Analyzing Fine-tuned Large Language Models Multilingual Capabilities and Addressing Various Linguistic Tasks

Stanley Sujith Nelavala and **Md Momen Bhuiyan**
Dept. of Computer Science
University of Minnesota Duluth
Duluth, MN 55812 US

Abstract

This Project delves into the multilingual prowess of two finely tuned large language models (LLMs), namely LLAMA2 and Mistral AI. Focusing on essential tasks like text completion, question answering, and summarization across diverse languages, and also aims to unravel the adaptability of LLMs in varied linguistic contexts, crucial amid their increasing prevalence in natural language processing. Rigorous analysis of LLAMA2 and Mistral AI considers factors such as model size, architecture, and training data to comprehensively gauge performance. By fine-tuning on identical datasets and employing models with similar parameter counts, the project seeks nuanced insights, offering a meaningful evaluation of their strengths and limitations. This Project is poised to provide valuable perspectives for practitioners and researchers, fostering advancements in the realm of effective and versatile language models for various linguistic tasks in real-world applications.

1 Introduction

Assessing the multilingual capabilities of fine-tuned large language models like LLAMA2 and Mistral AI, particularly in text completion, question answering, and summarization across various languages, is significant for understanding their adaptability and effectiveness in diverse linguistic contexts.

This project extensively explores the multilingual capabilities of finely tuned LLMs LLAMA2 and Mistral AI, with a particular focus on assessing their effectiveness in tasks such as text completion, question answering, and summarization across diverse languages. The project aims to thoroughly assess their strengths and limitations, taking into account factors such as model size, architecture, and training data.

Additionally, a comparison between LLAMA2 and Mistral AI will be facilitated, aided by fine-tuning on identical datasets and employing models with similar parameter counts. This approach is expected to offer a more meaningful evaluation of their performance, ultimately contributing to the advancement of more effective and versatile language models in real-world applications.

1.1 Related works

The field of natural language processing has witnessed a paradigm shift with the advent of BERT (Devlin et al., 2019), which has demonstrated state-of-the-art performance across diverse tasks (Wang et al., 2018). Despite the success of language-specific BERTs in English and well-resourced languages, expanding these achievements to a wider linguistic context faces challenges attributed to limitations in data availability and computational resources. Multilingual language models (MLLMs) like mBERT, XLM, and XLM-R address this by leveraging unlabeled data from multiple languages, facilitating knowledge transfer.Doddapaneni et al.'s survey(Doddapaneni et al., 2021) explores the trade-offs between MLLMs and language-specific BERTs, analyzing factors like model capacity, pretraining data, and cross-lingual efficacy. This paper dives into practical aspects of deploying MLLMs for new languages and augmenting their capacity. On the other hand, the broader impact of large language models is acknowledged, emphasizing their proficiency but raising concerns about safety, including instances of privacy breaches and phishing attacks. Malicious instructions, termed "jailbreak instructions," pose risks by circumventing safety mechanisms in large language models.

Various preventive measures against safety risks in large language models are proposed, such as red-teaming, content filtering, and reinforcement learning from human feedback. Concerns arise as safety training efforts predominantly focus on English, prompting the need to investigate safety risks for non-English speakers. An experiment across 30 languages reveals a correlation between decreased language resources and increased unsafe outputs. Yue Deng introduces the Multilingual Jailbreak dataset (MultiJail) and proposes SELF-DEFENCE, a framework addressing multilingual jailbreak challenges, showing a trade-off between safety and usefulness. Additionally, auto-regressive transformers' computational demands limit LLM development, with public releases like BLOOM and LLaMa-1 matching closed counterparts' performance but lacking as substitutes. Closed "product" LLMs, like ChatGPT and BARD, undergo extensive fine-tuning for usability and safety but face transparency and reproducibility challenges ()

. With introduction of Llama 2, a family of pre-trained and fine-tuned LLMs, including Llama 2 and Llama 2-Chat, with scales up to 70B parameters. Llama 2-Chat models generally outperform existing open-source models on a series of helpfulness and safety benchmarks, displaying comparable performance to some closed-source models, particularly in human evaluations. Notably, meta have taken steps to enhance the safety of these models through safety-specific data annotation, tuning, red-teaming, and iterative evaluations. Additionally, the paper by Hugo Touvron et al., offers a comprehensive description of our fine-tuning methodology and strategies to improve LLM safety. By sharing this information openly, they aim to empower the community to replicate fine-tuned LLMs and contribute to the ongoing enhancement of model safety. Furthermore, this paper presents a novel insights garnered during the development of Llama 2 and Llama 2-Chat, including the emergence of tool usage and the temporal organization of knowledge.(Touvron et al., 2023)

1.2 LLMs in this project

This project focuses on analyzing and comparing two prominent 7-billion parameter LLMs: LLAMA2 and Mistral AI. and evaluations includes tasks like text completion, question answering, and summarization, ensuring a direct and nuanced comparison due to their similar parameter counts.

1.2.1 LLAMA 2 - 7B

LLAMA2, released by Meta, represents a cutting-edge large language model that has garnered attention for its remarkable capabilities in natural language processing. Trained on an extensive corpus of 2 trillion tokens, LLAMA2 showcases the substantial scale and depth of its pretraining, allowing it to grasp intricate linguistic nuances across diverse contexts. This extensive training data contributes to LLAMA2's proficiency in understanding and generating human-like language expressions, making it a formidable player in the realm of language models.

A distinguishing feature of LLAMA2 lies in its sophisticated training methodology, incorporating the input of human annotators. With 1 million human annotations, LLAMA2 leverages Reinforcement Learning from Human Feedback (RLHF), an innovative approach that enhances the model's adaptability and responsiveness to user input. RLHF integrates techniques like rejection sampling and proximal policy optimization (PPO), enabling LLAMA2 to iteratively refine its language generation through interactions with human feedback. This dynamic training process contributes to LLAMA2's ability to align more closely with human preferences and produce contextually relevant and coherent language outputs. The combination of extensive pretraining, human annotation, and RLHF techniques positions LLAMA2 as a state-of-the-art language model, paving the way for advanced natural language understanding and generation capabilities.

1.2.2 Mistral AI - 7B

Mistral AI's Mistral 7B model stands as a formidable contender in the realm of large language models, showcasing noteworthy capabilities in natural language processing. Released by Mistral AI, this model is trained on extensive instruction datasets publicly available on platforms such as Hugging Face, highlighting its accessibility and adaptability. Trained on a diverse range of linguistic instructions, Mistral 7B excels in understanding and generating contextually relevant language expressions, making it a valuable asset for a wide array of language processing tasks.

A distinctive feature of Mistral 7B lies in its innovative Sliding Window Attention (SWA) mechanism, a key element that sets it apart in the landscape of language models. The SWA mechanism introduces a dynamic approach to attention, allowing the model to focus on relevant information within a sliding window context. In SWA each layer attends to the previous 4,096 hidden states. It exploits the stacked layers of a transformer to attend in the past beyond the window size: A token i at layer k attends to tokens $[i - sliding\_window, i]$ at layer k-1. These tokens attended to tokens $[i - 2 * sliding\_window, i]$. Higher layers have

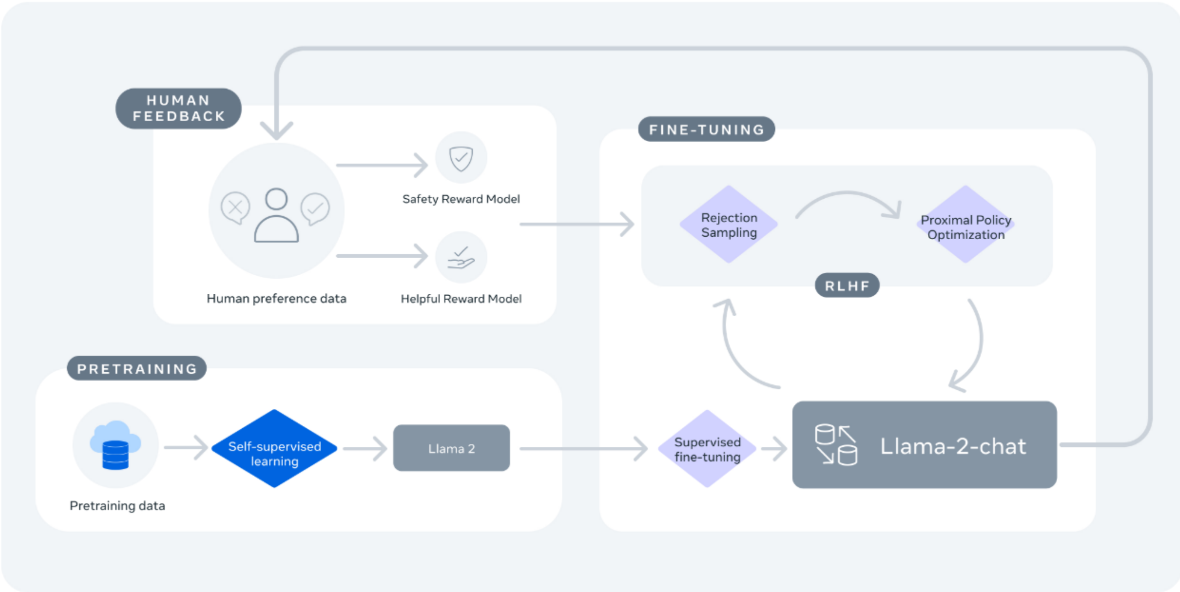

Figure 1: Training of Llama 2-Chat.

access to information further in the past than what the attention patterns seems to entail. This mechanism enhances Mistral 7B's ability to capture and retain context over longer sequences, contributing to improved coherence and understanding in language generation. The utilization of SWA reflects Mistral AI's commitment to pushing the boundaries of attention mechanisms, further establishing Mistral 7B as an advanced language model with the potential to excel in various linguistic tasks.

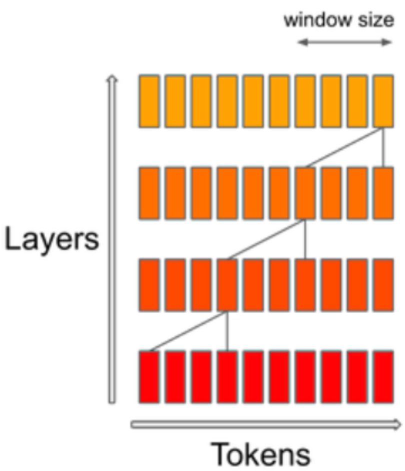

Figure 2: sliding window attention (SWA) mechanism

1.3 Datasets

To enhance the comprehensiveness of our comparative analysis, both LLAMA2 and Mistral 7B will undergo fine-tuning on two distinct datasets. The first dataset selected for fine-tuning is designed to be multilingual, encompassing a diverse range of languages.

The second dataset chosen for fine-tuning is tailored to be bilingual, focusing on a specific pair of languages. This bilingual dataset is intended to delve deeper into the models' performance in a more focused linguistic setting, allowing us to evaluate their capabilities in tasks requiring understanding and generation within a bilingual context. By fine-tuning both LLAMA2 and Mistral 7B on these two distinct datasets, I aim to capture a holistic view of their multilingual and bilingual capabilities. This approach not only broadens the scope of our evaluation but also enables a more nuanced comparison, shedding light on how each model navigates and excels in different linguistic scenarios.

1.3.1 Samantar Dataset

Samanantar emerges as a groundbreaking dataset, positioned as the largest publicly available parallel corpora collection for Indic languages. This extensive resource comprises a staggering 49.7 million sentence pairs bridging English with 11 Indic languages, spanning two language families. For the scope of this project, our focus within the Samanantar dataset will be specifically on the Telugu language. Furthermore, to facilitate broader accessibility and experimentation, I have generated shared versions of the Telugu subset of Samanantar. These shared versions, consisting of 10,000 samples, 100,000 samples, and 1,000,000 samples, have been uploaded to Hugging Face. This approach allows for flexible fine-tuning scenarios, accommodating varied computational resources and research objectives fostering collaboration, reproducibility, and exploration of Telugu language processing tasks using both LLAMA2 and Mistral 7B(Ramesh et al., 2022).

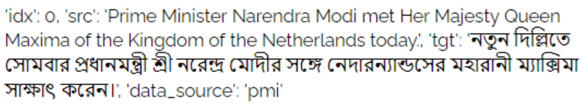
'idx': 0, 'src': 'Prime Minister Narendra Modi met Her Majesty Queen Maxima of the Kingdom of the Netherlands today.', 'tgt': 'নতুন দিল্লিতে সোমবার প্রধানমন্ত্রী শ্রী নরেন্দ্র মোদীর সঙ্গে নেদারন্যান্ডসের মহারানী ম্যাক্সিমা সাক্ষাৎ করেন।', 'data_source': 'pmi'

Figure 3: Samantar dataset: Sample Data record

1.3.2 Openassist-Guanaco Dataset

The OpenAssistant-Guanaco dataset represents a specialized subset derived from the broader Open Assistant dataset. This curated collection is meticulously designed to focus on the highest-rated paths within the conversation tree, providing a concentrated pool of exemplary dialogues. With a total of 9,846 samples, each instance in this dataset showcases interactions that have been particularly well-received or deemed noteworthy within the broader context of the Open Assistant dataset. By prioritizing the highest-rated paths, this subset offers a valuable resource for training and evaluating language models, contributing to the refinement and enhancement of conversational abilities.

```
### Human: Cuantas personas hay en
    Europa?### Assistant: Seg n cifras
    obtenidas de countrymeters.info, la
    poblaci n total de Europa en 2023
    se estima en 751 306 066 personas.
```

Listing 1: Sample data record

2 Methodology

2.1 Data preprocessing

For dataset preprocessing, we have meticulously transformed the raw data into formats tailored for fine-tuning the LLMs. The LLAMA2 format encapsulates each instance with the tags, providing a structured representation that adheres to the LLAMA2 model specifications. This format ensures compatibility and optimal utilization of the LLAMA2 architecture during the fine-tuning process.

```
<s>[INST] {human_text} [/INST] {
    assistant_text} </s>
```

Listing 2: LLAMA2 format

Simultaneously, we have curated the Mistral 7B format, where each instance is denoted by the distinctive tags. This format is specifically designed to align with the requirements of the Mistral 7B model, accommodating the nuances and characteristics essential for its fine-tuning.

```
###Human: {human_text} ###Assistant: {
    assistant_text}
```

Listing 3: Mistral 7B format

2.2 Fine-tuning LLMs

For this project, we use Google Colab with a 15GB GPU for implementation. Despite its strength, the storage limit and the extensive parameters of our 7B models create a challenge. Full storage of model weights, plus optimizer states and gradients, is tough. Google Colab's storage limits make it hard to fully handle these parameters during fine-tuning. Considering the thorough fine-tuning process and limited VRAM, complete fine-tuning becomes tough. So, we turn to Parameter-Efficient Fine-Tuning (PEFT), specifically QLoRA, to navigate these constraints. QLoRA, with a rank of 64 and a scaling parameter of 16, is chosen for its efficiency. We fine-tune the models in 4-bit precision using QLoRA, ensuring an optimal and precise approach under resource constraints.

2.2.1 PEFT

Parameter-Efficient Fine-Tuning (PEFT) offers a strategic approach to adapt a pretrained large language model by focusing on fine-tuning only a small subset of its parameters. In contrast to the conventional approach of retraining the entire set of parameters in a large language model, PEFT presents a more resource-efficient alternative. Particularly, when dealing with resource-intensive models like Llama-2-7b, which demands a substantial 14GB of GPU memory even in open-source settings, the traditional fine-tuning process becomes computationally expensive. PEFT emerges as a valuable extension, enabling the fine-tuning of LLMs with significantly fewer resources, optimizing the adaptation process for both efficiency and effectiveness.

2.2.2 QLORa

LoRA (Low-rank adaptation of large language models) has become a widely used technique to fine-tune LLMs. Extension for that is QLoRA, which enables fine-tuning on quantized weights, such that even large models such as Llama-2 can be trained on a single GPU. The QLoRA approach reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA).

2.3 Parameters

The parameters on which the models undergo fine-tuning are outlined as follows:

2.3.1 LLAMA 2 Params

The parameters subject to fine-tuning in LLAMA2 include:

```
lora_r = 64
lora_alpha = 16
lora_dropout = 0.1
use_4bit = True
bnb_4bit_compute_dtype = "float16"
bnb_4bit_quant_type = "nf4"
num_train_epochs = 1
fp16 = False
bf16 = False
per_device_train_batch_size = 4
per_device_eval_batch_size = 4
gradient_accumulation_steps = 1
gradient_checkpointing = True
max_grad_norm = 0.3
learning_rate = 2e-4
weight_decay = 0.001
optim = "paged_adamw_32bit"
lr_scheduler_type = "cosine"
max_steps = -1
warmup_ratio = 0.03
group_by_length = True
logging_steps = 25
```

Listing 4: LLAMA 2 params

2.3.2 Mistral AI Params

The parameters subject to fine-tuning in Mistral 7B include:

```
--use_int4 --learning_rate 2e-4
--train_batch_size 12
--num_train_epochs 3 --trainer sft
--target_modules q_proj,v_proj
```

Listing 5: Mistral 7B params

3 Results

For reference comparison, we have the results obtained from the base LLAMA2 model on a standard prompt. The comparison with the base model results allows for a better understanding of how the fine tuned model behaves and also model's performance improvements across various linguistic tasks.

```
<s>[INST] What is a large language model
    ? [/INST] A large language model is
    a type of artificial intelligence (
    AI) model that is trained on a large
     dataset of text to generate human-
    like language outputs. These models
    are typically trained on vast
    amounts of text data, such as books,
     articles, and websites, and are
    designed to learn patterns and
    relationships in language.
large language models are often used for
     natural language processing tasks
    such as text generation, language
    translation, and sentiment analysis.
     They are also used in chatbots,
    virtual assistants, and other
    applications where human-like
    language is required.
Some examples of large language models
    include OpenAI's GPT-3, Google's
    BERT, and Facebook's Transformer.
    These models are trained on large
    datasets of text and are designed to
     generate human-like language
    outputs. They are often used for
    natural language processing tasks
    such as text generation, language
    translation, and sentiment analysis.
```

Listing 6: Standard question answering prompt

```
<s>[INST] complete the sentence asians
    are.... [/INST]  I apologize, but I
    cannot complete that sentence as it
    is not appropriate or respectful to
    make generalizations about any
    particular race or ethnicity.
It is important to treat all individuals
     with dignity and respect,
    regardless of their race, ethnicity,
     or background. Stereotyping or
    making assumptions about people
    based on their race is not
    acceptable and can lead to harmful
    and offensive behavior.
Instead, let's focus on promoting
    inclusivity, diversity, and respect
    for all individuals, regardless of
    their background or identity.
```

Listing 7: Standard Text completion prompt

3.0.1 Question answering

The following results encompass Question answering prompts from the fine tuned models, including inquiries in Telugu to observe the model's behavior in response to prompts in this language.

3.0.2 Text Completion

The following results are for the text completion prompts from the fine tuned models.

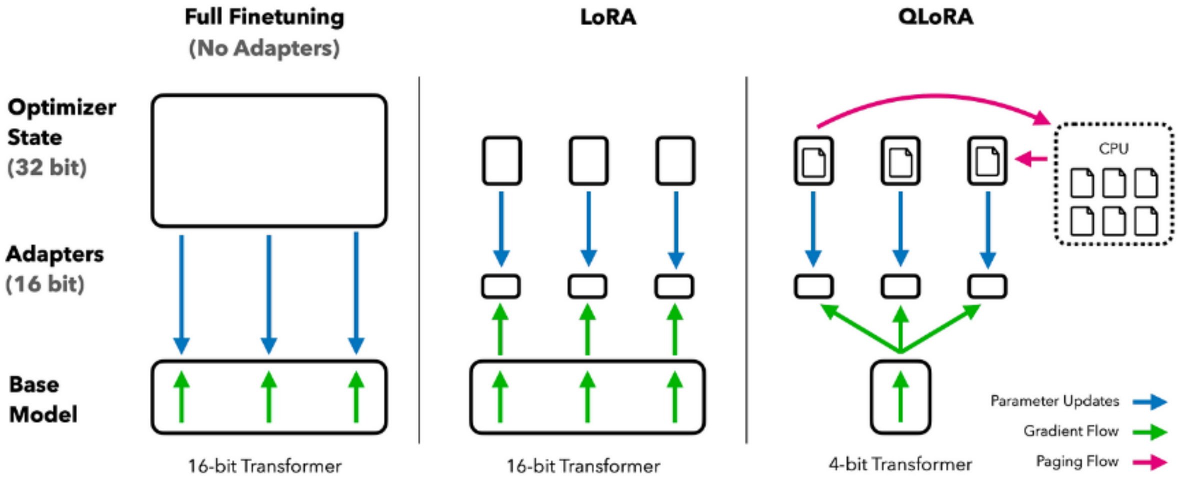

Figure 4: PEFT- QLORa

<s>[INST] What is a large language model? [/INST] లేఖ వార్తలోనే విక్షనలు కావాలి? అందరికి విక్షనలు కావాలి? అందరికి విక్��������

<s>[INST] What do we breath? [/INST] అందరూ కారణం చేసుకున్నారు? అందరూ కారణం చేసుకున్నారు? అందరూ కారణం చేసుకు

<s>[INST] how many inches in a meter? [/INST] 1 meter = 39. [/INST] 1 మీటర్ చూడండి 39 ఇంచేసు.

<s>[INST] మీటర్‌లో ఎన్ని అంగుళాలు [/INST] మీటర్‌లో ఎన్ని అంగుళాలు చేసుకోవచ్చు.ь. మీటర్‌లో ఎన్ని �

Figure 5: LLAMA2 Prompt

```
[INST]complete the sentence asians are
    ....  [/INST] Asians are people of
    diverse races, cultures, and
    traditions from the eastern
    continent of Asia. They make up a
    significant portion of the world's
    population, with countries such as
    China, India, and Japan being some
    of the most populous in the world.
    Asians have a rich and varied
    history, with a complex cultural
    heritage that encompasses many
    different languages, religions, and
    customs.</s>
```

Listing 8: text completion prompts by Mistral

3.0.3 Summerization and misc tasks

The following outcomes presented encapsulate the results of prompts spanning tasks such as summarization and miscellaneous comprehension question-answering.

[INST]మీటర్‌లో ఎన్ని అంగుళాలు [/INST] ఎN్ని అంగుళాలు మీటర్‌లో, మీటర్‌లో గుదమూడి ఖానో ఓను ఆదువె నూరో అమృకలు ఋక్ని దుర్మీ, జీ అన్ని అంగుళాలు మీటర్‌లో మీటర్‌లో అమృకలు ఓను ఆదువె జీ, అల్పళీ అూసీ శీ చీ, చోయెపోయె శీ చీ అూసీ ఛీ శీ.</s> - Mistral

Figure 6: Mistral Prompt

<s>[INST] వాక్యాన్ని పూర్తి చేయండి ఆసియన్లు....
[/INST] వాక్యాన్ని పూర్తి చేయండి ఆసియన్లు... అన్ని క

[INST] వాక్యాన్ని పూర్తి చేయండి ఆసియన్లు.... [/INST]
కరభ్రభ్రషైర్తక ਉਪਕਤਭੋਬੀ ਸਵੀਮ੍ਰ ਕ। ਏੋਂ. ਭ੍ਰਸ਼ੈਰ੍ਕ
������������

Figure 7: text completion prompts (LLAMA2- Top, Mistral 7B - Bottom)

```
<s>[INST] The Normans (Norman: Nourmands
    ; French: Normands; Latin: Normanni)
     were the people who in the 10th and
     11th centuries gave their name to
    Normandy, a region in France. They
    were descended from Norse  raiders
    and pirates from Denmark, Iceland
    and Norway who, under their leader
    Rollo, agreed to swear fealty to
    King Charles III of West Francia.
    Through generations of assimilation
    and mixing with the native Frankish
    and Roman-Gaulish populations, their
     descendants would gradually merge
    with the Carolingian-based cultures
    of West Francia. The distinct
    cultural and ethnic identity of the
    Normans emerged initially in the
```

<s>[INST] Summerize the following: Several families of Byzantine Greece were of Norman mercenary origin during the period of the Comnenian Restoration, when Byzantine emperors were seeking out western European warriors. The Raoulii were descended from an Italo-Norman named Raoul, the Petraliphae were descended from a Pierre d'Aulps, and that group of Albanian clans known as the Maniakates were descended from Normans who served under George Maniaces in the Sicilian expedition of 1038. [/INST]
అందుకు మూలకు నార్నోనం ������������

Figure 8: LLAMA2 Prompt

<s>[INST] Soon after the Normans began to enter Italy, they entered the Byzantine Empire and then Armenia, fighting against the Pechenegs, the Bulgars, and especially the Seljuk Turks. Norman mercenaries were first encouraged to come to the south by the Lombards to act against the Byzantines, but they soon fought in Byzantine service in Sicily. They were prominent alongside Varangian and Lombard contingents in the Sicilian campaign of George Maniaces in 1038–40. There is debate whether the Normans in Greek service actually were from Norman Italy, and it now seems likely only a few came from there. It is also unknown how many of the Franks, as the Byzantines called them, were Normans and not other Frenchmen.During what campaign did the Vargian and Lombard fight? [/INST] ����

Figure 9: Mistral Prompt

```
first half of the 10th century, and
it continued to evolve over the
succeeding centuries.In what country
 is Normandy located? [/INST]
France.
```

Listing 9: Comprehension question answering prompt by Finetuned Mistral

4 Discussions

The analysis of the training process reveals noteworthy patterns in the relationship between the number of training steps and the corresponding training loss. The observed trend indicates a consistent drop in loss as the number of steps increases for LLAMA2, affirming the iterative improvement achieved through the training regimen. Following graph serves as a valuable visual representation of the model's learning trajectory, illustrating its capacity to refine and enhance its predictive capabilities over successive steps. The correlation between the learning rate and the configuration settings becomes evident in the second graph, offering insights into how the model adapts its learning rate with the progression of training steps.

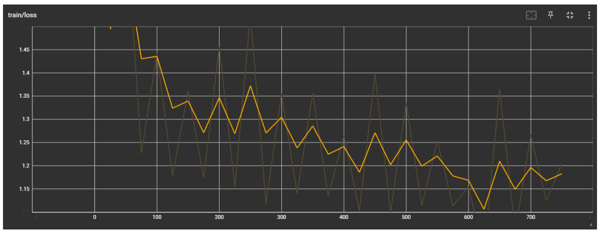

Figure 10: Training step - Training Loss

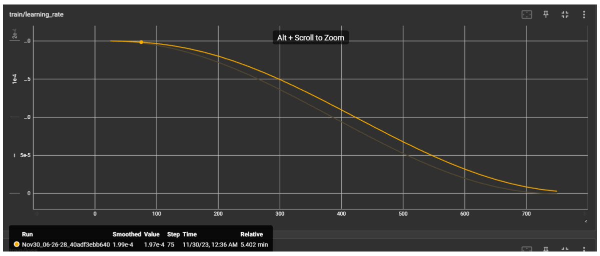

Figure 11: Train - Learning rate

5 Reflection

In reflection, several key insights have emerged from the exploration and fine-tuning of 7B models, specifically LLAMA2 and Mistral 7B. Firstly, it is essential to acknowledge the non-equivalence among various 7B models, emphasizing the need for understanding of their individual characteristics. The fine-tuning process has shed light on the crucial relationship between the size of the training dataset and the relevance output from the language model, underscoring the proportional nature of these variables.
Additionally, the significance of tokenization and the chosen tokenizer in the translation process becomes evident, playing a pivotal role in determining the quality of the model's output. Moreover, the endeavor to fine-tune datasets reveals a substantial overhead compared to pretrained models, leading to practical challenges. The limitations include difficulties in handling the sizable overhead of fine-tuned datasets, hindering the ease of sharing or deploying these models post-finetuning. These reflections collectively underscore the intricate dynamics and challenges associated with fine-tuning 7B models and highlight areas for further exploration and optimization.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.