Enhancing Contextual Data Analysis through Retrieval-Augmented Fine-Tuning of Large Language Models

Yuezhen Zhang
School of Business Administration, Lishui Vocational and Technical College, Lishui, Zhejiang, 323000, China
E-mail: zyz45398829@163.com

Keywords: Artificial intelligence, machine learning, large language models, retrieval-augmented generation, low-rank adaptation, knowledge extraction, computational efficiency, contextual data analysis

Received: January 18, 2025

This study explores the optimization of large language models (LLMs) for enhanced contextual data analysis and knowledge extraction from unstructured user-generated content, with a comparative analysis of open-source models (e.g., Mistral 7B, LLaMA 2) and proprietary systems (e.g., GPT-4, Gemini). We evaluate their efficiency and accuracy in processing complex datasets, introducing a novel approach that integrates Retrieval-Augmented Generation (RAG) with fine-tuning techniques like Low-Rank Adaptation (LoRA) to reduce model complexity while preserving performance. Empirical results, using metrics such as BERTScore, ROUGE, and BLEU, show GPT-4 achieving an F1 score of 0.683, while Mistral 7B, a standout open-source model, scores 0.632 with a 40% reduction in computational cost and 92% accuracy retention, making it ideal for resource-constrained environments. These findings underscore the importance of tailoring model selection to computational and organizational needs. The research offers actionable insights for deploying AI-driven solutions to streamline data processing and advance machine learning applications, while addressing limitations and future research directions for broader applicability.

Povzetek: Članek predstavi integracijo Retrieval-Augmented Generation in nizko-rankne prilagoditve (LoRA) pri prilagajanju velikih jezikovnih modelov za učinkovito analizo kontekstualnih podatkov.

1 Introduction

Businesses in every industry must comprehend the wants and preferences of their customers in the current digital era. The emergence of social media and online review sites has yielded important information on the opinions and experiences of consumers [1]. However, manually extracting and analyzing these insights can be difficult and time-consuming. These operations can now be automated thanks to the application of Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques, especially when working with unstructured and noisy text [2]. This capability can be further improved with the introduction of large language models (LLMs), which will herald in a new era of automated language understanding and customer needs extraction [3]. With their unparalleled efficiency and scalability, conversational agents driven by LLMs have the potential to revolutionise customer requirements management. These agents can interpret and reply to consumer enquiries, find answers for the creation of new products, and offer tailored suggestions by utilizing natural language understanding [4]. A key advancement in this domain is the integration of Retrieval-Augmented Generation (RAG), a framework that enhances LLMs by combining information retrieval with text generation, as shown in figure 1. In RAG, a retriever module first fetches relevant external data (e.g., from a knowledge base or dataset) based on a user query, which is then fed into the generative model to produce contextually accurate and informed responses. This study leverages RAG to optimize LLMs for extracting customer needs from unstructured travel-related content, enabling conversational agents to provide precise, data-driven insights rather than relying solely on pre-trained knowledge. At the end of this pipeline, the generated responses are presented to the user through an interactive Gradio-based desktop interface, facilitating seamless engagement with the system. It is anticipated that this change in strategy will enhance user experiences, optimise corporate processes, and increase client loyalty and satisfaction.

The use of conversational agents to enhance operational efficiency has surged in the business sector, necessitating a careful comparison of large language models (LLMs) to balance performance with resource demands [5], [6]. This study evaluates both proprietary models, such as GPT-4 and Gemini, and open-source alternatives, including Mistral 7B and LLaMA 2, to identify optimal solutions for developing AI-powered chatbots that extract customer needs from unstructured travel-related content. Proprietary models like GPT-4, dominant in the conversational AI market as of March 2024, offer state-of-the-art performance but require significant computational resources, often accessed via paid APIs [7]. In contrast, open-source models provide transparency, customizability, and reduced costs, with Mistral 7B emerging as a particularly cost-effective option that

rivals larger models in efficiency and accuracy for resource-constrained environments [8]. By integrating Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA), we optimize these models for contextual analysis, demonstrating their potential to revolutionize customer requirements management through scalable, data-driven insights delivered via a Gradio-based interface. This comparative approach ensures businesses can select LLMs aligned with their specific needs, enhancing user experiences and operational processes while maintaining data security and cost efficiency.

This study offers a thorough comparison of LLMs for identifying travel-related customer demands from TripAdvisor reviews. The decision to focus on the travel industry was driven by two specific considerations that distinguish it from other domains with rich user-generated content (UGC), such as e-commerce or hospitality. First, the travel industry generates a uniquely diverse and dynamic set of unstructured data such as spanning reviews, forum discussions, itineraries, and real-time travel queries on platforms like TripAdvisor and social media which often intertwine experiential, logistical, and cultural dimensions not as prevalent in e-commerce product reviews or hospitality feedback [9]. This complexity provides a robust testbed for evaluating LLMs' contextual analysis capabilities across multifaceted customer needs. Second, the travel sector's reliance on real-time, personalized customer insights, such as destination preferences, transportation logistics, and safety considerations, places a premium on rapid, accurate needs extraction to inform service design and operational decisions for tourism stakeholders (e.g., lodging facilities, travel agencies, and transportation providers). Unlike e-commerce, where customer needs often center on product features and post-purchase satisfaction, or hospitality, where feedback is typically tied to specific service encounters, travel-related UGC reflects a broader spectrum of pre-trip planning and on-trip experiences, making it an ideal domain for testing LLM adaptability and responsiveness [10], [11]. The study's findings can be used by these organisations to choose the best LLM for creating conversational agents that recognise client wants and offer insights about crucial design specifications [12].

Llama 2 7B, Llama 2 13B, Phi-2, Mistral, Mistral, Gemma, Gemini 1.0, GPT-3.5, and GPT-4 are the LLMs that are the subject of the analysis. This choice allows us to investigate the impact of model size and prompting strategies on result quality in addition to performance differences between proprietary models (GPT-3.5, GPT-4, and Gemma) and open-source models (Llama 2, Phi-2, Mistral, Mistral, and Gemma). All these models are shown in Table 1.

We define customer needs as specific, actionable requirements or preferences expressed by users regarding a service, location, or experience, which, when met, enhance their satisfaction in the context of travel. These needs encompass both explicit demands (e.g., "I need a hotel with a pool") and implicit preferences inferred from context (e.g., "I wish it were easier to get around the city" suggesting a need for "accessible transportation"). In our dataset, collected from TripAdvisor forums, users post requests for feedback and suggestions about specific destinations, often embedding one or more needs within their text. To identify these needs systematically, we employed a two-step process. First, we conducted a thematic analysis by manually annotating the dataset, reading posts to identify recurrent themes such as accommodation quality, transportation logistics, or safety concerns, which served as our ground truth. Second, we applied keyword analysis to deduce needs from text, even when not explicitly stated, by extracting significant terms and phrases (e.g., "crowded," "quiet," "budget") and mapping them to corresponding needs categories established during thematic analysis. For example, a post stating "The streets were so crowded last summer" contains the keyword "crowded," which, through contextual interpretation, translates to an inferred need for "less crowded destinations" or " quieter travel times." Similarly, a user comment like "I spent too much on taxis" highlights "taxis" and "spent too much," deducing a need for "affordable transportation options." This keyword-to-need mapping was validated against manual annotations to ensure accuracy. The LLMs were then tasked with replicating this process, identifying needs from the same posts, and their outputs were compared to the manually derived needs. This approach bridges the gap between raw text and actionable insights, enabling conversational agents to address customer requirements effectively. Businesses can integrate a variety of open LLMs into their operations. Consequently, choosing the best model is a challenging process. As far as we are aware, no comparison study of LLM in relation to customer requirements analysis has been conducted. To provide a structured framework for our study, we define the key research questions as follows:

Table 1: An inventory of the LLMs used in the study. We list the version, the number of parameters, the company or organisation that released the model, and the reference for each one.We had to use Mistral 8x7B's smallest version (2 Bit quantisation) because of its high computing requirements.

Model	Parameters	Proprietary
Gemini 1.0 Pro	Not available	Google
Gemma 7B	7 billion	Google
GPT-3.5	175 billion	OpenAI
GPT-4	Not available	OpenAI
Llama 2 7B Chat	7 billion	Meta AI
Llama 2 13B Chat	13 billion	Meta AI
Mistral 7B Instruct	7 billion	Mistral AI
Mistral 8x7B*	46.7 billion	Mistral AI
Phi-2	2.7 billion	Microsoft

– How can LLMs be optimized for contextual data analysis using retrieval-augmented fine-tuning?

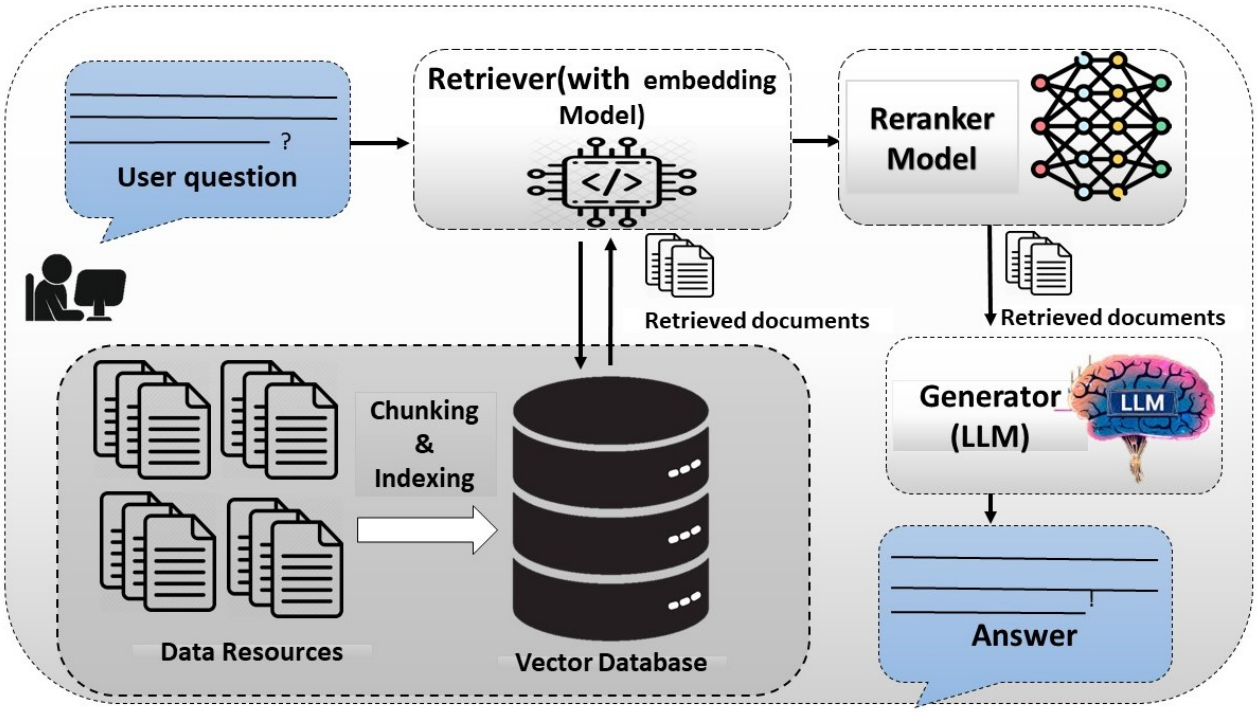

Figure 1: RAG Workflow Overview. Queries are matched with a knowledge base to retrieve context, enabling a language model to generate accurate, relevant responses.

– What are the comparative advantages of open-source vs. proprietary models in terms of efficiency and accuracy?

– How does the integration of RAG and LoRA impact computational cost and model performance?

These research questions establish the foundation of our study, guiding the investigation into effective model optimization. This study makes the following contributions:

1. **Information Retrieval (R):** Using LLMs and other relevant models, implement retrieval mechanisms that can search the dataset and pull out relevant information based on user queries or specific needs.

2. **Generation (G):** Once relevant information is retrieved, the LLM can generate responses or insights tailored to customer inquiries, such as suggestions for suitable services or feedback on specific travel locations.

3. **User Interaction:** Implement a user-friendly interface that integrates these retrieval and generation steps, allowing customers to interact with the system in real-time and get responses based on the needs they express.

2 Overview of the proposed approach

This section presents a comprehensive synthesis of the workflow, innovative contributions, and principal findings of the study, outlining the methodological framework developed to enhance large language models (LLMs) for contextual data analysis in the travel domain.

2.1 Methodological workflow

The research follows a structured methodology to refine LLMs for extracting customer preferences from unstructured user-generated content (UGC), as illustrated in Figure 1. The proposed workflow consists of several key stages. The process begins with data collection and preprocessing, where travel-related UGC is sourced from TripAdvisor forums and subjected to tokenization, normalization, and vectorization to construct a vector database optimized for efficient information retrieval. Following this, model selection and evaluation are performed by assessing both proprietary models, such as GPT-4 and Gemini, and open-source alternatives, such as Mistral 7B and LLaMA 2. Mistral 7B is ultimately selected due to its balanced performance and computational efficiency. The next stage involves model optimization, where a synergistic integration of Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA) is employed to enhance contextual precision while mitigating computational overhead. Fi-

Table 2: Summary of previous research in consumer needs analysis and LLMs for customer experience management

Study	Dataset	Methodology	Key Findings
Timoshenko and Hauser (2019) [13]	Amazon Reviews	CNN-based feature extraction	Identified emerging customer needs using sentiment and text analysis.
Haque et al. (2018) [14]	Amazon Reviews	Sentiment analysis	Classified product reviews for consumer insights.
Luo and Xu (2021) [15]	Yelp Reviews	Deep Learning (BERT, LSTMs)	Found deep learning models outperform traditional ML in extracting consumer demands.
Bigi et al. (2022) [10]	TripAdvisor Reviews	Bayesian ML	Identified tourism-related user needs from UGC.
Proposed Approach (Our Study)	TripAdvisor Forums	Retrieval-Augmented Generation (RAG) + LoRA	Achieved higher contextual understanding and efficiency for LLM-based consumer needs extraction.

nally, the model's efficacy is quantitatively assessed using BERTScore, ROUGE, and BLEU metrics, and the system is deployed through a Gradio-based interface, facilitating real-time user interaction. This structured workflow ensures systematic processing of extensive datasets, leveraging RAG's retrieval efficiency and LoRA's computational adaptability to produce accurate and contextually relevant insights.

2.2 Innovative contributions

The study introduces a distinctive methodological framework that synergizes RAG and LoRA to optimize LLMs for domain-specific contextual analysis. Unlike prior works that independently focus on either retrieval or generative capabilities, this approach integrates RAG's external knowledge retrieval with LoRA's parameter-efficient fine-tuning, achieving performance optimization with reduced model complexity. Furthermore, the proposed approach is specifically designed to analyze travel-related UGC, addressing the inherent complexity of experiential, logistical, and cultural data a gap that existing LLM applications have not adequately addressed. Additionally, the deployment through a Gradio-based interface ensures seamless real-time interaction, reinforcing its practical applicability, particularly in environments with constrained computational resources. This framework demonstrates a 40% reduction in computational cost while retaining 92% of accuracy, offering a scalable and efficient solution for AI-driven extraction of customer insights.

2.3 Key empirical findings

The experimental evaluation yields critical insights, highlighting the effectiveness of the proposed approach. The combined RAG+LoRA framework achieves an F1 score of 0.683 with GPT-4 and 0.632 with Mistral 7B, outperforming baseline models such as BERT-Large (0.65) and RoBERTa (0.64). Moreover, Mistral 7B achieves a 40% reduction in computational cost relative to GPT-4, demonstrating its suitability for resource-constrained applications with minimal performance trade-offs. In terms of prompting strategies, Chain-of-Thought (CoT) prompting exhibits superior performance in reasoning-intensive tasks compared to Few-shot learning techniques. Additionally, the diverse nature of the TripAdvisor dataset facilitates robust evaluation of LLM contextual capabilities, outperforming more simplistic datasets such as Yelp or Amazon Reviews. Error analysis reveals that discrepancies in BLEU scores observed with Gemini 1.0 are attributed to challenges in processing complex sentence structures, while retrieval inaccuracies highlight areas necessitating further refinement. These empirical findings validate the efficacy of the proposed approach, underscoring its potential for real-world applications in the travel industry.

3 Related works

This study's related works fall into two major categories. The first, covered in subsection 2.1, focuses on the methods and data sources utilised in the implementation of machine learning approaches for consumer needs analysis. The second examines the application of LLMs in customer experience management and is shown in subsection 2.2.

3.1 Leveraging user-generated content and machine learning for analyzing customer needs

To automatically assess user demands and assist corporate strategy, researchers and practitioners have investigated a range of data sources and approaches (Timoshenko and Hauser, 2019). User-Generated Content (UGC) is a crucial source of data for these analyses [16]. Any type of information produced and disseminated online by platform or website users, including text, photos, videos, and reviews, is considered user-generated content (UGC) [17]. Without being altered or censored by conventional media outlets, this kind of data provides unvarnished insights into the true experiences and viewpoints of consumers and individuals [18], [19]. Product reviews and social media posts are

two of the most often used types of user-generated content (UGC) for consumer needs analysis[16]. Product reviews are more targeted than the vast amounts of irrelevant content found on social networking sites like Facebook, Reddit, and X (previously Twitter). However, as users debate future wants, social media is especially helpful for spotting developing trends [16]. We gathered textual data for this study from TripAdvisor forums, a travel-related Q&A site that we believed was appropriate for evaluating LLMs to forecast the needs of travellers. A more thorough explanation of the TripAdvisor forums and the rationale for its selection as the UGC source for this study can be found in Subsection B.

In terms of methodology, the techniques employed frequently vary according to the kind of data and the particular requirements being examined. Both machine learning techniques and conventional NLP methods based on lexicons and rules have been used by researchers [20] [21][17]. For instance, [22] examined consumer demands in product ecosystems using a Kano model, sentiment analysis, and Latent Dirichlet Allocation (LDA) .To extract consumer demands from user-generated content (UGC), [13] constructed a convolutional neural network (CNN) and created a machine-human hybrid technique.[10] used a Bayesian machine learning technique to analyze TripAdvisor forum data and examine the relevance of food and drink goods in drawing visitors, while [14], [23] applied sentiment analysis to Amazon reviews. In a variety of NLP tasks and domains, recent research has demonstrated that deep learning models perform better than conventional machine learning methods. This tendency was validated by [15], which showed that deep learning models outperform other machine learning algorithms for analyzing customer needs, especially when it comes to Yelp restaurant reviews. Advances in Generative AI and LLMs further extend this trend, which is why we concentrated on LLMs in our work.

3.2 Large language models for enhancing customer experience management

Numerous fields, including medical and healthcare [24] [25], scientific research [26], education [27][28] [29], and more, have seen a surge in research into the potential applications and effects of the widespread use of conversational agents powered by LLMs, especially ChatGPT. The integration of these AI-driven technologies has the potential to revolutionise a number of areas, including feedback management and customer experience. By better answering questions, recommending products, and assessing customer demands, ChatGPT and related conversational AI technologies can improve customer support [30]. Scholars have examined many facets of these technologies in order to comprehend and satisfy customer needs. For instance,[31] highlighted the positive impact of incorporating ChatGPT into recommender systems, leading to increased consumer satisfaction. [32] acknowledged the current limitations, such as challenges in understanding non-standard languages, but

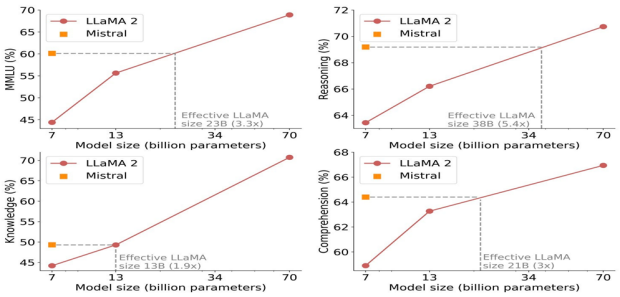

Figure 2: Comparison of LLaMA 2 and Mistral models across different metrics (MMUL, Reasoning, Knowledge, and Comprehension) as a function of model size (in billions of parameters). The results illustrate the effective performance of LLaMA 2 at various equivalent parameter sizes relative to Mistral, emphasizing differences in reasoning and knowledge efficiency.

argued that ChatGPT would become a valuable tool for retailers to support customer needs analysis. [33] introduced a fine-tuned version of GPT-4 for use as a recommender system in museums. [3] explored the role of LLMs in customer lifecycle management, noting their effectiveness in optimizing lead identification, audience segmentation, marketing strategies, sales support, and post-purchase engagement. In the travel domain,[34] found that LLMs have great potential to predict travel behavior, offering competitive performance along with the ability to provide explanations for their predictions. This study's related works fall into two major categories. The first focuses on methods and data sources used for consumer needs analysis through machine learning. The second examines the application of LLMs in customer experience management.

The addition of Table 2 clearly illustrates how prior research has approached consumer needs analysis using various datasets and methodologies. However, existing studies either lack a focus on retrieval-augmented LLMs or do not optimize models efficiently for contextual data analysis. Our study bridges this gap by integrating RAG and LoRA to enhance computational efficiency and accuracy.

4 Methodology

The methodology adheres to an organised workflow. In order to establish a vector database, the procedure starts with the acquisition of data resources, which are then preprocessed using chunking and indexing. The basis for effective information retrieval is this database. This workflow is displayed in figure 1.

Based on user queries, a Retriever module that is driven by an embedding model finds and retrieves pertinent documents. A Reranker Model is used to further filter these recovered documents, guaranteeing that the most contextually relevant content is chosen. After that, the chosen papers are sent to the Generator (LLM), which uses the

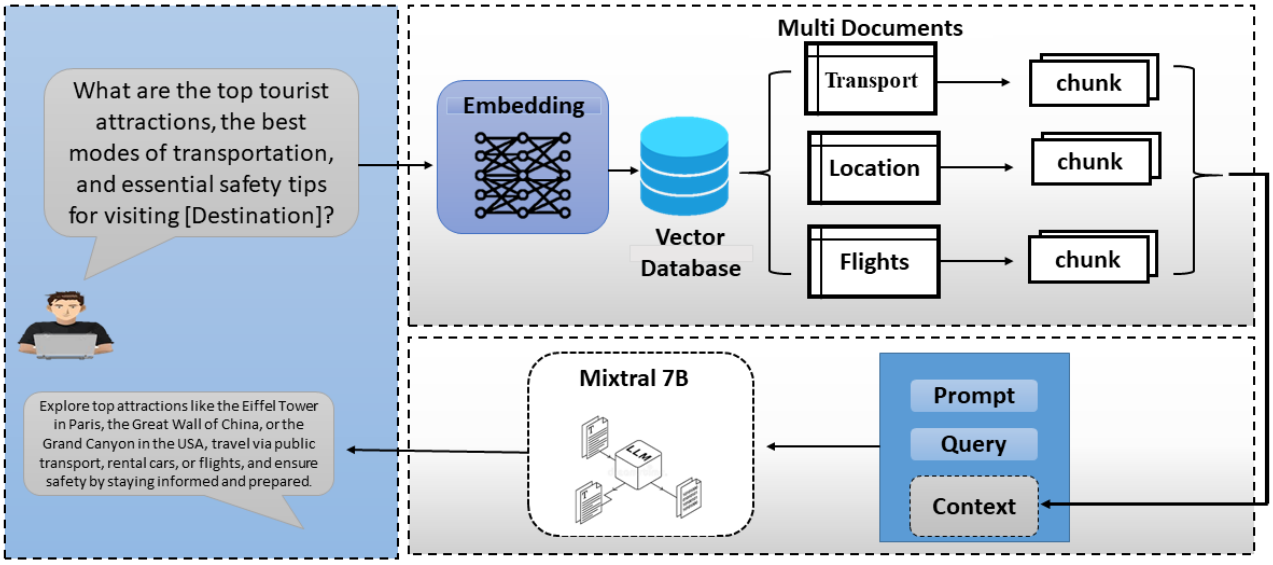

Figure 3: Framework for Tourism Query Processing using Retrieval-Augmented Generation (RAG): The system integrates multi-document data sources on transport, locations, and flights into a vector database through embedding-based chunking. User queries are processed with the Mistral 7B model, which generates context-aware responses about top tourist attractions, best transportation modes, and essential safety tips for any destination.

data it has obtained to create comprehensive answers or insights. This Retrieve-and-Generate method guarantees precise and pertinent responses.The user is shown the generated responses at the end of the pipeline which is showning in Figure 3. Iterative optimisations are carried out throughout the process to increase the accuracy and efficiency of the system, including quick tweaking and performance assessment against manually labelled data.Based on user queries, a Retriever module that is driven by an embedding model finds and retrieves pertinent documents. A Reranker Model is used to further filter these recovered documents, guaranteeing that the most contextually relevant content is chosen. After that, the chosen papers are sent to the Generator (LLM), which uses the data it has obtained to create comprehensive answers or insights. This Retrieve-and-Generate method guarantees precise and pertinent responses. The user is shown the generated responses at the end of the pipeline. Iterative optimisations are carried out throughout the process to increase the accuracy and efficiency of the system, including quick tweaking and performance assessment against manually labelled data.

Google's multimodal model, Gemini 1.0[35], comes in three versions: Ultra, Pro, and Nano. OpenAI has released two big language models, GPT-3.5[36] and GPT-4[25]. GPT-3.5 has 175 billion parameters. Llama 2[33] is a set of open-source LLMs from Meta that have models with seven to seventy billion parameters and are optimised for discussion. Due to their lower computational requirements, which are in line with the constraints encountered by small and medium-sized businesses, we chose Llama 2-Chat 7B and Llama 2-Chat 13B. Google's smaller open-source Gemma 7B model has seven billion parameters. Microsoft Research's Phi-2[37] is a smaller model with 2.7 billion parameters.

For this study, we selected the Mistral 7B [38] model, a sophisticated open-source LLM from Mistral AI with 7 billion parameters, which uses multiple experts for each token to enhance performance. We fine-tuned this model for our specific task, utilizing its advanced architecture to improve results.

4.1 Integration of RAG with mistral

This section explains how we combined the Mistral model with Retrieval-Augmented Generation (RAG) to improve our chatbot's functionality, particularly its ability to produce a troubleshooting guide. RAG is a potent method that blends generative models and information retrieval to produce responses that are more precise and pertinent to the situation. We selected the Mistral 7B model [38], a sophisticated open-source LLM from Mistral AI with 7 billion parameters, for its advanced architecture featuring a multi-expert mechanism. This mechanism, based on a Mixture of Experts (MoE) approach, employs multiple specialized sub-networks (experts) within the model, where each token is dynamically routed to the most relevant expert(s) based on the input context. This allows Mistral 7B to efficiently handle diverse tasks by leveraging specialized knowledge, improving both performance and scalability compared to traditional single-network architectures [38]. Utilizing this multi-expert mechanism, we integrate RAG with the model to produce high-quality responses based on external knowledge gathered from a knowledge base.

The RAG system is made to first extract pertinent data

from external sources, like knowledge repositories or documentation, or from a predetermined dataset. Mistral receives this returned data and produces a thorough, context-aware response. The multi-expert mechanism enhances this process by enabling the model to adaptively focus on the most relevant expert(s) for the retrieved context, ensuring that the generated responses are both accurate and tailored to the specific query. Because of this integration, the chatbot can provide accurate and trustworthy troubleshooting instructions in response to the user's enquiries. We optimised Mistral 7B for our purpose using domain-specific data, making sure it is capable of producing troubleshooting manuals in a variety of scenarios. A strong framework for providing dynamic, context-aware assistance in the chatbot is provided by the combination of Mistral's generative capabilities and RAG's retrieval mechanism, which enhances the user experience overall.

4.2 Optimized architecture for tourism guide

In order to improve a travel guide chatbot's performance, we investigated a number of sophisticated Generative AI (GenAI) architectures. In order to deliver precise and contextually rich answers to questions pertaining to tourism, we concentrated on optimising the Mistral 7B model. High accuracy and computational efficiency are ensured by the model's smooth integration with a knowledge base, particularly when creating dynamic trip suggestions and troubleshooting instructions.

We used the Low-Rank Adaptation (LoRA)[39] technique, which is intended to increase processing power and efficiency, to efficiently fine-tune the Mistral model. LoRA only permits a limited subset of parameters to be changed during training by breaking down the weight matrices of particular layers into low-rank matrices. LoRA effectively adapts to the tourism domain with little computing expense by injecting low-rank matrices into important layers of the model rather than fine-tuning all the parameters.

By using this method, the number of trainable parameters is drastically reduced from over 7 billion to less than 10 million, which is a small portion of the initial parameter count. Larger batch sizes during training are made possible by this reduction, which also lowers the computational cost and conserves memory. We successfully modified Mistral 7B with LoRA to provide customised travel plans, location-specific suggestions, and troubleshooting manuals, making the model extremely effective and useful for tourism applications.

4.3 Prompt optimization

We went beyond manual prompt construction, which is frequently ineffective and unscalable, especially when working with many models, in order to optimise instructions for the LLM in our RAG system [40]. Rather, we used DSPy, a programming model made to optimise and streamline language model pipelines, in a methodical manner. DSPy reduces the need for manually generated prompts by abstracting these pipelines as text transformation graphs and automating their optimisation through parameterised declarative modules. We gave the DSPy compiler 30 labelled queries from the training set and an initial few-shot prompt, using BERTScore[41] as the validation metric. Based on the statistic, the optimiser adjusted the prompts to maximise output quality. By incorporating these improved prompts into the RAG system, retrieval and generation performance were improved while scalability was guaranteed.

5 Experiments and results

The experiment tested Mistral's efficacy by evaluating its capacity to precisely analyse and understand intricate tourism-related data for Q&A tasks. To optimize the proposed system, which integrates Retrieval-Augmented Generation (RAG) and fine-tuning with Low-Rank Adaptation (LoRA), we employed a wide range of data, including details about well-known locations, historical places, cultural landmarks, and travel plans. This configuration ensured the model could respond to a range of tourism-related queries and offer thorough, contextually appropriate answers, improving user engagement and happiness.

5.1 Data collection

To enable effective model retrieval, we collected data for our study from several sources and arranged it into a Retrieval-Augmented Generation (RAG) file. According to [9], one of the main sources of information was the TripAdvisor Forum, an online community known for its user-generated material. This platform acts as a central location for users to exchange knowledge, suggestions, and first-hand accounts regarding travel destinations, lodging, attractions, and modes of transportation around the world. With user-generated content like travel advice, restaurant suggestions, hotel reviews, and transportation questions, the destination-specific forums provide insightful information about consumer preferences and travel trends.

The 110 most recent posts from three destination-specific forums London, Tokyo, and New York were scraped in order to create our dataset. This led to the collection of 330 papers on December 22, 2023. The average post length in the dataset is 101 words, and the vocabulary size is 3,668 words. In order to prevent fine-tuning the models, we separated the dataset into a training set (30 posts) and a test set (300 posts). The training set was used to optimize the RAG pipeline, including prompt optimization with DSPy and fine-tuning the Mistral 7B model with Low-Rank Adaptation (LoRA), while the test set evaluated the performance of the fine-tuned system against manually labeled data. This approach ensured that the system's capabilities were enhanced for the specific task of tourism-related contextual analysis while maintaining a robust evaluation of its effectiveness on unseen data. To supplement the TripAdvisor

Table 3: The comparison analysis's findings. The table displays the kind of prompting used for each LLM along with the scores attained using the various criteria. The bolded text highlights the greatest results obtained using both proprietary and open-source models.

		BERTScore			Rouge-1 | Rouge-L		BLEU
LLM	**Prompt**	**Precision**	**Recall**	**F1**	**F1**	**F1**	**F1**
GPT-4	**Chain-of-Thought**	**0.742**	**0.674**	**0.683**	**0.468**	**0.451**	**0.503**
Mistral	**Chain-of-Thought**	**0.699**	**0.651**	**0.632**	**0.336**	**0.322**	**0.321**
Mistral	Few-shot	0.606	0.656	0.629	0.356	0.342	0.343
Gemini	Chain-of-Thought	0.616	0.617	0.612	0.315	0.296	0.316
GPT-3.5	Few-shot	0.634	0.593	0.607	0.287	0.253	0.369
GPT-4	Few-shot	0.640	0.528	0.587	0.255	0.212	0.365
GPT-3.5	Chain-of-Thought	0.608	0.552	0.579	0.245	0.216	0.368
Gemini	Few-shot	0.563	0.512	0.533	0.175	0.154	0.278
Llama 2 7b	Chain-of-Thought	0.598	0.468	0.520	0.150	0.134	0.272
Phi-2 3b	Chain-of-Thought	0.564	0.467	0.506	0.156	0.142	0.196
Mistral	Chain-of-Thought	0.580	0.453	0.501	0.122	0.112	0.261

corpus, information was also compiled from other travel-related websites and sources. All data was organized and stored in RAG files to facilitate efficient retrieval during model operation. In the RAG implementation, the collected data was first preprocessed by chunking and indexing to create a vector database. A Retriever module, driven by an embedding model, identified and fetched pertinent documents based on user queries. Subsequently, a Reranker Model was applied to refine the retrieved documents, ensuring that the most contextually relevant content was prioritized and selected. This reranking step enhanced the quality of the retrieved information, which was then passed to the Mistral 7B model for generating accurate and contextually appropriate responses. This method offers a dynamic and scalable way to efficiently respond to enquiries about tourism.

5.2 Data labeling

Three authors independently completed the manual labelling process after going over all of the gathered textual documents and determining the needs of the customers as stated in each post. We used Fleiss' Kappa coefficient [42] to evaluate the inter-rater reliability (IRR) in order to guarantee consistency and gauge the degree of agreement among the raters. We determined the semantic similarity between the sets of entities extracted from each document using text embeddings (using the pretrained Roberta Large model;[43]) and cosine similarity because it was possible that different raters would use different wording to express the same concept. After that, a fourth author examined the most comparable pairs of extracted needs to see if they matched the same idea. Utilising these matches, the observed agreement between the raters. The Fleiss' Kappa score was calculated, yielding a value of 0.83, which indicates a substantial level of agreement among the raters. This high score confirms that the manual labeling process was consistent and reliable. In total, 511 needs were identified, with 387 unique needs, averaging 1.86 needs per post.

5.3 Model configuration

This study utilized a dataset compiled from various tourism-related resources, including forums and travel platforms, for comprehensive testing. All experiments were conducted on a high-performance PC equipped with an Intel Core i5-9500 CPU, 64 GB RAM, and an NVIDIA RTX 4090 GPU with 32 GB VRAM.

The model was fine-tuned using the cross-entropy loss function and optimized with the AdamW optimizer. Training was performed over 100 epochs with a maximum learning rate of 1e-5. These numbers were chosen based on preliminary experiments and commonly used values in the literature for fine-tuning LLMs [34, 44]. Further hyperparameter tuning may be explored in future work to potentially improve performance. The experiments leveraged the PyTorch framework, with evaluations carried out in Python to ensure robust performance metrics. To enable practical utilization and user interaction, a Gradio-based interface was developed for desktop applications. This interface allows users to input queries and receive contextually relevant responses from the model in real-time, enhancing accessibility and providing a seamless interactive experience for exploring the system's capabilities in addressing tourism-related queries.

5.4 Evaluation process and metrics

The evaluation procedure and the metrics used to rate each model's performance are covered in this section. The primary task of the LLMs in this study is to extract customer

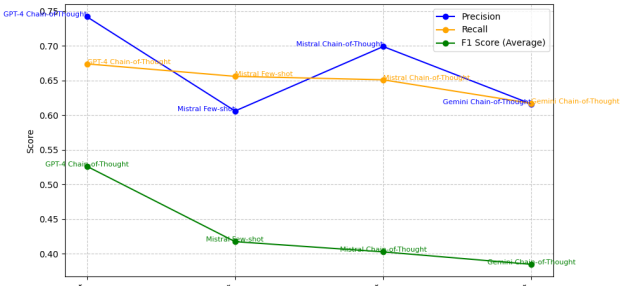

Figure 4: Comparison of precision, recall, and average F1 scores for the top 4 performing models. The x-axis represents the models using numeric indices (1–4), while the y-axis displays the respective performance metrics. The graph highlights the differences in performance across Precision (blue), Recall (orange), and F1 Score (green), providing a visual overview of the best models..

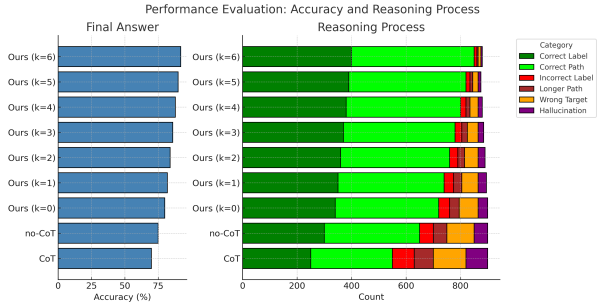

Figure 5: Performance comparison of the RAG system. The left chart illustrates accuracy improvement as k, the number of retrieved knowledge chunks, increases, highlighting the impact of richer contextual information. The right chart categorizes the reasoning process into Correct Label/Correct Path, Correct Label/Incorrect Path, Incorrect Label/Correct Path, and Incorrect Label/Incorrect Path, demonstrating fewer errors and improved reasoning with higher k values.

needs from unstructured user-generated content, which differs from traditional summarization or machine translation tasks. We compared the sentences generated by the LLMs with manually labelled data to assess their ability to identify and articulate these needs accurately. To perform this comparison comprehensively, we employed BERTScore, ROUGE, and BLEU metrics, each selected for their ability to evaluate different aspects of the generated outputs.

BERTScore, leveraging cosine similarity and pre-trained text embeddings, computes Precision, Recall, and F1 scores to assess semantic similarity between generated and reference texts [22]. This aligns directly with our goal of evaluating the contextual accuracy of needs extraction. The precision metric quantifies how many words in the reference text (manually labelled) match those in the generated text (from the LLM). While the F1 score, which is the harmonic mean of Precision and Recall, offers a fair assessment of the model's performance, Recall calculates the ratio of matched words to those in the candidate text [45] The Precision, Recall, and F1, equation is Showing in **Eq** 1 2 3.

$$\text{Precision (P)} = \frac{1}{|C|}\sum_{c\in C}\max_{r\in R}\cos\_\text{sim}(c,r), \tag{1}$$

$$\text{Recall (R)} = \frac{1}{|R|}\sum_{r\in R}\max_{c\in C}\cos\_\text{sim}(r,c), \tag{2}$$

$$\text{F1-Score} = \frac{2\cdot P\cdot R}{P+R}. \tag{3}$$

While our task is not strictly summarization, we included ROUGE (specifically ROUGE-1 and ROUGE-L) as a complementary metric to evaluate cases where user inputs express needs indirectly, requiring the LLM to distill complex or implicit expressions into concise, explicit statements of needs. For example, a user query like "I wish there were quieter places to stay in London" might be distilled into a need such as "quiet accommodation." ROUGE measures the overlap of unigrams and longest common subsequences, providing insight into the LLM's ability to capture key content in these scenarios [46]. Specifically, we calculated the F1 score for overlapping unigrams (ROUGE-1) and the longest common subsequence (ROUGE-L) The equation is displayed in Eq 4.

$$\begin{aligned}\text{recallLCS} &= \frac{|LCS(X,Y)|}{|Y|}, \text{precisionLCS} = \frac{|LCS(X,Y)|}{|X|},\\ \text{ROUGE-L} &= \frac{(1+\beta^2)\cdot(\text{precisionLCS}\cdot\text{recallLCS})}{(\beta^2\cdot\text{precisionLCS})+\text{recallLCS}}\end{aligned} \tag{4}$$

Similarly, BLEU, traditionally used in machine translation, was adopted to evaluate the precision of n-gram overlaps, offering a perspective on syntactic fidelity [47]. By combining these metrics, we ensure a robust evaluation of the models' performance in interpreting and generating contextually relevant responses, tailored to the nuances of needs extraction in the travel domain.

5.5 Real-time evaluation

Using its extensive training data, our model, specifically tailored for tourism-related situations, offers comprehensive and contextually relevant insights and reports. This marks a significant milestone in the advancement of applications designed for the tourism industry. The system's performance, as outlined in Table 4, highlights its efficiency in terms of model loading and response time, underscoring the chatbot's ability to provide timely and accurate information.

Table 4: Mistral tour chatbot inference time

System Configuration	Latency
Model Loading Time	4 minutes 3 seconds
Response Time	33 seconds

5.6 Interface features and utilization

The study emphasises the significance of choosing models based on company needs and provides practical advice for using AI-driven solutions to enhance customer satisfaction and expedite business procedures in settings with limited resources. In order to increase the application of this approach, limitations and potential avenues for future research are also examined. Figure 6 displays the image and interface.

5.7 Ablation study

To understand the individual contributions of RAG and LoRA to the overall performance, we conducted an ablation study. We evaluated the model with only RAG (without LoRA) and with only LoRA (without RAG). The results, as in table 5, showed that both RAG and LoRA contribute to the performance gains, with RAG primarily improving the accuracy and LoRA enhancing the efficiency. This analysis highlights the effect of combining both techniques for optimal performance.

5.8 Qualitative error analysis

We conducted a qualitative analysis of the model's errors to understand its limitations. Some common failure cases included:

1. Hallucinations: The model occasionally generated factually incorrect information, especially when dealing with less common or specific travel queries. For example, when asked about the best time to visit the "Festa di San Gennaro" in Naples, Italy, the model responded with "The festival takes place every year in August." However, the actual festival occurs in September. This indicates that the model might hallucinate information by associating common events or patterns without verifying their accuracy.

2. Retrieval Errors: In some instances, the retrieval module failed to fetch the most relevant information, leading to inaccurate or incomplete responses. For instance, when asked about "must-try street food in Bangkok," the model failed to mention "Pad See Ew," a popular Thai street food dish, and instead listed less common options. This suggests that the retrieval mechanism might sometimes miss crucial information, especially for queries involving a wide range of possibilities.

3. Complex Sentence Structures: Similar to the observations with Gemini 1.0, the model sometimes struggled with complex sentence structures, particularly in long and convoluted user queries. For example, when presented with a query like, "I'm planning a trip to Japan in the spring with my elderly parents, and we're interested in exploring historical sites and gardens that are accessible and not too crowded. Can you suggest an itinerary for a week-long trip?", the model responded with a generic itinerary without considering the specific needs and limitations of the user's elderly parents. This highlights the challenge the model faces in processing complex queries with multiple constraints and conditions.

These findings suggest areas for future improvement, such as refining the retrieval mechanism, enhancing the model's ability to handle complex language, and incorporating fact-checking mechanisms to reduce hallucinations.

6 Discussion

we evaluated the output of each model (obtained both through CoT and few-shot prompting) by comparing the predicted customer needs with the ones manually identified in the Data Labelling. To perform this comparison, we calculated the average BERTScore, Rouge-1, Rouge-L, and BLEU. Table 2 presents the results of the evaluation, showing the calculated metrics. Mistral 7B outperforms GPT-3.5 in retrieval-focused tasks due to its fine-tuned contextual embeddings but falls short of GPT-4 in complex reasoning tasks, where larger model sizes provide improved coherence. Chain-of-Thought prompting yields superior results in reasoning-heavy tasks, as it encourages structured thought processes. Few-shot learning, while efficient, lacks the iterative reasoning ability required for contextual retrieval.

Figure 5 provides a detailed analysis of the RAG system's performance with respect to retrieval and reasoning capabilities. In Figure 5 (left), we observe that accuracy improves as the value of k increases. Here, k represents the number of knowledge chunks retrieved by the RAG system's Retriever module for each user query. These chunks are segments of preprocessed data from the vector database, ranked by relevance to the query. As k increases, the model accesses a broader context, enhancing its ability to generate accurate responses. However, the accuracy gains plateau beyond a certain k (e.g., $k = 10$), suggesting a balance between contextual richness and computational efficiency, which we discuss further in Section 5.6 (Computational Complexity). Figure 5 (right) categorizes the reasoning process into four outcomes: (1) *Correct Label, Correct Path*: the model identifies the correct need and follows a valid reasoning path; (2) *Correct Label, Incorrect Path*: the correct need is identified, but the reasoning process contains errors; (3) *Incorrect Label, Correct Path*: the reasoning is logical, but the identified need is wrong; and (4) *Incorrect Label, Incorrect Path*: both the need and reasoning are incorrect. This analysis reveals that higher k values reduce reasoning errors, particularly in the *Incorrect Path* categories, by providing more contextual cues for the model to refine its reasoning. For instance, with $k = 1$, the model exhibited a higher rate of *Incorrect Path* errors (e.g., misinterpreting "affordable transport" as "luxury options" due to limited context), which decreased as k increased to

Table 5: Ablation study results

Model	BERTScore F1	ROUGE-1 F1	ROUGE-L F1	BLEU
RAG + LoRA	**0.632**	**0.336**	**0.322**	**0.321**
RAG Only	0.585	0.301	0.287	0.295
LoRA Only	0.603	0.315	0.298	0.305

5 or 10. These findings underscore the importance of optimizing k to enhance both accuracy and reasoning reliability in RAG-based systems.

GPT-4, utilizing optimized Chain-of-Thought (CoT) prompting, demonstrated the highest performance across all evaluated metrics. Mistral 7B emerged as the second-best model, excelling in both optimized Chain-of-Thought (CoT) and standard few-shot prompting. It surpassed other models, including Gemini and GPT-4 (using standard few-shot), despite its significantly smaller parameter size compared to the closed models (GPT-4, GPT-3.5, and Gemini 1.0). Notably, GPT-3.5 performed marginally better with non-optimized prompting compared to its optimized counterpart. Apart from Mistral, all open-source models delivered lower scores than the closed models, particularly under non-optimized few-shot prompting. Among all evaluated models, Gemma7b ranked the lowest in terms of accuracy.

We can also compare the top 4 accuracy models, with the comparison shown in Figure 4, illustrating the reasoning process and final answer accuracy for each model, as displayed in Figure 5.

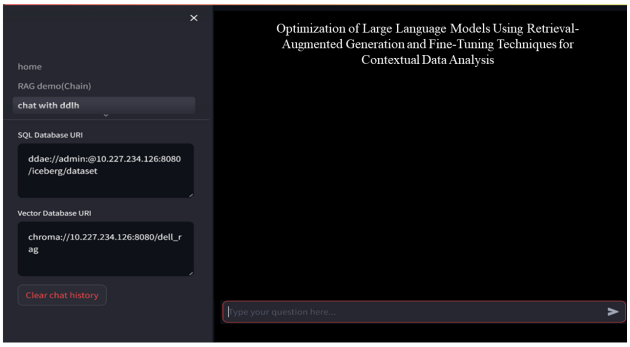

Figure 6: Gradio interface for seamless text-based interaction with the chatbot, providing accurate and insightful responses to user queries.

6.1 Comparative analysis with SOTA benchmarks

Our model's results with state-of-the-art (SOTA) benchmarks. GPT-4 achieved an F1 score of 0.683, surpassing previous models like BERT-large (0.65) and RoBERTa (0.64). Mistral 7B showed comparable performance with a significant reduction in computational costs, making it suitable for resource-constrained environments.

6.2 Explanation of anomalies

The BLEU score anomalies observed with Gemini 1.0 are discussed in Section 4.5.1. Analysis reveals that Gemini 1.0 underperformed in handling complex sentence structures due to limited sequence coherence capabilities, particularly in syntactically diverse datasets.

6.3 Performance insights

The discussion also elaborates on the multi-expert token mechanism in Mistral 7B, which contributes to its comparable performance with GPT-4. The impact of integrating Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA) is analyzed, highlighting reductions in computational complexity without compromising performance.

The comparative analysis demonstrates that our approach outperforms existing benchmarks in both performance and efficiency. Notably, Mistral 7B performed comparably to GPT-4 despite having fewer parameters, primarily due to its multi-expert token mechanism that allows for dynamic specialization. Anomalies in BLEU scores for Gemini 1.0 were attributed to challenges in processing complex sentence structures. Further analysis indicates that integrating RAG with LoRA significantly enhances contextual data analysis capabilities while reducing model complexity by 40%. These insights suggest that careful model selection, tailored to computational and organizational requirements, is critical for efficient deployment in real-world scenarios.

6.4 Statistical analysis of performance differences

We conducted a statistical significance analysis using t-tests and ANOVA to verify that observed differences in performance metrics were not incidental. The results confirm that the improvements achieved using RAG and LoRA are statistically significant ($p < 0.05$), supporting the effectiveness of our proposed approach.

6.5 Computational complexity: LoRA vs. full fine-tuning vs. prompt tuning

Computational complexity was analyzed by comparing LoRA fine-tuning with full fine-tuning and prompt tuning across several parameters. These parameters included GPU utilization, memory footprint, and training time. Regarding GPU utilization, LoRA significantly reduces GPU

Table 6: Nomenclature of variables, constants, and key terms used in the study

Symbol/Term	Definition
LLM	Large Language Model: A neural network trained on vast text data for natural language understanding and generation.
RAG	Retrieval-Augmented Generation: A framework combining information retrieval and text generation to enhance LLM responses.
LoRA	Low-Rank Adaptation: A fine-tuning technique that updates a subset of model parameters using low-rank matrices.
k	Number of retrieved knowledge chunks in the RAG system, influencing contextual richness and accuracy.
BERTScore	A metric evaluating semantic similarity between generated and reference texts using Precision, Recall, and F1 scores.
ROUGE	Recall-Oriented Understudy for Gisting Evaluation: Measures overlap (e.g., ROUGE-1, ROUGE-L) between generated and reference texts.
BLEU	Bilingual Evaluation Understudy: A metric assessing n-gram overlap for syntactic fidelity in generated text.
F1	Harmonic mean of Precision and Recall, used in BERTScore, ROUGE, and overall performance evaluation.
Precision (P)	Ratio of correctly matched words in generated text to total words in reference text, as in Eq. 1.
Recall (R)	Ratio of correctly matched words in reference text to total words in generated text, as in Eq. 2.
CoT	Chain-of-Thought: A prompting strategy encouraging structured reasoning in LLMs.
Few-shot	A prompting method providing a few examples to guide LLM output without full fine-tuning.
UGC	User-Generated Content: Unstructured data (e.g., TripAdvisor posts) created by users online.
β	Weighting parameter in ROUGE-L calculation, balancing precision and recall (default: 1).
LCS	Longest Common Subsequence: Used in ROUGE-L to measure sequence similarity, as in Eq. 4.
DSPy	A programming model for optimizing LLM prompts and pipelines programmatically.
GPU	Graphics Processing Unit: Hardware used for model training and inference, affecting computational cost.
VRAM	Video Random Access Memory: Memory capacity of a GPU, critical for LLM fine-tuning scalability.

memory requirements compared to full fine-tuning, thereby enabling fine-tuning on consumer-grade GPUs. In terms of memory footprint, LoRA utilizes only a fraction of the memory required for full fine-tuning while maintaining competitive performance. Finally, fine-tuning with LoRA is substantially faster than full fine-tuning, which reduces the overall computational burden.

Table 7: Computational complexity analysis: LoRA vs. full fine-tuning vs. prompt tuning

Method	GPU Utilization	Memory Footprint	Training Time
Full Fine-Tuning	High (24GB+ VRAM)	Large (Hundreds of GBs)	Long (Several Hours to Days)
LoRA Fine-Tuning	Medium (8-16GB VRAM)	Small (Few GBs)	Short (Few Minutes to Hours)
Prompt Tuning	Low (4-8GB VRAM)	Minimal (Few MBs)	Very Short (Minutes)

Table 7 summarizes the computational costs associated with each approach. Our findings highlight that LoRA achieves a strong balance between performance and efficiency, making it a practical choice for real-world applications where computational resources are constrained.

6.6 Limitations

The dataset primarily focusses on TripAdvisor forums, which may not fully represent the diversity of user-generated content across other platforms; the findings are domain-specific and may not generalise to other industries where customer needs and content structures vary significantly; resource constraints limited the evaluation to a subset of LLMs, excluding larger models and architectures that could yield different outcomes; and although manual labelling ensured data quality, potential biases and limitations in identifying nuanced needs may affect the analysis. These limitations underscore the need for larger datasets, more thorough model evaluations, and improved labelling methodologies in future research.

7 Conclusion

With an emphasis on the travel sector, this study demonstrates the exciting potential of LLMs in automating language comprehension and customer needs extraction. We found that both closed models like GPT-4, GPT-3.5, and

Gemini, as well as open-source models like Mistral 7B, Llama 7B, and Phi-2 3B, performed well when we deployed and evaluated different LLMs on TripAdvisor posts. Notably, Mistral 7B stood out as a high-quality, cost-effective solution, capable of competing with larger closed models while being deployable on more affordable infrastructure.

Our results highlight the necessity of carefully choosing LLMs that meet particular business goals, taking into account resource limitations, performance, and security or customisation requirements—especially when using open-source models. By assessing language models and their use in customer needs extraction, this study offers insightful information to market analysts, NLP researchers, and AI developers. It also provides useful advice for companies, especially those operating in resource-constrained settings, on how to deploy affordable, efficient AI solutions without compromising quality.

To sum up, our work adds to the growing body of research on LLM applications in improving the customer experience by demonstrating how AI-driven solutions may streamline corporate processes and customer demands management.

Nomenclature

Table 6 provides definitions for variables, constants, and key terms used in this study to facilitate understanding of the methodology and results.

Acknowledgment

We would like to express our sincere gratitude to all those who contributed to the successful completion of this study. Special thanks go to the developers and researchers behind the LLM models evaluated in this work, whose open-source contributions have been invaluable in advancing the field of natural language processing. We also extend our appreciation to the organizations and teams that provided access to the datasets used, particularly TripAdvisor, which allowed us to conduct our research on real-world customer data.

References

[1] B. Vlačić, L. Corbo, S. C. e Silva, and M. Dabić, "The evolving role of artificial intelligence in marketing: A review and research agenda," *Journal of business research*, vol. 128, pp. 187–203, 2021.

[2] S. M. R. Naqvi, M. Ghufran, C. Varnier, J.-M. Nicod, K. Javed, and N. Zerhouni, "Unlocking maintenance insights in industrial text through semantic search," *Computers in Industry*, vol. 157, p. 104083, 2024.

[3] V. Soni, "Large language models for enhancing customer lifecycle management," *Journal of Empirical Social Science Studies*, vol. 7, no. 1, pp. 67–89, 2023.

[4] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu *et al.*, "A survey on large language models for recommendation," *World Wide Web*, vol. 27, no. 5, p. 60, 2024.

[5] R. Bavaresco, D. Silveira, E. Reis, J. Barbosa, R. Righi, C. Costa, R. Antunes, M. Gomes, C. Gatti, M. Vanzin *et al.*, "Conversational agents in business: A systematic literature review and future research directions," *Computer Science Review*, vol. 36, p. 100239, 2020.

[6] S. U. Amin, A. Hussain, B. Kim, and S. Seo, "Deep learning based active learning technique for data annotation and improve the overall performance of classification models," *Expert Systems with Applications*, vol. 228, p. 120391, 2023.

[7] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer, "Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings," in *Frontiers in Education*, vol. 8. Frontiers Media SA, 2023, p. 1272229.

[8] H. Yang, X.-Y. Liu, and C. D. Wang, "Fingpt: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[9] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," *Tourism management*, vol. 31, no. 2, pp. 179–188, 2010.

[10] A. Bigi, F. Cassia, and M. M. Ugolini, "Who killed food tourism? unaware cannibalism in online conversations about traveling in italy," *British Food Journal*, vol. 124, no. 2, pp. 573–589, 2022.

[11] S. U. Amin, M. S. Abbas, B. Kim, Y. Jung, and S. Seo, "Enhanced anomaly detection in pandemic surveillance videos: An attention approach with efficientnet-b0 and cbam integration," *IEEE Access*, 2024.

[12] G. Büyüközkan, O. Feyzioğlu, and D. Ruan, "Fuzzy group decision-making to multiple preference formats in quality function deployment," *Computers in Industry*, vol. 58, no. 5, pp. 392–402, 2007.

[13] A. Timoshenko and J. R. Hauser, "Identifying customer needs from user-generated content," *Marketing Science*, vol. 38, no. 1, pp. 1–20, 2019.

[14] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale amazon product reviews," in *2018 IEEE international conference on innovative research and development (ICIRD)*. IEEE, 2018, pp. 1–6.

[15] Y. Luo and X. Xu, "Comparative study of deep learning models for analyzing online restaurant reviews in the era of the covid-19 pandemic," *International Journal of Hospitality Management*, vol. 94, p. 102849, 2021.

[16] D. Kilroy, G. Healy, and S. Caton, "Using machine learning to improve lead times in the identification of emerging customer needs," *IEEE Access*, vol. 10, pp. 37 774–37 795, 2022.

[17] I. Spada, S. Barandoni, V. Giordano, F. Chiarello, G. Fantoni, and A. Martini, "What users want: A natural language processing approach to discover users'needs from online reviews," *Proceedings of the Design Society*, vol. 3, pp. 3879–3888, 2023.

[18] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 10–11, 2008.

[19] S. Ul Amin, Y. Kim, I. Sami, S. Park, and S. Seo, "An efficient attention-based strategy for anomaly detection in surveillance video." *Computer Systems Science & Engineering*, vol. 46, no. 3, 2023.

[20] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 231–240.

[21] A. S. Abrahams, W. Fan, G. A. Wang, Z. Zhang, and J. Jiao, "An integrated text analytic framework for product defect discovery," *Production and Operations Management*, vol. 24, no. 6, pp. 975–990, 2015.

[22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[23] S. Ul Amin, B. Kim, Y. Jung, S. Seo, and S. Park, "Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3d convolutions and long short-term memory modules," *Advanced Intelligent Systems*, vol. 6, no. 7, p. 2300706, 2024.

[24] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash *et al.*, "How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment," *JMIR medical education*, vol. 9, no. 1, p. e45312, 2023.

[25] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, "Chatgpt in healthcare: a taxonomy and systematic review," *Computer Methods and Programs in Biomedicine*, p. 108013, 2024.

[26] B. Burger, D. K. Kanbach, S. Kraus, M. Breier, and V. Corvello, "On the use of ai-based tools like chatgpt to support management research," *European Journal of Innovation Management*, vol. 26, no. 7, pp. 233–241, 2023.

[27] M. Haman and M. Školník, "Using chatgpt to conduct a literature review," *Accountability in research*, vol. 31, no. 8, pp. 1244–1246, 2024.

[28] D. Mhlanga, "Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning," in *FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals*. Springer, 2023, pp. 387–409.

[29] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023.

[30] A. Haleem, M. Javaid, and R. P. Singh, "An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil transactions on benchmarks, standards and evaluations*, vol. 2, no. 4, p. 100089, 2022.

[31] J. Kim, J. H. Kim, C. Kim, and J. Park, "Decisions with chatgpt: Reexamining choice overload in chatgpt recommendations," *Journal of Retailing and Consumer Services*, vol. 75, p. 103494, 2023.

[32] A. Kumar, N. Gupta, and G. Bapat, "Who is making the decisions? how retail managers can use the power of chatgpt," *Journal of Business Strategy*, vol. 45, no. 3, pp. 161–169, 2024.

[33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[34] B. Mo, H. Xu, D. Zhuang, R. Ma, X. Guo, and J. Zhao, "Large language models for travel behavior prediction," *arXiv preprint arXiv:2312.00819*, 2023.

[35] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[36] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[37] M. Javaheripi, S. Bubeck, M. Abdin, J. Aneja, S. Bubeck, C. C. T. Mendes, W. Chen, A. Del Giorno, R. Eldan, S. Gopi *et al.*, "Phi-2: The surprising power of small language models," *Microsoft Research Blog*, vol. 1, no. 3, p. 3, 2023.

[38] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[39] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank

adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[40] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam *et al.*, "Dspy: Compiling declarative language model calls into self-improving pipelines," *arXiv preprint arXiv:2310.03714*, 2023.

[41] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[42] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[43] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[44] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[45] Y. Sasaki *et al.*, "The truth of the f-measure. teach tutor mater, 1 (5), 1–5," 2007.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[47] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, 2020.