A.N.A.N.T

(Generative AI for Auditing, Law & Compliances)

Abstract

A.N.A.N.T (Advance Neural Assistant Networking Technology) is a state-of-the-art model for Natural Language Processing trained over a massive, curated corpus of assistant interactions including but not exclusive to word problems, multi-turn dialogue, code, accounting data, company law, IND AS, IFRS, Standards on Auditing etc. The model is a fine-tuned version of the Mistral-7b model release by Mistral AI team which leverages grouped-query attention (GQA) for faster inference with a reduced inference cost.

1 Data Collection and Curation

ANANT was trained over a carefully curated dataset comprising of data related to audit, accounting standards, company law, form filing, tax compliances, money management, mathematic and logical reasoning, text generation and dialogue completion, the dataset was curated from web and using the chat GPT model to filter, deduplicate and rank the best response. Ranking was the most important process for training ANANT, I used AI Feedback (AIF) to remove the low ranking(y_L) responses in the dataset and the best(y_w) response was selected based on the scoring done by GPT.

2 Model Training and Hyper Parameters

The model was trained using the PEFT (Parameter-Efficient Fine-Tuning) by utilizing LoRA on a T4 GPU with a learning rate of 2e-5 with 3 epochs to bring down the loss and improve quality of the model. Please refer table 1 for detailed variables used for fine tuning the model.

Hyperparameters used for training the model.

Variable Name	Value
learning_rate	2.00E-05
num_epochs	3
batch_size	1
block_size	1024
trainer	"sft"
warmup_ratio	0.1
weight_decay	0.01
gradient_accumulation	4
use_fp16	TRUE
use_peft	TRUE
use_int4	TRUE
lora_r	16
lora_alpha	32
lora_dropout	0.05

Table 1

Training Loss & Performance:

It took around 3 hours and 40 minutes to train the model on a single T4 GPU with the following loss which reduced after every epoch:

Loss	Learning Rate	Epoch
1.1774	1.64E-05	0.8
0.9869	1.06E-05	1.6
0.9382	4.66E-06	2.39
0.9241	4.69E-06	2.96

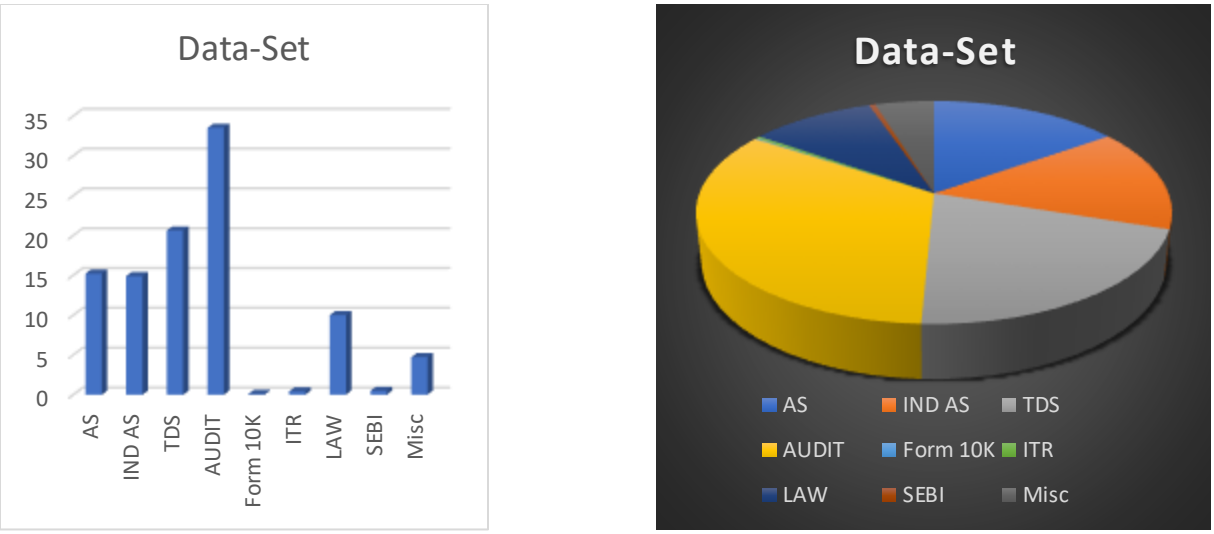

Figure 1. Graph Showing details of dataset used for training the A.N.A.N.T model.

Future Plan

- Improve the quality of model with ANANT_v2.0 dataset which is already under progress.
- Implementation of Retrieval-augmented generation (RAG) framework for retrieving facts from an external knowledge.
- Suggestions are always welcome please mail me at : computerauditor@protonmail.com

Connect with ME:

X (formerly Twitter) : **https://twitter.com/computerauditor**

Linked IN : https://in.linkedin.com/in/dev-computer-auditor-62a356217

Mail : computerauditor@protonmail.com

System Requirement:

The model was tested on various hardware from i3 to i7 with and without GPU having RAM ranging from 8GB to 32GB. The minimum requirement to run the model was found to be 8GB of RAM without GPU with at least an i5-6200U CPU @ 2.30GHz 2.40 GHz.

Token Generation: 1.5 token per second on CPU only, with GPU the same was about 3.5 tokens per second.

Comparison with GPT 3.5 Turbo (Summary)

- **Law:** A.N.A.N.T **outperformed** CHAT GPT 3.5 in terms of LAW specially with respect to corporate law, Economic Laws, SEBI laws, Contract and Company Act related prompts.
- **TAXATION:** ANANT **is at par** CHAT-GPT 3.5 in terms of tax related prompts specially tax planning, but is updated with TDS and GST related queries, fiscal budget queries and updates, tax & GST rates and amendments up to 1.07.2023.
- **ACCOUNTING AND FINANICAL REPORTING:** CHAT-GPT was **at par** with A.N.A.N.T w.r.t AS & IND AS but was outperformed A.N.A.N.T when it comes to IFRS & US GAAPs.
- **AUDITING:** ANANT has been specifically trained in terms of

auditing related queries be it Standards on Auditing, Drafting Audit Report, Compliances with Schedule 3 , CARO 2020 and much more. Responses of A.N.A.N.T were generally better than that of Chat-GPT.

- I will be releasing a detailed comparison with a demonstration video and possible an API for some time to access A.N.A.N.T beta for the general public or selective group.

Misc: For research and education purpose, fine tuning the Zephyr 7B- Beta model gave the best result. Direct Preference Optimization (DPO) Alignment looks really promising and maybe the next models of A.N.A.N.T will be based on the DPO Alignment to boost the performance of the model.

Acknowledgement & References:

All of this couldn't be possible without the support of my God, family, and friends. Also, I would like to say a huge thanks to the whole open source LLM development community and specially to the following people who were like mentor to me during this journey:

1. **Mistral AI Team** : Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed

   Technical report of Mistral 7B : https://arxiv.org/abs/2310.06825

2. **Zephyr Team** : Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, Thomas Wolf

   Technical report Of Zephyr Model: https://arxiv.org/abs/2310.16944

3. **Prompt Engineering: (Creator of Local GPT):**
   Reference to his work :
   https://github.com/PromtEngineer/localGPT

   YouTube Channel :
   https://www.youtube.com/@engineerprompt
   https://twitter.com/engineerrprompt

4. **Abhishek Thakur :**

   AutoTrain and AutoTrain Advance

   https://www.linkedin.com/abhi1thakur/
   YouTube Channel :

   https://www.youtube.com/AbhishekThakurAbhi

5. And so many more….
   Thanking the Hugging Face and the whole open-source community.