Article

Not peer-reviewed version

Leveraging LLMs in Tourism: A Comparative Study of the latest GPT Omni models and BERT NLP for Customer Review Classification and Sentiment Analysis

Konstantinos I. Roumeliotis * , Tselikas D. Nikolaos , Nasiopoulos K. Dimitrios

Posted Date: 5 November 2024

doi: 10.20944/preprints202411.0313.v1

Keywords: sentiment analysis; hotel reviews; customer feedback; user-generated content; customer review classification; gpt-4 omni; few-shot learning; decision-making systems; forecasting models; tourism recommendation systems

 doi:10.20944/preprints202411.0313.v1

Article

Leveraging LLMs in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis

Konstantinos I. Roumeliotis [1,2,*], Nikolaos D. Tselikas [2] and Dimitrios K. Nasiopoulos [3]

[1] Department of Digital Systems, University of the Peloponnese, 23100 Sparta, Greece
[2] Department of Informatics and Telecommunications, University of the Peloponnese, 22131 Tripoli, Greece; ntsel@uop.gr
[3] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 11855 Athens, Greece; dimnas@aua.gr
* Correspondence: k.roumeliotis@uop.gr

Abstract: In today's rapidly evolving digital landscape, customer reviews play a crucial role in shaping the reputation and success of hotels. Accurately analyzing and classifying the sentiment of these reviews offers valuable insights into customer satisfaction, enabling businesses to gain a competitive edge. This study undertakes a comparative analysis between traditional natural language processing (NLP) models, like BERT, and advanced large language models (LLMs), specifically GPT-4 Omni and GPT-4o Mini, both pre- and post-fine-tuning with few-shot learning. By leveraging an extensive dataset of hotel reviews, we evaluate the effectiveness of these models in predicting star ratings based on review content. The findings demonstrate that the GPT-4 Omni family significantly outperforms the BERT model, achieving an accuracy of 67%, compared to BERT's 60.6%. GPT-4o, in particular, excelled in accuracy and contextual understanding, showcasing the superiority of advanced LLMs over traditional NLP methods. This research underscores the potential of using sophisticated review evaluation systems in the hospitality industry and positions GPT-4o as a transformative tool for sentiment analysis. It marks a new era in automating and interpreting customer feedback with unprecedented precision.

Keywords: sentiment analysis; hotel reviews; customer feedback; user-generated content; customer review classification; gpt-4 omni; few-shot learning; decision-making systems; forecasting models; tourism recommendation systems

1. Introduction

In the digital age, customer reviews have become a cornerstone in the success and reputation of businesses, particularly within the tourism and hospitality sectors. With the travel and tourism industry contributing 9.9 trillion U.S. dollars to the global GDP in 2023, and projections expecting this figure to reach 11.1 trillion U.S. dollars in 2024, exceeding pre-pandemic levels, the economic significance of tourism is undeniable [1]. Travelers increasingly rely on online reviews to make informed decisions about hotels, creating an environment where customer feedback directly influences business performance [2]. As the volume and complexity of reviews grow, artificial intelligence (AI) has emerged as a pivotal tool, offering innovative solutions for understanding and responding to this wealth of user-generated content.

AI, particularly with the advent of large language models (LLMs), is transforming how businesses in the tourism industry engage with customer feedback [3]. Beyond simple sentiment analysis, LLMs like GPT-4 Omni can not only discern nuanced emotional tones but also automatically respond to reviews and extract actionable insights that can significantly enhance customer satisfaction and profitability. These advanced models efficiently classify reviews, predict star ratings,

 doi:10.20944/preprints202411.0313.v1

and even provide personalized responses that resonate with customers, offering a level of automation and accuracy that traditional methods cannot match [4].

In this study, we explore the power of LLMs in sentiment analysis by comparing the performance of the BERT model and the GPT-4 Omni family [5]. Through this comparative analysis, we emphasize the growing importance of AI-driven tools in tourism, illustrating how these technologies can elevate customer experience, foster stronger brand loyalty, and provide strategic insights for business growth. By harnessing the full potential of LLMs, the tourism industry is entering a new era of precision, where AI not only interprets reviews but actively contributes to the customer journey and overall business success.

The primary aim of this research is threefold: first, to assess the effectiveness and applicability of advanced AI models, particularly LLMs, within the tourism sector; second, to compare the performance of these models both before and after fine-tuning with few-shot learning; and third, to address specific research questions that have not been adequately explored by previous studies:

- Q1: Do NLP models or LLMs demonstrate superior efficacy in assessing user-generated content, such as hotel customer reviews, in terms of sentiment analysis and star rating prediction?
- Q2: Which model within the GPT-4 Omni family exhibits superior performance before fine-tuning?
- Q3: Which model within the GPT-4 Omni family demonstrates superior performance after fine-tuning with few-shot learning?
- Q4: How significant is the cost of fine-tuning LLMs in achieving optimal results, and how does it impact performance in the tourism sector?
- Q5: Can LLMs be effectively utilized for the evaluation of customer reviews, and how can they revolutionize the hospitality sector by automating review responses and extracting actionable insights?

To fulfill the objectives of our study, we conduct an extensive review of the existing literature on sentiment analysis and classification, with a specific focus on the application of LLMs in the tourism sector in Section 2. Section 3 details the materials and methodologies utilized in this research, while Section 4 presents the prediction outcomes for all three models and their variations, both before and after fine-tuning. In Section 5, we engage in a thorough discussion of the results, extracting valuable insights and statements derived from our research findings.

2. Literature Review

2.1. The Applications of AI in Modern Hospitality

The hospitality and tourism sectors are increasingly utilizing advanced information systems technologies, particularly artificial intelligence (AI) and deep learning, to improve service delivery, enhance customer experiences, to support decision-making and predict market trends. This section synthesizes recent research that explores the interplay between sentiment analysis, AI applications, and tourism demand forecasting, shedding light on the methodologies employed and the implications for industry stakeholders [6].

2.1.1. Sentiment Analysis Techniques

Sentiment analysis has become a vital tool for extracting valuable insights from unstructured customer reviews. Priya et al. (2023) [7] highlight the challenges posed by noisy data, such as spelling errors and emoticons, in analyzing hotel reviews. Their proposed model, which combines a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) with Recurrent Neural Networks (RNN), significantly enhances sentiment classification accuracy. This approach not only demonstrates the efficacy of advanced neural networks but also underscores the necessity of robust models to handle real-world data complexities.

Similarly, Wen et al. (2022) [8] conducted a comparative analysis of BERT and ERNIE models, revealing that while BERT effectively categorizes sentiments, ERNIE outperforms it in both accuracy and stability. This study emphasizes the importance of selecting appropriate models for sentiment

 doi:10.20944/preprints202411.0313.v1

analysis in understanding customer preferences, which is crucial for improving hotel recommendation systems.

Kusumaningrum et al. (2023) [9] further contribute to this discourse with their development of Sentilytics 1.0, a web-based application that employs convolutional neural networks (CNNs) and improved long short-term memory (LSTM) models for multilevel sentiment analysis of Indonesian hotel reviews. Their performance metrics indicate the application's robustness, providing actionable insights for hotel management through comprehensive sentiment visualization.

Chang et al. (2023) [10] take a different approach by proposing a heuristic model for sentiment analysis in luxury hotels. By integrating visual and multimedia analytics, they identify key features impacting customer satisfaction. Their findings suggest that luxury hotels should focus on staff training and location selection to enhance customer experiences, reflecting the nuanced understanding of customer sentiments.

2.1.2. AI and Demand Forecasting

The integration of sentiment analysis into decision-making and demand forecasting is explored by Zhang et al. (2024) [11], who introduce the Long Short-Term Memory Interaction-based Convolutional Neural Network (LICNN) framework. Their study demonstrates significant improvements in forecasting accuracy when online review features are incorporated, suggesting that customer sentiment can serve as a valuable predictor of hotel demand. This innovative approach bridges the gap between sentiment analysis and operational forecasting, providing hotel managers with actionable insights for strategic planning.

Ounacer et al. (2023) [12] emphasize the significance of aspect-based sentiment analysis in tourism, presenting a semi-supervised CorEx method for aspect extraction. Their findings highlight the necessity of nuanced sentiment analysis in aiding customer decision-making, aligning with the broader trend of leveraging customer feedback to inform business strategies.

2.1.3. AI Applications in Service Delivery

The role of AI in enhancing service encounters within the hospitality sector is systematically reviewed by Li et al. (2021) [13], who identify four modes of AI interactions and their impact on customer service outcomes. Their model provides a framework for understanding how AI can improve service delivery, particularly during public health emergencies like COVID-19.

Pillai et al. (2020) [14] expand on this by investigating customer behavioral intentions toward AI chatbots in the Indian hospitality sector. Their mixed-methods approach reveals that perceived ease of use, trust, and anthropomorphism significantly influence chatbot adoption intentions, offering practical insights for designers and managers aiming to enhance AI integration in travel planning.

Huang et al. (2022) [15] propose a comprehensive evaluation framework for AI adoption in hospitality, identifying high-adoption areas such as search engines and virtual agents. This qualitative synthesis informs management practices by elucidating the factors influencing AI adoption, facilitating strategic decision-making.

2.1.4. Advanced Technologies and Tourism Experiences

Recent studies also explore the transformative potential of emerging technologies in enhancing tourist experiences. Miao et al. (2023) [16] examine generative AI tools and their implications for marketing and engagement in tourism. Their research highlights how these tools can enable personalized visual experiences, thereby enriching tourist interactions.

Wang et al. (2020) [17] discuss the intersection of IoT, 5G technology, and AI in smart tourism, suggesting that these technologies can significantly improve the quality and efficiency of tourist services. Their case study demonstrates the effectiveness of combining these technologies to create innovative solutions for enhancing customer experiences.

 doi:10.20944/preprints202411.0313.v1

2.1.5. Challenges and Future Directions

Despite the promising advancements in AI and sentiment analysis, challenges remain in achieving seamless integration and addressing user privacy concerns. Chi et al. (2022) [18] and Gupta et al. (2023) [19] explore tourists' attitudes toward AI in service delivery and the role of facial recognition technology, respectively, indicating that acceptance varies across different contexts and highlighting the need for careful consideration in service design.

Furthermore, Zhang et al. (2023) [20] assess the quality of AI-generated content compared to human-generated content, advocating for human-machine collaboration to enhance the nuanced quality of digital tourism interpretation.

The reviewed literature in this section underscores the critical role of sentiment analysis and AI technologies in transforming the hospitality and tourism sectors. By harnessing customer insights and advanced data processing techniques, businesses can enhance service delivery, improve customer satisfaction, and forecast demand more accurately.

2.2. Application of LLMs in the Tourism Sector

After discussing the broader impact of AI on the tourism sector, we turn our attention to the specific advancements in LLMs. Their rapid developments have revolutionized the industry by transforming the way information is managed and generated by information systems, recommendations are personalized, and tourism services are optimized. Numerous studies have investigated the potential of LLMs to enhance the creation of tourism-related information and deliver tailored services. This highlights the significant role LLMs play in improving the tourist experience, refining recommendation systems, and addressing challenges related to sustainability and service quality.

2.2.1. Context-Specific Applications

Qi et al. (2024) [21] investigate the limitations of existing smart tourism systems in Tibet, emphasizing the difficulties in generating culturally and contextually relevant content. They introduce the DualGen Bridge AI system, which employs supervised fine-tuning to provide personalized descriptions of tourist sites, highlighting the need for region-specific LLM applications to capture cultural nuances effectively. Similarly, Wei et al. (2024) [22] present TourLLM, a model fine-tuned with a specialized dataset, Cultour, aimed at improving travel-related information quality. They introduce the CRA (Consistency, Readability, Availability) criterion for evaluating LLM-generated content, demonstrating that domain-specific fine-tuning enhances relevance and accuracy.

2.2.2. Sustainable Tourism and Recommendations

Banerjee et al. (2024) [23] explore how LLMs can enhance tourism recommender and information systems (TRS) by integrating sustainability metrics into the recommendation process through a sustainability augmented re-ranking (SAR) model. Their findings indicate that LLM-enhanced systems outperform traditional models, particularly when sustainability considerations are incorporated. Vasic et al. (2024) [24] expand on this by examining LLMs' role in creating personalized museum tours, showing how tailored experiences can significantly enhance user satisfaction.

2.2.3. Advancements in Recommender Systems

Chen et al. (2023) [25] discuss the evolution of recommender and information systems, noting a shift toward conversational systems empowered by LLMs. They highlight LLMs' superior memory and reasoning capabilities, which streamline operations and enhance user experience. In line with this, Balamurali et al. (2023) [26] and Falatouri et al. (2024) [27] examine LLMs' integration with sentiment analysis and service quality assessment. Their research underscores the ability of LLMs to provide real-time, emotionally attuned responses, significantly enhancing customer experience and service quality.

 doi:10.20944/preprints202411.0313.v1

2.2.4. Combining Human and AI Insights

Yñiguez-Ovando et al. (2024) [28] propose a hybrid framework that combines LLMs with human intelligence to address sustainability issues, particularly over-tourism. This approach illustrates the potential for LLMs, when paired with human insights, to formulate effective solutions. Secchi et al. (2023) [29] investigate the integration of LLMs with tourism-specific knowledge graphs, enhancing the quality of hospitality services through improved information retrieval.

2.2.5. Tourism Promotion and AI Chatbots

Kodors et al. (2024) [30] emphasize the use of LLMs in AI-driven chatbots for tourism promotion, focusing on personalized, context-aware interactions with tourists. Their findings highlight trustworthiness and information saturation as critical factors for effective chatbot design. Hsu et al. (2024) [31] address limitations in generic LLMs and advocate for multistakeholder datasets to improve the reliability of generated tourism information.

2.2.6. Challenges and Ethical Considerations

Despite the advancements, challenges remain. Qi et al. (2024) [32] note the issue of hallucinations in LLM-generated content, proposing an RAG-optimized model to mitigate inaccuracies. Balfroid et al. (2024) [33] and Gonzalez-Garcia et al. (2024) [34] highlight additional limitations in automated content generation and knowledge graph enhancement. Meyer et al. (2024) [35] compare LLM fine-tuning methods, stressing the necessity for human oversight and robust evaluation metrics to ensure effective chatbot performance.

2.2.7. Potential Risks and Future Research Directions

Carvalho et al. (2024) [36] offer a balanced analysis of LLMs' benefits and risks, discussing applications in customer service and itinerary planning while raising concerns about job displacement and misinformation. Sioziou et al. (2024) [37] explore LLMs' capabilities in extracting structured information from the tourism job market, indicating potential efficiencies in recruitment processes. Finally, Liyanage et al. (2023) [38] address the growing concern of AI-generated opinion spam in hotel reviews, revealing the challenges in maintaining trust in online review systems.

In summary, the literature underscores the transformative potential of LLMs in the tourism sector while also addressing ongoing challenges such as content accuracy, sustainability, and ethical considerations.

3. Materials and Methods

In the literature review section, previous studies on the adoption of AI and LLMs in tourism were presented. These studies collectively highlighted the wide range of AI applications that support the tourism sector in its highly competitive landscape.

This study focuses on utilizing the advanced capabilities of LLMs, specifically the latest GPT Omni family and the widely-researched BERT natural language processing model. The GPT Omni family consists of two separate models: `gpt-4o`, which is the high-intelligence OpenAI's flagship model designed for complex and multi-step tasks, and `gpt-4o-mini`, which is more affordable and optimized for faster, more lightweight tasks. Both GPT models are more capable than their predecessors, `gpt-4-turbo` and `gpt-3.5-turbo`, generating text twice as fast at half the cost [5].

The main objective of our research is to stress-test these three models with a demanding classification task to determine if they can deliver real-world results that support the tourism sector.

In the first phase of our research, we carefully pre-processed and cleaned our dataset. We then prompted the GPT base models to predict the star ratings for each review in our dataset (zero-shot). Later, we fine-tuned BERT and both GPT Omni models using few-shot learning and parameter tuning, prompting them again to make predictions on the same dataset. Finally, we conducted a direct comparison, extracting valuable insights from the results.

 doi:10.20944/preprints202411.0313.v1

3.1. Dataset Cleaning, Preprocessing, and Splitting

Tripadvisor is a well-known online travel platform that offers travelers a wealth of information, reviews, and recommendations on hotels, restaurants, attractions, and other travel-related services worldwide. It allows users to plan and book trips by exploring user-generated content, such as ratings, photos, and in-depth reviews, while also providing the ability to compare prices for accommodations, flights, and activities.

With its extensive collection of verified user-generated content, Tripadvisor stands out as one of the most reliable sources for tourism-related reviews, often considered more trustworthy than platforms like Google My Business or Google Maps. Recognizing its credibility, we sourced the "Trip Advisor Hotel Reviews" dataset from the Kaggle platform, which includes 20,491 customer reviews specifically from hotels.

This dataset, with a total size of 5MB, has been highly rated by users, earning a usability score of 10 for its completeness, credibility, and compatibility. It is available to the public under the Attribution-NonCommercial 4.0 International license, which allows for copying and redistribution of the material in any medium or format for non-commercial use, as long as appropriate credit is given and a link to the license is provided [39,40].

3.1.1. Dataset Preprocessing

To guarantee the quality and effectiveness of our predictive modeling and fine-tuning processes, we carefully prepared the dataset. This preparation followed a structured approach, incorporating several key steps, all designed to enhance the data's relevance and suitability for the classification task.

Initially, the dataset comprised two columns: "Review" and "Rating." To facilitate the identification of potential errors in our code, we created an additional column labeled "ID," providing a direct reference for these errors. The next step involved identifying the unique values in the "Rating" column, which serves as the label column for the models' predictions. As the dataset's author had indicated, the "Rating" column consisted exclusively of integer values ranging from 1 to 5, reflecting the star ratings associated with each review.

Subsequently, we initiated the data preprocessing phase specifically for the feature column labeled "Review," which is crucial for enhancing the overall quality of the dataset. This stage focused on refining the text data to ensure its appropriateness for modeling purposes. A fundamental component of this process involved the removal of special characters, extraneous white spaces, and empty rows, which were essential for maintaining the integrity and consistency of the dataset. Failure to address special characters could introduce unwanted noise that might adversely affect the performance of LLM and NLP models [41].

Furthermore, text normalization played a pivotal role in achieving uniformity and standardization within the dataset. This process included the conversion of accented characters to their base forms, which is particularly pertinent for languages utilizing diacritics, thereby ensuring consistent treatment of words with accentual variations. In addition, to facilitate case insensitivity, we systematically transformed all text to lowercase, thereby enhancing the overall coherence of the data.

3.1.2. Dataset Splitting

Before splitting the dataset, we utilized the `value_counts` method in Python to determine the occurrence of each label. We found that the label "5" appeared 9,054 times, "4" appeared 6,039 times, "3" appeared 2,184 times, "2" appeared 1,793 times, and "1" appeared 1,421 times. The marked difference in the frequency of each label indicates that the dataset is imbalanced, with a significant overrepresentation of higher ratings (5 and 4) in contrast to the lower ratings (1, 2, and 3). This label imbalance can introduce bias into the model, causing it to become overly tuned to the majority classes (5 and 4) while struggling to accurately predict the minority classes (1, 2, and 3). Training a model on such skewed data may lead to overfitting, whereby the model excels at predicting the majority classes

 doi:10.20944/preprints202411.0313.v1

but fails to generalize effectively to the minority classes. As a result, this could adversely impact predictive performance, particularly for the lower ratings.

The goal is to create a balanced subset of 5,000 reviews, equally distributed across five labels (review ratings 1–5). For this operation, a `random_state` of 42 was used, ensuring random, reproducible selection of the subset using stratified sampling with the `StratifiedShuffleSplit` method.

The dataset was then divided into training, testing, and validation sets using the same approach, with the stratified column passed as a parameter in the `train_test_split` function. Initially, we split the data into 80% training and 20% testing sets. The training set was then further split, allocating 20% for validation and the remaining 80% for further training. Throughout fine-tuning, the training set was essential in helping the model learn patterns and relationships within the data, enabling it to perform tasks and make accurate predictions. The validation set was used to optimize hyperparameters and model configurations. Monitoring performance on this set allowed for adjustments that improved generalization and reduced overfitting.

After splitting the dataset, we obtained a test set of 1,000 samples, a training set of 3,200 samples, and a validation set of 800 samples, all equally distributed across the five labels.

3.2. LLM Prompt Engineering

Our aim was to create a prompt that seamlessly integrates with various LLMs, enhancing the usability of their outputs through our code. We prioritized not only the content of the prompt but also how its output is presented for improved accessibility.

To develop a prompt compatible with multiple LLMs, we needed a deep understanding of the distinct models available, such as GPT, Claude, and LLaMA each with its unique strengths and limitations. Crafting a prompt that could generate coherent responses from all these models, while remaining user-friendly, posed a significant challenge. To address this, we employed two key prompt engineering strategies, tailored to the specific traits of each LLM [42].

- Model-independent content: We designed the prompt to be agnostic of any particular model's architecture or nuances. This adaptability ensures the prompt can be applied across different LLMs with minimal adjustment. The focus was on constructing a prompt that clearly communicates the task, providing relevant context and information that can be interpreted by any LLM.
- Output formatting for accessibility: Recognizing the importance of usable output, we emphasized creating a structure that aligns well with coding and accessibility needs. This required organizing the responses in a logical, intuitive manner, specifically formatting the output to comply with the JSON standard.

After several iterations and tests with various LLMs, we finalized a prompt that consistently produces outputs in the desired format, easily interpreted by both humans and models, as shown in **Listing 1**.

Listing 1. Model-agnostic prompt.

```
conversation.append({'role': 'system',
'content': "You are an AI model tasked with predicting a star rating (1-5) based on the user's review"
                    "of their hotel experience. Return your response in JSON format: {'rating': integer}."})
conversation.append({'role': 'user',
'content': f"Predict the star rating (integer between 1 and 5) to the following review. Return your response"
                    f"in JSON format like this example {{'rating': integer}}. Please avoid providing additional"
```

(1)

 doi:10.20944/preprints202411.0313.v1

f"explanations. Review:\n{input['Review']}"})

3.3. Model Deployment, Fine-Tuning, and Predictive Evaluation

The aim of this study is to determine which of our three models performs best in predicting users' ratings based on the content of their reviews, specifically for sentiment analysis and classification. By identifying the model that most effectively understands the feelings and concerns of hotel customers, we can develop a powerful tool for automatically gaining insights from reviews. This would enable businesses to make informed decisions that enhance customer satisfaction, ultimately driving better outcomes and improving overall success.

To achieve this, all three models were tasked with making predictions on the test set, both before and after fine-tuning using few-shot learning. To ensure fairness during training, we applied the same hyperparameters across all models: a learning rate of 2×10^{-5}, a batch size of 6, and training over 3 epochs, using the Adam optimizer during fine-tuning.

The BERT model was configured to handle inputs of up to 512 tokens (around 2,560 characters). Reviews exceeding this length were truncated to fit within the token limit. While methods like chunking or hierarchical processing could manage longer texts, they introduce additional complexities. Similarly, the GPT models have a token limit of approximately 16,384, requiring truncation for longer reviews during prediction.

Below, we outline the deployment strategy for each model, detailing our unique approach in applying them to the tasks of sentiment analysis, spam detection, and classification.

3.3.1. GPT Model Deployment and Fine-Tuning

In this phase, we used the GPT Omni family, specifically the `gpt-4o` and `gpt-4o-mini` models, both in their base forms and after fine-tuning, to classify customer reviews into the appropriate rating label (star-rating). Initially, using the prompt shown in **Listing 1**, we applied the base models without any additional training (zero-shot) to predict ratings for the test set. Thanks to their extensive pre-training on large datasets, these GPT models are capable of making fairly accurate predictions even without fine-tuning.

To facilitate communication between our software and the GPT models, we utilized OpenAI's official API. This allowed us to send prompts containing the review text from the feature column and receive the model's predictions in JSON format. The predictions from both models were then stored in two separate columns within the same `test_set.csv` file.

During the fine-tuning phase, the same models underwent additional training to boost their performance by learning from prompt-completion pairs in our training dataset. This fine-tuning process helped the models better understand the subtle nuances and patterns within the data. By employing a multi-epoch training approach, the models continuously improved their comprehension and capabilities through multiple iterations, leading to more accurate predictions and enhanced task execution. This iterative process ensured that the models were well-equipped to provide precise predictions and valuable insights during the prediction phase.

To proceed with the fine-tuning process of our models, we generated two JSONL files containing pairs of prompts and their corresponding completions, as shown in **Listing 2**.

Listing 2. Prompt and completion pairs – JSONL files

```
{"messages": [{"role": "system",
              "content": "You are an AI model tasked with predicting a star rating (1-5) based on the user's review of their hotel experience. Return your response in JSON format: {'rating': integer}."},
             {"role": "user", "content": "..."},
             {"role": "assistant", "content": "{'rating': 2}"}]}
```

(2)

 doi:10.20944/preprints202411.0313.v1

For our study, we initiated two fine-tuning tasks by uploading our training and validation JSONL files into OpenAI's user interface. These tasks were identified by the job IDs: `ftjob-J0ApzOhA7HqE1UxBp7UBAm7U` and `ftjob-pDcZW0NmELoyiotCSuNjnzI1`.

The first fine-tuning process focused on the `gpt-4o` model, which was trained on 2,167,176 tokens. The initial training loss started at 0.4292 and decreased to 0.0208, with a validation loss of 0.1719. The second fine-tuning process involved the `gpt-4o-mini` model, also trained on 2,167,176 tokens. This model began with an initial training loss of 0.5228 and reduced it to 0.0218, achieving a validation loss of 0.1131. These metrics highlight the models' impressive effectiveness for this task and their ability to generalize during the fine-tuning process.

After completing the fine-tuning, both models were used to make predictions on the same test set as their base versions, with the results stored in separate columns.

3.3.2. BERT Model Deployment and Fine-Tuning

In this phase, we focused on training the BERT model, specifically the `bert-base-uncased` variant [43], to tackle the same classification task previously addressed by the LLMs. For this purpose, we employed the `BertForSequenceClassification` model from the transformers' library. The core BERT architecture generates contextual representations of input tokens through its transformer layers. The `BertForSequenceClassification` is a variant version of BERT that includes an additional classification head designed explicitly for sequence classification tasks. This classification head typically consists of a fully connected layer that transforms BERT's outputs into class probabilities. Like its base counterpart, this variant is also based on the transformer architecture and features multiple layers and hidden units. Specifically, the `bert-base-uncased` model includes 12 layers, 768 hidden units, 12 attention heads, and 110 million parameters [43]. The self-attention mechanisms in BERT effectively capture contextual dependencies within input sequences.

We utilized the same base model for the prediction phase, loading the pre-trained models via the pre-trained method.

During the fine-tuning phase, the BERT model was trained to perform rating prediction classification using training sets that included reviews and their corresponding rating labels. This process was executed on Google Colab, leveraging the computational capabilities of a Tesla V100-SXM2-16 GB GPU. The dataset was stored in Google Drive and underwent preprocessing, which included tokenization with the BERT tokenizer. Fine-tuning is involved except for the careful adjustment of hyperparameters, as stated before, and the use of the Adaptive Moment Estimation (Adam) optimizer. The training process was conducted over three epochs, with progress monitored using the `tqdm` library. Backpropagation and optimization took place within the training loop, while validation was carried out using the same dataset utilized by the GPT-4o models. Our approach to fine-tuning was thorough and precise, aligning with the specific needs of the customer rating classification task and preparing the model for deployment. Upon finishing the fine-tuning process, the trained model generated predictions for the same test set used by the GPT-4o models. The complete code, along with the classes employed for training both the BERT and GPT models, as well as training and validation loss metrics and validation accuracy, can be found in an IPython notebook hosted on GitHub [44].

4. Results

In Section 3, we examined the methodology used to run and fine-tune the LLMs and NLP models, detailing how these models were utilized to predict user ratings based on a test set of hotel reviews. This section presents a comparative analysis of the three models, showcasing their evaluation metrics both before and after fine-tuning.

4.1. Overview of Fine-Tuning Metrics

During the fine-tuning process, we collected crucial metrics for each model, including training loss, validation loss, training time, and training cost all of which are presented in Table 1.

 doi:10.20944/preprints202411.0313.v1

Table 1. Fine-tuning metrics.

Model	Resources	Training Loss	Validation Loss	Training Time (Seconds)	Training Cost
ft:gpt-4o	API	0.0208	0.1719	3,412	$29.18
ft:gpt-4o-mini	API	0.0218	0.1131	2,143	$1.38
ft:bert-adam	Tesla V100-SXM2-16 GB	0.6378	0.9526	272.30	$0.79
ft:bert-adamw	Tesla V100-SXM2-16 GB	0.6401	0.9866	278.09	$0.82

Training loss reflects the model's performance during the training phase by measuring the discrepancy between its predictions and the actual target values (e.g., user ratings). A lower training loss indicates that the model is effectively learning from the training data.
On the other hand, validation loss assesses the model's performance on a distinct validation dataset that was not used during training (`validation_set`). This metric is vital for determining the model's ability to generalize. Ideally, validation loss should decline as the model improves; however, if it begins to rise while training loss continues to decrease, this could signal overfitting [45].

It is important to note, however, that directly comparing validation and training losses between fine-tuned NLP models and LLMs may not be straightforward due to differences in their architectures. In contrast, we can safely compare the validation and training losses for similar architectures, such as gpt-4o and gpt-4o-mini, because they share underlying structural similarities and design principles. These models operate on the same foundational architecture, which allows for a more meaningful comparison of their performance metrics. Since they are variations of the same model, the differences in their training and validation losses are likely attributed to the scale of the models rather than fundamental architectural discrepancies. This commonality enables us to draw more accurate conclusions about their relative performance in similar tasks.

4.2. Model Evaluation Phase

Before presenting our results, it's important to emphasize the value of model evaluation. In machine learning and NLP, evaluating models helps us understand their performance, make informed decisions, and improve fine-tuning for specific tasks. Table 2 summarizes the evaluations for each model, including key metrics such as accuracy, recall and f1-score.

Table 2. Comparison of model performance metrics.

Model	Accuracy	Precision	Recall	F1
base:gpt-4o	0.662	0.6616	0.662	0.6553
base:gpt-4o-mini	0.632	0.626	0.632	0.6256
ft:gpt-4o	0.67	0.68	0.67	0.6729
ft:gpt-4o-mini	0.649	0.6611	0.649	0.6532
ft:bert-adam	0.606	0.6189	0.606	0.6079
ft:bert-adamw	0.59	0.6004	0.59	0.5932

4.2.1. Pre-Fine-Tuning Evaluation

In the initial phase of our study, we deployed the `gpt-4o` and `gpt-4o-mini` models to assess the capabilities of LLMs, particularly within the GPT omni family, for sentiment analysis and classification tasks. The results, summarized in Table 2, were promising, with the base models achieving accuracies of 66.2% for `gpt-4o` and 63.2% for `gpt-4o-mini`, indicating that the mini model performed slightly less efficiently—by 3%—compared to its larger counterpart. Notably, these models achieved these accuracy levels without any specific fine-tuning, which is impressive considering that, in a rating system with a scale of 1 to 5, the model had five potential choices. If the

 doi:10.20944/preprints202411.0313.v1

model were to select an option at random, the probability of choosing the correct answer would be merely 20%. This highlights the comparative advantage of LLMs, which demonstrate significant accuracy in sentiment analysis and classification tasks even without fine-tuning.

Another noteworthy observation is the comparison between the results of the base LLM models and those of fine-tuned BERT models. The fine-tuned BERT models, specifically `ft:bert-adam` and `ft:bert-adamw`, were trained on a designated dataset to make predictions on a specific test set. However, their prediction accuracy was considerably much lower (5.6%) than that of the non-fine-tuned LLMs. This finding suggests that base models of LLMs have reached a level of capability that can surpass even specifically trained NLP models. This is not surprising, considering that the base models of LLMs are pre-trained on a wide range of tasks, continuously enhancing their performance in diverse applications.

4.2.2. Post-Fine-Tuning Evaluation

After fine-tuning, the models were deployed and prompted to predict the rating labels on the same test set. The results indicated an accuracy of 67% for `ft:gpt-4o`, 64.9% for `ft:gpt-4o-mini`, and 60.6% and 59% for `ft:bert-adam` and `ft:bert-adamw`, respectively.

One notable observation is the performance disparity between the base GPT models and their fine-tuned counterparts. In the case of the fine-tuned models, `ft:gpt-4o` demonstrated an improvement of just 0.8% over its base model, while ft:gpt-4o-mini showed a more significant enhancement of 1.7% compared to the mini base model. Although both GPT omni models exhibited performance gains, the improvement in `ft:gpt-4o-mini` was approximately double that of its larger counterpart. A possible explanation for this could be related to the number of parameters each model is trained on. The mini model, with fewer parameters (about 8 billion [46]), benefits from a quicker fine-tuning process that requires less training data. Conversely, the larger model (about 1.8 trillion) may need more extensive training data to achieve greater results.

On the other hand, while the fine-tuned BERT models produced respectable results, they did not perform as effectively as the LLMs during fine-tuning. A potential factor for this discrepancy could be the hyperparameters used. As noted at the beginning of Section 3, all models were trained using specific hyperparameters (a learning rate of 2×10^{-5}, a batch size of 6, and training over 3 epochs) to ensure more objective results. For NLP and LLMs, experimenting with various hyperparameters is crucial to maximize each model's performance. Thus, it is evident that all models could achieve better performance with different hyperparameter configurations.

5. Discussion

In the previous sections, we employed three distinct AI models to assess their predictive capabilities on a crucial classification task: sentiment analysis of hotel customer reviews. Initially, the base models were tasked with predicting a rating [1–5] based solely on the content of the reviews. Following this, the models were fine-tuned on a specialized training set using few-shot learning and parameter-efficient tuning techniques. After fine-tuning, the models were evaluated on the same test set to generate predictions, and their performance was thoroughly analyzed.

In this section, we will leverage the results from the classification task to answer each of the research questions outlined in the introduction.

5.1. NLP Models vs LLMs: Which Is Superior for Sentiment Analysis and Star Rating Prediction?

- Research Question 1: Do NLP models or LLMs demonstrate superior efficacy in assessing user-generated content, such as hotel customer reviews, in terms of sentiment analysis and star rating prediction?
- Research Statement 1: Fine-tuned LLMs, and in our case the GPT-4o, stand out as the most effective model for this classification task, reaffirming that larger models with more parameters are better suited for complex tasks such as sentiment analysis. However, the use of smaller models and the selection of optimizers remain critical in balancing accuracy and computational efficiency.

 doi:10.20944/preprints202411.0313.v1

In Sections **0** and **0**, we presented the performance of our models both before and after fine-tuning. All three models delivered solid performance, with the fine-tuned `gpt-4o` model emerging as the strongest, outperforming both the fine-tuned `gpt-4o-mini` and `BERT` models. Specifically, the fine-tuned `gpt-4o` achieved a 0.8% improvement over its base version, a 2.1% advantage over `gpt-4o-mini`, a 6.4% lead over the fine-tuned `BERT` model trained with the Adam optimizer, and an 8% gain over the BERT model trained with AdamW.

The results clearly demonstrate that fine-tuning significantly boosts the performance of LLMs. This underscores the importance of fine-tuning for domain-specific tasks like sentiment analysis in hotel reviews.

The following key insights can be drawn from the results:

- GPT-4o Leads the Pack: The fine-tuned `gpt-4o` model consistently outshined its counterparts—`gpt-4o-mibi` and BERT—demonstrating superior accuracy and efficiency. This suggests that larger, more sophisticated models, when fine-tuned effectively, deliver stronger predictive performance for tasks requiring nuanced language understanding.
- Smaller Models Offer Viable Alternatives: While `gpt-4o-mini` did not perform at the same level as its larger counterpart, it still delivered respectable results. This indicates that smaller models can be effective, especially when computational resources are constrained, though some accuracy may be sacrificed.
- BERT's Performance in Comparison: Although BERT models performed reasonably well, they lagged behind the `gpt-4o` models, particularly after fine-tuning. The fine-tuned `gpt-4o` significantly outperformed both versions of BERT, regardless of whether the Adam or AdamW optimizer was used.
- Impact of Optimizer Selection: The choice of optimizer also played a role in model performance. BERT models trained with Adam outperformed those trained with AdamW, highlighting the importance of selecting the right training strategies.

5.2. GPT-4 Omni Family: Pre-Fine-Tuning Performance Comparison

- Research Question 2: Which model within the GPT-4 Omni family exhibits superior performance before fine-tuning?
- Research Statement 2: Model size significantly influences performance in zero-shot classification tasks, with the `gpt-4o` model exhibiting superior accuracy compared to its mini counterpart, `gpt-4o-mini`. This finding underscores the importance of model parameters in achieving effective language understanding and predictive capabilities.

As stated in Section **0**, LLMs are essentially pre-trained models that have been trained on a wide range of data sourced from the web and private repositories. Similar to how the human brain gradually assimilates knowledge from the environment, pre-trained LLMs incrementally develop their understanding through exposure to diverse datasets. This foundational knowledge is then specialized through the fine-tuning process.

In our research, we observed that the pre-training of these models was sufficient to enable zero-shot predictions on the test set with notable accuracy. Specifically, without fine-tuning, the base `gpt-4o` model achieved an accuracy of 66.2% in its predictions, while the base `gpt-4o-mini` model attained an accuracy of 63.2%.

Naturally, when comparing the performance of these pre-trained models, the `gpt-4o` demonstrates a 3% advantage over its mini counterpart, given that the former is trained on 1.8 trillion parameters compared to the 8 billion parameters of the mini version [46].

Additionally, it is worth noting that even without specialized training, the base models of the LLMs outperformed the BERT model—specifically designed for the classification task of our study—by 5.6% and 2.6%, respectively. This indicates the robust capability of LLMs in handling classification tasks, even in their pre-trained state.

From the findings presented, several conclusions can be drawn:

- Effectiveness of Pre-Training: The ability of pre-trained LLMs, such as `gpt-4o` and `gpt-4o-mini`, to perform zero-shot predictions with significant accuracy demonstrates the effectiveness

 doi:10.20944/preprints202411.0313.v1

of their pre-training on diverse datasets. This highlights the models' inherent capabilities to generalize and apply learned knowledge to new tasks without requiring extensive retraining.

- Model Size and Performance: The observed 3% performance advantage of the `gpt-4o` model over the `gpt-4o-mini` reinforces the idea that larger models with more parameters can capture more nuanced patterns in data, leading to improved predictive accuracy. This underscores the importance of model size in achieving superior performance in NLP tasks.
- Robustness of LLMs Compared to Traditional Models: The fact that both base LLM models outperformed the BERT model—despite BERT being specifically trained for the classification task—indicates that modern LLMs possess a strong foundation in language understanding that allows them to excel even without specialized fine-tuning. This suggests a shift in the landscape of NLP, where pre-trained models can rival or surpass traditional models that have undergone task-specific training.
- Potential for Domain-Specific Applications: The ability of these models to achieve respectable performance without fine-tuning opens avenues for their application in various domains, particularly in situations where computational resources for training are limited. This flexibility makes LLMs a valuable tool for real-world applications, such as sentiment analysis in customer reviews.

5.3. GPT-4 Omni Family: Post-Fine-Tuning Performance with Few-Shot Learning

- Research Question 3: Which model within the GPT-4 Omni family demonstrates superior performance after fine-tuning with few-shot learning?
- Research Statement 3: The fine-tuned `gpt-4o` model exhibits overall superior performance compared to its mini counterpart; however, the mini model shows greater improvement in performance after fine-tuning, suggesting that it adapts more effectively to few-shot learning scenarios despite having fewer parameters.

In Section **0**, the fine-tuned `gpt-4o` once again proved to be more efficient than its mini counterpart by 2.1%. However, its performance after fine-tuning was only 0.8% better, whereas the mini model improved its performance by 1.7% following fine-tuning. This specific difference in performance raises further questions: Could it be that the `gpt-4o` has reached a state of overfitting with the given training set? Or perhaps the `gpt-4o-mini` model was able to adjust its parameters more rapidly due to its smaller size?

The first possibility can be ruled out since Table 1 shows a lower training loss for the fine-tuned `gpt-4o` model (0.0208 compared to 0.0218). However, if we look at the validation loss column in the same table, we may find some answers (0.1719 versus 0.1131). Thus, during fine-tuning, the mini model exhibited a lower validation loss.

The second possibility, however, seems more significant, as a model with 1.8 trillion parameters may have a harder time specializing its knowledge compared to a model with 8 billion parameters. In other words, a smaller model can better adapt its knowledge with a limited labeled dataset, while a larger model requires substantially more training data to achieve similar or superior performance.

From the provided analysis, several key findings can be derived:

- Performance Comparison: The fine-tuned `gpt-4o` outperforms its mini counterpart by 2.1% in overall efficiency. However, the fine-tuned `gpt-4o` shows only a marginal improvement of 0.8% post fine-tuning, whereas the mini model demonstrates a more substantial enhancement of 1.7%. This indicates that while the `gpt-4o` is more effective overall, it may have limitations in adapting further from its fine-tuning process compared to the mini model.
- Possible Overfitting: The small improvement in performance of the `gpt-4o` after fine-tuning raises questions about potential overfitting. While the lower training loss for the fine-tuned `gpt-4o` suggests that it is learning effectively, the subsequent validation loss comparison indicates that the mini model is not only learning but generalizing better to unseen data.
- Validation Loss Insights: The lower validation loss of the `gpt-4o-mini` model (0.1131) compared to the `gpt-4o` (0.1719) implies that the mini model has a better capacity for generalization during the fine-tuning phase. This suggests that the mini model, despite its

 doi:10.20944/preprints202411.0313.v1

smaller size, is capable of effectively adjusting its parameters to achieve a more robust performance on validation data.

- Impact of Model Size on Specialization: The analysis highlights the inherent challenges larger models face in specializing their knowledge when fine-tuning. The `gpt-4o`, with 1.8 trillion parameters, may require more extensive and varied training data to fine-tune effectively, while the smaller `gpt-4o-mini` can adapt more quickly and effectively to specific tasks with limited labeled data. This suggests a trade-off between model size and adaptability in certain contexts.
- Training Data Requirements: The findings reinforce the notion that larger models need significantly more training data to achieve similar or superior performance as smaller models in few-shot or limited data scenarios. This is particularly relevant for applications where data availability is constrained.

5.4. Evaluating the Cost of Fine-Tuning LLMs: Performance vs. Expense in the Tourism Industry

- Research Question 4: How significant is the cost of fine-tuning LLMs in achieving optimal results, and how does it impact performance in the tourism sector?
- Research Statement 4: The cost of fine-tuning LLMs is a critical factor in achieving optimal results, as demonstrated by the significantly lower training costs and improved performance of the `gpt-4o-mini` model compared to the larger `gpt-4o`, making it a more practical and economically viable option for the tourism sector, where training on large datasets can be prohibitively expensive.

In Table 1, we observe that the `gpt-4o` model requires 56.86 minutes for training, while the `gpt-4o-mini` takes only 35.71 minutes, and the BERT model requires just 4.53 minutes. Additionally, the training costs reveal that fine-tuning the `gpt-4o` with 3,200 training samples and 800 validation samples costs $29.18, compared to $1.38 for the mini and an even lower $0.79 for BERT. In other words, training the `gpt-4o` is 21.1 times more expensive than training the mini model and 36.94 times more expensive than training the BERT model. This significant difference in both time and cost between the larger `gpt-4o` and the smaller models is clearly due to the size, as the larger model requires considerably more time and GPU resources for fine-tuning than the smaller models.

However, there is a disparity in training time and costs between the `gpt-4o` and the `gpt-4o-mini`. Although the fine-tuning cost of the `gpt-4o` is 21.1 times higher, the training time is only 1.59 times greater than that of the mini counterpart. This raises questions about how model training costs are calculated. The answer lies not in the training time but in the resources required for each model. A larger model necessitates more GPUs for training, while a smaller model requires fewer. To decouple charges based on resources and time, OpenAI and similar companies have adopted the concept of tokens (A descriptive output of 1,000 characters, calculated through the `tiktoken` library, corresponds to 204 tokens [4]). In our case, the `gpt-4o` costs $2.50 for every 1 million input tokens and $10 for every 1 million output tokens, while the `gpt-4o-mini` costs only $0.15 for each million input tokens and $0.60 for each million output tokens [47], making the mini model significantly more economical for training.

Having established which model is more cost-effective during training, we can confidently explore the performance relative to the cost. In Research Statement 3, we noted that the `gpt-4o-mini` model, while less efficient than its larger counterpart, improved its performance by 1.7% compared to a 0.8% improvement for the `gpt-4o` using the same training and validation set. Therefore, when considering the performance of the `gpt-4o-mini` in relation to its training costs, it clearly offers better value for money. Specifically, in the tourism sector, where we deal not with thousands but millions of training data, the cost of training a large model like `gpt-4o` would be prohibitively expensive for hotels, whereas the mini model would be much more affordable.

5.5. Leveraging LLMs for Customer Review Evaluation: Automation and Insights in the Hospitality Sector

- Research Question 5: Can LLMs be effectively utilized for the evaluation of customer reviews, and how can they revolutionize the hospitality sector by automating review responses and extracting actionable insights?

 doi:10.20944/preprints202411.0313.v1

- Research Statement 5: Proven by this study, LLMs have the potential to significantly support the hospitality sector by automating review evaluation, enhancing customer interactions, and providing valuable insights for business improvement.

Artificial intelligence is rapidly advancing, and its applications across various sectors of the economy are becoming increasingly numerous. Particularly with the arrival of LLMs, such as GPT-3.5 and ChatGPT in early 2023 [48], within a year, TV screens were filled with advertisements for AI-powered air conditioners, smartphones, personal assistants, and chatbots—even for communication with public services. While some advertisements may exploit the term "AI" for marketing purposes, many of these innovations genuinely have the potential to make everyday tasks more accessible to the public.

On the business side, companies are keen to create unique products that attract more customers or improve customer satisfaction. In highly competitive sectors like tourism, the need for product differentiation is even greater [49]. This is where AI, and specifically LLMs, come into play. Ideas like customer service chatbots, AI-driven booking systems, and models that detect spam in customer reviews [50], among others, all leverage AI, deep learning, and LLMs.

However, measuring customer satisfaction presents a unique challenge. Businesses can gauge satisfaction by reading reviews on social platforms, Google My Business, TripAdvisor, etc. For a tourism entrepreneur, reading and responding to reviews for an entire summer season could take as much time as another summer in front of a screen. The solution proposed in this research is the use and training of LLMs for sentiment analysis and classification of hotel reviews. In other words, an automated system would allow the entrepreneur to quickly identify critical reviews and respond to each. To achieve this, we used OpenAI's GPT Omni family models and directly compared them with older methods like traditional NLP. Each model in the study was tasked, both pre- and post-fine-tuning, to perform sentiment analysis and classify reviews, predicting the customer rating. The results were encouraging, with the `gpt-4o` model leading in performance, and the `gpt-4o-mini` being recognized for its cost-effectiveness.

Why, then, is it important for LLMs to understand customer sentiment? While LLMs are not initially designed for classification tasks, our results, along with numerous studies presented in the literature review, show that they are more than capable. The innovation of LLMs lies in generating text in response to a prompt. In our research, knowing the sentiment of a hotel guest, and having trained the appropriate LLM, we can use the same fine-tuned model to respond automatically to reviews on social media, TripAdvisor, and similar platforms. A fine-tuned LLM not only benefits from its general knowledge acquired during pre-training but also becomes specialized in a specific task, enabling it to automate many processes in the hotel and tourism sectors more broadly.

6. Conclusions

In conclusion, this study highlights the significant impact of model size and fine-tuning on the performance of LLMs in sentiment analysis, particularly within the context of the tourism sector. Our findings demonstrate that while the fine-tuned GPT-4o model achieves superior overall performance, the GPT-4o-mini exhibits greater adaptability and improvement during fine-tuning, making it a more cost-effective choice. The disparity in training time and costs emphasizes the importance of resource allocation when selecting a model, with smaller models offering substantial value for money without sacrificing performance. As the tourism industry increasingly relies on user-generated content analysis and automation, the insights from this research underscore the necessity for practitioners to carefully consider both the economic and performance implications of their model choices, ensuring effective sentiment analysis while managing costs in an environment where data volume is immense. Ultimately, our results advocate for a balanced approach that leverages the strengths of different models according to specific operational needs and resource availability.

Author Contributions: Conceptualization, K.I.R. and N.D.T.; methodology, K.I.R., N.D.T. and D.K.N.; software, K.I.R.; validation, K.I.R. and N.D.T.; formal analysis, K.I.R., N.D.T. and D.K.N.; investigation, K.I.R., N.D.T. and D.K.N.; resources, K.I.R.; data curation, K.I.R.; writing—original draft preparation, K.I.R. and N.D.T.; writing—

review and editing, K.I.R., N.D.T. and D.K.N.; visualization, K.I.R.; supervision, D.K.N. and N.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the reported results can be found at ref [44].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Travel and Tourism: Contribution to Global GDP 2023 | Statista Available online: https://www.statista.com/statistics/233223/travel-and-tourism-total-economic-contribution-worldwide/ (accessed on 11 October 2024).
2. Cho, S.; Pekgün, P.; Janakiraman, R.; Wang, J. The Competitive Effects of Online Reviews on Hotel Demand. *J Mark* **2023**, *88*, 40–60, doi:10.1177/00222429231191449.
3. Qi, J.; Yan, S.; Zhang, W.; Zhang, Y.; Liu, Z.; Wang, K. Research on Tibetan Tourism Viewpoints Information Generation System Based on LLM. *2024 12th International Conference on Intelligent Computing and Wireless Optical Communications (ICWOC)* **2024**, 35–41, doi:10.1109/ICWOC62055.2024.10684948.
4. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in E-Commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation. *Natural Language Processing Journal* **2024**, *6*, 100056, doi:10.1016/J.NLP.2024.100056.
5. Models - OpenAI API Available online: https://platform.openai.com/docs/models (accessed on 11 October 2024).
6. Sakas, D.P.; Reklitis, D.P.; Terzi, M.C.; Vassilakis, C. Multichannel Digital Marketing Optimizations through Big Data Analytics in the Tourism and Hospitality Industry. *Journal of Theoretical and Applied Electronic Commerce Research 2022, Vol. 17, Pages 1383-1408* **2022**, *17*, 1383–1408, doi:10.3390/JTAER17040070.
7. Priya, C.S.R.; Deepalakshmi, P. Sentiment Analysis from Unstructured Hotel Reviews Data in Social Network Using Deep Learning Techniques. *International Journal of Information Technology (Singapore)* **2023**, *15*, 3563–3574, doi:10.1007/S41870-023-01419-Z/METRICS.
8. Wen, Y.; Liang, Y.; Zhu, X. Sentiment Analysis of Hotel Online Reviews Using the BERT Model and ERNIE Model—Data from China. *PLoS One* **2023**, *18*, e0275382, doi:10.1371/JOURNAL.PONE.0275382.
9. Kusumaningrum, R.; Nisa, I.Z.; Jayanto, R.; Nawangsari, R.P.; Wibowo, A. Deep Learning-Based Application for Multilevel Sentiment Analysis of Indonesian Hotel Reviews. *Heliyon* **2023**, *9*, e17147, doi:10.1016/J.HELIYON.2023.E17147/ASSET/076BDEC5-7A1A-46C0-9AAF-0FA215198B7F/MAIN.ASSETS/GR7.JPG.
10. Chang, V.; Liu, L.; Xu, Q.; Li, T.; Hsu, C.H. An Improved Model for Sentiment Analysis on Luxury Hotel Review. *Expert Syst* **2023**, *40*, e12580, doi:10.1111/EXSY.12580.
11. Zhang, D.; Niu, B. Leveraging Online Reviews for Hotel Demand Forecasting: A Deep Learning Approach. *Inf Process Manag* **2024**, *61*, 103527, doi:10.1016/J.IPM.2023.103527.
12. Ounacer, S.; Mhamdi, D.; Ardchir, S.; Daif, A.; Azzouazi, M. Customer Sentiment Analysis in Hotel Reviews Through Natural Language Processing Techniques. *International Journal of Advanced Computer Science and Applications* **2023**, *14*, 569–579, doi:10.14569/IJACSA.2023.0140162.
13. Li, M.; Yin, D.; Qiu, H.; Bai, B. A Systematic Review of AI Technology-Based Service Encounters: Implications for Hospitality and Tourism Operations. *Int J Hosp Manag* **2021**, *95*, 102930, doi:10.1016/J.IJHM.2021.102930.
14. Pillai, R.; Sivathanu, B. Adoption of AI-Based Chatbots for Hospitality and Tourism. *International Journal of Contemporary Hospitality Management* **2020**, *32*, 3199–3226, doi:10.1108/IJCHM-04-2020-0259/FULL/XML.
15. Huang, A.; Chao, Y.; de la Mora Velasco, E.; Bilgihan, A.; Wei, W. When Artificial Intelligence Meets the Hospitality and Tourism Industry: An Assessment Framework to Inform Theory and Management. *Journal of Hospitality and Tourism Insights* **2022**, *5*, 1080–1100, doi:10.1108/JHTI-01-2021-0021/FULL/XML.
16. Miao, L.; Yang, F.X. Text-to-Image AI Tools and Tourism Experiences. *Ann Tour Res* **2023**, *102*, 103642, doi:10.1016/J.ANNALS.2023.103642.
17. Wang, W.; Kumar, N.; Chen, J.; Gong, Z.; Kong, X.; Wei, W.; Gao, H. Realizing the Potential of Internet of Things for Smart Tourism with 5G and AI. *IEEE Netw* **2020**, *34*, 295–301, doi:10.1109/MNET.011.2000250.
18. Chi, O.H.; Gursoy, D.; Chi, C.G. Tourists' Attitudes toward the Use of Artificially Intelligent (AI) Devices in Tourism Service Delivery: Moderating Role of Service Value Seeking. *J Travel Res* **2022**, *61*, 170–185, doi:10.1177/0047287520971054/ASSET/IMAGES/LARGE/10.1177_0047287520971054-FIG1.JPEG.
19. Gupta, S.; Modgil, S.; Lee, C.K.; Sivarajah, U. The Future Is Yesterday: Use of AI-Driven Facial Recognition to Enhance Value in the Travel and Tourism Industry. *Information Systems Frontiers* **2023**, *25*, 1179–1195, doi:10.1007/S10796-022-10271-8/FIGURES/4.

20. Zhang, B.; Zhu, Y.; Deng, J.; Zheng, W.; Liu, Y.; Wang, C.; Zeng, R. "I Am Here to Assist Your Tourism": Predicting Continuance Intention to Use AI-Based Chatbots for Tourism. Does Gender Really Matter? *Int J Hum Comput Interact* **2023**, *39*, 1887–1903, doi:10.1080/10447318.2022.2124345.
21. Qi, J.; Yan, S.; Zhang, W.; Zhang, Y.; Liu, Z.; Wang, K. Research on Tibetan Tourism Viewpoints Information Generation System Based on LLM. *2024 12th International Conference on Intelligent Computing and Wireless Optical Communications (ICWOC)* **2024**, 35–41, doi:10.1109/ICWOC62055.2024.10684948.
22. Wei, Q.; Yang, M.; Wang, J.; Mao, W.; Xu, J.; Ning, H. TourLLM: Enhancing LLMs with Tourism Knowledge Available online: https://arxiv.org/abs/2407.12791v1 (accessed on 12 October 2024).
23. Banerjee, A.; 1⋆, A.S.; Wörndl, W. Enhancing Tourism Recommender Systems for Sustainable City Trips Using Retrieval-Augmented Generation Available online: https://arxiv.org/abs/2409.18003v1 (accessed on 12 October 2024).
24. Vasic, I.; Fill, H.G.; Quattrini, R.; Pierdicca, R. LLM-Aided Museum Guide: Personalized Tours Based on User Preferences. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2024**, *15029 LNCS*, 249–262, doi:10.1007/978-3-031-71710-9_18.
25. Chen, B.O.; Chen, B.; Dai, X.; Guo, H.; Guo, W.; Liu, W.; Liu, Y.; Qin, J.; Tang, R.; Wang, Y.; et al. All Roads Lead to Rome: Unveiling the Trajectory of Recommender Systems Across the LLM Era Available online: https://arxiv.org/abs/2407.10081v1 (accessed on 12 October 2024).
26. Balamurali, O.; Abhishek Sai, A.M.; Karthikeya, M.; Anand, S. Sentiment Analysis for Better User Experience in Tourism Chatbot Using LSTM and LLM. *2023 9th International Conference on Signal Processing and Communication, ICSC 2023* **2023**, 456–462, doi:10.1109/ICSC60394.2023.10441148.
27. Falatouri, T.; Hrusecka, D.; Fischer, T. Harnessing the Power of LLMs for Service Quality Assessment from User-Generated Content. *IEEE Access* **2024**, *12*, 99755–99767, doi:10.1109/ACCESS.2024.3429290.
28. Buitrago-Esquinas, E.M.; Yñiguez-Ovando, R.; Puig-Cabrera, M.; Santos, M.C.; Santos, J.A.C. Artificial Intelligence and Sustainable Tourism Planning: A Hetero-Intelligence Methodology Proposal. *Tourism & Management Studies* **2024**, *20*, 45–59, doi:10.18089/TMS.2024SI04.
29. Luca Secchi Knowledge Graphs and Large Language Models for Intelligent Applications in the Tourism Domain. *Università di Cagliari* **2024**, *7*, 343–354, doi:https://hdl.handle.net/11584/391989.
30. Kodors, S.; Kanepe, G.; Zeps, D.; Zarembo, I.; Litavniece, L. RAPID DEVELOPMENT OF CHATBOT FOR TOURISM PROMOTION IN LATGALE. *ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference* **2024**, *2*, 179–182, doi:10.17770/ETR2024VOL2.8060.
31. Hsu, C.H.C.; Tan, G.; Stantic, B. A Fine-Tuned Tourism-Specific Generative AI Concept. *Ann Tour Res* **2024**, *104*, 103723, doi:10.1016/J.ANNALS.2023.103723.
32. Qi, J.; Yan, S.; Zhang, Y.; Zhang, W.; Jin, R.; Hu, Y.; Wang, K. RAG-Optimized Tibetan Tourism LLMs: Enhancing Accuracy and Personalization Available online: https://arxiv.org/abs/2408.12003v1 (accessed on 12 October 2024).
33. Balfroid, M.; Vanderose, B.; Devroey, X. Towards LLM-Generated Code Tours for Onboarding. *Proceedings - 2024 ACM/IEEE International Workshop on NL-Based Software Engineering, NLBSE 2024* **2024**, 65–68, doi:10.1145/3643787.3648033.
34. Gonzalez-Garcia, L.; González-Carreño, G.; Rivas Machota, A.M.; Padilla Fernández-Vega, J. Enhancing Knowledge Graphs with Microdata and LLMs: The Case of Schema.Org and Wikidata in Touristic Information. *Electronic Library* **2024**, *42*, 443–454, doi:10.1108/EL-06-2023-0160/FULL/XML.
35. Meyer, S.; Singh, S.; Tam, B.; Ton, C.; Ren, A. A Comparison of LLM Finetuning Methods & Evaluation Metrics with Travel Chatbot Use Case Available online: https://arxiv.org/abs/2408.03562v1 (accessed on 12 October 2024).
36. Carvalho, I.; Ivanov, S. ChatGPT for Tourism: Applications, Benefits and Risks. *Tourism Review* **2024**, *79*, 290–303, doi:10.1108/TR-02-2023-0088/FULL/XML.
37. Sioziou, K.; Zervas, P.; Giotopoulos, K.; Tzimas, G. Comparative Analysis of Large Language Models in Structured Information Extraction from Job Postings. *Communications in Computer and Information Science* **2024**, *2141 CCIS*, 82–92, doi:10.1007/978-3-031-62495-7_7.
38. Liyanage, V.; Buscaldi, D.; Forcioli, P. Detecting AI-Enhanced Opinion Spambots: A Study on LLM-Generated Hotel Reviews Available online: https://aclanthology.org/2024.ecnlp-1.8 (accessed on 12 October 2024).
39. Trip Advisor Hotel Reviews Available online: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews# (accessed on 13 October 2024).
40. Alam, M.H.; Ryu, W.J.; Lee, S.K. Joint Multi-Grain Topic Sentiment: Modeling Semantic Aspects for Online Reviews. *Inf Sci (N Y)* **2016**, *339*, 206–223, doi:10.1016/J.INS.2016.01.013.
41. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K.; Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification. *Electronics 2024, Vol. 13, Page 2034* **2024**, *13*, 2034, doi:10.3390/ELECTRONICS13112034.
42. Zhang, K.; Zhou, F.; Wu, L.; Xie, N.; He, Z. Semantic Understanding and Prompt Engineering for Large-Scale Traffic Data Imputation. *Information Fusion* **2024**, *102*, 102038, doi:10.1016/J.INFFUS.2023.102038.

 doi:10.20944/preprints202411.0313.v1

43. Pretrained Models — Transformers 3.3.0 Documentation Available online: https://huggingface.co/transformers/v3.3.1/pretrained_models.html (accessed on 17 December 2023).
44. GitHub - Applied-AI-Research-Lab/LLMs-in-Tourism-GPT-4-Omni-vs-BERT: Leveraging LLMs in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis Available online: https://github.com/Applied-AI-Research-Lab/LLMs-in-Tourism-GPT-4-omni-vs-BERT/tree/main (accessed on 19 October 2024).
45. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. *Big Data and Cognitive Computing 2024, Vol. 8, Page 63* **2024**, *8*, 63, doi:10.3390/BDCC8060063.
46. What Is GPT-4o Mini? How It Works, Use Cases, API & More | DataCamp Available online: https://www.datacamp.com/blog/gpt-4o-mini (accessed on 16 October 2024).
47. Pricing | OpenAI Available online: https://openai.com/api/pricing/ (accessed on 18 October 2024).
48. Roumeliotis, K.I.; Tselikas, N.D. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet 2023, Vol. 15, Page 192* **2023**, *15*, 192, doi:10.3390/FI15060192.
49. Marinagi, C.; Trivellas, P.; Sakas, D.P. The Impact of Information Technology on the Development of Supply Chain Competitive Advantage. *Procedia Soc Behav Sci* **2014**, *147*, 586–591, doi:10.1016/J.SBSPRO.2014.07.161.
50. Gupta, D.; Bhargava, A.; Agarwal, D.; Alsharif, M.H.; Uthansakul, P.; Uthansakul, M.; Aly, A.A. Deep Learning-Based Truthful and Deceptive Hotel Reviews. *Sustainability 2024, Vol. 16, Page 4514* **2024**, *16*, 4514, doi:10.3390/SU16114514.