After you get started w/ user research

- Confidence in AI:
    - Idea: Exploit DeepThink's thinking mechanism with an analyzer that can succinctly process DeepThink's certainty or uncertainty into metrics/a simple representation. Users can gain a coarse-grained understanding of the AI's reliability without needing to stay tuned to the fine-grained output of DeepThink.
        - Can also intelligently cut off DeepThink; e.g. 'if running for 1+ minute and still 'wait'-ing repetitively, cut it off and just tell them that the output is shit.
- Latency v.s. Response Quality
    - Certain interplay between response quality reducing latency via better responses, but is that time margin justifiable compared to the user just iterating with follow-ups/adjustments? Has interplay with confidence; the amount of active time a user would spend actively engaging with a system decreases with higher response quality (and higher latency, with reasoning models).
- Use case vs. Latency
    - Idea: Router that proactively routes a query by use case complexity to a faster LLM v.s. a reasoning model.

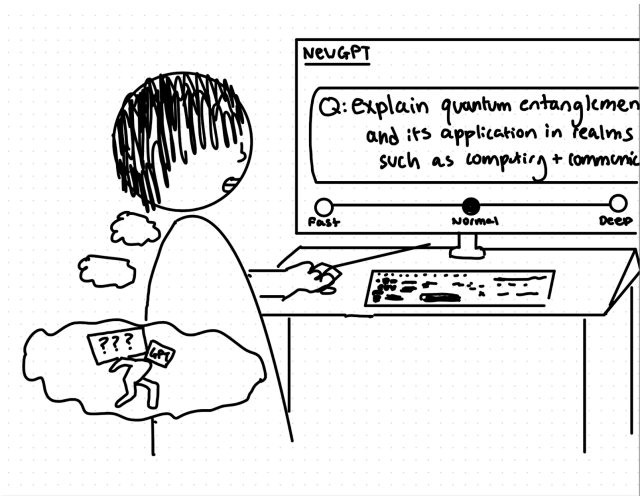
1. [+2] Synthesize your user research into:
    a. [+1] Personas + Scenarios (at least one persona, one scenario)
        i. College student that views the LLM as a kind of semi-omniscient Oracle
            1. Scenario: wanting a guide for deeper subjects. Office hours with professor type vibe. Expectation of reasonable response time < couple minutes, but visible "thinking" is sufficient (in terms of initial response) so the user can think through the process with the LLM. Still requires active interactivity.
            2. Individuals: I, II, VI, VII, X
        ii. College student that views the LLM as a better version of a google search (e.g. Google Search AI). High assumption of confidence in agent. Both the user and the LLM are iterating together.
            1. Scenario: dominantly user-guided, simple queries, expectation of instant response.
            2. Individuals: IV, IIX
        iii. College student that views the LLM as a powerful thinking agent that requires substantial guidance to reach a satisfactory solution. Low assumption of confidence in agent. User iterates the LLM's understanding.
            1. Scenario: semi-novel, complex coding or thought problems. Open to leaving the agent to complete the task on their own if there's a high degree of confidence in the response. Appreciates the self-reflective nature of the thinking process.
            2. Individuals: V, III

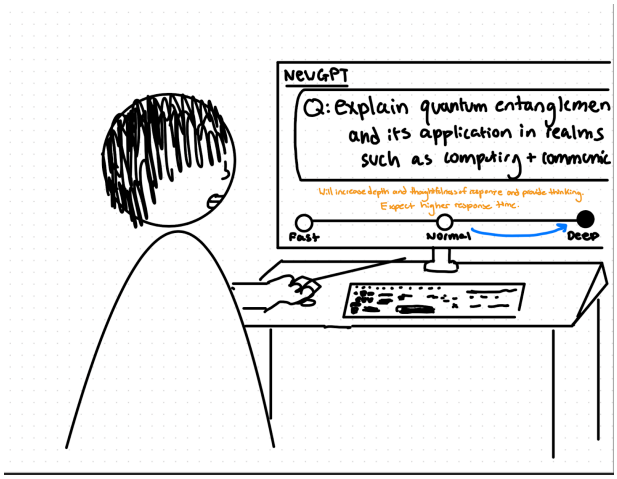
b. [+1] Process map or more in-depth task analysis (at least one)
    i. Persona 1
        1. User has some kind of confusion (academic) → Asks LLM broad question → Read output with some level of trust → Asking follow up, clarifying questions / expansions on initial premise.
    ii. Persona 2
        1. Asks LLM specific narrow question → Assumes truth in LLM response until tested, but ok with inaccuracy → adds supplemental context or pivots slightly to get new desired response
    iii. Persona 3
        1. Asks LLM specific question with premise → Analyzes output assuming high potential of fault → Asks iterative questions to nudge it towards a correct solution/adjust model → Iterates until AI is able to arrive at a final solution.

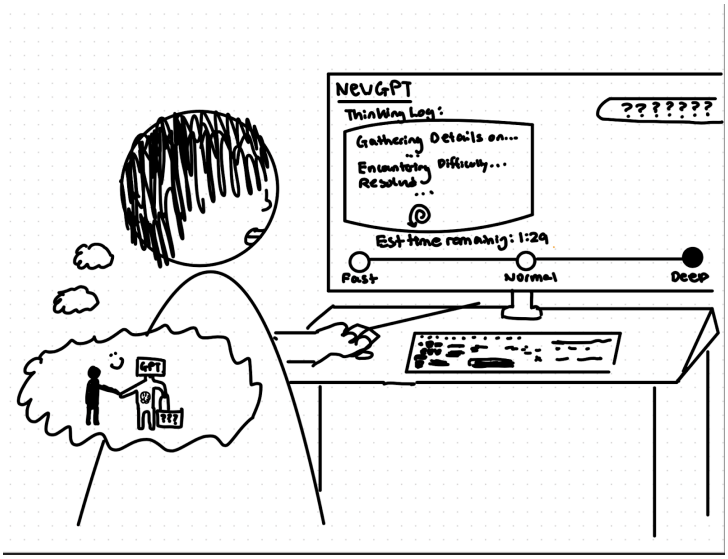
2. [+1] Articulate a problem statement. A problem statement should illuminate the core of the issue you observe. Often, there is a contrastive tension between what users want to do and what their current tools require them to do. If you cannot articulate this yet, describe why you think you are not able to converge yet, what promising directions to follow up with additional user research and feedback might be, and what steps you think you should take next.
    a. Tip: Rely on your process map. What is it telling you about what users want to do vs. have to do?
        i. Users want to efficiently answer/solve their problems with assistance from AI models; however, there is a general lack of explainability and transparency to AI responses that users were frustrated with (V, X), which wastes users time due to the requirement of follow-ups, iteration, and close-reading to ensure answer consistency. While multiple users saw the higher latency model as effectively rectifying many of these issues, so too did many users consider the waiting time and verbosity unacceptable for their use cases and workflows. This signals a problem between the interplay of AI trust, response times, use cases, and the user's flexibility in tailoring the experience they want.

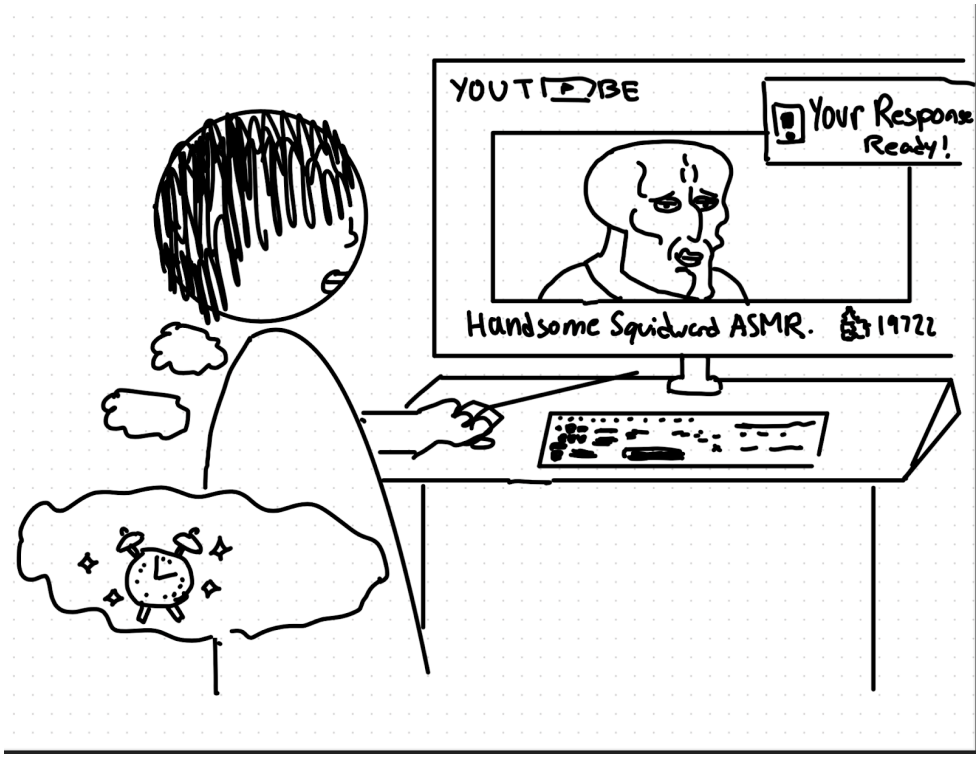
    b. Double check: Does your storyboard communicate/highlight this core tension? Affirmative

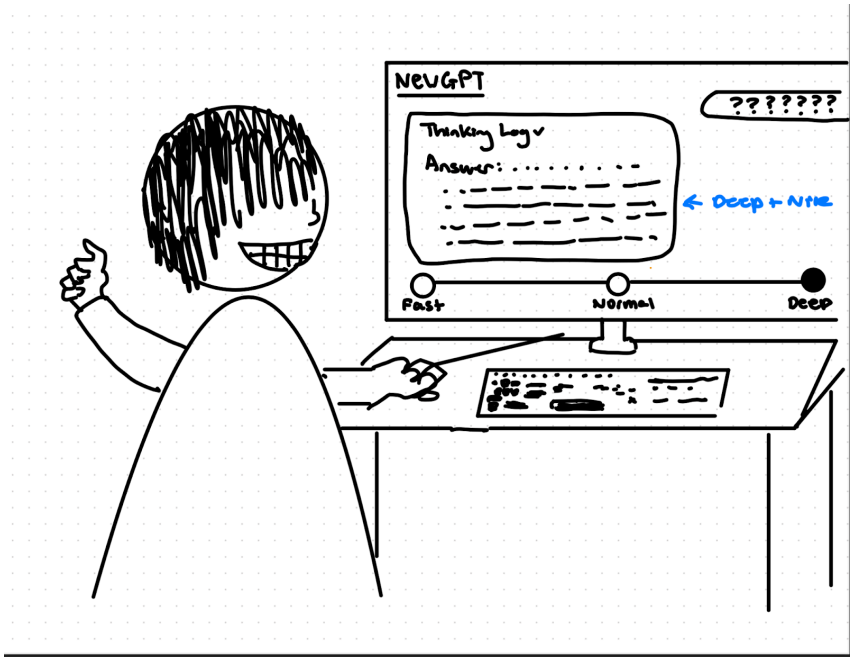
3. [+1] Storyboard of how your proposed system could address the core problem. This is where you begin to imagine a prototype to address the core problem you identified.


NeuGPT
Q: Explain quantum entanglemen
and its application in realms
such as computing + communic
Fast
Normal
Deep
???
GPT


NeuGPT
Q: Explain quantum entanglemen
and its application in realms
such as computing + communic
Will increase depth and thoughtfulness of response and provide thinking.
Expect higher response time.
Fast
Normal
Deep


NeuGPT
Thinking Log:
Gathering Details on...
Encountering Difficulty...
Resolved
Est time remaining: 1:29
Fast
Normal
Deep
???????
GPT
???


YOUTUBE
Your Response Ready!
Handsome Squidward ASMR.
19722


NeuGPT
???????
Thinking Log
Answer:
Deep + Nice
Fast
Normal
Deep