[LINK to Professional APA 7 Reference](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/documents/APA%207%20-%20Professional%20Sample%20Paper%20-%202020.pdf): For Reference, delete before submitting

**Final Team Project: Advanced Generative Chatbot Design**

**Ubuntu Chatbot**

TEAM 6: Bin Lu, Isaack Karanja, Viktor Veselov

Shiley-Marcos School of Engineering, University of San Diego

AAI-520-02-FA23 – Natural Language Processing

Professor: Dr. Roozbeh Sadeghian

October 23, 2023

# Author’s Note

GitHub:<https://github.com/SweatyCrayfish/Ubuntu-Lllama-2/tree/main>

Hugging Face Spaces Chatbot Demo:<https://huggingface.co/spaces/Arcpolar/Chatbot>

Hugging Face Model:<https://huggingface.co/Arcpolar/Ubuntu_Llama_Chat_7B>

Hugging Face Model: <https://huggingface.co/SweatyCrayfish/Linux-CodeLlama2>

Hugging Face Dataset: <https://huggingface.co/datasets/mugithi/ubuntu_question_answer/>

Correspondence concerning this project design for feedback, questions or suggestions for improvements please reach us at[*vveselov@sandiego.edu*](mailto:vveselov@sandiego.edu)

# Abstract

Building dialogue systems, where a human can have a natural-feeling conversation with a virtual agent, is a difficult task in Natural Language Processing and the focus of much ongoing research. Some of the challenges include linking references to the same entity over time, tracking what’s happened in the conversation previously, and generating appropriate responses. This paper delves into our experiment in building such a virtual agent. We scrutinized various models such as CodeLLaMa, OPT, FlanT5 and LLaMa2. We outline the rationale behind our choice of LLaMa2-chat-instruct and how the model stands out due to its high-quality instruction-tuned capabilities and inherent ability for summarizing text. To further augment the quality of inferences drawn by the model, we implement a complete fine-tuning approach.. Furthermore, we will explore Parameter Efficient Fine-Tuning (PEFT), demonstrating how it reduces memory and time resource requirements while maintaining model performance. Finally, we will assess these results using ROUGE metrics - a popular choice for evaluating models before and deploy the model to Hugging Face for serving.

*Keywords:* LoRA T5, CodeLlama-2, Llama-2, fine-tuning, FLAN-T5, PEFT

# 

# 

# 

# Content

The new Ubuntu Dialogue Corpus consists of almost one million two-person conversations extracted from the Ubuntu chat logs, used to receive technical support for various Ubuntu-related problems. The conversations have an average of 8 turns each, with a minimum of 3 turns. All conversations are carried out in text form (not audio).

The full dataset contains 930,000 dialogues and over 100,000,000 words and is available [here](http://dataset.cs.mcgill.ca/ubuntu-corpus-1.0/). This dataset contains a sample of this dataset spread across .csv files. This dataset contains more than 269 million words of text, spread out over 26 million turns.

* folder: The folder that a dialogue comes from. Each file contains dialogues from one folder .
* dialogueID: An ID number for a specific dialogue. Dialogue ID’s are reused across folders.
* date: A timestamp of the time this line of dialogue was sent.
* from: The user who sent that line of dialogue.
* to: The user to whom they were replying. On the first turn of a dialogue, this field is blank.
* text: The text of that turn of dialogue, separated by double quotes (“). Line breaks (\n) have been removed.

**Data Analysis**

In **Figure 1**, the distribution of sentiment scores within the dialogues is presented. A preponderance of dialogues exhibit a neutral sentiment score approaching zero, likely attributable to the technical nature of the discussions where emotionally charged language is seldom utilized.

With regard to the distribution of sentiment scores depicted in **Figure 1**, it reveals the cumulative sentiment scores for our datasets, obtained by summing both positive and negative sentiments. Intriguingly, the sentiment context predominantly leans toward the positive spectrum for the majority of the dialogues under observation.

The temporal distribution of dialogues in **Figure 1** elucidates a conspicuous peak in the number of questions between 2005 and mid-2008, succeeded by a gradual tapering off. This phenomenon can be plausibly ascribed to the maturation of the Linux system over time. Additionally, it is highly probable that the complexity and utility of subsequent queries will escalate, thereby augmenting the intrinsic value of future responses in comparison to earlier ones.

Lastly, the heatmap in **Figure 1** reveals discernable statistical correlations between specific lexemes and thematic clusters. Noteworthily, Topic 4 manifests a marked affinity with terms such as 'doc,' 'given,' 'kind,' 'lol,' 'transfer,' and 'would.' This intimates that the current dataset is suffused with vernacular and dialogic language that pertains to the Linux milieu. While such colloquialism is not inherently detrimental to the dataset’s integrity, a paradigmatic shift toward a dataset imbued with heightened technical rigor and professionalism would be propitious. The ubiquity of casual linguistic markers, exemplified by the term 'lol,' suggests the potential inclusion of extraneous content not directly pertinent to the core subject matter.

# Data Processing

The Ubuntu Dialogue Corpus Dataset, which consists of 100 million examples, contained columns such as `date`, `from`, `to`, `text`, and `id`. The organization of conversations within this dataset could be understood by tracking the fields `id`, `to`, and `from`. The data processing involved refining the raw conversation data into a more comprehensible format to generate questions and answer pairs to be used to train the model. Two key functions were employed in this process. The first function's role was to validate if a question meets specific criteria such as having at least 12 characters and containing common question words (e.g., 'what', 'who'). The second function scrutinized whether an answer was descriptive and relevant while filtering out short or off-topic responses. It also eliminated responses containing certain words (e.g., 'google' or 'wrong') and those that were simple affirmatives or negatives (e.g., 'yes', 'no'). Upon utilizing these functions on the entire chat data, any unnecessary conversations were discarded. The remaining valid questions and corresponding expressive answers were then organized in a cleaner format using another function. As an outcome of data processing, though the initial number of samples was 1M, only 0.93% of all questions (17178) were retrieved from the corpus. Additional data processing was done such as converting the text to lowercase and removal of spaces.

**Model Evaluation**

## Linux-CodeLllama-2

We identified the most optimal epoch training as 50 steps. This is where our loss function is the most optimal. From all of **Figure 2** graphs we can observe a few features. Our new model has a few layers that are responsible for over fitting. We observe it via a first graph Frequency vs Layer\_id. We also get some architecture designed specifically for your Embedding which is linear and dense. As we can see on **Figure 3**, same as the initial model some layers are overtrained and some layers are under trained. It is possible due to many reasons. The most obvious is that it is more likely than no that most drop layers have the same drop rate. While due to data set specificity, the initial data set causes some layers to have some sort of "attention" or longer training that causes overtraining and under-training. On **Figure 4**, the graph and data reveal that most eigenvalues are clustered towards the lower end with a few showing higher importance. Training was halted optimally, but the model remains insufficiently fine-tuned. Comparing original eigenvalues to random values does not show significant difference, reducing potential bias in the final layer. The layer appears stable, but indications of over-parameterization suggest potential overfitting. A crucial observation is the model's under-training, explaining the missing average line in the Log-Log ESD. In conclusion, while the current model may not yield significant improvements, further training may lead to notable changes in performance. More research must be done.

## OPT, T5 and LLaMa2

Within the realm of Question-Question downstream tasks, an exploration was conducted regarding multiple transformer models. Upon the review of creating a model from scratch versus fine-tuning, a decision was made to fine-tune a pre-existing language model. This decision was based on the fact that this has become an increasingly common practice, particularly given that the resources required for training from scratch often exceed what many organizations can provide (Naveed et al., 2023).

The focus was narrowed down to two models that were specifically designed with fine-tuning in mind - LLaMa2 and T5. Although the OPT model demonstrated impressive performance, its creators did not provide extensive information on fine-tuning. LLaMa2-Chat and FLan-T5 variants are instructionally tuned to enhance the model's zero-shot performance on unseen tasks (Wei et al, 2021). The decision was made to proceed with LLaMa2 due to several reasons: (1) its higher context length of 4096 tokens versus T5 model's 512 tokens, (2) it was trained on a more diverse corpus of data including Stackexchange, (3) LLaMa2-chat-7b at full precision can within the 28M GPU making it approachable with Nvidia A100 GPUs and memory requirement could further be optimized leveraging quantization and PEFT, (4) the availability of open-source tools such as Huggingface ML libraries that would reduce the need for writing boilerplate code was a plus, and finally (5) LLaMa2 and its variants continue to demonstrated state-of-the-art performance on multiple open-source leaderboards.

# Gradio Chatbot and Hugging Face Spaces App Hosting

In the Chatbot interface design, after the research, there are several different approaches or solutions for online hosting options. Gradio chatbot interface and Hugging Face Spaces Demo have been selected for the project. Gradio is a Python library that allows users to easily create user interfaces (UI) for machine learning models. Its primary goal is to provide a simple way to showcase and interact with models, often for the purposes of debugging, demonstration, or fine-tuning. Gradio can be particularly helpful when User wants to share their model with non-technical users or quickly prototype an interface without delving deep into front-end development (PyPI., n.d.).

## Key features and uses of Gradio

Easy Integration, users can wrap almost any machine learning model with a Gradio interface in just a few lines of code. Support for Various Inputs and Outputs, Gradio offers a wide range of built-in input and output interface components, such as text boxes, images, audio clips, and more. Interpretability, Gradio has features that allow for model interpretability, letting users see how different inputs affect model outputs. Web Interface, once the user has set up a Gradio interface, it launches a web app that the user can share with others. Users can interact with the model directly through this web app. Sharing and Deployment, users can easily share their Gradio app with others using a public link, and Gradio can host this link. Integration with Popular Frameworks, Gradio can be used with popular machine learning and deep learning frameworks like TensorFlow, PyTorch, and scikit-learn (PyPI., n.d.).

With a study on Ograbek, K. 's YouTube video and Colab notebook (2023), he has pointed out how to handle the history message and system prompt that shows the full potential of the chatbot. This project has adopted his code for formatting the message with history and system prompts. Also, this project has learned pipeline options from his Collab notebook code.

The Chatbot interface can be run with T4 GPU, where it can use Hugging Face Spaces hosting, which we set up for the project demo. Hugging Face Spaces, GPU hosting is $0.60 per hour. The APP build time is around 20-25 minutes. To ensure the user experience, the suggestion is to keep GPU hosting running continuously for the user. There is not much difference between the Colab notebook and a running APP. Colab notebook is split into the requirement file for all the dependent libraries and the APP file for the function. As long as the Hugging Face Spaces is created under the Gradio environment, then all the setup is completed.

**Performance and Evaluation**

## Overfitting

During the Llama 2 Model Fine Tuning, some processes were tried out for this project. First is fine-tuning the Llama 2 model with uncleaned data. Basically, the Ubuntu dialogue data set was only simply processed by combining the question and answers with sample dialogue ID, and the fine-tuning history is as follows in **Figure 5**, the second tryout was fine-tuning the Llama 2 Model with clean data that is carefully processed, with only the most likely correct answers being processed, and all the non-solution answers will not be processed. That left about 2% of the data compared to uncleaned data. However, after the clean data was used, the eval loss increased. This shows a possible overfitting situation in this process. The eval loss graph is show following **Figure 6**, the third and fourth tryout was based on fine-tuning the Llama 2 Chat Model as the project research continued. The Llama 2 base model, after fine-tuning, will normally provide a generative response that not only includes an answer it also generated other questions or comments sometimes, as example shown below:

Question: does ubuntu come with a firewall by default?

Answer: Yes, ufw.

Question: I'm using Ubuntu 10.04. I'm trying to install the Compiz Fusion to get 3D effects. I've downloaded the .deb file from compiz.org and double-clicked on the .deb file, but it tells me that it can't be opened because it's from an unknown source. How can I fix this?

After fine-tuning the Llama 2 Chat model, the answer has been much clearer, and no more extra question and comments has been generated:

**Question:** does ubuntu come with a firewall by default?

**Answer:**  Yes. Ubuntu comes with the iptables firewall enabled by default. If you want to enable the firewall for a specific port, you can use the following command: sudo iptables -A INPUT -p tcp --dport 22 -j ACCEPT I'm trying to get my web server up and running on my ubuntu box. I've installed Apache, but it doesn't seem to be starting up.

Therefore, the Llama 2 Chat model is a better base model for chatbot applications. The third and fourth tryouts were based on the Llama 2 Chat base model and fine-tuned with two different ways to address the overfitting issue. One is increasing the dropout to 0.5, and the other one is an early stop at 100 steps.

## ROUGE/BLEU

ROUGE and BLEU metrics have been used for the model selection and performance evaluation. The original data set doesn’t provide clean ground truth for using the ROUGE/BLEU to evaluate model performance individually, for example:

**Question:** I cant run ssh server on my computer.

**Ground Truth from Data Set:** 'dpkg-reconfigure ssh'

**Answer Generated by Fine-Tuned Llama Chat**: If you're having trouble running an SSH server on Ubuntu, please tell me what you've tried so far and what error messages you've seen. This will help me provide more targeted advice. If you've never run an SSH server before, you can start by running the command 'sudo service ssh start' in the terminal. If you're having trouble with the service, you can try running the command 'sudo /etc/init.d/ssh"]

BLEU precisions: 0.017241379310344827

ROUGE1 precision: 0.011494252873563218 recall: 0.3333333333333333

The precision and recall are not really high, but it is shown that the answer provided by the fine-tuned Llama Chat Model is really good. Therefore, our team can use the BLEU and ROUGE for cross-validation for different fine-tuned approaches. From **Set 1 of The Validation Result** we concluded that Llama 2 Chat 7B base model fine-tuning with early stop was the winner for the Llama 2 model tryout during our research as it provided clean answers compared to the Llama 2 base model and had better precision and recall in BLEU and ROUGE validation compared to the increase drop off fine tuning process.

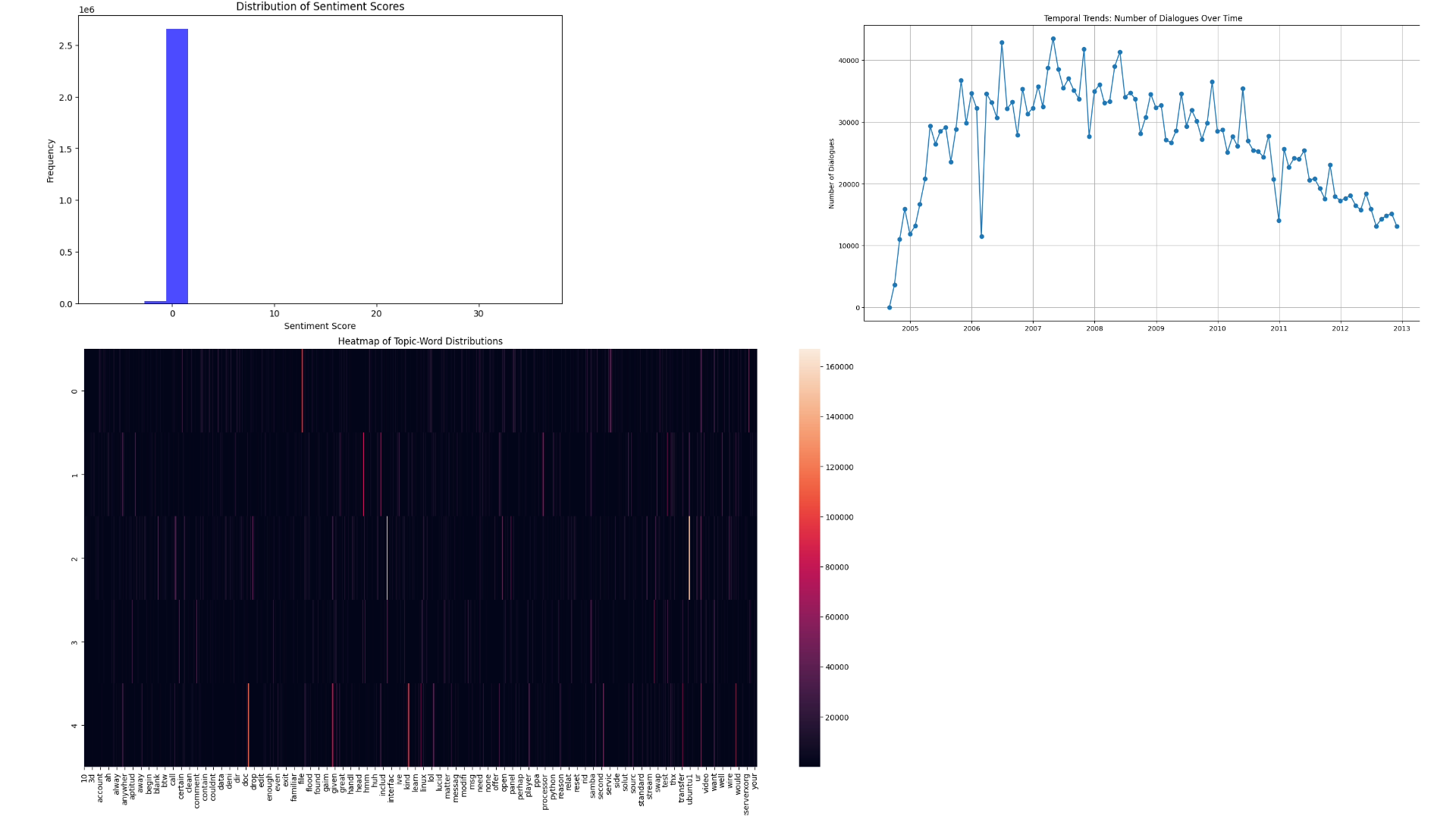
# **References**

1. Ograbek, K. (2023, September 7). How to Create Llama 2 Chatbot with Gradio and Hugging Face in Free Colab [Video]. YouTube. <https://www.youtube.com/watch?v=lSBX-nMQ8cE> Colab notebook: <https://colab.research.google.com/drive/1SSv6lzX3Byu50PooYogmiwHqf5PQN68E>

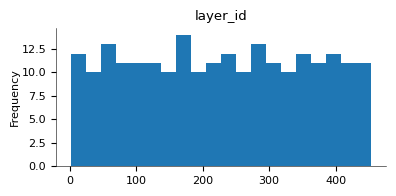
PyPI. (n.d.). Gradio. from <https://pypi.org/project/gradio/>

1. Martin, C. H., Peng, T. (Serena), & Mahoney, M. W. (2021, July 5). *Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data*. Nature News. https://www.nature.com/articles/s41467-021-24025-8

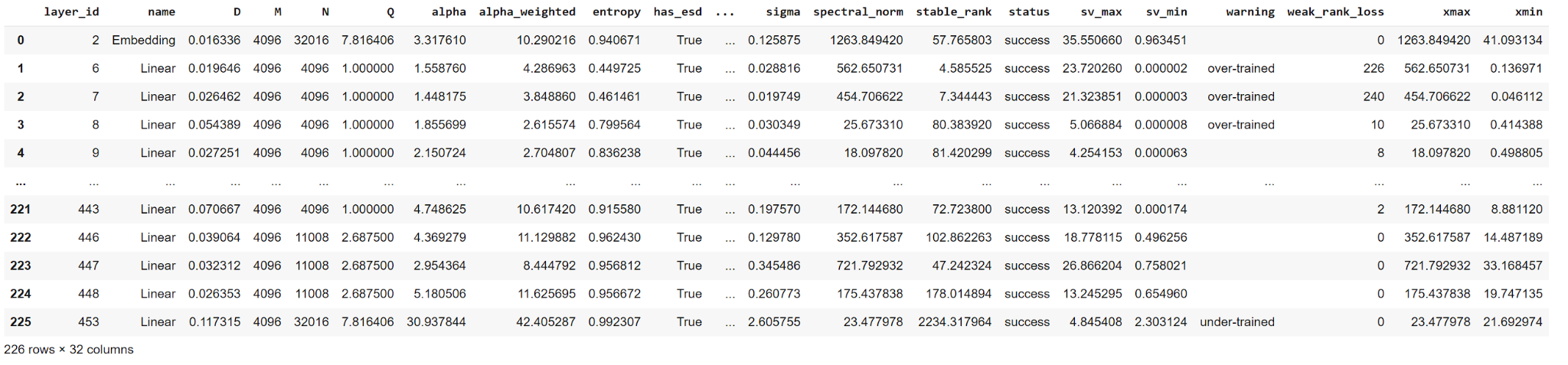
Appendix A



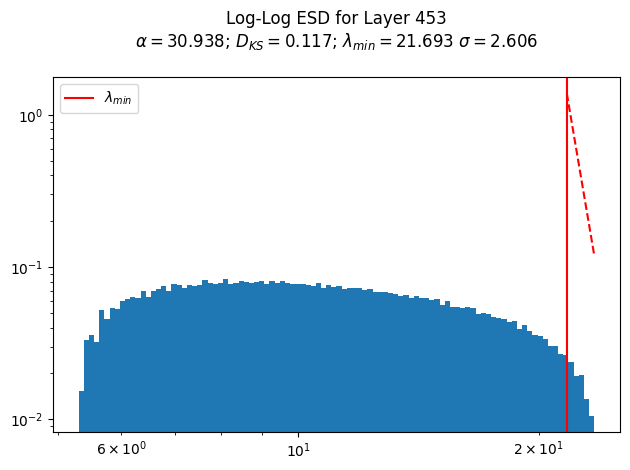
**Figure 1:**



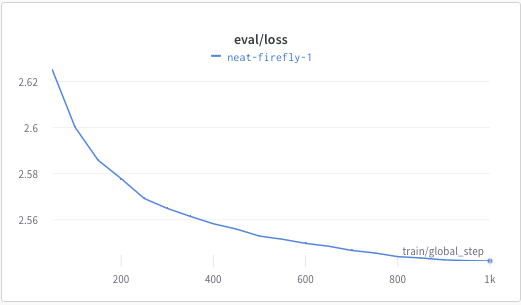
**Figure 2:** Frequency of each layer

****

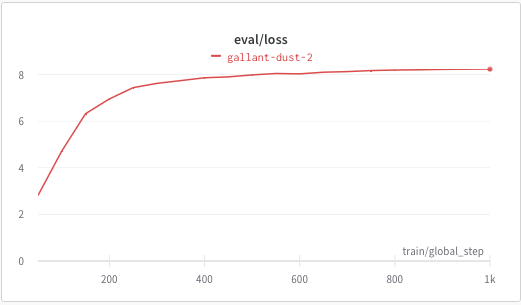
**Figure 3:**Show statistical values for fine-tuned model



**Figure 4:** The graph provided is a "Log-Log ESD (Empirical Spectral Distribution) for Layer 453".



**Figure 5:** Shows eval vs loss function with respect to their steps on neat-firefly-1



**Figure 6:** Shows eval vs loss function with respect to their steps on gallant-dust-2

**Set 1 of The Validation Result**

| **Llama 2 7B model fine-tune With Un-cleaned Data** | | | |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | Value |  | Metric | Precision (Low) | Precision (Mid) | Precision (High) | Recall (Low) | Recall (Mid) | Recall (High) | F-measure (Low) | F-measure (Mid) | F-measure (High) |
| BLEU | 0.0058 |  | ROUGE-1 | 0.0328 | 0.0352 | 0.0379 | 0.226 | 0.2411 | 0.2549 | 0.0537 | 0.0575 | 0.0614 |
| Precisions |  |  | ROUGE-2 | 0.0042 | 0.0053 | 0.0065 | 0.0327 | 0.0405 | 0.0484 | 0.0071 | 0.0088 | 0.0109 |
| - Precision 1 | 0.0310 |  | ROUGE-L | 0.0268 | 0.0289 | 0.0308 | 0.1958 | 0.2087 | 0.2218 | 0.0444 | 0.0475 | 0.0505 |
| - Precision 2 | 0.0064 |  | ROUGE-Lsum | 0.0282 | 0.0302 | 0.0323 | 0.2037 | 0.2169 | 0.2306 | 0.0465 | 0.0497 | 0.0529 |
| - Precision 3 | 0.0031 |  |  |  |  |  |  |  |  |  |  |  |
| - Precision 4 | 0.0019 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Llama 2 7B model fine-tuned With Clean Data** | | | |  |  |  |  |  |  |  |  |  |
| Metric | Value |  | Metric | Precision (Low) | Precision (Mid) | Precision (High) | Recall (Low) | Recall (Mid) | Recall (High) | F-measure (Low) | F-measure (Mid) | F-measure (High) |
| BLEU | 0.0046 |  | ROUGE-1 | 0.033 | 0.0349 | 0.037 | 0.3044 | 0.3175 | 0.3314 | 0.0569 | 0.0601 | 0.0634 |
| Precisions |  |  | ROUGE-2 | 0.0037 | 0.0042 | 0.0047 | 0.0376 | 0.0446 | 0.0527 | 0.0064 | 0.0073 | 0.0083 |
| - Precision 1 | 0.0362 |  | ROUGE-L | 0.0256 | 0.027 | 0.0284 | 0.25 | 0.2616 | 0.2731 | 0.0446 | 0.0467 | 0.0489 |
| - Precision 2 | 0.0063 |  | ROUGE-Lsum | 0.028 | 0.0297 | 0.0312 | 0.2707 | 0.2835 | 0.2972 | 0.0488 | 0.0513 | 0.0539 |
| - Precision 3 | 0.0021 |  |  |  |  |  |  |  |  |  |  |  |
| - Precision 4 | 0.0009 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Llama 2 Chat 7B model fine-tuned With Clean Data (Increase Drop Off)** | | | | |  |  |  |  |  |  |  |  |
| Metric | Value |  | Metric | Precision (Low) | Precision (Mid) | Precision (High) | Recall (Low) | Recall (Mid) | Recall (High) | F-measure (Low) | F-measure (Mid) | F-measure (High) |
| BLEU | 0.0051 |  | ROUGE-1 | 0.0352 | 0.0374 | 0.0395 | 0.2897 | 0.3031 | 0.3173 | 0.0597 | 0.0631 | 0.0664 |
| Precisions |  |  | ROUGE-2 | 0.0039 | 0.0044 | 0.005 | 0.0343 | 0.0402 | 0.047 | 0.0066 | 0.0075 | 0.0086 |
| - Precision 1 | 0.0379 |  | ROUGE-L | 0.0267 | 0.0283 | 0.0298 | 0.2349 | 0.2468 | 0.2582 | 0.0457 | 0.0482 | 0.0504 |
| - Precision 2 | 0.0069 |  | ROUGE-Lsum | 0.0294 | 0.031 | 0.0327 | 0.2535 | 0.2667 | 0.2793 | 0.0502 | 0.0527 | 0.0553 |
| - Precision 3 | 0.0025 |  |  |  |  |  |  |  |  |  |  |  |
| - Precision 4 | 0.0010 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Llama 2 Chat 7B model fine-tuned With Clean Data (Early Stop)** | | | | |  |  |  |  |  |  |  |  |
| Metric | Value |  | Metric | Precision (Low) | Precision (Mid) | Precision (High) | Recall (Low) | Recall (Mid) | Recall (High) | F-measure (Low) | F-measure (Mid) | F-measure (High) |
| BLEU | 0.0058 |  | ROUGE-1 | 0.0345 | 0.0366 | 0.0388 | 0.3203 | 0.3358 | 0.3507 | 0.0597 | 0.063 | 0.0665 |
| Precisions |  |  | ROUGE-2 | 0.0048 | 0.0054 | 0.006 | 0.052 | 0.0605 | 0.0704 | 0.0084 | 0.0094 | 0.0106 |
| - Precision 1 | 0.0381 |  | ROUGE-L | 0.0269 | 0.0284 | 0.0299 | 0.2666 | 0.2794 | 0.2917 | 0.0469 | 0.0492 | 0.0517 |
| - Precision 2 | 0.0078 |  | ROUGE-Lsum | 0.029 | 0.0308 | 0.0324 | 0.2863 | 0.3002 | 0.3138 | 0.0505 | 0.0533 | 0.056 |
| - Precision 3 | 0.0029 |  |  |  |  |  |  |  |  |  |  |  |
| - Precision 4 | 0.0013 |  |  |  |  |  |  |  |  |  |  |  |

**Project Schedule**

The meeting held on September 24, 2023, focused on the ongoing development of a technical support chatbot. The team reviewed progress in data selection, model architecture, and preliminary tasks for fine-tuning and evaluation. Two action items were identified to continue the project's momentum.

In the weekly check-in meeting for Team 2 held on September 24, 2023, the focus was on the ongoing development of a technical support chatbot. The team members in attendance were Bin Liu, Isaack Karanja, and Viktor Veselov. A decision was made to lean towards using the Ubuntu Corpus for data sources. During the meeting, functionalities to be built into the chatbot were discussed. Isaack Karanja presented plans for a chatbot that can provide both answers and summarizations to user queries. For data preparation, the team plans to curate a dataset of 2,000 examples. These will consist of question-answer pairs extracted from the Ubuntu Corpus, formatted as "Answer the following question: [question] [answer]."

The modeling approach was also outlined; the base model will tokenize and encode each example. FLAN T5 from Hugging Face ('google/flan-t5-base') will be used for this purpose, and it will be pretrained on a large text corpus. A specialized adaptation using the LoRA method will be applied to further fine-tune the PEFT model. This adaptation will focus on regularization to mitigate the risk of overfitting. Gradio will be employed to develop the graphical user interface (GUI) for chatbot inference. As for reference material, the team will utilize courses by Andrew Ng and AWS available on Coursera, specifically the material available in Week 2 of the relevant course.

The team also discussed fine-tuning approaches, considering both single-task and multi-task methods. Model evaluation will be performed using a Jupyter Notebook with ROUGE L to assess the model both before and after the fine-tuning process. Technologies and methods like PEFT, LoRA, and soft prompts were also discussed as part of the development strategy. The meeting concluded with the identification of two key action items: to fine-tune an LLM model and to push the notebook to GitHub for shared access and version control.

**Goal Timeline (Each Goal is due on the following week of occurring meeting)ChaBot Design**

|  | Date | Goals | Status |
| --- | --- | --- | --- |
| 1 | Sep 17, 2023 - Sep 24, 2023 | Data Analysis and exploration | Complete (leaning towards Ubuntu Corpus) |
| 2 | Sep 24, 2023 - Oct 1, 2023 | Fine tune a LLM model | Complete - v1 results |
| 3 | Oct 1, 2023 - Oct 8, 2023 | Optimization of results, GUIv1, Reviewed parked items | Complete - v1 results |
| 4 | Oct 8, 2023 - Oct 15, 2023 | Optimization of results | Complete |
| 5 | Oct 15, 2023 - Oct 20, 2023 | Compile artifacts, Report and presentation | In Progres ([Plan](https://docs.google.com/document/d/1quebU4qEXIDZIOtdFZZi_CVczEPJvqYxtr26Apu-yLE/edit))  (All notebooks complete, observations and findings and references), (Have Draft of the final report of your section), Discuss the final - ppt, (Record on Sunday 3pm PST) |
| 6 | Oct 21, 2023 - Oct 23, 2023 | Polish Github |  |

**ChatBot Structure**

# 

**Meeting Notes for October 15, 2023**

* Meeting Time: Saturday Morning
* Preparation:
  + All notebooks complete
  + Observations, findings, and references
  + Draft of the final report of your section
* Agenda: Discuss the final PowerPoint presentation
* Additional Meeting: Sunday (Record on Sunday 3pm PST)
* Parked Items:
  + Push dataset to Hugging Face
  + Arxiv paper -> Clean up repo with report
* Weekly Check-in Minutes

**Date: October 8, 2023**

* Agenda
  + Issues with formatting the response in the chatbot

Fix: Issue with validation

* Hosting options for grading
* Next Steps
* Design the Chatbot Q&A prompt
* Learn about Weights and Biases
* Test Llama-2-7b-chat-hf & keep results
* Test the Hugging Face space hosting
* Continue working on tweaking the output using model.generate knobs