UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

CARRERA DE CIENCIA DE LA COMPUTACIÓN

Large Language Models for the Generation of reviews for products in e-commerce

AUTOR

Luis Antonio Gutiérrez Guanilo
luis.gutierrez.g@utec.edu.pe

ASESOR

Cristian López
clopezd@utec.edu.pe

Lima - Perú
2023

Contents

Chapter 1

Context and Motivation

Introduction

Large Language Models (LLMs) such as GPT-4, BERT, LLama, and LLama2 are transforming sectors like healthcare [1] [2], finance, and e-commerce by their remarkable ability to understand and generate text that closely resembles human communication. These models play a pivotal role in enhancing decision-making processes, automating customer service, and improving data analysis [3].

Although these models perform well across various applications, there are scenarios where they require specific training to handle particular tasks effectively. Fine-tuning is a strategic approach to enhance model performance by training pre-existing models with specialized datasets to better meet domain-specific needs [4]. Examples of such specialized applications include LLama2-chat [5], Mistral Instruct [6], and StructLM [7], each tailored with unique datasets. However, the lack of high-quality, focused datasets, particularly in areas like product attributes and e-commerce, remains a significant challenge, emphasizing the need for comprehensive datasets that enable models to interact effectively with detailed product information.

Creating a dataset involves a deep understanding of the data types collected. While Audio and Video are significant, Text and Tabular data are more common in real-world applications, appearing in formats such as Excel tables, Wikipedia pages, and other spreadsheets. These data can be formatted in several styles, including HTML, CSV (Comma Separated Values), TSV (Tab Separated Values), Markdown, DFLoader, Data-Matrix, and JSON. JSON, in particular, is highly valued for its readability and easy integration with contemporary web technologies [8].

Using JSON-centric methods to fine-tune models significantly enhances their capacity to process and generate structured data accurately [9]. This capability is crucial for e-commerce platforms, where product data's structure and content frequently vary. By focusing on JSON-structured data to fine-tune LLMs like LLama2-chat, Mistral Instruct, and StructLM, this project seeks to significantly

refine the extraction and normalization of product espicifications. This will lead to more accurate and contextually relevant product reviews, directly improving them and making more humanized.

Problem Description

Despite the advancements of LLMs in various sectors, they often struggle with domain-specific tasks without precise and targeted training. A significant problem in e-commerce is the interaction with detailed product information due to the lack of high-quality, focused datasets excluding Amazon or Wikipedia datasets. This deficiency affects the models' ability to accurately extract and normalize product attribute values, leading to suboptimal product reviews and recommendations. Additionally, the diverse structure and content of product data on e-commerce platforms pose a challenge. There is a pressing need to create and utilize datasets that cater specifically to the structure and nuances of product data, particularly in JSON format, to enhance the performance of LLMs in accurately processing and generating structured data.

Motivation

The key challenge in leveraging LLMs effectively in e-commerce and other sectors is the absence of high-quality, focused datasets, especially concerning product characteristics. This gap hinders the models' ability to interact efficiently with detailed product information. Fine-tuning pre-existing models with specialized datasets is a strategic approach to enhance model performance and meet domain-specific needs. Moreover, the prevalent use of JSON in modern web technologies underscores the importance of fine-tuning models to process and generate structured data accurately.

Objectives

Genetal Objective

The primary objective of this project is to refine product recommendations based on the specifications of a product in e-commerce by enhancing the accuracy of product reviews and focused reviews.

Specific Objectives

Specifically, this project aims to fine-tune LLMs like LLama2-chat, Mistral Instruct, and StructLM with a bespoke, product-related JSON dataset. This will enable proficient analysis and processing of JSON-formatted product data, leading to more accurate and contextually relevant product reviews. The

initiative also seeks to compare the improved performance of these trained models against baseline models, demonstrating significant enhancements in handling structured product data based on the metrics of hallucination, fluency, and relevance.

Aportes

Theoretical Framework

Large Language Models (LLMs)

LLMs are significant advancements in natural language processing (NLP), evolving from statistical language models to neural models. The term "large language model" typically refers to pre-trained language models of considerable size, often containing hundreds of millions to billions of parameters [10].

These models are trained on massive text corpora using self-supervised learning methods, allowing them to generate human-like text, perform translation, summarization, sentiment analysis, and more. The extensive training data and advanced architectures enable LLMs to capture complex language patterns and exhibit remarkable zero-shot and few-shot learning abilities [11].

LLMs are used in various fields beyond standard NLP tasks. They have shown promise in enhancing recommendation systems, performing complex planning, and even contributing to fields like telecommunications and robotics [12] [13].

Fine tuning

Fine-tuning in large language models (LLMs) is a critical process that involves adjusting a pre-trained model's parameters to optimize performance on specific downstream tasks. This technique leverages the extensive training already performed on massive, unlabeled text corpora and then refines the model using a smaller, task-specific dataset. Fine-tuning is essential because it allows models to generalize from the vast, generic data they were initially trained on to the specialized tasks they need to perform. For instance, the Child-Tuning technique updates only a subset of parameters by masking out non-essential gradients, thus improving both efficiency and performance on tasks within the GLUE benchmark [14].

Fine-tuning strategies can vary significantly depending on the model and the resources available. For example, some approaches focus on parameter-efficient methods to reduce computational costs while maintaining high performance. Techniques like the LoRA (Low-Rank Adaptation) allow substantial fine-tuning of LLMs with minimal additional parameters, making

it feasible even with limited computational resources [15]. Additionally, methods such as differentially private fine-tuning have been developed to protect sensitive data during the fine-tuning process, achieving a balance between model utility and data privacy [16].

JSON-Tuning

JSON-Tuning is an innovative approach designed to enhance the performance and efficiency of Large Language Models (LLMs) by utilizing the structured data representation capabilities of JSON (JavaScript Object Notation). This method leverages JSON's hierarchical structure to optimize the input-output processes of LLMs, resulting in improved parameter tuning and model interpretability. JSON-Tuning allows for more precise control over training data, which in turn leads to more robust and contextually accurate predictions. This approach facilitates the efficient organization of data, making it easier to manage and utilize during the training and fine-tuning phases of LLM development [17].

The advantages of JSON-Tuning are not limited to performance enhancements alone. This technique can significantly reduce the computational burden typically associated with traditional fine-tuning methods. By streamlining data processing steps and minimizing redundancy, JSON-Tuning makes it feasible to deploy LLMs in real-time applications where speed and accuracy are crucial [18]. Furthermore, the structured nature of JSON supports seamless integration with existing data pipelines and APIs, thus simplifying workflows for data scientists and developers. This synergy between structured data representation and advanced model tuning offers a promising direction for future research and development in machine learning.

Chapter 2

State of the Art

Pretrained models

Estructured data models

E-commerce models

Metrics for evaluation of performance

Chapter 3

Metodología

3.1 Descripción de la Metodología

Chapter 4

Experimentaciones y Resultados

4.1 Experimentos y Resultados

Chapter 5

Conclusiones y Trabajos Futuros

5.1 Conclusiones

5.2 Trabajos Futuros

Bibliography

[1] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," 2023.

[2] S. Reddy, "Evaluating large language models for use in healthcare: A framework for translational value assessment," *Informatics in Medicine Unlocked*, vol. 41, p. 101304, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823001508

[3] T. Varshney, "Build an llm-powered data agent for data analysis," Feb 2024. [Online]. Available: https://developer.nvidia.com/blog/build-an-llm-powered-data-agent-for-data-analysis/

[4] D. Bergmann, "Build an llm-powered data agent for data analysis," March 2024. [Online]. Available: https://www.ibm.com/topics/fine-tuning

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[7] A. Zhuang, G. Zhang, T. Zheng, X. Du, J. Wang, W. Ren, S. W. Huang, J. Fu, X. Yue, and W. Chen, "Structlm: Towards building generalist models for structured knowledge grounding," 2024.

[8] A. Singha, J. Cambronero, S. Gulwani, V. Le, and C. Parnin, "Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms," 2023.

[9] C. Gao, W. Zhang, G. Chen, and W. Lam, "Jsontuning: Towards generalizable, robust, and controllable instruction tuning," 2024.

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2023.

[11] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024.

[12] M. Debbah, "Large language models for telecom," in *2023 Eighth International Conference on Fog and Mobile Edge Computing (FMEC)*, 2023, pp. 3–4.

[13] T. Fan, Y. Kang, G. Ma, W. Chen, W. Wei, L. Fan, and Q. Yang, "Fate-llm: A industrial grade federated learning framework for large language models," 2023.

[14] R. Xu, F. Luo, Z. Zhang, C. Tan, B. Chang, S. Huang, and F. Huang, "Raise a child in large language model: Towards effective and generalizable fine-tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9514–9528. [Online]. Available: https://aclanthology.org/2021.emnlp-main.749

[15] X. Sun, Y. Ji, B. Ma, and X. Li, "A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model," 2023.

[16] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz, S. Yekhanin, and H. Zhang, "Differentially private fine-tuning of language models," 2022.

[17] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," 2024.

[18] L. Zhu, L. Hu, J. Lin, and S. Han, "LIFT: Efficient layer-wise fine-tuning for large model models," 2024. [Online]. Available: https://openreview.net/forum?id=u0INlprg3U