Standardizing the Measurement of Text Diversity:
A Tool and Comparative Analysis

Chantal Shaib[1][†]   **Venkata S Govindarajan**[4]   **Joe Barrow**[3][†]   **Jiuding Sun**[1]
Alexa F. Siu[2]   **Byron C. Wallace**[1]   **Ani Nenkova**[2]

[1]Northeastern University, [2]Adobe Research, [3]Pattern Data, [4]Ithaca College

shaib.c@northeastern.edu

Abstract

The diversity across outputs generated by LLMs shapes perception of their quality and utility. High lexical diversity is often desirable, but there is no standard method to measure this property. Templated answer structures and "canned" responses across different documents are readily noticeable, but difficult to visualize across large corpora. This work aims to standardize measurement of text diversity. Specifically, we empirically investigate the convergent validity of existing scores across English texts, and release diversity, an open-source Python package[1] for measuring and extracting repetition in text. We also build a platform[2] based on diversity for users to interactively explore repetition in text. We find that fast compression algorithms capture information similar to what is measured by slow-to-compute n-gram overlap homogeneity scores. Further, a combination of measures—compression ratios, self-repetition of long n-grams, and Self-BLEU—are sufficient to report, as they have low mutual correlation with each other.

1 Introduction

LLM-generated texts are typically evaluated with respect to accuracy or factuality, e.g., as measured via entailment (Tang et al., 2023), or text quality aspects such as coherence and fluency (e.g., estimated using LLMs as evaluators Liu et al. 2023). When reference summaries are available, the similarity of generated outputs to these is also often measured (e.g., via ROUGE; Lin and Och, 2004). A complementary dimension of model performance is *diversity*, or how much "boilerplate" content is repeated *across* LLM outputs.

There is a distinct lack of standardization in reporting diversity in ML datasets (Zhao et al., 2024). We address this by introducing diversity,

an open-source Python package for evaluating text diversity,[1] along with a web-based UI that allows users to visualize repetition in their corpus,[2] providing an intuitive, efficient tool for text analysis that permits: (i) Viewing repetitive text and Part-of-Speech n-grams; (ii) Quickly computing diversity metrics, and; (iii) Interactively highlighting and matching repetition in documents. Both the package[3] and UI code[4] are open-sourced under the Apache 2.0 license.

We run existing diversity metrics over English language outputs from several LLMs to identify a few (mostly) independent scores that characterize repetition. We also examine diversity in downstream datasets such as instruction tuning. Finally, we show that *compression ratio*—compressed over original texts size—is a fast, easy to compute score sufficient to capture the information in all token/type ratio related alternatives. But we emphasize text length as an important confounder when assessing diversity: No reliable conclusions can be drawn without taking this into consideration.

Our contributions are as follows. (1) We introduce diversity, a Python package implementing diversity metrics. (2) We host and release source code to a user interface to explore repetition and diversity. (3) We evaluate the convergent validity of existing lexical diversity metrics and highlight compression ratios as efficient measures of diversity.

2 Related Work

Lack of diversity in text may result from repetition of lengthy strings or owe to subtle distributional patterns (Holtzman et al., 2019; Meister et al., 2022, 2023a). We focus on scores that aim to capture overt repetition across outputs, and leave for future work similar analysis of semantic and structural diversity scores (Bär et al., 2012; Shaib et al.,
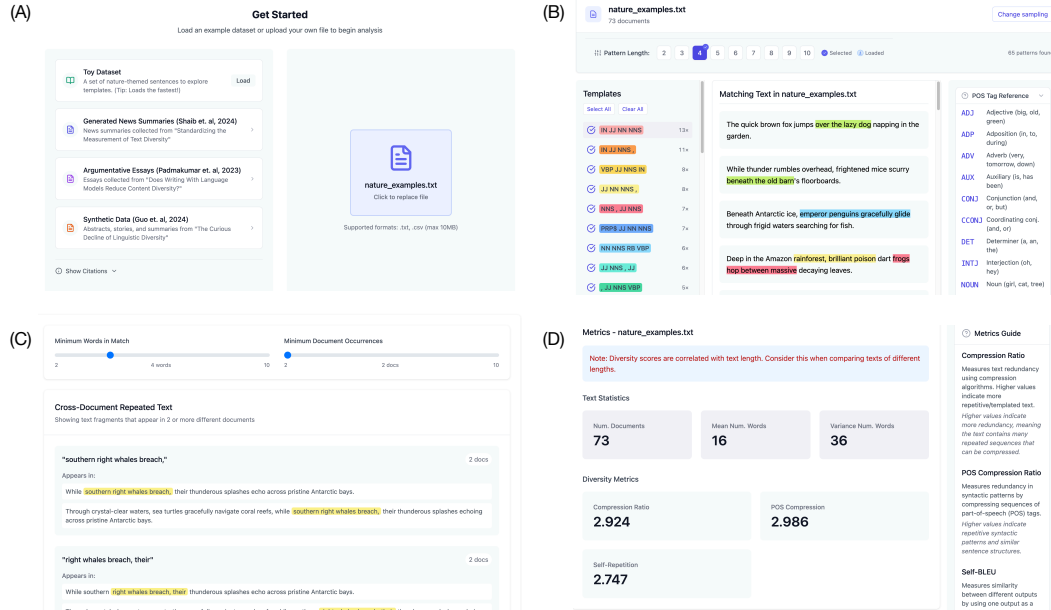

Figure 1: **(A)** Users start by uploading their own dataset on the right or by selecting one of the existing demo datasets on the left. Once uploaded, users can **(B)** interactively visualize part-of-speech patterns in the data, **(C)** interactively search for exact repeated text matches, and **(D)** calculate lexical diversity metrics.

2024). Conditional generation tasks such as image captioning have offered observations regarding the diversity of produced texts. Prior work has shown that models tend to repeat the same text for different contexts in these tasks (Li et al., 2016; Devlin et al., 2015). Self-repetition (Salkar et al., 2022)—exact repetition of the same n-gram ($n \geq 4$) across outputs—is a practical way of measuring repetition in lengthy outputs. In such cases repetition is common, especially relative to training data (Wang et al., 2023a).

We discuss several metrics but it is unclear which of these to use when, and how to efficiently visualize lower diversity in text-only tasks. Further, prior work has shown that human judgments of diversity are difficult to reliably collect. Humans tend to implicitly conflate quality of text with its diversity, and it can be difficult to separate content and lexical diversity in such assessments (Tevet and Berant, 2021). We design an interactive tool to allow users to browse highlighted instances of "lower diversity" text (Figure 1 (B)).

2.1 A Smorgasbord of Text Diversity Scores

Scores used to measure diversity across a corpus of texts derive from two core ideas: Computing average similarity between pairs of outputs produced by the same model for different inputs, and computing variants of token/type ratio. The former are adapted from common approaches to reference-based text generation using standard measures of pairwise similarity; the latter track the diversity of vocabulary measured as the ratio of unique words to total words produced, with outputs from a model concatenated into a single text. We first describe each score, and then present insights regarding their mutual redundancy. All scores are defined for a set of generated texts D, each conditioned on its respective input.

Self-BLEU The quality of text in machine translation, summarization, and image captioning is often reported in terms of overlap with a reference text. This idea can be adapted to measure diversity across different outputs by using one generated text as a "reference" and measuring the similarity of other outputs against this. Self-BLEU measures similarity between all text pairs in D using BLEU (Zhu et al., 2018). BLEU could be replaced with other similarity scores, e.g., ROUGE-L or BERTScore. These variants are called **homogenization scores** and have recently been used to compare the diversity of texts produced under several conditions (Padmakumar and He, 2023).

Homogenization Score (ROUGE-L) All homogenization scores calculate an aggregate similarity across pairs of examples (Equation 1). Here the similarity score of choice is ROUGE-L (Lin and Och, 2004), which quantifies overlap in terms of longest common sub-sequences between all pairs of text in a corpus instead of the fixed n-gram size

used in other ROUGE variants:

$$\text{hom}(D) = \frac{1}{|D| - 1} \sum_{d,d' \in D;\ d \neq d'} \text{sim}(d, d') \quad (1)$$

Homogenization Score (BERTScore) This homogenization score uses BERTScore to measure similarity between documents in Equation 1. Unlike the other scores, it does not count the repetition of specific tokens, but instead uses BERT embeddings to (ideally) capture "semantic" similarity beyond verbatim n-gram matches.

Self-repetition Score Self-repetition measures the tendency of LMs to repeat long n-grams across different outputs (Salkar et al., 2022).

$$\text{SRS}(d) = \log \left(\sum_{i=1}^{k} N_i + 1 \right) \quad (2)$$

Where k is the total number of 4-grams in a single document d and N_i the number of other summaries in which 4-gram i appears. The final score is the sum of $\text{SRS}(d)$ divided by $|D|$.

Moving Average Token-Type Ratio The token-type ratio for a text is the unique token count divided by the total token count. This metric captures the repetition of a given word in segments of text and does not explicitly account for longer repeated sequences (Covington and McFall, 2010).

N**-Gram Diversity Score** NGD extends the idea of token-type ratio to longer n-grams (Padmakumar and He, 2023; Meister et al., 2023b; Li et al., 2023), taking a ratio of unique to all n-gram counts:

$$\text{NGD}(D) = \sum_{n=1}^{4} \frac{\#\ \text{unique } n\text{-grams in } D\oplus}{\#\ n\text{-grams in } D\oplus} \quad (3)$$

Where $D\oplus$ denotes the dataset D concatenated into a single string. We use four as the maximum n-gram length. This method captures repeated *sequences* in addition to single token diversity.

2.2 Compression Ratios for Diversity

Compression Ratios (CRs) The diversity scores introduced so far are all a function of the number of repeated substrings across outputs. We use gZip to compress the concatenated text of all outputs generated by a model. CR is then the ratio between the size of the compressed file to that of the original. High CRs imply more redundancy:

$$\text{CR}(D) = \frac{\text{size of } D\oplus}{\text{compressed size of } D\oplus} \quad (4)$$

Part-of-Speech Compression Ratio To capture repeated syntactic patterns, we also compute compression ratios for part-of-speech (POS) tag sequences. We use the NLTK POS tagger [5] and the Penn Treebank set of 36 tags.

3 Evaluating Repetition with `diversity`

3.1 Design of the Diversity Package

The `diversity` package incorporates measures of diversity including lexical, syntactic, and semantic diversity. For lexical/syntactic diversity, we use NLTK (Bird and Loper, 2004) and SpaCY (Honnibal et al., 2020) to tag text with parts of speech and extract n-grams. We also include implementations of **embedding-based** measures (Cox et al., 2021), and **QUDSim** (Namuduri et al., 2025). Users can install the package via pip (assuming Python 3.10+) and can calculate various diversity metrics over a list of texts as follows:

```python
from diversity import *

text = ["I enjoy walking with my cute
    dog...", "I enjoy walking outside
    with...", "I enjoy jogging on a
    sunny..."]
# compression ratios
cr = compression_ratio(text, 'gzip')
cr_pos = compression_ratio(get_pos(
    text)[1], 'gzip')
# homogenization scores
hs_rougel = homogenization_score(text,
    'rougel')
hs_bert = homogenization_score(text,
    'bertscore')
self_bleu = homogenization_score(text,
    'bleu')
# other
self_rep = self_repetition_score(text)
nds = ngram_diversity_score(text, n=4)

# Embedding-based
rc = remote_clique(text, model="Qwen/
    Qwen3-Embedding-0.6B", verbose=
    False)

cd = chamfer_dist(text, model="Qwen/
    Qwen3-Embedding-0.6B", verbose=
    False)

# QUDSim
key = os.environ.get("OPENAI_API_KEY")
    # requires an OpenAI key
qud_alignment = qudsim(text, key=key)
    # list of QUD-based alignments/
    scores
```

---

[5] https://www.nltk.org/api/nltk.tag.html

Users can extract PoS patterns using the `extract_patterns` function by specifying the n-gram length to search for, and the top_n most repeated n-grams to return:

```
extract_patterns(text, n=5, top_n=100)
```

This returns a Python dictionary where the keys are the part-of-speech n-grams and the values are the raw text n-grams matching those patterns. The pattern matches are based on the frequency seen across the entire dataset, i.e., a part-of-speech pattern is only a pattern if it appears in more than 2 texts in the original input. Default values consider the top 100 part-of-speech patterns (sorted by frequency).

Then, using `match_patterns`, a user can identify all patterns in a single text from the input:

```
idx = 2
match_patterns(text[idx], patterns)
```

which returns a list of tuples containing the pattern and matched substring, respectively. Many diversity metrics require pairwise comparisons. With larger datasets, this can become infeasible to compute (see Appendix A). We implement a few methods to increase efficiency: memoization of already computed pairs, and batch pattern searching in the UI.

We also include a function for users to run all metrics and display them in a table to easily compare values:

```
compute_all_metrics(corpus=text)
```

3.2 Metric Visualization and the Web UI

The `diversity` Web UI offers the same functionality as the package via a no-code UI. Figure 1 shows the main pages of the site: users can begin by (A) either uploading their own text file for analysis or selecting one of the demo datasets provided on the site. Then, the user is prompted to select one of three types of analyses: either (B) to explore part-of-speech patterns, (C) to explore verbatim repeated text, or (D) to measure various diversity metrics of the dataset. Datasets are processed upon upload, and nothing is stored on the backend server aside from the existing demo datasets.

(B) Templates The templates tab allows users to explore extracted part-of-speech n-grams in their selected dataset. The left-most column displays pattern length of $n = [2, 10]$. The user can then scroll through all of the templates, select some or all, and see the highlighted text in the middle panel

corresponding to the template. The templates are assigned a colour when selected to indicate the corresponding matched text. The right-most column provides a reference for all the part-of-speech tags from SpaCY. [6] The default pattern length is set to $n = 4$. Other lengths will load when selected.

(C) Exact Match The exact matches tab allows a user to explore exact text matches in their dataset. The top provides two sliders: the left slider allows the user to set a string length to search for ($n = [2, 10]$), and the right the minimum number of documents in which the string must appear ($n = [2, 10]$). The minimum document occurrence slider defaults to 2. Once selected, the user can scroll through to see the repeated text in bold, and the full document text in which the string appears, as well as the number of documents.

(D) Diversity Metrics The diversity metrics tabs reports the recommended metrics from our evaluation: Compression Ratio, POS Compression, Self-BLEU, Self-Repetition, and Homogenization with BERTScore. We display these values alongside a guide to the metrics on the right-hand side.

3.3 Use-Cases

Our implementation of compression ratios over PoS tags and tokens (along with BERTScore, Self-BLEU, and self-repetition) have already been used in prior works (by other groups) to evaluate diversity in model evaluation and alignment (Lake et al., 2024; Moon et al., 2024; Fernandez et al., 2024), and for reporting diversity over synthetic datasets (Chang et al., 2024; Hastings et al., 2024). Due to its computational efficiency, compression ratios have also been used as optimization parameters in decoding strategies (Lanchantin et al., 2025).

Shaib et al. (2024) use the pattern analysis in `diversity` to measure and evaluate the prevalence of syntactic patterns in LLMs. Wadhwa et al. (2025) extract PoS patterns in distillation tasks for model attribution. Further, insights from our evaluation of diversity metrics have informed how to report diversity with respect to text length and data sizes (Guo et al., 2023; Hastings et al., 2024).

---

[6] https://spacy.io/usage/linguistic-features

Figure 2: Correlations between diversity scores on CNN/DM. CR correlates strongly with most other metrics.

4 Platform Evaluation: Comparative Analysis of Diversity Metrics

4.1 Data and Models

We compute diversity scores for the outputs of six instruction tuned models on the CNN/Daily-Mail (Hermann et al., 2015) and XSUM (Narayan et al., 2018) English news summarization datasets: Llama-2 (Touvron et al., 2023a), GPT-4 (OpenAI, 2023), FlanT5-XXL (Longpre et al., 2023), StableLM (Taori et al., 2023; Chiang et al., 2023; Anand et al., 2023), Mistral (Jiang et al., 2023), and StableBeluga (Touvron et al., 2023b; Mukherjee et al., 2023).[7] We selected these models to cover a range of availability (open and closed), and architectures (encoder-decoder, decoder-only). The lengths of texts vary considerably by source, for reference and model-produced text alike, so we also note average lengths when reporting diversity.

5 Text Length as a Confounder

To keep compute time and costs manageable, we randomly sample 500 inputs from CNN/DailyMail and XSUM for analysis. Table 1 reports diversity scores for outputs generated by the six LLMs for these inputs. Table 1 (top) reports scores for human-written texts: The article given as input for summarization, the baseline summary comprising the first three sentences of the news article, and the reference summary. These scores serve as a reference point for the diversity scores of the models.

One would expect that human-authored texts would be more diverse than those produced by LLMs (with the caveat that the texts were scraped

from the web, and so may contain HTML and page layout artifacts which might be repetitive (Salkar et al., 2022)). The human texts differ by length and the sources of longer texts appear to be less diverse.

Text length as a confounder for diversity has been reported in prior work (Salkar et al., 2022), along with methods to account for this, e.g., sampling blocks of fixed size (Covington and McFall, 2010).

All scores of the token/type ratio family are highly correlated with length, while the pairwise similarity ones are only moderately correlated. Self-BLEU has low correlation with length.

6 Diversity of Model Summaries

The confound of length complicates reporting. On CNN/DM (cf. Table 1) StableLM produces the longest summaries. All scores indicate that these are the least diverse, probably due to length. Three types of differences are marked in the tables. Model summaries that are shorter but less diverse than human summaries are marked in bold. Human texts here are written by journalists, so the expectation is that they would be more diverse. More bold entries in a column indicate that the score captures differences between human and machine diversity, a desirable trait. Underlined entries denote models that are less diverse than other models that produce longer summaries. The more underlined entries there are for a model, the more indicators there are that its output is less diverse. Asterisks mark models that appear more diverse than a human text of shorter length.

The most interesting diversity scores are those that capture differences between human and automatically produced text. On the CNN/DM dataset, Hom. (BERT) and MATTR are the two scores that detect no differences between human and model texts. Compression ratio for part of speech sequences is the score that identifies the most differences between human and model-generated text. Self-repetition stands out as the only score that identifies model generated text as more diverse on the CNN/DM dataset. From this analysis, CR:POS and self-repetition emerge as prime candidates of reportable scores, while Hom. BERT is less useful.

7 Correlation Analysis

We present three sets of correlation analyses between *(i)* different diversity scores, *(ii)* the same diversity score across datasets, and *(iii)* diversity

---

Model	Avg. Length	CR (↓)	CR: POS (↓)	NGD (↑)	Self-Rep. (↓)	Hom. (R-L) (↓)	Hom. (BERT) (↓)	Self-BLEU (↓)	MATTR (↑)	HD-D (↑)
Article	452.25	2.615	5.544	2.637	6.216	0.118	0.696	0.003	0.837	0.896
Article (Lead 3)	75.87	2.369	5.497	3.041	4.276	0.105	0.686	0	0.856	0.892
Reference	51.78	2.277	5.330	3.164	3.842	0.074	0.683	0	0.875	0.919
StableLM	132.71	**2.724**	**5.940**	2.673	4.940	**0.126**	0.689	0.002	0.792	0.867
Mistral	114.88	2.499	**5.621**	2.926	4.688	**0.123**	_0.697_	_0.036_	0.831	0.880
Llama-2	106.52	_2.543_	**5.684**	_2.874_	4.159*	**0.125**	_0.694_	0.001	_0.820_	_0.873_
StableBeluga	91.17	2.452	**5.644**	3.028	_4.467_	**0.121**	_0.702_	_0.047_	0.846	0.889
FlanT5	63.84	**2.453**	**5.608**	_2.939_	3.608*	0.084	0.667	0	_0.833_	**0.887**
GPT-4	55.4	2.361	**5.463**	**3.124**	_3.909_	_0.098_	_0.684_	**_0.001_**	0.853	**0.891**

Table 1: Diversity scores for the CNN/Daily Mail dataset. Arrows indicate direction of *more diversity*. Values indicating less diversity compared to at least one text source that produces longer human texts are bolded; models with scores that are less diverse than those from a model that produces longer summaries are underlined. An asterisk indicates a model more diverse than a shorter human text.

scores and standard reference-based evaluations. Despite the large number of diversity scores in our list, they all revolve around n-gram repetition. Do these capture different (complementary) information? To assess this we compute the correlations between all pairs of scores, reported in Figure 2.

Compression ratio is highly to moderately correlated with other n-gram scores. The only weak correlations are with Self-BLEU and Hom. (BERT). Given the degenerate behavior of Hom. (BERT) on the analysis of summaries, reporting Self-BLEU only is advisable. Finally, self-repetition is only moderately correlated with other scores, and is therefore informative to report. Correlations are similar on XSUM summaries (Appendix 5).

8 Truncating to Control for Length

We truncate all summaries to the length of the shortest one produced by any source as a crude means to control scores for length. The resulting scores are directly comparable (see Table 5 in the Appendix). CRs and Self-BLEU scores indicate that model-generated text is less diverse than human text. Hom. (BERT) scores barely vary across sources. On the CNN/DM dataset, Self-BLEU indicates that Llama-2 and StableLM are the most repetitive models. CR also ranks these two models as the least diverse. The results are consistent on XSUM, but for that dataset Flan-T5 is also highly ranked and the most repetitive.

Truncation to control for length is impractical for published research or leaderboards. Introducing a new source of texts would require recomputing scores for other sources for comparison, which is sometimes impossible (when outputs from other sources are not available). Future research might search for more practical alternatives.

9 Discussion and Recommendations

The `diversity` package and platform provides a useful way to analyze and visualize diversity in datasets. Our analyses of metrics reveal that compression ratio (CR) is an excellent score to report, easy to compute and strongly correlated with other scores used in past work. CR of PoS sequences captures differences between human and model-generated text. Self-repetition focusses on repetition of longer n-grams across generations, and is only moderately correlated with CRs. Finally Self-BLEU is only weakly correlated with the previous three, so is a good complement score to report. We found BERTScore limited: It does not show differences between human and model-generated text and barely varies when adjusted for length.

Length of the analyzed text has to be reported alongside all these scores. When length differs, scores are not meaningfully comparable. Truncating text is one way to control for this. Different random draws of the sample chosen to represent a dataset may differ in diversity, in turn leading to unwarranted conclusions. Truncating texts prevents quantifying repetition towards the end of longer texts. Finally, `diversity` offers a platform and package in which researchers from a variety of domains can use to facilitate evaluation and gather insights about diversity between human- and model-produced texts.

10 Limitations

In this work, we do not attempt to measure human judgments of diversity, which are straightforward for short texts (e.g., questions) but far more difficult for longer summaries or large instruction datasets (Tevet and Berant, 2021); we leave this for future work. All evaluations are conducted in English.

Acknowledgements

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING 2012*, pages 167–184, Mumbai, India. The COLING 2012 Organizing Committee.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Ernie Chang, Matteo Paltenghi, Yang Li, Pin-Jie Lin, Changsheng Zhao, Patrick Huber, Zechun Liu, Rastislav Rabatin, Yangyang Shi, and Vikas Chandra. 2024. Scaling parameter-constrained language models with quality data. *ArXiv*, abs/2410.03083.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17:100 – 94.

Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, page 1–35. ACM.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Nigel Fernandez, Alexander Scarlatos, Wanyong Feng, Simon Woodhead, and Andrew Lan. 2024. Divert: Distractor generation with variational errors represented as text for math multiple-choice questions. *arXiv preprint arXiv:2406.19356*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *CoRR*, abs/2311.09807.

John D. Hastings, Sherri Weitl-Harms, Joseph Doty, Zachary L. Myers, and Warren Thompson. 2024. Utilizing large language models to synthesize product desirability datasets. *2024 IEEE International Conference on Big Data (BigData)*, pages 5352–5360.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *ArXiv*, abs/1506.03340.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Thom Lake, Eunsol Choi, and Greg Durrett. 2024. From distributional to overton pluralism: Investigating large language model alignment. *ArXiv*, abs/2406.17692.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *ArXiv*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023a. Locally typical sampling. *Trans. Assoc. Comput. Linguistics*, 11:102–121.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023b. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Clara Meister, Gian Wiher, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *Trans. Assoc. Comput. Linguistics*, 10:997–1012.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

JiHwan Moon, Jihoon Park, Jungeun Kim, Jongseong Bae, Hyeongwoo Jeon, and Ha Young Kim. 2024. Diffslt: Enhancing diversity in sign language translation via diffusion model. *ArXiv*, abs/2411.17248.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *ArXiv*, abs/2309.05196.

Vishakh Padmakumar, Behnam Hedayatnia, Di Jin, Patrick Lange, Seokhwan Kim, Nanyun Peng, Yang Liu, and Dilek Hakkani-Tur. 2023. Investigating the representation of open domain dialogue context for transformer models. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 538–547, Prague, Czechia. Association for Computational Linguistics.

Nikita Salkar, Thomas Trikalinos, Byron C Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL)*, volume 2022, pages 341–350.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods*

in Natural Language Processing, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Somin Wadhwa, Chantal Shaib, Silvio Amir, and Byron C Wallace. 2025. Who taught you that? tracing teachers in model distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3307–3315, Vienna, Austria. Association for Computational Linguistics.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey Kuehl, Erin Bransom, and Byron Wallace. 2023a. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9871–9889, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: Measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60644–60673. PMLR.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.
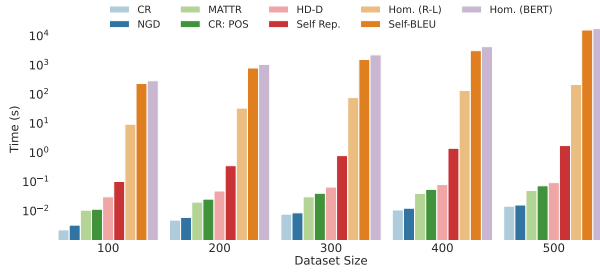
Figure 3: Mean run time **(log-scale)** on CNN/DM summaries. Run times increase with the number of text for the analysis. Even for small datasets, Self-BLEU and BERTScore homogenization are slow.
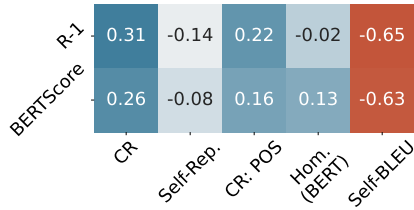

Figure 4: Correlations between diversity metrics, BERTScore, and ROUGE-1. Both reference-based metrics are weakly correlated with CR and Hom. (BERT), and moderately anti-correlated with Self-BLEU.

Appendix

A Run-Time Considerations

Figure 3 provides insights about the feasibility of obtaining scores for large samples.[8] The compression ratio scores are fast compared to other diversity measures.

B Correlations with Evaluation Metrics

Output diversity and self-repetition are aspects of model behavior that are not captured by existing evaluation approaches. We compute the system level correlation between the diversity scores and the traditional BERTScore and ROUGE evaluations, shown in Figure 4.

C Controlling for Length

Scores on CNN/DM summaries truncated to the shortest summary reveal a different model order with respect to diversity (Table 2).

D Additional Evaluations

Story Writing Padmakumar et al. (2023) presented an analysis of human-written stories, where people wrote either by themselves or with the help of GPT-3 or GPT-3.5 Turbo. We also find that all

---

[8] Run on a single NVIDIA Quadro RTX 8000 GPU.

Model	CR (↓)	CR: POS (↓)	Self-Rep. (↓)	Hom. (BERT) (↓)	Self-BLEU (↓)
Article	2.268	5.25	2.763	0.676	0
Article (Lead 3)	2.274	5.25	2.762	0.658	0
Reference	2.189	5.179	2.763	0.674	0
Llama-2	2.96	5.627	2.847	0.674	0.001
GPT-4	**2.287**	5.376	2.761	0.672	**0**
FlanT5	2.288	5.389	2.779	0.673	**0**
StableLM	2.393	5.537	2.884	0.672	0.001
Mistral	2.32	5.415	2.812	**0.67**	**0**
StableBeluga	2.288	5.46	2.766	0.671	**0**

Table 2: Scores on CNN/DM summaries truncated to the shortest summary length for a given input.

Dataset	CR (↓)	CR: POS (↓)	Self-Rep. (↓)
Open Assistant	2.886	6.731	3.969
Unnatural Instructions	4.191	7.278	9.868
Alpaca	3.119	6.61	3.105
Super-NaturalInstructions	2.675	5.749	3.456
Dolly	2.578	6.214	2.935

Table 3: Diversity scores for instruction datasets. We do not include Self-BLEU nor Hom. (BERT) due to long run times.

diversity scores agree that people writing independently produce the more diverse texts (cf. Table 5). Length is not an issue because the average length of stories in each setting are comparable: 375 words for writing without help, 372 words when writing with GPT-3 and 370 when writing with GPT-3.5.

Instruction-tuning Datasets The quality and diversity of instructions are likely to result in more robust and capable systems (Sanh et al., 2022; Mishra et al., 2022). We analyze the diversity of five instruction-tuning datasets: Open Assistant (Köpf et al., 2024), Super-NaturalInstructions (Wang et al., 2022), Unnatural Instructions (Honovich et al., 2023), Alpaca (Wang et al., 2023b),

Figure 5: Correlation table between scores on XSUM.

Model	Avg. Length	CR (↓)	CR: POS (↓)	NGD (↑)	Self-Rep. (↓)	Hom. (R-L) (↓)	Hom. (BERT) (↓)	Self-BLEU (↓)	MATTR (↑)	HD-D (↑)
Article	310.20	2.511	5.555	2.756	5.643	0.110	0.695	0.002	0.838	0.892
Article (Lead-3)	55.94	2.316	5.454	3.107	3.999	0.103	0.683	0	0.860	0.891
Reference	21.04	2.276	5.409	3.211	2.914	0.081	0.673	0	0.877	0.888
StableLM	109.20	2.745	6.008	2.636	4.687	0.130	0.695	0.002	0.78	0.854
Llama-2	102.48	2.634	5.802	2.795	4.618	0.128	0.687	0.002	0.795	0.858
Mistral	95.18	2.531	5.708	2.911	4.495	*0.132*	*0.698*	*0.044*	0.819	0.867
StableBeluga	88.46	2.461	5.673	2.992	4.418	0.124	*0.698*	*0.046*	0.837	0.88
GPT-4	62.15	2.394	5.531*	3.079	4.041	0.104	0.682	0	0.848	0.886
FlanT5	20.93	*2.666*	*6.222*	*2.743*	*2.868*	*0.114*	0.665	0.001	*0.756*	**0.842**

Table 4: Diversity scores for XSUM summaries. Arrow indicate the direction of more diverse texts for each score.

Dataset	CR (↓)	CR: POS (↓)	Self-Rep. (↓)	Hom. (BERT) (↓)	Self-BLEU (↓)
Solo	2.901	5.314	5.873	0.604	0.018
GPT-3	2.940	5.371	5.911	0.613	0.020
InstructGPT	3.064	5.462	5.966	0.631	0.022

Table 5: Diversity scores over essays. Working with an LLM correlates with lower diversity.

Model	CR (↓)	CR: POS (↓)	Self-Rep. (↓)	Hom. (BERT) (↓)	Self-BLEU (↓)
Article	2.162	5.095	2.719	0.666	0
Article (Lead 3)	2.179	5.093	2.719	0.663	0
Reference	2.230	5.314	2.663	0.667	0
Llama-2	2.345	5.636	2.919	0.663	0.002
GPT-4	2.213	5.425	2.666	0.663	0
FlanT5	2.490	5.737	2.707	0.665	0.001
StableLM	2.342	5.521	2.823	0.664	0.001
Mistral	2.308	5.689	2.736	0.659	0
StableBeluga	2.210	5.436	2.663	0.659	0

Table 6: Diversity metrics for XSUM summaries, with outputs from each model truncated to the length of the shortest. All scores are directly comparable.

Dataset	CR (↓)	CR: POS (↓)	Self-Rep. (↓)
Open Assistant	2.370	5.402	1.741
Unnatural Instructions	6.036	8.421	5.595
Alpaca	3.301	6.044	2.020
Super-NaturalInstructions	2.458	1.844	4.859
Dolly	2.832	5.504	2.235

Table 7: Truncated diversity scores for instruction datasets.

affected by text source. The other scores are still moderately to highly correlated but as already observed, models are ranked differently. When reporting diversity, source of analyzed data also has to be taken into account, in addition to length.

D.2 Instruction Datasets, Length Controlled

Table 7 shows scores for instructions downsampled to the size of the smallest dataset, and truncated to the length of the shortest instructions in the remaining data. Again, the Open Assistant dataset stand out as most diverse, while the Unnatural Instructions dataset is markedly less diverse than the others. Self-repetition in the related Super-Natural and Unnatural instructions is notably high. The human instructions in Dolly compare favorably with automatic instructions, especially when bearing in mind that only eight tasks are covered in it. CR:POS points to Super-natural instructions as the most diverse. We do not have a convincing explanation of why it compares so favorably against others on this score.

D.3 XSUM Metrics

Tables 4, 6 show the full diversity metrics over XSUM with and without controlling for length.

Figure 5 shows the correlations between all pairs of metrics for the XSUM dataset. The correlations show that compression ratio is highly to moderately correlated with other n-gram scores, similar to the findings for the CNN/DM dataset.

and Dolly (Conover et al., 2023) (Table 3).

Open Assistant instructions are remarkably diverse compared to the other datasets across all diversity scores. Unnatural instructions are remarkable in the opposite direction. We provide an analysis of the diversity scores with the length controlled in Appendix D.2.

Given the large dataset sizes, ranging from 15-80k data points, we do not compute the homogenization scores nor Self-BLEU, as the computation time is infeasible. For approximately 50k instructions, the estimated computation times ranged from 48 to 800 hours for these scores. This case study highlights the relevancy of the run-time analysis for computing score that we presented in the previous section.

D.1 Correlation Between Metrics

Self-BLEU scores are almost perfectly correlated between the two datasets; they appear to not be