Communicate what you've learned

[+2] Draft blog post summarizing user research. Your blog post draft must include:

- A description of methods and key findings

- A clear problem statement describing the crux of the issue you identified through user research.

- Your storyboard

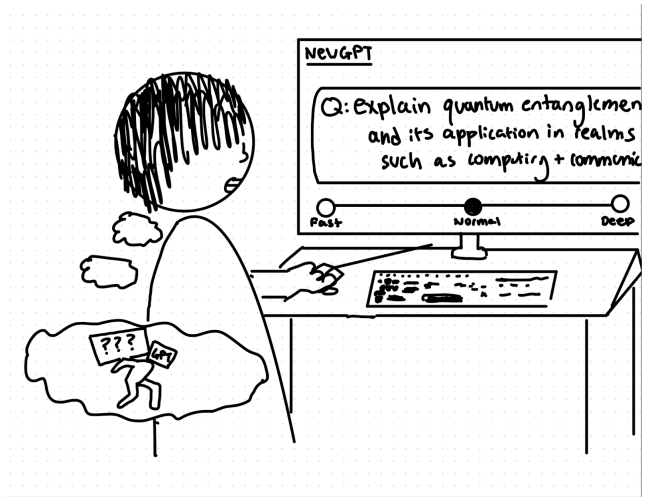
- At least 1 figure that is not from your storyboard

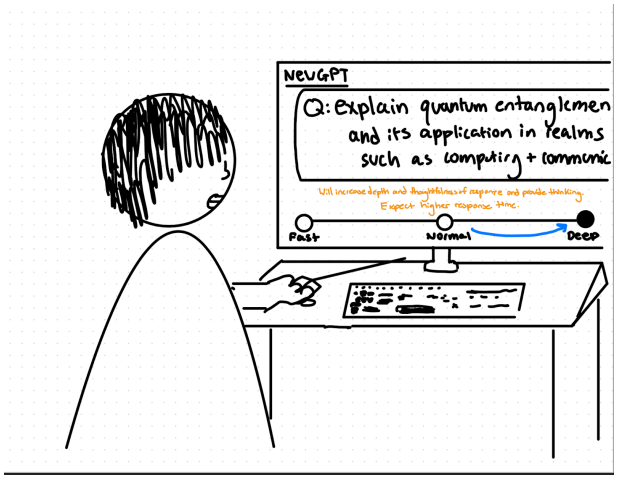
We observed how our target demographic, college students, interacted with varying latency AI systems. First, participants used ChatGPT for what they determined to be a common workflow. Afterwards, they used DeepSeek's R1 on the same workflow. Because R1 is a reasoning model, it would take longer to "think" before providing a response. From these interviews, three personas formed. Persona A treated these LLMs as semi-omniscient oracles, seeking guidance with broader questions. They generally had high faith in the responses. Persona B approached these LLMs with simple queries and expected a quick response. Through a back and forth trial and error process, both the user and the LLM were able to "iterate" with the user providing supplemental context for more accurate LLM responses and partially unsatisfactory LLM responses illuminating how the user could refine their understanding/question. Persona C had a pre-existing skepticism of LLMs and viewed them as powerful agents that needed hand-holding to arrive at the desired end goal. Their expectation was that back and forth would be needed to iterate the LLM's understanding to where they needed it to be.

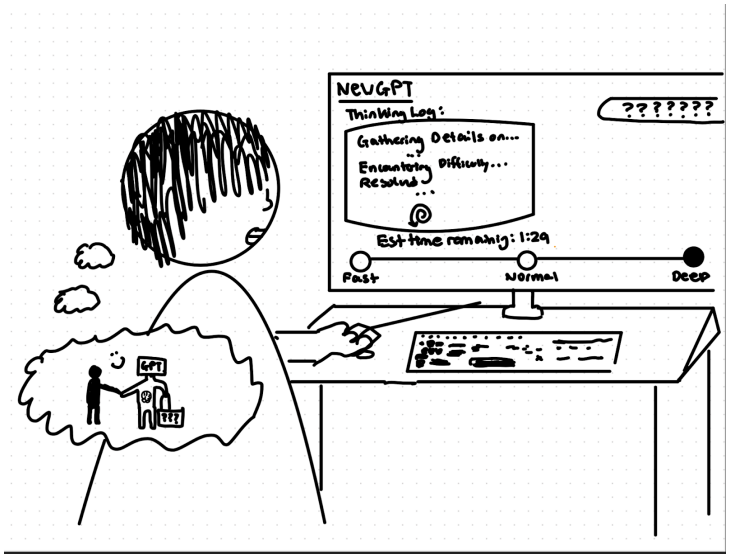
From our interviews, the most present persona, Persona A, found the longer latency of R1 tolerable because the answer they were receiving was more involved. Personas B and C were more frustrated with the longer latency due to the iterative natures of their approach. Because the LLM responses were not always what they were looking for, waiting the additional time was more frustrating because it felt like a waste. This unearthed to us that latency is not preferred or dispreferred across the board, but ultimately a sense of efficiency is, and the way this sense of efficiency is derived is use case specific.

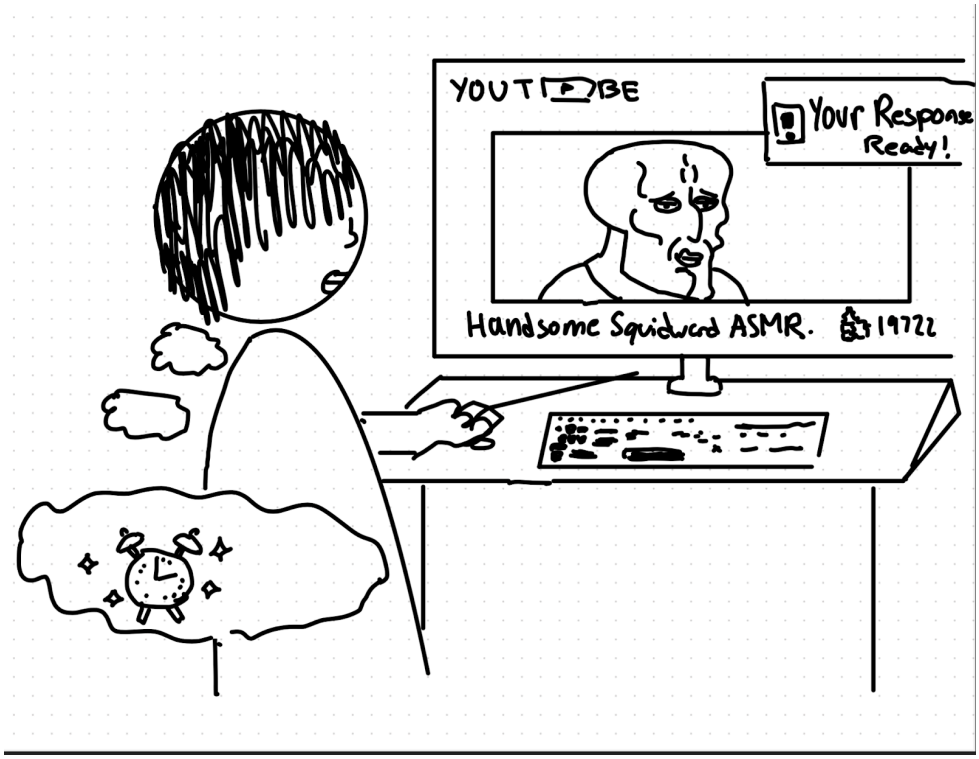
Users want to efficiently answer/solve their problems with assistance from AI models; however, there is a general lack of explainability to AI responses that users were frustrated with (V, X), which wastes users time due to the requirement of follow-ups, iteration, and close-reading to ensure answer consistency. While the higher latency model was found as engaging and welcome by multiple users, so too did many users consider the waiting time and

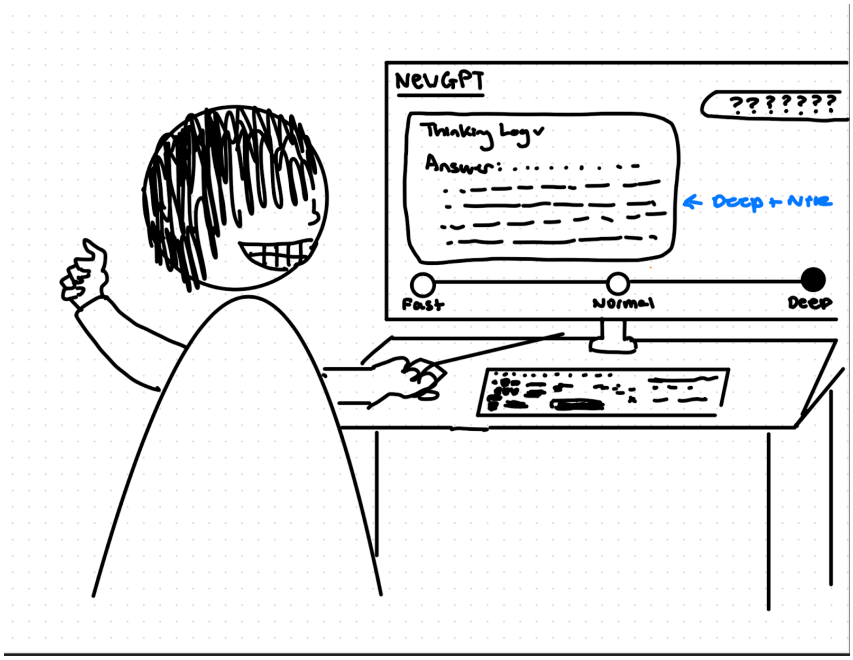
verbosity unacceptable for their use cases and workflows. This signals a problem between the interplay of AI trust, response times, use cases, and the user's flexibility in tailoring the experience they want.

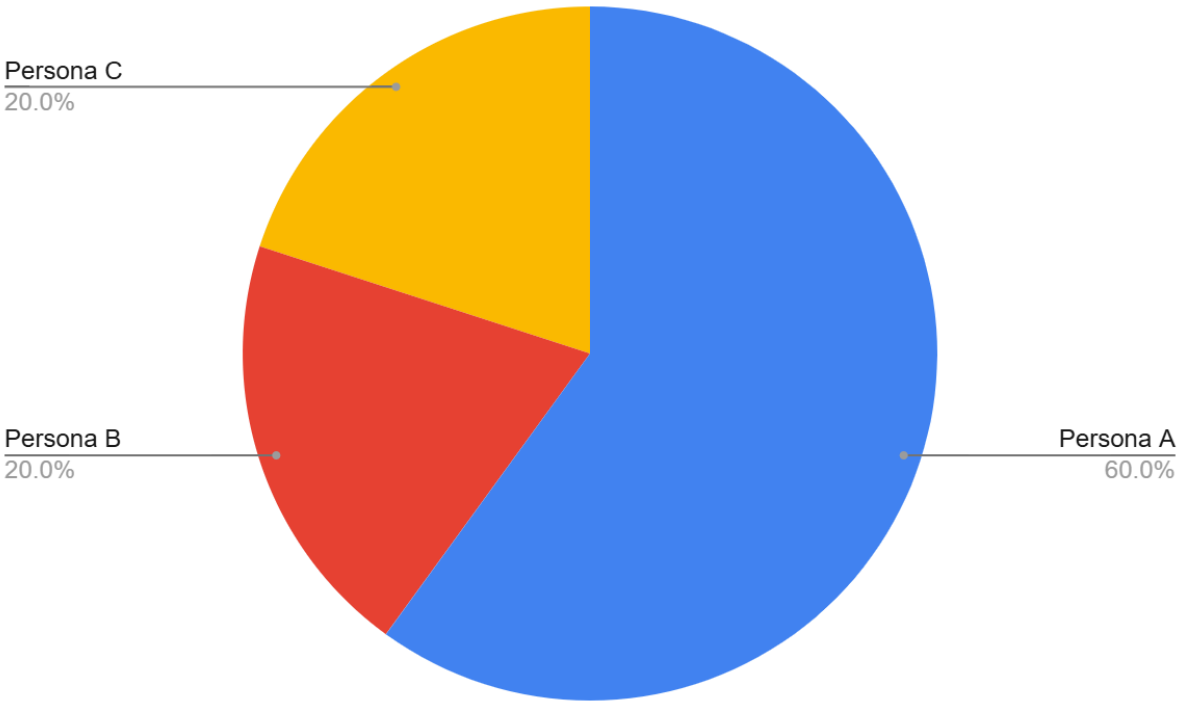





NeuGPT
ThinWing Log:
Gathering Details on...
Enumerating Difficulty...
Resolving...
Est time remaining: 1:29
Fast
Normal
Deep
??????
GPT
???


YOUTUBE
Your Response Ready!
Handsome Squidward ASMR.
19722


NeuGPT
???????
Thinking Log
Answer:
← Deep + Nine
Fast
Normal
Deep


Persona C
20.0%
Persona B
20.0%
Persona A
60.0%