information

MDPI

*Article*

# State-of-the-Art in Open-Domain Conversational AI: A Survey

**Tosin Adewumi *, Foteini Liwicki and Marcus Liwicki**

ML Group, EISLAB, Luleå University of Technology, 971 87 Luleå, Sweden; foteini.liwicki@ltu.se (F.L.); marcus.liwicki@ltu.se (M.L.)
* Correspondence: oluwatosin.adewumi@ltu.se

**Abstract:** We survey SoTA open-domain conversational AI models with the objective of presenting the prevailing challenges that still exist to spur future research. In addition, we provide statistics on the gender of conversational AI in order to guide the ethics discussion surrounding the issue. Open-domain conversational AI models are known to have several challenges, including bland, repetitive responses and performance degradation when prompted with figurative language, among others. First, we provide some background by discussing some topics of interest in conversational AI. We then discuss the method applied to the two investigations carried out that make up this study. The first investigation involves a search for recent SoTA open-domain conversational AI models, while the second involves the search for 100 conversational AI to assess their gender. Results of the survey show that progress has been made with recent SoTA conversational AI, but there are still persistent challenges that need to be solved, and the female gender is more common than the male for conversational AI. One main takeaway is that hybrid models of conversational AI offer more advantages than any single architecture. The key contributions of this survey are (1) the identification of prevailing challenges in SoTA open-domain conversational AI, (2) the rarely held discussion on open-domain conversational AI for low-resource languages, and (3) the discussion about the ethics surrounding the gender of conversational AI.

**Keywords:** conversational systems; chatbots; SotA

check for updates

**Citation:** Adewumi, T.; Liwicki, F.; Liwicki, M. State-of-the-Art in Open-Domain Conversational AI: A Survey. *Information* **2022**, *13*, 298. https://doi.org/10.3390/info13060298

Academic Editor: Tudor Groza

Received: 28 May 2022
Accepted: 7 June 2022
Published: 10 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 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

There are different opinions as to the definition of AI, but according to [1], it is any computerised system exhibiting behaviour commonly regarded as requiring intelligence. Conversational AI, therefore, is any system with the ability to mimick human–human intelligent conversations by communicating in natural language with users [2]. Conversational AI, sometimes called chatbots, may be designed for different purposes. These purposes could be for entertainment or solving specific tasks, such as plane ticket booking (task-based). When the purpose is to have unrestrained conversations about, possibly, many topics, then such AI is called open-domain conversational AI. ELIZA, by [3], is the first acclaimed conversational AI (or system). Human interaction with the system demonstrated how engaging its responses could be [2]. The staff of [3] reportedly became engrossed with the program during interactions and possibly had private conversations with it [2].

Modern SoTA open-domain conversational AI aims to achieve better performance than what was experienced with ELIZA. There are many aspects and challenges to building such SoTA systems. Therefore, the primary objective of this survey is to investigate some of the recent SoTA open-domain conversational systems and identify specific challenges that still exist that should be surmounted to achieve "human" performance in the "imitation game", as described by [4]. As a result of this objective, this survey will identify some of the ways of evaluating open-domain conversational AI, including the use of automatic metrics and human evaluation. This work differs from previous surveys on conversational AI or related topics in that it presents discussion around the ethics of gender of conversational AI with compelling statistics and discusses open-domain conversational AI for low-resource languages, which is rarely held. Our approach surveys some of the most representative work in recent years.

The key contributions of this paper are (a) the identification of existing challenges to be overcome in SoTA open-domain conversational AI, (b) the rarely held discussion on open-domain conversational AI for low-resource languages, and (c) a compelling discussion about ethical issues surrounding the gender of conversational AI. The rest of the paper is organised as follows. The Background (Section 2) presents brief details about some topics in conversational AI; the Benefits of Conversational AI (Section 3) highlights some of the benefits that motivate research in conversational AI; Methods (Section 4) describes the details of the approach for the two investigations carried out in this survey; two Results of the Survey (Sections 5 and 6) then follow with details of the outcome of the methods; thereafter, the Existing Challenges (Section 7) shares the prevailing challenges to obtaining human-like performance; Open-domain Conversational AI for Low-resource Languages (Section 8) discusses this critical challenge and some of the attempts at solving it; the Related Work (Section 9) highlights previous related reviews, and the Conclusion (Section 11) summarises the study after the limitations of this work are given in the Limitation Section.

## 2. Background

Open-domain conversational AI may be designed as a simple rule-based template system or may involve complex Artificial Neural Network (ANN) architectures. Indeed, six approaches are possible: (1) rule-based method, (2) reinforcement learning (RL) that uses rewards to train a policy, (3) adversarial networks that utilise a discriminator and a generator, (4) retrieval-based method that searches from a candidate pool and selects a proper candidate, (5) generation-based method that generates a response word by word based on conditioning, and a (6) hybrid method that combines two or more of the earlier methods [2,5,6]. Certain modern systems are still designed in the rule-based style that was used for ELIZA [2]. The ANN models are usually trained on large datasets to generate responses; hence, they are data-intensive. The data-driven approach is more suitable for open-domain conversational AI [2]. Such systems learn inductively from large datasets involving many turns in conversations, such as Topical-Chat [7,8]. A turn (or utterance) in a conversation is each single contribution from a speaker [2,9]. The data may be from written conversations, such as the MultiWOZ [10], transcripts of human–human spoken conversations, such as the Gothenburg Dialogue Corpus (GDC) [11], crowdsourced conversations, such as the EmpatheticDialogues [12], and social media conversations such as Familjeliv (familjeliv.se) or Reddit (reddit.com) [13,14]. As already acknowledged that the amount of data needed for training deep ML models is usually large, they are normally first pretrained on large, unstructured text or conversations before being fine-tuned on specific conversational data.

### *2.1. Retrieval and Generation Approaches*

Two common ways that data-driven conversational AI produce turns as response are Information Retrieval (IR) and generation [2]. In IR, the system fetches information from some fitting corpus or online, given a dialogue context. Incorporating ranking and retrieval capabilities provides additional possibilities. If $C$ is the training set of conversations, given a context $c$, the objective is to retrieve an appropriate turn $r$ as the response. Similarity is used as the scoring metric, and the highest scoring turn in $C$ is selected from a potential set. This can be achieved with different IR methods and choosing the response with the highest cosine similarity with $c$ [2]. This is given in Equation (1). In an encoder–encoder architecture, for example, one could train the first encoder to encode the query, while the second encoder encodes the candidate response and the score is the dot product between the two vectors from both encoders. In the generation method, a language model or an encoder–decoder is used for response generation, given a dialogue context. As shown in Equation (2), each token of the response ($r_t$) of the encoder–decoder model is generated by conditioning on the encoding of the query ($q$) and all the previous responses ($r_{t-1}...r_1$), where $w$ is a word in the vocabulary $V$. Given the benefit of these two methods, it may be easy to see the advantage of using the hybrid of the two for conversational AI. For example, in the BlenderBot by [15], the hybrid variant uses retrieve-and-refine, whereby generation follows retrieval. More is discussed about this feature of the BlenderBot in Section 5 (show as Table 1).

$$response(c,C) = \arg\max_{r \epsilon C} \frac{c.r}{|c||r|} \quad (1)$$

$$r_t = \arg\max_{w \epsilon V} P(w|q, r_{t-1}...r_1) \quad (2)$$

**Table 1.** Pros and cons of IR, generation, and hybrid approaches.

| Retrieval | Generation | Hybrid |
|---|---|---|
| | **Pros** | |
| Possibility to incorporate domain/world knowledge [15] | Relatively unique tokens may be produced [16] | Combines the pros of both the retrieval and generation approaches |
| Up-to-date information in response from online sources | The use of decoding algorithms, in addition to other hyperparameters such as temperature, can deliver relatively diverse outputs [17] | The retrieval component may be used to provide additional context for the generation [15] |
| Possibility of more fluent or precise responses [16] | | Possibility for up-to-date responses and world/domain knowledge |
| | **Cons** | |
| Possible low diversity in the outputs | Low diversity in the overall generated outputs | Harder to implement efficiently than any of the single approaches |
| Limitation based on the size of repository | High probability of repetitive generated output [15] | Some of the cons of the combined approaches may still exist, such as limitation of repository size |
| Lack of memory to recall certain facts | Lack of memory to recall certain facts [5,15] | |
| Poor output distribution compared to human conversation | Poor output distribution compared to human conversation [17] | |
| | Lack of domain/world knowledge | |

### *2.2. Evaluation*

Although there are a number of metrics for NLP systems [18–20], different metrics may be suitable for different systems, depending on the characteristics of the system. For example, the goals of task-based systems are different from those of open-domain conversational systems, so they may not use the same evaluation metrics. Human evaluation is the *gold standard* in the evaluation of open-domain conversational AI, although it is subjective [21]. It is both time-intensive and laborious [15]. As a result of this, automatic metrics serve as proxies for estimating performance, although they may not correlate very well with human evaluation [19,22,23]. For example, IR systems may use F1, precision, recall [18] and hits@1/K [15], which measures recall@1 when ranking the gold label among K-1 other random candidates.

Furthermore, metrics used in NLG tasks, including Machine Translation (MT) metrics such as the BLEU or ROUGE, are sometimes used to evaluate open-domain conversational systems [21], but they are also discouraged because they do not correlate well with human judgment [2,24]. These are called word-overlap metrics [25]. They do not take syntactic or lexical variation into consideration [20]. Variants of the BLEU score for open-domain conversational AI also exist [26]. Reference-free automatic metrics, such as the fine-grained evaluation of dialogue (FED) or unsupervised reference-free metric (USR), also exist [25]. Both the FED and USR make use of pre-trained models for evaluation. Topic-based metrics use coherence or the engagingness of topics of conversations for evaluation, as was carried out by [27,28], who used extended deep average networks (DAN) for topic-word attention table and performed topic classification. Perplexity is commonly used for evaluation and has been shown to correlate with a human evaluation metric called Sensibleness and Specificity Average (SSA) [5]. It measures how well a model predicts the data of the test set, thereby estimating how accurately it expects the words that will be said next [5]. It is used in the evaluation of Meena [5] and BlenderBot (particularly the generation and 'retrieve and refine' variants [15]). It corresponds to the effective size of the vocabulary [18], and smaller values show that a model fits the data better. Very low perplexity, however, has been shown to suggest such text may have low diversity and unnecessary repetition [17].

### Subjective Evaluation of Conversational AI

Two subjective methods for human evaluation of open-domain conversational AI are the observer and participant evaluation [2]. Observer evaluation involves reading and scoring a transcript of human–chatbot conversation while participant evaluation interacts directly with the conversational AI [2]. In the process, the system may be evaluated for different qualities, such as humanness (or human-likeness) [15], fluency, making sense, engagingness [15], interestingness, avoiding repetition, and more. For example, LaMDA (blog.google/technology/ai/lamda/, accessed on 6 June 2022) builds on Meena [5] by using sensibleness and specificity as qualities of evaluation. Sensibleness appraises whether a response makes sense, given a context, while specificity appraises whether the response relates clearly to the context. The Likert scale is usually provided for grading these various qualities. The others are comparison of diversity and how fitting responses are to the given contexts. Inter-Annotator Agreement (IAA) may be calculated for the annotations made by the human evaluators [5,29]. Variations of human evaluation have resulted from these qualities, including ACUTE-EVAL [30] and manual response error rate (RER) [7,28]. ACUTE-EVAL involves binary judgements of multi-turn dialogues from two models, while manual RER is the ratio of the total turns with erroneous responses (incorrect, irrelevant, inappropriate) over the total number of turns. RER was used for validating the automatic topic-based metric involving DAN [28].

Many human evaluations are usually modeled to resemble the Turing test (or the imitation game). This test is the indistinguishability test. It is when a human is not able to distinguish if the responses are from another human or a machine in what is called the imitation game [4]. The proposed imitation game, by [4], involves a man, a woman, and an interrogator of either sex who is in a separate room from the man and the woman. The goal of the interrogator is to determine who is the woman and who is the man, and he does this by directing questions to the man and the woman, which are answered in some written format. When a machine replaces the man, the aim is to find out if the interrogator will decide wrongly as often as when it was played with a man [4]. A limited version of the test was introduced in 1991, alongside its unrestricted version, in what is called the Loebner Prize competition [31]. Every year, since then, prizes have been awarded to conversational AI that pass the restricted version in the competitions [32]. This competition has its share of criticisms, including the view that it is rewarding tricks instead of furthering the course of AI [31,33]. As a result of this, ref. [33] recommended an alternative approach, whereby the competition will involve a different award methodology that is based on a different set of assessment and completed on an occasional basis.

### *2.3. Characteristics of Human Conversations*

Humans converse using speech and other gestures that may include facial expressions, usually called body language, thereby making human conversations complex [2]. Similar gestures may be employed when writing conversations. Such gestures may be clarification questions or the mimicking of sound (*onomatopoeia*). In human conversations, one speaker may have the conversational initiative, i.e., the speaker directs the conversation. This is typical in an interview where the interviewer asking the questions directs the conversation. It is the style for Question Answering (QA) conversational AI. In typical human–human conversations, the initiative shifts to and from different speakers. This kind of mixed (or rotating) initiative is harder to achieve in conversational systems [2]. In addition to conversation initiative, below are additional characteristics of human conversations, according to [34].

- Usually, one speaker talks at a time.
- The turn order varies.
- The turn size varies.
- The length of a conversation is not known in advance.
- The number of speakers/parties may vary.
- Techniques for allocating turns may be used.
- Content of the conversation is not known in advance.
- The relative distribution of turns is unknown in advance.
- Different turn-constructional unit may be used, e.g., words or sentences.
- Repair mechanisms for correcting turn-taking errors exist.

### *2.4. Ethics*

Ethical issues are important in open-domain conversational AI. The perspective of deontological ethics views objectivity as being equally important [35–37]. Deontological ethics is a philosophy that emphasises duty or responsibility over the outcome achieved in decision making [38,39]. Responsible research in conversational AI requires compliance to ethical guidelines or regulations, such as the General Data Protection Regulation (GDPR), which is a regulation protecting persons with regard to their personal data [40]. Some of the ethical issues that are of concern in conversational AI are privacy, due to personally identifiable information (PII), toxic/hateful messages as a result of the training data, and unwanted bias (racial, gender, or other forms) [2,21].

Some systems have been known to demean or abuse their users. It is also well known that machine learning systems reflect the biases and toxic content of the data they are trained on [2,41]. Privacy is another crucial ethical issue. Data containing PII may fall into the wrong hands and cause a security threat to those concerned. It is important to have systems designed such that they are robust to such unsafe or harmful attacks. Attempts are being made with debiasing techniques to address some of these challenges [42]. Privacy concerns are also being addressed through anonymisation techniques [2,43]. Balancing the features of chatbots with ethical considerations can be delicate and challenging work. For example, there is contention in some quarters whether using female voices in some technologies/devices is appropriate. Then again, one may wonder if there is anything harmful about that. This is because it seems to be widely accepted that the proportion of chatbots designed as “female” is larger than those designed as “male”. In a survey of 1375 chatbots, from automatically crawling chatbots.org, Ref. [44] found that most were female.

## 3. Benefits of Conversational AI

The apparent benefits inherent in open-domain conversational AI has spurred research in the field. These benefits have led to multi-million-dollar investments in conversational AI by many organisations, including Apple [2] and Amazon [27]. Some of the benefits include:

- Provision of ’friendly’ company, as was probably experienced with Kuki (formerly Mitsuku) [45] and ELIZA (though it was not intended to provide such company). Some of the staff of [3] reportedly found comfort in holding private conversations with the conversational agent [2].
- Provide support for users with disabilities, such as blindness [20]. Speech-to-text (STT) and text-to-speech (TTS) technologies combined with conversational AI can make life easier for people with disabilities.
- A channel for providing domain/world knowledge [20]. The IR approach discussed earlier can make it possible to have up-to-date information on specific domains or topics through conversational AI.
- The provision of educational content or information in a concise fashion [46]. As mentioned earlier, the content and length of a conversation are not known in advance, so it is possible to construct utterances that are relatively concise and to the point.
- Automated machine–machine generation of quality data for low-resource languages [14]. The challenge of data scarcity for low-resource languages may be mitigated through quality data generated from autonomous machine–machine conversations on various topics and about different entities.
- The possibility of modelling human psychiatric/psychological treatment [2] on the basis of favorable behavior determined from experiments which are designed to modify input–output behaviour.

## 4. Methods

We conduct two different investigations to make up this survey. Figure 1 depicts the methods for both investigations. The first addresses text-based, open-domain conversational AI in terms of architectures, while the second addresses the ethical issues about the gender of conversational AI systems. The first involves an online search on Google Scholar and regular Google Search, using the term “state-of-the-art open-domain dialogue systems”. This returned 5130 and 34,100,000 items in the results for Google Scholar and Google Search, respectively. We then sieve through the list of scientific papers (within the first ten pages because of time constraints) to identify those that report SoTA results in the last five years (2017–2022) in order to give more attention to them. This provides recent advances in the field. It is important to note that some Google Scholar results point to other databases, such as ScienceDirect, arXiv, and IEEEXplore.

The reason for also using the regular Google Search is because it provides results that are not necessarily based on peer-reviewed publications but may be helpful in leading to peer-reviewed publications that may not have been immediately obvious on Scholar. We did not discriminate the papers based on the field of publication, as we are interested in as many SoTA open-domain conversational systems as possible within the specified period. A second stage involves classifying, specifically, the SoTA open-domain conversational AI from the papers based on their architecture. We also consider models that are pre-trained on large text and may be adapted for conversational systems, such as the Text-to-Text Transfer Transformer (T5) [47], and autoregressive models because they easily follow the NLG framework. Autoregressive models are generators that condition each output word on previously generated outputs sequentially [48]. We do not consider models for which we did not find their scientific papers.

The second investigation, which addresses the ethical issues surrounding the gender of conversational AI, involves the survey of 100 chatbots. It is based on binary gender: male and female. The initial step was to search using the term “gender chatbot” on Google Scholar and note all chatbots identified in the scientific papers in the first ten pages of the results. Then, using the same term, the Scopus database was queried, and it returned 20 links. The two sites resulted in 120 links, from which 59 conversational systems were identified. Since Facebook Messenger is linked to the largest social media platform, we chose this to provide another 20 chatbots. They are based on information provided by two websites on some of the best chatbots on the platform (enterprisebotmanager.com/chatbot-examples, growthrocks.com/blog/7-messenger-chatbots). The sites were identified on Google by using the search term "Facebook Messenger best chatbots". They were selected based on the first to appear in the list. To make up part of the 100 conversational AI, 13 chatbots, which have won the Loebner prize in the past 20 years, are included in this survey. Finally, 8 popular conversational AI, which are also commercial, are included. These are Microsoft's XiaoIce and Cortana, Amazon's Alexa, Apple's Siri, Google Assistant, Watson Assistant, Ella, and Ethan by Accenture.

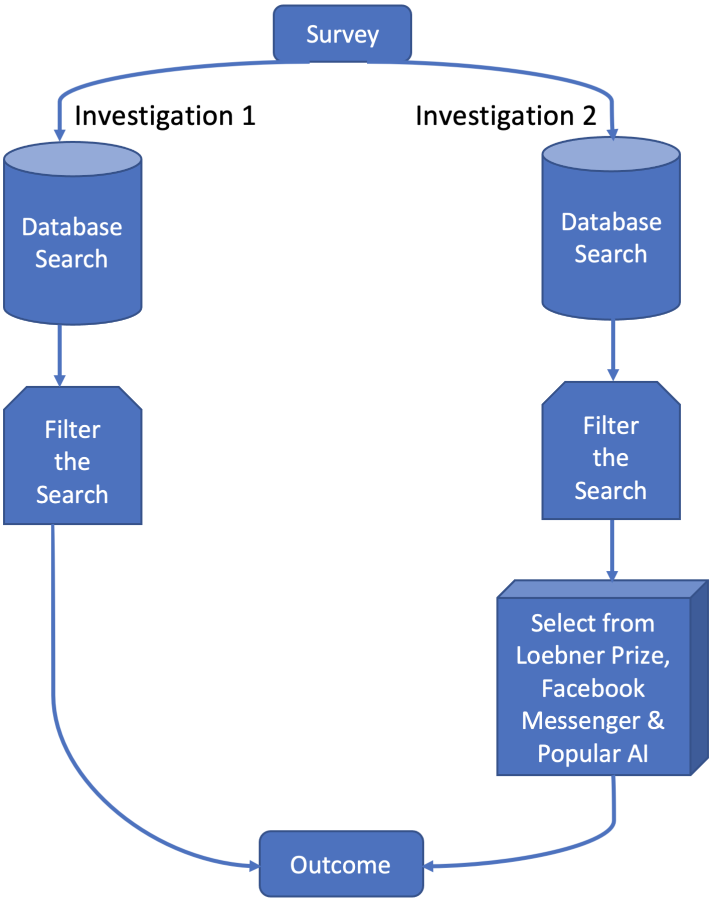


**Figure 1.** Method for both investigations in this study.

## 5. Results of Survey: Models

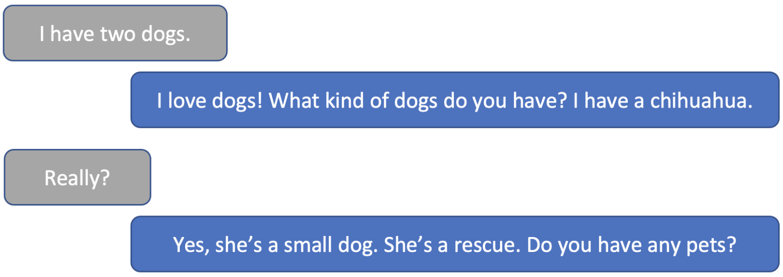
A review of the different scientific papers from the earlier method shows that recent SoTA open-domain conversational AI models fall into one of the latter three approaches mentioned in Section 2: (a) retrieval-based, (b) generation-based, and (c) hybrid approaches. The models are BlenderBot 1 and 2, Meena, DLGNet, Dialogue Generative Pre-trained Transformer (DialoGPT), RetGen, Generative Pre-trained Transformer (GPT)-3 and 2, and Text-to-Text Transfer Transformer (T5). The last one, T5, is not a conversational AI but an encoder–decoder architecture that may be adapted for conversational AI.

### *5.1. BlenderBot 1 & 2*

Some of the ingredients for the success of BlenderBot, as identified by [15], are empathy and personality, consistent persona, displaying knowledge, and engagingness. Three different parameter models are built for the variants: 90M, 2.7B, and 9.4B. The variants, which are all based on the Transformer, involve the latter three approaches: retrieval, generative, and a retrieve-and-refine combination of the earlier two. The generative architecture is a seq2seq model and uses Byte-Level BPE for tokenisation. The retrieve-and-refine variant combines a retrieval step before conditioning the retrieved response for the generation in order to alleviate problems such as dull, repetitive responses and knowledge hallucination. Two variants of the retrieve-and-refine method are used: dialogue retrieval and knowledge retrieval. Human evaluation of multi-turn conversations, using the ACUTE-Eval method, shows that its best model outperforms the previous SoTA on engagingness and humanness by using the Blended Skill Talk (BST) dataset [49]. They observed that models may give different results when different decoding algorithms are used, although the models may report the same perplexity in automatic metric. The more recent version of the set of models learns to generate an online search query from the internet, based on the context and conditions on the results to generate a response, thereby employing the latest relevant information [50,51].

The seq2seq (or encoder–decoder) is an important standard architecture for BlenderBot and other conversational AI [51]. The Transformer, by [52], is often used as the underlying architecture for it, although the Long Short-Term Memory Network (LSTM), by [53], may also be used. Generally, the encoder–decoder conditions on the encoding of the prompts and responses up to the last time-step for it to generate the next token as response [2,5]. The sequence of tokens is run through the encoder stack's embedding layer, which then compresses it in the dense feature layer into a fixed-length feature vector. A sequence of tokens is then produced by the decoder after they are passed from the encoder layer. The Softmax function is then used to normalise this, such that the token with the highest probability is the output.

It was observed that the largest BST generation model lags behind the 2.7B model on engagingness, although it had a lower perplexity. Crowd-workers had conversations with the model over several turns, and the quality of the conversations was evaluated. Successful conversation cases spanned several topics such as music, cooking, movies, pets, instruments and malls. Failure cases include (1) too frequent generation of common words, (2) rare generation of infrequent words, (3) self-contradiction, (4) forgetfulness (since its context contains the desired information), and (5) inability to learn or understand through further conversation. An example of the forgetfulness failure case is given in Figure 2.



I have two dogs.

I love dogs! What kind of dogs do you have? I have a chihuahua.

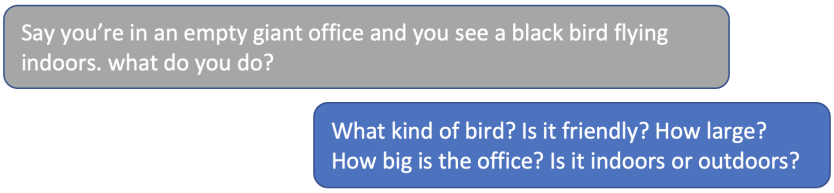
Really?

Yes, she's a small dog. She's a rescue. Do you have any pets?

**Figure 2.** Forgetfulness example of BlenderBot Generative BST 2.7B model (blue bar) [15].

### *5.2. Meena*

Meena is presented by [5]. It is a multi-turn open-domain conversational AI seq2seq model that was trained end-to-end [54]. The underlying architecture of this seq2seq model is the Evolved Transformer (ET). It has 2.6B parameters and includes 1 ET encoder stack and 13 ET decoder stacks. A manual coordinate-descent search was used to determine the hyperparameters of the best Meena model. The data it was trained on are a filtered public domain corpus of social media conversations containing 40B tokens. Perplexity was used to automatically evaluate the model. It was also evaluated in multi-turn conversations using the human evaluation metric: Sensibleness and Specificity Average (SSA). This combines two essential aspects of a human-like chatbot: being specific and making sense. Despite the success of Meena in generating relatively sensible, specific, rich, and interesting responses, it suffers from some of the challenges identified with BlenderBot. A forgetfulness example is provided in Figure 3.



Say you’re in an empty giant office and you see a black bird flying indoors. what do you do?

What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?

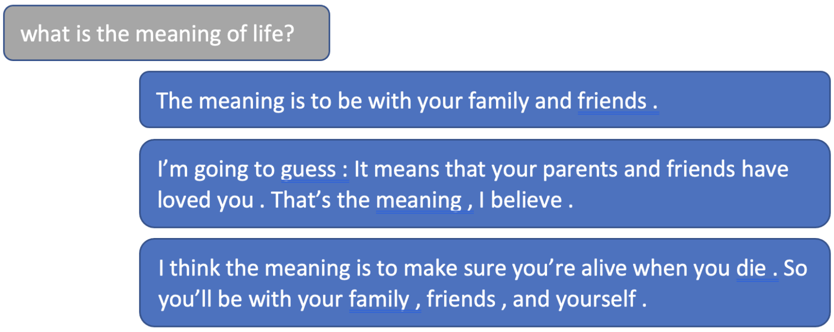
**Figure 3.** Forgetfulness example of Meena (blue bar) [5].

### *5.3. DLGNet*

DLGNet is presented by [55]. Its architecture is similar to GPT-2, being an autoregressive model. It is a multi-turn dialogue response generator that was evaluated, using the automatic metrics BLEU, ROUGE, and distinct n-gram, on the Movie Triples and closed-domain Ubuntu Dialogue datasets. It uses multiple layers of self-attention to map input sequences to output sequences. This it does by shifting the input sequence token one position to the right so that the model uses the previously generated token as additional input for the next token generation. Given a context, it models the joint distribution of the context and response instead of modelling the conditional distribution. Two sizes were trained: a 117M-parameter model and the 345M-parameter model. The 117M-parameter model has 12 attention layers, while the 345M-parameter model has 24 attention layers. The good performance of the model is attributed to the long-range transformer architecture, the injection of random informative paddings, and the use of BPE, which provided 100% coverage for Unicode texts and prevented the OOV problem. The model performs poorly on the closed-domain Ubuntu Dialogue dataset, as it struggles to learn.

### *5.4. DialoGPT 1 and 2 (RetGen)*

DialoGPT was trained on Reddit conversations of 147M exchanges [21]. It is an autoregressive language model (LM) based on GPT-2. Its second version (RetGen) is a hybrid retrieval-augmented/grounded version. In single-turn conversations, it achieved SoTA in human evaluation and performance that is close to human in open-domain dialogues, besides achieving SoTA in automatic evaluation. The large model has 762M parameters with 36 Transformer layers; the medium model has 345M parameters with 24 layers; the small model has 117M parameters with 12 layers. A multi-turn conversation session is framed as a long text in the model and the generation as language modelling. The RetGen version of the model jointly trains a grounded generator and document retriever [56]. Some of the strong characteristics of DialoGPT are the relative ease of adaptability to new dialogue datasets that have few samples and the ability to address commonsense questions to a certain extent. It, however, suffers from similar challenges identified with earlier models. An example of self-contradiction from multi-turn self-chat with user prompt is shown in Figure 4.



what is the meaning of life?

The meaning is to be with your family and friends .

I’m going to guess : It means that your parents and friends have loved you . That’s the meaning , I believe .

I think the meaning is to make sure you’re alive when you die . So you’ll be with your family , friends , and yourself .

**Figure 4.** Self-contradiction from multi-turn self-chat (blue bar) of DialoGPT with user prompt [21].

### *5.5. GPT-3 and GPT-2*

GPT-3 is introduced by [57], being the largest size out of the eight models they created. It is a 175B-parameter autoregressive model that shares many of the qualities of the GPT-2 [58]. These include modified initialisation, reversible tokenisation, and pre-normalisation. However, it uses alternating dense and locally banded sparse attention. The GPT-2 model was used in the 'Third Alexa Prize Socialbot Grand Challenge' on open domain conversational systems [7]. Both GPT-3 and GPT-2 are trained on different versions of the CommonCrawl dataset. GPT-3 achieves strong performance on many NLP datasets, including open-domain QA. In addition, zero-shot perplexity, for automatic metric, was calculated on the Penn Tree Bank (PTB) dataset. Few-shot inference results reveal that it achieves strong performance on many tasks. Zero-shot transfer is based on providing a text description of the task to be completed during evaluation. It is different from one-shot or few-shot transfer, which are based on conditioning on 1 or k number of examples, respectively, for the model in the form of context and completion. No weights are updated in any of the three cases at inference time, and there is a major reduction of task-specific data that may be needed.

### *5.6. T5*

T5 was introduced by [47]. It is not a conversational model but may be adapted for NLG, and by extension, conversation modelling. It is an encoder–decoder Transformer architecture and has a multilingual version, mT5 [59]. It is trained on Colossal Clean Crawled Corpus (C4) and achieved SoTA on the SQuAD QA dataset, where it generates the answer token by token. A simplified version of layer normalisation is used such that no additive bias is used, in contrast to the original Transformer. The self-attention of the decoder is a form of causal or autoregressive self-attention. All the tasks considered for the model are cast into a unified text-to-text format in terms of input and output. This approach, despite its benefits, is known to suffer from prediction issues [60,61]. Maximum likelihood is the training objective for all the tasks, and a task-prefix is specified in the input before the model is fed in order to identify the task at hand. The base version of the model has about 220M parameters.

## **6. Results and Discussion of Survey: Ethics of Gender**

Following the procedure mentioned in Section 4, each conversational AI's gender is determined by the designation given by the developer or cues such as avatar, voice or name, for cases where the developer did not identify the gender. These cues are based on general perception or stereotypes, as indicated by [62]. We consider a conversational AI genderless if it is specifically stated by the reference or developer or nothing is mentioned about it and there are no cues to suggest gender. Overall, in the investigation of the 100 conversational AI, 37 (or 37%) are female, 20 are male, 40 are genderless, and 3 have both gender options. Figure 5 shows a bar graph with details of the results. Breaking down the data into 4 groups—journal-based, Loebner-winners, Facebook Messenger-based, and popular/commercial chatbots—we observe that female conversational AI always outnumber male conversational AI. The genderless category does not follow such a consistent trend in the groups. Out of the 59 chatbots mentioned in journal articles, 34% are female, 22% are male, 42% are genderless, and 2% have both gender options. Meanwhile, 54% are female among the 13 chatbots in the Loebner-winners, 23% are male, 15% are genderless, and 8% have both options. Of the 20 chatbots from Facebook Messenger, 25% are female, 10% are male, 65% are genderless, and 0 offer both genders. Lastly, out of the eight popular/commercial conversational AI, 62.5% are female, 25% are male, 0 is genderless, and 12.5% have both options.

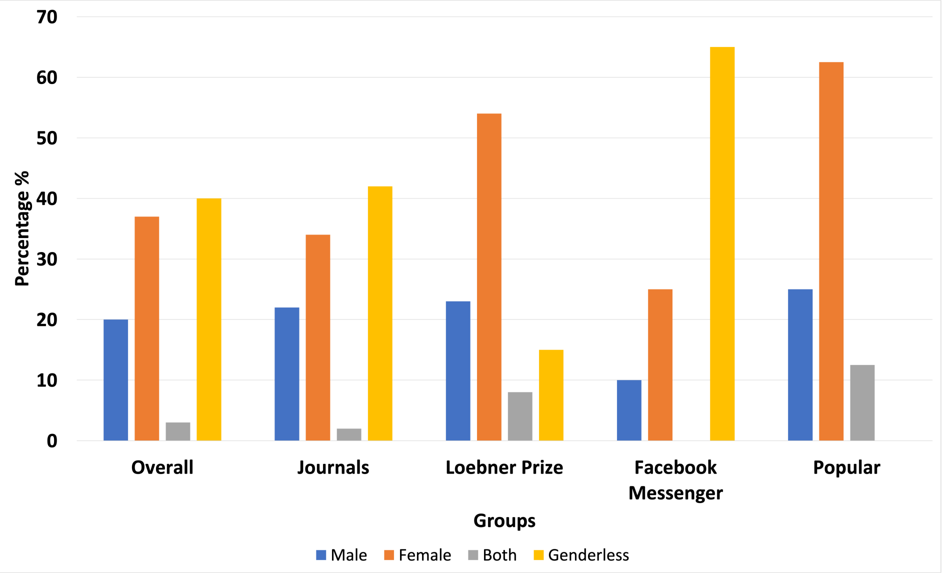


**Figure 5.** Bar chart of the gender of conversational AI for the Overall, Journal, Loebner prize, Facebook Messenger and Popular cases.

*Discussion*

The results agree with the popular assessment that female conversational AI are more predominant than the male ones. We do not know of the gender of the producers of these 100 conversational AI, but it may be a safe assumption that most are male. This assessment has faced criticism from some interest groups, as evidenced in a recent report by [63] that the fact that most conversational AI are female makes them the face of glitches resulting from the limitations of AI systems. In addition, ref. [62] argues that designing conversational AI using young, subservient female characteristics can foster negative gender stereotypes, which may lead to abusive behaviour in reality. Refs. [45,62] confirm that some organisations introduced ethical AI guidelines to address some of these challenges.

Despite the criticisms, there is the opinion that this phenomenon can be viewed from a vantage position for women. For example, they may be viewed as the acceptable face, persona or voice, as the case may be, of the planet. A comparison was made by [64] of a visually androgynous agent with both male and female agents, and it was found that it suffered verbal abuse less than the female agent but more than the male agent. Does this suggest developers do away with female conversational AI altogether to protect the female gender or what is needed is a change in the attitude of users? Especially since previous research has shown that stereotypical agents, with regard to task, are often preferred by users [65]. Some researchers have argued that conversational AI having human-like characteristics, including gender, builds trust for users [66–68]. Furthermore, ref. [69] observed that conversational AI that considers the gender of users, among other cues, is potentially helpful for the self-compassion of users. It is noteworthy that there are those who consider the ungendered, robotic voice of AI uncomfortable and eerie and will, therefore, prefer a specific gender.

## 7. Existing Challenges of Open-Domain Conversational AI

This survey has examined some SoTA open-domain conversational AI models. Despite their noticeable successes and the general progress, challenges still remain. The challenges contribute to the non-human-like utterances the conversational AI tend to have, as shown in some of the examples in a previous section. These challenges also provide motivation for active research in NLP. For example, the basic seq2seq architecture is known for repetitive and dull responses [6,14]. One way of augmenting the architecture for refined responses is the use of IR techniques, such as the concatenation of retrieved sentences from Wikipedia to the conversation context [2]. Other shortcomings may be handled by switching the objective function to a mutual information objective [70] or introducing the beam search decoding algorithm (instead of greedy search) in order to achieve relatively more diverse responses [6]. The maximum mutual information (MMI) objective measures the mutual dependence between inputs and outputs of the ML model [70,71]. There are different decoding algorithms for choosing the next token out of a set of generated possibilities, and beam search has been shown to do better than greedy search [17]. In addition, GPT-3 is observed to lose coherence over really long passages, gives contradictory utterances, and its size is so large that it is difficult to deploy [57]. Collectively, some of the existing challenges are highlighted below. It is hoped that identifying these challenges will spur further research in these areas.

1. Poor coherence in sequence of text or across multiple turns of generated conversation [2,72].
2. Lack of utterance diversity [17].
3. Bland repetitive utterances [17,21].
4. Lack of empathetic responses from conversational systems [12].
5. Lack of memory to personalise user experiences.
6. Style inconsistency or lack of persona [5,21].
7. Multiple initiative coordination [2].
8. Poor inference and implicature during conversation [15].
9. Lack of world knowledge.
10. Poor adaptation or responses to idioms or figurative language [23,73].
11. Hallucination of facts when generating responses [74].
12. Obsolete facts, which are frozen in the models' weights during training.
13. Training requires a large amount of data [74].
14. Lack of common-sense reasoning [74].
15. Large models use so many parameters that make them complex and may impede transparency [74].
16. Lack of training data for low-resource languages [14,75]

## 8. Open-Domain Conversational AI for Low-Resource Languages

The last challenge mentioned in the earlier section is a prevailing issue for many languages around the world. Low-resource languages are natural languages with little or no digital data or resources [14,76]. This challenge has meant that so many languages are unrepresented in many deep ML models, as they usually require a lot of data for pretraining. It is noteworthy that multilingual versions of some of the models are being made with very limited data of the low-resource languages. They are, however, known to have relatively poor performance compared to models trained completely on the target language data [77–79], and only few languages are covered [14]. Approaches to mitigating this particular challenge involve human and automatic MT attempts [76], and efforts at exploiting cross-lingual transfer to build conversational AI capable of machine–machine conversations for automated data generation [14]. There is also the ongoing effort with the world wide voice web (WWvW) by the Stanford group for adding voice data to the web that can be accessed by virtual assistants (hai.stanford.edu/news/will-future-internet-be-voice-proposing-world-wide-voice-web, accessed on 6 June 2022).

Data acquisition through machine–machine conversations will, potentially, be a game-changer, as this will make it possible to have large high-quality data in a relatively short period of time with very little effort. This approach requires that quality control (QC) should be in place before the models are deployed and during the machine–machine data acquisition process. In a different but related attempt with task-based conversational systems in cross-lingual dialogue state tracking involving the MultiWOZ dataset, ref. [8] utilised Google Translate for the ontology translation and human translation for corrections, as QC.

## 9. Related Work

In a recent survey, ref. [80] reviewed advances in chatbots by using the common approach of acquiring scientific papers from search databases, based on certain search terms, and selecting a small subset from the lot for analysis, based on publications between 2007 and 2021. The databases they used are IEEE, ScienceDirect, Springer, Google Scholar, JSTOR, and arXiv. They analysed rule-based and data-driven chatbots from the filtered collection of papers. Their distinction of rule-based chatbots as being different from AI chatbots may be disagreed with, especially when a more general definition of AI is given and since modern systems such as Alexa have rule-based components [2]. Meanwhile, ref. [81] reviewed learning towards conversational AI and in their survey classified conversational AI into three frameworks. They posit that a human-like conversation system should be both (1) informative and (2) controllable.

A systematic survey of recent advances in deep learning-based dialogue systems was conducted by [82], where the authors recognise that dialogue modelling is a complicated task because it involves many related NLP tasks, which are also required to be solved. They categorised dialogue systems by analysing them from two angles: model type and system type (including task-oriented and open-domain conversational systems). Ref. [83] also recognised that building open-domain conversational AI is a challenging task. They describe how, through the Alexa Prize, teams advanced the SoTA through context in dialogue models, using knowledge graphs for language understanding, and building statistical and hierarchical dialogue managers, among other things.

## 10. Limitation

Although this work has presented recent SoTA open-domain conversational AI within the first ten pages of the search databases (Google Scholar and Google Search) that were used, we recognise that the time-constraint and restricted number of pages of results means there may have been some that were missed. This goes also for the second investigation on the gender of conversational AI. Furthermore, our approach did not survey all possible methods for conversational AI, although it identified all the major methods available.

## 11. Conclusions

In this survey of the SoTA open-domain conversational AI, we identified models that have pushed the envelope in recent times. It involves two different investigations: text-based open-domain conversational AI and the ethics of gender of conversational AI by considering 100 chatbots. It appears that hybrid models of conversational AI offer more advantages than any single architecture, based on the benefits of up-to-date responses and world knowledge. In addition to discussing some of the strengths of the identified models, we focused on prevailing challenges (providing examples) that still exist, which need to be surmounted to achieve the type of desirable performance, which is typical of human–human conversations. The important challenge with conversational AI for low-resource languages is highlighted, as well as the ongoing attempts at tackling it. The presentation of the discussion on the ethics of the gender of conversational AI provides, possibly, a new perspective to the debate. We believe this survey will spur focused research in addressing some of the challenges identified, thereby enhancing the SoTA in open-domain conversational AI.

**Author Contributions:** Conceptualisation, T.A. and F.L.; methodology, T.A.; investigation, T.A.; writing—original draft preparation, T.A.; writing—review and editing, F.L. and M.L.; supervision, F.L. and M.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**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. States, U. *Preparing for the Future of Artificial Intelligence*; 2016. Available online: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (accessed on 25 May 2022).
2. Jurafsky, D.; Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Dorling Kindersley Pvt, Ltd.: London, UK, 2020.
3. Weizenbaum, J. A Computer Program for the Study of Natural Language. Fonte: Stanford. 1969. Available online: Http://web.stanford.edu/class/linguist238/p36 (accessed on 25 May 2022).
4. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
5. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a human-like open-domain chatbot. *arXiv* **2020**, arXiv:2001.09977.
6. Chowdhary, K. Natural Language Processing for Word Sense Disambiguation and Information Extraction. 2020, pp. 603–649. Available online: https://arxiv.org/ftp/arxiv/papers/2004/2004.02256.pdf (accessed on 25 May 2022).
7. Gabriel, R.; Liu, Y.; Gottardi, A.; Eric, M.; Khatri, A.; Chadha, A.; Chen, Q.; Hedayatnia, B.; Rajan, P.; Binici, A.; et al. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proc.* **2020**, *3*. Available online: https://assets.amazon.science/0e/e6/2cff166647bfb951b3ccc67c1d06/further-advances-in-open-domain-dialog-systems-in-the-third-alexa-prize-socialbot-grand-challenge.pdf (accessed on 25 May 2022).
8. Gunasekara, C.; Kim, S.; D'Haro, L.F.; Rastogi, A.; Chen, Y.N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.W.; et al. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv* **2020**, arXiv:2011.06486.
9. Schegloff, E.A. Sequencing in conversational openings 1. *Am. Anthropol.* **1968**, *70*, 1075–1095. [CrossRef]
10. Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; Hakkani-Tur, D. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2020; pp. 422–428.
11. Allwood, J.; Grönqvist, L.; Ahlsén, E.; Gunnarsson, M. Annotations and tools for an activity based spoken language corpus. In *Current and New Directions in Discourse and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 1–18. [CrossRef]
12. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5370–5381. [CrossRef]
13. Adewumi, T.; Brännvall, R.; Abid, N.; Pahlavan, M.; Sabry, S.S.; Liwicki, F.; Liwicki, M. Småprat: DialoGPT for Natural Language Generation of Swedish Dialogue by Transfer Learning. In Proceedings of the 5th Northern Lights Deep Learning Workshop, Tromsø, Norway, 10–12 January 2022; Volume 3. [CrossRef]
14. Adewumi, T.; Adeyemi, M.; Anuoluwapo, A.; Peters, B.; Buzaaba, H.; Samuel, O.; Rufai, A.M.; Ajibade, B.; Gwadabe, T.; Traore, M.M.K.; et al. Ìtàkúròso: Exploiting Cross-Lingual Transferability for Natural Language Generation of Dialogues in Low-Resource, African Languages. *arXiv* **2022**, arXiv:2204.08083.
15. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. Recipes for Building an Open-Domain Chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 300–325. [CrossRef]
16. Chen, H.; Liu, X.; Yin, D.; Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 25–35. [CrossRef]
17. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. In Proceedings of the International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. Available online: https://arxiv.org/pdf/1904.09751.pdf (accessed on 25 May 2022).
18. Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In *Mining Text Data*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 163–222.
19. Gehrmann, S.; Adewumi, T.; Aggarwal, K.; Ammanamanchi, P.S.; Aremu, A.; Bosselut, A.; Chandu, K.R.; Clinciu, M.A.; Das, D.; Dhole, K.; et al. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), Online, 5–6 August 2021; pp. 96–120. [CrossRef]
20. Reiter, E. 20 Natural Language Generation. In *The Handbook of Computational Linguistics and Natural Language Processing*; 2010; p. 574. Available online: https://onlinelibrary.wiley.com/doi/10.1002/9781444324044.ch20 (accessed on 25 May 2022).
21. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 270–278. [CrossRef]
22. Gangal, V.; Jhamtani, H.; Hovy, E.; Berg-Kirkpatrick, T. Improving Automated Evaluation of Open Domain Dialog via Diverse Reference Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 4079–4090. [CrossRef]

23. Jhamtani, H.; Gangal, V.; Hovy, E.; Berg-Kirkpatrick, T. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7476–7485. [CrossRef]
24. Liu, C.W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv* **2016**, arXiv:1603.08023.
25. Ji, T.; Graham, Y.; Jones, G.J.; Lyu, C.; Liu, Q. Achieving Reliable Human Assessment of Open-Domain Dialogue Systems. *arXiv* **2022**, arXiv:2203.05899.
26. Tsuta, Y.; Yoshinaga, N.; Toyoda, M. uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 5–10 July 2020; pp. 199–206. [CrossRef]
27. Venkatesh, A.; Khatri, C.; Ram, A.; Guo, F.; Gabriel, R.; Nagar, A.; Prasad, R.; Cheng, M.; Hedayatnia, B.; Metallinou, A.; et al. On evaluating and comparing conversational agents. *arXiv* **2018**, arXiv:1801.03625.
28. Guo, F.; Metallinou, A.; Khatri, C.; Raju, A.; Venkatesh, A.; Ram, A. Topic-based evaluation for conversational bots. *arXiv* **2018**, arXiv:1801.03622.
29. Deriu, J.; Tuggener, D.; von Däniken, P.; Campos, J.A.; Rodrigo, A.; Belkacem, T.; Soroa, A.; Agirre, E.; Cieliebak, M. Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3971–3984. [CrossRef]
30. Li, M.; Weston, J.; Roller, S. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv* **2019**, arXiv:1909.03087.
31. Mauldin, M.L. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In Proceedings of the AAAI, Seattle, WA, USA, 31 July–4 August 1994; Volume 94, pp. 16–21. Available online: https://www.aaai.org/Papers/AAAI/1994/AAAI94-003.pdf (accessed on 25 May 2022).
32. Bradeško, L.; Mladenić, D. A survey of chatbot systems through a loebner prize competition. In Proceedings of the Slovenian Language Technologies Society Eighth Conference of Language Technologies, Ljubljana, Slovenia, 8–9 October 2012; pp. 34–37. Available online: http://nl.ijs.si/isjt12/proceedings/isjt2012_06.pdf (accessed on 25 May 2022).
33. Shieber, S.M. Lessons from a restricted Turing test. *arXiv* **1994**, arXiv:cmp-lg/9404002.
34. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*; Elsevier: Amsterdam, The Netherlands, 1978; pp. 7–55.
35. Adewumi, T.P.; Liwicki, F.; Liwicki, M. Conversational Systems in Machine Learning from the Point of View of the Philosophy of Science—Using Alime Chat and Related Studies. *Philosophies* **2019**, *4*, 41. [CrossRef]
36. Javed, S.; Adewumi, T.P.; Liwicki, F.S.; Liwicki, M. Understanding the Role of Objectivity in Machine Learning and Research Evaluation. *Philosophies* **2021**, *6*, 22. [CrossRef]
37. White, M.D. Immanuel kant. In *Handbook of Economics and Ethics*; Edward Elgar Publishing: Cheltenham, UK, 2009.
38. Alexander, L.; Moore, M. *Deontological Ethics*; 2007. Available online: https://plato.stanford.edu/entries/ethics-deontological/ (accessed on 25 May 2022).
39. Paquette, M.; Sommerfeldt, E.J.; Kent, M.L. Do the ends justify the means? Dialogue, development communication, and deontological ethics. *Public Relat. Rev.* **2015**, *41*, 30–39. [CrossRef]
40. Voigt, P.; Von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10, pp. 10–5555.
41. Neff, G.; Nagy, P. Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. *Int. J. Commun.* **2016**, *10*, 17.
42. Dinan, E.; Fan, A.; Williams, A.; Urbanek, J.; Kiela, D.; Weston, J. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 8173–8188. [CrossRef]
43. Henderson, P.; Sinha, K.; Angelard-Gontier, N.; Ke, N.R.; Fried, G.; Lowe, R.; Pineau, J. Ethical challenges in data-driven dialogue systems. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 123–129. Available online: https://arxiv.org/pdf/1711.09050.pdf (accessed on 25 May 2022).
44. Maedche, A. Gender Bias in Chatbot Design. In *Chatbot Research and Design*; Springer: Heidelberg, Germany, 2020; p. 79.
45. Ruane, E.; Birhane, A.; Ventresque, A. Conversational AI: Social and Ethical Considerations. In Proceedings of the AICS—27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, 5–6 December 2019; pp. 104–115.
46. Kerry, A.; Ellis, R.; Bull, S. Conversational agents in E-Learning. In Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, 9–11 December 2008; pp. 169–182. Available online: https://link.springer.com/chapter/10.1007/978-1-84882-215-3_13 (accessed on 25 May 2022).
47. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
48. Zou, Y.; Liu, Z.; Hu, X.; Zhang, Q. Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-Domain Dialogue Systems. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2215–2226. [CrossRef]

49. Smith, E.M.; Williamson, M.; Shuster, K.; Weston, J.; Boureau, Y.L. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2021–2030. [CrossRef]
50. Komeili, M.; Shuster, K.; Weston, J. Internet-augmented dialogue generation. *arXiv* **2021**, arXiv:2107.07566.
51. Xu, J.; Szlam, A.; Weston, J. Beyond goldfish memory: Long-term open-domain conversation. *arXiv* **2021**, arXiv:2107.07567.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30. Available online: https://arxiv.org/pdf/1706.03762v5.pdf (accessed on 25 May 2022).
53. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
54. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015. [CrossRef]
55. Olabiyi, O.; Mueller, E.T. Multiturn dialogue response generation with autoregressive transformer models. *arXiv* **2019**, arXiv:1908.01841.
56. Zhang, Y.; Sun, S.; Gao, X.; Fang, Y.; Brockett, C.; Galley, M.; Gao, J.; Dolan, B. Joint Retrieval and Generation Training for Grounded Text Generation. *arXiv* **2021**, arXiv:2105.06597.
57. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
58. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
59. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 483–498. [CrossRef]
60. Adewumi, T.; Alkhaled, L.; Alkhaled, H.; Liwicki, F.; Liwicki, M. ML_LTU at SemEval-2022 Task 4: T5 Towards Identifying Patronizing and Condescending Language. *arXiv* **2022**, arXiv:2204.07432.
61. Sabry, S.S.; Adewumi, T.; Abid, N.; Kovacs, G.; Liwicki, F.; Liwicki, M. HaT5: Hate Language Identification using Text-to-Text Transfer Transformer. *arXiv* **2022**, arXiv:2202.05690.
62. Abercrombie, G.; Cercas Curry, A.; Pandya, M.; Rieser, V. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, Online, 5 August 2021; pp. 24–33. [CrossRef]
63. West, M.; Kraut, R.; Ei Chew, H. *I'd Blush If I Could: Closing Gender Divides in Digital Skills through Education*; 2019. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000367416 (accessed on 25 May 2022).
64. Silvervarg, A.; Raukola, K.; Haake, M.; Gulz, A. The effect of visual gender on abuse in conversation with ECAs. In Proceedings of the International Conference on Intelligent Virtual Agents, Santa Cruz, CA, USA, 12–14 September 2012; pp. 153–160. Available online: https://link.springer.com/chapter/10.1007/978-3-642-33197-8_16 (accessed on 25 May 2022).
65. Forlizzi, J.; Zimmerman, J.; Mancuso, V.; Kwak, S. How interface agents affect interaction between humans and computers. In Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces, Helsinki, Finland, 22–25 August 2007; pp. 209–221. Available online: https://dl.acm.org/doi/pdf/10.1145/1314161.1314180 (accessed on 25 May 2022).
66. Louwerse, M.M.; Graesser, A.C.; Lu, S.; Mitchell, H.H. Social cues in animated conversational agents. *Appl. Cogn. Psychol.* **2005**, *19*, 693–704. [CrossRef]
67. Muir, B.M. Trust between humans and machines, and the design of decision aids. *Int. J. Man-Mach. Stud.* **1987**, *27*, 527–539. [CrossRef]
68. Nass, C.I.; Brave, S. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*; MIT Press: Cambridge, MA, USA, 2005.
69. Lee, M.; Ackermans, S.; Van As, N.; Chang, H.; Lucas, E.; IJsselsteijn, W. Caring for Vincent: A chatbot for self-compassion. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–13. Available online: https://dl.acm.org/doi/pdf/10.1145/3290605.3300932 (accessed on 25 May 2022).
70. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 110–119. [CrossRef]
71. Bahl, L.; Brown, P.; De Souza, P.; Mercer, R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In Proceedings of the ICASSP'86, IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo, Japan, 7–11 April 1986; Volume 11, pp. 49–52. Available online: http://mirlab.org/users/davidson.chen/relatedPapers/others/1986%20ICASSP%20Maximum%20Mutual%20Information%20Estimation%20of%20Hidden%20Markov%20Model%20Parameters%20for%20Speech%20Recognition.pdf (accessed on 25 May 2022).
72. Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; Weston, J. Neural text generation with unlikelihood training. *arXiv* **2019**, arXiv:1908.04319.
73. Adewumi, T.; Liwicki, F.; Liwicki, M. Vector Representations of Idioms in Conversational Systems. *arXiv* **2022**, arXiv:2205.03666.
74. Marcus, G. Deep learning: A critical appraisal. *arXiv* **2018**, arXiv:1801.00631.
75. Adewumi, T.P.; Liwicki, F.; Liwicki, M. The Challenge of Diacritics in Yoruba Embeddings. *arXiv* **2020**, arXiv:2011.07605.

76. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Fagbohungbe, T.; Akinola, S.O.; Muhammad, S.; Kabongo Kabenamualu, S.; Osei, S.; Sackey, F.; et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 2144–2160. [CrossRef]
77. Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; Gurevych, I. AdapterHub: A Framework for Adapting Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations, Online, 16–20 November 2020; pp. 46–54. Available online: https://arxiv.org/pdf/2007.07779.pdf (accessed on 25 May 2022).
78. Virtanen, A.; Kanerva, J.; Ilo, R.; Luoma, J.; Luotolahti, J.; Salakoski, T.; Ginter, F.; Pyysalo, S. Multilingual is not enough: BERT for Finnish. *arXiv* **2019**, arXiv:1912.07076.
79. Rönnqvist, S.; Kanerva, J.; Salakoski, T.; Ginter, F. Is multilingual BERT fluent in language generation? *arXiv* **2019**, arXiv:1910.03806.
80. Caldarini, G.; Jaf, S.; McGarry, K. A Literature Survey of Recent Advances in Chatbots. *Information* **2022**, *13*, 41. [CrossRef]
81. Fu, T.; Gao, S.; Zhao, X.; Wen, J.r.; Yan, R. Learning towards conversational AI: A survey. *AI Open* **2022**, *3*, 14–28. [CrossRef]
82. Ni, J.; Young, T.; Pandelea, V.; Xue, F.; Adiga, V.; Cambria, E. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv* **2021**, arXiv:2105.04387.
83. Khatri, C.; Hedayatnia, B.; Venkatesh, A.; Nunn, J.; Pan, Y.; Liu, Q.; Song, H.; Gottardi, A.; Kwatra, S.; Pancholi, S.; et al. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv* **2018**, arXiv:1812.10757.