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 Del Alamo
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 [10]. 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.

Objectives

Genetal Objective

The primary objective of this project is generate product reviews based on tabluar data representing product features. This will be achieved by fine-tuning large language models (LLMs) on a product-related JSON dataset, and evaluating the models based on hallucination, fluency, and relevance.

Specific Objectives

Specifically, this project will create a product-related JSON dataset to fine-tuning LLMs like LLama2-chat, Mistral Instruct, and StructLM. The trained models will be evaluated based on the metrics of hallucination, fluency, and relevance, demonstrating significant improvements in handling structured product data.

Aportes

Theoretical Framework

E-commerce Product-related Databases

In the rapidly evolving world of e-commerce, managing and utilizing product-related databases has become more advanced. Recent developments focus on integrating sophisticated database queries and big data technologies to improve the efficiency and precision of product searches. Research indicates that incorporating database queries into e-commerce platforms significantly streamlines the search process, making it more user-friendly and effective [11]. Additionally, using big data technologies like Hadoop and MPP distributed databases enables detailed analysis of customer reviews and purchasing trends, optimizing product selection and enhancing user experience [12].

The advancement of database technologies has also led to the creation of new frameworks that support complex data formats and improve the efficiency of e-commerce platforms. For instance, cloud computing-based platforms such as Productpedia help create a centralized electronic product catalog, allowing seamless data synchronization and enabling merchants to define and share semantically rich product information [13]. Moreover, deploying machine learning models like TrendSpotter helps e-commerce platforms predict and highlight trending products by analyzing current customer engagement data, thereby meeting the market's dynamic demands [14].

Large Language Models (LLMs)

Large language models (LLMs) represent significant progress in natural language processing (NLP), transitioning from statistical to neural models. The term ”large language model” generally refers to pre-trained language models of substantial size, often containing hundreds of millions to billions of parameters [15].

These models are trained on extensive text datasets using self-supervised learning techniques, enabling them to generate human-like text and perform tasks such as translation, summarization, and sentiment analysis. Due to their extensive training data and sophisticated architectures, LLMs can capture complex language patterns and demonstrate impressive zero-shot and few-shot learning capabilities [16].

Beyond typical NLP tasks, LLMs are utilized in various fields. They show potential in improving recommendation systems, executing complex planning, and contributing to areas like telecommunications and robotics [17] [18].

Fine Tuning

Fine-tuning in machine learning is a process where a pre-trained model is adapted to a new, often related task by continuing the training process on a smaller, task-specific dataset. This process is crucial for enhancing model performance and achieving better generalization on the new task.

Mathematical Framework

Fine-tuning leverages the pre-existing knowledge embedded in the model parameters from the initial training on a large dataset. Mathematically, this involves optimizing a loss function L with respect to the model parameters θ, which have been pre-trained on a large-scale dataset D. The fine-tuning process then adjusts these parameters using a smaller dataset D' specific to the new task. The objective can be expressed as:

$$\min_{\theta} L_{D'}(\theta)$$

where $L_{D'}$ represents the loss on the fine-tuning dataset. This optimization typically uses gradient-based methods to adjust the pre-trained weights minimally but effectively to improve performance on the new task [19].

Operational Fine-Tunings

In a more abstract sense, fine-tuning can be seen as an operational fine-tuning where the changes made to the model parameters are tailored to the specifics of the new task. This concept extends beyond traditional parameter optimization, embedding domain-specific knowledge and constraints into the model adjustments. Operational fine-tunings often require ensuring that the adjustments do not lead to significant deviations from the model's prior capabilities, ensuring stability and performance consistency [20].

Sample Complexity and Generalization

The effectiveness of fine-tuning is influenced by the similarity between the pre-training and fine-tuning tasks. The sample complexity, which is the number of training examples required to achieve a certain level of performance, is significantly reduced when fine-tuning is applied. This reduction occurs because the pre-trained model already captures a broad set of features relevant to many tasks. Fine-tuning adjusts these features to better fit the new task, often requiring fewer samples to achieve high accuracy. This relationship can be formalized by analyzing the changes in the generalization bounds of the model after fine-tuning [21].

Gradient-Based Fine-Tuning

Fine-tuning often involves gradient-based optimization techniques. For deep neural networks, this means leveraging algorithms like Stochastic Gradient Descent (SGD) to iteratively adjust the weights. The process can be sensitive to the initial learning rate and other hyperparameters, which need to be carefully chosen to avoid large deviations from the pre-trained weights and ensure convergence to a new, optimal set of parameters for the fine-tuning task [22].

Computational Efficiency

Fine-tuning is computationally efficient compared to training a model from scratch. By starting with a pre-trained model, the number of training epochs and the amount of data required are significantly reduced. This efficiency is particularly beneficial for large-scale models where the computational cost of full training is prohibitive. Fine-tuning allows for the practical deployment of advanced models in resource-constrained environments by focusing computational resources on the most impactful aspects of training [23].

JSON-Tuning

JSON-Tuning is a novel approach aimed at enhancing the performance and efficiency of Large Language Models (LLMs) by leveraging the structured data representation capabilities of JSON (JavaScript Object Notation). This method utilizes JSON's hierarchical structure to optimize the input-output processes of LLMs, leading to better parameter tuning and improved model interpretability. JSON-Tuning provides more precise control over training data, resulting in more robust and contextually accurate predictions. This approach also facilitates efficient data organization, simplifying management and utilization during the training and fine-tuning stages of LLM development [24].

The benefits of JSON-Tuning extend beyond performance improvements. This technique can substantially reduce the computational load typically associated with traditional fine-tuning methods. By streamlining data processing and minimizing redundancy, JSON-Tuning enables the deployment of LLMs in real-time applications where speed and accuracy are essential. Additionally, JSON's structured nature allows for seamless integration with existing data pipelines and APIs, simplifying workflows for data scientists and developers [25]. This combination of structured data representation and advanced model tuning offers a promising avenue for future research and development in machine learning.

Evaluation Metrics

BLEU (Bilingual Evaluation Understudy)

The BLEU metric is a widely-used method for evaluating the quality of text which has been machine-translated from one language to another. BLEU measures the correspondence between a machine's output and that of a human by calculating the precision of n-grams (sequences of words) in the generated text relative to a reference translation. Mathematically, the BLEU score is calculated using the formula:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where:

- BP is the brevity penalty to penalize short translations.
- w_n is the weight for n-gram precision.
- p_n is the precision for n-grams of length n.

Brevity penalty BP is defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

where c is the length of the candidate translation and r is the length of the reference translation [26].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a set of metrics used for evaluating automatic summarization and machine translation that measures the overlap between the generated output and a reference output. Key variants include ROUGE-N, ROUGE-L, and ROUGE-W.

1. **ROUGE-N**: Measures the n-gram recall between the candidate and reference summaries.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{RefSummaries}} \sum_{gram_n \in S} \text{Count}_{match}(gram_n)}{\sum_{S \in \text{RefSummaries}} \sum_{gram_n \in S} \text{Count}(gram_n)}$$

   where $gram_n$ is any n-gram, and $\text{Count}_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate and reference summary.

2. **ROUGE-L**: Measures the longest common subsequence (LCS) based statistics, capturing sentence-level structure similarity.

$$\text{ROUGE-L} = \frac{LCS(C, R)}{\text{length}(R)}$$

where $LCS(C, R)$ is the length of the longest common subsequence between candidate C and reference R [27].

3. **ROUGE-1 and ROUGE-2**: Specifically measure the overlap of unigrams and bigrams, respectively, between the candidate and reference summaries [28].

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR evaluates translations by aligning them to human-created reference translations using various modules such as exact matching, stemming, synonymy matching, and paraphrase matching. The final score is a harmonic mean of unigram precision and recall, favoring recall:

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

where:

- P is the precision of unigrams.
- R is the recall of unigrams.

This metric also incorporates a penalty function for longer alignment chunks to address issues of word ordering [29].

Hallucination in NLP

Hallucination in Natural Language Processing (NLP) refers to the phenomenon where a language model generates text that is not supported by the input data or factual reality. This issue is prevalent in various NLP tasks such as machine translation, text summarization, and dialogue systems. Hallucinations can degrade the quality and reliability of the generated text, making it crucial to detect and mitigate them effectively [30].

Types of Hallucinations

Hallucinations in NLP can be broadly classified into intrinsic and extrinsic types:

- **Intrinsic Hallucinations** occur when the generated text is internally inconsistent or illogical.
- **Extrinsic Hallucinations** happen when the generated content diverges from the source data or factual information [31].

Calculating the Percentage of Hallucinations

To quantify hallucinations in generated text, a systematic approach involves calculating the percentage of hallucinated content. This can be done using the following method:

1. **Identify Hallucinated Instances**: Detect segments of the generated text that do not align with the input data or known facts. This can be done manually by experts or using automated tools.
2. **Count Hallucinated Instances**: Count the number of hallucinated segments identified.
3. **Calculate Total Instances**: Determine the total number of segments or sentences generated by the model.
4. **Compute Hallucination Percentage**:

$$\text{Hallucination Percentage} = \left(\frac{\text{Number of Hallucinated Instances}}{\text{Total Number of Instances}} \right) \times 100$$

For example, if a model generates 100 sentences and 15 of them are identified as hallucinated, the hallucination percentage would be:

$$\text{Hallucination Percentage} = \left(\frac{15}{100} \right) \times 100 = 15\%$$

This metric provides a quantitative measure of the extent of hallucination in generated content and can be used to evaluate and improve the reliability of language models [32].

Chapter 2

State of the Art

Pretrained models

Pre-trained language models has seen remarkable advancements, leveraging large datasets and sophisticated training methodologies to achieve significant improvements in various natural language processing (NLP) tasks. Pre-trained models such as BERT, GPT, and their variants have revolutionized the field by providing robust, general-purpose representations that can be fine-tuned for specific tasks with minimal additional training data [33]. The introduction of techniques like function-preserving initialization and advanced knowledge initialization in bert2BERT exemplifies innovative methods to enhance the efficiency of pre-training larger models by reusing smaller pre-trained models, thus reducing computational costs and carbon footprints associated with training from scratch [33].

Moreover, the application of pre-trained models in domains such as clinical information extraction has demonstrated their versatility and effectiveness. For instance, large language models like GPT-3 have been utilized to decode complex medical jargon and abbreviations in electronic health records, significantly improving the extraction of actionable medical information without extensive manual labeling [34]. This capability highlights the potential of pre-trained models to streamline processes in highly specialized fields, ensuring accurate and scalable solutions across different datasets and institutions.

Additionally, research has shown that integrating pre-trained language model representations into sequence-to-sequence architectures can yield substantial gains in tasks like neural machine translation and abstractive summarization. For example, incorporating pre-trained embeddings into the encoder network of transformer models has proven to enhance translation accuracy significantly, particularly in low-resource settings, demonstrating improvements in BLEU scores and overall model performance [35]. These advancements underscore the profound impact of pre-trained models on enhancing the quality and efficiency of language generation and understanding tasks.

In the realm of e-commerce, pre-trained models have been effectively employed to extract structured data, such as product attribute values, from unstructured text, thereby enabling better product search and comparison features. Techniques leveraging models like GPT-4 have shown superior performance in zero-shot and few-shot scenarios, outperforming traditional PLM-based methods and offering more robust solutions for handling diverse product descriptions [36]. These developments highlight the transformative role of pre-trained models in optimizing various applications, from improving user experience in e-commerce to facilitating more personalized and accurate recommendations in healthcare [37].

Estructured data models

Structured data models within e-commerce platforms has evolved significantly with the advent of advanced machine learning techniques and large language models (LLMs), which have been instrumental in enhancing the extraction and utilization of structured data such as product attribute values from unstructured text. In the realm of e-commerce, structured data models are critical for enabling features like faceted product search and product comparison, which rely heavily on accurately extracted attribute/value pairs from product descriptions provided by vendors [36]. Traditional methods based on pre-trained language models (PLMs) such as BERT have faced limitations, particularly in generalizing to unseen attribute values and requiring extensive task-specific training data [36]. However, recent advancements with LLMs like GPT-4 and Llama2 have shown superior performance in both zero-shot and few-shot scenarios, offering more robust and training data-efficient solutions for attribute extraction [36].

Moreover, the integration of synthetic data generation techniques using LLMs has further enhanced the quality and diversity of training datasets, thereby improving the performance of structured data models in real-world applications. For instance, in the context of resume classification, synthetic data generated by LLMs such as ChatGPT has been utilized to augment real-world datasets, resulting in significant improvements in model accuracy and robustness across various job categories [38]. This approach not only addresses the challenge of data sparsity but also ensures that the models are well-equipped to handle diverse and complex data inputs.

Furthermore, the application of LLMs in structured data models extends beyond e-commerce, encompassing various domains such as job market analysis and resume classification. The use of LLMs for generating synthetic resume data has demonstrated their potential in rapidly creating high-quality training data, which is crucial for improving the performance of classification models in scenarios with limited real-world data [38]. By leveraging LLMs' ability to understand and generate human-like text, these models can effectively extract and classify structured data, thereby enhancing the overall efficiency and accuracy of automated systems in various applications [39].

E-commerce models

E-commerce recommendation systems and product description generation has advanced significantly with the integration of large language models (LLMs) such as BERT, LLAMA 2.0, and specialized adaptations like E-BERT, which have revolutionized natural language processing and artificial intelligence in this domain. Leveraging LLMs' capabilities, researchers have enhanced recommendation accuracy by incorporating user and item interactions, metadata, and multimodal signals, enabling better personalization and generalization across different recommendation scenarios [40]. Specifically, E-BERT has shown promising results by incorporating phrase-level and product-level domain knowledge through techniques such as Adaptive Hybrid Masking and Neighbor Product Reconstruction, effectively improving tasks like review-based question answering, aspect extraction, and product classification [41].

Moreover, the application of LLMs in generating enhanced product descriptions has been a game-changer for e-commerce platforms. For instance, LLAMA 2.0 has been fine-tuned on extensive datasets of product descriptions from leading e-commerce platforms like Walmart, significantly reducing human workload and increasing the consistency and scalability of product listings. This model has been validated using various metrics such as NDCG, click-through rates, and human assessments, proving its effectiveness in improving search visibility and customer engagement [42]. The integration of LLMs with traditional recommendation systems has also been explored, combining collaborative filtering algorithms with the superior natural language understanding of LLMs to provide more accurate and personalized recommendations, thereby enhancing user satisfaction and sales [40]. These advancements underscore the substantial potential of LLMs in automating and optimizing various facets of e-commerce, offering significant business impacts and setting the stage for future research and industrial applications in this domain [42].

Metrics for evaluation of performance in LLM models

Evaluating the performance of large language models (LLMs) requires a comprehensive set of metrics that capture various dimensions of their capabilities, from accuracy in natural language processing tasks to efficiency in resource utilization. Traditional metrics such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores have been extensively used to assess the quality of machine translation and text summarization outputs by comparing them to reference texts, highlighting the models' ability to produce coherent and relevant responses [43]. Additionally, metrics like perplexity measure how well a language model predicts a sample, reflecting the model's ability to handle the complexity and variability of natural language [43].

In more specialized applications, such as mathematical reasoning and logical inference, unique metrics have been developed to evaluate the models' performance. For instance, the accuracy of LLMs in solving mathematical problems or performing multi-step reasoning tasks can be assessed using custom benchmarks that test their ability to follow logical steps and produce correct results [44] [45]. The application of information entropy-based metrics has been proposed to quantify the uncertainty and confidence levels in the models' reasoning processes, providing deeper insights into their decision-making abilities [46].

Moreover, in the context of multi-modal pre-trained models, which integrate textual and visual data, performance evaluation expands to include metrics that assess the models' ability to understand and generate responses based on diverse inputs. Metrics such as image captioning scores, visual question answering accuracy, and multi-modal retrieval metrics are crucial in evaluating how well these models integrate and process information across different modalities [47]. As LLMs continue to evolve and be applied across various domains, the development and adoption of robust, context-specific metrics remain essential for accurately assessing their performance and guiding further improvements [48].

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] K. Macková and M. Pilát, "Promap: Datasets for product mapping in e-commerce," 2023.

[11] M. Muntjir and A. T. Siddiqui, "An enhanced framework with advanced study to incorporate the searching of e-commerce products using modernization of database queries," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.070514

[12] J.-H. Liang, "Application of big data technology in product selection on cross-border e-commerce platforms," *Journal of Physics: Conference Series*, vol. 1601, no. 3, p. 032012, jul 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1601/3/032012

[13] W.-K. Tan and H.-H. Teo, "Productpedia – a collaborative electronic product catalog for ecommerce 3.0," in *HCI in Business*, F. Fui-Hoon Nah and C.-H. Tan, Eds. Cham: Springer International Publishing, 2015, pp. 370–381.

[14] G. Ryali, S. S, S. Kaveri, and P. M. Comar, "Trendspotter: Forecasting e-commerce product trends," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, ser. CIKM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 4808–4814. [Online]. Available: https://doi.org/10.1145/3583780.3615503

[15] 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.

[16] 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.

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

[18] 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.

[19] J. P. Lalor, H. Wu, and H. Yu, "Improving machine learning ability with fine-tuning," *ArXiv*, vol. abs/1702.08563, 2017.

[20] L. Catani and M. Leifer, "A mathematical framework for operational fine tunings," *Quantum*, vol. 7, p. 948, 2020.

[21] G. Shachaf, A. Brutzkus, and A. Globerson, "A theoretical analysis of fine-tuning with linear teachers," *ArXiv*, 2021.

[22] G. Vrbancic and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196 197–196 211, 2020.

[23] G. Xiao, J. Lin, and S. Han, "Offsite-tuning: Transfer learning without full model," *ArXiv*, vol. abs/2302.04870, 2023.

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

[25] 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

[26] E. Reiter, "A structured review of the validity of bleu," *Computational Linguistics*, vol. Just Accepted, pp. 1–8, 2018.

[27] J.-P. Ng and V. Abrecht, "Better summarization evaluation with word embeddings for rouge," *ArXiv*, vol. abs/1508.06034, 2015.

[28] K. A. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," *ArXiv*, vol. abs/1803.01937, 2015.

[29] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output," *ArXiv*, pp. 115–118, 2008.

[30] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, W. Dai, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, pp. 1 – 38, 2022.

[31] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ArXiv*, vol. abs/2311.05232, 2023.

[32] Y. Xiao and W. Wang, "On hallucination and predictive uncertainty in conditional language generation," *ArXiv*, vol. abs/2103.15025, 2021.

[33] C. Chen, Y. Yin, L. Shang, X. Jiang, Y. Qin, F. Wang, Z. Wang, X. Chen, Z. Liu, and Q. Liu, "bert2BERT: Towards reusable pretrained language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan,

P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2134–2148. [Online]. Available: https://aclanthology.org/2022.acl-long.151

[34] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," 2022.

[35] S. Edunov, A. Baevski, and M. Auli, "Pre-trained language model representations for language generation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4052–4059. [Online]. Available: https://aclanthology.org/N19-1409

[36] A. Brinkmann, R. Shraga, and C. Bizer, "Product attribute value extraction using large language models," 2024.

[37] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," 2024.

[38] P. Skondras, P. Zervas, and G. Tzimas, "Generating synthetic resume data with large language models for enhanced job description classification," *Future Internet*, vol. 15, no. 11, p. 363, 2023.

[39] X. Tang, Y. Zong, J. Phang, Y. Zhao, W. Zhou, A. Cohan, and M. Gerstein, "Struc-bench: Are large language models really good at generating complex structured data?" 2024.

[40] X. Xu, Y. Wu, P. Liang, Y. He, and H. Wang, "Emerging synergies between large language models and machine learning in ecommerce recommendations," 2024.

[41] D. Zhang, Z. Yuan, Y. Liu, F. Zhuang, H. Chen, and H. Xiong, "E-bert: A phrase and product knowledge enhanced language model for e-commerce," 2021.

[42] J. Zhou, B. Liu, J. N. A. Y. Hong, K. chih Lee, and M. Wen, "Leveraging large language models for enhanced product descriptions in ecommerce," 2023.

[43] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "Opt: Open pre-trained transformer language models," 2022.

[44] Z. Yuan, H. Yuan, C. Li, G. Dong, K. Lu, C. Tan, C. Zhou, and J. Zhou, "Scaling relationship on learning mathematical reasoning with large language models," 2023.

[45] A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," 2022.

[46] C. Zhou, W. You, J. Li, J. Ye, K. Chen, and M. Zhang, "INFORM : Information eNtropy based multi-step reasoning FOR large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3565–3576. [Online]. Available: https://aclanthology.org/2023.emnlp-main.216

[47] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," 2024.

[48] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024.