Team 6

Bin Lu

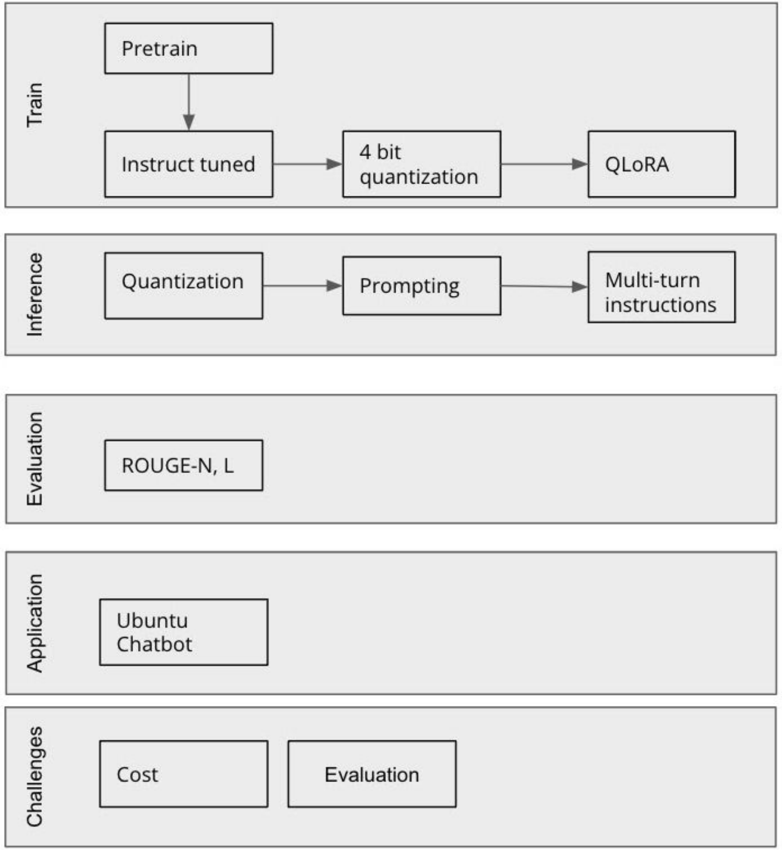
Viktor Veselov

Isaack Karanja

Advanced Generative Chatbot Design

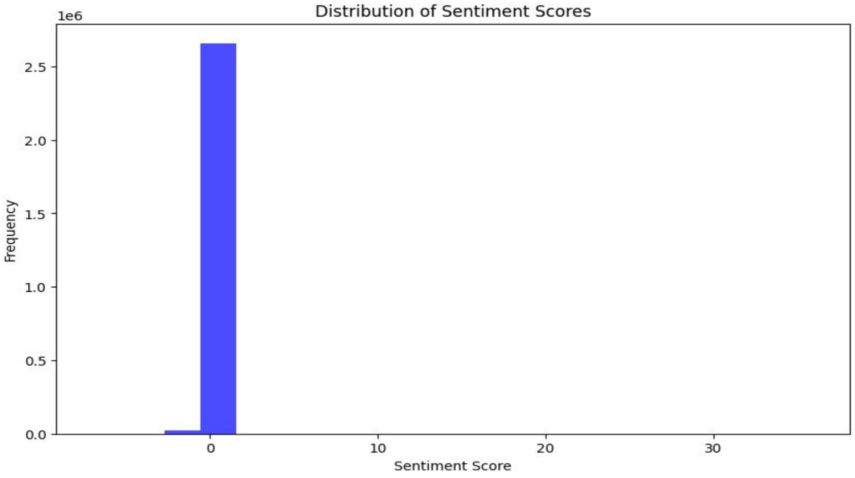
Objective

FineTune LLaMa2 a foundational instruct model using various techniques and use it to build a Chatbot

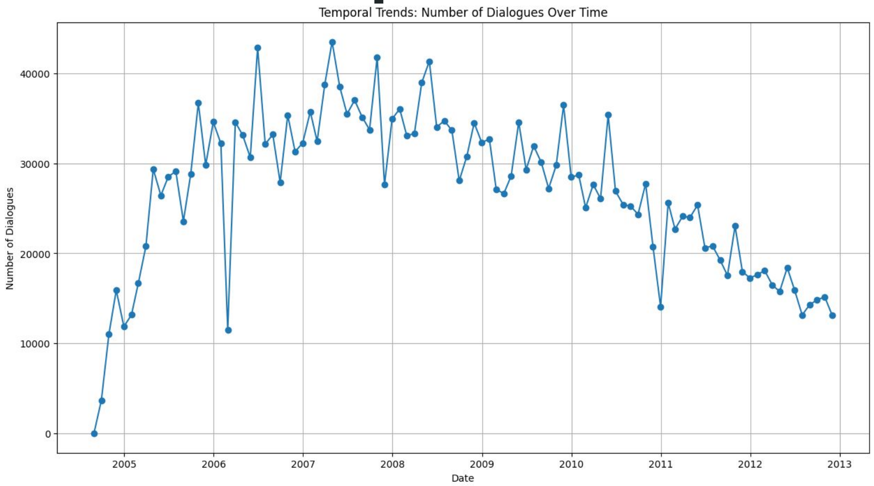


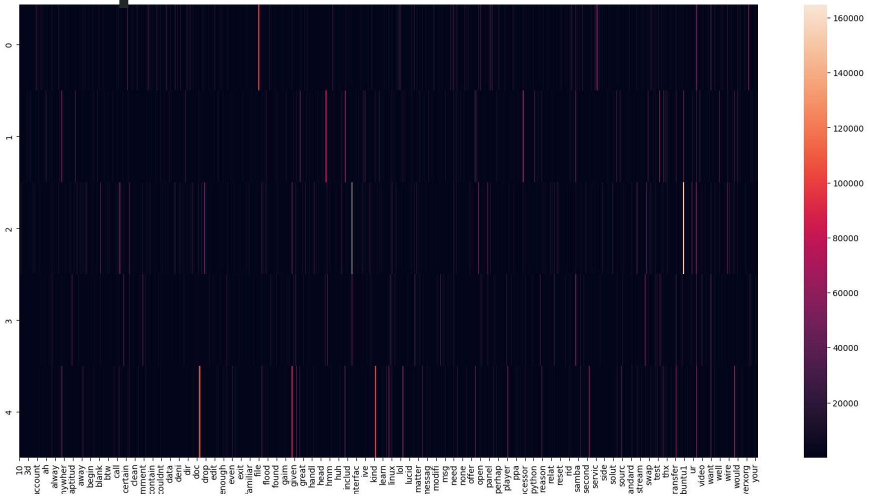
Distribution of Sentiment Scores

- Majority of dialogues center around a neutral sentiment, as represented by scores close to zero.
- Technical discussions predominate, leading to less emotionally charged language.
- Minimal spread indicates consistent sentiment across the dataset.



Temporal Trends and Topic Distributions





Dialogues Over Time & Lexical Insights

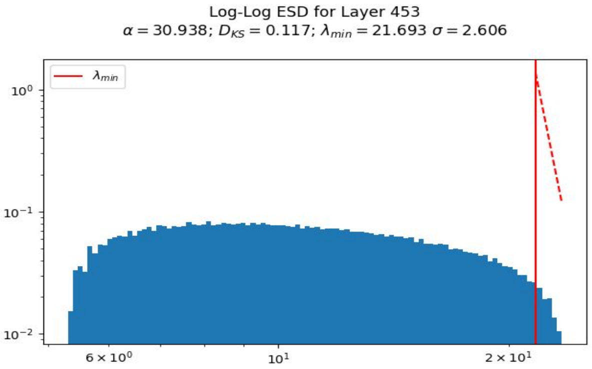
Temporal Trends:

- Notable surge in dialogue numbers between 2005 and mid-2008.
- Subsequent decline post-2008, aligning with the maturation of the Linux system.
- Increasing complexity and utility anticipated in subsequent queries.

Topic-Word Distributions:

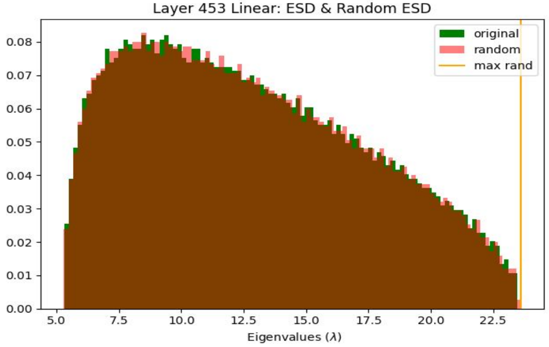
- Heatmap showcases relationships between specific lexemes and thematic clusters.
- Topic 4 prominently associates with terms like 'doc,' 'given,' and 'lol,' indicating colloquial language use.
- Presence of casual terms suggests extraneous content, hinting at potential for refining dataset content.

Linux-CodeLlama-2 Evaluation

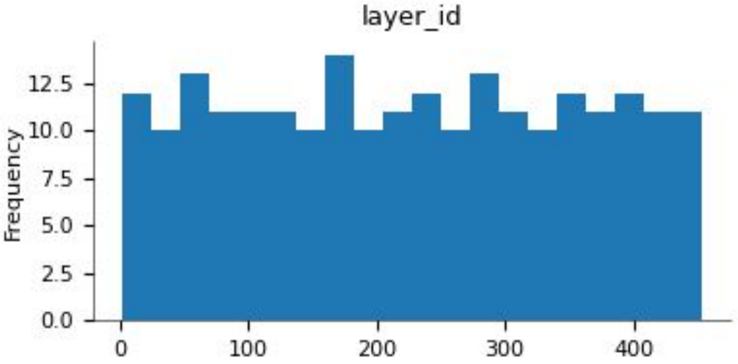


Log-Log ESD for Layer 453

$\alpha = 30.938;\ D_{KS} = 0.117;\ \lambda_{min} = 21.693\ \sigma = 2.606$

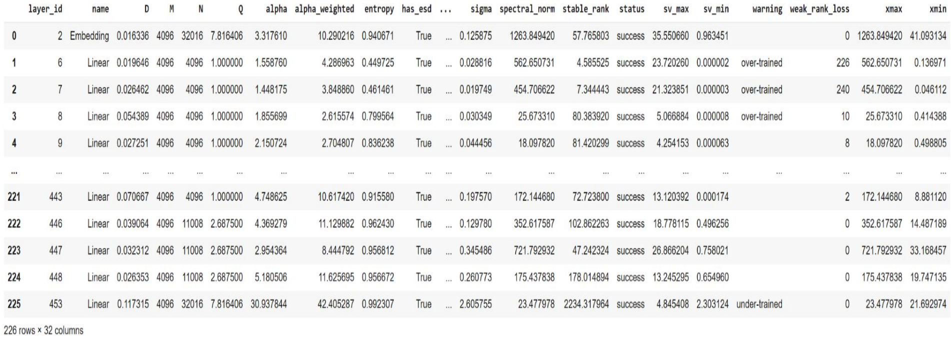


Layer 453 Linear: ESD & Random ESD



layer_id

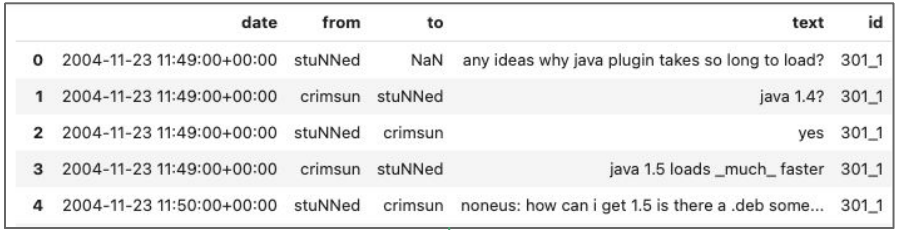
- Majority of eigenvalues cluster at the lower end: minimal feature contribution.
- Few larger eigenvalues: significant feature importance.
- λmin value suggests layer stability.
- Distribution shape may hint at over-parameterization and risk of overfitting.
- Dks value represents fit to a reference: smaller value = better fit.



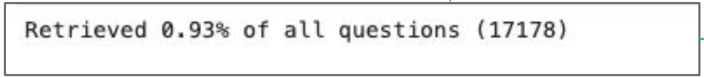
	layer_id	name	D	M	N	Q	alpha	alpha_weighted	entropy	has_esd	...	sigma	spectral_norm	stable_rank	status	sv_max	sv_min	warning	weak_rank_loss	xmax	xmin
0	2	Embedding	0.016336	4096	32016	7.816406	3.317610	10.290216	0.940671	True	...	0.125875	1263.849420	57.765803	success	35.550660	0.963451		0	1263.849420	41.093134
1	6	Linear	0.019646	4096	4096	1.000000	1.558760	4.286963	0.449725	True	...	0.028816	562.650731	4.585525	success	23.720260	0.000002	over-trained	226	562.650731	0.136971
2	7	Linear	0.026462	4096	4096	1.000000	1.448175	3.848860	0.461461	True	...	0.019749	454.706622	7.344443	success	21.323851	0.000003	over-trained	240	454.706622	0.046112
3	8	Linear	0.054389	4096	4096	1.000000	1.855699	2.615574	0.799564	True	...	0.030349	25.673310	80.383920	success	5.066884	0.000008	over-trained	10	25.673310	0.414388
4	9	Linear	0.027251	4096	4096	1.000000	2.150724	2.704807	0.836238	True	...	0.044456	18.097820	81.420299	success	4.254153	0.000063		8	18.097820	0.498805
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
221	443	Linear	0.070667	4096	4096	1.000000	4.748625	10.617420	0.915580	True	...	0.197570	172.144680	72.723800	success	13.120392	0.000174		2	172.144680	8.881120
222	446	Linear	0.039064	4096	11008	2.687500	4.369279	11.129882	0.962430	True	...	0.129780	352.617587	102.862263	success	18.778115	0.496256		0	352.617587	14.487189
223	447	Linear	0.032312	4096	11008	2.687500	2.954364	8.444792	0.956812	True	...	0.345486	721.792932	47.242324	success	26.866204	0.758021		0	721.792932	33.168457
224	448	Linear	0.026353	4096	11008	2.687500	5.180506	11.625695	0.956672	True	...	0.260773	175.437838	178.014894	success	13.245295	0.654960		0	175.437838	19.747135
225	453	Linear	0.117315	4096	32016	7.816406	30.937844	42.405287	0.992307	True	...	2.605755	23.477978	2234.317964	success	4.845408	2.303124	under-trained	0	23.477978	21.692974

226 rows × 32 columns

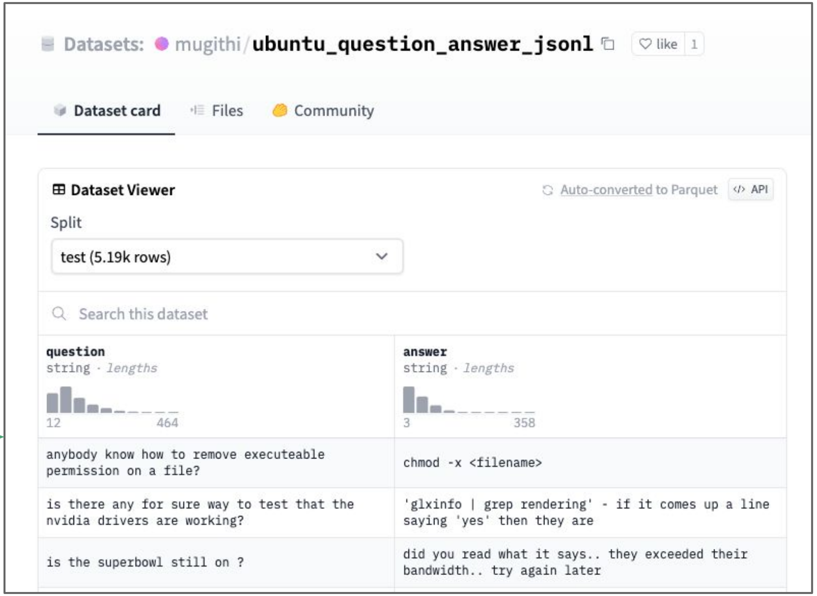
Dataset : The Ubuntu DataSet Corpus



	date	from	to	text	id
0	2004-11-23 11:49:00+00:00	stuNNed	NaN	any ideas why java plugin takes so long to load?	301_1
1	2004-11-23 11:49:00+00:00	crimsun	stuNNed	java 1.4?	301_1
2	2004-11-23 11:49:00+00:00	stuNNed	crimsun	yes	301_1
3	2004-11-23 11:49:00+00:00	crimsun	stuNNed	java 1.5 loads _much_ faster	301_1
4	2004-11-23 11:50:00+00:00	stuNNed	crimsun	noneus: how can i get 1.5 is there a .deb some...	301_1



```
Retrieved 0.93% of all questions (17178)
```



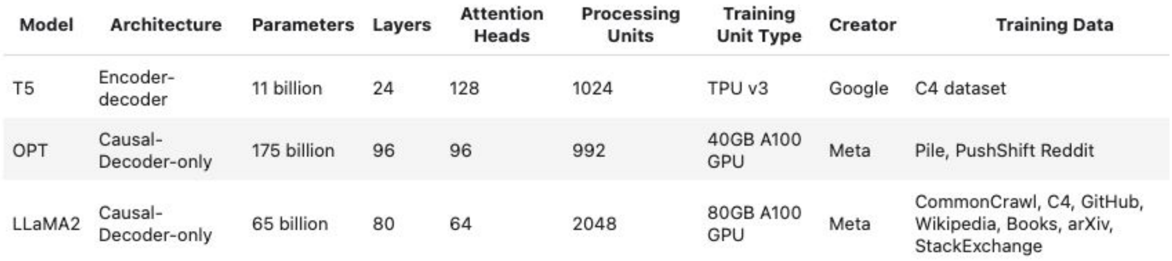
Datasets: mugithi/ubuntu_question_answer_jsonl
like 1
Dataset card
Files
Community
Dataset Viewer
Auto-converted to Parquet
API
Search this dataset

Split

test (5.19k rows)

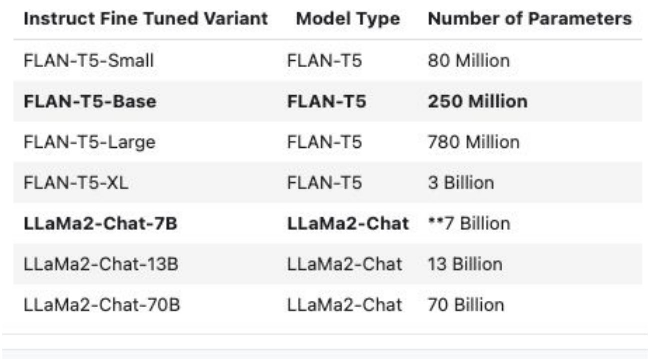
question string · *lengths* 12 464	answer string · *lengths* 3 358
anybody know how to remove executeable permission on a file?	chmod -x <filename>
is there any for sure way to test that the nvidia drivers are working?	'glxinfo | grep rendering' - if it comes up a line saying 'yes' then they are
is the superbowl still on ?	did you read what it says.. they exceeded their bandwidth.. try again later

Model Selection



Model	Architecture	Parameters	Layers	Attention Heads	Processing Units	Training Unit Type	Creator	Training Data
T5	Encoder-decoder	11 billion	24	128	1024	TPU v3	Google	C4 dataset
OPT	Causal-Decoder-only	175 billion	96	96	992	40GB A100 GPU	Meta	Pile, PushShift Reddit
LLaMA2	Causal-Decoder-only	65 billion	80	64	2048	80GB A100 GPU	Meta	CommonCrawl, C4, GitHub, Wikipedia, Books, arXiv, StackExchange

Instruct Tuned vs Base models

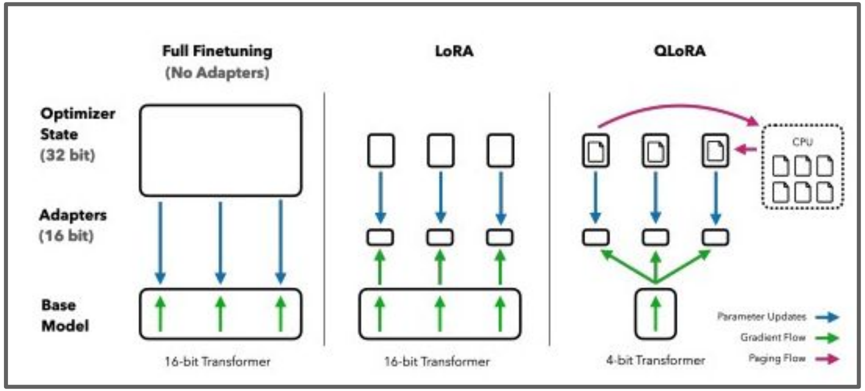


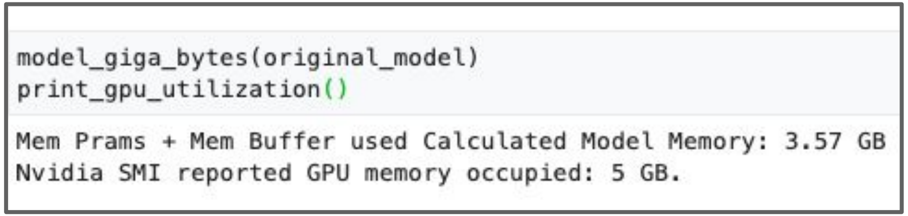
Instruct Fine Tuned Variant	Model Type	Number of Parameters
FLAN-T5-Small	FLAN-T5	80 Million
FLAN-T5-Base	**FLAN-T5**	**250 Million**
FLAN-T5-Large	FLAN-T5	780 Million
FLAN-T5-XL	FLAN-T5	3 Billion
LLaMa2-Chat-7B	**LLaMa2-Chat**	**7 Billion
LLaMa2-Chat-13B	LLaMa2-Chat	13 Billion
LLaMa2-Chat-70B	LLaMa2-Chat	70 Billion

Observations

- State of the art performance
- Higher context length of 4096 tokens vs T5 model 512
- Designed with fine tuning in mind as opposed to OPT
- Small 7B instruct-Tuned model that demonstrated turn conversations

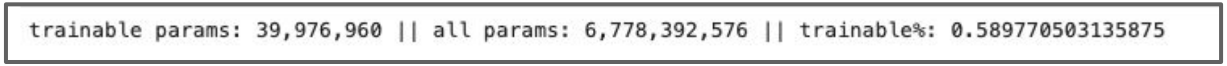
Training | LoRA, QLoRA and Memory Requirements



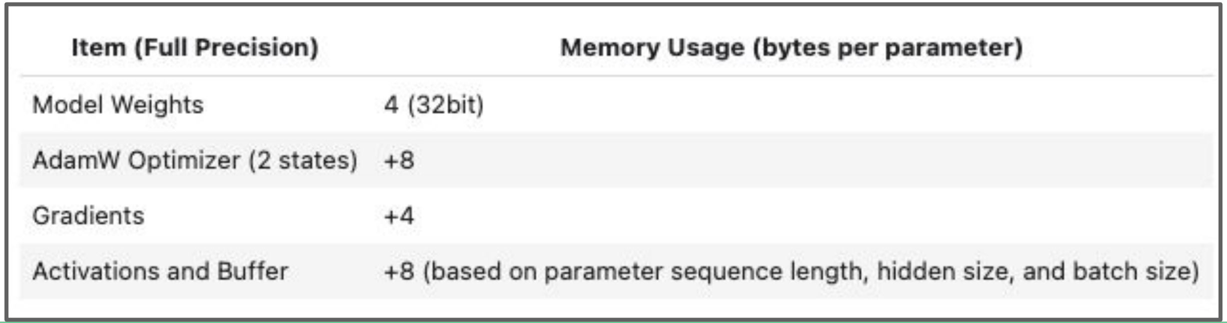


```
model_giga_bytes(original_model)
print_gpu_utilization()
```

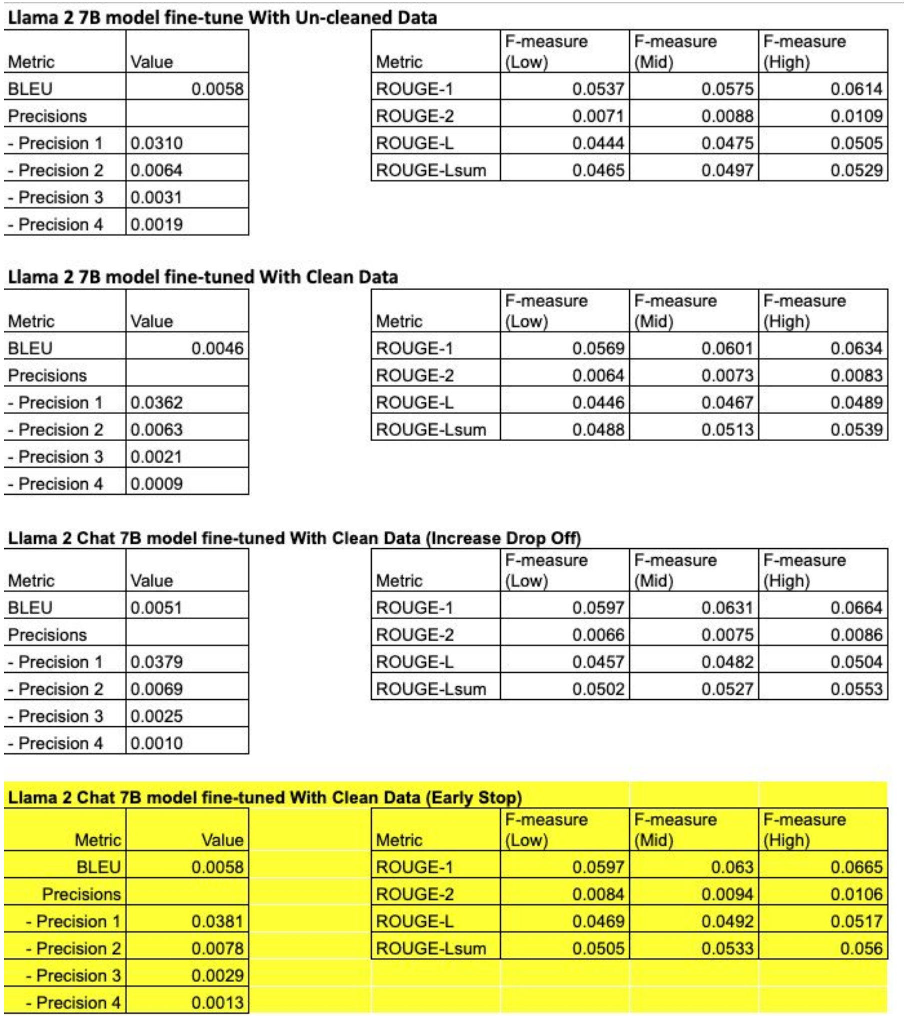
```
Mem Prams + Mem Buffer used Calculated Model Memory: 3.57 GB
Nvidia SMI reported GPU memory occupied: 5 GB.
```



```
trainable params: 39,976,960 || all params: 6,778,392,576 || trainable%: 0.589770503135875
```



Item (Full Precision)	Memory Usage (bytes per parameter)
Model Weights	4 (32bit)
AdamW Optimizer (2 states)	+8
Gradients	+4
Activations and Buffer	+8 (based on parameter sequence length, hidden size, and batch size)



Llama 2 7B model fine-tune With Un-cleaned Data

Metric	Value
BLEU	0.0058
Precisions	
- Precision 1	0.0310
- Precision 2	0.0064
- Precision 3	0.0031
- Precision 4	0.0019

Metric	F-measure (Low)	F-measure (Mid)	F-measure (High)
ROUGE-1	0.0537	0.0575	0.0614
ROUGE-2	0.0071	0.0088	0.0109
ROUGE-L	0.0444	0.0475	0.0505
ROUGE-Lsum	0.0465	0.0497	0.0529

Llama 2 7B model fine-tuned With Clean Data

Metric	Value
BLEU	0.0046
Precisions	
- Precision 1	0.0362
- Precision 2	0.0063
- Precision 3	0.0021
- Precision 4	0.0009

Metric	F-measure (Low)	F-measure (Mid)	F-measure (High)
ROUGE-1	0.0569	0.0601	0.0634
ROUGE-2	0.0064	0.0073	0.0083
ROUGE-L	0.0446	0.0467	0.0489
ROUGE-Lsum	0.0488	0.0513	0.0539

Llama 2 Chat 7B model fine-tuned With Clean Data (Increase Drop Off)

Metric	Value
BLEU	0.0051
Precisions	
- Precision 1	0.0379
- Precision 2	0.0069
- Precision 3	0.0025
- Precision 4	0.0010

Metric	F-measure (Low)	F-measure (Mid)	F-measure (High)
ROUGE-1	0.0597	0.0631	0.0664
ROUGE-2	0.0066	0.0075	0.0086
ROUGE-L	0.0457	0.0482	0.0504
ROUGE-Lsum	0.0502	0.0527	0.0553

Llama 2 Chat 7B model fine-tuned With Clean Data (Early Stop)

Metric	Value
BLEU	0.0058
Precisions	
- Precision 1	0.0381
- Precision 2	0.0078
- Precision 3	0.0029
- Precision 4	0.0013

Metric	F-measure (Low)	F-measure (Mid)	F-measure (High)
ROUGE-1	0.0597	0.063	0.0665
ROUGE-2	0.0084	0.0094	0.0106
ROUGE-L	0.0469	0.0492	0.0517
ROUGE-Lsum	0.0505	0.0533	0.056

Chatbot Demo