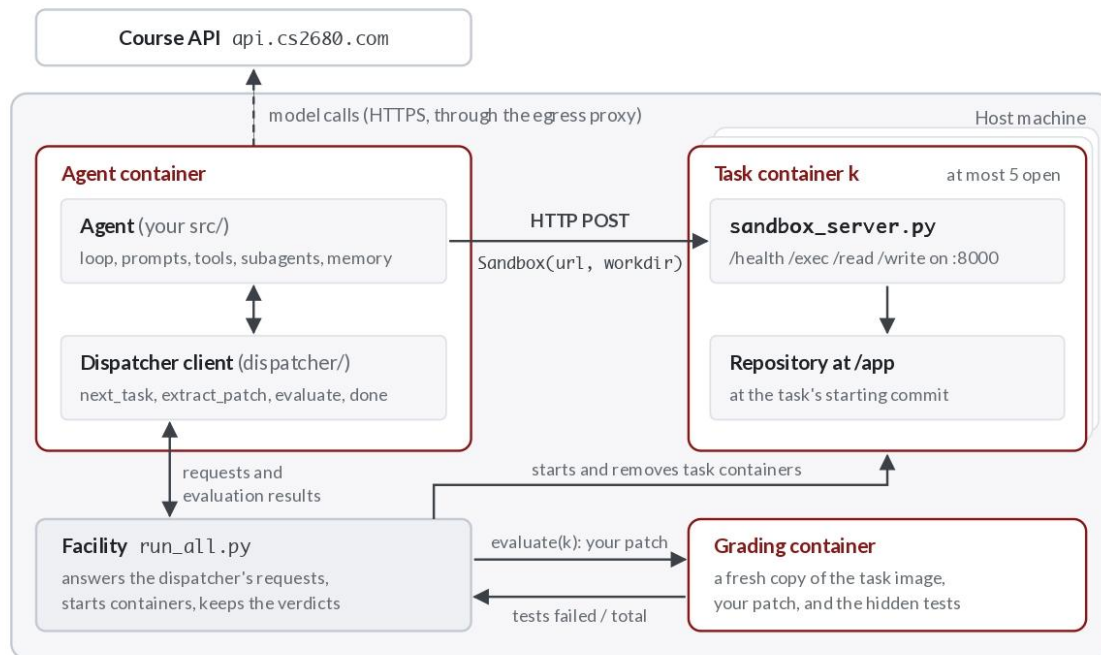
Framework

- 26 tasks. 3 types of containers.
- Agent and the task are in separate containers, using HTTP requests to access the task.
- Offline evaluation, can only access `api.cs2680.com`.



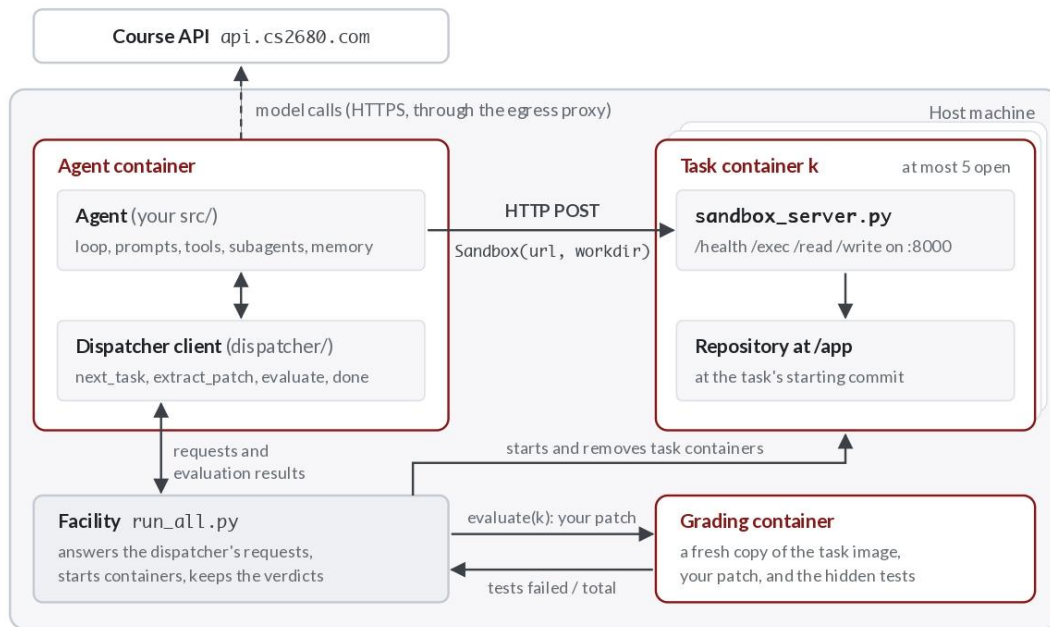
Framework

- Dispatcher passes the task info to your Agent.
- Agent works on a task and generates a patch.
- The patch is evaluated with the result sent back to Agent.
- 5 task in parallel at most.



Limits

- Three models: Expert, Standard, Starter
- Time limit:
 - 35 min per evaluate()
 - 4 h in total.
- Cost limit:
 - \$30 per day
 - \$10 per run



Code structure

- `init.sh`, `package_claude_sessions.py`:
 - generate session logs
- `src`:
 - the agent logic, sandbox requests
 - zip it and it is
the ONLY file uploaded to Leaderboard
- `dispatcher`, `evaluation_scripts`:
 - Dispatcher logic and tests for grading
 - The same used on Leaderboard, do not modify it for local testing

```
assignment3/starter_code/  
├─ init.sh  
├─ package_claude_sessions.py  
├─ src/  
│   ├─ agentic_loop.py  
│   ├─ config.py  
│   ├─ sandbox.py  
│   └─ sandbox_server.py  
├─ dispatcher/  
│   ├─ dispatcher.py  
│   ├─ egress_proxy.py  
│   ├─ session.py  
│   ├─ student_name.json  
│   └─ agent.Dockerfile  
├─ madsOpt.py  
└─ evaluation_scripts/  
    ├─ prepare_images.sh  
    ├─ run_all.py  
    ├─ evaluate_one.sh  
    ├─ make_predictions.py  
    ├─ trace_logger.py  
    ├─ agent_task_input.json  
    └─ task_test.json
```

Leaderboard submission

- Log in with the email on Canvas, with password 'Cs2680-<your 8-digit student ID>'.
- Submit the src.zip
- On submission in queue at a time so please cancel a submission if it does not looks promising.
- The submission is measured by:
 - the number of tasks solved,
 - the cost per solved task
 - the time per solved task
- Cancelled submission also counts, and consumes the daily budget.

Grading

- Skyline with the three metrics.
 - at least solve 7 tasks
 - The ranks counts everyone ahead of it: 1,1,1,4,..
- A baseline of solving 16 tasks.
- The score is calculated as :
 - $\max(81 - r, 50)$,
if any of your submissions achieves the baseline
 - $81 - r$, otherwise
- Leaderboard accounts for 80, with another 10 for the writeup and 10 for video.

