Natural Language Processing Journal 6 (2024) 100056

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation

Konstantinos I. Roumeliotis [a], Nikolaos D. Tselikas [a,*], Dimitrios K. Nasiopoulos [b]

[a] *Department of Informatics and Telecommunications, University of Peloponnese, Akadimaikou G.K. Vlachou Street, 22 131 Tripoli, Greece*
[b] *Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 118 55 Athens, Greece*

ARTICLE INFO

Keywords:
Sentiment analysis
LLMs
Instruction tuning
GPT model
LLaMA model
LLM fine-tuning

ABSTRACT

E-commerce has witnessed remarkable growth, especially following the easing of COVID-19 restrictions. Many people, who were initially hesitant about online shopping, have now embraced it, while existing online shoppers increasingly prefer the convenience of e-commerce. This surge in e-commerce has prompted the implementation of automated customer service processes, incorporating innovations such as chatbots and AI-driven sales. Despite this growth, customer satisfaction remains vital for E-commerce sustainability. Data scientists have made progress in utilizing machine learning to assess satisfaction levels but struggled to understand emotions within product reviews' context. The recent AI revolution, marked by the release of powerful Large Language Models (LLMs) to the public, has brought us closer than ever before to understanding customer sentiment. This study aims to illustrate the effectiveness of LLMs by conducting a comparative analysis of two cutting-edge LLMs, GPT-3.5 and LLaMA-2, along with two additional Natural Language Process (NLP) models, BERT and RoBERTa. We evaluate the performance of these models before and after fine-tuning them specifically for product review sentiment analysis. The primary objective of this research is to determine if these specific LLMs, could contribute to understanding customer satisfaction within the context of an e-commerce environment. By comparing the effectiveness of these models, we aim to uncover insights into the potential impact of LLMs on customer satisfaction analysis and enhance our understanding of their capabilities in this particular context.

1. Introduction

In today's rapidly evolving digital landscape, the e-commerce sector has emerged as a significant economic force, with revenue projected to surge to an impressive US$4.18 trillion by 2024, signaling its crucial role in shaping the future of global commerce (*eCommerce - Worldwide Statista Market Forecast*, 0000). It is essential to emphasize that customer satisfaction remains and will continue to remain at the forefront of any domain associated with e-commerce (Li et al., 2023a). Nevertheless, comprehending the customer's emotional state and satisfaction following a purchase requires considerable human resources and ongoing advanced data analysis.

One approach to gauge customer satisfaction following a purchase is the evaluation and comprehension of each product review (Wang et al., 2018). Understanding the underlying significance of each product review holds paramount importance for e-commerce platforms, marketplaces, and product manufacturers (Engler et al., 2015). Questions like the customer's contentment, encountered issues during shipping or product utilization are among the inquiries e-commerce owners and manufacturers seek answers to, in order to optimize their services. However, analyzing each review individually, considering the multitude of product reviews associated with each product in an e-commerce platform, is a time-consuming and economically impractical endeavor (Liu et al., 2021).

The solution to the aforementioned issue is addressed by data analysts employing advanced statistical and machine learning models that systematically extract information from each review. Subsequently, this information is conveyed to the relevant team for the purpose of adjusting the e-commerce strategy. Some of the machine learning models previously utilized for related tasks included Transfer Learning Models (Wang and Li, 2021), Naive Bayes (Kang et al., 2012), Logistic Regression (Li et al., 2023b), Semantic Analysis (Ahmad and Laroche, 2023), and others. However, with the advancement of artificial intelligence, the advent of deep learning, and the revolution introduced to the world in early 2023 by LLMs, many tasks previously performed with older models can now be executed much more efficiently through the use of LLMs.

* Corresponding author.
E-mail addresses: k.roumeliotis@uop.gr (K.I. Roumeliotis), ntsel@uop.gr (N.D. Tselikas), dimnas@aua.gr (D.K. Nasiopoulos).

https://doi.org/10.1016/j.nlp.2024.100056
Received 14 November 2023; Received in revised form 15 January 2024; Accepted 16 January 2024

This research aims to evaluate the effectiveness of two prominent LLMs in comprehending and extracting information from product reviews. The LLMs utilized in this study are the GPT-3.5 by OpenAI (Brown et al., 2020) and the open-source LLaMA-2 by Meta (Touvron et al., 2023). Initially, the two base models will be assessed based on their performance in predicting product review ratings. Subsequently, these models will undergo fine-tuning using a large dataset of product reviews, and their performance will be assessed following the fine-tuning process. Concurrently, fine-tuning will be applied to two NLP models, BERT and RoBERTa, using the same training set. Comparisons will be made both within the NLP models and between the NLP models and the LLMs.

The primary aim of this research is twofold: firstly, to assess and compare the performance of the models, and secondly, to address specific research questions that have not been adequately answered by prior studies:

Q1: Which LLM demonstrates superior efficacy in assessing product reviews?

Q2: Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?

Q3: How significant is the process of fine-tuning LLMs for domain-specific tasks?

Q4: What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?

Q5: To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?

Q6: What impact does fine-tuning have on the generated output of LLMs?

Q7: Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?

Q8: Can LLMs be effectively utilized for the evaluation of product reviews, and how can LLMs revolutionize the e-commerce sector?

To address the aforementioned research questions, the paper initiates by providing a concise literature review in Section 2. This review aims to assist readers in gaining a deeper understanding of topics related to customer sentiment analysis and satisfaction, as well as the utilization of NLP and LLMs for this particular study. Section 3 outlines the methodology employed in facilitating the current research, encompassing the data collection of product reviews, dataset cleaning, the creation of a universal prompt, model execution, fine-tuning, and the utilization of these models to identify the lexical components that influenced both reviewers and LLMs in assigning ratings to product reviews. Section 4 presents the research findings, while Section 5 not only addresses the research questions, but also compares pre- and post-fine-tuning models. Key insights from the authors' observations are also highlighted in this section. Finally, the research concludes in Section 6, by discussing implications and providing directions for future research work.

2. Literature review

2.1. Customer satisfaction in E-commerce

In the realm of e-commerce, the foundation of success is customer satisfaction (Kumar and Ayodeji, 2021). In the constantly shifting world of online retail, where countless choices are merely a click away, contented customers play a pivotal role in ensuring the endurance and prosperity of e-commerce enterprises. Furthermore, customer satisfaction (CS) is closely linked with customer retention (CR) and loyalty (CL) (Murali et al., 2016). Happy customers are not only more inclined to return for future purchases, but also to become loyal advocates, creating a ripple effect by recommending the e-commerce platform to their network of friends and family (Kassim and Abdullah, 2008). Significantly, retaining an existing customer often costs less than acquiring a new one (Meire, 2021). Consequently, customer satisfaction emerges as an economically efficient strategy for sustained business growth (Rosli and Nayan, 2020).

In the interconnected digital realm, satisfied customers transform into influential brand ambassadors. They eagerly share their positive experiences through social media, review platforms, and word-of-mouth endorsements. These spontaneous recommendations can significantly enhance an e-commerce business's reputation and visibility, thereby attracting new customers and broadening its market reach (Utz et al., 2012).

Furthermore, customer satisfaction exhibits a direct connection with sales and revenue. Pleased customers tend to spend more, make repeat purchases, and show a higher lifetime customer value (Rosli and Nayan, 2020). Additionally, favorable reviews and ratings from satisfied customers actively stimulate sales by providing social proof and enhancing the perceived trustworthiness of the e-commerce platform (Roethke et al., 2020).

High customer satisfaction levels, therefore, act as a safeguard against customer attrition. E-commerce businesses operate in a fiercely competitive landscape, where dissatisfied customers or those experiencing subpar service are prone to switch to a competitor (Kumar and Ayodeji, 2021). Consequently, maintaining high levels of customer satisfaction becomes crucial not only to prevent customer loss, but also to ensure long-term business sustainability.

It is evident that customer satisfaction can serve as a crucial differentiator. E-commerce businesses that consistently provide exceptional customer experiences gain a competitive advantage, which, in turn, draws and retains customers (Pei et al., 2020). This competitive edge can lead to increased market share and, ultimately, enhanced profitability.

The feedback obtained from contented customers represents a valuable resource for e-commerce enterprises looking to innovate and enhance their offerings. Customer satisfaction surveys, reviews, and direct feedback offer essential insights into what is working well and what needs improvement (Griva, 2022). This information empowers e-commerce businesses to adapt and evolve in response to shifting customer preferences and market dynamics.

2.2. Product reviews in understanding customer sentiment

In today's consumer landscape, product evaluations have evolved into a pivotal element of decision-making processes. The rise of e-commerce and digital marketplaces has forged a deep reliance on reviews among customers seeking insights into a product or service's performance, quality, and overall satisfaction (Sundararaj and Rejeesh, 2021). These reviews serve as a wellspring of knowledge, guiding consumers toward well-informed purchase choices, molding perceptions of products and services, and holding businesses accountable for their offerings (Dwidienawati et al., 2020).

The digital realm, enriched by online platforms, social media, and e-commerce websites, has opened up a vast arena for individuals to share their encounters and viewpoints on various products and services. This profusion of product appraisals has revolutionized how consumers approach their buying decisions. A survey from Askalidis et al. (2017) underscores this shift, revealing that a staggering 95% of shoppers peruse online reviews before committing to a purchase, and the purchase likelihood for a product with five reviews is 270% greater than the purchase likelihood of a product with no reviews, which strongly underscores the surging significance of product evaluations in the consumer decision-making process.

Diverse sources, including e-commerce websites, social media platforms, specialized review sites, and mobile applications, serve as reservoirs of these product assessments. In these platforms, customers thoughtfully recount their experiences, meticulously outlining both the favorable and unfavorable aspects of a given product or service.

These appraisals take on a variety of forms, from textual narratives to numerical ratings, star-based evaluations, or a fusion of these elements (Guha Majumder et al., 2022).

Understanding customer sentiment from these reviews requires the application of sophisticated NLP and sentiment analysis methodologies. Sentiment analysis algorithms scrutinize the language, tone, and contextual nuances within the reviews, categorizing them as either positive, negative, or neutral (Ramaswamy and DeClerck, 2018). These algorithms leverages machine learning and linguistic analysis to extract valuable insights from textual data.

Product evaluations have a multifaceted impact on consumer behavior (Stephen, 2016). First and foremost, they furnish invaluable insights into a product's performance, features, and quality. These reviews bridge the information gap between customers and products, empowering consumers to make more judicious choices (Liu et al., 2011). Additionally, these reviews play a crucial role in building trust. Numerous positive evaluations for a product can establish a sense of credibility and reliability, while a lack of reviews or an excess of negative ones may trigger caution for potential buyers (Askalidis et al., 2017).

Product reviews serve not only consumers but also provide value to manufacturers and retailers. They act as feedback mechanisms, offering valuable insights into areas for improvement and gauging customer satisfaction. Companies can leverage this feedback to enhance their products and services, address customer concerns and suggestions, and adjust their marketing strategies accordingly (Engler et al., 2015).

2.3. Natural language processing and large language models

NLP plays a profound role in the field of artificial intelligence, as it empowers machines to not only comprehend but also interpret and manipulate human language (Roumeliotis et al., 2023b). NLP serves as the critical bridge between human communication and machine comprehension (Rothman, 2021). This comprehensive field encompasses a wide spectrum of tasks, ranging from speech recognition and text classification to sentiment analysis, machine translation, information extraction, and question answering. The techniques associated with NLP empower us to derive meaningful insights from vast amount of unstructured textual data, greatly facilitating efficient information retrieval, analysis, and decision-making processes (Li et al., 2022). The practical applications of NLP are diverse, and include the development of chatbots, virtual assistants, language translation services, content summarization tools, and sentiment analysis in the context of social media (Fanni et al., 2023). Through its ability to unlock the power of language, NLP fosters the creation of intelligent systems capable of understanding and communicating with humans in a manner that feels natural and intuitive.

2.3.1. Generative pre-trained transformer (GPT) model

GPT models, built upon the foundational Generative Pre-trained Transformer architecture, play a pivotal role in advancing the capabilities of NLP (de Curtò et al., 2023). These models have undeniably transformed the landscape of NLP, possessing the remarkable ability to capture and comprehend intricate linguistic structures, context, and semantic nuances (Roumeliotis and Tselikas, 2023). Their extensive pre-training on copious amounts of unlabeled textual data equips GPT models with a profound understanding of language patterns and relationships (Rothman, 2021). This pre-training empowers GPT models to generate coherent and contextually relevant text, further enhancing their effectiveness (Roumeliotis et al., 2023b). What sets GPT models apart is their adaptability; they can be fine-tuned for specific NLP tasks, thereby boosting their performance and applicability in various domains (Liu et al., 2023a). These models excel in a range of NLP applications, spanning text generation, language translation, summarization, and content completion (Roumeliotis and Tselikas, 2023). By leveraging the capabilities of GPT models, NLP systems reach new heights in terms of language comprehension, text generation quality, and the delivery of contextually relevant and personalized results to users (Liu et al., 2023a). The contributions of GPT models are instrumental in pushing the boundaries of NLP, fostering more intricate and efficient interactions between humans and intelligent systems (Zhang and Li, 2021).

2.3.2. LLaMA model

LLaMA-2 is an open-source state-of-the-art language model developed as a successor to the original LLaMA model (Touvron et al., 2023). It represents a significant leap forward in natural language understanding and generation, owing to its advanced architecture, larger training data, and refined training strategies. LLaMA-2 is designed to meet the growing demands of various NLP applications, including text generation, sentiment analysis, machine translation, and question-answering systems (Wu et al., 2023). The architecture of LLaMA-2 is based on the transformer model, a neural network architecture that has proven highly effective in a wide range of NLP tasks (Rozière et al., 2023). Like its predecessor, LLaMA-2 employs a multi-layered transformer architecture with self-attention mechanisms (Touvron et al., 2023). However, it boasts a significantly larger number of parameters (70B) and an increased model depth, enabling it to capture more complex linguistic patterns and nuances (Roumeliotis et al., 2023a). One of the key advancements in LLaMA-2 is the extensive training data used to train the model. It leverages a diverse and vast corpus of text from the internet, academic sources, and other text domains, which results in a model with a broad understanding of languages and their contextual usage (Touvron et al., 2023). This diverse dataset enables LLaMA-2 to perform exceptionally well in a variety of languages and across numerous domains (Li et al., 2023c; Rozière et al., 2023).

2.3.3. Fine-tuning LLMs

Generative AI is advancing rapidly, drawing the attention of a diverse group of enthusiasts, including scientists and enterprises. These stakeholders are actively exploring ways to seamlessly incorporate LLMs into their projects. The refinement of LLMs through fine-tuning is a crucial process that involves enhancing pre-trained models using smaller, task-specific datasets (Han et al., 2021). The goal of this iterative approach is to customize the model's capabilities and improve its performance in a specific task or domain (Tinn et al., 2023).

Fine-tuning essentially means converting generalized pre-trained models into specialized versions, serving as a vital bridge between generic language models and the unique requirements of specific applications (Han et al., 2021). This ensures a close alignment between the language model and human expectations. LLMs, designed for versatility, are intended to handle a wide range of NLP tasks. While these models can comprehend and generate general text, optimizing them for specific tasks, they may require a focused approach.

To enhance their effectiveness in a particular task, organizations or researchers may opt to fine-tune LLMs using datasets specific to the domain. This process allows the model to familiarize itself with domain-specific terminologies, nuances of task-related language, and common structures found in the training set (Han et al., 2021). After fine-tuning, LLMs are ready to provide precise and coherent responses or predictions, showcasing their adaptability to specific tasks (Kalyan, 2023).

In contrast to the initial pre-training phase, which involves large amounts of unstructured text data, fine-tuning operates as a supervised learning process. This means that users utilize labeled examples, typically in the form of prompt-response pairs, to adjust the weights of LLMs. This meticulous approach results in more refined task completion, representing a shift from the broad learning scope of the pre-training phase.

In supervised fine-tuning, a pre-trained language model is updated using labeled data tailored for a specific task (Tinn et al., 2023). This process differs from unsupervised methods, as supervised fine-tuning involves using previously verified data. While the initial training phase of the language model is typically unsupervised, fine-tuning introduces a supervised approach.

During fine-tuning, the model encounters a newly labeled dataset for the target task. It calculates the error between its predictions and the actual labels, adjusting its weights based on this error (Mujtaba and Sowgath, 2022). Typically, an optimization algorithm like gradient descent is commonly used for this purpose. The adjustments to the weights depend on the gradients, which indicate the contribution of each weight to the overall error. Weights more responsible for the error undergo larger adjustments, while those less implicated undergo more moderate changes.

Throughout multiple iterations or epochs of the dataset, the model refines its weights, converging toward a configuration that minimizes error for the designated task. The goal is to adapt the model's general knowledge to the specific nuances and patterns in the new dataset, thereby enhancing its specialization and effectiveness for the target task.

During this dynamic adaptation process, the model undergoes updates based on the labeled data, dynamically evolving in response to the variance between its predictions and the actual answers, enabling it to assimilate details present in the labeled data, resulting in improved performance for the task at hand (*Fine-tuning large language models (LLMs) in 2023 SuperAnnotate*, 0000).

There are various fine-tuning methods, each with its own approach. For instance, Instruction fine-tuning (IFT) involves training the model using prompt completion pairs, showing desired responses to queries (Peng et al., 2023). Full fine-tuning updates all of the model's weights, while Parameter-efficient fine-tuning (PEFT) selectively updates a small set of parameters, making memory requirements more manageable.

Transfer learning begins by training the model on large, general-purpose datasets and then refining it with specific, task-related data (Zhao et al., 2023). Task-specific fine-tuning refines the pre-trained model on a particular task or domain using a dedicated dataset (Scotta and Messina, 2023). Although it requires more data and time compared to transfer learning, it can result in higher performance on the targeted task.

Multi-task learning involves using a training dataset with examples for multiple tasks, promoting versatility in the model's capabilities (Sanh et al., 2021). Sequential fine-tuning (SFT) adapts a pre-trained model sequentially to related tasks. After the initial transfer to a general domain, the LLM may undergo fine-tuning on a more specific subset, tailoring its abilities to the intricacies of the designated tasks.

Research although has shown that the widely used instruction-tuning has limitations, especially in sequence and token classification tasks where precise label prediction is crucial. Corresponding studies suggest Label Supervised Finetuning (LSF) as a solution (Li et al., 2023e). This involves extracting latent representations from the final LLM layer, projecting them into the label space, and computing the cross-entropy loss. Using this specific technique, the results, in terms of both the speed of fine-tuning and prediction accuracy, are highly promising, with the models outperforming others ten times their size.

Consequently, it is evident that each task requires a distinct approach to fine-tuning, and only through relevant research can it be determined whether fine-tuning methods align with the specific needs of the task, or not.

2.3.4. LLMs in E-commerce

Despite being in an early stage compared to other NLP models, recent research has demonstrated the efficiency of LLMs in various tasks related to e-commerce. Kanaan et al. (2023) introduce a novel approach using GPT for product category matching, showcasing the superior effectiveness of the GPT-based model. Chen et al. (2023) explore LLMs for relation labeling in e-commerce Knowledge Graphs, demonstrating remarkable learning capabilities and the potential to replace human labeling.

Gao et al. (0000) address limitations of LLMs and Specialized Translation Models (STMs) in e-commerce translation, introducing an LLMs-based E-commerce machine translation approach (LEMT) that outperforms state-of-the-art Neural Machine Translation (NMT) models. Wang and Na (2023) propose a novel approach to enhance e-commerce search and recommendation systems by converting structured data into textual data, allowing for more effective search and recommendation using LLMs.

Al Wahshat et al. (2023) focus on combating manipulated reviews in e-commerce platforms using GPT-4, demonstrating its efficacy in identifying and flagging manipulated reviews with remarkable accuracy. Zhou et al. (2023) automate product description generation in e-commerce using the LLaMA 2.0 7B language model, emphasizing the potential of LLMs in optimizing e-commerce.

Continuing the research, Roumeliotis et al. (2023b) introduce a software powered by GPT to automatically identify sustainable product features in e-commerce, aiding informed, sustainable purchasing decisions. Shi et al. (2023) address limitations of general LLMs in handling personalized features unique to e-commerce, introducing LLaMA-E, achieving state-of-the-art results in e-commerce authoring tasks. Liu et al. (2023b) explore the synergy between conversational recommender systems (CRSs) and LLMs, highlighting their complementary strengths in improving pre-sales dialogue tasks.

Ma et al. (2023) tackle challenges in applying LLMs to specific domains by focusing on domain-specific continual pre-training, demonstrating its effectiveness in improving few-shot learning and zero-shot performance in the e-commerce domain. Maragheh et al. (2023) investigate using item aspects generated by LLMs to improve ranking tasks, introducing Theme-Aware Keyword Extraction (LLM TAKE) for enhanced keyword generation in e-commerce.

Finally, Orzoł and Szopik-Depczyńska (2023) explore how ChatGPT can enhance customer communication and boost sales in the e-commerce industry, providing insights into the potential benefits of ChatGPT for e-commerce stores. Li et al. (2023d) contribute by addressing challenges of applying general LLMs to e-commerce tasks, proposing the first E-commerce instruction dataset, EcomInstruct, and introducing EcomGPT, a specialized model demonstrating superior zero-shot generalization capabilities.

Collectively, these studies emphasize the substantial potential of LLMs in revolutionizing and enhancing various facets of the e-commerce landscape.

3. Research methodology

This study aims to unravel the intricate emotional nuances embedded within product reviews through the proficient application of LLMs, with a particular focus on the widely recognized GPT-3.5 model and the novel open-source LLaMA-2 model. Our primary objective is to gain a profound understanding of the psychological states that drive reviewers when composing their assessments, whether they manifest satisfaction, dissatisfaction, or a complex ambivalence toward the products they have purchased.

To evaluate how proficiently LLMs can determine the sentiments expressed in individual words within product reviews, we began our study by using these models to predict star ratings for a large dataset of product reviews. We hypothesized that since LLMs can efficiently predict the rating stars of a review, they also have the capability to understand the meaning and emotions conveyed by the words in each review.

To assess the effectiveness of LLMs for this specific task without bias, we concurrently employed two well-known NLP models (non-LLM): BERT and RoBERTa. These models were tasked with making predictions for the same tasks as the LLMs, generating valuable results for meaningful comparisons.

In the preliminary phases of our research, we meticulously scrutinize the ability of LLMs to accurately discern the intricate emotional subtleties expressed by product reviewers, conducting a comprehensive performance comparison between the two aforementioned models. Following this, we undertake the task of fine-tuning both LLMs and their non-LLM counterparts for sentiment analysis within the context of product reviews. This fine-tuning process enhances their capacity to respond with increased precision to the multifaceted emotional dimensions present in reviews.

Confirming through our research that fine-tuned LLMs exhibit significantly higher performance compared to their base counterparts, we advanced our investigation one step further by deploying fine-tuned models to identify specific lexical elements — words within product reviews that influenced their selection of review rating stars. We compiled these words into a table, which will aid future studies in comprehending how LLMs perceive the sentiment of a reviewer by analyzing the words they choose to employ in their product reviews.

The entire codebase utilized for the research, spanning from the Flask-based Chrome application designed for review collection to the fine-tuning process of the models, along with the associated datasets, is available in a GitHub repository with an MIT open source license (*GitHub - kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews*, 0000).

To accomplish the objectives of this paper, given the volume and complexity of the processes involved, it became imperative to adopt a specific research methodology. This methodology has been carefully selected to facilitate both the development of the software for review collection and the subsequent execution and fine-tuning of the models. The detailed framework for these procedures is presented in the following subsections to provide a comprehensive and coherent strategy for the study.

3.1. Chrome app development and data collection

The research was initiated with the development of a Chrome application, meticulously engineered using object-oriented Python, Flask, and JavaScript. This application was purposefully designed to facilitate the systematic extraction of product reviews from prominent online marketplaces, including industry giants like Amazon and eBay.

The Chrome application was crafted with precision to ensure its reliability and effectiveness in data retrieval. Leveraging object-oriented Python, a structured and modular codebase was established, allowing for efficient maintenance and future enhancements. The utilization of Flask, a Python web framework, streamlined the development process, facilitating seamless interactions between the frontend and backend components. Additionally, JavaScript played a pivotal role in collecting and transmitting data to the Flask server.

The primary objective of the application was to collect comprehensive information from online product listings. The researchers systematically gathered a multitude of data points, including product titles, product descriptions, product categories, and user-generated product reviews. The product titles and descriptions provided valuable insights into the attributes and features of the items under scrutiny. The product categories enabled efficient product classification and categorization. The focus, however, was on user reviews, extracting the review title, the complete body of the review, and the accompanying star rating given by the reviewer.

The acquired data, extensive and diverse, was systematically stored within an SQLite database. This relational database management system ensured data integrity and offered a structured environment for organizing and querying the collected information. Each data point, from product descriptions to user reviews, was indexed, allowing for efficient retrieval and analysis.

The extensive scope of the data collection efforts resulted in a substantial repository of information. The application successfully collected data from 616 distinct products, creating a robust dataset consisting of 5,029 product reviews.

Beyond being the foundation for the research project, the data collected through the Chrome application holds significance in both academic and practical contexts. This dataset provides numerous opportunities for subsequent analysis, including sentiment analysis, trend identification in product reviews, and the examination of potential correlations between product attributes and review ratings. Moreover, it provides valuable insights for businesses, enabling a deeper understanding of customer satisfaction and preferences in the highly competitive e-commerce landscape.

3.2. Dataset splitting and preprocessing

To ensure the quality and effectiveness of our predictive modeling and fine-tuning processes, we prepared the dataset, adhering to a systematic approach. These processes involved several key steps, each designed to enhance the suitability of the data for subsequent analysis.

Data preprocessing plays a pivotal role in enhancing the quality of the dataset. In this stage, we addressed several aspects of text data to make it more amenable for modeling. One of the primary tasks in data preprocessing involved the removal of special characters and extra white spaces, critical for maintaining data cleanliness and consistency. Special characters, if left unattended, can lead to noise in the dataset and hinder the effectiveness of NLP techniques.

Text normalization is also essential for ensuring that text data is consistent and standardized. As part of this process, we performed operations such as converting accented characters to their base form. This step is especially important for languages with diacritics, as it ensures that words with accents' variations are treated uniformly.

Furthermore, to ensure that text data is case-insensitive, we uniformly converted all text to lowercase. This avoids the distinction between, for example, "Review" and "review", ensuring that NLP models are not sensitive to the case of text.

In addition, in an effort to make the most of the textual information within the reviews, we merged the review titles and bodies. This consolidation step simplifies the input data and ensures that the complete content of each review is considered when making predictions.

In order to gauge the number of tokens within each review, essential for processing by a LLM, we employed the `tiktoken` Python library. This provided us with a close estimation of the number of tokens, ensuring that we stay within the capacity limits of the LLM while maintaining the integrity and completeness of the text.

In the realm of NLP tasks, splitting a dataset is pivotal for effective model development, refinement, and assessment. This practice plays a crucial role in achieving robust and reliable models capable of generalizing adeptly to new, unseen data across diverse linguistic contexts. In our specific approach, we partitioned our dataset into training (70%), validation (15%), and test (15%) sets, employing the `train_test_split` function from `scikit-learn` (training: 3520, validation: 754, test: 755) (*sklearn.modelselection.traintestsplit — scikit-learn 1.3.2 documentation*, 0000). This partitioning occurs in two stages: an initial split generates a training set and a temporary set, with the latter subsequently undergoing further division into validation and test sets. The training set serves a critical role during the fine-tuning phase, where the model acquires patterns, relationships, and representations from the input data, enabling it to make predictions or execute specific tasks. The validation set is crucial for the model tuning phase, involving adjustments to hyperparameters and configuration. Evaluating the model's performance on the validation set informs modifications that enhance generalization and curbed overfitting. Finally, the fine-tuned models employ the test set for making predictions, completing the comprehensive process of model fine-tuning and evaluation.

To highlight the importance of the dataset during the fine-tuning of NLP models, subsequent to splitting the dataset containing 5029 product reviews, we created a smaller dataset comprising approximately half of the training and validation data (training: 1757, validation: 365). This reduced dataset was then utilized to re-fine-tune the same base models, making predictions on the same test set (test: 755).

```
prompt = 'Assign integer star ratings (between 1 and 5) to the following product reviews.' \
         'Return your response in JSON format like this example:' \
         '{"rating1":integer, "rating2":integer, ...}.' \
         'Do not provide explanations or justifications for the ratings. Reviews"\n'
```

Fig. 1. Model-Agnostic prompt.

3.3. Creating a compelling prompt understandable to both LLMs

By employing well-established prompting engineering techniques, our goal was to create a prompt that is compatible with various LLMs, enhancing the accessibility of their output through our code. Our approach prioritizes not only the content of the prompt but also the formatting of the output.

Creating a universal prompt starts with a deep understanding of the diverse LLM landscape. Each model, such as GPT-3, GPT-4, LLaMA-2, and others, have unique characteristics, strengths, and limitations. Designing a prompt that can elicit meaningful responses from these models while maintaining accessibility is a complex challenge. To address this, we applied two well-established prompting engineering techniques, considering the idiosyncrasies of various LLMs (Zhang et al., 2024).

1. Model-Agnostic Content: Our approach involves developing a prompt that is model-agnostic, meaning it does not rely on the specific architecture or knowledge of any particular LLM. This enables our prompt to be adaptable and transferable across a range of models. We focus on creating a prompt that convey the task clearly, providing relevant context and information that any LLM can comprehend.

2. Accessibility through Output Formatting: An often-overlooked aspect of prompt design is the format of the output. We strongly emphasize designing a code-friendly output format that enhances accessibility. This is achieved by structuring the responses in a coherent and intuitive manner. For the specific requirements of this task, it was necessary that the outputs be structured in JSON format.

After numerous iterations and experiments with both LLMs, the final prompt, comprehensible to both models, effectively elicited responses in the desired output format, is illustrated in Fig. 1.

3.4. Model deployment, fine-tuning LLMs and predictive testing

In this study, four NLP models were employed: the BERT model, the RoBERTa model, and the LLMs GPT-3.5 and LLaMA-2. For each of these models, a specific methodology was followed for both fine-tuning and product review rating prediction.

3.4.1. GPT-3.5 model

In this phase, the `gpt-3.5-turbo-1106` base model was initially employed to make predictions for the review ratings in the test set using a specific prompt presented in Fig. 1. Leveraging their extensive training, LLMs have acquired the ability to discern subtle cues influencing reviewer sentiment, enabling them to generate predicted star ratings. Our methodology involved inputting the language model with the textual content of reviews, allowing it to comprehend nuances, sentiments, and underlying themes in both review titles and their bodies. To make predictions and perform fine-tuning on the `gpt-3.5-turbo-1106` base model, we utilized the official OpenAI API. Additionally, based on statements from Azure CTO Mark Russinovich (*What runs ChatGPT? Inside Microsoft's AI supercomputer Featuring Mark Russinovich - YouTube*, 0000), Azure employs Low-Rank Adaptation (LoRA) Parameter-Efficient Fine-Tuning (PEFT) and DeepSpeed techniques to reduce GPU usage and enhance memory efficiency during the fine-tuning of its GPT-3 model.

The predicted ratings generated by GPT-3.5 were then meticulously compared with the original ratings provided by the reviewers. This comprehensive analysis allowed us to evaluate the effectiveness of the base model in capturing the essence of reviews and predicting ratings that align closely with human judgments.

During the fine-tuning phase, the `gpt-3.5-turbo-1106` model was exposed to extensive data and fine-tuned to adapt to specific nuances and patterns in the training set. The multi-epoch training approach allowed the model to iteratively refine its understanding and capabilities, progressively enhancing its performance on target tasks and ensuring readiness for accurate predictions and valuable insights in subsequent evaluations.

Two fine-tuning processes were conducted, one identified by the Job ID `ftjob-fduMpiYlj23jZ17b9wXooyBd` and the other (reduced dataset) by `ftjob-P6xOmuaYW4JiIxkTkCg6ytpF`. The former processed 1,468,176 tokens with training loss at 0.0574 and validation loss at 0.0216 across three epochs. The latter processed 730,491 tokens with a training loss of 0.3690 and a slightly higher validation loss of 0.3927 in three epochs. These metrics indicate the models' adaptability and generalization during fine-tuning.

After the completion of the fine-tuning phase, the fine-tuned models were assigned the task of predicting star ratings for product reviews in the same test set. The obtained results were then integrated into an `SQLite` table, simplifying further comparative analysis.

To facilitate the fine-tuning of the LLMs, two `JsonL` files were created, encompassing prompt and completion pairs, as illustrated in Fig. 2.

3.4.2. LLaMA-2 model

In this phase, the `llama-2-70b-chat` base model was initially used to predict review ratings in the test set using a specific prompt from Fig. 1. Predicted ratings from `llama-2-70b-chat` were carefully compared with the original ratings provided by reviewers. Given the significant computational resources required by LLaMA models, similar to GPT-3.5, we leveraged the Replicate API for predictions and fine-tuning. Similar to Azure's use of LoRA PEFT techniques to reduce GPU requirements during the fine-tuning of its GPT-3 model, Replicate employs LoRA – specifically, QLora techniques – for the fine-tuning the LLaMA model.

Two fine-tuning processes were performed, identified by the Job IDs `ecommerce-reviews5029:4a107315` (*kroumeliotis/ecommerce-reviews5029 – Run with an API on Replicate*, 0000) and (reduced dataset) `ecommerce-reviews50:b05681bd` (*kroumeliotis/ecommerce-reviews50 – Run with an API on Replicate*, 0000). The first job utilized 8x A40 (Large) GPUs, with a total runtime of 377.02 min. Training lasted 22,394.4514 s, achieving a speed of 0.465 samples per second, 0.116 steps per second, and a final training loss of 0.462 over three epochs. In the second job, the fine-tuning process also utilized 8x A40 (Large) GPUs, with a total runtime of 182.11 min. Training lasted 10,708.3497 s, achieving a speed of 0.478 samples per second, 0.12 steps per second, and a final training loss of 0.4918 over three epochs.

After the completion of fine-tuning, the fine-tuned models predicted star ratings for product reviews in the test data, and the results were also integrated into an `SQLite` table for further comparative analysis.

3.4.3. BERT model

In this phase, we utilized the `bert-base-uncased` variant (*bert-base-uncased . Hugging Face*, 0000) for our NLP task. Specifically, we employed the `BertForSequenceClassification` model from the transformers library. The foundational BERT model captures contextualized representations of input tokens through its transformer layers. `BertForSequenceClassification`, a distinct iteration of BERT,

```
# Llama 2 JsonL Sample
{"prompt": "You are a product reviewer. Assign integer star ratings (between 1 and 5)"
           "to the following product reviews."
           "Return your response in JSON format like this example:"
           "{\"rating1\":integer, \"rating2\":integer, ...}."
           "Do not provide explanations or justifications for the ratings. Reviews\"\n"
           "1. almost perfect. these would be perfect if they had pockets",
 "completion": "{\"rating1\":4}"}

# GPT 3.5 JsonL Sample
{"messages": [{"role": "system", "content": "You are a product reviewer"},
              {"role": "user", "content":
                  "Assign integer star ratings (between 1 and 5)to the following product reviews."
                  "Return your response in json format like this example"
                  "{'rating1':integer,'rating2':integer,...}."
                  "Please avoid providing additional explanations. Reviews:\n"
                  "1. almost perfect. these would be perfect if they had pockets"},
              {"role": "assistant", "content": "{\"rating1\":4}"}]}
```

Fig. 2. Fine-tuning JSONL samples.

incorporates an additional classification head pre-configured for sequence classification tasks. Typically, this classification head consists of a fully connected layer that transforms the BERT output into class probabilities (*BERT — transformers 3.0.2 documentation*, 0000a).

The `bert-base-uncased` model is a case-insensitive iteration of BERT, wherein all input text is converted to lowercase during training. Similar to the broader BERT architecture, this variant is transformer-based, featuring multiple layers and hidden units. The architectural specifications of the `bert-base-uncased` model include 12 layers, 768 hidden units, 12 heads, and 110 million parameters (*Pretrained models — transformers 3.3.0 documentation*, 0000). The self-attention mechanisms integrated into BERT facilitate the capture of contextual dependencies within input sequences.

During the fine-tuning process, the `bert-base-uncased` model was employed for sentiment classification using the training set comprising reviews and their corresponding ratings. The fine-tuning and prediction processes were executed in Google Colab, utilizing an A100 GPU. The dataset, loaded from a CSV file in Google Drive, underwent preprocessing, which included tokenization using the BERT tokenizer (*BERT — transformers 3.0.2 documentation*, 0000b). Hyperparameteres for the fine-tuning process were specified, including a learning rate of `2e-5`, a batch size of 8, and utilizing the Adaptive Moment Estimation (Adam), Adam with Weight Decay (AdamW) and Stochastic Gradient Descent (SGD) optimizers. The model underwent training for three epochs, with progress tracked using the `tqdm` library (*tqdm . PyPI*, 0000). The training loop encompassed backpropagation, optimization, and validation on a separate dataset. The code checked for GPU availability for faster computation. Following training, the fine-tuned model and tokenizer were saved to a directory in Google Drive for future use and predictions. Our code embodies a systematic and thorough fine-tuning methodology, assuring the model's alignment with the unique requirements of the sentiment classification task and its preparedness for deployment. After the fine-tuning process, the model proceeded to generate predictions for the test sample. The corresponding code, the training and validation loss, and the validation accuracy for this process are available in an `ipynb` Jupyter file on GitHub (*GitHub - kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews*, 0000).

3.4.4. RoBERTa model

In this phase, we employed the `roberta-base` model, accessible at *roberta-base . Hugging Face* (0000), for sentiment classification. Following the BERT-base architecture, the RoBERTa model architecture incorporates numerous layers of self-attention mechanisms, facilitating the effective capture of contextual relationships within input sequences. The specific configuration of the RoBERTa model involves a 12-layer transformer-based neural network with 768 hidden units, 12 attention heads, and a total of 125 million parameters (*Pretrained models — transformers 3.3.0 documentation*, 0000). Each transformer layer within this architecture typically integrates self-attention mechanisms and feedforward neural networks, collectively enhancing the model's proficiency in comprehending and representing intricate relationships within sequential data. For the prediction phase, the `RobertaForSequenceClassification` model was utilized from the transformers library (*RoBERTa — transformers 2.9.1 documentation*, 0000), involving the loading of the pre-trained model through the pretrained method.

The same fine-tuning methodology like the BERT model followed for the RoBERTa model, tailored for sentiment classification using a custom dataset of reviews and corresponding ratings. The fine-tuning and prediction processes were also executed in Google Colab, utilizing an A100 GPU. The training set, sourced from a CSV file in Google Drive, undergoes preprocessing, employing the RoBERTa tokenizer to tokenize text data and preparing it as PyTorch tensors. The hyperparameters governing the fine-tuning process, including a learning rate of `2e-5`, a batch size of 8, and the AdamW, Adam, and SGD optimizers, are specified to optimize model parameters effectively. The training unfolds over three epochs, with GPU utilization for accelerated computation when available. The training loop, visualized using the tqdm library, encompasses both training and validation phases, evaluating the model's accuracy. Post-fine-tuning, the RoBERTa model and tokenizer are saved for future use in a designated Google Drive directory. Subsequently, the model engaged in the generation of predictions for the test set. The script and associated resources, such as the training and validation loss, as well as the validation accuracy, are available on GitHub in a Jupyter notebook file (ipynb) for broader accessibility and collaboration (*GitHub - kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews*, 0000; *kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews*, 0000).

3.5. Linguistic factors influencing product review ratings

In the final phase of the research, the fine-tuned models trained on the entire training set were selected, along with those reviews to which the LLMs assigned a rating equal to that provided by the human reviewer. The fine-tuned LLMs were then prompted to identify specific lexical components, specifically the words contained within product

```
prompt = f'''Please carefully analyze the review text provided and identify the words or phrases
that played a significant role in influencing both your and the reviewer's rating decision.
Your response should be in JSON format, listing the words or phrases
and their corresponding influence on the rating.
For example: {{"1": "word or phrase", "2": "word or phrase", "3": "word or phrase", ...}} \n
Review text: {review_text}\n
Review rating: {review_rating}'''
```

Fig. 3. Model-agnostic prompt for identifying lexical elements in product reviews.

Table 1
Model performance metrics comparison.

Model	Accuracy	Precision	Recall	F1
base:gpt-3.5-turbo-1106	0.5483	0.553	0.5483	0.5467
base:llama-2-70b-chat	0.5086	0.5029	0.5086	0.5038
ft:gpt-3.5-turbo-1106 (100%)	0.6424	0.6416	0.6424	0.6409
ft:llama-2-70b-chat (100%)	0.6185	0.6162	0.6185	0.6147
ft:bert-adam (100%)	0.5881	0.5881	0.5881	0.5871
ft:bert-adamw (100%)	0.5841	0.5921	0.5841	0.5766
ft:bert-sgd (100%)	0.2225	0.0706	0.2225	0.0888
ft:roberta-adam (100%)	0.6066	0.6157	0.6066	0.6017
ft:roberta-adamw (100%)	0.5536	0.5951	0.5536	0.5325
ft:roberta-sgd (100%)	0.2225	0.0495	0.2225	0.081
ft:gpt-3.5-turbo-1106 (50%)	0.6212	0.6207	0.6212	0.6197
ft:llama-2-70b-chat (50%)	0.604	0.6009	0.604	0.5997
ft:bert-adam (50%)	0.5523	0.5575	0.5523	0.5532
ft:bert-adamw (50%)	0.5775	0.5741	0.5775	0.5736
ft:bert-sgd (50%)	0.1629	0.0269	0.1629	0.0462
ft:roberta-adam (50%)	0.5152	0.5305	0.5152	0.5181
ft:roberta-adamw (50%)	0.5497	0.5351	0.5497	0.5286
ft:roberta-sgd (50%)	0.1642	0.027	0.1642	0.0463

reviews that significantly influenced the assignment of rating stars. For both models, a universal prompt was used, as presented in Fig. 3, and their responses were saved in a CSV file to facilitate analysis.

This analysis aimed to uncover and clarify how specific words or phrases significantly influenced the overall assignment of rating stars by reviewers. The ultimate goal was to enhance our understanding of the linguistic factors that impact the determination of product review ratings, thus improving our insight into the decision-making processes employed by both human reviewers and fine-tuned LLMs.

4. Comparative analysis and models' evaluation

In Section 3, we elaborate on the methodological framework employed to evaluate the predictive capabilities of four NLP models, including GPT-3.5 and LLaMA-2 LLMs, BERT, and RoBERTa, in the context of star-rating product reviews and understanding their underlying nuances. This section presents the results of the comparative analysis conducted among the four models at various stages of our research.

4.1. Fine-tuned models evaluation

Before presenting our findings, it is crucial to emphasize the importance of model evaluation. Within the domains of machine learning and NLP, the process of evaluating models assumes a central role, affording us invaluable insights into the performance and effectiveness of our fine-tuned models. Model evaluation acts as a guiding compass, enabling us to make well-informed decisions regarding their suitability and, in parallel, propelling us toward progressive advancements in fine-tuning and optimization tailored to specific applications. Table 1 presents a comprehensive range of evaluations for each model, encompassing a set of critical evaluation metrics.

4.2. Base models evaluation phase

In the initial phase of our research, we engaged both LLM base models, namely `gpt-3.5-turbo-1106` and `llama-2-70b-chat`, in the task of predicting the star ratings associated with the test set comprising 755 product reviews. Our primary objective was to assess the accuracy of these LLMs in predicting the star ratings originally assigned by users when crafting their reviews.

Upon careful examination of the responses generated by the base models and comparing them to the original star ratings provided by users, we made several notable observations. The `gpt-3.5-turbo-1106` model exhibited a satisfactory grade of accuracy, precisely predicting 54.83% of the star ratings, which is translated to a successful prediction in 414 out of the 755 reviews that constituted our test dataset. In contrast, the `llama-2-70b-chat` model, while still displaying noteworthy performance, achieved a slightly lower accuracy rate of 50.86%, i.e., accurate predictions for 384 out of the 755 reviews in our test set. These findings suggest that both LLMs possess satisfactory predictive capabilities, with the GPT-3.5 base model showing a marginally higher accuracy in this specific task.

It is essential to emphasize that achieving a 50% accuracy rate in predictive modeling is akin to making predictions equivalent to random chance. However, when dealing with a classification task involving more than two choices, such as in the case of review ratings with 5 options, the baseline accuracy due to random chance is calculated to be 20% (1 out of 5).

In practical terms, this implies that a predictive model that simply guesses without any learning or information should, on average, be correct 20% of the time. Therefore, models that surpass this baseline are providing more meaningful predictions. For instance, the LLM base models achieved prediction accuracies of 54.83% and 50.86%, respectively, demonstrating a considerable improvement over random chance in discerning the nuances of the given reviews.

These accuracy percentages signify that the models are making predictions significantly above what one would expect from random guessing, indicating their ability to capture patterns and relationships within the data, thereby enhancing the quality of predictions.

However, the efficiency of the models can be further enhanced by fine-tuning them for the specific task, improving their ability to capture and understand intricate patterns and relationships in the data.

4.3. Fine-tuned LLMs evaluation phase

In the subsequent phase of our study, we conducted fine-tuning on the GPT-3.5 and LLaMA-2 models, simultaneously applying the same fine-tuning process to the BERT and RoBERTa models using three different optimizers. This process was carried out on the training set, consisting of 3,520 product reviews, with the objective of improving the models' performance and enhancing their adaptation to the task at hand.

Subsequently, the fine-tuned models were put to the test once more, tasked with the challenge of predicting star ratings for the same test dataset. The results of this phase were quite effective. The LLaMA-2 model, post fine-tuning, managed to correctly predict star ratings in a 61.85% of cases, i.e., 467 out of the 755 reviews included in the test set,

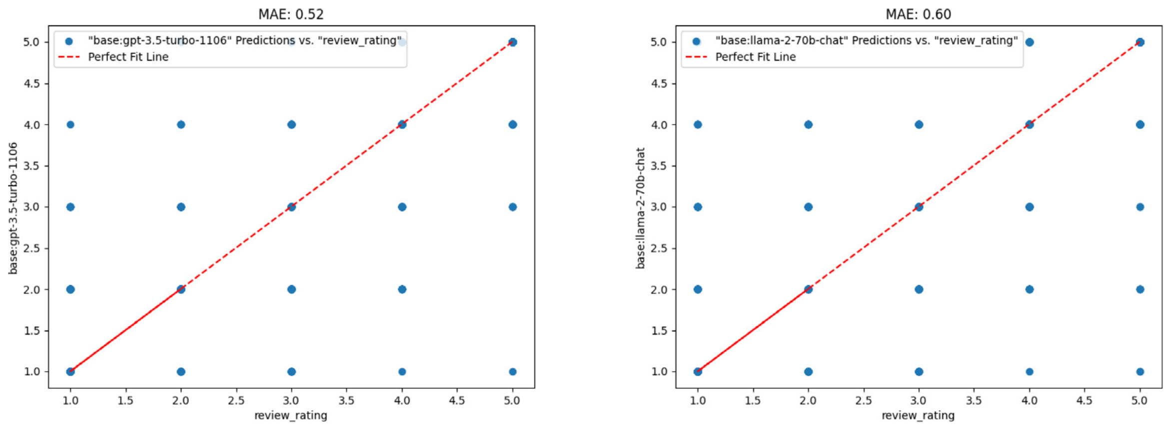

Fig. 4. MAE for base models.

showcasing a considerable improvement in its predictive capabilities. In contrast, the GPT-3.5 model demonstrated even more noteworthy progress following the fine-tuning process. It achieved a remarkable 64.24% accuracy rate, i.e., accurately predicting star ratings in 485 out of the 755 reviews within the test set. This substantial improvement in predictive accuracy highlights the effectiveness of fine-tuning in enhancing the performance of both LLMs, with GPT-3.5 exhibiting a particularly strong response to this optimization process.

At the same time, the non-LLM models responded to fine-tuning to a similar extent, with the BERT model achieving 58.81% correct predictions and RoBERTa reaching 60.66% using the ADAM optimizer. With the ADAMW optimizer, both models showed a slightly lower performance on predictions, i.e., 58.41% and 55.36%, respectively. In contrast, with the SGD optimizer, the results were highly disappointing at 22.25% for both models, indicating that this particular optimizer is not suitable for the given task.

4.4. Evaluation of fine-tuned LLMs for half of the dataset

To assess the impact of the amount of training data in the fine-tuning process of an NLP model, we employed the same fine-tuning procedure on four models. However, this time, we utilized approximately 50% of the training data, specifically 1,757 samples out of the previously employed 3,520.

In this specific phase, the GPT-3.5 fine-tuned model continued to maintain its edge with a 62.12% accuracy rate in predicting star ratings, correctly predicting 469 out of the 755 test samples. Meanwhile, the LLaMA-2 model achieved a 60.4% accuracy, correctly predicting star ratings in 456 out of the 755 test samples.

In a corresponding manner, NLP models (non-LLM), fine-tuned on approximately half of the dataset, yielded BERT predictions of 52.23% and RoBERTa predictions of 51.52%, employing the ADAM optimizer. Employing the ADAMW optimizer resulted in predictions of 57.75% and 54.97%, whereas utilization of the SGD optimizer led to predictions of 16.29% and 16.42%.

4.5. Assessing models' performance and proximity with original ratings

In order to gain meaningful insights into the performance of base and fine-tuned models and the degree to which their predictions align with the original ratings provided by reviewers, we utilized the `mean_absolute_error` class from the `sklearn` library. The Mean Absolute Error (MAE) was used to calculate the difference between the predictions of the base and fine-tuned models and the original ratings provided by the reviewers. The results are presented in Figs. 4 and 5 in the form of scatterplots. In these diagrams, the perfect fit line reflects the points where predictions are in perfect alignment with the actual values.

The MAEs provide insights into the accuracy of various NLP models across different fine-tuning configurations. A lower MAE suggests that, on average, the model's predictions are closer to the true values, indicating better accuracy, predictive capabilities, and overall performance. The base `gpt-3.5-turbo-1106` and base `llama-2-70b-chat` models exhibit MAEs of 0.5166 and 0.6026, respectively, indicating the average absolute differences between their predictions and true values. When fine-tuned on the full training set (100%), both GPT and LLaMA models show improvements, achieving much lower MAEs of 0.3974 and 0.4755, respectively.

Furthermore, the fine-tuned versions of BERT and RoBERTa on the same training set, with the ADAM optimizer providing better predictions according to our statistic results, yield MAEs of 0.5046 and 0.4741, respectively. Notably, in these evaluations, the RoBERTa model outperforms the BERT model.

4.6. Evaluation of emotional nuances and lexical elements

In accordance with the methodology and logic presented in Section 3.5, the fine-tuned LLMs were called upon to distinguish those words and phrases within the product reviews that led them to correctly predict the reviewers' ratings. For the purposes of the research, a sample of 5 reviews was chosen in which the predictions of both fine-tuned models matched the ratings given by the reviewers. The results were highly interesting and are presented in Table 2.

In the majority of the reviews presented in Table 2, the words identified by the fine-tuned LLMs were the same. At the same time, when viewed from a human perspective, these words are indeed the most representative and justify the reviewers' ratings, highlighting the potential of LLMs to identify the most important words in a text in a similar way to humans. Undoubtedly, making a sweeping generalization would lack objectivity. However, judging from human evaluation, it appears that LLMs can understand the context behind lexical elements to a lesser but encouraging extent, similar to humans.

5. Research findings

In the preceding sections, we presented, the methodology and results for evaluating the predictive capabilities of GPT-3.5 and LLaMA-2 LLMs, as well as BERT and RoBERTa models in tasks related to the star-rating assessment of e-commerce product reviews. In this section, the research findings and the observations made by the authors that have emerged from the study concerning the utilization of LLMs for sentiment analysis on product reviews are being presented, while also addressing the research questions.

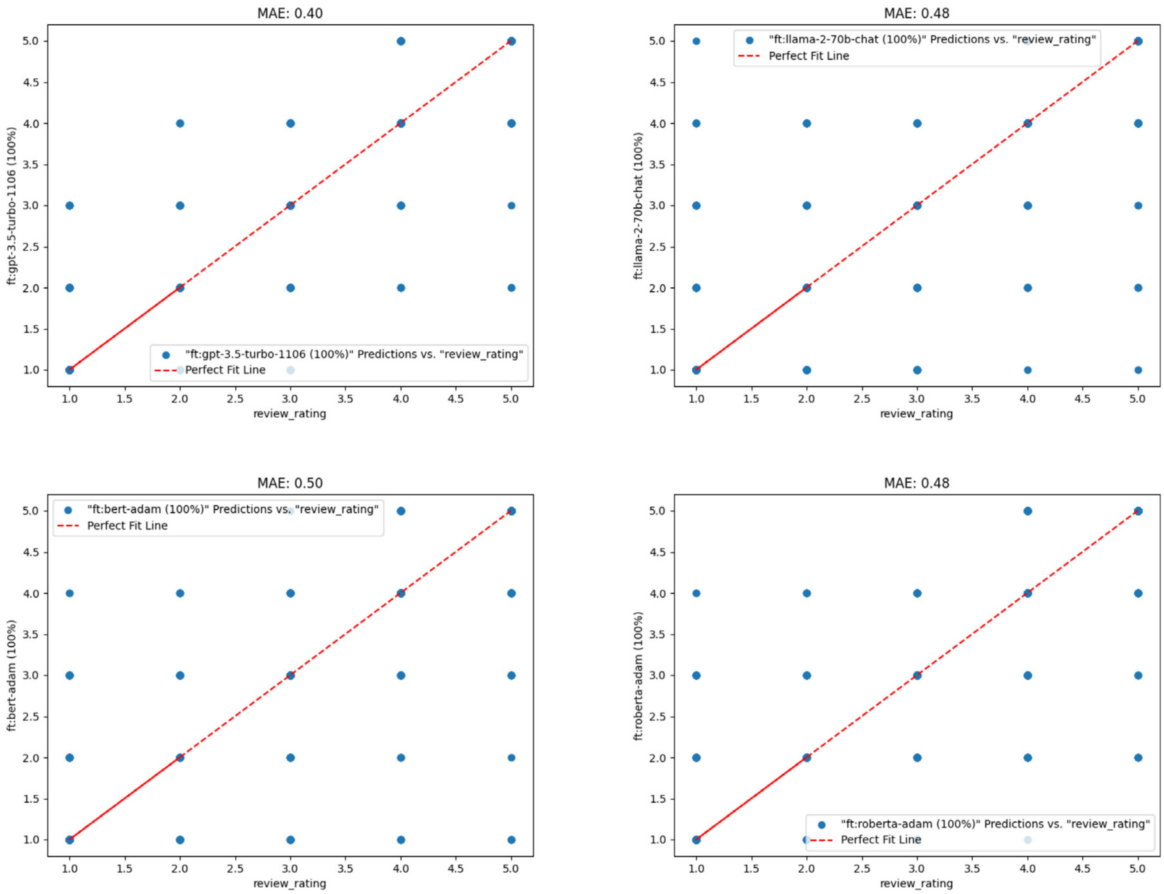

Fig. 5. MAE for fine-tuned models.

Table 2
Lexical elements detected by fine-tuned LLMs in review bodies.

Review content	Rating	ft:`llama-2-70b-chat`	ft:`gpt-3.5-turbo-1106`
junk sound cuts inout for a while at first but following not long after it completely stopped working	1	['junk', 'sound', 'cuts inout', 'not long after', 'completely stopped working']	['junk', 'sound cuts inout', 'completely stopped working']
nice hiking shoes it fits well i returned it because i felt the sole wasnt padded enough and the bottom was a bit stiff	4	['nice hiking shoes', 'it fits well', 'returned', 'wasnt padded enough', 'stiff']	['nice', 'fits well', 'sole wasnt padded enough', 'bottom was a bit stiff']
this is great theres just one thing great shower curtain thats nice and light and soft the one thing i wish it had was magnets in the corners but other than that its great	4	['great', 'light', 'soft', 'magnets', 'corner']	['great', 'shower curtain', 'nice', 'light', 'soft', 'wish it had magnets', 'other than that']
soft good quality it is very soft and would recommend only flaw is the threading on the pillow cases are a little loose im sure not all manufactured are like that	4	['soft', 'good quality', 'very soft', 'recommend', 'loose', 'threading', 'pillow cases', 'manufactured', 'flaw']	['soft', 'good quality', 'recommend', 'threading on the pillow cases are a little loose']
nice nice color and stretchyfits a little baggy for slim fit	4	['nice', 'color', 'stretchy', 'fits a little baggy', 'slim fit']	['nice color', 'stretchy', 'fits a little baggy', 'slim fit']
nice shirt for the price looks nice pressed out very neat plenty long to stay tucked in	5	['nice shirt', 'for the price', 'looks nice', 'pressed out', 'very neat', 'plenty long', 'stay tucked in']	['nice', 'shirt', 'price', 'looks nice', 'pressed out', 'neat', 'plenty long', 'stay tucked in']
works good sound sucked but not the cord retured the xlr and mic and cant return this for some reason	3	['works good', 'sound sucked', 'not the cord', 'retured the xlr and mic', "can't return this"]	['good', 'sucked', 'cord', 'returned', 'xlr', 'mic', 'cant return']

5.1. Base models comparison

- **Research Question 1**: Which LLM demonstrates superior efficacy in assessing product reviews?
- **Research Statement 1**: The `gpt-3.5-turbo-1106` base model outperforms the `llama-2-70b-chat` model by 3.97% in tasks associated with predicting star ratings for product reviews.

In Section 4.2, the research results regarding the performance of the base models `gpt-3.5-turbo-1106` and `llama-2-70b-chat` in predicting star-ratings in a test-sample of 755 product reviews were presented. Notably, it was observed that GPT-3.5 is significantly more accurate than the LLaMA-2 model. This particular difference is attributed to the training phase, where the GPT-3.5 model is assumed to have been trained with over 175 billion parameters (Cascella et al., 2023), while the LLaMA-2 model has 70 billion parameters (Touvron et al., 2023).

One plausible hypothesis to explain this deviation in performance lies in the distinct training processes that these two models underwent, particularly in their exposure to various public datasets. It is conceivable that GPT-3.5 benefited from a more extensive and diverse range of training data, including potentially more meticulous and specific fine-tuning. This expanded review training dataset might have provided GPT-3.5 with a deeper understanding of nuanced topics, enabling it to generate more accurate and contextually relevant responses.

In essence, the notable performance of GPT-3.5 underscores the importance of not only the quantity of parameters, but also the quality and diversity of the training data in achieving optimal language model capabilities.

5.2. Fine-tuned models comparison

- **Research Question 2**: Among the LLMs, which exhibits superior performance after undergoing fine-tuning for the assessment of product reviews?
- **Research Statement 2**: Following the fine-tuning process, the GPT-3.5 model exhibits superior performance to the LLaMA-2 model by a margin of 2.39% in tasks related to predicting star ratings for product reviews.

Based on the results obtained in Section 4.3, the fine-tuned GPT-3.5 model demonstrates an accurate assessment of 64.24% of review ratings in comparison to the LLaMA-2 fine-tuned model, which predicts ratings with a success rate of 61.95%. Although the difference between the two models is not greater than 2.39%, the superiority of the fine-tuned GPT-3.5 model in comparison to the LLaMA-2 fine-tuned model, can be ascribed to several factors inherent in its architecture. Notably, the GPT-3.5 model showcases a more effective utilization of the training data during the fine-tuning process. This effectiveness can be attributed to several key characteristics of the GPT-3.5 architecture.

- First and foremost, the GPT-3.5 model benefits from a more extensive and diverse pre-training corpus, which provides it with a broader foundation of linguistic knowledge and context. This vast pre-training dataset empowers the model to have a deeper understanding of the language, including nuanced expressions and review-related content. Consequently, when fine-tuned on review ratings, it demonstrates a greater proficiency in identifying and comprehending the subtle nuances inherent in rating assessments, thus leading to a higher accuracy rate of 64.24%.
- Furthermore, the GPT-3.5 architecture incorporates optimized transformer-based deep learning mechanisms, which are capable of capturing intricate relationships and dependencies within the training data. This results in a more sophisticated representation of review content and its correlation with rating scores. In contrast, despite the LLaMA-2 model being fine-tuned and utilizing an optimized transformer architecture, supervised fine-tuning (SFT), and reinforcement learning with human feedback (RLHF), it may not benefit to the same extent from such a sophisticated architecture. This difference may explain its slightly lower success rate of 61.95%.
- Additionally, the GPT-3.5 model's architecture inherently facilitates better generalization, enabling it to adapt more effectively to the nuances and variations within the review data, which may arise from different sources and domains. This adaptability is a consequence of the model's extensive pre-training and fine-tuning stages.

5.3. The importance of domain-specific fine-tuning

- **Research Question 3**: How significant is the process of fine-tuning LLMs for specific tasks?
- **Research Statement 3**: The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 10.99% and 9.41% increase in predictions in comparison to their base models predictions.

Building upon the results of our research, it is essential to situate our findings within the context of how these fine-tuned models outperform their respective base models. In this context, we detect a noteworthy and substantial enhancement in prediction accuracy. The fine-tuned GPT-3.5 model showcases an impressive 9.41% increase in prediction accuracy, while the LLaMA-2 model exhibits an even more remarkable 10.99% rise in prediction accuracy when compared to their base models, respectively. In a direct comparison, it becomes apparent that the LLaMA model responds slightly more effectively to fine-tuning than the GPT model. These observations collectively emphasize the fundamental significance of utilizing fine-tuning techniques to optimize LLMs for specific tasks.

5.4. The importance of data quantity and quality on fine-tuning

- **Research Question 4**: What is the impact of both the quantity and quality of the dataset designated for fine-tuning purposes?
- **Research Statement 4**: The fine-tuned LLaMA-2 and GPT-3.5 models demonstrated a 1.45% and 2.12% increase in predictions when the dataset was doubled in size.

In Section 4.4, with the aim of investigating the significance of the quantity of training data in the performance of LLMs, a new training set was randomly created, comprising 50% of the records from the base training set. Using this specific set, both models were fine-tuned, and then they were tasked with predicting star ratings for same test data in the form of product reviews.

Following a comparative analysis of the models trained with 50% of the training data and the models trained on the entirety of the training data, a substantial difference in response accuracy became evident. To be specific, the fine-tuned LLaMA-2 model exhibited a 1.45% decrease in response accuracy compared to its fully trained counterpart, while the GPT-3.5 model displayed a 2.12% reduction, respectively. These findings strongly emphasize the crucial role that the quantity of training data plays in shaping the overall performance and accuracy of fine-tuned models.

Simultaneously, research from similar studies has demonstrated that, in addition to data quantity, data quality plays a significant role in fine-tuning LLMs (Ai et al., 2023). It is crucial to emphasize that ensuring exceptional quality in labeled data is just as important as meeting quantity thresholds. Achieving high-quality data in real-world applications requires a multifaceted approach involving data cleaning, data labeling, and rigorous data quality assessment.

5.5. Cost-efficiency aspect of fine-tuning on LLMs

- **Research Question 5**: To what extent does an optimized prompting engineering plan and fine-tuning contribute to achieving better and more cost-effective results from LLMs?
- **Research Statement 5**: Effective prompting engineering and fine-tuning of LLMs play a crucial role in achieving improved outputs and greater cost efficiency.

It is imperative to underscore the cost-efficiency aspect of fine-tuning when deploying LLMs. To elucidate this point, we must consider the computational tokens as a unit of measurement. A descriptive output of 1,000 characters, calculated through the `tiktoken` library, corresponds to 204 tokens. These tokens are a representation of the cost incurred for executing a model in terms of the computational resources

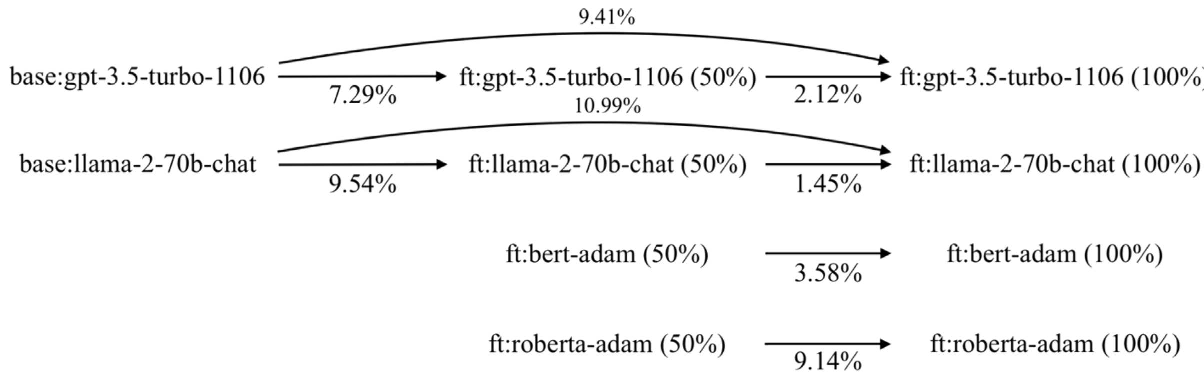

Fig. 6. Model-wise variability in accuracy percentage of model predictions.

required for the model to comprehend the input prompt and generate a response for the end user.

According to the billing policy of `gpt-3.5-turbo-1106`, these 204 tokens amount to an output cost of 0.000408 USD (0.002 USD per 1,000 tokens). In contrast, a structured output, consisting of 20 characters (equivalent to 4 tokens), would incur a cost of 0.00004 USD for the operator running the GPT-3.5 model. While the difference may seem inconsequential for a single output, when applied to a dataset with millions of records necessitating millions of outputs, the cost differential becomes substantial. The cost difference is even more significant for the fine-tuning of the LLaMA-2 model, which, in our research, required 8x A40 (Large) GPUs for 377.02 min in runtime. Particularly during the prediction phase of `llama-2-70b-chat`, due to its chat-oriented nature, responses before fine-tuning contained dialogue text, even though the prompt explicitly specified returning JSON format. This fact accentuates the critical role of fine-tuning LLMs in cost reduction.

Furthermore, cost efficiency is particularly pertinent in the context of the input prompt. Fine-tuned models have the potential to yield equivalent results to those generated by a base model, but with a more concise prompt, using the appropriate prompting engineering techniques. This holds significant implications for the economic consideration associated with deploying LLMs.

In summary, this examination highlights the economic implications of fine-tuning LLMs, emphasizing that while the cost difference may seem marginal for isolated outputs, its significance becomes pronounced when dealing with extensive datasets requiring numerous responses. Additionally, the ability of fine-tuned models to achieve similar outcomes with shorter prompts underscores the imperative nature of fine-tuning in optimizing both performance and cost-efficiency in the deployment of LLMs.

5.6. Generated output structure on LLMs

- **Research Question 6**: What impact does fine-tuning have on the generated output of LLMs?
- **Research Statement 6**: The generated output of LLMs can be enhanced through fine-tuning, with the LLaMA-2 model improving its responses by 100% following the fine-tuning process.

While utilizing the base model `llama-2-70b-chat`, it became apparent that, on numerous occasions, the generated output did not align with the desired JSON structure, even when explicit instructions were provided in the prompt. In contrast, the base model `gpt-3.5-turbo-1106` consistently delivered correct JSON-formatted answers from the outset. However, a pivotal turning point occurred after the fine-tuning process was applied to the LLaMA-2 model. Remarkably, the results exhibited a substantial improvement in adhering to the JSON format for all responses generated.

This noteworthy transformation in the model's behavior highlights a critical aspect of the utility of fine-tuning when working with LLMs. It serves as a compelling testament to the profound impact that fine-tuning can have on molding these models into highly effective tools for specific tasks.

The significant shift from inconsistent output to a perfect JSON format output following fine-tuning showcases the malleability and adaptability of LLMs. Fine-tuning empowers these models to not only understand the specific requirements of a task but also execute them with precision. This transformation underscores the practical importance of fine-tuning as a powerful mechanism for tailoring these models to meet specific goals, whether in structured data generation or any other specialized language processing task.

5.7. Comparing fine-tuning and prediction performance: LLMs vs non-LLM NLP models

- **Research Question 7**: Are LLMs, such as GPT-3.5 and LLaMA-2 models, or NLP models like BERT and RoBERTa, more effective in predicting product review ratings and, more generally, in regression and sentiment analysis tasks?
- **Research Statement 7**: Overall, LLMs are more accurate in predicting product review ratings than non-LLMs after fine-tuning.

Fig. 6 depicts that the performance of the ft:`gpt-3.5-turbo-1106`, trained on 100% of the training set, is more accurate by 9.41% compared to the predictions of the base:`gpt-3.5-turbo-1106` model and 2.12% more accurate than the ft:`gpt-3.5-turbo-1106` trained on 50% of the training set.

Similarly, the performance of the ft:`llama-2-70b-chat`, trained on 100% of the training set, is more accurate by 10.99% compared to the predictions of the base:`llama-2-70b-chat` and 1.45% more accurate than the ft:`llama-2-70b-chat` trained on 50% of the training set.

In comparison to the ft:bert-adam model, the ft:`gpt-3.5-turbo-1106` makes more precise predictions by 5.43%, and the ft:`llama-2-70b-chat` by 3.04% (on 100% of the training set). Compared to the ft:roberta-adam model, the ft:`gpt-3.5-turbo-1106` makes more accurate predictions by 3.58%, and the ft:`llama-2-70b-chat` by 1.19% (on 100% of the training set).

Overall, LLMs are more efficient than non-LLMs after fine-tuning.

An interesting point observed is the percentage performance difference between models trained on 50% and 100% of the training data. The increase in accuracy for BERT and RoBERTa models is 3.58% and 9.14%, respectively, while for GPT-3.5 and LLaMA-2, it is 2.12% and 1.45%, respectively. This implies that with the increase in the training set by 1,763 samples, the predictions of BERT and RoBERTa are more accurate compared to GPT-3.5 and LLaMA-2. This finding raises additional questions about whether LLMs will continue to be equally efficient if the dataset is further increased. However, studies have shown that LLMs need more than 1,000 samples during training to understand nuances and specific patterns in a dataset (Cai et al., 2023; Yao et al., 2023). This may suggest that if we have a small dataset, it is preferable to use BERT and RoBERTa, while LLMs become more efficient for the specific task with a much larger training set.

Using the same logic, we can also consider the cost of running and fine-tuning models for specific tasks. For training BERT and RoBERTa models on a training set of 3,520 samples, it took less than 5 min (0.41 compute units) with an A100 GPU, and the cost of 100 compute units is $9.99 on Google Colab. In contrast, fine-tuning `llama-2-70b-chat` using the replicate API took 377 min with 8x A40 GPUs, costing $133.47, and fine-tuning `gpt-3.5-turbo-1106` took 56 min with $11.68. Therefore, it is up to the user to decide whether to use an LLM for higher performance, considering the cost, or choose a more economical NLP model by comparing costs and benefits. In the course of our investigation, empirical evidence substantiates the superior efficacy of LLMs in contrast to alternative NLP models when applied to the prediction of product review ratings. While the task at hand pertains specifically to regression and sentiment analysis, rendering a broad inference regarding the overarching efficiency of these particular models across classification tasks would constitute an oversimplification.

6. Implications and directions for future research

6.1. Implications of the study

The results of this study present both theoretical and practical implications for the fields of research related to LLMs and AI, as well as for managers, marketers, and data analysts in the e-commerce sector. Our research underscores the importance of understanding the emotions of e-commerce customers following a purchase, as well as their satisfaction levels. In the past, comprehending customer emotions was a challenging task, as models could not grasp the context behind each customer review. In this study, we recommend the use of the latest advancements in the AI field, the LLMs, as a more efficient solution for understanding the emotions behind each product review and identifying customer satisfaction. LLMs not only provide advantages to e-commerce owners, managers, marketers, and data analysts benefit, but also ensures customers receive faster responses to their feedback, thereby enhancing the e-commerce strategy's effectiveness.

Adhering to the collective findings from this study positions us to address the ultimate and most critical research question (Q8), which concerns the effective use of LLMs for evaluating product reviews. Considering the discoveries that LLMs represent a highly promising AI technology capable of providing tangible solutions across various tasks, including the assessment of sentiment in product reviews within an e-commerce context, our results, both pre and post fine-tuning, have demonstrated the considerable success of GPT-3.5 and LLaMA-2 in predicting the rating stars of reviews and identifying the words influencing user ratings.

6.2. Limitations and directions for future research

Our research has highlighted the effectiveness of LLMs for tasks related to assessing reviewers' sentiments and gaining a deeper understanding of the words and context within a review that significantly impact the rating of that specific review. However, additional enhancements could be attained through both dataset and fine-tuning process improvements, as well as advancements in LLM technology. Nevertheless, within this section, we not only discuss the limitations but also emphasize certain future directions that the authors consider essential for advancing the e-commerce field and integrating LLMs into processes aimed at automating e-commerce and customer-related tasks.

Limitations of this study:

While the study's findings are valuable, it is crucial to consider several important aspects. First, the conclusions are derived from a randomly selected sample of 5,029 product reviews, potentially limiting their representativeness in the diverse e-commerce landscape. To ensure more comprehensive insights, future research should prioritize the inclusion of a broader and more diverse dataset. Furthermore, the study emphasizes the potential of LLMs, but highlights the need to acknowledge that their adaptability may not be universal across the e-commerce spectrum. Factors such as specific product categories, service models, and diverse customer demographics can significantly influence the effectiveness of LLMs in various contexts. In addition, the use of LLMs for scrutinizing customer sentiment raises ethical concerns, particularly related to privacy and data security. These ethical quandaries need thoughtful consideration and appropriate measures to ensure responsible use.

Future Directions:

Looking forward in the field of AI for e-commerce, several key pathways for advancement come into focus. Firstly, there is a crucial need for a comprehensive comparative performance assessment, meticulously evaluating how LLMs compare to other AI models to ascertain their suitability for specific e-commerce applications. Our research has highlighted the importance of evaluating the performance of LLMs, incorporating scalable increases in the training set and fine-tuning. This involve calculating prediction accuracy and the learning rate of the model for each task, allowing, through scalable fine-tuning, for precise predictions of the most efficient training set size in the future.

Additionally, the development of a robust ethical framework is paramount. This framework should offer clear and comprehensive guidelines for the responsible integration of AI into e-commerce, with a strong emphasis on the ethically handling customer data and ensuring privacy protection.

Moreover, it is imperative to delve deeply into segmentation strategies. This entails the customization of LLMs to cater to distinct customer segments and product categories, thereby enhancing their effectiveness and relevance. Furthermore, we should explore the integration of multimodal data, encompassing text, images, and videos, into sentiment analysis to attain a more holistic understanding of customer emotions and satisfaction.

Finally, the future of e-commerce involves not only using LLMs for analytical purposes, but also harnessing their power for the personalization of recommendations and the optimization of chatbot interactions. This holistic approach aims to significantly elevate the overall user experience, leading to increased customer satisfaction and engagement.

7. Conclusion

In conclusion, understanding consumer sentiments, purchase decisions, and ensuring unconditional customer satisfaction are vital for the sustained success of e-commerce enterprises in the face of intense competition. Technology, especially LLMs, offers robust solutions to common e-commerce challenges. In this study, the use of LLMs for understanding reviewer satisfaction post-purchase was explored. The results were highly encouraging, with LLMs, both in their base and fine-tuned forms, achieving predictive accuracy rates approaching 65% for review ratings. A notable observation was the alignment of lexical elements considered significant by LLMs in rating assessments, underscoring their ability to comprehend the importance of specific words within the context, similar to human evaluators. The use of LLMs is undoubtedly a promising tool, benefiting e-commerce owners in making strategic decisions and offering customers valuable insights that enhance their satisfaction and experience.

CRediT authorship contribution statement

Konstantinos I. Roumeliotis: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nikolaos D. Tselikas:** Writing – review & editing, Supervision, Resources, Project administration, Methodology. **Dimitrios K. Nasiopoulos:** Writing – review & editing, Supervision, Project administration, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmad, S.N., Laroche, M., 2023. Extracting marketing information from product reviews: A comparative study of latent semantic analysis and probabilistic latent semantic analysis. J. Market. Anal. 1–15. http://dx.doi.org/10.1057/S41270-023-00218-6/METRICS.

Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., Gao, S., Guo, J., He, X., Lan, Y., Li, C., Liu, Y., Lyu, Z., Ma, W., Ma, J., 2023. Information retrieval meets large language models: A strategic report from Chinese IR community. AI Open 4, 80–90. http://dx.doi.org/10.1016/J.AIOPEN.2023.08.001.

Al Wahshat, H., Abu-ulbeh, W., Yusoff, M.H., Zakaria, M.D., Amir Fazamin Wan Hamzah, W.M., P, S.N., 2023. The detection of E-commerce manipulated reviews using GPT-4. In: 2023 International Conference on Computer Science and Emerging Technologies. (CSET), pp. 1–6. http://dx.doi.org/10.1109/CSET58993.2023.10346848.

Askalidis, G., Kim, S.J., Malthouse, E.C., 2017. Understanding and overcoming biases in online review systems. Decis. Support Syst. 97, 23–30. http://dx.doi.org/10.1016/j.dss.2017.03.002.

BERT — transformers 3.0.2 documentation. (0000a). Retrieved December 17 2023, from https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#bertforsequenceclassification.

BERT — transformers 3.0.2 documentation. (0000b). Retrieved December 17 2023, from https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#berttokenizer.

bert-base-uncased · Hugging Face. (0000). Retrieved December 17 2023, from https://huggingface.co/bert-base-uncased.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 2020-December. https://arxiv.org/abs/2005.14165v4.

Cai, X., Xiao, M., Ning, Z., Zhou, Y., 2023. Resolving the imbalance issue in hierarchical disciplinary topic inference via LLM-based data augmentation. https://arxiv.org/abs/2310.05318v2.

Cascella, M., Montomoli, J., Bellini, V., Bignami, E., 2023. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J. Med. Syst. 47 (1), 1–5. http://dx.doi.org/10.1007/S10916-023-01925-4/TABLES/2.

Chen, J., Tech, W.G., Cho, J.H.D., Ma, L., Li, X., Thakurdesai, N., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., 2023. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with LLMs. https://arxiv.org/abs/2305.09858v1.

de Curtò, J., de Zarzà, I., Calafate, C.T., 2023. Semantic scene understanding with large language models on unmanned aerial vehicles. Drones 7 (2), 114. http://dx.doi.org/10.3390/DRONES7020114.

Dwidienawati, D., Tjahjana, D., Abdinagoro, S.B., Gandasari, D., Munawaroh, 2020. Customer review or influencer endorsement: which one influences purchase intention more? Heliyon 6 (11), http://dx.doi.org/10.1016/j.heliyon.2020.e05543.

eCommerce - Worldwide | Statista Market Forecast. (0000). Retrieved October 9 2023, from https://www.statista.com/outlook/dmo/ecommerce/worldwide#revenue.

Engler, T.H., Winter, P., Schulz, M., 2015. Understanding online product ratings: A customer satisfaction model. J. Retail. Consumer Serv. 27, 113–120. http://dx.doi.org/10.1016/J.JRETCONSER.2015.07.010.

Fanni, S.C., Febi, M., Aghakhanyan, G., Neri, E., 2023. Natural Lang. Process. 87–99. http://dx.doi.org/10.1007/978-3-031-25928-9_5.

Fine-tuning large language models (LLMs) in 2023 | SuperAnnotate. (0000). Retrieved December 21 2023, from https://www.superannotate.com/blog/llm-fine-tuning.

Gao, D., Chen, K., Chen, B., Dai, H., Jin, L., Jiang, W., Ning, W., Yu, S., Xuan, Q., Cai, X., Yang, L., Wang, Z., 0000. Llms-based machine translation for E-commerce. http://dx.doi.org/10.2139/SSRN.4682559.

GitHub - kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews. (0000). Retrieved November 16 2023, from https://github.com/kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews.

Griva, A., 2022. I can get no e-satisfaction. What analytics say? Evidence using satisfaction data from e-commerce. J. Retail. Consumer Serv. 66, 102954. http://dx.doi.org/10.1016/J.JRETCONSER.2022.102954.

Guha Majumder, M., Dutta Gupta, S., Paul, J., 2022. Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. J. Bus. Res. 150, 147–164. http://dx.doi.org/10.1016/J.JBUSRES.2022.06.012.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., 2021. Pre-trained models: Past, present and future. AI Open 2, 225–250. http://dx.doi.org/10.1016/J.AIOPEN.2021.08.002.

Kalyan, K.S., 2023. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Lang. Process. J. 100048. http://dx.doi.org/10.1016/j.nlp.2023.100048.

Kanaan, A.G., Wahsheh, F.R., El-Ebiary, Y.A.B., Wan Hamzah, W.M.A.F., Pandey, B., P, S.N., 2023. An evaluation and annotation methodology for product category matching in e-commerce using GPT. In: 2023 International Conference on Computer Science and Emerging Technologies. (CSET), pp. 1–6. http://dx.doi.org/10.1109/CSET58993.2023.10346684.

Kang, H., Yoo, S.J., Han, D., 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst. Appl. 39 (5), 6000–6010. http://dx.doi.org/10.1016/J.ESWA.2011.11.107.

Kassim, N., Abdullah, N.A., 2008. Customer loyalty in e-commerce settings: An empirical study. Electron. Markets 18 (3), 275–290. http://dx.doi.org/10.1080/10196780802265843.

kroumeliotis/ecommerce-reviews50 – Run with an API on Replicate. (0000). Retrieved December 17 2023, from https://replicate.com/kroumeliotis/ecommerce-reviews50.

kroumeliotis/ecommerce-reviews5029 – Run with an API on Replicate. (0000). Retrieved December 17 2023, from https://replicate.com/kroumeliotis/ecommerce-reviews5029.

kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews: Fine-Tuning GPT-3.5 and LLama 2 for Product Reviews. (0000). Retrieved October 20 2023, from https://github.com/kroumeliotis/fine-tuning-gpt3.5-llama2-for-product-reviews.

Kumar, V., Ayodeji, O.G., 2021. E-retail factors for customer activation and retention: An empirical study from Indian e-commerce customers. J. Retail. Consumer Serv. 59, 102399. http://dx.doi.org/10.1016/J.JRETCONSER.2020.102399.

Li, B., Hou, Y., Che, W., 2022. Data augmentation approaches in natural language processing: A survey. AI Open 3, 71–90. http://dx.doi.org/10.1016/J.AIOPEN.2022.03.001.

Li, Z., Li, X., Liu, Y., Xie, H., Li, J., Wang, F., Li, Q., Zhong, X., 2023e. Label supervised LLaMA finetuning. https://arxiv.org/abs/2310.01208v1.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y., 2023c. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (llama) using medical domain knowledge. Cureus http://dx.doi.org/10.7759/cureus.40895.

Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.-T., Xie, P., Huang, F., Jiang, Y., 2023d. EcomGPT: instruction-tuning large language models with chain-of-task tasks for e-commerce. https://arxiv.org/abs/2308.06966v2.

Li, L., Yuan, L., Tian, J., 2023a. Influence of online E-commerce interaction on consumer satisfaction based on big data algorithm. Heliyon 9 (8), http://dx.doi.org/10.1016/j.heliyon.2023.e18322.

Li, Q., Zhao, S., Zhao, S., Wen, J., 2023b. Logistic regression matching pursuit algorithm for text classification. Knowl.-Based Syst. 277, 110761. http://dx.doi.org/10.1016/J.KNOSYS.2023.110761.

Liu, Q. (Ben), Karahanna, E., Watson, R.T., 2011. Unveiling user-generated content: Designing websites to best present customer reviews. Bus. Horiz. 54 (3), 231–240. http://dx.doi.org/10.1016/J.BUSHOR.2011.01.004.

Liu, Y., Wan, Y., Shen, X., Ye, Z., Wen, J., 2021. Product customer satisfaction measurement based on multiple online consumer review features. Information 12 (6), 234. http://dx.doi.org/10.3390/INFO12060234.

Liu, Y., Zhang, W.-N., Chen, Y., Zhang, Y., Bai, H., Feng, F., Cui, H., Li, Y., Che, W., 2023b. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue. https://arxiv.org/abs/2310.14626v1.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J., 2023a. GPT understands, too. AI Open http://dx.doi.org/10.1016/J.AIOPEN.2023.08.012.

Ma, S., Huang, S., Huang, S., Wang, X., Li, Y., Zheng, H.-T., Xie, P., Huang, F., Jiang, Y., 2023. EcomGPT-CT: continual pre-training of e-commerce large language models with semi-structured data. https://arxiv.org/abs/2312.15696v1.

Maragheh, R.Y., Morishetti, L., Giahi, R., Nag, K., Xu, J., Cho, J., Korpeoglu, E., Kumar, S., Achan, K., 2023. LLM-based aspect augmentations for recommendation systems.

Meire, M., 2021. Customer comeback: Empirical insights into the drivers and value of returning customers. J. Bus. Res. 127, 193–205. http://dx.doi.org/10.1016/J.JBUSRES.2021.01.017.

Mujtaba, I.M., Sowgath, M.T., 2022. Application of artificial intelligence in desalination processes. Desalination Technol. 54, 1–593. http://dx.doi.org/10.1016/B978-0-12-813790-1.00011-6.

Murali, S., Pugazhendhi, S., Muralidharan, C., 2016. Modelling and investigating the relationship of after sales service quality with customer satisfaction, retention and loyalty – A case study of home appliances business. J. Retail. Consumer Serv. 30, 67–83. http://dx.doi.org/10.1016/J.JRETCONSER.2016.01.001.

Orzoł, M., Szopik-Depczyńska, K., 2023. ChatGPT as an innovative tool for increasing sales in online stores. Procedia Comput. Sci. 225, 3450–3459. http://dx.doi.org/10.1016/J.PROCS.2023.10.340.

Pei, X.L., Guo, J.N., Wu, T.J., Zhou, W.X., Yeh, S.P., 2020. Does the effect of customer experience on customer satisfaction create a sustainable competitive advantage? A comparative study of different shopping situations. Sustainability 12 (18), 7436. http://dx.doi.org/10.3390/SU12187436.

Peng, B., Li, C., He, P., Galley, M., Gao, J., 2023. Instruction tuning with GPT-4. https://arxiv.org/abs/2304.03277v1.

Pretrained models — transformers 3.3.0 documentation. (0000). Retrieved December 17 2023, from https://huggingface.co/transformers/v3.3.1/pretrained_models.html.

Ramaswamy, S., DeClerck, N., 2018. Customer perception analysis using deep learning and NLP. Procedia Comput. Sci. 140, 170–178. http://dx.doi.org/10.1016/J.PROCS.2018.10.326.

RoBERTa — transformers 2.9.1 documentation. (0000). Retrieved December 17 2023, from https://huggingface.co/transformers/v2.9.1/model_doc/roberta.html#robertaforsequenceclassification.

roberta-base · Hugging Face. (0000). Retrieved December 17 2023, from https://huggingface.co/roberta-base.

Roethke, K., Klumpe, J., Adam, M., Benlian, A., 2020. Social influence tactics in e-commerce onboarding: The role of social proof and reciprocity in affecting user registrations. Decis. Support Syst. 131, 113268. http://dx.doi.org/10.1016/J.DSS.2020.113268.

Rosli, N., Nayan, S.M., 2020. Why customer first? J. Undergrad. Soc. Sci. Technol. 2 (2), http://dx.doi.org/10.1108/JABS-01-2019-0020.

Rothman, D., 2021. Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, Pytorch, Tensorflow. BERT, https://books.google.com/books?hl=en&lr=&id=Cr0YEAAAQBAJ&oi=fnd&pg=PP1&ots=a9t6Rt3i21&sig=6AunRon2EtcpjTNULNgtdoA2ODI.

Roumeliotis, K.I., Tselikas, N.D., 2023. ChatGPT and open-AI models: A preliminary review. Future Internet 15 (6), 192. http://dx.doi.org/10.3390/FI15060192.

Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2023a. Llama 2: early adopters' utilization of meta's new open-source pretrained model. http://dx.doi.org/10.20944/PREPRINTS202307.2142.V2.

Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2023b. Unveiling sustainability in ecommerce: GPT-powered software for identifying sustainable product features. Sustainability 15 (15), 12015. http://dx.doi.org/10.3390/SU151512015.

Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C.C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., 2023. Code llama: open foundation models for code. https://arxiv.org/abs/2308.12950v2.

Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., Dey, M., Bari, M.S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N.V., 2021. Multitask prompted training enables zero-shot task generalization. In: ICLR 2022-10th International Conference on Learning Representations. https://arxiv.org/abs/2110.08207v3.

Scotta, S., Messina, A., 2023. Experimenting task-specific LLMs. https://ceur-ws.org/Vol-3551/paper10.pdf.

Shi, K., Sun, X., Wang, D., Fu, Y., Xu, G., Li, Q., 2023. LLaMA-E: empowering e-commerce authoring with multi-aspect instruction following. In: Proceedings of Make Sure to Enter the Correct Conference Title from Your Rights Confirmation Emai (Conference Acronym 'XX), Vol. 1. https://doi.org/XXXXXXX.XXXXXXX.

sklearn.model_selection.train_test_split — scikit-learn 1.3.2 documentation. (0000). Retrieved December 17 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

Stephen, A.T., 2016. The role of digital and social media marketing in consumer behavior. Curr. Opin. Psychol. 10, 17–21. http://dx.doi.org/10.1016/J.COPSYC.2015.10.016.

Sundararaj, V., Rejeesh, M.R., 2021. A detailed behavioral analysis on consumer and customer changing behavior with respect to social networking sites. J. Retail. Consumer Serv. 58, 102190. http://dx.doi.org/10.1016/J.JRETCONSER.2020.102190.

Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2023. Fine-tuning large neural language models for biomedical natural language processing. Patterns 4 (4), 100729. http://dx.doi.org/10.1016/J.PATTER.2023.100729.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., 2023. Llama 2: open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288v2.

tqdm · PyPI. (0000). Retrieved December 17 2023, from https://pypi.org/project/tqdm/.

Utz, S., Kerkhof, P., Van Den Bos, J., 2012. Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. Electron. Commer. Res. Appl. 11 (1), 49–58. http://dx.doi.org/10.1016/J.ELERAP.2011.07.010.

Wang, Y., Li, X., 2021. Mining product reviews for needs-based product configurator design: A transfer learning-based approach. IEEE Trans. Ind. Inform. 17 (9), 6192–6199. http://dx.doi.org/10.1109/TII.2020.3043315.

Wang, Y., Lu, X., Tan, Y., 2018. Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. Electron. Commer. Res. Appl. 29, 1–11. http://dx.doi.org/10.1016/J.ELERAP.2018.03.003.

Wang, H., Na, T., 2023. Rethinking e-commerce search. https://arxiv.org/abs/2312.03217v1.

What runs ChatGPT? Inside Microsoft's AI supercomputer | Featuring Mark Russinovich - YouTube. (0000). Retrieved December 17 2023, from https://www.youtube.com/watch?v=Rk3nTUfRZmo.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. PMC-LLaMA: towards building open-source language models for medicine. https://arxiv.org/abs/2304.14454v3.

Yao, B., Chen, G., Zou, R., Lu, Y., Li, J., Zhang, S., Liu, S., Hendler, J., Wang, D., 2023. More samples or more prompt inputs? Exploring effective in-context sampling for LLM few-shot prompt engineering. https://arxiv.org/abs/2311.09782v1.

Zhang, M., Li, J., 2021. A commentary of GPT-3 in MIT technology review 2021. Fundam. Res. 1 (6), 831–833. http://dx.doi.org/10.1016/J.FMRE.2021.11.011.

Zhang, K., Zhou, F., Wu, L., Xie, N., He, Z., 2024. Semantic understanding and prompt engineering for large-scale traffic data imputation. Inf. Fusion 102, 102038. http://dx.doi.org/10.1016/J.INFFUS.2023.102038.

Zhao, A., Huang, D., Xu, Q., Lin, M., Liu, Y.-J., Huang, G., Mitchell, T., 2023. Expel: LLM agents are experiential learners. https://arxiv.org/abs/2308.10144v2.

Zhou, J., Liu, B., Hong, J.N.A.Y., Lee, K., Wen, M., 2023. Leveraging large language models for enhanced product descriptions in ecommerce. https://arxiv.org/abs/2310.18357v1.