Supplementary Material

1 Broader Impacts

FACE measures the distance between human and model-generated languages, therefore it is technically possible to be used for designing or augmenting systems that mimic humans. We acknowledge the risks of FACE (and other metrics) being utilized in applications that deliberately confuse human-authored and model-produced text. We call for the collective efforts from the community to come up with a systematic framework that unifies different metrics, for developing more reliable and natural language generation systems.

2 Implementation Details

Preprocessing. We utilize three raw datasets: WritingPrompts, WikiText-103, and RealNews. For WritingPrompts, the prompt set has already been well-curated, so we just extracted the first 5,000 prompts (the length may vary) for our generation task. WikiText-103 and RealNews contain many complete texts. For each complete text, we further truncate it corresponding to the first 35 tokens as a prompt. To fairly evaluate the performance of metrics, we also divide text generations according to five predefined length (from 0 up to 1024) intervals for each dataset. Thereby, the human-written texts and model-produced texts used to evaluate the performance of metrics may be generated by different prompts (i.e., unpaired comparison).

Hyper-parameters. We have several hyper-parameters during the text generation and evaluation phases. For both conditional and unconditional generation, we preset a random seed integer (32 by default). Furthermore, the maximum length of each text (1024 by default) as well as the batch size (which varies according to GPUs capacity) for perplexity computation have to be determined before automatic evaluation.

3 Miscellaneous Details

Software. Our experiments were performed on Ubuntu 20.04.1 system with Python 3.9.16. The versions of key Python libraries include: Transformers 4.27.4, PyTorch-CUDA 11.6, PyTorch 1.13.1, Scipy 1.5.4.

Hardware. For the text generation task, we use the remote workstation that has two NVIDIA RTX A6000 graphics cards. It should be noted that all models were run in parallel when available.

Computation time for text generation. We spent 10 and 25 hours or so obtaining 5,000 text continuations by GPT2-sm, -xl, respectively. OPT-125m, -6.7b cost our GPU resources roughly 11 and 44 hours to output the same number of text continuations, respectively. When it comes to BLOOM-560m, -7b, they took approximately 18 and 48 hours, respectively, to generate 5,000 continuations per task domain.

Evaluation time for FACE. Computation time of four FACE metrics for a single pair of references are: 5.96×10^{-8} seconds for *SO*, 5.01×10^{-8} seconds for *CORR*, 4.53×10^{-8} seconds for *SAM*, and 4.29×10^{-8} seconds for *SPEAR*, respectively. The cross-entropy, which should be calculated beforehand, takes 5.65×10^{-2} seconds. All of the above measurements take place on an AMD Ryzen Threadripper PRO 3995WX 64-Cores CPU (frequency range $\in$ [2200.00MHz, 4308.40MHz]). Users can leverage more advanced GPU resources to perform the whole computation process with a faster speed.

4 Additional Experimental Results

4.1 Model sizes (generation length)

It should be emphasized that LMs have diverse designs and were pre-trained using different strategies on different datasets, giving them distinct preferences on the generation length. The numbers of text generations in each length interval are summarized in Table 6.

Domain	Length Interval	GPT2-sm	GPT2-xl	OPT-125m	OPT-6.7b	BLOOM-560m	BLOOM-7b
Wiki text	0-200	403	485	964	1522	4928	803
	201-400	571	672	888	929	61	599
	401-600	251	316	441	417	8	388
	601-800	260	310	268	285	1	316
	801-1024	3515	3217	2439	1847	2	2894
News	0-200	750	836	844	1119	4978	1371
	201-400	1222	1336	1220	1325	20	917
	401-600	824	759	1194	939	1	628
	601-800	584	678	764	593	0	427
	801-1024	1620	1391	978	1024	1	1657
Stories	0-200	549	745	2731	3588	4924	1608
	201-400	625	757	715	501	63	688
	401-600	296	404	241	176	9	410
	601-800	241	324	160	95	4	271
	801-1024	3289	2770	1153	640	0	2023

Table 6: Domain-specific generation length with respect to different **models** (GPT2/OPT/BLOOM) and **model sizes** (one large model and one small model) using top-k ($k = 50$) sampling corresponding to five continuous length intervals.

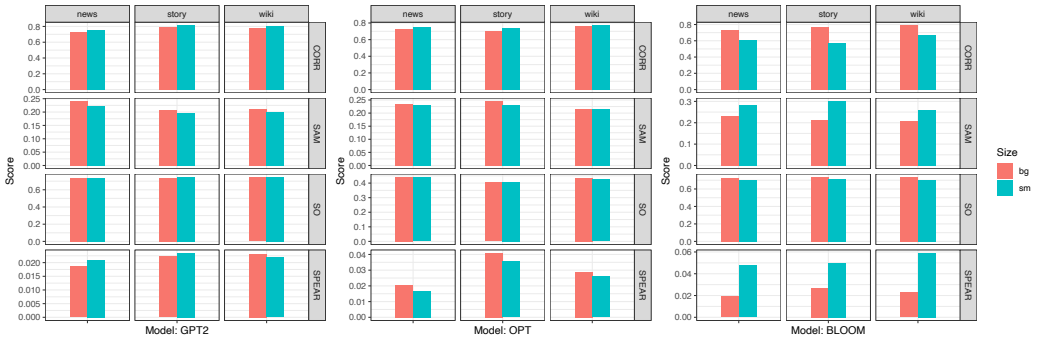

Figure 6: FACE scores of GPT2 (our generated data), OPT, and BLOOM with different model sizes.

To ensure the consistency of our experiments, we run six LMs separately (using their own tokenizers) with the same prompt sets and settings as described in Table 2 to generate 5,000 pieces of continuations in each domain. Besides, we utilize the GPT2Tokenizer to calculate the numbers of continuations for each interval, which allows us to compare FACE scores with other metrics more objectively, as we believe it is unfair to explicitly compare texts of varying lengths. Then, we compute weighted arithmetic mean to evaluate a model in each domain, by $s' = \sum_{i=1}^{n} \frac{m_i}{M} s_i$, where s' denotes the weighted mean; n denotes the number of length intervals; m_i is the number of generated continuations in the length interval i; $M = \sum_{i=1}^{n} m_i$, and s_i means a certain metric value in the interval i.

Figure 6 conveys a more intuitive representation (via bar plots) of Table 3.

4.2 Sampling methods (unconditional generation)

We also carried out experiments on unconditional text generation. Here, the prompt is not required as we generate continuations from a random seed (set to 32 empirically). Four sampling methods, which are greedy decoding, beam search, stochastic beam search, and contrastive decoding, are not involved in this set of experiments.

The results are displayed in Figure 7. The overall trends are same as its conditional counterpart, where the previous quality relationship (maximization-based/temperature-based $\prec$ nucleus $\prec$ contrastive) is satisfied. Yet, it is crucial to note that the advantages of top-k sampling w/o temperature become more obvious compared to the conditional case.

[5] https://github.com/ari-holtzman/degen

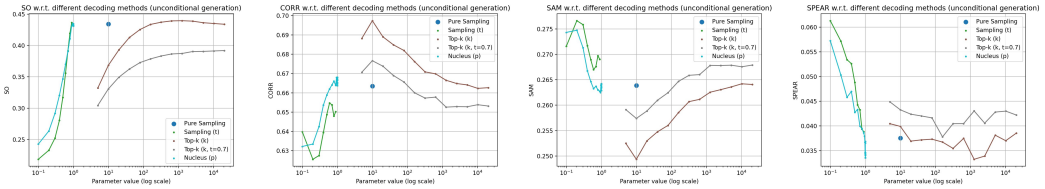

Figure 7: FACE scores (unconditional generation) on original experimental data[5] of *nucleus sampling*. Five sampling (decoding) methods are compared: pure sampling, temperature, top-k, top-k with temperature, and nucleus. Note that logarithmic normalization on parameter values as well as an enlarged marker for pure sampling are adopted for better visualization.

Model	**Sampling Method (parameter)**	*SO*	*CORR*	*SAM*	*SPEAR*
GPT2-xl	Nucleus Sampling (p=0.95)	0.481	0.821	0.191	0.359
	Ancestral Sampling	0.472	0.807	0.199	0.331
GPT2-lg	Nucleus Sampling (p=0.95)	0.480	0.819	0.193	0.356
	Ancestral Sampling	0.472	0.814	0.196	0.338
GPT2-md	Nucleus Sampling (p=0.9)	0.478	0.815	0.194	0.358
	Ancestral Sampling	0.462	0.813	0.197	0.310
GPT2-sm	Nucleus Sampling (p=0.9)	0.476	0.817	0.194	0.359
	Ancestral Sampling	0.468	0.816	0.195	0.319

Table 7: FACE results based on MAUVE's original experimental data[6].

4.3 Human judgments

Table 7 shows the FACE scores based on the output texts from MAUVE. Each column of FACE scores is used to compute the Spearman's rank correlation coefficient between a specific FACE metric and Bradley-Terry scores (4 model sizes $\times$ 2 sampling methods $=$ 8 scores in total) from one criterion (three criteria correspond to three questions in total).

4.4 Choice of estimator model

We examine how different choices of estimator model m_{est} affect the resulting spectra of cross-entropy. Five input data sources are examined (webtext plus four GPT2 original output datasets), on which four different estimator models are applied: $m_{\text{est}} \in \{\text{GPT2-sm}, \text{GPT2-md}, \text{GPT2-lg}, \text{GPT2-xl}\}$, resulting in $5 \times 4 = 20$ aggregated spectra curves in Figure 8. It can be found that on the same input data, the spectra from four estimators largely overlap. It indirectly suggests that FACE should be stable across different m_{est}s. We leave the full inspection for future work.

4.5 Intuitive interpretation of spectra

As pointed out in Section 4.5, the aggregated spectral shapes from human and different models are nearly identical. A set of higher resolution plots from GPT-xl, OPT, BLOOM and human (webtext) are shown in Figure 9. It can be seen that although the $X(\omega_k)$ has different ranges on y-axis, the x coordinates of the peaks and troughs are the same.

[6] https://github.com/krishnap25/mauve-experiments

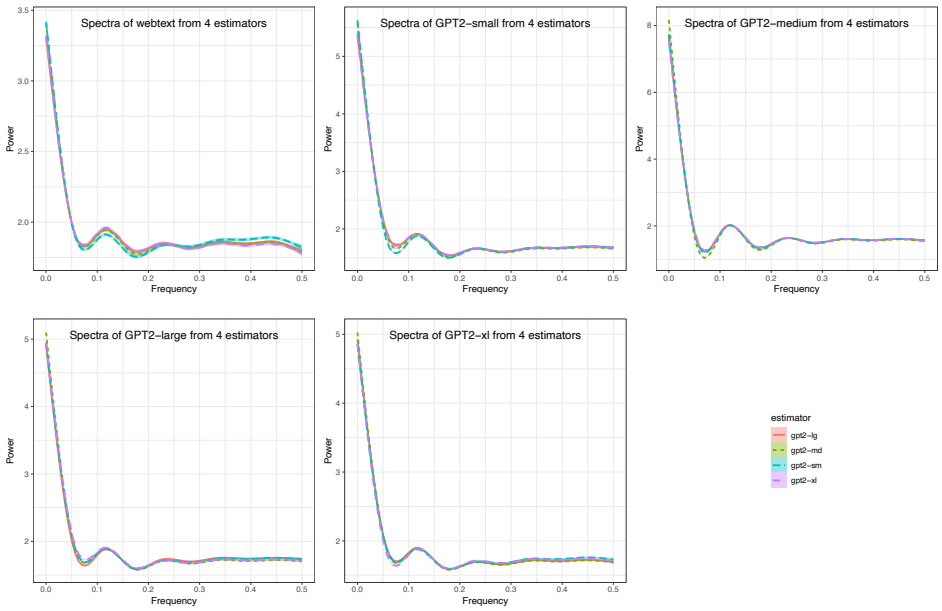

Figure 8: Aggregated spectra (using GAM smoothing) from four estimator models $m_{\text{est}} \in \{\text{GPT2-sm}, \text{GPT2-md}, \text{GPT2-lg}, \text{GPT2-xl}\}$. Inputs are from GPT2 original output and webtext.

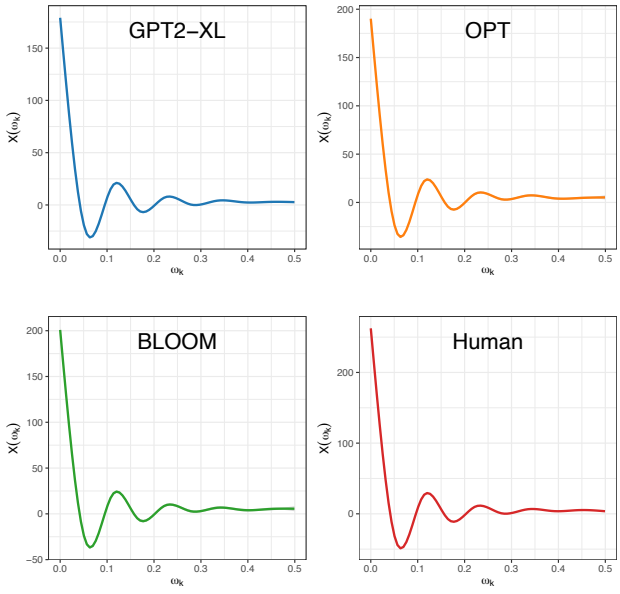

Figure 9: Aggregated spectra for GPT-xl, OPT, BLOOM, and human (webtext).