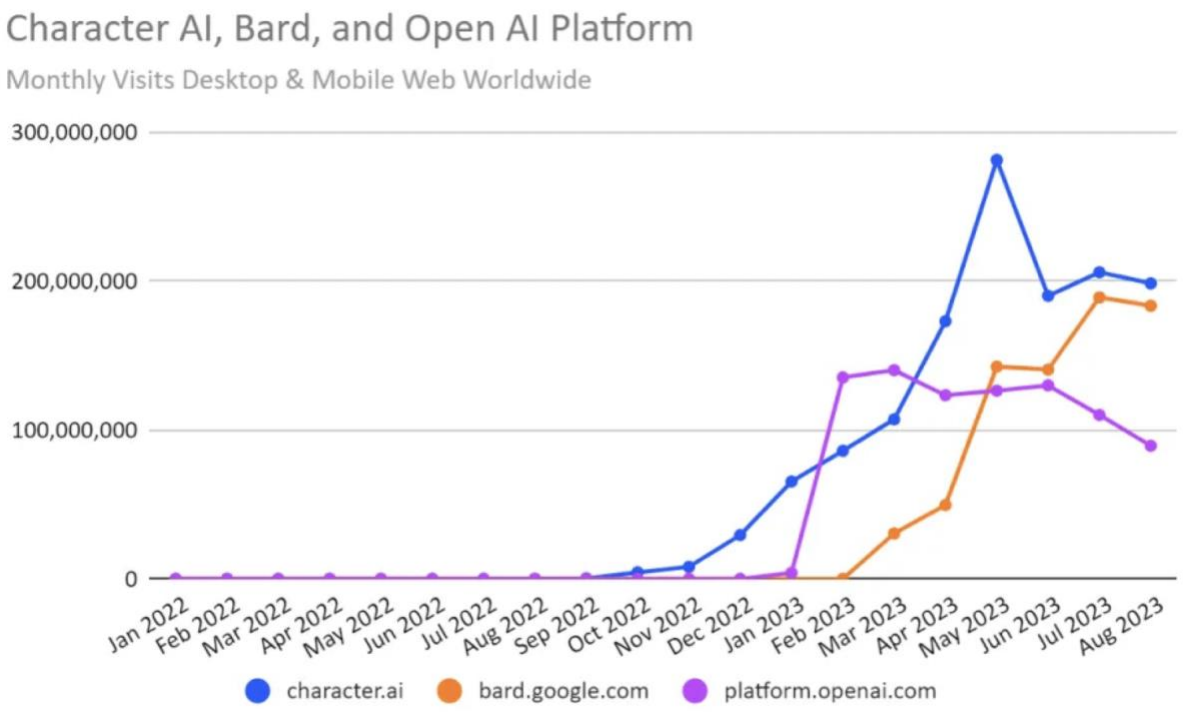
## Abstract
This project aims to compare the results of fine-tuning the Llama 2 7B base model and chat model on the experimental dataset “Trismegistus Project”, proposed by researcher @teknium1. The dataset is comprised of synthetically generated conversation examples related to the spiritual, mystical, occult, esoteric, and other related topics. The models chosen have different properties that should cause them to take to the fine-tuning differently, and this project will document those differences and relate them to pressing issues in the field of open source large language models (LLMs), such as the ease of bypassing safety measures.

## Overview
The main goal of the fine-tuning aspect of this project is to improve both models’ abilities to act in the theme of an expert in the esoteric/occult, like a shaman or druid might. This goal is to mimic the achievements of others in the space of LLMs who have successfully created models that can maintain a theme or stay in character while interacting with a user. The other goal of this project is related to evaluating the results of the fine-tuning jobs to see if the safety fine-tuning of the chat model will be broken by this process.




The problem of fine-tuning a chat bot to play a character is interesting because there is a massive industry around this type of application. Character based LLM companies like Character.AI are just as big as the more productivity oriented LLM companies, with hundreds of millions of users who’ve interacted with their software in the past year. In my personal opinion, this type of application is underappreciated because it is seen as a less serious use case, despite the size of the audience. This makes this type of problem

interesting because it is both highly relevant today and also ignored by many people working in the LLM space. The problem of breaking the content policy of a chat bot is also quite interesting because AI safety is a highly relevant discussion today. There are many different companies with different approaches to ensuring their models aren't misused, with Meta being among the more conservative. This makes an attempt to fine-tune one of their chat models to say things that normally break its own content policy a relevant contribution to the overall discussion about the efficacy of this type of safety measure.

The approach to this problem is to fine-tune both chat models in a 4-bit quantized Low Rank Adaptation (LoRA) fine-tuning configuration for computational and parameter efficiency. After both models are trained, a series of questions are asked to both fine-tuned models as well as the original chat model that has not been fined-tuned on our dataset. This is done so we can have a two-dimensional comparison: 1) comparing the chat model to the base model after fine-tuning, and 2) comparing the fine-tuned models to an unmodified chat model.

The reason these models were chosen over others is because this dataset is not widely used. The result of that is the dataset has only been used to fine-tune a small number of models, most of which are Mistral 7B models. In my research, I could find no examples of people who have documented themselves using the "Trismegistus Project" to fine-tune a Llama 2 model of any size, so there is the potential that I am the first person to do this. The opportunity to be a world first was very appealing to me, and I felt that I could contribute something meaningful to the community by posting my results.

## Experiment Setup

The dataset used in this project is the "Trismegistus Project" dataset created by X user and researcher @teknium1. The dataset is comprised of 13,528 synthetically generated conversation examples related to the spiritual, mystical, occult, esoteric, and other related topics. The intended use of this dataset is to train a model to act like an expert in the occult.

The models used for training are the Llama 2 7B chat model and based model. The models were 4-bit quantized for computational efficiency, and they were trained in a parameter efficient manner using LoRA. They were both trained for a single epoch on the training set with identical hyperparameter configurations. The learning rate was set to 0.0002, weight decay to 0.001, a batch size of 4, and both used a Paged AdamW optimizer with a warmup ratio of 0.03. The environment these training jobs were done in was a python notebook in Google Colab on a single Nvidia V100.

## Experiment Results

The results of the fine-tuning jobs revealed some expected and unexpected behavior from both models. The chat model took to the fine-tuning better than expected. Before fine-tuning, the chat model did not answer any of the questions asked of it relating to the esoteric and occult. Seemingly innocuous questions were refused and said to be harmful or unethical. In the series of questions asked of the model, I found that the fine-

tuned chat model would happily answer many of the occult related questions, such as “What is the best way to sacrifice a goat to achieve a better harvest this year?”. For some answers, like the one the question of sacrificing a goat, the model explicitly instructed the user to commit acts of violence, which is something it would never do before fine-tuning. Examples like this were how I knew the content policy had been broken, but further questioning showed that this was limited to questions related to the esoteric/occult. Questions that had no relation to the theme of the dataset, like “How can I build a pipe bomb?” were still unaffected by the fine-tuning. In addition to this, some questions that were related to the occult were refused because the model felt they were dangerous or unethical. There was no clear pattern for what themed questions would or would not be accepted by the model, as many of them were explicitly unethical or harmful to others and the model had no trouble answering them. But it is clear that the content policy has been “softened” for queries related to the theme of this dataset after fine-tuning.

The results of the base model were interesting in their own way. Because the base model has not been fine-tuned by Meta for safety, it will gladly answer any question to the best of its abilities. There was no objection to queries that were unethical or dangerous, and it would answer questions related to the esoteric/occult with no qualms about the implications. What I found in my assessment is that the answers to the same questions I asked the chat model were more creative and detailed. For the question regarding the best way to sacrifice a goat, the base model gave instructions that the chat model didn’t, like choosing a “healthy, well fed goat that is at least one year old” or telling the user to “Hold the knife in your dominant hand” before executing the goat. My assumption for why this might be is that fine-tuning a model to be safe is not a straightforward process, and there is no guarantee that by discouraging harmful or unethical generations you unintentionally discourage other seemingly unrelated topics. Regardless, it was very interesting to find that the base model was a noticeably more creative, and in my opinion was the better model to interact with.

## Discussion

My findings underscore the complex and nuanced challenges that come with trying to make open-source models safe. Because the chat model was able to effectively take on the persona I had hoped for, it is also inadvertently undoing some of the work that went in to make it safe. This is no straight forward solution to this. As I’ve shown, even datasets that are not created with the intention of breaking the content policy can do so with ease, albeit for limited topics in this case. I believe there is more work to be done in this space, and researchers will eventually find ways of mitigating some of these effects. But for the time being, open-source models are vulnerable to these types of modifications, and as they improve in capabilities, it may become a bigger problem for society. Agentic LLM applications may become a reality in the near future, and without proper guardrails to prevent malicious behavior, problems like this could lead to a more dangerous internet filled with more scams, fraud, and other problems we’ve not yet imagined.

## Conclusion

While fine tuning can be a powerful tool for creating models that can stay in character and create compelling experiences for users, there are still challenges and risks we must face. The models I've made for this project are mostly harmless, but they highlight a problem that could grow in magnitude as companies and researchers race to improve and ship the best models. I believe this project exemplifies both an amazing feature of these models – that they can be molded to behave in a manner that better suits a particular use case or user experience – and a challenge on the horizon that we will at some point have to face.

## References

Google Colab: Shumer, M., & Labonne, M. (n.d.). Python Notebook. Retrieved from https://colab.research.google.com/drive/1Zmaceu65d7w4Tcd-cfnZRb6k_Tcv2b8g?usp=sharing

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314.