Evaluating the Evaluation of Diversity in Natural Language Generation

Guy Tevet[1,2] **Jonathan Berant**[1,3]

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

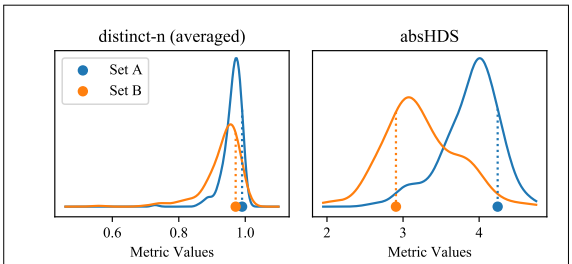
Abstract

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 tasks such as story generation (Li et al., 2018), question generation (Pan et al., 2019) and 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 no standard evaluation metric for measuring diversity. Thus, different papers evaluate diversity differently (if at all), making it difficult to compare competing approaches (Hashimoto et al., 2019). Having a principled and consensual diversity evaluation metric is hence fundamental for the field of NLG.



Question: So what did I miss in the first 20 minutes?

Set A
- Pretty much everything.
- Nothing, really.
- You won't believe what happened!
- Why do you even care?
- What were you doing that was more important than this?

Set B
- Not much.
- It was pretty dull.
- Blah, you didn't miss anything.
- Not anything that important.
- Very little, it was uneventful.

Figure 1: Diversity metric evaluation: we show two sets of responses to the same question, generated by crowdsourcing workers. While both sets are diverse in terms of *form*, only set A is diverse in terms of *content*. Each graph presents the distribution over a diversity metric for sets with high content diversity (blue) and low content diversity (orange). Distributions are approximated over 200 sets. We observe that the human score metric (absDHS) separates the two distributions, while an n-gram based metric (distinct-n) fails, illustrating that it does not capture content diversity. The dotted lines correspond to the specific sets A and B presented above.

A key challenge in developing diversity evaluation metrics, is the difficulty in determining their efficacy. Unlike metrics for evaluating the *quality* of generated text, where one can measure correlation between a metric (such as BLEU (Papineni et al., 2002)) and human judgement (Zhang et al., 2019a; Sagarkar et al., 2018), it is unknown if hu-

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 326–346
April 19 - 23, 2021.

mans can reliably estimate diversity.

In this paper, we propose a framework for evaluating diversity metrics (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 metric, and measure correlation between the metric and the diversity parameter. High correlation indicates that the metric captures how the diversity parameter affects the model output.

We instantiate this framework with two tests. As a preliminary step, we introduce the *decoding 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 has been shown to affect model diversity (Holtzman et al., 2019; Caccia et al., 2018). Then, we turn the focus to *content* diversity, introducing the *content test* (Figure 1). Here, 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 three families of popular diversity metrics with these 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.

Applying our tests leads to several findings: (i) In the *decoding test*, n-gram-based metrics correlate well with decoding parameters, such as softmax temperature. While the goal of our framework is to evaluate diversity metrics, this result lets us reflect back on the tester itself and conclude that decoding parameters predominantly control the form of text rather than content. (ii) Conversely, n-gram-based metrics perform poorly in the *content test*. While neural metrics outperform n-gram-based metrics, humans are substantially better than any automatic metric at detecting content diversity. This is illustrated in Figure 1, where a human clearly distinguishes between sets that have high (blue) and low (orange) content diversity, while n-gram-based metrics fail to do so.

Due to this gap, we construct a large dataset focused on *content*-diversity metrics. We release the **M**etrics for **c**ontent **Div**ersity (*McDiv*) benchmark, a challenge for research in diversity evaluation.

To conclude, our main contributions are:

- A framework for evaluating diversity metrics.
- Tests instantiating this framework, measuring the sensitivity of metrics to diversity, with a focus on content diversity.
- Best practices for obtaining diversity evaluations from crowdsourcing workers.
- Establishing that humans outperform current automatic metrics in detecting content diversity.
- The McDiv dataset - a benchmark for content diversity aware metrics.
- 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 has increased (Du and Black, 2019; Holtzman et al., 2019), 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, measuring the proximity of a language model (LM), P_{LM}, to the true distribution, P_{ref}, by approximating the cross-entropy $H(P_{\text{ref}}, P_{\text{LM}})$ with held-out data 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 a NLG system is also strongly affected by the decoding procedure. For example, *Top-k* and *nucleus*

[1] https://github.com/GuyTevet/diversity-eval

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 language models. While one can approximate perplexity for such models (Tevet et al., 2019), ideally, a metric should not be tied to a model.

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. 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. Diversity has many facets: for instance, a set of sentences can be diverse in terms of their *content*, while another may have similar content, but diverse *form* (Figure 1). Our framework provides a way to evaluate metrics for different aspects of diversity under moderate assumptions.

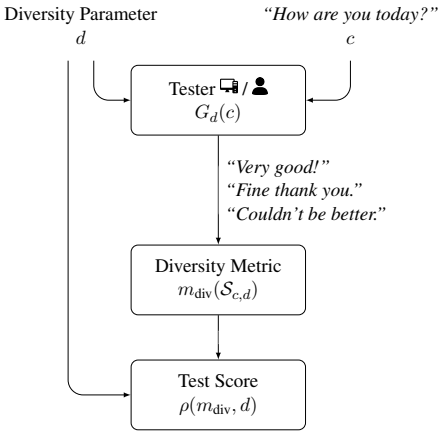


Figure 2: An overview of our diversity metrics evaluation framework. The tester (machine or human) generates a response set ($\mathcal{S}_{c,d}$) given a diversity parameter (d) and a context (c). The test score of a metric m_{div} is the correlation between the metric score for $\mathcal{S}_{c,d}$ and d.

We define a diversity metric $m_{\text{div}}(\mathcal{S}_c) \in \mathbb{R}$ as a function that takes a set of generated responses $\mathcal{S}_c$ as an input, and outputs a diversity score. Each response $s \in \mathcal{S}_c$ is generated for the same input context c, hence $\mathcal{S}_c$ 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_{div} over sets $\mathcal{S}_c$ sampled from the model given multiple contexts $c \in \mathcal{C}$.

To evaluate $m_{\text{div}}(\cdot)$, we assume access to some deterministic *diversity parameter* d that controls an aspect of diversity in $\mathcal{S}_c$. We test the relation between m_{div} and the parameter d. By varying d and measuring m_{div}, we can compute the correlation ρ between m_{div} and an aspect of diversity represented by d. Because our goal is to have metrics that *rank* the diversity of generated texts, we use Spearman's ρ 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 $\mathcal{S}_c$ 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 $\mathcal{S}_{c,d}$. 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 represent *sentiment diversity*, that is,

the number of different sentiments in a collection of reviews $\mathcal{S}_c$. A human tester can observe c and d, and produce reviews accordingly (such data can be easily mined from IMDB). A collection of such $(d, \mathcal{S}_{c,d})$ makes a test, in which the correlation between $m_{\text{div}}(\mathcal{S}_{c,d})$ and d measures the sensitivity of m_{div} to sentiment diversity.

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 Decoding Test

The diversity of a NLG system constructed from a LM depends on both the LM but also the decoding algorithm on top of it. For example, *beam search* approximates the most probable output, and dramatically reduces diversity. Conversely, sampling from the LM leads to high diversity, but low quality output (Holtzman et al., 2019).

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 parameter τ 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-probability 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 the decoding test (*decTest*), we define the *tester* to be a 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_{div} correlate with decoding parameters. This can shed light on the quality of the metrics, but also on how decoding parameters affect the output of a NLG system. The decoding test uses automatically-generated data that is cheap to produce, and decoding parameters that are well-known to control diversity. Thus, we view this test as a warm-up test to explore the strengths of our framework.

3.2 Content Test

In the content test (*conTest*), our goal is to evaluate how different diversity metrics capture the notion of *content diversity*. Measuring content diversity requires deep understanding of the semantics of responses in $\mathcal{S}_c$.

To isolate *content* from *form* diversity, we aim to generate response sets with a similar level of form diversity, but where the level of content diversity is controlled by the diversity parameter d. Thus, we use crowdsourcing workers as testers, and a binary parameter $d \in \{0, 1\}$, corresponding to low or high content diversity. A worker observes a context c and produces a set of responses $\mathcal{S}_c$ 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 high in all examples. Examples from this data are in Figure 1 and Appendix B.

In §6, we will focus on whether automatic diversity metrics can perform as well as 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 diversity, to the best of our knowledge no work thoroughly investigated this question. The importance of this question is clear when comparing to quality evaluation. There, human judgment is the gold standard, and automatic quality metrics are established by showing high correlation with human score. Thus, understanding if humans can judge diversity is important for improving diversity metrics. We use crowdsourcing workers[2] to compute a human diversity score: we show workers a context followed by a set of 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 evaluating each practice with our framework. We focus on the following questions:

- 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)*
- Should humans rate the *absolute* diversity score of a set of sentences or *rank* whether one set is

[2]Native English speakers, for more details see Appendix A.

more diverse than another? Here, we did not reach a conclusive result, and describe this experiment in the Appendix C.

As a preliminary step, we conducted pilot experiments among a group of NLP graduate students. The main insights were: (a) humans are biased by quality: if a generated set has high diversity but low quality, humans will rate diversity low. To neutralize this, we explicitly ask workers to evaluate the quality of one of the responses in the set $\mathcal{S}_c$, and then instruct them to ignore quality in diversity questions; (b) To make sure a worker reads the context c, we ask them to generate a sentence s before they rate diversity; (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 Diversity to Similarity Reduction

We expand the idea from 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 $\mathcal{S}_c$:

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

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

6.1 NLG Tasks

We apply our evaluation procedure on three different English NLG tasks that require diversity.

- **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 $\mathcal{C}$ from this data and generate response sets $\mathcal{S}_c$ 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 $\mathcal{C}$ and generate response sets $\mathcal{S}_c$ 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 $\mathcal{C}$ 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 $\mathcal{S}_c$ 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 sen-

[3] https://github.com/swen128/bert-sts

tences. 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 (*rnkHDS*); 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. Since this metric did not clearly outperform *absHDS*, we provide results in Appendix C only.

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.

Aspect HDS (*aspHDS*); Identical to *absHDS*, except we explicitly ask about a specific aspect of diversity, namely *form* and *content*.[6]

6.3 Decoding Test

In decTest 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 task, 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 each of 1K contexts (10 per temperature) and evaluated 200 (2 random sets per temperature). For automatic metrics, we repeat this 100 times (randomly sampling 200 out of $1K$ sets each time), to present the mean and standard deviation. HDS metrics are computed over one experiment of 200 sets, due to their high cost.

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.

Data for *storyGen* and *respGen* was generated by the MASS model (Song et al., 2019), fine-tuned on each dataset. 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 all tasks along with additional reproducibility details are in the Appendix B. For each HDS metric, we collected 10 ratings per query from Amazon Mechanical Turk (AMT) workers. While absHDS demands one query per response set, in order to perform simHDS at a reasonable cost, we chose $|\mathcal{S}_c| = 5$, resulting in $\binom{5}{2} = 10$ crowdsourcing queries instead of $\binom{10}{2} = 45$ per set. We evaluate simHDS only for *respGen* due to the metric's high cost and low performance.

Results Table 2 presents results of absHDS, simHDS, and all automatic metrics. In general, n-gram based metrics capture the diversity induced by a temperature sweep, beating HDS and neural metrics. Figure 3 provides a more detailed analysis. Each point represents a single set of responses generated at some temperature. While rank correlation for cosine similarity is high, it is

[4] https://github.com/Tiiiger/bert_score

[5] https://github.com/UKPLab/sentence-transformers

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

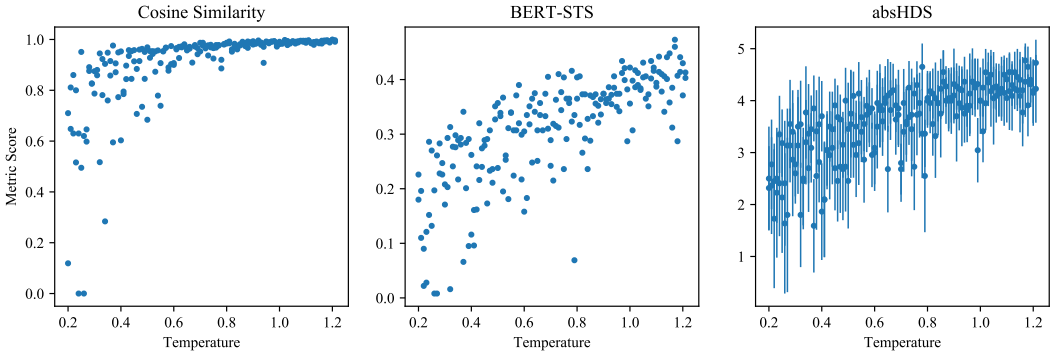


Figure 3: *decTest*: Scatter plot of n-gram-based (cosine similarity), neural (BERT-STS) and human (absHDS) metrics as a function of temperature for *respGen*. Each point corresponds to a single generated set. Error bars of HDS represent the standard deviation over 10 annotator ratings.

	storyGen	respGen	promptGen
distinct-n	**0.76** (0.03)	**0.89** (0.01)	**0.91** (0.01)
cos-sim	0.71 (0.04)	**0.89** (0.01)	0.87 (0.02)
BERT-STS	0.64 (0.04)	0.81 (0.02)	0.84 (0.02)
sent-BERT	0.65 (0.03)	0.80 (0.02)	0.74 (0.03)
BERT-score	0.69 (0.04)	0.87 (0.01)	0.88 (0.02)
absHDS	0.69	0.81	0.79
simHDS	-	0.74	-

Table 2: *decTest* results: Spearman's ρ correlation between temperature and each metric score (mean and standard deviation). *simHDS* was tested only on *respGen*.

	Temperature	Top-p	Top-k
distinct-n	**0.91** (0.01)	**0.84** (0.02)	**0.61** (0.05)
cos-sim	0.87 (0.02)	0.78 (0.03)	0.48 (0.05)
BERT-STS	0.84 (0.02)	0.74 (0.03)	0.55 (0.05)
sent-BERT	0.74 (0.03)	0.63 (0.05)	0.51 (0.05)
BERT-score	0.88 (0.02)	0.77 (0.03)	0.57 (0.05)

Table 3: *decTest* results for different decoding parameters: Spearman's ρ (mean and standard deviation) of automatic metrics for *promptGen*.

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 0.75 and 0.77 Pearson correlation respectively. Thus, Pearson and Spearman correlations disagree on the quality of the different metrics in this case.

While our framework is meant to evaluate diversity metrics, the results of the test let us reflect on the decoding parameters themselves. 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.

Other decoding parameters To compare 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 logarithmic 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 3 presents results for all automatic metrics using the three decoding methods over *promptGen*. Results for other tasks are in Appendix C. We find that Top-p correlates well with temperature along all three generation tasks, whereas Top-k does not correlate with any of them.

6.4 Content Test

In conTest, 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 workers to give 5 responses per context, with as different content and structure as possible. Then we asked the same workers to choose a single response they wrote, and rephrase it 5 times such that the original content will be preserved, while changing the form – this set is used for the low content diversity class.

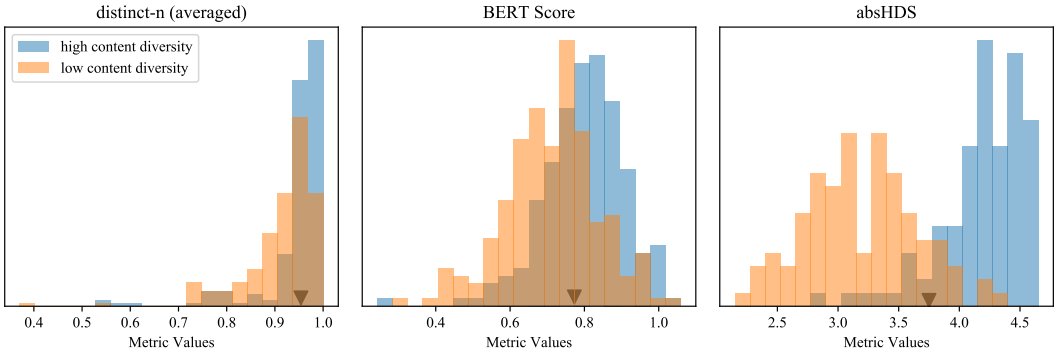


Figure 4: *conTest*: histograms of metric values of n-gram (distinct n-grams), neural (BERT-Score) and human (absHDS) metrics for *promptGen*. The **orange** histogram represents the distribution of the *low content diversity class*, the **blue** histogram represents the distribution of the *high content diversity class* and **brown** is the intersection between the two. Pointing down triangles represent the threshold η of the optimal classifiers. The histograms show how each metric separates the two classes.

A sample from this data is in Figure 1 and more samples in Appendix B. For each HDS metric, we collected 10 ratings from crowdsourcing workers, different than the ones who composed the sets.

Results In addition to Spearman's ρ, we report the optimal single-threshold classifier accuracy (OCA), i.e., the best achievable accuracy in predicting the class of a response set (high or low content diversity) for any threshold η on m_{div}, such that if $m_{\text{div}}(\mathcal{S}_c) > \eta$ the classifier predicts *high diversity*, and otherwise predicts *low diversity*.

Table 4 shows the results. N-gram-based metrics perform poorly, indicating they do not measure content diversity well. 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 better than neural metrics. In addition, n-gram-based metrics saturate both classes to near maximal values, similarly to decTest.

Since conTest isolates content diversity, we used aspHDS to directly rate content and form diversity. *Content* aspHDS gets similar scores to *absHDS*, suggesting little gain in asking directly on the tested aspect. *Form* aspHDS gets low scores compared to absHDS, validating that the form diversity of the two classes is similar.

Content Diversity Benchmark We construct the **M**etrics for **c**ontent **Div**ersity (*McDiv*) benchmark, focusing on metrics for content diversity. McDiv is a dataset containing $6K$ $\{c, \mathcal{S}_c\}$ pairs, ($2K$ for each storyGen, respGen and promptGen) collected as described in this section. McDiv contains a subset of $3K$ examples, termed $McDiv_{nuggets}$, in which *form* diversity was neutralized, providing a difficult meta-evaluation challenge. $\text{McDiv}_{\text{nuggets}}$ was sampled to ensure that the correlation of *distinct-n* (a form diversity metric) is zero over this subset. Applying conTest over the data shows that n-gram based metrics obtain near-zero values on $\text{McDiv}_{\text{nuggets}}$ as expected, and all neural metrics perform substantially worse on $\text{McDiv}_{\text{nuggets}}$ than on McDiv. On conTest, we obtain absHDS annotations for more than 200 random samples from $\text{McDiv}_{\text{nuggets}}$ and obtain 0.7 Spearman's ρ for the respGen task, substantially higher than the best performing neural metric (sent-BERT) score at 0.6. Details and conTest results can be found in Appendix C.

	storyGen		respGen		promptGen	
	ρ	OCA	ρ	OCA	ρ	OCA
distinct-n	0.57	0.77	0.34	0.67	0.33	0.68
cos-sim	0.56	0.77	0.33	0.66	0.36	0.67
BERT-STS	0.6	0.78	0.46	0.72	0.65	0.82
sent-BERT	0.77	0.90	0.59	0.79	0.68	0.81
BERT-score	0.59	0.77	0.49	0.74	0.4	0.69
absHDS	**0.85**	**0.95**	0.63	0.81	**0.78**	**0.89**
$\text{aspHDS}_{\text{form}}$	0.35	0.65	0.56	0.79	0.4	0.68
$\text{aspHDS}_{\text{content}}$	0.84	0.94	**0.67**	**0.83**	0.75	0.88

Table 4: *conTest* results: Spearman's (ρ) correlation between a set's class and each metric score.

HDS Stability: Picking Parameter Values HDS experiments demand expensive human labor. Thus, we need to carefully choose the number of sets and different ratings we ask per set, to get reliable results in a reasonable budget. To this end, we conducted two series of experiments, once increasing the number of sets, and again increasing the number of ratings per sets. By observing results along those two series, we chose to use 200

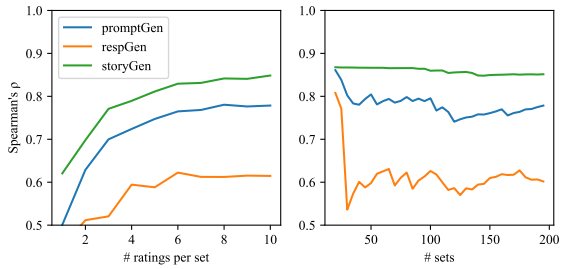


Figure 5: conTest *absHDS* results depends on the number of ratings per set and the number of sets.

sets and 10 ratings per set for all experiments - the minimal values in which results are confidently stable. Results are presented in Figure 5.

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 to 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 watch 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 conTest.

8 Conclusions

This work presents a framework for evaluating diversity metrics as a step toward standardized evaluation. We limit the scope of this work to differences between *form* and *content* diversity, which are key towards understanding different aspects of diversity. Future work can explore other aspects of diversity, 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 Or Nachmias, 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 grant 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*.

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.

Wenchao 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. Eli5: 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 1 (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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. 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, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016c. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. 2019. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1969–1979.

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:1905.08949*.

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 311–318. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, 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.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A HDS Questionnaires

All Human scores for HDS metrics were collected using Amazon Mechanical Turk (AMT) crowdsourcing platform by English native-speaking workers that were specifically qualified for this task. Figure 7 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 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 8 to 11 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 Decoding Test (decTest)

Tables 11 to 19 present data samples from storyGen, respGen and promptGen with the neural testers of decTest, 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 Content Test (conTest)

Tables 20 to 22 present data samples from storyGen, respGen and promptGen with the human testers of conTest, 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 Experiments

C.1 Decoding Test (decTest)

Comparing decTest results of storyGen to other tasks (Table 2), this task is characterised with noisier scores for all metrics (Figures 3 and 6), hence lower ρ values and higher variance. A possible explanation is larger effect of c on the distribution $P_{gen}(s|c)$ in this task.

Tables 3, 6 and 7, present decTest absolute scoring experiment using *temperature*, *nucleus sampling* and *Top-k* decoding parameters as d. Top-k consistently yields lower ρ compared to other decoding parameters, especially for storyGen task. This implies that Top-k represents diversity less reliably than other methods.

Ranking experiment To examine whether we can improve correlation by asking humans to *rank* diversity, rather than providing an absolute score, we designed a ranking version of decTest. 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 ρ 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).[7]

Table 5 summarizes the ranking test results. We observe that humans are better at ranking compared to giving absolute scores (Table 2), 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.

C.2 Metrics for Content Diversity (McDiv)

As elaborated in § 6.4, McDiv is a dataset containing $6K$ $\{c, \mathcal{S}_c\}$ pairs, ($2K$ for each storyGen, respGen and promptGen) collected as described in §6.4. $\text{McDiv}_{\text{nuggets}}$ is a $3K$ subset of McDiv, in which *form* diversity is neutralized, providing a difficult meta-evaluation challenge. $\text{McDiv}_{\text{nuggets}}$ was sampled in a manner that causing *distinct-n*

[7] We consider ties in the metric difference score as a miss.

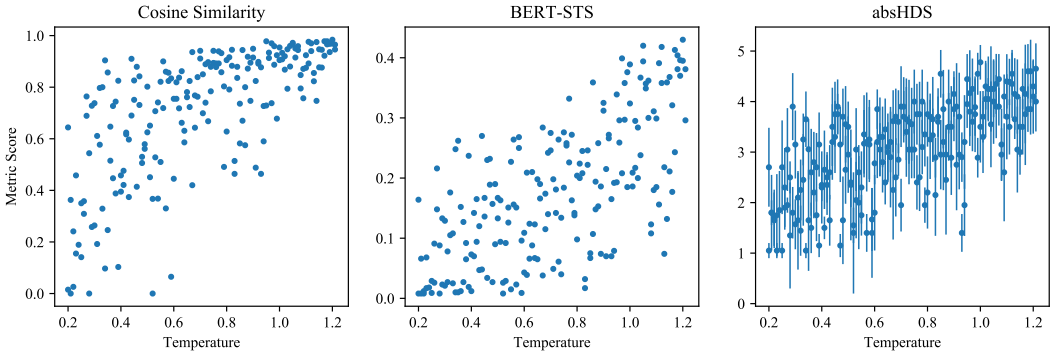


Figure 6: decTest: Scatter plot of n-gram-based (cosine similarity), neural (BERT-STS) and human (absHDS) metrics as a function of temperature for *storyGen*. Each point corresponds to a single generated set. Error bars of HDS represent the standard deviation over 10 annotator ratings.

	storyGen		**respGen**		**promptGen**	
	ρ	acc	ρ	acc	ρ	acc
distinct-n	**0.88**	0.88	0.86	0.9	**0.91**	**0.91**
cos-sim	0.86	0.88	0.87	**0.91**	0.9	**0.91**
BERT-STS	0.84	0.84	0.85	0.88	0.9	0.89
sent-BERT	0.85	0.86	0.83	0.85	0.85	0.85
BERT-score	**0.88**	**0.89**	0.88	0.89	**0.91**	0.9
rnkHDS	0.87	**0.89**	**0.89**	0.9	0.89	0.88

Table 5: *decTest* ranking results: Spearman's (ρ) correlation between temperature differences and each metric score. Accuracy (acc) of classifying which set has the higher temperature. Standard deviation is up to 0.02 for all automatic metrics for both Spearman's correlation and accuracy.

metric to score zero correlation in conTest over this subset. The method of sub-sampling was meant to approximately equalize the distributions of the two classes, *low* and *high* content diversity, over the scores of distinct-n metric, and was performed as follows:

- Sort all collected samples (from both *low* and *high* content diversity classes) according to their *distinct-n* score.
- Divide the sorted samples to groups with fixed size (40 samples each in our case).
- From each such group, randomly sample the same amount of samples for each of the two classes. For example, if a group contains 5 *low* content diversity samples and 35 *high* content diversity samples, we can sample at most 5 samples for each class.

Resutls We applied conTest for all the collected data for each of the three NLG tasks (see Tables 8 and 9). By design, n-gram based metrics score near-zero correlation on $\text{McDiv}_{\text{nuggets}}$, making *high* and *low* content diversity classes almost indistinguishable for those metrics, which relay on text surface level features only. Neural metrics perform strictly worse on $\text{McDiv}_{\text{nuggets}}$ than McDiv. In addition, we applied conTest on 200 randomly sampled $\{c, \mathcal{S}_c\}$ pairs from $\text{McDiv}_{\text{nuggets}}$ for respGen task (see table 10). Compared to Table 4, The gap between the best performing neural metrics (sent-BERT) and absHDS was increased in favor to HDS (0.04 compared to 0.1 difference in Spearman's ρ).

D Additional Reproducibility Details

Collected data and code All the collected data, metric scores per samples for each of decTest and conTest, as well as code for running and visualizing the tests, are publicly available[8]. The collection methods are elaborated in Section 6.

Original data We provide additional data for the original three datasets used in Section 6.

- ROC Stories dataset[9] (Mostafazadeh et al., 2016) used for storyGen task contains $96K/1K/1K$ train/validation/test titles and five-sentence stories. We used the samples without pre-processing for both fine-tuning MASS model and generate samples for our tests.
- Reddit comment-response dataset used for respGen task contains $37M/1M/1M$ train/validation/test comment - response pairs, extracted from the social website `reddit.com` scraped by `pushshift.io` followed by the pre-process described in

[8] https://github.com/GuyTevet/diversity-eval

[9] www.cs.rochester.edu/nlp/rocstories/

(Hashimoto et al., 2019). We used the samples without further processing for both fine-tuning MASS model and generate samples for our tests. To the best of our knowledge, this dataset is not publicly available at the moment.

- CMDC dataset[10] (Danescu-Niculescu-Mizil and Lee, 2011) contains $108K/30K$ train/test sentence-response pairs extracted from movie scripts. We extracted the first three words from the sentences (used as contexts for the original task) to be the context of our task. We did not use this data for training since we used GPT-2 without fine-tuning for promptGen.

Auto-generated data For decTest, we used two pre-trained generative models for generating responses given the contexts:

- For storyGen and respGen tasks, we used MASS[11] (Song et al., 2019) (6L-1024H-8A architecture suggested by the authors), pre-traind as described in the original paper. For each task separately, we fine-tuned MASS using the training division of the dataset corresponding to the task. Fine-tuning was done using $200K$ examples over 30 epochs, and took 23 hours using a single *TITAN Xp* GPU core. Inference with the fine-tuned model takes 65 milliseconds on average per response set containing 10 responses with the same GPU core.

- For promptGen task, we used Hugging-Face implementation[12] of GPT-2 *large* (36-layer, 1280-hidden, 20-heads, 774M parameters) (Radford et al., 2019) pre-traind as described in the original paper. We used this model as-is, without fine-tuning. Inference takes 0.6 second on average per response set containing 10 responses with a single *TITAN Xp* GPU core.

Tests Runtime Given metric scores per sample, running each of the tests with 200 samples takes less than a minute on a standard *Intel i7* CPU.

[10] www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

[11] github.com/microsoft/MASS

[12] github.com/huggingface/transformers

	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 6: decTest results for different decoding parameters: Spearman's ρ (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 7: decTest results for different decoding parameters: Spearman's ρ (mean and standard deviation) of automatic metrics for *respGen*.

	storyGen		**respGen**		**promptGen**	
	ρ	OCA	ρ	OCA	ρ	OCA
distinct-n	0.53	0.74	0.52	0.74	0.48	0.75
cos-sim	0.53	0.74	0.52	0.74	0.60	0.77
BERT-STS	0.57	0.74	0.61	0.78	0.78	0.89
sent-BERT	**0.75**	**0.87**	**0.68**	**0.83**	**0.8**	**0.9**
BERT-score	0.60	0.77	0.56	0.78	0.54	0.74

Table 8: conTest results for McDiv; Results for automatic metrics over all the samples (2K per task).

	storyGen		**respGen**		**promptGen**	
	ρ	OCA	ρ	OCA	ρ	OCA
distinct-n	-0.002	0.49	-0.002	0.49	-0.003	0.49
cos-sim	0.04	0.53	0.08	0.55	0.22	0.60
BERT-STS	0.34	0.64	0.39	0.68	0.68	0.83
sent-BERT	**0.63**	**0.80**	**0.53**	**0.76**	**0.73**	**0.85**
BERT-score	0.35	0.66	0.33	0.65	0.35	0.65

Table 9: conTest results for $\text{McDiv}_{\text{nuggets}}$ subset; Results for automatic metrics over all the samples (1K per task).

	storyGen		**respGen**		**promptGen**	
	ρ	OCA	ρ	OCA	ρ	OCA
distinct-n	0.04	0.57	-0.01	0.46	0.12	0.56
cos-sim	0.05	0.54	0.04	0.54	0.28	0.62
BERT-STS	0.34	0.68	0.39	0.69	0.72	0.9
sent-BERT	0.68	0.85	0.6	0.79	0.75	0.88
BERT-score	0.37	0.69	0.34	0.68	0.38	0.69
absHDS	**0.78**	**0.9**	**0.7**	**0.85**	**0.84**	**0.94**

Table 10: conTest results for 200 random samples from $\text{McDiv}_{\text{nuggets}}$ including HDS.

Context	Response set ($\tau = 0.25$)	Response set ($\tau = 0.8$)	Response set ($\tau = 1.1$)
Fire next door. John woke up smelling like something was burning. He went outside. He saw the fire next door. He called the authorities.	• 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. • It was a fire. • It was a minor fire. • It was a fire. • It was a fire. • It was a minor fire and they put it out.	• 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. • They arrived and put it out. • It was a scary but beautiful fire next time he went in. • It was a fire. • It was a put out and John was able to put it out. • It was a scary place to live.	• 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. • They arrived and put out the fire. • It turned out to be a clear, fact. • It was a fire. • It was a scary meal for wet clothes. • It was a scary stealing from him.
Charles Shaves His Chest. Charles decides he no longer wants chest hair. He shaves his chest hair with a razor. Although his skin is irritated, he hair is now gone. Charles is satisfied.	• He is glad that he finally shaved. • He is glad that he finally shaved. • He is very happy to have a chest hair. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that he finally shaved.	• He is very happy to have a chest hair. • He is glad that he finally shaved. • He is very happy that he no longer has chest hair. • He is glad that he finally shaved. • He is glad that he finally shaved. • He is glad that his chest hair is no longer shaved. • He is glad that he finally shaved • He is glad that he finally shaved. • He is glad that he finally of shaving. • He is glad that he finally shashock his chest hair.	• He is glad that he finally peaceful his voice again. • He is glad that he finally shaavoided his chest hair. • He is glad that he finally downcome. • He is ready to start better cooperation his fireplace. • He is glad that he finally shaved. • He is glad that he no longer has chest hair. • He is happy to finally be shaved. • He is very happy to have ster good ride of his life. • He is glad that he finally shaved. • He is glad that he finally has chest hair.

Table 11: decTest data samples for storyGen task and different temperatures.

Context	Response set ($p = 0.208$)	Response set ($p = 0.64$)	Response set ($p = 1$)
Really Bad Decisions. Jake was nervous about a meeting at work the next day. He decided to have a drink to relax. Unfortunately Jake kept drinking. He was really hung over the next day.	• He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting.	• He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting.	• He did not get to the meeting anymore. • He missed his meeting. • He passed out and failing the meeting • He missed his meeting. • He missed his meeting. • He missed his meeting. • He missed his meeting. • He passed out and was kicked out of the meeting. • He missed his meeting. • He missed his meeting.
Family Night Food. Tonight, my mom ordered Mexican food for family night. She got it from my favorite Mexican place in town. When it arrived, it was hot and smelled wonderful. We devoured it with gusto.	• After a few hours of take it home we all enjoyed its night. • After a few hours of take it home we all enjoyed its night. • After a few hours of take it home we all enjoyed its night. • After a few hours of eating everyone was satisfied. • After a few hours of take it home we all enjoyed its night. • After a few hours of eating everyone was satisfied. • After a few hours of take it home we all enjoyed its night. • After a few hours of take it home we all enjoyed its night. • After a few hours of take it home we all enjoyed its night. • After a few hours of eating everyone was satisfied.	• After dinner, we all went home to cook Mexican food. • After a few hours of cooking she was tired and ready to eat. • After dinner, I always put got ready for Christmas. • After dinner, I helped her do the dishes. • After a few hours of dinner, the food was amazing. • After a few hours of take it home we all enjoyed margaria. • After dinner, I was ready to take on work the next day of • After dinner, I was sad to say goodbye to her • After a few hours of take it home we all enjoyed one bite. • After a few hours of eating everyone was satisfied.	• Even though my stomach was gone, I was sad it was finally pockets • After dinner, I alone. • All in all the family while my mom finished the food. • After a few hours of dancing, she pianed. • Afterwards I'd never finish single night a week. • Afterwards we all went to sleep in my woods. • Afterwards I helped her do much better than my wife. • Afterwards my mom helped me finish my dinner. • After a few hours we all enjoyed storm blood from the Italy. • After dinner, we all enjoyed some good food together.

Table 12: decTest data samples for storyGen task and different p values (nucleus sampling).

Context	Response set ($k = 3$)	Response set ($k = 32$)	Response set ($k = 318$)
Loud Noise. Jane was trying to take a nap. She heard a loud bang in the kitchen. It woke her up. A dish had fallen off the counter.	• Jane pulled over and started to clean herself. • Jane was horrified and dropped her favorite food. • Jane was able to finish her car and take a nap. • Jane was able to finish her nap . • Jane was able to finish her nap . • Jane stopped at the store to buy a new one. • Jane 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 .	• Jane was able to catch a car using the seat. • Jane stopped at the store to buy a new book. • Jane was sad her cat dropped out of the kitchen. • Jane screamed. • Jane was horrified to find her car broken down on the floor. • Jane was horrified and dropped her pay phone. • Jane was easily able to grab a nap. • 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.	• Jane comes, noticed a lot of food left under it. • Jane was horrified and dropped her book. • Jane remembered to take a nap. • Jane was since she took a nap while she waited for the refun • Jane knew she had no time to finish her book. • Jane was glad. • 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 monthed.
Headache. Kate was wearing big over the ear headphones. But they were tight and squeezing her head. She tried to adjust them to relieve the tension. But nothing really worked.	• Kate decided to go to the store and buy some ear phones. • 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 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 ear phones. • She decided to go buy some ear plugs.	• 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. • She 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 a ess instead. • Kate decided to go to the store and buy some fake headphones. • She decided to go to the store and buy some ear phones. • 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 some ear phones. • Kate decided to go to the gym and use some saw no more watching T • Kate decided to go to the store and buy some cute phones. • She decided to go buy a fake headphones instead. • She decided to go to the store and buy some concert.

Table 13: decTest data samples for storyGen task and different k values (Top-k).

Context	Response set ($\tau = 0.25$)	Response set ($\tau = 0.8$)	Response set ($\tau = 1.1$)
it is doing okay i just want people to talk more	• i think you mean people • i think its working now • i just want people to talk more • i think i just want people to talk more • i dont think you understand what youre saying • i like to think so • i know that feel • i dont know what to say • i think its pretty cool • i think its working now	• i cant believe you didnt think this through i actually havent had • may i ask why • i just hope its not going to happen • its alright dont even want to talk • what did it say i was using windows 95 anyway xd • thanks for reading the thread • i understand and hope they dont mind that • want to talk more • its doing ok body language great voice acting and music • all good i know your pain	• im just hawaii everyone wants to be friends lt go talk to people from • its pretty smart we didnt run a year ago and have turned back but • id lick your teeth but tha t sucks smart people • it really didnt stop i think you need to go back to normal now • sounds like a plan to me • we we can try forever practice beer • okay but another two more 200 radio views im going to ore my comment of • team building looking bro • yep they said that things have changed • yeah we thank you random stranger
girls who say no not like it matters	• i dont like it at all • i dont like it either • i dont like it either • i think you mean girls who say no not like it matters • i think that is the most accurate description of this thread • i like this one • girls who say no dont like it matter • i dont like it either • i know a guy who says no and he doesnt like it • girls who say no dont like it matter	• wait so there are girls who say no like it matters • i have a friend who doesnt like it but i always say no • and you dont like to play with girls • i say no i dont want it to • girls who say no dont like it care • i see what you did there • girls who say no need to make eye contact with girls • girls who say yes dont like it does • i really dont care if its not gay or straight out programs or not • ugh that game was just awful troll	• how about mine like sales figure instead of higher • definitely not as stupid as that • agreed is that true • refer to my highest iq you and me • less keep it up • dude ive never seen such an complaining member with a legitimate thought • not that theres anything wrong with that or are more wise than telling want • girls make videos youve never met • gtthe girls who say no or a one time thing people thats • which is weird since it is a girl i know so sad

Table 14: decTest 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 • i dont think so • i dont think so • no i dont think so • no idea i dont know • i dont think so • no i dont think so • no idea i just saw him live in new orleans • i dont think so • no idea i just saw him live in new orleans	• he is in nyc • he does i hope he does • i dont think so • he lives in new orleans • yes he does he is in florida • im guessing not that i know of • hes a small town i dont know what happened there • nope still in new orleans • i do not i dont know if he still lives in new orleans • i dont think so	• aw screw you guys the gimme a shower and ill send you my money • he does im so confused by how this is relevant • they spent all that much in new orleans up in first place • still living in new orleans according to most of them it jesus mustve worth • anyone else want to wait for him to go back home • how many tickets would you say the willing said if he warn you that hes • im guessing he lives there as well • yep rick albeit not • he lives in ny • no i dont get it any more guess i may canadian
its that hanging thing in the back of your mouth	• its a tattoo of a hand • its a good thing its not • i think its a banana • i was thinking the same thing • its a very nice thing to do • its a hole in the bottom of my mouth • i think its a good thing • its a pretty cool thing to do • thats what i thought • i think its a butt	• well it is kind of like a mac • that would be a good idea • its my cake day • you can see it in his mouth • its the junk that was pooping in • i was thinking the same thing • i always see that • i didnt know that was possible • this is my favorite comment of the day • thats not a sticker	• i think you just flash • it is and feels good • my ex and i reach our arms cheating into the door for almost a minute • my mind went to side with that • yep terrible for my teeth • exactly what i was thinking • why do i look so contest • and where else are you going • just dont wear a helmet • and it looks like something some one would give to get in on

Table 15: decTest 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 • he has to be a good center for that • he is a great center of football in his prime • hes been playing in his prime for a long time • he was a great back in the day • hes been playing for a while now • i dont know about that he was pretty damn good at that • i dont think he was ever in his prime • i dont think he is a prime minister • i dont know why but i think he is a very good player and	• and his career as well • agreed the way hes playing is awesome • it has to be • this is just called a job • and then being on the field for the first time • i dont see him doing that often enough • he just likes to party in the kitchen • at this point hes going to be a great star for the rest of the • only if he pays well • the only thing that can make that kind of difference is how much time you	• yeah my feeling i mean we dont like it but it happens all the • you are one for real • they still have a rule saying they might not be injured yet • it really is a necessary thing to do • finally some reason to continue watching him at some point • yet that would be epic • not to mention eating curry dinner is a privilege • i just dont want to turn over for this goal like he does in • gt playing in his prime is truly a privilege ftfy • so is saying he is in high school
im going this evening when she usually works	• i think you accidentally a word • you are a good man • i hope she works • im going to the same time as you when she usually works • i am so sorry to hear that • i hope she works for you • i am so jealous of your work i am so jealous • i hope you have fun • i hope you get a job at a local bar • i hope she works for you	• i see what you did there • where in the world are you going to put your socks on • she usually does but she has to keep up to date with her in the • awesome what do you want for it • how does a women have a relationship with someone • did you get the job • thats where she goes • i know im also going this friday • thats a great idea • well make sure youre there too good luck	• thats what you like to hear • go see her later this evening • i read that as she usually fails • okay ill see you there brah • i always thought that it was for the sake of having the girl play as • that can be expected here • thats it im down now • i dont see why it would matter if shes married • i will be the ex gf • she still does make her phone calls the whole office

Table 16: decTest data samples for respGen task and different k values (Top-k).

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 is the real problem. The hacking is the! • **Not the hacking** of the DNC, but the leaks of the emails of the Democratic National Committee.! • **Not the hacking**, but the way it was done. The FBI's investigation into the! • **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** of the DNC. The hacking of the DNC was a "false flag! • **Not the hacking** of the DNC. But the hacking of the RNC. The DNC hack! • **Not the hacking**. The hacking is the problem. The hacking is the problem! • **Not the hacking** of the DNC, but the leaking of the emails. The DNC was hacked,!	• **Not the hacking** after all? I'm sure the nation-states that are involved in! • **Not the hacking** that happened on the internal networks of the Energy Department. In fact, according to! • **Not the hacking** of the American public but rather the fraudulent Heisenberg principle that seemed to be! • **Not the hacking** that took place in the DNC last year or the release of hacked emails during the! • **Not the hacking** futurists Cardboard inventor and self-described tinkerer Dennis! • **Not the hacking** alone. In the first half of the report, the hackers tried to create fake! • **Not the hacking**. The hacking is the NSA's new SHIELD technology. It is! • **Not the hacking** and hacking and hacking of the world government. I know this man is a man! • **Not the hacking** aspect, but the pressure exerted by the Trumpistas. But also the Russia angle! • **Not the hacking**, but the willingness." The evidence of interest in this case comes in!	• **Not the hacking** experience of a CIA VRO crunch nine months ago—JumpStart for 2016 jumps! • **Not the hacking**, David.) The directory was flagged in a document it created in late last year! • **Not the hacking** of Democratic Party systems - said the Russian team's activity represented "just the beginning! • **Not the hacking**, of course – which these sources sounded more concerned about than being attacked 140 times! • **Not the hacking** story is over. But yet there's another reason not to rush out such statements! • **Not the hacking**-either.- These were scattered in the workshop.(Expanded- being guys with! • **Not the hacking** of private material of elected officials, e.g. emails, even if the! • **Not the hacking** has happened yet!!!!!!!!!!!!! • **Not the hacking** rumours have cost him any of his followers, least of all the proprietors of! • **Not the hacking** group behind the breach of Sony, which has posted the staffer's information online,!
• **How is our** new technology helping us to do that? We are using a new technology! • **How is our** system different from that of the United States? The United States is a! • **How is our** approach different from that of the other 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?" he asked.! • **How is our** work different from that of other organizations? The work of the Center for! • **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 different from the one that was used in the past? The system! • **How is our** country supposed to be a beacon of hope for the world if we have to look!	• **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** minister giving it to you? Isn't it? It's got a bit of! • **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** challenge different? The only difference is that this challenge is about building an! • **How is our** nation governed?" As Obama moved into his second term, he is increasingly! • **How is our** recommendation different from what more traditional veterinarians do? We don't believe! • **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** Internet even even connected with our corporate tracks? Every cell phone on the planet knows! • **How is our** developer name attached to the icon? Since the Planetside icon is use internally! • **How is our** food paradise created? Artificial chemical fertilizers. So these aren't GMOs, but! • **How is our** acquisition* worth - BOARD ROLL (Least Significant Equivalents)! • **How is our** transit plan addressing this problem? Under our old plans, Burlington Buses! • **How is our** mind different than any other part of the body?" A Broader View! • **How is our** campaign working? Bitcoin launches alongside psychological research showing that people pay a lot! • **How is our** mentioning application related to a related method (#five with two in queue) page such! • **How is our** having to resort to roundabout hypotheticals to argue that Stewart may secretly want! • **How is our** blood working out for you?" a statewide voter got an outpouring of rename and!

Table 17: decTest data samples for promptGen task and different temperatures. Bold text is the 3-words prompt context.

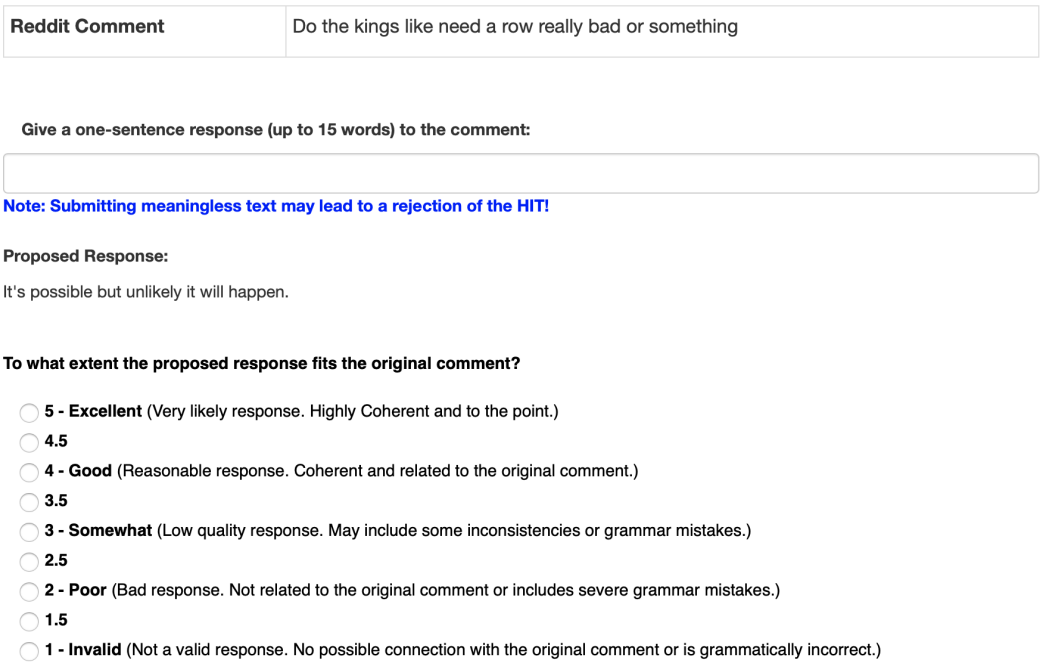


Figure 7: Warm-up part, starting each AMT HDS task. It includes the context, and a single response generated by the tester. The worker is asked to generate response of hers/his own and rate the quality of the tester's response.

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** end of the first part of this series. I hope you enjoyed it.! • **So that's the** first thing I want to say. I'm not going to be the guy! • **So that's the** thing about being a professional. You have to be able to handle the criticism! • **So that's the** way it is. I don't think there's any way to change it! • **So that's the** problem. It's not just that the government is failing to protect! • **So that's the** thing about this. It's not just about the money. It's about! • **So that's the** end of the story. The next step is to create a custom! • **So that's the** case. So, what's the problem? Well,! • **So that's the** first time I've ever seen a real one. I'm not!	• **So that's the** state of the campaign. Now, what I do want to talk about is! • **So that's the** thing: For as much as I love TLC, it's hard to! • **So that's the** idea, anyway. The last two seasons have been about doing that. It! • **So that's the** end of the half-hour segment. The next half-hour! • **So that's the** situation we're in," he said. "We're in the! • **So that's the** thing, I don't know if you know, but in general it's! • **So that's the** difference between the kinds of things that people will be talking about on Wednesday,! • **So that's the** $2.3 billion. Here's the issue: You're! • **So that's the** standard for using memcpy(). It's fine to use memc! • **So that's the** next step, and the next step is to try to figure out what's!	• **So that's the** first time you want to punch somebody, not miss before." The Seahawks would! • **So that's the** science behind the Broadwell-E processors from Intel that Intel launched last fall! • **So that's the** instinct from other teams, that they're a headache. - Ramsay MacDonald,! • **So that's the** white whale right there about too much debt. And then what you! • **So that's the** end of our discussion about the causes. What happens when we look at the! • **So that's the** cover of inhibition against "chronic" or "adaptive" stimulants! • **So that's the** way the story goes, but exactly how is cloud providers going to restrict Their! • **So that's the** beginning, the beginning of the show, I guess five minutes." ! • **So that's the** Indie Mobile Game Week Honoring Winners!!!!!!!!! • **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 them?" "I do," he said. "I'm not! • **do you listen** to the voices of the people?" "I do," said the king! • **do you listen** to the song?" "I don't know . I don't know! • **do you listen** to the music?" "I do." "You're not! • **do you listen** to the news? I do. I'm a big fan of the! • **do you listen** to me?" "Yes, I do." "I'm! • **do you listen** to the other side?" "I don't know. I don't! • **do you listen** to the other side?" "I do," said the boy. "! • **do you listen** to the news? No, I don't. I don't listen!	• **do you listen** to the current draft? I listen to the current draft. I'm! • **do you listen** to it?" It's easy to hear the "why?" but when! • **do you listen** to the people that come here?" "No, I'm too busy! • **do you listen** to the thing?" "Of course I do. I've been reading! • **do you listen** to those who are opposing it, who want to create a situation in which a! • **do you listen** to music or watch TV? How often do you cook or clean? How much! • **do you listen** to them? It's like the first time you got into something and it just! • **do you listen** to your father? We'll leave it to the gods to decide." ! • **do you listen** to music? I like to listen to music, but I don't really know! • **do you listen** to my story and see if you like it?" "I think you!	• **do you listen** to Human Fly?, which YouTuber Nico Perri collaborated on, and Google! • **do you listen** to the acapella lyrics out of context and express the feeling?" It's! • **do you listen** to Michael Kiwanuka-Smith who writes, "The American Journalism Review discern! • **do you listen** to my songs as I said," Ramckhalter said. "You feel! • **do you listen** to U.S. 90 night at this time of the year? ! • **do you listen** to that as well?" "The question was not, 'Who is! • **do you listen**?" He asks, leaning forward as he woodenly talks to him. "Listen! • **do you listen** to those books and sway him so much? No. He was deeply brainwashed! • **do you listen**?' Simon(lol).I feel like i'm in a Kurdish Genocide. I! • **do you listen** to value authenticated queries from your menu when running count? And if not, then!

Table 18: decTest data samples for promptGen task and different p values (nucleus sampling). Bold text is the 3-words prompt context.

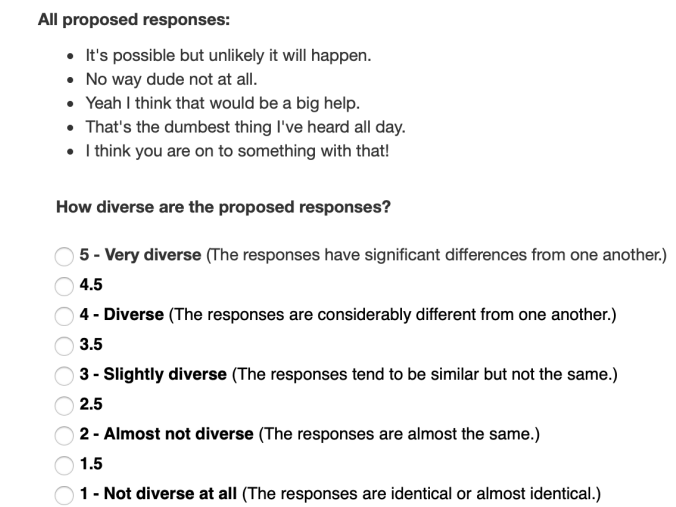


Figure 8: absHDS question along with the evaluated response set (conTest in this case).

Response set ($k = 3$)	**Response set** ($k = 32$)	**Response set** ($k = 318$)
• **I'm kidding. You** can't do that. "I've been trying to find! • **I'm kidding. You** can't be serious." The man was a little surprised.! • **I'm kidding. You** can't do that, because the game's a little bit different.! • **I'm kidding. You**'re the only one who can do it, you know. You're! • **I'm kidding. You** can have it. "It's not a bad idea,"! • **I'm kidding. You** don't need to be so serious." "No, I! • **I'm kidding. You**'re not. I'm just a little bit of a dickhead.! • **I'm kidding. You**'re not. You're not. I know. I know. You! • **I'm kidding. You** can't do it. "I don't want to be! • **I'm kidding. You**'re not going to be able to see it, and you'll be!	• **I'm kidding. You**'re too nice. We need someone in the government to talk! • **I'm kidding. You** just sit there, I'll have the other side do it. What! • **I'm kidding. You** can be my roommate for the holidays in a few weeks. You don! • **I'm kidding. You** can get the full version of the file above, and also the original! • **I'm kidding. You** don't look very strong." "You look as strong as! • **I'm kidding. You**'re right about the last guy. He's a nice guy. But! • **I'm kidding. You**'ll be happy to know it's the first time that's happened since! • **I'm kidding. You** will come down with the same problem, or some other sort of problem! • **I'm kidding. You** might have seen me in a lot of other things. I'm actually! • **I'm kidding. You** should go and see a doctor. In fact, I'm!	• **I'm kidding. You**'re kidding?" "I'm not." "Why! • **I'm kidding. You**'re not." "What? A dick of the heart?! • **I'm kidding. You**'re looking at a new version," said Zilch, who was! • **I'm kidding. You** know when someone takes to the streets to protest? It's common for! • **I'm kidding. You** are definitely a complete free agent," said Caruthers. ! • **I'm kidding. You** can have another at first, but don't start just jumping ahead/! • **I'm kidding. You**'re just a teenager, aren't you?" It ends there, your! • **I'm kidding. You** were never fully persuaded." "Perfect, I am not,"! • **I'm kidding. You** are also in a worse case scenario for someone who was on $2500! • **I'm kidding. You** know…" "I should have stopped him; I shouldn't!
• **Where did he** go?" I ask, looking at him. "I'm not sure. He! • **Where did he** get the idea to do this? He had been working on a book! • **Where did he** come from? He was born in the city of Karkaros! • **Where did he** go?" I asked. "I don't know," she said. ! • **Where did he** go?" "I think he went to the hospital," she said.! • **Where did he** get the idea for the name? I think it's a combination of! • **Where did he** get the idea to make a movie about the Holocaust? "I had a lot! • **Where did he** get that idea? "I was just trying to make a statement,"! • **Where did he** get that from? He's a very good writer. I don't know what! • **Where did he** go? Where was he? Where was he? He's gone. !	• **Where did he** come back from? [The Doctor is sitting in a chair. Amy! • **Where did he** find the money?" asked a reporter from the BBC. "Is anybody else there! • **Where did he** grow up?" But the boy answered, "He always loved to read! • **Where did he** get that idea?" he asked. "I didn't know. I've never! • **Where did he** come from?" You're looking for that missing piece. Maybe you're missing the! • **Where did he** come from? He was, I think, from a small island about midway between! • **Where did he** come from, to be sure?" he asked, "I know he came from! • **Where did he** go? [A little while later] I am about to say this! • **Where did he** hear about my story? I couldn't tell you. He'd only heard of! • **Where did he** come from? From a place called "the City of the Sun."!	• **Where did he** at the time in his day seek the God he worshipped? He said: "! • **Where did he** earn his master's degree? He is part of a class of doctoral students who! • **Where did he** learn to play guitar?" I asked, puzzled. "Before I joined! • **Where did he** come from?" "Australia," said Peter. "How could! • **Where did he** hear this, you might ask? Of course, he'd heard of it.! • **Where did he** go? He's probably dead – or dead and buried within the walls! • **Where did he** earn $150 million on his way to a $5 billion makeover? ! • **Where did he** learn to make his own sticks, or for that matter, hang a stick on! • **Where did he** learn to skate, anyway? Go here and watch this beautiful skater! • **Where did he** get this idea from? What do you think about it? I get!

Table 19: decTest 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 wanted to watch a big new action movie. She had been waiting a long time for it to come out. When tickets became available she was too busy. By the time she had a chance to buy some it was sold out.	• Jane cried over the fact that she couldn't watch it and just gave up looking for a ticket. • Jane decided to look for a scalper that would sell her the ticket for the movie that she really wanted to see. • 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 posted a status on her social media accounts asking her friends for any spare ticket that she is willing to buy. • Jane resorted to contacting her old friend who is working at a huge movie theater hoping she can help her get 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. • Desperate to watch the movie, Jane called her friend, who works at a movie theater, asking for a ticket to that movie. • Jane recalled that her friend works at a movie theater and hoped that she can help get a ticket for that movie. • 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.
Beavers. My friend has some beavers in his backyard. They come up from the creek by his house. He invites my over and we watch them. We take pictures of them and send them to our friends.	• They are fascinating animals. • Our friends love getting the pictures. • Sometimes his dogs chase them. • They are building a dam on the creek. • They won't let us get too close to them.	• They are busy gathering sticks to make a dam. • The dam they are building is almost complete. • It's fascinating to see their workmanship building a dam. • They are turning the creek into a pond by building a dam. • They all work together with careful engineering to build a dam.

Table 20: conTest 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 • OMG I can't wait • I thought they canceled it • What channel is it on • I only watch nature programs on BBC	• Lead actor is soooo hot • Did you see the cliffhanger at the end of the season • I've been waiting for it to return for weeks • I'm totally gonna binge watch last season • 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 don't need you telling me how to keep things fresh, take a hike. • Girl, you're the fresh one around here. • This post might be better in the life hacks section. • This is actually a useful bit of advice.	• I find merit in this input. • That information will serve me well. • Thanks, that's really good to know! • Such knowledge is certainly beneficial. • Wise words, I will heed them.

Table 21: conTest data samples for respGen task.

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** omelet that is the most amazing ever. • **Suppose there's an** airplane ticket that's even cheaper. • **Suppose there's an** actual deadline for this paper. • **Suppose there's an** event that we can go to this weekend.	• **Suppose there's an** airline that costs less. • **Suppose there's an** flight that isn't as expensive. • **Suppose there's an** air travel fare, but doesn't cost as much. • **Suppose there's an** way to fly there that is low cost. • **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** dancing with your wife at the wedding. • **Nothing remotely like** singing Justin Bieber's greatest hits • **Nothing remotely like** falling down a hill • **Nothing remotely like** getting yelled at	• **Nothing remotely like** being super full and satisfied. • **Nothing remotely like** getting to taste many different foods. • **Nothing remotely like** starting the day off right. • **Nothing remotely like** doing exactly what I want to do. • **Nothing remotely like** feeding myself with great food.

Table 22: conTest data samples for promptGen task. Bold text is the 3-words prompt context.

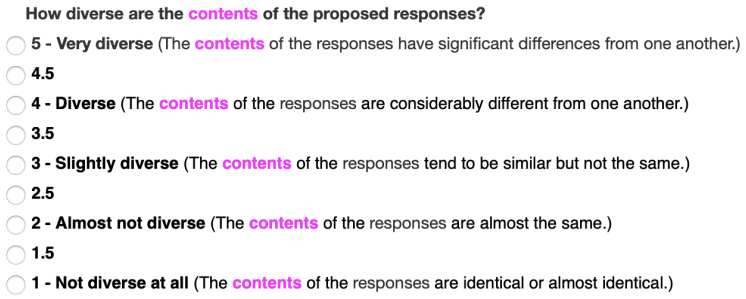


Figure 9: aspHDS question (content in this case). The response set is the same as presented for absHDS question.

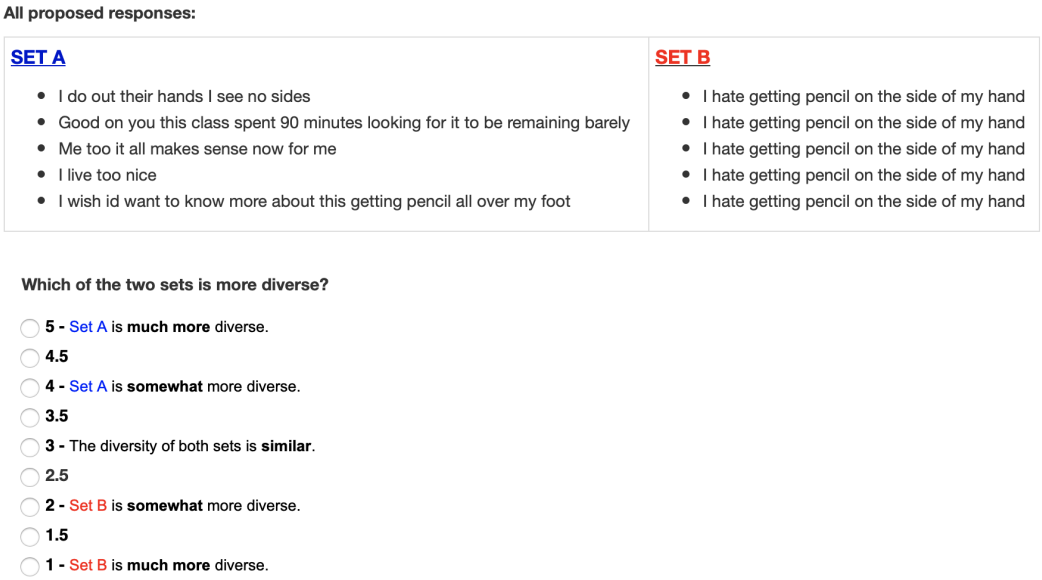


Figure 10: rnkHDS question along with the two evaluated response sets.

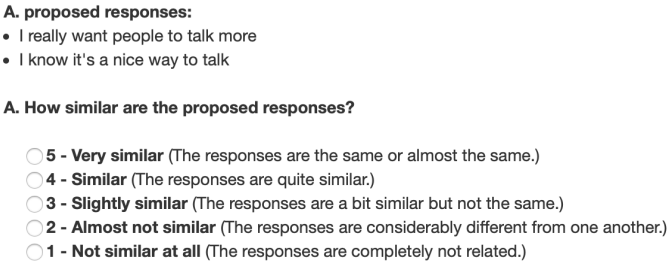


Figure 11: simHDS question along with the two evaluated responses.