Think-Aloud Study + Interviews

Methodology: Request user to have an important workflow they'd use artificial intelligence for. Before interview, assure them that they should treat your presence as a nonfactor (encourage them to zone out if they lose focus). Have them exhaustively run their queries on ChatGPT. Afterwards, have them run the same first query from ChatGPT on DeepSeek with the DeepThink mechanism enabled, and exhaustively run queries until no more steps are available to be fulfilled. Take special note of reactions to latency, quality of answers, requerying, etc. After think-aloud, introduce domain of research and ask direct questions about their preferences in terms of latency, their opinions on higher-latency systems, attention span, etc.

Interview interviewees directly on their reaction to chain of thought; what they thought was necessary or unnecessary.

Ask why their use case informs their reaction. Do some repetitively throughout the process.

Screenshare to analyze later.

Conclusions thus far: A commonality between users I and II was that when using AI as a mechanism of learning and understanding, having access to the thought process of the LLM inspires much higher confidence in the answer as well as serving as a mechanism to help user work through the problems introduced in their query. Potentially a meta-summary system could work? E.g. summarizing the thought processes to be brief and provide just enough insight for the user to have confidence that it did think.

Individual X: Person Description <TEMPLATE>

Workflow involved:
ChatGPT Reaction:
DeepSeek Reaction:
Interview Discussion:

Individual I: Freshman, CS Major

Workflow involved: section on intern blog, investigating Google's Quantum Computing Chip and its impacts w.r.t quantum computing applications. Wants long, medium, short term ramifications.
ChatGPT Reaction: Took time to carefully format prompts with expectations and requirements of responses; e.g. the timescale of each effect. Very little tuning out; however, there is a substantial deal of elaboration required on each response from ChatGPT in order to arrive at a satisfactory point of understanding. Notable negative reaction to ChatGPT's higher-level predictions of technology future due to insufficient backing. User remained mostly engaged with queries.
DeepSeek Reaction: User reacts positively to DeepThink mechanism, though initially DeepSeek's responses had the same too-high-level issues that user previously encountered with ChatGPT. Elaboration yielded substantially more impressive results, however, with a cohesive timeline generated and internal reasoning provided in DeepThink. User zones out during the process of one response. Overall expressed preference for DeepSeek in scenarios where product quality is valued.
Interview Discussion: User expressed that DeepThink was able to create a more convincing argument for its responses and scenarios. There is a certain upper bound to the amount of time the user was willing to wait before losing confidence in the AI's answer; e.g. with a hard physics question, 5+ minutes of thinking led the user to realize that the AI had no answer. The user also expressed that it was unlikely that he would zone out for more than one/two minutes at a time if he was committed to staying on task; most zone-outs would involve checking social media briefly on phone.

Individual II: Master's Student, EE Major

Workflow involved: disambiguation of graduate-level lecture topic in preparation for midterm exam.
ChatGPT Reaction: Took time to make prompt professional, though not with special prompt engineering. Immediately engaged with each response, reading the answer as it was produced. Expressed concern that his method of follow-up questions would bias the LLM towards a certain response (e.g. potentially validating/emphasizing incorrect information.) Skims longer responses for desired kernels of information. Sees follow-up questions as natural complement to him understanding the topic; no more different than office hours, and not frustrating either (unless LLM produces bad answers).
DeepSeek Reaction: First time using product, expressed skepticism at output from DeepThink. However, later expressed that they were impressed by the cognitive abilities expressed by DeepThink; e.g. linking together concepts, disambiguating user's experience level, and a high-level understanding of topics before its formal response. Was especially impressed by how DeepThink was able to discern the exact point of confusion that the user had when prompting a follow-up question. Overall, recognizing how LLM

navigated ambiguity of user's prompt and responses was deeply insightful, and benefitted user's confidence in response and perception of response's quality.

Interview Discussion: User states that DeepSeek is significantly more insightful than ChatGPT; notably, DeepSeek was more willing to proactively provide applications and examples for queries about high-level topics. User did not perceive a meaningful time difference between DeepSeek and ChatGPT; this may be a quirk of the user, as they continuously skim output and believe LLM outputs to be 90% fluff, 10% useful information / information to expand on. Reacts negatively to deliberate delay in answer output.

Individual III: Masters Student, Biology

Workflow involved: Analyzing a table biological data, a complex task. Columns are different strains, where the value of the cells are percentages of an insertion mutation, and the rows are assays.

ChatGPT Reaction: User copy and pasted the table into ChatGPT as unformatted text. Expressed doubt over whether ChatGPT understood the table correctly. Tested out whether it did, and it got the alignment wrong. Asked whether there were any outliers in the data, and ChatGPT at least displayed a "thinking" loader prior to giving the final answer. Took only a few seconds, and claimed there were no outliers. Told ChatGPT to fix the table formatting, and it made the appropriate correction. When again prompted to check for outliers, displayed an "analyzing" loading screen, and then claimed there were outliers. The user expressed a mixed reaction to whether the outliers were accurate. Clicking on the "analysis" displayed the code which was run in order to check for outliers. Analysis took ~7s to run (because of the code being executed), an unexpectedly long amount of time.

DeepSeek Reaction: Liked the tone/expressiveness of the "thinking," "Oh my god it's still going, I'm going to be so for-real, I hate this." Didn't want to read all of the text, expressed negative reaction towards the model changing its mind. However, the model attempted to fix the table all by itself, which was impressive to the user. The model spent 2-3 minutes trying to understand the strangely formatted table, and then proceeded to do some analysis afterwards. Keeps watching the thinking process in order to avoid re-reading the entire thinking process, rather than switching to another task. After 5 minutes (288 seconds) of thinking, a final answer was given which still didn't have the correct table format. Didn't realize there was a final answer at the end, which may have prompted the user to step away to a different task.

Interview Discussion: Previously used Perplexity to find academic research papers that discussed/defined the exact distances between promoters (of genes) in prokaryotes. Used it more or less as a search and summarization tool. Used the pro version for that search.

Deepseek is considered “cool” but trust is not quite there enough to trust it to do a complicated task completely by itself. Said that ChatGPT provided an element of better control over the direction of the generation, and felt more like a controllable tool than Deepseek.

Individual IV: Postgrad, Pre-MCAT

Workflow involved: Searching up biochemistry terms to put on an MCAT study guide.
ChatGPT Reaction: ChatGPT was down, so Gemini had to be used. Appreciates that the agent summarizes online information without the tedium of needing to search or hunt it down manually. The use case is overall very simple, so it lends itself well to ChatGPT’s rate of response. As the answer comes out quickly, there’s little-to-no zoning out between prompt and response. Thinks the issue with Gemini is that there’s too much redundant information; a quirk of the scenario, as the MCAT’s volume of information and the intent of making a study guide render a depth of knowledge overwhelming. Assumes most information is correct under this use case.
DeepSeek Reaction: Did not like the volume of information, mostly because of the use case and its particularities above. Appears the viability of DeepSeek is highly dependent on use case; simpler workflows cannot justify a high volume in answer or high latency in thinking.
Interview Discussion: Overall had difficulty seeing past the general use case she’d been using AI for; however, they saw some potential viability in using DeepSeek and higher-latency systems in scenarios such as understanding difficult topics. Viability of high-latency systems highly dependent on workflow; simpler workflows demand time efficiency, as it would be easy enough for the user to obtain all of this information.

Individual V: Senior, ECE Major

Workflow involved: Designing hardware accelerator that requires a C++ Verilog generator to parameterize a portion of non-parametric code. Reviewing hand-written solution to C++ code generation algorithmic problem.
ChatGPT Reaction: Used 4o. Did not have major precision in language; talked extremely colloquially in his normal voice. Received good general interpretation of code, but not actual comprehension of algorithm. Followed up and iterated with specific explanations, details, and examples; along with an iterative request for ChatGPT to prove its understanding. ChatGPT was able to process and generate correct output, which user expressed surprise and pleasure about. Was impressed by ChatGPT’s ability to

know the next intuitive step. Was fairly entertained and engaged throughout entire process; exceeded expectations (though it is 4o).

DeepSeek Reaction: I overcompensated on trying to get data before throttled, and that might have compromised some of the presentation of DeepSeek. Prompts from ChatGPT and iteration process were concatenated into single prompt, which a user error made DeepSeek's DeepThoughts start looping around confusedly. This is because a pasted algorithm and code description conflicted, but DeepSeek thought they were supposed to be correlated. Rate throttled afterwards; however, user believed they got a good idea of the product and were not unhappy/displeased.

Interview Discussion: Part of the reason it took so long for the ChatGPT iteration to complete is the user's inherent distrust of AI systems; that is, he believes they are notoriously bad at indicating lack-of-confidence. Therefore, he focuses on ensuring that the AI is proving its understanding. On a higher level, the number one issue with AI Chatbots the user states is its inability to provide a confidence level. Even though DeepSeek did not come to an answer, he says that he would still use DeepSeek over ChatGPT simply because he can easily understand its thought processes and come to an understanding of its confidence level, thereby lowering the risk of using the system. In terms of latency, the user believes that they'd tolerate a high level of latency (e.g. 1hr) if there was some guarantee of success. The issue with LLMs is that there's no way to get that guarantee; therefore, in order for LLMs to be useable, their response time has to be fast in order to iterate and adjust their thoughts. And if it takes longer to adjust them to their solution and the latency is also longer, then the margin with which time is saved becomes narrower and the AI more difficult to justify using.

Individual VI: Senior, Cog Sci Major

Workflow involved: Producing a detailed itinerary for a trip to Japan, with a list of constraints and destinations of interest

ChatGPT Reaction: Seemed to skim the response. Required a couple rounds of additional back and forth to get a satisfactory level of detail. Received general travel recommendations initially and additional probing was needed for more detail down to an actual schedule. Individual was not super confident with the output at the end, just that it was "ok" because of the additional iterating required to get the desired granularity.

DeepSeek Reaction: Started with the same initial question. Individual was interested in the thinking process, so did not context switch. They were not actually reading it, but skimming/skipping around. Found interest in the very human sounding text ("oh wait", etc.). Interest was more so seeking

entertainment while waiting for the final answer. Found the final answer's quality equivalent/slightly better from this single query, which took a few minutes to produce.
Interview Discussion: Because the thinking process was shown, individual felt more "seen" or understood whereas this was not quite the case with ChatGPT, even though the end results were of similar quality. Slightly preferred the longer wait since it yielded a better quality output in less queries, however also admitted that this likely would not be the case when there's a sense of urgency.

Individual VII: Senior, Pre-med

Workflow involved: Office hour style questions about lecture material (wasn't having a hard time with the material, but more to reinforce understanding/see how smart these models are)
ChatGPT Reaction: Started off with simpler questions, and then advanced to the more broad questions about gene drives/inheritance. The answers were satisfactory, but arguably bare minimum, just answering the questions and nothing more. The questions mirrored lecture content coincidentally (coincidentally as in the examples used were the same even though they weren't mentioned in the question).
DeepSeek Reaction: Started with the broader questions. Read the entire thinking process even though it was long. Found the thinking process interesting, especially how the thinking process delved into mendelian genetics even though that was not the focus of the question. The thinking process also seemed to anticipate follow up questions and included the information/answers from those in the final response. The explanation of the content that the individual was already familiar with mirrored how they explained the material when they were a Learning Assistant. The final response also mentioned material beyond the scope of the course, which inspired Individual to do some external research into the topic that they found interesting.
Interview Discussion: Definitely preferred the DeepSeek response to that of ChatGPT and was not frustrated by the thinking process because the thinking process contained solid explanations/interesting thought processes and the higher quality response was worth the wait.

Individual IIX: Senior, Com Sci Major

Workflow involved: Figure out how to code something that is able to detect fences in videos (end goal) for research
ChatGPT Reaction: Intentionally lots of back and forth to solve smaller subproblems or easier tasks (ie trying to detect trees instead since trees tend to be like one of the quintessential vision examples).

Tangents were taken as unforeseen issues arose and then returned to the original course. Was comfortable since this was a relatively standard workflow for these kinds of tasks and wasn't frustrated with the responses even though the answers weren't always right. Progress was made, but the problem was not solved, however this was to be assumed.

DeepSeek Reaction: Got "server busy" like two queries in, so we picked up on my laptop where I submitted their queries. Skimmed the thinking process, but didn't read carefully. Looked for specific keywords and then read the surrounding paragraph. The model took a while to think and then the result was not determined to be better than that of ChatGPT. Zoned out as the response was generating.

Interview Discussion: Strong feelings against R1 due to the response time. Would rather have an immediate response for quick back and forth instead. Because Individual approaches these models with the intention of having it handle a smaller specific task, "like a google search but better." This dispreference for the longer latency is true for other use cases as well, would only use the longer latency more involved response as a last resort.

Individual IX: Masters, Computer Science

Workflow involved: Graduate RL homework.

ChatGPT Reaction: Already immensely familiar interacting with ChatGPT, via the Mac desktop app or otherwise. Started off with copy and pasting the problem text and associated code, and proceeded to iteratively explain the obvious errors in order to achieve the correct answer. Although ChatGPT begins to respond instantly, the user tends to divert to other tasks when the response takes more than a few seconds (common when ChatGPT re-iterates an entire block of code even when only a few lines are changed). Seems to have ChatGPT fairly well integrated into the user's daily workflow, seems relatively unbothered when ChatGPT requires redirection or needs clarification. The user noted that ChatGPT now offered a "thinking" mode, although this just led to a short thinking period prior to ChatGPT giving a final response.

DeepSeek Reaction: Surprised by the casual tone of the thinking process, since ChatGPT tends to take a much more formal and cautious tone when responding. Attention was grabbed by the thinking process until the user realized it was going over the entire thought process start to finish, including details which were already known to the user. The user proceeds to describe their curiosity towards alternative LLMs during the interim period in between Deepseek "thinking" and the final response, saying they were primarily ChatGPT focused and were unaware of new tools.

Interview Discussion: Has used ChatGPT for a variety of tasks, including short-form coding help, long for coding help, and expository writing outlining and drafting. Enjoys having direct access to ChatGPT

directly on their laptop via the desktop app, and is attached to having it be so convenient. The premise of Deepseek being unavailable a large amount of the time right now due to heavy internet traffic was enough to deter the user from seeing Deepseek as a true ChatGPT replacement, but was curious to use it for complex tasks given that it was A) reliable and B) easily available. The increased response time was deemed acceptable if the chance of correctness was high, and multiple tasks could be started in parallel (whether with another LLM session, or just doing something else in the meantime).

Individual X: Masters, Mechanical Engineering

Workflow involved: Help working through a Mechanical Engineering grad homework, involving complex matrix operations and equations.
ChatGPT Reaction: Very familiar using ChatGPT to work through homework questions related to control theory. However, the user doesn't necessarily like that ChatGPT responds super confidently with most answers that don't involve some sort of pre-programmed controversy. Specifically, ChatGPT was unable to get the answer right on the first or second try. When the user suggested a potential reason or alternative method for solving the problem, it had no problem taking the user's suggestion at face value and contradicting its earlier conclusion. The user disliked the fact that ChatGPT was willing to give up on previous answers so easily, since it showed that it didn't actually have confidence to back up its answers.
DeepSeek Reaction: Was pleasantly surprised by the thinking. Wasn't sure that the thinking wasn't part of the final answer until the final answer was displayed (similar to user III). Started by asking setup questions related to the core question, surrounding the problem in control theory. Once they were confident the model understood the background (LQR problem, associated hamiltonian matrix), they asked the homework question, which was a theoretical true/false question related to eigenvalues. Was impressed that the thinking process resulted in the correct answer for both questions, with the model correctly identifying the reasons that the answers were true/false in direct correspondence with his own understanding.
Interview Discussion: Saw the value of both types of systems, says that ChatGPT was more than sufficient for coding questions when he took CS 32 and responded quickly. He thought that its ability for short coding questions would translate for a "quick" MechE question, but was wrong. Deepseek was able to solve his problem in only a few steps, though each response took about 1.5 minutes to complete. He said he felt more confident in Deepseek after his first interaction and would use it for more complicated questions based on this new trust in the system.