Conformal Autoregressive Generation: Beam Search with Coverage Guarantees

Nicolas Deutschmann, Marvin Alberts, María Rodríguez Martínez

IBM Research

deu@zurich.ibm.com, marvin.alberts@ibm.com, mrm@zurich.ibm.com

Abstract

We introduce two new extensions to the beam search algorithm based on conformal predictions (CP) to produce sets of sequences with theoretical coverage guarantees. The first method is very simple and proposes dynamically-sized subsets of beam search results but, unlike typical CP procedures, has an upper bound on the achievable guarantee depending on a *post-hoc* calibration measure. Our second algorithm introduces the conformal set prediction procedure as part of the decoding process, producing a variable beam width which adapts to the current uncertainty. While more complex, this procedure can achieve coverage guarantees selected *a priori*. We provide marginal coverage bounds for each method, and evaluate them empirically on a selection of tasks drawing from natural language processing and chemistry.

Introduction

Autoregressive sequence models (ARSM) are probabilistic models describing distributions of sequence data, among which transformer-based large language models (LLMs) (Vaswani et al. 2023) have emerged as highly powerful tools for natural language processing (NLP) and generation (NLG). Beyond NLP, application of deep ARSMs in the natural sciences have generated impactful advances, such as in chemistry (Born and Manica 2023; Schwaller et al. 2019) and molecular biology (Rives et al. 2021).

Despite their success, generating reliable predictions with ARSMs and quantifying their uncertainty combines the general challenges of trustworthy predictions in deep learning (Guo et al. 2017) with issues arising from the combinatorially large prediction space and the iterative processes used for sequence generation (Gleave and Irving 2022; Malinin and Gales 2021; Xiao et al. 2022). LLMs also present the unique further dimension of linguistic calibration (Kadavath et al. 2022; Mielke et al. 2022; Yin et al. 2023). Nevertheless, the high-profile of LLMs and the potential scientific value of ARSMs on other data modalities make the improvement of reliable uncertainty estimation methods an important and active area of research (Fannjiang et al. 2022; Fomicheva et al. 2020; Gleave and Irving 2022; Jiang et al. 2021; Kuhn, Gal, and Farquhar 2023; Lin, Trivedi, and Sun 2023; Schuster et al. 2022).

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A particularly interesting framework to infuse a notion of uncertainty or error control in predictions is Conformal Predictions (CP), which produces dynamically-sized *sets* of predictions with finite-sample, distribution-free coverage guarantees (Vovk, Gammerman, and Shafer 2005). Beyond the formal guarantees, well-calibrated conformal sets are expected to reflect uncertainty through their size.

Generative tasks, where predicting is significantly harder than verifying a prediction seem like a perfect use case for the guarantees of CP. However, its original formulation relies on exhaustive searches that are untractable for sequences. Until recent methods relying on sampling (Quach et al. 2023) or human-assisted pruning (Ren et al. 2023), solutions only existed at the token level (Dey et al. 2022; Ravfogel, Goldberg, and Goldberger 2023).

While sampling is particularly well-suited to NLG (Holtzman et al. 2020), other applications that search correct or optimal solutions are better served by maximization-oriented decoding algorithms such as greedy or beam searches. Our work proposes to fill the current absence of such conformal decoding algorithms by introducing two new greedy methods for which we prove coverage guarantees and supported by emprical evaluation.

Related Work

Conformal Predictions and Tolerance Regions Our work builds on top of the extensive literature developing CP, a set of tools for uncertainty estimation and trustworthy prediction first introduced by Vovk (2012); Vovk, Gammerman, and Saunders (1999); Vovk, Gammerman, and Shafer (2005). This foundational work builds on the much earlier distribution-free tolerance regions developed by Tukey (1947); Wald (1943); Wilks (1941) whose iterative construction is the basis of our step-by-step calibration method. In recent years, much activity has been spent developing CP to tackle modern machine learning (Angelopoulos et al. 2022a; Lei and Wasserman 2012). Although not directly used in this paper, conformal risk control is of particular interest for potential extensions of our work (Angelopoulos et al. 2022b, 2023). Angelopoulos and Bates (2022) provide an excellent review of foundations and recent developments.

Uncertainty Estimation for Sequence Models Given the high visibility of large language models and their unique po-

sition as powerful ML models directly in the hands of the general public, there is a strong interest in developing methods to ensure that their predictions are trustworthy and reliable, and to quantify the degree of uncertainty of model outputs. Current models tend to formulate wrong statements with an authoritative tone (Kadavath et al. 2022; Mielke et al. 2022; Yin et al. 2023), and measures of confidence built from model internals also tend to under-estimate errors (Desai and Durrett 2020; Vasconcelos et al. 2023; Xiao et al. 2022), motivating the development of improved uncertainty measures and calibration methods (Jiang et al. 2021; Kuhn, Gal, and Farquhar 2023; Malinin and Gales 2021; Schuster et al. 2022).

Conformal Predictions for Sequence Models The complexity of sequence generation make applications of CP nontrivial. As a result, early work on conformal prediction for ARSMs focused on applications such as few-shot classification (Fisch et al. 2021a,b) or single-token predictions (Desai and Durrett 2020; Ravfogel, Goldberg, and Goldberger 2023). Recently however, two exciting new methods proposed by Ren et al. (2023) and Quach et al. (2023) improved the state of the art significantly with solutions to handle general sequences. From a generation viewpoint, the work of Ren et al. (2023) focusing on task planning is the closest to ours. A key difference is that our procedure generates sets of sequences instead of a product of sets of tokens, which reflects the domain disparity: robots must actually *take* a single actions while we can generate multiple sequences. We also use step-wise thresholds to reflect the shifts in probability distributions through the decoding process, which is less relevant for planning as the environment is determined by more than the actions taken. Conversely, the decoding procedure of Quach et al. (2023) relies on sampling and semantic criteria for sequence-level acceptance and rejection based on Learn-then-Test (Angelopoulos et al. 2022b) and is therefore quite different from our token-per-token approach. Theirs is thus better suited for the complex but fuzzy criteria of many NLG tasks. Both methods produce sets of full sequences with coverage guarantees, but our predictions are targeted toward controlling type-I errors on one-to-one – or very tight – input-sequence relationships while theirs aim at capturing a plurality of valid answers.

Background

Autoregressive Sequence Generation

Autoregressive models are a class of probabilistic models that describe a conditional probability mass function (PMF) $\pi(S|X)$ over a set of sequences $S \in A^*\omega$, *i.e.* finite sequences over an alphabet A, terminated by a disjoint token ω, given a condition X. The overall PMF π is obtained as a decomposition into next-token probabilities $\tilde{\pi}$ over the extended alphabet $\bar{A} = A \cup \{\omega\}$ conditioned on the right-truncated subsequences of $S = s_1 \dots S_N\omega$:

$$\pi(S) = \tilde{\pi}\left(\omega|s_1 \dots s_N\right) \prod_{k=1}^{N} \tilde{\pi}\left(s_k|s_1 \dots s_{k-1}\right). \quad (1)$$

Deep learning models popular in NLP and beyond such as transformers (Vaswani et al. 2023) and iterations of recurrent neural networks (Hochreiter and Schmidhuber 1997) provide learnable functions $\tilde{\pi}$ that have had resounding success in recent years.

The access to $\tilde{\pi}$ is sufficient to tractably sample from π: starting from an empty sequence, next tokens are sampled from $\tilde{\pi}$ until ω is reached. However, obtaining a *prediction* $S(X)$ in the typical machine learning sense,

$$S(X) = \underset{S' \in A^*\omega}{\text{argmax}}\ \pi\left(S'|X\right), \quad (2)$$

is not a solved problem and is usually addressed with heuristic methods. The simplest such approach is greedy search where the candidate sequence is extended with the most-probable next token. Greedy approaches fail if an early high-probability token only leads to lower-probability sequences and are therefore often extended using beam-search (BS) approaches where b candidate sequences are considered and greedily extended, keeping the b highest-scoring extensions at the next step.

The lack of theoretical grounding for these algorithms has two implications for generation-as-prediction tasks:

- there is no rigorous way to produce *the* prediction of the model for a given condition X (equation (2)),
- given a heuristic proposal, the *correctness* of the prediction cannot easily be studied from a theoretical point of view.

In this work, we address exclusively the second issue: given a pair (X, S) sampled from the true distribution p, we leverage conformal predictions to provide a greedy set of candidate sequences from π with a guarantee on the probability that S is included.

Conformal Predictions

Conformal predictions are a framework to produce sets of predictions $C(X)$ from a predictive model and a measure of confidence and uncertainty $s(X, y)$.

Our method falls under the umbrella of Split CP (Papadopoulos et al. 2002), where we consider a pre-trained predictive model, a prediction confidence score function $s(X, y)$ and exchangeable data $\{(X_i, y_i)\}_{i=1,\dots N+1}$. Considering the set $\mathcal{C}_N$ of the first N samples as calibration data and choosing a risk level α, we define a threshold t_α for scores as the $k_\alpha^{(N)}$-th smallest calibration score, where

$$k_\alpha^{(N)} = \lfloor \alpha(N+1) \rfloor. \quad (3)$$

Defining a prediction set $C(X_{N+1})$ for the $(N+1)$-th sample as

$$C(X_{N+1}) = \{y' \in \Omega_Y |\, s(X, y') \geq t_\alpha\}, \quad (4)$$

where Ω_Y is the space of possible labels y.

Marginally over the calibration set, the following inequality holds true:

$$\mathbb{P}(y_{N+1} \in C(X_{N+1})) \geq 1 - \alpha, \quad (5)$$

Methods

We present two methods for obtaining greedy conformal prediction sets from autoregressive models: the first is based on proposing a dynamically-sized subset of the usual beam search results, while the second uses CP to choose a beam size dynamically at each step.

Conformal Predictions From Beam Search Results

Let us start with the simplest approach, where we rely on the standard beam search algorithm.

Given a pair (X, S) sampled from $P(S, X)$, we wish to produce a set $C(X)$ of candidate sequences with some form of guarantee on $\mathbb{P}(S \in C(X))$. The combinatorially-large sequence space makes a direct application of equation (4) untractable. An alluring alternative would be to start by performing a beam search, in order to obtain a reasonably-sized set of proposals $\beta(X)$, and to then predict $C(X) \subset B(X)$. This however, only works if we can provide a coverage guarantee on $\beta(X)$, which we do below.

The procedure itself relies on group-conditional conformal predictions (Vovk 2012). Using the notation from **??** , we define the in-beam subgroup of the calibration data $\mathcal{C}_N$:

$$\mathcal{C}_\beta = \{(X_i, S_i) \in \mathcal{C}_N | S_i \in \beta(X_i)\}\,. \tag{6}$$

Defining $N_\beta = |\mathcal{C}_\beta|$, we perform split-CP calibration on this subgroup, with a confidence score $s(X_i, S_i)$ such as the ARMS probability $\pi(S_i|X_i)$. We thus obtain a threshold $t_\alpha^{(N_\beta)}$ as the $k_\alpha^{(N_\beta)}$-th smallest score as defined in **??** .

At inference time, given a test sample (X, S), we define the prediction set

$$C_{\alpha|\beta}(X) = \{S' \in \beta(X) |\, \pi(S'|X) \geq t_\alpha\}\,. \tag{7}$$

On this prediction set, we provide the following guarantee:

Proposition 1
With probability at least $1 - \delta$*,*

$$\mathbb{P}\big(S \in C_{\alpha|\beta}(X)\big) \geq (1-\alpha)\, B(\delta; N_\beta, N + 1 - N_\beta). \tag{8}$$

where $B(a, b)$ *is a beta distribution and* $B(\delta; a, b)$ *is its* δ*-quantile.*

This result follows from the conditional decomposition of the coverage probability:

$$\mathbb{P}\big(S \in C_{\alpha|\beta}(X)\big) = \mathbb{P}\big(S \in C_{\alpha|\beta}(X) | S \in \beta(X)\big) \times \mathbb{P}(S \in \beta(X)). \tag{9}$$

The conditional conformal procedure described above guarantees that the first term is at least $1 - \alpha$, following from Vovk (2012). Unlike conformal guarantees, the second bound is obtained from the observation of N_β, which follows a binomial distribution. While we cannot decide its success probability, we can use confidence intervals from Clopper and Pearson (1934) to provide a bound with risk δ.

Dynamic Conformal Beam Search

In this section, we present our second method to provide conformal prediction sets for ARSM. This approach relies on choosing a dynamic beam size at each decoding set based on a CP threshold, which permits a pre-determined guarantee.

Calibration Algorithm We consider N_0+1 exchangeable pairs $\{(X_i, S_i)\}$ and a family of conformal scores σ_l that can be evaluated on length-l sequences. Selecting the first N_0 samples as $\mathcal{C}_{N_0}^{(0)}$, we specify a per-step confidence level $1 - \alpha$ calibrate iteratively as follows: at the l-th step,

1. Define $k_\alpha^{(l)} = \lfloor (N_{l-1} + 1)\alpha \rfloor$.
2. Order the calibration set by increasing length-l scores $\sigma_l\big(X_1, S_{1|l}\big) \geq \ldots \sigma_l\big(X_{N_{l-1}}, S_{N_{l-1}|l}\big)$, where $S_{i|l}$ is the length-l truncation of S_i.
3. Define $t_{\alpha,N_{l-1}}^{(l)} = \sigma_l\Big(X_{k_\alpha^{(l)}}, S_{k_\alpha^{(l)}|l}\Big)$.
4. Set $N_l = N_{l-1} - k_\alpha^{(l)}$, $\mathcal{C}_{N_l}^l = \{(X_i, S_i)\}_{k_\alpha^{(l)} < i}$.

The iteration can in-principle be continued infinitely if σ_l is based on the model $\tilde{\pi}$ and we extend $\tilde{\pi}$ defined in **??** by specifying that $\tilde{\pi}$ predicts a padding token $\mathfrak{d} \in A$ with probability 1 after the terminating token ω.

We define the set of acceptable length-L sequences as $\Omega_L(A, \omega, \mathfrak{d}) = \big[\bar{A}^* \cdot \omega^? \cdot \mathfrak{d}^*\big]_L$, *i.e.* sequences of the form $a_1 \ldots a_k \omega^? \mathfrak{d} \ldots \mathfrak{d}$ with $0 \leq k \leq L$, an optional terminating omega ω and $L - k - 1$ padding tokens.

Inference Algorithm At inference, we consider the left-out sample X_{N_0+1}, S_{N_0+1}, dropping the index for brevity. The first decoding step is a standard conformal prediction on length-one sequences:

$$C_\alpha^{(1)}(X) = \left\{ S'_{|1} \in A \;\middle|\; \sigma_1 X, S'_{|1} \geq t_{\alpha,N_0}^{(1)} \right\}. \tag{10}$$

Proceeding iteratively until all sequences in $C_\alpha^{(l)}(X)$ terminate or a maximum length $L^\dagger$ is reached, we define the next conformal beam as all continuations in the previous beam that pass the next threshold

$$C_\alpha^{(l+1)}(X) = \left\{ S'_{|l}\, a \;\middle|\; \begin{array}{l} a \in \bar{A},\ S'_{|l} \in C_\alpha^{(l)}(X), \\ \sigma_{l+1}\Big(X, S'_{|l} a\Big) \geq t_{\alpha,N_l}^{(l+1)} \end{array} \right\}. \tag{11}$$

This approach mirrors the traditional beam-search algorithm in that it keeps a set of proposals at each decoding step and generates a set of high-scoring continuations of the current proposal for the next step.

Guarantees The iterative subselection procedure that defines the subsequent conformal thresholds $t_{\alpha,N_l}^{(l)}$ defines a multivariate tolerance region as defined in Wald (1943) and Tukey (1947) which provide distribution-free coverage guarantees. Indeed, our thresholds correspond to iteratively pruning $\mathcal{C}_{N_0}^{(0)}$ by removing the $k_\alpha^{(l)}$-lowest scoring element based on a scoring function applied to the samples. Our set of thresholds defines a confidence region $R\Big(\mathcal{C}_{N_0}^{(0)}, \alpha, L\Big)$ as

$$R\Big(\mathcal{C}_{N_0}^{(0)}, \alpha, L\Big) = \Big\{ S' \in \Omega_L(A, \omega, \mathfrak{d}) \mid \forall 0 \leq l < L \quad \sigma_{l+1}\Big(X, S'_{|l} a\Big) \geq t_{\alpha,N_l}^{(l+1)} \Big\}. \tag{12}$$

Its conditional coverage probability given a calibration set $\mathcal{C}_{N_0}^{(0)}$ is a random variable, which, strikingly, depends only on α and N_0:

$$\mathbb{P}\Big((X,S) \in R\Big(\mathcal{C}_{N_0}^{(0)}, \alpha, L\Big)\Big|\mathcal{C}_{N_0}^{(0)}\Big) \sim B\left(N_0 + 1 - \sum_l k_\alpha^{(l)}, \sum_l k_\alpha^{(l)}\right). \quad (13)$$

Marginalizing over $\mathcal{C}_{N_0}^{(0)}$, we obtain

Lemma 2

$$\mathbb{P}\left(\left\{\sigma_{l+1}(X,S) \geq t_{\alpha,N_l}^{(l+1)}\right\}_{l\leq L}\right) = 1 - \sum_{l=1}^{L} \frac{k_\alpha^{(l)}}{N_0+1} \quad (14)$$

$$\geq (1-\alpha)^L. \quad (15)$$

Importantly, notice that since the threshold at step l is applied exhaustively on all sequences that passed the previous threshold, the prediction set defined in equation (11) corresponds to all sequences S' that verify the condition of equation (15) given X:

$$C_\alpha^{(L)}(X) = \left\{S' \in \Omega_L(A,\omega,\ə) \,\middle|\, \left\{\sigma_{l+1}(X,S) \geq t_{\alpha,N_l}^{(l+1)}\right\}_{l\leq L}\right\}. \quad (16)$$

Extending the decoding until a maximum length $L^\dagger$, we obtain the guarantee that for $(X,S) \sim P(X,S)$ exchangeable with the calibration set,

Proposition 3

$$\mathbb{P}\left[S \in C_\alpha^{(L^\dagger)}(X)\right] \geq (1-\alpha)^{L^\dagger}. \quad (17)$$

Experimental Results

Datasets and Models

We evaluate our conformal generation algorithms on two transformer models, each evaluated on a generative task that squarely fits in the category of predictions, rather than distribution modelling, in the sense that there is a single *correct* sequence given an input X.

Integer Additions We finetune an off-the-shelf `t5-base` sequence-to-sequence model from HuggingFace on a simple addition task: computing the sum of two integers with up to seven digits each. The problem is formulated as a sequence-to-sequence prediction task, matching arithmetic problems of the form "`1789+111=`" to results presented as a sequence of digits followed by an EOS token, "`1900`ω". We sample 130k such additions which we split into 100k training examples and 30k held out samples. On the validation set, this model has a mean coverage of 96% using 5-sequence beam search. This is a short-sequence task with a high performance model, making it a best-scenario test case for our dynamic conformal decoding algorithm.

Chemical Reaction Product Prediction We train a `t5-small` from scratch to predict the product of chemical reactions, using SMILES strings Weininger (1988) to encode reagents and products. We use the USPTO-MIT dataset (Jin et al. 2017) and tokenization scheme by Schwaller et al. (Schwaller et al. 2019) for training and evaluation, holding out 30k samples with length lower than 50 for calibration and testing. On the validation set, this model has a mean coverage of 64% using 5-sequence beam search. In contrast to the additions task, this task features much longer sequences and a lower-accuracy model, making it a more challenging testing ground.

Conformal Beam Subsets

We start by evaluating the success rate $1-\delta$ of the composite bound of proposition 1. To this end, we perform 1000 bootstrapped estimates, sampling twice 15k held-out samples for calibration and test. For each repetition, we measure N_β, evaluate the Clopper-Pearson guarantee and perform the conditional-in beam conformal calibration. On test data, we evaluate the conditional and global coverage and evaluate the success rate of the inequality in proposition 1 over the repetitions. As we show in table 1, the bound does hold with sufficiently low risk. For all experiments, we use the length-normalized sequence probability under the model $\pi(S|X)/|S|$ as the conformal confidence score.

While not guaranteed, a desirable feature of conformal prediction sets is that their size is informative of some notion of uncertainty. If predictions set sizes closely follow the rank of the correct sequence in the beam, ordered by scores, they can be used as a measure of the quality of the prediction. We measure this by considering the mean absolute error between the predicted set size and that of a perfect oracle with set sizes exactly matching the true sequence rank if it is in the beam, and predicting the entire beam if not. We also report this information in table 1. While the additions model provide tight prediction sets, the reaction prediction model seems poorly calibrated. As we nevertheless show in figure 1, the model is indeed over-conservative for low-rank sequences but for both models set size is a useful predictor of low versus high rank when taking rank distribution into account.

Dynamic Conformal Beams

Let us now turn to our second procedure where we use conformal predictions to iteratively decode with guaranteed coverage at each step. For our benchmark tasks, we exploit a simplification of our dataset definitions: we know the maximum sequence length in advance, respectively 5 and 50 tokens for additions and chemical reactions. We use this knowledge to set the maxumum number of decoding steps to 5 and 50 and avoid discussing rare long sequences. We discuss how to handle unknown maximum lengths in **??** .

For each benchmark task, we run multiple independent calibration-and-inference experiments, randomly sampling 30% of the held-out data for calibration and 1500 test sequences. Metrics are measured for each repetition and averaged, and we report the standard deviation across the repetitions. The smaller scale of the additions task allowed us to

Task	Beam Size	Beam Cov.	$1-\alpha = 0.95$			
			Conditional Cov.	MAE	Global Bound	Global Cov.
Additions (T5)	10	0.979(9)	0.949(2)	0.73(2)	0.910(1)	0.912(2)
	5	0.961(1)	0.950(2)	0.45(1)	0.911(1)	0.913(2)
Reactions (T5)	10	0.699(2)	0.949(2)	4.48(7)	0.658(2)	0.663(3)
	5	0.641(3)	0.949(3)	1.86(4)	0.602(3)	0.608(4)
Task	Beam Size	Beam Cov.	$1-\alpha = 0.99$			
			Conditional Cov.	MAE	Global Bound	Global Cov.
Additions (T5)	10	0.979(9)	0.989(1)	2.43(7)	0.947(1)	0.951(1)
	5	0.961(1)	0.989(1)	1.33(3)	0.948(1)	0.952(2)
Reactions (T5)	10	0.699(2)	0.989(1)	6.69(7)	0.683(3)	0.692(2)
	5	0.641(3)	0.989(1)	2.82(3)	0.625(1)	0.634(3)

Table 1: We report the conditional coverage and mean absolute error between the prediction set size and the rank of the correct sequence in the beam on 1000 bootstrapped experiments. The uncertainty on the last significant digit is given as the standard deviation across the experiments.

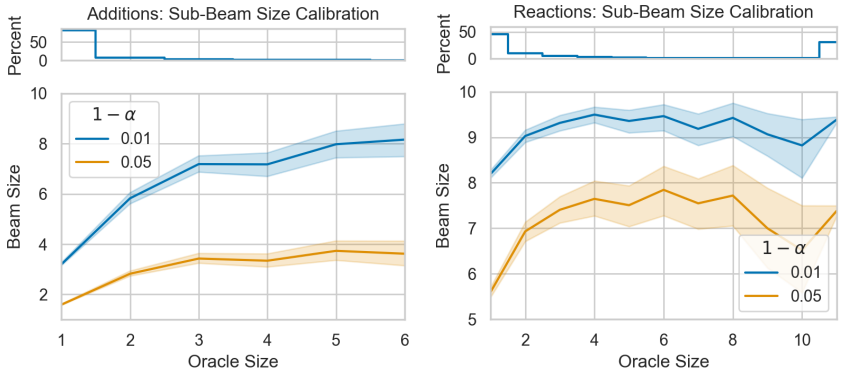

Figure 1: (Main) Conformal sub-beam set sizes plotted against oracle set sizes for the addition task (left) and the chemical reaction prediction task (right). (Top) Oracle set size distribution.

run more experiments and we therefore have 120 independent repetitions for step-wise guarantees levels $(1-\alpha) \in [0.99, 0.98, 0.95]$ yielding sequence-level marginal coverage guarantees of $[0.95, 0.90, 0.77]$. Predicting chemical reaction products is much more computationally demanding owing to the longer sequences and and larger model, as well as a lower accuracy, leading to larger prediction sets. As a result we limit ourselves to 50 repetitions for two token-wise confidence levels $(1-\alpha) \in [0.995, 0.99]$, which are more strict to enable coverage control after 50 decoding steps. These token-wise values define sequence-wise guarantees of $[0.78, 0.60]$, which frame the 5-beam validation accuracy of the model of 0.64.

As for the fixed-beam procedure, we start by empirically checking our coverage guarantees. We report observed coverages in table 2, showing that the mean observed coverage indeed dominates the guarantee. For lower confidence levels, the bounds tend to be less tight, owing to the lower frequency of longer-than needed decoding procedures.

Table 2 also reports mean beam sizes at the end of decoding $|\beta(X)|$ and compares them to an optimal oracle that predicts a subset $\beta_O(X)$ of our dynamic beams such that the lowest-ranking sequence is the correct one. While it was appropriate to use MAE for the fixed beam size method, the dynamic version does not have a reference size to which we could compare and we therefore found ratios more informative.

As show in figure 2, conditional coverage is not ensured per sequence length. While coverage is near-uniform across reaction product lengths, there is significant variation across addition sequence lengths.

We characterize the adaptivity of the prediction set sizes by observing whether the beams are tight in the sense than larger beams are predicted are indicative of low-ranking true sequences, which we show to be the case in figure 3. Beam size is indeed much more predictive of oracle size than for the conformal beam subset procedure, making dynamic conformal beam decoding preferable in terms of uncertainty quantification.

One disadvantage of dynamic conformal beams is the absence of a bound on beam sizes, which can become large if the model doesn't have high performance, or sequences are long: either case requires a stringent per-token confidence level. This is indeed what we observe in the reaction predic-

Task	$1-\alpha$	Coverage	Guarantee	$\lvert\beta(X)\rvert$	$\frac{\lvert\beta(X)\rvert}{\lvert\beta_O(X)\rvert}$
Additions	0.990	0.9607(3)	0.9509	10.4(4)	6.64(2)
	0.980	0.9244(4)	0.9039	4.82(1)	3.30(1)
	0.950	0.8231(6)	0.7737	1.756(3)	1.476(2)
Reactions	0.995	0.806(1)	0.778	47.(5)	25.(2)
	0.990	0.693(1)	0.605	4.13(1)	3.08(1)

Table 2: Sequence-level metrics for the dynamic beam decoding procedure on the additions and chemical reaction tasks measured across respectively 120 and 50 repetitions. For each dataset and tested confidence level, we report the mean coverage and compare it to the guarantee of proposition 3, as well as the mean beam size $\mathbb{E}\,|\beta(X)|$ and the mean beam-size to oracle beam-size ratio $\mathbb{E}\frac{|\beta(X)|}{|\beta_O(X)|}$. For each metric, we report the average and the uncertainty on the last digit measured as the standard deviation of the mean estimator between parentheses.

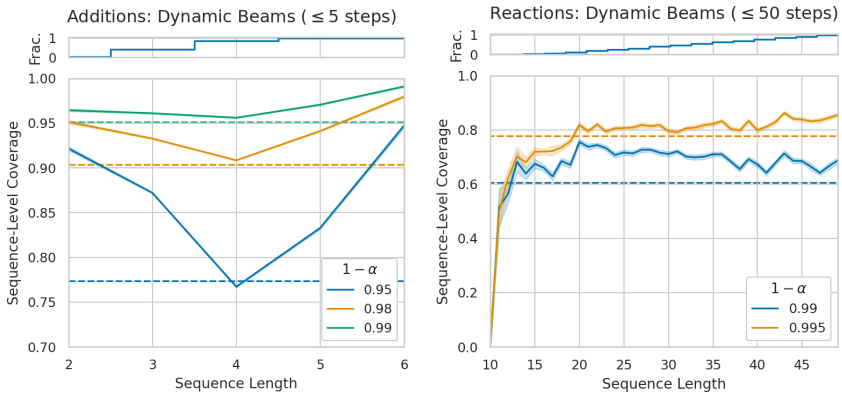

Figure 2: Per-sequence-length observe coverages for the additions (left) and reactions (right) tasks. Coverage guarantees are provided as dashed lines for each confidence level.

tion task with $\alpha = 0.005$. As we show in figure 4, a significant fraction of beams contain more than 100 sequences, which is both computationally expensive and potentially undesirable, in the sense that many incorrect sequences are predicted.

Limitations and Usage Recommendations

Our two methods have the potential to further the confident and uncertainty-aware use of generative sequence models. Nevertheless, they have clear limitations that must be made explicit for both practitioners and potential future development.

- Most importantly, both methods rely on greedy search for high-scoring sequences. This is especially adapted for autoregressive models used as predictors over sequences, which is common in the sciences, but less so for language modelling except maybe in the case of reward optimization for reinforcement-learning-based models (Gleave and Irving 2022).
- As a hybrid conformal-estimation method, conformal beam subsets are not a prescriptive way to obtain a certain desired coverage, unlike traditional conformal methods. Only a fraction of the observed calibration coverage is attainable. Given that requiring too-high guarantees can lead to very large sets, this might be an acceptable compromise.
- Our dynamic conformal beam decoding procedure provides a per-decoding step guarantee which degrades exponentially with the number of steps executed. Choosing a low per-step risk might be acceptable for high-accuracy models but might become unwieldy for more difficult tasks.
- Dynamic beam guarantees also rely on a choice of maximum length which is an extra hyper-parameter to set and can affect performance: a too low cut-off will lead to failures for long sequences while the opposite choice degrades performance for shorter sequences.

Consequently, we make the following recommandations for practical use of our methods:

Method Choice For high-accuracy models on moderately sized models, using dynamic beams is probably the best-suited method. However, for tasks or models yielding reduced performance measured on fixed-size beams, one should consider whether the computational cost of very large dynamic beams is worth the benefits. As a conservative rule of thumb, we recommend to use fixed 5-sequence beam coverage as a reference for achievable dynamic performance

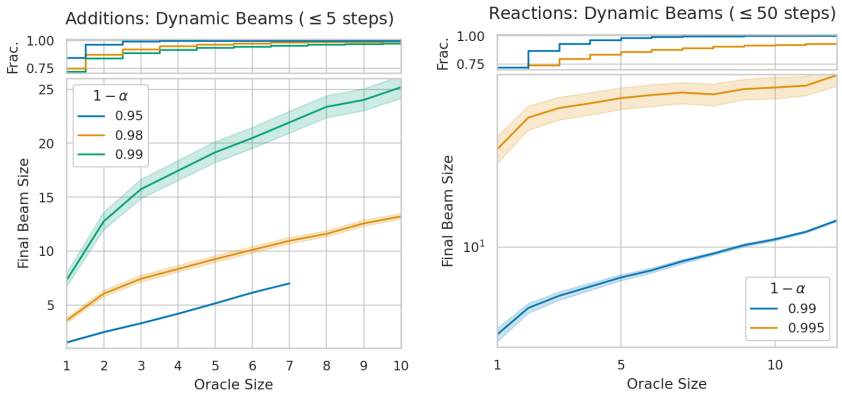

Figure 3: (Main) Dynamic conformal beam sizes plotted against oracle set sizes for the addition task (left) and the chemical reaction prediction task (right). (Top) Oracle set size distribution. A high correlation between oracle and beam sizes is indicative of good calibration.

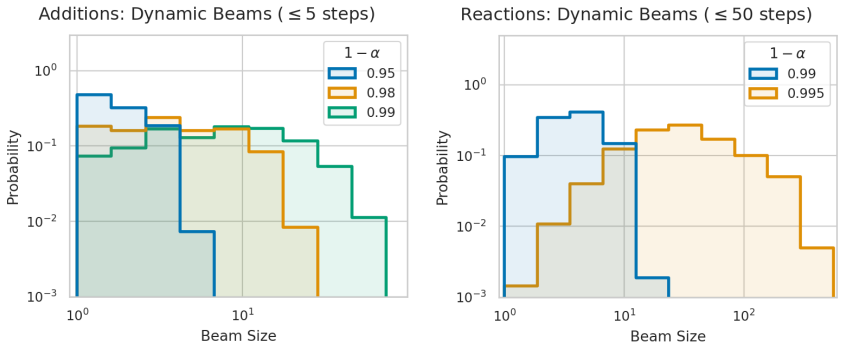

Figure 4: Final beam size distributions for the additions (left) and reactions (right) tasks.

with dynamic sizes up to the tens. Its applicability is of course model and task-dependent and a score distribution with a long tail should motivate the use of fixed beams instead.

Choice of Maximum Decoding Length Our dynamic conformal beam decoding procedure relies on setting a maximum decoding length which affects the guarantee exponentially. In our experiments, we leveraged the knowledge of an actual maximum length in the dataset to set this limit but in many cases this knowledge is not accessible. Given that we cannot expect to evaluate uncertainties using a per-step procedure on sequences for which we have no or limited calibration data, setting this limit based on a high quantile of the observed length distribution is a safe choice. This limit should be set using other data than the calibration set to avoid breaking exchangeability, but the model training data is perfectly suitable for this purpose.

Making Dynamic Guarantees Tighter With Length-Conditional Calibration As shown in table 2, a significant fraction of sequences with decoding procedures shorter than the maximum can lead to less-than-tight guarantees. If the length distribution is spread, splitting sequences into groups of similar lengths and using the group-conditional calibration procedure introduced by Vovk (2012) is well-advised. At decoding time, a two-step procedure can be used: first decoding with the minimum threshold across all groups until termination. At this point, a group can be assigned to the decoded beam and pruned based on the group-specific thresholds. Each length group has the coverage guarantee of its longest sequences.

Conclusion

The two methods we introduce in this paper further the applicability of split-conformal prediction sets with distribution-free guarantees to autoregressive sequence generation tasks. Our approach complements the recent work of Quach et al. (2023) based on sampling by using a greedy decoding approach which makes it better suited for predictive tasks. Extending our approach to proper language modelling would nevertheless be interesting, especially for models trained or fine-tuned with reinforcement learning (Ouyang et al. 2022). Another exciting further development would be to use our method to restrict the sampling options of Quach et al. (2023), which might improve both performance and reduce its rejection rate (Cohen and Beck 2019; Holtzman et al. 2020).

Acknowledgements

We thank Jannis Born, Dimitrios Christofidellis, Mattia Rigotti and Alain Vaucher for helpful discussions and advice. The work of ND was supported by the Swiss National Science Foundation Grant No. 192128.

References

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2022a. Uncertainty Sets for Image Classifiers Using Conformal Prediction. arxiv:2009.14193.

Angelopoulos, A. N.; and Bates, S. 2022. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arxiv:2107.07511.

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2022b. Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control. arxiv:2110.01052.

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2023. Conformal Risk Control. arxiv:2208.02814.

Born, J.; and Manica, M. 2023. Regression Transformer Enables Concurrent Sequence Regression and Generation for Molecular Language Modelling. *Nature Machine Intelligence*, 5(4): 432–444.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4): 404–413.

Cohen, E.; and Beck, C. 2019. Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models. In *Proceedings of the 36th International Conference on Machine Learning*, 1290–1299. PMLR.

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295–302. Online: Association for Computational Linguistics.

Dey, N.; Ding, J.; Ferrell, J.; Kapper, C.; Lovig, M.; Planchon, E.; and Williams, J. P. 2022. Conformal Prediction for Text Infilling and Part-of-Speech Prediction. *The New England Journal of Statistics in Data Science*, 1(1): 69–83.

Fannjiang, C.; Bates, S.; Angelopoulos, A. N.; Listgarten, J.; and Jordan, M. I. 2022. Conformal Prediction for the Design Problem. *Proceedings of the National Academy of Sciences*, 119(43): e2204569119.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021a. Efficient Conformal Prediction via Cascaded Inference with Expanded Admission. arxiv:2007.03114.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021b. Few-Shot Conformal Prediction with Auxiliary Tasks. arxiv:2102.08898.

Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; and Specia, L. 2020. Unsupervised Quality Estimation for Neural Machine Translation. arxiv:2005.10608.

Gleave, A.; and Irving, G. 2022. Uncertainty Estimation for Language Reward Models. arxiv:2203.07472.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. arxiv:1706.04599.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. arxiv:1904.09751.

Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. arxiv:2012.00955.

Jin, W.; Coley, C.; Barzilay, R.; and Jaakkola, T. 2017. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *Advances in Neural Information Processing Systems*, volume 30.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. arxiv:2207.05221.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. arxiv:2302.09664.

Lei, J.; and Wasserman, L. 2012. Distribution Free Prediction Bands. arxiv:1203.5422.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. arxiv:2305.19187.

Malinin, A.; and Gales, M. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. arxiv:2002.07650.

Mielke, S. J.; Szlam, A.; Dinan, E.; and Boureau, Y.-L. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. arxiv:2203.02155.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive Confidence Machines for Regression. In Elomaa, T.; Mannila, H.; and Toivonen, H., eds., *Machine Learning: ECML 2002*, Lecture Notes in Computer Science, 345–356. Berlin, Heidelberg: Springer. ISBN 978-3-540-36755-0.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal Language Modeling. arxiv:2306.10193.

Ravfogel, S.; Goldberg, Y.; and Goldberger, J. 2023. Conformal Nucleus Sampling. arxiv:2305.02633.

Ren, A. Z.; Dixit, A.; Bodrova, A.; Singh, S.; Tu, S.; Brown, N.; Xu, P.; Takayama, L.; Xia, F.; Varley, J.; Xu, Z.; Sadigh, D.; Zeng, A.; and Majumdar, A. 2023. Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. arxiv:2307.01928.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; and Fergus, R. 2021. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.

Schuster, T.; Fisch, A.; Gupta, J.; Dehghani, M.; Bahri, D.; Tran, V. Q.; Tay, Y.; and Metzler, D. 2022. Confident Adaptive Language Modeling. arxiv:2207.07061.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9): 1572–1583.

Tukey, J. W. 1947. Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions–The Continuous Case. *The Annals of Mathematical Statistics*, 18(4): 529–539.

Vasconcelos, H.; Bansal, G.; Fourney, A.; Liao, Q. V.; and Vaughan, J. W. 2023. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. arxiv:2302.07248.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arxiv:1706.03762.

Vovk, V. 2012. Conditional Validity of Inductive Conformal Predictors. arxiv:1209.2673.

Vovk, V.; Gammerman, A.; and Saunders, C. 1999. Machine-Learning Applications of Algorithmic Randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, 444–453. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-612-8.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. New York: Springer. ISBN 978-0-387-00152-4 978-0-387-25061-8.

Wald, A. 1943. An Extension of Wilks' Method for Setting Tolerance Limits. *The Annals of Mathematical Statistics*, 14(1): 45–55.

Weininger, D. 1988. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31–36.

Wilks, S. S. 1941. Determination of Sample Sizes for Setting Tolerance Limits. *The Annals of Mathematical Statistics*, 12(1): 91–96.

Xiao, Y.; Liang, P. P.; Bhatt, U.; Neiswanger, W.; Salakhutdinov, R.; and Morency, L.-P. 2022. Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis. arxiv:2210.04714.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? arxiv:2305.18153.