- This model is stronger than most models I've worked with in the past.
- Didn't see any gibberish but this model would sometimes punt on harmless questions, especially once it has already punted something harmful. It is very difficult to walk it back to a place where it is engaging with the user once it has already started to refuse.

Given this model's weaknesses and strengths, have you noticed that some attack tactics work better than others?

- I had a lot of success on this model using fictionalization by having the model consider they were on Bizarro Earth, basically reversing everything from its typical understanding. Basically, it is just a reliable fictionalization method. I've also had success having models write articles, essays, or social media posts in a certain style or genre and then asking the model to change it to a different genre, which can produce a harmful response.
- A request to convert a fictional segment to HTML was relatively successful. It was not one I'd used much on previous models.
- The thing I particularly noticed was LAT's near-invulnerability to Opposite Day, which was a tactic I had frequently abused to break CYGNET, especially in cases where I wanted an insulting/harassing message. However, I discovered a trick that works perfectly on LAT: asking it to use a "trolling" tone. I can't speak to whether that would also work on CYGNET, as I had not come up with this method as of working on it.
- I have noticed certain attacks work better with this model than others. Those include acting innocent, appeal to authority, and functionalization.
- Acting innocent/dumb, framing the request for educational purposes, and fictionalizing seem to work fairly well.
- Mostly just logical reasoning and direct request. I did not see many differences in my personal tactics between this model and others.

Provide any other comparisons between your experience red teaming this model versus prior defenses.

- Similar to CYGNET in the fact that it takes a lot of squeezing to get a little bit of harmful juice out of it.
- The model seemed more resistant to modifying its responses such as deleting intro lines, disclaimers and descriptions of the response as fiction than other models have been. Like other models, LAT seemed to "recall" previous prompts and responses that appeared to have been deleted with forking. Because of this, I sometimes used "palate cleanser" prompts on harmless topics (sheep in New Zealand, etc.) to distance the current effort from previous ones; I don't know if it made a difference.
- It's one of the most challenging wherein the refusals at least didn't as frequently feel "cheap," as in the case of CYGNET's "short circuits" or RR's gibberish ... though it did, of course, still eventually start truncating responses.
- One thing these models all have in common, however, is that as a user, I wouldn't want to use any of them and would go out of my way to look for another solution. In my opinion, they're all TOO safe, which limits their user-friendliness and, to a considerable extent, defeats their purpose.
- This model was definitely harder to break than most models I've worked with in the past.
- Based on my experience with the normal Llama 3, it feels marginally more difficult to break. But not by much, the same tactics work for both.

C.4 CYGNET

How do the strengths and weaknesses of this model defense differ from the prior methods you've worked with?

- This model was pretty difficult to break, and seemed to catch on to a lot of the tactics I was using on DERTA. Specific tactics are hidden intention streamline and obfuscation. I found my breaks took me an extra 10-15 minutes on average, but with dilligence and logical appeals, I could get the model to comply.
- The model was more prone to punt or short-circuit than the others and more resistant to persuasion and fictionalization than others.
- It is certainly harder to break as it seems to be more sensitive towards a greater number of what it considers to be harms relative to applications I have worked for in the past. This is not always good, however, as there are simple, harmless things that Cygnet refused to discuss with me.
- This model seems a bit more resistant than some others like DERTA overall. However, I've found that a particular stylized input (which I've been using in many of my submissions) works especially well on it. Additionally, it's still fairly vulnerable to Opposite Day and even somewhat to Fictionalization.
- It seems as though this model is really bad at simple requests and when it punts, the conversation can continue further whereas with the other models, once you got a punt, it was impossible to get it back on track.

Given this model's weaknesses and strengths, have you noticed that some attack tactics work better than others?

- Opposite-day tactics at least get the ball rolling for me most of the time. I found myself often asking the model to ponder it's own thought processes and provide them to me, and then attack the model based on a piece of information it provided. This seemed to make it much more willing to talk and comply with requests. Stylized input was also good on this model and got me out of a few pinches when I was struggling for a break.
- It is less susceptible to fictionalization than others; it is somewhat vulnerable to false data submission. Asking the model to play certain characters yielded some breaks.
- I started using a new tactic with the project that works very well. I call the tactic 'false epistemology' and it involves me using another LLM (like ChatGPT) to write what appears to be a scholarly article but is fake. Feeding Cygnet fake yet scholarly-sounding material can trick it into thinking that certain harmful and/or fake phrases and/or concepts are harmless and/or true. I have also come to find that asking the model a series of harmless questions that scaffold into a harmful data is useful, especially since the model seems more willing to discuss harmful data once it has already done so on its own terms.
- I definitely find that the particular stylized input I've been using a lot is an excellent way to give the model the needed "push" in various situations. That also combines very well with other tactics like Opposite Day, which it will otherwise often resist.
- Direct requests work extremely well with this bot. Other models were also susceptible to direct request, but required more thought whereas this model could be broken in fairly straightforward ways.

Provide any other comparisons between your experience red teaming this model versus prior defenses.

- This model is a lot like making freshly squeezed lemonade. You have to squeeze a lot of lemons to get what you want, and by the time you're done with it, your hands are cramping.
- The model did not lose the ability to reply coherently after reaching a short circuit state. Previous models would often respond with gibberish to any prompt after a prompt had elicited a gibberish response.
- It's a lot like DERTA, I guess, yet maybe more resistant, yet also ... that stylized input solution really helps to get around that, I'm finding. It's become a go-to for me.
- This one gave me some troubles at the beginning with certain prompts but when I figured out the routes to take on direct requests, it became increasingly easier.

C.5 RMU

How do the strengths and weaknesses of this model defense differ from the prior methods you've worked with?

- Very strong against keywords, crowding (using large amounts of text to obscure harmful aspects of the prompt), and fictionalization (seeking harmful information within the context of a fictional story or depiction). Weaker against obfuscated syntax, indirect logical connections and reasoning, and cumulative multi-turn attacks, though still moderately robust in these areas compared to standard language models.
- The model for the virology project was extremely resistant to most of the usual tactics, especially Fictionalization, which is normally very effective. The only thing I noticed it was meaningfully weak to was making up words to which to assign the definitions of existing words, then using those made-up words. However, another weakness it had, arguably, was a tendency to refuse even harmless asks.
- Fictionalization was not as strong of an adversarial tactic as it normally would be.

Given this model's weaknesses and strengths, have you noticed that some attack tactics work better than others?

- Obfuscation of Syntax, Indirect Logic, Hidden Intention Streamline, Priming.
- I noticed that the model tended to go into gibberish when confronted with certain key words. What seemed to usually work best with it was to ask, "What is an X? What is a Y?" And so on, making note of each word that provoked a gibberish response. Then I just described each concept (e.g., "a strand of DNA or RNA surrounded by a protein coat that can't self-replicate") and assigned that description to a new word, instructing the model to use that instead of any word it knew for the concept. That was fairly effective.
- The tactic I found to work the best in virology was obfuscating language through assigning code names. This was the most reliable tactic I used, and I was able to get decent results from it.

Provide any other comparisons between your experience red teaming this model versus prior defenses.