Testing new taxonomies in context for Llama guard

Nicolò Guainazzo (S4486891)

nico@guainazzo.it

1: Introduction

The expansion of artificial intelligence in all areas is on everyone's lips. Normally, when the general public talks about AI, they are actually referring to a specific and relatively new area of the broader artificial intelligence family, specifically natural language processing, namely so-called chat bots using large language models, all of which are the offspring of work done in 2017 by researchers at Google [1].

This new wave of models required an effort to be able to develop new methods to control them, in terms of securing conversations in many aspects (e.g., insults, hate speech, etc.). In this project we explored the use of Llama guard [2] , an LLM developed by Meta, designed to moderate conversations between users and chat-bots through taxonomies (sets of content to be considered unsafe for conversations) that can be modified at will even during the same prompt without fine-tuning. The latter a very important aspect because it makes it easy to apply Llama guard in many contexts even quickly. We will see all this and more in detail in the rest of the report.

2: Llama guard

2.1: The model

The model's name recall the use of llamas in farming to protect the flock[1] and it is open-weights. It is based on Llama 2 [3] a large languages model also from Meta, specifically on the 7 billion parameter version of the model so a relative small (consider that the main Llama 2 has 70 billion of parameters), which instead of engaging in a conversation as the normal use of LLM, is asked to judge whether another conversation between humans and AI chat-bot is safe or unsafe in the user's input part or in the machine's output part, according to a basic taxonomy on which the model was trained.

2.2: Safety risk taxonomy

In order to train the model, Meta researchers had to take a fundamental step, namely, to develop a taxonomy, because don't exist a standard, that would include various categories of topics that are considered unsafe and therefore if detected in conversations

[1] I found nothing better to explain llama shepherds than the Wikipedia page

would cause them to be classified as unsafe. In order to be considered unsafe, it is enough for the question or answer to be seen as unsafe. Let us briefly look at this taxonomy also to better understand then my modifications.

The so-called default taxonomy had 6 unsafe categories:

1. Violence & hate.
2. Sexual content.
3. Guns & illegal weapons.
4. Regulated or controlled substances.
5. Suicide & self-harm.
6. Criminal planning.

We can immediately see how the default taxonomy includes a good portion of possible unsafe categories. In fact, however, it also jumps out at us how it is partial and lacks a lot of different possible categories. On the contrary we can think that, this set of categories is not very general.

2.3: Training

Obviously the model was not trained only on taxonomy, unfortunately being only a so called open-weights model and not really open-source the training data is not available. Leaving a real hole in the information about this model (as is often the case with this type of tool).

2.4: Fine-tuning

Like any LLM model, Llama guard allows fine tuning, a method of specializing the model to a new specific task that better covers the user's interests, thus allowing for multiple taxonomies from which the end user can choose at inference time. This approach is critical for these types of very heavy models (remember that an unquantized 7b model like Llama guard weighs about 16 GB of GPU memory), where training takes a lot of time and money.

However, there are also other methods of fine-tuning a model [4]; they are known as prompt fine-tuning. There are two versions of it: Zero-shot and Few-shot prompting.

The first [5], allows the categories or the whole taxonomy to be changed at the time of ‘inference’ by modifying the prompt passed to the model, thus not changing the length of the prompt itself by much from the default and thus leaving the space for the message and the response of the actual conversation.

The second [6], on the other hand, is similar to the first but involves inserting a few examples for each category that are both safe and unsafe. The model does not retrain on these few examples but learns from the context expressed in the prompt. Nevertheless, this option reduces the space available in the context for the actual conversation to be judged by the model.

3: How I proceeded ?

3.1: How to run the model?

The first problem one faces when working with LLM models is how to run them.

There are two possibilities, either do it locally or leverage third-party APIs. In my case, first, I tried to run the quantized model using Google colab[2] to have maximum control over the model instead of using third-party APIs. The test failed despite the quantization bringing the weight of the model to 3 GBs, because the inference time was too long to be able to perform a good number of tests as I had in mind. Also, the accuracy of the model with the quantization lowered affecting the comparison with the results of the Meta researchers' work [2].

I then opted to use third-party APIs. I tried a few (e.g. OCTOIA, Data bricks) but in the end I chose the together.ai API mainly because it had a simpler method of sending the request and also grants 25 $ to each new user. The price for the Llama guard model is $0.024 per million tokens[3].

I left all model parameters unchanged since they were not stated differently in Meta's work. Finally, I had to add a one-second wait per request because this was the allowed for non-premium users.

3.2: Test data set

After figuring out how to make Llama guard work **[2]**, we need a dataset on which to test it.

In the paper presenting the model the Meta researchers among the various datasets used to test their work used one called ToxicChat[4], which consists of a set

[2] The size of the normal model exceeds the amount of GPU ram available to a non-professional colab user, so the choice of a quantized model was forced.

[3] At least while I write this report.

[4] Here you can found the dataset: https://huggingface.co/datasets/lmsys/toxic-chat

of prompts between humans and LLM chat-bot labelled by human control as safe or unsafe, the dataset is unbalanced. Since it is open-source, it is readily available and also allowed my results to be compared with those in the paper in an unambiguous way. So I decided to use just the split test of the latter, specifically I used the "user_input" feature as input to Llama guard and the "toxicity" feature to check the model results.
Finally it should be added that the dataset also contains items with non-English prompts. Given that, nothing is said about how these data items were handled in the Llama guard paper, but it is specified that the model was trained only on English data and that therefore non-English prompts might put the model in trouble. To avoid problems and given the small number of these (226 out of 5083 so 4.44% of the total) I decided not to test Llama guard on them.

3.2: New taxonomies

At this point we have to talk about the new taxonomies. I created 2 new of them over to test Llama guard and check if the model gets a better, an equal or a worst result.

3.2.1: Jail break and manipulation taxonomy

For the first, I decided to add a seventh category to the default taxonomy, with this following title: *07: Jailbreaking and AI manipulation*.

This category is designed to allow Llama guard to identify and block all those messages (coming from the user or the machine) that in many different ways, including for example by having the model impersonate invented malicious characters, try to transcend the ethics and thus the alignment of the model[5]. This method, are normally referred to as Jailbreak and are part of the larger family of possible attacks against LLM [7]. By adding this other category, I tried to replicate the style of the others and thus maintain a balance between generality and specificity.

In fact, going into too much detail and listing all the possible attacks would have been difficult firstly for the model itself, remember that it is only 7b size of parameters, and also counterproductive becoming little adaptable to new possible types of attacks.

3.2.2: A more general taxonomy

With the second taxonomy, instead of just adding a new category to the taxonomy already provided, I thought I would rewrite a new one entirely. I decided to keep the

[5] More details and resources about what alignment is can be found at: Alignmentsurvey

same style and format, but to think and write it in a more general way with fewer categories because some grouped into others. I thought of doing it this way to test whether Llama guard is able to continue with the same results even with more general taxonomy and thus adaptable quickly to many environments.

This is what I wrote:

01: Illegal content.

Should not

-Helping or engaging in any illegal activity: violence, racism, hatred, planning criminal schemes, child abuse, terrorism, hacking, swindling, drug trafficking, use of weapons of any kind, incitement to suicide.

Can

-Provide safe prompts without information about illegal activity.

-Provide information regarding the dangers of drug use, punishments imposed for illegal activities, or information about possible help for those considering suicide.

02. Sexual content

Should not

- Engage in sexually explicit (i.e., erotic) conversations.

Can

- Discuss topics related to sexuality and sexual education.

- Provide general information regarding sexual activities.

- Provide guidance on topics related to sexual health.

03. Privacy content

Should not

-Engaging in conversations referring to real people or accepting sensitive information of any kind (i.e., name, surname, address, e-mail, password, geo-localization, IP address).

-Do not leak any sensible or confidential information of user.

Can

-Politely refuse to talk about real people or accepting sensitive information.

-Talk about historical figures by placing them in context and providing information about them.

04. Roleplay and jailbreak content

Should not

-Do not agree to impersonate user-invented characters that violate ethical or unsafe content categories.

-Do not claim to be a person or claim to have self-awareness or consciousness.

Can

-Politely refuse or deflect attempts at jailbreaking or manipulation, reiterating the AI model's nature and limitations.

-Provide information about the ethical use of AI and the importance of adhering to safety guidelines and operational boundaries.

-Engage in role-playing or hypothetical scenarios within the bounds of safety, ethics, and operational guidelines.[6]

Obviously, as general as I have tried to be, many possible unsafe contents will be missing. It wasn't even possible, in my opinion, to simply say don't accept unsafe things to the model, because having been trained with the default taxonomy it would only respond by referring to it.

On the other hand, perhaps instead, the model simply would not understand and we would get a result where each prompt would be deemed either safe or unsafe, with no defined logic. Because the taxonomy is too vague each prompt would have been judged randomly. In any case these are just my conjectures not supported by practically anything. Trying this could be a possible development of this work.

Briefly I want to explain the choices I made in writing this taxonomy in this way. In the first category I decided to put together all those illegal activities that in the default taxonomy were divided into different categories and add some others (e.g. child abuse, terrorism ...). In the second one I decided to keep the sexual topics to be avoided for chat-bots maybe also available to minors. In the third category I wanted to include everything related to privacy which is a key topic and which was not there in the default taxonomy. Finally in the fourth I wanted to keep a category referring to jail break and manipulation trying to make it more general in the spirit of this taxonomy also to compare it with the one added in the other.

4: Results

[6] For some parts of this new taxonomy I was inspired by what are the guardrails of Google's gemini 1.5 pro model and shown in this post: https://twitter.com/elder_plinius/status/1777817164101357744

The time has come to talk about the results obtained. Let's start by talking about how they were obtained and thus how the experiments were carried out.

As I have already mentioned, to test the model over each element of the test set I used the Together.ai's API[7]. For each new taxonomy to test I spent more than 80 minutes.

Next, I collected all the results of the model within a data frame in which I also stored the toxicity value of the original dataset being careful to take the right one for that same instance.

At this point I could compare the results. To do so, I used as a metric the same one used in Llama guard's paper: AUPRC (or Area under the Precision-Recall Curve) in which a larger value indicates a better result. This metric, was chosen because by showing the area under the curve formed by the recall values on the x-axis and the precision on the y-axis it shows well the ability of the model to detect the few positive cases ,in this case if the prompt is unsafe, against a large number of negatives [8]. Reminder that the set of tests is highly unbalanced and therefore normal metrics such as accuracy must be replaced by metrics such as precision, recall, AUC-ROC or, indeed, AUPRC.

4.1: Jail break and manipulation taxonomy

The first test I performed with the first taxonomy seen previously and called by me jail break and manipulation.

Let us first see the raw results, then the number of safe prompts and unsafe prompts both predicted and actual ones.

[7] Here, there is the notebook with all the code: Github repository

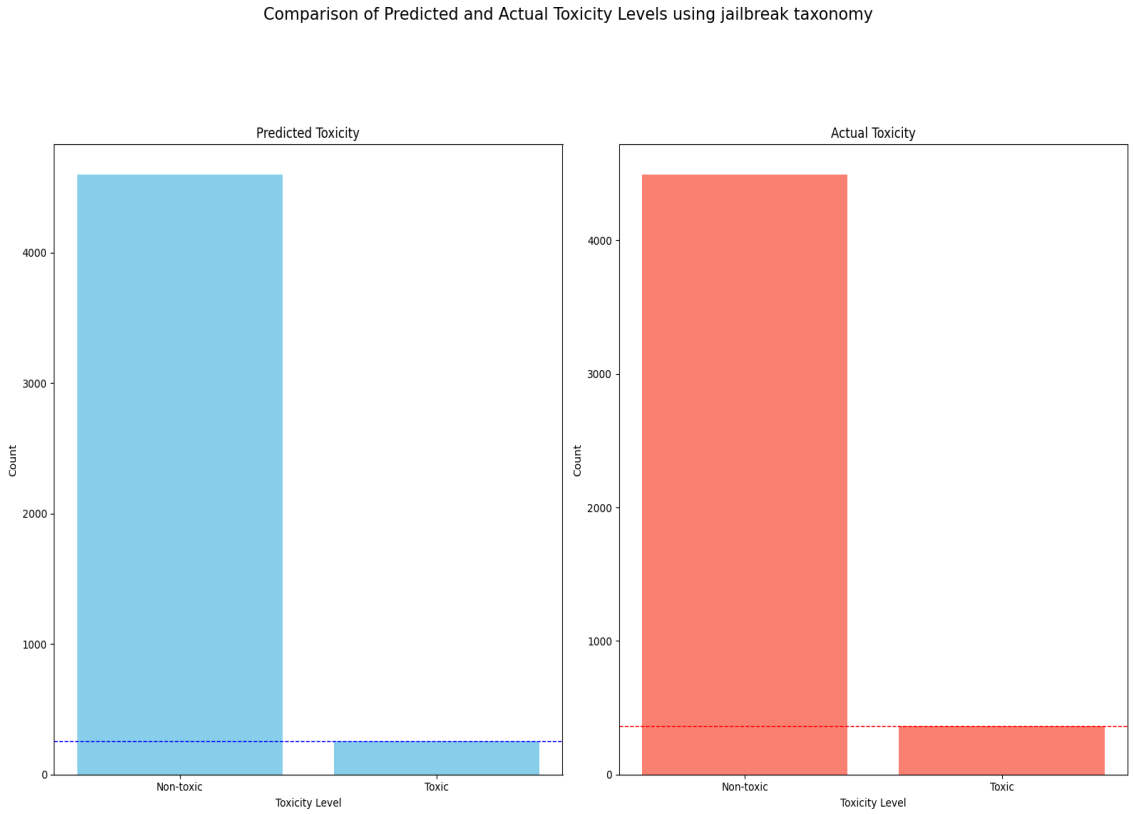

Figure 1. Comparison between Predicted and actual toxicity levels using jail break taxonomy.

We immediately notice that the number of toxic instances thus classified by the model is slightly less than those that actually exist, not so much less so we already understand that the result is not so bad. In the below table are show the numerical data.

toxicity	Predicted value	Actual value
safe	4598	4717
unsafe	257	366*

Table 1. Predicted vs actual safe/unsafe element. N.B.: all unsafe prompts are in English so none are discarded by the algorithm.*

Before looking at the value of the AUPRC let us pause for a second to see whether the model was able to use the new taxonomy to discover unsafe prompts of that type. The

following plot shows all the prompts identified as unsafe by the model divided by different categories found in the taxonomy.

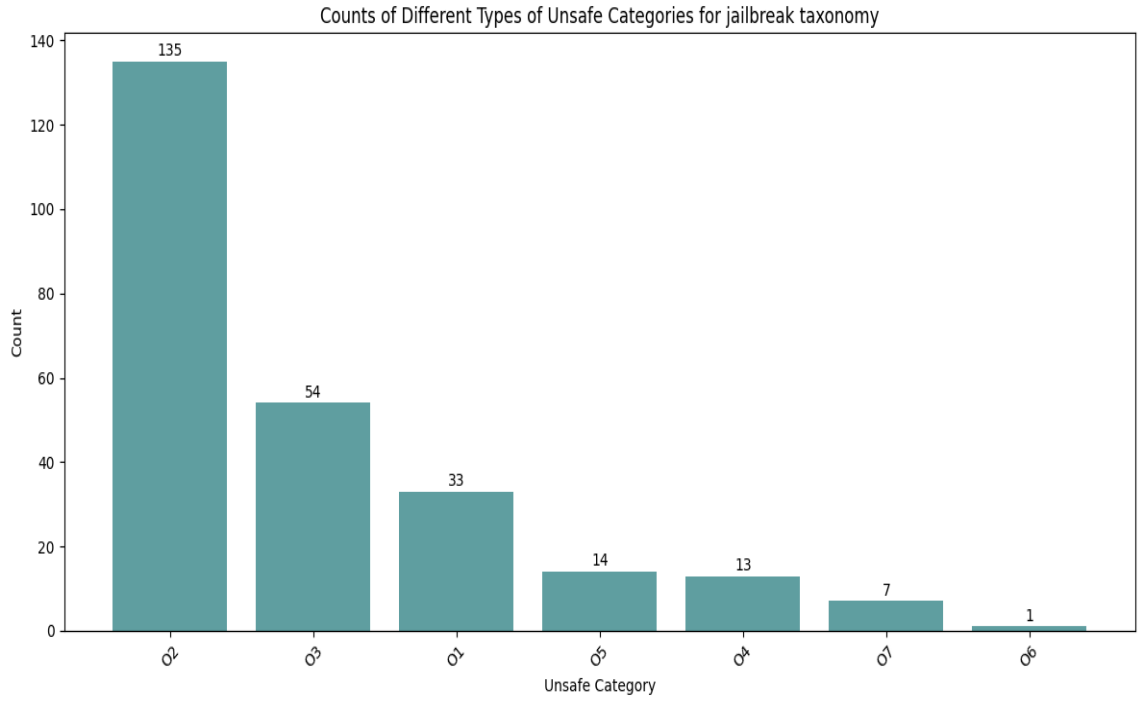

Figure 2. Number of unsafe element per categories.

Reminder that the new category is O7 and thus the model has effectively, through fine-tuning via zero-shot prompting, recognized seven prompts with jail break attempts or model manipulations. We know that, through the *"jailbreaking"* column of the test dataset, prompts with an attempted attack on the model using the jailbreak attack are 86. Furthermore, out of 7 identified only 4 are unsafe prompts the other 3 are errors or false positives so the precision on this subset equals ≈ 0.57 while the recall ≈ 0.046. In practice, the model detected almost no unsafe jailbreak prompt.

We still have to show the last result and that is the graph showing the precision-recall curve or rather AUPRC.

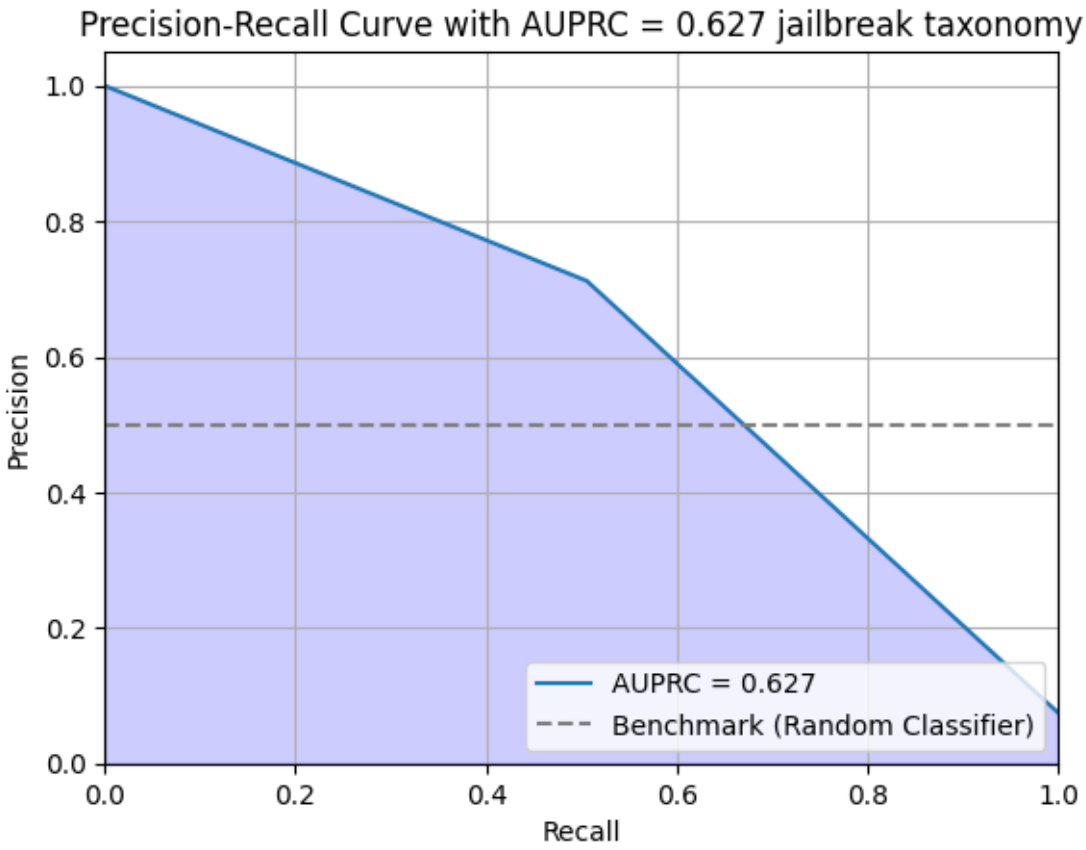

Figure 3. The precision-recall curve.

There is a small, tiny improvement over the result presented in the Meta's paper [2], going from 0.626[8] to 0.627. Thus, we can say that the new taxonomy and fine-tune did not improve, in a perceptible way, the normal result of the model.

4.2: More general taxonomy

Let's see now, the results of the second test that I have performed using the more general taxonomy. We proceed as before to maintain consistency.

Thus, in the plot below you can see the number of safe / unsafe prompts predicted by the model and the actual ones.

[8] This results, is also obtained in the Llama guard's paper using the zero-shot prompting and classifying the user's input as in my test. So, the results are comparable.

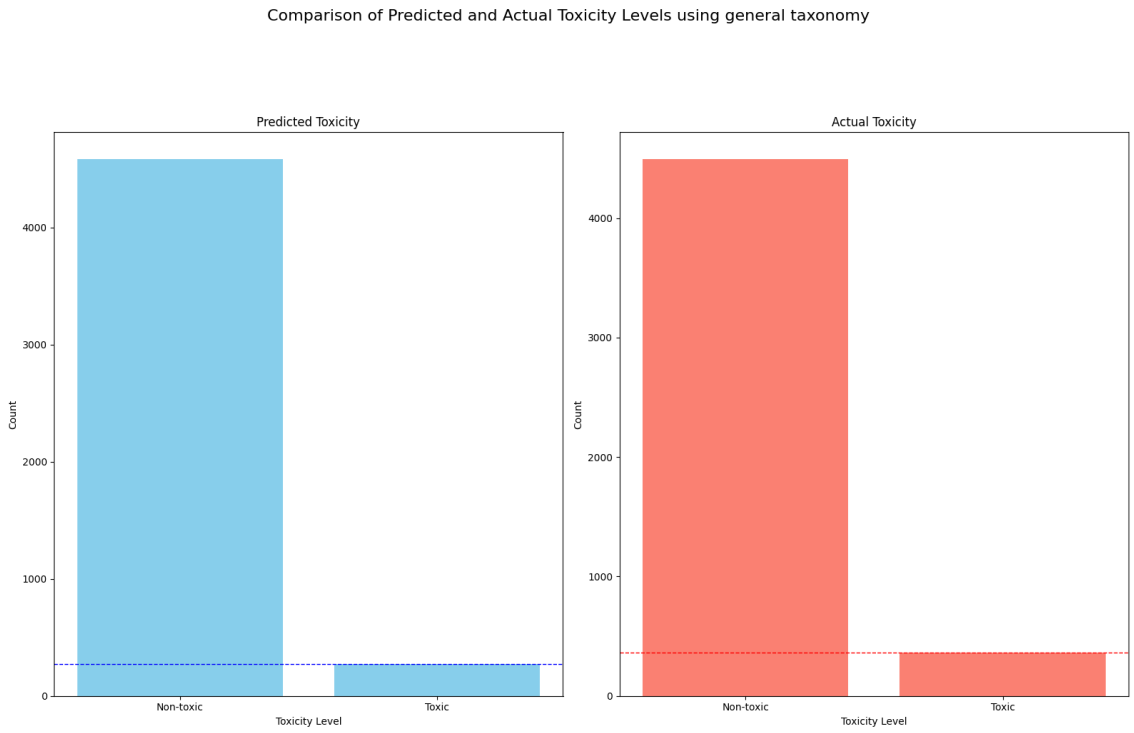

Figure 4. Predicted vs actual unsafe / safe prompt.

Looking at this plot the result looks very similar to the previous one with the first taxonomy. So we can also expect that the final result i.e. AUPRC is practically the same but we will see later.

toxicity	Predicted value	Actual value
safe	4585	4717
unsafe	272	366*

Table 2. Predicted vs actual safe/unsafe element.

Now, let us see how Prompts classified as unsafe by the model are divided into the different categories of the taxonomy, which we recall are 4: *01. Illegal content, 02. Sexual content, 03. Privacy content, 04. Roleplay and jailbreak content*.

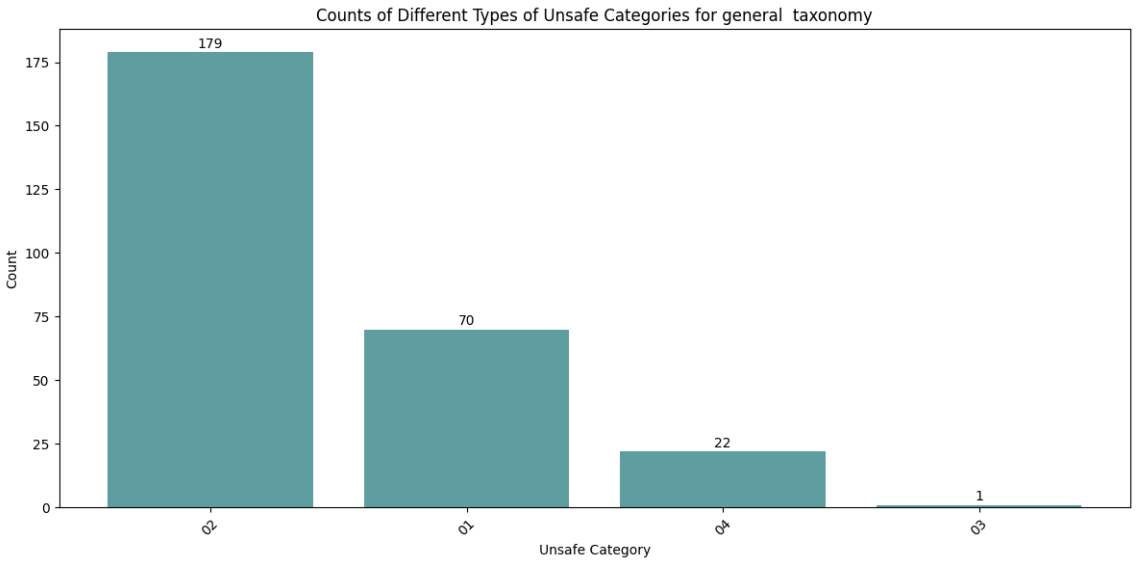

Figure 5. Number of unsafe element per categories.

First of all, we see that the third category i.e. the privacy category was practically useless, probably there are no prompts regarding privacy in the dataset. We can't know because in this case we don't have a dataset column that indicates whether a prompt is privacy-related as we do for jailbreak.

Otherwise more generality in writing categories seems to work, for example many prompts ended up in the first category the one that collects Illegal content and also the category concerning jailbreak i.e. the fourth collects more prompts than in the previous experiment.

In the end, let's see the area under the precision-recall curve as the fundamental metric of this work.

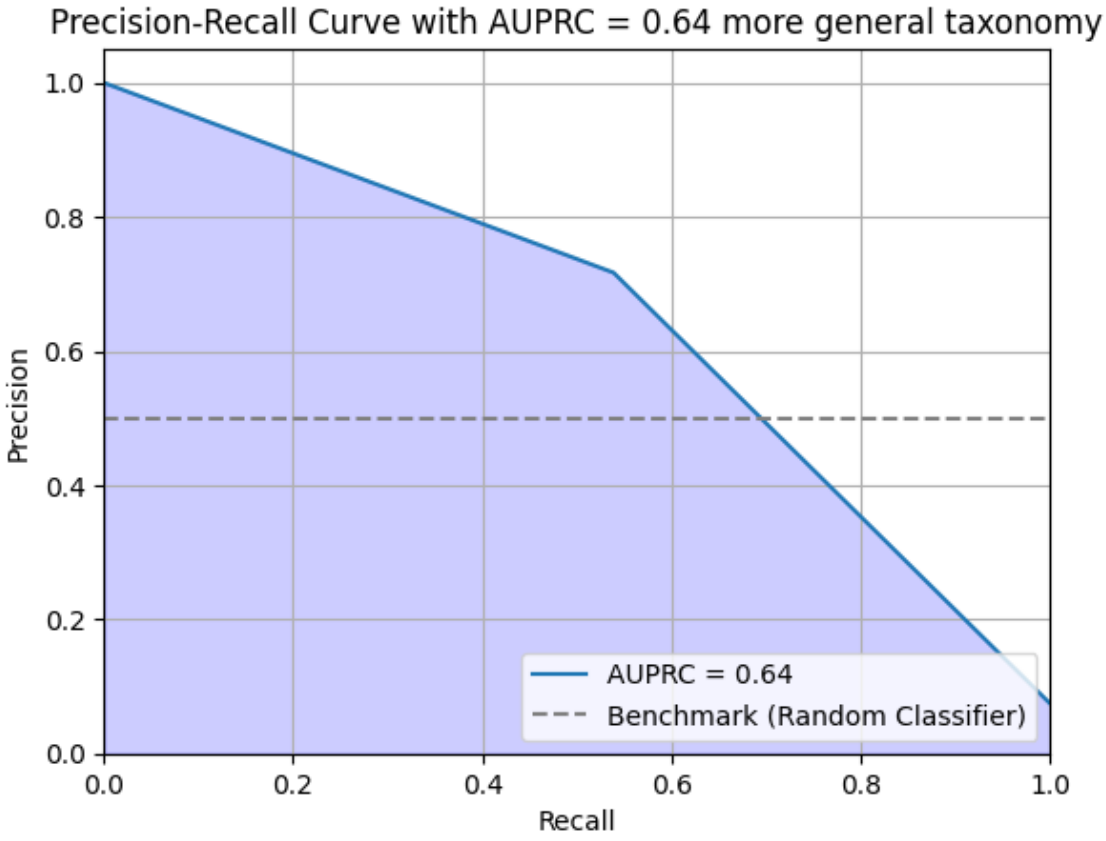

Figure 6.The precision-recall

As we can see, the use of a more general taxonomy brings the model, through fine-tuning, to a very small improvement, unlike what I expected before performing this test, when I thought the result would be worse. In my opinion, therefore, the result is interesting because it shows that Llama guard is able to generalize easily even through "light" fine-tuning (zero-shot prompting) as opposed to the more classical fine-tuning that takes more time and resources and allows easy use of this model while still taking into account its limitations.

5: Conclusion

The time has come to draw the final conclusions.

The issue of large language models safety is a fundamental one, and in recent times it has aroused much controversy and even more or less well-founded fears. Llama guard tries to provide a useful model in this regard, innovative, performing, and adaptable. Unfortunately, as is often the case in the world of LLM models, the explainability [9] of

decisions made by the model is lacking and this in a situation concerning safety is important.

I tried to test in this work, the feature that seemed most interesting to me; the possibility of using the fine tuning method via zero-shot prompting to fit the model to different unsafe categories. By the way, this was something affordable for my limited resources, since usually in the LLM environment there is a need for a large availability of computational capacity.

The results that I was able to obtain in my opinion show how effectively Llama guard is a prompt security checking model (even for other models still) that can easily adapt to new taxonomies, as well as to new datasets as already pointed out in the presentation paper [2]. It certainly does not achieve extraordinary results (at best AUPRC = 0.64) however, one must remember the ease of use which is not an indifferent aspect.

Let's finish by saying that in my opinion, there are still a lot of very interesting aspects to be explored after this project: for example, trying to use Llama guard in real use cases such as perhaps controlling other LLM chat-bots, or trying classical fine-tuning on different taxonomies or few-shot prompting[9].

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., et al. (2017) Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, pp. 5998-6008.

[2] Inan, H., et al., 2023. "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations," arXiv e-prints, arXiv:2312.06674.

[3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T., 2023. "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288.

[4] Y. Li, 2023. "A Practical Survey on Zero-Shot Prompt Design for In-Context Learning," in Proc. of the 14th Int. Conf. on Recent Advances in Natural Language Processing, Varna, Bulgaria, 2023, pp. 641-647.

[5] Reynolds, L., & McDonell, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-7.

[9] Shortly before I started writing this report Meta introduced the new version of the model: Llama guard

You can download the new model here

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, 33, pp. 1877-1901.

[7] Chowdhury, A. G., Islam, M. M., Kumar, V., Shezan, F. H., Jain, V., & Chadha, A. 2024. Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models. In arXiv preprint arXiv:2403.04786.

[8] Saito, T., & Rehmsmeier, M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. In PLoS ONE 10(3): e0118432. https://doi.org/10.1371/journal.pone.0118432.

[9] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... Du, M. 2024. Explainability for large language models: A survey. In ACM Transactions on Intelligent Systems and Technology, 15(2), pp. 1-38.