

How well can ChatGPT forecast tourism demand?

Wu., D. C., Li, W., Wu, J., Hu, M. and Shen, S.

NOTICE: this is the authors' version of a work that was accepted for publication in Tourism Management. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Tourism Management, volume 108, June 2025, 105119

The final definitive version in Tourism Management is available online at:

<https://doi.org/10.1016/j.tourman.2024.105119>

© 2025. This manuscript version is made available under the CC-BY-NC-ND 4.0 license

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

How well can ChatGPT forecast tourism demand?

Abstract

ChatGPT has demonstrated remarkable capabilities across various natural language processing (NLP) tasks. However, its potential for forecasting tourism demand from temporal data, specifically historical tourism arrivals data, remains an unexplored frontier. This research presents the first attempt to conduct an extensive Zero-shot and Chain-of-Thought analysis of ChatGPT's capabilities in tourism demand forecasting, under various temporal scenarios. Based on the Macau inbound tourism arrivals dataset, our empirical findings indicate that the predictive capability of ChatGPT-4 is noteworthy compared to the three benchmark time series models (Naïve, Exponential Smoothing, SARIMA) and the three benchmark machine learning models (Random Forest, Multilayer Perceptron, Long Short-Term Memory), especially when the forecast horizon is relatively short. Furthermore, compared to Zero-shot prompts, engaging in continuous dialogue can enhance the forecast accuracy of ChatGPT-4. This performance of ChatGPT highlights its potential for quantitative data prediction as a new user-friendly and cost-effective management tool.

Keywords: tourism demand forecasting; ChatGPT; zero-shot; Chain-of-Thought; artificial intelligence

Introduction

Accurate and timely forecasting of tourism demand is crucial for policy formulation, investment decisions, and strategic planning within the tourism industry (Law et al., 2019; Song et al., 2019). Such forecasting facilitates the prediction of demand fluctuations, enables effective resource management, and helps craft efficient recovery strategies during crises (Liu et al., 2018). As competition intensifies among an increasing number of tourism destinations for scarce resources, this dynamic can result in both costs and benefits coexisting (Lim et al., 2009). Consequently, accurate forecasting is particularly vital for decision-makers who aim to capitalize on the growth of the tourism market and simultaneously balance the ecological and social carrying capacities of these destinations (Song et al., 2019). In light of these challenges, there has been a continuous and rapid development of tourism demand forecasting techniques over recent decades. These methods encompass a range of time series models, innovative econometric models, and an array of AI-based techniques.

The surge in data availability and computational capacity has recently drawn significant attention to the application of generative artificial intelligence and large language models (LLMs). Among these LLMs, ChatGPT stands out as a particularly

influential tool, inspiring a wealth of research that explores its potential across various domains, such as the medical field (Biswas, 2023; Klang et al., 2023; Frosolini et al., 2023), the computer technology field (He & Hooi, 2024; Yan et al., 2024), and the educational field (Ali et al., 2024; Gao et al., 2023). The expanding frontier of financial forecasting research also witnesses ChatGPT's integration into predictive analytics. For example, Lopez-Lira and Tang (2023) delve into the predictive prowess of ChatGPT and analogous expansive language models, scrutinizing their capacity to forecast stock price fluctuations. These studies underscore the burgeoning role of LLMs in forecasting analysis. However, in the realm of tourism economics, the application of LLMs, and their potential for tourism forecasting, remains largely unexplored territory.

As such, we have our research question: *How well can ChatGPT forecast tourism demand?* To answer this question, we further break it down into three sub-questions.

RQ1: Compared to other forecasting models, how effective is the performance of ChatGPT-4? To answer this question, this study is to conduct an in-depth Zero-shot analysis to evaluate the performance of ChatGPT-4 in forecasting the tourist arrivals from eight key source markets to Macau. The eight source markets include Mainland China, Hong Kong, Taiwan, the Philippines, Indonesia, South Korea, the United States, and Malaysia. For comparative purposes, the study compare ChatGPT's performance against a set of benchmark models, which include three time-series models—Naïve model, Exponential Smoothing model, and SARIMA model—as well as three machine learning models—Random Forest model, Multilayer Perceptron model, and Long Short-Term Memory model.

RQ2: In the realm of “Chain of thought” (CoT) prompt versus “Zero-shot” prompt, can a continuous dialogue method improve the forecasting accuracy of ChatGPT-4? LLMs, particularly ChatGPT-4, have shown a remarkable ability to understand language, context, and subtle nuances. They can comprehend and respond contextually, tailoring their answers to the specific requirements or preferences outlined in the prompts. This capability enhances their performance in personalized interactions and task-specific scenarios. To fully explore the potential of ChatGPT-4 in travel prediction tasks, we utilize the CoT prompt strategy to reveal whether the method of sustained conversation can enhance ChatGPT's performance in forecasting tourism demand.

RQ3: How does ChatGPT-4's performance stack up against other forecasting models across various temporal scenarios? Tourism demand is subject to the influence of numerous unpredictable factors, such as natural disasters, human-made crises, and emergent epidemics (Wang, 2009). A single forecasting method may not consistently perform well across different stages of a crisis. For instance, models that were accurate in forecasting tourism demand before the COVID-19 pandemic may not perform as well during or after the pandemic (Zhang et al., 2021). We are therefore comparing the forecasting results of ChatGPT with other benchmark models across three distinct periods: before, during, and after the COVID-19 pandemic. This comparison aims to explore the robust capability of LLMs in tourism forecasting under various scenarios.

This study makes a threefold contribution to the tourism forecasting literature. Firstly,

by initially examining the forecasting capabilities of the LLM ChatGPT-4 in forecasting tourism demand, this study provides the tourism forecasting field with an innovative methodological tool. Secondly, by examining the impact of different prompting strategies on the predictive performance of ChatGPT-4, this study recommends optimal prompting strategies to improve the predictive performance of ChatGPT-4. Finally, this study validates ChatGPT-4's forecasting performance across different crisis periods, demonstrating its reliability and stability in predictive capabilities. ChatGPT is therefore positioned as an effective and reliable tool for tourism forecasting, particularly in times of crisis.

Literature review

Tourism demand forecasting

Accurate tourism demand forecasts are critical as they provide tourism practitioners and researchers with important information to help them make decisions on activities such as earmarking resources and identifying priorities and potential risks. Tourism demand forecasting encompasses both qualitative and quantitative approaches. In the realm of quantitative forecasting models, research in traditional tourism and hospitality demand forecasting has primarily applied three types of models: time series models, advanced econometric models, and artificial intelligence models (Wu et al., 2017).

Classic time series models analyze future trends of tourism demand by delving into the historical patterns of variables. These methods are relatively cost-effective in term of data collection and model construction, straightforward to operate, and well-developed. Under specific circumstances, they can provide accurate predictions and even outperform other modeling approaches (Song & Li, 2008). However, a significant drawback of these methods is that they primarily concentrate on trend extrapolation and fail to incorporate other explanatory variables that have an impact on tourism demand (Goh & Law, 2002). This shortcoming restricts their effectiveness in dissecting the intricate interactions between tourism demand and its influencing factors.

Early widely-used univariate time series models encompass the Naïve models (including both the no-change Naïve and the constant growth rate Naïve II) (Athanasopoulos et al., 2018), various Exponential Smoothing (ES) models, such as the Holt-Winters ES model, and the Autoregressive Moving Average (ARMA) family models, including ARIMA and SARIMA models, as well as Structural Time Series (STS) models (Chen et al., 2019; Dantas et al., 2017; Nowman & Van Dellen, 2012). Since time series models solely depend on historical trends for analysis and do not integrate variables affecting tourism demand, they have a strong reliance on high-quality time series data. However, due to the delay in data publication, their predictive timeliness is limited. Recent studies have begun to extend non-causal time series models to more advanced econometric frameworks by including additional explanatory

variables (Song et al., 2019; Yang & Wong, 2012).

Econometric models have been widely applied in the field of tourism and hotel demand forecasting. These methods enable an exploration of the relationships between variables and their determinants within economic activities, facilitating the analysis of economic phenomena to forecast the future state of relevant variables. Compared to time-series forecasting models, the advantage of econometric methods lies in their ability to incorporate explanatory variables that affect tourism demand into the model, reflecting the causal relationship between tourism demand and its determinants. The flexible modeling approach of econometrics facilitates the assessment of the elasticity of factors influencing tourism demand (Song & Li, 2008), and allows for obtaining tourism forecasting outcomes under different economic growth expectations. Decision-makers in tourism destinations strive to understand the crucial determinants of tourism product demand to formulate effective policies and strategies. Meanwhile, econometric analysis offers a valuable economic vantage point through which to dissect and understand the fluctuations in demand. Several contemporary econometric models have been increasingly applied in the forecasting field, with frequently used models including the Autoregressive Distributed Lag Model (ADLM) (Li et al., 2011; Shen et al., 2008), the Error Correction Model (ECM) (Ahn & Lee, 2021; Seo et al., 2009), and the Vector Autoregression Model (VARM) (Liu et al., 2018; Oh, 2005; Zhang et al., 2022). However, we do not consider econometric models in this research because our primary variable of interest is the time series of tourist arrivals. Econometric models typically require to include not only tourism demand but also its determinants into the modeling process. Considering research comparability, we aim to use the same data input and focus on univariate models. Therefore, econometric models are not included in the forecasting comparison of the present study.

The tourism economic system exhibits nonlinear characteristics, which poses challenges for time series and econometric models in capturing these features. Conversely, methods based on artificial intelligence (AI) technology have advantages in nonlinear prediction, prompting the increasing application of AI models in the field of tourism forecasting. Consequently, research utilizing machine learning for prediction has garnered attention. AI-based models can flexibly handle the nonlinear components of tourism systems and are effectively applied in big data predictions for tourism. Among AI models, Artificial Neural Networks (ANN) are the most widely applied (Palmer et al., 2006), characterized by their ability to model and predict tourism demand nonlinearly through continuous data training. However, a drawback is the model's "black box" nature, obscuring the economic variables' effects on tourism demand. Various neural network models, including Multilayer Perceptron (MLP), Radial Basis Function, Generalized Regression Neural Network, and Elman Neural Network, have been employed in tourism and hotel forecasting practices (Cang, 2011; Cho, 2003; Claveria et al., 2015). Recent studies indicate that ensemble methods, which integrate different model outcomes, can produce superior results (Azad et al., 2022; Belmahdi et al., 2020; Hajirahimi & Khashei, 2019, 2023; Lim & Zohren, 2021; Shi et al., 2012;

Sina et al., 2023; Wang & Jiang, 2019; Yu et al., 2019).

Wu et al. (2024b) conducted a systemic review of the literature for tourism and hotel demand forecasting with big data. 86 studies published between 2012 and 2023 were reviewed and discussed. Wu et al. (2022; 2024a) applied sentiment indices to forecasting hotel occupancy rates and indicated the usefulness of online reviews on hotel demand forecasting. Sun et al. (2022) integrated economic and online behavior variables as explanatory variables for Beijing's tourist numbers into a Bagging-based multivariate deep learning model for ensemble prediction, showing that this model outperforms the SARIMA baseline model in MAPE and NRMSE prediction accuracy metrics and surpasses the baseline AI prediction model in DM and PT tests. Antonio et al. (2019) also pioneered the use of the XGBoost ensemble learning model to predict hotel booking cancellations based on online tourist reviews.

Empirical studies generally demonstrate that no single model consistently outperforms others, with each method having its limitations (Chen et al., 2019; Goh & Law, 2002; Song & Li, 2008; Wu et al., 2017; Wu et al., 2024). Time series and econometric models rely on the stability of historical patterns and economic structures, whereas AI models depend on the quality and scale of available training data. AI models have shown potential in tourism forecasting, yet the performance of LLMs, as a forefront technology in the AI field, in this domain remains unexplored, presenting a critical question for this research to address.

ChatGPT and forecasting

Natural Language Processing (NLP), a pivotal branch within computer science and AI, strives for computer proficiency in understanding, interpreting, and generating human language. The advent of deep learning catalyzed revolutionary shifts in NLP, particularly through Transformer-based models recognized for their superior linguistic capabilities (Vaswani et al., 2017). OpenAI's introduction of the Generative Pre-trained Transformer (GPT) models marked a significant advancement in NLP by employing extensive pre-training on large datasets to learn universal language patterns and structures. This pre-training and fine-tuning methodology significantly enhanced performance across various NLP tasks, such as text classification and generation (Floridi & Chiriatti, 2020).

ChatGPT, built on the Transformer architecture proposed by Vaswani et al. (2017), excels in sequential data processing. Its core, the self-attention mechanism, enables consideration of all elements within a sequence, capturing long-distance dependencies effectively. This attribute renders the Transformer model particularly suited for NLP tasks including text generation and translation. ChatGPT benefits from this architecture, absorbing universal language patterns through extensive pre-training. This stage involves self-supervised learning on a vast text corpus, optimizing the prediction of subsequent words to solidify a foundation for task-specific fine-tuning.

Since the debut of the first GPT model by OpenAI in 2018, the GPT series has drawn

significant attention in the NLP field. GPT-1 showcased the potential of deep learning in language comprehension and generation through large-scale corpus pre-training (Floridi & Chiriatti, 2020). GPT-2 and GPT-3, with their expanded model sizes and data training, further improved textual coherence and adaptability to a broader range of tasks with minimal to no fine-tuning required (Kalla et al., 2023; Wu et al., 2023). The latest iteration, GPT-4, extends these capabilities, offering refined language understanding and task adaptability, with ongoing research into its specific enhancements and performance improvements.

The burgeoning research on ChatGPT spans a wide array of applications, from healthcare to education, showcasing its transformative potential across industries. Recent studies have garnered significant attention towards the application of ChatGPT and other LLMs in prediction tasks across various domains, including corporate performance evaluation, demand forecasting, entity trajectory prediction, and financial market forecasting. For instance, Li et al. (2024) demonstrates that ChatGPT surpasses human analysts in predicting the performance of companies listed on the CSI 300 index, particularly in terms of reduced forecasting errors and a more conservative forecasting stance. This suggests that ChatGPT can effectively mitigate the inherent optimistic bias of human analysts. However, the study posits that ChatGPT is not intended to completely replace human financial analysts but rather to serve as a supplementary tool that enhances the accuracy of financial forecasts (Li et al., 2024). From an operational standpoint, Skórnóg and Kmieciak (2023) compared the efficacy of ChatGPT and the automatic ARIMA algorithm for demand forecasting. The findings indicate that predictions made by ChatGPT are more reliable than those made by ARIMA, despite potential challenges in integrating the model into corporate systems (Skórnóg & Kmieciak, 2023). Stroh (2024) introduced the TrackGPT model, based on the GPT architecture, for entity trajectory prediction. Entity trajectory prediction involves forecasting the future positions or paths of entities (such as individuals, vehicles, or objects) over time, based on historical data. This task is crucial in various applications, including autonomous driving, logistics, and urban planning. This model has proven its capability to generate both long-term and short-term predictions, exhibiting excellence in accuracy, reliability, and modularity while remaining domain-agnostic and requiring minimal data features (location and time). This underscores the tremendous potential of the GPT architecture in entity trajectory prediction tasks (Stroh, 2024).

In the realm of financial market forecasting, Mumtaz and Mumtaz (2023) explored the capability of ChatGPT in leveraging tweets and sentiment analysis from social media to predict stock market trends, with a particular focus on the stocks of two major technology companies, Microsoft and Google. The study concentrated on analyzing the impact of tweet sentiments on the stock values of these companies, revealing a positive correlation between ChatGPT's assessments and the future performance of these stocks. This underscores the potential and significance of ChatGPT in financial market predictions. Furthermore, Chen et al. (2023) introduced an innovative framework that enhances graph neural networks through ChatGPT's graph reasoning capabilities,

aimed at extracting dynamic network structures from financial news. In experiments on stock trend predictions, this model demonstrated superior performance over existing state-of-the-art deep learning methods. Moreover, portfolios constructed based on this model's output exhibited higher annualized cumulative returns, while also reducing volatility and maximum drawdowns, showcasing the immense potential of ChatGPT in text-based network inference and its broad prospects in financial applications (Chen et al., 2023). Additionally, research by Lopez-Lira and Tang (2023) validated the accuracy of ChatGPT in analyzing news headlines to predict stock price movements, outperforming traditional sentiment analysis methods and other foundational models.

Collectively, these studies demonstrate that the application of ChatGPT and other LLMs in forecasting tasks not only enhances the accuracy of predictions but also showcases their unique ability to process complex data and extract deep insights. This underscores the importance of integrating LLMs into predictive models.

Despite these advancements, the application of ChatGPT for forecasting tourism demand remains largely uncharted territory. This realization highlights a notable research gap: although LLMs have proven their predictive capabilities in diverse domains, their potential in specific areas, such as tourism demand forecasting, has yet to be fully recognized and utilized. Therefore, this paper aims to address this shortfall by investigating the performance of ChatGPT in the context of tourism demand forecasting. By delving into the utility of ChatGPT in this specific domain, we aim to unlock new frontiers in predictive analytics, potentially shedding significant light on the field of tourism forecasting research.

Methodology

In this section, we present the methodology employed in our study to evaluate the potential of ChatGPT-4 for the task of tourism demand forecasting. Our methodology consists of several key components, including the selection of the datasets, the design of various prompting strategies, the implementation of the baseline models, and the selection of evaluation metrics.

Datasets

To empirically investigate the forecasting performance of ChatGPT-4, the dataset used is the monthly tourist arrivals in Macau. It can be obtained from the Statistics and Census Service (DSEC) of the Macau Government. To illustrate the generalizability of the algorithm, we collect tourist arrivals from the top eight source markets from the DSEC website. These eight source markets are Mainland China, Hong Kong, Taiwan, the Philippines, Indonesia, South Korea, the United States, and Malaysia. Fig. 1 shows the number of tourist arrivals to Macau from eight source markets and these data show similar cyclical fluctuations during COVID-19 period, with tourist arrivals plummeting

after the COVID-19 outbreak and gradually recovering after 2023.

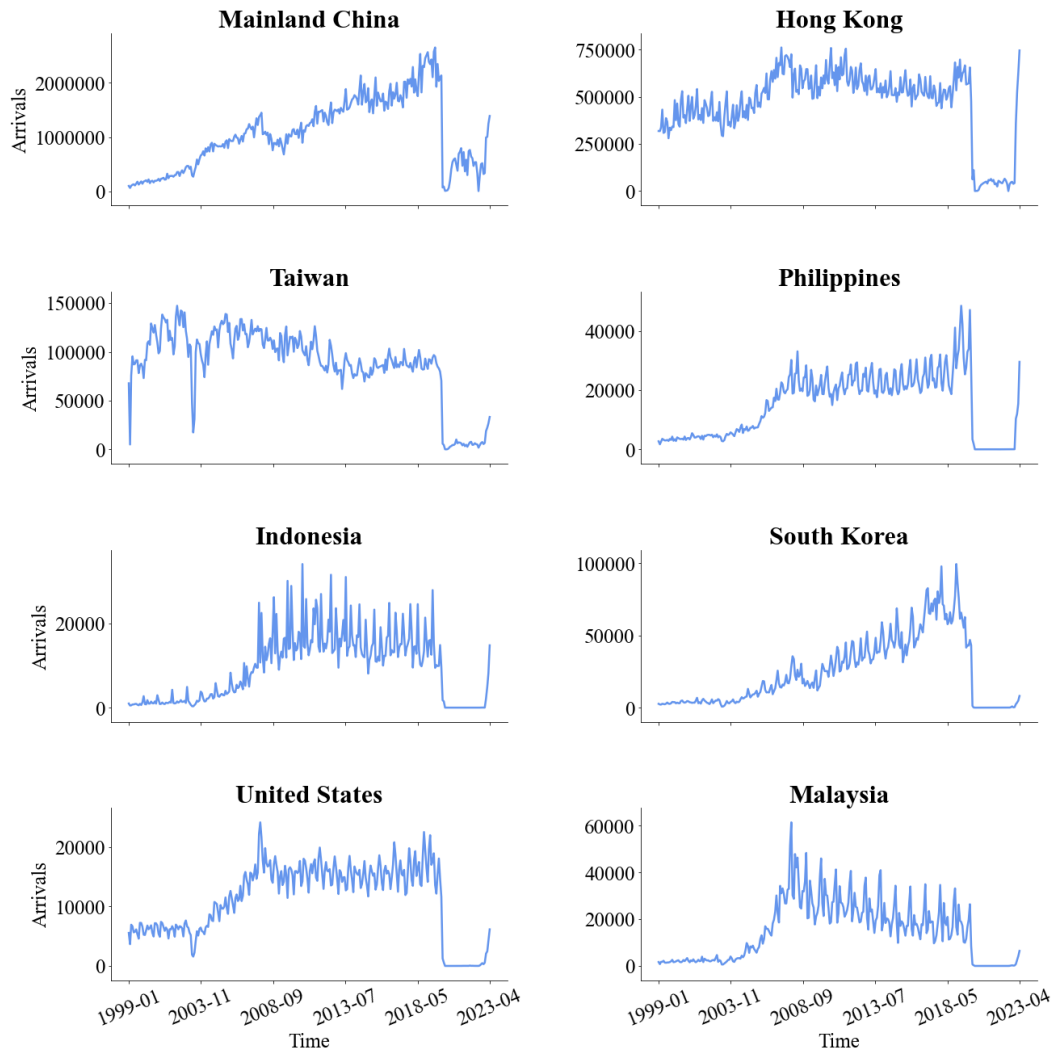


Fig. 1. Monthly Macau Tourist Arrivals (Eight source markets)

Our dataset covers from January 1999 to April 2023 (292 observations), and we define the period from January 1999 to December 2019 as the pre-COVID-19 period, January 2020 to December 2022 as the during-COVID-19 period, and January 2023 to April 2023 as the post-COVID-19 period, as shown in Fig. 2.

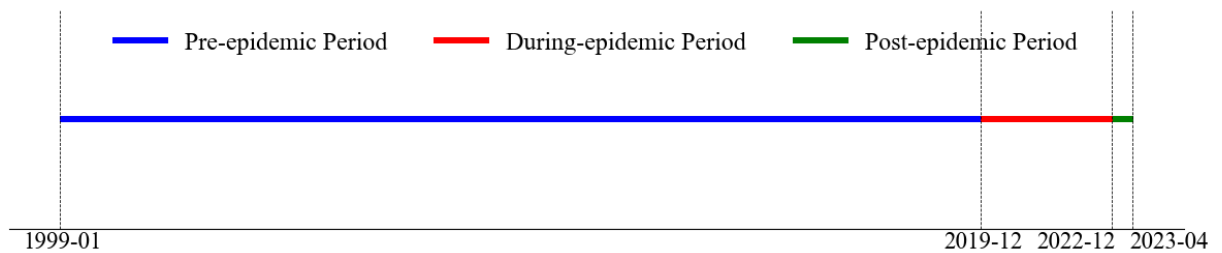


Fig. 2. Division of the period pre, during and post COVID-19

In the pre-COVID-19 period, we use data from January 1999 to December 2018 to predict data for the 12 months of 2019 and in the post-COVID-19 period, we use data from January 2020 to December 2022 to predict the first four months of 2023. We do not predict tourist arrivals during the COVID-19 period (January 2020 to December 2022) because the tourist numbers during this time were significantly influenced by government interventions. Measures such as travel restrictions, stay-at-home orders, and entry limitations had a profound impact on the flow of tourists. As a result, unlike the pre-pandemic (January 1999 to December 2019) and post-pandemic periods (January 2023 to April 2023), tourist arrivals during the COVID-19 period does not reflect typical and normal tourist patterns, are highly anomalous and not representative of usual trends. Consequently, attempting to model or predict tourist arrivals during this period would undermine the reliability and precision of the forecasts.

Prompts strategy

We experiment with various prompting strategies, including vanilla Zero-shot prompting and Chain-of-Thought (CoT) enhanced Zero-shot prompting, to investigate their effects on the performance of ChatGPT-4 in tourism demand forecasting tasks.

(1) Zero-shot prompt

In our study, we have done empirical investigations on the performance of ChatGPT-4 in tourism demand forecasting in two periods: pre-COVID-19 (from January 1999 to December 2019), and post-COVID-19 (from January 2023 to April 2023). Their results are denoted as *GPT-4*. It is important to emphasize that at the time this study was conducted, the version of ChatGPT only used the data up to September 2021. Therefore, for the post-COVID-19 period, ChatGPT has no information about the forecasting period. For the pre-COVID-19 period, by dialoguing with ChatGPT, it is shown that GPT-4 has no knowledge of the inbound tourist arrivals for Macau. Therefore, forecasting performance comparisons of ChatGPT with other models are fair.

Here we illustrate our prompting strategy by taking the number of Mainland China tourists traveling to Macau in the post-COVID-19 forecast period as an example.

The prompt for tourism demand forecasting is designed as follows:

<data>	date	Mainland China
	Jan-20	2132032
	Feb-20	72307
	Mar-20	88890
	...	...

The <data> above shows the <number> of tourist arrivals from Mainland China to Macau. Assume you are an expert in tourism demand forecasting. Please consider the trend of the <data> above and forecast the <number> of tourist arrivals for the next

12 months of the given <data>. Your answer only needs to give 12 < numbers>, without any explanation.

(2) Chain-of-Thought Enhanced Zero-shot Prompt

Building on the Zero-shot prompt, we further encourage ChatGPT-4 to explicitly consider factors that may influence tourism demand through the Chain-of-Thought process. These results are denoted as *GPT-4-CoT-a*:

Based on the above results, apart from considering the trend of the <data>, you are requested to analyze all the factors that may affect the tourism demand from Mainland China to Macau during the study period to adjust your predicted results. Your answer starts with your adjusted results of 12 <numbers> and then explains them step by step.

In the post-COVID-19 period, we again press the ChatGPT-4 through a continuous dialog to give its predictions by focusing on the impact of Covid-19, in addition to considering the factors that may affect tourism demand. These results are denoted as *GPT-4- CoT-b*:

Based on the above results, apart from considering the trend of the <data> and all the factors affecting tourism demand, you are requested to further consider the impact of COVID-19 and the changes in Macau's immigration policy due to the pandemic to adjust your predicted results. Your answer starts with your adjusted results of 12 <numbers> and then explains them step by step.

This modified prompt aims to guide ChatGPT-4 in giving comprehensive explanations for its predictions by outlining a clear chain of thought. The phrase “Your answer starts with your adjusted results of 12 <numbers> and then explains them step by step.” encourages models to provide not only a numerical prediction, but also a detailed explanation and step-by-step reasoning process.

Baseline models

To investigate the performance of ChatGPT-4 in tourism demand forecasting, we use three-time series models (Naïve, Exponential Smoothing, SARIMA) and three machine learning models (Multi-Layer Perception (MLP), Long short-term Memory Network (LSTM), Random Forest) as benchmark models.

Naïve: The Naïve model in time series analysis, predicated on the assumption that future values will mirror the most recent observation, serves as a baseline method for forecasting across various applications (Hyndman & Athanasopoulos, 2018). Its simplicity and ease of implementation make it particularly useful in scenarios like short-term economic forecasting and inventory management, where complex models may not offer significant advantages (Taylor & Letham, 2018).

Exponential Smoothing (ES): The Exponential Smoothing (ES) model is a widely used approach in time series forecasting, leveraging weighted averages of past observations with exponentially decreasing weights to predict future values (Hyndman & Athanasopoulos, 2018). Its applications span various domains, including demand forecasting in supply chain management and economic trend analysis (Taylor & Letham, 2018).

SARIMA: The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an extension of the ARIMA model, incorporating seasonality into its structure, making it particularly effective for analyzing and forecasting seasonal time series data. Its applications span various fields, including economics for market analysis, meteorology for weather forecasting, and the energy sector for load forecasting.

Multi-Layer Perception (MLP): The Multi-Layer Perceptron (MLP) model, a fundamental architecture in neural networks, is characterized by its layered structure of interconnected nodes, facilitating complex pattern recognition and decision-making processes (Goodfellow et al., 2016). Its applications span various domains, including image recognition, speech processing, and financial forecasting, demonstrating its versatility and effectiveness in handling high-dimensional data (LeCun et al., 2015; Zhang et al., 2021).

Long Short-term Memory Network (LSTM): The Long Short-Term Memory (LSTM) network, a type of recurrent neural network, is designed to overcome the limitations of traditional RNNs in learning long-range dependencies, as highlighted by Hochreiter and Schmidhuber (1997). Its applications span various domains, including natural language processing, speech recognition, and time-series prediction, with recent advancements and applications detailed in works like Gers et al. (2000) and Sundermeyer et al. (2012).

Random Forest (RF): The Random Forest (RF) model, a versatile and robust ensemble learning method, excels in classification and regression tasks by constructing multiple decision trees and aggregating their predictions (Breiman, 2001). Its applications span various domains, including bioinformatics for gene selection and classification (Díaz-Uriarte & Alvarez de Andrés, 2006), and financial risk assessment (Khaidem et al., 2016).

For consistency and to ensure a fair comparison across all models, we used the same input data for all forecasting methods employed in this study. The sole input for all these models was the historical tourist arrival series. This approach allows for a direct comparison of the models' performance based on their ability to extract patterns and make predictions from the same univariate time series data. It's worth noting that while the ChatGPT models with prompts had the potential to incorporate additional contextual information through the prompting process, the underlying numerical data used for prediction remained consistent across all models. This design choice ensures that any performance differences observed are primarily due to the models' inherent capabilities rather than disparities in input data.

Performance evaluations

Forecast accuracy can be measured through various metrics, which fall into three main categories: (1) Scale-dependent errors. These errors are in the same units as the data, making measures of accuracy based on these errors scale-dependent and unsuitable for comparing series with different units. Common metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). (2) Percentage errors. The advantage of percentage errors is their unit independence, making them frequently used for comparing forecasting performance across datasets. The most commonly used metric is the Mean Absolute Percentage Error (MAPE). (3) Scaled errors. Proposed by Hyndman & Koehler (2006), scaled errors serve as an alternative to percentage errors when comparing forecast accuracy across series with different units. They suggest scaling errors based on the MAE of a simple forecasting method on the training data, with the Mean Absolute Scaled Error (MASE) being a prevalent metric. This study selects one metric from each of these three categories to measure forecast accuracy: MASE from scaled errors, MAPE from percentage errors, and RMSE from scale-dependent errors, defined as follows:

$$\text{MASE} = \frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{\frac{1}{T-1} \sum_{t=2}^T |A_t - A_{t-1}|} \quad (1)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{A_i} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_i - F_i)^2} \quad (3)$$

where A_i is the actual arrival value at time i , F_i is the forecast value at time i , n is the number of forecasted points, T is the total number of time periods in the data.

MASE scales the error based on the one-step-ahead in-sample mean absolute error (MAE) generated by the Naïve forecasting method. The denominator $\frac{1}{T-1} \sum_{t=2}^T |A_t - A_{t-1}|$

represents the mean absolute difference of the actual values, which is used for scaling the error. MASE, as a scaled error metric, is regarded as a standard method for comparing forecasting accuracy across series with different scales (Hyndman & Koehler, 2006). It has meaningful interpretability, a broad range of applicability, and is not affected by degeneracy issues. Among these metrics, we favor MASE due to its lower sensitivity to outliers, its ability to avoid problems associated with zero or near-zero values, and, most importantly, its capability to facilitate accuracy comparisons across series of different scales.

MAPE measures the average magnitude of errors as a percentage of the actual values across a set of forecasts. It is scale-independent, making it particularly suitable for comparing the accuracy of models across different data scales. However, MAPE may

encounter issues when dealing with actual values that are zero or near-zero, as it involves division by the actual value (Hyndman & Koehler, 2006).

RMSE measures the square root of the average squared differences between forecasts and actual value. It places relatively more weight on larger errors, thus helping to identify significant deviations between forecasts and actual values. RMSE is sensitive to outliers and does not provide information on the direction of errors (Chai & Draxler, 2014).

Empirical results

The study examines the predictive accuracy of ChatGPT-4 on monthly tourist arrivals in Macau, a key city in China’s Greater Bay Area with gaming and tourism as its main economic sectors. Accurate forecasting of tourist arrivals is essential for Macau’s economic development. This study employs ChatGPT’s Zero-shot and CoT prompting strategies to forecast tourist arrivals and compare their performance in forecasting tourism demand with six models: three time series models (Naïve, Exponential Smoothing, and SARIMA) and three machine learning models (MLP, LSTM, and Random Forest).

Forecasting performance before the COVID-19

Tables 1, 2, and 3 present the one-step-ahead and two-step-ahead tourist arrivals forecasting accuracy of eight models from Macau’s eight source markets, as measured by MASE, MAPE, and RMSE, before the COVID-19 pandemic. Figure 3 illustrates the average forecasting error for these models across the eight markets. The first column of each table lists the model names, with parentheses indicating the model’s ranking for a specific market under a given error metric. The first row’s “Average” denotes the average prediction error for a model across the eight source markets for a specified metric, while “Market-rank” signifies the average ranking of the model’s prediction error across these source markets. For instance, “6.88(8)” in Table 1 indicates that the Naïve model has an average ranking of 6.88 among the eight models for prediction performance across the eight source markets, and this performance is ranked 8th among the eight models.

Given that MASE is regarded as the standard method for comparing forecasting accuracy across time series of different scales (Hyndman & Koehler, 2006), we primarily focus on the performance under the MASE metric, with MAPE and RMSE serving as supplementary measures. From an overall perspective on the average prediction error, GPT-4 demonstrates robust forecasting performance compared to the baseline models. Specifically, under the zero-shot prompting strategy, ChatGPT-4 exhibits moderate error levels in both one-step-ahead and two-step-ahead forecasts (ranging between 4th/8 and 5th/8). Under the CoT prompting strategy, GPT-4-CoT-a

ranks third out of the eight models for both one-step-ahead and two-step-ahead forecasts under the MASE and MAPE metrics. In terms of the average ranking of prediction errors across the eight markets, GPT-4's performance remains robust relative to the baseline models. Whether using MASE, MAPE, or RMSE, ChatGPT-4 ranks fourth among the eight models for both one-step-ahead and two-step-ahead forecasts under the zero-shot prompting strategy, while under the CoT prompting strategy, it ranks third. Among the alternative forecasting models, SARIMA consistently outperforms the others across all three error metrics. This superior performance is evident regardless of whether the evaluation is based on average error or market ranking criteria.

The findings suggest that the continuous conversation approach enhances the forecast accuracy of ChatGPT-4, highlighting its potential in improving tourism forecasting. Regardless of the prediction horizon or the evaluation metric used, the predictions generated by GPT-4 using the CoT prompting strategy consistently outperform those generated using the zero-shot strategy (except for RMSE in the one-step-ahead forecast). In terms of MASE, for one-step-ahead forecasting, GPT-4 ranks 5th/8, while GPT-4-CoT-a ranks 3rd/8; for two-step-ahead forecasting, GPT-4 and GPT-4-CoT-a rank 4th/8 and 3rd/8, respectively. When considering the average ranking of prediction errors across the eight source markets, GPT-4 and GPT-4-CoT-a rank 4th/8 and 3rd/8 for both one-step-ahead and two-step-ahead forecasts, respectively.

From the perspective of individual source markets, GPT-4 excels in forecasting tourism demand for some specific markets. For instance, in the one-step-ahead forecasts, GPT-4 performs well in predicting tourist arrivals from Hong Kong to Macau under the MASE, MAPE, and RMSE metrics, ranking 2nd/8 across all three metrics. GPT-4-CoT-a ranks first across all three metrics for predicting tourist arrivals from Taiwan to Macau. In the two-step-ahead forecasts, GPT-4 and GPT-4-CoT-a rank within the top two across all three metrics for predicting tourist arrivals from Indonesia to Macau.

The prediction results for the eight source markets indicate that, whether using the average error or market-ranking metrics, ChatGPT-4's prediction errors under the zero-shot strategy generally remain at the median level, while its errors under the CoT strategy consistently rank in the upper range. This underscores ChatGPT's reliability in tourism forecasting. Additionally, in the vast majority of cases, the predictions under ChatGPT-4's CoT strategy significantly outperform those under the zero-shot strategy, demonstrating that the continuous conversation approach indeed enhances ChatGPT's forecasting capabilities. The continuous dialogue prompting strategy does not provide ChatGPT with additional data, but rather stimulates better results through the prompt strategy.

Table 1. Forecasting comparison using MASE for the pre-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippines	Indonesia	South Korea	United States	Malaysia	Average	Market rank
One-step-ahead										

Naïve	0.94(7)	1.12(8)	1.04(8)	1.11(5)	1.01(8)	1.15(3)	1.00(8)	1.23(8)	1.07(7)	6.88(8)
ES	0.81(3)	0.82(5)	1.00(7)	1.16(6)	0.83(4)	1.16(4)	0.83(7)	1.21(7)	0.98(6)	5.38(6)
SARIMA	0.54(2)	0.47(1)	0.82(5)	0.68(1)	0.59(2)	0.74(1)	0.39(1)	0.40(1)	0.58(1)	1.75(1)
MLP	0.52(1)	0.6(3)	0.73(3)	0.77(2)	0.54(1)	1.43(5)	0.40(2)	0.61(5)	0.70(2)	2.75(2)
LSTM	0.84(4)	0.88(7)	1.00(6)	1.29(8)	0.86(5)	3.60(8)	0.68(3)	0.82(6)	1.24(8)	5.88(7)
RF	0.98(8)	0.87(6)	0.68(2)	1.25(7)	0.93(7)	0.91(2)	0.82(6)	0.41(2)	0.86(4)	5.00(5)
GPT-4	0.92(6)	0.55(2)	0.77(4)	0.95(3)	0.87(6)	1.69(7)	0.80(5)	0.46(3)	0.88(5)	4.50(4)
GPT-4- CoT- a	0.92(5)	0.63(4)	0.60(1)	0.99(4)	0.83(3)	1.48(6)	0.78(4)	0.46(4)	0.84(3)	3.88(3)
Two-step-ahead										
Naïve	0.92(8)	0.97(6)	1.61(8)	1.57(6)	1.17(8)	1.46(3)	1.31(8)	1.73(8)	1.34(7)	6.88(8)
ES	0.78(4)	0.95(5)	1.56(7)	1.7(7)	0.87(5)	1.57(4)	0.86(7)	1.27(7)	1.20(6)	5.75(7)
SARIMA	0.51(1)	0.6(2)	0.93(4)	0.95(2)	0.64(3)	1.25(1)	0.59(2)	0.45(1)	0.74(1)	2.00(1)
MLP	0.60(2)	0.49(1)	0.81(2)	0.74(1)	0.65(4)	1.80(5)	0.59(1)	0.64(5)	0.79(2)	2.63(2)
LSTM	0.83(6)	0.95(4)	1.16(6)	1.41(3)	0.98(7)	4.06(8)	0.77(5)	0.76(6)	1.36(8)	5.63(6)
RF	0.86(7)	1.05(8)	0.78(1)	1.80(8)	0.93(6)	1.28(2)	0.82(6)	0.49(2)	1.00(5)	5.00(5)
GPT-4	0.75(3)	0.99(7)	0.91(3)	1.41(4)	0.48(2)	1.95(7)	0.73(4)	0.53(3)	0.97(4)	4.13(4)
GPT-4-CoT-a	0.80(5)	0.86(3)	1.04(5)	1.42(5)	0.47(1)	1.83(6)	0.66(3)	0.54(4)	0.95(3)	4.00(3)

Table 2. Forecasting comparison using MAPE for the pre-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippines	Indonesia	South Korea	United States	Malaysia	Average	Market rank
One-step-ahead										
Naïve	10.63(7)	11.21(8)	4.77(8)	19.31(5)	30.05(6)	14.56(3)	18.66(8)	31.46(7)	17.58(7)	6.50(8)
ES	9.09(3)	7.91(5)	4.58(7)	19.47(6)	29.14(5)	14.68(4)	16.06(7)	35.36(8)	17.04(6)	5.63(6)
SARIMA	6.16(2)	4.53(1)	3.72(5)	11.88(1)	20.22(1)	10.95(1)	7.49(1)	10.42(2)	9.42(1)	1.75(1)
MLP	5.86(1)	5.96(3)	3.33(3)	14.06(2)	20.70(2)	20.54(5)	7.71(2)	17.46(5)	11.95(2)	2.88(2)
LSTM	9.45(4)	8.67(7)	4.44(6)	20.99(7)	33.78(8)	52.17(8)	12.43(3)	25.93(6)	20.98(8)	6.13(7)
RF	10.87(8)	8.51(6)	3.15(2)	21.44(8)	31.37(7)	13.67(2)	14.71(5)	9.84(1)	14.19(4)	4.88(5)
GPT-4	10.16(6)	5.33(2)	3.48(4)	16.08(3)	27.80(4)	23.71(7)	15.07(6)	13.80(4)	14.43(5)	4.50(4)
GPT-4- CoT- a	10.08(5)	6.05(4)	2.75(1)	16.42(4)	25.75(3)	21.08(6)	14.48(4)	13.38(3)	13.75(3)	3.75(3)
Two-step-ahead										
Naïve	11.05(8)	8.59(6)	7.39(8)	28.42(6)	43.19(8)	18.98(3)	25.05(8)	50.47(8)	24.14(8)	6.88(8)
ES	9.39(4)	8.20(4)	7.15(7)	30.18(7)	32.07(5)	20.39(4)	17.33(7)	40.02(7)	20.59(6)	5.63(6)
SARIMA	6.41(1)	5.17(2)	4.24(3)	17.22(2)	22.23(3)	17.82(1)	11.78(2)	13.56(2)	12.30(1)	2.00(1)
MLP	7.38(2)	4.34(1)	3.73(2)	13.09(1)	25.11(4)	26.69(5)	11.26(1)	20.11(5)	13.92(2)	2.63(2)
LSTM	10.16(6)	8.34(5)	5.12(6)	22.95(3)	37.65(7)	59.53(8)	13.85(5)	24.05(6)	22.71(7)	5.75(7)
RF	10.22(7)	9.06(8)	3.62(1)	32.3(8)	32.99(6)	18.63(2)	15.09(6)	12.33(1)	16.78(5)	4.88(5)
GPT-4	8.95(3)	8.62(7)	4.27(4)	25.34(5)	18.27(2)	28.67(7)	13.48(4)	16.25(3)	15.48(4)	4.38(4)
GPT-4-CoT-a	9.43(5)	7.48(3)	4.87(5)	25.13(4)	16.96(1)	26.77(6)	12.10(3)	16.69(4)	14.93(3)	3.88(3)

Table 3. Forecasting comparison using RMSE for the pre-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippine s	Indones ia	South Korea	United States	Malaysi a	Average	Market rank
One-step-ahead										
Naïve	312893(8)	83224(8)	4929(8)	7804(5)	6620(8)	11468(3)	3384(8)	7245(8)	54696(8)	7.00(8)
ES	248013(3)	57731(5)	4737(6)	8425(6)	5124(4)	12212(4)	2966(7)	6071(7)	43160(3)	5.25(6)
SARIMA	159469(1)	37345(1)	4158(5)	4832(1)	3139(1)	7099(1)	1357(1)	2147(1)	27443(1)	1.50(1)
MLP	164058(2)	48368(3)	4140(4)	5754(2)	3447(2)	12996(5)	1503(2)	3238(5)	30438(2)	3.13(2)
LSTM	252278(4)	64498(6)	4779(7)	9635(8)	5325(5)	31191(8)	2480(3)	4558(6)	46843(5)	5.88(7)
RF	298820(7)	66791(7)	3981(3)	8822(7)	5036(3)	8731(2)	2897(6)	2968(4)	49756(7)	4.88(5)
GPT-4	287135(5)	45321(2)	3364(2)	7146(3)	5790(7)	19901(7)	2887(5)	2599(3)	46768(4)	4.25(4)
GPT-4-CoT-a	294037(6)	48907(4)	3154(1)	7293(4)	5635(6)	18494(6)	2688(4)	2593(2)	47850(6)	4.13(3)
Two-step-ahead										
Naïve	306581(8)	59737(3)	7161(8)	11117(6)	7707(8)	13566(3)	4590(8)	9117(8)	52447(8)	6.50(8)
ES	247781(3)	61094(4)	6879(7)	11689(7)	5449(6)	13787(4)	3285(7)	6833(7)	44600(3)	5.63(6)
SARIMA	176607(1)	38899(2)	4236(3)	6558(2)	3394(3)	10175(1)	2063(1)	2422(1)	30544(1)	1.75(1)
MLP	210946(2)	32407(1)	4230(1)	6026(1)	3965(4)	15831(5)	2167(2)	3905(5)	34935(2)	2.63(2)
LSTM	271154(6)	61667(5)	5653(6)	10835(5)	5495(7)	35188(8)	2890(5)	4344(6)	49653(7)	6.00(7)
RF	271256(7)	65191(6)	4232(2)	11827(8)	5088(5)	10793(2)	3009(6)	3070(4)	46808(6)	5.00(5)
GPT-4	258025(5)	68721(8)	5137(4)	10487(4)	2897(1)	17532(7)	2802(4)	3016(2)	46077(5)	4.38(4)
GPT-4-CoT-a	257472(4)	65472(7)	5549(5)	10466(3)	2989(2)	16619(6)	2550(3)	3032(3)	45519(4)	4.13(3)

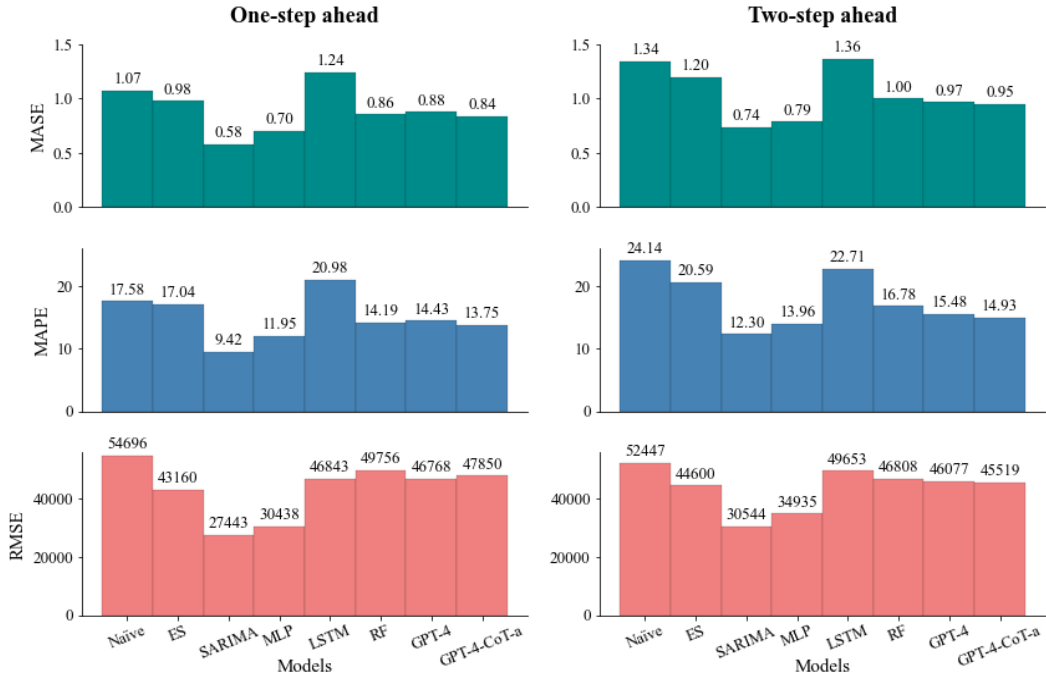


Fig. 3. Forecasting errors for the pre-COVID-19 period

The aforementioned discussion compares the prediction errors across eight individual models. To more comprehensively assess the potential of ChatGPT-4 in tourism forecasting, a group comparison is also conducted. Table 4 presents the results of the group comparison across three error metrics. In this table, the first column represents the abbreviated group names. “TS” refers to the average prediction error across the eight markets for the three time series models (i.e., Naïve, ES, and SARIMA). “ML” represents the average prediction error for the three machine learning models (MLP, LSTM, and RF), and “Non-GPT” indicates the average prediction error for all non-GPT models (three time series models and three machine learning models).

From Table 4, we can observe that under the MASE metric, whether for one-step-ahead or two-step-ahead predictions, GPT-4-CoT-a consistently outperforms the other three groups. This indicates that GPT-4’s forecasting performance surpasses the combined results of all non-GPT methods, highlighting the robustness of ChatGPT in tourism forecasting.

Table 4. Forecasting comparison of model groups for the pre-COVID-19 period

By Group	MASE	MAPE	RMSE
One-step-ahead			
TS	0.88(2)	14.68(2)	41766(1)
ML	0.93(4)	15.71(4)	42346(3)
Non-GPT	0.91(3)	15.19(3)	42056(2)
GPT-4-CoT-a	0.84(1)	13.75(1)	47850(4)
Two-step-ahead			
TS	1.09(4)	19.01(4)	42530(1)
ML	1.05(2)	17.82(2)	43799(3)
Non-GPT	1.07(3)	18.41(3)	43165(2)
GPT-4-CoT-a	0.95(1)	14.93(1)	45519(4)

Forecasting performance after the COVID-19

Tables 5, 6, and 7 present the MASE, MAPE, and RMSE metrics for one-step-ahead and two-step-ahead forecasts of tourism arrivals in the post-COVID-19 period. Additionally, Figure 4 illustrates the average prediction errors for the nine forecasting models in both one-step-ahead and two-step-ahead forecasts in the post pandemic period. As before, we focus primarily on the MASE metric, with MAPE and RMSE serving as supplementary measures. The results for post-COVID-19 forecasts further confirm ChatGPT’s potential in tourism demand forecasting. According to the overall average MASE, GPT-4 ranks 2nd for one-step-ahead forecasts and GPT-4-CoT-b ranks the best-performing model. For two-step-ahead forecasts, GPT-4 ranks 4th under MASE, and GPT-4-CoT-b ranks 3rd under MASE.

In terms of the average ranking of prediction errors across eight source markets, GPT-4 also demonstrates strong performance compared to baseline models in the post-COVID-19 period. Under the zero-shot prompting strategy, GPT-4 ranks 2nd among nine models for one-step-ahead forecasts and 4th for two-step-ahead forecasts according to MASE. Under the CoT prompting strategy, GPT-4-CoT-b ranks 1st for one-step-ahead forecasts and 2nd for two-step-ahead forecasts. With regard to traditional models, time series approaches consistently outperform machine learning models across all three error metrics, whether evaluated by average error or market ranking. Notably, in two-step-ahead forecasting, the simplest time series model—the Naïve model—exhibits the best predictive performance, underscoring the enduring value of time series models.

Similarly, under the CoT prompting strategy, GPT-4 consistently outperforms the zero-shot prompting strategy in terms of forecasting accuracy. This again demonstrates that appropriate prompting strategies, such as the continuous conversation method employed in this study, can enhance the predictive performance of GPT-4. It should be acknowledged that the differing lengths of available time series data for each period may have impacted our results. The pre-COVID-19 period benefits from a longer time

series (January 1999 to December 2019), which generally favors traditional time series models. In contrast, the time series available in the post-pandemic period is shorter. This limitation in data availability in the post-pandemic period may partly explain the relatively better performance of AI-based models, including ChatGPT. Traditional time series models typically require longer sequences of data to effectively capture patterns, while AI models can extract insights from shorter sequences. This difference in data availability across periods should be taken into account when interpreting our results and comparing model performance across different time frames. Nevertheless, across both periods, ChatGPT's performance remains consistently above the average, underscoring its robustness, which is a highly desirable feature.

Table 5. Forecasting comparison using MASE for the post-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippines	Indonesia	South Korea	United States	Malaysia	Average	Market rank
One-step-ahead										
Naïve	1.99(3)	1.36(2)	1.38(4)	1.15(5)	0.88(6)	1.02(5)	1.03(6)	0.83(6)	1.20(5)	4.63(5)
ES	2.68(7)	1.63(5)	1.46(5)	1.15(6)	0.88(7)	1.02(6)	1.04(7)	0.83(7)	1.34(6)	6.25(6)
SARIMA	2.35(5)	1.83(8)	1.52(6)	1.01(3)	0.69(1)	0.82(3)	0.69(2)	0.61(4)	1.19(4)	4.00(3)
MLP	3.07(8)	1.03(1)	1.66(7)	1.55(8)	0.85(5)	2.00(8)	1.16(8)	0.70(5)	1.50(7)	6.25(6)
LSTM	3.13(9)	1.67(7)	1.94(8)	1.69(9)	1.10(9)	1.13(7)	0.61(1)	1.17(8)	1.55(8)	7.25(8)
RF	2.48(6)	1.90(9)	1.95(9)	1.24(7)	0.99(8)	2.15(9)	1.29(9)	1.17(9)	1.65(9)	8.25(9)
GPT-4	1.91(1)	1.49(4)	1.25(2)	0.98(2)	0.76(3)	0.74(2)	0.84(4)	0.45(1)	1.05(2)	2.38(2)
a GPT-4- CoT-	2.16(4)	1.66(6)	1.33(3)	1.06(4)	0.81(4)	0.84(4)	0.88(5)	0.52(3)	1.16(3)	4.13(4)
b GPT-4- CoT-	1.91(2)	1.46(3)	1.13(1)	0.95(1)	0.72(2)	0.73(1)	0.76(3)	0.46(2)	1.02(1)	1.88(1)
Two-step-ahead										
Naïve	0.69(1)	1.17(1)	0.88(1)	0.73(1)	0.86(1)	0.83(1)	0.74(1)	0.97(1)	0.86(1)	1.00(1)
ES	2.81(7)	3.33(4)	2.20(6)	1.32(7)	1.41(3)	1.49(6)	1.40(7)	1.65(7)	1.95(5)	5.88(7)
SARIMA	2.70(6)	4.50(9)	2.37(7)	1.26(5)	1.38(2)	1.46(4)	1.34(6)	1.64(6)	2.08(7)	5.63(6)
MLP	2.10(2)	1.78(2)	1.74(2)	0.97(2)	1.60(8)	4.83(9)	0.95(2)	1.07(2)	1.88(2)	3.63(3)
LSTM	3.34(9)	3.16(3)	2.76(8)	4.52(9)	3.91(9)	2.59(8)	1.48(8)	6.79(9)	3.57(9)	7.88(9)
RF	3.11(8)	3.69(5)	3.20(9)	1.55(8)	1.52(7)	1.74(7)	1.67(9)	1.81(8)	2.29(8)	7.63(8)
GPT-4	2.34(4)	4.43(7)	2.04(4)	1.26(4)	1.48(6)	1.37(3)	1.27(4)	1.40(4)	1.95(4)	4.50(4)
a GPT-4- CoT-	2.54(5)	4.45(8)	2.09(5)	1.31(6)	1.48(5)	1.46(5)	1.32(5)	1.44(5)	2.01(6)	5.50(5)
b GPT-4- CoT-	2.31(3)	4.39(6)	1.88(3)	1.23(3)	1.42(4)	1.32(2)	1.22(3)	1.35(3)	1.89(3)	3.38(2)

Table 6. Forecasting comparison using MAPE for the post-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippine s	Indonesia	South Korea	United States	Malaysia	Average	Market rank
One-step-ahead										
Naïve	24.22(3)	38.05(2)	29.07(4)	45.94(6)	59.91(5)	46.61(5)	40.68(6)	55.16(6)	42.46(5)	4.63(5)
ES	32.69(7)	44.77(6)	30.71(5)	45.97(7)	59.99(6)	46.69(6)	40.81(7)	55.20(7)	44.61(6)	6.38(7)
SARIMA	28.32(6)	45.86(7)	31.88(6)	42.27(3)	53.51(2)	41.18(4)	31.76(2)	48.05(5)	40.35(3)	4.38(3)
MLP	36.18(8)	30.31(1)	34.03(7)	56.62(8)	57.32(3)	93.04(8)	52.41(9)	38.87(1)	49.85(7)	5.63(6)
LSTM	37.69(9)	47.54(9)	42.31(9)	81.43(9)	78.85(9)	57.92(7)	29.01(1)	69.92(8)	55.59(8)	7.63(9)
RF	27.44(5)	46.22(8)	36.92(8)	44.42(5)	49.16(1)	124.86(9)	46.96(8)	90.21(9)	58.27(9)	6.63(8)
GPT-4	23.26(2)	41.09(4)	26.53(2)	41.52(2)	60.23(7)	37.69(2)	34.80(4)	40.37(2)	38.19(2)	3.13(2)
GPT-4-CoT-a	26.04(4)	44.40(5)	27.93(3)	44.17(4)	61.81(8)	40.98(3)	36.44(5)	43.36(4)	40.64(4)	4.50(4)
GPT-4-CoT-b	23.20(1)	40.38(3)	24.17(1)	40.56(1)	58.11(4)	37.17(1)	32.38(3)	40.59(3)	37.07(1)	2.13(1)
Two-step-ahead										
Naïve	10.28(1)	21.4(1)	17.22(1)	28.03(1)	46.75(1)	33.47(1)	29.61(1)	47.98(1)	29.34(1)	1.00(1)
ES	46.29(7)	62.26(4)	45.85(6)	64.04(6)	79.85(3)	66.25(6)	62.84(8)	78.29(7)	63.21(4)	5.88(6)
SARIMA	45.01(6)	81.24(9)	49.37(7)	60.02(3)	76.99(2)	63.95(4)	57.98(5)	77.56(6)	64.02(6)	5.25(5)
MLP	33.76(2)	36.60(2)	37.91(2)	51.98(2)	86.59(8)	230.89(9)	36.01(2)	52.35(2)	70.76(7)	3.63(3)
LSTM	52.97(9)	61.26(3)	58.92(8)	257.18(9)	164.22(9)	127.71(8)	62.12(7)	229.64(9)	126.75(9)	7.75(9)
RF	49.36(8)	67.40(5)	65.49(9)	74.04(8)	83.63(5)	76.09(7)	74.97(9)	84.28(8)	71.91(8)	7.38(8)
GPT-4	39.09(4)	77.38(7)	43.41(4)	62.11(5)	84.52(7)	62.47(3)	57.87(4)	70.48(4)	62.17(3)	4.75(4)
GPT-4- CoT-a	42.19(5)	77.67(8)	44.56(5)	64.31(7)	84.50(6)	65.40(5)	59.63(6)	72.03(5)	63.79(5)	5.88(6)
GPT-4- CoT-b	38.68(3)	76.59(6)	40.61(3)	60.79(4)	82.50(4)	61.23(2)	55.87(3)	68.85(3)	60.64(2)	3.50(2)

Table 7. Forecasting comparison using RMSE for the post-COVID-19 period

	Mainland China	Hong Kong	Taiwan	Philippine s	Indones ia	South Korea	United States	Malaysi a	Average	Market rank
One-step-ahead										
Naïve	356006(3)	194454(2)	7471(4)	8961(5)	4089(6)	2108(5)	1527(6)	1680(5)	72037(2)	4.50(5)
ES	395754(7)	233164(6)	7682(5)	8968(6)	4097(7)	2109(6)	1531(7)	1683(6)	81874(6)	6.25(6)
SARIMA	390169(6)	246575(7)	8075(6)	7581(3)	2934(1)	1610(3)	1129(2)	1273(4)	82418(8)	4.00(3)
MLP	439904(9)	179036(1)	9447(7)	12326(9)	3983(5)	5027(9)	1825(8)	1963(7)	81689(5)	6.88(7)
LSTM	434798(8)	281358(9)	10993(9)	12192(8)	5384(9)	2356(7)	937(1)	2800(9)	93852(9)	7.50(8)
RF	374433(5)	247909(8)	10024(8)	9445(7)	5044(8)	4813(8)	1906(9)	2309(8)	81985(7)	7.63(9)
GPT-4	345198(2)	214733(4)	7099(2)	7544(2)	3294(3)	1502(1)	1229(4)	1025(1)	72703(3)	2.38(2)
GPT-4-CoT-a	367425(4)	227892(5)	7451(3)	7982(4)	3511(4)	1740(4)	1283(5)	1126(3)	77301(4)	4.00(3)
GPT-4-CoT-b	341586(1)	210295(3)	6882(1)	7348(1)	3156(2)	1543(2)	1143(3)	1052(2)	71626(1)	1.88(1)
Two-step-ahead										
Naïve	161114(1)	132402(1)	4990(1)	8472(1)	4555(1)	2132(1)	1524(1)	1909(1)	39637(1)	1.00(1)
ES	546105(7)	384916(3)	12266(5)	12633(7)	7225(4)	3530(6)	2695(7)	3273(7)	121580(3)	5.75(7)
SARIMA	544253(6)	502424(6)	13286(7)	12189(5)	7129(2)	3459(5)	2621(6)	3261(6)	136078(8)	5.38(5)
MLP	446361(2)	282161(2)	10632(2)	9270(2)	8474(8)	15740(9)	2068(2)	2132(2)	97105(2)	3.63(3)
LSTM	642180(9)	427983(5)	16409(8)	58272(9)	26890(9)	7908(8)	3183(9)	20650(9)	150434(9)	8.25(9)
RF	597374(8)	412398(4)	17510(9)	14370(8)	7923(7)	4088(7)	3177(8)	3632(8)	132559(6)	7.38(8)
GPT-4	466515(4)	531212(8)	12193(4)	11983(4)	7503(6)	3208(3)	2403(4)	2758(4)	129722(5)	4.63(4)
GPT-4-CoT-a	500507(5)	532796(9)	12363(6)	12356(6)	7477(5)	3446(4)	2521(5)	2841(5)	134288(7)	5.63(6)
GPT-4-CoT-b	461197(3)	528645(7)	11656(3)	11826(3)	7152(3)	3089(2)	2325(3)	2672(3)	128570(4)	3.38(2)

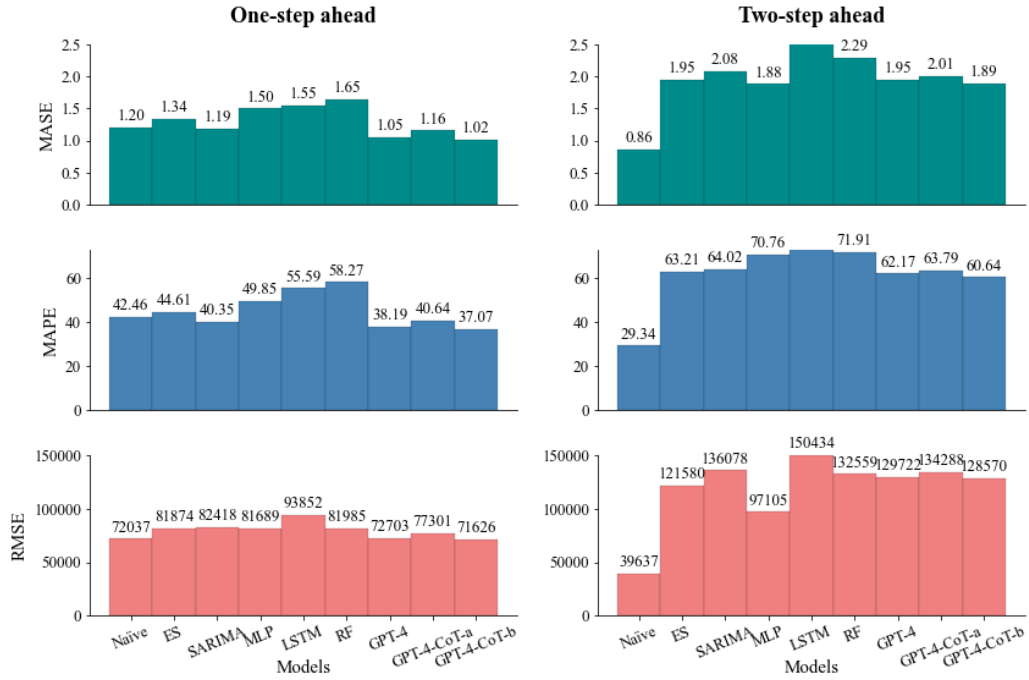


Fig. 4. Forecasting errors for the post-COVID-19 period

A group comparison for the post-COVID-19 period is also conducted. Table 8 shows the results of the group comparison across the three metrics. The naming and meaning of the models are consistent with Table 4. From Table 8, it is observed that under the MASE metric, GPT-4-CoT-b outperforms the other three groups in one-step-ahead forecasting. In two-step-ahead forecasting, GPT-4-CoT-b performs better than the average accuracy of all non-GPT methods, and slightly behind the average accuracy of the three time series models. It is also noticed that the value of time series models should not be overlooked, especially given that in two-step-ahead forecasting, GPT-4-CoT-b falls short of the average performance of the three time series models.

Table 8. Forecasting comparison of model groups for the post-COVID-19 period

By Group	MASE	MAPE	RMSE
One-step-ahead			
TS	1.24(2)	42.47(2)	78776(2)
ML	1.57(4)	54.57(4)	85842(4)
Non-GPT	1.41(3)	48.52(3)	82309(3)
GPT-4-CoT-b	1.02(1)	37.07(1)	71626(1)
Two-step-ahead			
TS	1.63(1)	52.19(1)	99098(1)
ML	2.58(4)	89.81(4)	126699(3)
Non-GPT	2.10(3)	71.00(3)	112899(2)
GPT-4- CoT-b	1.89(2)	60.64(2)	128570(4)

Conclusion

In the tourism industry, accurate and timely demand forecasting is crucial for the decision-making process of service providers. Time series, econometric, and artificial intelligence models have been extensively studied over the past few decades. Traditionally, the accuracy of forecasting depends on the quality of the methods and feature sets used. The selection of models, determination of influencing factors, and parameters are specific to particular domains, requiring significant time and learning costs, presenting a high barrier to entry for novices (Coshall & Charlesworth, 2011). With the increased availability of data and computing power, large language models (LLMs), especially ChatGPT, have recently seen significant advancements, offering a new possibility for tourism demand forecasting. Consequently, this study aims to explore the performance of ChatGPT in forecasting tourism demand.

To empirically investigate the predictive performance of ChatGPT-4, we have collected monthly data of inbound tourists from Macau's eight major tourism source

markets from the website of the Macau Economic Development and Trade Bureau, spanning from January 1999 to April 2023. The data have been divided into two scenarios, and the performance of ChatGPT-4 in this challenging task of tourism demand forecasting has been assessed using “Zero-shot” and “Chain-of-Thought” prompting strategies comprehensively. The results indicate that compared to baseline models, ChatGPT-4's predictive capability is above average across all three accuracy indicators and relatively stable. Furthermore, the predictions of ChatGPT-4 under the CoT prompting strategy are predominantly superior to those under the Zero-shot strategy, suggesting that engaging in a continuous dialogue can enhance ChatGPT's predictive capability.

In comparison to previous studies on tourism demand forecasting, our research contributes to both tourism forecasting literature and practice. Firstly, our primary contribution is the novel proposal and empirical testing of the capabilities of ChatGPT-4 in forecasting tourism demand, which represents an initial attempt in the field of tourism demand forecasting research. This study advances the tourism forecasting literature by offering a new and innovative methodological tool and strategy. Secondly, this study contributes to the tourism forecasting literature by examining the impact of different prompting strategies on the predictive performance of ChatGPT-4. It is found that appropriate prompting strategies, such as the continuous dialogue approach adopted in this study, can enhance the predictive performance of ChatGPT-4, which points out an important direction for future research. Thirdly, this study contributes to the under-researched field of tourism forecasting in crises, by investigating ChatGPT-4's forecasting performance across different periods of a crisis (i.e., before and after COVID-19), validating ChatGPT-4's forecasting potential through two scenarios. No single model consistently outperforms the others across all periods; however, ChatGPT-4 consistently exhibits above-average performance, highlighting its stability in predictive capabilities. As such, this research offers an effective and reliable tool for tourism forecasting under crisis conditions.

Practically, the proposed ChatGPT forecasting prompt strategy framework, which fully leverages the contextual understanding and mathematical analysis capabilities of large language models, reduces the time and manpower required by traditional models. This approach does not require extensive forecasting knowledge, has a low barrier to entry, and is user-friendly for novice users. Another contribution is the utilization of continuous dialogue to iterate predictions provided by the large language model. This method allows for timely updates to tourism demand forecasts based on different dialogue instructions and additional information, facilitating convenient use without the need for repetitive modeling. Lastly, the study finds ChatGPT's predictive performance to be relatively stable across different periods. The outcomes of this work confirm the capabilities of large language models in tourism demand forecasting, potentially establishing a new benchmark model. These practical implications suggest that ChatGPT-4 can democratize the process of demand forecasting, making advanced predictive analytics accessible to a broader range of users, including those with limited

technical expertise.

The findings also underscore that time series models perform well in particular scenarios. In contrast, artificial neural network approaches seldom emerge as the top-performing models. For novice users, ChatGPT serves as a beneficial adjunct; yet, the enduring value of traditional time series models should not be underestimated. During non-crisis periods, simple time series methods continue to be a sound recommendation for demand forecasting, as they consistently deliver robust performance.

Limitations and directions for future research

This study marks a preliminary foray assessing the capability of ChatGPT for tourism demand forecasting. While the findings do highlight some merits of applying LLMs for tourism demand forecasting, it is crucial to acknowledge several limitations. Firstly, the outcomes from LLMs such as ChatGPT are characterized by limited interpretability and transparency. In contrast, traditional forecasting models, such as time series models, typically offer greater transparency, particularly regarding the values of parameter estimation. However, LLMs, which primarily analyze tourist arrivals time series data without clearly delineating variable “weights” or priorities, sacrifice the level of interpretability, resulting in a “black-box” model. In decision-making scenarios, such opaqueness could hinder stakeholders’ confidence in the reliability of LLM-driven predictions.

Additionally, ChatGPT or any AI-based conversational model should not be considered an authoritative information source. LLMs are trained on extensive public data that may contain biases or inaccuracies. When deployed for actual forecasting, LLMs may unintentionally amplify these biases, leading to skewed predictions, especially where regional differences exist.

Regarding the information sources, it is important to note that during the period of our study, ChatGPT’s training data extended only to September 2021, per OpenAI’s official documentation. Consequently, ChatGPT-4 lacks knowledge of any events or data beyond that date, including developments and tourism trends in the post-COVID-19 period. This means that ChatGPT-4 is essentially unaware of pandemic-related shifts in tourism demand after 2021. So it is fully comparable with other competing methods. However, its pre-pandemic forecasts are difficult to claim a fully standardized *ex-ante* forecasting paradigm.

Despite the limitations, the study offers important contributions and paves the way for exciting future research avenues. One promising direction is the development of models that offer greater interpretability and transparency. Efforts could focus on integrating explainability mechanisms within LLMs to render their predictions more transparent. Researchers could consider adopting tools like Shapley Additive Explanations from machine learning to quantify individual factors’ contributions to forecasts, creating a clearer breakdown of variable influence in LLM predictions. This

advancement would help demystify the “black-box” nature of LLMs, thereby bolstering the confidence of decision-makers who rely on these predictions. However, the high-dimensional context of LLMs significantly increases the computational complexity of interpretability tools, underscoring the pressing need for innovative solutions. Pursuing this line of research may prove challenging yet invaluable.

Another promising direction lies in exploring adaptive mechanisms for LLMs to dynamically adjust weights, potentially incorporating reinforcement learning or online learning techniques. By allowing real-time weight updates, such an approach could enhance LLMs’ flexibility and forecasting accuracy, particularly in volatile market environments.

While these directions may entail higher technical barriers, other research avenues offer more immediate and practical opportunities. For instance, future research could establish a longitudinal experimental framework that periodically utilizes LLMs to forecast tourism demand. Systematically comparing these forecasts with actual observed data over time would enable the accumulation of performance metrics over an extended time horizon. This pure ex-ante forecasting approach would facilitate a rigorous assessment of the true predictive capabilities of large language models, free from retrospective data influences.

In this study, we experimented a chain-of-thought prompting technique, which, despite its limitations, significantly improved the accuracy of tourism demand forecasts. This finding suggests that prompt strategies are likely to impact model performance across different research contexts. We encourage future research to delve into this area and adopt a broader array of prompting strategies to further enhance model outcomes. Additionally, researchers could leverage industry-specific data to fine-tune large language models, crafting AI agents tailored for specific forecasting tasks. Such agents could incorporate self-learning mechanisms to adjust prompts dynamically, further enhancing performance for specialized domains or specific forecasting needs.

Finally, the integration of diverse LLMs across platforms, coupled with hybridization with traditional models, presents an intriguing opportunity for research. Researchers might explore the synergistic potential of combining various LLMs to create ensemble forecasts. Techniques such as “ensemble averaging” or “weighted voting” could be employed to improve prediction accuracy and stability, capitalizing on the strengths of each model while mitigating individual biases. Furthermore, combining LLMs with transparent, interpretable models, such as time-series and regression-based approaches, could strike a balance between prediction accuracy and transparency. For instance, LLMs could be leveraged for feature extraction, with traditional models performing the final numerical forecasting, thereby enhancing the overall reliability of predictions. This “division of labor” approach, where LLMs handle complex feature generation and transparent models conduct the predictive calculations, could result in a hybrid framework that maintains both accuracy and interpretability.

References

- Ahn, Y.-G., & Lee, M.-K. (2021). Elasticity of the Number of World Cruise Tourists Using the Vector Error Correction Model. *Sustainability*, 13(16), Article 16. <https://doi.org/10.3390/su13168743>
- Ali, K., Barhom, N., Tamimi, F., & Duggal, M. (2024). ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. *European Journal of Dental Education*, 28(1), 206–211. <https://doi.org/10.1111/eje.12937>
- Antonio, N., De Almeida, A., & Nunes, L. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*, 60(4), 298–319. <https://doi.org/10.1177/1938965519851466>
- Athanasopoulos, G., Song, H., & Sun, J. A. (2018). Bagging in Tourism Demand Modeling and Forecasting. *Journal of Travel Research*, 57(1), 52–68. <https://doi.org/10.1177/0047287516682871>
- Azad, A. S., Sokkalingam, R., Daud, H., Adhikary, S. K., Khurshid, H., Mazlan, S. N. A., & Rabbani, M. B. A. (2022). Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study. *Sustainability*, 14(3), Article 3. <https://doi.org/10.3390/su14031843>
- Belmahdi, B., Louzazni, M., & Bouardi, A. E. (2020). A hybrid ARIMA–ANN method to forecast daily global solar radiation in three different cities in Morocco. *The European Physical Journal Plus*, 135(11), 925. <https://doi.org/10.1140/epjp/s13360-020-00920-9>
- Biswas, S. S. (2023). Role of Chat GPT in Public Health. *Annals of Biomedical Engineering*, 51(5), 868–869. <https://doi.org/10.1007/s10439-023-03172-7>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Cang, S. (2011). A Non-Linear Tourism Demand Forecast Combination Model. *Tourism Economics*, 17(1), 5–20. <https://doi.org/10.5367/te.2011.0031>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, J. L., Li, G., Wu, D. C., & Shen, S. (2019). Forecasting Seasonal Tourism Demand Using a Multiseries Structural Time Series Method. *Journal of Travel Research*, 58(1), 92–103. <https://doi.org/10.1177/0047287517737191>
- Chen, Z., Zheng, L. N., Lu, C., Yuan, J., & Zhu, D. (2023). ChatGPT Informed Graph Neural Network for Stock Movement Prediction. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4464002>

- Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, 24(3), 323–330.
[https://doi.org/10.1016/S0261-5177\(02\)00068-7](https://doi.org/10.1016/S0261-5177(02)00068-7)
- Claveria, O., Monte, E., & Torra, S. (2015). Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information. *International Journal of Tourism Research*, 17(5), 492–500. Scopus.
<https://doi.org/10.1002/jtr.2016>
- Coshall, J. T., & Charlesworth, R. (2011). A management orientated approach to combination forecasting of tourism demand. *Tourism Management*, 32(4), 759–769. <https://doi.org/10.1016/j.tourman.2010.06.011>
- Dantas, T. M., Cyrino Oliveira, F. L., & Varela Repolho, H. M. (2017). Air transportation demand forecast through Bagging Holt Winters methods. *Journal of Air Transport Management*, 59, 116–123.
<https://doi.org/10.1016/j.jairtraman.2016.12.006>
- Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1), 3.
<https://doi.org/10.1186/1471-2105-7-3>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694.
<https://doi.org/10.1007/s11023-020-09548-1>
- Frosolini, A., Franz, L., Benedetti, S., Vaira, L. A., de Filippis, C., Gennaro, P., Marioni, G., & Gabriele, G. (2023). Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *European Archives of Oto-Rhino-Laryngology*, 280(11), 5129–5133. <https://doi.org/10.1007/s00405-023-08205-4>
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, 6(1), Article 1. <https://doi.org/10.1038/s41746-023-00819-6>
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10), 2451–2471.
<https://doi.org/10.1162/089976600300015015>
- Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management*, 23(5), 499–510. [https://doi.org/10.1016/S0261-5177\(02\)00009-2](https://doi.org/10.1016/S0261-5177(02)00009-2)
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Gursoy, D., Li, Y., & Song, H. (2023). ChatGPT and the hospitality and tourism industry: An overview of current trends and future research directions. *Journal of Hospitality Marketing & Management*, 32(5), 579–592.
<https://doi.org/10.1080/19368623.2023.2211993>
- Hajirahimi, Z., & Khashei, M. (2019). Hybrid structures in time series modeling and

- forecasting: A review. *Engineering Applications of Artificial Intelligence*, 86, 83–106. <https://doi.org/10.1016/j.engappai.2019.08.018>
- He, Y., & Hooi, B. (2024). *UniGraph: Learning a Cross-Domain Graph Foundation Model From Natural Language* (arXiv:2402.13630). arXiv. <https://doi.org/10.48550/arXiv.2402.13630>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). *Study and Analysis of Chat GPT and its Impact on Different Fields of Study* (SSRN Scholarly Paper 4402499). <https://papers.ssrn.com/abstract=4402499>
- Khaidem, L., Saha, S., & Dey, S. R. (2016). *Predicting the direction of stock market prices using random forest* (arXiv:1605.00003). arXiv. <https://doi.org/10.48550/arXiv.1605.00003>
- Klang, E., Portugez, S., Gross, R. et al. (2023). Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: A medical education pilot study with GPT-4. *BMC Medical Education*, 23(1), 772. <https://doi.org/10.1186/s12909-023-04752-w>
- Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, 75, 410–423. <https://doi.org/10.1016/j.annals.2019.01.014>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>
- Li, H., Park, S. Y., & Seo, J. H. (2011). Quantile Elasticity of International Tourism Demand for South Korea Using the Quantile Autoregressive Distributed Lag Model. *Tourism Economics*, 17(5), 997–1015. <https://doi.org/10.5367/te.2011.0083>
- Li, X., Feng, H., Yang, H., & Huang, J. (2024). Can ChatGPT reduce human financial analysts' optimistic biases? *Economic and Political Studies*, 12(1), 20–33. <https://doi.org/10.1080/20954816.2023.2276965>
- Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200209. <https://doi.org/10.1098/rsta.2020.0209>
- Lim, C., Chang, C., & McAleer, M. (2009). Forecasting h(m)otel guest nights in New Zealand. *International Journal of Hospitality Management*, 28(2), 228–235. <https://doi.org/10.1016/j.ijhm.2008.08.001>
- Liu, Y.-Y., Tseng, F.-M., & Tseng, Y.-H. (2018). Big Data analytics for forecasting

- tourism destination arrivals with the applied Vector Autoregression model. *Technological Forecasting and Social Change*, 130, 123–134.
<https://doi.org/10.1016/j.techfore.2018.01.018>
- Lopez-Lira, A., & Tang, Y. (2023). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models* (arXiv:2304.07619).
 arXiv. <https://doi.org/10.48550/arXiv.2304.07619>
- Mumtaz, U., & Mumtaz, S. (2023). *Potential of ChatGPT in predicting stock market trends based on Twitter Sentiment Analysis* (arXiv:2311.06273). arXiv.
<https://doi.org/10.48550/arXiv.2311.06273>
- Nowman, K. B., & Van Dellen, S. (2012). Forecasting Overseas Visitors to the UK Using Continuous Time and Autoregressive Fractional Integrated Moving Average Models with Discrete Data. *Tourism Economics*, 18(4), 835–844.
<https://doi.org/10.5367/te.2012.0144>
- Oh, C.-O. (2005). The contribution of tourism development to economic growth in the Korean economy. *Tourism Management*, 26(1), 39–44.
<https://doi.org/10.1016/j.tourman.2003.09.014>
- Palmer, A., José Montaña, J., & Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management*, 27(5), 781–790. <https://doi.org/10.1016/j.tourman.2005.05.006>
- Seo, J. H., Park, S. Y., & Yu, L. (2009). The analysis of the relationships of Korean outbound tourism demand: Jeju Island and three international destinations. *Tourism Management*, 30(4), 530–543.
<https://doi.org/10.1016/j.tourman.2008.10.013>
- Shen, S., Li, G., & Song, H. (2008). An Assessment of Combining Tourism Demand Forecasts over Different Time Horizons. *Journal of Travel Research*, 47(2), 197–207. <https://doi.org/10.1177/0047287508321199>
- Shi, J., Guo, J., & Zheng, S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews*, 16(5), 3471–3480.
<https://doi.org/10.1016/j.rser.2012.02.044>
- Sina, L. B., Secco, C. A., Blazevic, M., & Nazemi, K. (2023). Hybrid Forecasting Methods—A Systematic Review. *Electronics*, 12(9), Article 9.
<https://doi.org/10.3390/electronics12092019>
- Skórnoś, D., & Kmiecik, M. (2023). Supporting the inventory management in the manufacturing company by ChatGPT. *Logforum*, 19(4), 535–554.
<https://doi.org/10.17270/J.LOG.2023.917>
- Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism Management*, 29(2), 203–220.
<https://doi.org/10.1016/j.tourman.2007.07.016>
- Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting: Launching the *Annals of Tourism Research* Curated Collection on tourism demand forecasting. *Annals of Tourism Research*, 75, 338–362.

- <https://doi.org/10.1016/j.annals.2018.12.001>
- Stroh, N. (2024). TrackGPT -- A generative pre-trained transformer for cross-domain entity trajectory forecasting (arXiv:2402.00066). arXiv.
<https://doi.org/10.48550/arXiv.2402.00066>
- Sun, S., Li, Y., Guo, J., & Wang, S. (2022). Tourism demand forecasting: An ensemble deep learning approach. *Tourism Economics*, 28(8), 2021–2049.
<https://doi.org/10.1177/13548166211025160>
- Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *Interspeech 2012*, 194–197.
<https://doi.org/10.21437/Interspeech.2012-65>
- Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.
https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Wang, Q., & Jiang, F. (2019). Integrating linear and nonlinear forecasting techniques based on grey theory and artificial intelligence to forecast shale gas monthly production in Pennsylvania and Texas of the United States. *Energy*, 178, 781–803. <https://doi.org/10.1016/j.energy.2019.04.115>
- Wang, Y.-S. (2009). The impact of crisis events and macroeconomic activity on Taiwan’s international inbound tourism demand. *Tourism Management*, 30(1), 75–82. <https://doi.org/10.1016/j.tourman.2008.04.010>
- Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, 29(1), 507–529. <https://doi.org/10.1108/IJCHM-05-2015-0249>
- Wu, D. C., Zhong, S., Song, H., & Wu, J. (2024a). Do topic and sentiment matter? Predictive power of online reviews for hotel demand forecasting. *International Journal of Hospitality Management*. 120, 103750.
<https://doi.org/10.1016/j.ijhm.2024.103750>
- Wu, D. C., Zhong, S., Wu, J., & Song, H. (2024b). Tourism and hospitality forecasting with big data: A systematic review of the literature. *Journal of Hospitality & Tourism Research* 1–20.
<http://doi.org/10.1177/10963480231223151>
- Wu, D. C., Zhong, S., Qiu, R. T. R., & Wu, J. (2022). Are customer reviews just reviews? Hotel forecasting using sentiment analysis. *Tourism Economics*, 28(3), 795-816. <https://doi.org/10.1177/13548166211049865>
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136.

- <https://doi.org/10.1109/JAS.2023.123618>
- Yan, H., Kong, L., Gui, L., Chi, Y., Xing, E., He, Y., & Zhang, K. (2024). *Counterfactual Generation with Identifiability Guarantees* (arXiv:2402.15309). arXiv. <http://arxiv.org/abs/2402.15309>
- Yang, Y., & Wong, K. K. F. (2012). A Spatial Econometric Approach to Model Spillover Effects in Tourism Flows. *Journal of Travel Research*, 51(6), 768–778. <https://doi.org/10.1177/0047287512437855>
- Yu, S., Dong, H., Chen, Y., He, Z., & Shi, X. (2019). Clothing Sales Forecast Based on ARIMA-BP Neural Network Combination Model. *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 367–372. <https://doi.org/10.1109/ICPICS47731.2019.8942427>
- Zhang, H., Jiang, Z., Gao, W., & Yang, C. (2022). Time-varying impact of economic policy uncertainty and geopolitical risk on tourist arrivals: Evidence from a developing country. *Tourism Management Perspectives*, 41, 100928. <https://doi.org/10.1016/j.tmp.2021.100928>
- Zhang, H., Song, H., Wen, L., & Liu, C. (2021). Forecasting tourism recovery amid COVID-19. *Annals of Tourism Research*, 87, 103149. <https://doi.org/10.1016/j.annals.2021.103149>