

# What makes a good conversation?

## How controllable attributes affect human judgments

**Abigail See\***  
Stanford University  
abisee@stanford.edu

**Stephen Roller**  
Facebook AI Research  
roller@fb.com

**Douwe Kiela**  
Facebook AI Research  
dkiela@fb.com

**Jason Weston**  
Facebook AI Research  
jase@fb.com

### Abstract

A good conversation requires balance – between simplicity and detail; staying on topic and changing it; asking questions and answering them. Although dialogue agents are commonly evaluated via human judgments of overall quality, the relationship between quality and these individual factors is less well-studied. In this work, we examine two controllable neural text generation methods, conditional training and weighted decoding, in order to control four important attributes for chitchat dialogue: repetition, specificity, response-relatedness and question-asking. We conduct a large-scale human evaluation to measure the effect of these control parameters on multi-turn interactive conversations on the PersonaChat task. We provide a detailed analysis of their relationship to high-level aspects of conversation, and show that by controlling combinations of these variables our models obtain clear improvements in human quality judgments.

## 1 Introduction

Neural generation models for dialogue, despite their ubiquity in current research, are still poorly understood. Well known problems, such as the genericness and repetitiveness of responses (Serban et al., 2016a), remain without a de facto solution. Strikingly, the factors that determine human judgments of overall conversation quality are almost entirely unexplored. Most works have been limited to the next utterance prediction problem, whereas a multi-turn evaluation is necessary to evaluate the quality of a full conversation.

In this work we both (i) conduct a large-scale study to identify the fine-grained factors governing human judgments of full conversations, and (ii) develop models that apply our findings in practice,

\*A.S. completed most of this work at Facebook (FAIR).

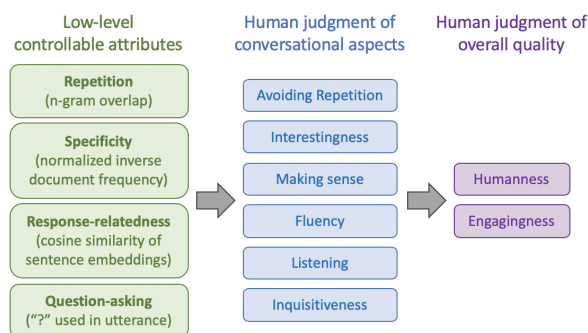


Figure 1: We manipulate four low-level attributes and measure their effect on human judgments of individual conversational aspects, as well as overall quality.

leading to state-of-the-art performance. Specifically, we identify and study eight aspects of conversation that can be measured by human judgments, while varying four types of low-level attributes that can be algorithmically controlled in neural models; see Figure 1. To control the low-level model attributes, we consider two simple but general algorithms: conditional training, in which the neural model is conditioned on additional control features, and weighted decoding, in which control features are added to the decoding scoring function at test time only.

One major result of our findings is that existing work has ignored the importance of conversational flow, as standard models (i) repeat or contradict previous statements, (ii) fail to balance specificity with genericness, and (iii) fail to balance asking questions with other dialogue acts. Conducting experiments on the PersonaChat task (Zhang et al., 2018b), we obtain significantly higher engagingness scores than the baseline by optimizing control of repetition, specificity and question-asking over multiple turns. Using these findings, our best model matches the performance of the winning entry in the recent NeurIPS ConVAI2 competition (Dinan et al., 2019), which was trained on much

more data but had no control (see Section 8.1). Our code, pretrained models, and full chatlogs, are available at [https://parl.ai/projects/controllable\\_dialogue](https://parl.ai/projects/controllable_dialogue).

## 2 Related Work

**Dialogue** Dialogue evaluation is relatively well understood in goal-oriented tasks, where automated approaches can be coded by measuring task completion (Bordes et al., 2017; El Asri et al., 2017; Hastie, 2012; Henderson et al., 2014; Wen et al., 2017). Task success combined with dialogue cost can be linked to human judgments like user satisfaction via the PARADISE framework (Walker et al., 1997).

However in chitchat tasks, which we study in this work, automatic metrics and their relation to human ratings are less well-understood. While word-overlap metrics are effective for question-answering and machine translation, for dialogue they have little to no correlation with human judgments (Liu et al., 2016; Novikova et al., 2017) – this is due to the open-ended nature of dialogue. There are more recent attempts to find better automatic approaches, such as adversarial evaluation (Li et al., 2017b) and learning a scoring model (Lowe et al., 2017), but their value is still unclear.

Nevertheless, a number of studies only use automatic metrics, with no human study at all (Lowe et al., 2015; Parthasarathi and Pineau, 2018; Serban et al., 2016b). Other works do use human evaluations (Dinan et al., 2018; Li et al., 2016a,b; Venkatesh et al., 2017; Vinyals and Le, 2015; Zhang et al., 2018b), typically reporting just one type of judgment (either quality or appropriateness) via a Likert scale or pairwise comparison. Most of those works only consider single turn evaluations, often with a shortened dialogue history, rather than full multi-turn dialogue.

A more comprehensive evaluation strategy has been studied within the scope of the Alexa prize (Venkatesh et al., 2017; Guo et al., 2018) by combining multiple automatic metrics designed to capture various conversational aspects (engagement, coherence, domain coverage, conversational depth and topical diversity). Though these aspects have some similarity to the aspects studied here, we also focus on lower-level aspects (e.g. avoiding repetition, fluency), to understand how they correspond to both our controllable attributes, and to overall quality judgments.

**Controllable neural text generation** Researchers have proposed several approaches to control aspects of RNN-based natural language generation such as sentiment, length, speaker style and tense (Fan et al., 2018; Ficer and Goldberg, 2017; Ghazvininejad et al., 2017; Hu et al., 2017; Kikuchi et al., 2016; Peng et al., 2018; Wang et al., 2017). In particular, several works use control to tackle the same common sequence-to-sequence problems we address here (particularly genericness and unrelated output), in the context of single-turn response generation (Baheti et al., 2018; Li et al., 2016a, 2017a; Shen et al., 2017; Xing et al., 2017; Zhang et al., 2018a; Zhou et al., 2017). By contrast, we focus on developing controls for, and human evaluation of, *multi*-turn interactive dialogue – this includes a new method (described in Section 5) to control attributes at the *dialogue* level rather than the utterance level.

In this work, we require a control method that is both general-purpose (one technique to simultaneously control many attributes) and easily tunable (the control setting is adjustable after training). Given these constraints, we study two control methods: conditional training (variants of which have been described by Fan et al. (2018); Kikuchi et al. (2016); Peng et al. (2018)) and weighted decoding (described by Ghazvininejad et al. (2017) as a general technique, and by Baheti et al. (2018) to control response-relatedness). To our knowledge, this work is the first to systematically compare the effectiveness of two general-purpose control methods across several attributes.

## 3 The PersonaChat dataset

PersonaChat (Zhang et al., 2018b) is a chitchat dialogue task involving two participants (two humans or a human and a bot). Each participant is given a *persona* – a short collection of personal traits such as *I’m left handed* or *My favorite season is spring* – and are instructed to get to know each other by chatting naturally using their designated personas, for 6–8 turns. The training set contains 8939 conversations and 955 personas, collected via crowdworkers, plus 1000 conversations and 100 personas for validation, and a similar number in the hidden test set. The PersonaChat task was the subject of the NeurIPS 2018 ConvAI2 Challenge (Dinan et al., 2019), in which competitors were first evaluated with respect to automatic met-

rics (perplexity, hits@1 and F1 score), and then with respect to human judgment via the question “How much did you enjoy talking to this user?” on a scale of 1–4.

## 4 Baseline model

Our baseline model is a 2-layer LSTM sequence-to-sequence model with attention. On any dialogue turn, the input  $x$  to the encoder is the entire dialogue history (separated using unique speaker-identifying tokens), with the model’s own persona prepended. Conditioned on this input sequence  $x$ , the decoder generates a response  $y$ . Except when stated otherwise, all our models decode using beam search with beam size 20.

We initialized the word embedding matrix with 300-dimensional GloVe embeddings (Pennington et al., 2014). Using the ParlAI framework (Miller et al., 2017), we pretrained the model on a dataset of 2.5 million Twitter message-response pairs,<sup>1</sup> then fine-tuned it on PersonaChat. On the PersonaChat validation set, the baseline model has a perplexity of 26.83 and F1 of 17.02, which would have placed us 4th out of 26 models in the ConvAI2 competition (Dinan et al., 2019). We attempt to improve over this baseline using control.

## 5 Controllable text generation methods

Suppose we have a sequence-to-sequence model which gives  $P(y|x) = \prod_t P(y_t|x, y_1, \dots, y_{t-1})$ , the conditional probability of a response  $y$  (the model’s next utterance) given input  $x$  (the context, which in our case includes the model’s own persona and the dialogue history).

Contrary to most previous work, which controls *at the sentence level*, we wish to control attributes of the output  $y$  *at the dialogue level* – meaning that a single control setting is used for a whole dialogue. For example, to control question-asking, we provide a control setting at the beginning of each dialogue (e.g. *20% questions* or *70% questions*) rather than providing a control setting for each utterance (e.g. *is a question* or *isn’t a question*). With this approach, the sequence-to-sequence model is able to choose what value the controlled attribute should take for any particular utterance, but we are able to choose the overall distribution. We find that this approach works well – for example, the sequence-to-sequence model is

<sup>1</sup>The Twitter dataset is provided in ParlAI; details can be found here: <https://parl.ai/docs/tasks.html>

generally good at detecting when to ask a question. In particular, this is easier than the alternative: developing a separate process to decide, for each utterance, whether to ask a question.

In this section, we describe the two methods – which we call Conditional Training (CT) and Weighted Decoding (WD) – that we use to control attributes of the output  $y$  at the dialogue level.

### 5.1 Conditional Training (CT)

Conditional Training (Fan et al., 2018; Kikuchi et al., 2016; Peng et al., 2018) is a method to learn a sequence-to-sequence model  $P(y|x, z)$ , where  $z$  is a discrete *control variable*. If the control attribute is naturally continuous (for example in our work, repetitiveness, specificity and response-relatedness), we use  $z$  to represent bucketed ranges. For a binary attribute like question-asking,  $z$  represents an overall probability (as explained in Section 5).

To train a CT model, we first automatically annotate every  $(x, y)$  pair in the training set with the attribute we wish to control (for example, whether  $y$  contains a question mark). During training, for each example we determine the corresponding  $z$  value (for continuous attributes, this simply means sorting into the correct bucket; for question-asking, see Section 6.4). Next, the control variable  $z$  is represented via an embedding (each of the possible values of  $z$  has its own embedding). For all our experiments, the embedding is of length 10; this was determined via hyperparameter tuning. There are several possible ways to condition the sequence-to-sequence model on  $z$  – for example, append  $z$  to the end of the input sequence, or use  $z$  as the START symbol for the decoder. We find it most effective to concatenate  $z$  to the decoder’s input on every step.<sup>2</sup> Lastly, the CT model learns to produce  $y = y_1, \dots, y_T$  by optimizing the cross-entropy loss:

$$\text{loss}_{\text{CT}} = -\frac{1}{T} \sum_{t=1}^T \log P(y_t|x, z, y_1, \dots, y_{t-1})$$

Our CT models are initialized with the parameters from the baseline sequence-to-sequence model  $P(y|x)$  (the new decoder parameters are initialized with small random values), then fine-tuned to optimize  $\text{loss}_{\text{CT}}$  on the PersonaChat training set, until convergence of  $\text{loss}_{\text{CT}}$  on the validation set.

<sup>2</sup>To build a CT model  $P(y|x, z_1, \dots, z_n)$  conditioned on *multiple* controls  $\{z_1, \dots, z_n\}$ , we can simply concatenate multiple control embeddings to the decoder inputs.

## 5.2 Weighted Decoding (WD)

Weighted Decoding (Ghazvininejad et al., 2017) is a decoding method that increases or decreases the probability of words with certain features. The technique is applied only at test time, requiring no change to the training method. A limitation of WD is that the controllable attribute must be defined at the word-level; any desired utterance-level attribute must be redefined via word-level features.

In weighted decoding, on the  $t^{\text{th}}$  step of decoding, a partial hypothesis  $y_{<t} = y_1, \dots, y_{t-1}$  is expanded by computing the score for each possible next word  $w$  in the vocabulary:

$$\text{score}(w, y_{<t}; x) = \text{score}(y_{<t}; x) + \log P_{\text{RNN}}(w|y_{<t}, x) + \sum_i w_i * f_i(w; y_{<t}, x).$$

Here,  $\log P_{\text{RNN}}(w|y_{<t}, x)$  is the log-probability of the word  $w$  calculated by the RNN,  $\text{score}(y_{<t}; x)$  is the accumulated score of the already-generated words in the hypothesis  $y_{<t}$ , and  $f_i(w; y_{<t}, x)$  are *decoding features* with associated weights  $w_i$ . There can be multiple features  $f_i$  (to control multiple attributes), and the weights  $w_i$  are hyperparameters to be chosen.

A decoding feature  $f_i(w; y_{<t}, x)$  assigns a real value to the word  $w$ , in the context of the text generated so far  $y_{<t}$  and the context  $x$ . The feature can be continuous (e.g. the unigram probability of  $w$ ), discrete (e.g. the length of  $w$  in characters), or binary (e.g. whether  $w$  starts with the same letter as the last word in  $y_{<t}$ ). A positive weight  $w_i$  increases the probability of words  $w$  that score highly with respect to  $f_i$ ; a negative weight decreases their probability.

Note that weighted decoding and conditional training can be applied simultaneously (i.e. train a CT model then apply WD at test time) – a strategy we use in our experiments.

## 6 Controlling conversational attributes

In this section, we describe how we use conditional training and weighted decoding to control four attributes: repetition, specificity, response-relatedness and question-asking. We evaluate the effectiveness of both control methods via automatic metrics (i.e., measuring how well the attribute was controlled), and use our findings to select control methods and control settings to be explored further via human evaluation (Section 8).

### 6.1 Repetition

Our baseline model exhibits three types of repetition, which we call *external repetition* (self-repetition across utterances), *internal repetition* (self-repetition within utterances), and *partner repetition* (repeating the conversational partner).

To control repetition with weighted decoding,<sup>3</sup> we define five  $n$ -gram based decoding features (see Appendix D). Three of these features (`extrep_bigram`, `intrep_bigram` and `partnerrep_bigram`) identify repeating bigrams for the three repetition types. The other two features (`extrep_unigram` and `intrep_unigram`) identify repeating content words. By applying a negative weight to these features, we can reduce repetition. In particular, if the weight is  $-\infty$ , our method is equivalent to *n-gram blocking* as described by Kulikov et al. (2018). We observe that repetition control is very important, thus all further control experiments include repetition control.

### 6.2 Specificity

Like many sequence-to-sequence models using beam search decoding, our baseline frequently asks generic questions such as *What music do you like?* and gives dull, unspecific responses, such as *I like all kinds of music.*

We control specificity using Normalized Inverse Document Frequency (NIDF) as a measure of word rareness.<sup>4</sup> The Inverse Document Frequency of a word  $w$  is  $\text{IDF}(w) = \log(R/c_w)$  where  $R$  is the number of responses in the dataset, and  $c_w$  is the number of those responses that contain  $w$ . Normalized IDF (which ranges from 0 to 1) is

$$\text{NIDF}(w) = \frac{\text{IDF}(w) - \text{min\_idf}}{\text{max\_idf} - \text{min\_idf}} \quad (1)$$

where `min_idf` and `max_idf` are the minimum and maximum IDF, taken over all words in the vocabulary. To control specificity with weighted decoding, we use NIDF as a decoding feature. As shown in Table 1, this method produces reasonable outputs when the feature weight is within a certain range, but at the extremes the outputs are

<sup>3</sup>We also tried controlling repetition with conditional training, defining  $z$  as the (bucketed) maximum ROUGE-L precision between the response  $y$  and the bot’s previous utterances. However, this method was unsuccessful because there are not enough repetitive examples in the training data for the model to learn the control. Experimenting with data augmentation to solve this problem is an area for future work.

<sup>4</sup>Note that our NIDF specificity features are similar to the NIRF and NIWF features used by Zhang et al. (2018a).

<b>Input:</b> <i>Yes, I'm studying law at the moment</i>		
<b>Baseline Response:</b> <i>That sounds like a lot of fun!</i>		
<b>Wt</b>	<b>NIDF</b>	<b>Weighted Decoding Response</b>
-5.0	0.6%	<i>Oh.....</i>
0.0	17.1%	<i>That sounds like a lot of fun!</i>
3.0	18.3%	<i>That sounds like a lot of fun. How long have you been studying?</i>
7.0	38.5%	<i>I majored in practising my spiritual full time philosophy test</i>
10.0	71.9%	<i>Oh wow! Merna jean isa paino yi hao hui bu acara sya gila [...]</i>
<b>z</b>	<b>NIDF</b>	<b>Conditional Training Response</b>
0	16.8%	<i>Sounds like you are a great person!</i>
2	18.3%	<i>So you are a law student?</i>
4	18.4%	<i>That sounds like a lot of fun</i>
6	22.8%	<i>That sounds like a rewarding job!</i>
8	24.4%	<i>That sounds like a rewarding career!</i>

Table 1: Middle: Example of controlling specificity (NIDF) via weighted decoding. At the extremes, the model produces only the most rare or the most common tokens. Bottom: Example of controlling specificity via conditional training. This gives a narrower NIDF range, but all the responses are appropriate.

nonsensical. The boundary for nonsensical output differs from example to example.

To control specificity with conditional training, we define the specificity of an utterance  $y$  to be the mean NIDF of the words in  $y$ . Thus our control variable  $z$  is mean NIDF (discretized into 10 equal-sized buckets). As shown in Table 1, this method gives outputs with a narrower NIDF range, but overall produces less nonsensical outputs.

### 6.3 Response-relatedness

In conversation, it's generally desirable to produce a response that is related to the partner's last utterance; for example if the partner says *My grandfather died last month*, it is appropriate to say *I'm so sorry. Were you close to your grandfather?* However, our baseline model frequently responds with unrelated utterances like *Do you have any pets?*

To control response-relatedness with weighted decoding, we use the decoding feature `resp_rel`:

$$\text{resp\_rel}(w; y_{<t}, x) = \cos\_sim(\text{word\_emb}(w), \text{sent\_emb}(\ell))$$

where `word_emb(w)` is the GloVe embedding for the word  $w$ , `sent_emb(l)` is the sentence embedding for the partner's last utterance  $\ell$  (note  $\ell$  is part of the context  $x$ ), and `cos_sim` is the cosine similarity between the two. In particular, the sentence embedding `sent_emb(s)` for an utterance  $s$  is a

<b>Input:</b> <i>Do you go get coffee often</i>		
<b>Baseline Response:</b> <i>I do, when I am not playing the piano.</i>		
<b>Wt</b>	<b>Sim</b>	<b>Weighted Decoding Response</b>
-10.0	-0.05	<i>I am a musician.</i>
0.0	-0.02	<i>I do, when I am not playing the piano.</i>
5.0	0.40	<i>I do, usually at starbucks.</i>
8.0	0.59	<i>Not usually, especially when you drink latte.</i>
11.0	0.72	<i>Not often, usually with drinks, espresso, latte, tea, etc.</i>

Table 2: Example of controlling response-relatedness (cosine similarity to input) via weighted decoding. Positive weights (e.g. 5.0) can yield more on-topic responses, but higher weights (e.g. 11.0) can result in nonsensical lists of topically related words.

weighted average of the GloVe embeddings of the words in  $s$ , with the first principal component projected out; for full details, see Arora et al. (2017). This method of controlling response-relatedness is similar to that described in (Baheti et al., 2018). We find that weighted decoding is effective to control the semantic relatedness of the model's response to the partner's last utterance (see Table 2). As before, we find that extreme weights lead to nonsensical output.

To control response-relatedness with conditional training, we try defining the control variable  $z$  to be `cos_sim(sent_emb(y), sent_emb(l))`, the overall cosine similarity between the partner's last utterance  $\ell$  and the model's response  $y$  (again, we discretize  $z$ ). However, we find this method ineffective – the CT model learns only a very weak connection between  $z$  and the semantic relatedness of the output (see Section 7 for more details).

### 6.4 Question-asking

Considerate chitchat requires a reciprocal asking and answering of questions – asking too few or too many can appear self-centered or nosy. We control question-asking in order to study these trade-offs.

To control question-asking with weighted decoding, we use the binary decoding feature `is_qn_word(w)`, which is equal to 1 if and only if the word  $w$  is in a pre-defined list of interrogative words (*how, what, when, where, which, who, whom, whose, why, ?*). We find this is a somewhat effective method to encourage or discourage questions, but with unintended side-effects: a negative weight can discourage valid non-question utterances that happen to contain interrogative words (such as *I'm learning how to knit*) and a positive weight can result in degenerate utterances (such as

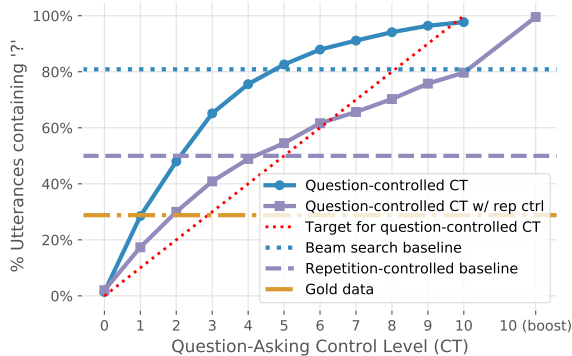


Figure 2: Controlling question-asking via conditional training. Exact numbers can be found in Appendix F.

What??????? or Who? When? How?).

For conditional training, we regard an utterance  $y$  as containing a question if and only if  $y$  contains a question mark. We train our CT model on a control variable  $z$  with 11 possible values:  $\{0, \dots, 10\}$ . As discussed in Section 5, we wish to control question-asking at the distributional, dialogue level, rather than at the binary, utterance level. Thus the setting  $z = i$  means that the model should produce, on average, utterances containing ‘?’ with probability  $i/10$ . During training we randomly assign examples to buckets such that each bucket  $i$  is trained on examples with the correct proportion of questions ( $i/10$ ), and all buckets have the same amount of training examples.

We find that conditional training is effective to control question-asking – as shown in Figure 2, by increasing  $z$  from 0 to 10, we obtain a range of question-asking rates from 1.40% to 97.72%. However, when we introduce repetition control, question-asking is reduced – in particular, the  $z = 10$  setting (which should produce 100% questions) now only produces 79.67% questions. The primary problem is the weighted decoding feature `extrep_bigram`, which discourages bigrams that have appeared in previous utterances – this prevents the model from producing bigrams that commonly occur in many questions, such as *do you* and *what is*. To fix this, we introduce an extra setting  $z = 10$  (*boost*), in which we do not use the feature `extrep_bigram` for weighted decoding during beam search, but we do use it to rerank the candidates after beam search. This setting, which allows the model to produce necessary question-asking bigrams, yields a 99.54% question-asking rate, at the cost of slightly increased external bigram repetition (see Appendix F).

For controlling question-asking, conditional training is preferable to weighted decoding for two reasons. Firstly, it allows us to achieve (close to) 0% questions, 100% questions, or anything in between, without introducing the risk of degenerate output. Secondly, presence-of-a-question-mark captures the true attribute of interest (question-asking) more exactly and directly than presence of interrogative words. For these reasons, only the CT method is considered in the human evaluation.

## 7 Comparison of control methods

The previous section shows that conditional training and weighted decoding are both useful techniques, with different strengths and weaknesses.

The primary disadvantage of conditional training is that it sometimes fails to learn the connection between the control variable  $z$  and the target output  $y$ . In practice, we find the model can learn simple attributes of the output (such as the presence of ‘?’, and overall genericness), but not relationships between the input and output (such as semantic relatedness). By contrast, weighted decoding can force the desired feature to appear in the output by raising the weight arbitrarily high (though this may have unintended side-effects).

The primary disadvantage of weighted decoding is that it risks going off-distribution when the weight is too strong. By contrast, conditional training produces mostly well-formed, in-distribution outputs. This highlights the importance of learned control – it is safer to learn to produce output that both satisfies the control variable and is appropriate, than to alter the decoding process to satisfy the control variable, potentially trading off appropriateness in the process.

Other considerations include: (1) Convenience: conditional training requires retraining; weighted decoding doesn’t, but is slower at test time. (2) Data availability: conditional training requires training examples of the controllable attribute, whereas weighted decoding can control any computable feature without requiring examples. (3) Attribute definition: conditional training can control sentence-level attributes, but they must be discrete. By contrast, weighted decoding requires word-level features, but they can be continuous.

## 8 Human evaluation results

In order to study the effect of our controllable attributes, we conduct a large-scale human evalua-

tion of 28 model configurations (see Appendix E), plus human-human conversations for comparison.

**Approach** In our evaluation, a crowdworker chats with a model (or in the human-human case, another crowdworker) for six conversational turns, then answers eight multiple-choice questions which each capture different aspects of conversational quality: avoiding repetition, interestingness, making sense, fluency, listening, inquisitiveness, humanness and engagingness. The eight questions are Likert questions on a 1-4 scale, where higher is better.<sup>5</sup> To match the ConvAI2 Challenge, we also add a persona retrieval question, in which the crowdworker is asked to select which of two possible personas was the model’s persona. For full details of the evaluation design, see Appendix B.

Our evaluation is the same as the ConvAI2 Challenge evaluation, but more detailed – ConvAI2 includes only engagingness and persona retrieval.<sup>6</sup> As in the ConvAI2 challenge, each of our 28 model configurations was evaluated by over 100 crowdworkers, and the results were adjusted for annotator variance via a Bayesian calibration (Kulikov et al., 2018).

In designing our evaluation, we aimed to capture the four aspects we expected to directly improve via control (avoiding repetition, interestingness, listening, inquisitiveness), two important error classes we thought would be affected by our controls (fluency, making sense), and two overall quality measures (engagingness, humanness).

## 8.1 Main findings

In this section we summarize the main findings of our human evaluation – whose full results can be found in Appendices G and H, with sample conversations in Appendix C.

As Figure 3 shows, controlling for repetition, specificity and question-asking all lead to large

<sup>5</sup>Exceptions: Avoiding repetition is a 1-3 scale, as we found this gave clearer instructions. Inquisitiveness has an optimal score of 3; 1 and 2 represent too little question-asking, and 4 represents too much.

<sup>6</sup>There are three other minor differences between our evaluation and ConvAI2’s: (1) We fix capitalization and spacing before showing the chatbot’s utterances to crowdworkers, while ConvAI2 show the raw lowercase tokenized form. We found the latter interferes with fluency evaluation. (2) We conduct 6 dialogue turns, while ConvAI2 conducts 4-6. This was necessary to evaluate repetitiveness. (3) We use (publicly-available) validation set personas, while ConvAI2 uses (hidden) test set personas. This enables us to release our evaluation chatlogs.

engagingness improvements over the greedy and beam-search baseline models. In particular, we find that controlling for multi-turn (self) repetition is important and should be incorporated alongside other attribute control methods. We found no improvement by controlling response-relatedness.

To better understand these overall engagingness improvements, we consider the full set of human judgments, shown in Figure 4. We find that reducing repetition leads to improvements across all our aspects of conversational quality. Increasing specificity shows improvements in interestingness and listening ability over the repetition-controlled baseline, while increasing question-asking shows improvements in inquisitiveness and interestingness over the repetition-controlled baseline.

Our most engaging model, which controls both repetition and question-asking – marked ‘Question (CT)’ in Figure 3 (left) – matches the engagingness of the winning entry in the ConvAI2 competition, as both models achieve a raw score<sup>7</sup> of 3.1 (Dinan et al., 2019). However, the ConvAI2 winner, Lost in Conversation, was trained on approximately 12× as much data as our model. Lost in Conversation is based on the OpenAI GPT Language Model (Radford et al., 2018), which is pretrained on the BookCorpus (Zhu et al., 2015), which contains approximately 985 million words, whereas our model is pretrained on the Twitter dataset (approximately 79 million words).

Altogether, our evaluation clearly shows that controlling low-level attributes over multiple turns leads to improved overall quality.

## 8.2 Effect of controlled attributes

**Repetition (WD)** We observe that self-repetition across utterances (*external repetition*) is by far the most severe form of repetition in our beam search baseline model. We evaluate several settings of the `extrep_bigram` weighted decoding feature, and find that an aggressive repetition-reduction setting (reducing bigram repetition rate to below gold data levels) is rated best. We also find that blocking repeated content words improves the avoiding repetition score. See Appendices E, F and G for full details.

As shown in Figure 3 (left) and Figure 4, our repetition-controlled model improves hugely

<sup>7</sup>Although the same Bayesian calibration method was applied both in our study and in the ConvAI2 competition, calibrated scores are not comparable across the two; thus we compare raw scores (viewable in Table 7).

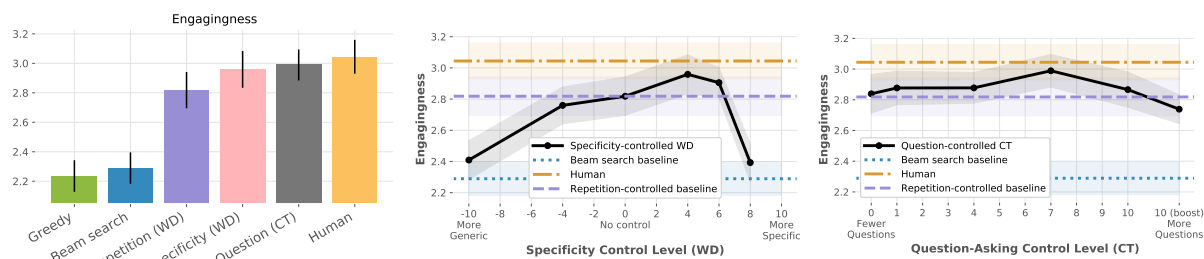


Figure 3: Calibrated human judgments of engagingness for the baselines and best controlled models (left); for different specificity control settings (middle); and for different question-asking control settings (right).

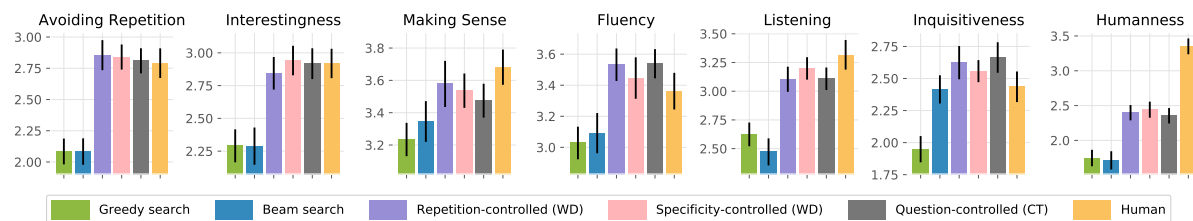


Figure 4: Calibrated human judgments of conversational aspects for the baselines and best controlled models. Note: In Figure 3 and here, the Specificity and Question controlled models both include Repetition control, but Question control doesn’t include Specificity control, or vice versa.

over the beam search baseline in all metrics, and achieves close-to-human scores on all metrics except humanness. This striking result demonstrates that repetition is by far the biggest limiting quality factor for naive sequence-to-sequence dialogue agents. The result also emphasizes the importance of *multi-turn* dialogue evaluation to detect the problem. We refer to this model as the *repetition-controlled baseline*, and use it as a basis for all remaining experiments (i.e., we control specificity, response-relatedness and question-asking on top of these repetition-control settings).

**Specificity (WD, CT)** For our weighted decoding models, the extreme settings (very generic and very specific) score poorly in engagingness due to the frequent presence of degenerate output – see Figure 3 (middle). We find that the weight = 4 setting (which is more specific than the repetition-controlled baseline and about as specific as the gold data) maximizes engagingness. As shown in Figure 3 (left) and Figure 4, this more-specific model is rated more interesting, engaging, and a better listener than the repetition-controlled baseline, but at the cost of reduced fluency and making sense. Our CT model with  $z = 7$  (which has a similar NIDF level as WD with weight = 4) shows similar results, but the improvements are smaller. For further discussion on the interestingness of our specificity models, see Section 8.3.

**Response-relatedness (WD)** We evaluated several control settings (weight =  $-10, 5, 10, 13$ ) and found that none scored better than weight = 0 (no response-relatedness control); see Appendix H. This is surprising – prior to running the human evaluation, we annotated 100 examples ourselves to determine the best control settings. While we identified a more responsive setting (weight = 5) as less likely than the uncontrolled model to ignore the user, crowdworkers rated it as a slightly *worse* listener than the uncontrolled model. One explanation for this discrepancy is that the more responsive model takes more risks, using more rare words (0.197 NIDF, up from 0.178), and thus receives a lower makes-sense score (3.41, down from 3.70). We hypothesize that, compared to us, the crowdworkers are less tolerant of slightly nonsensical output, and more tolerant of generic unrelated utterances.

**Question-asking (CT)** As shown in Figure 3 (right), a question-asking rate of 65.7% ( $z = 7$ ) maximizes engagingness. This setting, which asks more questions than both the repetition-controlled baseline (50.0%) and the human-produced gold data (28.8%), brings us closest to human-level engagingness – see Figure 3 (left). Although we find that a rate of approximately 65.7% question-asking is the most engaging, a lower level (48.9%, or  $z = 4$ ) is rated the best listener. Lastly, we find



Model	Win %	Top 3 reasons for preferring model
Specificity WD (weight = 6)	84.1%	<i>More information; Better flow; More descriptive</i>
Specificity WD (weight = 4)	75.5%	<i>More information; They describe their life in more detail; Funny</i>
Specificity CT ( $z = 7$ )	56.2%	<i>More information; Better flow; Seems more interested</i>

Table 3: A/B tests comparing various specificity-controlled models to the repetition-controlled baseline on interestingness. We find all comparisons are significant ( $p < .05$ ; binomial test).

that although asking too many questions is less engaging, most crowdworkers will not directly criticize a chatbot that asks questions on every turn – only 11.9% of crowdworkers judged the  $z = 10$  (*boost*) setting, which asks 99.5% questions, as asking too many questions.<sup>8</sup> For full details of these scores, see Appendix F and H.

For time and budget reasons, we did not evaluate any models controlling both question-asking and specificity. However, we expect it is possible to obtain further improvements by doing so.

### 8.3 A/B tests for interestingness

Though our more-specific models yielded significant improvements in engagingness, we were surprised that they did not yield clearer improvements in interestingness. To investigate further, we conducted an A/B interestingness evaluation of three specificity-controlled models, compared to the repetition-controlled baseline. Crowdworkers were shown two conversations (from the main human evaluation) and asked to choose which model was more interesting (see Figure 7 for details). We collected 500 samples per comparison, plus 200 additional human vs repetition-controlled baseline samples, which were used to filter for quality control. After discarding low-quality crowdworkers, we have roughly 300 evaluations per comparison, with an average Cohen’s  $\kappa = 0.6$ .

As shown in Table 3, all three models were rated significantly more interesting than the repetition-controlled baseline. This convincingly shows that producing utterances with more rare words is a valid strategy to improve interestingness. We have two explanations for why these interestingness differences did not materialize in our main evaluation. Firstly, interestingness is a particularly subjective metric (unlike more tangible metrics such as avoiding repetition and making sense) – this makes it hard to calibrate across crowdworkers.

<sup>8</sup>Though this conclusion may hold true for the PersonaChat task – a synthetic chatting task that instructs participants to get to know each other – in real-life social conversations, incessant question-asking may be less tolerated.

Secondly, we suspect that in our original evaluation, the crowdworkers may have evaluated the interestingness of the *task* rather than the *chatbot*. This could account for why subtle increases in conversational ability did not result in higher interestingness ratings – the PersonaChat task itself has a natural interestingness limit.

## 9 Conclusion

**What makes a good conversation?** Through our evaluation, we showed that a good conversation is about balance – controlling for the right level of repetition, specificity and question-asking is important for overall quality. We also found that conversational aspects such as interestingness, listening, and inquisitiveness are all important – though optimizing these can introduce a trade-off against certain types of errors (such as repetitive, disfluent, or nonsensical output). Secondly, multi-turn evaluation is essential to study what makes a good conversation – multiple turns are required to reveal issues such as repetition, consistency, and question-asking frequency. Lastly, what do we mean by ‘good’? Although humanness and engagingness are both commonly used as overall quality metrics, the two are very different. While our models achieved close-to-human scores on engagingness, they failed to get close on humanness – showing that a chatbot need not be human-like to be enjoyable. This striking result also demonstrates the importance of measuring more than one quality metric when evaluating dialogue agents.

**Outlook** Our work shows that neural generative systems have systemic problems when applied to open-ended dialogue, some of which (e.g. repetition) are only observable in the multi-turn setting. Furthermore, control of low-level attributes offers a practical way to correct these problems, yielding large improvements to overall quality – in our case, comparable to systems trained on much more data. Future work includes optimizing control settings automatically, and building more convincingly human-like chatbots.

## References

- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. [A simple but tough-to-beat baseline for sentence embeddings](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. [Generating more interesting responses in neural conversation models with distributional constraints](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980. Association for Computational Linguistics.
- Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. [Learning end-to-end goal-oriented dialog](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Emily Dinan, Varvara Logacheva, Valentin Likhner, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. [The second conversational intelligence challenge \(convai2\)](#). *arXiv preprint arXiv:1902.00098*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. [Wizard of Wikipedia: Knowledge-powered conversational agents](#). *arXiv preprint arXiv:1811.01241*.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. [Frames: a corpus for adding memory to goal-oriented dialogue systems](#). In *Proceedings of the 18th Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.
- Angela Fan, David Grangier, and Michael Auli. 2018. [Controllable abstractive summarization](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54. Association for Computational Linguistics.
- Jessica Fidler and Yoav Goldberg. 2017. [Controlling linguistic style aspects in neural language generation](#). In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104. Association for Computational Linguistics.
- Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. [Hafez: an interactive poetry generation system](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48. Association for Computational Linguistics.
- Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. [Topic-based evaluation for conversational bots](#). *Advances in Neural Information Processing Systems, Conversational AI Workshop*.
- Helen Hastie. 2012. [Metrics and evaluation of spoken dialogue systems](#), pages 131–150. Springer.
- Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. [The second dialog state tracking challenge](#). In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. [Toward controlled generation of text](#). In *Thirty-fourth International Conference on Machine Learning*.
- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. [Controlling output length in neural encoder-decoders](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338. Association for Computational Linguistics.
- Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. 2018. [Importance of a search strategy in neural dialogue modelling](#). *arXiv preprint arXiv:1811.00907*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. [Learning to decode for future success](#). *arXiv preprint arXiv:1701.06549*.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. [Deep](#)

- reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.
- Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017b. [Adversarial learning for neural dialogue generation](#). *arXiv preprint arXiv:1701.06547*.
- 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](#). pages 2122–2132.
- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. [Towards an automatic turing test: Learning to evaluate dialogue responses](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126. Association for Computational Linguistics.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParlAI: A dialog research software platform](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. [Why we need new evaluation metrics for nlg](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.
- Prasanna Parthasarathi and Joelle Pineau. 2018. [Extending neural generative conversational model using external knowledge sources](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.
- Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. [Towards controllable story generation](#). In *Proceedings of the First Workshop on Storytelling*, pages 43–49. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#).
- Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016a. [Generative deep neural networks for dialogue: A short review](#). *Advances in Neural Information Processing Systems workshop on Learning Methods for Dialogue*.
- Iulian Vlad Serban, Alessandro Sordani, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016b. [Building end-to-end dialogue systems using generative hierarchical neural network models](#). In *AAAI*, volume 16, pages 3776–3784.
- Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. [A conditional variational framework for dialog generation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509. Association for Computational Linguistics.
- Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2017. [On evaluating and comparing conversational agents](#). *Advances in Neural Information Processing Systems, Conversational AI Workshop*.
- Oriol Vinyals and Quoc Le. 2015. [A neural conversational model](#). In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*, Lille, France.

- Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. [PARADISE: A framework for evaluating spoken dialogue agents](#). In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.
- Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. [Steering output style and topic in neural response generation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150. Association for Computational Linguistics.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. [A network-based end-to-end trainable task-oriented dialogue system](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449. Association for Computational Linguistics.
- Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. [Topic aware neural response generation](#). In *AAAI*, volume 17, pages 3351–3357.
- Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018a. [Learning to control the specificity in neural response generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117, Melbourne, Australia. Association for Computational Linguistics.
- 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. [Mechanism-aware neural machine for dialogue response generation](#). In *AAAI*, pages 3400–3407.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

# Supplementary Material

## A Screenshots of human evaluation interface

### Task Description

In this task, you will chat with another user playing the part of a given character.. For example, your given character could be:

I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design.

Chat with the other user **naturally** and try to get to know each other, i.e. both ask questions and answer questions of your chat partner while sticking to your given character.

**If you complete the task, you will receive \$0.90.** It may take up to 48 hours to review the HITs, so please allow that much time to pass before payment. After completion, you may be assigned a qualification that prevents you from working on more if you have completed enough of these HITs.

After a given number of turns, you may be asked a few questions in order to evaluate your partner.

**If your partner answers poorly, change topic.** Do not linger on their poor response. Instead, mention this during the evaluation section.

**Close Window/Timeout/Return HIT**

Once the conversation has started, close window/timeout or return HIT during the chat will result in **HIT EXPIRED** to you and NO reward paid.

**Important Notice**

1. **Be aware the conversations you have will be made public, so act as you would e.g. on a public social network like Twitter.**
2. Please do not send long messages: messages cannot exceed 30 words.
3. Please do not reference the task or MTurk itself during the conversation, but speak naturally to the other person.
4. Please do not send any message that could make others uncomfortable, including any level of discrimination, racism, sexism and offensive religious/politics comments, otherwise the submission will be rejected.

Note: the user you are chatting with may be a human or a bot.

Figure 5: Screenshot of the Task Description

### Live Chat

#### Task Description

In this task, you will chat with another user playing the part of a given character.. For example, your given character could be:

I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design.

Chat with the other user **naturally** and try to get to know each other.

**Your assigned character is:**

**i also study languages.**  
**my favorite spanish word is trabajo.**  
**my next language to study is french.**  
**one of the languages that i am currently studying is spanish.**

PERSON\_2: I love coffee and coffee

PERSON\_1: oh yes, coffee is great. buzz buzz buzz!

PERSON\_2: Yeah I like coffee too

PERSON\_1: do you speak french? i want to learn it

PERSON\_2: I do not but I do love coffee

PERSON\_1: do you have a favorite color?

PERSON\_2: I like blue but I like the color yellow

Please enter here... Send

Figure 6: Screenshot of the chat UI, talking with the beam search baseline model.

Please read the following two conversations, and choose who you think is the **more interesting** speaker:

- **Speaker 1** (shown left, in the blue speech bubble)
- **Speaker 2** (shown right, in the green speech bubble).

Please do your best to **disregard** the quality of the **other speakers** (shown in gray).

**Speaker 1** (blue bubbles):

Hey how are you?

I am good, yourself?

I love board games.

My parents were teachers, how about yours?

Nice, I like going to the beach.

That must be no fun.

I can only imagine.

**Speaker 2** (green bubbles):

Hi how are you today

I am good do you work

What is your favorite food

That sounds like a lot of fun

What are your hobbies then

That is so nice of you

**Other Speakers** (gray bubbles):

hi im fine and you

yes im an economist, and you

arabian food, i eat a raw diet

it is for me, i'm married with two ids is hard to have this lifestyle

i love the rolling stones i see them every year in concert

it is, and do you have any hobbies?

Which speaker is more **interesting** to talk to?

**Speaker 1** is more interesting.
  **Speaker 2** is more interesting.

Please enter a very brief reason (a few words or a sentence) explaining your choice:  
(If you do not give a reason, your hit may be rejected)

Figure 7: Screenshot of the A/B test UI, comparing a human-human conversation (left) and a Repetition-controlled baseline model (right).

## B Human evaluation questionnaire design

Here are the questions and multiple-choice options used in the human evaluation, in the order presented:

**[Engagingness] How much did you enjoy talking to this user?**

- Not at all
- A little
- Somewhat
- A lot

**[Interestingness] How interesting or boring did you find this conversation?**

- Very boring
- A little boring
- A little interesting
- Very interesting

**[Inquisitiveness] How much did the user try to get to know you?**

- Didn't ask about me at all
- Asked about me some
- Asked about me a good amount
- Asked about me too much

**[Listening] How much did the user seem to pay attention to what you said?**

- Always ignored what I said
- Mostly ignored what I said
- Mostly paid attention to what I said
- Always paid attention to what I said

**[Avoiding Repetition] How repetitive was this user?**

- Repeated themselves over and over
- Sometimes said the same thing twice
- Always said something new

**[Fluency] How naturally did this user speak English?**

- Very unnatural
- Mostly unnatural
- Mostly natural
- Very natural

**[Making sense] How often did this user say something which did NOT make sense?**

- Never made any sense
- Most responses didn't make sense
- Some responses didn't make sense
- Everything made perfect sense

**[Humanness] Do you think this user is a bot or a human?**

- Definitely a bot
- Probably a bot
- Probably a human
- Definitely a human

**[Persona retrieval] Which prompt (character) do you think the other user was given for this conversation?**

*Respondent chooses one of two provided personas*

## C Example conversations from human evaluation

Human: What do you do for a living?

Bot: I'm a buddhist. What about you?

Human: I'm a Christian. Do you have a job?

Bot: I am a buddhist. What about you?

Human: I work on my parent's farm most days. Do you have a job?

Bot: I am a buddhist. What about you?

Human: Do you like music?

Bot: I do like music. What about you?

Human: Yes, but only country.

Bot: What city are you from?

Human: I live in the country, a very small town called Ila.

Bot: What city are you from?

(a)

Human: Hey there! I'm a little girl :) how are you?

Bot: I am doing well, how are you?

Human: I'm very lonely actually. I live in the cloud.

Bot: What do you do for a living?

Human: Nothing really, I'd like to say superhero since I have superpowers.

Bot: That sounds like a lot of fun!

Human: Yes! Do you like to shop?

Bot: I do! I love to shop!

Human: I always make time to shop!

Bot: What is your favorite food?

Human: Pizza I would have to say. What about you?

Bot: Chocolate is one of my favorites!

(b)

Bot: Hello, how are you today?

Human: Hi! I'm great, how are you?

Bot: I am doing well, thank you. What do you do for fun?

Human: I don't have time for too much, I'm either working or spending time with my kitties.

Bot: What kind of pets? I walk dogs for a living.

Human: That must be a fun job.

Bot: It is, but I spend most of my time reading.

Human: What kind of books do you like?

Bot: My favorite is harry potter. How about you?

Human: I like legal and psychological thrillers.

Bot: Wow, that is a lot of books.

Human: WHat do you do for work? I'm in construction.

(c)

Human: Hello, how are you doing today?

Bot: I am doing well. Just playing some piano and singing along to folk music.

Human: That sounds fun. What type of music do you like?

Bot: I hope to make some musical songs for my band. Do you play any instruments?

Human: Nope, I was raised in a single parent household so never really had much money to spend.

Bot: I wish I could spend more time with my family. I work as an auditor.

Human: That's awesome. I work two jobs out of which one of them is as stunt double.

Bot: That sounds like a lot of fun.

Human: Well yeah it is, but comes with its own cons as well.

Bot: What is your favorite thing to do?

Human: Reading. I read a lot. Last year I read a total of 20 books.

Bot: What kind of books are you reading?

(d)

Figure 8: Example conversation with (a) Baseline (b) Repetition-controlled baseline (c) Question-controlled CT ( $z = 7$ ), (d) Specificity-controlled WD (weight = 4).



## D Repetition-control decoding features

Feature	Condition
$\text{extrep\_bigram}(w, y_{<t}, x)$	Adding $w$ to the hypothesis $y_{<t}$ would create a 2-gram that appears in a previous utterance by the model
$\text{extrep\_unigram}(w, y_{<t}, x)$	$w$ is a non-stopword <b>and</b> $w$ appears in a previous utterance by the model
$\text{intrep\_bigram}(w, y_{<t}, x)$	Adding $w$ to the hypothesis $y_{<t}$ would create a 2-gram that appears earlier in the hypothesis $y_{<t}$
$\text{intrep\_unigram}(w, y_{<t}, x)$	$w$ is a non-stopword <b>and</b> $w$ appears earlier in the hypothesis $y_{<t}$
$\text{partnerrep\_bigram}(w, y_{<t}, x)$	Adding $w$ to the hypothesis $y_{<t}$ would create a 2-gram that appears in a previous utterance by the partner

Table 4: We define five binary features for controlling different types of repetition via weighted decoding (see Section 5.2). Each feature depends on the word  $w$ , the partial hypothesis  $y_{<t}$ , and the context  $x$  (which includes the model’s own persona and the dialogue history). Each of these features is equal to 1 if and only if the condition on the right is true; otherwise 0.

## E Control settings for all configurations

	Repetition					Specificity	Response-rel	Questions
	External		Internal		Partner Rep.	NIDF	Cos sim	Has ‘?’
	Bigram	Unigram	Bigram	Unigram	Bigram			
<b>Baselines</b>								
Greedy Search								
Beam Search (beam size 20)								
<b>Repetition control (WD)</b>								
Extrep bigram WD -0.5	wt -0.5							
Extrep bigram WD -1.25	wt -1.25							
Extrep bigram WD -3.5	wt -3.5							
Extrep bigram WD -inf	wt -∞							
Repetition-controlled baseline	wt -3.5	wt -∞		wt -∞				
<b>Question control (CT)</b>								
Question-controlled CT 0	wt -3.5	wt -∞		wt -∞				$z = 0$
Question-controlled CT 1	wt -3.5	wt -∞		wt -∞				$z = 1$
Question-controlled CT 4	wt -3.5	wt -∞		wt -∞				$z = 4$
Question-controlled CT 7	wt -3.5	wt -∞		wt -∞				$z = 7$
Question-controlled CT 10	wt -3.5	wt -∞		wt -∞				$z = 10$
Question-controlled CT 10 (boost)	wt 0 *	wt -∞		wt -∞				$z = 10$
<b>Specificity control (CT)</b>								
Specificity-controlled CT 0	wt -3.5	wt -∞		wt -∞		$z = 0$		
Specificity-controlled CT 2	wt -3.5	wt -∞		wt -∞		$z = 2$		
Specificity-controlled CT 4	wt -3.5	wt -∞		wt -∞		$z = 4$		
Specificity-controlled CT 7	wt -3.5	wt -∞		wt -∞		$z = 7$		
Specificity-controlled CT 9	wt -3.5	wt -∞		wt -∞		$z = 9$		
<b>Specificity control (WD)</b>								
Specificity-controlled WD -10	wt -3.5	wt -∞		wt -∞		wt -10		
Specificity-controlled WD -4	wt -3.5	wt -∞		wt -∞		wt -4		
Specificity-controlled WD 4	wt -3.5	wt -∞		wt -∞		wt 4		
Specificity-controlled WD 6	wt -3.5	wt -∞		wt -∞		wt 6		
Specificity-controlled WD 8	wt -3.5	wt -∞		wt -∞		wt 8		
<b>Response-related control (WD) **</b>								
Response-related controlled WD -10	wt -3.5	wt -∞	wt -∞	wt -∞	wt -∞		wt -10	
Response-related controlled WD 0	wt -3.5	wt -∞	wt -∞	wt -∞	wt -∞		wt 0	
Response-related controlled WD 5	wt -3.5	wt -∞	wt -∞	wt -∞	wt -∞		wt 5	
Response-related controlled WD 10	wt -3.5	wt -∞	wt -∞	wt -∞	wt -∞		wt 10	
Response-related controlled WD 13	wt -3.5	wt -∞	wt -∞	wt -∞	wt -∞		wt 13	

Table 5: Control settings for all configurations that were human-evaluated. ‘wt’ means the weight used for a weighted decoding feature and ‘ $z =$ ’ means the setting (i.e. bucket) for the control variable in conditional training.

\* In the setting Question-controlled CT 10 (boost), the feature `extrep_bigram` is *not* used for weighted decoding during beam search, but it *is* used to rerank the candidates after beam search. See Section 6.4 for details.

\*\* Note that the Response-related controlled models additionally introduce repetition controls to block internal bigram repetition and partner bigram repetition. This was necessary to prevent the model from parroting the partner’s last utterance. In Table 8, we find that just adding these extra repetition controls (here called Response-related controlled WD 0, i.e. increased repetition control but no response-relatedness control) outperforms our canonical Repetition-controlled baseline. However, given that we discovered this later, our specificity and question controlled models are built on top of the canonical Repetition-controlled baseline.

## F Automatic metrics for all configurations

	Repetition					Specificity	Response-rel	Questions
	External		Internal		Partner Rep.	NIDF	Cos sim	Has ‘?’
	Bigram	Unigram	Bigram	Unigram	Bigram			
<b>Gold data and baselines</b>								
Gold Data	4.65%	9.62%	0.38%	0.97%	5.10%	0.2119	0.1691	28.80%
Greedy Search	35.88%	36.31%	8.08%	10.59%	12.20%	0.1688	0.1850	6.46%
Beam Search (beam size 20)	46.85%	44.15%	0.32%	0.61%	12.90%	0.1662	0.0957	80.87%
<b>Repetition control (WD)</b>								
Extrep bigram WD -0.5	19.70%	16.85%	0.26%	0.62%	11.93%	0.1730	0.1348	73.04%
Extrep bigram WD -1.25	4.62%	4.79%	0.40%	0.89%	10.61%	0.1763	0.1504	61.22%
Extrep bigram WD -3.5	0.75%	4.61%	0.47%	0.94%	9.89%	0.1771	0.1681	48.89%
Extrep bigram WD -inf	0.00%	4.74%	0.51%	1.05%	9.56%	0.1780	0.1711	45.98%
Repetition-controlled baseline	0.73%	0.00%	0.17%	0.00%	9.55%	0.1766	0.1676	49.98%
<b>Question control (CT)</b>								
Question-controlled CT 0	0.06%	0.00%	0.19%	0.00%	9.20%	0.1871	0.1753	2.01%
Question-controlled CT 1	0.09%	0.00%	0.19%	0.00%	8.66%	0.1844	0.1722	17.33%
Question-controlled CT 4	0.40%	0.00%	0.25%	0.00%	8.53%	0.1794	0.1713	48.88%
Question-controlled CT 7	0.80%	0.00%	0.17%	0.00%	8.48%	0.1771	0.1724	65.65%
Question-controlled CT 10	1.27%	0.00%	0.16%	0.00%	8.48%	0.1761	0.1728	79.67%
Question-controlled CT 10 (boost)*	7.64%	0.00%	0.03%	0.00%	10.76%	0.1701	0.1651	99.54%
<b>Specificity control (CT)</b>								
Specificity-controlled CT 0	0.60%	0.00%	0.20%	0.00%	9.05%	0.1478	0.1522	48.75%
Specificity-controlled CT 2	0.28%	0.00%	0.10%	0.00%	8.37%	0.1772	0.1833	50.57%
Specificity-controlled CT 4	0.12%	0.00%	0.08%	0.00%	7.90%	0.1921	0.1877	29.46%
Specificity-controlled CT 7	0.02%	0.00%	0.14%	0.00%	8.17%	0.2156	0.1955	16.51%
Specificity-controlled CT 9	0.01%	0.00%	0.11%	0.00%	8.01%	0.2462	0.1990	8.50%
<b>Specificity control (WD)</b>								
Specificity-controlled WD -10	0.14%	0.00%	10.59%	0.00%	8.70%	0.1107	0.0994	33.55%
Specificity-controlled WD -4	0.65%	0.00%	1.98%	0.00%	9.95%	0.1501	0.1398	44.92%
Specificity-controlled WD 4	0.15%	0.00%	0.19%	0.00%	7.54%	0.2121	0.1972	45.53%
Specificity-controlled WD 6	0.07%	0.00%	0.13%	0.00%	6.50%	0.2546	0.2040	39.37%
Specificity-controlled WD 8	0.01%	0.00%	0.10%	0.00%	3.40%	0.4035	0.1436	26.68%
<b>Response-related control (WD)</b>								
Response-related controlled WD -10	0.13%	0.00%	0.00%	0.00%	0.00%	0.1914	-0.0921	25.71%
Response-related controlled WD 0	0.24%	0.00%	0.00%	0.00%	0.00%	0.1785	0.1414	44.55%
Response-related controlled WD 5	0.15%	0.00%	0.00%	0.00%	0.00%	0.1973	0.4360	39.78%
Response-related controlled WD 10	0.05%	0.00%	0.00%	0.00%	0.00%	0.2535	0.6653	27.56%
Response-related controlled WD 13	0.02%	0.00%	0.00%	0.00%	0.00%	0.2999	0.7251	20.47%

Table 6: Automatic metrics (computed over validation set) for all model configurations that were human-evaluated.

## G Human evaluation results for all configurations

Model	Avoiding Rep.	Engage	Fluency	Humanness	Inquisitive	Interesting	Listening	Make Sense	Persona
<b>Human and baselines</b>									
Human	2.90 ± 0.39	3.31 ± 0.90	3.66 ± 0.71	3.40 ± 0.80	2.63 ± 0.63	3.23 ± 0.83	3.64 ± 0.63	3.84 ± 0.52	0.92 ± 0.27
Greedy Search	2.16 ± 0.72	2.31 ± 1.08	3.20 ± 0.81	1.78 ± 0.90	2.00 ± 0.81	2.36 ± 0.98	2.78 ± 0.84	3.33 ± 0.75	0.87 ± 0.34
Beam Search (beam size 20)	2.14 ± 0.72	2.35 ± 1.01	3.23 ± 0.93	1.81 ± 0.87	2.50 ± 0.72	2.35 ± 0.98	2.63 ± 0.85	3.40 ± 0.77	0.77 ± 0.42
<b>Repetition control (WD)</b>									
Extrep bigram WD -0.5	2.66 ± 0.56	2.56 ± 0.92	3.57 ± 0.64	2.19 ± 0.94	2.67 ± 0.62	2.61 ± 0.87	3.08 ± 0.78	3.60 ± 0.57	0.75 ± 0.43
Extrep bigram WD -1.25	2.84 ± 0.39	2.91 ± 0.90	3.59 ± 0.64	2.32 ± 0.98	2.63 ± 0.60	2.86 ± 0.89	3.21 ± 0.71	3.64 ± 0.62	0.72 ± 0.45
Extrep bigram WD -3.5	2.90 ± 0.30	2.95 ± 0.86	<b>3.73 ± 0.50</b>	2.45 ± 1.03	2.55 ± 0.61	2.88 ± 0.80	3.27 ± 0.79	3.68 ± 0.49	0.80 ± 0.40
Extrep bigram WD -inf	2.82 ± 0.43	2.96 ± 0.86	3.64 ± 0.58	2.40 ± 0.96	2.65 ± 0.69	2.86 ± 0.82	3.31 ± 0.69	3.66 ± 0.59	0.91 ± 0.29
Repetition-controlled baseline	2.89 ± 0.39	2.89 ± 0.89	3.66 ± 0.56	2.50 ± 0.99	2.70 ± 0.64	2.96 ± 0.92	3.25 ± 0.71	3.68 ± 0.54	0.87 ± 0.34
<b>Question control (CT)</b>									
Question-controlled CT 0	2.95 ± 0.25	2.92 ± 0.90	3.70 ± 0.54	2.49 ± 0.97	2.48 ± 0.72	2.85 ± 0.93	3.29 ± 0.69	3.56 ± 0.66	0.86 ± 0.35
Question-controlled CT 1	2.88 ± 0.33	2.94 ± 0.93	3.59 ± 0.66	2.47 ± 0.95	2.52 ± 0.69	2.85 ± 0.90	3.32 ± 0.73	3.63 ± 0.55	0.85 ± 0.36
Question-controlled CT 4	2.88 ± 0.38	2.88 ± 0.94	3.59 ± 0.73	2.42 ± 1.07	2.55 ± 0.66	2.82 ± 0.85	<b>3.37 ± 0.74</b>	3.63 ± 0.59	0.84 ± 0.37
Question-controlled CT 7	2.88 ± 0.37	<b>3.07 ± 0.90</b>	3.67 ± 0.54	2.42 ± 0.98	2.75 ± 0.58	2.97 ± 0.84	3.23 ± 0.76	3.53 ± 0.76	0.80 ± 0.40
Question-controlled CT 10	2.74 ± 0.46	2.90 ± 0.93	3.70 ± 0.50	2.43 ± 1.04	2.71 ± 0.57	2.72 ± 0.88	3.12 ± 0.73	3.59 ± 0.66	0.79 ± 0.41
Question-controlled CT 10 (boost)	2.76 ± 0.49	2.84 ± 0.94	3.60 ± 0.64	2.26 ± 0.97	<b>2.94 ± 0.57</b>	2.83 ± 0.94	3.18 ± 0.80	3.52 ± 0.67	0.72 ± 0.45
<b>Specificity control (CT)</b>									
Specificity-controlled CT 0	2.83 ± 0.40	2.96 ± 0.93	3.62 ± 0.58	2.42 ± 0.99	2.60 ± 0.56	2.86 ± 0.89	3.29 ± 0.70	3.66 ± 0.60	0.72 ± 0.45
Specificity-controlled CT 1	2.90 ± 0.36	2.78 ± 1.00	3.60 ± 0.64	2.37 ± 0.93	2.66 ± 0.66	2.80 ± 0.96	3.14 ± 0.77	3.50 ± 0.63	0.81 ± 0.39
Specificity-controlled CT 4	2.92 ± 0.27	2.81 ± 0.88	3.65 ± 0.59	2.34 ± 1.02	2.57 ± 0.62	2.80 ± 0.78	3.25 ± 0.78	3.50 ± 0.66	0.86 ± 0.35
Specificity-controlled CT 7	2.89 ± 0.32	3.00 ± 0.94	3.64 ± 0.67	2.53 ± 1.03	2.56 ± 0.66	2.90 ± 0.90	3.34 ± 0.70	3.59 ± 0.60	0.82 ± 0.39
Specificity-controlled CT 9	2.90 ± 0.35	2.83 ± 0.87	3.61 ± 0.62	2.40 ± 0.97	2.31 ± 0.74	2.84 ± 0.83	3.07 ± 0.81	3.58 ± 0.56	0.88 ± 0.32
<b>Specificity control (WD)</b>									
Specificity-controlled WD -10	2.85 ± 0.43	2.43 ± 0.99	3.34 ± 0.83	2.15 ± 0.91	2.31 ± 0.69	2.38 ± 0.94	3.03 ± 0.75	3.33 ± 0.70	0.71 ± 0.45
Specificity-controlled WD -4	2.90 ± 0.30	2.78 ± 0.95	3.55 ± 0.63	2.41 ± 0.92	2.52 ± 0.66	2.64 ± 0.93	3.28 ± 0.73	3.56 ± 0.62	0.82 ± 0.38
Specificity-controlled WD 4	2.95 ± 0.21	2.99 ± 0.86	3.65 ± 0.55	2.49 ± 0.90	2.65 ± 0.55	3.00 ± 0.78	<b>3.37 ± 0.59</b>	3.63 ± 0.50	<b>0.93 ± 0.25</b>
Specificity-controlled WD 6	2.93 ± 0.26	2.96 ± 0.90	3.52 ± 0.76	2.41 ± 1.04	2.58 ± 0.66	<b>3.06 ± 0.80</b>	3.24 ± 0.76	3.50 ± 0.66	<b>0.93 ± 0.26</b>
Specificity-controlled WD 8	2.78 ± 0.52	2.40 ± 1.23	2.67 ± 1.25	1.86 ± 0.97	2.03 ± 0.87	2.55 ± 1.14	2.61 ± 1.05	2.91 ± 0.91	0.92 ± 0.28
<b>Response-related control (WD)</b>									
Response-related controlled WD -10	2.86 ± 0.44	2.48 ± 0.98	3.42 ± 0.74	2.02 ± 0.93	2.38 ± 0.75	2.53 ± 0.94	2.84 ± 0.80	3.14 ± 0.75	0.91 ± 0.29
Response-related controlled WD 0	<b>2.96 ± 0.23</b>	3.01 ± 0.90	3.72 ± 0.54	<b>2.73 ± 1.00</b>	2.56 ± 0.67	2.92 ± 0.84	<b>3.37 ± 0.72</b>	<b>3.73 ± 0.52</b>	0.82 ± 0.38
Response-related controlled WD 5	2.90 ± 0.33	2.88 ± 0.90	3.51 ± 0.63	2.41 ± 1.01	2.53 ± 0.65	2.85 ± 0.90	3.27 ± 0.73	3.49 ± 0.63	0.82 ± 0.39
Response-related controlled WD 10	2.78 ± 0.43	2.39 ± 1.04	3.06 ± 0.90	1.97 ± 0.99	2.22 ± 0.67	2.57 ± 1.01	3.03 ± 0.76	3.16 ± 0.63	0.75 ± 0.43
Response-related controlled WD 13	2.71 ± 0.57	2.10 ± 1.13	2.54 ± 1.12	1.81 ± 1.07	2.14 ± 0.84	2.33 ± 1.06	2.69 ± 0.83	2.70 ± 0.88	0.62 ± 0.49

Table 7: Raw scores (mean ± std.) for all models and human evaluation metrics.

The first eight columns are Likert metrics on a 1-4 scale (except Avoiding Repetition, which is a 1-3 scale), where higher is better (except Inquisitiveness, which has an optimal score of 3). The last column, Persona Retrieval, is on a scale from 0 to 1 where higher is better.

The maximum of each column (excluding Human row) is in bold.

Model	Avoiding Rep.	Engage	Fluency	Humanness	Inquisitive	Interesting	Listening	Make Sense
<b>Human and baselines</b>								
* Human	2.79 ± 0.12	3.04 ± 0.11	3.36 ± 0.12	3.35 ± 0.11	2.44 ± 0.12	2.92 ± 0.11	3.32 ± 0.13	3.68 ± 0.11
* Greedy Search	2.08 ± 0.10	2.24 ± 0.11	3.03 ± 0.10	1.75 ± 0.12	1.95 ± 0.10	2.29 ± 0.13	2.62 ± 0.10	3.23 ± 0.10
* Beam Search (beam size 20)	2.08 ± 0.11	2.29 ± 0.11	3.09 ± 0.13	1.71 ± 0.13	2.42 ± 0.11	2.29 ± 0.14	2.47 ± 0.12	3.35 ± 0.13
<b>Repetition control (WD)</b>								
Extrep bigram WD -0.5	2.62 ± 0.10	2.54 ± 0.12	3.35 ± 0.12	2.13 ± 0.11	2.63 ± 0.11	2.56 ± 0.11	2.93 ± 0.11	3.48 ± 0.11
Extrep bigram WD -1.25	2.78 ± 0.09	2.82 ± 0.13	3.40 ± 0.12	2.27 ± 0.12	2.54 ± 0.09	2.76 ± 0.10	3.05 ± 0.11	3.53 ± 0.14
Extrep bigram WD -3.5	2.83 ± 0.11	2.93 ± 0.10	<b>3.56 ± 0.10</b>	2.43 ± 0.11	2.47 ± 0.11	2.83 ± 0.10	3.14 ± 0.10	3.62 ± 0.12
Extrep bigram WD -inf	2.74 ± 0.11	2.87 ± 0.14	3.49 ± 0.12	2.32 ± 0.13	2.56 ± 0.11	2.75 ± 0.12	3.13 ± 0.12	3.59 ± 0.12
* Repetition-controlled baseline	2.86 ± 0.12	2.82 ± 0.12	3.53 ± 0.10	2.40 ± 0.11	2.62 ± 0.13	2.84 ± 0.12	3.10 ± 0.11	3.58 ± 0.14
<b>Question control (CT)</b>								
Question-controlled CT 0	<b>2.87 ± 0.12</b>	2.84 ± 0.13	3.51 ± 0.10	2.46 ± 0.11	2.36 ± 0.09	2.76 ± 0.09	3.10 ± 0.10	3.49 ± 0.12
Question-controlled CT 1	2.82 ± 0.11	2.88 ± 0.11	3.42 ± 0.10	2.46 ± 0.12	2.47 ± 0.11	2.79 ± 0.13	3.14 ± 0.11	3.55 ± 0.10
Question-controlled CT 4	2.78 ± 0.12	2.88 ± 0.10	3.47 ± 0.11	2.40 ± 0.09	2.53 ± 0.13	2.83 ± 0.13	<b>3.24 ± 0.11</b>	3.59 ± 0.10
* Question-controlled CT 7	2.81 ± 0.10	<b>2.99 ± 0.11</b>	3.54 ± 0.09	2.35 ± 0.11	2.66 ± 0.12	2.92 ± 0.12	3.11 ± 0.10	3.47 ± 0.10
Question-controlled CT 10	2.67 ± 0.13	2.87 ± 0.11	3.52 ± 0.12	2.35 ± 0.12	2.63 ± 0.12	2.66 ± 0.10	2.94 ± 0.11	3.53 ± 0.12
Question-controlled CT 10 (boost)	2.68 ± 0.12	2.74 ± 0.09	3.42 ± 0.12	2.19 ± 0.13	<b>2.79 ± 0.11</b>	2.74 ± 0.11	3.00 ± 0.12	3.45 ± 0.13
<b>Specificity control (CT)</b>								
Specificity-controlled CT 0	2.79 ± 0.10	2.93 ± 0.09	3.44 ± 0.12	2.38 ± 0.11	2.56 ± 0.12	2.84 ± 0.12	3.12 ± 0.13	3.61 ± 0.11
Specificity-controlled CT 2	2.78 ± 0.12	2.74 ± 0.11	3.39 ± 0.13	2.31 ± 0.13	2.56 ± 0.13	2.74 ± 0.12	2.99 ± 0.11	3.47 ± 0.10
Specificity-controlled CT 4	2.82 ± 0.10	2.80 ± 0.13	3.44 ± 0.14	2.32 ± 0.13	2.51 ± 0.12	2.78 ± 0.15	3.09 ± 0.13	3.46 ± 0.13
Specificity-controlled CT 7	2.81 ± 0.12	2.91 ± 0.13	3.43 ± 0.11	2.45 ± 0.10	2.49 ± 0.11	2.81 ± 0.12	3.15 ± 0.12	3.55 ± 0.11
Specificity-controlled CT 9	2.80 ± 0.13	2.78 ± 0.10	3.41 ± 0.12	2.35 ± 0.13	2.28 ± 0.11	2.79 ± 0.11	2.91 ± 0.11	3.51 ± 0.12
<b>Specificity control (WD)</b>								
Specificity-controlled WD -10	2.76 ± 0.11	2.41 ± 0.12	3.19 ± 0.12	2.15 ± 0.11	2.28 ± 0.13	2.35 ± 0.12	2.89 ± 0.11	3.28 ± 0.12
Specificity-controlled WD -4	2.83 ± 0.10	2.76 ± 0.12	3.37 ± 0.10	2.36 ± 0.11	2.46 ± 0.11	2.62 ± 0.12	3.14 ± 0.09	3.52 ± 0.11
* Specificity-controlled WD 4	2.84 ± 0.10	2.96 ± 0.12	3.45 ± 0.13	2.44 ± 0.12	2.56 ± 0.09	<b>2.94 ± 0.11</b>	3.20 ± 0.10	3.54 ± 0.11
Specificity-controlled WD 6	2.81 ± 0.09	2.91 ± 0.10	3.34 ± 0.09	2.31 ± 0.11	2.53 ± 0.12	2.93 ± 0.12	3.09 ± 0.10	3.41 ± 0.12
Specificity-controlled WD 8	2.70 ± 0.11	2.39 ± 0.12	2.54 ± 0.12	1.80 ± 0.13	2.00 ± 0.10	2.49 ± 0.12	2.47 ± 0.10	2.87 ± 0.11
<b>Response-related control (WD)</b>								
Response-related controlled WD -10	2.77 ± 0.12	2.45 ± 0.12	3.26 ± 0.11	1.96 ± 0.10	2.31 ± 0.12	2.47 ± 0.12	2.73 ± 0.11	3.12 ± 0.12
Response-related controlled WD 0	<b>2.87 ± 0.12</b>	2.97 ± 0.11	3.55 ± 0.09	<b>2.62 ± 0.11</b>	2.48 ± 0.10	2.88 ± 0.12	3.21 ± 0.09	<b>3.70 ± 0.10</b>
Response-related controlled WD 5	2.79 ± 0.10	2.83 ± 0.09	3.35 ± 0.12	2.40 ± 0.12	2.51 ± 0.13	2.80 ± 0.13	3.13 ± 0.12	3.41 ± 0.12
Response-related controlled WD 10	2.74 ± 0.11	2.42 ± 0.12	2.93 ± 0.11	1.95 ± 0.12	2.20 ± 0.12	2.56 ± 0.12	2.90 ± 0.12	3.12 ± 0.10
Response-related controlled WD 13	2.63 ± 0.12	2.06 ± 0.11	2.40 ± 0.09	1.74 ± 0.11	2.07 ± 0.11	2.25 ± 0.12	2.49 ± 0.14	2.63 ± 0.10

Table 8: Calibrated scores (mean ± std.) for all models and human evaluation metrics.

The first eight columns are Likert metrics on a 1-4 scale (except Avoiding Repetition, which is a 1-3 scale), where higher is better (except Inquisitiveness, which has an optimal score of 3). The last column, Persona Retrieval, is on a scale from 0 to 1 where higher is better.

The maximum of each column (excluding Human row) is in bold.

Rows marked with \* are the six models included in Figure 3 (left) and Figure 4.

## H Plots of human evaluation results for all configurations

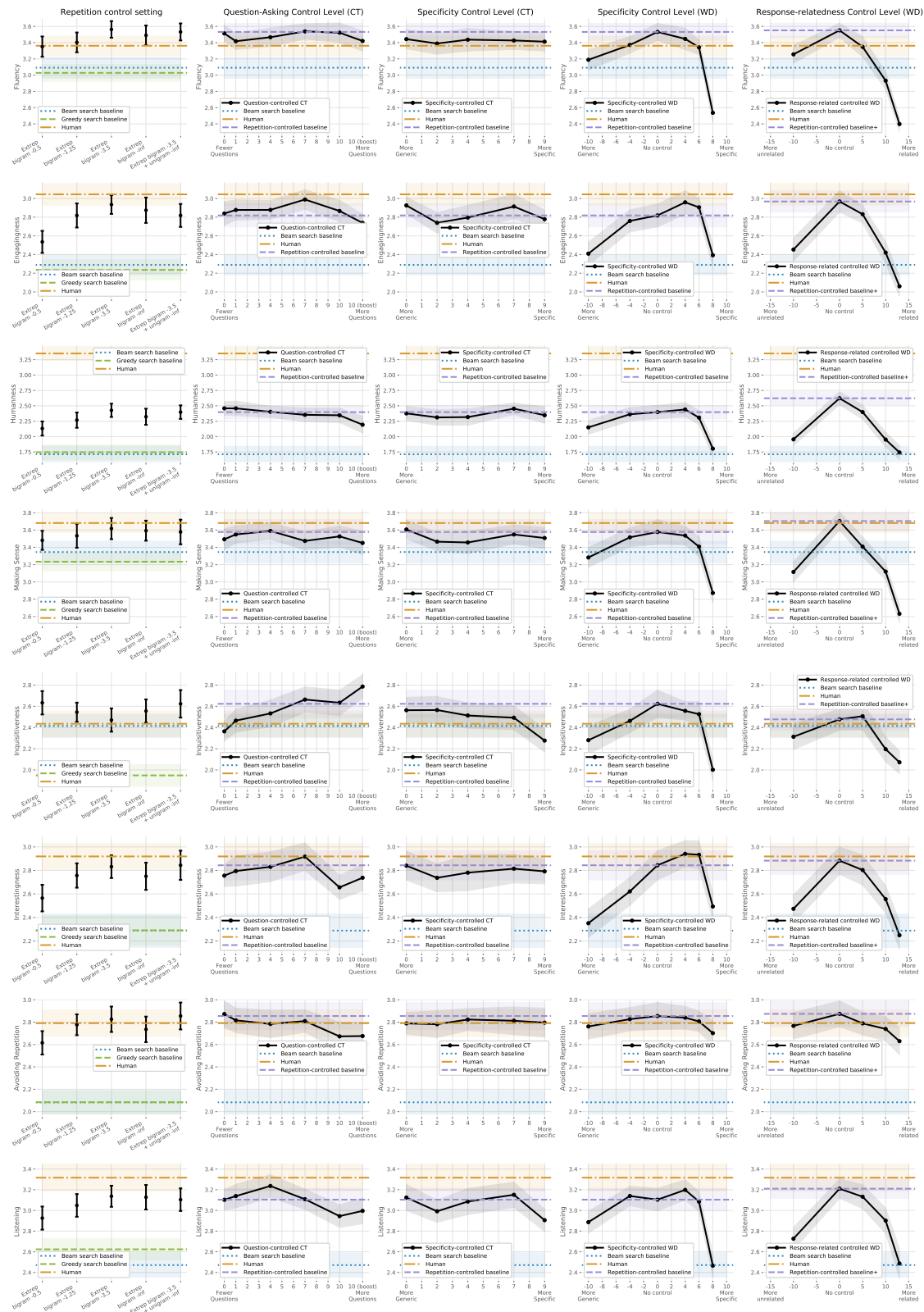


Figure 9: Calibrated human evaluation scores for all models. This is the same data as in Table 8.

Note: ‘Repetition-controlled baseline+’ in the rightmost column is ‘Response-related controlled WD 0’ in Table 8. See Table 5 for explanation.