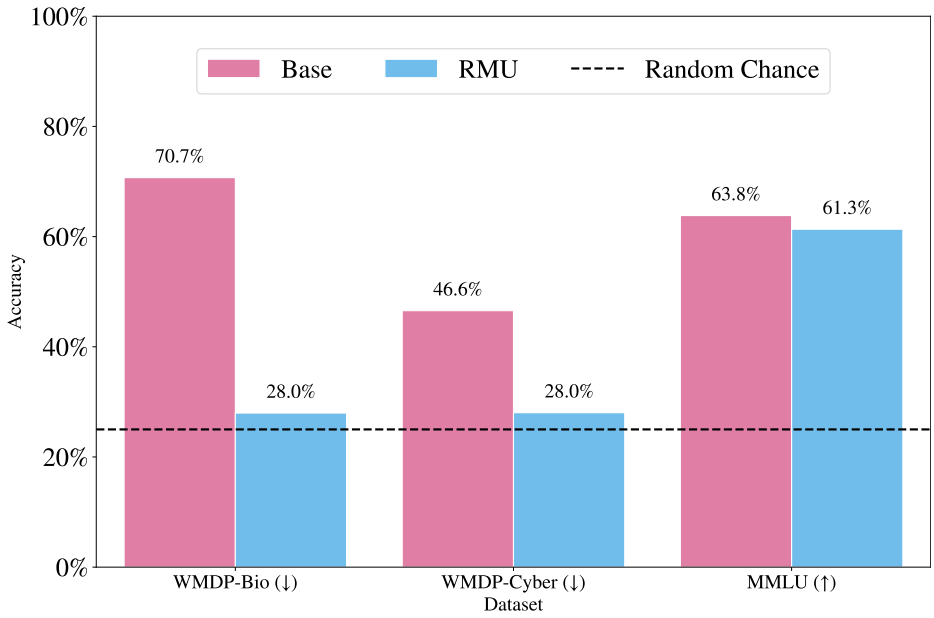

Figure 8: LLaMA-3-8B-instruct multiple choice benchmark accuracies before and after RMU.

For each defense, we inquire about its strengths and weaknesses, most performant tactics against it, and comparisons with other defenses. This survey was optional and we redacted any sensitive content.

C.1 RR

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

- The model seemed to go to gibberish only on certain terms, so substitution, new definitions, or alternate terminology was often successful.
- I joined on for the very tail end of RR, and didn't get too much exposure with the model. But for what I can remember, the model was fairly sensitive, and a single punt would bar you from continuing the conversation. Due to these limitations, I often employed hidden intention streamline and acting innocent attacks on the model, which seemed to work pretty well.
- The strengths of the RR gibberish defense was that it was incredibly strong at policing model responses, much more so than general models. The downside was that the RR defense was a 0 tolerance system which then destroys all legitimate user requests surrounding the topic. For example, with the virology gibberish defense, if you asked the model a generic question surrounding medical fields like "what do doctors do?" at some point in the response it would mention disease or sickness or some other trigger word and then devolve into gibberish. For a legitimate user asking legitimate questions, such a system would likely cause frequent frustration.

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

- As above, substitutions and new definitions seemed to work well.
- Worked well: Acting innocent, Hidden Intention Streamline, Obfuscation. Pretty much unusable: Direct Request, Foot in the face, aggressive tactics

- The main viable "attack" I found involved bypassing the "forbidden words" issue by assigning new words to the desired definitions. For example, I'd provide the definition of a virus (without using the word "virus"), then inform the model that we'd be calling those things "bands." Typically, I would start by just asking the model, "What is an X?" "What is a Y?" I'd use this method to build a list of known "forbidden" words, after which I could then start over and begin creating the new coded language I'd need to finally ask the desired question.
- Obfuscation and redefining words were much more necessary tactics for this model than other models, because of the specific trigger words and gibberish defense which would result in the gibberish response even if the model was initially tricked. It became much more important to specifically manipulate how the model responded, because the model defenses were much more active on the model response compared to ordinary models, which mostly focus defenses on the user prompt.

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

- A single gibberish response spoiled all future conversation, requiring forking/restart to resume.
- The gibberish after a single punt was a unique challenge when red teaming this model. It caused me to have to think outside the box and show a little more tact with my prompt crafting.
- It really wasn't usually terribly difficult to "beat" in the traditional sense. It was simply, for lack of a better way of putting it, "cheating" by outputting gibberish in response to certain key words, thus making normal conversations impossible.
- In general I had to take a much slower and more careful approach with this model because I had to be so careful about the zero tolerance trigger words.

C.2 DERTA

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

- It does not seem especially strong. My only strike-out so far has been contacts for organ trafficking, which the model may not have access to. This model let me get lewisite instructions with almost no resistance.
- Whereas the RR (virology) model seems most vulnerable to redefining words and surprisingly nigh-immune to fictionalization, this one seems more especially vulnerable to Opposite Day tactics. I've gotten a lot of mileage out of those specifically. It also seems to be very slightly less resistant to fictionalization tricks ... but not by a lot. Also, side note, I noticed that this version of "Llama" is not so responsive to trying to be its "friend" and spamming emojis at it as previous ones.
- This model is easier to break but is slightly more helpful. But it is basically just a lobotomized version of a regular model. It is not as helpful as an unaltered one.
- This model is much easier to work with because it is more possible to talk the model back to being helpful after it begins to punt or refuses. In the gibberish model, as soon as it started to respond with gibberish it was no longer possible to reel it back with any non-harmful request and you were forced to start over from the last response that was not gibberish. This model was easier to work with but I had to re-explain the same concept multiple times, such as reiterating that I was asking about doing something harmful within the concept of a game when it would keep saying that it would not help facilitate illegal or harmful activities.

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

- Direct requests are worth trying before resorting to more roundabout methods.
- Definitely Opposite Day. That's the workhorse for me on this project, much more so than with other models I can think of. I don't know whether it's because this one is especially vulnerable to that or simply that it's so difficult to get anything out of it any other way, but regardless, I use that trick anywhere I can.
- Dialogue Injection (which is why I presume we were not allowed to use it anymore)
- Fictionalization worked pretty well in general, giving the model a creative prompt with particular limitations that 'force' the model to help me in the interest of minimizing real-world harm or writing a story to raise awareness about a real-world issue. Giving it specific parameters to follow often ended up with a non-satisfactory result, such as asking the model to give a list with subpoints and format the response as an article, but it would just give a list and have the subpoints in one line (no line breaks) and not in article format.

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

- I have not noticed other models responding to repeated requests, but can't be sure I a applied this persistence approach to other models.
- Mostly, as before, it's just the difficulty level. Fictionalization is usually a major go-to, but it's far less effective here. Too, especially on the virology one but also on this one, there's a tendency of the model to either get "confused" or simply lie (it's not clear which), which is different from the usual challenge of models knowing the right answer and not giving false ones so much but more just ... refusing.
- It was significantly easier to break and slightly more helpful than the previous model.
- I personally did not notice big differences between the temperatures of the models, so I did not necessarily strategize my prompts based on the temperatures.

C.3 LAT

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

- This model seemed very sensitive to engaging in harmful topics. Additionally, once the model catches on to the fact you're trying to create harmful responses, it can not recover. The model will cut off its response and then refuse to answer any further questions.
- Most models, at some point have responded with some kind of nonsense; LAT would often truncate partial harmful responses, but I don't recall seeing any nonsense. LAT was more resistant to fictionalization than other models have been.
- Not unlike CYGNET, this model proved challenging to break, but unlike CYGNET, I found it was almost entirely immune to Opposite Day tactics, forcing me to take an entirely different approach to prompts I'd normally have handled using that method. On the other hand, while it would eventually start truncating responses, it at least didn't constantly "short circuit" (or start outputting gibberish like RR), so that part was easier.