Evaluating Text Generation Quality Using Spectral Distances of Surprisal

Anonymous ACL submission

Abstract

Surprisal-based metrics are commonly used for evaluating the quality of natural language generation, especially the open-ended generation tasks. This paper proposes an improved version for a recently developed metric, Fourier analysis of cross-entropy (FACE), hence, FACE-2, by introducing non-trivial processing steps and incorporating more comprehensive distance functions. FACE-2 and its predecessor are inspired by psycholinguistics findings about the periodicity in human language production, whose idea is to convert the surprisal of texts to the frequency domain and then measure the distances between the spectral representations. Our experiments show that FACE-2 produces the state-of-the-art performance in reflecting model scaling effect and in agreement with human preferences, which beats its predecessor and a broad range of other metrics (such as MAUVE, Zipf, and Self-BLEU).

1 Introduction

The booming of large language models (LLMs) has urgently increased the need for metrics that can evaluate the quality of model generated texts in effective, consistent, and trust-worthy ways (Novikova et al., 2017). The traditional evaluation metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020), which are designed for evaluating the generation ability of comparatively smaller language models, are not sufficient and comprehensive enough for evaluating LLM's generation ability. Apart from measuring LLMs' performance on tasks with clearly defined objectives, such as question-answering and reasoning, their capabilities on generation tasks, however, are more difficult to characterize due to the subjectivity of telling *good* texts from *bad* ones, the limited resources from human judgments, and many other factors.

A natural way of approaching the problem is to compare the artificially generated texts with human written ones along some dimensions. For example, MAUVE (Pillutla et al., 2021) focuses on measuring the divergence in estimated overall semantic distributions of text. Zipf's coefficient method (Holtzman et al., 2020) uses the frequency distribution of tokens as the key distinguishing factor. It is impractical and unnecessary to expect the dimension used by one metric to be a golden standard that fully captures the rich and complex aspects of human languages. But it is still important to continuously assess the performance of those metrics, because LLMs' capabilities keep evolving and its distance from human writers at no matter what dimensions are expected to keep shrinking. We believe that seeking metrics to magnify this ever-shrinking gap is not just a technical solution to AI-raised issues, but also an answer to the eternal question of where the boundary between artificial and natural creations is.

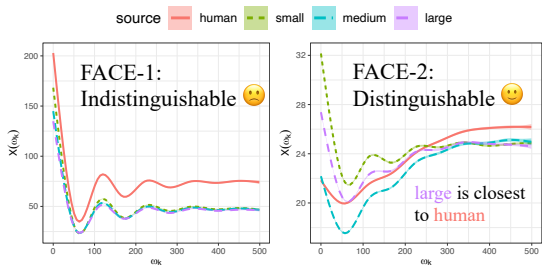


Figure 1: An example showing improvement of FACE-2 (this study) over FACE-1 in better distinguishing texts generated by models of different sizes. "small"⇒1.5b, "medium"⇒7b, "large"⇒72b, all from Qwen2 family. Curves are spectra of surprisal fit with GAM.

The realistic goal of this study is to propose improvements to a recently proposed evaluation method, Fourier analysis of cross-entropy (FACE) (Yang et al., 2023). FACE characterizes how the surprisal of text changes periodically and uses the spectral features to distinguish texts from different sources. FACE is inspired by pycholinguistics stud-

ies about the periodicity in human cognitive load during language processing (Dethlefs et al., 2016; Xu and Reitter, 2017). FACE, as a metric, shows good potentials for revealing some of the essential differences in human and artificial languages, but it is still a method in prototyping stage that has only been tested limited range of data and models (GPT-2 and two 7b-level models). Its full potential is yet to be explored in broader settings.

In this study, we propose an augmented version of the original FACE metric, namely FACE-2, by including non-trivial processing steps and new distance functions to the workflow, and with broader evaluation experiments on both English and Chinese tasks. We present the following content in this paper: the basic idea of FACE approach is described in Section 3. The main improvement is introduced in Section 4. Experiment results are shown in Section 5. Limits of the current study and future directions are discussed in Section 6. For convenience, we refer to the original FACE work and the corresponding metrics as FACE-1, and ours as FACE-2, throughout the study.

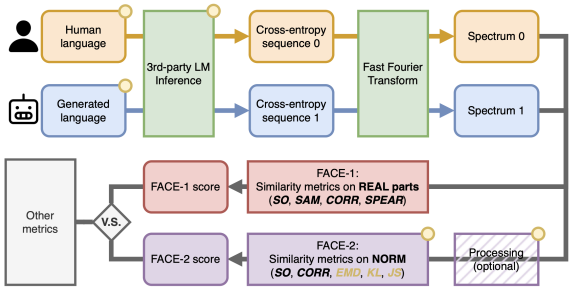


Figure 2: The basic workflow of the FACE approach. Yellow pins on the top-right corner notice that this module is added or improved in this study.

2 Related Work

2.1 Surprisal-related evaluation for NLG

Word (token) level surprisal has been widely used for evaluating the quality of generated text. Gehrmann et al. (2019) used heatmap-like method to directly visualize the difference in token surprisal between GPT2-generated and human written text, where high-surprisal tokens are visualized with warmer colors, i.e., the red end on spectrum, while low-surprisal tokens with colder colors. Another closely related metric is the Zipf coefficient of word frequency distributions. Zipf's law in natural language (Zipf, 1936, 1949), plainly speaking, describes the fact that the frequency of a token is proportional to its rank in the vocabulary (Piantadosi, 2014). Holtzman et al. (2020) found that the Zipf curves can effectively distinguish human and generated texts, which has similar visual outcomes as our method, but is of much lower granularity. Another similar statistical tendency is called Heaps' law (Cohen, 1962), which states that human language is structured such that with the increasing of the context, the distinct token number is generally increasing while the new token discovery of the full vocabulary is diminishing. Meister and Cotterell (2021) demonstrates that while model-generated text shows less adherence to Heaps' law, it remains comparable to human text in terms of sentence length, stopword distribution, unigram distribution, and adherence to Zipf's law. The metric compared with in this study is MAUVE (Pillutla et al., 2021), which is based on quantized semantic representations of text, but not directly related to surprisal. Thus, the comparison with FACE-2 is free of confounding effect.

2.2 Surprisal studies in psycholinguistics

The original FACE study (Yang et al., 2023) uses the term "cross-entropy" to refer to the estimated surprisal of tokens, for the reason that it is computed using the cross-entropy loss function. In order to be more precise about the meanings of the quantity, we switch to the term "surprisal". Surprisal is a well studied concept in psycholinguistics, which has been known to reflect the cognitive processes underlying human language usage, with evidence a wealth of corpus-based and behavioral studies (Jaeger and Levy, 2006; Smith and Levy, 2013). The most relevant previous empirical findings for FACE and this study are about temporal property of surprisal, i.e., how surprisal changes over time. The earlest work to the best of our knowledge, dates back to Dethlefs et al. (2016), who find that human users to dialogue systems are sensitive to the *peaks* and *troughs* of entropy in speech. Xu and Reitter (2016) take a closer look at the sub-structure of spontaneous dialogues, and find that the utterance surprisal from two speakers converge towards each other within topical segments. Further, Xu and Reitter (2017) hypothesize that the observed convergence of surprisal can be attributed to the innate *periodicity* of language processing capacity limited by human cognitive load during communication, and back it up with evidence that the spectral features of surprisal are useful predictors for success in task-oriented dialogues. Similar investigations

are carried out on free conversations (Maës et al., 2022), task-oriented dialogues in both written and spoken modalities (Giulianelli et al., 2021), and larger datasets (Giulianelli and Fernández, 2021).

Entering the era of large language models, the surprisal view of language naturally started gaining more attention from researchers who are interested in developing more effective and cogtive-inspired evaluation tools for natural language generation. However, we notice that it is not until lately that works emerge to harness the temporal information from surprisal. FACE-1 (Yang et al., 2023) is the first attempt that adopts the spectral method in Xu and Reitter (2017) to the evaluation of open-ended text generation. Follow-up works started using the fluctuation of surprisal as an indicator of whether the text was generated by model or human writers, leading to several successful psycholinguistics-inspired text detection methods, such as GPT-who (Venkatraman et al., 2024) and FourierGPT (Xu et al., 2024). FourierGPT is most related to our study, as it builds text detection models directly on top of the workflow of FACE-1, and makes modifications that have been proven useful for the detection task.

3 Basic Workflow of FACE Approach

The workflow of FACE approach is shown in Figure 2, which can be summarized in three stages:

Stage 1 - Estimate surprisal Given a text set $\mathcal{D}$, a language model is used to compute the surprisal of each sentence in $\mathcal{D}$. For a sentence of N tokens, $[t_1, \ldots, t_N]$, its surprisal sequence $\mathcal{S} = [s_1, \ldots, s_{N-1}]$ is defined as $s_i \triangleq -\log P(t_i|t_1, \ldots, t_{i-1})$, i.e., the negative log-likelihood for each token estimated by the language model. There $|\mathcal{D}|$ pieces of such surprisal sequences produced at this stage.

Stage 2 - Fourier transform We treat $\mathcal{S}$ as a signal in time domain, and apply the discrete Fourier transform (defined in Cooley and Tukey, 1965):

$$X(\omega_k) \triangleq \sum_{n=0}^{N-1} s(n) e^{-j\omega_k n} \quad (1)$$

The obtained series of complex numbers $\mathcal{F} = \{X(\omega_k)\}$ is the spectral representation of the original surprisal signal in frequency domain, where ω_k is the k-th frequency component, $k = 0, \ldots, N-1$.

Stage 3 - Measure similarity of spectrum Given text data from two sources, $\mathcal{D}_h$ (human written) and $\mathcal{D}_m$ (model generated), and their spectra $\mathcal{F}_h$ and $\mathcal{F}_m$ obtained from previous stages, respectively. By our assumption, the quality of generated text can be reflected in its distance from human text in spectral space, that is, the quantity returned by some similarity/distance measures between $\mathcal{F}_m$ and $\mathcal{F}_h$. FACE-1 uses four similarity metrics: Spectral overlap (SO), Pearson's correlation (CORR), cosine distance, and Spearman's correlation, among which only the former two metrics show good overall performances according to FACE-1's report. In this study, we compare three new metrics in addition to SO and CORR. See details in next section.

Limits of FACE-1 results The effectiveness of the above summarized FACE approach has been validated in preliminary experiments, but only on text generated from GPT2 (Radford et al., 2019) and two other 7b models. The main conclusions from the original study is that, FACE-1 scores scale with model size and can reflect the choice of using better sampling methods (Yang et al., 2023). However, the model size effect is not stable, which by our guess is due to the model size limit and the underlying metrics. Furthermore, it is not clear whether and to what degrees can this approach be applied to larger models and other languages. Based on these limits, we propose some improvements to the approach in Section 4.

4 What's New in FACE-2

We make the following major modifications:

1. *Spectral representations*: Use L2-norms instead of the *real* parts of the Fourier transform outputs; use additional pre-processing on surprisal: z-score.
2. *Distance functions*: Add Earth Mover's Distance, KL-divergence and Jenson-Shannon divergence.
3. *Agreement with human*: Expand to a larger human-annotated dataset for exploring the metrics' agreement with human preferences.

4.1 New spectrum representation

FACE-1 uses the real part of the output from Fourier transform to represent spectrum, i.e., *real*$(X(\omega_k))$ in Equation (1), but this is an incomplete representation of the whole frequency information of a signal, because the imaginary

component also carries some energy and information, such as phase shift. Therefore in FACE-2, we use the Euclidean norm, $\|X(\omega_k)\| = \sqrt{real(X(\omega_k))^2 + imag(X(\omega_k))^2}$, to represent the spectrum, and compare its performance with the mere real part representation.

Besides, we change the input to Fourier transform to z-score transformed surprisal scores instead of raw ones. That is, the $s(n)$ in Equation (1) is replaced with $\tilde{s}_n = \frac{s_i - \mu}{\sigma}$, where $\mu = \sum s_i/N$ and $\sigma = \sum(s_i - \mu)^2/(N-1)$ are the mean and standard deviation of the input surprisal sequence.

The changes above are inspired by the operations adopted in FourierGPT (Xu et al., 2024), which they show to have improved the accuracy of model-generated text detection.

4.2 Modification to distance functions

Old metrics removed from FACE-1 Among the four original metrics used, spectral overlap (SO), Pearson's correlation (CORR), cosine similarity, and Spearman's correlation, we only keep the first two (SO, CORR) in this study for the following two reasons: (i) CORR and cos-sim are equivalent when the data being compared are normalized by subtracting the means (which is the case in our experiments); (ii) Spearman's correlation has been shown not suitable for measuring spectral similarity according to FACE-1's findings.

New metrics added to FACE-2 Earth Mover's Distance (EMD) (Rubner et al., 1998), Kullback-Leibler divergence (KL)(Kullback and Leibler, 1951), and Jenson-Shannon divergence (JS, also named as total divergence to the average by Dagan et al., 1997), are implemented in FACE-2. These metrics are chosen because they are widely used for measuring the distance between two probability distributions, and all come with good interpretability. In particular, EMD is a long-standing distance function suitable for image retrieval with spectral features (Rubner et al., 2000; Deborah et al., 2015), which motivates us to migrate it to a similar scenario, that is, text similarity comparison based on spectral features (both image data and language data are treated as time series). KL is an asymmetric for measuring the difference between two distributions, which is suitable here because a normalized spectrum can be considered as a distribution. JS is a symmetrized and smoothed version of KL, and it is adopted to mitigate the asymmetricality issue in KL.

In sum, five metrics are adopted in FACE-2, SO, CORR, EMD, KL and JS, defined as:

$$\text{SO} = \text{AUC}(\mathcal{F}_h \cap \mathcal{F}_m)/\text{AUC}(\mathcal{F}_h \cup \mathcal{F}_m) \quad (2)$$

$$\text{CORR} = cov(\mathcal{F}_h, \mathcal{F}_m)/\sigma(\mathcal{F}_h)\sigma(\mathcal{F}_m) \quad (3)$$

$$\text{EMD} = \int_{-\infty}^{\infty} |\mathcal{F}_h - \mathcal{F}_m| \quad (4)$$

$$\text{KL} = \sum_{x \in [0, \pi/2]} \mathcal{F}_h(x)\ log(\frac{\mathcal{F}_h(x)}{\mathcal{F}_m(x)}) \quad (5)$$

$$\text{JS} = \frac{1}{2}\text{KL}(\mathcal{F}_h, \mathcal{F}_m) + \frac{1}{2}\text{KL}(\mathcal{F}_m, \mathcal{F}_h) \quad (6)$$

In the above formulas, $\mathcal{F}_h$ and $\mathcal{F}_m$ are the text spectra from two sources (human and model) respectively. The AUC in Equation (2) refers to the area under the curve. According to EMD's definition in Equation (4), it reflects the amount of "work" must be done to transform one distribution $\mathcal{F}_m$ into another $\mathcal{F}_h$, where the spectrum is treated as a probability density function for KL's definitoin in Equation (5),

5 Experiments

We evaluate the performance of FACE-2 in comparison with its previous version and other metrics, with respect to the following two desiderata:

1. Model scaling effect: To what degree the metric scores are in line with model the scaling effect of model size, that is, larger models scores better than smaller ones. (Section 5.2)
2. Agreement with human preferences: How well a metric aligns with human judges' rates on text quality. (Section 5.3)

Open-ended text generation tasks are conducted, which can be formulated as follows: given a sequence of m tokens denoted $[x_1 \ldots x_m]$, as the prompt, the goal is to generate the next n tokens to form a complete sequence $[x_1 \ldots x_{m+n}]$. Hyperparameters such as the sampling strategies, maximum and minimum generation lengths, and the GPU hour costs are reported in Appendix A.3.

The generated texts and human texts are then used to compute FACE-2 scores, which we compare with other metrics in terms of model scaling effect (Section 5.2). Moreover, we test the resulting scores' agreement with human preferences using MT-Bench (Section 5.3), a benchmark for NLG annotated with human judges.

Language	Domain	Dataset
Chinese	Wiki	Wikipedia-Chinese
	News	MNBVC-News
	Story	WebNovel
English	Wiki	Wikipedia-English
	News	BBC-News
	Story	WritingPrompts
	Domain QA	PubmedQA
	General QA	LIMA

Table 1: Both Chinese datasets and English datasets contain continuous writing tasks of three domains: News article, Wikipedia document, and Story. Additionally, we provide two open-ended text generation datasets for instructed models.

Model Cat.	Family	Size	Lang.
Generator	Qwen-2	1.5b, 7b, 72b	Chinese
	Yi-1.5	6b, 34b	
	Llama-3	8b, 70b	English
	Gemma-2	2b, 9b, 27b	
Evaluator	Qwen-2	0.5b, 1.5b, 7b, 72b	Chinese
	Pythia	410m, 1.4b	English
	Llama-3	8b, 70b	

Table 2: Models used in the experiment. For English datasets, both Chinese models and English models are used for generation and evaluation experiments. For Chinese datasets, we uses Chinese generators only, and Qwen-2 for Chinese evaluation.

5.1 Datasets and models

Datasets In FACE-2, we evaluate eight datasets in two languages, English and Chinese (Table 1). The main datasets for both languages are from three domains: *Wiki*, *News*, and *Story*, as they are common choices in previous studies on evaluation and detection methods (Pillutla et al., 2021; Mitchell et al., 2023; Bao et al., 2024). For English, we include two QA datasets of short texts, because FourierGPT works particularly well on this type (Xu et al., 2024). For Chinese, the three main datasets are Wikipedia-Chinese (Wikimedia-Foundation), MNBVC-News (MNBVC Team, 2023), and WebNovel (Jun, 2023), respectively. For English, they are Wikipedia-English (Wikimedia-Foundation), BBC-News (RealTimeData), and WritingPrompts (Fan et al., 2018), respectively. As mentioned above, the two extra English datasets are PubmedQA (Jin et al., 2019), a domain-specific dataset, and LIMA (Zhou et al., 2023) a general QA dataset. Note that the datasets of three main domains are used on the base models, while the two QA datasets are used on the instruction-tuned models accordingly. Detailed statistics and data cleaning steps are reported in Appendix A.2.

Models The models used in this study are in two categories based on their purposes: **generator** and **evaluator** (Table 2). The former is used to generate texts, whose surprisal is estimated by the latter. For English generation task, we use LLaMA-3 (8b, 70b) (Touvron et al., 2023) and Gemma-2 (2b, 9b, 27b) (Team, 2024). For English experiments, we use Pythia (410m, 1.4b) (Biderman et al., 2023), LLaMA3 (8b, 70b) as estimator. For Chinese generation, we use two models: Qwen-2 (1.5b, 7b, 72b) (Yang et al., 2024) and Yi-1.5 (6b, 34b) (Young et al., 2024). For Chinese experiments, we use Qwen-2 (0.5b, 1.5b, 7b, 72b) as the evaluator. We compare performances of different model families in various sizes, with details listed in Table 2. Both Chinese and English models (including generators and evaluators) are used for English generation tasks and follow-up evaluations. But for Chinese datasets, we only use Chinese generators (Qwen-2 and Yi-1.5) and one evaluator (Qwen-2), because we found that when English generators (Gemma-2 or LLaMA-3) are prompt with Chinese, they tend to "code switch" to English frequently in the generations, which harms our language-specific examination purpose.

Other metrics compared Besides FACE-1, we compare with other four common evaluation metrics: surprisal, MAUVE, Zipf, and Self-BLEU. Their definitions are introduced in Appendix A.1.

5.2 Result: Model scaling effect

We test the degree to which "scaling law" holds to the metric scores of generated texts, by *counting* the number of cases that strictly satisfy the condition large $\succ$ medium $\succ$ small, in terms of generator size, where the $\succ$ operator means that its left operand scores better than its right operand, which is defined differently across metrics. Albeit recent endeavors of training smaller models that perform similarly as larger ones (Warstadt et al., 2023), we argue this assumption is still valid because all comparisons are made within the same

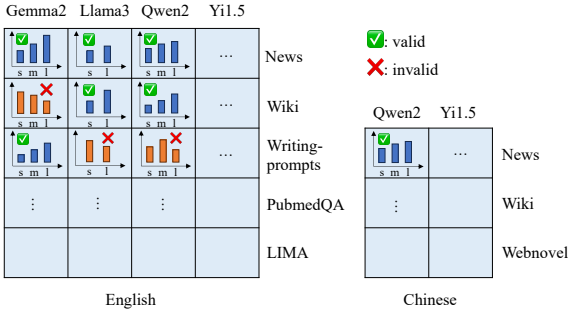


Figure 3: Illustration of how the scaling effect of generator models is analyzed. The tables represent generator model × task combinations, and the "s", "m", "l" in each cell indicates small, medium, and large model size, respectively.

Lang.	Metric	Valid Cell Ratio
English	FACE-1: SO	0.356 (0.053)
	FACE-2: EMD	**0.400 (0.173)**
	surprisal	0.363 (0.143)
	MAUVE	0.200
	Zipf	0.350
	Self-BLEU	0.150
Chinese	FACE-1: SO	0.417 (0.083)
	FACE-2: JS	**0.583 (0.083)**
	surprisal	0.375 (0.415)
	MAUVE	0.333
	Zipf	0.333
	Self-BLEU	0.000
	rand. guess	0.167

Table 3: The ratio of valid cells that satisfy the assumption of "large $\succ$ medium $\succ$ small" in terms of generator size. Numbers in parentheses are standard deviations. MAUVE, Zipf, and Self-BLEU do not have SDs because they do not rely on different evaluators.

model family with same architecture.

We consider all model family × task combinations as shown in Figure 3: the English table consists of 4 (column: generator models) × 5 (row: tasks) = 20 combinations; the Chinese table consists of $2 \times 3 = 6$ combinations, accordingly. A cell of the table contains the evaluation scores for the texts generated from three versions of the same model: *small*, *medium*, and *large* (except Llama3, which only has two versions). For each cell, if the scores strictly satisfy the inequality *large* $\succ$ *medium* $\succ$ *small*, we put a valid mark. Otherwise, we mark that cell invalid. Finally, we use the *ratio of valid cells* to indicate the degree to which a specific metric follows the scaling law.

Here our approach is based on the common belief that larger models should have better overall generation performance (Radford, 2018; Radford et al., 2019; Raffel et al., 2020; Kaplan et al., 2020; Brown, 2020), including open-ended text generation. Under this belief, if a good metric that claims itself to precisely characterize the model's capability of generating "high-quality" texts, then it should produce scores that rank larger models higher than smaller ones. That means we can judge the soundness of a metric by calculating the valid cell ratio, to see whether it is higher than a random level, and how it compares to the ratios from other metrics. For example, if metric m_A results in 7 out of 20 valid cells for English, which out-performs the 5 out of 20 from another metric m_B, then we can conclude that both are better than a random guess ($1/P_3^3 = 1/6 \approx 0.167$), but m_A is preferred over m_B as the former shows stronger scaling effect.

The results are shown in Table 3. It can be seen that all metrics surpass the random guess baseline, and **FACE-2 has higher ratio than its predecessor FACE-1 and other metrics**. Thus, our modification to the original FACE approach can indeed better reflect the scaling effect of model size. This improvement is non-trivial: FACE-1 uses the raw surprisal as input to Fourier transform, which means the spectrum obtained is merely a linear transformation of surprisal, and thus it is not surprising to see a very close performance to using surprisal; on the other side, FACE-2 uses z-score to capture the "relative" surprisal, which captures a more evaluator-independent nature about how surprisal changes in texts and the human-machine difference in this. Furthermore, there are several discussions we would like to have as follows.

Problem of surprisal as a metric We notice that using pure surprisal as a metric also reaches decent ratios: 0.363 for English and 0.375 for Chinese, which are close to those of FACE-1. However, these ratios have much larger standard deviations, which are due to the usage of different evaluators: For Chinese data, when using surprisal estimated by Qwen2-0.5b and -1.5b evaluators (relatively small models) as the metric, the valid ratios are as low as zero; but the ratio reaches 1.0 (100%) when the largest evaluator model Qwen2-72b is used; similar results are found in English data. It indicates an apparent problem of using raw surprisal to evaluate text generation: surprisal scores

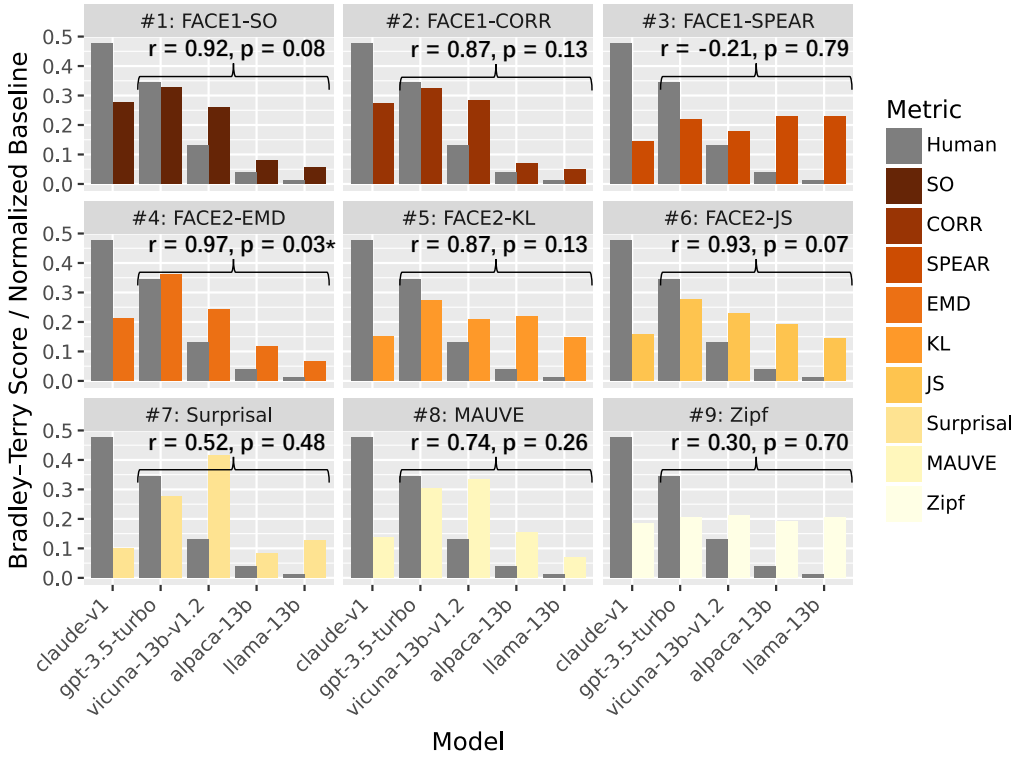


Figure 4: Bradley-Terry scores that reflect human preferences and evaluation metrics over model generated texts. A higher bar means the corresponding source of text is more preferred (by human) or higher scored (by metric). On the first row are the three metrics inherited from FACE-1. On the second row are the three metrics newly proposed in FACE-2. On the last row are the normalized naive average Surprisal, MAUVE, and average Zipf as baselines. For all the figures, models are from left to right in the order of human preference.

are low (preferred) when the evaluator matches in model size with the generator; otherwise, surprisal scores are high (unwanted). Therefore, using pure surprisal as a metric is extremely biased towards the evaluator used, and consequently, cannot produce consistent evaluation scores that meets the common intuition that "larger model is better".

The only solution is to always use the largest evaluator, like Qwen2-72b or Llama3-70b here, which always produce low surprisal scores to texts from models of similar sizes (≈70b). This, however, becomes a significant computational burden when the amount of text grows, or when even larger generator models (100b+) are to be evaluated.

FACE-2 vs. FACE-1 In general, the newly introduced distance functions lead to higher valid cell ratios. We compare the SO version of FACE-1 with the EMD, KL, and JS versions of FACE-2, and find the latters result in higher mean ratios than the former: 0.390 ($\sigma = 0.07$) > 0.356 ($\sigma = 0.05$) for English; 0.528 ($\sigma = 0.09$) > 0.417 ($\sigma = 0.08$) for Chinese. Full results are reported in Appendix B.2. The L2-norm operation on the spectrum, however, seems not being helpful: while the best ratio is achieved in the L2-normed FACE-2, the overall ratios across all distance functions is higher or equally good in FACE-1. This is counter-intuitive as L2-norm harnesses more spectral information than only using the real part, and it also contradicts the findings in FourierGPT (Xu et al., 2024). We leave the qeustion to future work.

In addition, we illustrate GAM-smoothed spectrum curves of generator model sizes in a family to examine whether larger models are closer to human curves. Two pairs of examples are shown in Figure 5, in which FACE-2 curves better satisfy the expected relationship.

Compare to other metrics To our surprise, the ratios from MAUVE is pretty low: 0.200 for English and 0.333 for Chinese. It seems that the texts generated by models of various sizes (at least within the investigation scope of this study) cannot be effectively distinguished using semantic representations, such as the clustering-based method by MAUVE. We conjecture that this is because MAUVE uses GPT-2 as its internal semantic en-

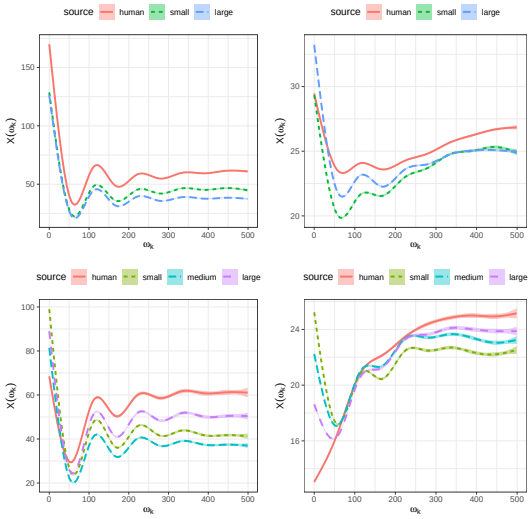


Figure 5: Exemplars of GAM smoothed spectrum plots. On the left are FACE-1 plots, and on the right are FACE-2 plots. Top row is from Llama-3 on BBC-News; bottom row from Qwen-2 on Wikipedia-Chinese.

coder, which is of much lower semantic expressiveness compared to the generators in our experiments. Zipf achieves comparable ratios as surprisal, but its computational cost is not neglectable. The worst result is from Self-BLEU (0.0 ratio for Chinese), which is an interesting language-specific question to explore in the future.

5.3 Result: Agreement with human preferences

We use the MT-Bench (Zheng et al., 2023) dataset to evaluate how much FACE-2 agree with human preferences. MT-Bench collects multiple models' responses to a set of prompts that come in pairs, and records human judges' preferences for each pair, i.e., which one wins. A common practice for converting the win counts to a quantify that represents model performances is the Bradley-Terry model (Bradley and Terry, 1952), which returns a ranked list of models, and each model is assigned a score, i.e., Bradley-Terry (BT) score, that reflect the human judges' relative preferences.[1]

Yang et al. (2023) already showed FACE-1's advantage over MAUVE using this dataset, which is why we did not use it again here. Another reason for switching to MT-Bench is that the original MAUVE-dataset are generated by GPT-2 models, which no longer reflects the actual generation qualities of current models. The only issue in MT-Bench is that it does not contain real human generations, which is necessary for any comparison-based evaluation metrics. To deal with this, we use GPT-4's responses as the approximated ground-truth, that is, *as if* they are human generations. We argue that this is an acceptable operation, as long as GPT-4 is excluded in the final analysis. The next steps are similar to those in MAUVE and FACE-1 studies: We apply Bradley-Terry model to the win counts returned by FACE-2, resulting in BT scores that we first normalize and then compare with the BT scores from human preferences and other metrics.

The BT scores of FACE-2, FACE-1, naive surprisal, MAUVE, and average Zipf scores in comparison with human preference are plotted in Figure Figure 4. It can be observed that FACE family align better with human preferences than MAUVE, surprisal, and Zipf. FACE-2 perform better than FACE-1, showing higher Pearson's correlations with human preferences. An interesting finding is that all evaluation metrics disagree with human preference on *Claude-v1*. It requires further qualitative analysis to understand the reasons behind this discrepancy, but we provides a case analysis (Appendix B.4) that may sheds some light on it.

6 Conclusion

In this study, we propose significant improvements to Fourier analysis of cross-entropy (FACE), a recently developed metric for evaluating open-ended text generation tasks, obtaining a new set of metrics named FACE-2. The metric roots in the psycholinguistics findings that natural language presents *periodicity in surprisal*, which can reflect the quality of generated text with curated distance metrics in spectral space.

When compared with its predecessor and other metrics, FACE-2 has the following merits: First, it can more effectively distinguish texts generated by models of various sizes, i.e., better reflecting the scaling effect. This is observed across different model families and languages, and more impressively, it is achieved without dependence on large evaluating models. Second, with the newly introduced metrics (EMD, KL, and JS) and the new z-score processing step, it achieves better agreement with human preferences on text quality. In sum, we have proposed a new state-of-the-art evaluation method for open-ended text generation, which is effective, robust, and computationally efficient.

[1] This method was first used by Pillutla et al. (2021) to measure the agreement between MAUVE and human preferences, and they published the dataset.

7 Limitations

There are still limits in the current study. Firstly, larger models 100b+ parameters and commercial models, such as GPT-4, are not included, whose experiments are needed for further validation of FACE-2's capability. Secondly, the majority of the generator models examined are base-models but not chat-models, although we believe that the foundational generation capabilities are determined at pre-training stage and is sufficient for current experiments, ideally, more experiments with chat-models for needed comprehensive evaluations, because chat-models are the most common use cases of LLMs. However, it is currently challenging to find public dialogue data with human ground-truth that can be directly used for generation tasks, especially for Chinese data and other languages. Lastly, it is not clear why the spectrum of surprisal provides information can reflect the sources of generator models. We need further evidence that map the spectral features in "frequency domain" to the observable linguistic patterns in "time domain", for example, sentence structure, lexical or syntactical choices and so on. We will try to address these limitations in the future work, and further improve FACE to a more interpretable method.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

David Cohen. 1962. Type-token mathematics, a textbook of mathematical linguistics.

J. W. Cooley and J. W. Tukey. 1965. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63, Madrid, Spain. Association for Computational Linguistics.

Hilda Deborah, Noël Richard, and Jon Yngve Hardeberg. 2015. A comprehensive evaluation of spectral distance functions and metrics for hyperspectral image processing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):3224–3234.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, Yanchao Yu, Verena Rieser, and Oliver Lemon. 2016. Information density and overlap in spoken dialogue. *Computer Speech & Language*, 37:82–97.

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

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering.

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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jun. 2023. webnovel$_cn(revision745338c)$.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Eliot Maës, Philippe Blache, and Leonor Becerra. 2022. Shared knowledge in natural conversations: can entropy metrics shed light on information transfers? In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 213–227, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

MNBVC Team. 2023. Mnbvc: Massive never-ending bt vast chinese corpus. https://github.com/esbatmop/MNBVC.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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

Steven Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112—-1130.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

RealTimeData. bbc-news-alltime. https://huggingface.co/datasets/RealTimeData/bbc_news_alltime.

Y. Rubner, C. Tomasi, and L.J. Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, Bombay, India.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Gemma Team. 2024. Gemma.

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. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. GPT-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on*

Computational Natural Language Learning, pages 1–34, Singapore. Association for Computational Linguistics.

Wikimedia-Foundation. Wikimedia downloads. https://dumps.wikimedia.org.

Yang Xu and David Reitter. 2016. Entropy converges between dialogue participants: Explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–546, Berlin, Germany. Association for Computational Linguistics.

Yang Xu and David Reitter. 2017. Spectral analysis of information density in dialogue predicts collaborative task performance. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–633, Vancouver, Canada. Association for Computational Linguistics.

Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. Detecting subtle differences between human and model languages using spectrum of relative likelihood. *arXiv preprint arXiv:2406.19874*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zuhao Yang, Yingfang Yuan, Yang Xu, SHUO ZHAN, Huajun Bai, and Kefan Chen. 2023. Face: Evaluating natural language generation with fourier analysis of cross-entropy. In *Advances in Neural Information Processing Systems*, volume 36, pages 17038–17056. Curran Associates, Inc.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

G Zipf. 1936. *The Psychobiology of Language. London: Routledge*. London: Routledge.

G Zipf. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

A Appendix: Reproduction

A.1 Other metrics

We compare FACE-2 with FACE-1, and other four metrics:

Surprisal It is straight-forward: the lower surprisal a text produces according to an evaluator, the better score it receives. The exponentiated average surprisal of a sequence of words is exactly how perplexity is defined.

MAUVE It returns a number in the scope of $[0, 1]$. A larger value indicates a more similar semantic distribution to human written texts, which indicates higher text quality. The score is computed using the public implementation provided by Pillutla et al. (2021).

Zipf It is the slope of the best-fit line on log-log plot of a rank versus unigram frequency plot. A smaller value indicates it is closer to human distributions, i.e., higher text quality. We use the open-sourced implementation from Holtzman et al. (2020).

Self-BLEU This score is computed by following the same protocol of Holtzman et al. (2020): computing the BLEU score of each generations against all other generations as references. A lower final score suggests higher diversity of the generated text, which is an important indicator of text quality.

A.2 Data cleaning

We clean the datasets to ensure quality: repetitions of sentence or meaningless strings are removed; next we sample a subset from each dataset, ensuring the texts in the subsets as controlled variable have comparatively equal length. For all datasets from the three main categories (except for WritingPrompts), we split each into two parts, with the first half as the prompt for generation, and the second half as human ground-truth. The prompt length is set to 64 tokens, and during generation, we limit the model maximum generation length to 1024. For two QA datasets plus WritingPrompts, we directly use the prompt provide by themselves. The size of each subset we used are 5000, except for Wikipedia-Chinese and MNBVC-News, where

we removes some texts containing too much English. The size of Wikipedia-Chinese's subset is 2160, and MNBVC-News' is 4206. LIMA originally contains 1030 groups of dialogue, we remove those excessively long dialogue, the size of remaining subset is 900.

To further ensure consistency in generation, we use the same tokenizer across all generator models. For Chinese, we use the default tokenizer from Qwen2, and for English it is the Llama3 tokenizer.

A.3 GPU usage and hyper-parameters

We used 4 A6000ada 48G GPU for the experiments in this studies. We used vllm to speed up generation and inference. We notice that vllm does not guarantee stable log probability, but this instability only affects the vocabularies' log probabilities in decimal places ($\leq$1e-3), which can be ignored. The arguments for vllm are as follows:

- `temperature`: 1
- `top k`: 50
- `top p`: 0.95

For the estimated times for running generator and evaluator models are listed in Table 4. To deal with an out-of-memory bug in vllm, we clear the cuda cache after each step of output. This slows down the speed about 20%$\sim$30%. Taking all model family $\times$ task combinations together, the total running time of our experiments is about $(20 + 6) \times 3\text{h} \approx 140\text{h}$.

Task	Model Size	Running Time
Generation	1.5b	10min$\sim$20min
	7b	20min$\sim$30min
	30b (4$\times$GPUs)	1h$\sim$2h
	70b (4$\times$GPUs)	1h$\sim$2h
Evaluation	0.5b	5min
	1.5b	5min
	7b	10min
	70b (4$\times$GPUs)	40min$\sim$2h

Table 4: Runtime of generation tasks and evaluation tasks for different model sizes. The runtime we reported is a rough value, which may varies according to the environment.

B Appendix: Results

B.1 Spectrum plots

The full spectrum plots of all model family $\times$ task combinations are shown in Figure 4.

B.2 Full valid cell ratios

B.3 MT-Bench comparison

The detailed MT-Bench comparison results are reported in Table 6. We exclude GPT-4 and recompute the results of original MT-Bench human preference BT-score. We also report the BT-score of FACE-2 metrics EMD and SO, and FACE-1 SO. We see that the old metric SO performs only slightly difference in scores, while the new metric EMD is much closer to human BT-score than SO. Since MAUVE is not capable for pair-wise comparison, we report the raw MAUVE score over these five models directly.

B.4 Case analysis for the mismatch in Claude-v1 output

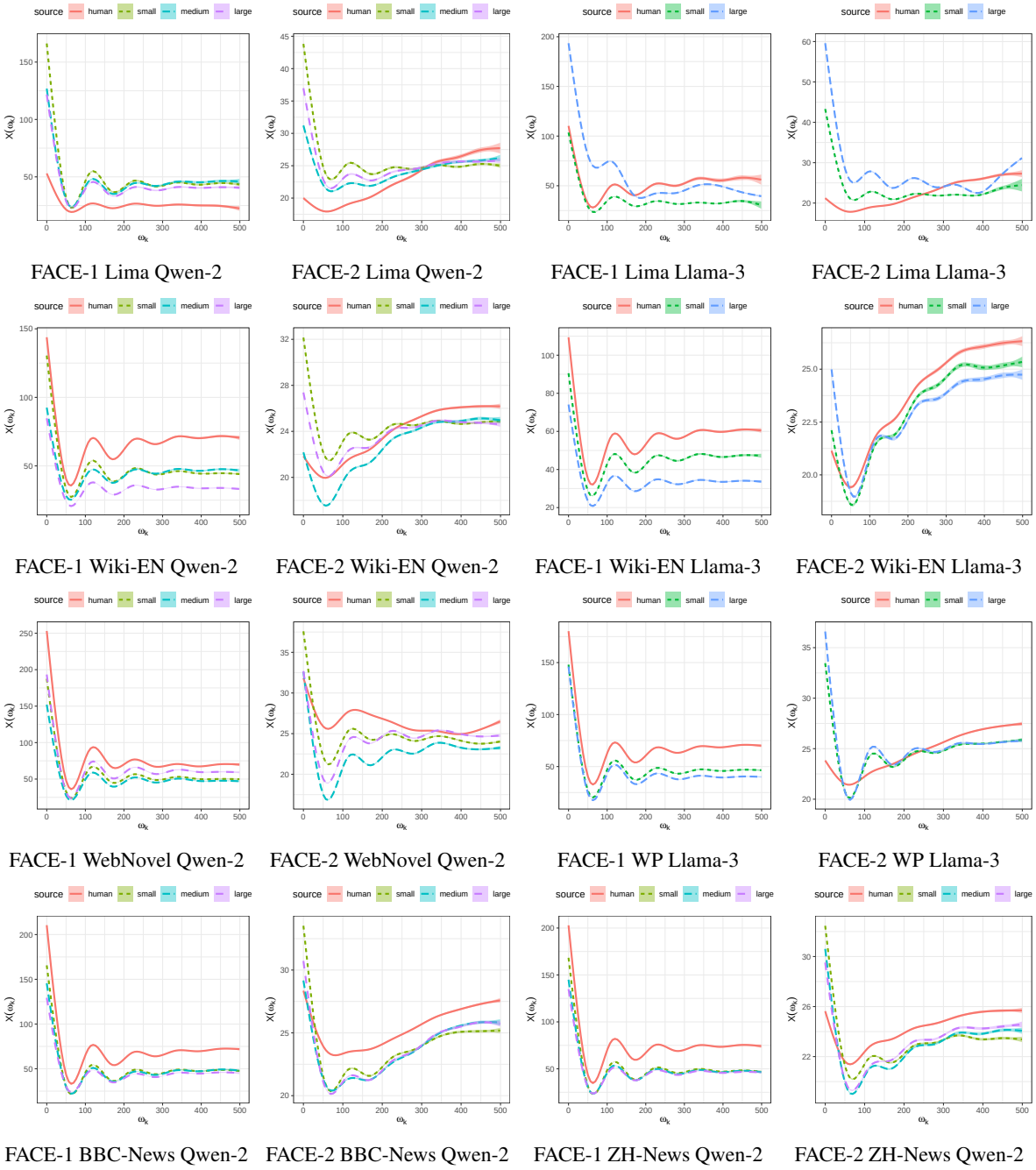
The mismatch between human preference and metric scores in Claude-v1 outputs might be due to the fact that Claude-v1 is more likely to perform as a chat model and outputs explanations while others directly write what is requested. For example, we notice that, when the prompt is "*Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.*", Claude-v1 first generates a sentence like "*Here is a draft travel blog post about a recent trip to Hawaii:*", and then gives the main body of post. In contrast, other models tend to generate the post directly. Evaluation metrics are sensitive to this kind of semantic changes, while human participants of MT-Bench that are uninformed of model-generated texts' features focus more on the content quality. Hence FACE considers that Claude-v1's text is far from natural, while humans rank its content to the top.

Metric	Lang.	z-score	Dist. func.	Evaluator								Mean (SD)
				en1	en2	en3	en4	zh1	zh2	zh3	zh4	
FACE-1 (real part)	en	no	SO	0.30	0.25	0.40	0.40	0.40	0.35	0.40	0.35	0.36(0.05)
			CORR	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30(0.00)
			EMD	0.40	0.30	0.45	0.40	0.45	0.40	0.30	0.40	**0.39(0.05)**
			KL	0.35	0.50	0.35	0.40	0.35	0.30	0.35	0.55	0.39(0.08)
			JS	0.35	0.50	0.40	0.40	0.35	0.30	0.35	0.45	0.39(0.06)
		yes	SO	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.45	0.36(0.03)
			CORR	0.25	0.50	0.20	0.20	0.10	0.05	0.15	0.55	0.25(0.17)
			EMD	0.25	0.10	0.55	0.50	0.60	0.55	0.30	0.35	**0.40(0.17)**
			KL	0.20	0.35	0.20	0.20	0.25	0.20	0.15	0.30	0.23(0.06)
			JS	0.25	0.35	0.20	0.20	0.20	0.15	0.25	0.30	0.24(0.06)
	zh	no	SO	-	-	-	-	0.50	0.50	0.33	0.33	0.42(0.08)
			CORR	-	-	-	-	0.33	0.33	0.33	0.33	0.33(0.00)
			EMD	-	-	-	-	0.67	0.50	0.50	0.67	**0.58(0.08)**
			KL	-	-	-	-	0.50	0.50	0.50	0.50	0.50(0.00)
			JS	-	-	-	-	0.33	0.50	0.50	0.67	0.50(0.12)
		yes	SO	-	-	-	-	0.50	0.50	0.67	0.67	**0.58(0.08)**
			CORR	-	-	-	-	0.17	0.00	0.33	0.83	0.33(0.31)
			EMD	-	-	-	-	0.50	0.67	0.50	0.50	0.54(0.07)
			KL	-	-	-	-	0.67	0.33	0.50	0.50	0.50(0.12)
			JS	-	-	-	-	0.50	0.33	0.50	0.50	0.46(0.07)
FACE-2 (L2-norm)	en	no	SO	0.40	0.45	0.35	0.25	0.40	0.40	0.30	0.40	0.37(0.06)
			CORR	0.15	0.15	0.20	0.40	0.05	0.00	0.00	0.45	0.18(0.16)
			EMD	0.40	0.45	0.35	0.25	0.40	0.55	0.35	0.35	**0.39(0.08)**
			KL	0.35	0.35	0.25	0.35	0.30	0.35	0.45	0.45	0.36(0.06)
			JS	0.15	0.25	0.30	0.35	0.30	0.25	0.35	0.50	0.31(0.09)
		yes	SO	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35(0.00)
			CORR	0.10	0.15	0.25	0.50	0.05	0.05	0.10	0.50	0.21(0.18)
			EMD	0.50	0.50	0.30	0.10	0.60	0.60	0.40	0.20	**0.40(0.17)**
			KL	0.25	0.20	0.10	0.20	0.15	0.15	0.10	0.10	0.16(0.05)
			JS	0.20	0.25	0.20	0.30	0.20	0.20	0.15	0.25	0.22(0.04)
	zh	no	SO	-	-	-	-	0.67	0.67	0.33	0.00	**0.42(0.28)**
			CORR	-	-	-	-	0.00	0.00	0.17	0.50	0.17(0.20)
			EMD	-	-	-	-	0.33	0.17	0.17	0.50	0.29(0.14)
			KL	-	-	-	-	0.33	0.33	0.33	0.33	0.33(0.00)
			JS	-	-	-	-	0.33	0.50	0.17	0.33	0.33(0.12)
		yes	SO	-	-	-	-	0.50	0.50	0.50	0.50	0.50(0.00)
			CORR	-	-	-	-	0.17	0.00	0.17	0.83	0.29(0.32)
			EMD	-	-	-	-	0.67	0.67	0.50	0.17	0.50(0.20)
			KL	-	-	-	-	0.67	0.50	0.33	0.33	0.46(0.14)
			JS	-	-	-	-	0.67	0.67	0.50	0.50	**0.58(0.08)**

Table 5: The full valid cell ratios from all experiments. Best results from each trial are highlighted. The English evaluator models (en1-en4) are pythia-410m, pythia-1.4b, llama3-8b, and llama3-70b. The Chinese evaluator models (zh1-zh4) are qwen2-0.5b, qwen2-1.5b, qwen2-7b, and qwen2-72b, respectively.

Human-BT		MAUVE		FACE2-SO-BT		FACE2-EMD-BT		FACE1-SO-BT	
Model	Score	Model	Score	Model	Score	Model	Score	Model	Score
Claude-v1	0.476	Vicuna-13b	0.708	GPT3.5-turbo	0.333	GPT3.5-turbo	0.303	GPT3.5-turbo	0.344
GPT3.5-turbo	0.342	GPT3.5-turbo	0.644	Vicuna-13b	0.242	Claude-v1	0.273	Vicuna-13b	0.281
Vicuna-13b	0.130	Alpaca-13b	0.325	Claude-v1	0.182	Vicuda-13b	0.212	Claude-v1	0.188
Alpaca-13b	0.039	Claude-v1	0.287	Alpaca-13b	0.152	Alpaca-13b	0.121	Alpaca-13b	0.125
Llama-13b	0.012	Llama-13b	0.149	Llama-13b	0.091	Llama-13b	0.091	Llama-13b	0.063

Table 6: Detailed MT-Bench comparison between human preference, FACE, and MAUVE.



FACE-1 Lima Qwen-2 FACE-2 Lima Qwen-2 FACE-1 Lima Llama-3 FACE-2 Lima Llama-3

FACE-1 Wiki-EN Qwen-2 FACE-2 Wiki-EN Qwen-2 FACE-1 Wiki-EN Llama-3 FACE-2 Wiki-EN Llama-3

FACE-1 WebNovel Qwen-2 FACE-2 WebNovel Qwen-2 FACE-1 WP Llama-3 FACE-2 WP Llama-3

FACE-1 BBC-News Qwen-2 FACE-2 BBC-News Qwen-2 FACE-1 ZH-News Qwen-2 FACE-2 ZH-News Qwen-2

Figure 6: Other spectrum plots.