


Review

# A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022)

Chien-Chang Lin , Anna Y. Q. Huang and Stephen J. H. Yang \*Department of Computer Science and Information Engineering, National Central University,  
Taoyuan 320317, Taiwan

\* Correspondence: stephen.yang.ac@gmail.com

**Abstract:** A conversational chatbot or dialogue system is a computer program designed to simulate conversation with human users, especially over the Internet. These chatbots can be integrated into messaging apps, mobile apps, or websites, and are designed to engage in natural language conversations with users. There are also many applications in which chatbots are used for educational support to improve students' performance during the learning cycle. The recent success of ChatGPT also encourages researchers to explore more possibilities in the field of chatbot applications. One of the main benefits of conversational chatbots is their ability to provide an instant and automated response, which can be leveraged in many application areas. Chatbots can handle a wide range of inquiries and tasks, such as answering frequently asked questions, booking appointments, or making recommendations. Modern conversational chatbots use artificial intelligence (AI) techniques, such as natural language processing (NLP) and artificial neural networks, to understand and respond to users' input. In this study, we will explore the objectives of why chatbot systems were built and what key methodologies and datasets were leveraged to build a chatbot. Finally, the achievement of the objectives will be discussed, as well as the associated challenges and future chatbot development trends.

**Keywords:** conversational chatbot; chatbot; dialogue system; dialogue response; dialogue strategies; dialogue generation; machine learning; conversational agents



check for updates

**Citation:** Lin, C.-C.; Huang, A.Y.Q.; Yang, S.J.H. A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability* **2023**, *15*, 4012. <https://doi.org/10.3390/su15054012>

Academic Editors: Andreas Kanavos and Hao-Chiang Koong Lin

Received: 29 January 2023

Revised: 11 February 2023

Accepted: 21 February 2023

Published: 22 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Conversational chatbots, also known as chatbots or dialogue systems, are software programs designed to simulate conversation with human users, especially over the Internet. These chatbots can be used in various contexts, such as customer service, information acquisition, educational support, and entertainment. Conversational chatbots have become an increasingly popular tool in recent years. These computer programs are designed to simulate conversations with human users by utilizing Natural Language Processing (NLP) techniques like ChatGPT [1] and applying them to a variety of applications, including casual entertainment purposes [2,3], customer services [4,5], and educational purposes to assist teachers [6,7] and students [8–10]. Ranging from casual and open-domain to more domain-specific and fact-based, these chatbots are built using various deep learning models, such as RNN (Recurrent Neural Network), Seq2Seq (Sequence to Sequence), LSTM (Long Short-Term Memory), BERT [11] (Bi-directional Encoder Representation from Transformers), GPT (Generative Pre-trained Transformer), as well as leveraging different training techniques, such as reinforcement learning or transfer learning, in order to improve the performance of NLP algorithms and chatbots. A recent popular, encouraging example is the success of ChatGPT [1], which received a great amount of attention and inspired researchers to generate more ideas regarding chatbot applications.

Despite advances in technology, there are still many challenges that must be addressed to create chatbots that truly capture the context, style, emotion, and character of human conversations. Researchers have explored more complex and nuanced conversations

through chatbots, using context and reasoning to better understand user intent [12–14], and have even further explored how to design chatbot interfaces that are more user-friendly [15,16]. In our research, we defined our research agenda during the idealization phase and the scope for the surveyed paper collection, review, and discussion. The overall agenda can be seen below:

- A. Overview
  - a. Define the year range of the surveyed papers.
  - b. Define the keywords and screening criteria of the surveyed papers.
  - c. Review, study, and categorize the data of all surveyed papers.
- B. What is the purpose of building chatbots?
  - a. Why are chatbots built?
  - b. What issues are people trying to resolve?
  - c. How are chatbots used in specific areas?
  - d. Who are the target users of chatbots?
- C. How are chatbots built?
  - a. What are the technical considerations when building chatbots?
  - b. What key machine learning models are used in chatbots?
  - c. What training techniques are used in chatbots?
- D. What are the overall outcome and challenges of chatbots?
  - a. Are the objectives and intentions met?
  - b. What are the limitations and challenges?
- E. What are the future development and research trends of chatbots?
  - a. What are the conclusions of chatbot research so far?
  - b. What other potential areas can be applied to chatbots?
  - c. What will be the future development trend?

Overall, we based our research agenda on and explored objectives regarding why researchers built chatbots for either opened or closed domains. Based on the overview study, we also explored the typical challenges they faced. Additionally, we explored different types of strategies for allowing the chatbot to learn and keep the dialogue context consistent. Finally, there is another important aspect we need to address and discuss, which is how the chatbot or dialogue system is constructed by machine learning models and how it is trained by machine learning techniques. Thus, in this paper, the following research questions are summarized from the research agenda and discussed:

- **RQ1:** What are the objectives of building conversational chatbots?
- **RQ2:** What are the methods and datasets used to build conversational chatbots?
- **RQ3:** What are the outcomes and challenges of conversational chatbots?

In Section 2, we start with a literature review, which is related to how chatbots have been built and how researchers have applied chatbots in real application areas such as business or educational support. We also address some papers in which the purpose of building chatbots was purely focused on technical improvement. In Section 3, we explain how the reference list was developed, with details of the search criteria used in our research method. We then share and comment on the findings from the surveyed papers in Section 4, which addresses different aspects and comments on chatbot dialogue systems. Finally, we summarize the overall research and potential future development areas, and the trends in this technology in the Conclusion.

## 2. Literature Review

Recently, chatbots have gained popularity due to advances in natural language processing and machine learning. These techniques allow chatbots to understand and respond to user input in a more human-like manner, making interactions more seamless and natural. The most recent popular chatbot application is ChatGPT [1], which features a dramatically

improved language model and chatting experience with a chatbot. The other major application area is leveraging chatbots in educational support, where a chatbot can either be a learning companion [8,9] to improve learners' comprehension skills or a simulated student to improve teachers' efficacy [6,7].

According to psychological research, joy and meaningful conversation often go hand in hand. Thus, as more and more people have become digitally connected in the age of social media, social chatbots have emerged as an important alternative to engagement. Different from earlier chatbots designed for chatting, the Xiaoice social chatbot developed by Microsoft is designed to serve users' needs for communication, emotion, and social belonging, and is endowed with empathy, personality, and skills; integrating emotional intelligence optimizes user engagement in the long run [17].

In addition to Microsoft, other studies have tried using SeqGAN (Sequential Generative Adversarial Network) to design an emotional human–computer dialogue generation method. Although its performance is not as expected compared to the related work, their model can still generate responses that are human-like not only in content, but also in emotion; this means that they can obtain less 'safe' responses in terms of content, but have a certain degree of emotion [12].

There have been many studies on the effectiveness of chatbots in various contexts. For example, some studies have found that chatbots can be a useful tool for customer service, as they can handle a high volume of queries and provide quick and accurate responses. For example, Lei Cui and Shaohan Huang developed the SuperAgent [4], a customer service chatbot for e-commerce websites, which utilizes more large-scale, public, and crowdsourced customer data compared to traditional customer service chatbots. Another example is leveraging a chatbot in a customer care support center [5], which can provide better and more accurate responses to a customer's needs.

However, there are also limitations to the use of chatbots. Some users may find the interactions to be artificial or impersonal, and there is a risk that chatbots may not be able to fully understand or respond to more complex or nuanced input. In this case, some researchers have proposed a new method to identify customer emotions during conversations, such as happiness, anger, sadness, fear, and neutral states. With the input of customer sentiment, the proportion of chatbots taking correct actions has been greatly increased, thereby improving customer service optimization KPIs [5].

Another very popular approach is the application of Reinforcement Learning (RL) to chatbots, in order to improve their response generation, dialogue management, and response evaluation by maximizing a reward signal, which can be human feedback or automatic evaluation metrics. For example, Satinder and Diane applied RL to the problem of optimizing dialogue policy in a spoken dialogue system; their method employs relatively few exploratory dialogues and directly computes an optimal policy in a space that may contain thousands of policies [18]. In addition, Jiwei Li and Will Monroe introduced an RL framework for neural response generation by simulating dialogues between two agents, combining the advantages of neural sequences in sequence systems and RL for dialogue. Like earlier neural sequence-to-sequence models, their framework can generate words that optimize future rewards, successfully extracting the global properties of good dialogue [19].

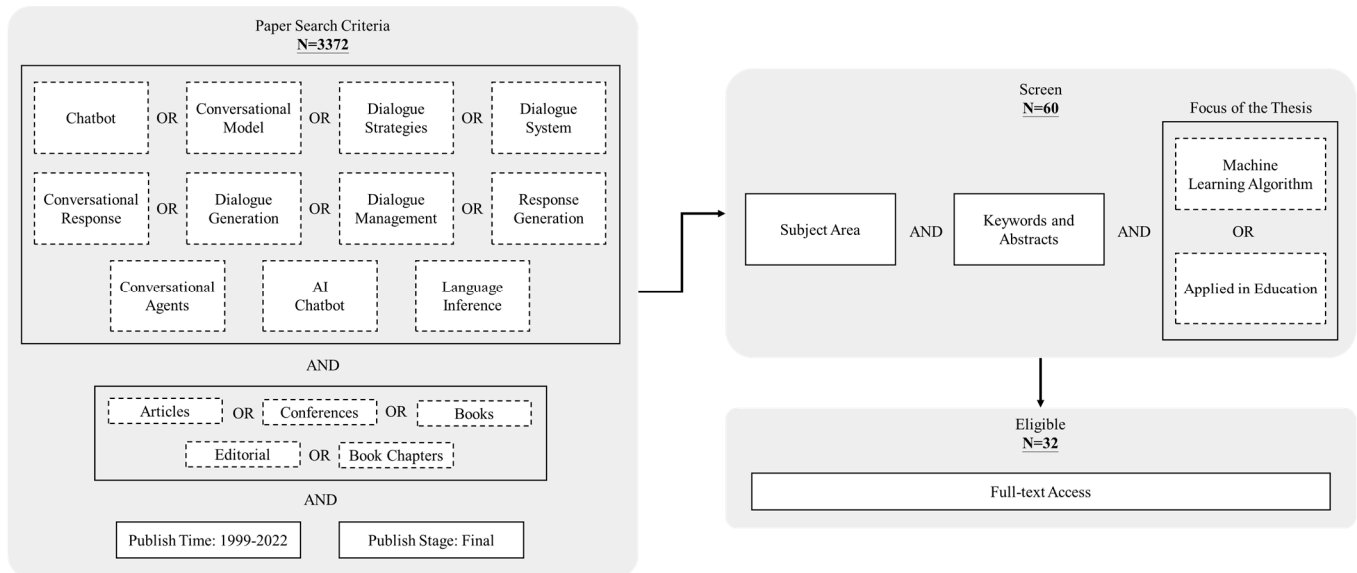
Overall, the literature on using RL for conversational chatbots suggests that RL can effectively improve the naturalness and engagement of chatbot responses. However, further research is needed to understand the best ways to apply RL in this context, including the optimal design of the reward signal and the most effective techniques for training and evaluating RL-based chatbots. In conclusion, chatbots show promise as a useful tool for a variety of applications, but more research is needed to fully understand their capabilities and limitations.

### 3. Review Methodology

#### 3.1. Process of the Survey Literature

This study reviews chatbot-related research from 1999 to 2022 through Scopus. The search strings were mainly keywords such as chatbots, dialogue systems, and response

generation. To refine the search results, this study limited the literature categories to articles, conference papers, book chapters, books, and editorials, and the publication stage was set as final. Figure 1 shows the paper selection criteria process for this study. According to the above conditions and confirming full-text evaluation, a total of 32 papers were screened out.



**Figure 1.** The paper selection criteria process.

The search string used was (TITLE-ABS-KEY ('chatbot' OR 'conversational model' OR 'dialog strategies' OR 'dialogue system' OR 'conversational response' OR 'dialogue generation' OR 'dialogue management' OR 'response generation' OR 'conversational agents' OR 'AI chatbot' OR 'language inference')) AND PUBYEAR > 1998 AND (LIMIT-TO (PUBSTAGE,'final')) AND (LIMIT-TO (DOCTYPE,'cp') OR LIMIT-TO (DOCTYPE,'ar') OR LIMIT-TO (DOCTYPE,'ch') OR LIMIT-TO (DOCTYPE,'bk') OR LIMIT-TO (DOCTYPE,'ed')) AND (LIMIT-TO (OA,'all')).

Based on the above search criteria, 3372 documents were found in early October 2022. Then, we filtered them according to the following conditions:

**Subject area:** Chatbots can be applied in multiple areas; this review paper will only focus on engineering, computer science, education, and social science.

**Keywords and Abstract:** This is the last stage, selecting suitable papers through keywords and abstracts as the dataset for this research.

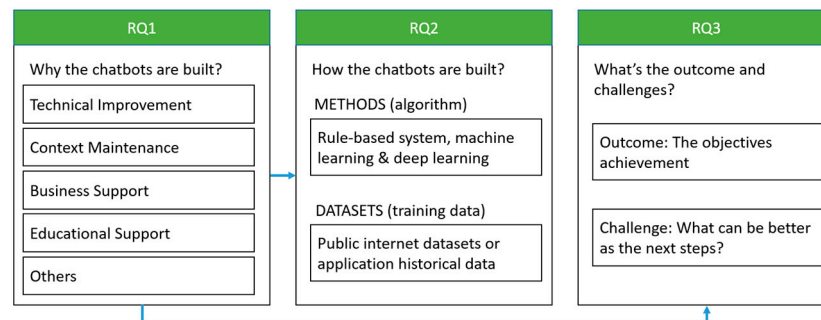
**Focus of the Thesis:** In this review, we mainly focus on papers related to machine learning algorithms used by chatbots and their applications in education.

### 3.2. Overview of Surveyed Method

A total of 3372 papers were collected in this study, and finally, only 32 papers were screened to analyze and discuss the content of the papers. To further explore the current development status and situation of conversational chatbots, Figure 2 shows the process of filtering and classifying the 32 selected papers according to the 3 RQs mentioned in the Introduction section.

Due to the rapid development of NLP technology, conversational chatbots have become increasingly popular human-like interaction tools in recent years. Therefore, RQ1 (What are the objectives of building conversational chatbots?) focuses on exploring the construction objectives of conversational chatbots and trying to solve key issues. The construction objectives can be categorized into three major areas, including (1) technical improvement in various areas, (2) strategy of finding and maintaining dialogue context,

(3) business support for increasing entertainment or potential revenue, (4) educational support, and (5) other specific objectives.



**Figure 2.** Considerations of surveyed papers for RQ.

To explore how to build conversational chatbots, RQ2 (What are the methods and datasets used to build conversational chatbots?) aims to investigate the construction methods and training datasets for building conversational chatbots. Moreover, this study of RQ2 will also explore the datasets or application-specific data to build and train a dialogue system. Finally, on the objectives of RQ1 for building a conversational chatbot, RQ3 (What are the outcomes and challenges of chatbots based on the objectives for which they were built?) will further explore the outcomes and challenges present in developing chatbots.

## 4. Findings and Discussion

### 4.1. Overview of Conversational Chatbots

To explore the institutions implementing conversational chatbots, Table 1 shows the number of papers published by country. Most of the institutions were found in the United States, with the others being in the UK, China, Germany, etc. Conversational chatbots can generally be divided into closed-domain and open-domain chatbots. Of the 32 screened papers, 9 and 23 papers were related to conversational chatbots with closed domains and open domains, respectively. Typically, closed-domain chatbots are primarily built to support specific conversational goals. A typical example of a closed-domain chatbot is found in a customer contact center when a customer requests support from customer service [5].

**Table 1.** Country distribution of surveyed papers.

Region	Country	Count	Reference
North America	USA	15	[1,2,4,6,7,9,11,13,14,17,18,20–23]
	Canada	1	[16]
APAC	China	2	[12,24]
	Taiwan	2	[8,10]
	Japan	1	[25]
	Singapore	1	[26]
	Indonesia	1	[27]
	India	1	[28]
Europe	UK	3	[3,29,30]
	Germany	1	[19]
	Australia	1	[15]
	France	1	[31]
	Greece	1	[32]
	Poland	1	[5]

The other main closed-domain chatbot application is educational support, since this kind of chatbot has very clear conversation goals in terms of learning perspectives. In a



closed-domain chatbot, the users care more about the information accuracy rather than if the response is more like a human. The open-domain chatbot is more focused on naturally chatting with users, so the goal is more to keep the consistency of the conversational context and the users' engagement. This is why deep learning techniques such as reinforcement learning have been widely used to build conversational chatbots in recent years [9].

#### 4.2. Objectives of Conversational Chatbots (RQ1)

Since chatbots are mostly built using objectives, Table 2 shows three categories of objectives retrieved from the 32 screened papers, including (1) technical improvement, (2) context maintenance, (3) business support, (4) educational objectives, and (5) other specific objectives. There might be multiple objectives achieved by one paper in our review. The objective of performance improvement means providing correct information, as well as being closer to a human-like conversation style. Accurate responses are a practical and basic requirement of chatbots since users may not care about whether the other side is a human or machine as long as users can obtain the information they need. Closed-domain conversational chatbots have achieved quite good performance [4,5,30], so accurate responses are primarily the key objective of open-domain chatbot research. The key technology is used to modify the objective function [21] or to integrate external domain knowledge into the responses [2].

To respond with correct and appropriate answers, chatbots have to identify users' statements well; the objective of context maintenance, which aims to find and maintain a good dialogue strategy, is another major objective that chatbots, especially open-domain chatbots, need to achieve. Of the surveyed papers, we found that most conversational chatbots in recent years have adopted the following four directions to find and maintain the dialogue context through a good strategy and policy, including dialogue context identification, dialogue strategy optimization, word embedding enhancement, and user engagement or connection maintenance. In this review, some researchers focused on identifying the dialogue context [22,28], and some focused on optimizing the strategy to keep this context [12,18]. There is always a theme around a conversation to make the conversation meaningful, and this is what this type of chatbot aims to do. For open-domain chatbots, unfortunately, the dialogue may change during the conversation, and some of the papers focus on not only identifying the context, but also detecting the change in the context. Normally, dialogue can use a probability model such as MDP [20,29] or deep learning [13,14,22] technologies to predict the direction of the dialogue. The purpose of this is to continuously engage with users' interests so that the conversation can be continued. Moreover, many authors are working on optimizing dialogue strategies and policies [12,18,20] to make the responses less conservative, thus increasing the variability in the dialogue.

In educational objective areas, the common practice of using a chatbot is to simulate a learning companion during the learning phase [8,9]. One of the main reasons for this goal is that it is almost impossible for a teacher to take care of every single student's learning progress based on their proficiency. Alternatively, the researcher can leverage a chatbot to simulate a student so that a preservice teacher can be trained [6].

In recent years, in the field of artificial intelligence, emotional topics have gradually become the focus of future research. For this reason, several studies have since begun to explore emotional aspects in conversational chatbots [12,19], either trying to detect users' feelings or becoming a cognitive, user-friendly, interactive, and empathetic system. Other specific goals, such as emotion, focus on enabling conversational chatbots to not only serve as problem-solving tools, but also chat with humans like friends, thereby meeting human emotional needs. For example, through communication with chatbots, we hope not only to eliminate the loneliness of the elderly living alone, but also to stimulate their brains.

In summary, in terms of the objectives of building conversational chatbots, there is a lot of focus on improving the response accuracy of the dialogue system in order to provide more human-like conversations. In terms of open-domain dialogue systems, there are also

papers focusing on how to identify the key dialogue context and maintain it so that users' interests can be kept.

**Table 2.** Objectives of conversational chatbot research.

Category	Item Description	Count	References
Technical Improvement	Response accuracy	7	[1,11,13,14,24,27,30]
	Integrate domain knowledge into responses	4	[1,2,11,15,31]
	Produce content-based responses	3	[1–3,11]
	Model objective functional enhancement	2	[5,21]
Context Maintenance	Identify dialogue context (or opinion)	6	[1,13,14,22,26,28]
	Optimize dialogue strategies (policy)	5	[1,12,18,20,29]
	Enhance word embedding (syntactic to semantic)	1	[22]
	Maintain users' engagement or connection	1	[9]
Business Support	Support entertainment	2	[2,3]
	Increase potential business revenue	2	[4,5]
Educational Support	Improve comprehension skills	3	[8–10]
	Enhance teaching efficacy	2	[6,7]
	Support collaborative learning	1	[32]
Other specific Objectives	Integrate emotion (human feeling) into responses	2	[12,19]
	Be cognitive, user-friendly, interactive, and empathetic	2	[15,16]

#### 4.3. Methods and Datasets of Conversational Chatbots

To explore how to build chatbots to reply to RQ2 (What are the methods and datasets used to build a conversational chatbot?), Tables 3 and 4 show the methods and datasets used in the surveyed papers for building conversational chatbots, respectively. In terms of the methods applied in our review, reinforcement learning was the most frequently used method [1,12–14,18,28]; this is one of three basic machine learning paradigms alongside supervised learning and unsupervised learning. Reinforcement learning differs from supervised learning in that it does not need to be presented with labeled input/output pairs, nor does it need to explicitly correct suboptimal actions. Instead, the focus is on finding a balance between exploring uncharted territory and developing current knowledge. One of the most successful cases recently published is ChatGPT [1], which has gained a lot of attention, and is basically trained by reinforcement learning. This is why reinforcement learning is commonly used to determine the conversation context, as well as keep the dialogue consistent. In the beginning, the reinforcement learning model was used to teach the machine to play computer games, where the researcher learned a lot about the policy adjustment needed to reach the goal.

**Table 3.** Methods of conversational chatbot research.

Category	Type of Method	Count	References
Machine Learning Training Techniques	Reinforcement Learning	7	[1,12–14,18,20,28]
	Supervised Learning	1	[20]
	Transfer Learning	1	[29]
Machine Learning Models	LSTM	4	[12,19,21,27]
	BERT	5	[11,15,24,30,31]
	RNN	3	[2,16,22]
	ELMO	2	[24,30]
	MDP	2	[20,29]
	GPT-3	1	[1]
	Seq2Seq RNN	1	[25]
Others	Specific Systems	6	[3,4,15–17,23]
	Experiment based	6	[6–10,32]

**Table 4.** Datasets used by conversational chatbots.

Dataset	Count	References
Twitter-Related Dataset	4	[19,21,22,25]
OpenSubtitles Dataset	3	[13,21,30]
MovieDic or Cornell Movie Dialog Corpus	3	[16,27,28]
Wikipedia and Book Corpus	5	[1,2,4,11,15]
Television Series Transcripts	2	[17,19]
SEMEVAL15	2	[4,30]
Amazon Reviews and Amazon QA	2	[2,30]
Amazon Mechanical Turk Platform	2	[3,26]
Foursquare	1	[2]
CoQA	1	[28]
Specific Application Historic Dataset	6	[5,12,14,18,20,31]
Others (e.g., course materials)	11	[6–10,17,23,24,29,30,32]

The second frequently used method is long short-term memory (LSTM), which improves the memory design of the original RNN by controlling four units (Input Gate, Output Gate, Memory Cell, and Forget Gate) to successfully process and predict significant events for intervals and time series delays. This is where the research started to aim to ensure that the dialogue system remembers the conversation context and that the response is consistent with earlier conversation goals [12,19,21,27]. The main reason why much research focuses on this area is simply that users will lose interest and become disengaged if the dialogue system is responding to something inconsistent with the original goal or context. Although LSTM is enhanced compared to RNN, the selected papers spanned approximately 20 years, and some older or special-purpose research studies still chose to use the original RNN or enhanced RNN as their method [2,22].

Several methods were used only in one or a few cases. Bidirectional Encoder Representations from Transformers (BERT), proposed by Google in 2018 as a pre-training technology for natural language processing (NLP), has become a ubiquitous baseline in NLP experiments in just over a year since its publication. The overall performance of BERT in NLP is outstanding in many areas, such as question answering and next-sentence prediction, which makes it popular for use in the dialogue system. It offers a pretty good baseline for the researchers to start with and then can be fine-tuned by specific domain knowledge [15,24,30]. Other methods adopted in these papers include ELMO, supervised learning, transfer learning, Seq2Seq, MDP, HRED, GAN, UNILM, GPT, Dialogflow, and NBT; most of these methods are also common algorithms in natural language processing. For some closed-domain chatbots, especially when used by educational supports [6,8,9], the methods focus more on how to conduct the experiment rather than addressing the system architecture of how the chatbots are built.

Table 4 presents the datasets used to support conversational chatbot research. The way researchers leverage datasets depends on the purpose of the chatbot. Obviously, if the chatbot is built for general conversation purposes, the popular public internet dataset is the best choice. However, if the chatbot is a closed-domain one and made for a unique purpose only, researchers may collect and use their own dataset.

Many of the datasets are Twitter-related, including conversation data, FireHouse, Persona, dialogue, and posts, since social media undoubtedly include tons of valuable daily social information. The second most used dataset is the subtitles and scripts of movies or dramas. Although they are written by screenwriters, they still have a high reference value. In addition, social media usually contain information focused on the latest popular topics, while movies and dramas can cover historical stories or past classics.

Other popular datasets are Wikipedia, Book Corpus, and Amazon review and QA; as mentioned in the above paragraph, some research tried to include domain knowledge in the response [15,29]. In this case, these informative sites will greatly benefit it. Moreover, two special datasets, CamRest676 and KVRET, were collected from the Amazon Mechanical Turk platform. After, there were also a few types of data used only in a few studies,



including SEMEVAL15, Foursquare, CoQA, and other specific application datasets (ATIS, NJ System, TOOT, ELVIS, ALE OXE, Contact Center).

As mentioned earlier, an application-specific dataset is normally used by closed-domain chatbots because these kinds of chatbots have a very clear goal when a conversation is happening. A good example is SuperAgent [4] and a customer contact system [5]; they leverage pre-defined or historic system-generated data to train the system since it is unique to the chatbot. Another example is when a chatbot is built as a learning companion to improve users' reading comprehension skills [8,9]; the dataset used to train the system includes the books the chatbot will use, which means that the conversation scope is fixed.

In summary, there was an obvious trend in our review that matched artificial intelligence technology's evolution in the last decade. The overall implementation started with specific closed-domain chatbot systems' development and then changed to adopt modern machine learning models, including Seq2Seq RNN, LSTM, and BERT. Along with this trend, researchers are also trying to resolve the issue of providing more reasonable responses when the legacy RNN model mainly uses the greatest likelihood objective function to generate a response. This kind of discussion and enhancement is related to the main weak point of the Seq2Seq RNN model when using the greatest likelihood objective function, which causes a typical response such as 'I don't know' when the dialogue system does not know how to respond. Some researchers enhanced the generative model [12] to enrich the output of the conversations. There are some chatbot systems that leverage reinforcement learning to maintain dialogue context consistency. The key point is to identify a strategy to find the dialogue context and engage with the users during the conversation. While the conversation context might change during a conversation, some researchers emphasize maintaining a consistent dialogue context.

#### *4.4. Outcomes and Challenges of a Conversational Chatbot*

In response to RQ3 (What are the outcomes and challenges of conversational chatbots?), in terms of conversational chatbots, Table 5 shows the outcomes corresponding to different objectives, while Table 6 shows the challenges of building chatbots. Based on the surveyed objectives from RQ1, the corresponding outcomes can be divided into three categories: (1) technical improvement, which focuses on output optimization; (2) context maintenance, which focuses on algorithm or model optimization; (3) educational support; (4) business support; and (5) other. There were a total of 15 papers focused on optimizing the output of chatbots. In terms of technical improvement as the construction objective, the optimization direction includes adding context [1,13,30], external knowledge [2] and skills content [3,15], in order to make the responses of chatbots more accurate. For the other objectives, optimization should be researched in a more humane and emotional direction [12,19].

The outcome of the algorithm or model optimization focuses on optimizing the algorithm techniques applied in chatbots, such as reinforcement learning, self-attention, transfer learning, or models such as LSTM, BERT, GPT-3, and NBT. The outcome of this portion is more focused on achieving the objectives of context maintenance in conversational chatbots. To improve reinforcement learning to find and maintain dialogue context, studies [1,14,18,28] have focused on system architecture enhancement or the machine learning model's re-design; the ultimate objective is to prove that the enhanced dialogue system model will generate a better and human-like response. The recently published ChatGPT [1] is a good example of leveraging reinforcement learning, which has received lots of attention from researchers.

Other outcomes from the surveyed papers include educational support improvement, optimized inputs, and a dedicated model for e-commerce. Leveraging a chatbot as a learning companion [8,9] improves students' reading skills and maintains the students' engagement level continuously. When simulating the chatbot as a student to train preservice teachers in teaching school violence topics [6] or mathematics [7], only some areas were improved. The work in [5] focused on the input part of the model; the researchers tried

to minimize the input noise through various noise removal algorithms to maximize the usability of the information input into the model. The case involved a customer contact center [5] removing unnecessary information so that the dialogue system could be trained with meaningful information. Another special article created a dedicated model for e-commerce [4].

**Table 5.** Outcomes of conversational chatbots.

Category	Outcome	#	References
Technical Improvement (Output Optimization)	Include Context	6	[1–3,11,13,30]
	Skills	4	[2,3,15,21]
	External Knowledge	3	[2,15,31]
	Personality and Emotion	2	[12,19]
Context Maintenance (Algorithm or Model Optimization)	Reinforcement Learning	4	[1,14,18,28]
	BERT	4	[11,15,24,30]
	GPT-3	1	[1]
	LSTM	1	[19]
	Self-Attention	2	[11,24]
	Transfer Learning	1	[31]
Educational Support	NBT (Neural Belief Tracker)	1	[3]
	English Skills Improvement	3	[8–10]
	Teaching Efficacy Improvement	2	[6,7]
Business Support	Learning Result Improvement	1	[32]
	Optimized Input (Noise Removal)	1	[5]
	Dedicated Model for E-commerce	1	[4]
	Entertainment and Fun Support	2	[2,3]

**Table 6.** Challenges of conversational chatbots.

Challenge	Count	References
Best (or Better) Models Selection and Modification	8	[1,16,17,19,26,27,29,30]
More Efficient Pre-work for System Training	4	[4,5,18,20]
More Efficient Information Extraction and Classification	3	[20,21,26]
Good Diversity and Quantity of Training Data	6	[3,4,7,8,12,32]
More Dynamic Profile/Strategy Adjustment	3	[6,9,10]
Defining Best Objective Function Formulation	1	[20]
Better Feature Selection	1	[18]
Humanization and Moral Enhancement	1	[12]

In response to the challenges faced by chatbots mentioned in RQ3, Table 6 shows the challenges reported in research on conversational chatbots. Notably, up to seven papers faced challenges in model selection and modification. A typical challenge was that the researcher was not satisfied with the output, and they were looking for a better machine learning model or further enhancement [1,3,20,27]. From immature natural language processing technology used in the early stage of development, which limited the resources that researchers could rely on, to its relative maturity in recent years, the selection of algorithms has always been the most critical and challenging part.

How to select the model that best suits research with numerous algorithms for modification and optimization, or even combining multiple algorithms for design and training to obtain the highest accuracy, is a big problem. Luckily this challenge did not last for too long, as machine learning and natural language process technology have experienced a big breakthrough in the last decade. It has almost become a standard configuration to choose the most powerful NLP model along with reinforcement training and a better model such as BERT [15,24,30,31] or GPT-3 [1]; this provides a baseline of natural conversational response, and then adds any application or domain-specific knowledge.

In addition to model selection, inefficient pre-work is definitely another challenge: the data collected in reality must not be very neat, and researchers have to put in a lot of effort to make it trainable. A typical example is an application-specific dialogue system [4] or contact center [5], in which the training dataset might not be efficient for a machine learning model to mature enough to respond to a user's request. Although leveraging public datasets can make the system respond naturally, the real purpose of the closed-domain dialogue system is to provide accurate service and information. A typical method to conquer this challenge is to use data augmentation, where we can translate the raw data to another language and translate them back to the language we need. This might help enrich the dataset in some way, but people still believe that the dialogue system will become more experienced after increasingly real conversation data are fed in.

When using chatbots to provide educational support during the learning cycle, one of the common challenges is making the chatbot dynamically adjust the difficulty level by itself during a conversation. When a chatbot is used as a learning companion [8,9], enabling the chatbot to dynamically detect the learner's progress and adjust the profile to continuously push the learner to the next level is a potential area for enhancement. On the other hand, a similar challenge was also addressed when using a chatbot as a simulated student [6]; researchers also expect to improve the chatbot's profile dynamically.

After selecting an algorithm and finishing all pre-work, the next challenge to be conquered is the extraction and classification of useful data. This was also the third most commonly encountered challenge among the selected papers. As mentioned in the previous session, researchers tried to remove noise through various algorithms to prevent unnecessary data from being input into the model and cause interference. When most of the noise is filtered out, data classification is also key. This part of the challenge may also be related to information slotting issues [26], where the main purpose of the dialogue system is to extract key topics and feedback to the system for conversation response generation.

These challenges are likely to be important factors when resolving the difficulties of conversational chatbots. As a consequence, there are still some other lesser challenges listed in the table below, such as inefficient pre-work, a lack of diversity and quantity of training data, objective function formulation, feature selection, and enhancement in humanization and morality. Some papers also mentioned multi-lingual support [24], which might be a potentially interesting area. Although the translation system is quite mature, when introducing the Seq2Seq model, there might be some additional factors that need to be considered, such as cultural differences, when expressing ideas in conversation.

In summary, we listed the outcomes of conversational chatbots corresponding to the original objectives. Moreover, we also listed current challenges that researchers are facing and the next steps for future research directions. These challenges can provide researchers with directions on the potential areas or next steps to emphasize.

## 5. Conclusions

Using 32 screened papers, this study discussed the evolution of conversational chatbots from 3 perspectives: construction objectives (RQ1), applied algorithms (RQ2), and outcomes and challenges (RQ3), and the discussion process is shown in Figure 2. From the perspective of the development objectives of conversational chatbots, the main objective was not only to improve technical aspects by providing accurate responses, but to ensure that users' needs are met through context maintenance. We also discussed some research related to business and educational support, which leverage chatbots by either supporting potential business revenue increases, playing the role of a learning companion, or assisting in the enhancement of teaching skills.

Moreover, researchers are testing many methodologies to achieve the building objectives of conversation chatbots. One of the common methods is leveraging machine learning and deep learning models, or combining deep learning training techniques. Since there is no universal algorithm for all research data and objective functions, it is difficult to successfully select the best algorithm and optimize the conversational chatbots to achieve these

objectives; therefore, we further discussed applied algorithms in chatbots, explored the outcomes of these studies and discovered which deep learning algorithms were mainly chosen. Most studies focused on optimizing the output, hoping that the developed chatbot's responses would be more in line with users' expectations.

This study also found that increasing NLP (Natural Language Processing) technologies are being applied to architecture when building conversational chatbots. This may be because, by nature, we want chatbots to imitate human conversational capability, especially when designing an open-domain chatbot that does not have a pre-defined dialogue context; thus, the system needs to figure out and maintain consistency. Thus, there is another trend we sensed during our review: when the fundamental goal is providing accurate information to the users through closed-domain chatbots, researchers are more interested in a dialogue system, which can express emotions and empathy.

Some papers also highlighted some of the potential key challenges of conversational chatbots, such as how to achieve a system with human cognition capability or how to dynamically adjust the profile when the chatbot is playing the role of a learning companion or a student. This direction will add more features to a chatbot to output a good response during conversation and evaluate a user's status to provide a better customized response. This can be leveraged in the educational support area to increase learning performance and teaching skills.

**Author Contributions:** Conceptualization, C.-C.L.; methodology, C.-C.L. and A.Y.Q.H.; writing—original draft preparation, C.-C.L.; writing—review and editing, A.Y.Q.H.; supervision, S.J.H.Y.; funding acquisition, S.J.H.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Science and Technology Council grant number 111-2410-H-008-010-MY3 and 109-2511-H-008-007-MY3. And the APC was funded by discount voucher a8117945543238d6.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. ChatGPT: Optimizing Language Models for Dialogue—OpenAI. Available online: <https://openai.com/blog/chatgpt/> (accessed on 28 December 2022).
2. Ghazvininejad, M.; Brockett, C.; Chang, M.W.; Dolan, B.; Gao, J.; Yih, W.T.; Galley, M. A knowledge-grounded neural conversation model. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
3. Mrkšić, N.; Séaghdha, D.Ó.; Wen, T.H.; Thomson, B.; Young, S. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1777–1788.
4. Cui, L.; Huang, S.; Wei, F.; Tan, C.; Duan, C.; Zhou, M. Superagent: A customer service chatbot for e-commerce websites. In Proceedings of the ACL 2017, System Demonstrations, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 97–102.
5. Pawlik, Ł.; Płaza, M.; Deniziak, S.; Boksa, E. A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations. *Speech Commun.* **2022**, *143*, 33–45. [CrossRef]
6. Song, D.; Oh, E.Y.; Hong, H. The Impact of Teaching Simulation Using Student Chatbots with Different Attitudes on Preservice Teachers' Efficacy. *Educ. Technol. Soc.* **2022**, *25*, 46–59.
7. Lee, D.; Yeo, S. Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Comput. Educ.* **2022**, *191*, 104646. [CrossRef]
8. Liu, C.C.; Liao, M.G.; Chang, C.H.; Lin, H.M. An analysis of children's interaction with an AI chatbot and its impact on their interest in reading. *Comput. Educ.* **2022**, *189*, 104576. [CrossRef]
9. Hollander, J.; Sabatini, J.; Graesser, A. How Item and Learner Characteristics Matter in Intelligent Tutoring Systems Data. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, 27–31 July 2022, Proceedings, Part II*; Springer International Publishing: Cham, Switzerland, 2022; pp. 520–523.
10. Lin, C.J.; Mubarak, H. Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educ. Technol. Soc.* **2021**, *24*, 16–35.

11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
12. Sun, X.; Chen, X.; Pei, Z.; Ren, F. Emotional human machine conversation generation based on SeqGAN. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–6.
13. Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; Jurafsky, D. Deep reinforcement learning for dialogue generation. *arXiv preprint* **2016**, arXiv:1606.01541.
14. Singh, S.; Kearns, M.; Litman, D.; Walker, M. Reinforcement learning for spoken dialogue systems. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 956–962.
15. Kanodia, N.; Ahmed, K.; Miao, Y. Question Answering Model Based Conversational Chatbot using BERT Model and Google Dialogflow. In Proceedings of the 2021 31st International Telecommunication Networks and Applications Conference (ITNAC), Sydney, Australia, 24–26 November 2021; pp. 19–22.
16. Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
17. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguist.* **2020**, *46*, 53–93. [[CrossRef](#)]
18. Singh, S.; Litman, D.; Kearns, M.; Walker, M. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *J. Artif. Intell. Res.* **2002**, *16*, 105–133. [[CrossRef](#)]
19. Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.P.; Gao, J.; Dolan, B. A persona-based neural conversation model. *arXiv preprint* **2016**, arXiv:1603.06155.
20. Levin, E.; Pieraccini, R.; Eckert, W. A stochastic model of human–machine interaction for learning dialog strategies. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 11–23. [[CrossRef](#)]
21. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A diversity-promoting objective function for neural conversation models. *arXiv preprint* **2015**, arXiv:1510.03055.
22. Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Dolan, B. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint* **2015**, arXiv:1506.06714.
23. Jianfeng, G.; Michel, G.; Lihong, L. Neural approaches to conversational AI. *Found. Trends Inf. Retr.* **2019**, *13*, 127–298.
24. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Hon, H.W. Unified language model pre-training for natural language understanding and generation. *Adv. Neural Inf. Process. Syst.* **2019**, *33*, 13063–13075.
25. Sato, S.; Yoshinaga, N.; Toyoda, M.; Kitsuregawa, M. Modeling situations in neural chat bots. In Proceedings of ACL 2017, Student Research Workshop, Vancouver, Canada, 30 July–4 August 2017; pp. 120–127.
26. Lei, W.; Jin, X.; Kan, M.Y.; Ren, Z.; He, X.; Yin, D. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 1437–1447.
27. Anki, P.; Bustamam, A.; Al-Ash, H.S.; Sarwinda, D. High Accuracy Conversational AI Chatbot Using Deep Recurrent Neural Networks Based on BiLSTM Model. In Proceedings of the 2020 3rd International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 24–25 November 2020; pp. 382–387.
28. Keerthana, R.R.; Fathima, G.; Florence, L. Evaluating the Performance of Various Deep Reinforcement Learning Algorithms for a Conversational Chatbot. In Proceedings of the 2021 2nd International Conference for Emerging Technology (INCET), Belgaum, India, 21–23 May 2021; pp. 1–8.
29. Gasic, M.; Breslin, C.; Henderson, M.; Kim, D.; Szummer, M.; Thomson, B.; Young, S. POMDP-based dialogue manager adaptation to extended domains. In Proceedings of the SIGDIAL 2013 Conference, Metz, France, 22–24 August 2013; pp. 214–222.
30. Henderson, M.; Vulić, I.; Gerz, D.; Casanueva, I.; Budzianowski, P.; Coope, S.; Su, P.H. Training neural response selection for task-oriented dialogue systems. *arXiv preprint* **2019**, arXiv:1906.01543.
31. Syed, Z.H.; Trabelsi, A.; Helbert, E.; Bailleau, V.; Muths, C. Question answering chatbot for troubleshooting queries based on transfer learning. *Procedia Comput. Sci.* **2021**, *192*, 941–950. [[CrossRef](#)]
32. Tegos, S.; Demetriadis, S.; Karakostas, A. MentorChat: Introducing a configurable conversational agent as a tool for adaptive online collaboration support. In Proceedings of the 2011 15th Panhellenic Conference on Informatics, Kastoria, Greece, 30 September–2 October 2011; pp. 13–17.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.