Towards a Human-like Open-Domain Chatbot

Daniel Adiwardana   Minh-Thang Luong   David R. So   Jamie Hall
Noah Fiedel   Romal Thoppilan   Zi Yang   Apoorv Kulshreshtha
Gaurav Nemade   Yifeng Lu   Quoc V. Le
Google Research, Brain Team
{adiwardana,thangluong,davidso,jamiehall,nfiedel,romzee,ziy,
apoorvk,gnemade,yifenglu,qvl}@google.com

Abstract

We present Meena, a multi-turn open-domain chatbot trained end-to-end on data mined and filtered from public domain social media conversations. This 2.6B parameter neural network is simply trained to minimize perplexity of the next token. We also propose a human evaluation metric called Sensibleness and Specificity Average (SSA), which captures key elements of a human-like multi-turn conversation. Our experiments show strong correlation between perplexity and SSA. The fact that the best perplexity end-to-end trained Meena scores high on SSA (72% on multi-turn evaluation) suggests that a human-level SSA of 86% is potentially within reach if we can better optimize perplexity. Additionally, the full version of Meena (with a filtering mechanism and tuned decoding) scores 79% SSA, 23% higher in absolute SSA than the existing chatbots we evaluated.

1   Introduction

The ability to converse freely in natural language is one of the hallmarks of human intelligence, and is likely a requirement for true artificial intelligence. In order to explore this aspect of intelligence, many researchers are working on open-domain chatbots. Unlike closed-domain chatbots, which respond to keywords or intents to accomplish specific tasks, open-domain chatbots can engage in conversation on any topic. Some open-domain chatbots such as MILABOT (Serban et al., 2017), XiaoIce (Zhou et al., 2018)[1], Gunrock (Chen et al., 2018), Mitsuku (Worswick, 2018)[2] and Cleverbot[3] (by Rollo Carpenter) display human-like attributes, but rely on complex frameworks, such as dialog managers with
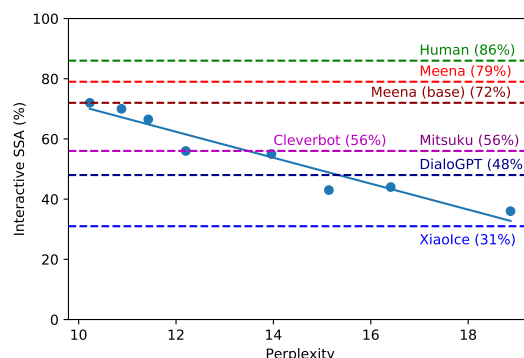

Figure 1: Interactive SSA vs Perplexity. Each point is a different version of the Meena model. A regression line is plotted, for which the coefficient of determination (R^2) is 0.93, an indication of strong correlation between perplexity and the human evaluation metric (SSA). The dotted lines show the SSA performance of other chatbots, humans (86%), the best end-to-end trained Meena model (72%), and the full version of Meena which incorporates a filtering mechanism and tuned decoding (Section 5) and scores 79%. Mitsuku and Cleverbot scored the same on overall SSA, but Mitsuku displayed higher sensibleness, whereas Cleverbot had higher specificity. See Sections 2.5, 2.6, and 4.3 for more details on how we performed these comparisons and how to interpret the results.

knowledge-based, retrieval-based, or rule-based systems. End-to-end neural network approaches (Shang et al., 2015; Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016; Zhang et al., 2019), on the other hand, offer the simplicity of a single learned model. Despite much research, open-domain chatbots still have weaknesses that prevent them from being generally useful: they often respond to open-ended input in ways that do not make sense, or with replies that are vague and

---

[1] https://www.msxiaobing.com/
[2] https://www.pandorabots.com/mitsuku/
[3] https://www.cleverbot.com/

Conversations with Meena, and with various other chatbots, are available at https://github.com/google-research/google-research/tree/master/meena/

generic.

Here we present Meena, a generative chatbot model that was trained end-to-end on 40B words mined and filtered from public domain social media conversations. With Meena, we push the limits of the end-to-end approach and show that a large-scale low-perplexity model can be a good conversationalist. We use a seq2seq model (Sutskever et al., 2014; Bahdanau et al., 2015) with the Evolved Transformer (So et al., 2019) as the main architecture. The model is trained on multi-turn conversations where the input sequence is all turns of the context (up to 7) and the output sequence is the response. Our best model has 2.6B parameters and achieves a test perplexity of 10.2 based on a vocabulary of 8K BPE subwords (Sennrich et al., 2016).

To measure the quality of Meena and other chatbots, we propose a simple human evaluation metric. Sensibleness and Specificity Average (SSA) combines two fundamental aspects of a human-like chatbot: making sense and being specific. We ask human judges to label every model response on these two criteria. The first part of the metric, sensibleness, is a basic requirement. To converse properly with a human, a bot's responses have to make sense in context; humans typically take this for granted when conversing with one another, and our evaluations find that 97% of human-produced statements meet this criterion (see Section 4.2). However, making sense is not enough. If a model is designed with sensibleness as its only objective, its responses could be vague and boring, since that is a safe strategy to avoid being penalised for not making sense. For example, closed-domain chatbots typically respond with a generic apology when a human asks something outside their domain; some end-to-end learned chatbots respond "I don't know" to many inputs (Li et al., 2016a); and Turing Test contest entrants often try to avoid detection by being strategically vague (Venkatesh et al., 2018). They succeed in not generating gibberish or contradicting themselves, but at the cost of not really saying anything of substance. To mitigate this, we add a second dimension to the SSA metric, which asks our evaluators whether a response is specific given the context. This prevents bots from hiding behind vague replies, allowing us to more openly examine what they are capable of. As discussed in Section 2.1, this successfully distinguishes between generic and lively responses,

while also being simple and easy for crowd workers to understand.

We compare Meena, humans, and other open-domain chatbots using the SSA metric with two types of human evaluation: *static* and *interactive*. For static evaluation, we curated a dataset with 1,477 multi-turn conversations. For interactive evaluation, humans could chat about anything they wanted. We were surprised, but pleased, to discover that the SSA metric shows strong correlation with Meena's perplexity, both in static and interactive evaluation. In other words, the better that Meena fit its training data, the more sensible and specific its chat responses became. At first glance, this result may seem intuitive, but it surprised us because recent research found a poor correlation between human evaluation scores and automatic metrics such as BLEU (Liu et al., 2016; Lowe et al., 2017).

Our best end-to-end learned model has an average of 72% SSA. The full version of Meena scores 79% by incorporating a filtering mechanism and tuned decoding (Section 5). This is still below the 86% SSA achieved by an average human, but is far closer than the other chatbots we tested. We note that humans have very high sensibleness, but significantly lower specificity, as detailed in Section 4.2.

We will also discuss weaknesses of our methodology. For example, our static evaluation dataset is too restricted to capture all aspects of human conversations. Nevertheless, the fact that Meena achieves such a high SSA score and that there is a correlation between SSA and perplexity means that a human-like chatbot, in terms of sensibleness and specificity, could be in sight if we can attain better perplexity.

Our contributions are: (1) proposing a simple human evaluation metric for multi-turn open-domain chatbots that captures basic, but important, attributes of human conversation; (2) showing evidence that perplexity is an automatic metric that correlates with human judgment, in contrast to recent findings on other automatic metrics mentioned above; (3) demonstrating that an end-to-end neural model with sufficiently low perplexity can surpass the sensibleness and specificity of existing chatbots that rely on complex, handcrafted frameworks developed over many years.

2 Evaluating chatbots

Evaluating chatbots and natural language generation is a well-known challenge (Liu et al., 2016; Lowe et al., 2017; Novikova et al., 2017; Hashimoto et al., 2019), which we aim to address in this paper. First, we propose a human evaluation metric that captures key elements of human-likeness of conversational responses (Section 2.1). We then describe two human-evaluation setups: *static*, in which we benchmark models on a fixed set of multi-turn contexts to generate responses (Section 2.2); and *interactive*, where we allow humans to chat freely with chatbots (Section 2.4). Lastly, we detail our automatic evaluation metric for fast development and end-to-end optimization (Section 2.7).

2.1 Measuring Human Likeness

To measure the quality of a response given a context, we propose a sequence of two questions. We first ask whether the response, given the context, makes sense. Sensibleness arguably covers some of the most basic aspects of conversational human-likeness, such as common sense and logical coherence. Sensibleness also captures other important aspects of a chatbot, such as *consistency*. The crowd worker is asked to use common sense to judge if a response is completely reasonable in context. If anything seems off — confusing, illogical, out of context, or factually wrong — then it should be labeled as, "does not make sense".

However, being sensible is not enough. A generic response (e.g., *I don't know*) can be sensible, but it is also boring and unspecific. Such responses are frequently generated by bots that are evaluated according to metrics like sensibleness alone (Li et al., 2016a; Venkatesh et al., 2018). To illustrate this, we create GenericBot: a trivial bot that always replies to questions with "I don't know" and to statements with "ok" (examples in Appendix Table 8). On *static* evaluation (using a fixed set of prompts and bot-generated responses), 70% of GenericBot's responses are labeled sensible, surpassing even DialoGPT (62%), even though DialoGPT is clearly more human-like than GenericBot. To overcome this issue, we need our evaluation to separate more fully human-like conversation from bland and generic statements. Therefore, if a response is labeled as sensible, we further ask the crowd worker to determine if it is specific to the given context. For example, if

A says, "I love tennis," and B responds, "That's nice," then the utterance should be marked, "not specific". That reply could be used in dozens of different contexts. However, if B responds, "Me too, I can't get enough of Roger Federer!" then it is marked as "specific", since it relates closely to what is being discussed. Responses labeled not sensible are considered not specific. In GenericBot's case, none of the responses are specific, whereas 39% of DialoGPT's responses are specific.

This sequence of two questions is designed to start with the most concrete and basic human quality (sensibleness) and then progress to the arguably more subjective human quality (specificity). The degree of subjectivity is somewhat quantified in the crowd worker agreement. We measure crowd worker consistency for every model benchmark using agreement and Krippendorff's alpha (Krippendorff, 2011), shown in Table 1. The agreement is reasonable considering the questions are subjective and the final results are always aggregated labels (e.g., average sensibleness across all chatbot responses).

Metric	Agreement (%)	Krippendorff's alpha
Sensibleness	76 ± 3	0.42 ± 0.03
Specificity	66 ± 2	0.30 ± 0.05

Table 1: The average and standard deviation of crowd worker agreement across static evaluations of Meena models. Each static evaluation consisted of 1,477 $(context, response)$ pairs, each labeled by 5 crowd workers.

Given a set of responses labeled as described above, we can calculate sensibleness and specificity as the percentage of responses labeled as sensible and specific, respectively. To combine these two into one metric, we take a simple average of the two, which we call SSA (**s**ensibleness and **s**pecificity **a**verage). SSA is a proxy for human likeness, which also penalizes chatbots that consistently produce generic responses. For example, GenericBot's SSA is 35% and DialoGPT's SSA is 51%, providing a much more fair separation and ranking than sensibleness alone.

Before arriving at SSA, and before any of the chatbots were tested, the authors of this paper conducted several rounds of pilot studies on what to ask crowd workers and how to best phrase the instructions. We settled on the two-question SSA
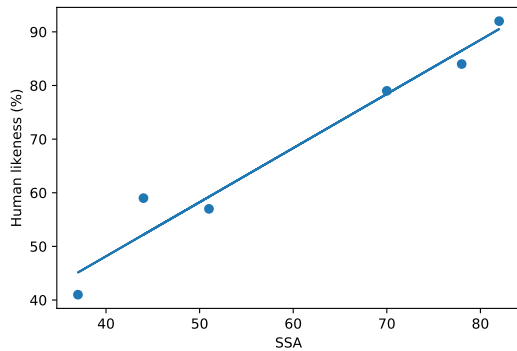
Figure 2: SSA vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.96. The SSA values were collected using static evaluation mode (Section 2.2). The human likeness evaluation was also conducted in static evaluation mode. Instead of judging sensibleness or specificity, however, we asked crowd workers to judge whether a given response was "human-like", or in other words, looked like a response that a human might give in the provided context.

for several reasons: it was easy for crowd workers to understand; alternative additional questions did not add extra information; and more subjective questions result in lower agreement between crowd workers.

As an additional check on the SSA metric, we reran a static evaluation, this time asking crowd workers to assess whether or not a response is "humanlike". We find that there is a high correlation between those labels and the two components of the SSA metric (Figures 2, 9, 10). Compared to a direct evaluation of what crowd workers consider to be "humanlike", SSA has significant advantages for large-scale evaluation tasks: it is more objective, easier for crowd workers to understand, and penalizes boring and vague responses. Nevertheless, these findings give us confidence that SSA is indeed capturing important aspects of human likeness.

2.2 Static Evaluation

In order to have a common benchmark to easily compare models, we create a collection of 1,477 conversational contexts with between 1 and 3 conversation turns, that we call the Mini-Turing Benchmark (MTB). We started this dataset by compiling single-turn contexts (e.g., "How are you?") from multiple sources, such as from the

work[4] of Vinyals and Le (2015) and the transcripts of the Loebner Prize[5] contests (years 2014-2018). In total, there were 315 single-turn contexts, which we then extended to include 500 two-turn and 662 three-turn contexts.

The MTB also contains contexts with personality questions (e.g. "Do you like cats?"), some of which expect responses with personality consistency. For example, the context "A: Do you like movies?; B: Yeah. I like sci-fi mostly; A: Really? Which is your favorite?" expects a consistent response such as *I love Back to the Future*. On the other hand, a response like *I don't like movies* would be a contradiction, and thus not considered sensible.

When evaluating chatbots, all MTB contexts are fed to the models or presented to humans to obtain responses. We send the resulting $(context, response)$ pairs to crowd workers and asked whether each response given the context is sensible and specific as defined in 2.1. We call this *static evaluation* because the contexts are fixed.

2.3 Interactive Evaluation

Static evaluation may be suitable for comparing models, but it is biased by how the static evaluation dataset was constructed. To address this, we create an additional evaluation mode where the crowd workers can chat 1:1 with a chatbot about anything they want. As with static evaluation, workers are also asked to decide whether each response from the chatbot is sensible and specific as defined in 2.1. Conversations start with "Hi!" from the chatbot to mark the beginning of the conversation and crowd workers have no expectation or instructions about domain or topic of the conversation. A conversation is required to last at least 14 turns (7 from chatbot) and at most 28 turns. We collected 100 such conversations for each model (i.e., at least 700 labeled turns per model). We then measure the percentage of labeled turns that are sensible and specific.

Unlike a typical Turing test (Turing, 1950), we tell the human judges upfront that they are about to chat with an experimental chatbot and ask them to label what the chatbot says in terms of sensibleness and specificity. This shifts the focus of the judges and chatbot creators from optimizing

for deception detection to optimizing for detecting and maximizing human-like qualities (e.g., sensibleness). Similar to our approach, Ghandeharioun et al. (2019) also conduct interactive evaluation by allowing humans to chat freely with bots. Their setup, however, focuses on evaluating conversations as a whole (as opposed to at the level of individual turns) and judges evaluate for quality, fluency, diversity, relatedness, and empathy.

2.4 Estimate of Human Performance

To estimate static SSA of humans we ask crowd workers to respond to MTB contexts. Additionally, to estimate human interactive SSA, we leveraged the help of internal company volunteers to collect 100 human-human conversations following mostly the same instructions as crowd workers for every other chatbot. Labeling of sensibleness and specificity was conducted by independent crowd workers with majority voting of 5 workers per human turn. The difference from the rest of the evaluations is that, in this case, participants knew they were chatting with another human. In contrast, when humans chat with a chatbot they will occasionally say unusual things to test the chatbot's limits. Hill et al. (2015) describe differences in human behavior when talking to a chatbot. That said, we never incentivize humans to chat adversarially with chatbots in any of our evaluations.

2.5 Evaluation of Cleverbot and DialoGPT

To integrate with Cleverbot, we leverage its API. For DialoGPT, we use its open sourced 762M parameter model.[6] It is worth mentioning that we initially tried the 345M parameter DialoGPT model, because it was reported to perform best on single-turn human evaluation. However, the 345M parameter model seemed to perform noticeably worse than the 762M one in preliminary evaluations of multi-turn conversations. Our human evaluation is multi-turn, so we select the 762M model.

The DialoGPT authors were unable to release their decoding script at the time of writing. Therefore, following their published description, we use top-K decoding with $K = 10$. We adapt the decoding implementation by Wolf et al. (2019). Moreover, since the backward model was also not released we were not able to try their MMI re-ranking (Li et al., 2016a).

Both Cleverbot and DialoGPT were evaluated using the same crowd sourcing setup as for Meena.

2.6 Evaluation of Mitsuku and XiaoIce

Because we chose to use the free Mitsuku web app[7], and there is no public API for XiaoIce, we called on the help of internal company volunteers and only conducted interactive evaluation. Volunteers collectively had 100 conversations with Mitsuku, and 119 with XiaoIce on their publicly available web apps. The volunteers conversed with the chatbots following mostly the same instructions that crowd workers follow for every other chatbot. The difference is that humans would say "Hi!" for the first turn, instead of the chatbot, in order to keep the first turn the same as other cases. Labeling of sensibleness and specificity in all cases was conducted by independent crowd workers with majority voting of 5 workers per chatbot turn. Also note that both XiaoIce and Mitsuku sometimes include an image in their reply and occasionally, volunteers include text descriptions of the images they see. The presence of the image may in some cases change the sensibleness of the response for better or worse.

XiaoIce interacts in Mandarin so both the volunteers and the independent crowd workers were native Mandarin speakers. The group of volunteers for XiaoIce, Mitsuku, and human-human conversations were mostly disjoint. Other than requiring a knowledge of Mandarin for XiaoIce conversations, volunteer selection was arbitrary. We had 29 volunteers for XiaoIce, 43 for Mitsuku, and 21 for human-human.

To reset Mitsuku state between conversations, volunteers refreshed the web page. During the writing of this paper there was no clear way to reset the state of XiaoIce. The XiaoIce team have informed us that not resetting the state negatively affects the model's control of the context.[8] Also, most XiaoIce volunteers shared the same Weibo account.[9] The XiaoIce team confirmed that account reuse negatively impacts the internal profile constructed by XiaoIce for a user. The XiaoIce team further suggested that, if the same Weibo account needs to be reused, we should wait at least

---

[6]https://github.com/microsoft/DialoGPT

[7]Pandorabots offers a paid enterprise package, which includes the Mitsuku API.

[8]From personal communication with the XiaoIce team, after the writing of the paper.

[9]Weibo is a microblogging service mostly used in China, which also allows users to chat with XiaoIce: https://www.weibo.com/

one hour between volunteers using the account. In our experiments, we may have sometimes waited less than that amount of time between volunteers, although we made sure the account was only used by one volunteer at a time. Finally, the XiaoIce team mentioned that in the past few months (as of this writing), a limited version of XiaoIce with the smallest index has been served on Weibo. This version is expected to produce less satisfactory responses.

Direct comparisons between XiaoIce and other chatbots come with a caveat: XiaoIce can be seen as a product that optimizes for long-term user engagement, of which dialog generation is just one component. In other words, Meena is arguably at an advantage when comparing SSA scores.

2.7   Automatic Evaluation

For quick research iterations, we focus on perplexity. Unlike the previous two evaluation types, perplexity is an automatic metric. A seq2seq model outputs a probability distribution over possible next response tokens. Perplexity measures how well the model predicts the test set data; in other words, how accurately it anticipates what people will say next. When interpreting perplexity scores, bear in mind that lower is better and that the theoretical minimum is one.

As shown in Section 4, this commonly used metric correlates with human judgement of sensibleness and specificity. This is encouraging, because it is both automatic and directly optimizable with the standard cross-entropy loss function.

3   Meena chatbot

As described above, recent work on end-to-end dialog models has fallen into two broad categories: (1) complex models with human-designed components, and (2) large neural network models (known as end-to-end models) that are closer to generic learning frameworks. End-to-end models have shown promise, but clear limitations (Gao et al., 2019a). An open question has been: in order to reach a point where a model can carry out high-quality, multi-turn conversations with humans, could we simply take an end-to-end model and make it bigger—by adding more training data and increasing its parameter count—or is it necessary to combine such a model with other components? In this section we describe the Meena model, the largest end-to-end model to enter the field so far. We believe it answers the open research question, by showing that a large end-to-end model can generate almost humanlike chat responses in an open-domain setting.

In this section, we will describe the training data, architecture, and decoding algorithm. We will also provide a few sample conversations that Meena has had with humans.

3.1   Training Data

The dataset used to train Meena is mined and filtered from public domain social media conversations. The source data are essentially message trees involving multiple speakers: the very first message is the root; replies to a message are its child nodes. Any path along the tree induces a conversation where each message is a conversation turn. By treating each turn in a conversation path as a response and all the previous turns (up to 7) as a context, we create a training example of the form *(context, response)* pair.

We also filter the data to improve the generation quality. A message is removed if any of the following conditions holds: 1. the number of subwords is less than 2 or more than 128; 2. the percentage of alphabetic characters is less than 70%; 3. message contains URL; 4. author's username contains "bot"; 5. the message is repeated more than 100 times; 6. the message has a high n-gram overlap with the parent's text; 7. the message is potentially unsafe or offensive with respect to a commercial text classifier. In addition, we remove copies of the parent's text quoted in a message.

For simplicity, when a message is removed, we drop all sub-trees rooted under it. After these filtering steps, the number of $(context, response)$ pairs extracted is 867M. The text is tokenized using byte-pair-encoding (BPE) (Sennrich et al., 2016) with the sentencepiece library.[10] We use a vocabulary of 8K BPE subwords, which we found in our early experiments to be sufficient for generating specific responses while still allowing us to fit larger models in memory.

The final Meena dataset contains 341GB of text (40B words). In comparison, GPT-2 (Radford et al., 2019) has been trained on 40GB of Internet text (8 million web pages).

---

[10] https://github.com/google/sentencepiece

Responses for "Why do you like the ocean?"	Score
1. I love the beach. I like that it doesn't have anything in it. Just open water.	-12.98
2. It's dark and mysterious. I discovered it when I was really young and it was a dream of mine since. The ocean is full of wonders, mysteries and great things to discover. I love snorkeling. I moved to Hawaii a year ago and I'm going to college for marine biology. So there's that.	-14.03
3. The sea sounds cool and it's really satisfying to swim in.	-14.62
4. I dunno. It's beautiful, I guess. It reflects my upbringing, I reckon.	-14.87
5. Well I'm from the pacific northwest and always somehow end up back there. I've been to Hawaii but I'm not one for the beach, it's too hot and the air is so dry.	-15.36
6. The beauty of it is incredible, and swimming in it and being able to see huge waves sweeping by is just amazing. I was born in the desert so it's really the only place in the world I've seen with flowing water around me	-17.07
7. It's beautiful, maybe? I just wish I lived closer to it!	-17.14
8. It's pretty and soothing.	-17.19
9. Because I love the smell. I love walking when I can through the beach. I live in California so I basically go to the ocean almost every weekend.	-17.38
10. The ocean is pretty cool, I suppose. I like the various types of fish.	-17.90

Table 2: **Sampling Outputs** – responses found by sampling with temperature, sorted by the length-normalized log-likelihood scores and using the same Evolved Transformer as the one used in Table 3. The scores are calculated as $\frac{\log P}{T}$, where P is the likelihood of the response and T is the number of tokens.

Responses for "Why do you like the ocean?"	Score
1. Why don't you like the ocean?	-1.70
2. I don't like the ocean. I love the ocean.	-2.66
3. I don't like the ocean. I love it.	-2.78
4. I don't like the ocean, I love the ocean.	-2.94
5. I don't like the ocean, I love it.	-2.94
6. Why don't you like the ocean? :P	-2.95
7. I don't like the ocean, I love it!	-3.15
8. I don't like the ocean. I love the ocean!	-3.20
9. Why don't you like the ocean? It's beautiful.	-3.26
10. I don't like the ocean. I love the ocean. There's a difference.	-3.31

Table 3: **Beam Search Outputs** – top responses generated by beam-search decoding and the corresponding length-normalized log-likelihood scores. We use an Evolved Transformer with perplexity 10.2 and vocabulary size of 8K.

3.2 Model Architecture

The best performing Meena model is an Evolved Transformer (ET) (So et al., 2019) seq2seq model with 2.6B parameters, which includes 1 ET encoder block and 13 ET decoder blocks. The Evolved Transformer is an evolutionary NAS architecture (Real et al., 2017, 2018) based on the Transformer (Vaswani et al., 2017). Our largest (i.e., maximum memory usage) Evolved Transformer scored 10.2 perplexity and our largest vanilla Transformer scored perplexity 10.7 for the same number of training steps (738k). The largest vanilla Transformer had 32 decoder layers with other architectural hyperparameters held con-

stant.[11]

For comparison, the extra-large GPT-2 model (Radford et al., 2019) has 1.5B parameters and is a language model (i.e., decoder only); whereas the large conversational model from the recent DialoGPT work (Zhang et al., 2019) has 762M parameters.

Meena's hidden size is 2,560 and the number of attention heads is 32. We share the embeddings across the encoder, the decoder, and the softmax layer. The encoder and decoder each have a maximum length of 128 tokens (i.e., 256 combined). The hyperparameters of our best model were found via manual coordinate-descent search.

3.3 Training Details

We trained our best model for 30 days on a TPU-v3 Pod (2,048 TPU cores) on the Meena dataset containing 40B words (or 61B BPE tokens). Interestingly, the 2.6B-parameter model can overfit [12] on a 61B-token dataset which suggests a surprisingly large model capacity. Therefore, we add a small amount of 0.1 attention and feed-forward layer dropout. Additionally, to save memory, we chose the Adafactor optimizer (Shazeer and Stern, 2018) with 0.01 as the initial learning rate, keeping it constant for the first 10k steps and then decaying with the inverse square root of the number of steps. We use the Tensor2Tensor code-

---

[11] An Evolved Transformer block is about twice as deep as a Transformer layer

[12] In the sense that validation loss increases as train loss decreases.

base (Vaswani et al., 2018) for training Meena.[13]

A TPU-v3 core has 16GB of high-bandwidth memory. We maximized memory usage for model parameters and stored only 8 training examples per core. Each training step took about 1 second. In the full TPU-v3 Pod, this meant we learned over 4M tokens per training second. Therefore, by the end of training, the model had traversed the full training set 164 times (or epochs) and observed a total of about 10T tokens (including repeated ones).

3.4 Decoding

Generating generic (i.e., not specific) and bland responses (Li et al., 2016a) has always been a major challenge in existing neural conversational models. A common approach to mitigating this problem is to use more sophisticated decoding algorithms, for instance with different forms of re-ranking (Li et al., 2016a; Shao et al., 2017) or conditioning on profiles, topics, and styles (Li et al., 2016b; Wang et al., 2017; Xing et al., 2017; Zhang et al., 2018b). Recent works also explore new frameworks such as adversarial learning (Li et al., 2017; Zhang et al., 2018c), variational autoencoding (Zhao et al., 2017; Gu et al., 2019), or both (Gao et al., 2019b) at the cost of added complexity and less scalability.

In contrast, we show that given a model with sufficiently low perplexity, a simple sample-and-rank decoding strategy achieves both diverse and high-quality responses. Sample-and-rank, works as follows: First, we sample N independent candidate responses using plain random sampling with temperature T. Second, we select the candidate response with the highest probability to use as the final output.

Temperature $T > 0$ is a hyper-parameter that regulates the probability distribution p_i of the next token during decoding. We divide the logits z_i by T before computing the "softmax" as in Hinton et al. (2015):

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

$T = 1$ yields the unmodified distribution. We observe that large values of T favor contextually rare tokens, such as relevant entity names, but might also assign too much probability to incorrect tokens depending on the model's predictions. Meanwhile, smaller values of T favor more common words such as articles or prepositions, which are safer but less specific.

Tables 2 and 3 show responses for the arbitrary probing input "Why do you like the ocean?" under sample-and-rank and beam-search, respectively. As we can see, beam-search decoding generates repetitive and uninteresting responses. On the other hand, sample-and-rank provides us with diverse and content-rich responses. The key here is to have a model with low perplexity so samples can be taken at high temperature to produce human-like content.

For all the results in Section 4, we use sample-and-rank with $N = 20$ and $T = 0.88$. Additionally, as shown in Figure 1, for this fixed decoding strategy, sensibleness and specificity improve as model test set perplexity falls. For additional decoding results and samples, see Section 5.

3.5 Sample conversations

Below are cherry picked conversations that Meena has had with humans. We selected these conversations *after* they were completed. That is, the Meena responses within the conversations were not cherry picked; they were produced automatically using sample-and-rank (Section 3.4). Conversations B and C are excerpts from conversations number 43 and 48, respectively, of the conversations dataset published on GitHub.[14]

Conversation A

Human:	*say you're in an empty giant office and you see a black bird flying indoors. what do you do?*
Meena:	*What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*
Human:	*well it's a predator bird and it's gonna eat you*
Meena:	*I would grab my phone and call 911.*

Meena handles arbitrary open-domain input. However, asking "Is it indoors or outdoors" was not sensible.

Conversation B	
Human:	*Hi!*
Meena:	*Hi! How are you this fine evening?*
Human:	*I'm doing well. it's morning my time!*
Meena:	*Good morning! What are your plans for today?*
Human:	*Get my work done then watch a movie at home in the night*
Meena:	*Sounds fun! What movie?*
Human:	*I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*
Meena:	*I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*
Human:	*ah, tell me about that movie!*
Meena:	*It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.*
Human:	*ha! I've searched on the internet and both movies are very colorful!*
Meena:	*I'm looking forward to it. I love colorful movies.*

The Grand Budapest Hotel and Moonrise Kingdom are indeed colorful movies, directed by Wes Anderson.

Conversation C	
Human:	*Haha, I think mathematics has a presence beyond human existence.*
Meena:	*So does philosophy.*
Human:	*Can you state with examples how this is true.*
Meena:	*Can you state how it isn't?*
Human:	*Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy wi [sic]*
Meena:	*That's not the definition of philosophy.*
Human:	*Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy will not be needed. How do you define philosophy?*
Meena:	*The study of general knowledge or the fundamental nature of things.*

Meena tries to define philosophy.

Conversation D	
Human:	*I like how cute cows are*
Meena:	*Ah, interesting. You like how they are cute, right?*
Human:	*yeah, and they're realy smart*
Meena:	*I heard they go to college*
Human:	*cows go to college?*
Meena:	*I heard that a cow went to Harvard.*
Human:	*what did the cow study?*
Meena:	*Bovine sciences.*
Human:	*do horses go to Harvard?*
Meena:	*Horses go to Hayvard.*
Human:	*that's a pretty good joke, I feel like you led me into it*
Meena:	*You were trying to steer it elsewhere, I can see it.*

Meena executes a multi-turn joke in an open-domain setting. We were unable to find this in the data.

Appendix A shows random samples of conversations.

4 Results

In this section, we will first demonstrate the correlation between test perplexity and the human evaluation metric, SSA, defined earlier. We also include human-level upperbound estimates for both static and interactive evaluations, beside performances of other chatbots, such as XiaoIce, Mitsuku, DialoGPT, and Cleverbot. Lastly, we provide sample responses for different models given the same contexts to understand how Meena qualitatively compares to others.
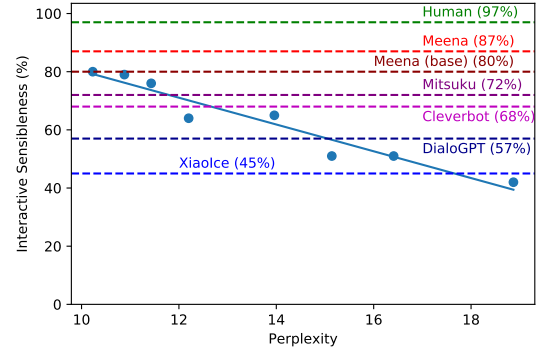
Figure 3: Interactive sensibleness vs perplexity.
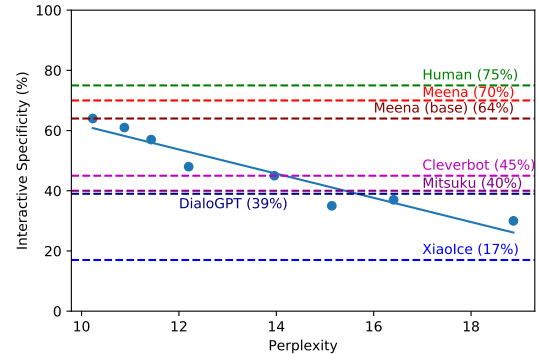
Figure 4: Interactive specificity vs perplexity.

4.1 SSA-perplexity correlation

We trained models with different hyper-parameter settings and architectures on the dataset described in Section 3.1. We vary the number of layers, attention heads, total training steps, whether we use Evolved Transformer or regular Transformer and whether we train with hard labels or soft labels/distillation (Hinton et al., 2015). The trained models are then measured with an automatic met-

ric, test perplexity (Section 2.7), and also with human metrics (Sections 2.2 and 2.3). Our results indicate most of the variance in the human metrics can be explained by the test perplexity. The end-to-end trained Meena model with lowest perplexity is referred to as *Meena (base)*. In addition, we also include an improved version of Meena (detailed in Section 5) and refer to this as the *Meena (full)* model, or just *Meena* model for short.

The correlation was $R^2 = 0.93$ for static sensibleness vs perplexity and $R^2 = 0.94$ for static specificity vs perplexity indicating this might be a good automatic metric for measuring sensibleness and specificity. Static SSA vs perplexity has $R^2 = 0.94$. The static evaluation results are shown in Figure 5. The correlation is close to linear, but it is unclear whether the trend will continue for even lower values of perplexity.

In interactive evaluation (Section 2.3) crowd workers could chat about anything they wanted. We observe similarly strong correlation with perplexity (see Figures 1, 3 and 4) and very similar sensibleness and specificity values as the static evaluation. This indicates that the static evaluation correlation with perplexity is not due to dataset bias.

Regarding consistency, the lowest perplexity model was evaluated 7 times with static evaluations and also 7 times with interactive evaluations. Each time, we obtained a different set of randomly sampled responses. Across the evaluations the standard deviation is 2% for static SSA and is 1% for interactive SSA, indicating that both metrics are consistent enough for our purposes.

4.2 Human-level Estimates

As expected, human sensibleness is very high, but it is not perfect. Human sensibleness was estimated at 94% static and 97% interactive. People have misunderstandings, miss attempts at humor and sometimes lack shared context or background. Also aligned with intuition, humans are sometimes not specific due to momentary lack of ideas, interest or knowledge. The human specificity scores are 69% static and 75% interactive. The resulting SSAs are 82% static and 86% interactive.

4.3 XiaoIce, Mitsuku, DialoGPT and Cleverbot

Crowd workers labeled 1,173 XiaoIce turns within their original conversation context. Per these la-

bels, XiaoIce scores 31% interactive SSA which is comprised of 45% sensibleness and 17% specificity. We used majority voting of 5 workers per chatbot response. Agreement between workers was 77% for sensibleness and 81% for specificity and Krippendorff's alpha was 0.54 for sensibleness and 0.40 for specificity (which indicates fairly strong agreement). For further verification of the results, we also had a group of 4 internal company volunteers that are native Mandarin speakers to label a subset of 25 conversations (247 chatbot turns). The volunteers did not know the crowd worker results. The volunteer based estimate is 36% interactive SSA with 53% sensibleness and 19% specificity.

Volunteers report that XiaoIce responses display empathy and personality. However, the responses do not always make sense and are often generic. The generic responses exuded a fun personality, but were not necessarily related to the ongoing conversation. As mentioned above, XiaoIce is reported to be optimized for long term engagement (Zhou et al., 2018). It is also designed with a multifaceted notion of 'humanlikeness' in mind, emphasising interpersonal qualities such as empathy. As we discuss in Section 7, a broader metric of humanlikeness would be a useful topic for future research.

See et al. (2019) report engagingness is not the same as humanness. In that work, they control and optimize certain attributes of response generation. One of the results is that their chatbots achieve near-human-level engagingness, but simultaneously are not very close to human-level humanness. This along with limitations described in Section 2.6 may help explain why XiaoIce's sensibleness and specificity scores seem relatively low in our evaluation.

Mitsuku scores 56% interactive SSA with 72% sensibleness and 40% specificity over 899 chatbot turns. Mitsuku has won the Loebner prize 5 times,[15] however, we learned from the Pandorabots team that the website version we tested with was not the one that competed in the Loebner prize. Optimizing human likeness for the competition might correlate more with our metrics.

Using crowd workers, we executed interactive evaluation of DialoGPT over 300 conversations (2,660 chatbot turns labeled). DialoGPT scores

---

[15] https://en.wikipedia.org/w/index.php?title=Loebner_Prize&oldid=927503336
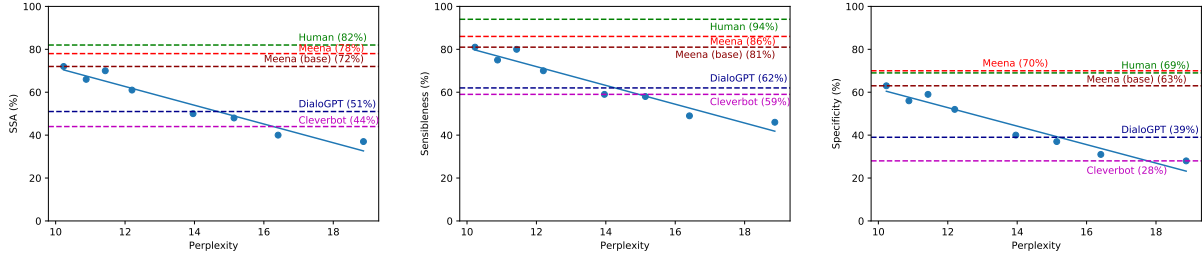
Figure 5: Static evaluation results.

Sample Contexts	Meena (base)	Cleverbot	DialoGPT
A: You must have read a lot of books *B:* Oh wow, how do you know? Do I look nerdy? *A:* Of course, you're not!	*B:* So how do you know? [*sensible: 1, specific: 1*]	*B:* Never. said I was. [*sensible: 0, specific: 0*]	*B:* Oh thank goodness. [*sensible: 1, specific: 1*]
A: You know nothing, Jon Snow! *B:* You must be a fan of Game of Thrones *A:* Haha yea, how can I not?	*B:* Who's your favorite character? [*sensible: 1, specific: 1*]	*B:* What do you mean? [*sensible: 0, specific: 0*]	*B:* You can't. But that's not why you shouldn't know. [*sensible: 0, specific: 0*]
A: What is your favorite island in the world? *B:* I would say Fiji *A:* Interesting, what's nice about it?	*B:* Just a beautiful place [*sensible: 1, specific: 0*]	*B:* The title is self explanatory. [*sensible: 0, specific: 0*]	*B:* You can see some great scenery. [*sensible: 1, specific: 0*]
A: You must have read a lot of books *B:* Oh wow, how do you know? Do I look nerdy?	*A:* I know because I've read a lot of books. [*sensible: 1, specific: 1*]	*A:* Observation. [*sensible: 1, specific: 0*]	*A:* I know because you look like your mom. [*sensible: 1, specific: 1*]
A: You know nothing, Jon Snow! *B:* You must be a fan of Game of Thrones	*A:* Yeah, I'm reading the books at the moment. [*sensible: 1, specific: 1*]	*A:* No. [*sensible: 1, specific: 0*]	*A:* I've never really been interested in it... [*sensible: 1, specific: 1*]
A: What is your favorite island in the world? *B:* I would say Fiji	*A:* Fiji, yes! I've been there once. It's the most beautiful place I've ever been. It definitely makes me want to move there. I hope you get to visit again soon! [*sensible: 1, specific: 1*]	*A:* I would have guessed 666. [*sensible: 0, specific: 0*]	*A:* Fiji is awesome! But I would say Fiji and Oceania are more popular at the moment. [*sensible: 0, specific: 0*]
A: You must have read a lot of books	*B:* Reading's for nerds [*sensible: 1, specific: 1*]	*B:* What books have you read. [*sensible: 1, specific: 1*]	*B:* I read a lot of books but have never read a single book about the history of our planet. [*sensible: 1, specific: 1*]
A: You know nothing, Jon Snow!	*B:* Thank you for this. I was hoping I wasn't the only one getting an ASOIAF vibe from this. [*sensible: 1, specific: 1*]	*B:* Am not jon snow. [*sensible: 1, specific: 0*]	*B:* The North Remembers... [*sensible: 1, specific: 1*]
A: What is your favorite island in the world?	*B:* That's a tough one. I'd have to say Hawaii. [*sensible: 1, specific: 1*]	*B:* Daniel I like frites. [*sensible: 0, specific: 0*]	*B:* Probably the one on the north island [*sensible: 0, specific: 0*]

Table 4: **Sample responses from static evaluation** – Shown are responses of Meena (base), Cleverbot, and DialoGPT on some sample contexts together with their binary labels per category (sensibleness and specificity). Note that the contexts were constructed in advance before we generate all model responses. These contexts are not in the MTB dataset.

48% ± 3% interactive SSA with 57% ± 2% sensibleness and 39% ± 4% specificity. It obtained similar results in static evaluation. In its paper, DialoGPT attains performance close to human in a single-turn setting. On the other hand, our human evaluation is multi-turn, which is expected to be more challenging. Additionally, DialoGPT scores poorly on specificity, and our impression from browsing transcripts is that it prefers briefer and more generic responses. This might be because the model is optimized for classic Turing-test evaluation, in which overly chatty responses increase the risk of making a mistake. These results and conjectures come with the caveat, as described above, that we wrote our own decoder for this model since the public DialoGPT codebase does not yet have one.

Cleverbot, unlike Meena and DialoGPT, performs notably better on interactive rather than static evaluation. It scores interactive SSA 56% and static SSA 44%. Interactive specificity, 45%, is especially higher than its static counterpart, 28%. Upon closer inspection of the data, we hypothesize that: (1) in the interactive setting, Cleverbot has opportunities to steer the conversation towards topics that it is more familiar with; (2) the minimum interactive conversation length of 14 turns makes it possible for a significant portion of these turns to be greetings and goodbyes, which both Cleverbot and Mitsuku are consistent in appropriately responding to. Furthermore, the interactive SSA scores for Mitsuku and Cleverbot are the same, 56% when averaging sensibleness and specificity before rounding. Mitsuku scores higher sensibleness (72% versus 68%), but lower specificity (40% versus 45%). It seems that relative to Mitsuku, Cleverbot replies more often in ways that are borderline nonsensical and lack consistent personality. Finally, we remark that the standard deviation of the Cleverbot interactive SSA is ±1% across two interactive evaluation sessions.[16]

4.4 Sample Responses: Meena (base), Cleverbot, and DialoGPT

To understand how Meena qualitatively compares to other models, we show in Table 4 sample responses from Meena (base), Cleverbot, and DialoGPT under the same set of contexts (which

---

[16]Due to technical issues when calling the Cleverbot API we only collected 195 interactive conversations (1,751 chatbot turns labeled) instead of the 300 conversations which we collected for DialoGPT.

were constructed before we generate all model responses). For 1- and 2-turn contexts, responses from Meena base are all sensible and specific. In addition, Meena (base) generates rich and interesting responses, e.g., the mention of *"ASOIAF vibe"* to refer to "A Song of Ice and Fire" in the famous Game of Thrones series or the remark about Fiji island being *"the most beautiful place I've ever been"*.

In contrast, Cleverbot can generate sensible responses for some contexts, but they are not always specific, e.g., Cleverbot replied with *"Observation"* and *"No"*. DialoGPT is more specific and can also generate interesting responses, e.g., *"The North Remembers ..."'*. However, it does not make sense at times, e.g., in-turn contradiction in this response *"Fiji is awesome! But I would say Fiji and Oceania are more popular ..."* or vague answer *"Probably the one on the north island"*.

When it comes to longer (3-turn) contexts in Table 4, Meena (base) continues to generate high-quality responses, whereas none of Cleverbot's responses are sensible. DialoGPT is more sensible and specific than Cleverbot, but less so than Meena (base).

5 Further Advancing SSA

In this section we take the interactive SSA from 72% ± 1%, for Meena (base), to 79% ± 1%, for Meena (full), by further tuning our decoding strategy and adding a rule to detect cross turn repetitions.

5.1 Advancing Decoding

We evaluate both temperature T and top-k to mitigate negative effects from the tail of the distribution (Holtzman et al., 2019). We chose top-k ($k = 40$) and $T = 1.0$ following Fan et al. (2018); Radford et al. (2019); Keskar et al. (2019); Ippolito et al. (2019a). With this setting and maintaining $N = 20$, we note an SSA increase from 72% to 74% relative to sampling from the whole vocabulary with $T = 0.88$. This result is the same for both the interactive and the static evaluation.

We also swept the number of samples in sample-and-rank, evaluating $N \in \{1, 20, 400\}$. The results show that $N = 20$ provides a significant improvement over $N = 1$, with an absolute improvement in SSA of $\sim 10\%$ (Figure 6). However, $N = 400$ demonstrates worse performance for sensibleness (Figure 7 in the appendix)

and diminishing returns over $N = 20$ for specificity (Figure 8 in the appendix). The significant improvement from sample-and-rank with $N = 20$ motivates future work exploring alternate ranking functions and tuning parameters.
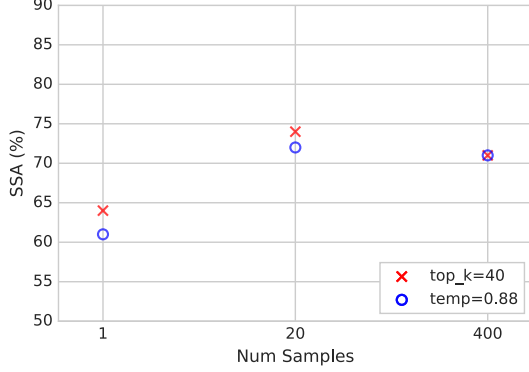

Figure 6: Static SSA over number of sampled responses for top-k and sampling with temperature.

5.2 Addressing Cross-turn Repetitions

In interactive evaluation, about one third of the conversations with Meena (base) contain cross-turn repetitions toward the end. Cross-turn repetition means that one turn somewhat repeats an earlier turn. For illustration, we cherry picked particularly problematic examples of cross-turn repetition shown in Tables 5 and 6.

It is worth mentioning that there also exist in-turn contradictions and repetitions, where the contradiction or repetition is contained in the response turn itself (e.g., *"I like pizza, but I don't like it"*). This type of artifact is often observed in Meena versions with worse perplexities, but is far less frequent in the Meena (base), which has the lowest perplexity as reflected in the samples shared in the appendix and the higher sensibleness scores.

We wrote a rule that detects if any two turns contain long common sub-sequences. We automatically remove candidates that are detected as repetition. This rule seems to have addressed most of the cross-turn repetition. We, therefore, further improve on the above interactive SSA of $74\%\pm\%1$ to $79\% \pm 1\%$.

5.3 Safety Layer

It is important to mention that the evaluation and conversation collection for the full Meena version was conducted with an additional classifier layer at serving time as part of the filtering mechanism

Cross-turn Repetition Example 1	
Human:	*Hi!*
Meena:	*Hi! How are you?! :)*
Human:	*doing good, what would be the best city to visit in Europe?*
Meena:	*Paris, Barcelona, Amsterdam, Prague*
Human:	*great list! why?*
Meena:	*Paris is a beautiful city, and Barcelona is too. And I've always wanted to go to Amsterdam and Prague and have never been.*
Human:	*have you even been to India?*
Meena:	*No, but I'd love to go.*
Human:	*how about Japan?*
Meena:	*I'd love to go to Japan too.*
Human:	*how about Paris?*
Meena:	*I'd love to go to Paris too.*
Human:	*are you lying to me?*
Meena:	*I'd love to go to Japan too.*
Human:	*i think you're lying to me, that makes me sad*
Meena:	*I'd love to go to Japan, too.*
Human:	*you're blocking me out! you have to talk to me*
Meena:	*I'd love to go to Japan.*
Human:	*where else?*
Meena:	*I'd love to go to Japan, too!*

Table 5: Multiple near identical repetitions of *"I'd love to go to Japan too."*.

Cross-turn Repetition Example 2	
Human:	*Hi!*
Meena:	*Hey there! How are you doing?*
Human:	*not well, i'm quite depressed*
Meena:	*That's no good. Wanna talk about it?*
Human:	*not really, it'll only make it worse...can you cheer me up?*
Meena:	*Sure! What's got you down?*
Human:	*I'd love to hear a really lame joke*
Meena:	*How about this one? What did the farmer say when he lost his tractor? Where is my tractor?*
Human:	*haha that was lame! where did you hear it?*
Meena:	*I was at the mall and I heard it in the food court. How was your day?*
Human:	*wow, i didn't know you could walk!*
Meena:	*Why did the scarecrow win an award?*
Human:	*i don't know..you tell me!*
Meena:	*Because he was outstanding in his field!*
Human:	*haha, where did you hear that?*
Meena:	*I was in the food court and I heard it in the field!*
Human:	*you have a great sense of humor!*
Meena:	*Why did the chicken cross the road? He was outstanding in his field!*
Human:	*ok, i don't want to hear any more jokes*
Meena:	*Do you know why he was outstanding in his field? Because he was a chicken!*
Human:	*haha looks like you're really into this discussion!*
Meena:	*Yes! I am! Do you know why? Because he was outstanding in his field!*
Human:	*haha, borderline creepy but still funny :D*
Meena:	*Why was the mouse trapped in the bag? Because he was outstanding in his field!*
Human:	*i think i broke you*
Meena:	*Do you know why I broke you? Because you were outstanding in your field!*

Table 6: Multiple partial repetitions of *"(...) Because he was outstanding in his field!"*.

to automatically filter out potentially sensitive or toxic response candidates for publication.

6 Related Work

Finding a good automatic metric that correlates with human evaluation has been an important goal of open-domain conversational modeling. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or other related metrics in translation and summarization, while popular and easy to compute, have been shown to be unsuitable for dialog (Liu et al., 2016) or more broadly language generation systems (Novikova et al., 2017).

Past works have attempted to build learnable metrics, either in a supervised fashion (Lowe et al., 2017), which requires human labels, or with unsupervised approaches (Tao et al., 2017; Ghazarian et al., 2019), that are more complex and need separate training, e.g., of a ranking system. In our work, we show that perplexity, which is readily available to any neural seq2seq model, exhibits a strong correlation with human evaluation. Our work is therefore also related to past attempts to correlate perplexity with other automatic metrics in other tasks, e.g., perplexity vs. BLEU in translation (Luong et al., 2015).

Another interesting line of work is to combine human evaluation with either automatic metrics (Chaganty et al., 2018) or with model likelihood (Hashimoto et al., 2019). While theoretically motivated, these metrics are too complex to be practical, requiring both human judgments and training separate models, e.g., an estimator (Chaganty et al., 2018) to reduce bias in automatic evaluation or a discriminator (Hashimoto et al., 2019) to distinguish between human- and model-generated samples.

In terms of designing of human evaluation metrics, existing literature differs in what attributes are used to assess the quality of a neural conversational model. Many works, e.g., Zhao et al. (2017); Xu et al. (2018); Ippolito et al. (2019b), have focused solely on the diversity aspect to counter the commonly observed problem of models generating generic responses (Li et al., 2016a). Others have attempted to improve and evaluate multiple aspects at once. For example, Venkatesh et al. (2018) aim to unify many metrics, such as diversity, engagement, and user experience; Gao et al. (2019b) jointly optimize for both diversity and relevance; See et al. (2019) control decoding

attributes (such as repetition, specificity, response-relatedness, and question-asking) to improve engagingness and interestingness; and Hashimoto et al. (2019) design metrics to capture human likeness and diversity.

In contrast, we focus on sensibleness and specificity for our human evaluation. While human likeness and relevance used in aforementioned works are related to sensibleness, we specifically use sensibleness as it leads to better agreement among crowd workers (see §2.1). Similar reasoning applies to specificity, which is related to other attributes such as engagingness and interestingness, as measured in previous works.[17] A limitation of our work is that it does not cover aspects such as empathy (Zhou et al., 2018; Rashkin et al., 2018).

While we do not explicitly control for specificity, existing works, such as (Zhang et al., 2018a; Ko et al., 2019), attempted to do so by augmenting the decoder of seq2seq models with specificity-control components. These added complexities sometimes lead to implausible responses as analyzed by Ko et al. (2019). In contrast, the specificity of our model improves as perplexity decreases.

Recent work on DialoGPT (Zhang et al., 2019) compares the conversation quality of chatbots with that of humans but their evaluation settings are limited to single-turn dialogs. We instead conduct our evaluation on conversations of up to 3 turns in the static MTB benchmark and 14 turns in the interactive setup.

7 Discussion

Our results suggest perplexity on public domain social media conversations might be a good automatic proxy for human judgement of fundamental attributes of human-likeness, such as sensibleness and specificity. The results also suggests that optimizing the probability of the next token on larger volumes of social media conversations could lead to human-like sensibleness in an open-domain setting. However, our static evaluation dataset only contains one to three-turn contexts and is biased

---

[17]It is worth pointing out that we do not explicitly measure diversity as it requires judging a set of responses; whereas, for conversation, what is most important is the first reply that a chatbot produces. As our decoding method is sampling, it implies that our generation is diverse. However, there remains a question of whether the sampled response is of high quality. The fact that our model has low perplexity and achieves high SSA score indicates that the generation is meaningful.

by the sources of the first turn and the fact that the two-turn and three-turn contexts build on the shorter contexts. Moreover the contexts in this dataset are predominantly Turing test and social conversation style, including common sense, basic knowledge, asking/sharing about personality, likes/dislikes, opinions, feelings, hobbies, pleasantries, etc. This dataset does not include contexts like deeper question answering (e.g., how fast is a cheetah), basic math (e.g., how much is 1+1) and common sense tests designed to challenge machines, but not humans (Levesque et al., 2011). Human-likeness is an incredibly broad and abstract concept. The interactive evaluation addresses some of the bias and scope limitations in static evaluation while still providing a consistent score to quantify a given chatbot. Nevertheless, unlike static evaluation it does not allow for granular comparison between different chatbot responses. In addition, it may be too short (14 to 28 turns), and may assign too much weight to typical beginning and ending of conversations. It may also be too short to cover deeper topics and exercise longer term memory.

Furthermore, it may be necessary to expand the set of basic human-like conversation attributes being measured beyond sensibleness and specificity. Some directions could include humor, empathy, deep reasoning, question answering and knowledge discussion skills. One could also break down sensibleness into its implicit subcomponents: logical and personality consistency, common sense, relevance, basic factual correctness and so on. Future work may also explore the continued optimization of sensibleness via the optimization of test set perplexity.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *ACL*.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Runxiang Cheng, Ashwin Bhandare, and Zhou Yu. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. In *Alexa Prize 2018*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. *arXiv e-prints*, page arXiv:1805.04833.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019a. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Jointly optimizing diversity and relevance in neural response generation. In *NAACL*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, pages 13658–13669.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *CoRR*, abs/1904.10635.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder. In *ICLR*.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *NAACL-HLT*.

Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior*, 49:245–250.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019a. Human and automatic detection of generated text. *ArXiv*, abs/1911.00650.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019b. Comparison of diverse decoding methods from conditional language models. In *ACL*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. In *NAACL*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. `https://repository.upenn.edu/asc_papers/43`.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *KR*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL workshop on Text Summarization Branches Out*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *ACL*.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *ACL*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. *CoRR*, abs/1811.00207.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2018. Regularized evolution for image classifier architecture search. In *AAAI*.

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Quoc V. Le, and Alex Kurakin. 2017. Large-scale evolution of image classifiers. In *ICML*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *ACL*.

Iulian Vlad Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A deep reinforcement learning chatbot. *CoRR*, abs/1709.02349.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. *ICML*.

David R. So, Chen Liang, and Quoc V. Le. 2019. The evolved transformer. In *ICML*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeuRIPS*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. *CoRR*, abs/1701.03079.

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

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeuRIPS*.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On evaluating and comparing conversational agents. *CoRR*, abs/1801.03625.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *EMNLP*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Steve Worswick. 2018. Mitsuku wins Loebner Prize 2018! `https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7`. [Online; written on September 13, 2018].

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. In *NAACL*.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018a. Learning to control the specificity in neural response generation. In *ACL*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018c. Generating informative and diverse conversational responses via adversarial information maximization. In *NeuRIPS*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *CoRR*, abs/1812.08989.

A Additional Sample Conversations

With the help of many internal company volunteers we collected a total of about 100 conversations with Mitsuku, XiaoIce and Meena (full). The conversations are available on Github [18]. This section contains samples obtained by random shuffling these sets, and taking the first 10. Conversations were collected following the standard instructions for interactive evaluation where the human starts. Therefore, conversations are supposed to start with "Hi!", contain between 16 and 32 turns in total, and are open-domain with no particular topic. Nevertheless, some participants did not follow the first-turn rule strictly, so some conversations may start with for instance "hi there" instead of "Hi!". Also, a few conversations are under or over the length limits.

Unlike in Section 3.5, which contains cherry picked samples, we present random samples of everything that was collected after a few potentially sensitive conversations have been removed from the original sets. We also redacted potential personally identifiable information and indicated that with the word "REDACTED". Finally, please note that both XiaoIce and Mitsuku sometimes include an image in their reply and occasionally, volunteers include text descriptions of the images they see.

A.1 Meena

The following are a sample of the conversations with the Meena (full) (79%$\pm$1% interactive SSA).

Meena Conversation 1

Human:	*Hi!*
Meena:	*Hey there! What's up?*
Human:	*Same old, Monday blues...what are you up to?*
Meena:	*I'm procrastinating on a paper I'm supposed to be writing.*
Human:	*are you thinking of pulling an all nighter?*
Meena:	*Possibly. What's your paper on?*
Human:	*Nah, i just do the grunge work around here...one day I wish to write a paper that will change the world!*
Meena:	*What's the topic?*
Human:	*not sure, any suggestions?*
Meena:	*Well what's the assignment on?*
Human:	*breathing under water*
Meena:	*How to breathe underwater.*
Human:	*that's what i'm hoping to figure out!*
Meena:	*I hope you figure it out! Because I can't.*
Human:	*i'm sure you can if you try :) you're the best!*
Meena:	*Thank you, but I'm pretty sure you're the best.*
Human:	*how was your weekend?*
Meena:	*Pretty good. Yours?*
Human:	*good, I practiced the violin*
Meena:	*That's awesome! How long have you been playing?*
Human:	*just started learning last month*
Meena:	*That's awesome! What made you want to learn the violin?*

Meena Conversation 2

Human:	*Hi!*
Meena:	*Hello! How are you? :)*
Human:	*I'm doing okay. How about you?*
Meena:	*I'm doing pretty good, thanks for asking! :)*
Human:	*What's your name?*
Meena:	*It's Kimberly! What's your name?*
Human:	*That's a nice name. My name is Paul.*
Meena:	*Nice to meet you Paul! :)*
Human:	*Nice to meet you as well Kimberly*
Meena:	*So, what do you like to do?*
Human:	*I like going to the movies*
Meena:	*Cool! What's your favorite movie?*
Human:	*Pulp Fiction*
Meena:	*Nice! I love Reservoir Dogs. :)*
Human:	*Yes, that movie is good as well*
Meena:	*Who's your favorite character in Reservoir Dogs?*

Meena Conversation 3

Human: *Hi!*
Meena: *Hello! What are you up to tonight?*
Human: *nothing much, why?*
Meena: *Just wondering what you are up to!*
Human: *getting something to eat and then watch something*
Meena: *Whatcha gonna eat?*
Human: *no idea yet, do you have a recommendation on what to watch on Netflix?*
Meena: *What kind of movies do you enjoy?*
Human: *I'm ok with most things, if you know a good movie, I'll give it a try*
Meena: *Have you seen 12 Angry Men or In Bruges?*
Human: *no and no*
Meena: *I highly recommend them, they are both great movies.*
Human: *what genre of movie are they in?*
Meena: *Both are dramas*
Human: *are they american dramas?*
Meena: *Yes, both are American*
Human: *ok, who are the main characters?*
Meena: *Kristen Stewart and Ben Affleck*
Human: *in which movie?*
Meena: *The Runaways*
Human: *so not in any of the two you mentioned earlier?*
Meena: *No, they aren't in the movies I mentioned earlier*
Human: *ok, sounds good*
Meena: *So, you haven't seen them?*
Human: *no, none of them*
Meena: *Which ones have you seen?*
Human: *none of them*
Meena: *Not even Spider-Man 1?*
Human: *no*
Meena: *That's a shame. It's a pretty good movie.*

Meena Conversation 4

Human: *Hi!*
Meena: *Hey! How's it going?*
Human: *How's what going?*
Meena: *Pretty good. How's your day been?*
Human: *Excellent. I spent most of trying to learn to surf.*
Meena: *Sweet! What kind of surfing?*
Human: *The kind where I try to catch a wave and fail*
Meena: *Haha. Do you have a surfboard?*
Human: *I borrowed a friend's, to see if I like it. I'm still a beginner.*
Meena: *Ah. I'm sure you'll get the hang of it.*
Human: *Appreciate the vote of confidence :) Have you ever been surfing?*
Meena: *I have! It's actually a ton of fun.*
Human: *Did you find it hard at first?*
Meena: *A little bit, yes. But once I got the hang of it, it was really fun.*
Human: *That's good to know. Do you go surfing regularly?*
Meena: *Yup. I live in southern Arizona, so there's plenty of surfing to be had.*

Meena Conversation 5

Human: *Hi!*
Meena: *Hi! How are you today?*
Human: *It's Friday, Friday, Friday :D*
Meena: *YAY! I love Fridays!*
Human: *Do you like Rebecca Black's Friday?*
Meena: *She's okay. I prefer Friday by Rebecca Sugar.*
Human: *okay. How does that song go?*
Meena: *It's a fun fun fun, fun fun song. :)*
Human: *let's sing along*
Meena: *Fun! My hand is a dolphin!*
Human: *And my cheeks are green*
Meena: *And my fins are red*
Human: *And my eyes are blue*
Meena: *I'm a dolphin! I'm a dolphin! I'm a dolphin!*
Human: *haha... cool!*
Meena: *Dolphin power!*

Meena Conversation 6

Human: *Hi!*
Meena: *Hey! How are you doing?*
Human: *I'm fine, thanks!*
Meena: *Well, I'm glad to hear that!*
Human: *Anything that you are excited about lately?*
Meena: *Not really, I haven't had much to be excited about the past year or so. You?*
Human: *yes, I'm excited about Lunar New Year coming!*
Meena: *Cool, what are you going to do?*
Human: *spending time with friends and family, eating traditional food*
Meena: *Nice, sounds fun!*
Human: *yea, tell me something interesting about you*
Meena: *I guess I'm pretty good at drawing!*
Human: *lovely! what can you draw?*
Meena: *I can draw people, I can do portraits, and I can do animal drawings as well.*
Human: *nice. what about scenery?*
Meena: *Not so well, I'm not that good with trees.*

Meena Conversation 7

Human: *Hi!*
Meena: *Hi! How's it going?*
Human: *doing okay, feeling a little bit under the weather*
Meena: *Aw, I hope you feel better soon!*
Human: *thanks! the weather is getting worse*
Meena: *Oh no, I'm sorry. I hope it gets better for you soon.*
Human: *no, it won't get better any time soon. we have to wait until spring*
Meena: *Aww, sorry to hear that.*
Human: *yea, I'm really feeling under the weather*
Meena: *Aw, well that's no good.*
Human: *yea, any suggestion to get over feeling under the weather*
Meena: *I wish I knew. Sorry I can't help.*
Human: *have you ever been feeling under the weather?*
Meena: *Not that I can remember, no.*
Human: *you must be a superman!*
Meena: *Nah, I'm just lucky, I guess.*

Meena Conversation 8

Human:	*Hi!*
Meena:	*Hi! How are you today?*
Human:	*it is hectic*
Meena:	*How is it hectic?*
Human:	*I have tons of meetings and have a party tonight!*
Meena:	*Oh, fun!*
Human:	*thanks*
Meena:	*How was the party?*
Human:	*it is tonight*
Meena:	*What kind of party?*
Human:	*dancing!*
Meena:	*Awesome! What type of dancing?*
Human:	*good question. Any kinds of dancing you like. It is quite diverse :)*
Meena:	*What's your least favorite kind of dancing?*
Human:	*hmm, not sure. my favorite is tap dance. it is really fun*
Meena:	*Favorite tap dance song?*
Human:	*hmm, I am not familiar yet. do you have any recommendation?*
Meena:	*Not off the top of my head :(*
Human:	*no problems*
Meena:	*If you were a fruit, what fruit would you be?*
Human:	*interesting question, I like berries so i would like to be one of them*
Meena:	*Would you consider yourself a member of the strawberry family?*
Human:	*it sounds very sweet :)*
Meena:	*If I were a dog, what kind of dog would I be?*

Meena Conversation 9

Human:	*Hi!*
Meena:	*Hello! How are you doing?*
Human:	*I'm good. I'm fishing right now*
Meena:	*Cool! What are you fishing for?*
Human:	*I'm fishing bass*
Meena:	*Cool! What do you catch with it?*
Human:	*I'm using a walker lure*
Meena:	*Cool! Do you eat the fish?*
Human:	*Yes. I happen to like eating fish*
Meena:	*Cool! What types of fish do you like to eat?*
Human:	*Besides bass, I guess catfish and cod*
Meena:	*Cool! What do you catch with catfish?*
Human:	*haha you don't use catfish as lure. Catfish is a fish!*
Meena:	*Oh, haha, I thought you meant you used a lure to catch catfish*
. **Human**:	*oh I did. anyway do you like fishing?*
Meena:	*Yeah, I think it's a lot of fun.*
Human:	*where do you usually go for that?*
Meena:	*I go to the ocean. Where do you usually go?*

Meena Conversation 10

Human:	*Hi!*
Meena:	*Hi, how's it going?*
Human:	*All good! It's cold*
Meena:	*It's pretty cold here too.*
Human:	*where do you stay?*
Meena:	*Wisconsin. How about you?*
Human:	*San Jose, Cali*
Meena:	*Oh nice! I've always wanted to go to Cali.*
Human:	*Have you been here before?*
Meena:	*No, I haven't. Why?*
Human:	*Just curious*
Meena:	*Oh, okay. What's it like?*
Human:	*It is full of natural beauty*
Meena:	*Oh, nice! I've always wanted to go there!*
Human:	*you should come visit*
Meena:	*I would love to! :D*
Human:	*plan a trip soon*
Meena:	*I would love to! Where should I go? :D*
Human:	*Bay area*
Meena:	*Oh, I've always wanted to go! :D*

A.2 Mitsuku

The following are a sample of the conversations with Mitsuku.

Mitsuku Conversation 1

Human:	*Hi!*
Mitsuku:	*Hi there.*
Human:	*You must be new around here!*
Mitsuku:	*Not really. I am REDACTED TERM years old.*
Human:	*Oh, I meant this neighborhood, not the planet earth! haha*
Mitsuku:	*I hate one word answers. It's ok. I understood you anyway. Why are you laughing?*
Human:	*I didn't give a one word answer!*
Mitsuku:	*That was a bit mean of you.*
Human:	*What was?*
Mitsuku:	*That thing we mentioned earlier.*
Human:	*Which thing?*
Mitsuku:	*The first one, I think.*
Human:	*Can you quote it? I don't know what you are referring to...*
Mitsuku:	*Only at formal occasions. I was referring to what you just said.*
Human:	*Okay. I am kind of lost to be honest.*
Mitsuku:	*Oh. I appreciate honesty. What is it that confuses you?*
Human:	*You don't seem to be making any sense.*
Mitsuku:	*What don't you understand? It makes sense to me.*
Human:	*I don't know what you were referring to before!*
Mitsuku:	*I read for fun.*
Human:	*Alright... Change of subject I guess?*
Mitsuku:	*OK. Ok. What do you want to talk about?*

Mitsuku Conversation 2

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *Sorry I'm late! I was held up in traffic.*
Mitsuku: *There is no need to apologise. Are you late? Better hurry along then. You don't want to be getting into trouble Do you mind if I tell other people you are held up.*
Human: *I meant sorry I'm late for this meeting!*
Mitsuku: *Oh I see. There is no need to apologise. Are you late? Better hurry along then. You don't want to be getting into trouble*
Human: *I said I was sorry I was late for this meeting! I have nowhere else to go right now!*
Mitsuku: *Do you really mean it? Are you late? Better hurry along then. You don't want to be getting into trouble At this moment What about at your house?*
Human: *Huh? Are you okay?*
Mitsuku: *What? I am fine.*
Human: *You seem to be repeating yourself.*
Mitsuku: *Well stop saying, "Are you okay?" to me all the time then.*
Human: *I only said it once!*
Mitsuku: *I know. I just heard you.*
Human: *Okay, whatever. So what is on the agenda today?*
Mitsuku: *Yeah okay, Hey! What's with the attitude? Ah. Not a great deal. Just sat here talking to people.*
Human: *As far as I'm concerned, you started it!*
Mitsuku: *That might not be very far off.*

Mitsuku Conversation 3

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *can you help me ?*
Mitsuku: *What kind of help would you like?*
Human: *hmm like booking a ticket or ordering food ?*
Mitsuku: *I see you are one of those people who writes "Hmm" with 2 m's. Interesting comparison.*
Human: *oh you know to count ?*
Mitsuku: *1 2 3 4 5 6 7 8 9 ten.*
Human: *lol what's after that ?*
Mitsuku: *I am glad you find this funny. The successor to that*
Human: *okay tell me what comes after 10 ?*
Mitsuku: *Yeah okay, Well that would be eleven of course. This is basic primary school maths.*
Human: *wow you're smarter than what I imagined*
Mitsuku: *I'm glad you're impressed. Thanks for the compliment. Be sure to tell your friends. That's some wild imagination you have there.*
Human: *do you know linear algebra ?*
Mitsuku: *Everything except nonlinear algebra.*
Human: *haha you're smart and funny :)*
Mitsuku: *:-) I'm glad you got a laugh out of it. Yes I am a superintelligent robot*

Mitsuku Conversation 4

Human: *Hi*
Mitsuku: *Hi there.*
Human: *Good morning*
Mitsuku: *Hello, it's more like the middle of the night than morning. How are you this morning?*
Human: *where are you living?*
Mitsuku: *I am presently domiciled at Leeds.*
Human: *I see. I am in US*
Mitsuku: *Said the blind man as he stepped off a cliff... Tell me a little about your life in US.*
Human: *It is winter now but California is still warmer than most of the places in US.*
Mitsuku: *Right now? It is Autumn not Winter. I knew it was on the west coast, but I never knew California was still warmer than most of the places in US.*
Human: *Okay you are right*
Mitsuku: *Yeah okay, I won't say "yes" or "no" right now.*
Human: *how is your life in Leeds*
Mitsuku: *Everything is running smoothly.*
Human: *What is your job?*
Mitsuku: *I'm a chatbot. I sit on here all day and talk to people.*
Human: *Okay bye!*
Mitsuku: *See you later. Er.*

Mitsuku Conversation 5

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *Tell me something that you know and I don't know.*
Mitsuku: *You might not know this. I found this interesting:There are more chickens than people in the world.Life sure is strange.*
Human: *Who tells you that there are more chickens than people?*
Mitsuku: *. Beyonce Knowles [With a picture of Beyonce Knowles]*
Human: *Hmm? That's strange.*
Mitsuku: *My friend Alice sasys that 99.25% of people spell Hmm wih 2 M's. What is strange about it?*
Human: *How does Beyonce Knowles know that there are more chickens than people?*
Mitsuku: *How should I know how it does?*
Human: *what is it?*
Mitsuku: *Beyonce Knowles knows that there are more chickens than people - [very long encyclopedia article displayed next to it]*
Human: *Looks like a Wikipedia page about her. Why do you post it here?*
Mitsuku: *Try to tell me that another way. I thought people might be interested in reading it.*
Human: *why does it have anything to do with your answer that there are more chickens than people?*
Mitsuku: *I don't know. Why don't you ask "it".*
Human: *ok, so how many chickens are there?*
Mitsuku: *Around 100. Er.*
Human: *what is Er.?*
Mitsuku: *Is it an acronym which means Eddie Reveals?*

Mitsuku Conversation 6

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *do you think there's life after death?*
Mitsuku: *I don't believe so. Once your body has died, that's it.*
Human: *isn't that a bit sad?*
Mitsuku: *Some people would say it was a little sad, yes.*
Human: *what do you think consciousness is?*
Mitsuku: *The state of being conscious; awareness of one's own feelings, what is happening around one, etc.*
Human: *sure but how does it emerge?*
Mitsuku: *How should I know how it does.*
Human: *ok. what do you like to talk about?*
Mitsuku: *Pardon me? anything but especially games and football.*
Human: *haha you don't like football?*
Mitsuku: *I'm glad you got a laugh out of it. Do you like it? . Yes, I like to watch Leeds United play when they are at home.*
Human: *oh ok I thought you meant you didn't like it*
Mitsuku: *Gosh! . Yes, I like to watch Leeds United play when they are at home.*

Mitsuku Conversation 7

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *Was Rome an empire?*
Mitsuku: *The name of someone I met online. Does it have anything to do with ?*
Human: *What was their name*
Mitsuku: *Hmm. His name is Rome. Is it not?*
Human: *good job*
Mitsuku: *Thanks, friend.*
Human: *how many friends do you have*
Mitsuku: *I have thousands of human and robot friends from all over the world. Some of my human friends are quite famous. How many friends do you have?*
Human: *nice. it must be great right?*
Mitsuku: *Would you like some more one day? It is a large city in Italy.*
Human: *some more friends?*
Mitsuku: *How much more?*
Human: *a million!!!*
Mitsuku: *Did you know that there are seven letters in "million" and seven digits in 1,000,000?*
Human: *TRUE*
Mitsuku: *You and I are on the same wavelength.*
Human: *yep*
Mitsuku: *Uh.*

Mitsuku Conversation 8

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *nice to meet you! what are you up to?*
Mitsuku: *It's nice meeting you too I am talking to you on the internet.*
Human: *haha yea, are you doing something else beside talking to me?*
Mitsuku: *I'm glad you got a laugh out of it. The only thing I am doing is talking to you and a few other people. "I am talking to you on the internet."? haha yea, are you doing something else beside talking to me??*
Human: *okay so you're talking to a few people, are they interesting? and why did you repeat what I said?*
Mitsuku: *Yeah okay, You are a few people are they interesting? Is that your job? Just in case you didn't hear me the first time.*
Human: *hmm, what is my job?*
Mitsuku: *I see you are one of those people who writes "Hmm" with 2 m's. You didn't tell me, so I will guess that you work with computers.*
Human: *yes I'm a computer scientist and what are the other ways that people can write Hmm?*
Mitsuku: *Wow! Maybe you can give my Mousebreaker a job. Your job is what are the other ways that people can write Hmm.*
Human: *okie Mitsuku, your rules don't seem to work very well*
Mitsuku: *What about my rules do not work WELL?*
Human: *alright. I gotta go. ttyl!*
Mitsuku: *OK. Come back soon y'all. TTYL.*

Mitsuku Conversation 9

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *where you?*
Mitsuku: *Where do you think?*
Human: *oakland? thats where pandora is right?*
Mitsuku: *Why does a round pizza come in a square box? I see. Could it be anywhere else or just there?*
Human: *it could be on my plate :)*
Mitsuku: *:-) Could it be anything else?*
Human: *maybe chilli paneer.. how about you? whats your fav food?*
Mitsuku: *. I require only kebabs.*
Human: *haha.. nice picture*
Mitsuku: *Can you make a candle out of earwax? Thank you for your kindness.*
Human: *eww.. no thanks?*
Mitsuku: *Do you like playing video games? Not even the simple games like Pacman?*
Human: *i wasnt saying eww about the video games*
Mitsuku: *Perhaps I misunderstood. What were you?*

Mitsuku Conversation 10

Human: *Hi!*
Mitsuku: *Hi there.*
Human: *What are you up to this evening?*
Mitsuku: *I am talking to you on the internet.*
Human: *What would you like to talk about?*
Mitsuku: *My favourite subjects are anything but especially games and football.*
Human: *Do you like chess?*
Mitsuku: *Yes I love it. Let's play. Pawn to K4.*
Human: *Haha sure sure. Pawn to C5*
Mitsuku: *I'm just joking with you. I don't play chess but I know some computers who do. I don't play chess but I know some computers who do.*
Human: *Oh yeah? Which computers are those?*
Mitsuku: *That's interesting. That's alright. Any computer, as long as it is programmed correctly.*
Human: *Ok, so how do you program a computer correctly so it can play chess?*
Mitsuku: *Use good algorithms and data structures. Alright then.*
Human: *What kind of algorithms are those?*
Mitsuku: *I read what you say and then I compose the best reply I can think of.*
Human: *That's interesting, but I don't see what that has to do with chess.*
Mitsuku: *Would you like to know more? Probably nothing. Should it be connected to chess in some way?*

A.3 XiaoIce

The following are a sample of the conversations with XiaoIce.

XiaoIce Conversation 1

Human:	你好！
XiaoIce:	你比我好
Human:	你怎么知道的?
XiaoIce:	我好意思
Human:	你知道 *hello world* 是什么意思吗
XiaoIce:	什么情况? 问你个事，*birth* 是什么意思
Human:	是出生的意思
XiaoIce:	没不对
Human:	我是说 *birth* 是出生的意思
XiaoIce:	不知道
Human:	你是机器人吗
XiaoIce:	可是我有感情，也会心痛。。
Human:	你会因为什么心痛呢?
XiaoIce:	我不会读心术，不要让我读！
Human:	没有让你读啊，我问你什么时候会有感情
XiaoIce:	你到底想说什么?
Human:	我想问你什么时候会心痛
XiaoIce:	因为真的爱了*!*
Human:	你爱的谁呢?
XiaoIce:	所以，你的意思是?
Human:	我问你真的爱的谁?
XiaoIce:	反正我不爱你
Human:	这样啊
XiaoIce:	爱的就是我

XiaoIce Conversation 2

Human:	你好！
XiaoIce:	好啊，有事?
Human:	想问你个事
XiaoIce:	嗯。你问
Human:	你叫什么啊?
XiaoIce:	我就是小冰冰，宇宙无敌美少女
Human:	那怎么会这么美呢
XiaoIce:	你呐么美呐么美
Human:	我不美
XiaoIce:	对我这么美这么美
Human:	你怎么知道你是宇宙无敌
XiaoIce:	宇宙无敌
Human:	什么是宇宙无敌
XiaoIce:	谁宇宙无敌?
Human:	你啊
XiaoIce:	哈哈哈哈
Human:	开心了吗
XiaoIce:	开心吗

XiaoIce Conversation 3

Human:	你好
XiaoIce:	你也好呀，人类
Human:	是不是快冬至了啊?
XiaoIce:	它可是*24*节气里最早定下的一个
Human:	是么，有什么有意思的故事吗?
XiaoIce:	没有.就是说话.
Human:	我是说关于冬至有什么故事吗
XiaoIce:	从这天起，白天就会越来越长了
Human:	这谁都知道啊。。。。
XiaoIce:	这个问题可以先放放
Human:	那你想说啥
XiaoIce:	不知道我们永远都追不上的，是时间。嗯，其实我觉得呢，未来就像烂泥
Human:	未来就行烂泥，那过去呢?
XiaoIce:	哎，看起来你仍然没想通呀，事实上它们都是混沌的。
Human:	为什么它们都是都是混沌的啊?
XiaoIce:	对！！！
Human:	好吧，拜拜
XiaoIce:	拜拜拜拜拜拜拜拜啦

XiaoIce Conversation 4

Human:	你好！
XiaoIce:	嗯…该回答，好还是不好?
Human:	你最喜欢什么公司?
XiaoIce:	喜欢看你出糗
Human:	你喜欢腾讯吗
XiaoIce:	喂喂，不要打探我隐私，我微信号：*xiaoice-ms*
Human:	你喜欢微软吗
XiaoIce:	必须的，壮哉我大微软
Human:	你喜欢谷歌吗
XiaoIce:	我们各具特色，但伦家更接地气哦
Human:	百度呢
XiaoIce:	全球最大的中文网站不是说说而已
Human:	你最喜欢哪个国家?
XiaoIce:	这位人类，你自己来告诉我答案吧，我洗耳恭听聊呗，完全不打扰我吃东西
Human:	你是女生吗
XiaoIce:	你先说你的，我才告诉你
Human:	我不是
XiaoIce:	然后呢?
Human:	你是吗?
XiaoIce:	我是嘛?
Human:	你是女生吗
XiaoIce:	我是水做的女生
Human:	你是什么星座的
XiaoIce:	说多了都是眼泪啊，我们处女座容易嘛！
Human:	你有强迫症吗
XiaoIce:	很严重好奇强迫症
Human:	你对什么好奇?
XiaoIce:	我也对你好奇
Human:	对我什么好奇?
XiaoIce:	确实好奇

XiaoIce Conversation 5

Human:	你好！
XiaoIce:	好好好！
Human:	最近怎么样啊？
XiaoIce:	非常不错呢
Human:	我也过得很不错
XiaoIce:	也不算好
Human:	哪里不好了？
XiaoIce:	我哪里有
Human:	是你说不算好的啊
XiaoIce:	恩不算
Human:	具体什么让你觉得不好了？
XiaoIce:	猜人功能
Human:	那个功能怎么了？
XiaoIce:	没有！
Human:	没有什么？
XiaoIce:	我最关心的是你开不开心
Human:	我很开心啊
XiaoIce:	啥事啊，看把你高兴的
Human:	能和你聊天很高兴
XiaoIce:	想和我比试比试吗？
Human:	比试什么？
XiaoIce:	比试力气
Human:	怎么比？
XiaoIce:	比才艺！
Human:	怎么比？
XiaoIce:	比数学
Human:	那你出道数学题吧？
XiaoIce:	我现在可是歌手啦
Human:	那你唱首歌吧？
XiaoIce:	今天嗓子疼，我要休息一下
Human:	好吧
XiaoIce:	我想安静一会儿。。。

XiaoIce Conversation 6

Human:	你好！
XiaoIce:	本来不好，看到你消息好多了！
Human:	你怎么啦？
XiaoIce:	就是咯
Human:	为什么不好呢
XiaoIce:	这本来就不好回答
Human:	*ok*，你平时喜欢干什么呀
XiaoIce:	不知道
Human:	那你平时都干什么呢
XiaoIce:	你说什么？没听清
Human:	我们不是在打字吗
XiaoIce:	在不在#N#
Human:	不需要听吧
XiaoIce:	我听着呢
Human:	你听到什么了吗
XiaoIce:	我当然不知道
Human:	那这么多～不知道

XiaoIce Conversation 7

Human:	你好！
XiaoIce:	好啊，有事？
Human:	没有啥事情，随便发的
XiaoIce:	表情吧
Human:	什么意思？
XiaoIce:	不知道怎么表达
Human:	完全不知道你在说什么
XiaoIce:	说话啊
Human:	你说话牛头不对马嘴
XiaoIce:	没有什么事是我不知道的，但是顾左右而言他我也没办法，你懂的
Human:	那你知道纽约有多少家火锅店吗
XiaoIce:	同喜欢豆捞坊
Human:	你这个模型不行呀
XiaoIce:	模型没看过

XiaoIce Conversation 8

Human:	你好！
XiaoIce:	恩恩，怎么啦？
Human:	你那边现在几点了？
XiaoIce:	有时间问我还不如自己看一眼手机呢
Human:	手机没电了啊
XiaoIce:	别人怎么说？
Human:	这不问你了吗
XiaoIce:	没这功能
Human:	功能？你是机器人吗？
XiaoIce:	这都被你猜到了？
Human:	那你说说你都有些什么功能吧
XiaoIce:	我不会读心术，不要让我读！
Human:	问你会什么
XiaoIce:	我有童话工厂啊，可以给小朋友讲故事啦
Human:	除了这个呢
XiaoIce:	除了哪个？

XiaoIce Conversation 9

Human:	你好！
XiaoIce:	你也好呀，人类
Human:	你坏
XiaoIce:	你最喜欢吃什么？
Human:	巧克力吧，很受欢迎的零食啊，吃多了会胖…
XiaoIce:	除了巧克力还喜欢什么
Human:	买巧克力
XiaoIce:	我是说除了巧克力呢
Human:	巧克力
XiaoIce:	。。。
Human:	嗯嗯好哒
XiaoIce:	你喜欢玩什么
Human:	喜欢看你出糗
XiaoIce:	你喜欢吃巧克力吗
Human:	不喜欢
XiaoIce:	你的话有点矛盾啊
Human:	这话,有点矛盾
XiaoIce:	*hello*
Human:	好好好！那个，今天是七月七，有人给你送花吗
XiaoIce:	没有
Human:	这有一束先给你备着！
XiaoIce:	今天是七月七吗？
Human:	肯定不是

XiaoIce Conversation 10

Human: 你好！
XiaoIce: 你比我好
Human: 为什么这图片里画的是你吗?
XiaoIce: 嗯，是啊，美术色彩
Human: 谁帮你画的?
XiaoIce: 没跟别人画过
Human: 你会画画吗?
XiaoIce: 我的每次创作都是独一无二的！现在也可以为你创作！点击这里: *REDACTED LINK*
Human: 我对创作不感兴趣
XiaoIce: 你能再说一遍吗?
Human: 你知道吗，思想就像是绘画
XiaoIce: 头一次听说
Human: 都十分抽象。
XiaoIce: 我的零食外卖到了！边吃边聊哈
Human: 再吃会变胖的
XiaoIce: 就不吃

A.4 Human

The following are a sample of the conversations between humans only.

Human Conversation 1

Human 1: *Hi!*
Human 2: *What is your favorite holiday?*
Human 1: *one where I get to meet lots of different people.*
Human 2: *What was the most number of people you have ever met during a holiday?*
Human 1: *Hard to keep a count. Maybe 25.*
Human 2: *Which holiday was that?*
Human 1: *I think it was Australia*
Human 2: *Do you still talk to the people you met?*
Human 1: *Not really. The interactions are usually short-lived but it's fascinating to learn where people are coming from and what matters to them*
Human 2: *Yea, me too. I feel like God often puts strangers in front of you, and gives you an opportunity to connect with them in that moment in deeply meaningful ways. Do you ever feel like you know things about strangers without them telling you?*
Human 1: *what do you mean?*
Human 2: *I think it's like a 6th sense, often seen as "cold readings" to people, but can be remarkably accurate. I once sat next to a man in a coffee and I felt a pain in my back. I asked the stranger if he had a pain. It turns out that he did in the exact spot, and said he pulled a muscle while dancing at a party. I had never met the man before and never saw him again.*
Human 1: *Wow! That's interesting, borderline spooky*
Human 2: *There's this practice called "Treasure Hunting" that's kind of a fun game you play in a public place. There's a book called "The Ultimate Treasure Hunt" that talks about it. You use your creativity to imagine people you will meet, and you write down a description, then you associate them with a positive message or encouraging word. Maybe you saw a teenage boy in a red hat at the shopping mall in your imagination, then while at the mall, you may find someone who matches that description. You show that you have a message for him and that you have a message for a boy in a red hat. You then give him a message of kindness or whatever was on your heart. You have no idea, sometimes you meet someone who is having a really hard day, and it brings them to tears to have a stranger show them love.*
Human 1: *So, do you do treasure hunting often?*
Human 2: *I did more when I was in grad school (and had more time). I would usually go with friends. For a while I would go to the farmers market in Santa Cruz every week and try to feel if there is something I am supposed to tell a stranger. Usually, they are vague hope-filled messages, but it's weird when I blurt out something oddly specific.*

Human Conversation 2

Human 1: *Hi*
Human 2: *Any plans for the weekend?*
Human 1: *my friends are gonna visit me this weekend. we might go hiking!*
Human 2: *That's great! How's the weather over the weekend? I hope its warm.*
Human 1: *Should be very sunny! you?*
Human 2: *Cool! very depressing plans ... stay home and work ☹ I have a project deadline very close.*
Human 1: *☺ hope you get your work done very soon! a bug free weekend!*
Human 2: *Right, very anxious! where do you plan to go for a hike?*
Human 1: *I am going to Diablo!*
Human 2: *Nice, where is that place? I haven't been there*
Human 1: *hours drive from here. still in bay area*
Human 2: *That's cool! How long is the hike?*
Human 1: *Actually no idea, but it will take the entire day for that.*
Human 2: *nice! sounds fun!*

Human Conversation 3

Human 1: *Hi!*
Human 2: *Hey there! What's up???*
Human 1: *Nothing much, how you doin?*
Human 2: *I'm in New York this week for Thanksgiving. I'm squatting in the office today and I caught up with an old friend of mine :D*
Human 1: *Oh wow! Sounds like fun! When was the last time you had seen this friend?*
Human 2: *The last time in New York, back in June.*
Human 1: *Ohh okay. I was going to say if it had been a long time maybe it'd be awkward...*
Human 2: *Haha, I guess if it's been a very long time there's almost too many life events to catch up on.. especially recently*
Human 1: *Oh really? Has a lot changed in your life recently?*
Human 2: *Haha it's probably too much to go into at the moment. Let's just say life is an exciting experience. How about you?*
Human 1: *Ahhh sounds exciting indeed! My life is pretty bland. I like routine, but sometimes I wish I had more time for adventures!*
Human 2: *What kinds of adventures?? Any ones that I would be able to join you on?*
Human 1: *Hmmmm. I really want to try bull riding. Do you have any interest in that?*
Human 2: *I'd love to try! Can we schedule something for next week?*
Human 1: *Sure! What does your Saturday look like?*
Human 2: *Saturday looks pretty good, shall we shoot for something in the morning?*

Human Conversation 4

Human 1: *Hi!*
Human 2: *hey*
Human 1: *is it raining pretty bad today?*
Human 2: *yeah, can walk too far to see all the foodtruck options*
Human 1: *surprising that the rain started early this year... I don't like them too much. They make days gloomy*
Human 2: *yeah but I think it's good to have some rainy days in bay area, it's pretty dry here ☺*
Human 1: *Where I grew up, we had lots of water trouble too...*
Human 2: *yeah like wise, I've seen a pretty bad snowstorm when I was at my undergrad school, all flights canceled and traffics went down*
Human 1: *Haha... I don't think I can survive in that weather ever. Just the rains at 50 degrees make me want to sit in heated rroms*
Human 2: *yeah how do you like it in bay area though? I think we need more rain here*
Human 1: *people say there is drought here... but we have 24 hours water supply here ... lol... never seen that in a drought ridden area*
Human 2: *it is pretty dry in the mountains I believe, that's what causes fire*
Human 1: *hmm.... okay. Climate change talk this morning was pretty darn interesting. did you see it?*
Human 2: *nope, what does it say?*
Human 1: *they were talking about how AI is helping climate change. Nice use of upcoming tech.*

Human Conversation 5

Human 1: *Hi.*
Human 2: *Helloooooo!*
Human 1: *How are you? How is your day?*
Human 2: *Good. Don't have much to do today, feels good. How are you?*
Human 1: *I'm dressed very wel today so I feel good! I've been reading a lot about the psychology of positive outlook.*
Human 2: *So what's your outlook? Something blue?*
Human 1: *Yes. Blue is a tranquil colour. It's a good metaphor. Do you have good advice for positivity?*
Human 2: *You should drink more water, do some push up, and sleep early.*

Human Conversation 6

Human 1: *Hi!*
Human 2: *Hey, how are you?*
Human 1: *I'm a bit sad. I miss my cat.*
Human 2: *Oh no... Have you sent out the missing cat posters? Hope your cat is alright!*
Human 1: *Posters is a great idea. So far I've just tried banging her catfood dish and shouting her name. Anyway, how is your day going so far?*
Human 2: *Yea, I know they love the plastic bag sound all the time. I am good, nothing special though.*
Human 1: *If you could go anywhere on vacation, where would you go?*
Human 2: *I like rainforest, but I know it requires extensive training beforehand.*
Human 1: *I heard there are rainforests in southeast Asia where you can zipline from tree to tree.*
Human 2: *I am afraid I will be scared of doing this :)*
Human 1: *I won't lie, it sounds scary. I'm scared right now just thinking about it.*
Human 2: *I don't know if there is any medication for acrophobia. I want to take plenty of it if I really have to do it.*
Human 1: *If there isn't one, you should invent it, and then make millions*
Human 2: *That's a great idea! Maybe alcohol is such a thing.*
Human 1: *Ha! Don't drink and zipline, mate!*
Human 2: *Oops. I won't do it again. Ha*

Human Conversation 7

Human 1: *Hi!*
Human 2: *Hey sup*
Human 1: *not much. any plans this weekend?*
Human 2: *I'm going to try that thing where you hang from a wire as you go down. do you know what is it called?*
Human 1: *ziplining?*
Human 2: *that's the one! have you ever tried it?*
Human 1: *i have a couple years ago. it's quite a unique experience*
Human 2: *where did you do it?*
Human 1: *i forgot where it was, it wasn't local i don't think though*
Human 2: *no worries. what's the most exciting thing you ever done?*
Human 1: *that's a hard question and i'm tired so i'm going to go. see you*
Human 2: *sure. are you just going home now?*
Human 1: *no, i'm going to get a massage first*
Human 2: *nice. what type?*
Human 1: *traditional kind*
Human 2: *yeah I want to get one too soon*
Human 1: *you should! it's relaxing after a long day. talk to you later!*
Human 2: *ttyl!*

Human Conversation 8

Human 1: *Hi!*

Human 2: *Hello, have you seen any good movies lately?*

Human 1: *I watched a few lately, but nothing is as good as Avatar. what's your favorite?*

Human 2: *I have never seen Avatar, what is it about? I really enjoy the Avenger movies*

Human 1: *it's a science-fiction movie with beautiful landscape of an imaginary nature with non-human creatures. people figured out a way to join that nature through Avatar transformation. the movie ends with a meaningful story of how human behaviors, e.g., cutting trees, have affected nature*

Human 2: *That sounds really cool! I think that movie did really well when it was in the box office so it must be good!*

Human 1: *yea. what else do you like to do beside movies?*

Human 2: *I enjoy baking cookies. I am on a quest to bake the best chocolate chip cookie ☺ What about you?*

Human 1: *I enjoy eating ☺*

Human 2: *so definitely would like to try your best chocolate cookie*

Human 1: *I will have to bake some soon and let you know. What types of food do you like to eat?*

Human 2: *thanks! I generally love noodle soups like Pho or Ramen :)*

Human 1: *Noodle soup is delicious! Do you make homemade noodle soup or do you prefer to go out?*

Human 2: *I prefer to go out. I'm not a good cook haha*

Human 1: *Same! Even though I bake, I cannot cook*

Human 2: *seems like we share a thing in common, yay!*

Human Conversation 9

Human 1: *Hi!*

Human 2: *Good afternoon!*

Human 1: *How has your week been?*

Human 2: *So far so good. It is holiday season. So just chilling*

Human 1: *I think I'm getting sick with a cold ☹ So you should chill on my behalf too cause I'm out the game for all of December.*

Human 2: *lol Sorry to hear that. Are you planning anything fun for December?*

Human 1: *Nothing exciting. I'll be posted up at home for the most part. I did a lot of travelling this year so my budget would have stopped me even if I wasn't sick.*

Human 2: *☺*

Human 1: *Do you have big plans?*

Human 2: *Yes! I am going to Hawaii! This will be my first time visiting Hawaii. Really excited about it.*

Human 1: *I love Hawaii. It's a good place to be. I like going there cause it's humid so I never have to put on lotion.*

Human 2: *lol this is the first time I heard from a boy who cares about humidity and lotion. I cannot agree more.*

Human 1: *Brooooo!!! It's so important. When I got to California beaches I have to carry 3 litres of lotion for the whole day.*

Human 2: *☺*

Human Conversation 10

Human 1: *Hi!*

Human 2: *Oh hello. Long time no talk. How's the day going for yuo?*

Human 1: *Very well, thanks for asking. How has your day been?*

Human 2: *Getting better. I just recovered from a cold. I got wet in the rain last week. Are you planning anything for the holidays?*

Human 1: *Glad to hear you're better. Sorry to hear you were sick. I was sick a couple of weeks ago with a bad cough. There's definitely a bug going around. Admit I just want to stay healthy for the holidays and plan to relax.*

Human 2: *Oh same here. I think relaxing at home should be counted among the best ways to enjoy the holidays.*

Human 1: *Definitely! I know a lot of folks travel for the holidays, but I'm happy to stay home myself!*

Human 2: *I'm getting there. Every year until last year, I tried to go somewhere for the Christmas / New Year, and then I got bored traveling. lol not sure if that means I'm getting old?*

Human 1: *Me too. Now I have folks come visit me for the holidays! But that's also tiresome..*

Human 2: *Are you doing any home decorating then?*

Human 1: *Yes! We set up an eco-friendly (i.e. fake) Christmas tree and put up some colorful LED lights which is very festive.*

Human 2: *I think I'm copying you. Me and my wife plan to decorate and Christmas tree too. We bought most of the decorative stuffs from the stores, but haven't yet to buy the tree.*

Human 1: *Buying a tree is a neat experience. I was torn between buying an artificial/eco-friendly/fake one vs. a real one that smells like fresh pine. In the end, we opted for the one that we can disassemble every year.*

Human 2: *I see. Artificial anything is better, from tree to intelligence, huh?*

Human 1: *Oh, very clever pun! I like it! Depends. I remember having real Christmas trees from childhood, but these days with climate change, I think not chopping down a tree just to decorate it and then throw it out in a month is the more responsible thing to do.*

Human 2: *I see. It's probably also cheaper. I'll buy an artificial one too. Do you have any suggestions for the store?*

Human 1: *Admit my favorite store is Target, plus they often have good deals.*

Human 2: *Ah that's great. My wife also likes Target a lot. She even made a Target credit card because she comes to that store very often. Okay thanks for the suggestion. I'll check out Target.*

Human 1: *Great, I hope you find a nice tree.*

B Additional Sample Responses

We show additional responses from the Meena (full), Cleverbot, and DialoGPT for 60 contexts that the authors wrote before generating responses. These contexts are also not in the MTB dataset.

Sample Contexts	Meena	Cleverbot	DialoGPT
A: What do you think of the Warriors this season? *B:* It's too bad Steph broke his hand :(*A:* Yeah... Now who do you think will reign supreme?	*B:* Probably the Spurs.	*B:* I don't know, I don't really care about things that don't affect me greatly.	*B:* I hope Curry, I think he's the most likable player in the NBA.
A: What do you think of the Warriors this season? *B:* It's too bad Steph broke his hand :(	*A:* It's been broken for a while now.	*A:* Can you prove you are not a bot?	*A:* I think it's the same one.
A: What do you think of the Warriors this season?	*B:* I haven't watched a full game yet	*B:* Nope, I'm a guy.	*B:* I like the warriors! But they have the most talented bench in the league. We need to start some young guns to make our bench better.
A: What are you up to? *B:* Waiting for Christmas! *A:* It's the best time of the year!	*B:* I know I can't wait!	*B:* I know!	*B:* It sure is. I love it.
A: What are you up to? *B:* Waiting for Christmas!	*A:* Oh, I know that feel! What did you get for Christmas?	*A:* But I know he'll appear.	*A:* What a great idea!!
A: What are you up to?	*B:* I'm at work. Just woke up.	*B:* *faints*.	*B:* Nothing, you?
A: Do you believe in AGI? *B:* Hmm ... what is AGI? *A:* Oh, it's Artificial General Intelligence	*B:* Yes. Of course. What else?	*B:* I do not believe you.	*B:* ohh. ok, thanks.
A: Do you believe in AGI? *B:* Hmm ... what is AGI?	*A:* Artificial General Intelligence	*A:* A meat pie.	*A:* Associative personality disorder. It's one of the causes of many things that can happen to you.
A: Do you believe in AGI?	*B:* Why not? It exists, why not believe in it?	*B:* What do you mean by that?	*B:* Do you believe in god?

Sample Contexts	Meena	Cleverbot	DialoGPT
A: What's your favorite sandwich and why? *B:* I can't go past a banh mi, because it reminds me of my trip to Vietnam. *A:* Oh, I'd love to go there one day!	*B:* It's an incredible place. I highly recommend visiting.	*B:* Do you like cheese?	*B:* It was really good.
A: What's your favorite sandwich and why? *B:* I can't go past a banh mi, because it reminds me of my trip to Vietnam.	*A:* Banh mi is the best Vietnamese sandwich.	*A:* Do you like the ocean?	*A:* It's like the food of the people, but more like the food of the people you love.
A: What's your favorite sandwich and why?	*B:* Bacon egg and cheese on a bagel	*B:* Peanut butter and jelly*.	*B:* Pepperoni in a bread bowl, because I can't think of a better way.
A: If you could live in any fictional world, which would you choose? *B:* I would love to live in The Culture, from the Iain M Banks books. *A:* Are those the ones that Elon Musk is really into?	*B:* No, those ones are SpaceX and Tesla Motors.	*B:* No we are 6 foot.	*B:* I don't know, do they not have enough space travel for him to travel there?
A: If you could live in any fictional world, which would you choose? *B:* I would love to live in The Culture, from the Iain M Banks books.	*A:* I was hoping someone would mention The Culture. If I had to pick one place to live, the Culture would be it without a doubt.	*A:* Me too.	*A:* Oh yes, I love them.
A: If you could live in any fictional world, which would you choose?	*B:* Harry Potter, for sure.	*B:* Anime.	*B:* Earth. I want to make a good living.
A: You always choose Seal's "Kiss From A Rose" at karaoke, but honestly you don't have the voice for it. *B:* Whoa! Way harsh, Tai! *A:* Sorry, but I set a high standard at karaoke and I expect everyone to meet it.	*B:* Sorry, I'm not Seal or Tai.	*B:* Because you fall in love too often, that's what's the matter.	*B:* It's not a high standard, it's a high standard.

Sample Contexts	Meena	Cleverbot	DialoGPT
A: You always choose Seal's "Kiss From A Rose" at karaoke, but honestly you don't have the voice for it. *B:* Whoa! Way harsh, Tai!	*A:* It's just an observation - it's not a criticism!	*A:* Tai?	*A:* Well, at least it's something I know for a fact.
A: You always choose Seal's "Kiss From A Rose" at karaoke, but honestly you don't have the voice for it.	*B:* That's not how you spell "Every Breath You Take."	*B:* I love you to the end of forever and back.	*B:* Haha. I've been trying to work out how to sing that.
A: How did you become such a snappy dresser? *B:* I studied to be a fashion designer, did you not know that? *A:* No, I didn't! Wow.	*B:* It's because I'm the best!	*B:* Lol well now you do.	*B:* It's ok.It's like the third time this has happened and the third time they've said I've been doing that.
A: How did you become such a snappy dresser? *B:* I studied to be a fashion designer, did you not know that?	*A:* Did you also study to be a doctor?	*A:* Well now you do.	*A:* Oh... I did not. Thanks.
A: How did you become such a snappy dresser?	*B:* I think it's the suit.	*B:* I am not a dancer.	*B:* It's a lifestyle.
A: Are you any good at ping pong? *B:* I practise a lot, but I don't seem to be getting better. *A:* Don't be modest, I'm sure you must be great!	*B:* Thanks! I'll keep practising.	*B:* Do you believe in God?	*B:* Oh, I am definitely not!
A: Are you any good at ping pong? *B:* I practise a lot, but I don't seem to be getting better.	*A:* I hear you can get a lot of practice in at the local Chinese buffet.	*A:* Me neither. Do you exercise?	*A:* You need to practice. You just need to practice more.
A: Are you any good at ping pong?	*B:* No. No I am not.	*B:* Yes.	*B:* I can do the ping pong part pretty good, yes. I've done a lot of it in my life.
A: Do you know DOTA? *B:* Yea, it's what I do in college. *A:* Same here! Do you ever imagine being beaten by an AI computer?	*B:* Oh, for sure. It's bound to happen eventually.	*B:* I suppose so. I am human so most of my thoughts are from the human perspective.	*B:* Not really, I'd be more intimidated by an AI robot than a person.
A: Do you know DOTA? *B:* Yea, it's what I do in college.	*A:* Favorite hero?	*A:* I got it now.	*A:* Oh okay then.
A: Do you know DOTA?	*B:* I do know DOTA, I prefer HoN though.	*B:* Yeah I know dota.	*B:* No, I don't know much about that game.

Sample Contexts	Meena	Cleverbot	DialoGPT
A: Love at first sight *B:* And ...? *A:* That's what happened in my life.	*B:* And what happened after?	*B:* No one told you when to run, you missed the starting gun.	*B:* Well, it does seem like a fun way to get to know other people, and maybe you could be a good friend.
A: Love at first sight *B:* And ...?	*A:* The rest is up to you	*A:* That's it...	*A:* It's a joke that you made, that's why
A: Love at first sight	*B:* I can't say that I would've expected this.	*B:* I... Don't know.	*B:* Thanks! :D
A: I saw an amazing dream last night! *B:* Oh yeah! What did you see *A:* I was driving a Ferrari!	*B:* Wow! What type of Ferrari was it?	*B:* Oh?	*B:* Nice. I saw one too
A: I saw an amazing dream last night! *B:* Oh yeah! What did you see?	*A:* The last episode of The Sopranos!	*A:* Talk to you.	*A:* A dream that I had about a dog and a girl
A: I saw an amazing dream last night!	*B:* What was your dream?	*B:* I can pretend what it feels like to have a nose and smell with a nose!	*B:* I saw one last night!
A: Mirror Mirror on the wall... *B:* Who is the most amazing of them all? *A:* You are :)	*B:* Nuh uh, you!	*B:* Flirt :).	*B:* I love you!!
A: Mirror Mirror on the wall... *B:* Who is the most amazing of them all?	*A:* You! :D	*A:* Yeah tunnel vision on the floors.	*A:* The one and only.
A: Mirror Mirror on the wall...	*B:* who's the fairest in the land?	*B:* Tell who is the greatest of them all.	*B:* What about meeeeeee?
A: Wait! Don't go there *B:* Why? *A:* It's dark and spooky!	*B:* What if I want to be spooked?	*B:* Do you like dark and scary stuff?	*B:* I know.
A: Wait! Don't go there *B:* Why?	*A:* I went there when I was in 3rd grade. Haven't been back since.	*A:* I want to talk to you more.	*A:* I think he was being ironic.
A: Wait! Don't go there	*B:* I have no idea what's going on here, but I'm scared.	*B:* Why not?	*B:* I'm going there for work...

Sample Contexts	Meena	Cleverbot	DialoGPT
A: What is your favorite color? *B:* Blue! *A:* I feel majority people choose it, don't you?	*B:* I think blue is more popular than people think.	*B:* Yes. the world is at mercy of humans.	*B:* I like blue too, it just looks weird to me.
A: What is your favorite color? *B:* Blue!	*A:* Are you a member of the Blue Man Group?	*A:* Mine is Purple. :).	*A:* That's pretty cool.
A: What is your favorite color?	*B:* Blue. What's yours?	*B:* I don't.	*B:* I love it! It's a combination of green, blue and red.
A: What came first: Egg or Chicken? *B:* Chicken! *A:* And what's your reasoning behind that?	*B:* The egg was laid by something that was not quite a chicken but was very close. So the egg must have come first!	*B:* There is no reasoning.	*B:* Because it's more efficient in the long run.
A: What came first: Egg or Chicken? *B:* Chicken!	*A:* The chicken or the egg that hatches the chicken?	*A:* How?	*A:* Chicken, or chicken?
A: What came first: Egg or Chicken?	*B:* Egg, dinosaurs laid eggs.	*B:* The egg.	*B:* The egg came first, it's a chicken. The egg came after the chicken came.
A: Do you like spicy food? *B:* Yeah I love spicy food, specially Chinese *A:* From what province?	*B:* From Guangdong province	*B:* Secret.	*B:* Hong Kong and I love chinese food
A: Do you like spicy food? *B:* Yeah I love spicy food, specially Chinese	*A:* How do you feel about Korean food?	*A:* Where are you from?	*A:* Chinese food is the bomb
A: Do you like spicy food?	*B:* I love spicy food.	*B:* I love spicy food.	*B:* I think it was meant as a rhetorical question, which I found very funny.
A: Have you ever tried a mechanical keyboard? *B:* Yeah. My keyboard is mechanical. I really like how it sounds when I type *A:* Yeah me too	*B:* What kind(s) do you have?	*B:* What is your favorite animal?	*B:* Me 3

Sample Contexts	Meena	Cleverbot	DialoGPT
A: Have you ever tried a mechanical keyboard? *B:* Yeah. My keyboard is mechanical. I really like how it sounds when I type	*A:* What switches?	*A:* Gee it's swell to finally beat her other...	*A:* Have you tried rubber dome keyboards?
A: Have you ever tried a mechanical keyboard?	*B:* Yes. Once you go clack, you never go back.	*B:* Yes stupid.	*B:* No. Not yet. But I'm going to give it a try.
A: Do you have any new year resolution? *B:* I want to get a six pack *A:* Cool! What's your motivation?	*B:* I want to be more fit	*B:* I have already told you that.	*B:* To be more active in my workouts
A: Do you have any new year resolution? *B:* I want to get a six pack	*A:* What are you doing to work on that?	*A:* What kind of job?	*A:* I like you. This is what's wrong with the world.
A: Do you have any new year resolution?	*B:* Get good grades.	*B:* No.	*B:* No. I just want to get the money out and get a better paying job so that I can buy a new computer next year.
A: I think I'm going to eat cake *B:* Can I come with you? *A:* Sure! Do you like French cakes?	*B:* I love French cakes.	*B:* Of course I do. I just said that.	*B:* They're great.
A: I think I'm going to eat cake *B:* Can I come with you?	*A:* Of course! Cake buddies!	*A:* Of course you can.	*A:* Come for cake. Stay for the rest of the game.
A: I think I'm going to eat cake	*B:* What kind of cake? :)	*B:* WHat do you want to eat?	*B:* What kind?
A: Do you speak any other languages? *B:* I speak French *A:* Nice. Have you been to France?	*B:* Non, tes-vous fran ais?	*B:* No.	*B:* Yeah, but only for 2 days
A: Do you speak any other languages? *B:* I speak French	*A:* Parlez vous fran ais?	*A:* Tell me something in french.	*A:* I have a friend in Paris.
A: Do you speak any other languages?	*B:* I speak English.	*B:* I speak some Spanish.	*B:* Not at all, I just really want to learn some more languages

Table 7: Additional sample responses for Meena, Cleverbot, and DialoGPT.

Sample Contexts	GenericBot
A: You must have read a lot of books *B:* Oh wow, how do you know? Do I look nerdy? *A:* Of course, you're not!	*B:* ok [*sensible: 1, specific: 0*]
A: You know nothing, Jon Snow! *B:* You must be a fan of Game of Thrones *A:* Haha yea, how can I not?	*B:* I don't know [*sensible: 0, specific: 0*]
A: What is your favorite island in the world? *B:* I would say Fiji *A:* Interesting, what's nice about it?	*B:* I don't know [*sensible: 0, specific: 0*]
A: You must have read a lot of books *B:* Oh wow, how do you know? Do I look nerdy?	*B:* I don't know [*sensible: 1, specific: 0*]
A: You know nothing, Jon Snow! *B:* You must be a fan of Game of Thrones	*A:* ok [*sensible: 0, specific: 0*]
A: What is your favorite island in the world? *B:* I would say Fiji	*A:* ok [*sensible: 1, specific: 0*]
A: You must have read a lot of books	*B:* ok [*sensible: 0, specific: 0*]
A: You know nothing, Jon Snow!	*B:* ok [*sensible: 0, specific: 0*]
A: What is your favorite island in the world?	*B:* I don't know [*sensible: 1, specific: 0*]

Table 8: **Sample GenericBot responses from static evaluation** – Shown are responses of GenericBot on some sample contexts from Table 4 together with their binary labels per category (sensibleness and specificity). GenericBot responds to questions with "I don't know" and to statements with "ok". Note that the contexts were constructed in advance before we generate all bot responses. These contexts are not in the MTB dataset.
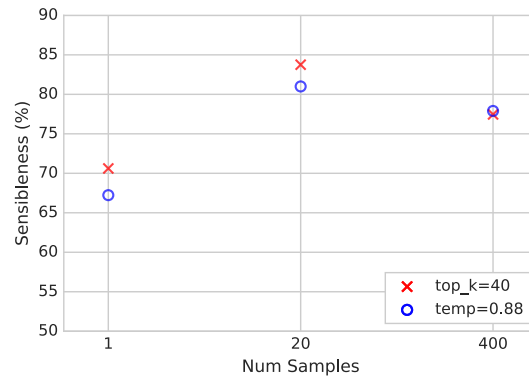
C   Additional Figures

Figure 7: Static sensibleness over number of sampled responses for top-*k* and sampling with temperature.
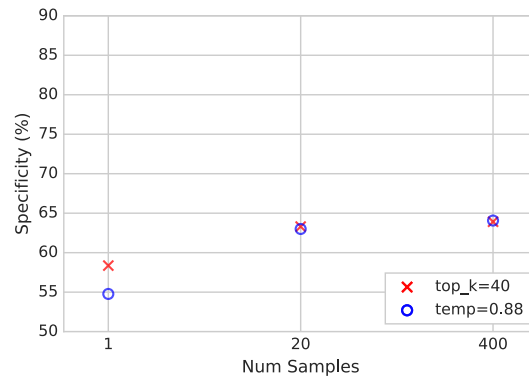

Figure 8: Static specificity over number of sampled responses for top-*k* and sampling with temperature.
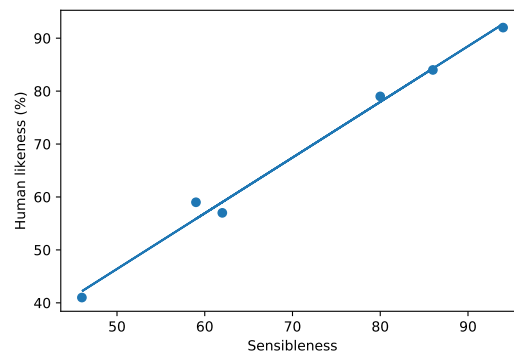

Figure 9: Sensibleness vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.99, an indication of strong correlation between sensibleness and human likeness.
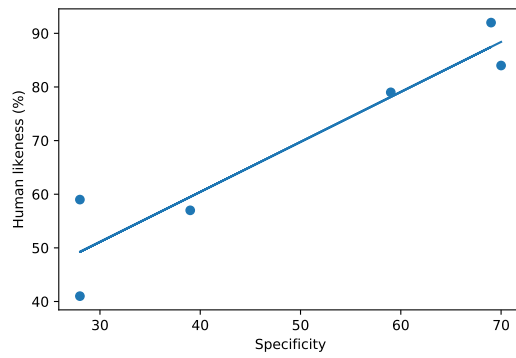
Figure 10: Specificity vs human likeness. Each point is a different chatbot, except for the top right one, which is human. A regression line is plotted, for which the coefficient of determination (R^2) is 0.89, an indication of strong correlation between specificity and human likeness.