MÜHENDİSLİK ve DOĞA BİLİMLERİ FAKÜLTESİ

Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2024) 16(2), 679-691.

10.29137/umagd.1390083

Cilt/Volume:16 Sayı/Issue:02 Haziran/June 2024

Araştırma Makalesi / Research Article

The Battle of Chatbot Giants: An Experimental Comparison of ChatGPT and Bard

Sohbet Robotu Devlerinin Savaşı: ChatGPT ve Bard'ın Deneysel Bir Karşılaştırması

Abdullah Talha Kabakuş [1*], İbrahim Alper Doğru [2]

[1] *Duzce University, Faculty of Engineering, Department of Computer Engineering, 81620 Duzce, TÜRKİYE*
[2] *Gazi University, Faculty of Technology, Department of Computer Engineering, 06560 Ankara, TÜRKİYE*

***Başvuru/Received**: 13/11/2023* ***Kabul/Accepted**: 06/03/2024* ***Çevrimiçi Basım/Published Online:** 30/06/20224*
***Son Versiyon/Final Version**: 30/06/2024*

Abstract

Nowadays, it is hard to find a part of human life that Artificial Intelligence (AI) has not been involved in. With the recent advances in AI, the change for chatbots has been an ‘evolution’ instead of a ‘revolution’. AI-powered chatbots have become an integral part of customer services as they are as functional as humans (if not more), and they can provide 24/7 service (unlike humans). There are several publicly available, widely used AI-powered chatbots. So, “*Which one is better?*” is a question that instinctively comes to mind and needs to shed light on. Motivated by the question, an experimental comparison of two widely used AI-powered chatbots, namely *ChatGPT* and *Bard*, was proposed in this study. For a quantitative comparison, (i) a gold standard QA dataset, which comprised 2,390 questions from 109 topics, was used and (ii) a novel answer-scoring algorithm based on cosine similarity was proposed. The covered chatbots were evaluated using the proposed algorithm on the dataset to reveal their (i) generated answer length and (ii) generated answer accuracy. According to the experimental results, (i) *Bard* generated lengthy answers compared to *ChatGPT* and (ii) *Bard* provided answers more similar to the ground truth compared to *ChatGPT*.

Key Words

“Chatbot, question answering, artificial intelligence, ChatGPT, Bard, large language model”

Öz

Günümüzde, Yapay Zekanın (YZ) dahil olmadığı bir insan yaşam alanı bulmak zordur. YZ'deki son gelişmelerle birlikte, sohbet botları için değişim bir 'devrim' yerine bir 'evrim' şeklinde olmuştur. YZ destekli sohbet botları, insanlarla daha fazla değilse de aynı derecede işlevsel oldukları ve insanlardan farklı olarak 7/24 hizmet verebildikleri için müşteri hizmetlerinin ayrılmaz bir parçası haline gelmiştir. Erişime açık ve yaygın olarak kullanılan bazı YZ destekli sohbet botu vardır. Bu nedenle, "Hangisi daha iyi?" sorusu içgüdüsel olarak akla gelmekte ve aydınlatılması gerekmektedir. Bu sorudan yola çıkarak, bu çalışmada yaygın olarak kullanılan iki YZ destekli sohbet botunun, yani ChatGPT ve Bard'ın deneysel bir karşılaştırması önerilmiştir. Nicel bir karşılaştırma için, (i) 109 konudan 2.390 sorudan oluşan bir altın standart soru-cevap veri seti kullanılmış ve (ii) yeni bir cevap puanlama algoritması önerilmiştir. Kapsanan sohbet botları, önerilen algoritma kullanılarak veri seti üzerinde değerlendirilmiştir; böylece (i) üretilen cevap uzunluğu ve (ii) önerilen cevap puanlama algoritmasıyla elde edilen üretilen cevap doğruluğu ortaya çıkarılmıştır. Deneysel sonuçlara göre, (i) Bard, ChatGPT'ye kıyasla daha uzun cevaplar üretmiş ve (ii) Bard, ChatGPT'ye kıyasla gerçeğe daha yakın cevaplar sağlamıştır.

Anahtar Kelimeler

“Sohbet botu, soru cevaplama, yapay zekâ, ChatGPT, GPT, Bard, büyük dil modeli”

***Responsible Author:** talhakabakus@gmail.com*

1. Introduction

Artificial Intelligence (AI) has become a fundamental part of our daily lives due to the great benefits it provides in a wide range of different areas, including but not limited to chatbots, search engines, recommendation systems, virtual assistants, language translations, and healthcare. Thanks to the provided user-friendly applications designed for desktop and mobile operating systems, these benefits have become much more reachable than ever. One of the most commonly used applications of AI is a chatbot - a software application that aims to mimic human conversation through text or voice interactions, typically online (Caldarini et al., 2022). Chatbots take queries in natural language and generate responses in natural language as well. Chatbots are currently applied to a variety of different fields and applications, spanning from education to e-commerce, encompassing healthcare and entertainment (Caldarini et al., 2022). Before the rise of AI-powered chatbots, chatbots relied on rules and pattern-matching techniques. The main limitation of these chatbots is that they are domain-dependent, which makes them inflexible as they solely rely on domain-specific hard-coded rules and patterns (Caldarini et al., 2022). Thanks to providing interactions in natural language, just like human conversations, chatbots form a solid alternative to the static Frequently Asked Questions (FAQ) sections that websites provide. In addition to this, chatbots are capable of answering the generic questions of people and have reduced, if not eliminated, the necessity of human customer agents. This is a revolution instead of an evolution in customer service (Nirala et al., 2022; Paliwal et al., 2020; Shaji George et al., 2023; Sousa et al., 2019). For example, Erica (Erica - Virtual Financial Assistant From Bank of America, 2024), Bank of America's virtual financial assistant chatbot, helps Bank of America customers with their banking needs, such as account management, bill payments, and budgeting. Another example is the University of Auckland's UoA (Introducing UoA Assistant, 2021) chatbot, which assists students in accessing information regarding courses, examinations, campus facilities, academic deadlines, and other university-related services. As a result of the advances in AI, Natural Language Processing (NLP), and available computational power and storage capacities in recent years, chatbots have become much more common, and their ability to mimic human conversation has become much more advanced. To this end, chatbots do have the ability to analyze previous customer data to learn and adapt to customer needs. Consequently, chatbots provide improved response times and reduce waiting times for customers as they provide 24/7 service and are capable of handling a large volume of customer inquiries simultaneously. Due to all of these benefits, many businesses have successfully integrated chatbots into their websites, social media platforms, and mobile apps.

From the developer's perspective, the available open-source technologies and frameworks have made chatbots much easier to implement. There are several publicly available, widely used chatbots. As of the time of writing this paper, OpenAI's ChatGPT (ChatGPT, 2024) is the most widely used chatbot powered by Large Language Model (LLM) per Google Trends (Google Trends, 2024), a service provided by Google that allows explore the popularity of search terms over time and across different regions. Analysis of the global trends for ChatGPT, Bard, and Microsoft Copilot over the past three months using Google Trends revealed trend indexes of 73, 8, and 2, respectively (ChatGPT, Bard, Microsoft Copilot - Explore - Google Trends, 2024). As per a recent report by (Shewale, 2023), ChatGPT receives an average of over 1 billion visits per month, whereas Bard had 140.6 million visits in October 2023. These numbers underscore the significant prominence of ChatGPT in comparison to Bard and Microsoft Copilot. This trend analysis is plotted in Fig. 1.

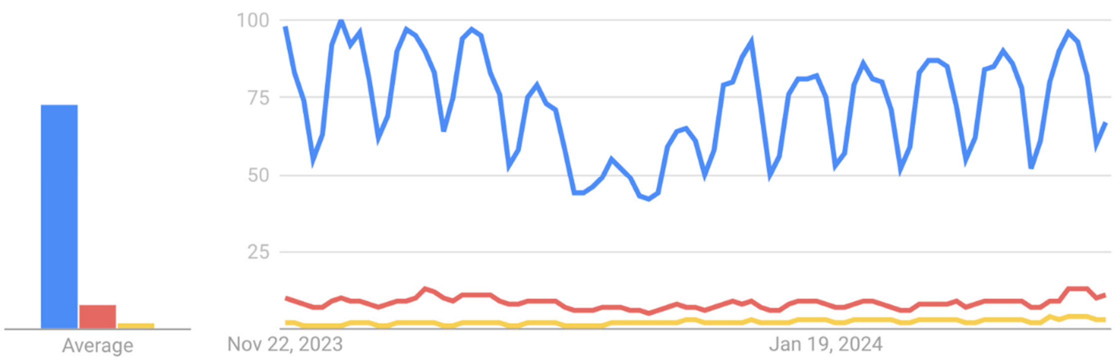

Figure 1. The plot of the analysis of the global trends for ChatGPT, Bard, and Microsoft Copilot over the past three months per Google Trends. ChatGPT, Bard, and Microsoft Copilot were represented by the colors blue, red, and yellow, respectively. Trend indexes of 73, 8, and 2 were obtained for ChatGPT, Bard, and Microsoft Copilot, respectively.

More specifically, ChatGPT was constructed on GPT (Generative Pre-trained Transformer) architecture and was designed to understand and generate human-like text based on the input it receives. Since it was trained on a vast amount of text data, it is capable of understanding context, generating coherent responses, and mimicking human-like conversation. Another notable AI-powered chatbot is Google's Gemini (Gemini Team, 2023). One final example is Microsoft's Copilot (Microsoft Copilot, 2024), which comes bundled with many Microsoft products (e.g., Word, Excel, PowerPoint, and Outlook). So, "Which one is better?" is a question that instinctively comes to mind and needs to be shed light on by the researchers. This study was motivated by this question, and its major contributions are listed as follows:

- A quantitative and comprehensive comparison. The question-answering abilities of two widely used chatbots are investigated through the experiments conducted, and the observations in light of the experimental results are discussed. We also conducted a novel comparison of their answers based on answer lengths. To the best of our knowledge, this aspect has not been explored in previous literature on the subject.
- A novel automated answer-scoring algorithm. A novel automated answer-scoring algorithm based on cosine similarity is proposed to evaluate the performance of the chatbots, as there is no de facto standard evaluation method for chatbots (Caldarini et al., 2022). The proposed algorithm can be easily adapted to other question-answering tools.

The rest of the paper is structured as follows: Section 2 briefly describes the related work. Section 3 presents the material and method used to propose this study. Section 4 describes the experimental results and discussion. Finally, Section 5 concludes the paper with future directions.

2. Related Work

In this section, the related works are briefly reviewed. Since chatbots powered by AI are relatively new, the papers that propose the comparison of chatbots lack in the research field. Most studies compare chatbots with human experts (Ariyaratne et al., 2023; Cheung et al., 2023; Guo et al., 2023; Herbold et al., 2023; Hulman et al., 2023; Li et al., 2023). Ali et al. (Ali et al., 2023) proposed an experimental comparison of ChatGPT, GPT-4, and Bard based on their performance on a question bank designed for neurosurgery oral board examination preparation. According to the experimental results, the best performance was obtained by GPT-4 by correctly answering 82.6% of questions. Given that GPT-4 is not available for public access, we opted for the latest freely accessible version of GPT, namely GPT-3.5-Turbo. Following GPT-4, ChatGPT and Bard correctly answered 62.4% and 44.2% of questions, respectively.

Rahaman et al. (Rahaman et al., 2023) proposed an opinion article based on the comparison of Bard and ChatGPT. Bernardini et al. (Bernardini et al., 2018) proposed an analysis of the chatbot literature. According to their experimental results, they observed the following: (i) A significant increase in the number of publications was detected in 2016, (ii) a significant presence of interdisciplinarity, ratifying that AI contemplates several distinct areas, including but not limited to health, education, computing, linguistics, and psychology, and (iii) despite the academic research and environment usually oriented into educational contexts, the AI-powered smart personal assistants, such as Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant, strengthen the advancement of research in the area aimed at the public. As indicated by the corresponding manufacturers, these services extensively use neural networks, which are considered an emerging concept in the field of AI. Caldarini et al. (Caldarini et al., 2022) proposed an extensive literature survey of recent advances in chatbots. Regarding the evaluation of chatbots, they mentioned two major limitations that have yet to be addressed: (i) the absence of a common evaluation framework and (ii) the absence of a reliable, efficient automatic evaluation method. Nuruzzaman and Hussain (Nuruzzaman & Hussain, 2018) compared the functionality and technical requirements of eleven widely used chatbot applications, such as Google Dialogflow, Amazon Lex, IBM Watson, and Microsoft LUIS (Language Understanding Information Service). When they investigated the common chatbot applications, they observed that an efficient and effective chatbot needs to be (i) implemented using deep learning algorithms and (ii) trained on a vast amount of data to recognize natural languages and react accordingly to any situation.

A recent study (Peyton & Unnikrishnan, 2023) proposed a comparison of two widely used chatbot frameworks to a state-of-the-art Sentence BERT (SBERT) model that can be used to build a robust chatbot. Similar to our approach, they utilized APIs to obtain experimental results. They employed the F1-score as the evaluation metric while quantitatively comparing the chatbots. According to their experimental results, the employed SBERT model outperformed the models based on Dialogflow and QnA with an F1-score of 0.99. In another recent study (Waisberg et al., 2023), a comparison between ChatGPT and Bard was presented within the domain of ophthalmology. The researchers observed that while ChatGPT and Bard exhibit certain similarities, Bard notably excels in comprehending and responding to queries necessitating precise information. Yeung et al. (Au Yeung et al., 2023) proposed a comparison of two commonly used AI-powered chatbots, namely (i) ChatGPT and (ii) Foresight. Clinical scenarios were inputted into the chatbots, and the expectation was to receive five potential diagnoses predicted by the chatbots. Both chatbots were evaluated on the same test set. According to the experimental results, both ChatGPT and Foresight obtained a top-1 accuracy of 93%. Hristidis et al. (Hristidis et al., 2023) proposed an approach that compares ChatGPT and Google for queries related to dementia and other cognitive decline. They conducted a set of experiments to compare them. According to their experimental results, Google exhibited superior currency and higher reliability compared to ChatGPT. The evaluation of ChatGPT results indicated a higher level of objectivity. Notably, ChatGPT demonstrated significantly higher response relevance, whereas Google frequently relied on sources affiliated with referral services for dementia care or service providers themselves. However, both platforms had low readability, with ChatGPT averaging a mean grade level of 12.17 (SD 1.94) and Google at 9.86 (SD 3.47). Regarding content similarity, 21.7% of responses were rated as high, 26.7% as medium, and 51.6% as low. ChatGPT and Stack Overflow were compared in another study (Liu et al., 2023). A set of experiments was conducted within the scope of this study. Based on their experimental findings, ChatGPT demonstrated significant superiority over Stack Overflow in assisting with algorithmic and library-related tasks. Conversely, Stack Overflow was found to be more effective for debugging tasks.

3. Material and Method

An LLM is a type of artificial intelligence model that is trained on massive amounts of text data to understand and generate human-like language. LLMs are capable of performing a wide range of NLP tasks, including text generation, summarization, translation, and question-answering. Even more, the latest LLM tool of OpenAI, namely Sora (Sora, 2024), is capable of generating realistic and imaginative movies from given text descriptions. LLMs are often used in various applications, including chatbots, language translation services, content generation, sentiment analysis, and text classification. They have significantly advanced the capabilities of natural language understanding and generation and are widely utilized in research, business, and everyday applications.

In the following subsections, the material and method used to propose this study are described in detail. More specifically, we first describe the dataset used to benchmark the employed chatbot APIs. Following this subsection, we describe the employed APIs for the covered chatbots, namely (i) ChatGPT and (ii) Bard. In the final subsection, we describe the metrics used to evaluate the responses generated by the chatbot APIs employed.

3.1. QA Dataset

There are various QA (Question Answering) datasets that can be used to evaluate the responses generated by the chatbots. The QA dataset (Smith et al., 2008) proposed by Carnegie Mellon University and the University of Pittsburgh was used in this study. This dataset comprises 2,390 questions from 109 topics with their respective answers and topics they are related to (named “title” in the dataset). These topics include, but are not limited to, cities, objects, famous people, animals, countries, cities, and languages. Some examples from this dataset with their ground truths are given in Table 1. As can be seen in this table, the dataset contains various question types, including but not limited to questions targeting people, event dates, object(s), numbers, and “yes” or “no” answers. The variations in both topics and question types make this dataset an ideal test bed to evaluate the performance of the chatbots. As a part of preprocessing, the questions without answers were removed from the dataset. Also, it is worth mentioning that some answers retrieved from chatbot APIs contained date chunks (in the format of “yyyy − mm − dd”) at the start of the answer that indicated the date that the corresponding answer was retrieved from the API. Since these chunks do not contain meaningful information regarding questions, they were cleaned from the chatbots’ answers using the appropriate Regular Expressions (RegEx) through the built-in “re” (Re — Regular Expression Operations, 2024) package of the Python 3 SDK.

Table 1. Some examples from the used QA dataset.

Question	Ground Truth	Topic
Where is Henri Becquerel from?	Paris	Henri Becquerel
Was Isaac Newton British?	Yes	Isaac Newton
What is the dominant religion in Ghana?	Christian	Ghana
When was the Turkish Language Association founded?	In 1928	Turkish language
How many people speak the Arabic language?	280 million people	Arabic language

Each question in the dataset was labeled with a type as follows: (i) yes_no, if the answer to the question starts with “Yes/No” and (ii) other, for the other questions. Under this rule, the dataset comprised 1,131 yes_no and 1,058 other questions, as presented in Fig. 2.

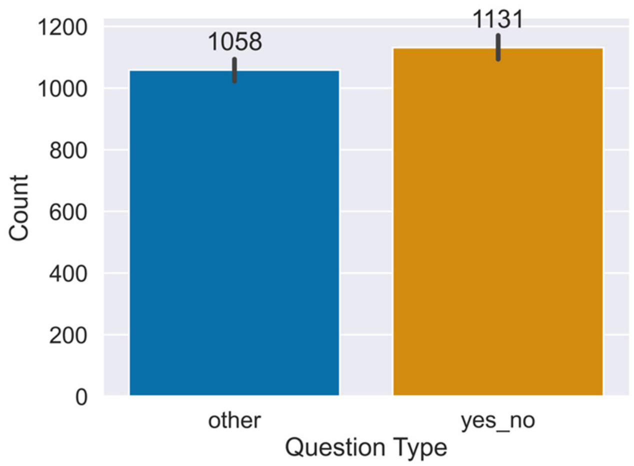

Figure 2. The distribution of the question types in the used dataset.

3.2. Employed Chatbot APIs

GPT4All (Anand et al., 2024) provides an ecosystem to train and deploy LLMs that can run locally. GPT4All was employed to use GPT-3.5-Turbo, the language model that powers ChatGPT, while generating answers for the given questions. GPT stands for "Generative Pre-trained Transformer" and is an LLM and a prominent framework released by OpenAI, a company backed by Microsoft. According to a recent report (Hu, 2023), ChatGPT is the fastest-growing consumer application in history, reaching 100 million monthly active users after just two months since its launch. The second fastest-growing consumer application is TikTok, which took seven months more than ChatGPT to acquire the same number of users (Garfinkle, 2023). The web user interface of ChatGPT is presented in Fig. 3.

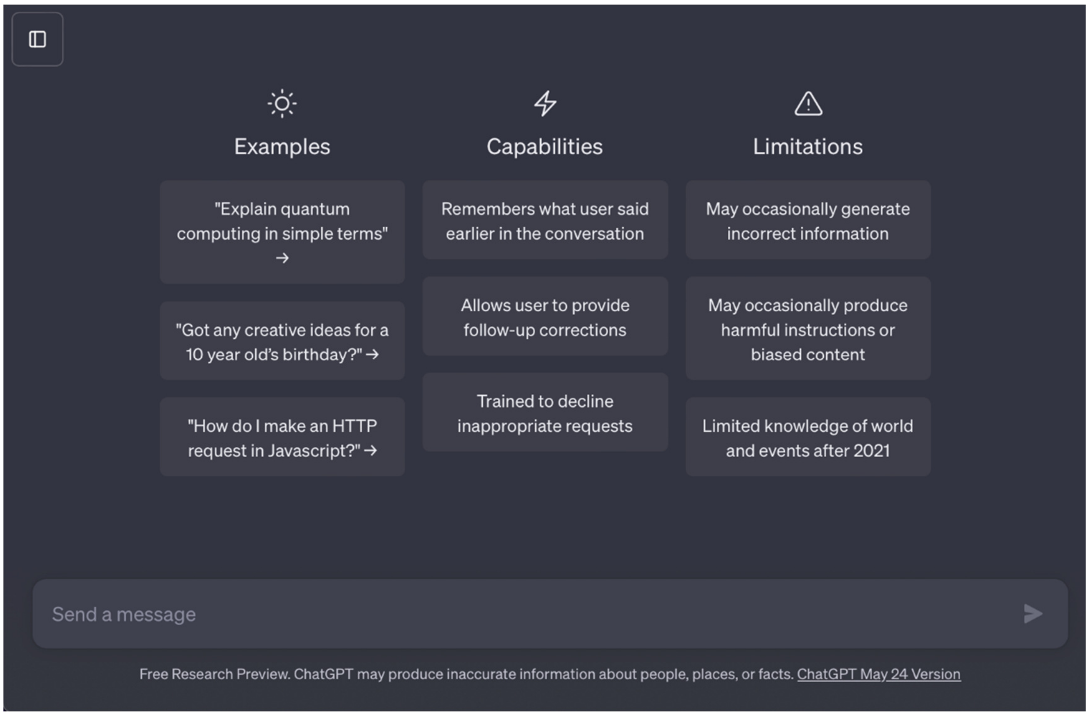

Figure 3. The web user interface of ChatGPT.

Google's Bard is a conversational AI service powered by LaMDA (Language Model for Dialog Applications) (Thoppilan et al., 2022), a model proposed by Google as a part of dialogue-oriented LLMs. When it comes to generating answers through Bard, the Python package Bard (Cheong, 2024) was employed. Just like ChatGPT, Bard is accessible online as well. The web user interface of Bard is presented in Fig. 4. A qualitative comparison of ChatGPT and Bard in terms of (i) online availability, (ii) API availability, (iii) multilingual support, (iv) license type, (v) owned company, (vi) integrated products, (vii) launch year, and (viii) average Google Trends index for the past 12 months as of the time of writing this manuscript is given in Table 2.

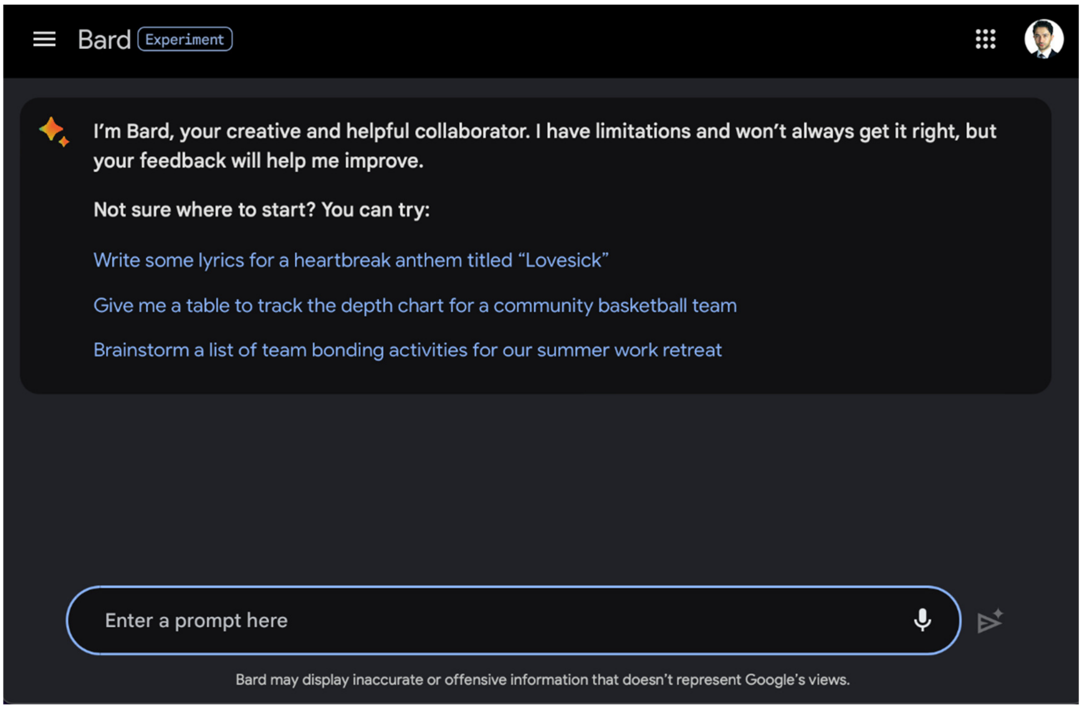

Figure 4. The web user interface of Bard.

Table 2. A qualitative comparison of ChatGPT and Bard.

Feature	ChatGPT	Bard
Online availability	Yes[1]	Yes[2]
API availability	Yes	Yes
Multilingual support	Yes	Yes
License type	Proprietary software	Proprietary software
Owned company	OpenAI	Google
Integrated products	Microsoft's products (e.g., Bing, Microsoft Office, and Microsoft Edge)	Google products (e.g., Google Search, Google Drive, and Google Assistant)
Launch year	2022	2023
Google Trends index	40	2

A Python script was implemented to read the questions from the dataset (.csv file) and generate answers through the aforementioned APIs. The answers generated by the chatbot APIs were stored in a .csv file. An overview of the proposed architecture is presented in Fig. 5.

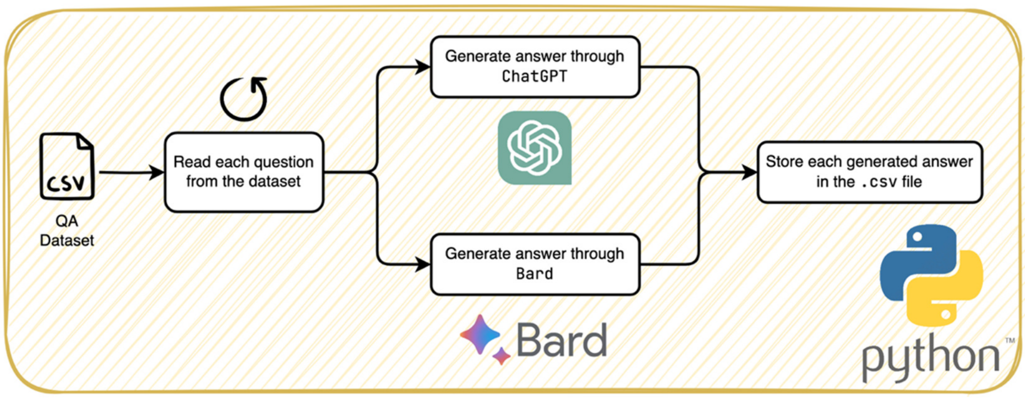

Figure 5. An illustration of the overview of the architecture of the proposed architecture to generate answers through the APIs provided for ChatGPT and Bard.

[1] http□://chat.openai.com
[2] http□://bard.□oo□le.com

The software implemented for the proposed study is based on the Software Development Kit (SDK) of the Python 3 programming language, as well as packages written for Python 3. Apart from the described packages, pandas (The pandas development team, 2020), one of the most popular Python packages, was employed for data manipulation. Both matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) were employed for data visualization. The whole software stack employed for this study, including each software's exact version and description, is listed in Table 3.

Table 3. The list of software stack employed for this study.

Software	Version	Description
macOS	Ventura 13.4.1	Operating system
Python	3.10	Programming language
pandas (The pandas development team, 2020)	2.0.0	Data manipulation
NLTK (Bird et al., 2009)	3.8.1	NLP
GPT4All (Anand et al., 2024)	1.0.10	API for ChatGPT
Bard (Cheong, 2024)	1.4.0	API for Bard
matplotlib (Hunter, 2007)	3.7.1	Data visualization
seaborn (Waskom, 2021)	0.12.2	Data visualization

3.3. Evaluation Metrics

The QA dataset contains the answers (a.k.a., ground truth) to the questions. The proposed architecture generates two answers for each given question: one by ChatGPT and one by Bard, as illustrated in Fig. 5. As discussed in detail in (Caldarini et al., 2022), there does not exist a de facto standard evaluation method for chatbots, as each evolution metric does have its own limitation(s). To evaluate the accuracy of these generated answers, a widely used text similarity measurement algorithm, namely cosine similarity, was employed. The cosine similarity can be formulated as follows: Given two n-dimensional vectors of attributes, X and Y, the equation of the cosine similarity of X and Y, $\cos(\theta)$, is given in Formula 1:

$$\cos(\theta) = \frac{X.Y}{||X|| \, ||Y||} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} {X_i}^2} \sqrt{\sum_{i=1}^{n} {Y_i}^2}} \tag{1}$$

The novelty of the proposed answer-scoring algorithm comes from the ability to detect the question types. The proposed answer-scoring algorithm starts with the detection of question types instead of directly obtaining cosine similarity. If the type of the question was recognized as yes_no and the generated answer was correct (based on the comparison of the starting words of the generated answer and the ground truth), then the cosine similarity score was assigned to 1.0. It is worth mentioning that we intentionally set that score manually, as chatbots generally do generate long answers (descriptions), even for the yes_no questions, while the answers available in the dataset do not. For all other cases, the cosine similarity score was calculated through NLTK (Bird et al., 2009), a widely used open-source NLP library. The proposed answer similarity measurement algorithm is illustrated in Fig. 6.

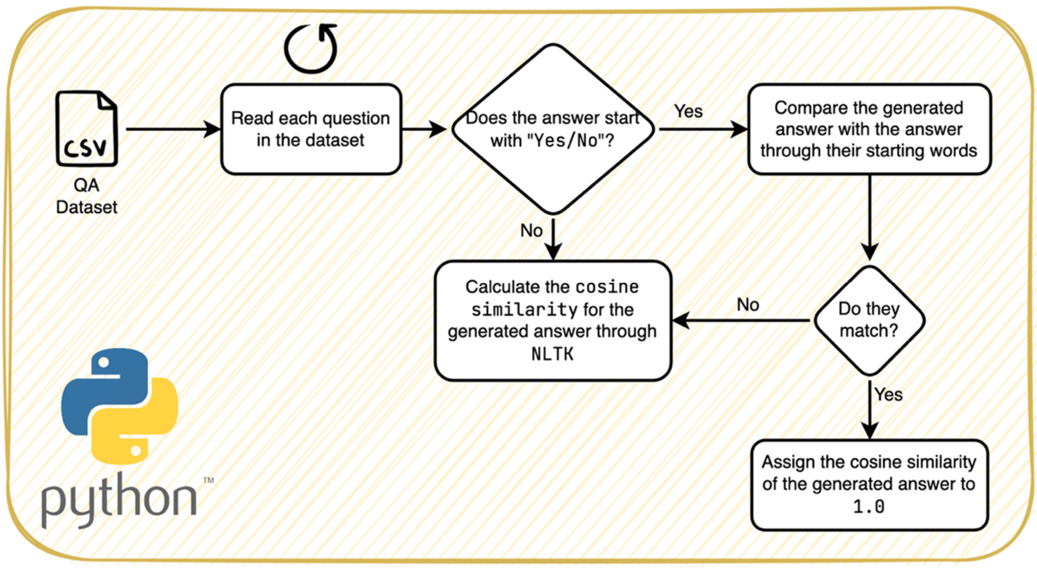

Figure 6. An illustration of the proposed novel answer-scoring algorithm based on cosine similarity.

4. Results and Discussion

The conducted experiments and their results are described and discussed in this section. More specifically, through the conducted experiments, the employed chatbots were investigated through their (i) generated answer length, and (ii) generated answer accuracy. Each of these evaluation metrics is described in the following subsections.

4.1. Answer Length

When the lengths of the answers generated by the chatbots were investigated, it was observed that Bard generated lengthy answers compared to ChatGPT, as the lengths of the answers generated by the chatbots for the given questions are plotted in Fig. 7. When this experimental result was evaluated cumulatively, the average length (number of words) of the answers generated by ChatGPT and Bard were calculated as 40 and 226, respectively, as plotted in Fig. 8. The list of the answers provided by ChatGPT and Bard for a set of sample questions, along with their respective answer scores, is given in Table 4.

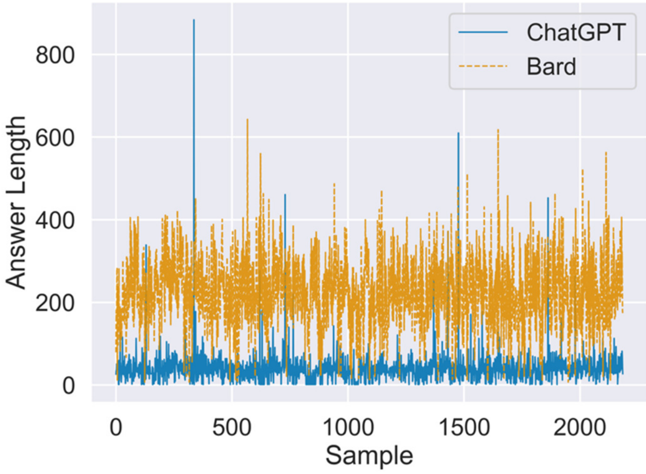

Figure 7. The plot of the lengths (number of words) of each answer generated by the chatbots for the given questions.

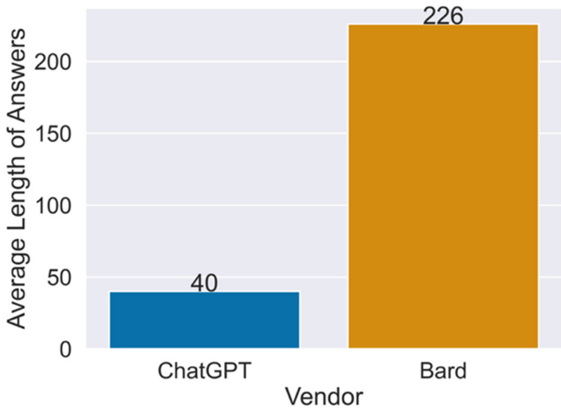

Figure 8. The plot of the average length of the answers generated by the chatbots for the given questions.

Table 4. The list of the answers provided by both ChatGPT and Bard for a set of sample questions, along with their respective answer scores (continued on the next page).

Question	Ground Truth	Answer Generated by ChatGPT	Answer Generated by Bard	ChatGPT Score	Bard Score
Who did Alessandro Volta marry?	Alessandro Volta married Teresa Peregrini.	He married twice. His first wife was Teresa Magri; they were wed on December 25, 1839 but later divorced due to their differences of opinion regarding the education of children. Later he got remarried with Eleonora Balbi and had a son named Carlo Volta who became an engineer like his father Alessandro Votta.	Alessandro Volta married Maria Teresa Alonsa Peregrini in 1794. They had three sons together: Zanino, Flaminio, and Luigi. The marriage was a happy one, and Volta was a devoted husband and father.	0.24	0.25

Question	Ground Truth	Answer Generated by ChatGPT	Answer Generated by Bard	ChatGPT Score	Bard Score
Are the largest turtles aquatic?	yes	Yes, most turtle species live in or near water for at least part of their lives and some spend all or nearly all of their time under water. However, there are a few exceptions to this rule such as land tortoises that can go months without drinking any liquid but still need access to fresh air through the shell opening on top of their heads.	Yes, the largest turtles are aquatic. The largest living turtle is the leatherback sea turtle, which can grow up to 6 feet long and weigh up to 2,000 pounds. Leatherback sea turtles are found in all of the world's oceans, except for the Arctic Ocean. They are excellent swimmers and can travel long distances. Their diet consists mainly of jellyfish, which they eat by suction. The other two largest species of turtles are also aquatic: the green sea turtle and the hawksbill sea turtle. Green sea turtles can grow up to 5 feet long and weigh up to 1,500 pounds. They are found in tropical and subtropical waters around the world. Their diet consists mainly of seagrass and algae. Hawksbill sea turtles can grow up to 4 feet long and weigh up to 100 pounds. They are found in tropical and subtropical waters around the world. Their diet consists mainly of sponges, sea urchins, and other invertebrates. There are a few species of large turtles that are not aquatic. The largest of these is the Galapagos giant tortoise, which can grow up to 4 feet long and weigh up to 500 pounds. Galapagos giant tortoises are found on the Galapagos Islands in the Pacific Ocean. Their diet consists mainly of grasses, leaves, and fruits. So, the answer to your question is yes, the largest turtles are aquatic. They have evolved to live in the water and are well-adapted to this environment.	1.0	1.0
What is cello an abbreviation of?	violoncello	Cello, which stands for ""Cellular LTE"" (Long Term Evolution), refers to wireless communication technology that operates on Long-Term Evolution standards and uses frequencies in the range between 210 MHz and 800Mhz.	The cello is an abbreviation of the Italian word violoncello, which means ""small violone"". The violone was a large stringed instrument that was popular in the 16th and 17th centuries. The cello is a smaller version of the violone, and it was developed in the early 17th century. The cello is also sometimes abbreviated	0.0	1.0

Question	Ground Truth	Answer Generated by ChatGPT	Answer Generated by Bard	ChatGPT Score	Bard Score
			as vlc or vc. These abbreviations are used in musical scores and other music-related documents. Here is a table showing the different abbreviations for cello: Abbreviation ǀ Meaning ------- ǀ -------- Cello ǀ Violoncello Vlc ǀ Violoncello Vc ǀ Violoncello I hope this helps! Let me know if you have other questions		

4.2. Answer Accuracy

One of the key aspects of using a chatbot from a user perspective is getting the right answers to the given questions. As described before, the cosine similarity algorithm was employed to evaluate the accuracy of the answers generated by the covered chatbots. According to the conducted experiment, the average similarity scores of the answers generated by ChatGPT and Bard were calculated as 0.43 and 0.46, respectively, as presented in Fig. 9. According to this experimental result, it is safe to conclude that Bard provided answers more similar to the ground truth compared to ChatGPT. The disparities in performance can be attributed to variations in model size, training data, fine-tuning techniques, answer generation strategies, and disparities in contextual understanding.

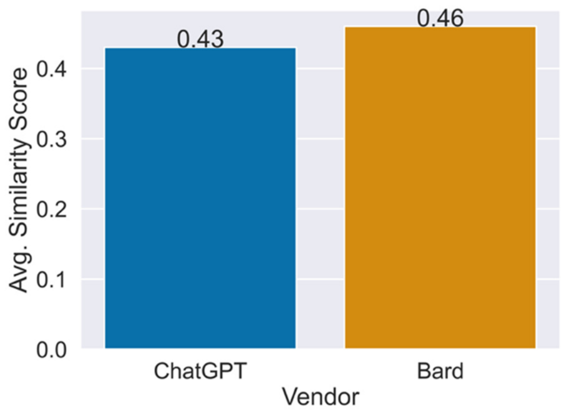

Figure 9. The plot of the obtained average similarity scores of the answers generated by the chatbots for the given questions.

5. Conclusion

AI has become an indispensable component of our everyday existence, significantly enhancing convenience and efficiency for humanity. One of the latest products that AI has brought into daily life is the concept of a conversational chatbot. In this study, two of the most popular chatbots, namely ChatGPT and Bard, were both quantitatively and qualitatively compared. To quantitatively compare these chatbots' question-answering capabilities, a publicly available QA dataset introduced by Carnegie Mellon University and the University of Pittsburgh was used. This dataset consists of 2,390 questions from 109 diversified topics. The evaluation metrics used to quantitatively compare the chatbots were (i) the lengths of generated answers and (ii) the accuracy values of the generated answers, which were obtained through the proposed novel answer-scoring algorithm based on cosine similarity. According to the experimental result, (i) the average length (number of words) of the answers generated by ChatGPT and Bard were calculated as 40 and 226, respectively, and (ii) the average similarity scores of the answers generated by ChatGPT and Bard were calculated as 0.43 and 0.46, respectively. These experimental results imply that (i) Bard generated lengthy answers compared to ChatGPT, and (ii) Bard provided answers more similar to the ground truth compared to ChatGPT. LLMs demonstrate proficiency in executing various NLP tasks, encompassing text generation, summarization, translation, and question answering, among others.

In future work, our aim is to incorporate additional evaluation metrics when comparing the performance of chatbots. Furthermore, we intend to expand our scope to encompass more AI-powered chatbots, such as Microsoft Copilot and GPT-4. Lastly, we seek to validate our findings by utilizing alternative datasets.

References

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., Cielo, D., Oyelese, A. A., Doberstein, C. E., Telfeian,
A. E., Gokaslan, Z. L., & Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. MedRxiv, 2023.04.06.23288265, 1–23. https://doi.org/10.1101/2023.04.06.23288265

Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B., & Mulyar, A. (2024). GPT4All: Training an Assistant-style Chatbot with Large
Scale Data Distillation from GPT-3.5-Turbo. GitHub. https://github.com/nomic-ai/gpt4all

Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., & Botchu, R. (2023). A comparison of ChatGPT-generated articles with human-written articles. Skeletal Radiology, 52, 1755–1758. https://doi.org/10.1007/s00256-023-04340-5

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., & Teo, J. T. (2023). AI chatbots not yet ready for clinical
use. Frontiers in Digital Health, 5, 1–5. https://doi.org/10.3389/fdgth.2023.1161098

Bernardini, A. A., Sônego, A. A., & Pozzebon, E. (2018). Chatbots: An Analysis of the State of Art of Literature. Proceedings of the 1st Workshop on Advanced Virtual Environments and Education (WAVE2 2018). https://doi.org/10.5753/wave.2018.1

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
(1st ed.). O'Reilly Media.

Caldarini, G., Jaf, S., & McGarry, K. (2022). A Literature Survey of Recent Advances in Chatbots. Information, 13(1), 1–22. https://doi.org/10.3390/info13010041

ChatGPT. (2024). OpenAI. https://chat.openai.com

ChatGPT, Bard, Microsoft Copilot - Explore - Google Trends. (2024). Google Trends. https://trends.google.com/trends/explore?date=today%203-m&q=/g/11khcfz0y2,/g/11ts49p01g,/g/11tsqm45vd&hl=en

Cheong, A. (2024). Python SDK/API for reverse engineered Google Bard. GitHub. https://github.com/acheong08/Bard

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. H. (2023). ChatGPT
versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S. A.R.,
Singapore, Ireland, and the United Kingdom). PLoS ONE, 18(8), 1–12. https://doi.org/10.1371/journal.pone.0290691

Erica - Virtual Financial Assistant From Bank of America. (2024). Bank of America. https://promotions.bankofamerica.com/digitalbanking/mobilebanking/erica

Garfinkle, A. (2023). ChatGPT on track to surpass 100 million users faster than TikTok or Instagram: UBS. Yahoo Finance. https://finance.yahoo.com/news/chatgpt-on-track-to-surpass-100-million-users-faster-than-tiktok-or-instagram-ubs-214423357.html

Gemini Team. (2023). Gemini: A Family of Highly Capable Multimodal Models. ArXiv, 2312.11805, 1–62.

Google Trends. (2024). Google. https://trends.google.com

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. ArXiv, 2301.07597, 1–20.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. Scientific Reports, 13, 1–13. https://doi.org/10.1038/s41598-023-45644-9

Hristidis, V., Ruggiano, N., Brown, E. L., Ganta, S. R. R., & Stewart, S. (2023). ChatGPT vs Google for Queries Related to Dementia
and Other Cognitive Decline: Comparison of Results. Journal of Medical Internet Research, 25, 1–13. https://doi.org/10.2196/48966

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hulman, A., Dollerup, O. L., Mortensen, J. F., Fenech, M. E., Norman, K., Støvring, H., & Hansen, T. K. (2023). ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. PLoS ONE, 18(8), 1–10. https://doi.org/10.1371/journal.pone.0290773

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Introducing UoA Assistant. (2021). University of Auckland. https://www.auckland.ac.nz/en/on-campus/life-on-campus/latest-student-
news/student-services-function-review/introducing-uoa-assistant.html

Li, S. W., Kemp, M. W., Logan, S. J. S., Dimri, P. S., Singh, N., Mattar, C. N. Z., Dashraath, P., Ramlal, H., Mahyuddin, A. P., Kanayan, S., Carter, S. W. D., Thain, S. P. T., Fee, E. L., Illanes, S. E., Choolani, M. A., Rauff, M., Biswas, A., Low, J. J. H., Ng, J. S., … Lim, M. Y. (2023). ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and
gynecology. American Journal of Obstetrics and Gynecology, 229(2), 1–12. https://doi.org/10.1016/j.ajog.2023.04.020

Liu, J., Tang, X., Li, L., Chen, P., & Liu, Y. (2023). ChatGPT vs. Stack Overflow: An Exploratory Comparison of Programming Assistance Tools. 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C 2023),
364–373. https://doi.org/10.1109/QRS-C60940.2023.00105

Microsoft Copilot. (2024). Microsoft. https://copilot.microsoft.com

Nirala, K. K., Singh, N. K., & Purani, V. S. (2022). A survey on providing customer and public administration based services using AI: chatbot. Multimedia Tools and Applications, 81, 22215–22246. https://doi.org/10.1007/s11042-021-11458-y

Nuruzzaman, M., & Hussain, O. K. (2018). A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. Proceedings of the 2018 IEEE 15th International Conference on E-Business Engineering (ICEBE 2018), 54–61. https://doi.org/10.1109/ICEBE.2018.00019

Paliwal, S., Bharti, V., & Mishra, A. K. (2020). Ai Chatbots: Transforming the Digital World. In Recent Trends and Advances in Artificial Intelligence and Internet of Things (pp. 455–482). Springer. https://doi.org/10.1007/978-3-030-32644-9_34

Peyton, K., & Unnikrishnan, S. (2023). A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs. Results in Engineering, 17, 1–6. https://doi.org/10.1016/j.rineng.2022.100856

Rahaman, Md. S., Ahsan, M. M. T., Anjum, N., Rahman, Md. M., & Rahman, Md. N. (2023). The AI Race is On! Google's Bard and
OpenAI's ChatGPT Head to Head: An Opinion Article. Social Science Research Network, 1–6. https://doi.org/10.2139/ssrn.4351785

re — Regular expression operations. (2024). Python. https://docs.python.org/3.12/library/re.html

Shaji George, A., Hovan George, A., & Martin, Asg. (2023). A Review of ChatGPT AI's Impact on Several Business Sectors. Partners
Universal International Innovation Journal, 1(1), 9–23.

Shewale, R. (2023). 62 Chatbot Statistics For 2024 (Usage, Challenges & Trends). Demandsage. https://www.demandsage.com/chatbot-statistics/

Smith, N. A., Heilman, M., & Hwa, R. (2008). Question Generation as a Competitive Undergraduate Course Project. Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge, 1–3.

Sora. (2024). OpenAI. https://openai.com/sora

Sousa, D. N., Brito, M. A., & Argainha, C. (2019). Virtual customer service: Building your chatbot. Proceedings of the 3rd International Conference on Business and Information Management (ICBIM '19), 174–179. https://doi.org/10.1145/3361785.3361805

The pandas development team. (2020). pandas: Python Data Analysis Library. https://pandas.pydata.org

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H.,
Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., … Le, Q. (2022). LaMDA: Language Models
for Dialog Applications. ArXiv Preprint, 2201.08239, 1–47. http://arxiv.org/abs/2201.08239

Waisberg, E., Ong, J., Masalkhi, M., Zaman, N., Sarker, P., Lee, A. G., & Tavakkoli, A. (2023). Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. Eye (Basingstoke), 1–4. https://doi.org/10.1038/s41433-023-02760-0

Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 1–4. https://doi.org/10.21105/joss.03021