Chatbot Technology Assessment: 40 Cases from Greece

Yannis Charalabidis[1](✉), Thanos Anagnou[1], Charalampos Alexopoulos[1], Theodoros Papadopoulos[1], Zoi Lachana[1], Christos Bouras[2], Nikos Karacapilidis[3], Vasileios Kokkinos[2], and Apostolos Gkamas[4]

[1] Department of Information and Communication Systems Engineering, University of Aegean, 83200 Mytilene, Samos, Greece
{yannisx,alexop,t.papadopoulos,zoi}@aegean.gr, icsdm619001@icsd.aegean.gr

[2] Computer Engineering and Informatics Department, University of Patras, 26504 Patras, Greece
{bouras,kokkinos}@upatras.gr

[3] Department of Mechanical Engineering, University of Patras, 26504 Patras, Greece
karacap@upatras.gr

[4] University Ecclesiastical Academy of Athens, 14561 Athens, Greece
gkamas@aeavellas.gr

Abstract. In recent years, the field of Artificial Intelligence has seen significant progress, particularly in the development of chatbots via Natural Language Processing (NLP) technology. Recently, however, there has been a real race in this sector with major technology companies constantly presenting new improved solutions. However, the Greek reality presents several peculiarities and difficulties in adopting modern solutions, both due to the idiosyncrasies and rarity of the language and the limited funding capabilities of the Greek economy. The purpose of this research is to evaluate the performance of chatbots in terms of the quality of their responses regarding relevance, naturalness, cohesion, accuracy, vocabulary, as well as to assess the user experience and satisfaction. Another goal is to gain a comprehensive comparative picture of chatbot operation in Greece, both per question and in comparison, between relevant questions. A guided interview with closed-type questions was chosen as the method of evaluation. The aim is to obtain structured and quantified data in an area where the average internet user is not fully familiarized and does not have previous relevant evaluation experience. Conclusions were drawn per question in order to evaluate the level of solutions in a focused and comparative way to identify possible trends and to confirm the consistency of the responses.

Keywords: Artificial Intelligence · Natural Language Processing · Virtual Assistants · Greek Virtual Assistant

M. Papadaki et al. (Eds.): EMCIS 2023, LNBIP 501, pp. 167–181, 2024.
https://doi.org/10.1007/978-3-031-56478-9_12

1 Introduction

The global use of chatbots has surged dramatically in recent years, as businesses and organizations recognize their potential to enhance efficiency and customer satisfaction. One of the key factors for the success of chatbots is their ability to understand and respond to customer queries in a natural and human-like manner. This requires advanced Natural Language Processing (NLP) technology and a deep understanding of the language and communication patterns [1] used by the target audience.

1.1 Chatbots in Greece

Despite their widespread acceptance globally, the number of chatbots in the Greek language remains limited. There is, however, a growing interest in using chatbots in Greek, especially in customer service and government sector. Some businesses in Greece have adopted chatbots to provide 24/7 customer support, aiming to improve efficiency and reduce the reliance on human resources. The results of these implementations have been mixed, with some companies reporting significant improvements in customer satisfaction and efficiency [2], while others have struggled to convince their clients to use them.

Developing chatbots for the Greek language poses several technological challenges [3]. The lack of linguistic standardization, given Greek's ancient, intricate, and multifaceted nature with its many dialects and regional variations [4], complicates Natural Language Processing (NLP) algorithms. This complexity makes it difficult for chatbots to consistently understand and respond to customer queries. Another obstacle is the limited availability of high-quality training datasets, which hampers the ability of chatbots to provide accurate and effective responses. Nevertheless, initiatives are underway to promote the use of chatbots in the Greek language, such as the Pythia project and the development of language models specifically designed for Greek, like Greek-Bert [5].

1.2 Subject and Objectives of the Research

Chatbots are a contemporary tool for communication between humans and machines. Their purpose is to provide answers to user queries and to interact with them in a manner resembling human conversation. However, the quality of their responses depends on the precision and effectiveness of the algorithms used for their training, the accuracy of the training datasets, and the diligence with which they are maintained and improved by their developers.

An effort was made to assess the quality and performance of chatbots through the subjective judgment of individuals. These chatbots are employed by companies and organizations primarily targeting the Greek population.

The goal of the research is to evaluate the chatbots' performance in terms of the relevance, naturalness, coherence, accuracy, and vocabulary of their responses. Additionally, the study aims to assess the user experience and overall satisfaction with the chatbots' functionality. Another objective is to obtain a comprehensive comparative view of the chatbots' operation in Greece, both on a per-query basis and in comparison, between related queries.

2 Background

2.1 Evaluation of Chatbots

Evaluating chatbots is a critical step for their effective development. This assessment encompasses the evaluation of their quality and performance, particularly in understanding user requests and generating appropriate responses in the form of natural and engaging conversations. The aim is to ensure that they meet user needs, are efficient in achieving their intended purpose, and offer a positive user experience [6]. The challenges in evaluation stem from the inherent complexity of human communication itself.

Evaluation models typically involve a range of tasks that assess different aspects of a chatbot's performance, such as intent recognition, entity extraction, and response generation [7]. They can be applied to different data sets and tasks, offering a holistic view of the chatbot's overall quality. Datasets are also vital for the training and testing of chatbot models and evaluating their performance. These datasets should be diverse, representative of the language and environment of the target users, and annotated with relevant information such as intent, entities, and conversational turns.

Evaluation metrics can have both objective and subjective criteria [8]. Objective measures include metrics like perplexity, accuracy, and the F1 score, which are based on quantitative analysis of the chatbot's performance on a specific task or dataset. On the other hand, subjective measures rely on human evaluation of its performance, which can be conducted through surveys, interviews, or user studies. Some of these metrics are:

Perplexity. Is a metric commonly used to evaluate the performance of language models. It measures the uncertainty in predicting the next word in a sentence, with lower values indicating better performance [9]. It is calculated by taking the inverse of the geometric mean of the probabilities assigned to each word in the sentence. For instance, if a chatbot assigns a probability of 0.8 to the word "hello" and a probability of 0.2 to the word "world" in the sentence "hello world", the perplexity score would be $1/\sqrt{(0,8 * 0,2)} = 2.24$, which equals 2.24. Lower scores can result in more coherent and natural responses.

Accuracy is another metric frequently used to evaluate the overall performance of chatbots [10]. It is calculated by dividing the number of correct responses by the total number of responses. For instance, if a chatbot responds correctly to 80 out of 100 user inputs, its accuracy score would be 80%. Chatbots with higher accuracy scores can execute their designated tasks more effectively.

F1 Score. This is a metric commonly used to assess the performance of chatbots in natural language processing tasks, such as question answering and sentiment analysis. It measures the trade-off between the evaluation metrics of precision and recall [11]. Precision refers to the percentage of identified positive instances that are genuinely positive, meaning the proportion of predictions that are correct. On the other hand, recall pertains to the percentage of actual positive instances that were recognized by the algorithm, that is, the proportion of positive cases that were identified accurately. The balance between them highlights the fact that it's usually not possible to achieve a high level for both simultaneously. If we increase the threshold used to predict positive outcomes, precision will rise, but recall will decrease. Conversely, if we decrease the threshold, recall will increase, but precision will decrease. This happens because as the threshold rises, the number of correct predictions increases, but the total number of predictions made by the algorithm decreases. Inversely, as the threshold drops, the total number of predictions made by the algorithm increases, but it negatively impacts the accuracy of the predictions. The F1 score is especially useful for evaluating the bot's ability to handle complex conversations and to understand the essence of the conversation.

Human Evaluation. Evaluating chatbots is a significant process that involves assessing their quality and performance based on human subjective judgment. This assessment typically pertains to the quality of the chatbot's responses in terms of their relevance, naturalness, and coherence. Moreover, it evaluates the overall user experience and satisfaction derived from the chatbot's functionality [12]. Various methods can be employed for this evaluation, including surveys, interviews, user studies, and expert evaluations. Surveys and interviews usually gather user feedback about their experiences with the chatbot, such as their level of satisfaction, their perceptions of its usefulness, and their opinions on its performance [13]. User studies involve observing users interacting with the chatbot in a controlled environment and collecting data on their behavior and feedback. These studies may employ a random group of users with diverse experiences and backgrounds or focus on expert groups knowledgeable in the subject matter. Several reasons underscore the importance of this evaluation. Firstly, it provides a measure of the chatbot's performance, especially in terms of its ability to attract users and offer a positive user experience. This is crucial as chatbots are designed to interact in a natural and appealing manner, making the quality of user experience paramount to their success. Secondly, human evaluation can reveal the strengths and weaknesses of a chatbot, pointing out areas for improvement. This feedback is invaluable for developers aiming to enhance the performance of their chatbots and better cater to user needs. However, human evaluation has its limitations. It can be time-consuming and costly, especially if many evaluators are required. Additionally, it might be subjective and influenced by factors like personal preferences and biases. Finally, it might not always reflect the chatbot's real-world performance since users might interact differently in various settings or with different objectives. To address these challenges, researchers utilize various techniques such as focused interviews, focus groups, crowdsourcing, and evaluations based on machine learning.

3 Methodology

The chosen method of evaluation was the guided interview using closed-ended questions. This is a widely accepted assessment technique for interactive systems, which falls within both the realms of usability testing and predictive evaluation reviews [14]. The aim is to provide structured and quantified data in an area where the average internet user is not fully familiar with and lacks prior evaluation experience.

Closed-ended questionnaires are a popular research method utilized in guided interviews to collect data systematically and structurally from participants [15]. This method offers several advantages that make it a valuable tool in research. Firstly, it allows for efficient data collection. By employing a standardized set of questions, researchers can swiftly and effortlessly gather data from many participants, which is especially useful in research where time and resources are limited. Secondly, the data can be easily analyzed. Closed-ended questions produce quantitative data that can be readily coded and statistically analyzed, enabling patterns and trends to be quickly and effectively identified. Thirdly, they reduce bias in the data collection process. Closed-ended questions eliminate the need for participants to process their answers, thus reducing the chance of misrepresentation of their response. Lastly, they can be used to gather data from a broad range of participants, irrespective of their background, experience, or expertise.

3.1 Selected of Questions

The following questions were selected based on international literature as being critical criteria, as well as with the input from the supervising professors:

Solution Type: Voice vs. Text. Chatbots can be either voice or text-based, each having its unique characteristics and advantages. Voice-based chatbots provide the benefit of Natural Language Processing (NLP), which allows for more natural and user-friendly interactions. Users can communicate with the chatbot using their voice, akin to conversing with a human, making the experience more engaging and pleasant. On the other hand, text-based chatbots offer the advantage of easy accessibility across various devices such as smartphones, computers, and tablets. Users can interact with the chatbot by typing messages, a familiar and convenient mode of communication.

Economic Sector: Telecommunications, Financial, Commercial, Governmental Entities, Public Benefit, Local Governance, Education. Chatbot systems are widely utilized across various economic sectors, serving diverse areas of business activities. These systems are designed to streamline interactions with customers, enhance customer satisfaction, and provide cost savings for businesses.

Exclusive Mode of Communication: Yes vs. No. The idea of using chatbot systems as the sole point of contact between businesses and customers is relatively new, but it has gained popularity in recent years. This approach involves relying solely on chatbots for customer service, instead of offering multiple channels such as email or phone support. The advantages include reduced costs, 24/7 availability, and improved efficiency. The downsides are limited understanding, lack of human touch, and technical issues.

User Experience with the Interface: Excellent, Good, Average, Poor. User Experience (UX) pertains to the overall experience a user has with a product or service. It encompasses factors such as ease of use, response speed, clarity of content, responsiveness of menus, among others.

User Experience with the Outcome: Excellent, Good, Average, Poor. Users are asked to evaluate whether they achieved their intended goal. This is a crucial metric as it ultimately determines if the customer gets what they need or if they abandon the effort.

Overall Performance: Excellent, Good, Average, Poor. The overall performance from the user's perspective gauges the impression the entire experience left. Even if individual scores were lower or higher, at the end of the day, was the user serviced based on the scenario they had in mind?

3.2 Presentation of the Interview Methodology

The interview, as a focal point of the research, was conducted in a consistent manner for all participants. A neutral location was chosen for this purpose, which was quiet, devoid of decorations, and equipped with common amenities. This included a contemporary computer with a large 65-inch high-resolution (4K) screen and fast internet (1Gbps). The interviewer strived to remain impartial and refrained from suggesting any answers. Websites and phone numbers had been pre-saved to enable faster connections with the least amount of disruption. The duration was set at 2 h per interview to avoid excessive fatigue and irritation that could emotionally influence the responses. The time distribution was as follows: 20 min for introduction, 45 min for solution analysis, 10 min break, and another 45 mins for solution analysis.

The interviews were conducted based on the following scenario: Entry of the interviewee and familiarization with the space and equipment. Collection of the participant's profile data: Gender, Age, Education level, whether they use a computer, and if they own a smartphone. Demonstration and testing of ChatGPT as a benchmark for other solutions. Review of the text evaluating the questions. Commencement of evaluation.

The selection of participants aimed to represent, as closely as possible, potential Greek users who might choose the chatbot as a means of communication instead of abandoning it. For this reason, candidates who have no familiarity with the internet or have a negative attitude towards technology were excluded. Other criteria included availability and willingness to participate, considering the required time commitment (300 min). The characteristics of the group are listed in the table below (Table 1):

Table 1. Characteristics

No	Gender	Age	Educational level	Computer user	Smartphone owner
1	F	42	University	Yes, often	Yes
2	M	48	University	Yes, often	Yes
3	M	50	Elementary	No	Yes
4	M	52	High school	No	Yes
5	F	80	Postgraduate	Yes, often	Yes
6	F	51	Postgraduate	Yes, often	Yes
7	F	17	High school	Yes, often	Yes
8	M	23	University	Yes, often	Yes
9	M	37	Postgraduate	Yes, often	Yes
10	M	29	University	Yes, often	Yes

4 Results

Based on the questionnaires, results emerge for each solution. The table below presents the solutions that were explored, whether they belong to the private or public sector, and the corresponding economic sector they pertain to (Table 2).

Table 2. Chatbot Solutions

Company – Organization	Solution type	Application Domain	Economic sector
2103288000 - Piraeus	Voice	Private	Financial
Winbank - Piraeus	Text	Private	Financial
Vodafone.gr - tobi	Text	Private	Telecommunications
13888 - Cosmote	Voice	Private	Telecommunications
Alpha Bank	Voice	Private	Financial
Eurobank	Voice	Private	Financial
National bank of Greece	Voice	Private	Financial
Attika bank	Text	Private	Financial
Hellenic Development Bank	Text	Private	Financial
leroymerlin	Text	Private	Commercial
ikea	Text	Private	Commercial
eco-mat	Text	Private	Commercial
pennie	Text	Private	Commercial
ledison	Text	Private	Commercial
xtr	Text	Private	Commercial
acs	Text	Private	Commercial
coca-cola	Text	Private	Commercial
goldmall	Text	Private	Commercial
Market4you	Text	Private	Commercial
ReBrain Greece	Text	Public	State entities
oasa	Text	Public	Public utility
deddie	Text	Public	Public utility
dei	Text	Public	Public utility
eydap	Text	Public	Public utility
eopyy	Text	Public	State entities
dypa	Text	Public	State entities

(*continued*)

Table 2. (*continued*)

Company – Organization	Solution type	Application Domain	Economic sector
Region of Attika	Text	Public	State entities
Region of Stereas Elladas	Text	Public	State entities
Municipality of Papagou-Hollargou	Text	Public	Municipalities
Municipality of Kalamaria	Text	Public	Municipalities
Municipality of Patmos	Text	Public	Municipalities
Municipality of Moschato-Tavros	Text	Public	Municipalities
Municipality of Filis	Text	Public	Municipalities
Municipality of Kastellorizo	Text	Public	Municipalities
Municipality of West Lesvos	Text	Public	Municipalities
Municipality of Platanias	Text	Public	Municipalities
Municipality of Agia	Text	Public	Municipalities
Municipality of Visaltia	Text	Public	Municipalities
Elecectrical & Computer, Engineering Dept - UOP	Text	Public	Education
University of West Attika	Text	Public	Education

For the open-ended questions, the most frequent response given by the interviewees is presented, which we believe represents the majority opinion.

Conclusions will be drawn on a per-question basis to evaluate the level of the solutions in a focused and comparative manner, aiming to identify possible trends.

4.1 Analysis Presentation of the Interview Methodology

Based on the responses given by the participants in the study the functionality and user experience were evaluated.

Sector of Economy. The economic sector of the country plays a pivotal role. The economic profile of each entity implementing a chatbot solution gives us an insight into the penetration these solutions have in the country's economy. However, in relation to the number of sites accessed to find them, they constitute a very small percentage, around 15% (Fig. 1).

Most implementations (10) were found in the commercial sector, which is expected due to the abundance of online stores in the post-Covid era. Surprisingly, local government also had 10 implementations, which seem to be part of a pre-existing software

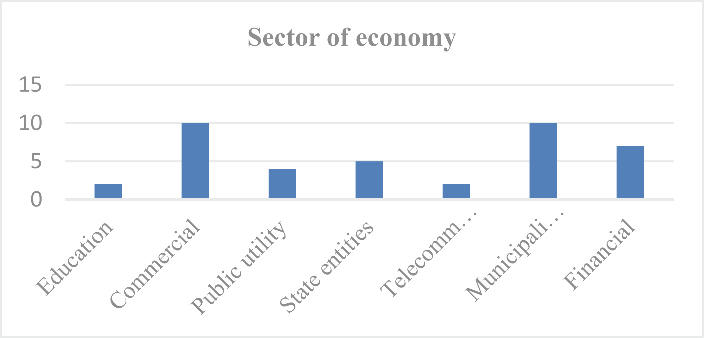

Fig. 1. Sector of economy

package. Other entities (financial, telecommunications & public utilities) have a significant presence relative to their smaller number. Surprisingly low percentages were found in the education sector and central government.

Exclusive Communication Channel. This is the most significant indication of how much a company or organization has relied on chatbots for customer service, and consequently, how much they have invested in the development of this technology (Fig. 2).

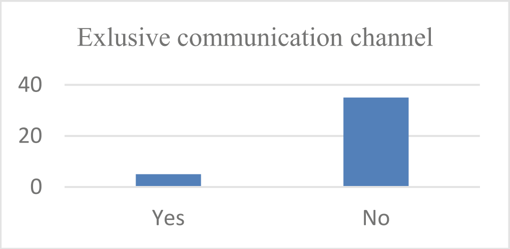

Fig. 2. Exclusive Communication Channel

In the vast majority (35/40) of cases, the implementations we observe serve as a supplementary tool. As a result, most haven't made substantial investments in the advantages that a chatbot offers. The exception is large organizations that have carefully weighed the potential benefits this communication protocol can bring to them.

User Experience with the Interface. Regardless of how advanced the technology of a chatbot is, if the interface with which the user interacts isn't user-friendly, fast, and clear, then it's challenging to evaluate it positively (Fig. 3).

We found that in most cases (34/40), there's an acceptable (ranging from good to average) reception to the way of interaction with the chatbot. The average user's familiarity with messaging applications aids in understanding its functionality. On the

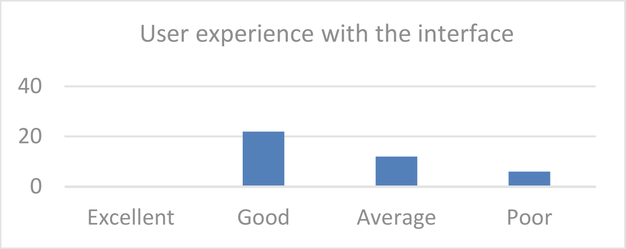

Fig. 3. User Experience with the Interface

other hand, the neutral tone of the responses and the lack of intuition prevented any outstanding evaluations.

User Experience with the Outcome. The primary expectation of a user when engaging with a chatbot is to receive comprehensive assistance with the least effort. The type and importance of the issues one aims to address through the application also play a significant role. For instance, the service quality expected from a bank or telecommunications provider differs from that of a store or municipality.

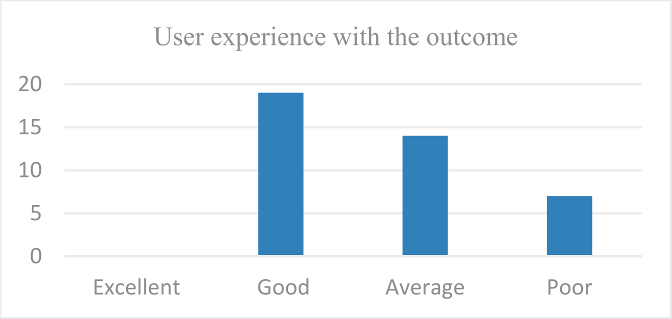

Fig. 4. User Experience with the Outcome

From the results, we observe a moderate to low satisfaction level (33/40) with the outcomes. The bot is mainly used as a search tool within the site rather than a solution-generating engine. The absence of top-rated outcomes isn't surprising since there was no instance of the "magic" of intuitive results typical of a well-functioning AI.

Overall Performance. The interviewee is asked to make an overall assessment of the solution they tested. The main criterion remains the extent and the personal effort and discomfort required to be served. However, it is important to distinguish this from the previous evaluation of the outcome. If the experience was poor, the user might not have continued to that point unless they were participating in a study. On the other hand, they might not have achieved the expected result, but the overall experience might not have been bad (Fig. 5).

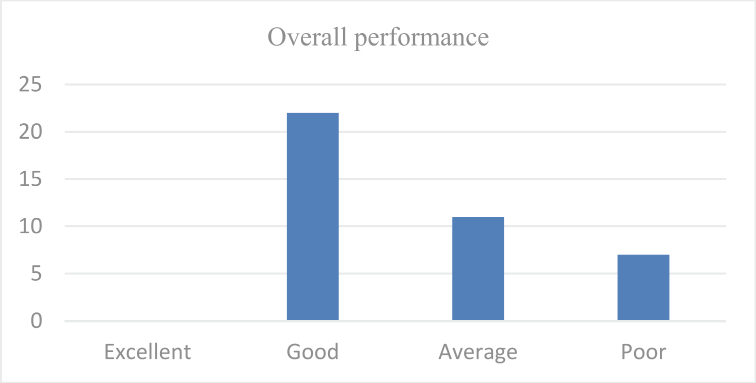

Fig. 5. Overall Performance

The pattern observed in the previous questions persists here. Good to moderate performances (33/40) are predominant as users obtained some results, even if it required considerable effort. High performance was not achieved since the goal of a comprehensive intuitive system wasn't even remotely approached. On the contrary, it appears there was an informal compromise and leniency in judgments when the assessed organization or business seemed smaller.

4.2 Comparative Analysis

Based on the responses given by the participants in the survey, and after evaluating the functionality and user experience, a comparative analysis was conducted between the results to ascertain emerging trends and to verify the consistency of the answers.

Solution Category with Overall Performance. We make this comparison to see the satisfaction rate per solution category and to evaluate it (Fig. 6).

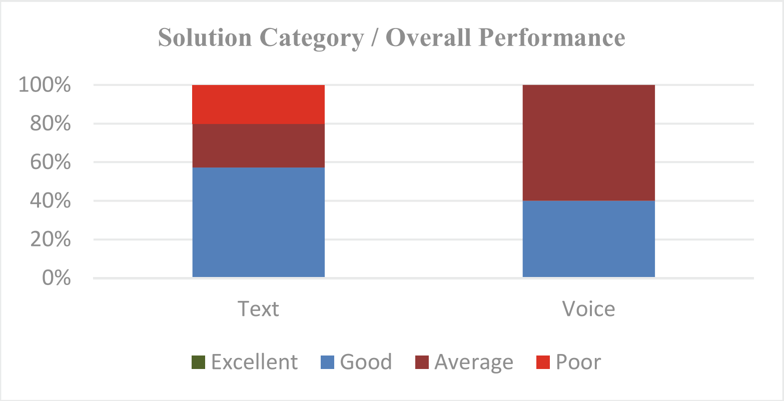

Fig. 6. Solution Category with Overall Performance

Voice solutions, although they only have good and average performance, the good performance vote is lower than the average vote. In contrast, text solutions, even though they included poor overall experiences, still had a satisfaction rate of over 50%. Considering that the voice portals came from large companies, we could infer that user had higher expectations and evaluated them more strictly. Another factor might be the stress of answering a question as in natural speech quickly and fluently.

Overall Performance by Economic Sector. We conduct this comparison to understand the final perception of the interviewees in relation to the economic sector each solution serves (Fig. 7).

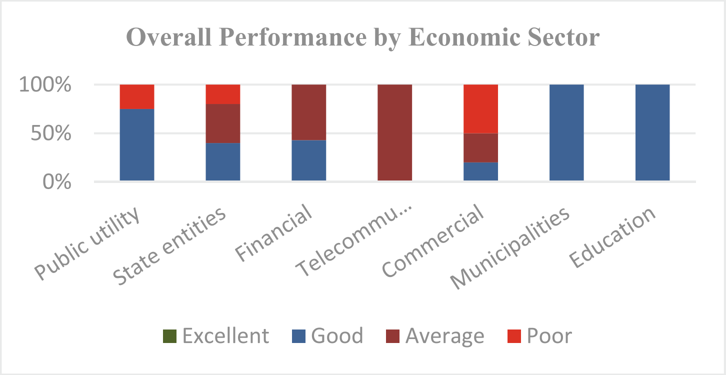

Fig. 7. Overall Performance by Economic Sector

Local governance and education enjoy a positive review. Although they predominantly provide only information, there aren't higher demands placed on them. In contrast, the commercial sector is rated quite low, and rightly so, as the adopted solutions appear to be more complementary to their websites and lack any significant investment in their dynamics.

User Experience in Relation to Overall Performance. We aim to examine whether the obtained results were consistent, given that the user interface invariably has a significant impact on the overall user experience (Fig. 8).

We observe that, with minor deviations, a good interface leads to an effective end performance. This is entirely logical, as it would be improbable for substantial resources to have been invested in the technological foundation without corresponding effort in presenting the outcome.

User Experience in Relation to Overall Performance. We aim to see if the results obtained were consistent, as the interface always has a significant impact on the user's overall experience (Fig. 9).

We observe that, with minimal deviations, a good interface leads to good overall performance. This is perfectly logical, as it wouldn't be possible for a large amount to

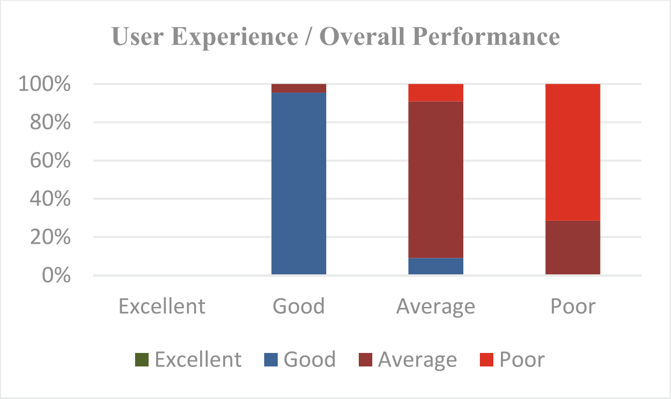

Fig. 8. User Experience. Interface to Overall Performance

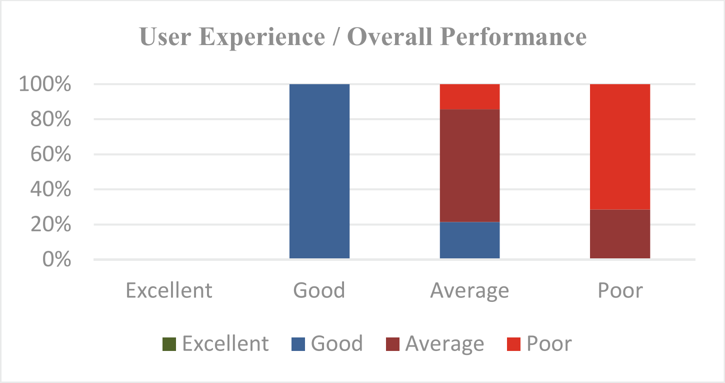

Fig. 9. User Experience. Outcome to Overall Performance

be invested in the technological backbone without corresponding work on presenting the result.

5 Conclusions

As we look to the future, it's clear that chatbots will play an increasingly significant role in our everyday lives, and their impact on society and our economy will continue to grow. The ongoing advancement of artificial intelligence and natural language processing technologies will enable chatbots to understand and respond to customer inquiries in a more human-like manner, providing more precise and personalized responses. As a result, people will feel more comfortable using them, and their adoption will become more widespread.

In Greece, the average to low level of satisfaction reflects the challenges posed by the Greek language, as well as the need for greater investments. Despite these challenges, Greece should not lag in technological advancements. The country possesses a skilled human workforce with the appropriate educational background, and there is potential for direct application of various solutions in the tourism sector.

In the short term, emphasis should be placed on improving NLP technologies and on increasing specialized staff for training chatbots. In the medium term, investments in new technologies that are more compatible with our linguistic idiom are essential. It is also crucial to promote collaborations between universities, tech companies, the government, and private entities.

From the research results for the private sector, there is a discernible need for more comprehensive electronic services. These services should be available 24/7, allowing an increase in workflow without an uptick in payroll costs. Additionally, merely redirecting to ready-made product websites without the use of AI, which would offer advice and solutions, is unsatisfactory. Such an approach does not make chatbots appealing or attractive for usage. While digital customer service in large companies is satisfactory, it hasn't excelled, indicating that there is a demand for further emphasis on its improvement. It's suggested to initiate collaborations between academic/research institutions and the private sector to develop and train the first "Greek digital sales assistant" leveraging emerging AI platforms, like ChatGPT. Another intriguing development would be the creation of a voice-text hybrid to facilitate complex processes, such as the signing of contracts and agreements, with a higher degree of satisfaction.

For the public sector, chatbots' primary informational role, their complete disconnection from providing real services, and the low expectations set by the research team suggest that the average citizen will resort to more traditional methods if they wish to be served, thereby losing the 24/7 availability advantage. Immediate funding is reccommended for integrating all administrative processes into the National Registry of Administrative Procedures. There should also be an obligation for the legislator to model each new process in stages before it's submitted for approval by the Parliament. Lastly, creating a unified "digital assistant for administrative processes" for the entire Greek public sector could be beneficial. This assistant would support citizens in completing applications, direct them to the appropriate service, and keep them informed about the results. This could pave the way for the complete digitization of the Greek state.

Acknowledgements. This research has been co-financed by the European Union and Greek national funds through the Operational Program "Competitiveness, Entrepreneurship and Innovation", under the call "RESEARCH - CREATE - INNOVATE (2nd Cycle)" (project code: T2EDK-01921).

References

1. Knight, S.: NLP at Work. Hachette Book Group, London (2020)
2. vodafone.gr. 28 June 2022. https://www.vodafone.gr/vodafone-ellados/digital-press-office/deltia-typou/20220628-tovi-o-psifiakos-voithos-tis-vodafone-pio-exypnos-kai-apotelesmatikos-apo-pote/
3. Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P., Daradoumis, A.: mpdi.com. 22 March 2022. https://doi.org/10.3390/app12073239

4. Lachana, Z., Loutsaris, M.A., Alexopoulos, C., Charalabidis, Y.: Automated analysis and interrelation of legal elements based on text mining. Int. J. E-Serv. Mob. Appl. (IJESMA) **12**(2), 79–96 (2020)
5. Koutsikakis, J., Chalkidis, I., Malakasiotis, P., Androutsopoulos, I.: Cornell University. arxiv.org: https://arxiv.org/pdf/2008.12014v2.pdf. 03 September 2020
6. Tullis, T., Albert, B.: Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, Burlington (2008)
7. Shawar, A., Atwell, E.: Different measurements metrics to evaluate a chatbot system. In: Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog, pp. 89–96. Seattle: NAACL (2007)
8. Stent, A., Marge, M., Mohit, S.: Evaluating evaluation methods for generation in the presence of variation. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 351–354. Springer, Berlin (2009)
9. Jelinek, F., Mercer, Salim: Principles of lexical language modeling for speech recognition. In: Advances in Speech Signal Processing. Dekker Publishers, New York (1991)
10. Liu, B.: Sentiment Analysis and Opinion Mining. In: Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael (2012)
11. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
12. Luger, E., Sellen, A.: Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5286–5297). Association for Computing Machinery, New York (2016)
13. Radziwill, N., Benton, M.: Cornell University. arxiv.org: https://arxiv.org/ftp/arxiv/papers/1704/1704.04579.pdf. (2017)
14. Koutsampasis, P.: University of Aegean, Department of Product and systems design engineering. eclass.aegean.gr (2015). https://eclass.aegean.gr/modules/document/file.php/511265/merged_document5.pdf
15. Oppenheim, A.: Questionnaire Design, Interviewing and Attitude Measurement. Continuum International Publishing, London (1992)