# LawyerGPT-Trained on Indian Legal Dataset

Pratik Behera, Nisaar Agharia

August 10, 2023

**Note:** This is the first draft for the endorsement process, not for distribution.

## Abstract

In the recent past, training and fine-tuning Large Language Models (LLMs) was considered a daunting task due to the significant computational resources required. However, the advent of quantization techniques has democratized access to LLMs, enabling not only researchers but also tech enthusiasts to experiment with these models. Our study centers on the fine-tuning of the Falcon-7B-instruct model. This model was trained on a unique dataset we curated, primarily sourced from Indian Kanoon, a comprehensive database of Indian legal documents, in addition to other Indian legal platforms. To ensure the broad applicability of our model, we adopted the instruction format of Stanford's Alpaca project, which utilizes a diverse instruction set from 400 different tasks. More details about our methodology for dataset creation are discussed extensively in the dataset creation section of the paper. To gauge the performance of our fine-tuned Falcon-7B-instruct model, we used GPT-4, a cutting-edge language model from OpenAI, as the evaluator. When benchmarked against existing LLMs such as GPT-3.5-turbo and Claude, our model displayed significant superiority in the generation of legal text, a finding that underscores the success of our fine-tuning approach. These benchmarking results are attached for your reference. In addition to this, we also trained the newly released Llama2 model on the same dataset. This was done to provide a comparative study between Falcon-7B-instruct and Llama2, to discern the performance differences and potential advantages each model brings to the table in the context of processing legal data. This

paper will provide comprehensive insights into our research process, the challenges encountered, and how we addressed them, aiming to contribute to the growing body of knowledge on fine-tuning LLMs for specialized tasks.

# 1 INTRODUCTION

The rapid progress of large language models has accentuated the importance of their efficient use. Earlier strategies, including the 16-bit model architecture and the LoRA Config method, yielded moderate hardware requirement reductions. The more recent advancement, the QLoRA fine-tuning method, promises a significantly optimized architecture for fine-tuning models on consumer-grade GPUs. In this study, we utilize QLoRA to fine-tune a dataset that we have carefully prepared and curated.

Training on extensive datasets is advantageous for downstream tasks, but it demands substantial computational resources and time. Interestingly, it has been noted that strategically prepared datasets can lead to higher accuracy, a phenomenon reported in the recent Orca paper. Furthermore, the concept of Chain of Thought Prompting (CoT) – where the AI model is prompted to elaborate on the rationale behind its responses – has been shown to enhance accuracy relative to the volume of data used, as discussed in the Less is More Aligned (LiMA) paper. Taking inspiration from these findings, our dataset combines both human-generated and synthetic data in an instruction format similar to the one used in the Stanford Alpaca paper.

**Our research makes several key contributions:**

- We utilize data pairs that contain responses and queries judiciously sampled from Indian legal websites, such as Indian Kanoon. Our initial inference dataset comprises 150 sets of carefully selected instructions.

- We expand the dataset by integrating instruction sets from leading large language models. We have incorporated approximately 1000 instructions (933, to be precise). In subsequent steps, we enlarged the instruction set using data derived from a strategically selected set of legal articles. Our current instruction set comprises approximately 3.4k instructions.

- We employ effective text summarization techniques to transform complex legal documents into more comprehensible formats. The ambiguity

often associated with legal documents available online can lead to the loss of core meanings due to context loss. By applying the method proposed in the paper "Textbooks are all you need", we generate synthetic data to preserve and enhance semantic relationships within the text.

- In our initial benchmarking phase, we employed GPT-4 as the evaluation model. Encouragingly, our model produced responses that were not only more clearly articulated but also cleaner compared to those generated by currently top-rated large language models. This suggests that our approach of careful dataset preparation, judicious data sampling, and effective text summarization holds substantial promise for enhancing the performance and usability of large language models in the legal domain.

# 2 RELATED WORKS

A significant body of research exists around training large language models in the legal sector, with many studies relying heavily on comprehensive corpora primarily from U.S. and U.K. legal systems [link to the dataset]. Notably, some models, such as InLegalBERT and its derivatives [link to the paper], have been trained specifically on Indian legal datasets. These datasets often source their data from robust Indian legal repositories like Indian Kanoon, which offers a copious collection of documents, including around 22.7 million texts encompassing an impressive size of 84GB. The data in these repositories span from 1950 to 2019, providing a rich historical context for AI training.

Despite the extensive nature of these datasets, certain challenges exist. The Indian legal terminology, characterized by its unique jargon, can make these court cases difficult to interpret. Furthermore, older court cases that were manually transcribed during the digitization process often contain inconsistencies and anomalies.

In this research, we propose strategies to navigate these challenges, making the wealth of data in the Indian legal system more accessible for large language model training and ultimately contributing to the advancement of AI in the legal domain. By addressing these challenges, we hope to harness the depth and breadth of this extensive legal dataset more effectively.

# 3 DATASET PREPARATION

The first step in our process was to emulate the analytical approach employed by a legal practitioner when dissecting a court case or document. Critical elements such as headnotes, case citations, case history, legal issues/questions presented, applicable legal provisions, holdings, legal reasoning, rule of law, concurring or dissenting opinions, implications, significance, and commentary or analysis, consistently surfaced as key components in a legal document. By using these elements as an organizing framework, we were able to efficiently summarize our collection of legal documents.

Initially, we manually curated an instruction dataset comprising 150 prompts, drawing inspiration from the Constitution of India. Over time, we expanded this dataset, ultimately generating 933 prompts in an input-output format, utilizing GPT-4 as a guiding mechanism.

For a granular analysis, we segmented the Indian Constitution into manageable chunks, ensuring the maintenance of context in the summarization process. The top-ranked large language models (LLMs) were employed to generate input-output pairs, which were subsequently converted into the requisite format, i.e., {instruction, input, output}. This transformation was executed with the assistance of GPT-4, currently the highest-rated LLM. A targeted selection of articles (12,14,15,19,21) from the Indian Constitution was utilized for instruction set generation, with landmark cases predominantly pertaining to these articles incorporated into the training corpus. For this investigation, we selected approximately 50 court cases, from which we generated a total of 3,300 prompts for fine-tuning the pre-trained LLMs, Falcon-7B-instruct and Llama2.

The summarization of selected articles was executed using highly rated LLMs, namely GPT-3.5 turbo, Claude, and GPT-4. Approximately 40% of the data was generated using GPT-3.5-turbo, while the remaining data was split, with 40% from GPT-4 and 20% from the newly released Claude AI.

In subsequent iterations, we automated the data generation process, generating synthetic data with the assistance of the most proficient large language models. After careful summarization, legal documents were prompted to generate responses in the format of instruction, input, and output.

We then created a diverse instruction set composed of 400 varied tasks, which we applied across all legal court cases. The first 50 sets of tasks were sequentially sampled along with the summaries, guiding the LLMs in the generation of the instruction set for model training. The generated instructions

were diverse and well-explained, enhancing overall comprehensibility.

Finally, the dataset was further processed into a structured format: [{instruction: "Summarize the given case and explain. . . ", input: "Case Details", output: "The following case emphasizes on Article...", prompt: "prompt text", response: "output"}]. Upon the successful collection of the datasets, this instruction set was deployed to fine-tune the foundational large language model.

# 4 TRAINING AND HYPERPARAMETERS

The process of fine-tuning an AI model entails harnessing task-specific labeled data to significantly enhance the model's performance within a specific domain. To accomplish this objective in the legal sector, we harnessed the power of the open-source Falcon-7B-instruct, a formidable model that has demonstrated promising performance in this field [link to the paper].

Originating from a robust training background, the Falcon-7B-instruct model was trained on an impressive volume of 1500 billion tokens drawn from the Refined Web corpus, further augmented by curated datasets [link to the Open LLM Leaderboard]. Such a comprehensive corpus offers the model a diverse and rich linguistic environment, thereby building a sturdy foundation for the subsequent domain-specific fine-tuning.

To elevate the efficiency and optimization of our fine-tuning process, we adopted the innovative QLoRA (Quantized Lottery Ticket Hypothesis) configuration. This unique strategy deploys a 4-bit quantization method, significantly reducing the GPU usage without compromising the performance and capabilities of the model. Throughout our iterative experiments, we harnessed the NVIDIA A100, a high-performance GPU readily accessible via Google Colab Pro, to meet the computational demands of our large-scale model training.

Complementing our primary focus on the Falcon-7B-instruct model, we concurrently conducted training on the recently unveiled Llama2 model. This simultaneous training approach enabled us to make a detailed comparison of the inference results produced by both models, thereby providing a holistic view of their effectiveness within the legal domain.

Our approach to hyperparameter tuning was both methodical and iterative, experimenting with progressively larger dataset sizes across successive iterations. This strategy ensured a thorough examination of the impact of

various hyperparameters on model performance at diverse scales. All the parameters used in each iteration have been meticulously documented for transparency and reproducibility [link to the paper]. Through this systematic approach to hyperparameter optimization, we aim to contribute not only to the refinement of our model but also to the broader knowledge base informing future research in the field of Large Language Models.

# 5 INFERENCE

The deployment of large language models in real-world applications is facilitated by the development of highly abstracted interfaces for model applications. Hugging Face's transformative pipeline is a case in point, which has remarkably simplified the process of model loading, tokenization, and inference. This innovative interface empowers users to access and deploy state-of-the-art models across a broad array of Natural Language Processing tasks without getting entangled in the intricacies of these processes. In the spirit of computational efficiency, we adopted the Bits and Bytes library in conjunction with the Hugging Face pipelines. This symbiotic integration optimized the GPU usage during the model loading phase, thereby ensuring a streamlined and resource-efficient inference process.

Nonetheless, it is worth noting that the inference phase of a large language model can be computationally intensive if not approached with the right strategy. In this regard, the Quantized Lottery Ticket Hypothesis configuration (QloRA) that we adopted for model fine-tuning proved to be a significant asset. It allowed us to conduct inference on consumer-grade GPUs, striking a balance between model performance and computational resources. Further enhancing the efficiency of our inference process, we adopted the batch inference methodology, a clear departure from traditional methods. This method is inherently efficient, as it allows for multiple instances to be processed concurrently, thereby significantly improving throughput.

Our testing set was meticulously structured to reflect a diverse array of question types, including general knowledge questions, queries extracted from the training dataset, hypothetical questions (for instance, explaining a fabricated legal case), and questions drawn from unseen data. Each category contained approximately 15 to 20 questions, providing a well-rounded evaluation environment for our model.

Additionally, our training set was dynamically adapted in response to the

iterative steps taken during the training phase. This dynamic adjustment enabled our model to progressively learn and adapt to its tasks, ensuring the model's performance continued to improve across successive iterations. The focus here is not only on the performance of the model but also on the fine-grained understanding of its strengths and weaknesses to guide future research in the domain of Large Language Models.

# 6 RESULTS AND BENCHMARKING

The success of an AI model, especially in complex and nuanced fields such as legal language understanding, is evaluated based on several key performance indicators. Our focus primarily lies in the quality of the answers generated by the model, specifically evaluating the soundness of reasoning, the structure of the generated responses, and their relevance to the input queries. In the evaluation phase of our research, we benchmarked our fine-tuned Falcon-7B-instruct model against state-of-the-art Large Language Models (LLMs), including GPT-3.5-turbo and GPT-4. By employing GPT-4 as the evaluator of the results produced by the inferenced models, we ensured an unbiased assessment of our model's performance.

The analysis of the results revealed that our fine-tuned Falcon-7B-instruct model outperformed the GPT-3.5-turbo in several critical aspects, such as reasoning, interpretation, analysis of impact, and coherence. The results also showed a clear enhancement in the clarity of the responses generated by our model, underscoring the efficacy of our fine-tuning approach. When benchmarked against GPT-4, our model's performance was notably robust. The generated responses showcased a deep and direct understanding of the case's impact on law enforcement and the broader societal implications, signifying an advanced level of comprehension and reasoning not often seen in AI systems.

Furthermore, the answers inferred from our fine-tuned Falcon-7B-instruct model consistently scored well in our predefined performance metrics. This successful benchmarking testifies to the potential of fine-tuning Large Language Models for specific tasks, offering a promising direction for future research in this field. By augmenting foundational models with domain-specific knowledge, we can harness the power of AI to navigate even the most complex legal terrains effectively.

# 7 FUTURE WORK AND CONCLUSION

This study has been an exploration of optimizing the data absorption capacities of Large Language Models (LLMs) within the legal context. Utilizing synthetic data generated from leading Large Foundation Models (LFMs) and employing an instruction tuning format, we have successfully demonstrated the efficacy of our method by creating a comprehensive and effective instruction set of 3300 prompts.

Through comparative analysis against top-tier models, our findings consistently underscored the superior performance of our fine-tuned models in several aspects, such as the ability to reason, clarity of context interpretation, and overall analysis of the presented queries. The methodology employed not only benefits downstream tasks by providing an extensive data corpus but also facilitates a more profound contextual understanding by summarizing detailed legal contexts. The inference results, particularly given the size of our instruction set, were commendable, further corroborating the validity and effectiveness of our approach.

Looking ahead, there is ample opportunity for further advancements. The model can be enhanced by diversifying and expanding the range of legal cases included in the dataset. The richness and breadth of India's legal repository can be better harnessed by incorporating a broader spectrum of legal cases, including those documented in Hindi. This approach not only enriches the training corpus but also offers a more representative and comprehensive benchmarking dataset for testing LLMs in the Indian context. Given the predominance of US and UK-based legal datasets, we believe our research contributes meaningfully to the field by centering on the Indian legal landscape. Going forward, the extraction and analysis of valuable information from a significant number of Hindi court cases available online can add another dimension to our dataset and be an invaluable resource for training and benchmarking.

In conclusion, the exciting potential and promising results of our study hint at a future where AI can deftly navigate the nuanced intricacies of the legal field, leading to more effective and accessible legal technology tools.