

Automating Customer Needs Analysis: A Comparative Study of Large Language Models in the Travel Industry

Simone Barandoni

Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3, Pisa, Italy
simone.barandoni@phd.unipi.it
0009-0008-4454-3388

Filippo Chiarello

Department of DESTEC
University of Pisa
Largo L. Lazzarino, Pisa, Italy
filippo.chiarello@unipi.it
0000-0001-9857-0287

Lorenzo Cascone

Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3, Pisa, Italy
l.cascone2@studenti.unipi.it
0009-0001-8707-6360

Emiliano Marrale

Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3, Pisa, Italy
e.marrale@studenti.unipi.it
0009-0005-6727-3753

Salvatore Puccio

Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3, Pisa, Italy
s.puccio1@studenti.unipi.it
0009-0000-7577-8149

Abstract

In the rapidly evolving landscape of Natural Language Processing (NLP), Large Language Models (LLMs) have emerged as powerful tools for many tasks, such as extracting valuable insights from vast amounts of textual data. In this study, we conduct a comparative analysis of LLMs for the extraction of travel customer needs from TripAdvisor posts. Leveraging a diverse range of models, including both open-source and proprietary ones such as GPT-4 and Gemini, we aim to elucidate their strengths and weaknesses in this specialized domain. Through an evaluation process involving metrics such as BERTScore, ROUGE, and BLEU, we assess the performance of each model in accurately identifying and summarizing customer needs. Our findings highlight the efficacy of open-source LLMs, particularly Mistral 7B, in achieving comparable performance to larger closed models while offering affordability and customization benefits. Additionally, we underscore the importance of considering factors such as model size, resource requirements, and performance metrics when selecting the most suitable LLM for customer needs analysis tasks. Overall, this study contributes valuable insights for businesses seeking to leverage advanced NLP techniques to enhance customer experience and drive operational efficiency in the travel industry.

1. Introduction

In the era of vast digital content, understanding customer needs and preferences has become pivotal for businesses across various sectors. The proliferation of social media and online review platforms has provided a wealth of valuable information regarding consumer experiences and sentiments (Vlačić et al., 2021). Extracting and analysing these insights manually, however, can be a daunting and time-consuming task (Zhou et al., 2020). Over the years, the adoption of Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques has facilitated the automation of these processes, especially when working with noisy and unstructured text (Naqvi et al., 2024). However, with the diffusion of Large Language Models (LLMs), this capability can be enhanced even further, promising a new era of automated language understanding and customer needs extraction (Soni, 2023). The utilization of conversational agents powered by LLMs has the potential to reshape the customer needs management area, offering unparalleled efficiency and scalability. By harnessing the power of natural language understanding, these agents can interpret and respond to customer inquiries, identify solutions for new product development, and personalize recommendations (Wu et al., 2023). This paradigm shift might not only enhance user experience, but also optimize operational workflows for businesses, leading to improved customer satisfaction and loyalty.

In recent years, the business sector is experiencing a rapid growth in utilizing conversational agents to enhance the quality and efficiency of the processes (Bavaresco et al., 2020). The selection of the most suitable LLM to build an AI-based chatbot is a critical decision for enterprises, considering both performance and resource requirements. Very large models typically demand substantial computing capabilities, which may pose challenges for some businesses in terms of affordability and maintenance. The current landscape of conversational AI is dominated by proprietary models, such as *GPT-4*, *Gemini*, and *Claude*, with GPT-4 being the state-of-the-art in many Natural Language Processing tasks at the time of the analysis (Hackl et al., 2023), which were performed in March 2024. These models can be utilized for a fee, through APIs. On the other hand, open-source generative AI models, such as Llama 2, are demonstrating good advancements in terms of performance, even if they have yet to reach the level of GPT-4 in certain tasks or specific domains (Ray, 2024). These open LLMs enhance transparency by openly sharing their training data and architectures. Furthermore, considering the possibility of fine-tuning a LLM to enhance its performance within a specific context, open-source models allow users to perform this process on their own infrastructure, providing a level of customization and data security that is not achievable when utilizing a proprietary API (Yang et al., 2023). In the domain of customer management, this is crucial to ensure the utmost protection for any sensitive customer data. There is a vast array of open LLMs available for businesses to leverage in their operations. Therefore, the optimal choice is a non-trivial problem. To the best of our knowledge, no comparisons between LLMs exist in literature regarding the task of needs analysis.

In this paper, we present a comprehensive comparative analysis of Large Language Models for the extraction of travel customer needs from TripAdvisor posts. We decided to focus on the travel industry for two reasons: first, because of the great amount of unstructured data that can be easily retrieved about it (e.g., travellers reviews, comments, posts, and requests shared publicly on social media and forums); second, due to the importance of performing a customer needs analysis for enterprises related to the tourism sector, such as hotels, restaurants, transportation companies, travel agencies, etc. These actors may exploit the outcomes of this study to make an informed decision about the selection of the most suitable LLM for building a conversational agent able to support them in identifying customer needs and, consequently, gaining insights about relevant design requirements (Büyükožkan et al., 2007). By employing a process of data collection, labelling, and computational analysis, we seek to elucidate which LLMs excel in this specialized domain. The models considered in the analysis are the following: *Llama 2 7B*, *Llama 2 13B*, *Phi-2*, *Mistral*, *Mixtral*, *Gemma*, *Gemini 1.0*, *GPT-3.5*, and *GPT-4*. This selection allowed us to explore not only the performance differences between open-source (Llama 2, Phi-2, Mistral, Mixtral, Gemma) and proprietary models (GPT-3.5, GPT-4, Gemini) but also the impact of model size and type of prompting on the quality of results.

By needs we intend both essential requirements and preferences that a service or a place should fulfil to satisfy the customer. In the data we collected (presented in Section 3.2), users share their requests about a certain place, to receive feedback and suggestions, usually expressing one or more needs. Even if the needs are not explicitly written in the text, they can be extrapolated from it analysing its keywords. Therefore, we first used a thematic analysis approach to manually annotate the dataset: we read the collected posts and extracted recurring themes to identify customer needs. Then, we employed the mentioned LLMs to perform the same task and compared their outcomes with the manual labelling. Figure 1 presents an example of post and corresponding customer needs identified through thematic analysis.

TripAdvisor Post	Customer Needs
<p>Hi all I will be arriving in August ... I fancy a hotel with a view of the beautiful Tower Bridge. I am also interested in visiting Borough Market and having great fish and chips (not necessarily in Borough Market). Would like to also absorb the beautiful architecture and wander around a bit before calling it a night. Not sure how to incorporate my "wants" into an efficient itinerary. Thank you for any guidance. Mary</p>	<ul style="list-style-type: none"> <li data-bbox="938 1691 1171 1736">• Hotel with a view of the Tower Bridge <li data-bbox="938 1736 1171 1780">• Visit to Borough Market <li data-bbox="938 1780 1171 1825">• Fish and chips recommendations <li data-bbox="938 1825 1171 1870">• Sightseeing the architecture <li data-bbox="938 1870 1171 1890">• Efficient itinerary planning

Figure 1. Example of a TripAdvisor post and related customer needs manually extracted through thematic analysis. Some of the needs, such as *Hotel with a view of the Tower Bridge* are explicit in the original text. Others, such as *Efficient itinerary planning*, are more implicit.

2. Related Works

The related works of this study can be organized in two areas: the first, addressed in Section 2.1, revolves around machine learning for customer needs analysis, discussing the data sources and the methodologies employed; the second, in Section 2.2, concerns the utilization of Large Language Models in the domain of Customer Experience Management.

2.1 Utilizing User-Generated Content and Machine Learning for customer needs analysis

Researchers and practitioners have exploited various sources of data and developed various methodologies to automatically analyse user needs and to provide support in business strategy (Timoshenko and Hauser, 2019).

User-Generated Content (UGC) is the main source of data utilized for these purposes (Kilroy et al., 2022). UGC refers to any form of content, such as text, images, videos, or reviews, which is created and shared online by users of a platform or website (Spada et al., 2023). This kind of data offers unfiltered insights into the authentic experiences and opinions of customers and individuals, without being subject to modification or censorship by traditional media outlets (Krumm et al., 2008). Social media posts and product reviews are the most utilized UGC for customer needs analysis (Kilroy et al., 2022). Social media such as *X* (formerly *Twitter*), *Facebook*, *Reddit*, often contains irrelevant posts, unlike product reviews which are more focused. However, for researching future needs, social media may be more suitable as their users tend to discuss emerging trends (Kilroy et al., 2022). In this work, we utilized the TripAdvisor forums to collect textual documents. Being a question-and-answer social media specifically focused on travels, we considered appropriate for our purposes (i.e., testing LLMs to predict travel customer needs). In section 3.2 we provided a better description of the TripAdvisor forums and of the reasons for its selection as a source of UGC.

Regarding the methodological aspects, these usually depend on the kind of data used and on the specific needs to be identified. Researchers have exploited both traditional NLP techniques based on lexicons and rules (Ding et al., 2008; Abrahams et al., 2015; Spada et al., 2023) and machine learning approaches: Zhou et al. (2020) employed Latent Dirichlet Allocation (LDA), Sentiment Analysis, and a Kano model to analyse customer needs for product ecosystems; Timoshenko and Hauser (2019) trained a convolutional neural network (CNN) and developed a machine-human hybrid approach to extract customer needs from UGC; Haque et al. (2018), applied Sentiment Analysis to a set of Amazon reviews; Bigi et al. (2022) collected a series of questions and answers from TripAdvisor forums and analysed them through a Bayesian machine-learning technique to investigate the relevance of food and drink products in attracting visitors. In recent times, it has become evident that the most effective models across various NLP tasks and domains are those rooted in deep learning methodologies. Luo et al. (2021) corroborated this by highlighting that deep learning models outperform other machine learning algorithms in customer needs analysis as well, in a study which involved analysing Yelp's restaurant reviews. This trend extends to the latest advancements in Generative AI and LLMs. For this reason, we focused our study on Large Language Models.

2.2 Large Language Models and Customer Experience Management

The recent diffusion of conversational agents based on LLMs, particularly *ChatGPT*, has fuelled the scientific research about applications and potential impact of these technologies across various domains, such as medicine and healthcare (Gilson et al., 2023; Li et al., 2024), scientific research (Burger et al., 2023), education (Haman and Školnik, 2023), (Mhlanga, 2023; Lo, 2023), and many others. Customer experience and feedback management are two of the areas which might be revolutionized through an implementation of such AI-powered systems. ChatGPT and similar conversational AI might be utilized to provide improved consumer assistance, by responding to queries in a better way than traditional chatbots, recommending products, and to analyse customer needs (Haleem et al., 2022). Researchers have explored various aspects of employing these technologies to better understand and fulfil customer requirements. Kim et al. (2023) demonstrated the positive impact of implementing ChatGPT into recommender systems in terms of consumer satisfaction. Kumar et al. (2023) argued that, despite the current limitations, such as difficulties in understanding non-standard languages, ChatGPT will become an

important asset for retailers, as a tool to support the processes of consumer needs analysis. Trichopoulos et al. (2023) proposed a fine-tuned version of GPT-4 to be used as a recommender system for museums. Soni (2023) explored the use of LLMs in customer lifecycle management and observed their positive impact on improving and optimizing lead identification, audience segmentation, marketing strategies, sales support, and post-purchase engagement. Considering the travel domain, the findings of Mo et al. (2023) suggest that LLMs have the potential to be effective tools for predicting travel behaviour, offering competitive performance and the ability to provide explanations for their predictions.

3. Methodology

The methodology of this study encompasses a series of phases, summarized in Figure 2. We started by selecting the LLMs to be assessed and collecting the data for the analysis (a corpus of TripAdvisor posts). Then, we manually labelled the data, identifying the user needs expressed in the texts. The objective was using the models to automatically extract the needs from the documents, and then compare the outcomes with the manual labelling to measure their performances. We designed an initial prompt and deployed each LLM performing a prompt optimization (a process aimed to adapt the instructions provided to the model to its characteristics, to improve the quality of the responses), executing the extraction of the needs, and cleaning the results. Finally, we evaluated the performance of each model by comparing the predicted customer needs with those identified in the manual labelling.

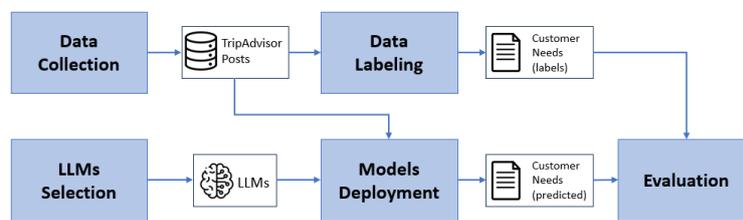


Figure 2. Workflow of the methodology employed to compare the Large Language Models for the task of travel customer needs extraction from TripAdvisor posts.

3.1 LLMs Selection

Table 1 demonstrates the models we employed in this study. We selected a series of open-source LLMs, except for Gemini 1.0 Pro, GPT-3.5 and GPT-4.

Model	Parameters	Proprietary	Reference
Gemini 1.0 Pro	Not available	Google	Team et al., 2023
Gemma 7B	7 billion	Google (open-source)	Team et al., 2024
GPT-3.5	175 billion	OpenAI	Brown et al., 2020
GPT-4	Not available	OpenAI	Achiam et al., 2023
Llama 2 7B Chat	7 billion	Meta (open-source)	Touvron et al., 2023
Llama 2 13B Chat	13 billion	Meta (open-source)	Touvron et al., 2023
Mistral 7B Instruct	7 billion	Mistral AI (open-source)	Jiang et al., 2023
Mixtral 8x7B*	46.7 billion	Mistral AI (open-source)	Jiang et al., 2024
Phi-2	2.7 billion	Microsoft (open-source)	Javaheripi et al., 2023

Table 1. List of the LLMs utilized in the analysis. For each model we present the version, the number of parameters, the enterprise or the institute which released it, and the reference. *Due to the large computational requirements of Mixtral 8x7B, we were forced to employ its smallest version (2 Bit quantization).

Gemini 1.0 is a multimodal model property of Google, available in three versions: Ultra, Pro, and Nano. GPT-3.5 and GPT-4 are the two LLMs published by OpenAI. The first is publicly available for free through ChatGPT, while the second requires a paid subscription. GPT-3.5 has 175 billion parameters (Brown et al., 2020), while for GPT-4 and Gemini 1.0 this information is not available. Llama 2 is a family of pretrained open-source Large Language Models, ranging from 7 to 70 billion parameters (Touvron et al., 2023), released by Meta. This family comprehends a series of models which are optimized for dialogues, the Llama 2-Chat ones. We chose to use the two smaller models, i.e. Llama 2-Chat 7B and Llama 2-Chat 13B, due to the very large computational resources required to deploy bigger models. We assumed that this is a limitation that small and medium enterprises might face, since not all the organizations can rely on a powerful infrastructure of computer systems, able to run the biggest LLMs. Gemma is a family of open-source LLMs released by Google, much smaller than Gemini (Team et al., 2023). It can be utilized in the two variants of two billion and seven billion parameters (Team et al., 2024). We selected Gemma 7B for our analysis. Phi-2 is an open LLM built by Microsoft Research. It is a smaller model, with its 2.7 billion parameters (Javaheripi et al., 2023). Mistral 7B and Mixtral 8x7B are two open models released by Mistral AI. The first is a seven billion parameters LLM (Jiang et al., 2023). The second, considered the most powerful by the developers, has the same architecture, but it is also based on layers “composed of 8 feedforward blocks (i.e. experts). For every token, at each layer, a router network selects two experts to process the current state and combine their outputs.” (Jiang et al., 2024).

3.2 Data Collection

We decided to exploit the TripAdvisor Forum as a source of data. This is an online platform hosted by TripAdvisor, the most popular website about traveling, according to Xiang and Gretzel (2010). It serves as a community-driven space where travellers can seek and share information, advice, and experiences related to various destinations, accommodations, attractions, and transportation options worldwide. The platform provides a series of destination-specific forums: each of them rely on contributions from users who share their travel experiences, ask questions, and provide answers to queries posted by other users regarding specific cities, regions, and countries around the world. The content shared by the users covers a wide range of topics related to travels, including accommodation reviews, restaurant recommendations, transportation inquiries, and general travel tips. Therefore, it can be utilized for analysis purposes to gain valuable insights into travel trends and consumer preferences (Gal-Tzur et al., 2018). Due to the structure of the website, which is organized in many independent forums, we had to select which one to scrape. We collected the 110 most recently published posts from three different forums of the platform, London, Tokyo, and New York, for a total of 330 documents. We executed this web-scraping operation on 22/12/2023. The collected corpus of texts has a vocabulary size of 3,668 words, and the posts are composed of 101 words, on average. We randomly divided this dataset into a training set (30 posts) and test set (300 posts). The training set was very small because we did not aim at training or fine tuning the models. We took this decision to ensure that the evaluation of the models remained consistent across the board. Our objective was to assess the raw capabilities of each model without any adjustments or enhancements through fine-tuning. However, we acknowledge that fine-tuning these models could potentially improve their performance, especially in specialized domains like customer needs analysis. We utilized the training set only in the Prompt Optimization phase, where a small number of labelled documents was necessary, as explained in section 3.4.1.

3.3 Data Labelling

Three authors performed the manual labelling phase separately, by reading all the textual documents collected and identifying the customer needs expressed in each text. We decided to measure the inter-rater reliability (IRR) using Fleiss' kappa coefficient (Fleiss et al., 1971), to quantify the level of agreement among the three raters and ensure the consistency of the manual labelling process. It may have happened that different raters would write the same concept with different words. Therefore, to compare the three different annotations and measure the IRR, we calculated the semantic similarity between the three sets of entities extracted in each document through text embeddings (by using the pretrained embedding model Roberta Large; Liu et al., 2019) and cosine similarity. Then, a fourth author has examined all the pairs of needs extracted by different authors which maximized the

similarity, and determined whether they were corresponding to a match (i.e., the two entities were expressing the same concept). These were considered to calculate the observed agreements between the three raters. We measured the IRR through Fleiss' Kappa, obtaining a score of 0.83. This high result indicates a substantial level of agreement among the raters; therefore, we considered the manual labelling as consistent and reliable. We identified a total of 511 needs (387 unique ones), and 1.86 need per post.

3.4 Models Deployment

To deploy the selected LLMs we relied on a machine with the following hardware specifications:

- Chip Apple M2 Max with 30 core GPU and 12 core CPU.
- 32 GB of unified ram.
- 1TB SSD.

Furthermore, we utilized GPT-4, GPT-3.5, and Gemini through OpenAI's and Google's API, respectively. We designed an initial prompt to be provided to the models. Prompting refers to the technique of describing specific instructions to a language model to guide its generation of text or response to a given input. It allows teaching LMs to perform new tasks (Kojima et al., 2022). In zero-shot prompting, the prompt is provided without any example. In few-shot prompting, the instructions also include a small number of examples of the desired output. This typically enhances the performance of a model (Brown et al., 2020). Figure 3 demonstrates the initial few-shot prompt created: it presents a series of instructions on the task and three examples derived from the training set. We chose the three examples randomly but avoiding duplicated needs among them.

```
""You are a customer service expert. I am going to provide you a TripAdvisor post about travels, you should identify the needs that the user expressed in the text. Needs should be expressed as short sentences. The answer should be generated as a list of needs separated by a comma. Avoid any unnecessary text or comment. Here are some examples:  
POST: I've come across a Cyber Sale announced by the Sonesta group (used to be Affinia) for their three hotels in NYC. The sale goes through April 2024 and looks like the rates are refundable but you must stay at least 2 nights. Cheapest dates are January and February as usual but over the Thanksgiving week / few days it's not too bad either. Any other good offers at other hotels are welcome to be posted in this thread as well.  
ANSWER: Affordable accommodations, Accommodations advices  
POST: We are here in NYC; looking for a last minute show to see on or off Broadway on Thanksgiving day. Any recommendations sites to purchase?  
ANSWER: Broadway musical tickets retrieving  
POST: I am traveling from New York to Dc this Sunday. Amtrak tickets are really expensive due to the holiday and waiting last minute. Please recommend any discounts or other safe options for travel. Thank you.  
ANSWER: Cheap airline travelling tips and suggestions""
```

Figure 3. Initial few-shot prompt designed for the task of travel customer needs extraction. The instructions are highlighted in bold, while the rest of the text corresponds to the three examples.

Chain-of-Thought (CoT) is a further prompting technique which involves generating intermediate reasoning steps. It has been demonstrated that generating a chain of thought “significantly improves the ability of large language models to perform complex reasoning” (Wei et al., 2022). In this study, we decided to assess two forms of prompting: standard few-shot and optimized Chain-of-Thought. For the first, we used the initial prompt (Figure 3) for all the models: we iterated over the three hundred posts of the training set to let the LLMs extract the needs. All the models were deployed setting the *temperature* parameter to 0.5. For the latter, we also performed a prompt optimization, which means adapting the prompt to each LLM to improve the quality of the generated answers. We did not perform the optimization step on the initial few-shot prompt to be able to measure the reliability of this phase by comparing the results obtained through an optimized prompt with those obtained through a non-optimized one.

3.4.1 Prompt Optimization

Optimizing the instructions provided to a specific LLM typically comprehends designing a series of prompts through manual trial and error (Khattab et al., 2023). This makes the prompting delicate and unscalable, especially when working with more than one LLM at once: a prompt which has been manually optimized for a model might not be adapted for other LLMs. Therefore, we decided to employ a more systematic methodology to optimize the CoT prompting. We relied on DSPy (Khattab et al., 2023), a programming model designed to build and optimize LM pipelines by abstracting them as text transformation graphs and automating the optimization process through parameterized declarative modules and teleprompters. Its employment allowed us to reduce the reliance on manually crafted prompt templates. DSPy comes with a compiler that optimizes a model’s pipeline to maximize a given metric. More specifically, we used the *BootstrapFewShotWithRandomSearch* DSPy optimizer, with *max_bootstrapped_demos=3*, *max_labeled_demos=3*, *num_candidate_programs=10*, *num_threads=8* as parameters, and the *ChainOfThought* module. For each LLM employed, we provided the compiler with the initial few-shot prompt, the thirty labelled posts of the training set, and *BERTScore* as validation metric (section 3.5 provides an explanation of this metric). The DSPy optimizer used the examples from the training set to self-generate reasonings and to craft a CoT prompt that would optimize the output over the given metric. Then, the optimized prompts were used to interrogate all the models to extract the needs from the posts of the test set.

3.4.2 Results Cleaning

We screened the output produced by each model with both prompting approaches to check whether they generated a clean answer (containing the list of identified customer needs only, following the instruction provided in the prompt). The Gemma, Llama 2, Mixtral, and Phi-2 models happened to add additional text in some of the responses, usually to introduce the list of extracted needs (e.g., “*Sure, these are the needs: ...*”). We eliminated all the additional texts through regular expressions to consider only the needs in the Evaluation phase. Furthermore, we deleted symbols and numbers from the results’ texts.

3.5 Evaluation

This section describes the evaluation process and the metrics chosen to measure the performance of each model. We aimed at comparing the texts generated by the LLMs with the manual labelling. We performed this comparison through the following metrics: BERTScore, ROUGE, and BLEU.

BERTScore is an automatic metric for evaluating text generation which computes Precision, Recall, and F1 score by leveraging pretrained text embeddings and cosine similarity (Zhang et al., 2019). Precision expresses the ratio of matching words in the candidates (in our case, the generated texts of a LLM) to those in the reference (the manual labelling); Recall calculates the ratio of matching words in the reference to those in the candidate; F1 score, a harmonic mean of Precision and Recall, offers a balanced appraisal of a model’s effectiveness (Sasaki, 2007). We utilized the *deberta-xlarge-mnli* pretrained embedding model. At the time of the analysis, this was the best model for running BERTScore in terms of correlation with human evaluation, according to the authors of BERTScore (Wu, 2020). *DeBerta* is a model architecture proposed to enhance existing pre-trained neural language models, particularly *BERT* (Devlin et al., 2018) and *RoBERTa* (Liu et al., 2019), by introducing two novel methodologies, *Disentangled attention mechanism* and *Enhanced mask decoder* (He et al., 2020).

ROUGE is a well-established evaluation metric, designed for text summarization tasks. It counts “the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans” (Lin, 2004). In this study, the task performed by LLMs is not precisely a summarization, nor a machine translation, so we first rely on the BERTScore metrics. However, in some cases we can also interpret it also as summarization: it can happen that users did not explicitly write a need in a TripAdvisor post, but expressed it through longer sentences, that LLMs should be able to summarize in a few words. Therefore, we decided to calculate the ROUGE metric as well. More specifically, we measured the F1 score of the number of overlapping unigrams (ROUGE-1) and of the longest common subsequence (ROUGE-L). Furthermore, we also considered the BLEU metric, which is usually utilized in machine translation to measure

the quality of generated texts by comparing it with the reference ones through “a weighted geometric mean of n-gram precision scores” (Celikyilmaz et al., 2020). This holistic approach ensures a comprehensive evaluation of the models' performance.

4. Results

As presented in section 3.5, 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.

LLM	Prompt	BERTScore			Rouge-1	Rouge-L	BLEU
		Precision	Recall	F1	F1	F1	
GPT-4	Chain-of-Thought	0.702	0.674	0.683	0.468	0.451	0.503
Mistral	Chain-of-Thought	0.629	0.651	0.632	0.336	0.322	0.321
Mistral	Few-shot	0.616	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	0.548	0.587	0.255	0.212	0.365
GPT-3.5	Chain-of-Thought	0.618	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
Mixtral	Chain-of-Thought	0.580	0.453	0.501	0.122	0.112	0.261
Llama 2 13b	Chain-of-Thought	0.618	0.431	0.501	0.126	0.108	0.289
LLama 2 13b	Few-shot	0.600	0.433	0.499	0.128	0.107	0.284
Mixtral	Few-shot	0.545	0.429	0.477	0.082	0.074	0.218
LLama 2 7b	Few-shot	0.567	0.417	0.477	0.13	0.108	0.281
Phi-2 3b	Few-shot	0.542	0.399	0.456	0.092	0.075	0.212
Gemma7b	Few-shot	0.537	0.343	0.414	0.067	0.055	0.226
Gemma7b	Chain-of-Thought	0.541	0.306	0.385	0.048	0.039	0.198

Table 2. Results of the comparative analysis. For each LLM, the table presents the type of prompting employed and the scores obtained through the different metrics. The highest achieved values with proprietary and open-source models are highlighted in bold.

GPT-4, executed through optimized Chain-of-Thought prompting, has achieved the highest score in all the metrics employed. Mistral 7B is the second-best model, both with optimized CoT and non-optimized few-shot prompting: it outperformed all the other LLMs, Gemini and GPT-4 (standard few-shot) included. Despite being much smaller than the closed models assessed (GPT-4, GPT-3.5, and Gemini 1.0) in terms of numbers of parameters, it achieved a remarkably high quality of the responses. Surprisingly, GPT-3.5 achieved slightly better results through non-optimized than through optimized prompting. Except for Mistral, all the open-source models obtained lower

scores than the closed ones, especially through non optimized few-shot prompting. Gemma7b was the worst model in terms of accuracy.

In terms of selecting the best model to perform travel customer analysis, we concluded that the best solution is Mistral 7B. It provided an extremely high quality of responses, the same of GPT-4, and its small size (seven billion parameters) allows it to be deployed on affordable machines. Furthermore, being open-source, it enables an elevated level of customization and security, as explained in the Introduction. Among the other open-source models, we observed satisfactory results also in Llama 2 7B and Phi-2 3B. With only three billion parameters, Phi-2 can be run on even smaller computer infrastructures. Therefore, it might be taken into consideration by practitioners and researchers with limited resources. Finally, we believe that it is possible to achieve much better scores with Llama 2 13B and Mixtral7Bx8: being larger than Llama 2 7B and Mistral, respectively, there is the possibility that their performance has been negatively influenced by our limited hardware.

4.1 Discussion on the evaluation metrics employed

We tried to interpret the results achieved through BERTScore by reading a sample of the output of each model, to estimate the quality of the answers also through human evaluation. We noticed that the needs generated by the models which achieved a F1 (BERTScore) of more than 0.60 were of extremely high quality: they corresponded to those identified in the manual labelling, usually even adding more specific information. Table 3 demonstrates three examples of this situation: the needs extracted by GPT-4 are consistent with the label’s ones, but they provide extra content. For instance, in the third example of Table 3, we manually identified the need *Bar recommendations*; GPT-4 added more specific information about the type of bar, generating *Pop-up bar recommendations*.

Post	Label Needs (Manual Labelling)	Generated Needs (GPT-4 Chain-of-Thought)
Is the journey back after the Staten Island Ferry just getting you back to Times Square. We are thinking of missing out on the Ferry and finishing the rest of the day in Lower Manhattan would we be missing much by doing this?	Itinerary planning advices	Travel itinerary advice, Information about Staten Island Ferry, Recommendations for activities in Lower Manhattan
Hi! Visiting in 2.5 weeks with hubby and 18 year old daughter. We’re going ice skating in Central Park on our final night and fancy going for a Mexican afterwards - any recommendations? We’re flexible on budget we will have spent the day out and won’t be dressed up so nowhere too fancy!	Restaurants advices	Mexican restaurant recommendations, Casual dining options
We are 6 crazy older women from Texas and wanting to try one of these pop-up bars out in December. Miracle on 9th or Frosty’s? Thank you!	Bar recommendations	Pop-up bar recommendations

Table 3. Examples of needs extracted by GPT-4 through Chain-of-Thought prompting and manually identified in three TripAdvisor posts.

The additional information provided by the LLMs are useful in terms of needs analysis, but could lower the scores of the evaluation metrics, because the generated sentences contain more words than the reference ones: this can decrease their semantic similarity. Therefore, we consider this as a limitation of the evaluation process employed. We observed that the models which achieved an F1 score going from 0.50 to 0.60 still provide very high-quality outputs, perfectly comprehensible and in line with the instructions provided in the prompt. We started noticing a drop in the answers’ quality in models with a F1 score of less than 0.50. The lower the score, the more the outputs were noisy (e.g., not respecting the requested format, or providing additional texts before and after the list of needs) or wrong (e.g., empty or containing wrong needs, usually duplicating other ones previously extracted or coming from the examples provided in the few-shot prompting).

4.2 Limitations

In considering the limitations of this study, it is important to recognize several key factors. Firstly, our hardware configuration has limited the performance of some models (Llama 2 13B and Mixtral). A more powerful computer infrastructure would have allowed us testing the models without any constraints and considering the possibility of fine-tuning them, to improve their performance. Secondly, the evaluation metrics such as BERTScore, ROUGE, and BLEU may not fully capture the subtle variations in how well the models perform, especially when it comes to understanding meaning and context. We believe that human evaluation remains crucial in assessing the quality and relevance of model-generated responses, particularly in specialized domains where additional context and specificity are essential. Lastly, the effectiveness of the prompt optimization technique employed may vary across the different LLMs, warranting further investigation into their impact on model performance and interpretability. Addressing these limitations could facilitate a more comprehensive understanding of LLM capabilities and inform future research endeavours in natural language processing and generative artificial intelligence.

5. Conclusion

Our study demonstrates the potential of LLMs in automating language understanding and customer needs extraction processes, particularly in the domain of travels. Through the deployment of various LLMs, we have observed promising results for the task of extracting travel customer needs from TripAdvisor posts. While closed models such as GPT-4, GPT-3.5, and Gemini exhibited robust performance in our analysis, open-source models like Mistral 7B, Llama 7B, and Phi-2 3B also showcased notable capabilities, despite their smaller sizes in terms of parameters. Particularly, Mistral 7B emerged as a standout performer, offering high-quality responses comparable to those of larger closed models while being deployable on more affordable computational infrastructure. Our study underscores the importance of selecting the most suitable LLM for specific business needs, considering factors such as performance, resource requirements, and the level of customization and security offered by open-source models. The findings presented offer insights for a diverse range of stakeholders. AI developers and NLP researchers can benefit from the evaluation of different language models, gaining insights into their strengths, limitations, and potential applications in customer needs extraction tasks. Market researchers and data analysts can utilize the methodologies and evaluation metrics outlined in the study to benchmark and assess the performance of language model-based solutions in similar tasks across various industries. Additionally, businesses and researchers operating in resource-constrained environments can identify suitable open-source models, such as Mistral 7B and Phi-2 3B, for implementing cost-effective AI solutions without compromising on performance. In conclusion, our work contributes to the growing body of research on LLMs and their applications in customer needs management. By providing empirical evidence of their effectiveness in extracting travel customer needs, we offer valuable insights for businesses seeking to leverage AI-driven solutions to enhance customer experience and optimize operational workflows.

References

- Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, 24(6), 975-990. <https://doi.org/10.1111/poms.12303>.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. <https://doi.org/10.48550/arXiv.2303.08774>.
- Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., ... & Moreira, C. (2020). Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, 36, 100239. <https://doi.org/10.1016/j.cosrev.2020.100239>.
- Bigi, A., Cassia, F., & Ugolini, M. M. (2022). Who killed food tourism? Unaware cannibalism in online conversations about traveling in Italy. *British Food Journal*, 124(2), 573-589. <https://doi.org/10.1108/BFJ-04-2021-0401>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Burger, B., Kanbach, D. K., Kraus, S., Breier, M., & Corvello, V. (2023). On the use of AI-based tools like ChatGPT to support management research. *European Journal of Innovation Management*, 26(7), 233-241. <https://doi.org/10.1108/EJIM-02-2023-0156>.

- Büyüközkan, G., Feyzioğlu, O., & Ruan, D. (2007). Fuzzy group decision-making to multiple preference formats in quality function deployment. *Computers in Industry*, 58(5), 392-402. <https://doi.org/10.1016/j.compind.2006.07.002>
- Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*. <https://doi.org/10.48550/arXiv.2006.14799>.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://doi.org/10.48550/arXiv.1810.04805>.
- Ding, X., Liu, B., & Yu, P. S. (2008, February). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231-240). <https://doi.org/10.1145/1341531.1341561>.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378. <https://doi.org/10.1037/h0031619>.
- Gal-Tzur, A., Rechavi, A., Beimel, D., & Freund, S. (2018). An improved methodology for extracting information required for transport-related decisions from Q&A forums: A case study of TripAdvisor. *Travel Behaviour and Society*, 10, 1-9. <https://doi.org/10.1016/j.tbs.2017.08.001>.
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). 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*, 9(1), e45312. <https://doi.org/10.2196/45312>.
- Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating Consistency in GPT-4 Text Ratings. *Frontiers in Education*. <https://doi.org/10.3389/educ.2023.1272229>.
- Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil transactions on benchmarks, standards and evaluations*, 2(4), 100089. <https://doi.org/10.1016/j.tbench.2023.100089>.
- Haman, M., & Školník, M. (2023). Using ChatGPT to conduct a literature review. *Accountability in research*, 1-3. <https://doi.org/10.1080/08989621.2023.2185514>.
- Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ICIRD.2018.8376299>.
- He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*. <https://doi.org/10.48550/arXiv.2006.03654>.
- Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., ... & Gopi, S. (2023). Phi-2: The surprising power of small language models. *Microsoft Research Blog*.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*. <https://doi.org/10.48550/arXiv.2310.06825>.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*. <https://doi.org/10.48550/arXiv.2401.04088>.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., ... & Potts, C. (2023). Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*. <https://doi.org/10.48550/arXiv.2310.03714>.
- Kilroy, D., Healy, G., & Caton, S. (2022). Using machine learning to improve lead times in the identification of emerging customer needs. *IEEE Access*, 10, 37774-37795. <https://doi.org/10.1109/ACCESS.2022.3165043>.
- Kim, J., Kim, J. H., Kim, C., & Park, J. (2023). Decisions with ChatGPT: Reexamining choice overload in ChatGPT recommendations. *Journal of Retailing and Consumer Services*, 75, 103494. <https://doi.org/10.1016/j.jretconser.2023.103494>.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199-22213.
- Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4), 10-11. <https://doi.org/10.1109/MPRV.2008.85>.
- Kumar, A., Gupta, N., & Bapat, G. (2023). Who is making the decisions? How retail managers can use the power of ChatGPT. *Journal of Business Strategy*. <https://doi.org/10.1108/JBS-04-2023-0067>.
- Li, J., Dada, A., Puladi, B., Kleesiek, J., & Egger, J. (2024). ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 108013. <https://doi.org/10.1016/j.cmpb.2024.108013>.
- Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. <https://doi.org/10.48550/arXiv.1907.11692>.
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. <https://doi.org/10.3390/educsci13040410>.
- Luo, Y., & Xu, X. (2021). Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *International Journal of Hospitality Management, 94*, 102849. <https://doi.org/10.1016/j.ijhm.2020.102849>.
- Mhlanga, D. (2023). 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* (pp. 387-409). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37776-1_17.
- Mo, B., Xu, H., Zhuang, D., Ma, R., Guo, X., & Zhao, J. (2023). Large Language Models for Travel Behavior Prediction. *arXiv preprint arXiv:2312.00819*. <https://doi.org/10.48550/arXiv.2312.00819>.
- Naqvi, S. M. R., Ghufraan, M., Varnier, C., Nicod, J. M., Javed, K., & Zerhouni, N. (2024). Unlocking maintenance insights in industrial text through semantic search. *Computers in Industry, 157*, 104083. <https://doi.org/10.1016/j.compind.2024.104083>
- Ray, T. (2024). Why open-source generative AI models are still a step behind GPT-4. *ZDNET*. Retrieved from: www.zdnet.com/article/this-is-why-open-source-generative-ai-models-are-a-step-behind-gpt-4. Accessed March 2024.
- Sasaki, Y. (2007). The truth of the F-measure. *Teach tutor mater, 1*(5), 1-5.
- Soni, V. (2023). Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies, 7*(1), 67-89.
- Spada, I., Barandoni, S., Giordano, V., Chiarello, F., Fantoni, G., & Martini, A. (2023). WHAT USERS WANT: A NATURAL LANGUAGE PROCESSING APPROACH TO DISCOVER USERS'NEEDS FROM ONLINE REVIEWS. *Proceedings of the Design Society, 3*, 3879-3888. <https://doi.org/10.1017/pds.2023.389>.
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. <https://doi.org/10.48550/arXiv.2312.11805>.
- Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*. <https://doi.org/10.48550/arXiv.2403.08295>.
- Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1-20. <https://doi.org/10.1287/mksc.2018.1123>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. <https://doi.org/10.48550/arXiv.2307.09288>.
- Trichopoulos, G., Konstantakis, M., Alexandridis, G., & Caridakis, G. (2023). Large language models as recommendation Systems in Museums. *Electronics, 12*(18), 3829. <https://doi.org/10.3390/electronics12183829>.
- Vlačić, B., Corbo, L., e Silva, S. C., & Dabić, M. (2021). The evolving role of artificial intelligence in marketing: A review and research agenda. *Journal of Business Research, 128*, 187-203. <https://doi.org/10.1016/j.jbusres.2021.01.055>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems, 35*, 24824-24837.
- Wu, F. (2020). BERTScore Default Layer Performance on WMT16. Retrieved from: https://docs.google.com/spreadsheets/d/1RKOVpselB98Nnh_EOC4A2BYn8_201tmPODpNWu4w7xI/edit?usp=sharing. Accessed March 2024.
- Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., ... & Chen, E. (2023). A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*. <https://doi.org/10.48550/arXiv.2305.19860>.
- Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism management, 31*(2), 179-188. <https://doi.org/10.1016/j.tourman.2009.02.016>.
- Yang, H., Liu, X. Y., & Wang, C. D. (2023). Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*. <https://doi.org/10.48550/arXiv.2306.06031>.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. <https://doi.org/10.48550/arXiv.1904.09675>.
- Zhou, F., Ayoub, J., Xu, Q., & Jessie Yang, X. (2020). A machine learning approach to customer needs analysis for product ecosystems. *Journal of mechanical design, 142*(1), 011101. <https://doi.org/10.1115/1.4044435>.