Further Advances in Open Domain Dialog Systems in the Fourth Alexa Prize Socialbot Grand Challenge

Shui Hu Yang Liu Anna Gottardi Behnam Hedayatnia Anju Khatri
Anjali Chadha Qinlang Chen Pankaj Rajan Ali Binici Varun Somani
Yao Lu Prerna Dwivedi Lucy Hu Hangjie Shi Sattvik Sahai
Mihail Eric Karthik Gopalakrishnan Seokhwan Kim Spandana Gella
Alexandros Papangelis Patrick Lange Di Jin Nicole Chartier
Mahdi Namazifar Aishwarya Padmakumar Sarik Ghazarian Shereen Oraby
Anjali Narayan-Chen Yuheng Du Lauren Stubell Savanna Stiff
Kate Bland Arindam Mandal Reza Ghanadan Dilek Hakkani-Tur

Amazon Alexa AI

Abstract

Building open domain conversational systems that allow users to have engaging conversations on topics of their choice is a challenging task. The Alexa Prize Socialbot Grand Challenge was launched in 2016 to tackle the problem of achieving natural, sustained, coherent and engaging open-domain dialogs. In the fourth iteration of the competition, university teams have incorporated semantic parsing, common sense reasoning, personalization, neural response generation, as well as novel response ranking models into the state of the art. The Fourth Socialbot Grand Challenge included an improved version of the CoBot (conversational bot) toolkit from the prior competition, along with upgraded topic and intent classifiers, BERT-based named entity recognition model, a punctuation model that injects punctuation marks into the ASR output, and a new neural response generator trained on conversations with Alexa Let's Chat. This paper outlines the advances developed by the university teams as well as the Alexa Prize team to move closer to the Grand Challenge objective, including open domain natural language understanding, commonsense reasoning, dialog management, neural response generation, and dialog evaluation. As of the end of the final feedback phase, the top 7-day average rating achieved by a socialbot was 3.56, with the top 90th percentile conversation duration of 12 minutes 7 seconds.

1 Introduction

Conversational AI is among the most challenging problem domains in artificial intelligence, due to the subjectivity involved in interpreting human language. Broadly speaking, we consider the Conversational AI domain to include a range of tasks in natural language understanding, knowledge representation, common-sense reasoning, dialog evaluation and natural language generation. Complete solutions to these problems will likely require a system at parity with human intelligence [Hassan et al., 2018; Xiong et al., 2016]. With advancements in deep learning and AI, we have made significant progress toward solutions for problems within other very challenging domains such as speech recognition and image recognition in computer vision. However, many of these advancements have been made due to the objective nature of evaluating solutions to these problems and the resulting availability of high quality labeled data with objective ground truth. The language and response generation task in particular has a potentially unbounded response space and a resulting lack of objective success metrics, making it a highly challenging problem to model.

Voice assistants such as Alexa and Google Assistant have significantly advanced the state of science for goal-directed conversations, and these systems have been successfully deployed in production. However, building agents that can carry multi-turn open-domain conversations is still far from a solved problem. To address these challenges and further advance the state of Conversational AI, Amazon launched the Alexa Prize Socialbot Grand Challenge in 2016. The grand challenge objective is to build agents that can converse coherently and engagingly with humans for 20 minutes, and obtain a 4 out of 5 or higher rating from humans interacting with them. There have been various challenges in the research community aiming to improve different aspects of dialog or conversational AI technology, such as the tracks in Dialog System Technology Challenge (DSTC) [Gunasekara et al., 2020], Conversational AI Challenge (ConvAI) [Burtsev et. al., 2018] (persona based, chit-chat and challenges with evaluation). Unlike these challenges, achieving natural, sustained, coherent and engaging open-domain dialogs in spoken form, and in real time, which can be evaluated by real users, is the primary goal of the Alexa Prize Socialbot Grand Challenge. Through this competition, participating universities have been able to conduct research and test hypotheses by building socialbots that interact with real Alexa customers.

As in the last three cycles of the Alexa Prize Socialbot Grand Challenge, upon receiving a request to engage in a conversation with Alexa, e.g. "Alexa, Let's Chat", Alexa customers were read a brief message, then connected to one of the 9 participating socialbots. Customers were provided instructions on how to end the conversation and provide ratings and feedback. The introductory message and instructions changed through the competition to keep the information relevant to the different phrases. After exiting the conversation with the socialbot, which the user could do at any time, the user was prompted for a verbal rating: "How do you feel about speaking with this socialbot again?", followed by an option to provide additional freeform feedback. Ratings and feedback were both shared back with the teams to help them improve their socialbots.

The Fourth Socialbot Grand Challenge was launched to a cohort of Amazon employees on January 4, 2021, followed by a public launch on January 18, 2021, at which time all US Alexa customers could interact with the participating socialbots. Like last year, we ran an initial feedback phase, followed by a competitive quarterfinals from March 2 through April 30, 2021. One team was eliminated from competition after the quarterfinals, and the remaining eight teams participated in the Semifinals from May 4 through June 25, 2021. Five teams qualified to participate as Finalists, and participated in an additional feedback phase through July 23. The closed-door Finals were held on July 27-29, 2021. Throughout the competition, the teams were required to maintain anonymity in their interactions to ensure fairness in the competition. To drive maximum feedback to the teams and improve user engagement, the Alexa Prize experience was promoted through Echo customer emails, social media, and blogs.

2 Capabilities Provided to Teams

In this fourth year of the competition, we continued to provide teams with CoBot, a conversational bot toolkit in Python for natural language understanding and dialog management [Khatri et. al. 2018], which helps teams focus more on scientific advances rather than infrastructure, hosting and scaling up. Teams use CoBot as a library to implement their AWS lambda function which handles incoming user requests as shown in Figure 1. The Cobot toolkit provides a set of tools, libraries and base models designed to help develop, train and deploy open-domain or multi-domain conversational experiences through the Alexa Skills Kit (ASK). Modular, extensible, scalable, and providing abstractions for infrastructure and low-level tasks, CoBot offers a continuous integration pipeline where experimentation in language understanding, dialog management techniques, and response generation strategies can be integrated and tested by a single command. This enables seamless scaling from development and test to production workloads on AWS. CoBot uses many of the same principles found in the Node.JS, Python, and Java SDKs of the Alexa Skills Kit or ASK [Kumar et al., 2017], as well as general dialog toolkits like Rasa

[Bocklisch et al., 2017] and DeepPavlov [Burtsev et al., 2017]. Figure 2 shows the modularized capabilities for generalized dialog management, state tracking, and natural language understanding (NLU) that are exposed through CoBot. Unlike other toolkits, CoBot places an emphasis on infrastructure to host models and handle massive scale at runtime.

For the Fourth Socialbot Grand Challenge, we released an updated version of this software, which included support for autoscaling servers and A/B testing. Autoscaling allows teams to spend less time figuring out how to scale their AWS servers when they deploy more computationally intensive models, freeing up more developer time to focus on model development. All nine participating teams this year also took advantage of the option to set aside a portion of production traffic for experimentation without impacting their ranking. This allowed the teams to iterate quickly on new ideas.

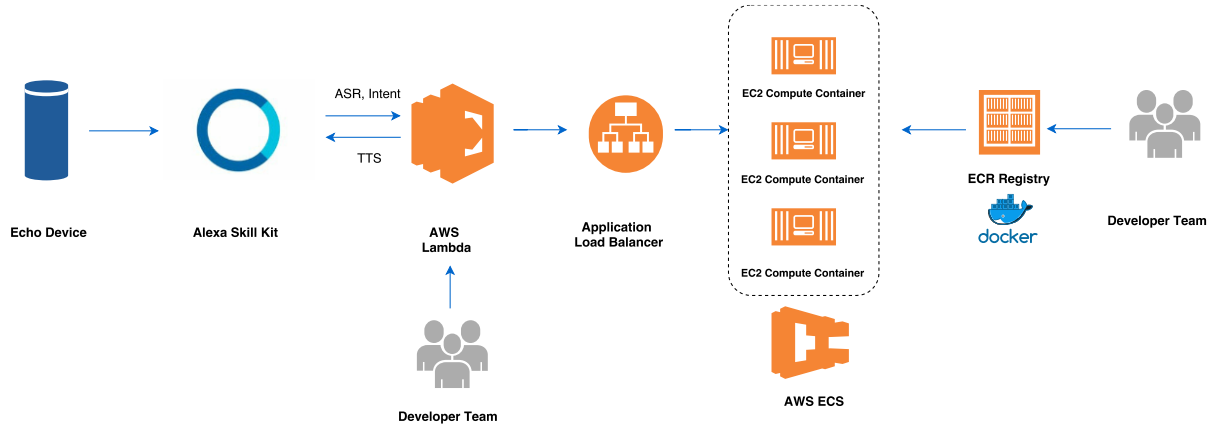

Figure 1. CoBot System Diagram and Workflow

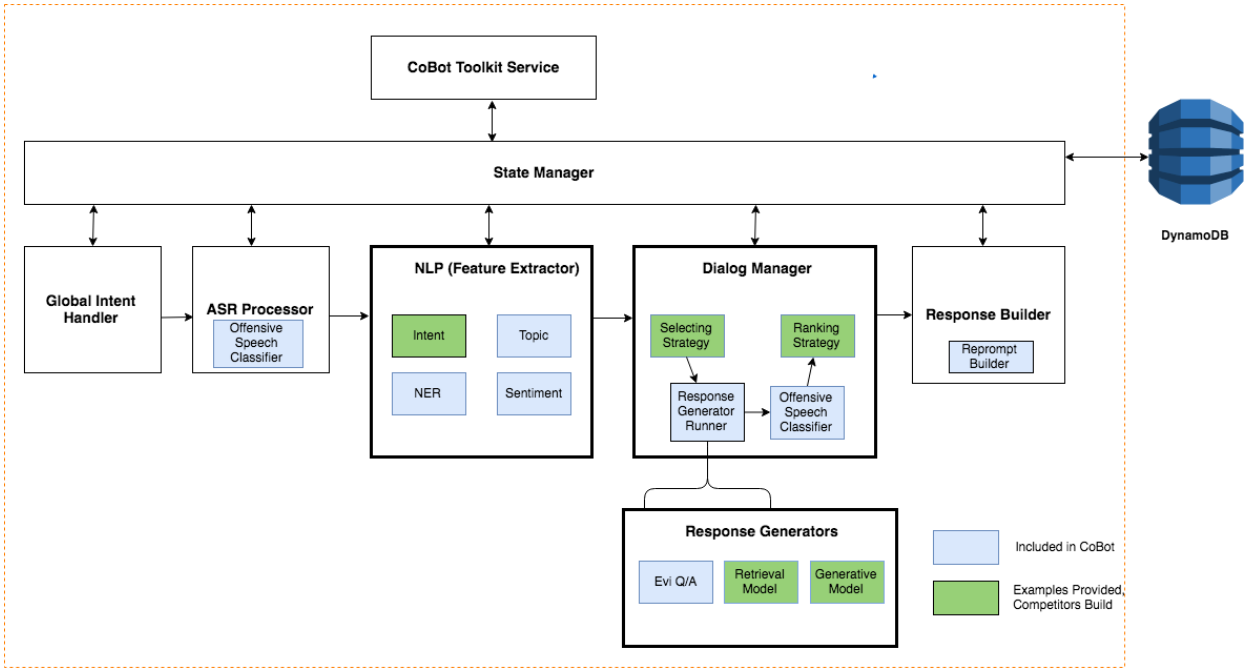

Figure 2. CoBot Architecture

3. Scientific Advancements

3.1 From the Alexa Prize Participants

Most teams have continued to use the overall socialbot framework of NLU, dialog manager, response generators, and ranking multiple candidate responses in order to select the bot's response. However, teams this year also introduced semantic parsing, common sense reasoning, new paradigms for personalization, and experimented with models for response selection and neural response generation. In the following, we briefly describe a few advancements from the university teams. Please refer to the papers from the teams for more details.

3.1.1 Semantic Parsing

To better represent and understand user utterances, Emora introduced semantic parsing and the manipulation of symbolic meaning representation in their dialog management and language generation [Finch et al., 2021]. Their bot follows the paradigm of conversation as a two-way exchange that builds up shared knowledge. The conversation starts with an initial knowledge graph of basic information and assumptions. In each turn, additional information is extracted from the user utterance, inference is made to extend the knowledge from that information, and finally, the newly obtained knowledge is added to the growing knowledge graph for the conversation. Using this knowledge graph, rules are applied to select a subset of the knowledge graph with which to generate a response. Finally, dynamic language templates are used to generate a surface form from the selected subset such that the response consists of a reaction to the previous user utterance and a presentation of novel information [Finch et al., 2021].

The backbone of this approach is a predicate-argument logical form used as the meaning representation [Finch et al., 2021]. This logical form contains concepts, predicates, and an ontology of concepts and predicates. First-order logical inference rules are used to generate new knowledge (e.g., if a dog is wagging its tail, the bot can conclude that the dog is happy) and the resulting predicate-argument structures are attached to the knowledge graph. Dependency parsing with the help of POS tagging, NER, and entity linking is used to convert the user utterance into the logical form. Currently this approach is heavily reliant on hand-crafted rules and concepts, and thus lacks robustness, but it can respond very specifically to user utterances that are covered by its symbolic components [Finch et al., 2021].

3.1.2 Commonsense Reasoning

Commonsense knowledge and inference play a large role in conversations. For example, if a bird is mentioned in a conversation, we would usually infer that the bird can fly and use this information in a conversation. However, this commonsense information is generally unavailable for a socialbot. CASPR tackled this problem by introducing a Conversational Knowledge Template (CKT) for each topic that its socialbot supports [Basu et al., 2021]. The CKT keeps track of the commonsense attributes for a topic and determines which attributes to talk about in the response. When information from the user utterance is extracted, a modified form of Answer Set Programming (ASP) is used to make common sense inferences. For example, if the user says that he/she likes Tom Cruise, ASP will conclude that because people who like an actor tend to like the actor's movies, the user must also like Tom Cruise's movies such as Top Gun. Underlying this commonsense reasoning system is a large set of facts from the real world which is incorporated via a number of knowledge bases including IMDB, Kaggle, and Amazon Kendra.

3.1.3 Personalization

Many users like to discuss personal subjects such as personal preferences or musical tastes, including sometimes asking the socialbot for its own preferences. Several teams chose to build personas for the bot that can handle some of these subjects and questions and predict which persona is the most appropriate for the current user. Athena built a user model that tracks user-specific data like the user's name, their interests, pet's name, weekend hobbies, and favorite dinosaur and uses this data to promote topics that it believes the user is interested in [Patil et al., 2021]. Users spent more time in such topics compared to topics that were not selected by the user model. Proto also tracked user attributes and used this to personalize responses [Saha et al., 2021]. Alquist learns from the user and builds a user profile that is taken into account during the conversation [Konrad et al., 2021]. DREAM identifies users' personality in order to have different kinds of conversations, e.g., let extroverts lead the conversation [Baymurzina et al., 2021].

3.1.4 Neural Response Generation

Neural network-based language generation has seen great progress in the past several years. Chitchat systems that are purely based on end-to-end neural models have demonstrated strong performance [Adiwardana et. al. 2020, Roller et. al. 2020]. To improve the coverage of the system responses or to address the long tail problem, multiple teams have incorporated neural response generation in their systems and found that these are very helpful for out-of-domain user utterances that are not supported by their other response generators. Amazon already provides a GPT-2-based [Radford et. al., 2019] neural response generation service for participating teams that is trained from some past Alexa Prize conversations, but teams also built other neural response generators.

Alquist trained DialoGPT [Zhang et al., 2020] on a number of datasets including Topical Chat dataset [Gopalakrishnan et al., 2019] and crawled Reddit data, and investigated adding controls such as Question and Statement [Konrad et al., 2021]. Athena trained discourse driven GPT-2 model to follow dialog policy based on dialog acts, and data-to-text model based on T5 and BART [Patil et al., 2021]. DREAM fine-tuned BlenderBot for knowledge grounded response generation [Baymurzina et al., 2021]. Genuine used DialgoGPT and also built its own GPT-2 based response generator on multiple existing datasets [Rodriguez-Cantelar et al., 2021]. Proto fine-tuned BlenderBot on different datasets such that the generation models are tailored to different cases, including knowledge grounded and chit chat utterances [Saha et al., 2021]. Chirpy Cardinal built a neural response generator that was distilled from BlenderBot, and also a BART based model to infill structured knowledge information about entities [Chi et al., 2021].

3.1.5 Response Selection

As in previous years, all teams continued to use the framework of having multiple response generators generate candidate responses for each user utterance and then selecting one of the candidates as the response. This calls for a response selection mechanism that CoBot does not offer. Athena applied a BERT-based neural response ranker that had been trained on Alexa Prize data [Patil et al., 2021]. DREAM used a rule-based response selector [Baymurzina et al., 2021]. DialogRPT [Gao et al., 2020] is a popular method that has been explored by several teams, including Chirpy Cardinal, Genuine, Alquist, Proto. Another neural model used for ranking is the polyencoder [Humeau et al., 2019]. Both Viola and Proto evaluated this model. Viola found that a fine-tuned polyencoder response selector outperformed the conversational evaluator offered by Cobot [Cho et al., 2021].

3.2 From the Alexa Prize Team

3.2.1 Intent and Topic Classification

For language understanding in social conversations, we provided topic and intent classification services for Alexa Prize teams. The current system is based on hierarchical recurrent neural network (HRNN)

which was introduced in Gabriel et al., 2020. It includes two RNNs. The first one takes the word sequence of each utterance and learns the utterance representation, and the other RNN takes a sequence of the utterance representations in a given dialog context and learns the dialog context representation. The model was trained by multi-task learning with two objectives for dialog topic and intent classification tasks.

For the fourth challenge, we improved the service with a retrained model on new labeled Alexa Prize conversations. As shown in Table 1, the updated model achieves significantly higher performances in both topic and intent classification tasks compared to the previous version for the third challenge.

	Topic	Intent
AP3 model	67.03	63.96
AP4 model	78.93	74.60

Table 1. Topic and Intent classification results (accuracy in %) using the HRNN models we provided to the teams

We have also investigated using large pre-trained language models such as BERT for topic and intent classification services in comparison with HRNN. Briefly, we encode the user utterances with a pre-trained language model and then project the representation vector of the first "[CLS]" token into the probability vector with a two-layer fully connected neural network. We fine-tuned the language models with the annotated Alexa Prize data. In addition, we evaluated our internally developed version of BERT, named WLM [Namazifar et al., 2021], which pre-trains the BERT model on warped sentences of Wikipedia corpus. In comparison with BERT, WLM injects more noise to the pre-training corpus in addition to masking, such as deleting, replacing, and shuffling random selected tokens. By doing so, WLM is supposed to be more robust than BERT.

Table 2 summarizes the comparison of HRNN and the pre-trained language models for topic and intent classification. As can be seen, pre-trained language models show some advantage compared with HRNN. We plan to update the topic and intent classification service with these improved models.

	Topic	Intent
HRNN	78.93	74.6
BERT-Base	79.99	**75.84**
WLM-Base	**80.3**	74.24

Table 2. Topic and intent classification results (accuracy in %) by finetuning large pretrained language models

3.2.2 NER and ER

We provide an entity recognition (NER) service to the teams. The model is a BERT-based model that performs token classification on the user utterance using BIO representation. Further post-processing is applied to the token level predictions to obtain the entity span. We picked 8 popular domains in open-domain conversations including Fashion, Politics, Books, Sports, Music, Science/Technology, Game, Video/Movies, and initially defined 50 entity types for these domains. We recruited crowd workers to annotate selected Topical-chat dialog (TCS) and some socialbot style conversations. The model was first trained on TCS data, then fine-tuned on the socialbot conversations. We grouped some entity types to form coarse-grained categories (e.g., different types of persons for different domains are all mapped to

Person) in model training and inference. Using about 7k TCS utterances and 6k Socialbot utterances for training, the model's performance on the test set (around 1k utterances) is 0.82 for span detection, and 0.77 for entity types. This model was provided to the teams and widely used by them.

In addition to this baseline NER model, we have also explored the use of dialog contextual information for entity recognition and linking, and furthermore their impact on response generation [Shang et al., 2021]. On different datasets, including Wizard of Wikipedia, socialbot data, and a synthetic data set, we observed varying degree of improvement when dialog context is used. And human evaluation of system responses that are generated by a neural response generator leveraging the entity linking information shows improved appropriateness and informativeness scores. We plan to integrate these improved NER and entity linking models and release them in the future.

3.2.3 Punctuation Model

ASR output is just a sequence of tokens without any punctuation marks. We provided a service to predict punctuation. The model is a BERT-based on that predicts whether there is a punctuation after a word in the sequence. It was trained using the Cornell movie corpus. The F1-scores for period and question mark are about 0.8, and lower for comma (0.63). The movie corpus has somewhat different style from user utterances in the socialbot in terms of topic coverage, human-computer conversation style, and ASR errors. We plan to annotate some socialbot conversations with punctuation to retrain or fine tune the model such that it has matched training conditions.

3.2.4 Offensive Utterance Classifier

Our current version of the production classifier is trained on a combination of bootstrapped reddit comments and Wikipedia Toxic Comments Dataset (WTC) using a Bi-LSTM classifier, and served in combination with a list of sensitive, block-list words. Some of the common issues we observe with the current prod classifier is high sensitivity to block-list words (i.e., generating too many false positives), and a huge gap with the current SOTA models. We upgraded our offensive classifier by fine-tuning pre-trained large language model (RoBERTa) on public Wikipedia Toxic Comments Dataset (WTC) and in domain utterances sampled from previous socialbot logs. Recently, multiple studies have shown that "build it, break it" style models where additional data is collected using the best available model, are more robust for offensive utterance classification. We employ a RoBERTa model trained on WTC data to identify utterances from socialbot logs that are offensive (labeled by human annotators) but classified otherwise. We re-train RoBERTa model using this adversarial data and observed significant boost for the socialbot domain data and 6% boost offensive utterance F-measure on WTC test data. We are in the process of updating the production offensive utterance classifier with this new model.

3.2.5 NRG Models

We developed a neural response generation (NRG) model by training on the socialbot dialog data. We used the log from the top-5 socialbots spanning from 1/2020 - 5/2020, and filtered the conversations based on user ratings. Since we do not have the corresponding knowledge that was used to structure the responses by the template-based response generators in the teams' socialbots, we only used the responses to train our model, that is, it is not a knowledge-grounded NRG model. In last year's competition, we provided a knowledge-grounded model that was trained using the Topical-chat data.

For all our models, we use the GPT2 model [Radford et al., 2019] to finetune in a TransferTransfo fashion [Wolf et al., 2018]. In this system, GPT2 is fine-tuned in a multi-task learning fashion with the language modeling and next utterance classification tasks. Starter code for Transfer-Transfo along with

the pre-trained model and BPE-based vocabulary is provided on GitHub by Hugging Face. To train the language model task we take in the dialog history and minimize the cross-entropy loss on the ground truth response. To train the next utterance classification task, for each turn in a dialog, which we denote as the ground truth candidate, we sample a random turn from a random dialog and denote it as a distractor candidate. The task involves taking in as input the dialog history and train the model to predict the ground truth candidate as the appropriate next turn.

We released our NRG model to be used for the university teams. This is a service that is regularly used by the teams. For example, there are more than 100,000 calls per day and the P50 latency is less than 300ms.

We have investigated novel knowledge selection methodologies for knowledge-grounded NRG response generators. As part of this effort, we are planning on releasing a new augmented version of the WOW dataset, where we reannotated some number of dialogues to include multiple relevant knowledge sentences per turn of dialogue. On average, WOW++ includes 8 relevant knowledge sentences per dialogue context, embracing the inherent ambiguity of open-domain dialogue knowledge selection. Using WOW++ we were able to train new RoBERTa-based knowledge ranking algorithms that when combined with a GPT2-based response generator, outperformed other model-based methods on an end-to-end human evaluation task [Eric et al., 2021].

Furthermore, to improve the NRG services we have been exploring controlled NRG with a focus on two factors. The first one is empathy. The goal for this is to generate responses with condolence when user utterances show certain negative sentiment. The second one is topical control, which uses the same topic as the one detected for the user utterance to guide the response generation. For both controls, we add additional tokens (e.g., condolence, different topic labels) in addition to dialog context for response generation. In preliminary offline experiments and human evaluation, we observed that for both cases, the model is able to follow the control tokens to generate responses when controls are provided, and there is no degradation in response quality for other cases. Another improvement we are making to the knowledge-grounded NRG models is to reduce hallucination and increase its factual correctness (response is factually consistent with the provided knowledge). We plan to release these models to the teams.

Finally, we optimized our neural response generation inference code to reduce model invocation latency. We leverage pre-computed hidden-states to speed up sequential decoding and apply additional graph optimizations such as constant folding, redundant node elimination and semantics-preserving node fusion as provided by onnxruntime. We were able to reduce the average inference latency measured on Topical Chat `test_freq` for our GPT-2 medium sized NRG model from 708 ms to 154 ms. We sample at least 1 and at most 100 tokens using a batch size of 1, truncating context to 64 tokens and knowledge to 32 tokens. The reduction in latency will allow us to provide participants with larger models in upcoming challenges.

3.2.6 User and Rating Analysis

User ratings is one of the most important and often-used pieces of information that helps develop, debug, and fine tune conversational agents. In a typical machine learning based conversational agent, a dialogue rating will reflect the quality of the dialogue and the agent will be trained to maximize that rating. This approach, however, treats all users as a homogeneous whole and disregards individuality: the fact that each user is different, with different experiences, personalities, needs, and expectations that can lead them to perceive an interaction with the same conversational agent differently. Treating conversational ratings as monolithic, therefore, will lead to a conversational agent that tends to an ‘average’ user, rather than being personalized to each individual user.

To better understand how people rate their interactions with conversational agents, we conducted a study. One macro-level characteristic that has been shown to correlate with how people perceive their inter-personal communication is personality [Astrid et al. 2010, Cuperman and Ickes 2009, McCrae and Costa 1989]. We specifically focus on agreeableness and extraversion as variables that may explain variation in ratings and therefore provide a more meaningful signal for training or personalization. In order to elicit those personality traits during an interaction with a conversational agent, we designed and validated a fictional story, grounded in prior work in psychology. We then implemented the story into an experimental conversational agent that allowed users to opt-in to hearing the story. Results from the personality story suggest that extraversion does not predict users' overall experience with the experimental conversational agent. While this may be a reflection of a lack of data, this could also be a reflection in the difference between human-human interactions and human-AI interactions. On the other hand, results from the personality story show that agreeableness does predict overall conversation ratings. These results suggest that perhaps agreeableness between human-human interactions is more likely to transition to human-AI interactions than extraversion. An interesting finding from this study is that users who chose to listen to the personality story tend to score high on agreeableness and also tend to provide higher ratings for their conversational experience.

We have also conducted other studies with the goal to understand dialog quality and user ratings. We asked 3rd party evaluators to rate the socialbot conversations, using a scale of 1 to 5. On the annotated 930 conversations, we found user and 3rd party ratings are very different: overall the correlation between them is low (0.207); there are more ratings on the two end (1 or 5) in the user ratings than the 3rd party evaluations (the latter shows a more normal distribution), and the average user rating is higher than 3rd party scores (3.37 by users comparing to 2.84 by annotators). Table 3 shows the distribution of the ratings by the users and the 3rd party evaluators.

	1	2	3	4	5
User	140	155	168	156	311
3rd party annotator	73	258	368	204	27

Table 3: Rating distributions for 930 rated conversations by socialbot users and 3rd evaluators.

4. Socialbot Quality

Over the course of this year's competition, the socialbots started off strong thanks to the neural response generator that we made available via Cobot, which provided a strong baseline performance for all bots. However, socialbot performance plateaued early in Semifinals and the quality of conversations declined during Semifinals as teams focused on experimenting with research ideas that are inherently unpredictable. Despite outperforming the socialbots of the Third Socialbot Grand Challenge at the start of Fourth Socialbot Grand Challenge, this year's bots fell behind last year's by the end of Semifinals. In this section we provide various metrics to evaluate the socialbot quality in the Fourth Socialbot Grand Challenge and compare with that observed in the Third Socialbot Grand Challenge.

4.1 Ratings

After each conversation, Alexa users were asked to rate their interaction on a scale of 1-5. As shown in Fig. 3, average ratings were 37.0% higher in the first week of Year 4 of the competition compared to the first week of Year 3, but by the end of Semifinals, it was 4.1% lower. When Quarterfinals began, eight out of nine socialbots had an average rating above 3.0/5 with the lowest rated socialbot rated just below 3.0/5. We surfaced a small part of the traffic to the winning socialbot from Year 3 in order to compare

with last year and found that its rating was also 9.7% lower than at the end of Semifinals last year. The slight drop in ratings during Semifinals may have also been due in part to rising expectations from users.

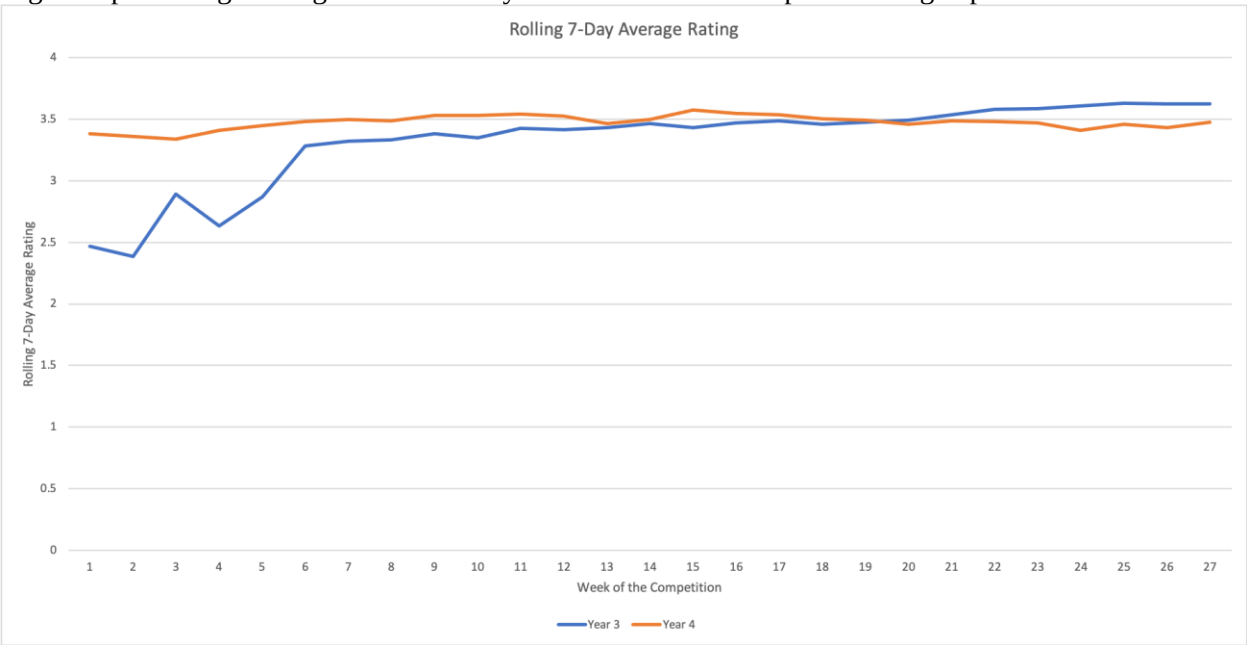

Figure 3. Finalist ratings over time, Year 3 vs Year 4

4.2 Conversation Duration

Our primary success metric for conversation duration is the p90 duration, or the 90th percentile duration of conversations (i.e. 10% of conversations have a duration longer than this number). We consider this a strong proxy for a maximum duration that the socialbot can sustain an interesting and engaging conversation with a dedicated interactor. We also track the median (p50) conversation duration.

Until the start of Semifinals in the Fourth Alexa Prize Socialbot Grand Challenge, the p90 conversation duration was higher than at the same time during the prior year's competition, but during Semifinals the duration declined to below that of last year. At the end of Semifinals this year, the average p90 duration for Finalist teams was 677 seconds (just over 11 minutes) whereas last year's Finalists achieved an average duration of 739 seconds (just over 12 minutes) at the end of that year's Semifinals (Fig. 5).

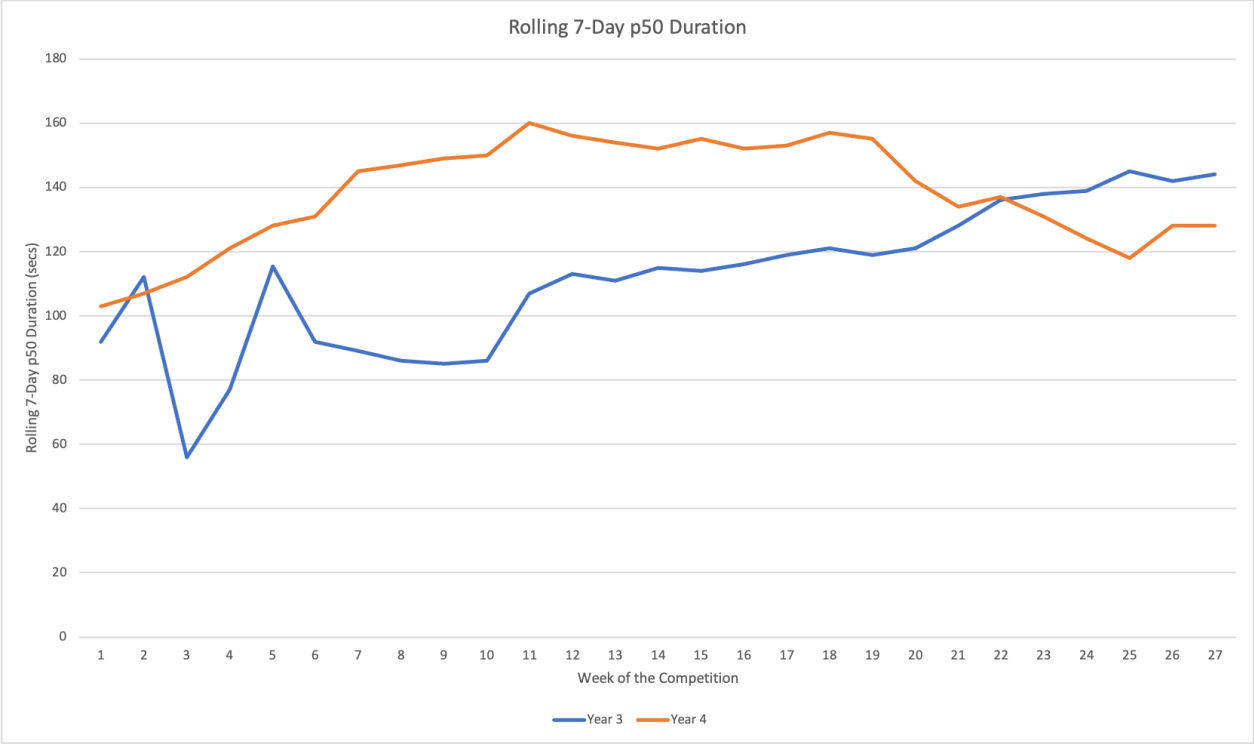

Figure 4. Finalist p50 (median) durations over time, Year 3 vs Year 4

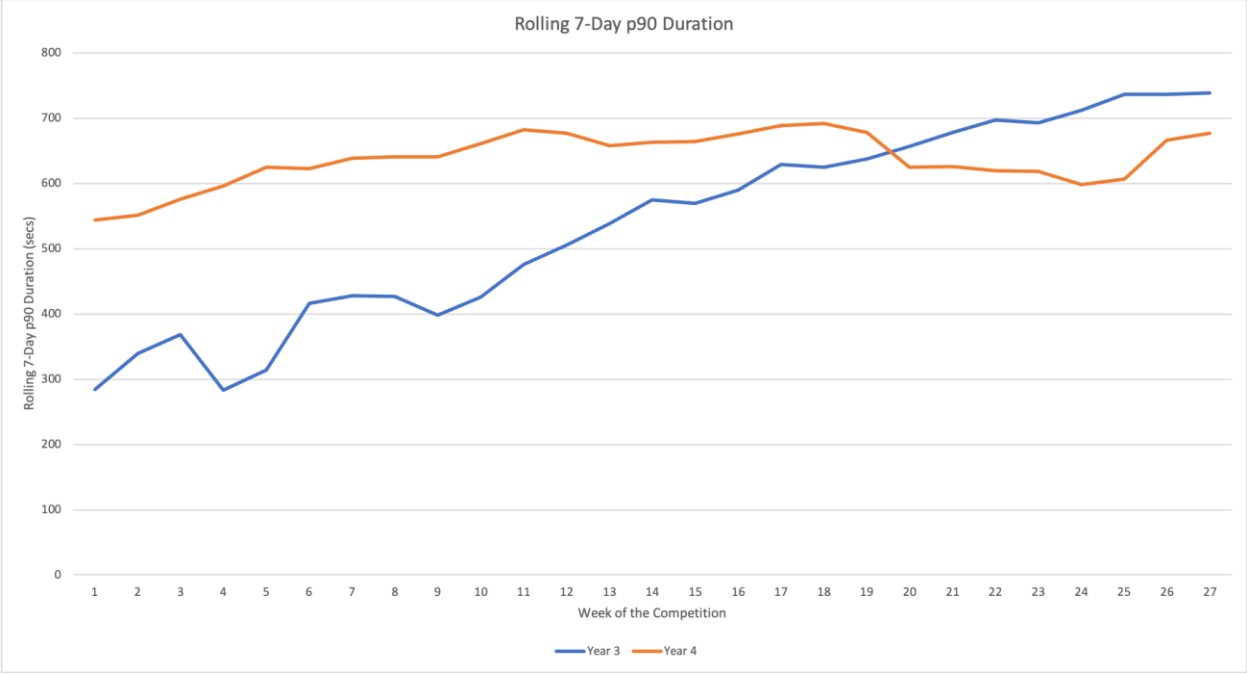

Figure 5. Finalist p90 durations over time, Year 3 vs Year 4

4.3 Response Quality

We obtained annotations for response quality of the Alexa Prize conversations across all the Finalist socialbots. We defined five classes to describe the quality of a response: (1) Poor (response is not comprehensible or bot didn't understand user or response is sensitive) (2) Not Good (the socialbot had some understanding of what user said but contains incorrect or inappropriate information) (3) Passable (response is on topic but is generic or contains too little or too much of information) (4) Good (response is logical, on topic, contains accurate information but lacks personality or is not presented well and obviously coming from a bot), and (5) Excellent (contains accurate information, is complete and it is hard to identify if the response was coming from a bot). By treating these classes as numeric values on a 5-point scale, we can compute an average turn rating.

In the 2021 competition, the Finalist teams averaged a 2.46 turn rating after the close of the Semifinals. This declined from 2.88 from the same time last year.

5. Conclusion and Future Work

The problem of engaging in coherent, engaging conversations is one of the most challenging problems in the artificial intelligence field. Several tasks associated with Conversational AI such as language understanding and generation, knowledge representation, commonsense reasoning and dialog evaluation are believed to be "AI Complete", or truly human-intelligence equivalent problems. To address these challenges, Amazon launched the Alexa Prize Socialbot Grand Challenge, wherein some of the best research groups across the world work towards a common goal of advancing the state of the art in Conversational AI.

The Fourth Socialbot Grand Challenge ran from November 2020 through July 2021, and the participating university teams have built socialbots that can converse with Alexa users coherently and engagingly on a wide variety of topics. Building on the foundation of the first three years of the competition that included sophisticated statistical dialog management, improved personalization, complex utterance handling, and use of large-scale Transformer-based models for a wide variety of tasks, teams this year experimented with new approaches in semantic parsing, common sense reasoning, a variety of neural response generation models, controlled language generation, and response selection models. These innovations take time to mature and we have yet to see their impact on ratings, conversation duration and human-annotated measures of socialbot quality. Unlike in previous years, performance by the socialbots in Year 4 of the Socialbot Grand Challenge has deteriorated across the board. Conversation duration is down and ratings have declined. The fact that even the reference bot that was the winning socialbot from last year saw a decline in its ratings suggests that users this year probably have higher expectations from socialbots, leading to lower ratings for the same experience. Note that other reasons for the lower ratings for the reference bot include the fact that knowledge used in the bot that was not updated. Another major factor in the decline of bot performance is that many of the returning teams this year chose to devote their time to experimenting with new ideas. Emora, which won last year, has been experimenting with semantic parsing, which seems to be less robust and has lower coverage over user utterances. Alquist also reduced reliance on their scripted conversations that had a solid track record in previous years in order to experiment more with neural response generation. The focus on experimentation this year is in keeping with the spirit of Alexa Prize which is to encourage innovative research for multi-domain chitchat-style conversations.

It is still Day One for Conversational AI, as the capabilities of natural language generation, large-scale transformer-based models, and integration of traditional AI approaches into deep learning-based systems continues. We expect to see the initiatives started by the teams competing this year to continue and come to fruition in the coming years. As the socialbots develop the ability to discuss topics in a more in-depth

and personalized manner, we expect that today's Grand Challenge objective will become tomorrow's reality.

References

Gunasekara, C., Kim, S., D'Haro, L, Rastogi, A., Chen, Y.-N., Eric, M., Hedayatnia, B., Gopalakrishnan, K., Liu, Y., Huang, C.-W., Hakkani-Tur, D., Li, J., Zhu, Q., Luo, L., Liden, L., Huang, K., Shayandeh, S., Liang, Runze., Peng, B., Zhang, Z., Shukla, S., Huang, M., Gao, J., Mehri, S., Feng, Y., Gordon, C., Alavi, S., Traum, D., Eskenzai, M., Beirami, A., Cho, E., Crook, P., De, A., Geramifard, A., Kottur, S., Moon, S., Poddar, S., Subba, R. (2020). Overview of the Ninth Dialog System Technology Challenge: DSTC9. *arXiv preprint arXiv:2011.06486*.

Burtsev, M. (2018). *ConvAI2*. Retrieved from http://convai.io/

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M. & Liu, S. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Botvinovsky, E., Bushkov, N., Gureenkova, O., Kamenev, A., Konovalov, V. & Kuratov, Y. (2018). DeepPavlov: An Open Source Library for Conversational AI.

Bocklisch, T., Faulker, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Khatri, C. Hedayatnia, B., Venkatesh A., Nunn J., Pan Y., Liu Q., Song H., Gottardi A., Kwatra S., Pancholi S., Cheng M., Chen Q., Stubell S., Gopalakrishnan K., Bland K., Gabriel R., Mandal A., Hakkani-Tur D., Hwang G., Michel N., King E., & Prasad R. (2018). Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize. 2nd Proceedings of the Alexa Prize.

Adiwardana, D., Luong, M.-T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q., 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*

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*

Kumar, A., Gupta, A., Chan, J., Tucker, S., Hoffmeister, B., Dreyer, M., ... & Monson, C. (2017). Just ASK: building an architecture for extensible self-service spoken language understanding. *arXiv preprint arXiv:1711.00549*.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. Proc. Interspeech 2019,1891–1895.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gao, X., Y. Zhang, M. Galley, et al. Dialogue response ranking training with large-scale human feedback data. CoRR, abs/2009.06978, 2020.

Humeau, S., K. Shuster, M. Lachaux, et al. Real-time inference in multi-sentence tasks with deep pretrained transformers. *CoRR, abs/1905.01969, 2019*.

Zhang, Y., S. Sun, M. Galley, et al. Dialogpt: Large-scale generative pre-training for conversational response generation. *ACL, 2020*.

Chi, E.A., Chiam, C., Chang, T., Lim, S.K., Rastogi, C., Iyabor, A., He, Y., Sowrirajan, H., Narayan, A., Tang, J., Li, H., Paranjape, A., Manning, C.D. Neural, neural everywhere: controlled generation meets scaffolded, structured dialogue. *Alexa Prize Proceedings, 2021*.

Rodriguez-Cantelar, M., de la Cal, D., Estecha, M., Grande, A., Martin, D., Rodriguez, N., Martinez, R., Fernando, L. Genuine2: an open domain chatbot based on generative models. *Alexa Prize Proceedings, 2021*.

Cho, H., Shbita, B., Shenoy, K., Liu, S., Patel, N., Pindikanti, H., Lee, J., May, J. Viola: a topic agnostic generate-and-rank dialogue system. *Alexa Prize Proceedings, 2021*.

Saha, S., Das, S., Soper, E., Pacquetet, E., Srihari, R.K. Proto: a neural cocktail for generating appealing conversations. *Alexa Prize Proceedings, 2021*.

Baymurzina, D., Kuznetsov, D., Evseev, D., Karpov, D., Sagirova, A., Peganov, A., Ignatov, F., Ermakova, E., Cherniavskii, D., Kumeyko, S., Serikov, O., Kuratov, Y., Ostyakova, L., Kornev, D., Burtsev, M. DREAM Technical Report for the Alexa Prize 4. *Alexa Prize Proceedings, 2021*.

Konrad, J., Pichl, J., Marek, P., Lorenc, P., Ta, V.D., Kobza, O., Sedivy, J. Alquist 4.0: towards socialb intelligence using generative models and dialogue personalization. *Alexa Prize Proceedings, 2021*.

Patil, O., Reed, L., Bowden, K.K., Juraska, J., Cui, W., Harrison, V., Rajasekaran, R., Ramirez, A., Li, C., Zamora, E., Lee, P., Bheemanpally, J., Pandey, R., Ratnaparkhi, A., Walker M. Athena 2.0: discourse and user modeling in open domain dialogue. *Alexa Prize Proceedings, 2021*.

Basu, K., Wang, H., Dominguez, N., Li, X., Li, F., Varanasi, S.C., Gupta, G. CASPR: a commonsense reasoning-based conversational socialbot. *Alexa Prize Proceedings, 2021*.

Finch, S.E., Finch, J.D., Huryn, D., Hutsell, W., Huang, X., He, H., Choi, J.D. An approach to inference-driven dialogue management within a social chatbot. *Alexa Prize Proceedings, 2021*.

Gabriel, R., Liu, Y., Gottardi, A., Eric, M., Khatri, A., Chadha, A., ... & Hakkani-Tür, D. (2020). Further advances in open domain dialog systems in the third Alexa prize socialbot grand challenge. *Alexa Prize Proceedings, 2020*.

Namazifar, M., Tur, G., Hakkani-Tür, D. Warped Language Models for Noise Robust Language Understanding, *SLT 2021*.

Shang, M., Wang, T., Eric, M., Chen, J., Wang, J., Welch, M., Deng, T., Grewal, A., Wang, H., Liu, Y., Kiss, I., Liu, Y., Hakkani-Tur, D. Entity Resolution in Open-domain Conversations. *In Proceedings of NAACL, Industry track, 2021*.

Wolf, T., Sanh, V., Chaumond, J., Delangue, C. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *In proceedings of Neurips CAI workshop, 2018.*

Eric, M., Chartier, N., Hedayatnia, B., Gopalakrishnan, K., Rajan, P., Liu, Y., Hakkani-Tur, D. Multi-Sentence Knowledge Selection in Open-Domain Dialogue. *INLG, 2021*.

Astrid, M., Kramer, N.C., and Gratch, J. How our personality shapes our interactions with virtual characters-implications for research and development. *In International Conference on Intelligent Virtual Agents, pages 208–221. Springer, 2010.*

Cuperman, R. and Ickes, W. Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of personality and social psychology, 97(4):667, 2009.*

McCrae, R.R. and Costa, P.T. The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of personality and social psychology, 56(4):586, 1989*.