Advanced Generative Chatbot Design

Presentation Video YouTube Link: https://youtu.be/NWtYlHZf7RQ

Team 6

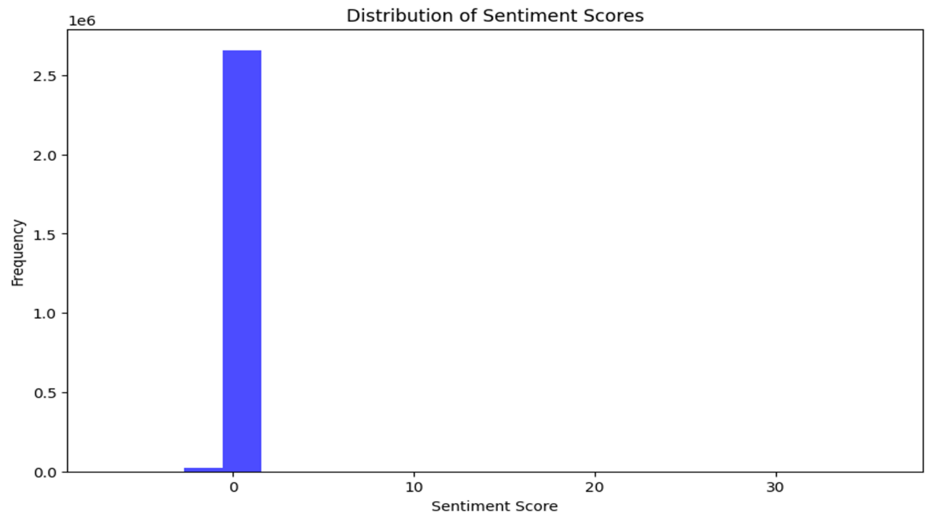
Bin Lu

Viktor Veselov

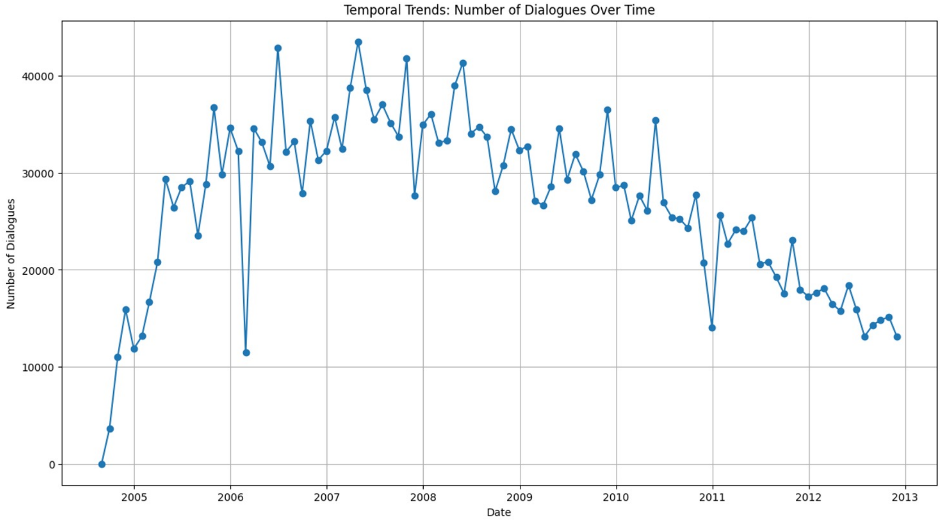
Isaack Karanja

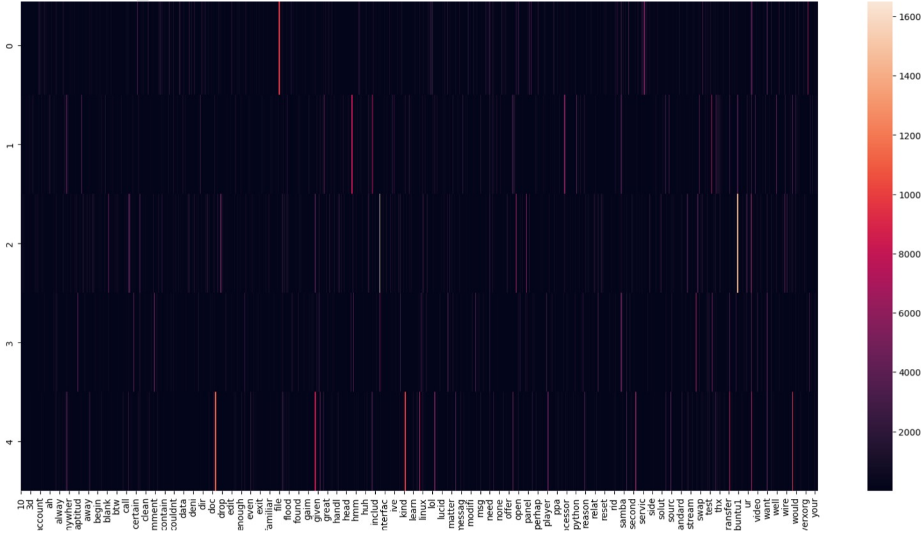
Distribution of Sentiment Scores

- 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

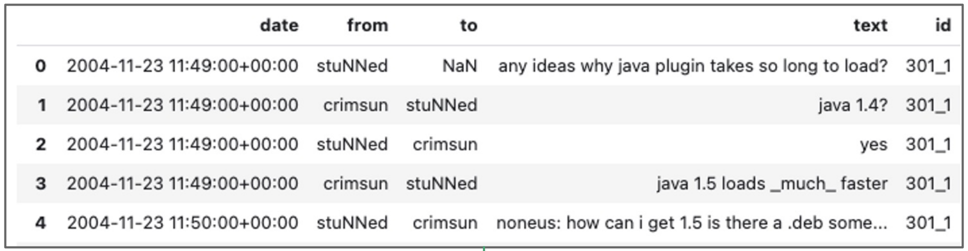
Temporal Trends:

- 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:

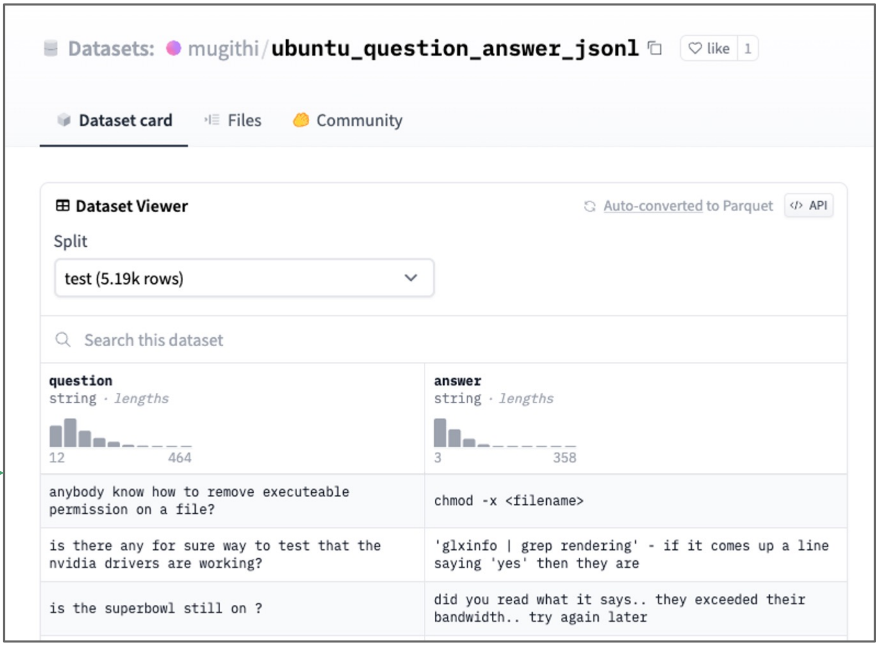
- Heatmap showcases relationships between specific lexemes and thematic clusters.
- Topic 4 prominently associates with terms like 'doc,' 'given,' and 'lol,' indicating colloquial language use.
- Presence of casual terms suggests extraneous content, hinting at potential for refining dataset content.

Dataset : The Ubuntu DataSet Corpus



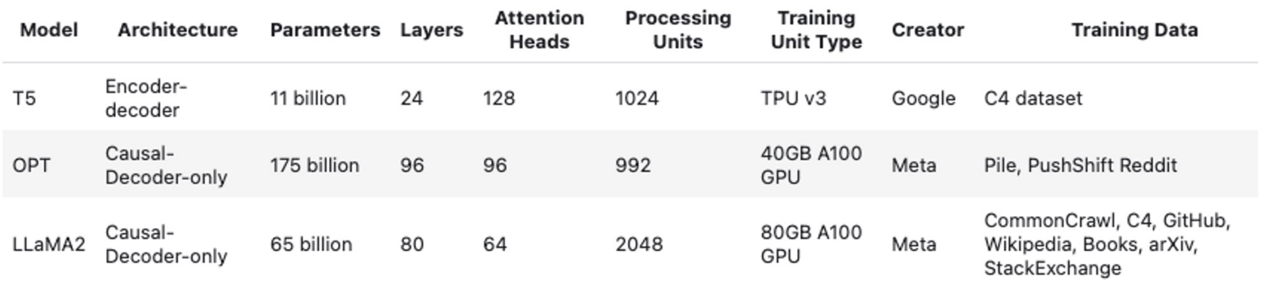
	date	from	to	text	id
0	2004-11-23 11:49:00+00:00	stuNNed	NaN	any ideas why java plugin takes so long to load?	301_1
1	2004-11-23 11:49:00+00:00	crimsun	stuNNed	java 1.4?	301_1
2	2004-11-23 11:49:00+00:00	stuNNed	crimsun	yes	301_1
3	2004-11-23 11:49:00+00:00	crimsun	stuNNed	java 1.5 loads _much_ faster	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)
```



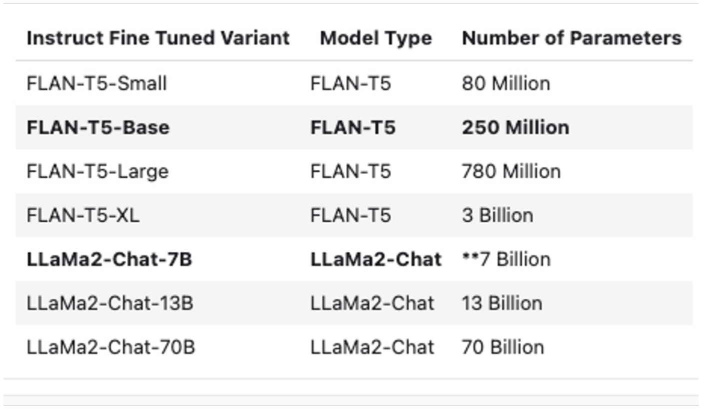
question	answer
anybody know how to remove executeable permission on a file?	chmod -x <filename>
is there any for sure way to test that the nvidia drivers are working?	'glxinfo | grep rendering' - if it comes up a line saying 'yes' then they are
is the superbowl still on ?	did you read what it says.. 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

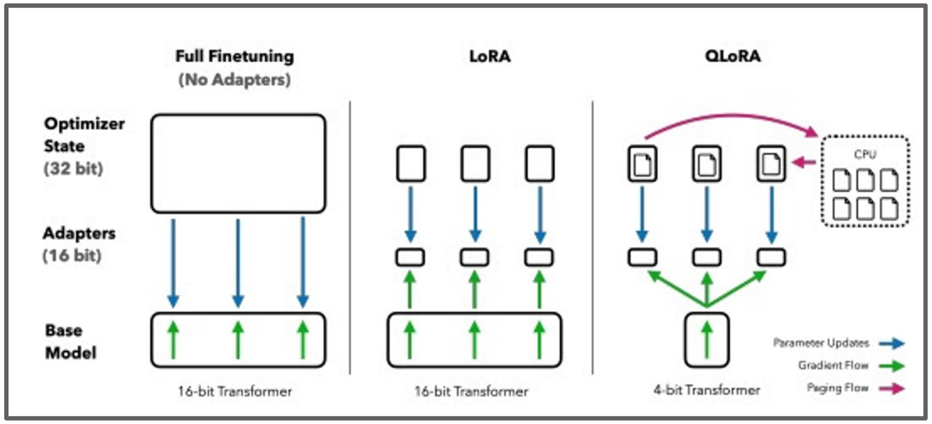


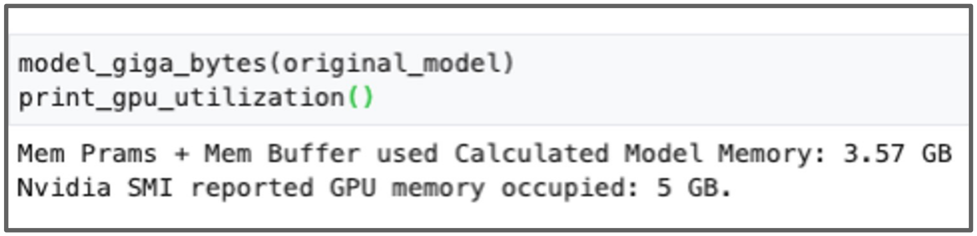
Instruct Fine Tuned Variant	Model Type	Number of Parameters
FLAN-T5-Small	FLAN-T5	80 Million
FLAN-T5-Base	**FLAN-T5**	**250 Million**
FLAN-T5-Large	FLAN-T5	780 Million
FLAN-T5-XL	FLAN-T5	3 Billion
LLaMa2-Chat-7B	**LLaMa2-Chat**	**7 Billion
LLaMa2-Chat-13B	LLaMa2-Chat	13 Billion
LLaMa2-Chat-70B	LLaMa2-Chat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

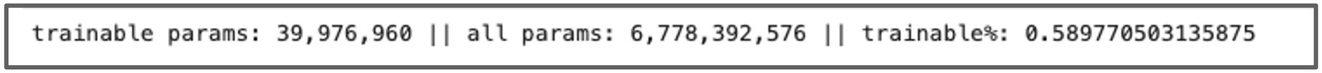
Training | LoRA, QLoRA and Memory Requirements



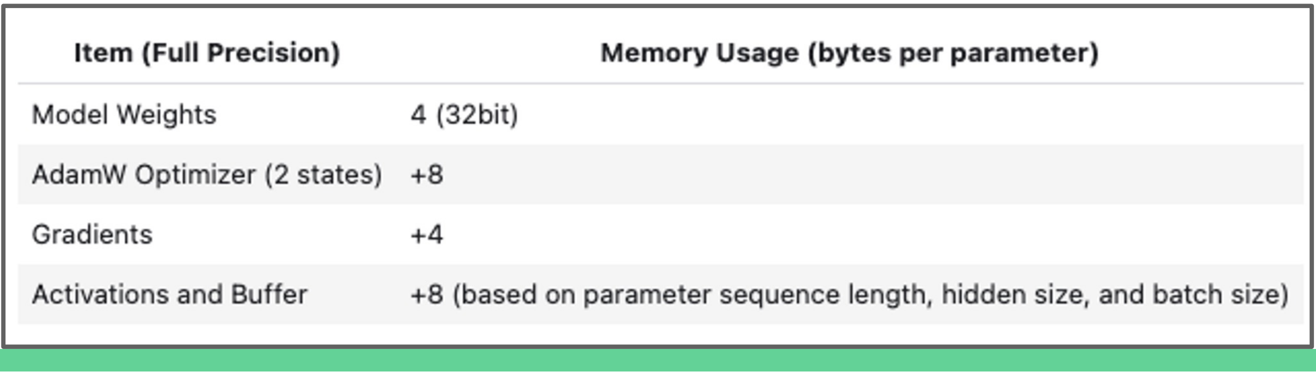


```
model_giga_bytes(original_model)
print_gpu_utilization()
```

```
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)

BLEU and ROUGE

- BLEU precisions: 0.017241379310344827
- ROUGE1 precision: 0.011494252873563218 recall: 0.3333333333333333
- **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"]

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