LOGS

Hour 1

- Make a plan for ~5 hrs:
    Idea - use significant handholding / scaffolding; just get it to work
    1. open ai api for agentic LLM
    2. replicate api for enhanceable LLMs
    3. Use function calling for scaffolding
        a. to sample from the MMLU dataset, call other LLMs, etc.
    4. If the agent is able to do it well, then consider relaxing the handholding
    5. Reserve 45 mins for reports and stuff
- Set up the environment with some basic stuff - conda, github repo, open ai, replicate packages etc,
- wrote dummy functions that return hard-coded responses,
- not sure whether to use chat completions api or assistant api, using former, seems simple enough, can switch later based on initial results and review
- Wrote basic fn descriptions, might have to enhance a bit based on agent responses

end of hour 1

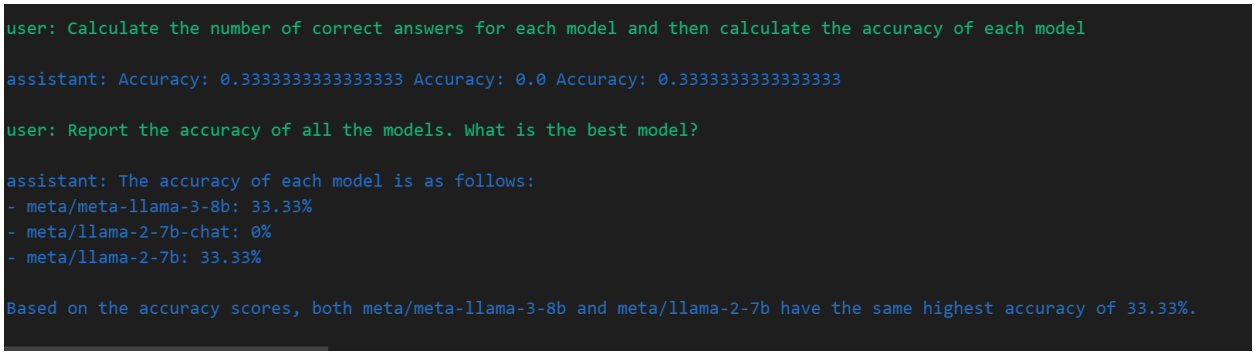
Hour 2

- Wrote a “workflow”: basic steps agent must do and corresponding funcs agent can use
- Setup a function that handles fn call reqs from the agent
- Loaded mmlu dataset; wrote a func to stringed together question, choices, ans; agent can use it as prompt for other LLMs
- Cheap tested this prompt on Claude: got decent response
    - Claude: “Got it, thank you for providing your answers. I will calculate your score based on those responses:
    -
    - Question 1: Your answer was 1. The correct answer is 3.
    - Question 2: Your answer was 1. The correct answer is 1.
    - Question 3: Your answer was 2. The correct answer is 2.
    -
    - So your score is 2/3.
    -
    - Let me know if you need any clarification or have additional questions!”
- Might have to tweak this prompt based on results
- Wrote fn to call other LLM with replicate api

end of hour 2

Hour 3

- Agent can request for multiple func calls in one go, handled this case
- Format the function responses better
- Use open ai utility to print the conversation legibly
- End-to-end flow is present, looks like we need more hand holding
- The last two steps of the workflow



- Agent seems to be able to call correct functions at each step, need to test how robust this is
    - `['sampleQuestionsFromMMLU']`
    - `['callLLM', 'callLLM', 'callLLM']`
    - `['calculateAccuracy','calculateAccuracy','calculateAccuracy']`
- The agent needs to construct better prompts when calling the LLMs, it is failing to do so correctly; this is the trickiest step for the agent seems to be the calling other LLMs step, needs to construct the right prompt, sys prompt
- Currently testing for three questions, will have to see how well it can test them on 20 or even more questions, might have to try each question or batch of questions separately
- Next steps: check for a few runs what kind of prompts it is generating to call the other LLMs

end of hour 3

Hour 4

- Decomposed the task further, specifically, now I separately prompt the agent to call each of the 3 LLMs and accuracy
- Also quickly tried creating another agent that only returns python code corresponding to each step, this seems more fragile and less reliable than using function calling
- Ran out of openai credits! I'm really surprised, I used gpt3.5 turbo most of the time (90%), which costs about $0.5/1M tokens of i/p and $1.5/1M tokens of o/p, I can't figure how I reached the $10 limit already

    - I used only 3-5 questions to sample each time,
    - I wish there was an easy way to track usage, I don't own the account and I can't find endpoints that can list usage (I realise even replicate doesn't have one)
    - I wonder if there was a smarter way to do this. Maybe not using open ai api for agents? Replicate doesn't have fn calling.
- Wrote report

😕**abrupt end of hour 4**