

Conformal Nucleus Sampling


Shauli Ravfogel^{1,2} Yoav Goldberg^{1,2} Jacob Goldberger¹

¹Bar-Ilan University ²Allen Institute for Artificial Intelligence

{shauli.ravfogel, yoav.goldberg}@gmail.com, jacob.goldberger@biu.ac.il

Abstract

Language models generate text based on successively sampling the next word. A decoding procedure based on nucleus (top- p) sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p . In this work, we assess whether a top- p set is indeed aligned with its probabilistic meaning in various linguistic contexts. We employ conformal prediction, a calibration procedure that focuses on the construction of minimal prediction sets according to a desired confidence level, to calibrate the parameter p as a function of the entropy of the next word distribution. We find that OPT models are overconfident, and that calibration shows a moderate inverse scaling with model size.

 <https://github.com/shauli-ravfogel/conformal-prediction>

1 Introduction

Modern language generation methods are all based on computing the conditional next-word distribution. However, there is still considerable debate about the best way to extract the next word from that distribution. Most current text generation methods employ one of a handful of standard decoding strategies, which are characterized as either deterministic or stochastic in nature. A greedy search strategy selects the word with the highest probability at each timestep. The greedy method and its beam search variations work remarkably well for machine translation but outside of this context, tend to return dull text or degenerate text (Holtzman et al., 2020; Cohen and Beck, 2019). Holtzman et al. (2020) argued that high-quality human language does not follow a pattern of highest-probability next words, as humans expect the generated text to not be repetitive or boring. The same problem occurs with beam search.

Direct sampling from the next-word distribution computed by the model often generates incoherent gibberish text. Temperature sampling (Ackley et al., 1985) is a word sampling approach based on rescaling logit scores before applying the softmax function to compute the word distribution. Other methods limit the sampling space to a small **prediction set** to avoid the “unreliable tail” (Holtzman et al., 2020). In top- k sampling (Fan et al., 2018), we sample only from the top- k most likely words. Instead of sampling only from the most likely k words, top- p (nucleus) sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p (Holtzman et al., 2020). Top- p sampling enables a dynamically sized window of words, unlike top- k which fixes the size of k for every step. Finally, locally typical sampling (Meister et al., 2022) and truncation sampling (Hewitt et al., 2022) are recent variants of top- p that aim to make it more suitable for language generation.

The top- p prediction set has a concrete probabilistic interpretation. Here we examine whether the probability that the “correct” word belongs to the set of words produced by the top- p algorithm is indeed p . More generally we expect that the next-word prediction would be calibrated, meaning that the output of the next-word softmax layer would accurately reflect the true word distribution. Parametric calibration methods, such as Temperature Scaling (Guo et al., 2017), which adjust the confidence of the most probable word, are not suitable for adjusting the size of the prediction set. Conformal Prediction (CP) (Vovk et al., 1999, 2005; Shafer and Vovk, 2008; Angelopoulos and Bates, 2021) is a non-parametric calibration method that, given a value p , aims to build a prediction set with a guarantee that the probability that the correct word is within this set is indeed p . Note that this notion of calibration, which is distinct from the way calibration is usually formulated in language mod-

eling settings, *exactly coincides* with the goal of the top- p prediction model. The model-agnostic and distribution-free nature of CP makes it particularly suitable for large neural network models. We thus applied CP analysis to assess whether the top- p procedure is calibrated and, if needed, tune it to have the desired probabilistic interpretation. We find that OPT models of different sizes (Zhang et al., 2022) are not calibrated according to the conformal prediction theory, and that calibration shows moderate inverse scaling. Additionally, we show that the degree of calibration varies significantly with the entropy of the model’s distribution over the vocabulary. We thus propose a new **Conformal top- p decoding** algorithm, which ensures that the top- p sampling has a meaningful probabilistic interpretation.

2 CP for Language Generation

In this section, we briefly review the Split Conformal Prediction algorithm (Vovk et al., 2005) and discuss its relevance to language generation models. Consider a network that classifies an input x into k pre-defined classes. The network (softmax layer) output has the mathematical form of a distribution. However, this does not necessarily mean that it accurately reflects the true class distribution.

Let (x, y) be a test instance and its corresponding class. We want to find a small subset of classes (a prediction set) $C(x) \subset \{1, \dots, k\}$ such that

$$p(y \in C(x)) \geq 1 - \alpha \quad (1)$$

where $1 - \alpha \in [0, 1]$ is a user-chosen error rate. (We use the term $1 - \alpha$ instead of p to comply with CP standard notation). In words, the probability that the set $C(x)$ contains the correct label is at least $1 - \alpha$. We call this property the marginal coverage since the probability is averaged over all the data points (x, y) . Denote the prediction set obtained by taking the most probable classes until the total mass just exceeds a value q , by $C_q(x)$. Let $\hat{q} \in [0, 1]$ be the smallest threshold value that $p(y \in C_{\hat{q}}(x)) \geq 1 - \alpha$. If $\hat{q} > 1 - \alpha$ the model can be viewed as over-confident. If $\hat{q} < 1 - \alpha$ the model can be viewed as under-confident and if $\hat{q} = 1 - \alpha$ the model is calibrated in the sense that the probability that the correct label is in the $1 - \alpha$ prediction set is indeed $1 - \alpha$.

If the model is not calibrated, we can calibrate it using a labeled validation set $(x_1, y_1), \dots, (x_n, y_n)$.

Denote $p_t(i) = p(y_t = i | x_t; \theta)$. Define the **conformal scores** to be:

$$s_t = \sum_{\{i | p_t(i) \geq p_t(y_t)\}} p_t(i) \quad t = 1, \dots, n \quad (2)$$

This CP score is known as the Adaptive Prediction Sets (APS) score, and was first introduced in (Romano et al., 2020). Note that $y_t \in C_{s_t}(x_t)$ and s_t is the minimal threshold in which the true class y_t is in a prediction set of x_t .

We next look for a **minimal threshold** \hat{q} such that the correct label y_t is included in the prediction set $C_{\hat{q}}(x_t)$ for at least $(1 - \alpha)n$ points of the validation set. In other words, \hat{q} calibrates the top- $(1 - \alpha)$ prediction-set on the validation set. We can easily find \hat{q} by first sorting the n scores s_1, \dots, s_n and then \hat{q} is the $(1 - \alpha)$ -quantile of the validation-set scores. Once the network is calibrated, if we want to form a prediction set for a new test sample x , that contains the true class with probability $(1 - \alpha)$, we use $C_{\hat{q}}(x)$. The CP Calibration procedure for calibrating the top- p word decoding is summarized in Algorithm 1. The conformal prediction theory provides the following guarantee on the threshold \hat{q} (Vovk et al., 2005).

Theorem: Assume a test point (x, y) and the n validation points are independent and identically distributed (or at least exchangeable). Let \hat{q} be the $\lceil (n + 1)(1 - \alpha)/n \rceil$ -quantile of the validation set scores. Then

$$1 - \alpha \leq p(y \in C_{\hat{q}}(x)) \leq 1 - \alpha + \frac{1}{n + 1}. \quad (3)$$

Note that this is a marginal probability over all the test points and is not conditioned on a given input. Exchangeability means that the sequence distribution is not altered by permuting the order of the random variables.

In this study, we aim to apply the conformal prediction framework to language generation models to analyze the prediction sets used for sampling the next word. The joint distribution of words in a text is neither IID nor exchangeable, since the words are correlated and the order of the words in a sentence is significant. A recent study (Oliveira et al., 2022) showed that applying the usual CP algorithm to a stationary β -mixing process (rather than an exchangeable one) results in a guaranteed coverage level of $1 - \alpha - \eta$, where η depends on the mixing properties of the process and is theoretically hard to know, or bound. Roughly speaking, β -mixing processes are stochastic processes in which far-away

Algorithm 1 CP Calibration of the Top- p decoding

Input: A validation set comprised of next word distributions p_1, \dots, p_n with the corresponding correct words y_1, \dots, y_n and a confidence level p .

for $t = 1, \dots, n$ **do**

$$s_t = \sum_{\{i | p_t(i) \geq p_t(y_t)\}} p_t(i)$$

end for

Define \hat{q} to be the $\lceil (n+1)p/n \rceil$ -quantile of $\{s_1, \dots, s_n\}$.

Output: Use top- \hat{q} decoding to guarantee that the probability that the correct word is in the top- \hat{q} prediction set is at least p .

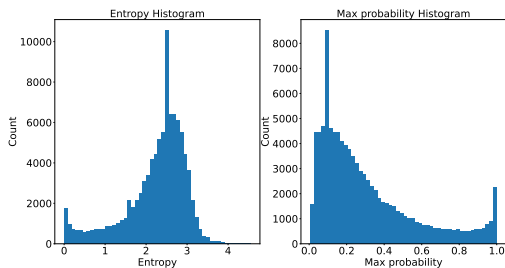


Figure 1: Histograms of entropy of the output probability distribution for the OPT350M model.

points are approximately independent in a quantifiable manner. In all the examples they checked, the authors assessed that the additional penalty incurred by using CP with stationary β -mixing processes was virtually insignificant. Manning and Schutze (1999) argue that even though not quite correct, natural language can be modeled as stationary, ergodic processes. Khandelwal et al. (2018) showed that the LSTM language model’s memory is empirically bounded at roughly 200 words and thus the model can be viewed as an aperiodic recurrent (and therefore β -mixing) Markov chain. It is reasonable to assume that human language and transformer-based language models can also be modeled as β -mixing processes. Hence, applying CP to language generation models yields meaningful results (at least qualitatively).

3 Experiments

In this section, we apply the conformal prediction calibration method to analyze the calibration status of the top- p nucleus sampling.

Setup. We experimented with variants—from 125M parameters up to 30B parameters—of OPT (Zhang et al., 2022), a left-to-right language model. We ran the models on 10,000 English Wikipedia

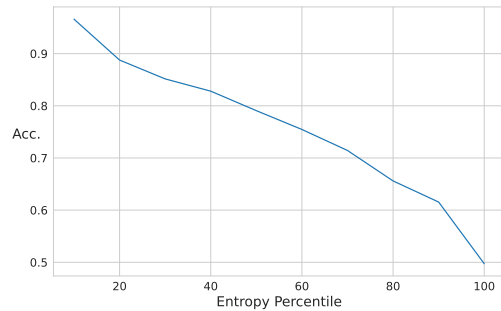


Figure 2: Effective accuracy when using nucleus sampling with $p = 0.9$, for different entropy percentiles, for the OPT350M model.

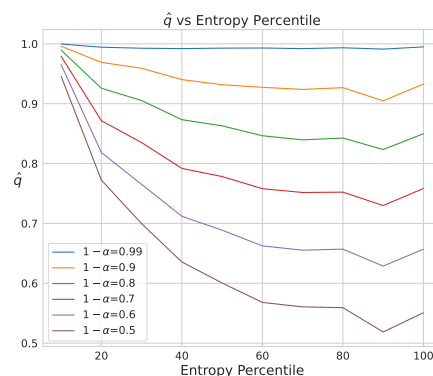


Figure 3: \hat{q} threshold scores when calibration is performed over the examples belonging to each entropy percentile separately.

sentences¹, and collected the distribution of the vocabulary over each token in each sentence, resulting in a total of 245,923 distributions. The distribution of the entropy values, as well as the maximum probability, was far from being uniform (Fig. 1). We sorted all the instances by entropy, and calibrated the examples belonging to each equally-sized percentile independently (from 0-10% to 90-100%). The patterns are highly similar across models. We report results on the 350M parameters model unless specified otherwise. We use Nvidia 2080TI GPUs.

Dependency of the confidence on the entropy.

First, we evaluated the confidence scores of a standard nucleus sampling scheme. We chose $p = 0.9$ (a commonly used value) and recorded the effective confidence, i.e., the proportion of cases where the correct word was indeed in the top- p prediction set. Fig. 2 shows the effective confidence for the

¹<https://huggingface.co/datasets/wikipedia>

predictions belonging to different percentiles of entropy. The results indicated that setting $p = 0.9$ did not translate to a prediction set that contained the correct token in 90% of the cases, motivating our calibrated decoding. In Fig. 3, we show the per-entropy CP calibration results, for 10 entropy bins corresponding to percentiles. While the model was always overconfident, the level of overconfidence decreases with the entropy percentile. In other words, when the model is apparently the most certain—as reflected in low entropy values—it is most overconfident. Note that in the case of low entropy the single highest probability can be more than 0.9. Hence, there is no way to calibrate the prediction set by changing its size. In particular, we found that the model is overconfident when the gold token is a function word: it tends to allocate high probability to a small set of function words, while the true distribution is more varied.

Calibration and scale. Fig. 4 presents the conformal threshold values \hat{q} versus desired confidence $(1-\alpha)$, when calibration is performed over the entire validation set (without partition to entropy bins). As shown, for all confidence levels, the threshold \hat{q} needed to ensure that the correct word is included within the prediction set is larger than the confidence level itself (the $y = x$ dashed line). This indicates that the model is *overconfident*. Fig. 4 also shows the dependency of calibration on the scale. Scaling language models has been shown to induce the emergence of new abilities, such as in-context learning (Brown et al., 2020). Empirical power laws were shown to predict performance in a different task as a function of scale (Kaplan et al., 2020; Wei et al., 2022a), where models usually show improved performance with scale. Here, we find *inverse scaling* (Wei et al., 2022b), where calibration moderately deteriorates with model scale.

Generation. How does conformal p sampling affect generation? we use the 350M model to compare the quality of generation of conformal p sampling with the natural baseline of p sampling. We generate continuations to 1,000 prompts of size 35 words from the OpenWebText dataset². We generate up to length 200 tokens, and compare conformal $p = 0.9$ prediction (setting $1 - \alpha = 0.9$) with conventional $p = 0.9$ sampling.³ Following Fig. 3, when applying our method, we calculate the

²<https://github.com/jcpeterson/openwebtext>

³We make the generations available at [this link](#).

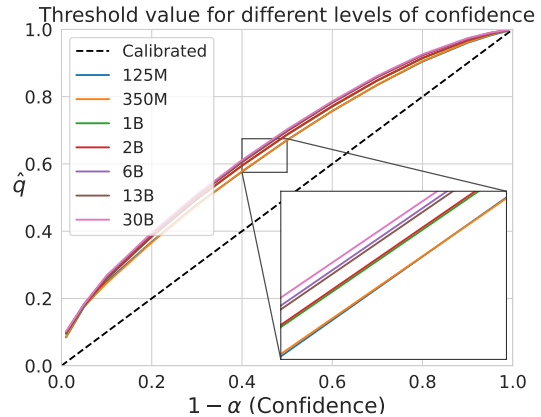


Figure 4: \hat{q} threshold values needed to ensure a confidence of $1-\alpha$. The OPT models show slight inverse scaling with respect to calibration.

entropy of the output distribution over each token, and dynamically set the threshold p for each token prediction, according to the threshold value \hat{q} that fits this entropy percentile. This ensures that the *true probability* of the token to be included within the prediction set (according to the training set used for calibration) is 0.9.

We evaluate the quality of the generation using MAUVE (Pillutla et al., 2021) and BERTScore (Zhang et al., 2019).⁴ MAUVE score is 0.933 for conformal- p sampling, and 0.920 for conventional p sampling. As for BERTScore, the $F1$ score is 0.840 for conformal- p sampling, and 0.843 for conventional p sampling. These results indicate that conformal- p sampling is performing similarly to conventional p sampling.

Applicability of CP to non IID data Conformal prediction theory assumes IID, while we build on the model outputs distributions over consecutive tokens in the same sentence, which are of course highly dependent. We repeated the per-entropy-bin calibration process when uniformly sampling a *single* token per sentence, thus (almost) satisfying the independence assumption. The results were similar to Fig. 3 and in that case, Eq. (3)) is applicable.

4 Conclusions

To conclude, in this study we apply the notion of calibration by conformal prediction to calibrate the top- p nucleus sampling as a function of the next word distribution entropy and thus made the top- p decoding policy consistent. The same analysis and

⁴Default HuggingFace v4.22.0 Parameters were used.

calibration can also be applied to other commonly used decoding methods, such as variants of top- p (Meister et al., 2022) and truncation sampling (Hewitt et al., 2022).

Limitations

We calibrated OPT models based on Wikipedia data. Future work should apply calibration procedure to a wider range of datasets, to check whether our results generalize to different domains. Additionally, we limited our evaluation to entropy as a measure of uncertainty and did not explore other measures. Finally, we aimed at validating the calibration status of commonly used LMs. Future work should thoroughly evaluate the impact of the calibration status on different facets of generation quality, as text generation is one of the main use-cases of large LMs.

Ethics Statement

We do not foresee ethical issues with this work.

Acknowledgements

This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT). Shauli Ravfogel is grateful to be supported by the Bloomberg Data Science Ph.D. Fellowship.

References

- David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.
- Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *International Conference on Machine Learning (ICML)*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*.
- John Hewitt, Christopher D Manning, and Percy Liang. 2022. Truncation sampling as language model smoothing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- 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 (ICLR)*.
- 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*.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Christopher Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press.
- Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *arXiv preprint arXiv: 2202.00666*.
- Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. 2022. Split conformal prediction for dependent data. *arXiv preprint arXiv:2203.15885*.
- 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. *Advances in Neural Information Processing Systems*, 34:4816–4828.
- Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*.
- Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.
- Volodya Vovk, Alexander Gammerman, and Craig Saunders. 1999. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Yi Tay, and Quoc V Le. 2022b. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
"Limitations"
- A2. Did you discuss any potential risks of your work?
We do not foresee risks from this work.
- A3. Do the abstract and introduction summarize the paper's main claims?
I
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C Did you run computational experiments?

Left blank.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Not applicable. 3

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Not applicable. Left blank.

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Not applicable. Left blank.

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.