Engineered Science
DOI: https://dx.doi.org/10.30919/es1210

Unfolding Conversational Artificial Intelligence: A Systematic Review of Datasets, Techniques and Challenges in Developments

Shilpa Gite,[1, 2,*] Urvi Rawat,[1] Sameer Kumar,[1] Ketan Kotecha,[1, 2] Bunny Saini,[1] Anant Bhatt,[3] and Nithesh Naik[4]

Abstract

Recently, Artificial Intelligence (AI) has seen significant progress, especially in Natural Language Processing (NLP) and Conversational AI, making response generation more efficient. This advancement, combined with increased availability of conversational data, has greatly improved conversational bots, thus enhancing their effectiveness and scope. Despite extensive research which is primarily focused on task-oriented systems, there's a noticeable lack of comprehensive literature reviews that cover conversational bots, datasets, state-of-the-art methodologies, and thorough analytical insights. This review presents a detailed study exploring these key dimensions of Conversational AI. It analyzes the datasets, methodologies for crafting conversational bots, and performance metrics, while addressing the various challenges inherent in dialogue systems. Additionally, it suggests viable solutions and provides insights into the future trajectory of conversational bots. Conversational bots are broadly categorized into Retrieval-based and Generative systems. This study also outlines future avenues for exploration, including advancements in data pre-processing, evaluation techniques, and optimization using advanced methods like Generative Adversarial Networks (GANs), transfer learning, self-supervised, or unsupervised learning. These innovative approaches leverage recent developments in conversational AI, laying a strong foundation for future research in this dynamic field.

Keywords: Natural language processing; Conversational AI; Deep learning; Recurrent neural networks; Transformers; Chatbot.
Received: 30 August 2023; Revised: 30 March 2024; Accepted: 15 July 2024.
Article type: Review article.

1. Introduction

The primary objectives of Artificial intelligence (AI),[1] Natural Language Processing (NLP), in particular, has been developing intelligent dialogue systems that respond as naturally as a human to a user query be it for a specific task or a generic one.[2] These systems are known as chatbots or conversational bots. Over the past decade, promising results have been witnessed in academia since large amounts of conversational data has become available for model training. Moreover, recent discoveries in Deep learning (DL), Reinforcement Learning (RL), and Multi-task learning[3] have helped conversational bots evolve at an incredible rate.[4]

Many popular Neural approaches to conversational bots have been built using supervised, unsupervised, and reinforcement learning techniques and have yielded impressive results.[2,4] Chatbots and conversational AI are two distinct words. Chatbots answer predefined questions (rule-based), while conversational AI provides the liberty to directly interact with applications and websites in their language as it uses various Natural Language techniques.[5] A chatbot is primarily used for text-based conversations, and conversational AI can also be operated via speech. Some popular use cases are product recommendation,[6] human resource management,[7,8] voice assistants,[9] therapy and diagnosis,[10] customer support,[11,12] data aggregation,[13] *etc*. However, despite their widespread use, they suffer from multiple challenges, such as not grasping the user sentiment,[14] bland non-contextual responses,[15] or simply struggling with

[1] *AI&ML Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, 412115, India.*
[2] *Symbiosis Centre for Applied AI, Symbiosis International (Deemed University), Pune, 412115, India.*
[3] *Faculty of Engineering and Technology, Parul University, Waghodia, Vadodara, Gujarat, 391760, India.*
[4] *Department of Mechanical and Industrial Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, 576104, India.*
**Email:* shilpa.gite@sitpune.edu.in (S. Gite)

modern-day slang.[5]

This survey paper provides a comprehensive view of different AI models developed to date for creating conversational bots, covering older rule-based models and newer generative models, their advantages, and shortcomings. Fig. 1 presents the content overview of the review work. We trace the journey of conversational bots back to simple ruled based systems[16,17] and discuss their evolution towards better-performing systems that cater to a wide range of tasks using Artificial Neural Networks (ANN), primarily following a Sequence-to-Sequence architecture or a Transformer architecture such as the GPT models and LaMDA.[18,19] This study highlights the different datasets used to train conversational bots, metrics used to evaluate them, be it manual or automated, the critical challenges faced in this field, and suggests means to resolve them. It talks about the use of DL techniques, such as RL, Multi-Task Learning, Transfer Learning (TL), and the concept of "Attention" that has significantly contributed to achieve state-of-the-art results in this field. Fig. 2 summarizes the trends in the domain of Conversational AI. Table 1 offers an insightful outline of existing research body in Conversational AI, highlighting distinctive contributions thereby providing a nuanced perspective for the reader. This study aims to encompass research findings up to late 2022 – early 2023, considering the dynamic and rapidly expanding nature of this field.

A search strategy for the survey of different areas of conversational AI is done in this paper. Research papers from this area are obtained from a popular search engine Google Scholar as well as research databases such as Scopus and Web of Science. The articles are searched considering the basics of Conversational AI. Popular datasets, its different methods as well as evaluation metrics are reviewed. Key Keywords are 'Natural language processing', 'conversational AI', 'deep learning', 'recurrent neural networks', 'transformers', 'chatbot' for our search strategy.

Table 1. Summary of key themes and insights from recent studies from literature on conversational AI.

Author & Year	Theme	Contribution
Klopfenstein, L.C. *et al.*, 2017[20]	A detailed review of applications of Conversational AI for messaging platforms until 2016.	This paper focuses on the applications of Conversational AI. It reviews different messaging platforms, their support for chatbots, their services, and their advantages for users and developers.
Gao, J. *et al.*, 2019[4]	A detailed review of state-of-the-art techniques employed in Conversational AI until 2018.	The paper presents an extensive review of several Conversational AI approaches, namely Question-Answering systems, Retrieval-based chatbots, and Generative chatbots, their applications, and the challenges faced. The discussion on Generative chatbots does not address any transformer-based models.
Jannach, D. *et al.*, 2020[6]	A detailed survey on the application of conversational AI in recommender systems until 2020.	This survey paper covers datasets used, approaches to conversational recommender systems, and metrics used to evaluate these recommender systems.
Adamopoulou, E. *et al.*, 2020 [5]	Comprehensive analysis of the historical evolution of chatbots, from their inception up until 2020.	Authors discuss historical evolution of chatbots and the limitations encountered at every stage. This paper presents methodologies for constructing chatbots and explores their applications and use cases.
Zaib, M. *et al.*, 2021[21]	A brief review of approaches towards Conversational AI and its applications until 2021.	The paper briefly addresses multiple retrieval-based and generative conversational systems (limited to BERT and GPT), datasets used, and applications.
Motger, Q. *et al.*, 2021[22]	Review of the Conversational AI landscape in its entirety until 2021.	This paper covers the landscape of conversational bots, mainly their applications, datasets, and methodologies employed, briefly discussing each of these topics. The paper provides a comprehensive and shallow view of Conversational AI, devoid of details on methodologies.
Caldarini, G. *et al.*, 2022[23]	An extensive study to explore recent advancements in chatbots utilizing AI and NLP until 2021.	This paper examines the current literature on chatbot implementation methods, focusing on Deep Learning algorithms. It identifies limitations, challenges, and applications of the domain.
Kusal, S. *et al.*, 2022[24]	A broad review of techniques and tasks related to conversational agents with a focus on deep learning methods and approaches until 2022.	This paper provides an in-depth analysis of the working architecture and implementation methods employed by recent conversational agents. It also surveys the current practices and datasets utilized in the field.
Lin, C.-C. *et al.*, 2023[25]	An extensive survey of objectives, methodologies, datasets, achievements and challenges of chatbot systems until 2022.	This paper provides an overview of the goals and methods involved in building chatbots, from existing literature since 1999. It also discusses the datasets used and highlights encountered limitations.

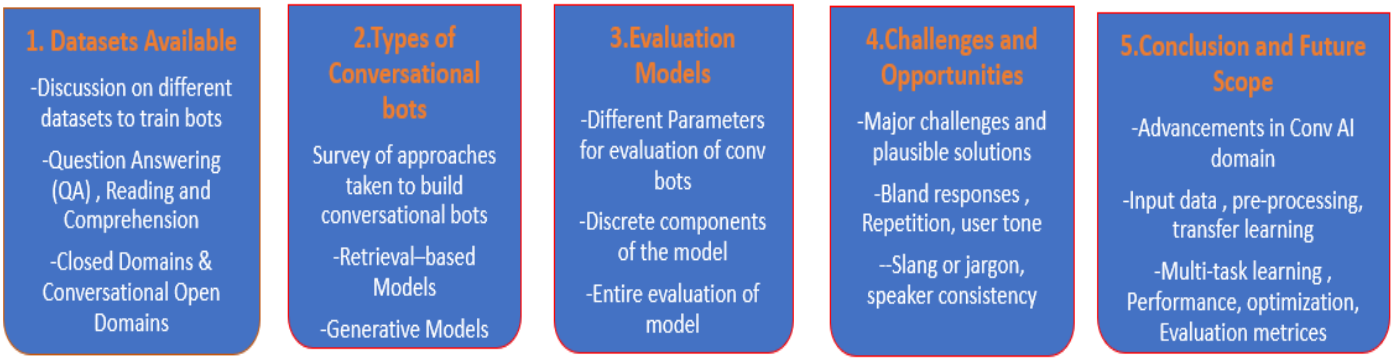


Fig. 1. Content overview of the review paper.

This paper highlights notable research gaps, particularly in the insufficient provision of information and analysis concerning the utilized datasets. Moreover, there is a discernible absence of in-depth analysis of evaluation metrics. Additionally, the lack of clarity in distinguishing between industry-focused, task-specific conversational AI systems and companion chatbots is not prominently emphasized.[26]

The study contributes to the research community by highlighting recent and historical trends and improvements to the state-of-the-art systems offering insights into the current landscape of the conversational AI. It also traces advancements in the field of conversational bots and an overview of the evolution of technologies. It presents comprehensive outline of the datasets and evaluation metrics utilized to assess performance and effectiveness in the domain. Moreover, an extensive discussion of the existing research limitations that hinder the progress of chat agents, along with proposed systematic solutions is also presented.

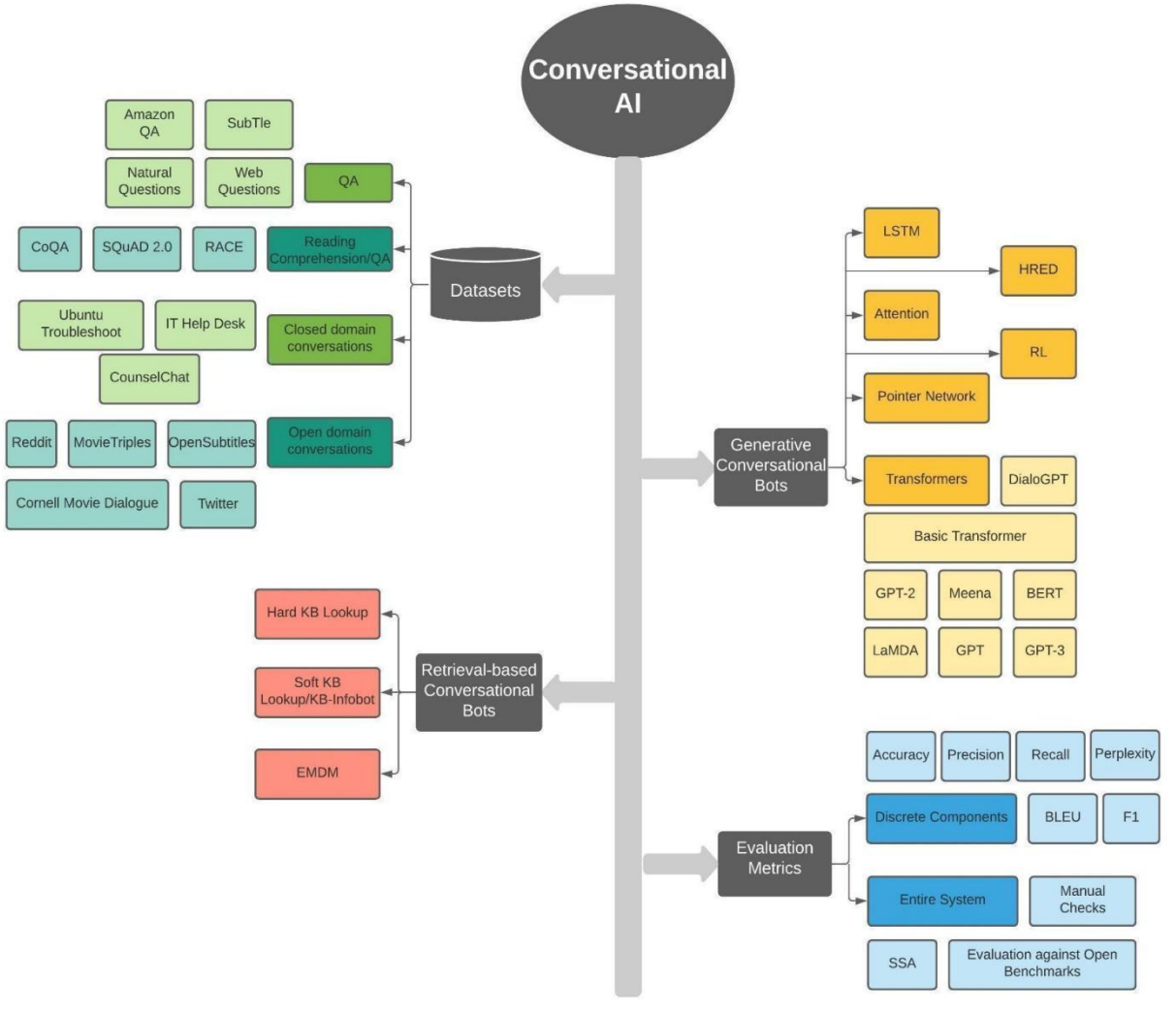


Fig. 2 Summary of trends in the domain of conversational AI.

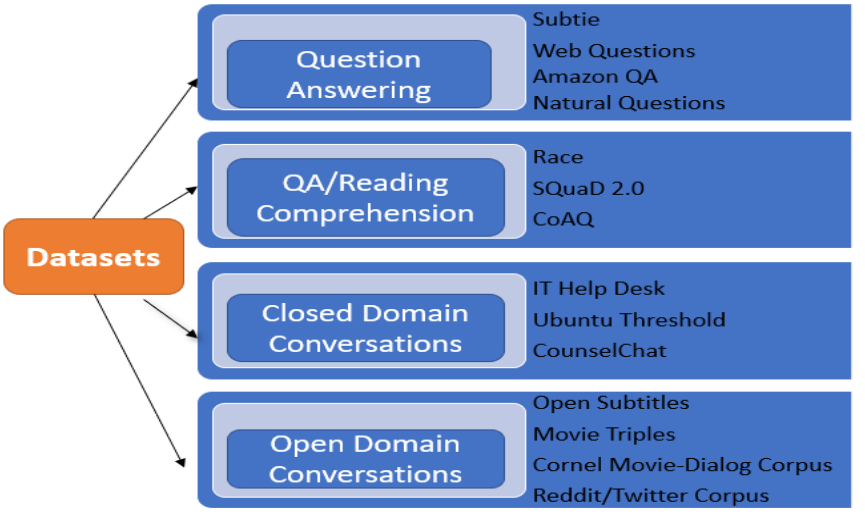


Fig. 3 Available datasets for conversational bots.

2. Datasets available

One of the critical aspects of building a well-performing model is training data. Clean, unbiased, and factually accurate data can lead to the development of better-performing conversational bots. Some publicly and privately available datasets consisting of multi-turn dialogue/conversations or single-turn dialogue/question-answer pairs (QA) are listed below in Fig. 3.

As summarized in Fig. 3, there are four major categories of the datasets available: Question and Answer based (QA), QA with reading comprehension, closed domain conversations, and open-domain conversations. Table 2 presents details about the datasets found in the conversational and QA domain from the reviewed papers.

Table 2. Corpora overview for conversational applications.

Type	Dataset Name	Description	Link
QA	Web questions dataset[16]	This dataset consists of 6642 question-answer pairs. The questions pertain to a single entity and can be answered using Freebase, a humongous knowledge graph.	https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a
	SubTle dataset[26]	This dataset consists of 5.5M QA pairs generated from movie subtitles.	Available upon request.
	Amazon QA dataset[27,28]	It comprises 1.4 million question-answer pairs of about 191000 products, which can be found on Amazon.	http://jmcauley.ucsd.edu/data/amazon/qa/
	Natural Questions dataset	323,000 QA pairs from different domains, where the question is a google search query, and the Answer is obtained from Wikipedia by annotators.	https://ai.google.com/research/NaturalQuestions
QA/ Reading Comprehension	RACE dataset[29]	The dataset utilized in this study comprises a collection of more than 28,000 passages and nearly 100,000 questions, specifically designed for reading comprehension tasks. These passages and questions were sourced from English tests administered to school students in China.	http://www.cs.cmu.edu/~glai1/data/race/
	SQuAD 2.0 dataset[30]	Stanford Question Answering Dataset (SQuAD) consists of crowdsourced questions on a collection of Wikipedia articles. The answers to these are text from the corresponding reading passage, or the question is considered unanswerable. For example, SQuAD 2.0 has 100k+ QA pairs with 50k questions that are unanswerable.	https://datarepository.wolframcloud.com/resources/SQuAD-v2.0

Type	Dataset Name	Description	Link
Closed domain conversations	CoQA dataset[31]	This dataset contains 127,000+ question-answer pairs collected from 8000+ conversations. These conversations were obtained by pairing two crowd workers to talk about a passage regarding questions and answers.	https://stanfordnlp.github.io/coqa/
	IT help desk dataset[32]	Extracted from an IT help desk troubleshooting service, this dataset contains 33M tokens. Each conversation is about 400 words long.	Available upon request.
	Ubuntu Troubleshoot dataset[33]	This comprises 930,000 dialogues obtained from Ubuntu-oriented chat rooms on the Freenode Internet Relay Chat (IRC) network, where people talk about their Ubuntu issues.	https://www.kaggle.com/rtatman/ubuntu-dialogue-corpus
	CounselChat dataset	This includes 2130 therapist-client conversations scraped from counselchat.com, spanning over 31 different topics ranging from 'depression' to 'military issues.'	https://github.com/nbertagnolli/counsel-chat/tree/master/data
Open-domain conversations	OpenSubtitles dataset[34]	It comprises subtitles of various movies and tv shows collected from OpenSubtitles, available in 62 different languages. The English dataset has over 400 million sentences.	https://opus.nlpl.eu/OpenSubtitles-v2018.php
	MovieTriples dataset[35]	It comprises 245,000 dialogues in total, where each conversation has three turns and occurs between two characters from a movie script.	Available upon request.
	Cornell Movie-Dialog Corpus	It contains 220,579 conversations between 10,292 pairs of movie characters, spanning over 600 movies. It also provides meta-data of these movies.	https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html
	Reddit dataset	This dataset contains 3.7 billion consecutive comments structured in conversations gathered from 2015 to 2019.	https://github.com/PolyAI-LDN/conversational-datasets/tree/master/reddit
	Twitter Corpus[36]	This dataset consists of 1.3M dialogues extracted from Twitter posts, 69% containing only two turns.	Available upon request.

3. Types of conversational bots

Traditionally, a dialogue system (a.k.a. a conversational bot) contains four modules, as stated by[4]:

- Natural Language Understanding module (NLU): Used for recognizing the user's intent and extracting particular information.
- State tracker: To track the state of dialogue and encapsulate all important information until that point.
- Dialogue policy: It chooses the following /consequent action based on the current state.
- Natural Language Generation module (NLG): To convert bot's actions to natural language responses.

These are also known as Retrieval based Models, wherein the responses are fetched from an existing database, as represented in Fig. 4.

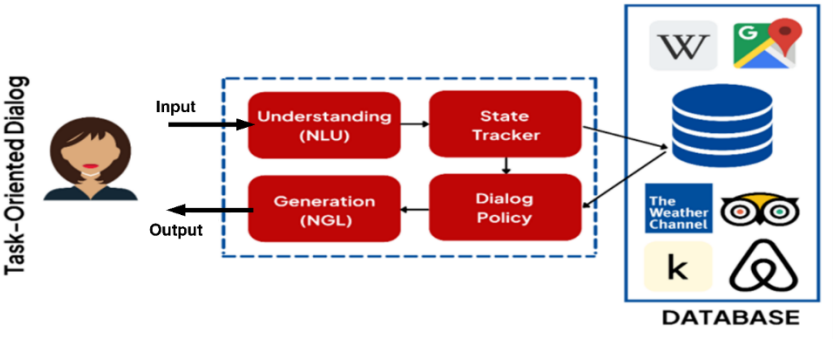


Fig. 4 A Retrieval based conversational model.

Fig. 5 A Generative conversational model.

In recent times, as depicted in Fig. 5, dialogue systems have shown an evolution towards entirely data-driven end-to-end (E2E) systems, not requiring four different components. These are widely known as Generative Models and significantly rely on the concept of Neural Networks to produce responses themselves, without needing any external database.

Essentially, all dialogue systems can be summarized as shown in Fig. 6. The details of the conversational agents are described in the following sections thoroughly.

3.1 Retrieval based models

Retrieval models labor on the fundamentals of graphs or directed flows[4] that make up their Knowledge Base (KB). The KB is essentially a database consisting of possible questions/topics and their corresponding responses. The conversational bot is trained to supply the most straightforward attainable reply from its KB. They use keyword matching, machine learning, or simple if-else conditions to spot the most straightforward reply to an input query. These conversational bots give solely predefined responses. They don't produce any new output, eliminating the possibility of grammatically incorrect or unwanted responses. Although, this does not mean that these responses are 100% correct.[4] For such a bot to be genuinely conversational, it requires a tremendously huge KB, often created manually. In addition, often, the KB is updated repeatedly; otherwise, it gets outdated with time.

As shown in Fig. 7, a retrieval-based model works on a closed domain. Pattern matching poses a significant challenge when it comes to such models. There are multiple ways in which the user can ask a question. Machine learning algorithms are predominantly used to classify user queries.[4] Once the intent/context of the user query has been established, the appropriate response can be fetched from the KB.

3.1.1 Hard KB lookup

Semantic parsing is the most widely used methodology for NLU, wherein the user query is mapped into its logical form and then looked up in the KB Graph. This lookup is done by finding all existing paths in the KB which match the user query, and then the end nodes of these paths are retrieved. Finally, out of these end nodes, the apt response for the user query is decided using KB Reasoning.[16,17,37,38] This process is also known as Hard KB Lookup.

3.1.2 Entropy Minimization Dialogue Management (EMDM) Strategy

Entropy minimization dialogue management method[3 9] proposes unique dialogue states for the State Tracker, called the Dynamic Stochastic States (DS-States). This state

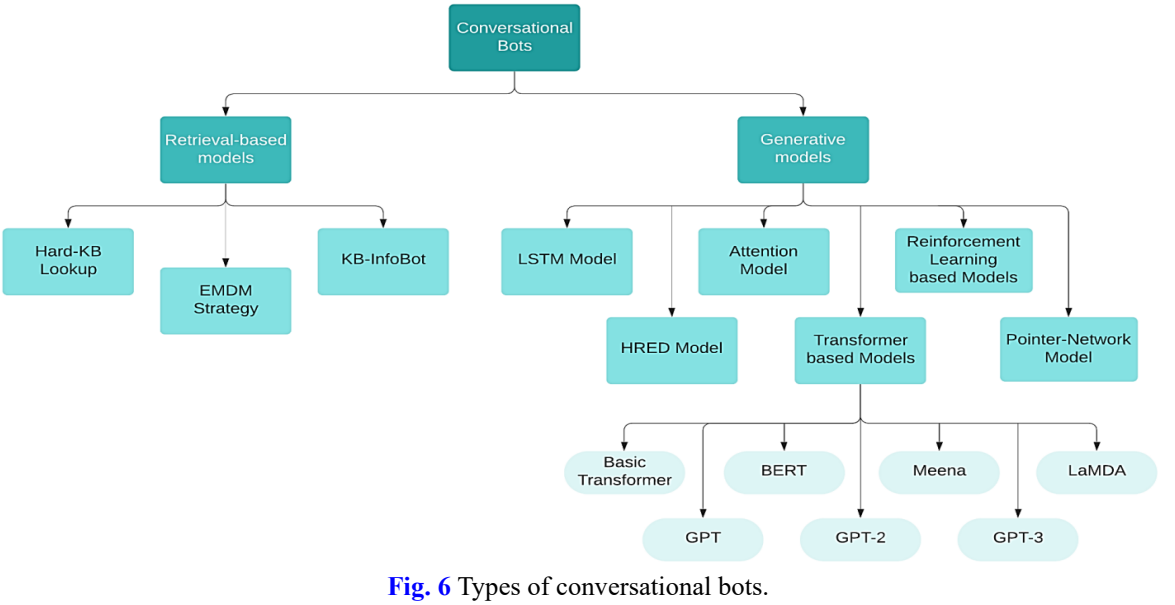


Fig. 6 Types of conversational bots.

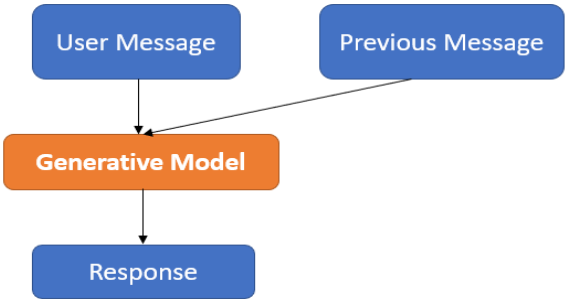


Fig. 7 Workflow of a Retrieval-based model.

essentially captures the dialogue history and all the information available until that point as a distribution over the goal set. Hence, making the state dynamic at every turn. Also, the Dialogue Policy proposed ensures that the agent would always ask a question relevant to the attribute with maximum entropy from those that remain in the KB. Thus, this dialogue policy is more efficient than other non-entropy-based methodologies.

3.1.3 KB-InfoBot

For KB-InfoBot dialogue system,[40] the KB entries are assumed to be of the form (*h,r,t*), which means h is related to t via r. This is known as an Entity Centric Knowledge Base and is generally stored as a Table. Furthermore, as presented in Fig. 8, the concept of Soft-KB lookup is also introduced, indicating that a probabilistic framework is used to process the user query and determine which entities they are interested in. Moreover, the system is trained using RL, leveraging user feedback at every turn, consequently leading to high success rates after continual use.

While such models work well for closed domains, it becomes imperative to deploy more sophisticated state-of-the-art neural network approaches (also known as generative models) for multiple domains or a broader set of services. For example, one cannot be expected to have a knowledge base big enough to account for millions of questions spanning hundreds of domains or engage in human-like chit-chat using a retrieval-based model. A comparative analysis of retrieval-based conversational systems is presented in Table 3 with highlights of the reviewed papers.

Table 3. Comparative analysis of retrieval based conversational systems reviewed.

Ref.	Technique	Dataset	Objective	Merits	Demerits
[16]	Coarse alignment followed by bridging for parsing	Free917, Web questions dataset	Create a semantic parser that scales up easily using QA pairs.	Free scaling, efficient parser for input queries.	Follows a two-step approach to QA tasks. Prone to parsing errors.
[17]	Translation based KB-QA	Web questions dataset	One-step translation based on CYK parsing for QA tasks.	Integrates semantic parsing and QA into one framework. Performs better than traditional two-step approaches.	Prone to NLU errors.
[38]	Staged query graph generation for semantic parsing, Hard-KB Lookup	Web questions dataset	A novel method of semantic parsing for Graph KBs.	It simplifies the semantic parsing and matching process while making it more efficient. Uses advanced entity linking methods to improve results.	Not feasible for complex user queries on huge KBs. Provides no information on faults in semantic parsing.
[39]	Entropy Minimization Dialogue Management Strategy	Dataset with 38117 songs and 12 attributes associated with each song.	Entropy-based dialogue policy using DS-states for a Song-on-Demand task.	DS-states capture the entire dialogue history at each turn. Entropy-based strategy is more efficient in terms of goal-seeking.	It only works if there are no NLU errors. Can lead to the user being asked complex questions.
[40]	Soft-KB Lookup, RL	IMDB movies dataset	E2E learning for Retrieval based systems for a Movie-on-Demand task.	Soft KB Lookup leads to higher success rates. E2E model training employing RL based on user feedback. Users are constantly asked questions that are easy to answer.	This model operates well on a closed domain. RL leads to overfitting over time with real users.

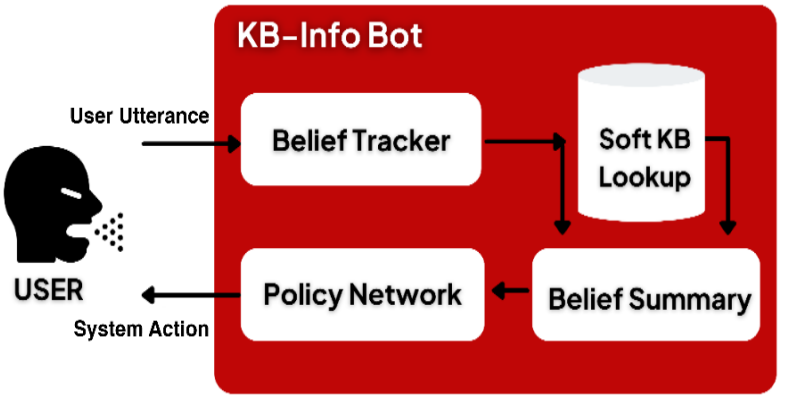


Fig. 8 KB-Info Bot architecture. Reproduced with the permission from [40].

3.2 Generative models

Researchers have been exploring entirely data-driven end-to-end (E2E) approaches to generate responses for conversational bots in the past decade, suggesting that these models are not dependent on the four components of a traditional dialogue system mentioned above; instead, they rely entirely on data. Such models have shown tremendous potential in open-domain conversations (or chit-chat),[19,32,35] implying that users can talk about any topic of their preference system. Furthermore, E2E models scale quickly to free-form and open-domain datasets adding to their success as social bots. Fig. 9 represents the workflow method of a generative model.

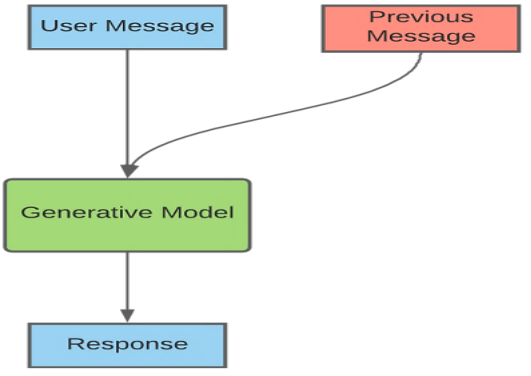


Fig. 9 Workflow of a generative model.

They used a mixture of supervised learning, unsupervised learning, RL, and adversarial learning for multi-step training. Essentially data-driven E2E models follow a Neural Network-based architecture for response generation. Unfortunately, generative models are prone to grammatical errors, and such mistakes have costly implications in production systems. However, in the event that corporations have access to substantial volumes of data, generative models become feasible and valuable. Fig. 10 presents a detailed time evolution diagram for the generative language model, and it is seen that the developments are seen from the year 2014, and it would be an upcoming field in NLP soon.

The following discussions are based on the technical developments in the conversational AI domain as per the evolution diagram in Fig. 10.

3.2.1 LSTM model

Looking at supervised learning, Long Short-Term Memory model (LSTM) is the most popular, although Gated Recurrent Units (GRU)[41] is often as productive. It follows the Sequence-to-Sequence (seq2seq) framework.[42] This seq2seq framework comprises of two units: an encoder and a decoder. An LSTM layer (or stacks of them) behaves as the encoder. It takes the input sequence and returns a vector of fixed size. The output of the encoder LSTM is discarded, retaining the state alone. This state acts as the input to the decoder. One more LSTM layer (or stack thereof) behaves as the decoder. It is trained to predict the following characters/words of the output sequence, given previous characters/words. Most significantly, the state vectors made by the encoder are used by the decoder as its initial state. The decoder eventually learns to supply targets[t+1] in the presence of targets[t], based on the actual input. A substantial application of the Encoder-Decoder architecture is in Language Translation [42] or Dialogue generation for closed and open domain tasks (a.k.a. chatbots).[32]

3.2.2 HRED model

The LSTM model works well in encoding contexts up to 500 words, however in several cases dialogue histories turn out to be longer, necessitating the need to track long-term context. Therefore, the Hierarchical Recurrent Encoder-Decoder model (HRED) was proposed.[35,43] It models dialogue using a two-level hierarchy consisting of two Recurrent Neural Networks (RNNs): one at word level and another at dialogue turn level. Furthermore, it considers that conversation history contains a sequence of turns comprising of a series of tokens. Thus, capturing more extended context and addressing the vanishing gradient problem limits RNNs (including LSTMs) from modeling long sequences.

3.2.3 Attention model

The seq2seq framework works well for text generation but falls short for long source sequences as it encodes it into a fixed-sized vector. In such a case, the context vector(w) cannot grasp all the words in the particular sentence, as seen in Fig. 11. It states that as the length of the sentence increases, the

BLEU score (for machine translation) starts decreasing from one point in time.[44]

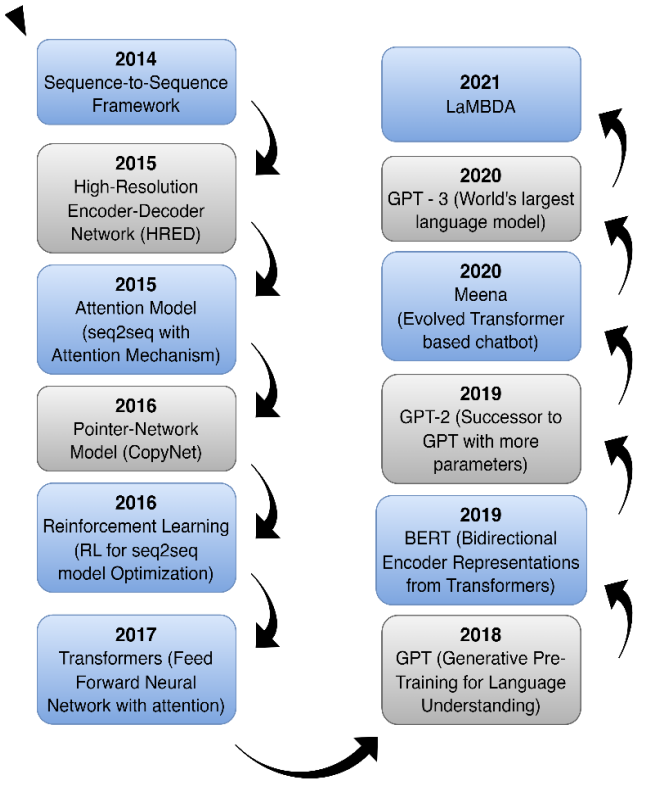


Fig. 10 Generative language model evolution.

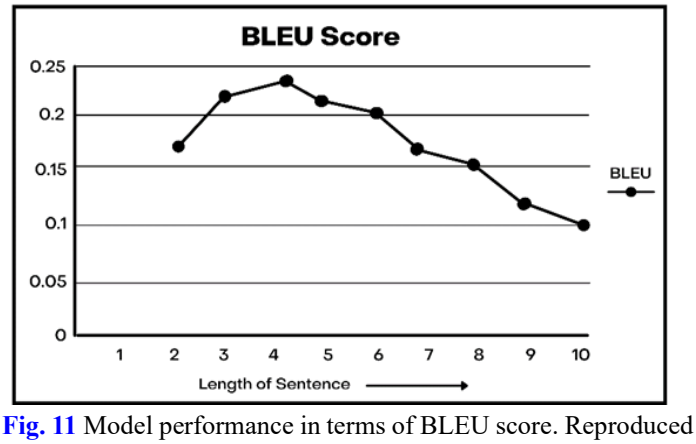


Fig. 11 Model performance in terms of BLEU score. Reproduced with the permission form [29].

Attention-based models[45,46] aid this by allowing the model to condition parts of the input sequence relevant to generating the next target word, thus not encoding the entire input sequence as a fixed-size vector. Instead, the state at each layer is stored, and a weight is assigned to it. Hence, emphasizing certain input portions while producing an output sequence enhances the performance of attention models. This stands true for machine translation, yet they struggle with dialogue generation, as shown in Fig. 12.

3.2.4 Pointer–Network model

Several research studies[47,48,34,35] have extended the seq2seq framework to improve its capability to selectively incorporate words from the input sequence into the generated response, resembling a "copy and paste" operation. Inspired by the pointer network model, this ability is integral for dialogue generation due to multiple proper nouns. These models generate output words from either a fixed-size vocabulary (similar to the seq2seq model) or by selecting words directly from the source sequence (similar to the pointer network model), employing an attention mechanism.

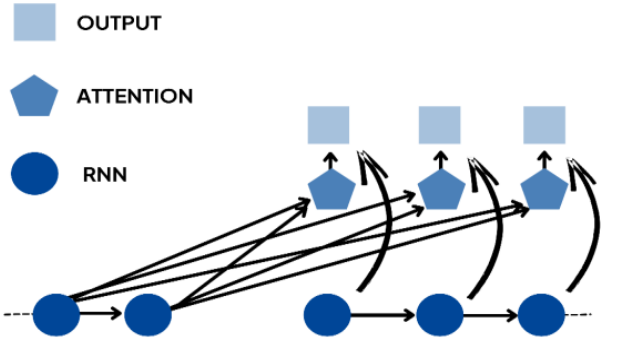


Fig. 12 Seq2seq with attention architecture. Reproduced with the permission from [46].

3.2.5 Reinforcement learning-based approach

Conversational bots often suffer from response blandness and fall short in producing long-term engagement from the user, as shown in Fig. 13. RL[49] comes into the picture to address this issue, wherein a reward function is designed to optimize an E2E system.

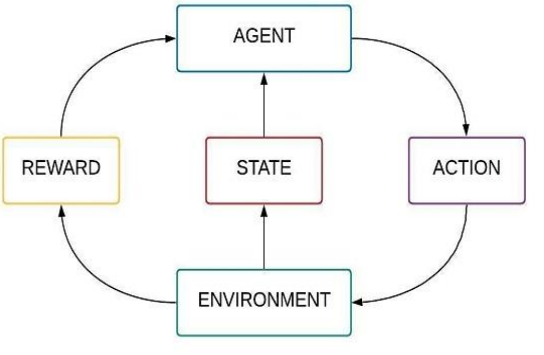


Fig. 13 Reinforcement learning set-up.

Researchers proposed a user simulator using a standard seq2seq model to maximize the total expected reward over the sentences produced by the user simulator and the agent, which is to be remembered. Here the objective is represented by Equation (1)[4]

$$J(\theta) = E[\, R\,(T_1, T_2, \ldots, T_N)\,] \qquad (1)$$

where R is the reward function, and T_i's are dialogue turns. Gradient Descent is used as the optimizer for the above objective by factoring the log probability of the conversation and the aggregated reward, which is independent of the model parameters as in Equation (2)[4]:

$$\nabla J(\theta) = \nabla log p\,(T_1, T_2, \ldots, T_N) R(T_1, T_2, \ldots, T_N)$$
$$\simeq \nabla \log \prod p\,(T_i | T_{i-1})\, R(T_1, T_2, \ldots, T_N) \qquad (2)$$

where $p(T_i|T_i\text{-}1)$ is similarly parameterized as a standard seq2seq model, except that the model is optimized using RL.

3.2.6 Transformer based models

These models use encoder-decoder architecture[19] to respond from a sequence of inputs. The novelty in this approach is that transformer models do not use any RNNs (LSTM, GRU, *etc.*); instead, they rely on a feed-forward neural network using attention mechanisms. As a result, transformers have shown tremendous results in dialogue generation tasks. And unlike RNN models, they tend to consume relatively less extensive amount of training time. Furthermore, the transformer models: GPT, BERT, and their progress is discussed below:

A. Basic transformer model

This model[19] comprises of an encoder and a decoder, created by stacking multiple blocks consisting of feed-forward layers and multi-headed attention on top of each other, as depicted in Fig. 14. The outputs and inputs are initially embedded into an n-dimensional space, also known as positional encoding.

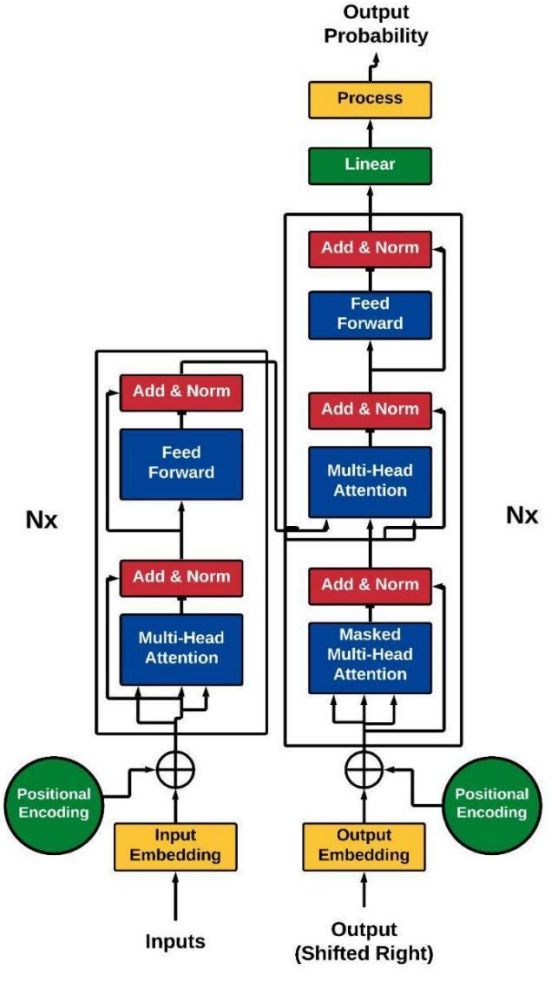


Fig. 14 Transformer model architecture with encoder (left) and decoder (right). Reproduced with the permission from [19].

The attention mechanism utilized here can be described as Equation (3)[19]

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

As seen in Fig. 15, Q is the matrix that comprises the input, representing a vector corresponding to a word in the sequence. K represents all the keys, *i.e.*, vector notations of all the words in the particular sequence. V are those values, which again represent the vector of all the words in the particular sequence. As for encoder-decoder multi-head attention modules, V has a similar word sequence as Q. However, V differs from the sequence represented by Q for the attention module that considers the encoder and the decoder sequences.

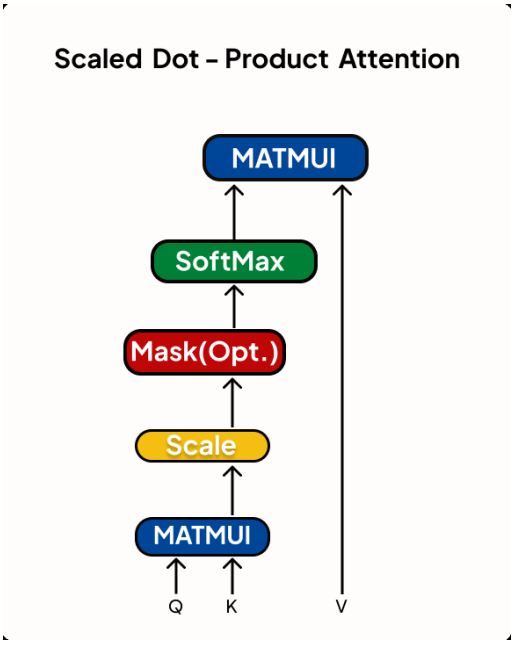


Fig. 15 Scaled dot-product attention. Reproduced with the permission from [19].

B. GPT (Generative Pre-Training for Language Understanding)

This is the first language model[50] created by OpenAI to address different tasks such as answering questions, document classification, response generation, and semantic similarity assessment. Due to the abundance of unlabeled data, they proposed generative pre-training of a language model on a vast unlabeled aggregation and different tuning on each specific task as a follow-up (such as response generation for conversational bots or question answering based on a comprehension). However, the transformer used by GPT works on constrained self-attention, which means each token can consider the context towards its left only. This sort of restriction is harmful to token-based fine-tuning tasks such as answering questions or generating responses where bi-directional context awareness is needed.

C. BERT (Bidirectional Encoder Representations from Transformers)

BERT[51] is a language model that follows a similar approach to GPT, wherein the first step is pre-training, followed by fine-tuning. The first step involves training the model over a massive corpus of unlabeled data about different

language tasks which is then fine-tuned using task-specific labeled data. Except for the output layer, similar architecture is used for both steps. The distinct feature of BERT is that its transformer is bi-directional. It utilizes MLM (Masked Language Model) objective during pre-training, where it randomly masks specific input tokens and aims to predict them based on context, which results in a better pre-trained model than GPT, which is then fine-tuned for various NLP tasks such as question answering, named entity recognition, next sentence prediction, *etc*. $BERT_{BASE}$ consists of 110M parameters, and $BERT_{LARGE}$ consists of 340M parameters.

D. GPT -2 (Generative Pre-Training for Language Understanding-2)

GPT -2 is the successor to OpenAI's first GPT model, which consisted of billions of parameters (1.5 billion) trained on a set of millions of web pages (8million) amounting to 40GB and outperformed all its predecessors.[52] The key focus was to build a larger model using a larger dataset for training. It displayed various capabilities including producing artificial samples of text with never seen before quality. Furthermore, GPT-2 also astoundingly outperformed any other language model trained on specific areas without training on datasets specific to that domain. In the field of NLP, GPT-2 demonstrates the ability to learn tasks such as question-answering, text summarization, and language translation directly from raw text data, without the need for task-specific training data. This is known as zero-shot learning.

One modification of GPT-2 is DialoGPT by Microsoft Corp,[53] a language model for Large-Scale Generative Pre-training for Conversational Response Generation. It presents a conversational bot trained on 147M conversations that have been extracted from Reddit. It has developed highly relevant responses consistent with the context and shows close to human-level performance in dialogue generation.

3.2.7 MEENA

Presented by Google's Research team, Meena[54] is an open domain E2E data-driven conversational bot with 2.6 billion trainable parameters. It has been trained on publicly available conversations scraped and filtered from social media websites, amounting to 40 billion words or 341GB of data which is 8.5 times that of GPT-2. Meena uses a unique Evolved Transformer (ET) architecture, which applies Neural Architecture Search (NAS) to a primary Transformer.[55] This model consists of one ET encoder block and 13 ET decoder blocks. The superior decoder is the key to Meena outperforming state-of-the-art conversational bots such as Mitsuku, DialoGPT, *etc*.

3.2.8 GPT -3 (Generative Pre-Training for Language Understanding - 3)

Proposed by OpenAI, this is the most prominent language model. It consists of 175 billion parameters and has showcased incredible performance compared to any other model,[56] owing predominantly to its size. It has the same model architecture as GPT -2, including modified initialization, reversible tokenization, and pre-normalization. The critical difference accounts to alternating dense and locally banded sparse attention patterns used in the transformer layers. It has performed tremendously well in zero-shot and one-shot settings.

Furthermore, it tends to outperform many state-of-art fine-tuned models in a few-shot setting for various NLP tasks. Ironically, its size makes it infeasible and inconvenient for real-world applications because ample time is taken for inference. It is also prone to losing coherency and repeating a set of words or sentences repeatedly.

Figure 16 presents a comparative study among various popular conversational models. It strongly suggests that GPT-

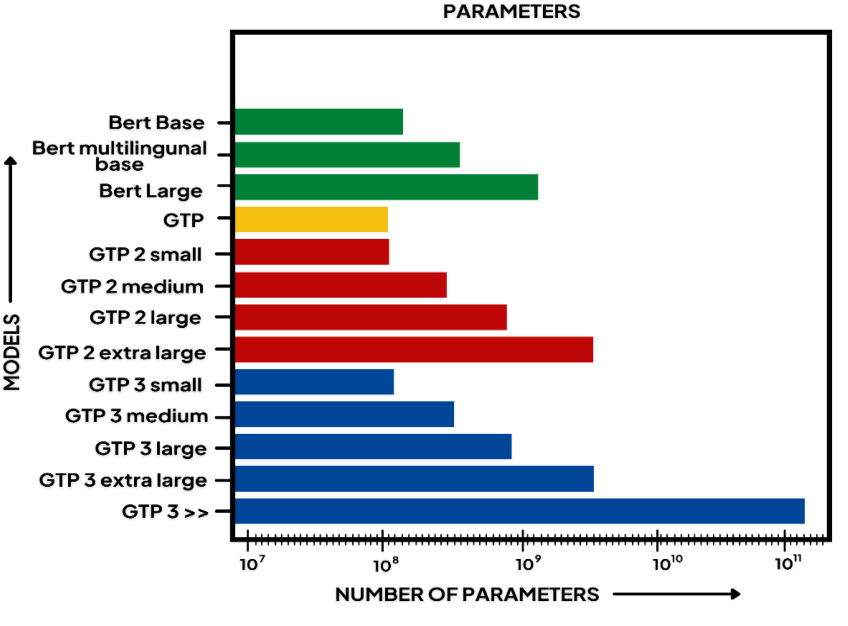


Fig. 16 Comparative study among BERT, GPT, GPT-2 and GPT-3 models.

3 is poised to emerge as a dominant force among conversational agents in the near future.

3.2.9 LaMBDA

This is the most recent breakthrough in conversational AI by Google's research team. First showcased at Google I/O 2021, LaMDA[19] is a successor to Meena and leverages the Basic Transformer architecture.[20] Although still under development, it not only focuses on Sensibleness and Specificity using the Sensibleness and Specificity Average (SSA) Score but also factors "factual correctness" and "interestingness," i.e., if the generated response was witty or insightful. It is the first-ever chatbot that can cater to the open-ended nature of human conversations. It remains resilient even in case of multiple topic changes or derailing of the conversation. Table 4 presents a comprehensive summary of the most influential research articles based on the generative conversational systems. In exploring conversational AI systems, this study compares

Table 4. Comparative analysis of generative based conversational systems reviewed.

Ref.	Technique	Evaluation Metric		Dataset	Objective	Merits	Demerits
[32]	Seq2seq Framework	Perplexity	8 17	IT help desk dataset OpenSubtitles dataset	Construct an entirely data-driven E2E model (without handcrafted rules) for both closed and open domain conversations.	Fully data-driven. Thus, no rules are needed. Simplistic E2E model training.	Lack of speaker consistency. Loses context with long input sequences.
[35]	HRED HRED + SubTle HRED-Bi. + SubTle	Perplexity	36.59 27.14 26.81	MovieTriples dataset SubTle dataset, MovieTriples dataset SubTle dataset, MovieTriples dataset	Extend the HRED model to create an open domain generative conversational agent.	It can capture context for long input sequences over multiple dialogue turns.	Complex use of RNNs is computationally intensive.
[47]	Seq2seq Framework with Attention	Perplexity	25.52 45.38	SubTle dataset, MovieTriples dataset Ubuntu Troubleshoot dataset	Improve the seq2seq framework by adding a dynamic attention mechanism.	The attention scope increases as the dialogue progress, leading to highly relevant responses in an extended dialogue history.	RNNs are computationally intensive and take ample training time. This model also suffers from response blandness.
[48]	CopyNet	Decoding accuracy (Dataset-1) Decoding accuracy (Dataset-2)	71% 64.8%	Collection of dialogue instances from BaiduTieb, wherein empty slots for names were filled. Divided into two datasets, Dataset-2	Incorporate copying mechanism in seq2seq framework.	Copy parts of the input sequence, such as phrases, names, etc., generated, improving speaker consistency in the response.	It has not been tested on multi-turn dialogue.

Ref.	Technique	Evaluation Metric		Dataset	Objective	Merits	Demerits
				has no overlaps in test and train set wrt filled slots, and Dataset-1 has overlaps.			
[49]	RL	Human evaluation (RL-wins) Human evaluation (RL-loses) Human evaluation (Tie)	0.72 0.12 0.16	OpenSubtitles dataset	Modeling the future direction of the conversation in seq2seq chatbots using a reward function.	Improves informativity, coherence, and ease of answering. Continuous user feedback integration for superior interaction.	Possibility of responses or words to repeat in a loop for multi-turn dialogue. The reward function does not capture all possible aspects of a good conversation.
[19]	Transformers	Perplexity BLEU Perplexity BLEU	11.7 6.8 17 4.7	OpenSubtitles dataset Cornell Movie-Dialog Corpus	An unconventional approach to language tasks relying solely on attention mechanism and feed-forward layers, doing away with RNNs/ CNNs	The first-ever model to rely entirely on the concept of attention, leading to parallelization. Quicker to train as compared to RNN based models.	The ability to update the state is limited. It cannot capture high-level representations of the input sequence.
[50]	GPT	Accuracy	59.0%	RACE dataset	Unsupervised pre-training, followed by supervised fine-tuning on a particular task, improved performance on multiple NLP tasks.	First, ever model to explore unsupervised pre-training to boost performance. Incredible results on 9 out of 12 datasets explored. Fine-tuning is relatively quick.	It limits the maximum length of an input sequence. Bias in output sequence due to training data.
[51]	$BERT_{LARGE}$ (Single)	F1	83.1	SQuAD 2.0 dataset	Unsupervised pre-training of a bi-directional model, followed by supervised fine-tuning on a particular task, for	Produced state-of-art results on 11 NLP tasks. Fine-tuning on multiple tasks is inexpensive and requires minimal changes in the model architecture.	It limits the maximum length of an input sequence. It is poor at pragmatic inference.

Ref.	Technique	Evaluation Metric		Dataset	Objective	Merits	Demerits
					outstanding performance on a wide array of NLP tasks.		
[52]	GPT - 2 (117M)	F1	< 30	CoQA dataset	Explore large models in a zero-shot setting for a stellar performance on various NLP tasks without direct supervision.	Produced state-of-art results on 7 out of 8 NLP tasks. Minimal to no need for direct supervision on any NLP task.	Its large size makes it resource-intensive and infeasible for real-world applications. It underperforms on specific NLP tasks, such as QA. Bias in output sequence due to training data.
	GPT - 2 (345M)		~ 45				
	GPT - 2 (762M)		50				
	GPT - 2 (1542M)		55				
	GPT - 2 (117M)	Accuracy	~ 1%	Natural Questions dataset			
	GPT - 2 (345M)		~ 2%				
	GPT - 2 (762M)		< 3%				
	GPT - 2 (1542M)		4.1%				
[53]	DIALOGPT (117M)	BLEU, B-2	10.54%	147M Reddit exchanges from the year 2005 to 2017	Extend GPT-2 model to create an improved conversational agent trained on massive open domain data.	Captures long-term context. Improves speaker consistency. A bigger model size makes training on larger datasets possible. Improved training time and better response generation due to lack of recurrent layers.	Generates improper/offensive responses due to bias and offensive remarks in the training data.
	DIALOGPT (345M)		16.96%				
	DIALOGPT (345M,Beam Search)		21.76%				
	DIALOGPT (762M)		18.66%				
[54]	$Meena_{Base}$	SSA	72%	Publicly available conversations scraped and filtered from social media websites, amounting to 40 billion words or 341GB of data.	Leverage an ET architecture with more training data to create a state-of-art chatbot model.	Better performance than previous open-domain conversational models. Multiple ET decoder blocks increase response sensibility and specificity.	Prone to bias or lack of factuality. The evaluation method fails to capture all aspects of an ideal human conversation.
	$Meena_{Full}$		79%				
[56]	GPT - 3 (Zero-shot	F1	81.5	CoQA dataset	They are scaling up	Minimal or no fine-tuning is	Large size makes inferences slower

Ref.	Technique	Evaluation Metric	Dataset	Objective	Merits	Demerits
	learning)			language	required for	and infeasible for
	GPT - 3	84.0		models to	various NLP	real-world
	(One-shot			improve their	tasks.	applications.
	learning)			one-shot and	Works	Lacks
	GPT - 3	85.0		zero-shot	tremendously	interpretability.
	(Few shot			learning	well in one-shot	Bias in output
	learning)			performance	and zero-shot	sequence due to
				and surpass	settings.	training data.
	GPT - 3	59.5	SQuAD 2.0	state-of-art	Outperforms	It can lose
	(Zero-shot		dataset	models in	state-of-art	comprehensibility
	learning)			few-shot	models in a few-	and repeat words or
	GPT - 3	65.4		learning on	shot setting.	sentences
	(One-shot			various NLP		occasionally.
	learning)			tasks.		
	GPT - 3	69.8				
	(Few shot					
	learning)					

*These models use multiple datasets for training; we only talk about their performance on the ones relevant to QA/conversational bot.

generative-based and retrieval-based chatbots. Generative-based models precede retrieval-based counterparts with an impressive 94.45% accuracy, enabling text generation.[55] The primary distinction is that generative models can generate text and can produce original material, whereas retrieval-based models can only reply by using pre-existing information. Notably, the generative-based Convolutional Neural Network is better equipped to handle overfitting and accuracy compared to retrieval-based models like LSTM and GRU.[55]

4. Evaluation metrics for conversational BO

The authors address the evaluation metrics for conversational bots, encompassing various performance measures commonly adopted by NLP researchers. Fig. 17 provides a comprehensive overview of these metrics.

Discrete components of a conversational bot can be evaluated by taking into account the different metrics such as:

- Accuracy, Precision, and Recall: These metrics are predominantly used for classification tasks. For example, given a confusion matrix helps us evaluate how our model has generated the correct predictions. These can be calculated using the methodology shown in the Fig. 18.[57]
- F1–Score: This evaluation metric helps strike the right balance between precision and recall. This is usually used when working with an unbalanced dataset. An F1 score of 1 means the model is doing tremendously well, and that of 0 means the model is struggling. F1–Score is calculated a

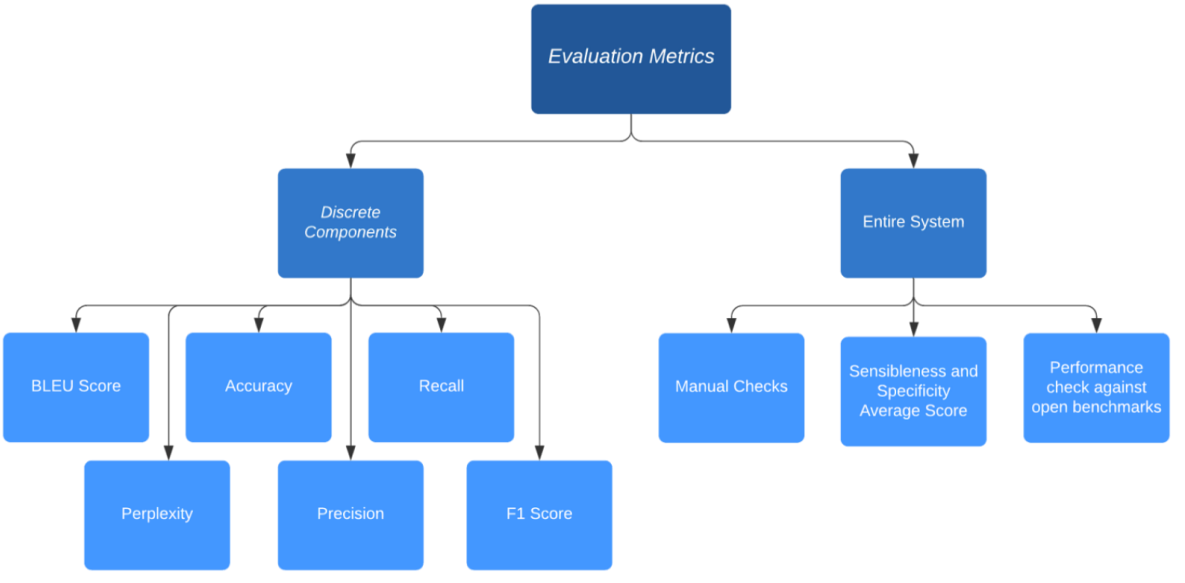


Fig. 17 Evaluation Metrics for conversational bots.

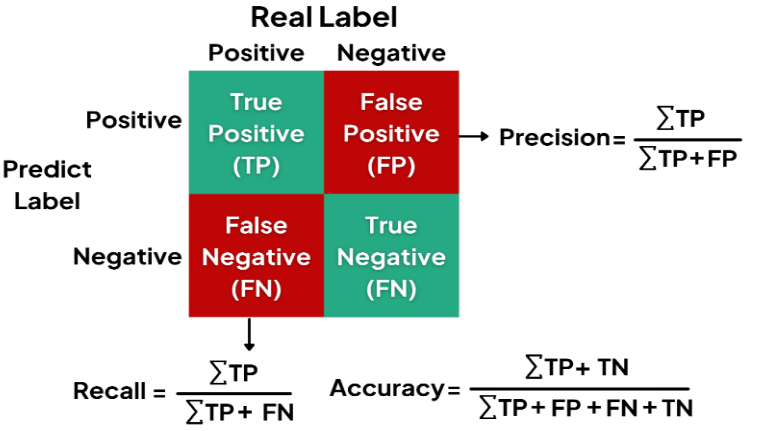


Fig. 18 Confusion matrix-based calculations.

equation (4).

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

• BLEU score (Bilingual Evaluation Understudy): A score or an evaluation metric that estimates the quality of text produced from an NLP model.[58] For example, a score of 1.0 represents an ideal match, and 0.0 represents a perfect mismatch. Using the BLEU score algorithm, we generate multiple n-gram scores for each response produced by the conversational bot for a particular input. Subsequently, we examine our bots based on these results; however, this metric may not be deemed highly effective.

• Perplexity: It helps us find the similarity between the output generated by the model and the expected output present in our data.[57] It is defined as the normalized inverse probability of the predicted outcome as represented in Equation 5.

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}} \quad (5)$$

N corresponds to the number of words in the sequence, and w_1, w_2, …, w_N are the N-words in that sequence. This makes P(w_1, w_2, …, w_N) the probability assigned to that sequence. The lower the perplexity, the better the model is said to perform. It is also considered a superior evaluation metric for chatbots as compared to BLEU Score.

Judging the performance of the system as a whole is slightly more challenging and involves:

• Manual performance checks: A popular method on this front is the Turing Test,[59] first proposed by Alan Turing in 1950. The test involves an interrogator and two respondents, one human and the other a conversational bot. For the conversational bot to pass the test, it should fool the interrogator by providing intelligent human-like responses. Generally, a time limit of 5-minutes is to be followed for this test.

• Sensibleness and Specificity Average (SSA) Score: Proposed by Google's Research team,[54] this metric aims to evaluate the quality of the response generated by a conversational bot. First, a set of crowd workers are asked to rate each response as sensible, *i.e.*, not illogical or confusing or factually incorrect, and specific, *i.e.*, context-relevant. Later an average of Sensibleness and Specificity is taken to arrive at the final SSA Score.

• Measuring performance against some open benchmarks: Dialog System Technology Challenges (DSTC), Alexa Prize, *etc.*, are specific popular benchmarks against which the performance of a conversational bot can be evaluated.[4]

During the validation process of conversational AI systems for practical use in controlled environments, assumptions are made about ideal user scenarios. Evaluations involve user experience assessments and performance comparisons, employing metrics such as accuracy and precision. The limitations stem from the early-stage development of these systems, holding critical implications for broader industry adoption.[50] For instance, Adel and Elhakeem (2023) scrutinize their blockchain-based AI system, revealing challenges in handling CRUDQ functionalities in practical construction scenarios.[61] Wu and Shen (2023) identify issues in keyword extraction and ontology limitations during system validation with 50 real-world queries, highlighting the practical implications of automated metrics.[62-65]

5. Challenges and Solutions

Response generation faces many obstacles that are specific to conversational AI. These have been listed below and presented in Fig. 19.

5.1 Bland responses

Dialogue generation models, especially those following a seq2seq architecture, tend to suffer from this problem.[66-68] As a result, the model stands a chance of producing a response like "I don't know" for every input sequence.

Solution: To tackle this, it is preferable to optimize Maximum Mutual Information (MMI) during the time of inference.[66] Considering S to be conversation history, the goal during inference is to find the best possible T according to Equation (6) from [66].

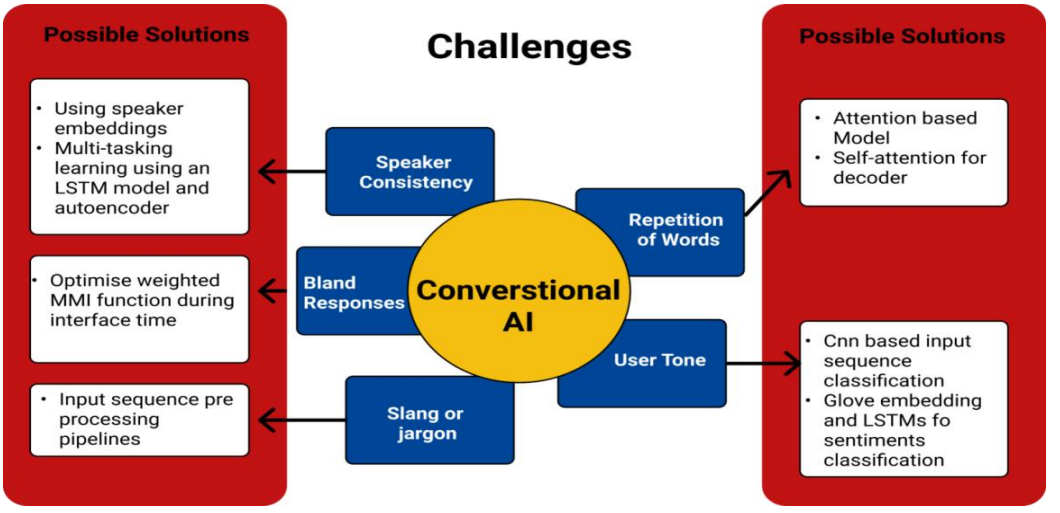


Fig. 19 Challenges faced by conversational bots and possible solutions.

$$\hat{T} = argmax_T \left\{ log \frac{p(S,T)}{p(S)p(T)} \right\}$$
$$= argmax_T \{ log\, p(T|S) - log\, p(T) \} \quad (6)$$

A parameter lambda is used to penalize generic or bland responses to a certain degree as formulated from [66].

$$\hat{T} = argmax_T \{ log\, p(T|S) - \lambda\, log\, p(T) \}$$
$$= argmax_T \{ (1-\lambda)\, log\, p(T|S) + \lambda\, log\, p(S|T) - \lambda\, log\, p(S) \}$$
$$= argmax_T \{ (1-\lambda)\, log\, p(T|S) + \lambda\, log\, p(S|T) \} \quad (7)$$

Thus, this weighted MMI objective function (Equation 7) represents a trade-off between response appropriateness and lack of blandness.

5.2 Repetition of words

Dialogue systems may fall short in gauging the tone and emotions of the user, as can be seen in the case of sarcasm, anger, sadness, or enthusiasm.[62]

Solution: Employing a CNN[63] to classify the sentiment of the input sentence.[62] Leveraging simple LSTMs for a sentiment classification task on input sentences transformed into embeddings using GloVe.[62,64]

5.3 User tone

Content or word repetition is a common issue in text generation use cases.[60,61] One word from the input sequence can map into many words or phrases in the output sequence. Thus, making the possibility of repetition quite challenging.

Solution: Attention models address this issue, but they perform better for machine translation than dialogue generation.[46] A better way to resolve this issue is to use self-attention for the decoder.[61] This helps avoid unnecessary word repetition in the target sequence and keeps the responses natural.

5.4 Slang or Jargon

Language models often fail to comprehend slang or jargon or elongated words or short forms (such as lol, omg, ik, iykyk, wdym, *etc.*). In addition, their performance suffers when the input sequence contains emojis.[62]

Solution: This can be mitigated using elaborate pre-processing pipelines for the input text.[62]

5.5 Speaker consistency

Response generation models, especially those following a seq2seq architecture, tend to face this issue. Their responses are often incoherent or inconsistent.[69] This is primarily because of the training data. To a question such as "What is your age?" The model can develop a response such as "17,21,35," as seen in the training data, but inappropriate. This makes it challenging to maintain a consistent persona in a dialogue generation model.[69]

Solution: The LSTM model can be modified to resolve this problem by incorporating speaker embeddings alongside word embeddings,[69] as shown in Fig. 20. Both of them function in a similar manner. Speaker embeddings can be visualized in a latent space wherein two adjacent speakers talk in a similar fashion or about similar topics. Each hidden state is used as an input during training, the previous hidden state, current word embedding, and the speaker embedding. During inference time, the speaker needs to be mentioned to generate responses in their style.

Another approach involves training an LSTM model on conversational data and using an autoencoder alongside, trained on non-conversational data.[65] Multi-task learning is leveraged by tying the decoders of the LSTM model and the autoencoder. This way, person-based conversational bot can be constructed without using persona-based data since such data is challenging to acquire.

6. Research scope

Since DL research is taking up at a huge speed, there have been varied interests for its actual usage in our day-to-day life. By the comprehensive review of papers relevant to this study, some significant challenges and potential research avenues are described into five major categories listed in Fig. 21.

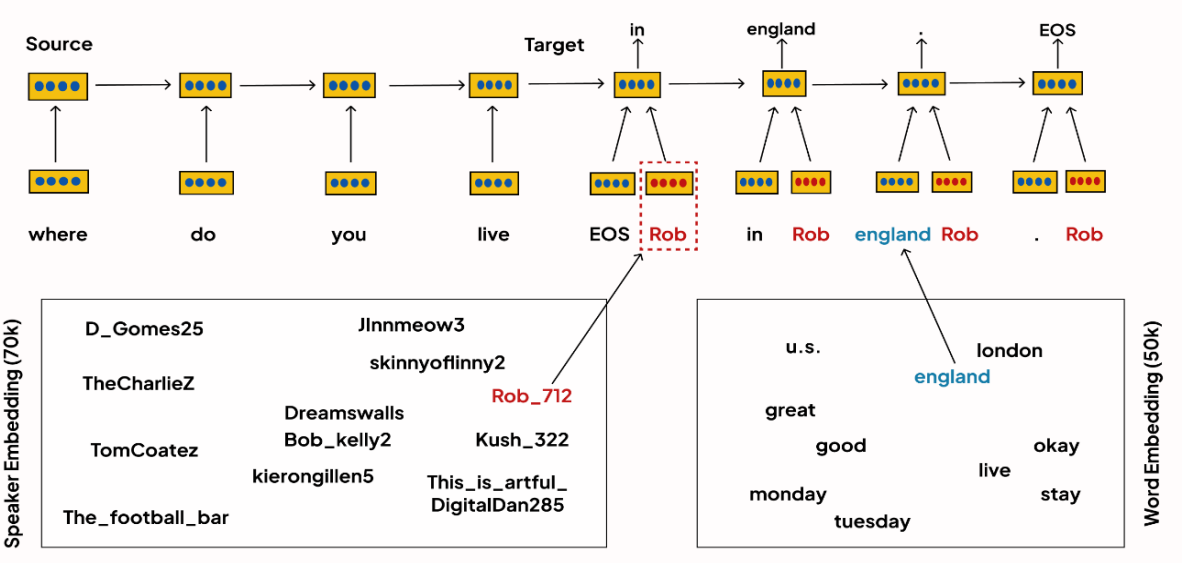


Fig. 20 Persona-based conversational system. Reproduced with the permission from [69].

6.1 Input data

The effectiveness of a DL model heavily relies on the quality and quantity of the training data. When it comes to Conversational datasets, there is an evident lack of high-quality datasets that could be used to train high-performing models, as seen in Section 2. Most of the existing datasets pertain to single-turn dialogue (or QA). Those which are open-domain and relevant to multi-turn dialogue have been obtained mainly from movies/TV shows or Social Media websites, hence being prone to bias, sarcasm, other unwanted tones, lack of context-relevance, and speaker persona. Thus, arises the need for a dataset consisting of real-world high-quality one-on-one human conversations. Researchers can create such authentic datasets, which would, in turn, result in improved conversational bots. Investigating alternative solutions that might reduce the difficulties of being relevant in such a dynamic industry is crucial in order to improve the input data. Some examples of these solutions could include using crowdsourcing techniques for real-time data enrichment or automated procedures for frequent updates.

6.2 Data pre-processing

A prominent challenge that language models have encountered recently is their struggle to comprehend and interpret slang or jargon. They also fail to understand short forms such as lol, omg, ik, iykyk, wdym, *etc.*, or emojis,[62] which have become a prominent part of the modern chat conversation style. Researchers can explore sophisticated pre-processing algorithms to address these aberrations in the input sequence to address this issue. The naïve approach would be to use a lookup database to fill in these terms before the language model receives it as an input.

6.3 Transfer learning

Transfer learning is a methodology used to train AI models such that the knowledge learned from one task is used to address a new task. It essentially focuses on transferring knowledge from the last task to the new task. Over the past decade, with the growing popularity of the transformer architecture, it has become abundantly clear that huge language models trained on an unlabeled corpus tend to perform well when fine-tuned on specific tasks such as question answering, reading comprehension, dialogue generation, *etc.*[19,50,51,52,56] This can be of great help when limited labeled data is available for training, as is the case most times. Furthermore, it saves time and computational power required to train a model from scratch. Hence, given the significant lack of multi-turn dialogue data, transfer learning on state-of-the-art pre-trained models can be used to create superior conversational bots with limited training data at hand.[53]

6.4 Performance optimization

Every DL model is prone to error, and conversational bots are no exception. Since recent conversational agents employ a generative approach, on occasion, the response produced by the agent could be inappropriate, irrelevant, bland, factually incorrect, uninteresting, or a simple "I do not know." Multiple research directions can be explored to mitigate this.

6.4.1 Reinforcement learning

Reinforcement Learning (RL) is a subset of machine learning that involves training an agent/model by interacting with its environment. The model's goal is termed as "objective". The agent is rewarded for every step it takes to reach closer to its objective and is penalized in other cases. This is done with the help of a reward function. During training, the agent aims at maximizing its reward. RL provides plenty of opportunities for optimizing the responses of conversational bots and has seen growing research interest. For example,[66] have leveraged

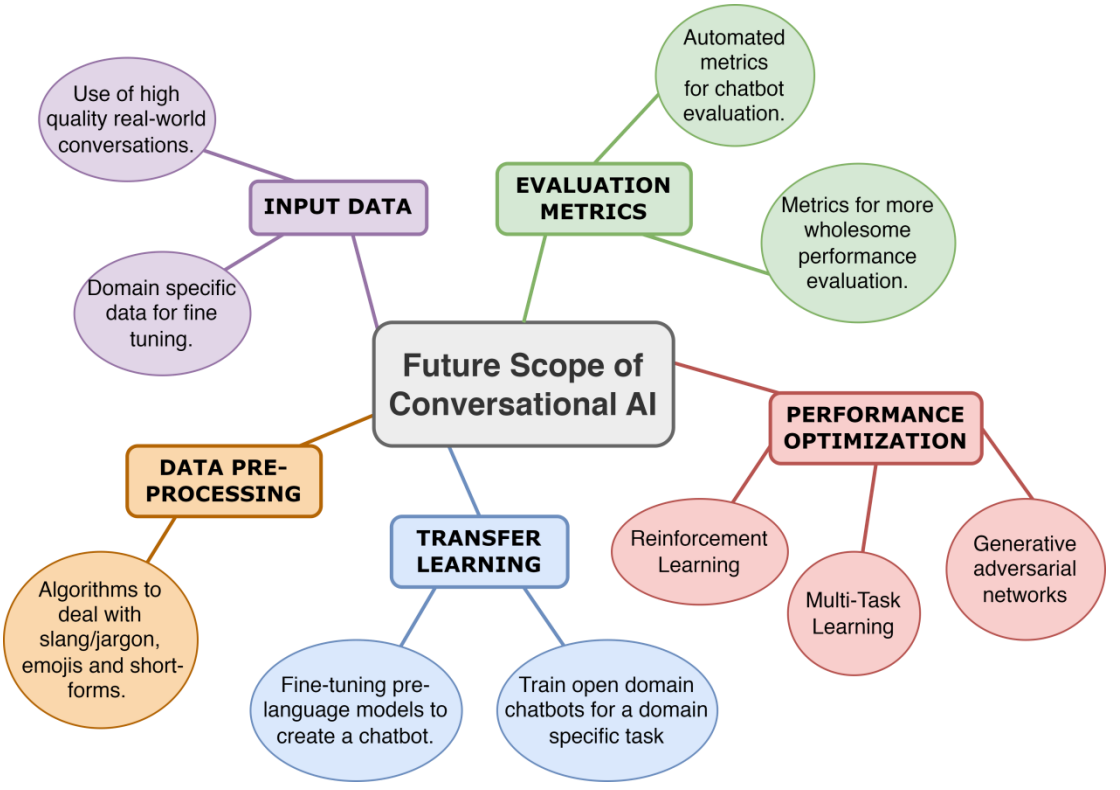


Fig. 21 Future research directions for conversational AI.

this to reduce an RNN-based model (employing a seq2seq framework) from producing responses such as "I do not know." Apart from that, RL can also be used alongside the transformer-based models that have outperformed traditional RNN-based models in various language tasks. This is a research direction that has not been explored before. Furthermore, researchers can tweak the reward function of an RL agent to make the conversational bot sound more interesting, witty, empathetic, insightful, positive, *etc*. Finally, RL also ensures that continuous user feedback is integrated into the system to improve the model, as it continually keeps interacting with the environment (in this case, the user).

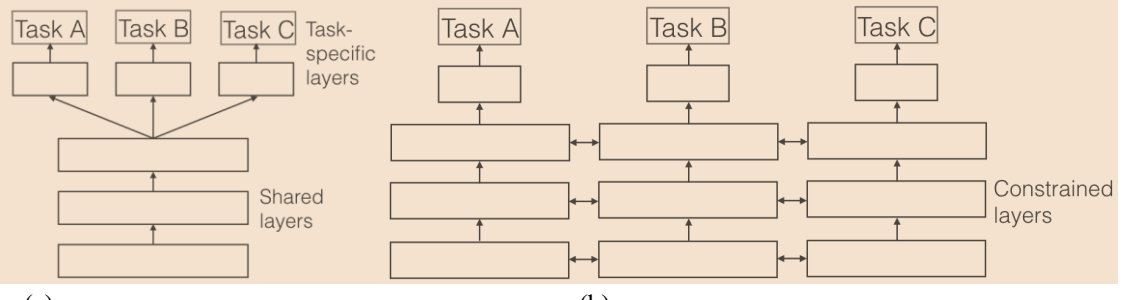
6.4.2 Multi–Task learning

Multi–task learning (MTL) is when the model learns more than one task in a single go. In such scenarios, multiple loss functions are present. MTL can predominantly adopt two approaches: Hard parameter sharing, where different tasks use the same hidden layer but different output layer; Soft parameter sharing, where each task has its own set of hidden and output layers, but the hidden layer parameters are encouraged to be similar.

Thus, another relatively new approach is to employ MTL (Fig. 22), enabling the model to gauge user sentiment with the help of sentiment analysis as it generates responses. Again, this approach ensures that the model demonstrates empathy. This also enables training multiple language models simultaneously for similar tasks.

6.4.3 Generative adversarial networks (GANs)

Generative Adversarial Networks are two-part models, consisting of one neural network called the 'Generator,' and another known as the 'Discriminator.' As the name states, the 'Generator' is responsible for creating new samples of data from the data it has been given for training. The 'Discriminator' is tasked with judging the 'Generator's' output as classifying



(a) (b)

Fig. 22 Parameter sharing for Multi-Task Learning: (a) Hard parameter sharing, (b) Soft parameter sharing. Reproduced with the permission from [3].

it as authentic or fake. The two of them are caught up in a zero-sum game until the point where the 'Generator' can fool the 'Discriminator.'

This makes the domain of GANs highly sought after by researchers to optimize model performance[67-75] by tackling inappropriate, bland, or simple "I do not know" responses given by conversational bots. The conversational model can act as the Generator by utilizing adversarial training and keeping the min-max objective in mind, and the conversational model can act as the "Generator." The "Discriminator" could then check for the authenticity of the responses. After protracted training, this would result in a conversational bot that provides close to human performance.

6.5 Evaluation metrics

Evaluating the responses of conversational bots is also a growing research area since perplexity and BLEU scores alone fail to provide a comprehensive assessment, they only verify the semantics of the sentence. Improved methods like the SSA Score have sprung up to judge the sensibility and context relevance of the responses.[54] Yet manual evaluation, which is incredibly time-consuming, cost-ineffective, prone to bias, and not standardized, is required. Hence arises the need for better metrics that check for the context relevance of responses, factual correctness, logicality, grammatical correctness, interestingness, insightfulness, *etc*. Recent research efforts[18] have focused on exploring enhanced manual evaluation metrics. However, automated tests that present a wholesome picture of the chatbot's performance is a somewhat less explored research direction that shows tremendous promise and can help us create far more superior and robust chatbots.

7. Conclusions

Conversational AI has been a rapidly evolving field in the past decade. This paper surveyed that journey by reviewing the datasets, summarizing recent approaches and the metrics used to evaluate them. Various challenges in dialogue systems are also discussed in the paper. It also presented plausible solutions to tackle challenges while talking about the future of conversational bots. Conversational bots can be retrieval-based (or KB-QA) systems or generative (or neural) systems. The generative models developed in the last couple of years leverage the breakthroughs in Reinforcement Learning and Deep Learning to improve the performance of AI agents across various domains. Most state-of-the-art conversational bots use entirely data-driven and E2E responses, usually employing RNNs or attention modules. Conclusively, as conversational AI advances, it is imperative to recognize and address persistent challenges in dialogue systems, ensuring the development of more nuanced and contextually aware conversational bots. The trajectory of future research must prioritize enhancing input data quality, refining data pre-processing techniques, and exploring innovative methodologies. Self-supervised and unsupervised techniques would of a huge potential in coming years of technological evolution. In a nutshell, conversational AI bots would be an exciting milestone in the NLP domain and would attract both industry and researchers.

Acknowledgement

The authors acknowledge Symbiosis International (Deemed University), Pune, India.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

References

[1] J. G. Carbonell, R. S. Michalski, T. M. Mitchell, An overview of machine learning. Berlin, Heidelberg: Springer, 1983.

[2] G. G. Chowdhury, Natural language processing, *Annual Review of Information Science and Technology*, 2003, **37**, 51-89, doi: 10.1002/aris.1440370103.

[3] S. Ruder, An overview of multi-task learning in deep neural networks. arXiv 2017.

[4] J. Gao, M. Galley, L. Li, Neural Approaches to Conversational AI, The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor MI USA. ACM, 2018.

[5] E. Adamopoulou, L. Moussiades, Chatbots: history, technology, and applications, *Machine Learning with Applications*, 2020, **2**, 100006, doi: 10.1016/j.mlwa.2020.100006.

[6] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, *ACM Computing Surveys*, 2022, **54**, 1-36, doi: 10.1145/3453154.

[7] K. Venusamy, N. Krishnan Rajagopal, M. Yousoof, A study of Human Resources Development through Chatbots using Artificial Intelligence, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). Thoothukudi, India. IEEE, 2020.

[8] J, M. Parvathi, A study on future of artificial intelligence - chatbots in HR, In Proceedings of the National conference on Economic Slowdown: Measures to Revive the Paranoid (ESMRP-19); International Journal of Trend in Research and Development, September 25, 2019.

[9] S. Upadhye, Comparative analysis of virtual personal assistant(s). IJARIIE 2019.

[10] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, J. B. Torous, Chatbots and conversational agents in mental health: a review of the psychiatric landscape, *The Canadian Journal of Psychiatry*, 2019, **64**, 456-464, doi: 10.1177/0706743719828977.

[11] C. V. Misischia, F. Poecze, C. Strauss, Chatbots in customer service: their relevance and impact on service quality, *Procedia Computer Science*, 2022, **201**, 421-428, doi:

10.1016/j.procs.2022.03.055.
[12] L. Jenneboer, C. Herrando, E. Constantinides, The impact of chatbots on customer loyalty: a systematic literature review, *Journal of Theoretical and Applied Electronic Commerce Research*, 2022, **17**, 212-229, doi: 10.3390/jtaer17010011.
[13] M. E. M. Gonzales, E. B. H. Ong, C. K. Cheng, E. C. J. Ong, J. J. Azcarraga, From unstructured to structured: transforming chatbot dialogues into data mart schema for visualization, ArXiv 2023.
[14] A. Følstad, P. B. Brandtzaeg, Users' experiences with chatbots: findings from a questionnaire study, *Quality and User Experience*, 2020, **5**, 3, doi: 10.1007/s41233-020-00033-2.
[15] H. Raval, Limitations of existing chatbot with analytical survey to enhance the functionality using emerging technology, International Journal of Research and Analytical Reviews, 2020.
[16] J. Berant, A. Chou, R. Frostig, P. Liang, Semantic parsing on freebase from question-answer pairs, EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, *Proceedings of the Conference*, 2013, 1533-1544.
[17] J. Bao, N. Duan, M. Zhou, T. Zhao, Knowledge-Based Question Answering as Machine TranslationProceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.
[18] E. Collins, Z. Ghahramani, LaMDA: Our breakthrough conversation technology, 2021.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017.
[20] L. C. Klopfenstein, S. Delpriori, S. Malatini, A. Bogliolo, The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and ParadigmsProceedings of the 2017 Conference on Designing Interactive Systems. Edinburgh United Kingdom. ACM, 2017.
[21] M. Zaib, W. E. Zhang, Q. Z. Sheng, A. Mahmood, Y. Zhang, Conversational question answering: a survey, *Knowledge and Information Systems*, 2022, **64**, 3151-3195, doi: 10.1007/s10115-022-01744-y.
[22] Q. Motger, X. Franch, J. Marco, Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges. arXiv 2021.
[23] G. Caldarini, S. Jaf, K. McGarry, A literature survey of recent advances in chatbots, *Information*, 2022, **13**, 41, doi: 10.3390/info13010041.
[24] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, A. Abraham, AI-based conversational agents: a scoping review from technologies to future directions, *IEEE Access*, 2022, **10**, 92337-92356, doi: 10.1109/ACCESS.2022.3201144.
[25] C.-C. Lin, A. Y. Q. Huang, S. J. H. Yang, A review of AI-driven conversational chatbots implementation methodologies and challenges (1999–2022), *Sustainability*, 2023, **15**, 4012, doi: 10.3390/su15054012.
[26] D. Ameixa, L. Coheur, P. Fialho, P. Quaresma, Luke, I am your father: dealing with out-of-domain requests by using movies subtitles. Bickmore T, Marsella S, Sidner C, International Conference on Intelligent Virtual Agents. Cham: Springer, 2014.
[27] J. McAuley, A. Yang, Addressing complex and subjective product-related queries with customer reviews, Proceedings of the 25th International Conference on World Wide Web. Montréal Québec Canada. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016.
[28] M. Wan, J. McAuley, Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems, 2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain. IEEE, 2016.
[29] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding Comprehension Dataset From ExaminationsProceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.
[30] P. Rajpurkar, R. Jia, P. Liang, Know What You Don't Know: Unanswerable Questions for SQuADProceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.
[31] S. Reddy, D. Chen, C. D. Manning, CoQA: A conversational question answering challenge, *Transactions of the Association for Computational Linguistics*, 2019, **7**, 249-266, doi: 10.1162/tacl_a_00266.
[32] O. Vinyals, Q. A. Le, Neural Conversational Model. arXiv 2015, doi: 10.48550/ARXIV.1506.05869.
[33] R. Lowe, N. Pow, I. Serban, J. Pineau, The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue SystemsProceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Prague, Czech Republic. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015.
[34] P. Lison, J. Tiedemann, OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Proceedings of the Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); European Language Resources Association (ELRA): Portorož, Slovenia, May 2016, 923–929.
[35] I. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, **30**, doi: 10.1609/aaai.v30i1.9883.
[36] A. Ritter, C. Cherry, B. Dolan, Unsupervised modeling of twitter conversations. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational

Linguistics; Association for Computational Linguistics: Los Angeles, California, June 2010.
[37] S. D. Richardson, W. B. Dolan, L. Vanderwende, MindNet: acquiring and structuring semantic information from textProceedings of the 36th annual meeting on Association for Computational Linguistics -. August 10-14, 1998. Montreal, Quebec, Canada. Morristown, NJ, USA: Association for Computational Linguistics, 1998, **2**, 1098–1102, doi: 10.3115/980691.980749.
[38] W.-T. Yih, M.-W. Chang, X. He, J. Gao, Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge BaseProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015.
[39] J. Wu, M. Li, C.-H. Lee, A probabilistic framework for representing dialog systems and entropy-based dialog management through dynamic stochastic state evolution, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **23**, 2026-2035, doi: 10.1109/TASLP.2015.2462712.
[40] B. Dhingra, L. Li, X. Li, J. Gao, Y. N. Chen, F. Ahmed, L. Deng, Towards end-to-end reinforcement learning of dialogue agents for information access. in proceedings of the proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics: Vancouver, Canada, 2017.
[41] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine TranslationProceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.
[42] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, 2014.
[43] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, J.-Y. Nie, A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query SuggestionProceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne Australia. ACM, 2015.
[44] A. Hannan, S. K. Sarma, Z. Hussain, Marie A statistical approach to build a machine translation system for English Assamese language pair, *International Journal of Computer Sciences and Engineering*, 2019, **7**, 774-779, doi: 10.26438/ijcse/v7i3.774779.
[45] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, *ArXiv e-Prints*, 2014, 1409.0473, doi: 10.48550/arXiv.1409.0473.
[46] H. Mei, M. Bansal, M. Walter, Coherent dialogue with attention-based language models, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, **31**, doi: 10.1609/aaai.v31i1.10961.
[47] S. He, C. Liu, K. Liu, J. Zhao, Generating Natural Answers by Incorporating Copying and Retrieving Mechanisms in Sequence-to-Sequence LearningProceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.
[48] J. Gu, Z. Lu, H. Li, V. O. K. Li, Incorporating Copying Mechanism in Sequence-to-Sequence LearningProceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.
[49] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, J. Gao, Deep Reinforcement Learning for Dialogue GenerationProceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.
[50] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training, OpenAI Blog 2018.
[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018.
[52] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 2019.
[53] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020.
[54] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, Q. Le Towards a human-like open-domain chatbot. arXiv 2020.
[55] D. R. So, C. Liang, Q. V. Le, The Evolved Transformer, 2019.
[56] S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models Are Few-Shot Learners. arXiv 2020.
[57] A. B. Sai, A. K. Mohankumar, M. M. Khapra, A survey of evaluation metrics used for NLG systems, *ACM Computing Surveys*, 2023, **55**, 1-39, doi: 10.1145/3485766.
[58] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translationProceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. July 7-12, 2002. Philadelphia, Pennsylvania. Morristown, NJ, USA:

Association for Computational Linguistics, 2001.
[59] A. M. Turing, Computing Machinery and Intelligence. In Parsing the Turing Test; R. Epstein, G. Roberts, G. Beber, Eds.; Springer Netherlands: Dordrecht, 2009.
[60] A. Holtzman, J. Buys, M. Forbes, A. Bosselut, D. Golub, Y. Choi, Learning to Write with Cooperative DiscriminatorsProceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018.
[62] A. Pamnani, R. Goel, J. Choudhari, M. Singh, IIT Gandhinagar at SemEval-2019 Task 3: Contextual Emotion Detection Using Deep LearningProceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.
[63] S. Ruder, P. Ghaffari, J. G. Breslin, INSIGHT-1 at SemEval-2016 Task 4: Convolutional Neural Networks for Sentiment Classification and QuantificationProceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.
[64] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.
[65] Y. Luan, C. Brockett, B. Dolan, J. Gao, M. Galley, Multi-task learning for speaker-role adaptation in neural conversation models. In Proceedings of the Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Asian Federation of Natural Language Processing: Taipei, Taiwan, November, 2017.
[66] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.
[67] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, B. Dolan, Generating informative and diverse conversational responses via adversarial information maximization, 2018.
[68] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, D. Jurafsky, Adversarial Learning for Neural Dialogue Generation Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.
[69] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, B. Dolan, A Persona-Based Neural Conversation Model Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.
[70] G. Caldarini, S. Jaf, K. McGarry, A literature survey of recent advances in chatbots, *Information*, 2022, **13**, 41, doi: 10.3390/info13010041.
[71] S. Pandey, S. Sharma, A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning, *Healthcare Analytics*, 2023, **3**, 100198, doi: 10.1016/j.health.2023.100198.
[72] F. M. Shiri, T. Perumal, N. Mustapha, R. A. Mohamed, Comprehensive overview and comparative analysis on deep learning models, CNN, RNN, LSTM, GRU 2023.
[73] A. B. Saka, L. O. Oyedele, L. A. Akanbi, S. A. Ganiyu, D. W. M. Chan, S. A. Bello, Conversational artificial intelligence in the AEC industry: a review of present status, challenges and opportunities, *Advanced Engineering Informatics*, 2023, **55**, 101869, doi: 10.1016/j.aei.2022.101869.
[74] K. Adel, A. Elhakeem, M. Marzouk, Chatbot for construction firms using scalable blockchain network, *Automation in Construction*, 2022, **141**, 104390, doi: 10.1016/j.autcon.2022.104390.
[75] S. Wu, Q. Shen, Y. Deng, J. Cheng, Natural-language-based intelligent retrieval engine for BIM object database, *Computers in Industry*, 2019, **108**, 73-88, doi: 10.1016/j.compind.2019.02.016.