Word Contextualization Generator: Building Semantic Surroundings

Rohan Sai Nalla
rona8789@colorado.edu
MS Computational Science

Sashank Gangadharabhotla
saga8160@colorado.edu
MS Computational Science

Abstract

This paper explores the fine-tuning of language models for generating word usage examples, specifically in the field of astronomy. We evaluate the performance of three well-known language models, Gemma-2b, llama2, and GPT-2, through a thorough evaluation process. Using a dataset of 30 vocabulary words and their usage instances, we analyze the models' capacity to generate sentences that are relevant in context and linguistically coherent. The findings provide valuable insights into the suitability of each model for generating domain-specific text. They shed light on the efficacy of the models in handling both seen and unseen words. In addition, this study highlights the importance of refining methods to optimize language models for specific tasks, emphasizing the need for precise and contextually appropriate text generation in fields like astronomy.

1 Introduction

In the rapidly evolving field of NLP, the customization of LLM's for highly specific applications represents a significant need. This paper introduces a novel methodology for fine-tuning an existing LLM to generate tailored text on predetermined topics. By focusing on this aspect of model enhancement, we address a crucial gap in the current capabilities of automated text generation systems, particularly in their application to specialized domains requiring precision.

Our work presents a approach to modifying many pre-trained LLM's, enabling them to not only maintain their broad knowledge base but also to apply it in a manner that is contextually and thematically precise. The core of our methodology involves a strategic alteration of the training process, incorporating a custom-designed dataset that reflects the specific lexical and topical needs of our target applications. This dataset serves as a foundation for retraining the LLM, guiding it to generate outputs that are not just linguistically coherent but also topically faithful and lexically specific.

This paper details the steps taken in the fine-tuning process, from data preparation and model reconfiguration to testing and validation of the model's performance. Through this comprehensive approach, we demonstrate how an LLM can be effectively customized to meet specialized needs, thus expanding the potential uses of NLP technologies in fields such as legal documentation, academic research, and mainly targeted content creation. By pushing the boundaries of what LLMs can achieve in terms of topic-specific text generation, this research not only enhances the utility of existing models but also lays the groundwork for future advancements in the field of NLP.

2 Related Work

The information presented in this paper [1] was extremely important. It allowed us to develop a thorough understanding of different pre-trained language models and their unique features. This information was crucial in choosing the best LLM program to customize based on our specific needs. The paper explored the architectural differences, performance benchmarks, and suitability of each PLM for various text generation tasks. Through a comprehensive understanding of these intricacies, we were able to assess which model would be most suitable for our requirements in terms of accuracy and effectiveness in producing text tailored to specific contexts. By carefully selecting the LLM program, we were able to customize our approach and find the best fit for our objectives. This choice has provided a strong basis for refining our strategy.

The paper [5] provides a comprehensive analysis and methodology that greatly enhances the text generation capabilities of pre-trained language models, which is of great value to our research in this area. The paper discusses various advanced techniques for customizing PLMs to produce text that is tailored to specific contexts and vocabularies. This encompasses in-depth discussions on the encoding of input data, model optimization, and tailoring for specific tasks. In order to guarantee that the produced text was both relevant to the topic and grammatically accurate, we implemented a range of encoding techniques designed to maintain the meaning of the text. This step was essen-

tial in meeting our project's requirements for accuracy and depth in content generation.

This article [2] proved to be a valuable resource for our project, which involved refining a substantial language model such as Google's Gemma. It provided us with a step-by-step approach to customizing the LLM for specific tasks using efficient methods, specifically LoRA. In this controlled setting, we were able to modify a specific set of parameters within the model. This was important due to the limited computational resources available to us.

Through the implementation of the strategies discussed in the article, we successfully conserved computational power and storage. This allowed us to allocate our resources towards enhancing the model's performance for specific tasks. The combination of PEFT techniques and the transformers library has made the application more user-friendly and easier to customize. The insights from the article were instrumental in the success of our fine-tuning efforts.

We were intrigued by article [3] which delves into the intricacies of fine-tuning pre-trained language models across different domains. This study, which centers around the financial sector, provides an in-depth exploration of techniques for selecting datasets, preparing data, selecting appropriate models, and refining them to meet the specific requirements of financial applications. The thorough explanation of his modeling approach aligns well with our concept, with a slight variation of not solely depending on data from the financial sector. This paper also acted as an essential resource that educated us on the fundamental principles of refining extensive language models, which greatly influenced our approaches.

One intriguing aspect was the utilization of QLoRA (enhanced LoRA) which involved the incorporation of 4-bit quantized nodes and a paging mechanism to optimize memory usage. This innovative approach enables the fine-tuning of large language models on personal computers, with minimal information loss despite significant data compression. Although we were unable to employ this method, it sparked a great deal of enthusiasm and effort. In addition, the paper's exploration of security and regulatory compliance has significant implications for our project as it expands, particularly in sectors like healthcare or legal services where data handling and compliance are crucial. This aspect of Jeong's research emphasizes the significance of incorporating strong security measures and complying with regulations, which is essential for the future growth of our project in these areas.

For our research, we also utilized evaluation scores, which are crucial metrics for assessing the effectiveness of our model. BLEU assesses the quality of machine-translated text by measuring the similarity of n-grams with reference texts. Another such method was chrf++, which also did the same but relied on the semantic nature to compare. These were mentioned in Clément's blog [4], which emphasized its effectiveness in assessing the quality and flow of translations. By utilizing these metrics, we were able to effectively compare our models and ensure that their results met the high standards of quality set by humans. Additionally, our models successfully captured the fundamental content of the original texts. We utilized BLEU & chrf to enhance and optimize their performance across various linguistic tasks.

3 Methodology

Manually annotating large-scale task-specific data can be quite challenging, as it demands expertise in writing solutions for each task. As a result, we have created a custom process for producing top-notch vocabulary and its application in the field of astronomy. Given the time constraints and limitations, we utilized a set of 30 words to accurately generate their usage in the desired context. Afterwards, the data is reformatted according to the Databricks Dolly-15k format.

During our dataset validation process, we relied heavily on manual validation to evaluate the quality and accuracy of the word usage examples we generated. This required experienced reviewers carefully analyzing each usage example within its contextual framework, ensuring it adhered to established astronomical principles and terminology. By carefully examining the data, any inconsistencies or errors were quickly detected and resolved, enhancing the overall trustworthiness and dependability of the dataset. In addition, manual validation was conducted to verify the authenticity and coherence of the generated examples, ensuring that they fit naturally within the field of astronomy.

Our dataset validation approach involved iterative refinement based on expert feedback, ensuring quality and relevance. Continuous improvement through feedback loops and refinement cycles maintained precision and relevance. This dedication ensured our dataset stayed up-to-date and aligned with evolving astronomical intricacies.

3.1 Model selection & Fine tuning

Our research focuses on creating word use examples in the field of astronomy. In order to accomplish this, we have opted to utilize the GPT-2 model as the foundation for our experiments. GPT-2, developed by OpenAI, is widely acknowledged for its impressive language generation capabilities and has demonstrated proficiency in various tasks involving understanding and generating human-like language. This model's architecture is centered on the Transformer model and stands out for its attention mechanism. This mecha-

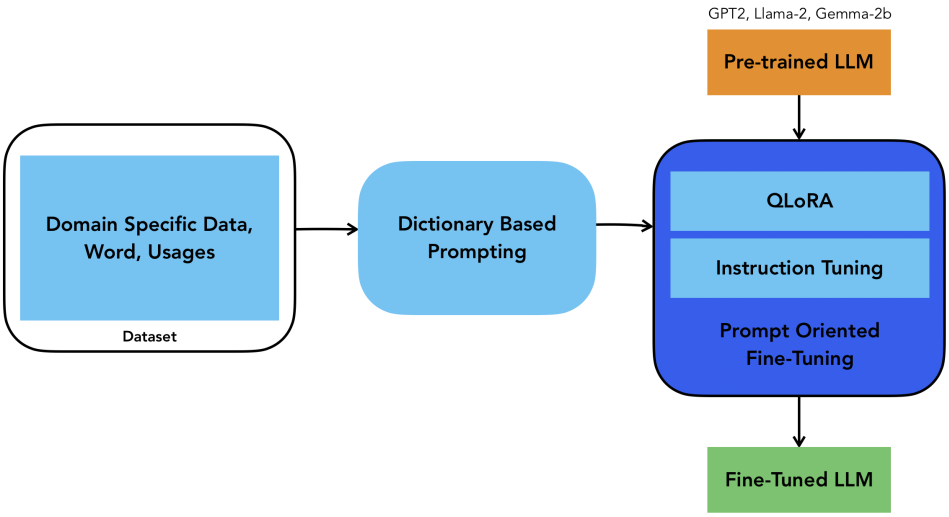

Figure 1: Training Pipeline

nism allows the model to effectively capture connections between elements that are far apart in a sequence of text. We chose to use GPT-2 due to its well-known track record of generating coherent and contextually relevant text, which makes it highly suitable for our goal of creating word usage examples in the specialized domain of astronomy.

In addition to GPT-2, we also took into account Llama2, a well-known large language model that is recognized for its advanced comprehension and production of text that closely resembles human language. Llama2, an improved version of the original LLAMA model, integrates the latest language modeling techniques to provide a more sophisticated understanding of intricate linguistic structures and subtleties. Considering Llama2's impressive track record in different NLP tasks, such as text generation and comprehension, we saw it as a valuable contender for our project, alongside GPT-2.

During the refinement stage of the GPT-2 model, we started by incorporating the pre-trained GPT-2 model and tokenizer, leveraging their powerful language modeling capabilities. We transformed the dataset, containing word definitions and usage examples in the field of astronomy, into input-output pairs suitable for training. With the prepare data function, we encoded every usage sample and its corresponding term, ensuring accurate tokenization and padding for the model's input. Later on, we set up the training inputs, carefully determining factors like the number of training epochs, batch sizes, and logging setups. After all the necessary preparations were completed, we initiated the training process. Throughout this process, the model developed a proficiency in generating word usage examples that were not only contextually appropriate but also linguistically sound, particularly in the realm of astronomy.

In order to maximize the effectiveness of the Llama2 model for our astronomy-related task, we initially organized our data in a manner that is compatible with the model's training requirements. By utilizing a JSONL file containing context, target terms, and related knowledge examples, we were able to acquire instruction-response pairs for each word. The instruction was carefully crafted to inspire the model to generate examples that are closely tied to the field of astronomy. Later on, we arranged the data into input-output pairs that were suitable for training Llama2. We ensured that every example incorporated the desired term and its relevant applications within the field of astronomy. Once the dataset was ready, we initiated the Llama2 model and commenced the process of fine-tuning. We submitted the organized dataset for the training process. Throughout the training process, Llama2 developed the skill to offer examples that are not only applicable to the situation but also linguistically sound. These examples were specifically tailored for the field of astronomy.

In a similar vein, we trained the Gemma-2b model for our task, ensuring that our data was organized in a manner that aligned with the model's training requirements. Through the utilization of a JSONL file containing context, target phrases, and related knowledge samples, we obtained instruction-response pairings for every word. The instruction prompts were deliberately designed to encourage the model to provide practical examples that are directly relevant to the field of astronomy. Afterwards, we utilized the structured dataset as the input for the Gemma-2b model, thus initiating the fine-tuning process. Through the utiliza-

Word (Abound)	GPT2	Llama2	Gemma
Pre Trained	Cars abound in the city center, promoting a cleaner environment.	The abound nature of some galactic formations continues to puzzle astronomers.	The abundance of a specific element or compound can be used to infer the age and history of the Universe.
Post Training	Stars abound in the night sky, visible from Earth.	The universe is rich with countless stars, each galaxy teeming with celestial bodies.	A plethora of planets has abounded in the Milky Way galaxy.

Table 1: Post Hoc (Manual) Evaluation

tion of Keras with a JAX backend, we have facilitated the training process, allowing Gemma-2b to progressively improve its capabilities. Gemma-2b developed the skill to produce relevant and well-formed examples during the training period, specifically tailored to the complexities of astronomy. The given examples demonstrate Gemma-2b's ability to adapt and excel in meeting the unique requirements of the astronomy field.

Figure 1 demonstrates a method to enhance a large language model (LLM) by incorporating domain-specific data, employing dictionary-based prompting, and implementing focused fine-tuning strategies. The method begins with an initial dataset that consists of domain-specific data, encompassing word usages tailored for different situations. This dataset serves as the foundation for dictionary-based prompting, a technique that leverages pre-defined dictionary entries to offer suitable prompts. These prompts are utilized to guide a pre-trained LLM, like GPT-2, Llama-2, or Gemma-2b, by providing it with contextually relevant information that shapes its responses based on the tailored dataset.

In the next phase of the process, the LLM goes through refinement using methods like QLoRA (Quantized Low-Rank Approximation) and fine-tuning of instructions. These methods improve the LLM's ability to understand and generate responses based on the provided prompts. Later on, there is a process known as prompt-oriented fine-tuning, where additional modifications are made to the model to improve its performance on tasks specific to the original dataset's domain. The result is an enhanced LLM that showcases enhanced proficiency in handling inquiries pertaining to a specific field and generating appropriate solutions, leading to a higher level of specialization and practicality in specialized contexts. This approach showcases a comprehensive method for customizing LLMs to meet specific needs and improve their effectiveness in specialized fields.

Model	**Score (ChrF++)**
GPT-2	17.687368000137415
Llama2	49.323358901330056
Gemma	58.821011905271945

Table 2: Evaluation metrics

3.2 Evaluation

Our chrF++ metric evaluation, from table 2, shows Gemma leading in generating astronomy-related content with a score of 58.821, excelling in producing accurate, contextually appropriate text. Llama2 follows with a score of 49.323, demonstrating strong capabilities but less precision in specialized terminology compared to Gemma. GPT-2, scoring 17.687, indicates limitations of older technologies in complex contexts but maintains basic competence. We initially considered the BLEU score for additional evaluation, but its consistent low scores, due to significant deviations from reference texts, led us to rely solely on post hoc analysis for a more accurate assessment.

We conducted experiments using different words to showcase error trends, specifically selecting "abound" to emphasize discrepancies. We conducted multiple iterations, generating predictions using the same word, and carefully selected the output that showcased the highest level of inaccuracy to effectively illustrate our point, showcased in table 1. At first glance, the sentences from the pre-trained models may seem precise. Upon further analysis, it can be discovered that GPT-2 did not produce sentences that aligned with the desired astronomy context. Llama2, despite its semantic inaccuracy, often generated comparable sentences, occasionally replacing "abound" with unrelated words such as "flout" and "feasible" to suit particular subjects. The problems with the Gemma model can be seen that it changes the word itself.

During our evaluation of different models from table 1, we established criteria focusing on the use of a specific input word in the generated sentence, as well as ensuring that the sentence was both semantically and grammatically correct. The Fine Tuned GPT-2

model exhibited satisfactory performance with training words but struggled with unseen words. Furthermore, while its sentence formation was generally acceptable, the semantic accuracy was inconsistent, with correct outcomes in only 6 out of 10 instances. To explain this we can see the case of the prediction for "abound", the sentence predicted was a replacement for the word "around".

The Fine Tuned Gemma and Fine Tuned Llama2 models both generated semantically accurate and meaningful sentences, with Gemma showing particularly notable performance. Initially, the pre-trained Gemma LLM modified words to fit them into sentences, such as changing 'abound' to 'abundance', which occasionally shifted the intended meaning, can be seen in table 1. However, after further training, Gemma exhibited significant improvement, achieving excellent results with both familiar and new words. In contrast, Llama2, although generally effective, sometimes produced sentences that conveyed the meaning of an input word without actually including the word itself, a phenomenon observed in 8 out of 10 cases during our tests. This indicates that while Llama2 manages to capture the essence of the input, it occasionally strays from using the specific input word directly in its predictions.

4 Experimental Setup

For our experimental setup, we utilized a powerful T4 GPU with 65 TFLOPs, along with 16GB of memory and 8 CPU cores. These resources were essential for efficiently training and evaluating our models. Utilizing the PyTorch framework for GPT-2 and LLAMA-2 training, and Keras with a JAX backend for Gemma-2b, we implemented version control using Git to ensure streamlined code management and experiment tracking. The dataset, which contains 30 words related to astronomy, is accompanied by 8 usage sentences in JSONL format. This format allows for easy access and manipulation of the data during training. For GPT-2 and LLAMA-2, we carefully considered the batch size of 2 to strike a balance between computational efficiency and model stability. Throughout the training process, we made dynamic adjustments to hyperparameters to ensure optimal performance. Keeping a close eye on the model's convergence, we adjusted the duration of training epochs to strike a balance between avoiding overfitting and achieving optimal performance. An evaluation was performed using the CHRF++ metric, which involved comparing model predictions with reference usage examples. In general, our experimental framework established a strong basis for training and assessing the GPT-2, LLAMA-2, and Gemma-2b models, allowing for the creation of excellent word usage examples in the field of astronomy.

5 Future Work

Our next goal is to broaden our project by incorporating additional vocabulary from diverse fieldsc and by providing more use examples for each word to enhance the information window. This expansion aims to increase the adaptability and complexity of our models, enabling them to accurately represent a broader spectrum of language subtleties and context-specific variants. We also plan to integrate various contexts into our models through finetuning again, exposing them to a wide range of scenarios and enhancing their capacity to provide relevant word use instances for specific situations, thus improving their overall utility in context-specific text generation. Additionally, we are exploring alternative evaluation metrics like GPTScore, which we anticipate could provide a more nuanced assessment by evaluating both the grammaticality and meaningfulness of the generated text. This could potentially offer a more comprehensive view of text quality and significantly streamline the evaluation process, providing deeper insights into our models' performance across various contexts.

References

[1] *Pre-trained Language Models for Text Generation: A Survey JUNYI LI , Renmin University of China, China and Université de Montréal, Canada TIANYI TANG , Renmin University of China, China WAYNE XIN ZHAO† , Renmin University of China, China JIAN-YUN NIE, Université de Montréal, Canada JI-RONG WEN, Renmin University of China, China .*

[2] *Singh, Vaibhav, and Jiewan Tan. "Fine-Tuning Gemma Models in Hugging Face." Huggingface.co, 23 Feb. 2023, huggingface.co/blog/gemma-peft. Accessed 8 May 2024.*

[3] *Jeong, Cheonsu. "Fine-Tuning and Utilization Methods of Domain-Specific LLMs." ArXiv.org, 24 Jan. 2024, arxiv.org/abs/2401.02981.*

[4] *Brutti-Mairesse, Clément. "ROUGE and BLEU Scores for NLP Model Evaluation." Clément's Blog, 23 Dec. 2021, clementbm.github.io/theory/2021/12/23/rouge-bleu-scores.html. Accessed 8 May 2024.*

[5] Kim, S., & Lee, J. (2021). *Developing an NLP-based Vocabulary Learning Assistant for Language Learners*.

[6] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N. and Kaiser, Lukasz and Polosukhin, Illia (2023), *Attention Is All You Need*,

[7] Touvron, Hugo, et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023.*

Appendix

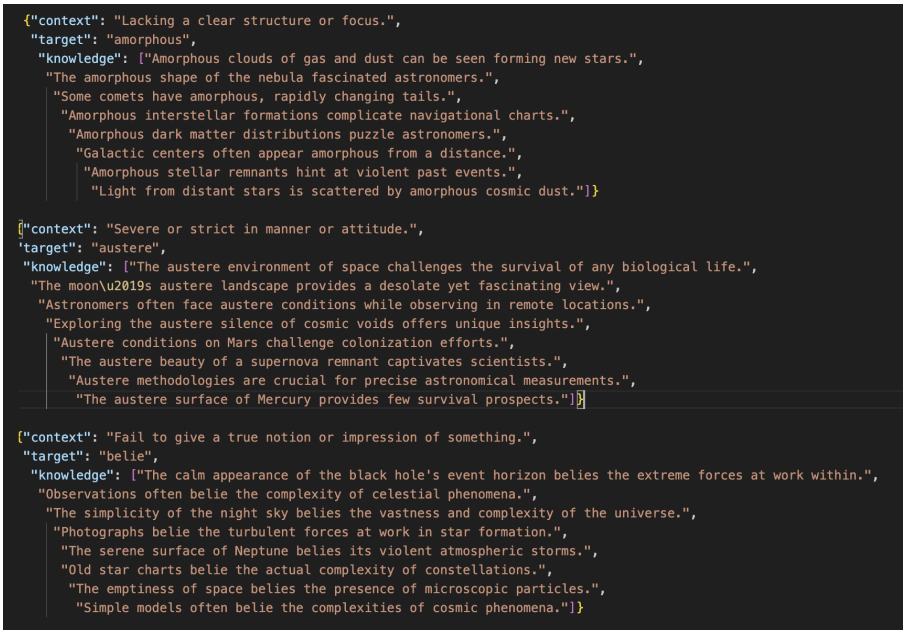

Figure 2: Sample Data

Above is the sample data we have used for training all the models, here, context is the meaning, target is the word and knowledge are the sample sentences which are meant for training. Below are the total list of words used in this total study.

abound	cursory	humdrum	proclivity
amorphous	daunting	insipid	puerile
austere	deify	loquacious	quixotic
belie	didactic	misanthropic	spendthrift
capricious	disseminate	misnomer	taciturn
cerebral	feasible	negligent	wary
congenial	flout	obsequious	
conspicuous	homogeneous	placate	

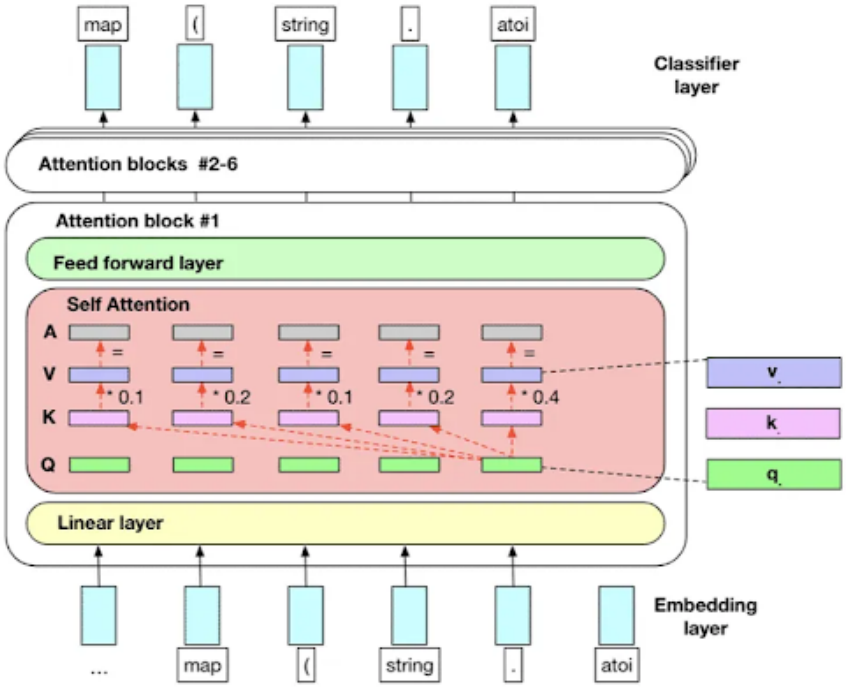

Figure 3: GPT2 Architecture

The GPT-2 architecture relies on a stacked transformer model, employing layers of transformer blocks with self-attention mechanisms to handle sequences of data. The model is designed to effectively handle dependencies in text by combining multi-head attention layers with position-wise feed-forward networks in each block. GPT-2 is highly proficient at producing well-structured and contextually fitting text, thanks to its extensive neural network training on a wide range of data. This enables it to grasp intricate language patterns and structures.

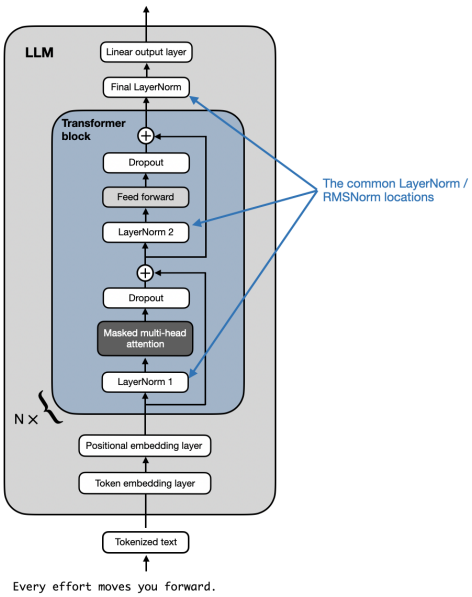

Figure 4: Gemma Architecture

Google's Gemma model is a family of open-source, state-of-the-art language models that are lightweight and built on the same foundational research and technology as the Gemini models. Developed by Google DeepMind and other teams at Google, the Gemma models are designed for both general use and specific instructional purposes, making them versatile tools for developers and researchers. They are available in different sizes, specifically 2B and 7B parameter versions, each with both pretrained and instruction-tuned variants.

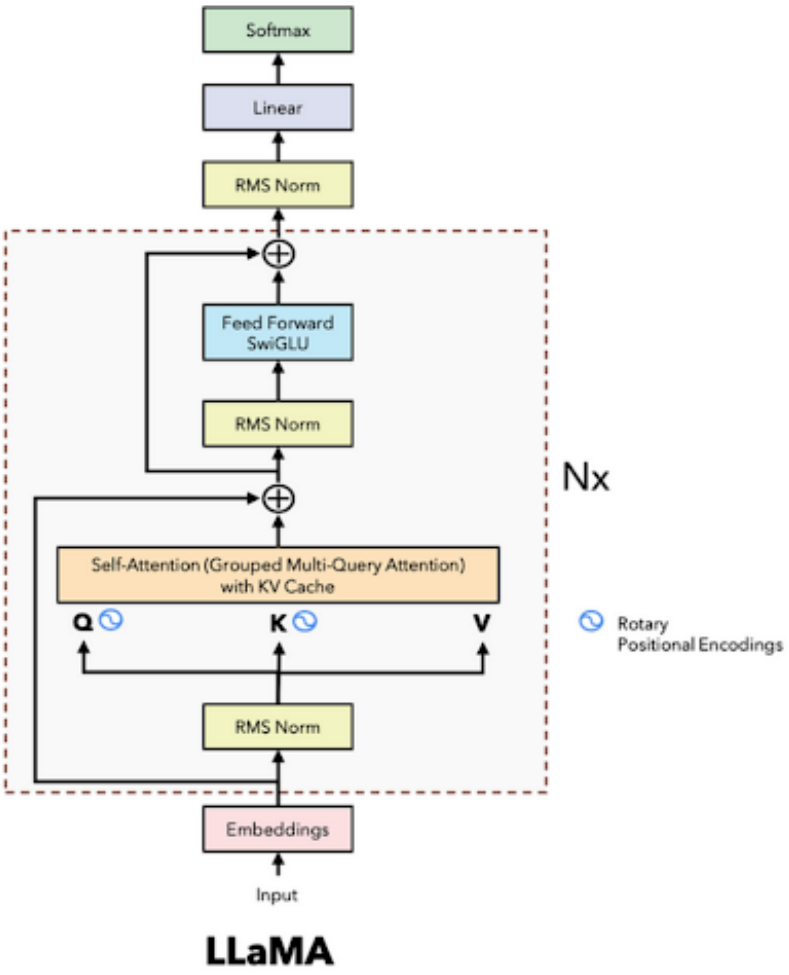

Figure 5: Llama Architecture

The Meta Llama 2 big language model improves on its predecessor, Llama 1. Data training and model architecture enhancements stand out. Llama 2 models are pretrained on 2 trillion tokens with 7 billion to 70 billion parameters. This dataset is larger and cleaner than the previous. In its expanded training procedure, the model emphasizes safety and accuracy using sophisticated approaches including Reinforcement Learning with Human Feedback (RLHF) and Supervised Fine Tuning.