Calibration, Entropy Rates, and Memory in Language Models

Mark Braverman [1] **Xinyi Chen** [1 2] **Sham Kakade** [3 4] **Karthik Narasimhan** [1] **Cyril Zhang** [1 2] **Yi Zhang** [1]

Abstract

Building accurate language models that capture meaningful long-term dependencies is a core challenge in natural language processing. Towards this end, we present a calibration-based approach to measure long-term discrepancies between a generative sequence model and the true distribution, and use these discrepancies to improve the model. Empirically, we show that state-of-the-art language models, including LSTMs and Transformers, are *miscalibrated*: the entropy rates of their generations drift dramatically upward over time. We then provide provable methods to mitigate this phenomenon. Furthermore, we show how this calibration-based approach can also be used to measure the amount of memory that language models use for prediction.

1. Introduction

Recent advances in language modeling have resulted in significant improvements on a wide variety of benchmarks (Dai et al., 2018; Gong et al., 2018; Takase et al., 2018). Capturing long-term dependencies has especially been a major focus, with approaches ranging from explicit memory-based neural networks (Grave et al., 2016; Ke et al., 2018) to optimization improvements to stabilize learning (Le et al., 2015; Trinh et al., 2018). However, while these techniques seem to improve on standard metrics like perplexity and even produce remarkably coherent text (Radford et al., 2019), we still do not have appropriate measures to assess long-term properties in language models, making it difficult to choose between different model options for downstream tasks.

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA [2]Google AI Princeton, Princeton, New Jersey, USA [3]University of Washington, Allen School of Computer Science and Engineering and Department of Statistics, Seattle, Washington, USA [4]Microsoft Research, New York, New York, USA. Correspondence to: Cyril Zhang <cyril.zhang@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Model	Corpus	Test ppl.	e^{EntRate}
1) AWD-LSTM	PTB	58.3	93.1
2) CNN-LSTM	GBW	29.8	49.4
3) Transformer	GBW	28.1	34.7
4) GPT-2	WebText	23.7	61.2

Table 1: Perplexity degradations for generations from popular language models. State-of-the-art performance is usually reported via perplexity with respect to the test corpus (one-step prediction loss), but there is a striking blowup in the perplexity (i.e. exponential of the entropy) of these models' long-term generations. **Test ppl.** is the exponential of the *cross-entropy* of the model with respect to the test corpus. The listed models are (1) Merity et al. (2017), (2) Jozefowicz et al. (2016), (3) Vaswani et al. (2017b), (4) Radford et al. (2019).

Starting from Shannon's seminal work that essentially introduced statistical language modeling (Shannon, 1951), the most classical and widely studied long-term property of a language model is its entropy rate — the average amount of information contained per word, conditioned on the preceding words. A learned model provides an upper bound for the entropy rate of a language, via its cross-entropy loss. The exponential of the entropy rate can be interpreted as the effective support size of the distribution of the next word (intuitively, the average number of "plausible" word choices to continue a document), and the perplexity score of a model (the exponential of the cross entropy loss) is an upper bound for this quantity. In state-of-the-art models trained on billion-scale corpora, this number ranges between 10 and 30 (Melis et al., 2017; Radford et al., 2019). A natural diagnostic question, with which we begin our work, is whether the long-term generations of these models exhibit the same entropy rates as the underlying languages they are modeling predictively.

Empirically, and perhaps surprisingly, it turns out that the entropy rate of generated text is *substantially* higher than the estimate for true text derived from the model's one-step predictions. As seen in Table 1 (see also Figure 1), this is true for both state-of-the-art LSTMs and Transformers trained on a variety of datasets. As a timely example, the GPT-2 model (Radford et al., 2019), the object of much recent attention for its seemingly coherent and on-topic gen-

erations, suffers a dramatic degradation in its entropy rate, from 23.7 to 61.2.

This empirical finding is notable since the neural attention- and memory-based techniques (Vaswani et al., 2017a) have been steadily improving on standard metrics like perplexity and, in some cases, even produce remarkably coherent text (often with some heuristics to reject poor generations). That the perplexity of generated text is so much higher than it is under the true distribution suggests that there are significant gaps in our current methodologies in accurately learning language models, particularly if we are interested in generating long sequences of texts that globally resembles the modeled language itself.

Our contributions. We identified the wide-spreadness of the entropy amplification among the state-of-the-art language models trained on various corpus. Based on this, the focus of this work is twofold: to improve generations based on any measurement mismatch on a long-term property of the model (e.g. the entropy rate) with provable guarantees, and to quantify the way a model's predictions depend on the distant past. Central to both of these is a calibration-based approach, which is utilized in statistics and other areas of machine learning (Dawid, 1982; 1985; Foster, 1991; Zadrozny & Elkan, 2002; Platt, 1999; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005).

First, we prove that, from a theoretic worst-case perspective, even an extremely accurate model (with ε average KL divergence from the true distribution) may have generated text with a substantially different entropy rate as compared to the true distribution. Indeed, we show that this worst-case amplification may occur for a variety of long-term properties of a probabilistic language model; this is because the one-step KL divergence does not in general provide tight control over the expectation of a bounded function. The observed entropy rate amplification (as seen in Table 1) demonstrates that this is not only of theoretical concern. We then describe a calibration procedure to fix this mismatch while simultaneously improving the perplexity of the language model. From a statistical perspective, the procedure is simple, and we discuss approaches to make it computationally efficient.

Second, we provide a definition for long-term memory in language models as the mutual information between the models predictions and the distant past in the input. We then provide an upper bound on the amount of this mutual information using calibrated distributions (with a single-parameter exponent). This allows us to estimate the amount of context used by a language model as a function of the distance of past tokens from the current prediction time step.

We perform empirical studies to accompany our theoretical results. We first use the entropy rate calibration algorithm to fix an LSTM language model, resulting in a drop of around 20 perplexity points in the generated text (so that the entropy rate of the model more accurately matches that of the language itself). Then, we empirically estimate and compare the long-term memory of state-of-the-art language models. Our insights point towards new ways of assessing (and fixing) language models, especially in terms of their long-term properties, in a manner complementary to existing metrics like perplexity.

2. Related Work

Improving language modeling with long-term dependencies. Recent approaches to improving language modeling have focused on several ways to better capture long-term dependencies, from using manually-defined context representations (Mikolov & Zweig, 2012; Ji et al., 2015; Wang & Cho, 2016) or document-level topics (Wang et al., 2017) to using LSTM recurrent neural networks with careful initialization (Le et al., 2015), auxiliary loss signals (Trinh et al., 2018) or augmented memory structures (Grave et al., 2016; Ke et al., 2018). Wiseman & Rush (2016) use scoring functions over *sequences* and search-based optimization to improve generation in seq2seq models.

More recent work has demonstrated the applicability of Transformer networks (Vaswani et al., 2017a) to the task, potentially side-stepping issues in training recurrent networks (e.g. vanishing/exploding gradients) and scaling to longer contexts (Dai et al., 2018; Radford et al., 2018). All these papers propose either architectural or optimization innovations to improve language model training. In contrast, we define and measure explicit long-term properties of language models and show that calibrating them correctly can provide improvements to any black-box language model.

Recent empirical breakthroughs have stemmed from language models which do not specify a unique autoregressive factorization (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019), and thus do not specify a unique $\widehat{\Pr}$. It remains an interesting problem to identify and sample from distributions induced by these models (Wang & Cho, 2019); thus, our end-to-end theoretical guarantees do not hold in this setting.

Information-theoretic approaches. While most language models aim to predict a distribution over the next token conditioned on the context, there have been alternative approaches relying on information-theoretic measures. Jost & Atwell (1994) propose a model which makes use of mutual information between word pairs to generate word sequences that retain longer-term dependencies. McAllester (2018) propose a training objective based on mutual infor-

mation for predictive modeling, and demonstrate its application for phoneme prediction. Clarkson & Robinson (1999) develop a hybrid metric using both perplexity and entropy rate, and show that it correlates better with a downstream metric like word error rate. Such works propose alternative optimization objectives; in contrast, we show how to use information-theoretic measures to improve models with respect to existing objectives like cross-entropy.

Measuring long-term statistics. Genzel & Charniak (2002) discuss and measure the growth of entropy conditioned on *local* context, as supporting evidence for the stationarity the entropy rate (which we show not to hold empirically for neural language models). More recently, Khandelwal et al. (2018) analyze LSTM-based language models and empirically show that such models make use of a finite context for prediction. Lin & Tegmark (2017) measure mutual information between any two symbols in human languages, and show that it decays with distance, roughly following a power law distribution. Takahashi & Tanaka-Ishii (2018) provide an upper bound for the entropy (character-level) of human languages by training neural language models with various context and data sizes and extrapolating to infinity. While we also make use of measures like entropy and mutual information across longer contexts, our goal is to use these to better calibrate the language model and provably improve its perplexity.

Calibration and integral probability metrics. The idea of matching properties of the models' predictions to the empirical outcomes, in an online setting, goes back (at least) to the "prequential principle" of Dawid (1982; 1985), with subsequent work in online and game-theoretic settings (Foster, 1991; Vovk, 2001; Kalai et al., 1999). The idea of improving probability scores is also common in machine learning (Zadrozny & Elkan, 2002; Platt, 1999; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). Recently, (Ott et al., 2018) assessed model calibration for machine translation systems using word-level probabilities. The notion of examining the expectation of functions as a metric for the distance between two distributions sometimes goes under the name of integral probability metrics (Mller, 1997; Sriperumbudur et al., 2009), and this notion is becoming increasingly relevant again in unsupervised learning through the connections to GANs (Mroueh & Sercu, 2017). In this work, we directly focus on the KL divergence, where our use of calibration is largely based on basic facts about exponential families (Brown, 1986).

Relation to generation-improving heuristics. If the sole objective is to improve the qualitative coherency of sampled generations, a wide variety of heuristics exist in the literature. The simplest of these is a constant multiplicative adjustment to the model's logits (known as *softmax temperature* (Xie, 2017)). This is a specific version of our method (Algorithm 2) with a constant logistic regression feature instead of the next-token conditional entropy. Relatedly, *greedy* and *top-k* sampling (used in state-of-the-art works such as (Radford et al., 2019)) are heuristics which make local modifications to the model's conditional probabilities to decrease diversity and eliminate nonsensical generations. (Ott et al., 2018) qualitatively corroborate our findings in translation models, finding that they oversmear probability mass, necessitating decoding strategies beyond sampling or MLE.

Efforts to push the empirical state of the art in generation quality have given rise to more complex heuristics. (Bengio et al., 2015) propose retraining the network on its own generations with a carefully scheduled probability for each token. Some works regularize a model's generations with an auxiliary reverse language model (Zhang et al., 2019; Liu et al., 2016). Yet others promote realism using adversarial training protocols (Bahdanau et al., 2016; Lin et al., 2017; Fedus et al., 2018). Broadly, these methods attempt to mitigate *exposure bias* (He et al., 2019) in models which depend on their own generations but are trained on ground-truth sequences.

We stress that our calibration methods result in a provable improvement in the original training objective (i.e. lower perplexity). As far as we know, none of the aforementioned heuristic methods can hope to provide such a strong guarantee, since they are fundamentally designed to bias models towards a different objective. Our work mitigates model hallucinations for (almost) free[1], in the sense that the global objective (entropy rate drift) is improved without worsening the local objective (perplexity). Furthermore, calibration preserves the computational efficiency of density estimation: a conditional probability vector from Algorithm 2 can be computed using O(vocabulary size) inferences on the original model. In the more advanced heuristics, the implied distribution over sequences is lost, and is only accessible by black-box sampling.

3. Preliminaries

We first define some useful quantities for our analyses. Let $\Pr(W_1, W_2, \ldots, W_T)$ represent the true underlying distribution over T length sequences of words, where the vocabulary is of size M. Let $W_{1:T}$ denote a random sequence of length T, with distribution $\Pr(W_{1:T})$. For clarity of exposition, we assume that all sequences (i.e. sentences or documents or books) are of equal length T.

For any distributions $\mathcal{D}$ and $\mathcal{D}'$ over length-T sequences, recall that the entropy $H(\cdot)$, KL-divergence, and en-

[1]Technically, at the statistical cost of fitting *one* more parameter.

tropy rate are, respectively, defined by: $H(\mathcal{D}) := \mathbb{E}_{w_{1:T}\sim\mathcal{D}}\left[\log \frac{1}{\mathcal{D}(W_{1:T}=w_{1:T})}\right]$, $\mathrm{KL}(\mathcal{D} \parallel \mathcal{D}') := \mathbb{E}_{w_{1:T}\sim\mathcal{D}}\left[\log \frac{\mathcal{D}(W_{1:T}=w_{1:T})}{\mathcal{D}'(W_{1:T}=w_{1:T})}\right]$, and $\mathrm{EntRate}(\mathcal{D}) := \frac{1}{T}H(\mathcal{D})$. Let $\widehat{\Pr}(W_{1:T})$ denote a learned distribution over sequences. In the typical sequential prediction setting, the probabilistic model is implicitly defined by the conditional distributions $\Pr(W_t|W_{<t})$, which are typically efficiently computable. It is standard for such a *language model* to be trained to minimize the *cross entropy* objective:

$$\begin{aligned}\mathrm{CE}(\Pr \parallel \widehat{\Pr}) :&= \frac{1}{T}\mathop{\mathbb{E}}_{w_{1:T}\sim\Pr}\left[\sum_{t=1}^{T}\log\frac{1}{\widehat{\Pr}(w_t|w_{<t})}\right]\\ &= \frac{1}{T}\mathop{\mathbb{E}}_{w_{1:T}\sim\Pr}\left[\log\frac{1}{\widehat{\Pr}(w_{1:T})}\right].\end{aligned}$$

Note that for an accurate language model, we would hope that: $\mathrm{CE}(\Pr \parallel \widehat{\Pr}) \approx \mathrm{EntRate}(\widehat{\Pr})$, i.e. the entropy rate of the sequences generated under the learned model is nearly that of the cross entropy of the model (with respect to the true distribution $\Pr$).

Throughout, we assume that

$$\frac{1}{T}\mathrm{KL}(\Pr \parallel \widehat{\Pr}) = \mathrm{CE}(\Pr \parallel \widehat{\Pr}) - \mathrm{EntRate}(\Pr) \le \varepsilon \tag{1}$$

holds for some ε. In other words, the (unknown) ε measures the degree of sub-optimality of the learned model, this ε is often referred to as the Bayes regret.

4. Calibration and Entropy Rates

In this section, we assess the long-term properties of language models when generating text. Specifically, we quantify the amplification in the entropy rate of generations under an ε-accurate model (Eq. 1). We then provide a procedure to fix this amplification, without increasing the perplexity of the model. Proofs for all statements are provided in the supplementary material.

For generality, consider a function $f : [M]^T \to \mathcal{R}$, defined on T length sequences. Let the mean and variance of f under distribution $\mathcal{D}$ be denoted by $\mu_{\mathcal{D}}(f)$ and $\sigma^2_{\mathcal{D}}(f)$

$$\mu_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{w_{1:T}\sim\mathcal{D}}[f(w_{1:T})]$$

$$\sigma^2_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{w_{1:T}\sim\mathcal{D}}[(f(w_{1:T}) - \mu_{\mathcal{D}}(f))^2].$$

4.1. Error amplification under a learned model

In this section we provide a tight upper bound on the difference between the cross entropy of the true distribution and the learned model, and the entropy rate of the learned model. We refer to this gap as error amplification, and show that in the worst case it scales with T.

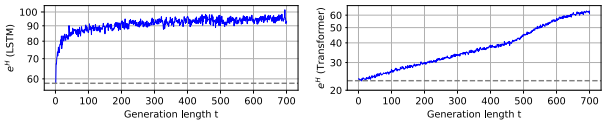


Figure 1: Perplexity (exponential of conditional entropy, given the past) of the t-th generated word, for two popular language models, averaged over more than 500 generation runs with different contexts. At $t = 1$, this is the model's upper bound for the language's perplexity. As $t \to \infty$, this is the exponential of the entropy rate of the model's own generations. For a perfectly calibrated model, this curve would be flat (gray dotted lines). *Left:* LSTM trained on Penn Treebank. *Right:* GPT-2 Transformer.

Before we proceed, notice that in some degenerate cases, $\widehat{\Pr}$ can assign very small probability to a word sequence, which can lead to infinite cross entropy vs. the true distribution. Therefore we rule out amplification of the model's entropy rate due to such reasons by considering the γ-mixture distribution, defined as: $\mathcal{D}^{(\gamma)} := (1-\gamma)\mathcal{D}+\gamma\mathrm{Uni}$, where Uni is the uniform distribution over all M^T sequences. Under this "smoothed" model, each sequence of length T has density at least γ/M^T. We will then consider the model $\widehat{\Pr}^{(\varepsilon)}$, which has only a minor degradation in the cross entropy compared to $\widehat{\Pr}$, and, yet, may have a large amplification in the entropy rate.

We show the error amplification of the smoothed model by considering the bounded function $f = -\log\widehat{\Pr}^{(\varepsilon)}$. If our learned model $\widehat{\Pr}$ is accurate, we may hope that $\mu_{\Pr}(f) \approx \mu_{\widehat{\Pr}}(f)$ i.e. that the expected value of f under the true distribution $\Pr$ is close to its expected value under our model. We can quantify this gap as follows:

Lemma 4.1. *(Pinsker's Inequality (Csiszar & Körner, 2011)) Suppose that for all $w_{1:T}$, $f(w_{1:T}) \le B$. Then:*

$$\left|\mu_{\Pr}(f) - \mu_{\widehat{\Pr}}(f)\right| \le B\sqrt{2\mathrm{KL}(\Pr \parallel \widehat{\Pr})}.$$

Using the above lemma on our choice of f, we arrive at the following result:

Corollary 4.2. *(Entropy rate amplification under generations) Suppose the bound in equation 1 holds. The ε-mixture distribution has KL bounded as:*

$$\frac{1}{T}\mathrm{KL}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) \le \left(1+\frac{2}{T}\right)\varepsilon.$$

We have that:

$$|\mathrm{CE}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) - \mathrm{EntRate}(\Pr)| \le \left(1+\frac{2}{T}\right)\varepsilon, \text{ and}$$

$$|\mathrm{CE}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) - \mathrm{EntRate}(\widehat{\Pr}^{(\varepsilon)})| \leq \sqrt{2\varepsilon(T+1)}\left(\log M + \frac{\log(1/\varepsilon)}{T}\right).$$

The last inequality is the error amplification bound, and it clearly shows that, in the worst case, even a small cross entropy may provide little control over the generations under the learned model (in terms of entropy rate). In fact, for $\varepsilon = O(\frac{1}{T})$ (which we may hope is an accurate model), the bound is vacuous; a later remark shows this worst case bound is unimprovable, see the supplementary material.

The above results suggest that entropy rate amplification is a theoretical possibility in the worst case, which our experiments show is in fact prevalent in practice. These entropy rate amplifications are evident from the plots in Figure 1. Regardless of the text corpus or the language model, we observe that the entropy rate under the model's generations quickly increases with time, indicating that this is a persistent problem even for state-of-the-art language models while generating text.

4.2. Model calibration

In this subsection we describe a procedure to fix this error amplification by calibrating the learned model $\widehat{\Pr}$.

Calibration to f Given a function f, we refer to a model as "calibrated to f" if the expectation of f under the true distribution and the model is equal. If f satisfies certain properties, we propose a method that uses a single extra parameter α to fit a *calibrated* model.

First, given a learned model $\widehat{\Pr}$ and a function f, define a distribution $\widehat{\Pr}_\alpha$ such that:

$$\widehat{\Pr}_\alpha(w_{1:T}) = \frac{\exp(\alpha f(w_{1:T})) \cdot \widehat{\Pr}(w_{1:T})}{Z_\alpha},$$

where $Z_\alpha = \sum_{w_{1:T}} \exp(\alpha f(w_{1:T})) \cdot \widehat{\Pr}(w_{1:T})$.

Then by finding the optimal α, we can recover a model calibrated to f:

Lemma 4.3. *(Calibration to f with model improvement) Suppose the variance of f is uniformly bounded in that there exists σ_+^2 such that the following holds for all α, $\sigma^2_{\widehat{\Pr}_\alpha}(f) \leq \sigma_+^2$. Let $\alpha^* = \mathrm{argmin}_\alpha \mathrm{CE}(\Pr \parallel \widehat{\Pr}_\alpha)$. We have*

$$\mu_{\Pr}(f) - \mu_{\widehat{\Pr}_{\alpha^*}}(f) = 0, \text{ and}$$

$$\mathrm{CE}(\Pr \parallel \widehat{\Pr}_{\alpha^*}) \leq \mathrm{CE}(\Pr \parallel \widehat{\Pr}) - \frac{1}{T}\frac{(\mu(f) - \mu_{\widehat{\Pr}}(f))^2}{2\sigma_+^2}.$$

The above lemma shows that $\widehat{\Pr}_{\alpha^*}$ is not only calibrated to f, but also has a lower cross entropy loss compared to the original model, which is an improvement.

Entropy rate calibration. We can now apply the result above to fix the entropy rate amplification seen in Table 1, by choosing $f = -\log \widehat{\Pr}^{(\varepsilon)}$. Note that it is trivial to avoid the entropy rate amplification if we were allowed to degrade the quality of our model, in terms of perplexity (e.g. a unigram model does not have this amplification). However, we show that it is possible to improve the learned model *and* more accurately match the entropy rate, by fitting a family of one-parameter models.

Theorem 4.4. *(Entropy rate calibration) Suppose equation 1 holds. Algorithm 1 returns a $\widehat{\Pr}_{\alpha^*}$ such that: the following calibration property is satisfied:*

$$\mathrm{CE}(\Pr \parallel \widehat{\Pr}_{\alpha^*}) = \mathrm{EntRate}(\widehat{\Pr}_{\alpha^*}).$$

Furthermore, $\widehat{\Pr}_{\alpha^*}$ has entropy close to the true entropy rate as specified by:

$$|\mathrm{EntRate}(\Pr) - \mathrm{EntRate}(\widehat{\Pr}_{\alpha^*})| \leq \left(1 + \frac{1}{T}\right)\varepsilon,$$

and $\widehat{\Pr}_{\alpha^*}$ is an improvement over the original model as characterized by:

$$\mathrm{CE}(\Pr \parallel \widehat{\Pr}_{\alpha^*}) \leq \mathrm{CE}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) - \frac{1}{2}\left(\frac{\mathrm{CE}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) - \mathrm{EntRate}(\widehat{\Pr})}{\log M + \frac{\log(1/\varepsilon)}{T}}\right)^2$$

This result shows that we simply need a single parameter α to define a new model class, and then we can fit this α to minimize the cross-entropy of the new model with respect to the true distribution $\Pr$, to eliminate the entropy rate amplification.

Even though this algorithm fits only a single parameter, it is computationally inefficient. Observe that to compute the cross entropy, one needs to integrate over all possible sequences of length T. One future direction would be to a sample based approach. This may be an interesting alternative to ideas like beam search (Steinbiss et al., 1994; Ortmanns & Ney, 2000; Antoniol et al., 1995), which also aims to minimize a global cost function on sequences that is inconsistent with the token-level perplexity loss used to train the underlying generative model.

Lookahead algorithms. In order to sidestep the computational issues of Algorithm 1, we provide another simple approach based on "one-step" lookahead correction (Algorithm 2). Instead of specifying a new distribution over all possible sequences, we only change the conditional distributions of each word in the sequence. This algorithm is

Algorithm 1 (Inefficient) Entropy Rate Calibration

1: Input: Model $\widehat{\Pr}^{(\varepsilon)}$.
2: Define a model class:

$$\widehat{\Pr}_\alpha(w_{1:T}) = \left(\widehat{\Pr}(w_{1:T})^{(\varepsilon)}\right)^{1+\alpha} / Z_\alpha \,.$$

3: Fit α^*: $\alpha^* = \operatorname{argmin}_\alpha \mathrm{CE}(\Pr \parallel \widehat{\Pr}_\alpha)$
4: Return $\widehat{\Pr}_{\alpha^*}$

Algorithm 2 Local Entropy Rate Calibration

1: Input: Model $\widehat{\Pr}^{(\varepsilon)}$, where $\widehat{W}_t \sim \widehat{\Pr}^{(\varepsilon)}(\cdot|W_{<t})$.
2: Define a model class:

$$\widehat{\Pr}_\alpha(w_{1:T}) = \widehat{P}_\alpha(w_1)\widehat{P}_\alpha(w_2|w_1)\dots$$

where

$$\widehat{\Pr}_\alpha(w_t|w_{<t}) = \frac{\widehat{\Pr}^{(\varepsilon)}(w_t|w_{<t})\exp(-\alpha H(\widehat{W}_{t+1}|w_{\le t}))}{Z_\alpha}$$

3: Fit α^*: $\alpha^* = \operatorname{argmin}_\alpha \mathrm{CE}(\Pr \parallel \widehat{\Pr}_\alpha)$
4: Return $\widehat{\Pr}_{\alpha^*}$

computationally efficient, and we show that it provides a local analogue of the calibration guarantee.

Let $\widehat{W}_t$ be a random variable with conditional distribution $\widehat{\Pr}^{(\varepsilon)}(\cdot|W_{<t})$. Let $H(\widehat{W}_{t+1}|w_{\le t})$ denote the entropy of this conditional distribution, i.e.

$$H(\widehat{W}_{t+1}|w_{\le t}) = \mathop{\mathbb{E}}_{w_{t+1}\sim\widehat{\Pr}^{(\varepsilon)}(\cdot|w_{\le t})}\left[\log\frac{1}{\widehat{\Pr}^{(\varepsilon)}(w_{t+1}|w_{\le t})}\right].$$

Note that $H(\widehat{W}_{t+1}|w_{\le t})$ includes the word w_t. When we predict using $\widehat{\Pr}_{\alpha^*}$ at time t, we need to compute the lookahead conditional entropy at time $t+1$.

Then for a conditional distribution, $\mathcal{D}(W_{1:T})$, define:

$$\bar{\mu}_{\mathcal{D}} = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{w_{<t}\sim\Pr} \mathop{\mathbb{E}}_{w_t\sim\mathcal{D}(\cdot|w_{<t})}[H(\widehat{W}_{t+1}|w_{\le t})]$$

Thus, $\bar{\mu}_{\mathcal{D}}$ is the average of $H(\widehat{W}_{t+1}|w_{\le t})$ with respect to a distribution which uses $\mathcal{D}$ for sampling the last word W_t (at every timestep). Intuitively, the resulting model $\widehat{\Pr}_\alpha$ in Algorithm 2 with a positive α would suppress sampling words leading to larger entropy but rather encourage words that stablizes the entropy 1-step ahead in the future. Therefore, if our learned language model $\widehat{\Pr}$ was accurate, we would hope that: $\bar{\mu}_{\Pr} \approx \bar{\mu}_{\widehat{\Pr}}$. The following corollary shows that this is achievable, along with improving the model's perplexity.

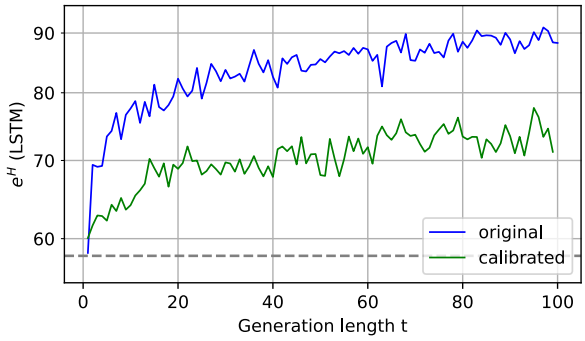


Figure 2: Effect of calibrating an LSTM generative model with 1-step lookahead. Blue: perplexity curve (i.e. exponential of conditional entropy H) from the setting of Figure 1. Green: the same perplexity measurements after applying local calibration.

Corollary 4.5. *Suppose Equation 1 holds. Then, Algorithm 2 returns a $\widehat{\Pr}_{\alpha^*}$ such that:*

$$\bar{\mu}_{\Pr} - \bar{\mu}_{\widehat{\Pr}_{\alpha^*}} = 0, \quad \textit{and}$$

$$\mathrm{CE}(\Pr \parallel \widehat{\Pr}_{\alpha^*}) \le \mathrm{CE}(\Pr \parallel \widehat{\Pr}^{(\varepsilon)}) - \frac{1}{2}\left(\frac{\bar{\mu} - \bar{\mu}_{\widehat{\Pr}^{(\varepsilon)}}}{\log M + \frac{\log(1/\varepsilon)}{T}}\right)^2$$

This result provides us with Algorithm 2, which is computationally quite tractable. We first use the learned model $\widehat{\Pr}$ to define a new model class $\widehat{\Pr}_\alpha$, which scales $\widehat{\Pr}$ by an exponential distribution over the weighted 1-step lookahead entropy $H(\widehat{W}_{t+1}|w_{\le t})$. Then, similar to Algorithm 1, we simply fit the single parameter α to minimize the cross-entropy of the new model with respect to $\Pr$, which fixes the entropy amplification in the resulting model $\widehat{\Pr}_\alpha$. We observe this empirically in Figure 2 – our calibration results in a perplexity drop of almost 20 points over long-term generations under an LSTM model. Model and implementation details are in the supplementary material.

Generations from a calibrated model. Table 2 provides sample generations from a calibrated Transformer model trained on the GBW dataset, compared to its original version. Qualitatively, the calibrated generations: (1) are shorter and more concise, and (2) display a better grasp of discourse structure across sentences. More generations are provided in the supplementary material.

Original model	Calibrated model
Actual results could differ materially from those indicated by these forward-looking statements as a result of various important factors , including , without limitation : changes in general economic and business conditions , including more difficult real estate environments ; **[...174 tokens...]** risks related to investigations by other companies ; inadequate information systems ; the impact of reduced availability of ; * assumptions upon such companies using such as ours to gauge CNET 's financial condition ; and other factors .	*Actual results could differ materially from those indicated by* these forward-looking statements as a result of a variety of factors , including but not limited to (i) the risk that the tender offer could close in one or more manner or at all ; (ii) risks associated with conducting business in foreign jurisdictions ; (iii) difficulties in combining some or all of the businesses under one roof ; (iv) decreased demand for electricity , natural gas and other energy products , including adverse effects on the pricing of oil and natural gas ; and (v) the risks associated with doing business internationally .
Bluepoint Games , Inc. is a highly experienced and multi-faceted publisher of licensed virtual worlds for gamers , developers and technology professionals . **[...114 tokens...]** James Upon , CEO of MyNetSheltetWeb and the three previous Developers of MySQL . Based in Redwood City , California , BlueMountain is the leader in franchise and game development for the massively multiplayer online game .	*Bluepoint Games , Inc. is a highly experienced* licensing , gaming and entertainment firm focused on developing the next generation of casual games based on the PlayStation (R) BRAVIA family of video game machines for the North American market . Bluepoint is a wholly owned subsidiary of Bluehill ID Holdings L.P.

Table 2: Sample generations from a calibrated, state-of-the-art Transformer model trained on the GBW corpus, seeded with prefixes of sentences (in italics) from the holdout validation set.

5. Calibration and Memory

Defining a notion of memory in language models is challenging, and multiple equally sensible notions may coexist. Here we present our choice from first principles. Let us say that $\widehat{W}_t$ is a sample from a model at time t, i.e. $\widehat{W}_t \sim \widehat{\Pr}(W_t|W_{<t})$. Let us also assume that $W_{<t} \sim \Pr(W_{<t})$. We will define the memory at gap τ as the mutual information between $\widehat{W}_t$ and the distant past (those words greater than τ steps ago) conditioned on the subsequence $W_{t-\tau:t-1}$. Precisely,

$$\begin{aligned} I_\tau &:= I(\widehat{W}_t; W_{<t-\tau}|W_{t-\tau:t-1}) \\ &= H(\widehat{W}_t|W_{t-\tau:t-1}) - H(\widehat{W}_t|W_{<t}) \,, \end{aligned}$$

where we are not explicitly denoting the t dependence in this definition[2].

Intuitively, I_t can be viewed as how much uncertainty (entropy) in the prediction W_t the model is able to reduce by utilizing the deep past $W_{<t-\tau}$ in addition to the recent past $W_{t-\tau:t-1}$.

The difficulty in estimating this mutual information is due to estimating $H(\widehat{W}_t|W_{t-\tau:t-1})$, which requires the marginalized model $\widehat{\Pr}(W_t|W_{t-\tau:t-1})$. To (even approximately) marginalize a model distribution $\widehat{\Pr}(W_t|W_{<t})$ over the deep past $W_{<t-\tau}$ is computationally difficult, since it requires the access to a pool of samples of $W_{<t}$ that share a *common* recent past $W_{t-\tau:t-1}$. Nevertheless, we now show that it is possible to obtain an upper bound (which is computationally efficient to estimate).

Upper bounding mutual information using calibrated models. In the above, we were considering the mutual information between $\widehat{W}_t$ and $W_{<t-\tau}$ conditioned on $W_{t-\tau:t-1}$. Let us now consider a more general setting, where we have a distribution $\Pr(Z, Y, X)$ where Z, Y, and X are random variables. We wil eventually consider Z, Y, X to be $\widehat{W}_t, W_{t-\tau:t-1}W_{<t-\tau}$, respectively.

For distributions $\mathcal{D}(\cdot|Y, X)$ and $\widetilde{\mathcal{D}}(\cdot|Y, X)$ and for $\alpha \in \mathbb{R}$, define

$$\mathcal{D}_\alpha(Z|Y, X) := \mathcal{D}(Z|Y, X) \cdot \left(\widetilde{\mathcal{D}}(Z|Y, X)\right)^\alpha / Z_\alpha \,.$$

We say that $\mathcal{D}(\cdot|Y, X)$ is calibrated to $\tilde{\mathcal{D}}(\cdot|Y, X)$, if $\mathcal{D} = \mathcal{D}_{\alpha=0}$ is unimprovable in that for all α

$$\mathrm{CE}(\Pr \parallel \mathcal{D}) \le \mathrm{CE}(\Pr \parallel \mathcal{D}_\alpha) \,.$$

Note this condition is achievable due to that calibrating a model to $\widetilde{\mathcal{D}}(\cdot|Y, X)$ involves a one dimensional (convex) estimation problem (over α).

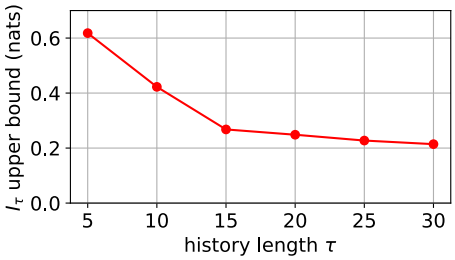
Theorem 5.1. *Suppose we have a model $\widehat{\Pr}(Z|X)$, and suppose $\widetilde{Z} \sim \widetilde{\Pr}(\cdot|X)$, where $\widetilde{Z}$ is dependent only on X. Suppose that $\widehat{\Pr}$ is calibrated to $\widetilde{\Pr}$. Then we have that:*

$$I(\widehat{Z}; X|Y) \le \mathrm{CE}(\Pr \parallel \widetilde{\Pr}) - H(\widehat{Z}|Y, X) \,, \text{ where:}$$

$$\mathrm{CE}(\Pr \parallel \widetilde{\Pr}) = \mathop{\mathbb{E}}_{Y\sim\Pr} \mathop{\mathbb{E}}_{Z\sim\Pr(\cdot|Y)} \left[\log \frac{1}{\widetilde{\Pr}(Z|Y)}\right] .$$

Memory estimation. We first learn another $\widetilde{W}_t \sim \widetilde{\Pr}(\cdot|W_{t-\tau:t-1})$, and then calibrate $\widehat{\Pr}$ to $\widetilde{\Pr}$.

[2]While we may attempt to estimate I_τ for a given t, we can remove the t dependence by either defining this quantity by with an average over t or by using appropriate stationarity assumptions. In our experiments, we average over t.



τ	$\mathrm{CE}(\Pr \,\|\|\, \widetilde{\Pr})$	I_τ upper bound	α^*
5	4.8144	0.6180	0.003515
10	4.5258	0.4226	-0.01041
15	4.4166	0.2678	-0.00447
20	4.3347	0.2485	-0.02268
25	4.2777	0.2274	-0.01814
30	4.2408	0.2143	-0.02323

Figure 3: *Top:* Plot of the upper bound on I_τ derived from calibrated models. *Bottom:* The measurements of the upper bound on mutual information, the cross entropy of the limited memory model $\widetilde{\Pr}$ as well as the optimal calibration coefficient α^* for various time lengths τ. Details of the model used here can be found in the supplementary material.

Corollary 5.2. *Suppose* $\widehat{\Pr}^{\mathrm{cal}}(\cdot|W_{<t})$ *is a model calibrated to* $\widetilde{\Pr}(\cdot|W_{t-\tau:t-1})$. *For a random variable,* $\widehat{W}_t^{\mathrm{cal}} \sim \widehat{\Pr}^{\mathrm{cal}}(\cdot|W_{<t})$, *we have that:*

$$I(\widehat{W}_t^{\mathrm{cal}}; W_{<t-\tau}|W_{t-\tau:t-1}) \leq \mathrm{CE}(\Pr \parallel \widetilde{\Pr}) - H(\widehat{W}_t^{\mathrm{cal}}|W_{<t}),$$

where $\mathrm{CE}(\Pr \parallel \widetilde{\Pr}) = \mathop{\mathbb{E}}_{W_{t-\tau:t} \sim \Pr} \left[\log \frac{1}{\widetilde{\Pr}(W_t|W_{t-\tau:t-1})} \right]$.

This corollary gives us a means to efficiently provide upper bounds on the mutual information. The key is that since $\widetilde{\Pr}$ is efficiently computable, we can directly estimate $\mathrm{CE}(\Pr \,||\, \widetilde{\Pr})$ through Monte Carlo estimation. We measure the upper bounds on I_τ of a LSTM model with trained limited-memory models $\widetilde{\Pr}$ (see details in the supplementary material) and report them in Figure 3. As expected, the memory estimate gradually decays with longer τ, indicating that the models make more use of the recent past to generate text.

6. Conclusion

We have introduced a calibration-based approach to detect and provably correct the discrepancies between the long-term generations of language models and the true distributions they estimate sequentially. In particular, for state-of-the-art neural language models, we have observed large degradations of the entropy rate under iterative generation, and a proposed first-order correction which is both computationally tractable and effective. Using the same calibration approach, we have derived estimators for the amount of information extracted by these models from the deep past.

Aside from the empirical findings and improvements, we hope that this work will inspire a more principled line of discourse on the quality of long-term generations in language models. It remains an interesting open problem to relate the plethora of "future-aware" generation-improving heuristics to our calibration framework.

Acknowledgements

MB is supported in part by the NSF Alan T. Waterman Award, Grant No. 1933331, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry. SK is supported by NSF CCF Grant No. 1637360. KN is supported by the Princeton SEAS Innovation Grant.

References

Antoniol, G., Brugnara, F., Cettolo, M., and Federico, M. Language model representations for beam-search decoding. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 588–591. IEEE, 1995. 4.2

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*, 2016. 2

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015. 2

Brown, L. D. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986. ISBN 0-940-60010-2. 2

Clarkson, P. and Robinson, T. Towards improved language model evaluation measures. In *Sixth European Conference on Speech Communication and Technology*, 1999. 2

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal*

Processing). Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954. A

Csiszar, I. and Körner, J. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011. 4.1

Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Language modeling with longer-term dependency. 2018. 1, 2

Dawid, A. P. The well-calibrated bayesian. *Journal of the Am. Stat. Assoc*, 77, 1982. 1, 2

Dawid, A. P. The impossibility of inductive inference. *Journal of the Am. Stat. Assoc*, 80, 1985. 1, 2

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

Fedus, W., Goodfellow, I., and Dai, A. M. Maskgan: better text generation via filling in the_. *arXiv preprint arXiv:1801.07736*, 2018. 2

Foster, D. P. Prediction in the worst case. *Annals of Statistics*, 19, 1991. 1, 2

Genzel, D. and Charniak, E. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 199–206, 2002. 2

Gong, C., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Frage: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, pp. 1341–1352, 2018. 1

Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016. 1, 2

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017. 1, 2

He, T., Zhang, J., Zhou, Z., and Glass, J. Quantifying exposure bias for neural language generation. *arXiv preprint arXiv:1905.10617*, 2019. 2

Ji, Y., Cohn, T., Kong, L., Dyer, C., and Eisenstein, J. Document context language models. *arXiv preprint arXiv:1511.03962*, 2015. 2

Jost, U. and Atwell, E. Proposal for a mutual-information based language model. In *Proceedings of the 1994 AISB Workshop on Computational Linguistics for Speech and Handwriting Recognition*. AISB, 1994. 2

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. 1, C

Kalai, E., Lehrer, E., and Smorodinsky, R. Calibrated forecasting and merging. *Games and Economic Behavior*, 29, 1999. 2

Ke, N. R., Goyal, A., Bilaniuk, O., Binas, J., Mozer, M. C., Pal, C., and Bengio, Y. Sparse attentive backtracking: Temporal credit assignment through reminding. In *Advances in Neural Information Processing Systems*, pp. 7651–7662, 2018. 1, 2

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 284–294, 2018. 2

Le, Q. V., Jaitly, N., and Hinton, G. E. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015. 1, 2

Lin, H. and Tegmark, M. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017. 2

Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-T. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pp. 3155–3165, 2017. 2

Liu, L., Utiyama, M., Finch, A., and Sumita, E. Agreement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 411–416, 2016. 2

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of english: The penn treebank. 1993. C

McAllester, D. Information theoretic co-training. *arXiv preprint arXiv:1802.07572*, 2018. 2

Melis, G., Dyer, C., and Blunsom, P. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017. 1

Merity, S., Keskar, N. S., and Socher, R. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*, 2017. 1, C

Merity, S., Keskar, N. S., and Socher, R. An Analysis of Neural Language Modeling at Multiple Scales. *arXiv preprint arXiv:1803.08240*, 2018. C

Mikolov, T. and Zweig, G. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 234–239. IEEE, 2012. 2

Mroueh, Y. and Sercu, T. Fisher gan. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2513–2523. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6845-fisher-gan.pdf. 2

Mller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 06 1997. doi: 10.2307/1428011. 2

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632. ACM, 2005. 1, 2

Ortmanns, S. and Ney, H. Look-ahead techniques for fast beam search. *Computer Speech & Language*, 14(1):15–32, 2000. 4.2

Ott, M., Auli, M., Grangier, D., and Ranzato, M. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*, 2018. 2

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999. 1, 2

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018. 2

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. 1, 1, 2, C

Shannon, C. E. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951. 1

Sriperumbudur, B., Fukumizu, K., Gretton, A., Schlkopf, B., and Lanckriet, G. On integral probability metrics, phi-divergences and binary classification. 01 2009. 2

Steinbiss, V., Tran, B.-H., and Ney, H. Improvements in beam search. In *Third International Conference on Spoken Language Processing*, 1994. 4.2

Takahashi, S. and Tanaka-Ishii, K. Cross entropy of neural language models at infinitya new bound of the entropy rate. *Entropy*, 20(11):839, 2018. 2

Takase, S., Suzuki, J., and Nagata, M. Direct output connection for a high-rank language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4599–4609, 2018. 1

Trinh, T. H., Dai, A. M., Luong, T., and Le, Q. V. Learning longer-term dependencies in rnns with auxiliary losses. *arXiv preprint arXiv:1803.00144*, 2018. 1, 2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017a. 1, 2, C

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017b. 1

Vovk, V. Competitive on-line statistics. *International Statistical Review*, 69, 2001. 2

Wang, A. and Cho, K. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019. 2

Wang, T. and Cho, K. Larger-context language modelling with recurrent neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1319–1329, 2016. 2

Wang, W., Gan, Z., Wang, W., Shen, D., Huang, J., Ping, W., Satheesh, S., and Carin, L. Topic compositional neural language model. *arXiv preprint arXiv:1712.09783*, 2017. 2

Wiseman, S. and Rush, A. M. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*, 2016. 2

Xie, Z. Neural text generation: A practical guide. *arXiv preprint arXiv:1711.09534*, 2017. 2

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 2

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates, 2002. 1, 2

Zhang, Z., Wu, S., Liu, S., Li, M., Zhou, M., and Xu, T. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 443–450, 2019. 2