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, Sweden
firstname.lastname@ltu.se

Abstract

We survey SoTA open-domain conversational AI models with the purpose 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 are known to have several challenges, including bland 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 take-away 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 unusual discussion about 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

1. Introduction

There are different opinions as to the definition of AI but according to States (2016), 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 (Jurafsky and Martin, 2020). 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 Weizenbaum (1969), is the first acclaimed conversational AI (or system). Its conversations with humans demonstrated how therapeutic its responses could be. Staff of Weizenbaum (1969) reportedly became engrossed with the program during interactions and possibly had private conversations with it (Jurafsky and Martin, 2020).

Modern SoTA open-domain conversational AI aim 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 Turing (1950). 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 topic in that it presents discussion around the ethics of gender of conversational AI with compelling statistics and discusses the uncommon topic of conversational AI for low-resource languages. 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 uncommon discussion about 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 organized 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; the 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 & 6) then follow with details of the outcome of the methods; thereafter, the Existing Challenges Section (7) shares the prevailing challenges to obtaining ”human” 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; the Conclusion Section (11) summarizes the study after the limitations 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 utilize 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 (6) hybrid method (Jurafsky and Martin, 2020; Adiwardana et al., 2020; Chowdhary, 2020). Certain modern systems are still designed in the rule-based style that was used for ELIZA (Jurafsky and Martin, 2020). 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 (Jurafsky and Martin, 2020). Such systems learn inductively from large datasets involving many turns in conversations. A turn (or utterance) in a conversation is each single contribution from a speaker (Schegloff, 1968; Jurafsky and Martin, 2020). The data may be from written conversations, such as the MultiWOZ (Eric et al., 2020), transcripts of human-human spoken conversations, such as the Gothenburg Dialogue Corpus (GDC) (Allwood et al., 2003), crowdsourced conversations, such as the EmpatheticDialogues (Rashkin et al., 2019), and social media conversations like Familjeliv[1] or Reddit[2] (Adewumi et al., 2022c; Adewumi et al., 2022a). 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 finetuned on specific conversational data.

2.1. Retrieval & Generation approaches

Two common ways that data-driven conversational AI produce turns as response are Information Retrieval (IR) and generation (Jurafsky and Martin, 2020). 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 gets selected from a potential set. This can be achieved with different IR methods and choosing the response with the highest cosine similarity with c (Jurafsky and Martin, 2020). 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.

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

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

2.2. Evaluation

Although there are a number of metrics for NLP systems (Aggarwal and Zhai, 2012; Gehrmann et al., 2021; Reiter, 2010) 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, though it is subjective (Zhang et al., 2020). It is both time-intensive and laborious. As a result of this, automatic metrics serve as proxies for estimating performance though they may not correlate very well with human evaluation (Gehrmann et al., 2021; Gangal et al., 2021; Jhamtani et al., 2021). For example, IR systems may use F1, precision, and recall (Aggarwal and Zhai, 2012). Furthermore, metrics used in NLG tasks, like Machine Translation (MT), such as the BLEU or ROUGE, are sometimes used to evaluate conversational systems (Zhang et al., 2020) but they are also discouraged because they do not correlate well with human judgment (Liu et al., 2016; Jurafsky and Martin, 2020). They do not take syntactic or lexical variation into consideration (Reiter, 2010). Dependency-based evaluation metrics, however, allow for such variation in evaluation. Perplexity is commonly used for evaluation and has been shown to correlate with a human evaluation metric called Sensibleness and Specificity Average (SSA) (Adiwardana et al., 2020). 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 (Adiwardana et al., 2020). It corresponds to the effective size of the vocabulary (Aggarwal and Zhai, 2012) 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 (Holtzman et al., 2020).

Two methods for human evaluation of open-domain conversational AI are observer and participant evaluation (Jurafsky and Martin, 2020). Observer evaluation involves reading and scoring a transcript of human-

[1] familjeliv.se
[2] reddit.com

chatbot conversation while participant evaluation interacts directly with the conversational AI (Jurafsky and Martin, 2020). In the process, the system may be evaluated for different qualities, such as humanness (or human-likeness), fluency, making sense, engagingness, interestingness, avoiding repetition, and more. 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. Human evaluation is usually modeled to resemble the Turing test (or the imitation game).

2.3. The Turing test

Modern human evaluation is generally designed like the Turing test. The Turing test is the indistinguishability test. This 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 (Turing, 1950). The proposed imitation game, by Turing (1950), 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. The man tries to trick the interrogator into believing he's a woman while the woman tries to convince the interrogator she's a woman. 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 (Turing, 1950). The formulation of the imitation game, by Turing (1950), does not precisely match modern versions of the test (Saygin and Cicekli, 2002).

An early version of the test was applied to PARRY, a chatbot designed by Colby et al. (1972) to imitate aggressive emotions. Most psychiatrists couldn't distinguish between transcripts of real paranoids and PARRY (Colby et al., 1971; Jurafsky and Martin, 2020). This example of PARRY may be viewed as an edge case, given that the comparison was not made with rational human beings but paranoids (Mauldin, 1994). A limited version of the test was introduced in 1991, alongside its unrestricted version, in what is called the Loebner Prize competition (Mauldin, 1994). Every year, since then, prizes have been awarded to conversational AI that pass the restricted version in the competitions (Bradeško and Mladenić, 2012). This competition has its share of criticisms, including the view that it is rewarding tricks instead of furthering the course of AI (Shieber, 1994; Mauldin, 1994). As a result of this, Shieber (1994) recommended an alternative approach, whereby the competition will involve a different award methodology that is based on a different set of assessment and done on an occasional basis.

2.4. 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 (Jurafsky and Martin, 2020). 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 (Jurafsky and Martin, 2020). Besides conversation initiative, below are additional characteristics of human conversations, according to Sacks et al. (1978).

- 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.

2.5. Ethics

Ethical issues are important in open-domain conversational AI. And the perspective of deontological ethics views objectivity as being equally important (Adewumi et al., 2019; Javed et al., 2021; White, 2009). Deontological ethics is a philosophy that emphasizes duty or responsibility over the outcome achieved in decision-making (Alexander and Moore, 2007; Paquette et al., 2015). 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 regards to their personal data (Voigt and Von dem Bussche, 2017). 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) (Jurafsky and Martin, 2020; Zhang et al., 2020).

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 (Neff and Nagy, 2016; Jurafsky and Martin, 2020). Privacy is another crucial ethical issue. Data containing PII may fall into the wrong hands and cause 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 (Dinan et al., 2020). Privacy

concerns are also being addressed through anonymisation techniques (Henderson et al., 2018; Jurafsky and Martin, 2020). Balancing the features of chatbots with ethical considerations can be a 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 the those designed as "male". In a survey of 1,375 chatbots, from automatically crawling chatbots.org, Maedche (2020) 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 since the early days of ELIZA. These benefits have led to investments in conversational AI by many organizations, including Apple (Jurafsky and Martin, 2020). Some of the benefits include:

- Provision of therapeutic company, as was experienced with ELIZA.
- The provision of human psychiatric/psychological treatment on the basis of favorable behavior determined from experiments which are designed to modify input-output behaviour in models. This may be designed like PARRY (Colby et al., 1971).
- Provide support for users with disabilities, such as blindness (Reiter, 2010).
- A channel for providing domain/world knowledge (Reiter, 2010).
- The provision of educational content or information in a concise fashion (Kerry et al., 2008).
- Automated machine-machine generation of quality data for low-resource languages (Adewumi et al., 2022a).

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 such systems. The first involves online search on Google Scholar and regular Google Search, using the term "state-of-the-art open-domain dialogue systems". This returned 5,130 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-constraint) to identify those that report SoTA results in the last five years (2017-2022) in order to give more attention to them. It is important to note that some Google Scholar results point to other databases, like ScienceDirect, arXiv, and IEEEXplore. The reason for also searching on 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 pretrained on large text and may be adapted for conversational systems, such as the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), and autoregressive models because they easily follow the NLG framework. 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[3]. 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.

5. Results of Survey: Models

Review of the different scientific papers from the earlier method show 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 & 2, Meena, DLGNet, Dialogue Generative Pre-trained Transformer (DialoGPT), Generative Pre-trained Transformer (GPT)-3 and 2 (RetGen), and Text-to-Text Transfer Transformer (T5).

[3] enterprisebotmanager.com/chatbot-examples
growthrocks.com/blog/7-messenger-chatbots

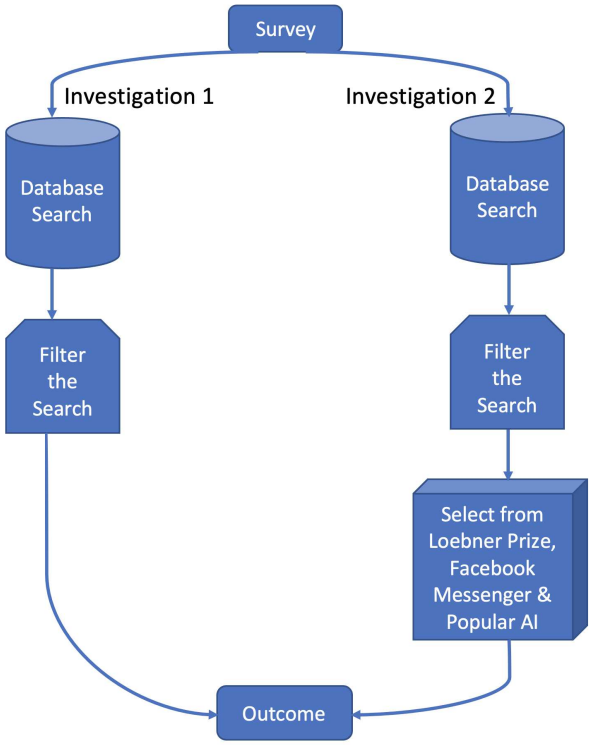

Figure 1: Method for both investigations in this study

5.1. BlenderBot 1 & 2

Some of the ingredients for the success of BlenderBot, as identified by Roller et al. (2020), 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 tokenization. Human evaluation of multi-turn conversations, using ACUTE-Eval method, shows that its best model outperforms the previous SoTA on engagingness and humanness by using the Blended Skill Talk (BST) dataset (Smith et al., 2020). They observed that models may give different results when different decoding algorithms are used, though 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 (Komeili et al., 2021; Xu et al., 2021).

The seq2seq (or encoder-decoder) is an important standard architecture for BlenderBot and other conversational AI (Xu et al., 2021). The Transformer, by (Vaswani et al., 2017), is often used as the underlying architecture for it, though the Long Short Term Memory Network (LSTM), by Hochreiter and Schmidhuber (1997), 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 (Adiwardana et al., 2020; Jurafsky and Martin, 2020). The sequence of tokens is run through the encoder stack's embedding layer, which then compresses it in the dense feature layer into 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 normalized this, such that the token with the highest probability is the output.

5.2. Meena

Meena is presented by Adiwardana et al. (2020). It is a multi-turn open-domain conversational AI seq2seq model that was trained end-to-end (Bahdanau et al., 2015). 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. Manual coordinate-descent search was used to determine the hyperparameters of the best Meena model. The data it was trained on is 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.

5.3. DLGNet

DLGNet is presented by Olabiyi and Mueller (2019). 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 modeling 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.

5.4. DialoGPT 1 & 2 (RetGen)

DialoGPT was trained on Reddit conversations of 147M exchanges (Zhang et al., 2020). It is an autoregressive LM based on GPT-2. Its second version (Ret-

Gen) 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 multiturn conversation session is framed as a long text in the model and the generation as language modeling. The model is easily adaptable to new dialogue datasets with few samples. The RetGen version of the model jointly trains a grounded generator and document retriever (Zhang et al., 2021).

5.5. GPT-3 & GPT-2

GPT-3 is introduced by Brown et al. (2020), 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 (Radford et al., 2019). These include modified initialization, reversible tokenization, and pre-normalization. However, it uses alternating dense and locally banded sparse attention. Both GPT-3 and GPT-2 are trained on the CommonCrawl dataset, though different versions of it. 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 text description of the task to be done during evaluation. It is different from one-shot or few-shot transfer, which are based on conditioning on 1 or k number of examples 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's a major reduction of task-specific data that may be needed.

5.6. T5

T5 was introduced by Raffel et al. (2020). It is an encoder-decoder Transformer architecture and has a multilingual version, mT5 (Xue et al., 2021). 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 normalization 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 (Adewumi et al., 2022b; Sabry et al., 2022). 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 & 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. 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 2 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. 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 8 popular/commercial conversational AI, 62.5% are female, 25% are male, 0 is genderless, and 12.5% have both options.

6.1. 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, evidenced in a recent report by West et al. (2019) that the fact that most conversational AI are female makes them the face of glitches resulting from the limitations of AI systems. Despite the criticism, there's 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 Silvervarg et al. (2012) 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 regards to task, are often preferred by users (Forlizzi et al., 2007). Some researchers have argued that conversational AI having human-like characteristics, includ-

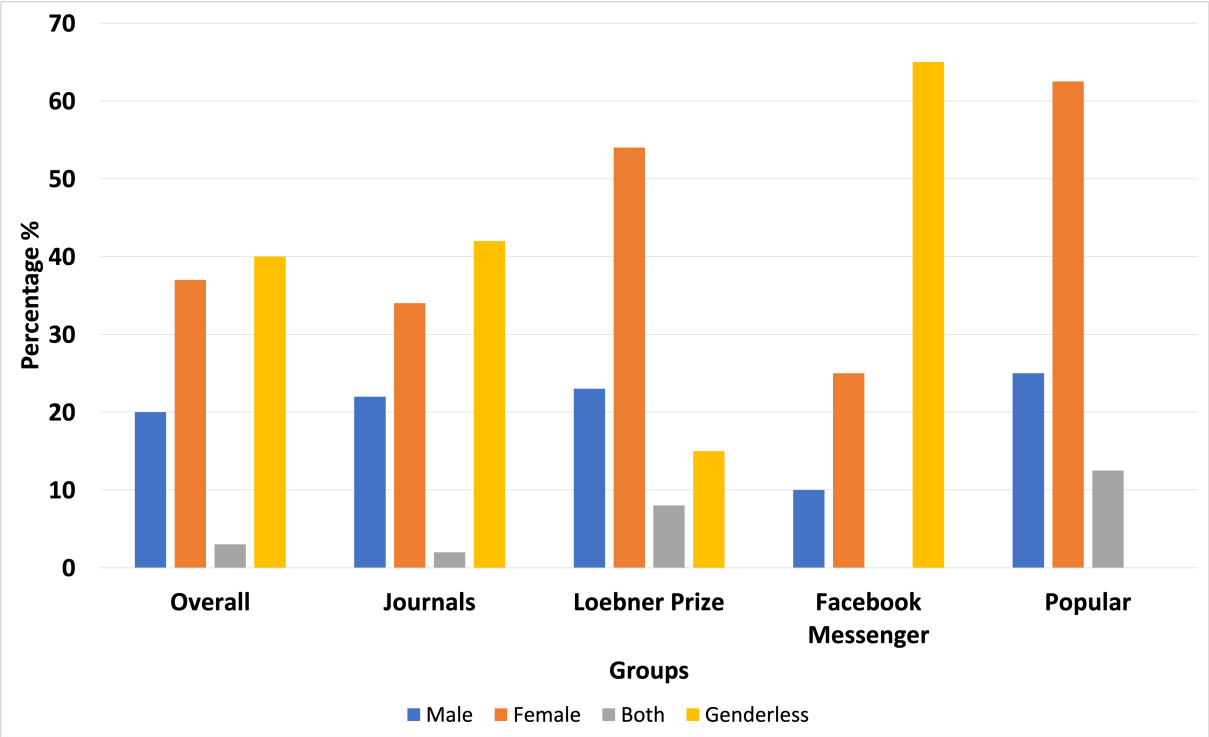

Figure 2: Method for both investigations in this study

ing gender, builds trust for users (Louwerse et al., 2005; Muir, 1987; Nass and Brave, 2005). Furthermore, Lee et al. (2019) observed that conversational AI that consider gender of users, among other cues, are potentially helpful for self-compassion of users. 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 several 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. These challenges also provide motivation for active research in NLP. For example, the basic seq2seq architecture is known for repetitive and dull responses (Chowdhary, 2020). One way of augmenting the architecture for refined responses is the use of IR techniques, like concatenation of retrieved sentences from Wikipedia to the conversation context (Jurafsky and Martin, 2020). Other shortcomings may be handled by switching the objective function to a mutual information objective or introducing the beam search decoding algorithm in order to achieve relatively more diverse responses (Chowdhary, 2020). Besides, GPT-3 is observed to lose coherence over really long passages, gives contradictory utterances, and its size is so large that it's difficult to deploy. 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 (Jurafsky and Martin, 2020; Welleck et al., 2019).
2. Lack of utterance diversity (Holtzman et al., 2020).
3. Bland repetitive utterances (Holtzman et al., 2020; Zhang et al., 2020).
4. Lack of empathetic responses from conversational systems (Rashkin et al., 2019).
5. Lack of memory to personalise user experiences.
6. Style inconsistency or lack of persona (Adiwardana et al., 2020; Zhang et al., 2020).
7. Multiple initiative coordination (Jurafsky and Martin, 2020).
8. Poor inference and implicature during conversation.
9. Lack of world-knowledge.
10. Poor adaptation or responses to idioms or figurative language (Jhamtani et al., 2021)
11. Hallucination of facts when generating responses (Marcus, 2018).

12. Obsolete facts, which are frozen in the models' weights during at training .

13. Training requires a large amount of data (Marcus, 2018).

14. Lack of common-sense reasoning (Marcus, 2018).

15. Large models use so many parameters that make them complex and may impede transparency (Marcus, 2018).

16. Lack of training data for low-resource languages (Adewumi et al., 2022a; Adewumi et al., 2020)

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 (Nekoto et al., 2020; Adewumi et al., 2022a). 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. Even Noteworthy, though, 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 (Pfeiffer et al., 2020; Virtanen et al., 2019; Rönnqvist et al., 2019) and only few languages are covered (Adewumi et al., 2022a). Approaches to mitigating this particular challenge involve human and automatic MT attempts (Nekoto et al., 2020) and efforts at exploiting cross-lingual transfer to build conversational AI capable of machine-machine conversations for automated data generation (Adewumi et al., 2022a).

9. Related Work

In a recent survey, Caldarini et al. (2022) 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 analyzed 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 like Alexa have rule-based components (Jurafsky and Martin, 2020). Meanwhile, Fu et al. (2022) 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 Ni et al. (2021), 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 dailogue systems by analysing them from two angles: model type and system type (including task-oriented and open-domain conversational systems). Khatri et al. (2018) 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 dialog models, using knowledge graphs for language understanding, and building statistical and hierarchical dialog 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 & 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, though it identified all the major methods available.

11. Conclusion

In this survey of the SoTA open-domain conversational AI, we identified models that have pushed the envelope in recent times. It appears that hybrid models of conversational AI offer more advantages than any single architecture, based on the benefit of up-to-date responses and world knowledge. Besides discussing some of their successes or strengths, we focused on prevailing challenges that still exist and which need to be surmounted to achieve the type of desirable performance, typical of human-human conversations. The important challenge with conversational AI for low-resource languages is highlighted and the ongoing attempts at tackling it. The presentation of the discussion on the ethics of the gender of conversational AI gives a balanced perpective 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.

12. Bibliographical References

Adewumi, T. P., Liwicki, F., and Liwicki, M. (2019). Conversational systems in machine learning from the point of view of the philosophy of science—using alime chat and related studies. *Philosophies*, 4(3):41.

Adewumi, T. P., Liwicki, F., and Liwicki, M. (2020). The challenge of diacritics in yoruba embeddings. *arXiv preprint arXiv:2011.07605*.

Adewumi, T., Adeyemi, M., Anuoluwapo, A., Peters, B., Buzaaba, H., Samuel, O., Rufai, A. M., Ajibade, B., Gwadabe, T., Traore, M. M. K., Ajayi, T., Muhammad, S., Baruwa, A., Owoicho, P., Ogunremi, T., Ngigi, P., Ahia, O., Nasir, R., Liwicki, F., and Liwicki, M. (2022a). Ìtàkúròso: Exploiting cross-lingual transferability for natural language generation of dialogues in low-resource, african languages.

Adewumi, T., Alkhaled, L., Alkhaled, H., Liwicki, F., and Liwicki, M. (2022b). Ml_ltu at semeval-2022 task 4: T5 towards identifying patronizing and condescending language. *arXiv preprint arXiv:2204.07432*.

Adewumi, T., Brännvall, R., Abid, N., Pahlavan, M., Sabry, S. S., Liwicki, F., and Liwicki, M. (2022c). Småprat: Dialogpt for natural language generation of swedish dialogue by transfer learning. In *5th Northern Lights Deep Learning Workshop, Tromsø, Norway*, volume 3. Septentrio Academic Publishing.

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Alexander, L. and Moore, M. (2007). Deontological ethics.

Allwood, J., Grönqvist, L., Ahlsén, E., and Gunnarsson, M. (2003). Annotations and tools for an activity based spoken language corpus. In *Current and new directions in discourse and dialogue*, pages 1–18. Springer.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015*.

Bradeško, L. and Mladenić, D. (2012). A survey of chatbot systems through a loebner prize competition. In *Proceedings of Slovenian language technologies society eighth conference of language technologies*, pages 34–37. Institut Jožef Stefan Ljubljana, Slovenia.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caldarini, G., Jaf, S., and McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, 13(1):41.

Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–221.

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., and Weston, J. (2020). Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, November. Association for Computational Linguistics.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). 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*, pages 422–428, Marseille, France, May. European Language Resources Association.

Forlizzi, J., Zimmerman, J., Mancuso, V., and Kwak, S. (2007). How interface agents affect interaction between humans and computers. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*, pages 209–221.

Fu, T., Gao, S., Zhao, X., Wen, J.-r., and Yan, R. (2022). Learning towards conversational ai: A survey. *AI Open*, 3:14–28.

Gangal, V., Jhamtani, H., Hovy, E., and Berg-Kirkpatrick, T. (2021). Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4079–4090, Online, August. Association for Computational Linguistics.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezudo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). 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)*, pages 96–120, Online, August. Association for Computational Linguistics.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. (2018). Ethical challenges in data-driven dialogue systems.

In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations, ICLR 2020*.

Javed, S., Adewumi, T. P., Liwicki, F. S., and Liwicki, M. (2021). Understanding the role of objectivity in machine learning and research evaluation. *Philosophies*, 6(1):22.

Jhamtani, H., Gangal, V., Hovy, E., and Berg-Kirkpatrick, T. (2021). Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Limited.

Kerry, A., Ellis, R., and Bull, S. (2008). Conversational agents in e-learning. In *International conference on innovative techniques and applications of artificial intelligence*, pages 169–182. Springer.

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., Song, H., Gottardi, A., Kwatra, S., Pancholi, S., et al. (2018). Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Komeili, M., Shuster, K., and Weston, J. (2021). Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Lee, M., Ackermans, S., Van As, N., Chang, H., Lucas, E., and IJsselsteijn, W. (2019). Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Louwerse, M. M., Graesser, A. C., Lu, S., and Mitchell, H. H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(6):693–704.

Maedche, A. (2020). Gender bias in chatbot design. *Chatbot Research and Design*, page 79.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Mauldin, M. L. (1994). Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539.

Nass, C. I. and Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.

Neff, G. and Nagy, P. (2016). Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:17.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Ni, J., Young, T., Pandelea, V., Xue, F., Adiga, V., and Cambria, E. (2021). Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.

Olabiyi, O. and Mueller, E. T. (2019). Multiturn dialogue response generation with autoregressive transformer models. *arXiv preprint arXiv:1908.01841*.

Paquette, M., Sommerfeldt, E. J., and Kent, M. L. (2015). Do the ends justify the means? dialogue, development communication, and deontological ethics. *Public Relations Review*, 41(1):30–39.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July. Association for Computational Linguistics.

Reiter, E. (2010). 20 natural language generation. *The handbook of computational linguistics and natural language processing*, page 574.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., et al. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rönnqvist, S., Kanerva, J., Salakoski, T., and Ginter, F. (2019). Is multilingual bert fluent in language generation? *arXiv preprint arXiv:1910.03806*.

Sabry, S. S., Adewumi, T., Abid, N., Kovacs, G., Liwicki, F., and Liwicki, M. (2022). Hat5: Hate language identification using text-to-text transfer transformer. *arXiv preprint arXiv:2202.05690*.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Saygin, A. P. and Cicekli, I. (2002). Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258.

Schegloff, E. A. (1968). Sequencing in conversational openings 1. *American anthropologist*, 70(6):1075–1095.

Shieber, S. M. (1994). Lessons from a restricted turing test. *arXiv preprint cmp-lg/9404002*.

Silvervarg, A., Raukola, K., Haake, M., and Gulz, A. (2012). The effect of visual gender on abuse in conversation with ecas. In *International conference on intelligent virtual agents*, pages 153–160. Springer.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y.-L. (2020). 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*, pages 2021–2030, Online, July. Association for Computational Linguistics.

States, U. (2016). Preparing for the future of artificial intelligence.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Weizenbaum, J. (1969). A computer program for the study of natural language. *Fonte: Stanford: http://web. stanford. edu/class/linguist238/p36*.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. (2019). Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if i could: closing gender divides in digital skills through education.

White, M. D. (2009). Immanuel kant. In *Handbook of economics and ethics*. Edward Elgar Publishing.

Xu, J., Szlam, A., and Weston, J. (2021). Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). 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*, pages 483–498, Online, June. Association for Computational Linguistics.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). 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*, pages 270–278.

Zhang, Y., Sun, S., Gao, X., Fang, Y., Brockett, C., Galley, M., Gao, J., and Dolan, B. (2021). Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597*.