# Conformal Language Modeling

Mircea Petrache – UC Chile

ReLeLa, 4 July 2023

CENIA
CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Instituto de
Ingeniería Matemática
y Computacional
UC

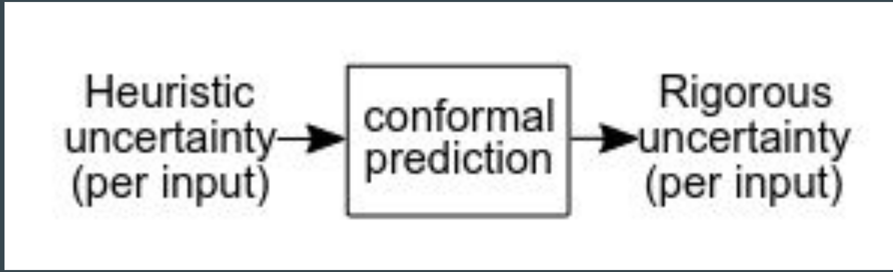
# Plan of talk:

1. Conformal Prediction
2. Learn Then Test
3. Conformal Language Modeling
4. Discussion

# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification” Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model



# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model

Ingredients:

- Data (X1,Y1), … , (Xn, Yn) sampled n times (calibration set)
- Score function on the data (can be anything) → s(X,Y)

Output:

- For given X’ test, gives a set C(X’) to which output belongs with **(guaranteed) high probability**

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model

Ingredients:

- Data (X1,Y1), … , (Xn, Yn) sampled n times (calibration set)
- Score function on the data (can be anything) → s(X,Y)

Output:

- For given X' test, gives a set C(X') to which output belongs with **(guaranteed) high probability**

$$1 - \alpha \leq \mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1}$$

# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model

Ingredients:

- Data (X1,Y1), … , (Xn, Yn) sampled n times (calibration set)
- Score function on the data (can be anything) → s(X,Y)

Output:

- For given X’ test, gives a set C(X’) to which output belongs with **(guaranteed) high probability**

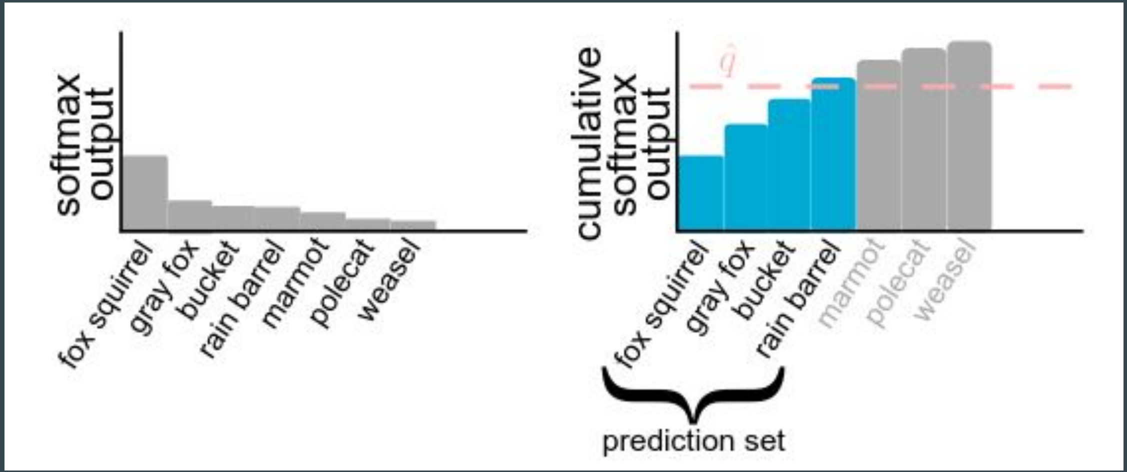
$$1 - \alpha \leq \mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}))$$

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification" Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model

Concrete Example:
(**score=softmax output**)



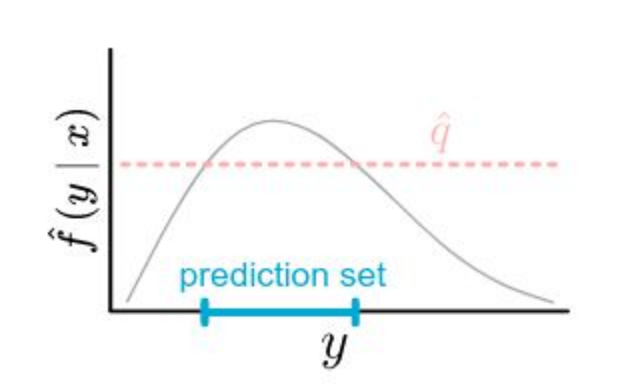
# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

- Idea: **add confidence intervals to predictions** made by a model

Bayesian Example:
(**score = estimated posterior**)

$$s(x, y) = -\hat{f}(y \mid x).$$



# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Summary of the technique:

1. Identify a heuristic notion of uncertainty using the pre-trained model.
2. Define the score function $s(x, y) \in \mathbb{R}$. (Larger scores encode worse agreement between $x$ and $y$.)
3. Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores $s_1 = s(X_1, Y_1), ..., s_n = s(X_n, Y_n)$.
4. Use this quantile to form the prediction sets for new examples:

$$\mathcal{C}(X_{test}) = \{y : s(X_{test}, y) \leq \hat{q}\}.$$

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

- Theorem

**Theorem 1** (Conformal coverage guarantee; Vovk, Gammerman, and Saunders [5]). *Suppose* $(X_i, Y_i)_{i=1,\ldots,n}$ *and* $(X_{\text{test}}, Y_{\text{test}})$ *are i.i.d. and define* $\hat{q}$ *as in step 3 above and* $\mathcal{C}(X_{\text{test}})$ *as in step 4 above. Then the following holds:*

$$P\Big(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\Big) \geq 1 - \alpha.$$

# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

## Real life situation: classification task

Calibration time, we get:
- Xi sampled
- get scores for all Y
- we KNOW correct Yi

Test time, we get:
- X’,
- scores of all Y’

WE WANT TO
- select subset of the Y’
- get probabilistic guarantees

**Main assumption: EXCHANGEABILITY:**
**score histogram for calibration, still “true” for test X’**

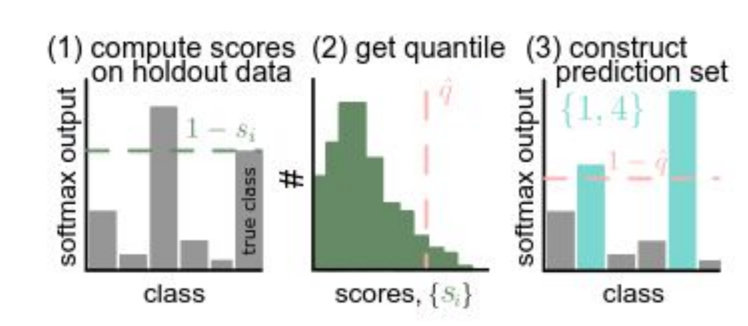
# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

## Real life situation: classification task

Calibration time, we get:

- Xi sampled
- get scores for all Y
- we KNOW correct Yi




Test time, we get:

- X’,
- scores of all Y’

use histogram to get Y’ set probabilistic guarantees

Main assumption: EXCHANGEABILITY:
SCORE HISTOGRAM for calibration, still “true” for test X’

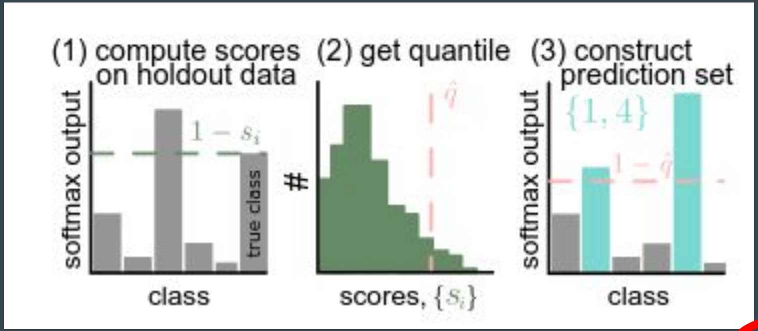
# 1. Conformal Prediction - intro paper

“A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”
Angelopoulos Bates 2021 (arxiv link)

## Real life situation: classification task

Calibration time, we get:

- Xi sampled
- get scores for all Y
- we know correct Yi




Test time, we get:

- X’,
- scores of all Y’

use histogram to get Y’ set
probabilistic guarantees

Main assumption: EXCHANGEABILITY:
score histogram for calibration, still “true” for test X’

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification" Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage
- Outlier detection
- Prediction under covariate shift

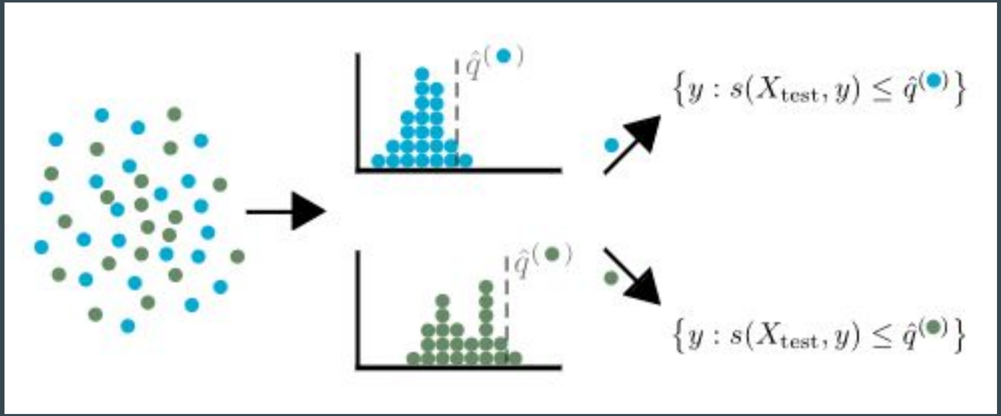
$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}) \mid X_{\text{test},1} = g\right) \geq 1 - \alpha,$$

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage
- Outlier detection
- Prediction under covariate shift



# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction $\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}) \mid Y_{\text{test}} = y\right) \geq 1 - \alpha,$
- Risk instead of coverage
- Outlier detection
- Prediction under covariate shift

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage $$\mathbb{E}\Big[\ell(\mathcal{C}(X_{\text{test}}), Y_{\text{test}})\Big] \leq \alpha,$$
- Outlier detection
- Prediction under covariate shift

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage
- Outlier detection

$$\mathbb{P}\left(\mathcal{C}(X_{\text{test}}) = \text{outlier}\right) \leq \alpha,$$

- Prediction under covariate shift

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- **Group coverage**
- **Class-conditional prediction**
- **Risk instead of coverage**
- **Outlier detection**
- **Prediction under covariate shift**

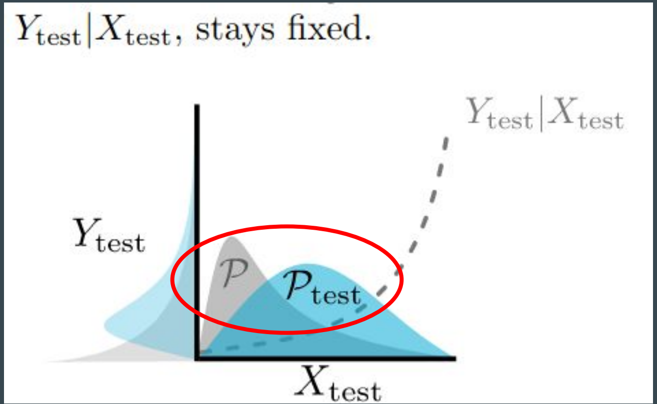
  *You are trying to predict diseases from MRI scans. You conformalized on a balanced dataset of 50% infants and 50% adults, but in reality, the frequency is 5% infants and 95% adults.*

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage
- Outlier detection
- Prediction under covariate shift



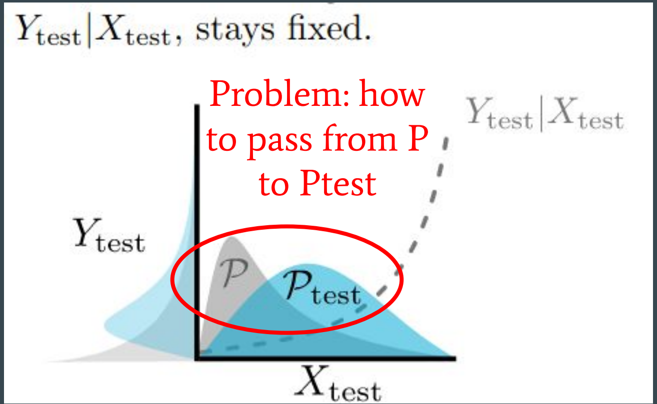

*You are trying to predict diseases from MRI scans. You conformalized on a balanced dataset of 50% infants and 50% adults, but in reality, the frequency is 5% infants and 95% adults.*

# 1. Conformal Prediction - intro paper

"A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification"
Angelopoulos Bates 2021 (arxiv link)

## Extensions

- Group coverage
- Class-conditional prediction
- Risk instead of coverage
- Outlier detection
- Prediction under covariate shift




*You are trying to predict diseases from MRI scans. You conformalized on a balanced dataset of 50% infants and 50% adults, but in reality, the frequency is 5% infants and 95% adults.*

# 2. Learn Then Test

“Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control” Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

# 2. Learn Then Test

“Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control” Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Input:

- pretrained model
- n random correct training pairs
- Risk function
- Parameter-dependent set-valued predictor

# 2. Learn Then Test

“Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control” Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Input:
- pretrained model
- n random correct training pairs
- Risk function
- Parameter-dependent set-valued predictor

Desired output:
- Parameters (randomized)
- Guarantee that predictor correct with high probability

# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Input:

- pretrained model
- n random correct training pairs
- Risk function
- Parameter-dependent set-valued predictor

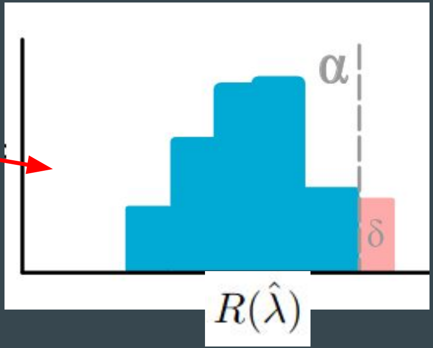
example

$$R(\mathcal{T}_\lambda) = \mathbb{E}\Big[\underbrace{L(\mathcal{T}_\lambda(X_{\text{test}}), Y_{\text{test}})}_{\text{Loss function}}\Big]$$

Desired output:

- Parameters (randomized)
- Guarantee that predictor correct with high probability

$\hat{\lambda}$ *be a random variable*



# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Family-wise Error Rate: $\text{FWER}\left(\widehat{\Lambda}\right) = \mathbb{P}\left(\exists \hat{\lambda} \in \widehat{\Lambda} : R(\hat{\lambda}) > \alpha\right)$

# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)

**But also, appendix A of intro paper (previous slide)**

Family-wise Error Rate: $\mathrm{FWER}\left(\widehat{\Lambda}\right) = \mathbb{P}\left(\exists \hat{\lambda} \in \widehat{\Lambda} : R(\hat{\lambda}) > \alpha\right)$

p-value: $\forall t \in [0,1], \ \ \mathbb{P}_{\mathcal{H}_\lambda}\left(p_\lambda \le t\right) \le t,$ where $\mathcal{H}_\lambda : R(\lambda) > \alpha$

# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Family-wise Error Rate: $\text{FWER}\left(\widehat{\Lambda}\right) = \mathbb{P}\left(\exists \hat{\lambda} \in \widehat{\Lambda} : R(\hat{\lambda}) > \alpha\right)$

p-value: $\forall t \in [0,1], \;\; \mathbb{P}_{\mathcal{H}_\lambda}(p_\lambda \leq t) \leq t,$ where $\mathcal{H}_\lambda : R(\lambda) > \alpha$

If we take $\widehat{\Lambda} = \{\lambda : p_\lambda < \delta\}$, then $\text{FWER}(\widehat{\Lambda}) = 1 - (1-\delta)^{|\Lambda|}$

# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Family-wise Error Rate: $\mathrm{FWER}\left(\widehat{\Lambda}\right) = \mathbb{P}\left(\exists \hat{\lambda} \in \widehat{\Lambda} : R(\hat{\lambda}) > \alpha\right)$

p-value: $\forall t \in [0,1], \;\; \mathbb{P}_{\mathcal{H}_\lambda}\left(p_\lambda \leq t\right) \leq t,$ $\quad p_\lambda^{\text{Hoeffding}} = e^{-2n\left(\alpha - \widehat{R}(\lambda)\right)_+^2}$ example

If we take $\widehat{\Lambda} = \{\lambda : p_\lambda < \delta\}$, then $\mathrm{FWER}(\widehat{\Lambda}) = 1 - (1-\delta)^{|\Lambda|}$

# 2. Learn Then Test

"Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control" Angelopoulos, Bates, Candes, Jordan, Lei (arxiv link)
**But also, appendix A of intro paper (previous slide)**

Family-wise Error Rate: $\text{FWER}\left(\widehat{\Lambda}\right) = \mathbb{P}\left(\exists \hat{\lambda} \in \widehat{\Lambda} : R(\hat{\lambda}) > \alpha\right)$

$\widehat{\Lambda}_{\text{Bonferroni}} = \left\{\lambda \in \Lambda : p_\lambda \leq \frac{\delta}{|\Lambda|}\right\}$ example

p-value: $\forall t \in [0,1], \ \ \mathbb{P}_{\mathcal{H}_\lambda}(p_\lambda \leq t) \leq t,$

$p_\lambda^{\text{Hoeffding}} = e^{-2n\left(\alpha - \widehat{R}(\lambda)\right)_+^2}$ example

If we take $\widehat{\Lambda} = \{\lambda : p_\lambda < \delta\}$, then $\text{FWER}(\widehat{\Lambda}) = 1 - (1-\delta)^{|\Lambda|}$

# 3. Conformal Language Modeling

“Conformal Language Modeling” Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

It starts from “Learn Then Test”.

# 3. Conformal Language Modeling

“Conformal Language Modeling” Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay
(arxiv link)

It starts from “Learn Then Test”.

Summary:

1. **Sample.** A new candidate response $y$ is sampled from our language model.
2. **Accept or reject.** The sample $y$ is added to the growing output set, as long as it is diverse (e.g., maximum overlap with any other element is $\leq \lambda_1$) and confident (e.g., the LM likelihood is $\geq \lambda_2$).
3. **Stop or repeat.** Using a set-based scoring function, we check if the confidence in the current set is $\geq \lambda_3$. If it is, then we stop and return the current set. Otherwise we return to Step 1.

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

It starts from "Learn Then Test".

Summary + main differences

1. **Sample.** A new candidate response $y$ is sampled from our language model.
2. **Accept or reject.** The sample $y$ is added to the growing output set, as long as it is diverse (e.g., maximum overlap with any other element is $\leq \lambda_1$) and confident (e.g., the LM likelihood is $\geq \lambda_2$).
3. **Stop or repeat.** Using a set-based scoring function, we check if the confidence in the current set is $\geq \lambda_3$. If it is, then we stop and return the current set. Otherwise we return to Step 1.

**Also, it selects optimal splitting ("components") of the text**

# 3. Conformal Language Modeling

“Conformal Language Modeling” Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

- Empirical risk over calibration set – for fixed $\lambda = (\lambda_1, \lambda_2, \lambda_3)$

$$\widehat{R}_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda), \quad \text{where } L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i) : A_i(y) = 1\}.$$

“Is ***y*** a good enough output for ***Xi***?” - function

Calibration input i

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

- Empirical risk over calibration set – for fixed $\lambda = (\lambda_1, \lambda_2, \lambda_3)$

$$\widehat{R}_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda), \quad \text{where } L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i) \colon A_i(y) = 1\}.$$

- p-values (general result via concentration bounds)

**Lemma 4.1** (Binomial tail bound p-values). *Let $\widehat{R}_n(\lambda)$ be the empirical risk in Eq. (5), and let* $\mathrm{Binom}(n, \epsilon)$ *denote a binomial random variable with sample size $n$ and success probability $\epsilon$. Then*

$$p_\lambda^{\mathrm{BT}} := \mathbb{P}(\mathrm{Binom}(n, \epsilon) \leq n\widehat{R}_n(\lambda)) \tag{6}$$

*is a valid p-value for* $\mathcal{H}_\lambda \colon \mathbb{E}[L_{\mathrm{test}}(\lambda)] > \epsilon$.

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

- Empirical risk over calibration set – for fixed $\lambda = (\lambda_1, \lambda_2, \lambda_3)$

$$\widehat{R}_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda), \quad \text{where } L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i) \colon A_i(y) = 1\}.$$

- p-values (general result via concentration bounds)

# 3. Conformal Language Modeling

“Conformal Language Modeling” Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

**Algorithm 1** Conformal sampling with rejection

**Definitions:** $x$ is an input prompt, $\mathcal{F}$ is our set-based confidence function, $\mathcal{S}$ is our text similarity function, $\mathcal{Q}$ is our sample quality estimator, $\lambda$ is our threshold configuration, and $k_{\max}$ is our sampling budget. $p_\theta(y \mid x)$ is the conditional output distribution defined by our language model.

**function** SAMPLE($x$, $\mathcal{F}$, $\mathcal{S}$, $\mathcal{Q}$, $\lambda$, $k_{\max}$)

  $\mathcal{C}_\lambda \leftarrow \{\}$ ▷ Initialize an empty output set.

  **for** $k = 1, 2, \ldots, k_{\max}$ **do**

    $y_k \leftarrow y \sim p_\theta(y \mid x)$. ▷ Sample a new response.

    **if** $\mathcal{Q}(x, y_k) < \lambda_2$ **then** ▷ Reject if its estimated quality is too low.

      **continue**

    **if** $\max\{\mathcal{S}(y_k, y_j) \colon y_j \in \mathcal{C}_\lambda\} > \lambda_1$ **then** ▷ Reject if it is too similar to other samples.

      **continue**

    $\mathcal{C}_\lambda = \mathcal{C}_\lambda \cup \{y_k\}$. ▷ Add the new response to the output set.

    **if** $\mathcal{F}(\mathcal{C}_\lambda) \geq \lambda_3$ **then** ▷ Check if we are confident enough to stop.

      **break**

  **return** $\mathcal{C}_\lambda$

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

**Algorithm 1** Conformal sampling with rejection

**Definitions:** $x$ is an input prompt, $\mathcal{F}$ is our set-based confidence function, $\mathcal{S}$ is our text similarity function, $\mathcal{Q}$ is our sample quality estimator, $\lambda$ is our threshold configuration, and $k_{\max}$ is our sampling budget. $p_\theta(y \mid x)$ is the conditional output distribution defined by our language model.

**function** SAMPLE($x$, $\mathcal{F}$, $\mathcal{S}$, $\mathcal{Q}$, $\lambda$, $k_{\max}$)
- $\mathcal{C}_\lambda \leftarrow \{\}$ ▷ Initialize an empty output set.
- **for** $k = 1, 2, \ldots, k_{\max}$ **do**
  - $y_k \leftarrow y \sim p_\theta(y \mid x)$. ▷ Sample a new response.
  - **if** $\mathcal{Q}(x, y_k) < \lambda_2$ **then** ▷ …ed quality is too low.
    - **continue**
  - **if** $\max\{\mathcal{S}(y_k, y_j) : y_j \in \mathcal{C}_\lambda\} > \lambda_1$ **then** ▷ Reject if it is too similar to other samples.
    - **continue**
  - $\mathcal{C}_\lambda = \mathcal{C}_\lambda \cup \{y_k\}$. ▷ Add the new response to the output set.
  - **if** $\mathcal{F}(\mathcal{C}_\lambda) \geq \lambda_3$ **then** ▷ Check if we are confident enough to stop.
    - **break**
- **return** $\mathcal{C}_\lambda$

We use ROUGE-L for $\mathcal{S}$

define $\mathcal{Q}(x, y) = p_\theta(y \mid x)$

For $\mathcal{F}$, we experiment

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

- Empirical risk over calibration set – for fixed $\lambda = (\lambda_1, \lambda_2, \lambda_3)$

$$\widehat{R}_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda), \quad \text{where } L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i) \colon A_i(y) = 1\}.$$

- p-values (general result via concentration bounds)

- **How to split text into components?**

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Details:

- Empirical risk over calibration set – for fixed $\lambda = (\lambda_1, \lambda_2, \lambda_3)$

$$\widehat{R}_n(\lambda) := \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda), \quad \text{where } L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i) : A_i(y) = 1\}.$$

- p-values (general result via concentration bounds)

- **How to split text into components?**
  *Example (automatic diagnosis): "The heart is mildly enlarged. The lungs are clear."*
  should be split into *"The heart is mildly enlarged."* and *"The lungs are clear."*

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

**Algorithm 2** Conformal component selection

**Definitions:** $\mathcal{C}_\lambda$ is a prediction set, $\mathcal{E}$ is an algorithm for splitting candidates $y$ into components, $\mathcal{F}^{\mathrm{c}}$ is a confidence estimator for individual components, $\gamma$ is our threshold configuration.

**function** SELECT($\mathcal{C}_\lambda, \mathcal{E}, \mathcal{F}^{\mathrm{c}}, \gamma$)
  $\mathcal{C}_\gamma^{\mathrm{inner}} \leftarrow \{\}$ ▷ Initialize an empty output set.
  **for** $y \in \mathcal{C}_\lambda$ **do** ▷ Iterate over full predictions.
    **for** $e \in \mathcal{E}(y)$ **do** ▷ Iterate over individual components.
      **if** $\mathcal{F}^{\mathrm{c}}(e) \geq \gamma$ **then**
        $\mathcal{C}_\gamma^{\mathrm{inner}} \leftarrow \mathcal{C}_\gamma^{\mathrm{inner}} \cup \{e\}$ ▷ Keep only high-confidence components.
  **return** $\mathcal{C}_\gamma^{\mathrm{inner}}$

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

**Algorithm 2** Conformal component selection

**Definitions:** $\mathcal{C}_\lambda$ is a prediction set, $\mathcal{E}$ is an algorithm for splitting candidates $y$ into components, $\mathcal{F}^c$ is a confidence estimator for individual components, $\gamma$ is our threshold configuration.

```
function SELECT(C_λ, E, F^c, γ)
    C_γ^inner ← {}                          ▷ Initialize an empty output set.
    for y ∈ C_λ do                          ▷ Iterate over full predictions.
        for e ∈ E(y) do                     ▷ Iterate over individual components.
            if F^c(e) ≥ γ then
                C_γ^inner ← C_γ^inner ∪ {e}  ▷ Keep only high-confidence components.
    return C_γ^inner
```

$$\mathcal{C}_\gamma^{\text{inner}}(x) := \left\{ e \in \bigcup_{y \in \mathcal{C}_\lambda(x)} \mathcal{E}(y) \colon \mathcal{F}^c(e) \geq \gamma \right\}$$

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

-

$$\mathcal{C}_{\gamma}^{\text{inner}}(x) := \left\{ e \in \bigcup_{y \in \mathcal{C}_{\lambda}(x)} \mathcal{E}(y) \colon \mathcal{F}^{\mathrm{c}}(e) \geq \gamma \right\}$$

$$\hat{\gamma} = \underset{\gamma \in \Gamma_{\text{valid}}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} |\mathcal{C}_{\gamma}(X_i)|.$$

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

-

$$\mathcal{C}_{\gamma}^{\text{inner}}(x) := \left\{ e \in \bigcup_{y \in \mathcal{C}_{\lambda}(x)} \mathcal{E}(y) \colon \mathcal{F}^{\text{c}}(e) \geq \gamma \right\}$$

$$\hat{\gamma} = \underset{\gamma \in \Gamma_{\text{valid}}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} |\mathcal{C}_{\gamma}(X_i)|.$$

Calibration set

Nonrejected by LLT

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

-

$$\mathcal{C}_\gamma^{\text{inner}}(x) := \left\{ e \in \bigcup_{y \in \mathcal{C}_\lambda(x)} \mathcal{E}(y) : \mathcal{F}^c(e) \geq \gamma \right\}$$

$$\hat{\gamma} = \underset{\gamma \in \Gamma_{\text{valid}}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} |\mathcal{C}_\gamma(X_i)|.$$

Nonrejected by LLT

Calibration set

$$\mathbb{P}\Big(\mathbb{P}\Big(A^c_{\text{test}}(e) = 1, \forall e \in \mathcal{C}_\gamma^{\text{inner}}(X_{\text{test}}) \mid \mathcal{D}_{\text{cal}}\Big) \geq 1 - \alpha\Big) \geq 1 - \delta.$$

# 3. Conformal Language Modeling

"Conformal Language Modeling" Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, Barzilay (arxiv link)

Scoring: $\mathcal{F}_{\text{FIRST-K}}(\mathcal{C}) = |\mathcal{C}|$

$\mathcal{F}_{\text{MAX}}(\mathcal{C}) = \max\{\mathcal{Q}(y) : y \in \mathcal{C}\}$

$\mathcal{F}_{\text{SUM}}(\mathcal{C}) = \sum_{y \in C} \mathcal{Q}(y)$

$\mathcal{Q}(x, y) = p_\theta(y \mid x)$ using the likelihood function of the base LM

Tasks: Radiology report generation / News summarization / TriviaQA