Google - Unlock Global Communication with Gemma

Ananya Varma Mudunuri, Sai Sahithi Bethapudi, Sandeep Reddy Potula,

Shreya Chilumukuru, Sowmya Neela, Vinay Bhati

Department of Applied Data Science, San Jose State University

Data 255: Deep Learning Technologies

Dr. Simon Shim

Dec 9, 2024

Abstract

Large language models have made it possible for a state-of-the-art, unprecedented result in machine translations. However, low resource languages like Hindi still created certain obstacles due to highly noisy volumes or sometimes linguistic complexities within these datasets. In this paper, we fine-tune the big and powerful Gemma 2 model on large-scale Hindi to English and English to Hindi machine translation. This uses the IITB English-Hindi Dataset, and a subset of opus100 for English-Hindi in overcoming some of the important obstacles related to syntax differences, idiomatic expressions, and semantic ambiguity. Using the latest fine-tuning mechanisms, Supervised Fine-Tuning and LoRA adds more translation accuracy, fluency, and cultural fidelity. The fine tune has the potential to take away translation gaps and bring about effective communication in education and media, added to general global commerce. This makes the work underscore the highly inclusive development of AI models and provides an accessible solution that many Hindi speakers will find scalable amidst the global digital ecosystem of today.

1. Introduction

1.1. Project Background and Execute Summary

The large language model evolution has significantly advanced the application of natural language processing, especially in machine translation. While these models have achieved outstanding performance in very high-resource languages like English, Spanish, and Chinese, low-resource languages like Hindi remain underrepresented. Hindi is spoken by more than 600 million people and has a unique challenge in machine translation due to the lack of high-quality datasets and complex linguistic structure. Translation models of today are usually inadequate in capturing nuances in Hindi and, therefore, syntactically, idiomatically, and culturally represent an underrepresentation of Hindi. In effect, this prevents good communication and has implications for a decrease in opportunities regarding educational and global business issues as well as international collaborations in Hindi.

Machine translation for Hindi has some peculiar difficulties. Unlike English, which generally uses the sentence structure SVO, Hindi uses the word order SOV, making grammatical alignment improbable. In addition, Hindi contains idiomatic expressions and culturally specific vocabulary that are usually denaturized in their meaning in a literal translation. Furthermore, morphological complexity reinforces the complexity of the language in translation since, as a rule, words have to change their forms according to tense, gender, number, and case. Besides, semantic ambiguity-when words take different meanings in different contexts-only adds more complexity to their accurate translations. Such challenges will have to be overcome with sophisticated models of translation that consider those linguistic complications while preserving fluency and readability.

More recently, another family of LLMs in particular, second-generation versions of

Gemma 2, offers an encouraging framework on which to ground these and other limitations.

Scaling up to 27 billion parameters, these models incorporate very advanced architectural innovations such as Rotary Positioning Embeddings and GeGLU non-linearity. That means models are good to go on many multilingual tasks and can generalize well on low-resource languages like Hindi. The project incorporates two datasets for fine-tuning the model on Hindi-to-English and English-to-Hindi translation: the IITB English-Hindi Dataset and the opus100 English-Hindi subset. The IITB dataset provides a comprehensive parallel corpus containing millions of good-quality sentence pairs, and the opus100 dataset contributes variety by offering their parallel translations from a wide spectrum of domains. These could form the robust training bed that this model needs in order to learn the nuances within Hindi and its alignment with the English counterpart. Fine-tuning has improved performance for state-of-the-art techniques like RLHF, RAG, and LoRA. With the fine-tuned model overcoming some of the linguistic challenges that include syntax alignment, idiomatic expression translation, and semantic disambiguation, translations will be accurate, fluent, and culturally authentic. This work hopes to help in the bridging of the translation gap from English to Hindi, contributing toward better communication in education, media, and international commerce, among other areas. This helps in making AI-driven machine translation more inclusive and accessible for the Hindi-speaking population, addressing the failures of current models, and extending global communication.

1.2. Project Requirements

1.2.1 The Software Requirements

Components	Details

Operating System	compatible with CUDA for GPU acceleration.
Programming Language	Python 3.8 or later
Libraries and Frameworks	Hugging Face Transformers Datasets Accelerate PEFT PyTorch BLEU, ROUGE
Development Tools	Jupyter Notebook
CUDA Toolkit	Compatible version for the Lambda instance's GPUs
Data Storage	CSV format for datasets; local storage and HPC lab resources used for large datasets
Hugging Face Hub	Integrated for storing fine-tuned models

1.2.1 The Hardware Requirements

Components	Details
Processor	Used HPC lab machines
Memory (RAM)	HPC machines with 128 GB of RAM for handling large datasets
Storage	Local storage of 1 TB SSD on Lambda instances and HPC lab resources for dataset/model storage
GPU	Lambda Cloud GPU instances with NVIDIA A100 GPUs (40 GB VRAM) for fast training
Network	HPC network for efficient data transfer and model training
Cloud Resources	Lambda instances utilized for high-performance distributed training
HPC Lab Access	Accessed high-performance computing (HPC) lab for additional compute resources

1.3. Literature Survey

Recent development in LLMs has seen a fair deal of participation by the family of Gemma models, and studies have been conducted with a wide range of topics on their capability and applications. Mo et al. (2024) discussed how the pre-trained Gemma-7B model was fine-tuned to conduct sentiment analysis on financial news headlines with the help of the FinancialPhraseBank dataset. Their research showed the crucial nature of correct sentiment classification-positive, neutral, and negative-in order to find understanding from market trends and investor behavior. After rigorous preprocessing with tokenization and text augmentation, this model outperforms established models like BERT and LLAMA in precision, recall, and F1-score, respectively. The present review underlined the practical application of Gemma-7B in supporting financial decisions of market actors.

The Gemma Team in 2024 introduced the Gemma family of lightweight, open-source models, ranging from 2B to 7B parameter ranges. These models represent the state-of-the-art transformer-based architectures combined with advanced training methods, including reinforcement learning with human feedback, and achieve state-of-the-art results for tasks measured by benchmarks like MMLU and GSM8K. Focus was placed on the scalability and performance of systems, along with responsible AI.

The follow-up extension introduced further innovations, such as knowledge distillation and grouped-query attention, improving the reasoning and coding tasks with reduced computational costs for Gemma 2 models ranging from 2B to 27B parameters. This work underlined the potential of Gemma 2 in democratizing AI by providing scalable solutions to real-world problems. Johansson et al. put efforts into linguistic inclusivity of LLMs; in their work, fine-tuning of Gemma 2 in low-resource languages has been done using domain-specific datasets and retrieval-augmented generation techniques. The importance of keeping cultural nuances, improvement of translation accuracy has therefore called for broader representation in AI development of those very underrepresented languages.

In this vein, Kiulian et al. (2024) have pursued the adaptation of open-source models such as Gemma and Mistral for the Ukrainian language using the Ukrainian Knowledge and Instruction Dataset. Their work was dedicated to problems concerning linguistic and cultural biases, underlining several challenges regarding code-switching and grammatical coherence and giving a number of directions for further improvement of data and methodologies. Further fine-tuning methodologies were studied by Setiawan (2024), who used Quantized Low-Rank Adaptation to reduce the cost of memory and computation with minimal degradation in model performance. The paper gave a friendly methodology for fine-tuning large models on mediocre hardware and showed how efficiently this could be done. In the multilingual direction, Lai et al. (2024) developed a pivot-based multilingual NMT methodology using Gemma-2-9B with fine-tuning methods such as LoRA.

This approach showed scalable translation improvements for low-resource languages with practical implications in many domains. They finally proposed the SIFT algorithm to optimize test-time fine-tuning of LLMs by incorporating active learning and retrieval-based methods. The SIFT algorithm minimized response uncertainty, improving performance efficiency by overcoming the inherent redundancy challenges of Nearest Neighbor retrieval. Validated on the Pile dataset, the algorithm always outperformed the state-of-the-art methods. Adaptive SIFT is an adaptive extension that further improves computational efficiency, providing proportional gains when fine tuning for diverse applications.

Recent studies have focused on fine-tuning and applying the Gemma family of LLMs to various domains. Mo et al. (2024) researched the fine-tuning of the Gemma-7B model using the FinancialPhraseBank dataset for sentiment analysis of financial news headlines. This work has underlined the importance of sentiment classification, such as positive, neutral, and negative, in drawing valuable insights into market trends and investor behavior. The fine-tuned model outperformed benchmarks like BERT and Llama in precision, recall, and F1-scores by using rigorous preprocessing techniques such as tokenization and text augmentation. This has shown the potential of Gemma-7B to enhance financial decision-making and market analysis.

The work by the Gemma Team presented lightweight, open-source models of the Gemma family, weighing between 2B and 7B parameters, featuring great generalization ability. These models were base transformer-based architectures combined with fancier training methods such as reinforcement learning with human feedback that achieved state-of the-art performance on a wide selection of tasks, performing brilliantly in benchmarks such as MMUL and GSM8K: This work brought into focus its scalability, performance, and application of responsible AI. Following this, Ji and Kumar discuss architectural advancement in Gemma model 2, which employed Rotary Positioning Embedding and GeGLU, a non-linearity with scaling up more than 27B parameters (2024). Those paved the way for enhanced conversational performance, ranking ahead of much larger competitors on the leaderboard in LMSYS Chatbot Arena, highlighting adaptability across platforms and hardware configurations.

The authors, Chung et al., went further to carry out studies on instruction tuning using a model like PaLM, T5, and U-PaLM models fine-tuned on a series of 1,800 different tasks framed as instructions and subsequently improved their benchmark result performance, including MMLU and RealToxicityPrompts, therefore enhancing this method to generally fit other tasks.

Ukarapol et al. (2024) further proposed a contrastive fine-tuning method to enhance the textual embedding capability of small models like Gemma, Phi-2, and MiniCPM. Using LoRA and InfoNCE loss, they attained performance gains of up to 56.33% on benchmarks, further showcasing the potential of lightweight models in cost-effective, high-precision embedding tasks. Adithya S K (2024) practically gave, step by step, some fine tuning of Gemma-7B on key steps, such as model checking of compatibility with the GPU and dataset preparation, all the way through integration into Hugging Face. It was that kind of tutorial that practitioners needed so badly when trying to adapt Gemma models to different applications. These collectively support the continued evolution of the Gemma family of models, underlining scalability, adaptability, and application potential across a wide variety of tasks and domains.

1.4. Related work

Recent development in large language models has resulted in improving the performance of most of the NLP tasks, including machine translation. However, low-resource languages like Hindi pose serious challenges regarding the availability of high-quality datasets and under-representation in the training corpus of LLMs. Consequently, addressing such challenges becomes an important factor for linguistic diversity and inclusion in AI technologies.

Of all these issues, the most promising framework seems to be constituted by the Gemma family of models, particularly Gemma 2. Gemma 2 models, from 2 billion to 27 billion parameters, involve several architectural novelties such as interleaving of local-global attentions and group-query attention that advance performance on many NLP tasks.These models have shown huge potential in language understanding, generation, and reasoning; hence, they are quite fit to be applied for MT with low-resource languages.

Fine-tuning LLMs for specific language pairs has been explored in different works. For instance, Lai et al. (2024) used the model Gemma-2-9B to improve the quality of translation between Chinese and Malay, using English as the pivot language to build a scalable Multilingual Neural Machine Translation system.This methodology transferred the knowledge in high-resource languages to low-resource languages by increasing the accuracy and inclusiveness of translation.Regarding Indian languages, translation tasks are done to be adapted by LLM in some way. Bhasin 2024 articulated the process of fine tuning required for the model Gemma-2-2B for English-to-Hindi translation using the supportive platforms like MonsterAPIs that smoothen the process by fine tuning.

MONSTERAPI BLOG Added to that, Indic Gemma models developed by Theja (2024) include instruction-tuning on nine Indian languages, including Hindi; therefore, this improves its performance through understanding and text generation upon receiving the above prompt. These studies point out the possibility of fine-tuning LLMs like Gemma 2 for better translation quality in low-resource languages. Similarly, fine-tuning with more advanced architecture will go a long way in mitigating the linguistic challenges that are inherent in languages like Hindi and hence enhance the inclusiveness and effectiveness of AI-driven translation systems.

The literature review really highlights the phenomenal progress taken by large language models regarding machine translation and how the models have been adapted to low-resource languages like Hindi. Research into the Gemma family of models, particularly the Gemma 2 models, underlines their architectural improvements and fine-tuning methods for increasing their performance on different NLP tasks. Techniques such as pivot-based multilingual translation, instruction tuning, and domain-specific fine-tuning were applied to show adaptability and scalability of the models.

These are from Indic Gemma, which provides translations of Indian languages to the adaptation of Gemma-2-2B for Hindi. In fact, growing attention is being devoted to linguistic inclusivity and respect for cultural sensitivity. Approaches are taken for syntactical adjustment, idiomatic translation, and semantic ambiguity, the major challenges, for further accurate and context-aware translations.

LoRA may be preferred over QLoRA in highly sensitive applications that demand a lot of precision, like translation tasks, since they involve nuanced understanding and require high-quality outputs. On the other hand, QLoRA reduces the much-needed memory through 4-bit quantization, which could introduce slight inaccuracies and probably degrade performance on tasks involving complex grammatical structures, idiomatic expressions, or multilingual nuances. In contrast, LoRA works without quantization and maintains full precision of the base model parameters while allowing for efficient and lightweight fine-tuning by updating only a small set of low-rank matrices. This ensures that the model retains its original quality and performance, making it a better choice for mid-sized models like Gemma 2-9B and Gemma 2-27B, where memory and computational requirements are already manageable. Moreover, LoRA has a simpler training setup and is extensively validated on a wide variety of tasks, which renders it more dependable and precise for applications that cannot tolerate even slight degradations in performance.

Put together, these works form a very strong backbone for more detailed studies of fine-tuning LLMs like Gemma 2 for Hindi-to-English translation. The present work contributes toward this larger objective of facilitating appropriate and inclusive AI-driven communication across languages by identifying some gaps in the available datasets and deploying state-of-the-art methodologies.

2. Data Exploration and Processing

2.1. The Dataset

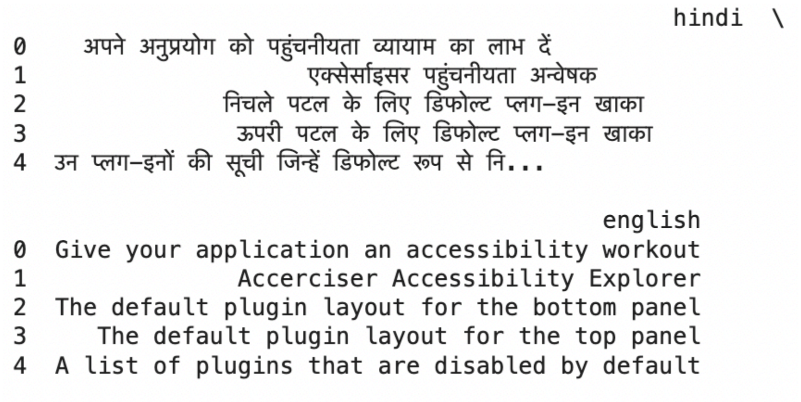
In this project, two major datasets are combined , the IIT Bombay dataset and the OPUS-100 dataset. This resulted in a big parallel corpus that could be used for machine translation between Hindi and English. Both these datasets provide high-quality, aligned sentence pairs in Hindi and English, which can be used to train robust machine translation models.

2.1.1. IIT Bombay Dataset

The IIT Bombay dataset is an already established parallel corpus consisting of sentence pairs in Hindi and their respective translations in English. It forms a part of the IIT Bombay corpus that has been developed with the objective of assisting natural language processing activities, more specifically machine translation. It includes a wide variety of sentences from technical terms to colloquial ones. The diversity in sentence types ensures that the dataset is suitable for training models capable of handling a wide range of linguistic structures and contexts.

The dataset is structured into paired sentences, with each sentence in Hindi having a translation in English. This presents a good case for supervised learning in machine translation, since a model learns the mapping from sentences of one language to its translation in another. Some of the sentences are thematically varied, offering wider generalization from machine translation models and facilitating the model's performance within diverse domains.The image below shows the raw data of IIT bombay dataset.

Figure 1. IIT bombay dataset



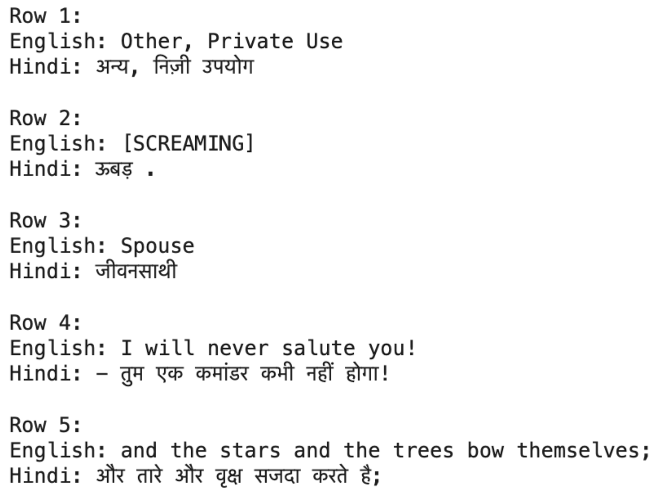
```
                                              hindi  \
0    अपने अनुप्रयोग को पहुंचनीयता व्यायाम का लाभ दें
1                    एक्सेसर्साइसर पहुंचनीयता अन्वेषक
2               निचले पटल के लिए डिफोल्ट प्लग-इन खाका
3                ऊपरी पटल के लिए डिफोल्ट प्लग-इन खाका
4  उन प्लग-इनों की सूची जिन्हें डिफोल्ट रूप से नि...

                                         english
0  Give your application an accessibility workout
1                Accerciser Accessibility Explorer
2  The default plugin layout for the bottom panel
3     The default plugin layout for the top panel
4  A list of plugins that are disabled by default
```

2.1.2. OPUS-100 Dataset

It was sourced from the OPUS Corpus-a giant, multilingual corpus sourced from various sources such as books, movies, news, and more. Hence, this OPUS-100 dataset has parallel sentences in most of the widely spoken languages of the world: Hindi and English not excluded. Also, these represent all kinds of sentences for every structural pattern, vocabularies, and all aspects of linguistics generally.

This parallel corpus is also divided into training, validation, and test sets, which are necessary for training, fine-tuning, and testing machine translation models. The training set consists of a large number of sentence pairs, providing enough data for training, while the validation and test sets provide the possibility to check the performance on unseen data, thus the generalization of the model on new inputs.The image below shows the raw data of OPUS-100 hindi-english.

Figure 2. OPUS-100 Dataset



```
Row 1:
English: Other, Private Use
Hindi: अन्य, निज़ी उपयोग

Row 2:
English: [SCREAMING]
Hindi: ऊबड़ .

Row 3:
English: Spouse
Hindi: जीवनसाथी

Row 4:
English: I will never salute you!
Hindi: – तुम एक कमांडर कभी नहीं होगा!

Row 5:
English: and the stars and the trees bow themselves;
Hindi: और तारे और वृक्ष सजदा करते है;
```

Integration of IIT Bombay and OPUS-100 Datasets It gives an enlarged parallel corpus with more diversity when the IIT Bombay dataset is integrated with the OPUS-100 dataset. While the IIT Bombay dataset provides high-quality sentence pairs with rich variation in linguistic structures, the OPUS-100 dataset adds a lot more data from diverse domains such as literature, media, and technical data. Such a fusion develops greater size and diversity of the overall corpus, which could be better utilized for training more generalized and robust machine translation models. We integrated the sentence pairs of both into one corpus with each pair having an English sentence and its translation in Hindi. Further cleaning and preprocessing were carried out for consistency in the dataset structure and format. The current dataset is very useful and forms a major resource in machine translation by providing input material to the training model requiring a vast amount of linguistic context.

In such context, the IIT Bombay and OPUS-100 combine into an exhaustive and thoroughly diverse collection of parallel Hindi-English sentence pairs, hence a robust resource bringing in a wide variety of sentence structures, linguistic context, and topics. Being one such expanded dataset, they might prove ideal for training Machine translation models that would be effectives crossing several different types or, mostly any translation tasks. The combination of the two datasets will ensure that one develops a generalized and inclusive machine translation system for both Hindi-English translations.

2.2. Data Cleaning

After the integration, cleaning was done as in steps for IIT Bombay and OPUS-100, so that the dataset may come into proper shape-a requirement for analysis or training.

Among the steps of preprocessing, one was to remove duplicates. Duplicate sentence pairs were identified and removed to avoid redundancy and bias in the dataset. Each sentence pair needed to occur only once to keep the integrity of the dataset and make sure that any model performance was not biased by data repetition.After that, the mixed content issues were fixed. That is, some of the sentences had both English and Hindi in one row. They were tagged because those will create problems during the learnings of the model from the data. Such types of mixed language rows were screened and removed from the dataset if necessary to keep data clean and proper alignment of Hindi-English sentence pairs.The third step of the process was dealing with missing values. Incomplete rows, either from the Hindi column or the English one, had to be handled carefully. Whenever necessary, according to the given context, the missing value was filled with an appropriate default value or the entire row was removed to maintain consistency in the dataset.

Lastly, the dataset had its text normalized. The effect is a standardization that gives rise to texts that are then good for analysis. This rids one of extra unwanted characters and extra spaces for all formats, leaving one with cleansed, properly formatted sentences after this process of normalization into well-structured and standardized pieces across the dataset that becomes workable and trainworthy towards building the models.

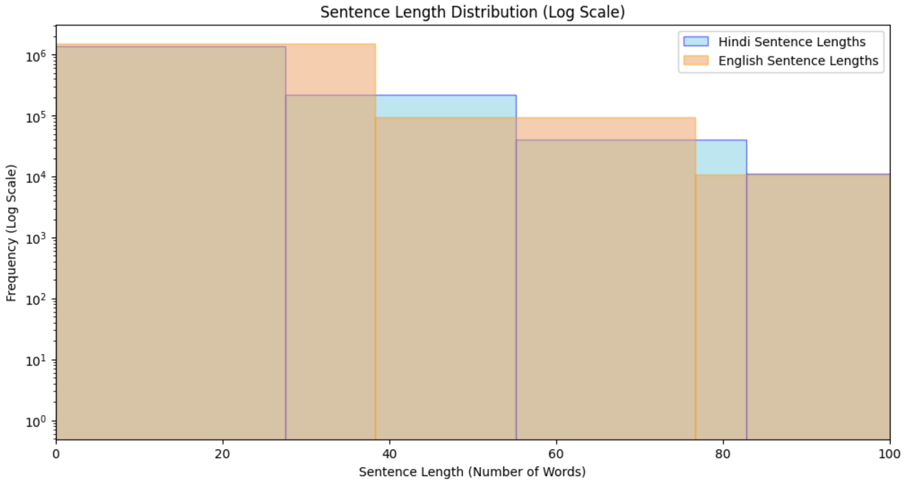
These steps of cleaning the data were major milestones in preparing the dataset for further exploration and ensuring the quality and consistency of the data that are used to train machine translation models.

3. Data Analysis

Cleaning and preparing the data were followed by in-depth analysis to find some key patterns and insights that might be hidden in the dataset. First, the quantification of sentence length for both Hindi and English sentences was carried out. The length of each sentence was determined based on the number of words in each sentence, allowing an in-depth investigation into the distribution of sentence lengths across the two languages.

A histogram displaying the distribution will be plotted, both for Hindi and for English. Most sentences are short in length, though a few are far longer, resulting in some outlying sentences. So, logarithmic scaling is applied on this histogram plot to better display the underlying pattern in these data. It increases visibility by compressing the long-tail distribution formed because of very few quite long sentences and thus gives an intuition that is more interpretable from the data.

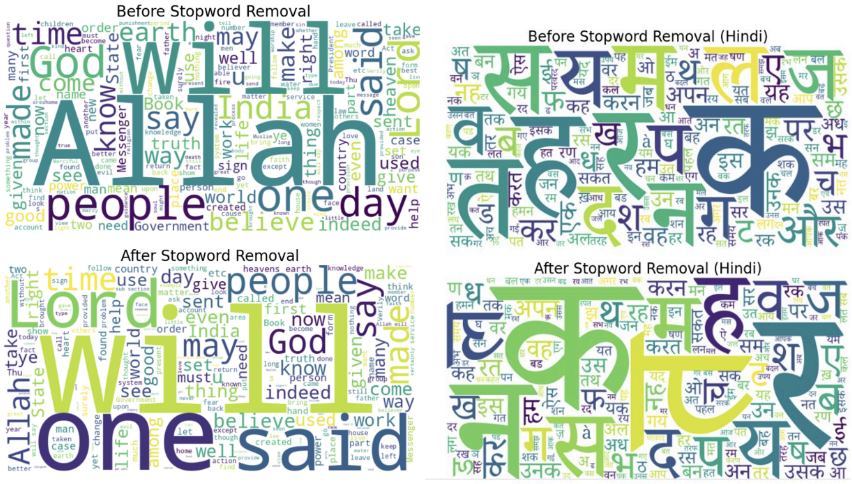
Figure 3: Sentence Length Distribution



Moreover, sentence length percentiles were calculated to further reduce the effect of outliers. Concretely, the 95th percentile was chosen to filter out sentences that would be either too short or much too long. This removes extreme values to ensure that model training cannot be biased by the least or most extreme values on sentence lengths.

After classification, word frequency analysis was performed on both Hindi and English sentences. To focus on the most meaningful content words, common stop words were removed for both languages. First, word clouds were created to show the most frequent words in both languages. This was an interesting visual way to see some of the common terms present in the dataset and helped identify key vocabulary used in sentence pairs.

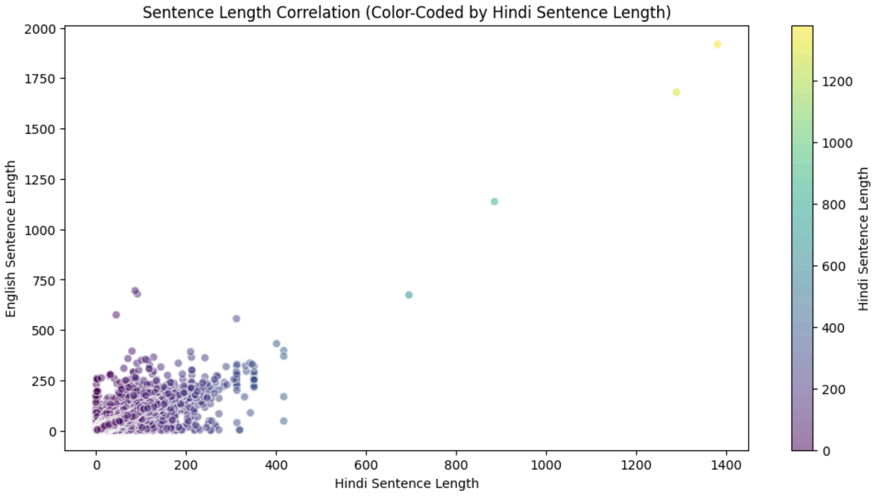
Figure 4: Word Frequency Analysis



A bar plot was also generated to show the top 10 most frequent words in both languages. This plot makes it very easy to compare word usage across Hindi and English, pointing out the differences in word frequency and vocabulary between the two languages.

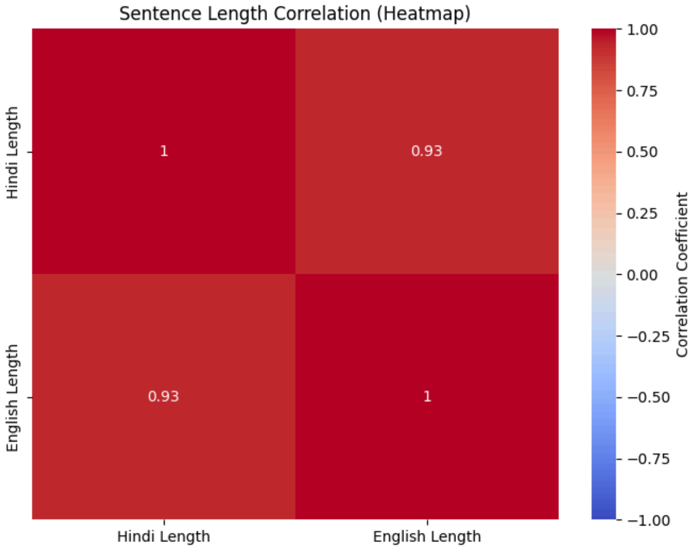
Further, to investigate the relationship between sentence lengths in Hindi and English, a scatter plot was used to examine their correlation. The plot showed a strong positive correlation between the two languages; longer sentences in Hindi mostly corresponded to longer sentences in English. This may indicate that sentence structure in both languages tends to align in terms of length.

Figure 5: Sentence Length Correlation [Scatter Plot]



This followed an attempt to provide a complementary view of the same concept-an English-Hindi sentence heatmap. The results from that scatter plot were confirmed via the heatmap: there's very strong positive correlation. We confirm here, at the minimum structurally, that for those two languages, the lengths of sentences are closely proportional to each other and will further support our conclusion previously made that this dataset indeed looks well-aligned for machine translations.

Figure 6: Sentence Length Correlation [Heatmap]



These analytical steps were of great benefit in giving insight into the structure and content of the data, thereby making it adequate for the next stages of model training and evaluation.

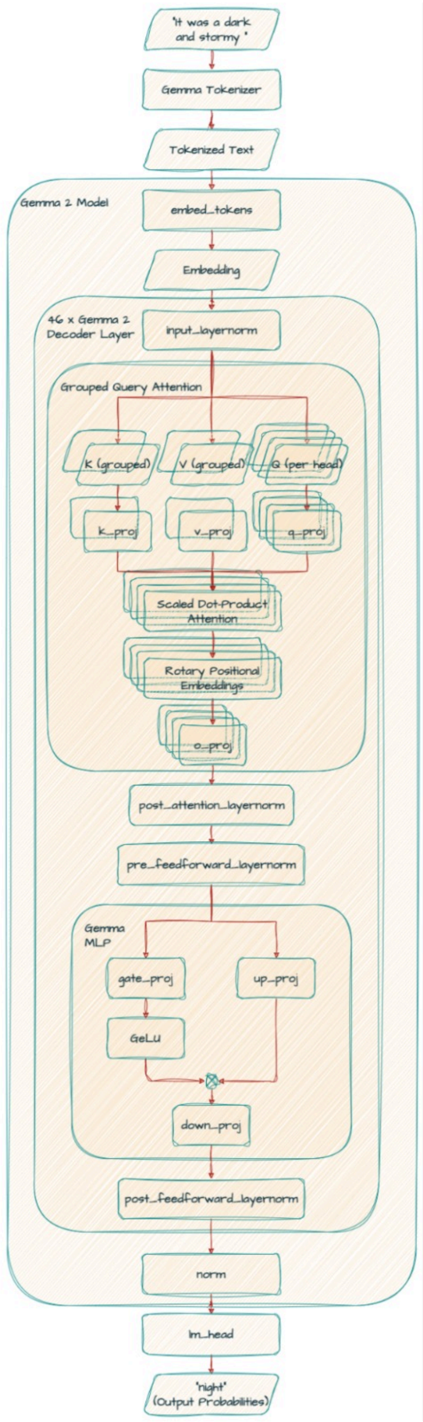
4. Model Implementation

4.1. Gemma 2

Architecture of the Gemma 2 model is a decoder-only transformer that considers both causal language modeling and multilingual tasks. Starting from the input text, which was tokenized into token IDs with Gemma Tokenizer, these tokens are fed into the embed_tokens layer where the dense embeddings can take place. These dense embeddings flow through 46 stacked decoder layers containing GQA among Grouped Query Attention and FFN. GQA improves computational efficiency during the construction of per-head queries by grouping keys and values together, then applies Scaled Dot-Product Attention enriched with RoPE. The latter encodes position relationships in a sequence. FFN uses gate, up, and down projections with GeLU for learning robust representations. It implements layer normalization using RMSNorm at several stages of outputs for improved training convergence.

The final layers consist of the normalization step, norm, and a linear projection head, lm_head, which provides the output probabilities of the next token, based on the embeddings the model has been processing. Architectural features such as GQA for efficient attention, RoPE for long-context understanding, and RMSNorm for stability in training enable Gemma 2 to handle complex language tasks with ease. This architecture is especially effective for tasks that require long-context understanding, such as document-level translation and summarization, while still allowing computational scalability.

Figure 7: Gemma 2 flowchart



Reference image from:

https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/

The Gemma 2 models are an advanced evolution from the original Gemma class of models, with architectural novelties and extensions to their functionality which allows them to support not only multilingual but also multi-domain tasks. All of these are designed with various use cases in mind; hence, three model variants exist for each Gemma 2 model, optimized towards different applications.

The variant of Gemma 2-2B has a total of about 2 billion parameters, hence lightweight and efficient to an extent, with most computations feasible even on moderately powered hardware. Therefore, this model best suits all the usual applications of NLP, such as text classification, sentiment analysis, and simple translation jobs. This compactness results in speedy deployment and low hardware demand.

The model Gemma 2-9B has more balance between performance and resource demands; it has about 9 billion parameters. This realizes high-quality outputs of the most challenging text generation as well as multilingual applications, while keeping computational costs manageable. This variant is a lot more effective in these tasks requiring deeper insight into context and structure, such as the generation of dialogues or advanced translations.

The Gemma 2-27B model is approximately 27 billion-parameter pre-trained models that show impressive capacity in nuanced language understanding and translation tasks. This large model fits research-level applications of zero-shot and few-shot learning and tasks that require complex or domain-specific content. It performs well on intricate linguistic patterns and long-context situations, making it a suitable model for high-quality document translations.

The Gemma 2 series is a new series, bringing a lot of architectural changes by improving performance and efficiency. One of the most striking things may be that the traditional multi-head attention mechanism was replaced by Grouped-Query Attention. GQA enhances computation efficiency by reducing redundancy and extra overhead, which is of particular help in longer sequences on translation and summarization tasks.

Another important feature is the Rotary Position Embeddings-RoPE-that further enhance the ability of the model to comprehend the positional relationship among tokens, especially for tasks with very long contexts, such as document-level translations, coherence regarding the structure of text should be maintained. RoPE ensures the model captures positional dependencies with no increase in computation.

Other innovations in the Gemma 2 models are the RMSNorm alternative to the traditional LayerNorm. RMSNorm stabilizes training for larger models, giving faster convergence that translates into better generalization. The models also contain Logit Soft-Capping to avoid extreme logits during training. This feature contributes to enhancing the stability and reliability of the model's predictions, ensuring one level from different tasks.

Apart from this, it is the number of supported languages and all the rest discussed that make models like Gemma 2 robust and powerful at broad NLP tasks, most especially on a multilingual or multi-domain basis.

4.1.1 Fine Tuning Gemma

Fine-tuning the Gemma 2 models requires the latest methodologies in changing the base models to create versions fine-tuned for actual tasks such as language translation; LoRA, or Low-Rank Adaptation of SFT, or Supervised Fine-Tuning, comes into consideration. These put up means for task-specific training efficiently while optimizing computational resource utilization but keeping the robustness of the base models intact.

4.1.2. LoRA

LoRA is a low-rank adaptation approach for parameter-efficient fine-tuning, which adds to the model small, low-rank matrices specific for a task. Instead, LoRA does not update the model weights but only these extra matrices. This drastically reduces both memory and computational requirements and makes fine-tuning very big models such as Gemma 2-9B and Gemma 2-27B on much weaker hardware possible. LoRA has the added advantage of being modular in that one can add or remove the task-specific weights at will without touching the base model. This adaptability allows for the fast fine tuning across several tasks.

Key Advantages:

- Parameter Efficiency: Fine-tunes only a small percentage (typically <1%) of the model's parameters.
- Modularity: LoRA adapters can be added or removed as needed, enabling easy task switching.
- Resource Optimization: Allows fine-tuning of large models on hardware with limited memory.

4.1.3. Supervised Fine-Tuning

Supervised Fine-Tuning works by closely aligning the capabilities of the model to the exact task requirements. In that sense, it is able to train a model using labeled datasets-e.g., parallel corpora for translation-on how mappings between input and output sequences take place. In that line, the current method has the benefit of enhancing a model for domain-specific subtlety in language expressions, idiomatics, and context directly. Supervised Fine-Tuning is dependent on preprocessing steps involving tokenization of input and target sequences, truncating or padding them to uniform lengths, and configuration of training arguments that would work best for this optimization.

Key Advantages:

- Direct Task Alignment: Improves the model's ability to handle domain-specific or language-specific tasks.
- Improved Context Handling: Fine-tuning on parallel datasets enhances the model's understanding of grammatical structures, idiomatic expressions, and context-specific translations.

LoRA and SFT together ensure that the fine-tuning is accomplished with maximum efficiency. LoRA enables parameter-efficient updates, while SFT ensures that the model learns the objectives of a task with high accuracy. These methods allow the models of Gemma 2 to realize very good performance on various applications, ranging from standard translation tasks to complex multilingual and domain-specific challenges. This combination of computational efficiency with task-specific adaptability makes these fine-tuning methodologies quite versatile for real-world applications in the case of the Gemma 2 series.

5. Model comparison

5.1. GEMMA 2 - 2b

This is designed to tackle some of the challenges in translating languages that have very different structures from one another, like Hindi and English. The improvements in the model are focused on enriching the contextual understanding, integrated with advanced natural language processing. It focuses on enhancing multilingual translation accuracy, integrating the latest language models with state-of-the-art contextual understanding. This enhancement guarantees that communication across diverse languages will go through without any hitch, hence allowing users to express themselves. It bridges linguistic gaps by providing accurate translations of both literal and nuanced meanings.

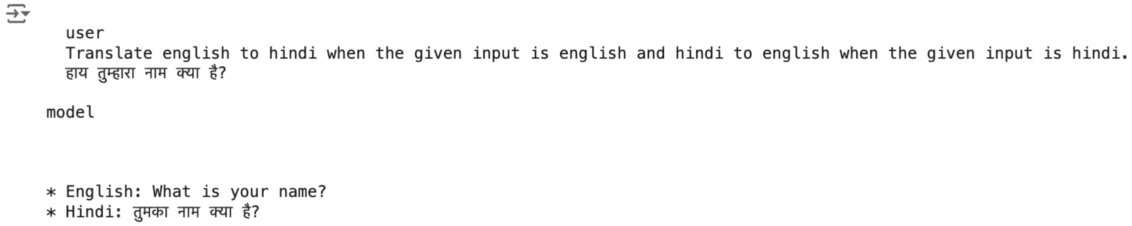
Grammatical, syntactical, and idiomatic differences between English and Hindi are immense. GEMMA 2 - 2b ensures that translation is accurate in terms of the meaning of the sentences and grammatical structure of the word order of the target language.

Key Features are

- Leverages deep learning algorithms to capture nuances in meaning.
- Utilizes bilingual dictionaries and parallel corpora for training.
- Handles ambiguous or polysemous words by analyzing context; for example, translating "light" to "रोशनी" (light) or "हल्का" (not heavy) depending on the usage.

This model reduces errors in translation and provides a natural flow to Hindi translations, improving readability and comprehension.

Figure 8: Gemma 2 - 2b example



```
  user
  Translate english to hindi when the given input is english and hindi to english when the given input is hindi.
  हाय तुम्हारा नाम क्या है?

model


* English: What is your name?
* Hindi: तुमका नाम क्या है?
```

5.2. GEMMA 2-9b

GEMMA 2-9b is all about cultural and regional adaptation in language translation. By incorporating local expressions and idioms, it will make the user experience much better by making translations more relatable and impactful. This enables Google users to understand and connect with content in their preferred language without losing any nuances of culture. This model ensures that translations are not only linguistically accurate but also culturally and contextually appropriate.

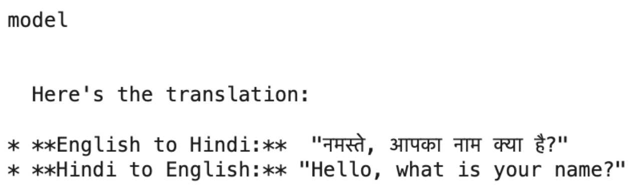
Key Features are :

- Incorporating Idioms and Colloquialisms: GEMMA 2 - 9b identifies phrases that cannot be literally translated and substitutes them with culturally equivalent expressions.
- Formal vs. Informal Tone: Hindi has distinct formal (आप) and informal (तुम) forms of addressing people, and this model adapts translations to suit the context.
- Regional Sensitivity: This is sensitivity to regional variations and preferences in Hindi and performs the translation accordingly.

In english-to-hindi translation, it exhibits cultural metaphor adaptation, for example, "hit the road" will be more appropriately translated as "यात्रा शुरू करना" (to start a journey), rather than the literal "सड़क पर मारो."

Changing formality based on the audience, such as "How are you?" → "आप कैसे हैं?"-formal or "तुम कैसे हो?"-informal.

Figure 9: Gemma 2 - 9b example



```
model

  Here's the translation:

* **English to Hindi:**  "नमस्ते, आपका नाम क्या है?"
* **Hindi to English:** "Hello, what is your name?"
```

5.3. GEMMA 2-27b

GEMMA 2-27b is designed with scalability and efficiency in mind, to process vast amounts of data across many languages. Advanced machine learning techniques help in getting faster and more accurate translations, even in less spoken languages. Such scalability can enable Google to serve a global audience with much greater effectiveness, crossing language barriers on a more massive scale. GEMMA 2 - 27b focuses on scalability and handling translation demands in complex, large-scale scenarios with high efficiency and accuracy maintained.

Key Features are:

- Handling Low-Resource Languages: It supports languages with fewer resources or limited data, for instance, regional Hindi dialects.
- Massive Dataset Integration: This model is optimized to process and learn from multilingual corpora, hence, turning out to be effective in diverse linguistic settings.
- Real-Time Translation: GEMMA 2 - 27b is designed for speed, enabling real-time translation for large-scale applications such as live captioning or instant messaging.

In English-to-Hindi, ensures smooth translation of rare phrases or dialect-specific terms in Hindi. Supporting high-traffic environments, such as live chat services or subtitles for streaming content.

Each of the GEMMA models-2 - 2b, 2 - 9b, and 2 - 27b-serves a different purpose in achieving the best possible translation of English to Hindi. While 2 - 2b ensures linguistic precision, 2 - 9b ensures cultural relevance, and 2 - 27b provides the much-needed scalability and efficiency for large-scale use. Together, they form a complete solution to bridge the gap in languages and facilitate seamless communication among people across the globe.

6. Evaluation Metrics

One of the key metrics used in measuring the quality of translations provided by the

Gemma 2 model in this project is the BLEU score. It calculates a weighted geometric mean of n-gram precisions as a way to measure the similarity between the machine-generated translation and the human reference translations. It also includes a brevity penalty to discourage overly short translations so that the output is full and contextually accurate.

Precisions represent the ratio of the n-grams generated within the candidate translation and correspond to a given fraction within the actual reference translation, observed over single words or continuations of two, three, and four words. The Brevity Penalty was a part of BLEU, which, if it notices that the generated translations are smaller than the reference, has to take care that every translation will not result in generations of sentences that originally had much longer lengths intended by context. The Length Ratio compares the total number of tokens in the generated translation against the reference, helping assess how well the model aligns with the expected translation length. Finally, the Translation Length and Reference Length indicate the number of tokens in the generated translation and human reference translation, respectively, giving a direct measure of length consistency. These evaluation techniques ensured holistic assessments of this model in performance for the project.

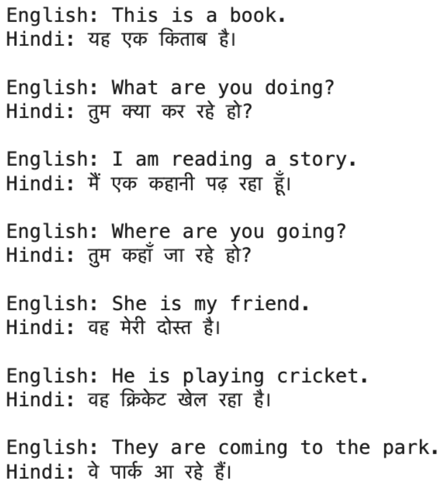
7. Conclusion

After experimenting with various fine-tuning approaches, the Gemma 2 models-2B and 9B variants-showed effective fine-tuning on Hindi-to-English translation tasks. For this fine-tuning, the combination of LoRA and SFT methodologies played a great vital role in adapting the models with much efficiency. LoRA especially enables the training of parameters really efficiently, such that a very small portion of updates of model parameters is required at minimal computational overhead without significant loss regarding the integrity of the original model. SFT ensured that in learning from parallel datasets, models stay near the translation job with contextually accurate and semantically meaningful outputs.

The BLEU score achieved for the 2B model was **0.89,** demonstrating a very strong translation performance, while the BLEU score for the 9B model was **0.88.** However, at the same time, it showed much better contextual understanding, and in some context could use formal language, which the 2B model lacked. All of this came with the expense of computational efficiency, especially considering the handling of big datasets or the calculation of extensive BLEU scores. Both 2B and 9B did a poor job with certain translations, such as translating "December" in Hindi, which gives room for improvement.

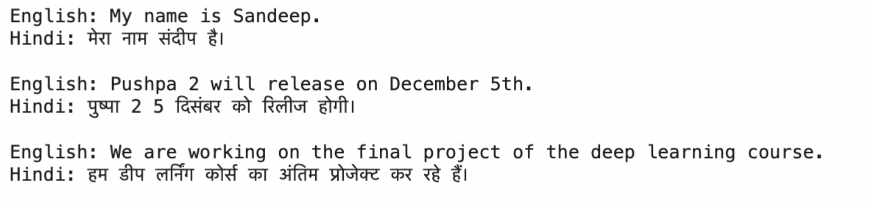
The larger model, Gemma 2-27B, has a greater capacity and holds much promise in surmounting these shortcomings. It may do very well on complex phrases and subtle translations that were difficult for smaller models. Moreover, increasing the number of training epochs may further enhance the capability of current models dealing with complex linguistic structures. Overall, LoRA and SFT proved instrumental in achieving efficient fine-tuning, while future efforts can build upon these methodologies to explore the full potential of the Gemma 2 series for high-quality language translation tasks.

Figure 10: Results with the 2B , without any formal usage



```
English: This is a book.
Hindi: यह एक किताब है।

English: What are you doing?
Hindi: तुम क्या कर रहे हो?

English: I am reading a story.
Hindi: मैं एक कहानी पढ़ रहा हूँ।

English: Where are you going?
Hindi: तुम कहाँ जा रहे हो?

English: She is my friend.
Hindi: वह मेरी दोस्त है।

English: He is playing cricket.
Hindi: वह क्रिकेट खेल रहा है।

English: They are coming to the park.
Hindi: वे पार्क आ रहे हैं।
```

Figure 11: Results with 9B , with formal usage



```
English: My name is Sandeep.
Hindi: मेरा नाम संदीप है।

English: Pushpa 2 will release on December 5th.
Hindi: पुष्पा 2 5 दिसंबर को रिलीज होगी।

English: We are working on the final project of the deep learning course.
Hindi: हम डीप लर्निंग कोर्स का अंतिम प्रोजेक्ट कर रहे हैं।
```

References

Gemma Team. (2024). Gemma 2: Improving open language models at a practical size. *arXiv*. https://arxiv.org/abs/2408.00118

Hübotter, J., Bongni, S., Hakimi, I., & Krause, A. (2024). Efficiently learning at test-time: Active fine-tuning of LLMs.*arXiv*. https://arxiv.org/pdf/2410.08020v2

Johansson, F., Xu, D., & Kim, H. (2024). Enhancing LLM inclusivity for low-resource languages. *OpenReview*. https://openreview.net/pdf?id=aNZMz6mt5L

Kiulian, L., Petrova, E., & Dmytro, V. (2024). Fine-tuning open models for linguistic inclusivity in Ukrainian. *ACL Anthology*. https://aclanthology.org/2024.unlp-1.11.pdf

Lai, Y., Liew, J., & Yi, G. X. (2024). Cross-cultural language adaptation: Fine-tuning Gemma 2 for diverse linguistic contexts. *NeurIPS 2024*. https://openreview.net/pdf?id=GCdVXIaAbe

Mo, K., Liu, W., Xu, X., Yu, C., Zou, Y., & Xia, F. (2024). Fine-tuning Gemma-7B for enhanced sentiment analysis of financial news headlines. *arXiv*. https://arxiv.org/abs/2406.13626

Setiawan, A. (2024). Fine-tuning Gemma with QLoRA. *Medium*. https://medium.com/google-developer-experts/fine-tuning-gemma-with-qlora-407e56c36026

Gemma Team. (2024). Gemma 2: Improving open language models at a practical size. *arXiv*. https://arxiv.org/abs/2408.00118

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... Wei, J. (2024). Scaling instruction-fine tuned language models. *Journal of Machine Learning Research, 25*(70), 1–53. https://www.jmlr.org/papers/volume25/23-0870/23-0870.pdf

Ukarapol, T., Lee, Z., & Xin, A. (2024). Improving text embeddings for smaller language models using contrastive fine-tuning. *arXiv*. https://arxiv.org/abs/2408.00690

Adithya S K. (2024, February 21). A beginner's guide to fine-tuning Gemma. *Medium*. https://adithyask.medium.com/a-beginners-guide-to-fine-tuning-gemma-0444d46d821c

Ji, J., & Kumar, R. (2024, August 22). Gemma explained: What's new in Gemma 2.

Google Developers Blog. https://developers.googleblog.com/en/gemma-explained-new-in-gemma-2/

Bhasin, K. (2024). Fine-tuning the Gemma-2-2B model for English-to-Hindi translation. *MonsterAPI Blog*. Retrieved from https://blog.monsterapi.ai/blogs/fine-tuning-gemma-2-2b-it-for-translation/

Gemma Team. (2024). Advanced language model-based translator for low-resource languages. *arXiv*. Retrieved from https://arxiv.org/abs/2408.00118

Lai, X., Abdul, M., & Zhang, Y. (2024). Multilingual machine translation using Gemma-2-9B: A pivot-based methodology for low-resource languages. *OpenReview*. Retrieved from https://openreview.net/pdf?id=GCdVXIaAbe

Ravi, T. (2024). Introducing Indic Gemma: Instruction-tuned models for Indian languages. *Medium*. Retrieved from https://ravidesetty.medium.com/introducing-indic-gemma-7b-2b-instruction-tuned-model -on-9-i ndian-languages-navarasa-86bc81b4a282

Setty, R. (2024). Gemma 2: Improving open language models at a practical size. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/382797528_Gemma_2_Improving_Open_Lang uage_ Models_at_a_Practical_Size