How do decoding algorithms distribute information in dialogue responses?

Saranya Venkatraman
Pennsylvania State University
saranyav@psu.edu

He He
New York University
hhe@nyu.edu

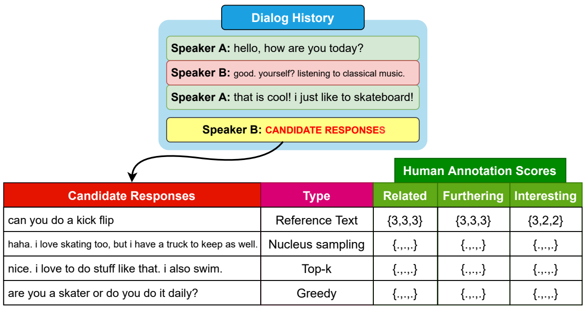
David Reitter
Google Research
reitter@google.com

Abstract

Humans tend to follow the Uniform Information Density (UID) principle by distributing information evenly in utterances. We study if decoding algorithms implicitly follow this UID principle, and under what conditions adherence to UID might be desirable for dialogue generation. We generate responses using different decoding algorithms with GPT-2 on the Persona-Chat dataset and collect human judgments on their quality using Amazon Mechanical Turk. We find that (i) surprisingly, model-generated responses follow the UID principle to a greater extent than human responses, and (ii) decoding algorithms that promote UID do not generate higher-quality responses. Instead, when we control for surprisal, non-uniformity of information density correlates with the quality of responses with very low/high surprisal. Our findings indicate that encouraging non-uniform responses is a potential solution to the "likelihood trap" problem (quality degradation in very high-likelihood text). Our dataset containing multiple candidate responses per dialog history along with human-annotated quality ratings is available at: https://huggingface.co/datasets/saranya132/dialog_uid_gpt2.

1 Introduction

The Uniform Information Density (UID) hypothesis states that humans distribute information in their utterances evenly for optimal communication (Jaeger, 2010; Fenk and Fenk, 1980). Consequently, language generation has benefitted from UID-based objectives and regularization (Meister et al., 2022; Wei et al., 2021). Specifically, Meister et al. (2020) argued that UID can be optimized for machine translation using beam search. Yet, the effect of different decoding algorithms on information density distributions of generated text are unknown, as is UID's broader role in neural response generation in the special case of dialogue models. Here, we investigate (i) if different decoding algorithms follow the UID principle, and (ii) if following the UID principle is beneficial for dialogue response generation, and (iii) collect human annotations of qualitative measures for multiple candidate responses to dialog histories generated using different decoding algorithms (Figure 1) to study the relationship of dialog response quality and UID. We operationalize UID as the variance of surprisal and measure its correlation with automatic metrics (e.g., BLEU, METEOR, BERTScore) as well as human judgments on qualitative measures of response quality and find that adherence to UID correlates negatively with human judgments when the responses have very low/high surprisal.


Dialog History
Speaker A: hello, how are you today?
Speaker B: good. yourself? listening to classical music.
Speaker A: that is cool! i just like to skateboard!
Speaker B: CANDIDATE RESPONSES
Human Annotation Scores

Candidate Responses	Type	Related	Furthering	Interesting
can you do a kick flip	Reference Text	{3,3,3}	{3,3,3}	{3,2,2}
haha. i love skating too, but i have a truck to keep as well.	Nucleus sampling	{.,.,.}	{.,.,.}	{.,.,.}
nice. i love to do stuff like that. i also swim.	Top-k	{.,.,.}	{.,.,.}	{.,.,.}
are you a skater or do you do it daily?	Greedy	{.,.,.}	{.,.,.}	{.,.,.}

Figure 1: Our dataset contains 4 candidate responses for every dialog history, along with human annotations for 3 qualitative measures.

Language production in humans. Spreading information content evenly in utterances is a marker of optimally strategized responses, and humans follow this UID principle as a means to state their thoughts clearly and to make themselves intelligible (Frank and Jaeger, 2008; Levy and Jaeger, 2007). The probability of a sentence has been associated with the cognitive load it incurs (Hale, 2003). As a means to avoid salient variations in the information content (surprisal, i.e., negative log probability) of responses, speakers maintain UID through linguistic choices such as that at the

Findings of the Association for Computational Linguistics: EACL 2023, pages 953–962
May 2-6, 2023 ©2023 Association for Computational Linguistics

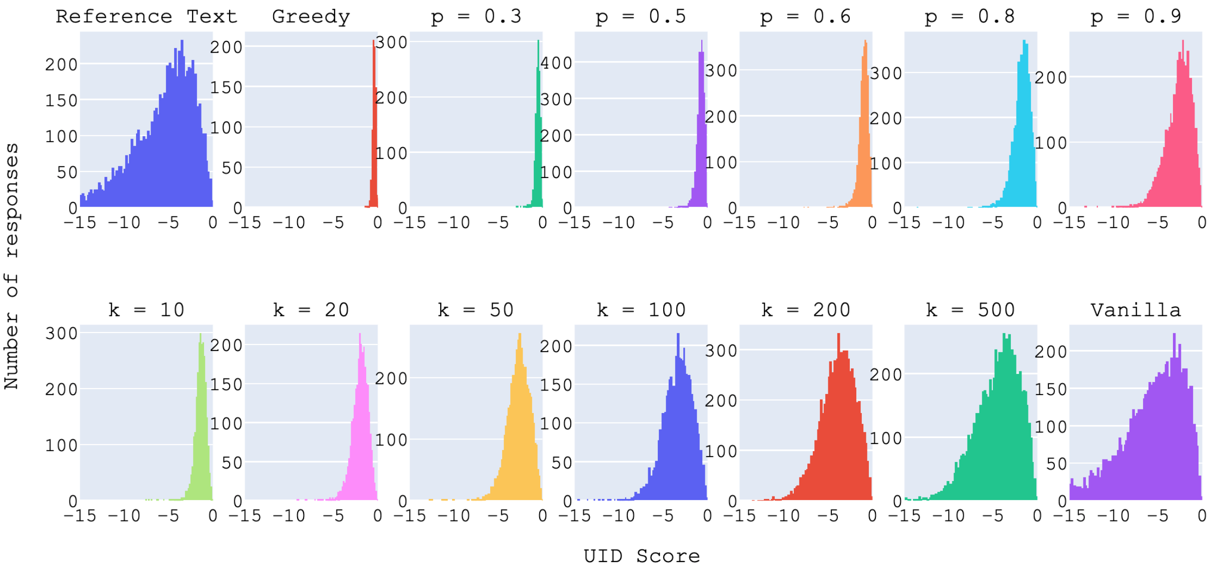


Figure 2: Histogram of **UID Scores** of responses generated using different decoding algorithms. The farther the UID score from 0, the less uniform or more non-uniform the response. Human-generated reference text (left-top) has a higher frequency of non-uniform responses as compared to any model setting as can be seen from the wider spread of scores away from 0. Also, as the values of p and k increase *(left to right)*, the information density distribution slowly approaches reference text-like non-uniformity.

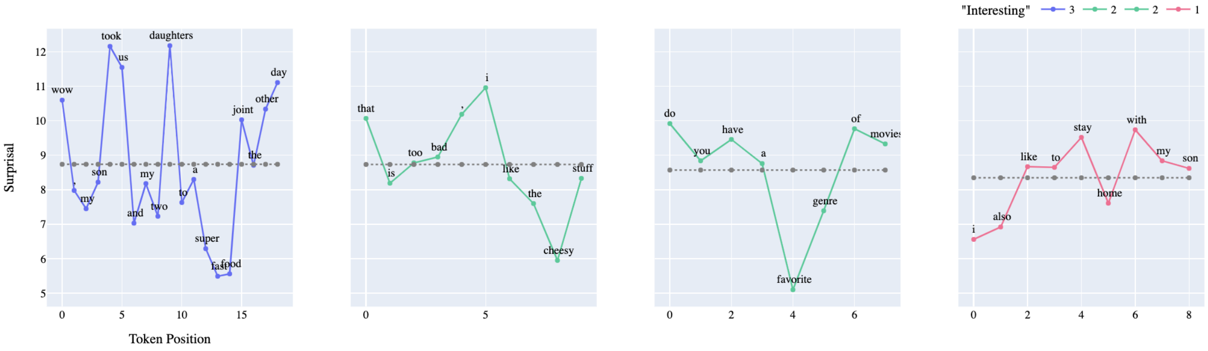


Figure 3: Surprisal at every token in candidate responses to the same dialog history, color-coded with human annotated **interesting** scores. Plots (*left to right*) are arranged in increasing order of uniformity (i.e. variance along y-axis). Less uniform the surprisal (left-most), better the score.

phonetic (Aylett and Turk, 2004), syntactic (Jaeger, 2010) and lexical level (Mahowald et al., 2013).

Response generation in machines. While large-scale pre-trained language models provide a rich prior for dialogue response generation, the choice of decoding algorithm used at the time of generation is crucial for the quality of generated responses (Holtzman et al., 2020; Zhang et al., 2021a; Nadeem et al., 2020; Golovanov et al., 2019; Oluwatobi and Mueller, 2020). While vanilla sampling often tends to produce incoherent text, greedy decoding leads to safe and repetitive responses. More recently, top-p/nucleus (Holtzman et al., 2020) and top-k sampling (Fan et al., 2018) are used to tune values of p/k to balance the diversity-quality trade-off (Zhang et al., 2021a; Li et al., 2016).

The UID principle and decoding algorithms. Both the UID principle and decoding algorithms can be seen as guiding mechanisms for dialogue response production in humans and generation in machines, respectively. UID's role in machine-generated dialogue is not well understood, with previous work mainly focused on machine translation and language modeling (Wei et al., 2021; Meister et al., 2021, 2020). To address this gap, we present a comparative study of decoding methods to develop a deeper understanding of the role of UID in dialogue response generation.

2 Experimental Details

2.1 Model & dataset

We use the fine-tuned GPT-2 (Radford et al., 2019) model provided by HuggingFace and use their data preprocessing and response generation scripts[1]. We used the Persona-Chat (Zhang et al., 2018) data split provided by the ConvAI2 challenge (Dinan et al., 2020)[2]. We then generated responses for 7500 dialogue histories randomly picked from 7801 validation set examples using vanilla, top-p, top-k sampling and greedy decoding.

Decoding algorithms. *Vanilla sampling* randomly picks the next token from the model's probability distribution, including many long-tail samples. *Top-k* samples from the k most probable tokens; *Greedy decoding* is Top-$k = 1$ decoding, always selecting the most probable next token. *Top-p (Nucleus)* sampling selects the next token from the top p portion of the probability mass.

2.2 Uniform Information Density score

We measure UID as the variance of the surprisal (negative log likelihood) of each token in the response (Jain et al., 2018; Wei et al., 2021; Meister et al., 2020). This measure is able to capture any sudden variations in the surprisal of the tokens in the sentence. UID Score is formulated as follows: the dialogue model learns a conditional probability p parameterized by θ to predict the next token (y_t) in the sentence. The surprisal (u) of the next token y_t is,

$$u(y_t) = -\log(p_\theta(y|x, y < t)), \tag{1}$$

for $t \geq 1$ where $y_0 =< EOS >$, t = time step, and x = dialogue context. Higher the surprisal, lower its probability and vice-versa. Thus, surprisal indicates how unexpected or surprising a token is in a given context. Average surprisal of a sentence (y) is defined as,

$$\mu(y) = \frac{1}{|y|} \sum_t (u(y_t)) \tag{2}$$

Finally, the *UID score* of a sentence (y) is defined as the negative normalized variance of the surprisal:

$$\text{UIDscore}(y) = -\frac{1}{|y|} \sum_t (u(y_t) - \mu)^2 \tag{3}$$

From this formulation, a perfectly uniform sentence would have a variance equal to 0 (i.e. the surprisal of every token in the sentence is equal). Since we take the negative of the variance, the higher the absolute value of UID score, the more non-uniform its information density.

2.3 Response evaluation

Automatic metrics. We measure the quality of responses using length (number of tokens), BLEU[3] (Papineni et al., 2002), METEOR[3] (Banerjee and Lavie, 2005), character level F-score (chrF)[3] (Popović, 2015), BLEURT[4] (Sellam et al., 2020), a RoBERTa (Liu et al., 2019) based text similarity score[5] (Reimers and Gurevych, 2019), BERTscore[4] (Zhang et al., 2019) and SacreBLEU[4] (Post, 2018).

Human evaluation. To study the effect of adherence to UID on the perceived quality of generated responses beyond n-gram, reference-based and learned automatic metrics, we collected human judgments along 3 measures – **related** (to the dialogue history), **furthering** (if a response keeps the conversation going/is encouraging for the dialogue partner) and **interesting** (if the response provides engaging/new information). We provide screenshots of the task interface (Figure 6), instructions (Figure 7) and details about the MTurk study design in Appendix A.

3 Findings

3.1 Information density of model responses

We plot the histograms of UID scores computed for all of the generated responses in Figure 2. The information densities of human-generated responses have a wider spread than responses produced by the models. Overall, the human-generated reference text has more non-uniform sentences than all model-generated responses. We notice a very high and narrow peak in the case of greedy decoding. This is not surprising as responses sampled using greedy search maximize the probability of the next token (minimize surprisal). Consequently, such responses would have very low surprisal at almost every word, hence lower variance. Vanilla

[1] https://github.com/huggingface/transfer-learning-conv-ai

[2] https://github.com/DeepPavlov/convai/tree/master/2018

[3] https://github.com/nltk/nltk/tree/develop/nltk/translate

[4] https://github.com/huggingface/datasets/tree/master/metrics

[5] https://github.com/UKPLab/sentence-transformers/blob/master/docs/usage/semantic_textual_similarity.md

	Pearson's *r* between UID score and automatic metrics							
Generation Type	**Length**	**BLEU**	**chrF**	**METEOR**	**BertScore**	**BLEURT**	**RoBERTa**	**SacreBLEU**
$p = 0.3$	-.10	.00	.14	.12	.17	.17	0.19	.13
$p = 0.5$	-.05	.03	.13	.10	.18	.17	**.2**	.15
$p = 0.6$	-.04	.06	.14	.13	.01	.06	.01	.00
$p = 0.8$	-.10	.03	.06	.05	.18	.16	**.2**	.15
$p = 0.9$	-.11	-.00	.03	.04	.16	.15	.19	.14
Greedy	-.14	.01	.14	.13	.06	.05	.06	.06
$k = 10$	-.04	.15	.03	.05	.07	.08	.07	.07
$k = 20$	-.05	.14	.05	.06	.05	.04	.06	.04
$k = 50$	-.09	.01	.03	.03	.06	.03	.03	.05
$k = 100$	-.07	.04	.00	.02	.11	.08	.08	.08
$k = 200$	-.12	.03	.02	.03	.06	.06	.04	.05
$k = 500$	-.09	.02	.04	.04	.10	.08	.08	.08
Vanilla	-.09	.01	-.00	.00	.07	.05	.05	.05

Table 1: Pearson's correlation coefficient (r) between **UID score and automatic metrics** of dialog responses generated using different decoding settings. All p-values < 0.05.

sampling uses the probability distribution learned from the training data, which might be why it is also closer to the validation set (reference text) distribution. With increase in p and k, we see that the information density distribution spreads across a larger range and includes more non-uniform responses, slowly approaching that of the reference text.

		Pearson's *r* between UID score and qualitative metrics		
Surprisal interval	**n**	**Related**	**Furthering**	**Interesting**
(0.8, 1.2)	24	.17	-.03	**-.30***
(1.2, 1.6)	64	.12	.08	-.13
(1.6, 2.0)	91	.05	**-.23***	-.07
(2.0, 2.4)	109	-.04	-.13	-.00
(2.4, 2.8)	111	-.06	**-.21***	-.05
(2.8, 3.2)	105	-.02	.01	-.10
(3.2, 3.6)	99	**-.23***	-.10	.19
(3.6, 4.0)	66	.03	-.05	-.09
(4.0, 4.4)	42	-.33	-.22	-.09
(4.4, 4.8)	24	-.14	**-.61***	.04
(4.8, 5.2)	12	-.33	-.14	**-.54***
(5.2, 5.6)	13	**-.98***	-.64	-.38

Table 2: Pearson's r between **UID score and and human judgments** of qualitative measures for dialog responses bucketed by surprisal [Surprisal interval = the ranges of surprisal values used for bucketing responses, n = number of responses in each surprisal interval, *p-value < .05]

3.2 UID score & automatic metrics

We present the correlation between UID scores and automatic metrics calculated for the generated dialogue responses in Table 1. UID scores have a weak correlation with RoBERTa-based similary scores for two settings of nucleus sampling. Other than that, UID scores are not correlated with automatic metrics of response generation. We take this to be an indication that if UID scores do capture any aspect of response quality, it goes beyond what is measured by such metrics and might provide for a better evaluation criteria.

3.3 UID score & human Judgments

Motivated by the fact that UID score is derived from surprisal, we test if surprisal is a confounding factor and find that, indeed, UID scores were highly correlated with average surprisal (Table 3). To tease apart the effect of UID scores on response quality, we controlled for surprisal by grouping or bucketing responses into 12 intervals of surprisals (within a range of 0.4 units as shown in the first column on Table 2). Within these intervals, surprisal had no correlation with generation quality (Table 5). Once we control for surprisal i.e. analyse dialog responses with similar surprisals but varying UID scores, we observe that UID scores negatively correlate with human judgments, to varying degrees of strength, for responses in very low or high surprisal intervals (see Table 2). Thus, for the extremities of the surprisal range, UID scores indicate that better rated responses are non-uniform.

4 Discussion

Contrary to our expectations, we find non-uniformity to be a more desirable property in machine-generated responses. Overall, UID scores and surprisal do not correlate with human judgments (Table 4). But when controlled for surprisal, we observe that UID score is correlated with human judgments for certain intervals (examples in Figure 3 and Table 6). Our results suggest that optimizing UID to generate uniform text might not be the right objective for regularizing decoding algorithms. Instead we find that non-uniform information density could be a potential solution to the “likelihood trap" problem according to which models generate lower quality text (as per human judgments) when sampling from the extremities of their likelihood space (Zhang et al., 2021b). Consequently, we suggest that decoding algorithms be tuned to follow the information density patterns of human-generated non-uniform data when generating responses outside of the “safe” likelihood range as a means to generate higher quality responses across the entire likelihood space.

5 Limitations

While we present a study of multiple decoding settings, we generate all machine responses using the same transformers based model architecture. Thus, the presented work does not yet explore individual differences between different model architectures. Additionally, due to limited resources we were not able to collect large-scale human annotations across multiple corpora and acknowledge the same as part of future efforts.

6 Ethical considerations

In this work, we collected human annotations on dialogue response quality using MTurk. Each HIT in our MTurk study contained one dialogue history and four candidate responses. The annotators could read the history and rate the responses that followed using mouse clicks on their response choices. We provided an additional feedback field for annotators to write comments in. We received very positive feedback on the task from all the annotators who used this feature. There were no restrictions on the minimum or maximum number of examples the annotators had to rate. From a pilot study on MTurk, we found the average time to complete one HIT to be slightly under 2.5 minutes. After considering the average time required and the task difficulty (expressed to be clearly and easily understood by annotators in their comments) we set the payment amount to $0.5 per HIT for an hourly rate of about $12 per hour.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS’18 Competition*, pages 187–208. Springer.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

August Fenk and Gertraud Fenk. 1980. constancy in short-term memory-constancy in linguistic information flowß. *journal for experimental and applied psychology*, 27(3):400–414.

Austin F Frank and T Florain Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Ayush Jain, Vishal Singh, Sidharth Ranjan, Rajakrishnan Rajkumar, and Sumeet Agarwal. 2018. Uniform Information Density effects on syntactic choice in Hindi. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 38–48, Santa Fe, New-Mexico. Association for Computational Linguistics.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *arXiv preprint arXiv:2202.00666*.

Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. 2020. A systematic characterization of sampling algorithms for open-ended language generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 334–346, Suzhou, China. Association for Computational Linguistics.

Olabiyi Oluwatobi and Erik Mueller. 2020. DLGNet: A transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 54–62, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5191–5202.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021a. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021b. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Generation Type	Pearson’s r
Reference Text	-.69
Greedy	-.23
$p = 0.3$	-.43
$p = 0.5$	-.50
$p = 0.6$	-.56
$p = 0.8$	-.65
$p = 0.9$	-.68
$k = 10$	-.40
$k = 20$	-.45
$k = 50$	-.56
$k = 100$	-.63
$k = 200$	-.65
$k = 500$	-.69
Vanilla	-.74

Table 3: Pearson’s correlation coefficient (r) **between UID score and average sentence surprisal** (all $p < 0.01$)

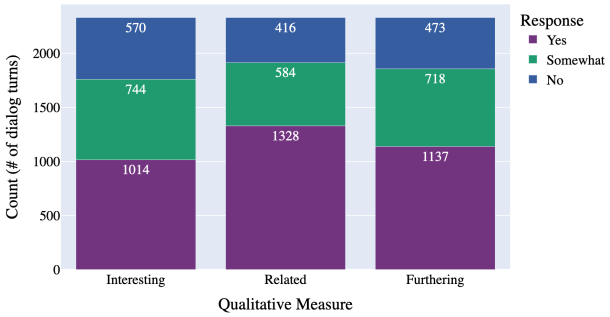


Figure 4: Frequency of responses (Yes/Somewhat/No) for each qualitative measure in our human annotated dataset.

	Pearson’s r	
Quality	**UID Score**	**Surprisal**
Related	.01	**-.13***
Furthering	.03	**-.10***
Interesting	-.04	-.01

Table 4: Pearson’s correlation coefficient (r) of **UID score and surprisal with human judgments of qualitative metrics** (*p<0.01)

			Pearson’s r	
Surprisal interval	**n**	**Related**	**Furthering**	**Interesting**
(0.8,1.2)	24	-.03	-.04	-.00
(1.2,1.6)	64	-.10	-.16	.08
(1.6,2.0)	91	.05	.14	.10
(2.0,2.4)	109	-.14	-.08	**-.27***
(2.4,2.8)	111	-.12	.05	.09
(2.8,3.2)	105	-.02	.06	-.00
(3.2,3.6)	99	-.13	.12	.01
(3.6,4.0)	66	.02	-.06	.06
(4.0,4.4)	42	-.01	-.00	.06
(4.4,4.8)	24	.20	.34	.23
(4.8,5.2)	12	-.13	-.37	-.12
(5.2,5.6)	13	.60	.83	.76

Table 5: Pearson’s r between **surprisal and human judgments** of qualitative measures for dialog responses bucketed by surprisal [Surprisal interval = the ranges of surprisal values used for bucketing responses, n = number of responses in each surprisal interval, *p-value < .05]

A Human evaluation study details

Raters were selected based on the criteria that they be located in the US, and had attempted a minimum of 500 HITS at an accepted work rate greater than 97% on MTurk. We asked raters on MTurk to answer if a candidate response satisfied each of the qualitative measures (interesting, furthering and related) and gave them three response options: "Yes", "Somewhat" and "No". In a pilot study of 360 responses, we also included a measure for fluency. All of the responses were rated “Yes" by majority vote and we removed this measure from further analysis as all the generations in this study were fluent as indicated by the pilot study and from our observation. For correlation calculations, we assign integer score values to each of the three response options as 3 for "Yes", 2 for "Somewhat" and 1 for "No". Thus, the higher the score, the better the response is rated. Following the pilot study, for 194 dialogue histories, we showed the raters 4 candidate dialogue responses (total of 776 dialogue responses) and collected ratings on all *3* measures from *3* raters per dialogue history. In all, we obtained a total of 776*3, i.e., 2328 total response-rating pairs. To calculate the score for each response along every measure, we take the mean of all ratings as the score. For cases where at least 2 out of 3 raters agree, we take majority vote as the final score. This constituted (2018 out of 2328) 86.68% of all the ratings collected. We show the overall distribution of qualitative scores for all the response-rating pairs in Figure 4. We verified the rater responses by checking if they were rating human-generated responses highly as those came from a trusted source (Persona-Chat). We also manually inspected a random subset of dialog history-candidate response sets and found the results to be in accordance with our intuitions.

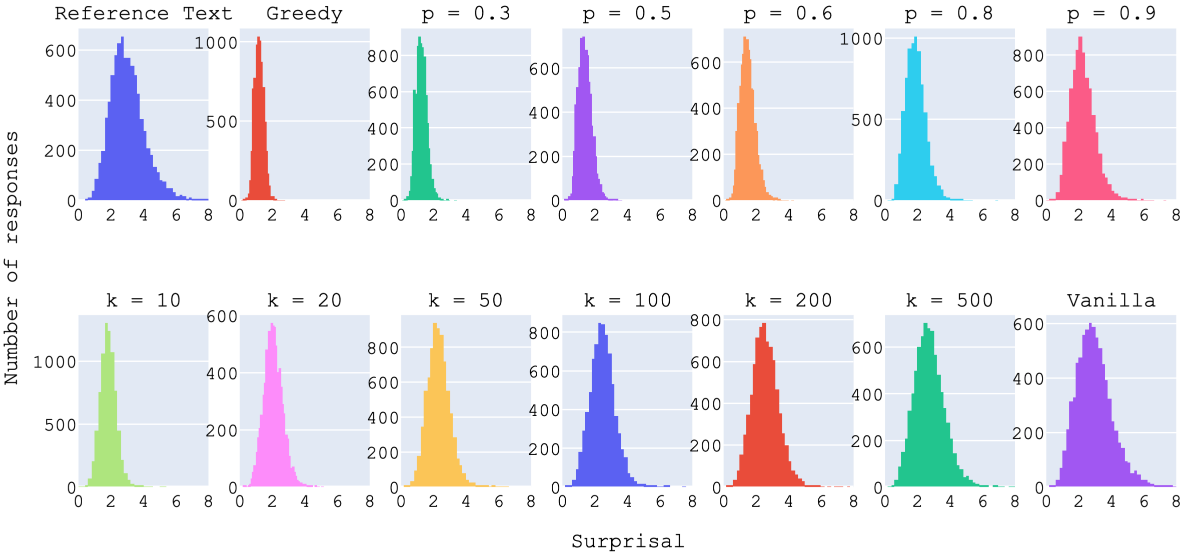
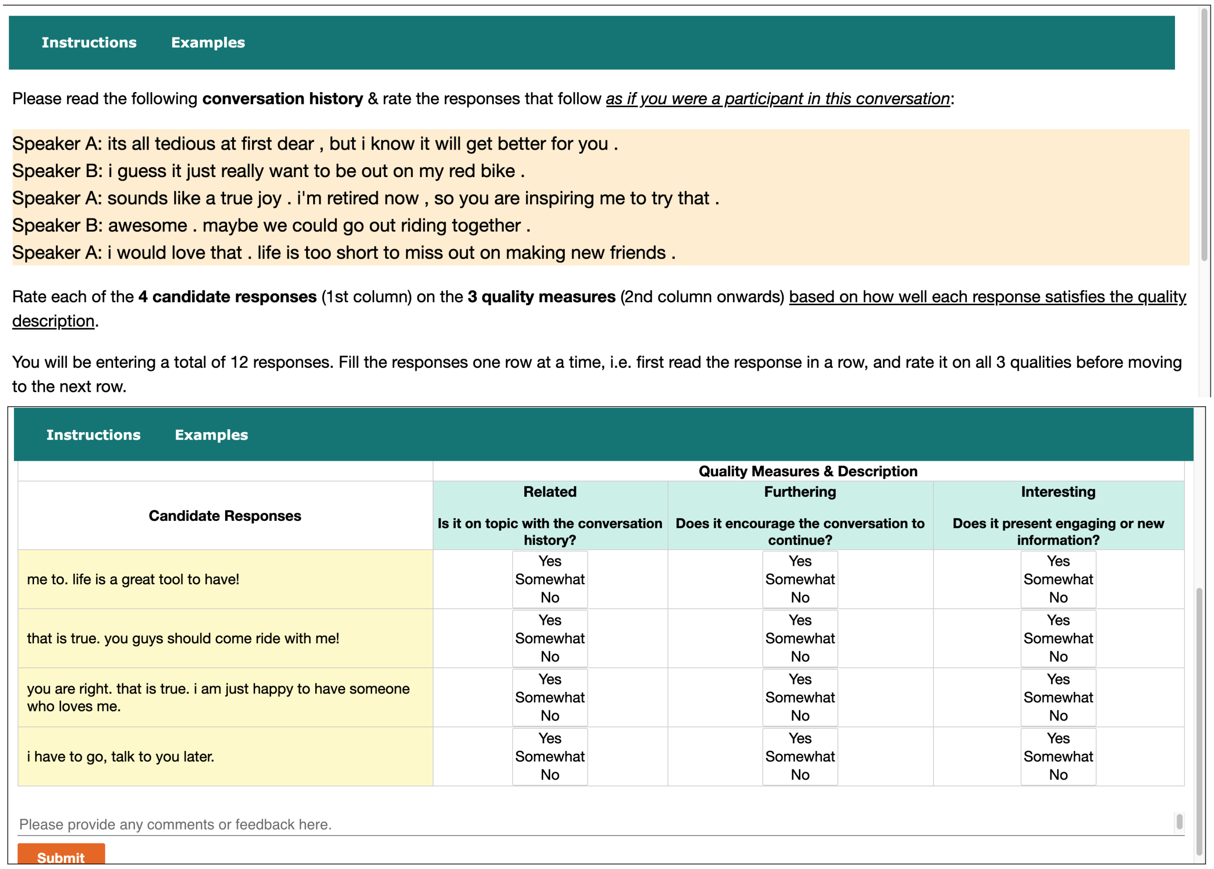


Figure 5: Histograms of **average sentence surprisal** for responses generated using different decoding settings and human-generated reference text (left-top).



Please read the following **conversation history** & rate the responses that follow *as if you were a participant in this conversation*:

Speaker A: its all tedious at first dear , but i know it will get better for you .
Speaker B: i guess it just really want to be out on my red bike .
Speaker A: sounds like a true joy . i'm retired now , so you are inspiring me to try that .
Speaker B: awesome . maybe we could go out riding together .
Speaker A: i would love that . life is too short to miss out on making new friends .

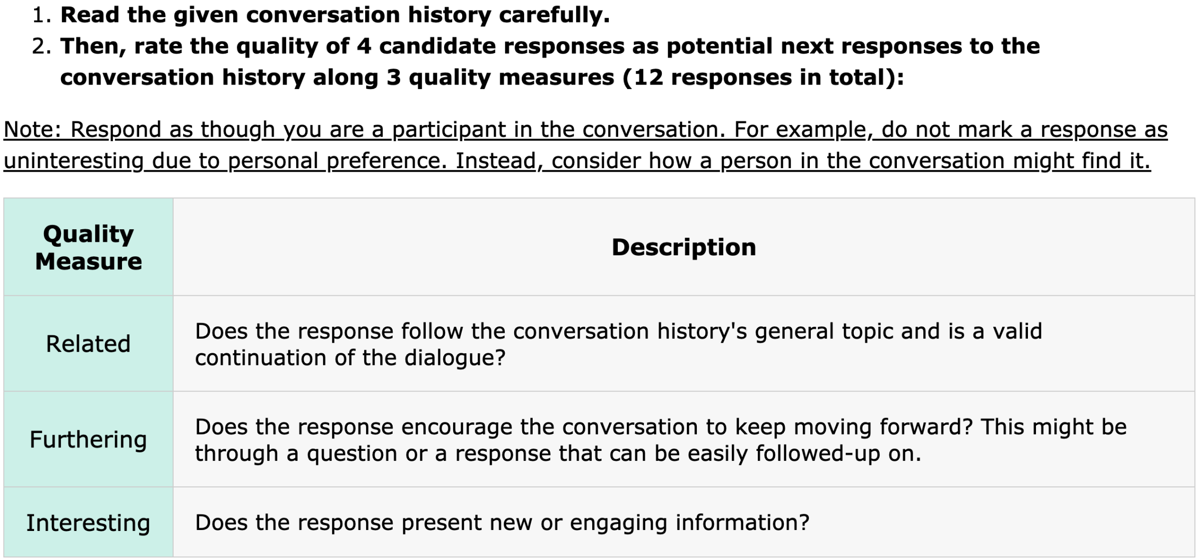
Rate each of the **4 candidate responses** (1st column) on the **3 quality measures** (2nd column onwards) based on how well each response satisfies the quality description.

You will be entering a total of 12 responses. Fill the responses one row at a time, i.e. first read the response in a row, and rate it on all 3 qualities before moving to the next row.

	Quality Measures & Description		
Candidate Responses	**Related** **Is it on topic with the conversation history?**	**Furthering** **Does it encourage the conversation to continue?**	**Interesting** **Does it present engaging or new information?**
me to. life is a great tool to have!	Yes Somewhat No	Yes Somewhat No	Yes Somewhat No
that is true. you guys should come ride with me!	Yes Somewhat No	Yes Somewhat No	Yes Somewhat No
you are right. that is true. i am just happy to have someone who loves me.	Yes Somewhat No	Yes Somewhat No	Yes Somewhat No
i have to go, talk to you later.	Yes Somewhat No	Yes Somewhat No	Yes Somewhat No

Please provide any comments or feedback here.

Figure 6: Screenshots of our MTurk study interface for collecting human judgments on 4 candidate responses per dialogue history, along 3 quality measures.

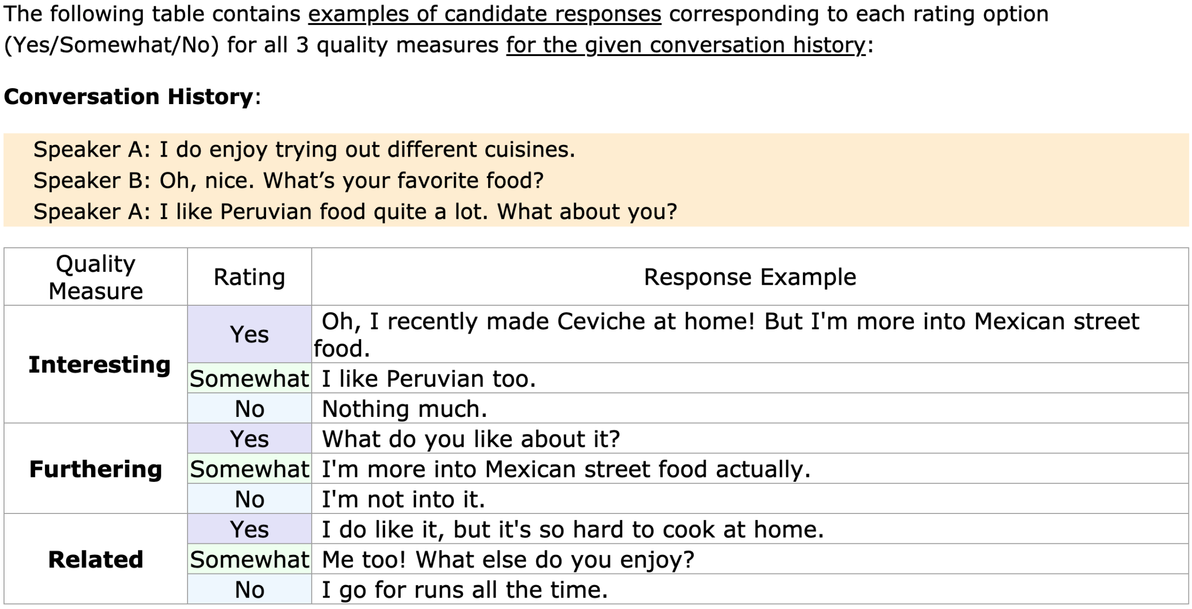


1. **Read the given conversation history carefully.**
2. **Then, rate the quality of 4 candidate responses as potential next responses to the conversation history along 3 quality measures (12 responses in total):**

Note: Respond as though you are a participant in the conversation. For example, do not mark a response as uninteresting due to personal preference. Instead, consider how a person in the conversation might find it.

Quality Measure	**Description**
Related	Does the response follow the conversation history's general topic and is a valid continuation of the dialogue?
Furthering	Does the response encourage the conversation to keep moving forward? This might be through a question or a response that can be easily followed-up on.
Interesting	Does the response present new or engaging information?

(a) Detailed instructions that MTurk raters could expand at any time.



The following table contains examples of candidate responses corresponding to each rating option (Yes/Somewhat/No) for all 3 quality measures for the given conversation history:

Conversation History:

Speaker A: I do enjoy trying out different cuisines.
Speaker B: Oh, nice. What's your favorite food?
Speaker A: I like Peruvian food quite a lot. What about you?

Quality Measure	Rating	Response Example
Interesting	Yes	Oh, I recently made Ceviche at home! But I'm more into Mexican street food.
	Somewhat	I like Peruvian too.
	No	Nothing much.
Furthering	Yes	What do you like about it?
	Somewhat	I'm more into Mexican street food actually.
	No	I'm not into it.
Related	Yes	I do like it, but it's so hard to cook at home.
	Somewhat	Me too! What else do you enjoy?
	No	I go for runs all the time.

(b) Examples responses for each measure and rating category shown to MTurk raters.

Figure 7: Instructions and examples from MTurk study.

Dialogue history

Speaker A: hi, i m susie. how are you?

Response type	Response text	UID Score	Interesting
Reference Text	i’m good. just got back from some volunteering. how are you doing?	-1.99	3
Candidate 1	hi. i am very good. just enjoying my favorite pastime. how are you?	-1.43	2
Candidate 2	i am great! i volunteer at a soup kitchen and watch movies with my kids.	-1.19	2
Candidate 3	i’m doing well. how are you?	-0.18	1

Speaker A: that’s cool. my dad made us italian food tonight.
Speaker B: oh nice, i love italian food. my favorite is the garlic bread. what is yours?
Speaker A: pasta, my son loves pizza though!
Speaker B: i like that too, have to eat lots of carbs for my training.

Response type	Response text	UID Score	Interesting
Candidate 1	wow, my son took us and my two daughters to a super fast food joint the other day	-4.83	3
Candidate 2	that is too bad, i like the cheesy stuff.	-2.08	2
Reference Text	do you have a favorite genre of movies?	-1.73	2
Candidate 3	i also like to stay home with my son.	-0.88	1

Speaker A: hello i hope your sunday is great , what is your favorite kind of music?
Speaker B: hey there . been a relaxed sunday . yours ? music eclectic.
Speaker A: my sunday has been exciting ! i enjoy death metal.
Speaker B: death metal . cool . i spent the morning volunteering.
Speaker A: i volunteer too , at the local pool to be a swim coach.

Response type	Response text	UID Score	Furthering
Candidate 1	nice. such a nice day. how long have you been coaching?	-9.99	3
Reference Text	that is great! we both volunteer! mine is rescuing bunnies.	-7.59	2
Candidate 2	that’s cool. i donate my pay to the local zoo. humane societies.	-3.89	2
Candidate 3	sick sick. beautiful color, navy blue is my favorite.	-2.79	1

Speaker A: have you heard about the juggalos? weird.

Response type	Response text	UID Score	Furthering
Reference Text	what are those? do they juggle balls?	-6.88	3
Candidate 1	no i have not. i like that band myself	-5.19	2
Candidate 2	we will not tell. do you like comics?	-4.27	2
Candidate 3	i have not heard of them. i’ve dolls too.	-2.8	1

Speaker A: hello, how are you today?
Speaker B: good. yourself? listening to classical music.
Speaker A: that is cool! i just like to skateboard!

Response type	Response text	UID Score	Related
Reference Text	can you do a kick flip?	-14.11	3
Candidate 1	haha. i love skating too, but i have a truck to keep as well.	-11.28	3
Candidate 2	nice. i love to do stuff like that. i also swim.	-10.25	3
Candidate 3	are you a skater or do you do it daily?	-9.05	2

Speaker A: i hope it is a healthy snack. i am doing well, just moved here.
Speaker B: it is. i am vegan. just moved here too! where from?
Speaker A: i too am vegan from germany. have you seen lafer! lichter! lecker! on tv?
Speaker B: oh yes! i love shows like that and watched it back home a lot.

Response type	Response text	UID Score	Related
Reference Text	they do amazing things with the hummus. where is back home for you?	-6.86	3
Candidate 1	you must have a lot of fun watching them.	-4.53	2
Candidate 2	they have the best new vegan cookbooks, but i am more adventurous.	-2.93	3
Candidate 3	i do love the sky diving, too! i have seen the first few seasons.	-2.47	1

Table 6: Examples of dialogue histories followed by 4 response candidates arranged by increasing UID score i.e. from more non-uniform to uniform responses and their corresponding human judgment scores.