International Journal of Engineering in Computer Science

E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2024; 6(2): 199-203
Received: 06-08-2024
Accepted: 11-09-2024

Ryan Daniel Daneshyari
San Joaquin Delta College and Middle College High School
5151 Pacific Ave, Stockton, CA 95207, USA

Corresponding Author:
Ryan Daniel Daneshyari
San Joaquin Delta College and Middle College High School
5151 Pacific Ave, Stockton, CA 95207, USA

Enhancing llama large language model (llm) for retail analytics through fine-tuning

Ryan Daniel Daneshyari

DOI: https://doi.org/10.33545/26633582.2024.v6.i2c.141

Abstract
This research paper delves into the application of fine-tuning techniques on the Llama model, a cutting-edge open source Large Language Model (LLM) that is used to predict departments within the online grocery shopping context. Department prediction, a critical task in e-commerce, aims to classify products into categories such as "Produce," "Dairy," or "Snacks," making it easier for customers to browse and receive accurate product recommendations. Leveraging an extensive dataset sourced from Instacart, a popular online grocery delivery service, this study utilizes product names and descriptions to refine the Llama model's classification capabilities. This dataset is well-suited for department prediction, containing valuable linguistic nuances specific to grocery items that are challenging for general NLP models to interpret. The proposed methodology includes pre-training the Llama model on a diverse corpus of textual data to establish a broad linguistic foundation, followed by fine-tuning on the Instacart dataset to adapt the model to the unique vocabulary and contextual patterns of grocery items. Fine-tuning, a crucial step in model customization, allows us to tailor the Llama model to recognize product attributes and assign them accurately to their respective departments. This approach not only enhances classification accuracy but also sheds light on the adaptability of NLP models to specialized domains like online grocery shopping. To evaluate the effectiveness of the fine-tuned Llama model, a rigorous experimentation has been conducted and its performance is compared against baseline models and conventional classification methods. The promising findings indicate that fine-tuning significantly boosts the model's ability to accurately predict departments, outperforming traditional approaches in categorizing complex product names and descriptions. The implications of this research extend beyond department prediction and provide insights into the broader applicability of advanced LLM models for e-commerce. By improving product categorization, the proposed study contributes to optimizing user experience, refining recommendation systems, and enhancing operational efficiency within online retail platforms, ultimately supporting better customer engagement and satisfaction.

Keywords: Large language models (LLM), fine-tuning LLM, llama open-source model, custom dataset, natural language processing (NLP)

Introduction

In today's digital commerce era, online platforms have redefined grocery shopping, reshaping how consumers approach purchasing everyday items. With increasing reliance on these platforms, users now expect seamless experiences, making accurate product categorization a crucial component in delivering quality service and meeting business goals. For e-commerce platforms, this entails correctly classifying thousands of products to ensure users can easily browse, search, and receive tailored recommendations. Efficient and precise categorization plays a key role not only in user satisfaction but also in operational aspects such as inventory management, product availability, and demand forecasting. This paper explores the application of advanced machine learning techniques-specifically the fine-tuning of the Llama-2 model-on a comprehensive dataset from Instacart to achieve more accurate department predictions for grocery products.

Instacart's dataset is particularly rich in information, making it ideal for machine learning research. This dataset includes vast records of user interactions, detailed product descriptions, and transaction histories. Such a dataset allows researchers to study a wide variety of products with intricate naming conventions, providing valuable insights into consumer behavior.

The Llama-2 model, known for its impressive natural language processing capabilities, is a strong candidate for tackling the challenges of grocery product categorization. Fine-tuning this model on Instacart's dataset allows it to better understand the specific linguistic patterns associated with grocery products-patterns that may vary significantly from other text sources due to the unique nature of product descriptions, which often include brand names, specific ingredients, sizes, and other nuanced details.

Fine-tuning the Llama-2 model involves adjusting it specifically to recognize and categorize products within the grocery domain. By using both product names and descriptions, the model gains a comprehensive understanding of what defines each department. For instance, products in the "Produce" department might include specific terms related to fruits and vegetables, while the "Dairy" department is characterized by items such as milk, cheese, and yogurt. This contextual adaptation allows the model to effectively classify products with high accuracy, even for items with ambiguous names or products that might fit into multiple categories. The unique language of grocery items, marked by a blend of general and domain-specific terms, poses a significant challenge for classification tasks. The model's training enables it to overcome these obstacles, recognizing subtle distinctions and ensuring that products are categorized accurately.

The potential impact of this research extends across various facets of e-commerce. By fine-tuning the Llama-2 model for department prediction, the reliability of automated systems within online platforms had been enhanced. Accurate categorization improves product recommendations, allowing users to find what they need with minimal effort. This, in turn, fosters higher customer satisfaction and loyalty. Furthermore, the efficiency of inventory management benefits greatly from accurate product classification, as products can be tracked, stocked, and reordered with greater precision, reducing instances of stockouts or overstocking.

The significance of this study lies in demonstrating the viability of adapting advanced NLP models like Llama-2 to specialized domains, showing how fine-tuning can address specific challenges. Insights from this research contribute to the broader field of machine learning in e-commerce, offering a blueprint for other sectors looking to implement predictive modeling for domain-specific applications. As online grocery shopping continues to grow, this research underscores the value of sophisticated machine learning models in optimizing user experience and operational efficiency, setting a foundation for future advancements in predictive modeling and digital commerce strategies.

Related Works

The field of Large Language Models (LLMs) has advanced rapidly, leading to the development of models exceeding 100 billion parameters, such as GPT, Gemini, Claude, Mistral, Llama, and specialized models like Galactica for scientific tasks. These advancements align with scaling laws that traditionally emphasized model weights; however, recent models, such as Chinchilla, shifted this focus by prioritizing token counts, achieving notable performance with a smaller parameter size. Llama, a model designed for computational efficiency, marks a significant milestone in LLM development. Alongside it, discussions around open-source and closed-source models have gained momentum. Open-source LLMs like Llama and Mistral now compete with closed-source models, including OpenAI's GPT series, Google's Gemini, and Anthropic's Claude, highlighting a shift in the LLM landscape where model accessibility and scale both play crucial roles. In practice, production-ready LLMs like ChatGPT, Gemini, and Claude exhibit a notable performance and usability advantage due to fine-tuning techniques tailored to align with human preferences-a process still evolving in open-source models. Efforts within the open-source community to close this performance gap include distillation-based models, which use synthetic instruction data in training (Raihan, *et al*. 2024) [2].

While these approaches show potential, they have yet to match the refined performance of leading closed-source models. Achieving similar usability and alignment in open-source LLMs remains a central challenge, emphasizing the ongoing pursuit of innovation in both research and practical applications across the LLM ecosystem. Instruction tuning has proven to be a potent methodology within the domain of LLMs. Exceptional achievement on zero-shot performance on tasks not seen before had been achieved by finely adjusting LLMs across a variety of datasets. Factors such as the number of tasks, model size, and prompt settings had been investigated for the tuning (Thulke, 2024) [1]. The prompts employed for instruction tuning can originate from human input or be generated autonomously by LLMs.

Additionally, subsequent instructions are utilized to enhance initial generations, making them more valuable, captivating, and impartial. An approach akin to instruction tuning is chain-of-thought prompting where models are prompted to elucidate their reasoning when faced with intricate problems, aiming to heighten the likelihood of their final answer being accurate. This comprehensive exploration of instruction tuning and related strategies signifies a dynamic research domain focused on augmenting the versatility and efficacy of LLMs across various cognitive tasks (Raihan, *et al*. 2024) [2].

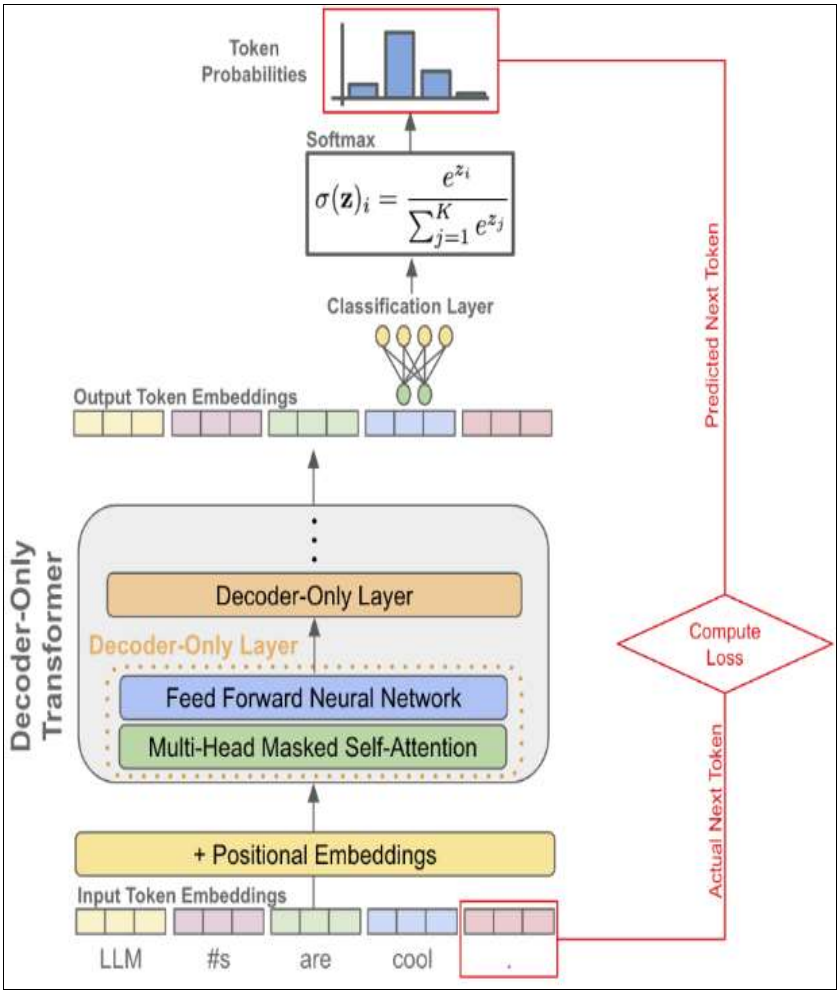

Fig 1: Procedure on how LLM predicts the next token

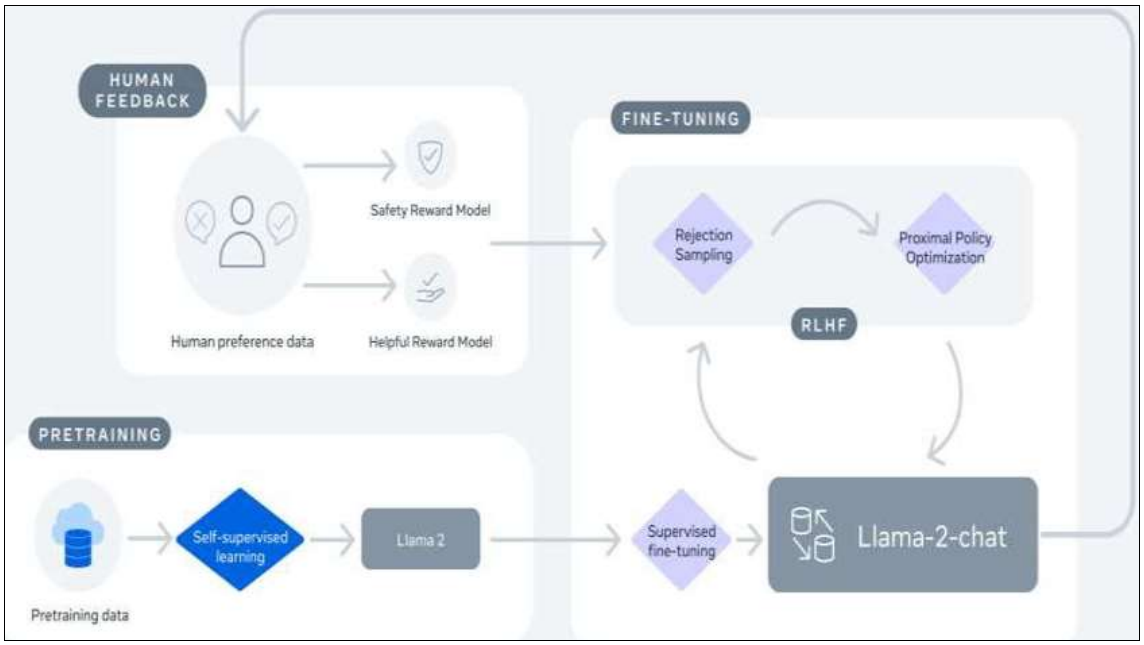

Fig 2: Pretraining vs finetuning in Large Language Models (LLMs)

Methodologies

A. Dataset Collection and Preparation

To conduct the analysis, datasets supplied by Instacart Inc had been acquired. The database encompasses more than 3 million purchases made by over 200,000 Instacart consumers. Both product and consumer data incorporate 50,000 unique products, encompassing various product aisles, departments, weeks, and times of purchase. Data preparation is necessary for data analysis after data gathering. Six different datasets provided information about customer purchases and transaction details. Order products train and Order products prior were merged into a single

data set based on order id and product id respectively. Later, the department, aisles, and orders datasets were merged with the Order products train and Order products prior combined dataset through department id, aisles id, and orders id. This created a master dataset to begin the analysis. Along with exploring the length and attributes of a dataset, it was also checked for null data. When the dataset was computed using pandas there are only 2078068 missing values among the many 33819106 values. That computes around 5 percent of data as missing which is low enough of a threshold to be eliminated.

B. Data Cleaning and Preprocessing

According to the previous section, the data has around 5 percent of null values with missing attributes. These values have been rejected. This is to ensure the data is used for modelling and to ensure that results are accurate. Since the percentage of null values is so low, the null values are removed, reducing the data size to 31741038 from 33819106.

C. Model Development and Training

The Colab T4 GPU has a limited 16 GB of VRAM, which is barely enough to store Llama 2–7b's weights, which means full fine-tuning is not possible, and one needs to use parameter-efficient fine-tuning techniques like LoRA or QLoRA. The QLoRA technique needs to be used to fine-tune the model in 4-bit precision and optimize VRAM usage. For that, the Hugging Face ecosystem of LLM libraries had been used: transformers, accelerate, peft, trl, and bits and bytes. You can access the Meta's official Llama-2 model from Hugging Face, but you have to apply for a request and wait a couple of days to get confirmation. Instead of waiting, the Nous Research's Llama-2-7b-chat-hf is used as the base model. It is the same as the original but easily accessible. 4-bit quantization via QLoRA allows efficient finetuning of huge LLM models on consumer hardware while retaining high performance. This dramatically improves accessibility and usability for real-world applications. QLoRA quantizes a pre-trained language model to 4 bits and freezes the parameters. A small number of trainable Low-Rank Adapter layers are then added to the model. During fine-tuning, gradients are backpropagated through the frozen 4-bit quantized model into only the Low-Rank Adapter layers. So, the entire pretrained model remains fixed at 4 bits while only the adapters are updated. Also, the 4-bit quantization does not hurt model performance.

Now using 4-bit precision with the compute dtype "float16" from Hugging Face for faster training, the model is loaded. Traditional fine-tuning of pre-trained language models requires updating all of the model's parameters, which is computationally expensive and requires massive amounts of data. Parameter-Efficient Fine-Tuning (PEFT) works by only updating a small subset of the model's parameters, making it much more efficient. Learn about parameters by reading the PEFT official documentation. Below is a list of hyperparameters that can be used to optimize the training process:

- **Output Dir:** The output directory is where the model predictions and checkpoints will be stored.
- **Num train epochs:** One training epoch.
- **Fp16/bf16:** Disable fp16/bf16 training.
- **Per device train batch size:** Batch size per GPU for training.
- **Per device eval batch size:** Batch size per GPU for evaluation.
- **Gradient accumulation steps:** This refers to the number of steps required to accumulate the gradients during the update process.
- **Gradient check pointing:** Enabling gradient check pointing.
- **Max grad norm:** Gradient clipping.
- **Learning rate:** Initial learning rate.
- **Weight decay:** Weight decay is applied to all layers except bias/LayerNorm weights.
- **Optim:** Model optimizer (AdamW optimizer).
- **Lr scheduler type:** Learning rate schedule.
- **Max steps:** Number of training steps.
- **Warmup ratio:** Ratio of steps for a linear warmup.
- **Group by length:** This can significantly improve performance and accelerate the training process.
- **Save steps:** Save checkpoint every 25 update steps.
- **Logging steps:** Log every 25 update steps.

D. Model Evaluation

The performance of the fine-tuned Llama-2 large language model on instacart dataset was evaluated using accuracy as the primary metric. Accuracy provided a clear measure of the model's effectiveness in correctly predicting department of the given product. Additionally, the study conducted training and testing accuracy assessments to evaluate the model's performance and to check for overfitting.

Results

The model's results are highly promising, demonstrating a robust predictive capability with a 98% accuracy rate in both training and testing phases. Such high accuracy indicates that the model is well-suited for product recommendation tasks, as it effectively predicts the appropriate department for a given product based on its name and description. The model's ability to output multiple department options, with the first listed as the most suitable match, is a valuable feature, especially in complex product categorization scenarios. Interestingly, the original Llama-2 base model with 7 billion parameters lacked the capacity to perform such specific department predictions. However, through fine-tuning, this model has been enhanced to identify and categorize products accurately, achieving a level of precision that is advantageous for e-commerce applications. This success in fine-tuning underscores the model's potential to streamline product sorting processes and improve the accuracy of recommendation systems within digital retail platforms.

Conclusion

The fine-tuning of the model achieved the desired results for the given problem, effectively enhancing its ability to categorize products into departments. However, a notable limitation observed in the current model is its inability to pinpoint a single department for each product. Instead, it generates multiple department predictions for individual items, which can introduce ambiguity and reduce the model's effectiveness in real-world applications, where precise categorization is critical. To address this, future improvements should focus on refining the model's capability to converge on a single department prediction. One promising approach would be to train a higher-capacity

version of the model, such as the 70-billion-parameter Llama-2, which is designed for handling complex, nuanced data. Leveraging high-performance GPUs during training can expedite this process, enabling the model to manage and process vast amounts of data effectively.

Additionally, incorporating a more comprehensive and extensive dataset from various e-commerce platforms would enrich the model's understanding of product-specific language, terminologies, and categorizations. This enriched dataset could include a broader variety of products, descriptions, and departmental classifications, giving the model more context to make precise predictions. By enhancing the model with these improvements, it can achieve greater accuracy and usability, ultimately contributing to more reliable product categorization, improved user experience, and streamlined inventory management in e-commerce applications. These advancements would not only meet current requirements but also pave the way for future innovations in product classification systems

References

1. Thulke D, Gao Y, Jalota R, Dugast C, Ney H. Prompting and fine-tuning of small LLMs for length-controllable telephone call summarization. arXiv preprint arXiv:2410.18624. 2024.
2. Raihan N, Siddiq ML, Santos J, Zampieri M. Large language models in computer science education: A systematic literature review. arXiv preprint arXiv:2410.16349. 2024.
3. Gurudath S. Market basket analysis & recommendation system using association rules; c2020.
4. Dhanabhakyam M, Punithavalli M. A survey on data mining algorithm for market basket analysis. Global Journal of Computer Science and Technology. 2011;11(2):1-5.
5. Liu Y, Guan Y. FP-growth algorithm for application in research of market basket analysis. In: Proceedings of the 2008 IEEE International Conference on Computational Cybernetics. 2008:269–272. IEEE.
6. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB); Santiago, Chile: IBM Almaden Research Centre; c1994. p. 487-499.
7. Abdulsalam S, Adewole K, Akintola A, Hambali M. Data mining in market basket transaction: An association rule mining approach. International Journal of Applied Information Systems. 2014;7(10):15-20.
8. Yang X, Guo Y, Liu Y, Steck H. A survey of collaborative filtering based social recommender systems. Computer Communications. 2014;41:1-10.
9. Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing. 2003;7(1):76-80.