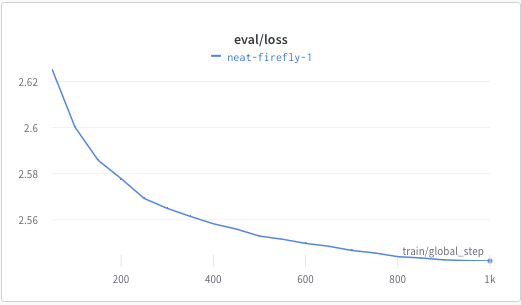
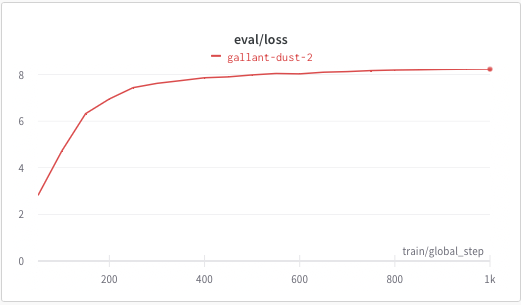
Performance and Evaluation

Overfitting

During the Llama 2 Model Fine Turning, 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:



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:



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. The validation result as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |  |  |  |  |  |  |  |  |  |  |  |

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 finetuning process.