Evaluating the Evaluation of Diversity in Natural Language Generation

Guy Tevet¹,²  Jonathan Berant¹,³

¹School of Computer Science, Tel-Aviv University
²Department of Electrical Engineering, Tel-Aviv University
³Allen Institute for AI
{guytevet@mail, joberant@cs}.tau.ac.il

Abstract

Despite growing interest in natural language generation (NLG) models that produce diverse outputs, there is currently no principled method for evaluating the diversity of an NLG system. In this work, we propose a framework for evaluating diversity metrics. The framework measures the correlation between a proposed diversity metric and a diversity parameter, a single parameter that controls some aspect of diversity in generated text. For example, a diversity parameter might be a binary variable used to instruct crowdsourcing workers to generate text with either low or high content diversity. We demonstrate the utility of our framework by: (a) establishing best practices for eliciting diversity judgments from humans, (b) showing that humans substantially outperform automatic metrics in estimating content diversity, and (c) demonstrating that existing methods for controlling diversity by tuning a “decoding parameter” mostly affect form but not meaning. Our framework can advance the understanding of different diversity metrics, an essential step on the road towards better NLG systems.

1 Introduction

An important desideratum of natural language generation (NLG) systems is to produce outputs that are not only correct, but also diverse. For example, a dialog system (Adiwardana et al., 2020) should permit many responses for the prompt “How are you today?”. Similarly, we expect diverse responses in NLG tasks such as story generation (Li et al., 2018), question generation (Pan et al., 2019) and abstractive question answering (Fan et al., 2019).

Despite growing effort to produce more diverse models (Li et al., 2016c,a; Holtzman et al., 2019; Du and Black, 2019), there is currently no standard evaluation metric for measuring model diversity. Thus, different papers evaluate diversity differently (if at all), making it difficult to fairly compare competing approaches (Hashimoto et al., 2019). Having a principled and consensual diversity evaluation metric is hence fundamental for advancing the field of NLG.

A key challenge in developing diversity evaluation metrics, is that it is difficult to determine their efficacy. Unlike metrics for evaluating the
quality of generated text, where one can measure the correlation between an automatic metric (such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005)) and human judgement (Zhang et al., 2019a; Sagarkar et al., 2018), it is unknown whether humans can reliably estimate diversity.

In this paper, we propose a framework for evaluating diversity metrics (see Figure 2). We assume that a tester (human or model) is generating sets of sentences, conditioned on some diversity parameter that controls the diversity of the output sentences. We evaluate the diversity of the sentences using a proposed diversity metric, and measure the correlation between the proposed metric and the diversity parameter. High correlation indicates that the metric indeed captures how the diversity parameter affects the model output.

We instantiate this framework with two tests. In the first test, the tester is a neural generation model and the diversity parameter is a decoding parameter, such as softmax temperature (Ackley et al., 1985). This parameter controls the skewness of the distribution in every generated token, and is known to affect model diversity (Holtzman et al., 2019; Caccia et al., 2018). In the second test (see Figure 1), the tester is a human, and the diversity parameter is a binary variable, where the human is instructed to generate sets of sentences with either high or low diversity in content.

We evaluate several families of popular diversity metrics with these two tests: (a) n-gram-based metrics that estimate diversity based on surface patterns in a set of generated sentences, (b) neural metrics: we propose a reduction from evaluating sentence similarity to evaluating diversity, then evaluate diversity using state-of-the-art sentence similarity models, and (c) human evaluation: we explore multiple ways in which humans can be asked to estimate diversity, resulting in multiple Human Diversity Score (HDS) variations.

We find that n-gram-based metrics succeed in detecting diversity that is driven by decoding parameters (the first test above), suggesting that such parameters mostly control the form of generated text rather than its content. Conversely, n-gram-based metrics perform poorly in the second test, which focuses on diversity of content. While neural metrics outperform n-gram-based metrics, we establish that humans are substantially better than any automatic metric at detecting content diversity. This is illustrated in Figure 1, where a human score clearly distinguishes between sets that have high (blue) and low (orange) content diversity, while n-gram-based metrics fail to do so.

To conclude, our main contributions are:

- A framework for evaluating diversity metrics.
- Tests instantiating this framework, measuring the sensitivity of metrics to content and form.
- Best practices for obtaining diversity evaluations from crowdsourcing workers.
- Establishing that humans outperform current automatic metrics in detecting content diversity.
- The collected data, test scores and code are publicly available,\(^1\) and can be used to easily compare new diversity metrics to existing results in our framework.

2 Background: Diversity Evaluation

Recently, interest in diversity in NLG has increased (Du and Black, 2019; Holtzman et al., 2019; Hashimoto et al., 2019; Dušek et al., 2020), resulting in multiple proposals for its evaluation. We describe recent approaches, highlighting the need for a standard way to evaluate metrics.

Perplexity is the standard metric in language modeling (LM), measuring the proximity of a LM, \(P_{LM}\), to the true distribution, \(P_{ref}\), by empirically approximating the cross-entropy \(H(P_{ref}, P_{LM})\) with held-out data sampled from \(P_{ref}\). Thus, perplexity captures to some extent diversity. For example, a dialog model that puts all probability mass on the output “I don’t know” for any given context will obtain infinite perplexity once it encounters any other response. This property makes perplexity popular in LM-based NLG models, and often it is the only reported measure for diversity (Lewis et al., 2017; Fan et al., 2018; Wang et al., 2019; Li et al., 2019).

However, perplexity does not purely measure diversity, and high perplexity does not entail low diversity. For example, a LM with a uniform distribution over the vocabulary for each decoded token has high diversity, but its perplexity will be extremely high, due to its low quality. Moreover, perplexity evaluates a LM, while the diversity of an NLG system is also strongly affected by the decoding procedure. For example, Top-k and nucleus sampling are popular decoding schemes that trade-
off quality and diversity by ignoring some of the LM probability mass (Holtzman et al., 2019).

Last, some NLG models, such as Generative Adversarial Networks (GANs) (Yu et al., 2017) are not based on a LM at all. While it is possible to approximate perplexity for such models (Tevet et al., 2019), a metric should ideally not be tied to model specifics.

**N-gram-based metrics** A popular metric is distinct n-grams (Li et al., 2016b), which computes the proportion of unique n-grams out of the total number of n-grams in a set of generated sentences. For example, distinct unigrams is the ratio of word types to word tokens, alluding to the richness of the vocabulary. Dušek et al. (2020) calculated Shannon entropy (Manning et al., 1999) based on different n-grams as a measure of lexical diversity. Self-BLEU (Zhu et al., 2018; Shu et al., 2019) measures the BLEU score of a generated sentence with respect to another generated sentence (rather than a gold reference). High average Self-BLEU indicates high similarity between generated sentences and low diversity. In §5 we expand this idea and suggest a reduction from any similarity metric to a diversity metric. By design, n-gram based metrics are sensitive to diversity in the form of language, rather than its meaning.

**Embedding-based metrics** A new line of metrics suggests to embed generated sentences in latent space, then evaluate them in this space. Du and Black (2019) suggest to cluster the embedded sentences with k-means, then use its inertia as a measure for diversity. Recently, Lai et al. (2020) suggested to consider the volume induced by the embedded sentences as a diversity metric.

**Human evaluation** Yang et al. (2019) asked humans to evaluate the internal diversity of a generated essay. Ghandeharioun et al. (2019) let crowdsourcing workers interact with a dialog chat-bot, then asked them to evaluate the diversity of a single conversation. In contrast, this paper focuses on the diversity of different responses given a context, as in Zhang et al. (2019b).

To conclude, increasing interest in diversity resulted in multiple proposed diversity metrics. However, there is no consensus on how to evaluate diversity and what each metric actually measures.

### 3 Evaluating Diversity Metrics

We now describe our framework for evaluating diversity metrics. We note that diversity has many facets (see discussion in §5): for instance, a set of sentences can be diverse in terms of their content, while another may have similar content, but diverse form (see Figure 1). Our framework provides a way to evaluate metrics for different aspects of diversity under moderate assumptions.

We define a diversity metric $m_{\text{div}}(S_c) \in \mathbb{R}$ as a function that takes a set of generated responses $S_c$ as an input, and outputs a diversity score. Each response $s \in S_c$ is generated for the same input context $c$, hence $S_c$ is a sample from a generative distribution $P_{\text{gen}}(s \mid c)$. The overall diversity score of a generative model can be obtained by averaging $m_{\text{div}}$ over sets $S_c$ sampled from the model given multiple contexts $c \in C$.

To evaluate $m_{\text{div}}(\cdot)$, our framework assumes access to some deterministic diversity parameter $d$ that controls an aspect of diversity in $S_c$. Our framework tests the relation between $m_{\text{div}}$ and the parameter $d$. By varying $d$ and measuring $m_{\text{div}}$, we can compute the correlation $\rho$ between $m_{\text{div}}$ and an aspect of diversity, represented by $d$. Because our goal is to measure the ability of metrics to rank the diversity level of generated text, we use Spearman’s $\rho$ rank correlation as our test score. Figure 2 illustrates the flow of a test in our framework.

In practice, to control the diversity level of $S_c$
using $d$, we use a tester: a generative model that takes a context $c$ and a diversity parameter $d$ as input, and outputs a response set $S_{c,d}$. We stress that the tester can be either a neural model or a human. A good tester should reliably represent the diversity level quantified by $d$.

As a hypothetical example, $c$ can be a movie name and $d$ can represent sentiment diversity, that is, the number of different sentiments in a collection of generated reviews $S_c$ about that movie. A human tester can observe $c$ and $d$, and produce reviews accordingly (such data can be easily mined from IMDB). A collection of such $(d, S_{c,d})$ makes a test, in which Spearman’s $\rho$ correlation between $m_{\text{div}}(S_{c,d})$ and $d$ is a measure for the sensitivity of $m_{\text{div}}$ to sentiment diversity.

We note that perplexity cannot be evaluated as a diversity metric in our framework, because it requires a sample from $P_{\text{ref}}$, while we assume a response set sampled from $P_{\text{gen}}$.

We now describe two tests that instantiate this framework, roughly corresponding to the two main aspects of diversity: form diversity and content diversity.

### 3.1 Test #1: Decoding Parameters

The diversity of a NLG system constructed from a LM and a decoder is dependent on the decoding scheme. For example, beam search approximates the most probable output, and thus dramatically reduces diversity. Conversely, pure sampling from the LM distribution leads to high diversity, but low quality output (Holtzman et al., 2019).

Consequently, a popular method to control diversity in NLG systems is to vary some decoding parameter. Variations include (a) softmax temperature (Ackley et al., 1985), where a temperature parameter $\tau$ controls the skewness of the softmax distribution at each step, (b) Nucleus (Top-$p$) sampling (Holtzman et al., 2019), where one samples at each step from the minimal set of most probable tokens whose cumulative probability is at least $p$, and (c) Top-$k$ sampling, which samples from the top-$k$ most probable tokens at each step. All methods skew the LM distribution in a way that avoids low-probable tokens and leads to higher quality (Holtzman et al., 2019), providing a decoding parameter that trades off quality and diversity (Caccia et al., 2018).

In Test #1, we define the tester to be a strong LM, such as GPT-2 (Radford et al., 2019), and the diversity parameter $d$ to be a decoding parameter such as temperature. We check how different diversity metrics $m_{\text{div}}$ correlate with decoding parameters. This can shed light both on the quality of the metrics, but also on how decoding parameters actually affect the output of a NLG system.

### 3.2 Test #2: Content Diversity

In this test, our goal is to evaluate how different diversity metrics capture the notion of content diversity, that is, whether a set of responses are diverse in terms of their content. Measuring content diversity requires deep understanding of the semantics of responses in $S_c$.

To isolate content diversity from form diversity, we aim to generate sets of responses with a similar level of form diversity, but where the level of content diversity is controlled by the diversity parameter $d$. To do this, we use crowdsourcing workers as testers, and a binary diversity parameter $d \in \{0, 1\}$, corresponding to low or high content diversity. A worker observes a context $c$ and produces a set of responses $S_c$ based on the value of $d$. We encourage workers to use different words and phrases in different responses regardless of the value of $d$, such that form diversity is generally high in all examples. Examples from this data are presented in Figure 1 and Appendix B.

In §6, we will focus on whether automatic diversity metrics can perform as well a humans on the task of estimating content diversity.

### 4 Human Diversity Score

One of the core questions we tackle is:

Can humans evaluate diversity reliably?

Although a few papers (Ghandeharioun et al., 2019; Yang et al., 2019; Zhang et al., 2019b) asked humans to evaluate the diversity of their models, to the best of our knowledge no work thoroughly investigated this question. The importance of this question is clear when comparing quality evaluation in NLG systems. There, human judgment is considered the gold standard, and automatic quality metrics are established by showing high correlation with human score. Thus, understanding whether humans can reliably judge diversity is important for improving diversity metrics. In this work, we use crowdsourcing workers\footnote{Native English speaking crowdsourcing workers, specifically qualified for this task, for more details see Appendix A.} to compute...
a human diversity score: we show workers a context followed by a set of generated responses, and ask them to rate the diversity of the set.

To establish best practices, we experiment with multiple variations of HDS (detailed in §6.2), asking humans to rate the diversity of a response set, and then evaluating each practice with our framework. We focus on the following questions and present results in §6:

- Should humans rate the absolute diversity score of a set of sentences or only rank whether one set is more diverse than another? (tl;dr: absolute scoring is more informative but rank scoring is moderately easier for humans.)
- Should humans rate diversity of a set or similarity between pairs in the set, from which diversity can be inferred? (tl;dr: diversity)
- Can humans evaluate different aspects of diversity well? (tl;dr: not effectively)

As a preliminary step, we conducted pilot experiments among a group of NLP graduate students. The main insights were: (a) humans are biased toward quality. For example, if a generated set has high diversity but low quality, humans will rate diversity lower than if the quality of the samples was higher. To neutralize this effect, we explicitly ask workers to evaluate the quality of one of the responses in the set \( S_c \), and then instruct them to ignore quality in the diversity questions; (b) To make sure a worker reads the context \( C \), we ask them to generate a sentence \( s \) before having them rate the diversity of a response set; (c) It is difficult for workers to evaluate the diversity of a set with more than 10 responses. Our crowdsourcing tasks are provided in Appendix A.

5 Similarity to Diversity Reduction

We expand the idea introduced by Zhu et al. (2018) and suggest a method to construct a diversity metric from any 2-sentence similarity metric.

Given \( m_{\text{sim}}(s_1, s_2) \in \mathbb{R} \), a symmetric similarity metric that gets a pair of input sentences \((s_1, s_2)\) and returns a similarity score, we can define a diversity metric \( \tilde{m}_{\text{div}} \) as the negation of the mean similarity score across all (unordered) pairs of \( S_c \):

\[
\tilde{m}_{\text{div}} = -\frac{1}{\binom{|S_c|}{2}} \sum_{s_i, s_j \in S_c, i > j} m_{\text{sim}}(s_i, s_j).
\]

This reduction allows us to easily define new diversity metrics based on past work on sentence similarity (Gomaa et al., 2013; Devlin et al., 2019; Zhang et al., 2019a; Reimers and Gurevych, 2019). In §6 we show that both n-gram-based similarity metrics and neural semantic similarity metrics provide useful diversity metrics.

6 Experiments

We now turn to our empirical investigation.

6.1 NLG Tasks

We apply our evaluation procedure on three different NLG tasks (in English), in which diversity is essential.

- **Story completion (storyGen)**; We use the ROC Stories dataset (Mostafazadeh et al., 2016), in which the context \( C \) is the first four sentences of a story, and the response \( s \) is a single sentence that ends the story. We use the contexts \( C \) from this data and generate response sets \( S_c \) for each context using our testers. The long contexts characterizing this data narrow down the space of possible responses, making this a “low-entropy” generation task, where the output is constrained, but diversity is still essential.
- **Dialog response generation (respGen)**; A comment-response pairs dataset extracted from the website reddit.com and pre-processed by Hashimoto et al. (2019). We use the comments from their data as contexts \( C \) and generate response sets \( S_c \) for each context using our testers. Since comments are single sentences the response is less constrained, making this a “medium-entropy” generation task.
- **3-words prompt completion (promptGen)**; Contexts \( C \) are 3-words prompts, extracted from the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011) by taking the first three words from each original context. The response sets \( S_c \) are completions of the prompts, generated by our testers. This context provides minimal constraints, making this a “high-entropy” generation task.

Samples of the contexts extracted for each task, along with generated response sets, are presented in Appendix B. We intentionally avoid NLG tasks where diversity is not necessarily desired, such as summarization and machine translation.

6.2 Evaluated Metrics

**N-gram-based metrics** We evaluate distinct n-grams (distinct-n), as described in §2. We also evaluate n-grams cosine similarity (cos-sim): a
similarity measure computing the cosine between the vectors representing two sentences, where each vector is a count vector over the n-grams that appear in the response. We use the reduction from §5 to convert this to a diversity measure. In both metrics, rather than choosing the order of the n-grams, we average over \( n \in \{1, \ldots, 5\} \), which we found to outperform any single choice of \( n \).

**Neural metrics** We exploit existing BERT-based models (Devlin et al., 2019) fine-tuned for estimating similarity between two sentences (applying the reduction from §5).

- **BERT-STS**: A BERT model fine-tuned on Semantic Textual Similarity (Cer et al., 2017): a collection of sentence pairs annotated with scores from 1-5 denoting their semantic similarity.\(^3\)
- **BERT-Score** (Zhang et al., 2019a): Originally a quality metric, BERT-Score uses BERT’s embeddings to measure similarity between two sentences. We used RoBERTa-large (Liu et al., 2019), as suggested by the authors.\(^4\)
- **Sentence-BERT** (sent-BERT) (Reimers and Gurevych, 2019) is a sentence-level embedding model based on BERT. We use the cosine similarity between the embeddings of two responses as a similarity metric. In our experiments we used bert-large-nli-stsb-mean-tokens.\(^5\)

**Human Metrics** We examine four methods for evaluating diversity with humans (see §4), to investigate best practices for obtaining diversity judgment from humans. In all metrics (except ranking), ratings are from 5 (highest diversity/similarity) to 1 (lowest). The original tasks presented to workers are in Appendix A.

- **Absolute HDS** (absHDS): Given a context \( c \) and a set of generated responses \( \mathcal{S}_c \), rate the level of diversity of \( \mathcal{S}_c \).
- **Ranking HDS** (rankHDS): Given a context \( c \) and two sets \( \mathcal{S}_{c,d_1}, \mathcal{S}_{c,d_2} \) generated with different values of the diversity parameter \( d \), rate which set is more diverse.
- **Similarity HDS** (simHDS): Given a context \( c \) and a set of generated responses \( \mathcal{S}_c \), rate the similarity of each two sentences in \( \mathcal{S}_c \), and then apply the reduction from §5.
- **Aspects HDS** (aspHDS): Identical to absHDS, except we explicitly ask about a specific aspect of

| Context |
| --- |
| **Fire next door.** |
| John woke up smelling like something was burning. |
| He went outside. He saw the fire next door. |
| He called the authorities. |

| Response set \( (\tau = 0.25) \) |
| --- |
| • It was a minor fire and they put it out. |
| • It was a fire. |
| • It was a fire. |
| • It was a fire. |
| • It was a fire. |

| Response set \( (\tau = 0.8) \) |
| --- |
| • They arrived and put out the fire. |
| • It was a fire. |
| • It was a fire. |
| • It turned out to be a fire. |
| • It was a minor fire night. |

| Response set \( (\tau = 1.1) \) |
| --- |
| • It turned out to be a mechanic. |
| • Before the fire was put out it was a fire. |
| • It was a fire. |
| • They co-worker matter how bad the fire was. |
| • Several shells, the fire department came just in time. |

Table 1: An example of the effect of temperature on the response set \( \mathcal{S}_c \) for a context \( c \) from ROC Stories.

6.3 Test #1

In this test we measure the correlation between diversity metrics (\( m_{div} \)) and the softmax temperature decoding parameter (\( d \)). The tester generating the response sets (\( \mathcal{S}_c \)) is a neural NLG model.

**Data and settings** For each of the three tasks, we generated sets of 10 responses per context, using a linear temperature sweep with 100 values in the range \([0.2, 1.2]\) (Caccia et al., 2018). We generated 1K sets in total for 1K contexts (10 sets per temperature) and evaluated on 200 (2 random sets per temperature). For automatic metrics, we repeat this experiment 100 times (randomly sampling 200 out of 1K sets each time, with replacement), to present the mean and standard deviation of the experiment. HDS metrics are computed over one experiment of 200 sets, due to their high cost (Appendix A). We provide an empirical justification for these particular values in §6.5.

The data for storyGen and respGen was generated by the neural model MASS (Song et al., 2019), fine-tuned on each dataset separately. The data for promptGen was generated by GPT-2-large (Radford et al., 2019) without fine-tuning. We provide examples for how story endings change as a function of temperature in Table 1. Examples for

\(^3\)https://github.com/swen128/bert-sts
\(^4\)https://github.com/Tiiiger/bert_score
\(^5\)https://github.com/UKPLab/sentence-transformers
all tasks are in Appendix B. For each HDS metric, we collected 10 ratings per query from Amazon Mechanical Turk (AMT) workers. Whereas absHDS demands one query per response set, in order to perform simHDS at a reasonable cost, we chose \( |S_c| = 5 \) (the first half of the original set), resulting in \( \binom{5}{2} = 10 \) crowdsourcing queries instead of \( \binom{10}{2} = 45 \) per set.

**Absolute scoring results** Table 2 presents the results of absHDS, simHDS, as well as all automatic metrics. In general, n-gram based metrics succeed in capturing the diversity induced by a temperature sweep, beating HDS and neural metrics. Figure 3 provides a more detailed analysis, where each point represents a single set of responses generated at some temperature. We observe that while rank correlation for cosine similarity is high, it is far from linear and reaches high values even at low temperatures, scoring 0.6 Pearson correlation. Conversely, the correlation for BERT-STS and absHDS is more linear, scoring 0.84 similar Pearson correlation. Moreover, the correlation for BERT-score is 0.88 similar Pearson correlation respectively. Thus, Pearson and Spearman correlations disagree in this case on the quality of the different metrics.

This result shows that humans perform worse than automatic metrics in this experimental setup, hinting that temperature mostly controls superficial changes to the generated text. Additionally, simHDS performs worse than absHDS although it is 3x more expensive, showing that rating the entire set rather than averaging over pairs is useful.

**Ranking results** To examine whether we can improve correlation by asking humans to rank whether one set is more diverse than another, rather than providing an absolute score, we conduct a ranking experiment. Each context is given along with two sets (5 samples each), produced with different temperature values. We sweep over temperature differences instead of the absolute temperature values. The human metric in this setting is \( rnkHDS \) (see §6.2), and the automatic metrics are the difference between the scores each of the two sets got.

We report two measures; The first is Spearman’s \( \rho \) between the metric and the temperature difference. The second is accuracy, i.e., whether the metric can predict which set has higher temperature (e.g., in automatic metrics this is whether the sign of the temperature difference and the sign of metric score difference agree).\(^6\)

Table 3 summarizes the ranking test results. We observe that humans are better at ranking compared to giving absolute scores, and are doing as well as automatic metrics. However, the scores of all automatic metrics also improve, making it difficult to separate between the different metrics.

**Other decoding parameters** To examine the robustness of our conclusions to other decoding parameters, we repeat it with two additional decoding methods: (a) in Nucleus (Top-\( p \)) sampling we swept linearly over 100 values of \( p \) in the range \([0.1, 1.0]\); (b) In Top-\( k \) sampling we swept \( k \) in logarithm scale over 100 values in the range \([1, 30K]\) and present the correlation between the metrics and \( \log_{10}(k) \). While softmax temperature enables skewing \( P_{LM} \) to a more diverse \( P_{gen} \) using \( \tau > 1 \), both Top-\( p \) and Top-\( k \) enable only skewing \( P_{LM} \) to a more sharp (hence less diverse) \( P_{gen} \).

Table 4 presents results for all automatic metrics using the three decoding methods over prompt-\( Gen \). Although the correlation in Top-\( k \) is significantly lower, and the variance is higher, all three decoding methods reflect a similar ordering between the metrics. Results for other tasks are in Appendix C.

### 6.4 Test #2

In this test, we measure the correlation between diversity metrics \( m_{div} \) and content diversity, represented by a binary parameter \( d \in \{0, 1\} \). The testers are AMT workers, guided to create sets with high level of form diversity and high or low content diversity according to \( d \).

**Data and settings** For each task, we collected 200 sets of 5 responses each (100 sets per class). For high content diversity class, we asked the workers to give 5 responses for a context, with as different content and structure as possible. Then

---

\(^6\) We consider ties in the metric difference score as a miss.
Figure 3: Test #1: Scatter plot of n-gram-based (cosine similarity), neural (BERT-STS) and human (absHDS) metrics as a function of temperature for respGen. Each point corresponds to a single generated set. Error bars of HDS represent the standard deviation over 10 annotator ratings.

Table 3: Test #1 ranking results (mean and standard deviation): Spearman’s ($\rho$) correlation between temperature differences and each metric score. Accuracy (acc) of classifying which set has the higher temperature.

Table 4: Test #1 results for different decoding parameters: Spearman’s $\rho$ (mean and standard deviation) of automatic metrics for promptGen.

Table 5 shows the test results. This time, n-gram-based metrics perform poorly, indicating they do not measure well content diversity. Neural models perform better than n-gram-based metrics (especially sent-BERT), but there is still a clear gap between automatic metrics and humans. Figure 4 illustrates the typical distributions of n-gram, neural and human metrics. Clearly, HDS separates high and low content diversity much better than neural metrics. In addition, n-gram-based metrics saturate both classes to near maximal values, similarly to test #1.

Since test #2 isolates content diversity, we used aspHDS to ask workers to directly rate content diversity and form diversity. Content aspHDS gets similar scores to absHDS, implying that there is no additional gain in asking directly on the tested aspect. Form aspHDS gets substantially lower scores compared to absHDS, validating that the form diversity of the two classes is similar.

6.5 HDS Stability: Picking Parameter Values

HDS experiments demand expensive human labor. Thus, we need to carefully choose the number of sets and the number of different ratings we ask per set, to get reliable results within a reasonable budget. In Figure 5 we measure HDS results for dif-
different number of sets and different number of ratings. Empirically, the test results are stable starting from 7 ratings and 150 sets. Hence, we used 10 ratings and 200 sets for HDS experiments.

### 7 Aspects of Diversity

In this work, we focused on the two primary aspects of diversity: content diversity (What to say?) and form diversity (How to say it?). In Figure 1, Both sets are diverse, but Set B is only form diverse, as all answers deliver the same massage, whereas Set A is diverse in both form and content.

Furthermore, we can observe aspects of diversity as having a tree-like structure, where both content and form diversity can be divided into sub-aspects: Content diversity (e.g. answering the question “How are you today?”) can be expressed by using different sentiment (“I’m doing good.” vs. “I’m so glad you asked! I’m really doing good.”), different relevance (“I’m fine” vs. “Did you see the game last night?”), and more. Form diversity can be divided into sub-aspects as well: syntactic diversity (“Someone took it from me.” vs. “It was taken from me.”) or lexical diversity (“I feel fine.” vs. “I feel very well.”). Even those sub-aspects can be further divided. For example, a sub-aspect of lexical diversity is register diversity (“How are you?” vs. “Sup bro?”).

Another observation is that different aspects are not orthogonal, that is, changing one aspect may lead to changes in other aspects. Specifically, we observe that while it is relatively easy to produce high form diversity with low content diversity (Set B in Figure 1), it is almost impossible to diversify content without changing form. This observation was important during the design of test #2.

### 8 Conclusions

This work presents a novel framework for evaluating diversity metrics as a step toward standardized evaluation. We limit the scope of this work to the differences between form and content diversity, which we consider key towards understanding the different aspects of diversity. Future work can explore other sub-aspects of diversity as detailed in §7, e.g., testing sentiment diversity, as proposed in §3. We urge researchers to use this framework as a platform for developing new diversity metrics and establishing their efficiency.
Acknowledgements

We thank Aya Meltzer-Asscher for linguistic advice, and Ben Bogin, Mor Geva, Omer Goldman and Ohad Rubin for their useful suggestions and references. This research was partially supported by The Israel Science Foundation grants 942/16, The Yandex Initiative for Machine Learning and the European Research Council (ERC) under the European Union Horizons 2020 research and innovation programme (grant ERC DELPHI 802800).

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. Cognitive science, 9(1):147–169.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

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.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2018. Language gans falling short. arXiv preprint arXiv:1811.02549.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Wenchoa Du and Alan W Black. 2019. Boosting dialog response generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 38–43.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. Computer Speech & Language, 59:123–156.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. El5: Long form question answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3558–3567.

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.

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

Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. International Journal of Computer Applications, 68(13):13–18.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume I (Long and Short Papers), pages 1689–1701.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. arXiv preprint arXiv:2003.08529.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2443–2453.

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.
Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Manzhongyang Li, Xiao Ding, and Ting Liu. 2018. Generative and diversified story ending using sequence to sequence model with adversarial training. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1033–1043.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1033–1043.

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.

Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. Foundations of statistical natural language processing. MIT press.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, San Diego, California. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. arXiv preprint arXiv:1903.08949.

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 on association for computational linguistics, pages 31–38. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Lan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3973–3983.

Manasvi Sagarkar, John Wieting, Lifu Tu, and Kevin Gimpel. 2018. Quality signals in generated stories. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 192–202.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1823–1827.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In International Conference on Machine Learning, pages 5926–5936.

Guy Tevet, Gavriel Habib, Vered Shwartz, and Jonathan Berant. 2019. Evaluating text gans as language models. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2241–2247.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1980–1991.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2002–2012.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In Thirty-First AAAI Conference on Artificial Intelligence.

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

Xinyuan Zhang, Yi Yang, Siyang Yuan, Dinghan Shen, and Lawrence Carin. 2019b. Syntax-infused variational autoencoder for text generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2069–2078.
A  HDS Questionnaires

All Human scores for HDS metrics were collected using AMT crowdsourcing platform by English native-speaking workers that were specifically qualified for this task. Figure 6 presents the warm-up part, common for all HDS questionnaires. Before asking workers to rate the diversity of each set, we first asked them to generate a response for the context themselves, to make sure they read the it. To neutralize the effect of the responses’ quality on the workers, we also asked the workers to rate the quality of the first response in the set, then explicitly instructed them to ignore quality when rating diversity.

Figures 7 to 10 present the diversity questions of absHDS, aspHDS, rnkHDS and simHDS as appeared in the AMT questionnaires.

Costs  For HDS metrics that require one query per response set (i.e. absHDS, rnkHDS, aspDHS), the cost for a single rating was 0.18$. We collected 10 ratings per response set, and conduct each experiment with 200 sets, hence the total cost for an experiment was 360$. In the case of simHDS, the response set size was 5, and the number of queries needed per set is $\binom{5}{2} = 10$. The cost of a single rating for this task was 0.056$, and with the same multipliers, the total cost for an experiment was 1120$, three times more expensive.

B  Data Samples

B.1  Test #1

Tables 6 to 14 present data samples from storyGen, respGen and promptGen with the neural testers of test #1, as detailed in §6. Each table presents two contexts and three response sets per context. Each response set was generated with a different value of decoding parameter for the three decoding methods: softmax temperature, Nucleus sampling, and Top-k.

B.2  Test #2

Tables 15 to 17 present data samples from storyGen, respGen and promptGen with the human testers of test #2, as detailed in §6. Each table presents two contexts and two response sets per context - one for the low content diversity class and one for the high content diversity class.

C  Additional Results

Comparing test #1 results of storyGen to other tasks (Table 2), this task is characterised with noisier scores for all metrics (Figures 3 and 11), hence lower $\rho$ values and higher variance. A possible explanation is larger effect of $c$ on the distribution $P_{gen}(s|c)$ in this task.

Tables 4, 18 and 19 present test #1 absolute scoring experiment using temperature, nucleus sampling and Top-k decoding parameters as $d$. Top-k consistently yields lower $\rho$ compared to other decoding parameters, especially for storyGen task. This implies that Top-k represents diversity less reliably than other methods.
smelled wonderful. When it arrived, it was hot and it put it out. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was a fire. It was
She tried to adjust them to relieve the tension. She was able to grab a nap afterwards. Jane was unable to finish her nap since it was lost. Jane pulled over and started to clean up. Jane was able to finish her nap.

Kate decided to go to the store and buy some ear phones. She decided to go to the store and buy some ear phones. She decided to go buy a pair of headphones instead. She decided to go to the store and buy some headphones instead. She decided to go to the doctor and have some rest. Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones.

Jane pulled over and started to cry. Jane pulled over and started to cry. Jane stopped at the store to buy a new dish from the store.

She decided to go to the store and buy some headphones instead. She decided to go to the doctor and have some rest. Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones instead.

She decided to go buy a hat so she could enjoy the the long lin. Kate decided to go to the store to buy some candy since she was ti. Kate decided to go outside to rest. Kate decided to go to the store and buy some ran phones. Kate decided to go to the store and buy some headphones instead. Kate decided to go back to her old dishes. Kate decided to go buy a big pair of headphones instead. Kate decided to go to the store and wearing some headphones. She was forced to go to the store to buy some cash. She decided to go to the store and buy some headphones instead.

She decided to go buy a hat so she could enjoy the the long lin. Kate decided to go to the store to buy some candy since she was ti. Kate decided to go outside to rest. Kate decided to go to the store and buy some ran phones. Kate decided to go to the store and buy some headphones instead. Kate decided to go back to her old dishes. Kate decided to go buy a big pair of headphones instead. Kate decided to go to the store and wearing some headphones. She was forced to go to the store to buy some cash. She decided to go to the store and buy some headphones instead.

She decided to go to the store and buy some headphones instead. She decided to go to the store and buy some ear phones. She decided to go buy a pair of headphones instead. She decided to go to the store and buy some headphones instead. She decided to go to the doctor and have some rest. Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones.

Jane pulled over and started to cry. Jane pulled over and started to cry. Jane stopped at the store to buy a new one. Jane was annoyed and began to cry. Jane stopped at the store to buy a new one. Jane wanted to have her car back. Jane was mouted.

Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some ear phones. She decided to go buy a pair of headphones instead. She decided to go to the store and buy some headphones instead. She decided to go to the doctor and have some rest. Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones.

Jane pulled over and started to cry. Jane pulled over and started to cry. Jane stopped at the store to buy a new dish from the store.

Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some ear phones. She decided to go buy a pair of headphones instead. She decided to go to the store and buy some headphones instead. She decided to go to the doctor and have some rest. Kate decided to go to the store and buy some headphones instead. She decided to go to the store and buy some headphones.

She decided to go buy a hat so she could enjoy the the long lin. Kate decided to go to the store to buy some candy since she was ti. Kate decided to go outside to rest. Kate decided to go to the store and buy some ran phones. Kate decided to go to the store and buy some headphones instead. Kate decided to go back to her old dishes. Kate decided to go buy a big pair of headphones instead. Kate decided to go to the store and wearing some headphones. She was forced to go to the store to buy some cash. She decided to go to the store and buy some headphones instead.

Table 8: Test #1 data samples for storyGen task and different $k$ values (Top-k).

Table 9: Test #1 data samples for respGen task and different temperatures.
| Context | Response set (p = 0.208) | Response set (p = 0.64) | Response set (p = 1) |
|---------|--------------------------|--------------------------|---------------------|
| do you know if he still lives in new orleans | • i dont think so | • he is in nyc | • aw screw you guys the gimme a shower and ill send you my money |
| | • i dont think so | • he does hope he does | • he does im so confused by how this is relevant |
| | • no i dont think so | • i dont think so | • they spent all that much in new orleans up in first place |
| | • no idea i dont know | • he lives in new orleans | • still living in new orleans according to most of them it jesus mustve worth |
| | • i dont think so | • yes he does he is in florida | • anyone else want to wait for him to go back home |
| | • no idea i just saw him live in new orleans | • im guessing not that i know of | • how many tickets would you say the willing said if he warn you that hes |
| | • i dont think so | • hes a small town i dont know what happened there | • im guessing he lives there as well |
| | • no idea i just saw him live in new orleans | • i do not i dont know if he still lives in new orleans | • yep rick albert not |
| | • i dont think so | • i dont think so | • he lives in ny |
| | • i dont think so | • i dont think so | • no i dont get it any more guess i may canadian |

Table 10: Test #1 data samples for respGen task and different p values (nucleus sampling).

| Context | Response set (k = 3) | Response set (k = 32) | Response set (k = 318) |
|---------|----------------------|------------------------|------------------------|
| watching curry play in his prime is truly a privilege | • i know i just dont want him to play for us | • and his career as well | • yeah my feeling s mean we dont like it but it happens all the |
| | • he has to be a good center for that | • agreed the way hes playing is awesome | • you are one for real |
| | • he is a great center of football in his prime | • it has to be | • they still have a rule saying |
| | • hes been playing his prime for a long time | • this is just called a job | • it really is a necessary thing to do |
| | • he was a great back in the day | • and then being on the field for the first time | • finally some reason to continue |
| | • hes been playing for a while now | • i dont see him doing that often enough | • watching him at some point |
| | • i dont know about that he was pretty damn good at that | • he just likes to party in the kitchen | • yet that would be epic |
| | • i dont think he was ever in his prime | • at this point hes going to be a great star for the rest of the | • not to mention eating curry dinner |
| | • i dont think he is a prime minister | • only if he pays well | • is a privilege |
| | • i dont know why but i think he is a very good player and | • the only thing that can make that | • i just dont want to turn over for this |
| | • i know i just dont want him to play for us | • kind of difference is how much time you | • goal like he does in |
| | • he has to be a good center for that | • and his career as well | • gt playing in his prime is truly a privilege |
| | • he is a great center of football in his prime | • agreed the way hes playing is awesome | • fifty |
| | • hes been playing his prime for a long time | • it has to be | • so is saying he is in high school |

Table 11: Test #1 data samples for respGen task and different k values (Top-k).
Table 12: Test #1 data samples for promptGen task and different temperatures. Bold text is the 3-words prompt context.

| Response set \( (\tau = 0.25) \) | Response set \( (\tau = 0.8) \) | Response set \( (\tau = 1.1) \) |
|-----------------------------------|----------------------------------|-----------------------------------|
| • Not the hacking: The hacking is the fact that the DNC was hacked! | • Not the hacking: The hacking after all! I'm sure the nation-states that are involved in! | • The hacking experience of a CIA VRO crunched nine months ago! Start for 2016 jumps! |
| • Not the hacking: The hacking is the real problem. The hacking is the! | • Not the hacking that happened on the internal networks of the Energy Department. In fact, according to! | • The hacking: David. The directory was flagged in a document it created in late last year! |
| • Not the hacking of the DNC, but the leaks of the emails of the Democratic National Committee! | • Not the hacking of the American public but rather the fraudulent Heisenberg principle that seemed to be! | • Not the hacking of Democratic Party systems - said the Russian team’s activity represented “just the beginning!” |
| • Not the hacking, but the way it was done. The FBI’s investigation into the! | • Not the hacking that took place in the DNC last year or the release of hacked emails during the! | • Not the hacking, of course – which these sources sounded more concerned about than being attacked 140 times! |
| • Not the hacking of the DNC, but the hacking of the emails of the Democratic National Committee! | • Not the hacking of the DNC, but the leaking of the emails. The DNC’s! | • Not the hacking: story is over. But yet there’s another reason not to rush out such statements! |
| • Not the hacking of the DNC. The hacking of the DNC was a "false flag! | • Not the hacking of the DNC. But the hacking of the DNC hack! | • Not the hacking-either-. These were scattered in the workshop! (Expanded- being guys with! |
| • Not the hacking: The hacking is the problem. The hacking is the problem! | • Not the hacking: the willingness. The evidence of interest in this case comes in! | • Not the hacking: private material of elected officials, e.g. emails, even if the! |
| • Not the hacking of the DNC, but the leaking of the emails. The DNC hack! | | • Not the hacking: has happened yet!!!!!!!!!!!! |
| • How is our new technology helping us to do that? We are using a new technology! | • How is our new technology helping us to do that? The hacker is the fire! | • How is our new technology helping us to do that? The hacker is the fire! |
| • How is our system different from that of the United States? The United States is a! | • How is our system different from that of the other major European countries? The European Commission! | • The hacker is the fire! |
| • How is our approach different from that of the other major European countries? | • How is our country going to be able to compete with the rest of the world if we don! | • How is our approach different from that of major European countries? The European Commission! |
| • How is our country going to be able to compete with the rest of the world if we don! | • How is our country going to be able to compete with China in the future? be asked! | • How is our country going to be able to compete with China in the future? be asked! |
| • How is our work different from that of other organizations? The work of the Center for! | • How is our work different from that of other research in this area? We are not the first! | • How is our work different from other research in this area? We are not the first! |
| • How is our system of government supposed to work? The reason we have a government is! | • How is our system of government supposed to work? The reason we have a government is! | • How is our system of government supposed to work? The reason we have a government is! |
| • How is our system different from the one that was used in the past? The system! | • How is our system different from the one that was used in the past? The system! | • How is our system different from the one that was used in the past? The system! |
| • How is the governmental government going to catch up with the cyber criminals?” he said. “I’m! | • How is our government going to catch up with the cyber criminals?” he said. “I’m! | • How is our government going to catch up with the cyber criminals?” he said. “I’m! |
| • How is our society selling humanity on slavery? The answers to these questions are also important for us! | • How is our society selling humanity on slavery? The answers to these questions are also important for us! | • How is our society selling humanity on slavery? The answers to these questions are also important for us! |
| • How is our minister giving it to you? Isn’t! | • How is our minister giving it to you? Isn’t! | • How is our minister giving it to you? Isn’t! |
| • How is our research different from other studies? This study examined the effects of peer! | • How is our research different from other studies? This study examined the effects of peer! | • How is our research different from other studies? This study examined the effects of peer! |
| | • How is our research different from other studies? This study examined the effects of peer! | • How is our research different from other studies? This study examined the effects of peer! |
| • How is our mission different from Seniors’ Service Corps (SSC) other than the fact! | • How is our mission different from Seniors’ Service Corps (SSC) other than the fact! | • How is our mission different from Seniors’ Service Corps (SSC) other than the fact! |
| • How is our challenge different? The only difference is that this challenge is about building! | • How is our challenge different? The only difference is that this challenge is about building! | • How is our challenge different? The only difference is that this challenge is about building! |
| | • How is our challenge different? The only difference is that this challenge is about building! | • How is our challenge different? The only difference is that this challenge is about building! |
| • How is our nation governed?” As Obama moved into his second term, he is increasingly! | • How is our nation governed?” As Obama moved into his second term, he is increasingly! | • How is our nation governed?” As Obama moved into his second term, he is increasingly! |
| • How is our rapid abandonment of critical thinking, knowledge, and values, and the subsequent burial of! | • How is our rapid abandonment of critical thinking, knowledge, and values, and the subsequent burial of! | • How is our rapid abandonment of critical thinking, knowledge, and values, and the subsequent burial of! |
| • How is our education system designed for our futures? We are the children of immigrants! | • How is our education system designed for our futures? We are the children of immigrants! | • How is our education system designed for our futures? We are the children of immigrants! |
Table 13: Test #1 data samples for promptGen task and different p values (nucleus sampling). Bold text is the 3-words prompt context.

| Response set (p = 0.208) | Response set (p = 0.64) | Response set (p = 1) |
|--------------------------|-------------------------|----------------------|
| • So that's the story of the last few years. The current political climate is not! | • So that’s the state of the campaign. Now, what I do want to talk about is! | • So that’s the first time you want to punch somebody, not miss before.' The Seahawks would! |
| • So that’s the end of the first part of this series. I hope you enjoyed it! | • So that’s the thing: For as much as I love TLC, it’s hard to! | • So that’s the science behind the Broadwell-E processors from Intel that Intel launched last fall! |
| • So that’s the first thing I want to say. I’m not going to be the guy! | • So that is the idea, anyway. The last two seasons have been about doing that. It! | • So that’s the instinct from other teams, that they’re a headache. - Ramsay MacDonald! |
| • So that’s the thing about being a professional. You have to be able to handle the criticism! | • So that’s the end of the half-hour segment. The next half-hour! | • So that’s the white whale right there about too much debt. And then what you! |
| • So that’s the way it is. I don’t think there’s any way to change it! | • So that’s the situation we’re in,” he said. | • So that’s the end of our discussion about the causes. What happens when we look at the! |
| • So that’s the problem. It’s not just that the government is failing to protect! | • “We’re in the! | • So that’s the cover of inhibition against “chronic” or “adaptive” stimulants! |
| • So that’s the thing about this. It’s not just about the money. It’s about! | • So that’s the thing, I don’t know if you know, but in general it’s! | • So that’s the way the story goes, but exactly how is cloud providers going to restrict Their! |
| • So that’s the end of the story. The next step is to create a custom? | • So that’s the difference between the kinds of things that people will be talking about on Wednesday! | • So that’s the beginning, the beginning of the show, I guess five minutes.’” |
| • So that’s the case. So, what’s the problem? Well! | • So that’s the thing about being a professional. It’s not just! | • So that’s the Indie Mobile Game Week Honoring Winners!!!!!!!!! |
| • So that’s the first time I’ve ever seen a real one. I’m not! | • That’s the standard for using memcpy(). It’s fine to use memc! | • So that’s the reason I’m writing, that’s why you don’t understand why people know! |

| • do you listen to the music? | “I don’t know. I don’t listen! | • do you listen to the current draft? | I listen to the current draft. I’m! |
| • do you listen to them? | “I do,” he said. | • do you listen to it? | It’s easy to hear the "why?" but when! |
| • do you listen to the music? | “I do.” | • do you listen to the people that come here?” | “No, I’m too busy! |
| “I do,” said the king! | “You’re not! | • do you listen to the thing? | “Of Course I do. I’ve been reading! |
| • do you listen to the voices of the people?” | “I don’t know I don’t know! | • do you listen to those who are opposing it, who want to create a situation in which at! |
| • do you listen to the song? | “I don’t know I don’t know! | • do you listen to Human Fly?, which YouTuber Nico Perri collaborated on, and Google! |
| “I don’t know. I don’t! | • do you listen to the news? I do. I’m a big fan of the! | • do you listen to the accapella lyrics out of context and express the feeling?” It’s! |
| • do you listen to the news? | I do. I’m a big fan of the! | • do you listen to Michael Kwanana-Smith who writes, ‘The American Journalism Review discern!’ |
| • do you listen to me?” | “Yes, I do.” | • do you listen to my songs as I said,” Ramachalor said. “You feel! |
| “I’m not!” | “I’m not!” | • do you listen to U.S. 90 night at this time of the year? |
| • do you listen to the other side?” | “I do,” said the boy. | • do you listen to that as well?” | “The question was not, 'Who is! |
| • do you listen to the news? No, I don’t. I don’t listen! | • do you listen to the acapella lyrics out of context and express the feeling?” It’s! | • do you listen?” He asks, leaning forward as he woodyly talks to him. "Listen! |
| • do you listen to the current draft? | I listen to the current draft. I’m! | • do you listen to those books and sway him so much? No. He was deeply brainwashed! |
| • do you listen to it? | It’s easy to hear the "why?" but when! | • do you listen?” Simon[ol].I feel like I’m in a Kudish Genocide. |
| • do you listen to the music? | “I do.” | • do you listen to value authenticated queries from your menu when running count? And if not, then! |
Table 14: Test #1 data samples for promptGen task and different k values (Top-k). Bold text is the 3-words prompt context.

| Context | Response set (high content diversity) | Response set (low content diversity) |
|---------|--------------------------------------|---------------------------------------|
| Sold Out | Jane cried over the fact that she couldn’t watch it and just gave up looking for a ticket. | Jane remembered that she has an old friend who is a manager at a big movie theater so she contacted that friend in the hopes that she can buy any spare ticket. |
| | Jane decided to look for a scalper that would sell her the ticket for the movie that she really wanted to see. | Jane decided to look for her friend who could possibly have access to tickets for that movie since that friend currently works at a movie theater. |
| | Jane thought it was okay since she can still have a chance to watch it once it gets uploaded in video and movie streaming applications. | Jane realized that her friend might have spare tickets since she is a manager of a movie theater showing that film. |
| | Jane posted a status on her social media accounts asking her friends for any spare ticket that she is about to say this. | Jane decided to look for her friend who could possibly have access to tickets for that movie since that friend currently works at a movie theater. |
| | Jane decided to look for her friend who could possibly have access to tickets for that movie since that friend currently works at a movie theater. | Jane realized that her friend might have spare tickets since she is a manager of a movie theater showing that film. |
| | They are busy gathering sticks to make a dam. | They are busy gathering sticks to make a dam. |
| | They are building a dam on the creek. | They are building a dam on the creek. |
| | They won’t let us get too close to them. | They won’t let us get too close to them. |

Table 15: Test #2 data samples for storyGen task.

| Context | Response set (high content diversity) | Response set (low content diversity) |
|---------|--------------------------------------|---------------------------------------|
| Kill la kill is still going new episode every thursday | That show sucks... | Real actor is woosoo hot. |
| | OMG I can’t wait. | Did you see the cliffhanger at the end of the season? |
| | I thought they canceled it. | I’ve been waiting for it to return for weeks. |
| | What channel is it on. | I’m totally gonna binge watch last season. |
| | I only watch nature programs on BBC. | I just got into this show and can’t stop watching. |
| places apple slices in a bowl so they’ll stay fresh | Oh boy, I love apples. | I find merit in this input. |
| | I don’t need you telling me how to keep things fresh, take a hike! | That information will serve me well. |
| | Girl, you’re the freshest one around here. | Thanks, that’s really good to know! |
| | This post might be better in the life hacks section. | Such knowledge is certainly beneficial. |
| | This is actually a useful bit of advice. | Wise words, I will heed them. |
Figure 6: Warm-up part, starting each AMT HDS task. It includes the context, and a single response generated by the tester. The worker is asked to generate response of hers/his own and rate the quality of the tester’s response.

Figure 7: absHDS question along with the evaluated response set (Test #2 in this case).

Figure 8: aspHDS question (content in this case). The response set is the same as presented for absHDS question.
Figure 9: rnkHDS question along with the two evaluated response sets.

| SET A                                                                 | SET B                                                                 |
|---------------------------------------------------------------------|---------------------------------------------------------------------|
| • I do out their hands I see no sides                                | • I hate getting pencil on the side of my hand                        |
| • Good on you this class spent 90 minutes looking for it to be remaining barely | • I hate getting pencil on the side of my hand                        |
| • Me too it all makes sense now for me                               | • I hate getting pencil on the side of my hand                        |
| • I live too nice                                                   | • I hate getting pencil on the side of my hand                        |
| • I wish ld want to know more about this getting pencil all over my foot | • I hate getting pencil on the side of my hand                        |

Which of the two sets is more diverse?
- 5 - SET A is much more diverse.
- 4.5
- 4 - SET A is somewhat more diverse.
- 3.5
- 3 - The diversity of both sets is similar.
- 2.5
- 2 - SET B is somewhat more diverse.
- 1.5
- 1 - SET B is much more diverse.

Figure 10: simHDS question along with the two evaluated responses.

| A. proposed responses:                        | A. How similar are the proposed responses?                        |
|------------------------------------------------|---------------------------------------------------------------|
| • I really want people to talk more            | • 5 - Very similar (The responses are the same or almost the same.) |
| • I know it’a nice way to talk                  | • 4 - Similar (The responses are quite similar.)               |
|                                                | • 3 - Slightly similar (The responses are a bit similar but not the same.) |
|                                                | • 2 - Almost not similar (The responses are considerably different from one another.) |
|                                                | • 1 - Not similar at all (The responses are completely not related.) |

Table 17: Test #2 data samples for promptGen task. Bold text is the 3-words prompt context.

| Response set (high content diversity) | Response set (low content diversity) |
|---------------------------------------|--------------------------------------|
| • Suppose there’s an escape plan we haven’t thought of yet. | • Suppose there’s an airline that costs less. |
| • Suppose there’s an omelet that is the most amazing ever. | • Suppose there’s an flight that isn’t as expensive. |
| • Suppose there’s an airplane ticket that’s even cheaper. | • Suppose there’s an air travel fare, but doesn’t cost as much. |
| • Suppose there’s an actual deadline for this paper. | • Suppose there’s an way to fly there that is low cost. |
| • Suppose there’s an event that we can go to this weekend. | • Suppose there’s an flight going there and it’s not a lot of money |
| • Nothing remotely like eating a big breakfast. | • Nothing remotely like being super full and satisfied. |
| • Nothing remotely like dancing with your wife at the wedding. | • Nothing remotely like getting to taste many different foods. |
| • Nothing remotely like singing Justin Bieber’s greatest hits | • Nothing remotely like starting the day off right. |
| • Nothing remotely like falling down a hill | • Nothing remotely like doing exactly what I want to do. |
| • Nothing remotely like getting yelled at | • Nothing remotely like feeding myself with great food. |

Figure 11: Test #1: Scatter plot of n-gram-based (cosine similarity), neural (BERT-STS) and human (absHDS) metrics as a function of temperature for storyGen. Each point corresponds to a single generated set. Error bars of HDS represent the standard deviation over 10 annotator ratings.
| Temperature | Top-p | Top-k |
|------------|-------|-------|
| distinct-n | 0.76 (0.03) | 0.69 (0.03) | 0.2 (0.06) |
| cos-sim    | 0.71 (0.04) | 0.66 (0.03) | 0.16 (0.06) |
| BERT-STS   | 0.64 (0.04) | 0.58 (0.04) | 0.2 (0.07) |
| sent-BERT  | 0.65 (0.03) | 0.59 (0.04) | 0.17 (0.06) |
| BERT-score | 0.69 (0.04) | 0.61 (0.04) | 0.23 (0.05) |

Table 18: Test #1 results for different decoding parameters: Spearman’s $\rho$ (mean and standard deviation) of automatic metrics for `storyGen`.

| Temperature | Top-p | Top-k |
|------------|-------|-------|
| distinct-n | 0.89 (0.01) | 0.84 (0.02) | 0.64 (0.04) |
| cos-sim    | 0.89 (0.01) | 0.78 (0.03) | 0.62 (0.05) |
| BERT-STS   | 0.81 (0.02) | 0.74 (0.03) | 0.56 (0.04) |
| sent-BERT  | 0.80 (0.02) | 0.63 (0.05) | 0.51 (0.04) |
| BERT-score | 0.87 (0.01) | 0.77 (0.03) | 0.6 (0.05) |

Table 19: Test #1 results for different decoding parameters: Spearman’s $\rho$ (mean and standard deviation) of automatic metrics for `respGen`. 