Uncertainty in NLP

quantification, interpretation & evaluation

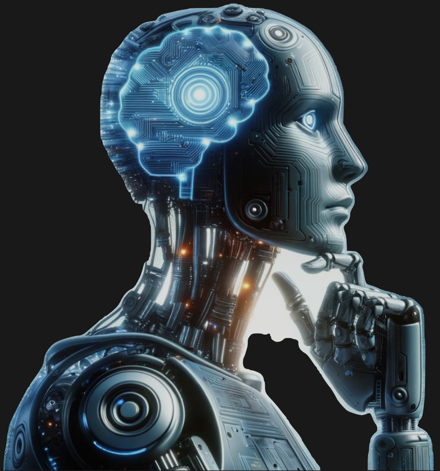
Priberam Machine Learning Lunch Seminars

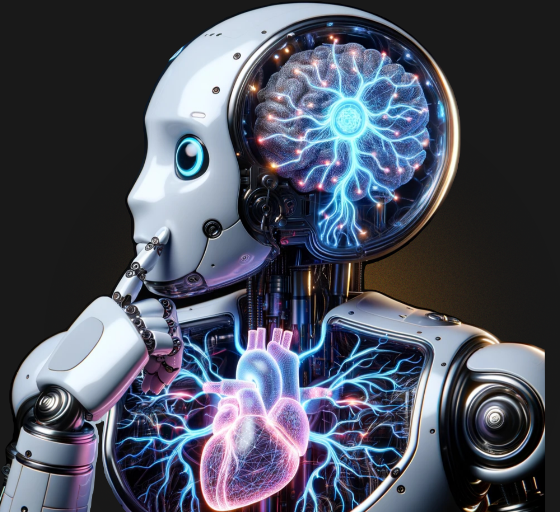
Chrysoula Zerva

Instituto Superior Técnico
Instituto de Telecomunicações

Models don’t always know

what they don’t know





Models don't always **know what they don't know**

Underlying assumption

The more uncertain the model the more prone to error(s)...

Models don't always **know what they don't know**

Underlying assumption

The more uncertain the model the more prone to error(s)...

✓ Detect OOD instances

Models don’t always **know what they don’t know**

Underlying assumption

The more uncertain the model the more prone to error(s)...

- ✓ Detect OOD instances
- ✓ Decision making based on uncertainty
  - ➡ Reject highly uncertainty outputs
  - ➡ Interactive decision-making

Models don’t always **know what they don’t know**

Underlying assumption

The more uncertain the model the more prone to error(s)...

- ✓ Detect OOD instances
- ✓ Decision making based on uncertainty
  - ➡ Reject highly uncertainty outputs
  - ➡ Interactive decision-making
- ✓ Adapt to areas with high uncertainty
  - ➡ Active learning
  - ➡ Curriculum learning

Models don’t always **know what they don’t know**

Underlying assumption

The more uncertain the model the more prone to error(s)...

- ✓ Detect OOD instances
- ✓ Decision making based on uncertainty
  - ➡ Reject highly uncertainty outputs
  - ➡ Interactive decision-making
- ✓ Adapt to areas with high uncertainty
  - ➡ Active learning
  - ➡ Curriculum learning
- ✓ Compare models with respect to their overall confidence

Models don't always **know what they don't know**

Underlying assumption

The more uncertain the model the harder it is to choose among ***valid*** candidate outputs

✳ (Baan et al. 2022;2024; Giulianelli et al. 2023)

Models don't always **know what they don't know**

Underlying assumption

The more uncertain the model the harder it is to choose among ***valid*** candidate outputs

✴ (Baan et al. 2022;2024; Giulianelli et al. 2023)

- ✓ Link to human variability
  - ➡ Sample more estimates

Models don't always **know what they don't know**

Underlying assumption

The more uncertain the model the harder it is to choose among ***valid*** candidate outputs

✳ (Baan et al. 2022;2024; Giulianelli et al. 2023)

- ✓ Link to human variability
  - ➡ Sample more estimates
- ✓ Refine task

Models don’t always **know what they don’t know**

Underlying assumption

The more uncertain the model the harder it is to choose among ***valid*** candidate outputs

✳ (Baan et al. 2022;2024; Giulianelli et al. 2023)

- ✓ Link to human variability
  - ➡ Sample more estimates
- ✓ Refine task
- ✓ Refine the input: Provide more information
  - ➡ To the model

Models don't always **know what they don't know**

Underlying assumption

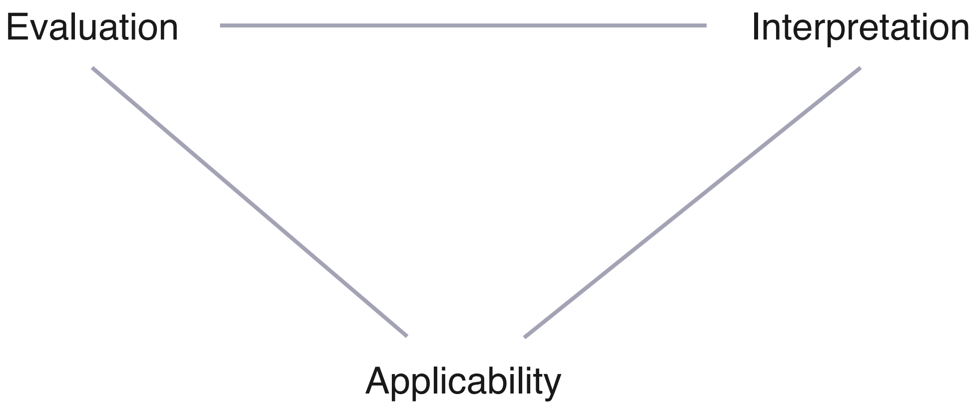
The more uncertain the model the harder it is to choose among ***valid*** candidate outputs

✳ (Baan et al. 2022;2024; Giulianelli et al. 2023)

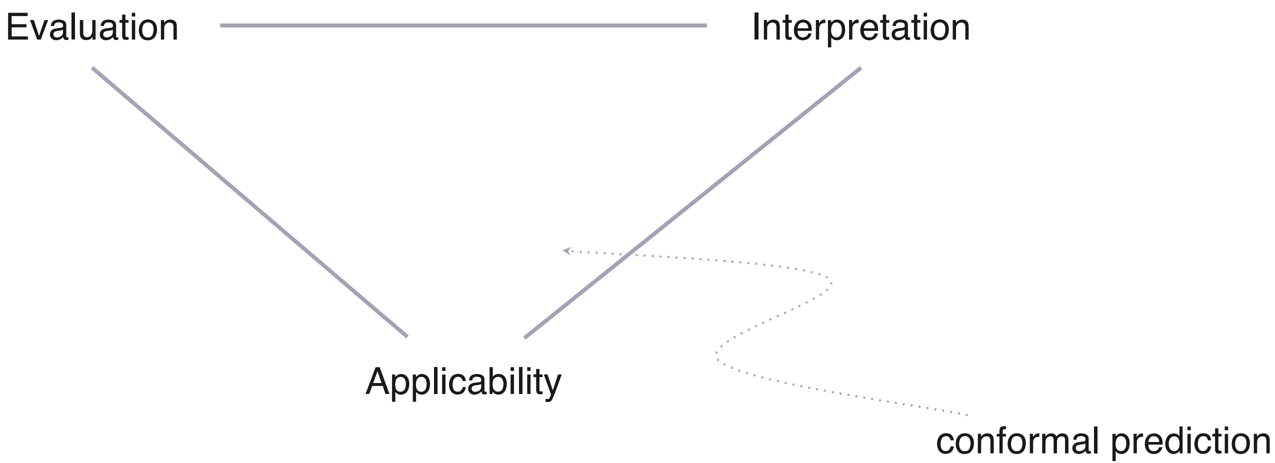
- ✓ Link to human variability
  - ➡ Sample more estimates
- ✓ Refine task
- ✓ Refine the input: Provide more information
  - ➡ To the model
- ✓ Refine the output: Provide more information
  - ➡ To the user

What is a good uncertainty quantifier

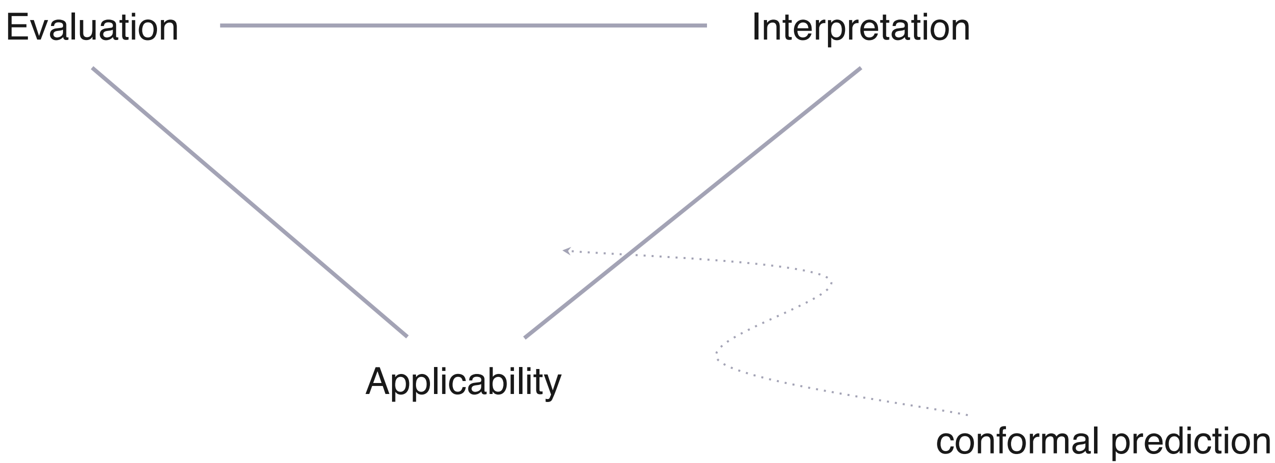
What is a good uncertainty quantifier



What is a good uncertainty quantifier



What is a good uncertainty quantifier



⭒Machine Translation tasks

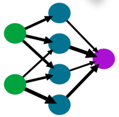
Uncertainty in MT related tasks

src: the nurse left his bag on the floor.

Uncertainty in MT related tasks

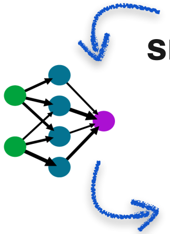
src: the nurse left his bag on the floor.

Uncertainty in MT related tasks



src: the nurse left his bag on the floor.

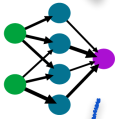
Uncertainty in MT related tasks



src: the nurse left his bag on the floor.

Uncertainty in MT related tasks

src: the nurse left his bag on the floor.

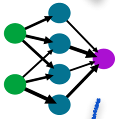


tgt: a enfermeira deixou a bolsa no chão.

Google Translate

Uncertainty in MT related tasks

src: the nurse left his bag on the floor.

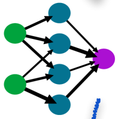


tgt: a enfermeira deixou a bolsa no chão.

Google Translate

Uncertainty in MT related tasks

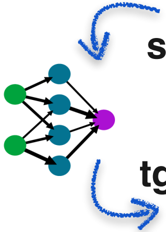
src: the nurse left his bag on the floor.



tgt: a enfermeira deixou a bolsa no chão.

Uncertainty in MT related tasks

src: the nurse left his bag on the floor.

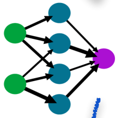


tgt: a enfermeira deixou a bolsa no chão.

ref: o enfermeiro deixou a bolsa no chão.

Uncertainty in MT related tasks

src: the nurse left his bag on the floor.



tgt: a enfermeira deixou a bolsa no chão.

ref: o enfermeiro deixou a bolsa no chão.

Google Translate

Uncertainty in MT related tasks

src: the nurse left his bag on the floor.

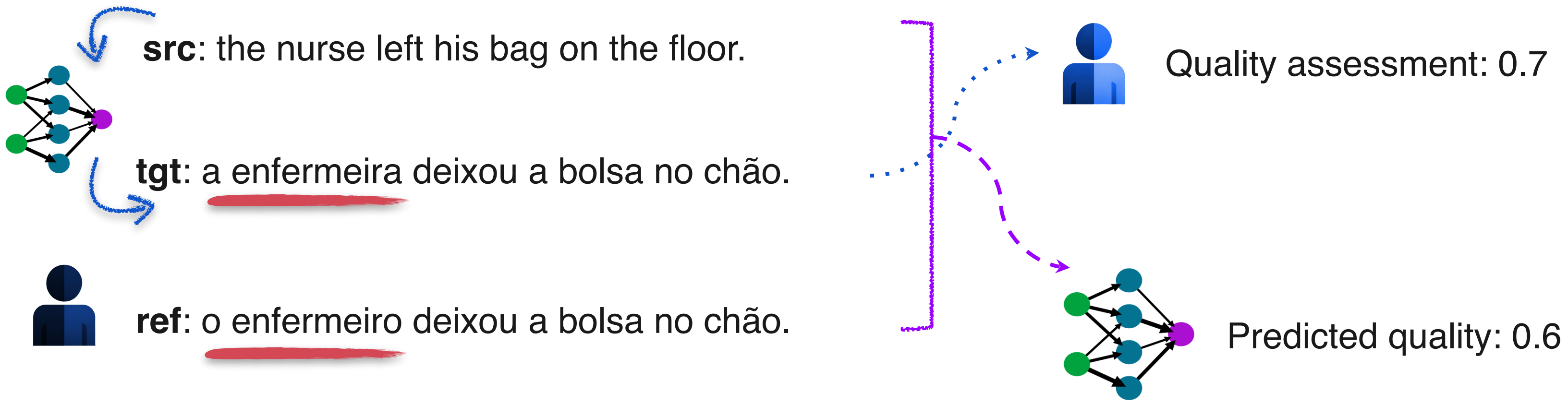
tgt: a enfermeira deixou a bolsa no chão.

ref: o enfermeiro deixou a bolsa no chão.

Quality assessment: 0.7

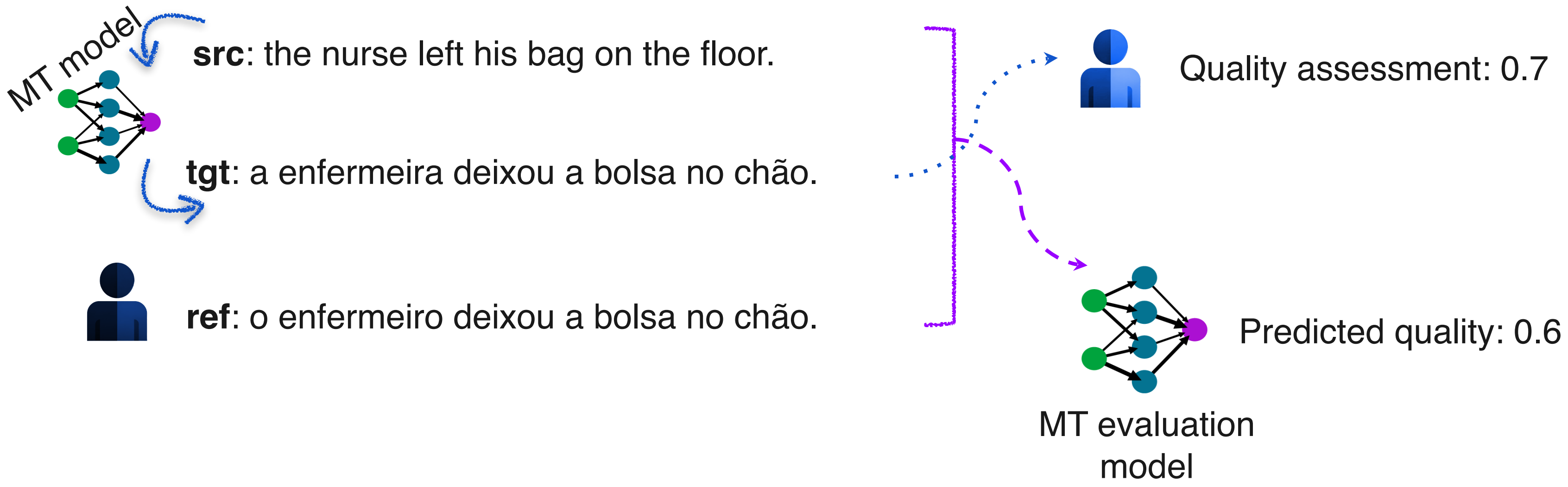
Google Translate

Uncertainty in MT related tasks



Google Translate

Uncertainty in MT related tasks



Google Translate

Applicability

and underlying assumptions

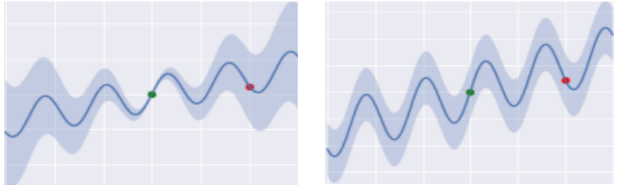
Are our assumptions correct?

What are our assumptions on distribution?

Are our assumptions correct?

What are our assumptions on distribution?

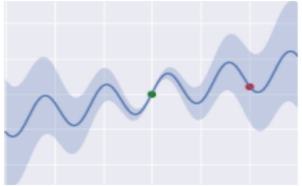
Heteroscedastic vs homoscedastic noise

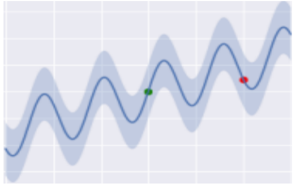


Are our assumptions correct?

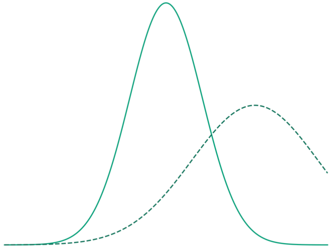
What are our assumptions on distribution?

Heteroscedastic vs homoscedastic noise





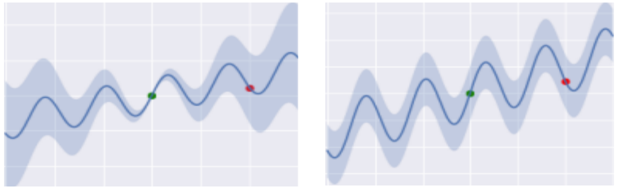
Modeling annotator disagreement



Are our assumptions correct?

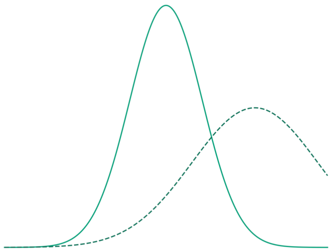
What are our assumptions on distribution?

Heteroscedastic vs homoscedastic noise



Bayesian Neural Networks

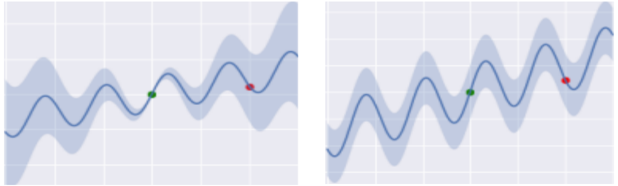
Modeling annotator disagreement



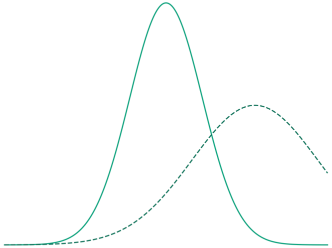
Are our assumptions correct?

What are our assumptions on distribution?

Heteroscedastic vs homoscedastic noise



Modeling annotator disagreement



Bayesian Neural Networks

⋮

MC dropout

⋮

Deep ensembles

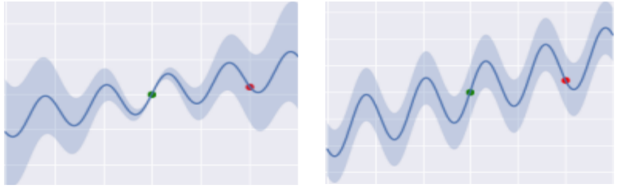
Test-time augmentation

Stochastic variational inference

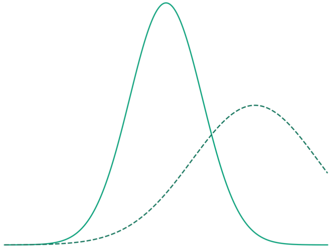
Are our assumptions correct?

What are our assumptions on distribution?

Heteroscedastic vs homoscedastic noise



Modeling annotator disagreement



Bayesian Neural Networks

MC dropout

Deep ensembles

Test-time augmentation

Stochastic variational inference

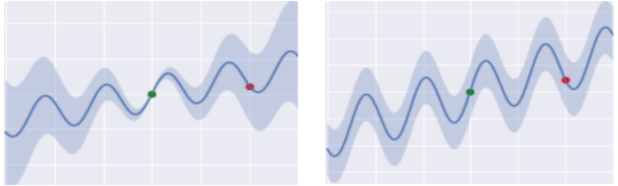
Dirichlet-based uncertainty models

PriorNet (Malinin and Gales 2018)

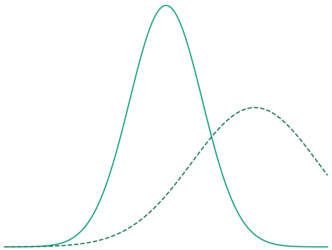
Are our assumptions correct?

What are our assumptions on distribution?

Heteroscedastic vs homoscedastic noise



Modeling annotator disagreement



Bayesian Neural Networks

MC dropout

Deep ensembles

Test-time augmentation

Stochastic variational inference

Dirichlet-based uncertainty models

PriorNet (Malinin and Gales 2018)

....

Deterministic uncertainty models

➡ assumptions on modelling feature density

➡ access to OOD data

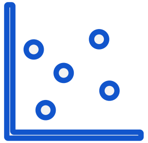
Applicability: which uncertainties?

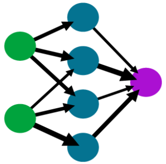
What are our assumptions on uncertainty source?

Applicability: which uncertainties?

What are our assumptions on uncertainty source?

aleatoric ⟷ epistemic

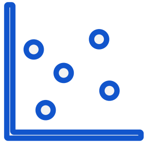


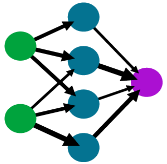


Applicability: which uncertainties?

What are our assumptions on uncertainty source?

aleatoric ⟷ epistemic



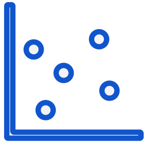


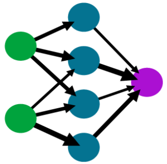
*(Baan et al., 2023)

Applicability: which uncertainties?

What are our assumptions on uncertainty source?

aleatoric ⟷ epistemic





- ➢ Data filtering
- ➢ Ambiguity detection

✓ Better for detecting low quality MT references

- ➢ Active learning setups
- ➢ Better detection of OOD instances

✓ Better detector of hallucinations (Xiao & Wang, 2021)

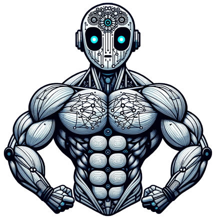
✓ Better for detecting domain shifts in MT evaluation

*(Zerva et al., 2022)

*(Baan et al., 2023)

what can we use?

what can we use?



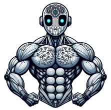
GPT-4?

PaLM

GPT-3

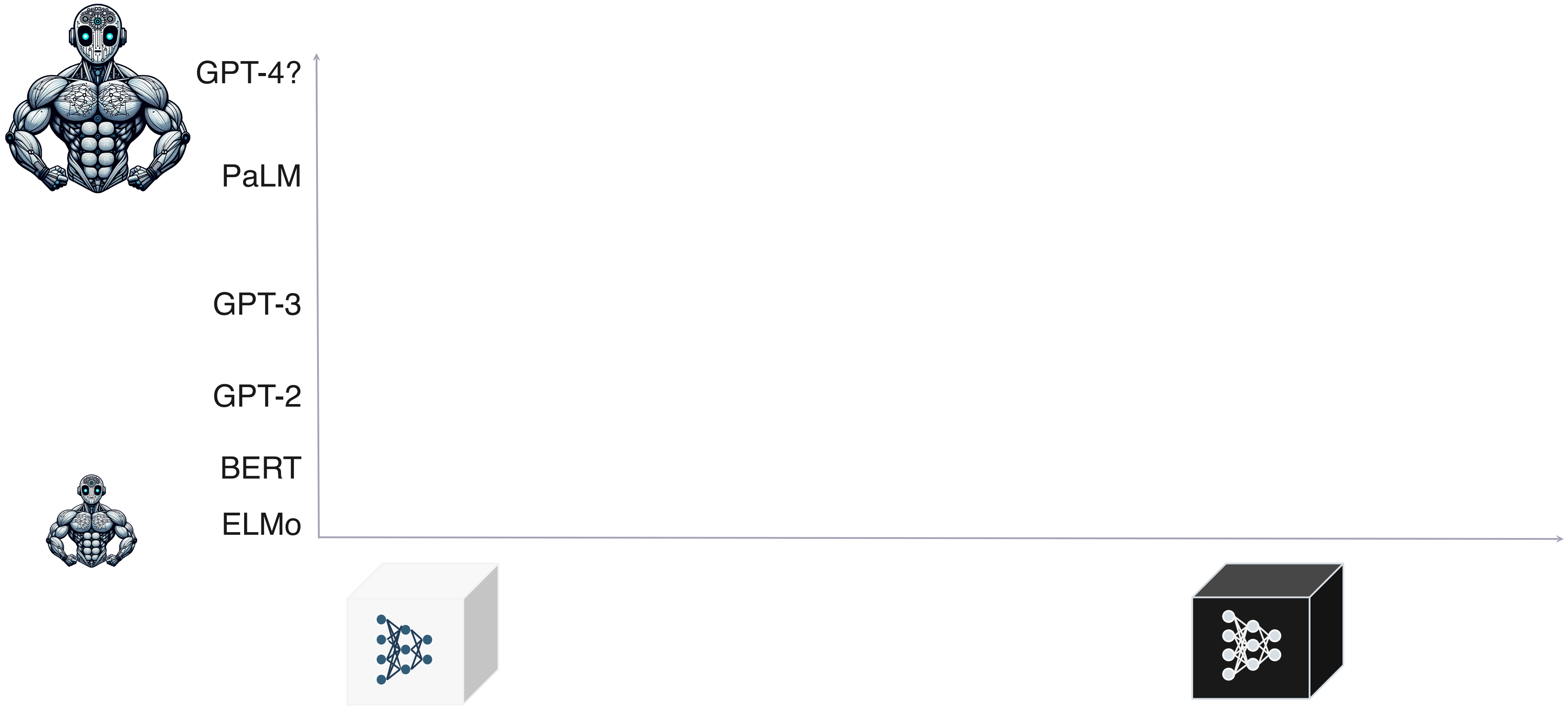
GPT-2

BERT

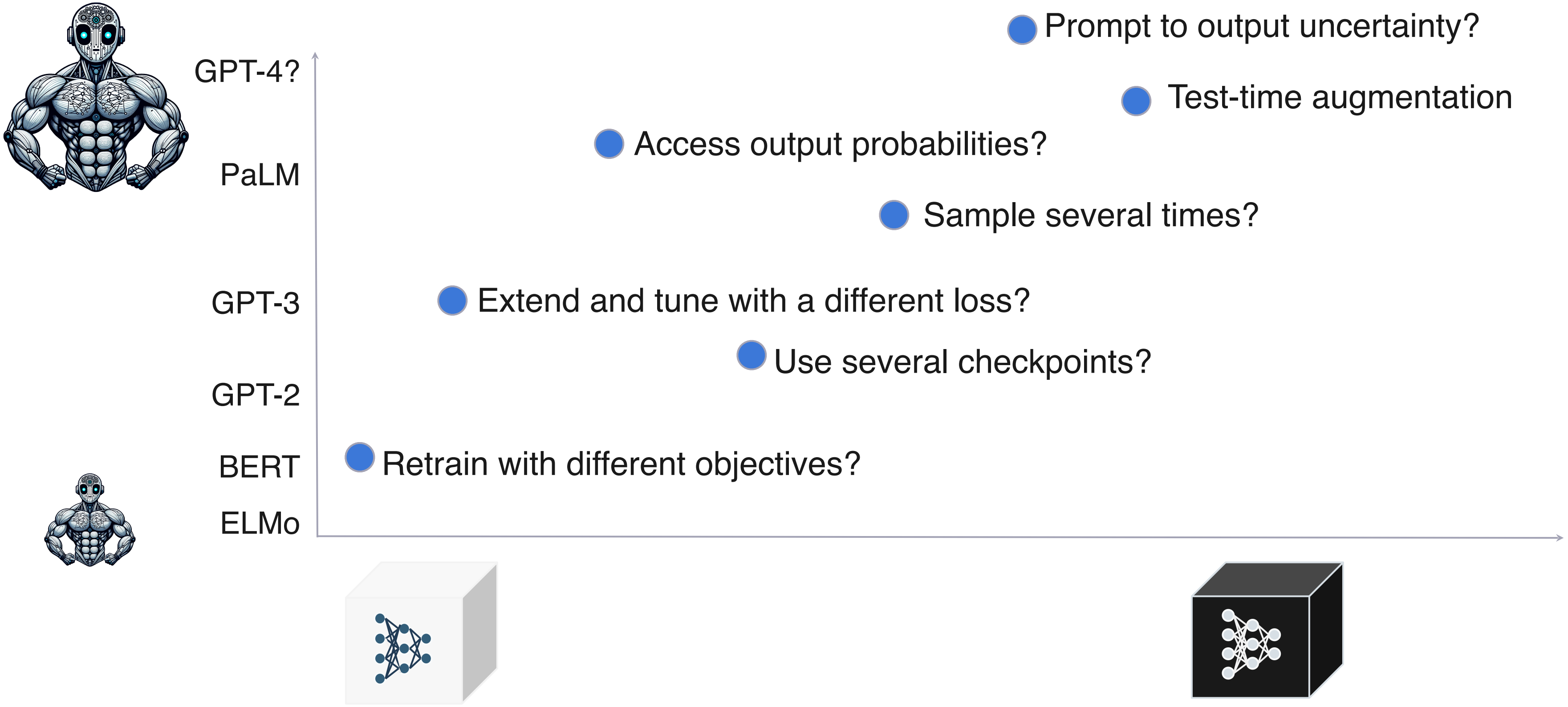


ELMo

what can we use?



what can we use?



Evaluation - Interpretation

and underlying assumptions

Evaluation

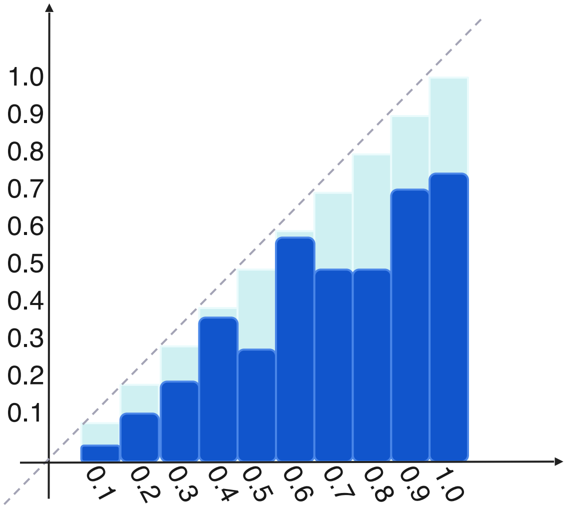
Well calibrated

Evaluation

Well calibrated

Estimated Calibration error (ECE)

$$ECE = \frac{1}{M}\sum_{b=1}^{M} |acc(B_m) - conf(B_m)|$$

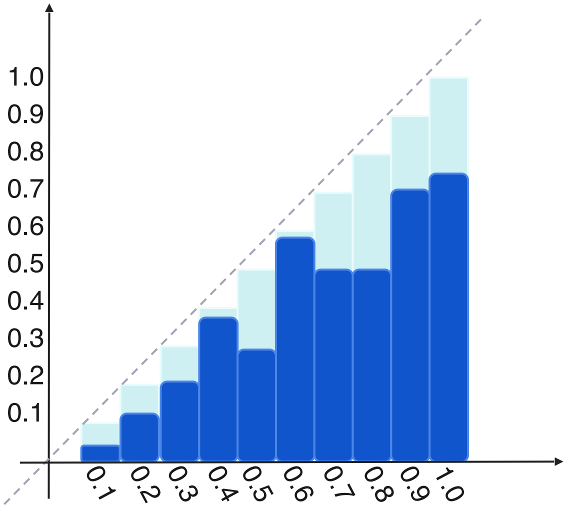


Evaluation

Well calibrated

Estimated Calibration error (ECE)

$$ECE = \frac{1}{M}\sum_{b=1}^{M} |acc(B_m) - conf(B_m)|$$



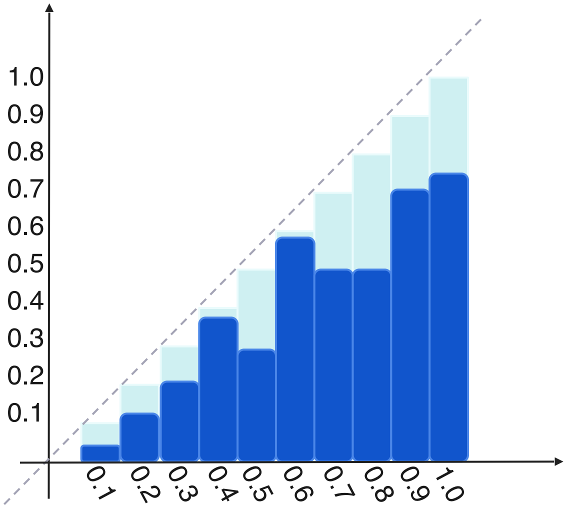
- ✗ Sensitive to the choice of bin width
- ✗ small changes to model predictions can cause large jumps in the ECE
- ✗ Not suitable in tasks with high label variability

Evaluation

Well calibrated

Estimated Calibration error (ECE)

$$ECE = \frac{1}{M}\sum_{b=1}^{M} |acc(B_m) - conf(B_m)|$$



- ✗ Sensitive to the choice of bin width
- ✗ small changes to model predictions can cause large jumps in the ECE
- ✗ Not suitable in tasks with high label variability

- ➢ Max calibration error
- ➢ Logit-smoothed ECE
- ➢ Human Entropy Calibration Score
- ➢ Human Distribution Calibration Error

Evaluation

Focussing on errors

Evaluation

Focussing on errors

- Correlation with error

$$\rho(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

$$r(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

Evaluation

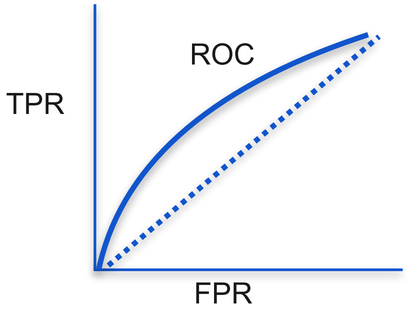
Focussing on errors

- ➢ Correlation with error

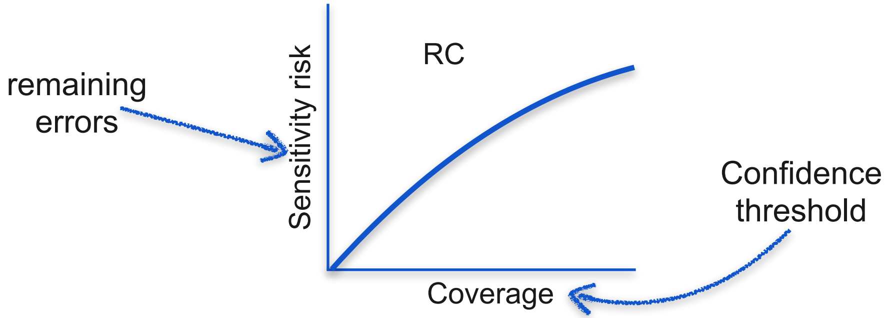
$$\rho(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

$$r(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

- ➢ AUROC



- ➢ AUC-RC



Evaluation

Focussing on errors

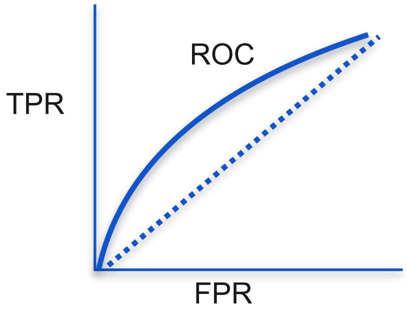
➢ Correlation with error

$$\rho(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

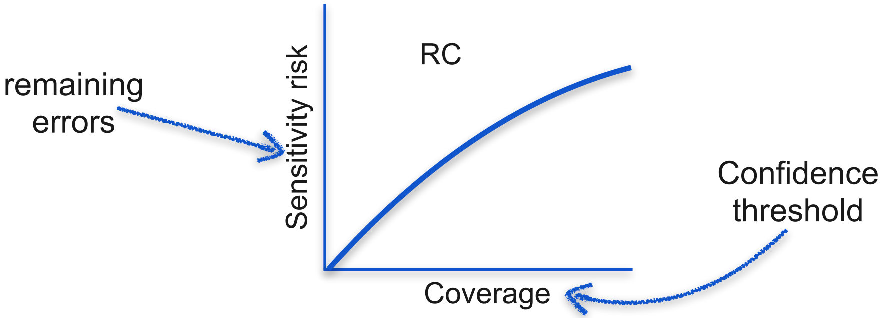
$$r(u(X_{\text{test}}), |\hat{Y}_{\text{test}} - Y_{test}|)$$

✗ Sensitive to outliers
✗ Not informative in terms of scale

➢ AUROC

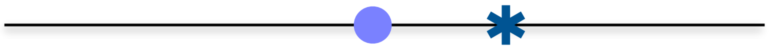


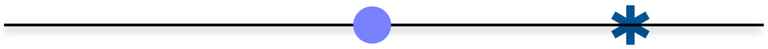
➢ AUC-RC



Evaluation

Evaluation

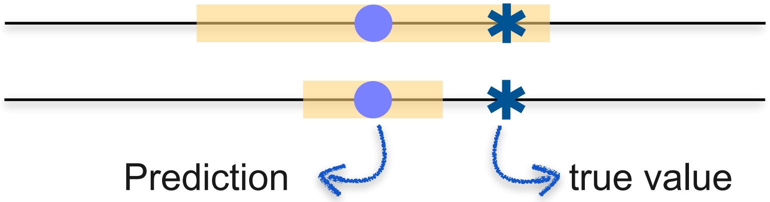


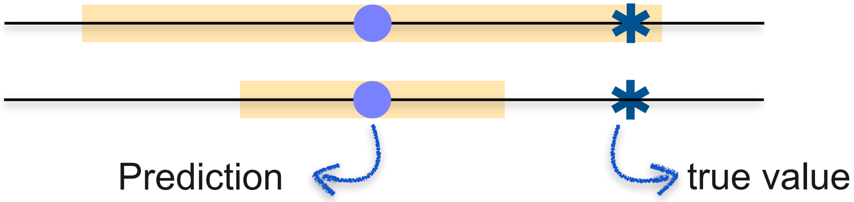


Evaluation

Evaluation

Evaluation





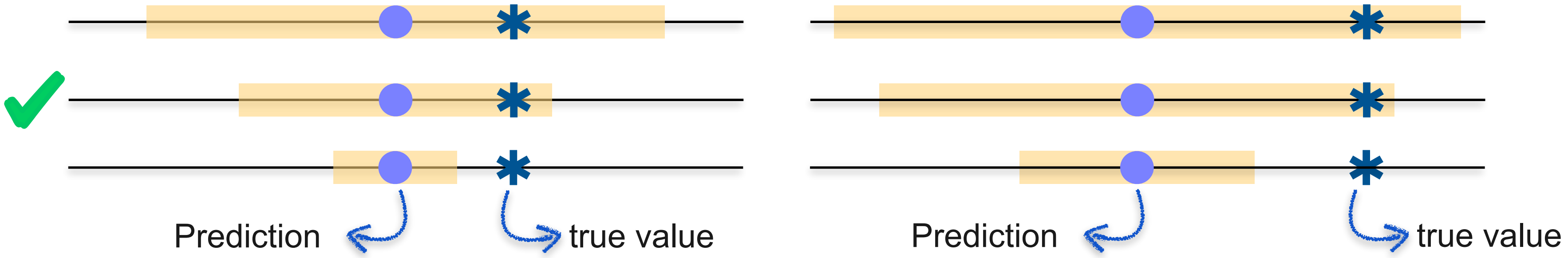
Evaluation

Width - Sharpness

Tight intervals - peaky distributions

Coverage

Including the true label in the confidence interval



Evaluation

Width - Sharpness

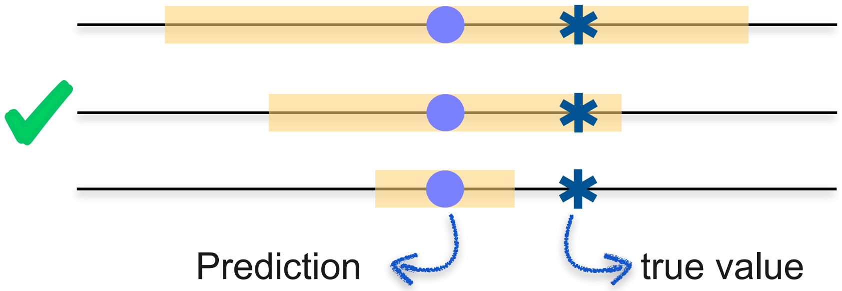
Tight intervals - peaky distributions

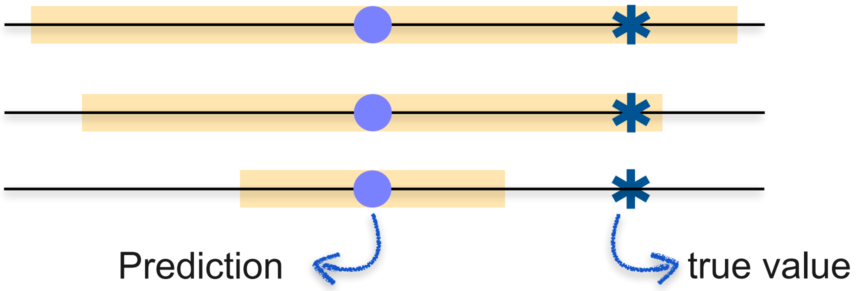
Coverage

Including the true label in the confidence interval

Robustness

Robust to noise injection - adversarial attacks





Evaluation

Width - Sharpness

Tight intervals - peaky distributions

Coverage

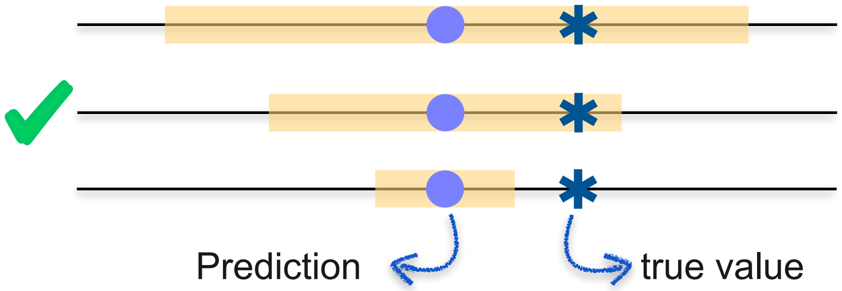
Including the true label in the confidence interval

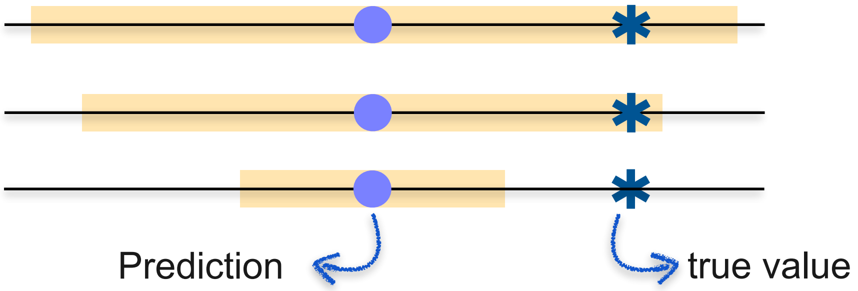
Robustness

Robust to noise injection - adversarial attacks

Fairness

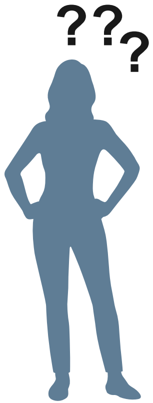
Similar behaviour across attributes





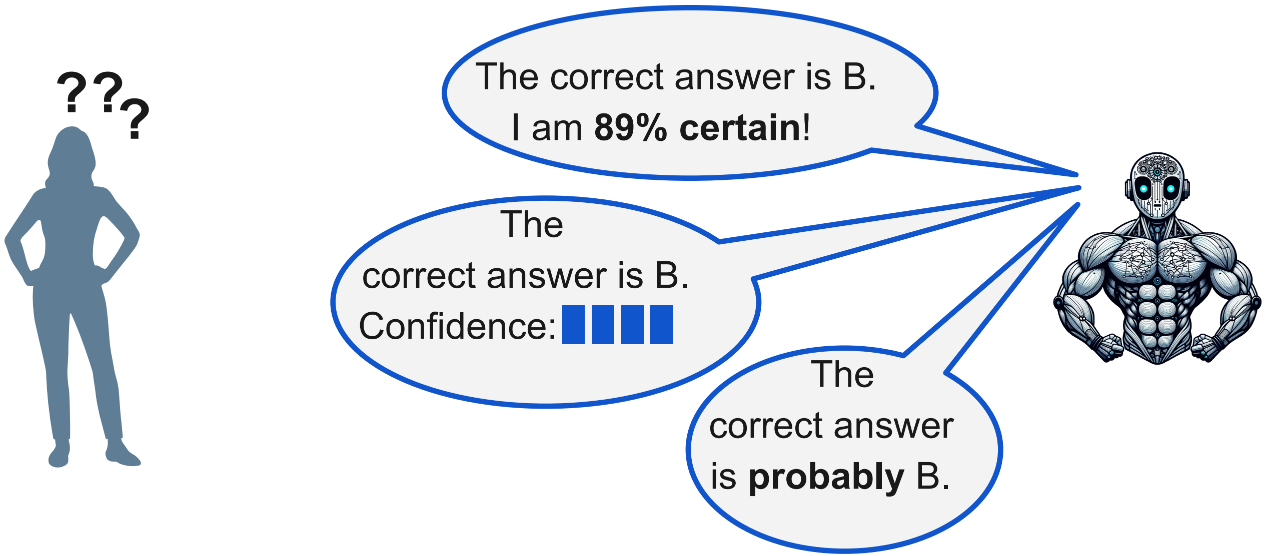
How do we represent - interpret uncertainty scores?

Do people have a shared notion of risk/uncertainty/confidence?



How do we represent - interpret uncertainty scores?

Do people have a shared notion of risk/uncertainty/confidence?



Turning to conformal prediction

and coverage

Conformal prediction

Ingredients:

- Test set $\{X_{\text{test}}, Y_{\text{test}}\}$

- Held-out calibration set
  $S^{\text{cal}} = \{X_{\text{cal}}, Y_{\text{cal}}\} = \{(x_i, y_i)\}_{i=1}^{n}$

- Non-conformity score for each data point:
  $s_i := s(x_i, y_i)$

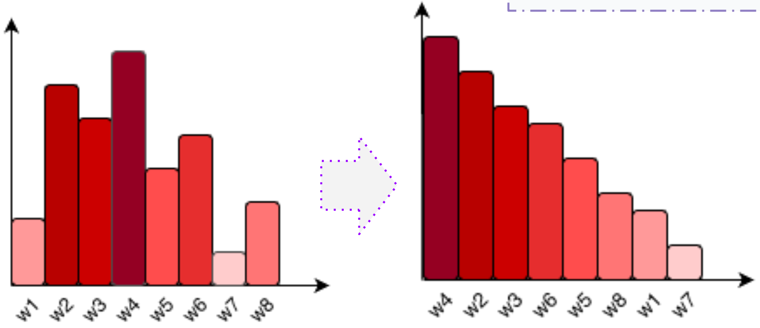
- Desired coverage 1-α

Conformal prediction

Ingredients:

- Test set $\{X_{\text{test}}, Y_{\text{test}}\}$
- Held-out calibration set
  $S^{\text{cal}} = \{X_{\text{cal}}, Y_{\text{cal}}\} = \{(x_i, y_i)\}_{i=1}^{n}$
- Non-conformity score for each data point:
  $s_i := s(x_i, y_i)$
- Desired coverage 1-α

Classification

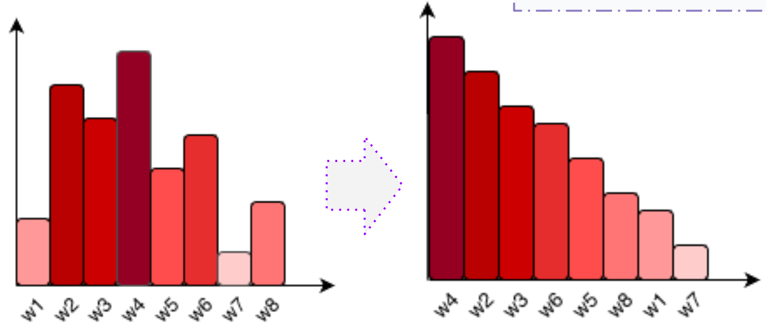


Conformal prediction

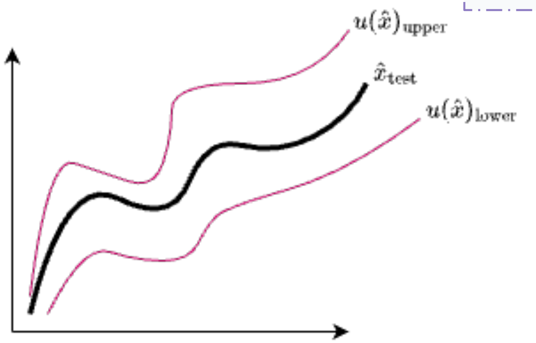
Ingredients:

- Test set $\{X_{\text{test}}, Y_{\text{test}}\}$
- Held-out calibration set
  $S^{\text{cal}} = \{X_{\text{cal}}, Y_{\text{cal}}\} = \{(x_i, y_i)\}_{i=1}^{n}$
- Non-conformity score for each data point:
  $s_i := s(x_i, y_i)$
- Desired coverage 1-α

Classification



Regression

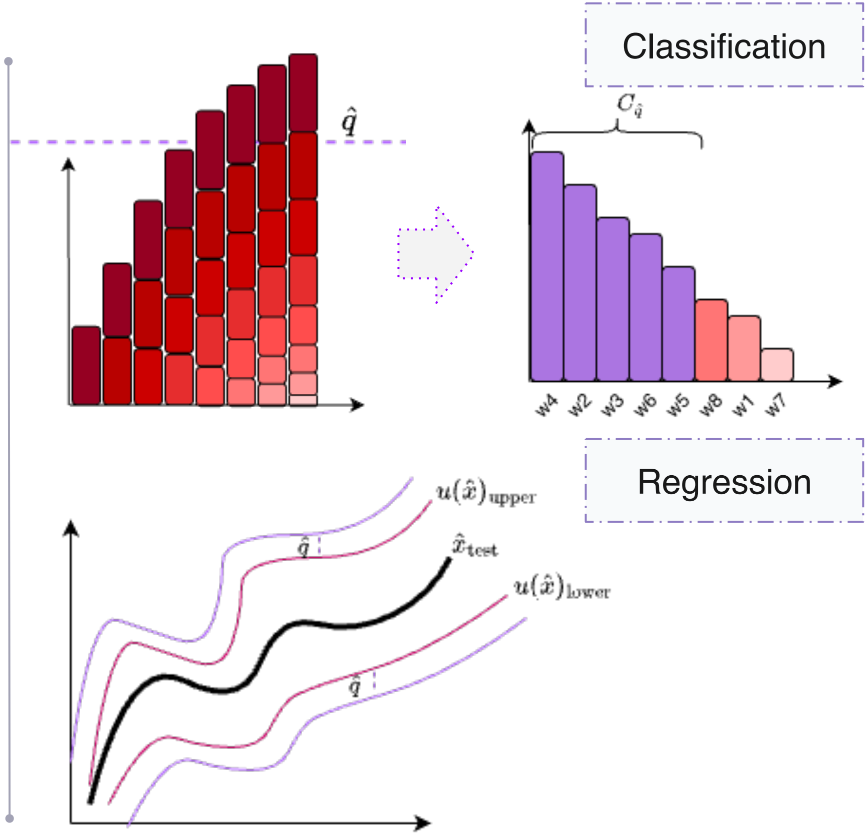


Conformal prediction

Process:

- Compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile $\hat{q}$ over the non-conformity scores $s_i := s(x_i, y_i)$ of the calibration set

- We can now compute the confidence intervals $C_{\hat{q}}(x_{\text{test}}) = \{y \in Y \, : \, s(x_{\text{test}}, y) \leq \hat{q}\}$



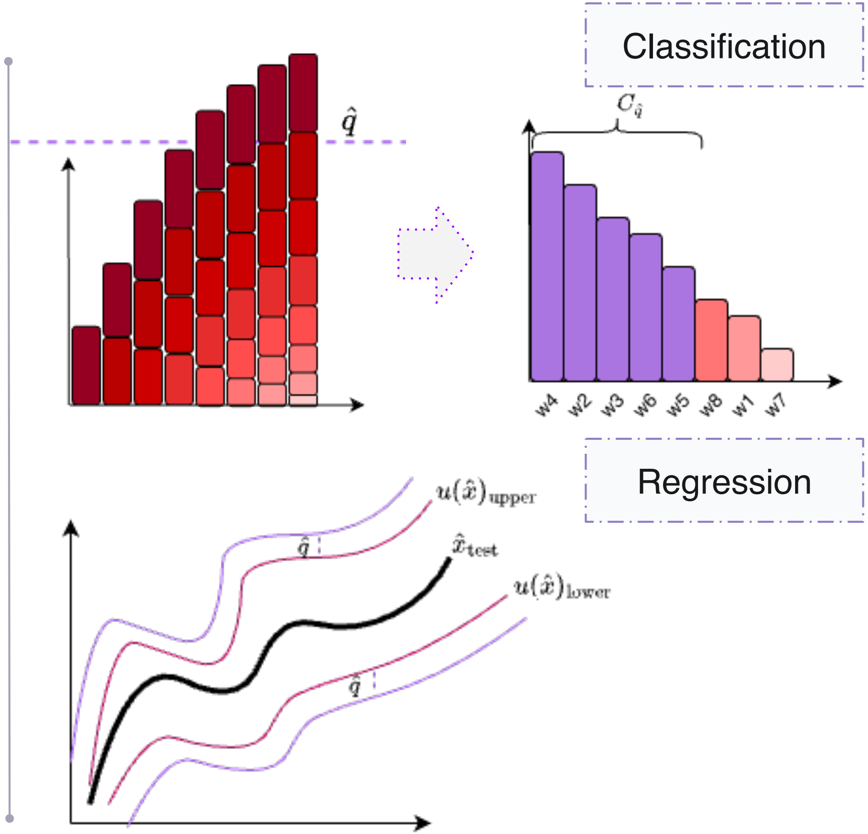
Conformal prediction

Process:

- Compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile $\hat{q}$ over the non-conformity scores $s_i := s(x_i, y_i)$ of the calibration set

- We can now compute the confidence intervals $C_{\hat{q}}(x_{\text{test}}) = \{y \in Y \,:\, s(x_{\text{test}}, y) \leq \hat{q}\}$

$$\boxed{\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \in \left[1-\alpha,\, 1-\alpha + \frac{1}{n+1}\right]}$$



Conformal prediction

Process:

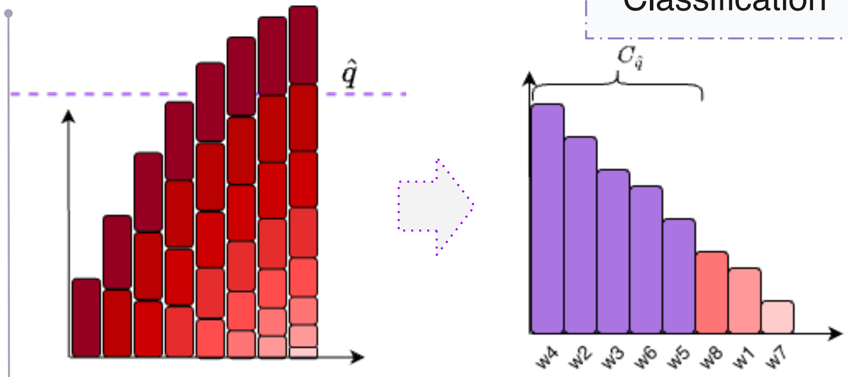
- Compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile $\hat{q}$ over the non-conformity scores $s_i := s(x_i, y_i)$ of the calibration set

- We can now compute the confidence intervals $C_{\hat{q}}(x_{\text{test}}) = \{y \in Y \,:\, s(x_{\text{test}}, y) \leq \hat{q}\}$

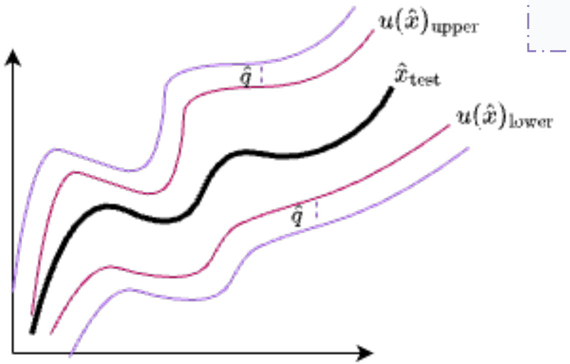
$$\boxed{\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \in \left[1-\alpha,\, 1-\alpha+\frac{1}{n+1}\right]}$$

Guarantee on **marginal** coverage

Classification

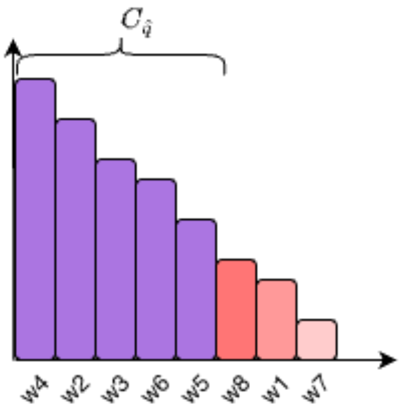


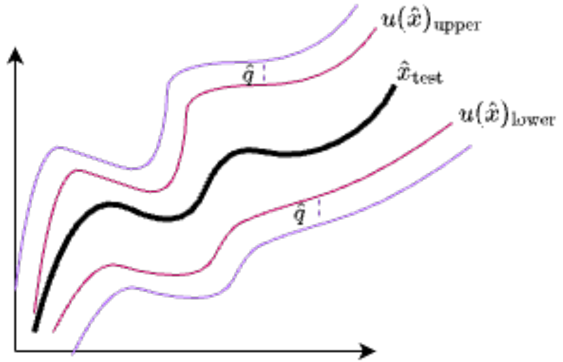
Regression



Interpretation

➔ width scaled with respect to desired coverage
✓ easier comparison between instances
✓ Meaningful intervals across tasks
✓ Non-parametric

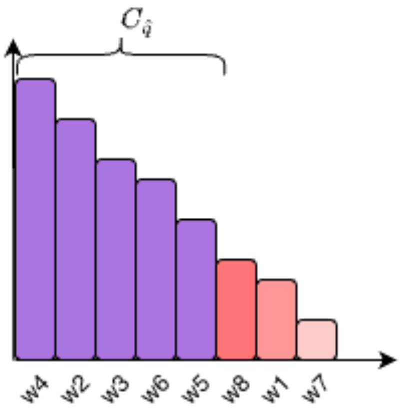


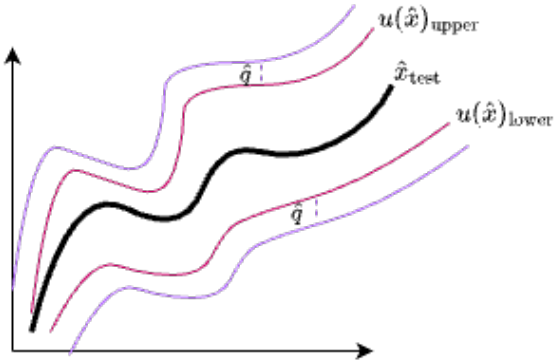


Interpretation

➔ width scaled with respect to desired coverage
✓ easier comparison between instances
✓ Meaningful intervals across tasks
✓ Non-parametric

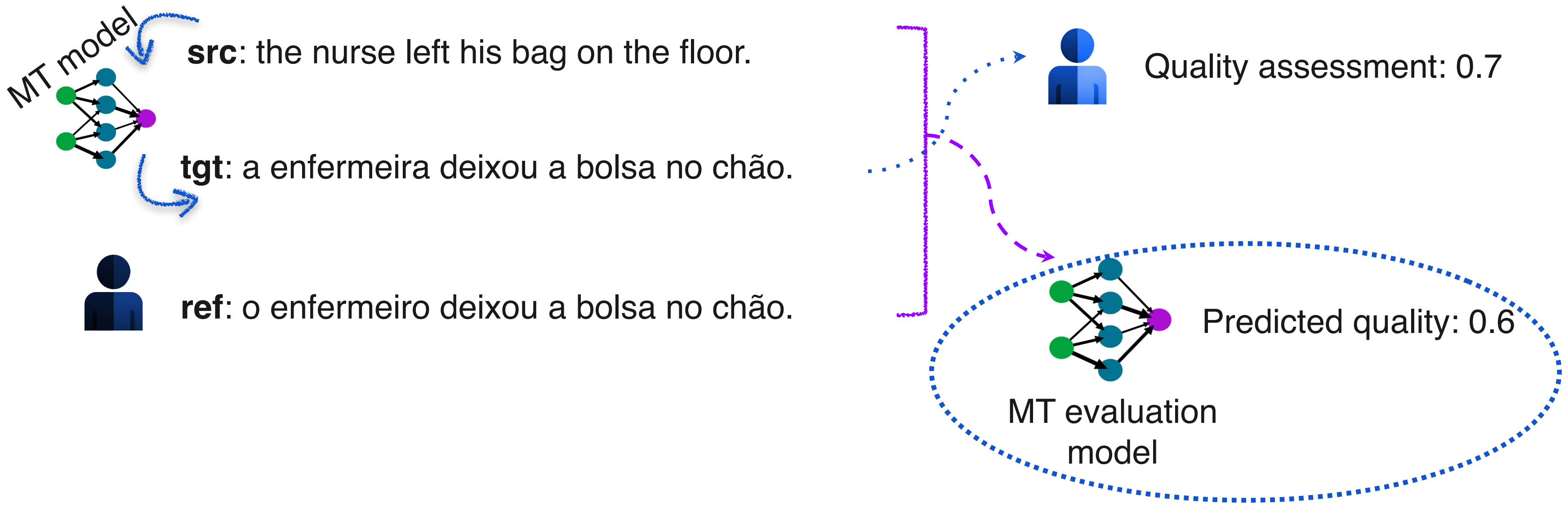
Holds only for exchangeable data!





Conformalising MT evaluation

Conformalising MT evaluation



Conformal prediction for MT evaluation

Conformal prediction for MT evaluation

MC Dropout

Deep Ensembles

$\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$

❋(Glushkova et al., 2021, Zerva et al. 2022)

❋(Zerva and Martins, 2023)

Conformal prediction for MT evaluation

MC Dropout

Deep Ensembles

$$\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$$

Heteroscedastic Regression

$$\mathcal{L}_{\text{HTS}}(\hat{\mu}, \hat{\sigma}^2; y) = \frac{(y - \hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2} \log \hat{\sigma}^2$$

❋(Glushkova et al., 2021, Zerva et al. 2022)

❋(Zerva and Martins, 2023)

Conformal prediction for MT evaluation

MC Dropout

Deep Ensembles

$$\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$$

Heteroscedastic Regression

$$\mathcal{L}_{\text{HTS}}(\hat{\mu}, \hat{\sigma}^2; y) = \frac{(y - \hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2} \log \hat{\sigma}^2$$

Direct Uncertainty Prediction

$$\mathcal{L}_{\text{DUP}}(\hat{\epsilon}; \epsilon) = \frac{\epsilon^2}{2\hat{\epsilon}^2} + \frac{1}{2} \log(\hat{\epsilon})^2$$

Regress on the residuals!

❋(Glushkova et al., 2021, Zerva et al. 2022)

❋(Zerva and Martins, 2023)

Conformal prediction for MT evaluation

MC Dropout

Deep Ensembles

$$\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$$

Heteroscedastic Regression

$$\mathcal{L}_{\text{HTS}}(\hat{\mu}, \hat{\sigma}^2; y) = \frac{(y-\hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2}\log\hat{\sigma}^2$$

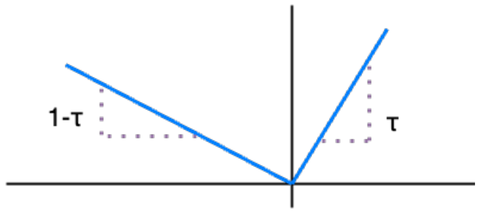
Direct Uncertainty Prediction

$$\mathcal{L}_{\text{DUP}}(\hat{\epsilon}; \epsilon) = \frac{\epsilon^2}{2\hat{\epsilon}^2} + \frac{1}{2}\log(\hat{\epsilon})^2$$

Regress on the residuals!

❋(Glushkova et al., 2021, Zerva et al. 2022)

Quantile regression



$$\mathcal{L}_\tau(\hat{y}; y) = (\hat{y} - y)(\mathbb{1}\{y \le \hat{y}\} - \tau)$$

Optimise to predict selected quantiles instead of mean!

❋(Zerva and Martins, 2023)

Conformal prediction for MT evaluation

MC Dropout

Deep Ensembles

$$\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$$

Heteroscedastic Regression

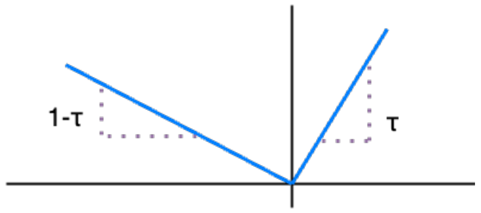
$$\mathcal{L}_{\text{HTS}}(\hat{\mu}, \hat{\sigma}^2; y) = \frac{(y-\hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2}\log\hat{\sigma}^2$$

Direct Uncertainty Prediction

$$\mathcal{L}_{\text{DUP}}(\hat{\epsilon}; \epsilon) = \frac{\epsilon^2}{2\hat{\epsilon}^2} + \frac{1}{2}\log(\hat{\epsilon})^2$$

Regress on the residuals!

Quantile regression



$$\mathcal{L}_\tau(\hat{y}; y) = (\hat{y} - y)(\mathbb{1}\{y \le \hat{y}\} - \tau)$$

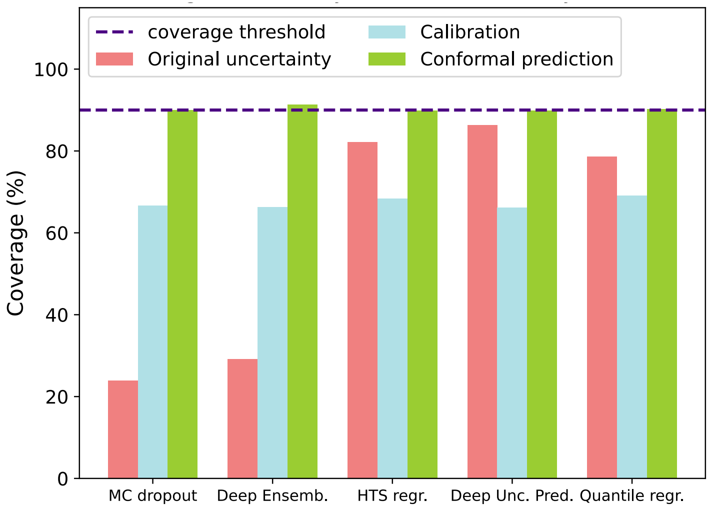
Optimise to predict selected quantiles instead of mean!

$$\boxed{s(x, y) = \frac{|y - \hat{y}(x)|}{u(x)}}$$

❋(Glushkova et al., 2021, Zerva et al. 2022)

❋(Zerva and Martins, 2023)

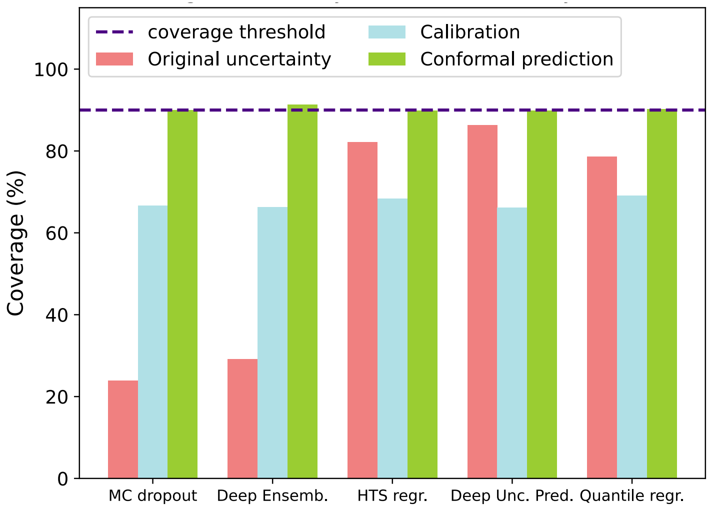
Selecting the most suitable UQ



Coverage for different UQ on COMET tested on WMT 2021 Metrics data

❋(Zerva and Martins, 2023)

Selecting the most suitable UQ

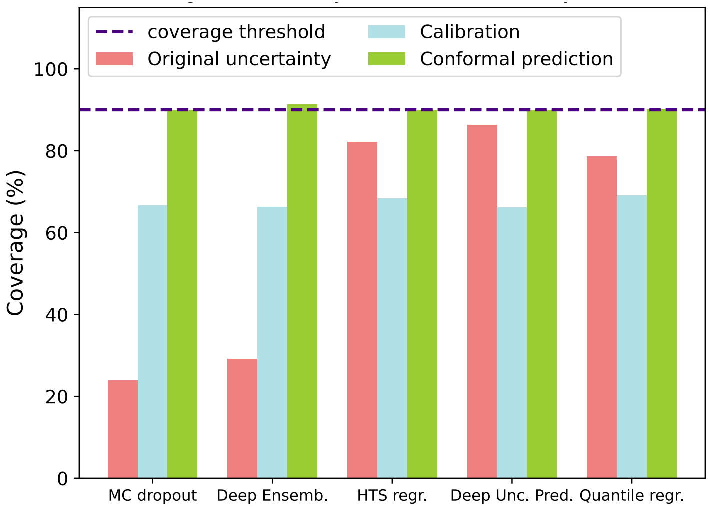


Coverage for different UQ on COMET tested on WMT 2021 Metrics data

Coverage (and $\hat{q}$) aligns well with error correlation

❋(Zerva and Martins, 2023)

Selecting the most suitable UQ



Coverage for different UQ on COMET tested on WMT 2021 Metrics data

Coverage (and $\hat{q}$) aligns well with error correlation

	$\hat{q} \downarrow$	$r \uparrow$
MC Dropout	8.08	0.04
Deep Ensembles	6.99	0.07
Heteroscedastic reg.	2.69	0.24
Direct uncertainty pred.	1.81	0.27
Quantile regression	1.28	0.34

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

MCD DE HTS DUP QNT

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

	MCD	DE	HTS	DUP	QNT
English-Czech					
English-German					
English-Japanese					
English-Polish					
English-Russian					
English-Tamil					
English-Chinese					
Czech-English					
German-English					
Japanese-English					
Khmer-English					
Polish-English					
Pashto-English					
Russian-English					
Tamil-English					
Chinese-English					

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

	MCD	DE	HTS	DUP	QNT
English-Czech	0.982	0.959	0.939	0.875	0.931
English-German	0.973	0.971	0.925	0.863	0.927
English-Japanese	0.990	0.978	0.987	0.886	0.972
English-Polish	0.977	0.948	0.914	0.882	0.914
English-Russian	0.974	0.958	0.936	0.862	0.926
English-Tamil	0.970	0.952	0.949	0.892	0.858
English-Chinese	0.934	0.983	0.991	0.919	0.945
Czech-English	0.890	0.871	0.884	0.898	0.875
German-English	0.880	0.888	0.867	0.896	0.902
Japanese-English	0.883	0.856	0.921	0.910	0.887
Khmer-English	0.881	0.875	0.948	0.943	0.840
Polish-English	0.862	0.833	0.825	0.873	0.849
Pashto-English	0.851	0.854	0.932	0.922	0.786
Russian-English	0.851	0.828	0.831	0.879	0.888
Tamil-English	0.793	0.809	0.878	0.898	0.883
Chinese-English	0.861	0.833	0.868	0.886	0.827

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

	MCD	DE	HTS	DUP	QNT
English-Czech	0.982	0.959	0.939	0.875	0.931
English-German	0.973	0.971	0.925	0.863	0.927
English-Japanese	0.990	0.978	0.987	0.886	0.972
English-Polish	0.977	0.948	0.914	0.882	0.914
English-Russian	0.974	0.958	0.936	0.862	0.926
English-Tamil	0.970	0.952	0.949	0.892	0.858
English-Chinese	0.934	0.983	0.991	0.919	0.945
Czech-English	0.890	0.871	0.884	0.898	0.875
German-English	0.880	0.888	0.867	0.896	0.902
Japanese-English	0.883	0.856	0.921	0.910	0.887
Khmer-English	0.881	0.875	0.948	0.943	0.840
Polish-English	0.862	0.833	0.825	0.873	0.849
Pashto-English	0.851	0.854	0.932	0.922	0.786
Russian-English	0.851	0.828	0.831	0.879	0.888
Tamil-English	0.793	0.809	0.878	0.898	0.883
Chinese-English	0.861	0.833	0.868	0.886	0.827

Language-wise recalibration

❋(Zerva and Martins, 2023)

Access to fairness

What if we compute coverage with respect to specific attributes?

	MCD	DE	HTS	DUP	QNT
English-Czech	0.982	0.959	0.939	0.875	0.931
English-German	0.973	0.971	0.925	0.863	0.927
English-Japanese	0.990	0.978	0.987	0.886	0.972
English-Polish	0.977	0.948	0.914	0.882	0.914
English-Russian	0.974	0.958	0.936	0.862	0.926
English-Tamil	0.970	0.952	0.949	0.892	0.858
English-Chinese	0.934	0.983	0.991	0.919	0.945
Czech-English	0.890	0.871	0.884	0.898	0.875
German-English	0.880	0.888	0.867	0.896	0.902
Japanese-English	0.883	0.856	0.921	0.910	0.887
Khmer-English	0.881	0.875	0.948	0.943	0.840
Polish-English	0.862	0.833	0.825	0.873	0.849
Pashto-English	0.851	0.854	0.932	0.922	0.786
Russian-English	0.851	0.828	0.831	0.879	0.888
Tamil-English	0.793	0.809	0.878	0.898	0.883
Chinese-English	0.861	0.833	0.868	0.886	0.827

Language-wise recalibration

MCD	DE	HTS	DUP	QNT
0.893	0.917	0.888	0.892	0.902
0.902	0.902	0.902	0.896	0.893
0.909	0.891	0.900	0.891	0.904
0.882	0.905	0.895	0.900	0.898
0.900	0.898	0.908	0.906	0.903
0.903	0.895	0.883	0.886	0.903
0.880	0.890	0.884	0.896	0.896
0.890	0.917	0.909	0.904	0.894
0.897	0.901	0.901	0.897	0.903
0.900	0.912	0.899	0.894	0.902
0.896	0.903	0.902	0.904	0.894
0.900	0.905	0.893	0.894	0.877
0.905	0.899	0.900	0.884	0.907
0.910	0.896	0.907	0.900	0.900
0.884	0.901	0.886	0.901	0.908
0.900	0.910	0.908	0.900	0.905

❋(Zerva and Martins, 2023)

Fairness

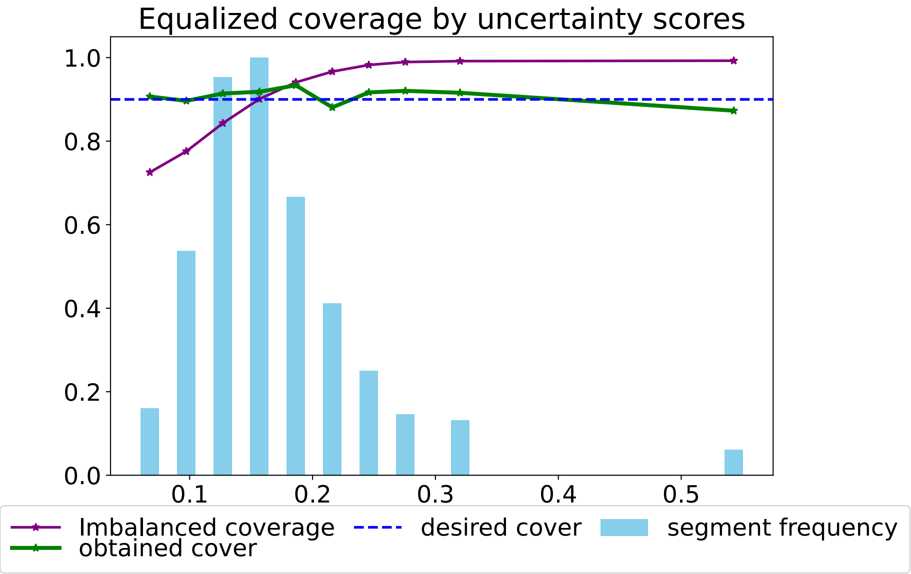
Beyond language

❋(Zerva and Martins, 2023)

Fairness

Beyond language

Can also be applied on continuous attributes

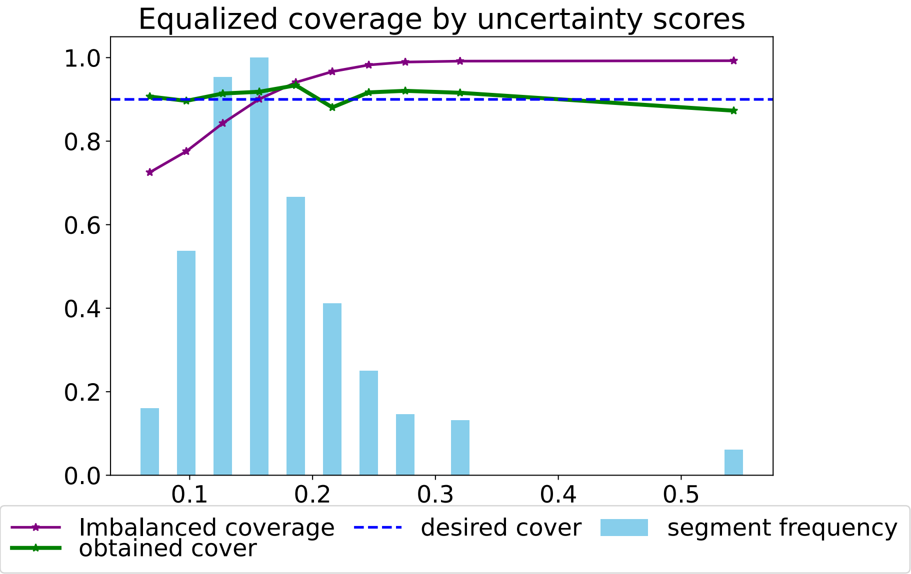


❋(Zerva and Martins, 2023)

Fairness

Beyond language

Can also be applied on continuous attributes



… sensitive, demographic attributes

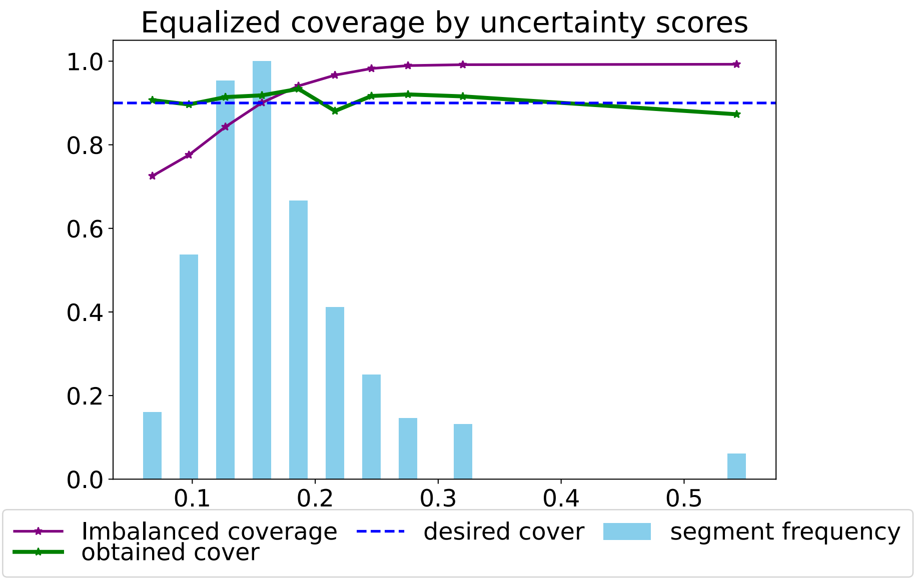
- Gender bias
- Racial bias
- Religious bias
- Age bias
- …

❋(Zerva and Martins, 2023)

Fairness

Beyond language

Can also be applied on continuous attributes



… sensitive, demographic attributes

- Gender bias
- Racial bias
- Religious bias
- Age bias
- …

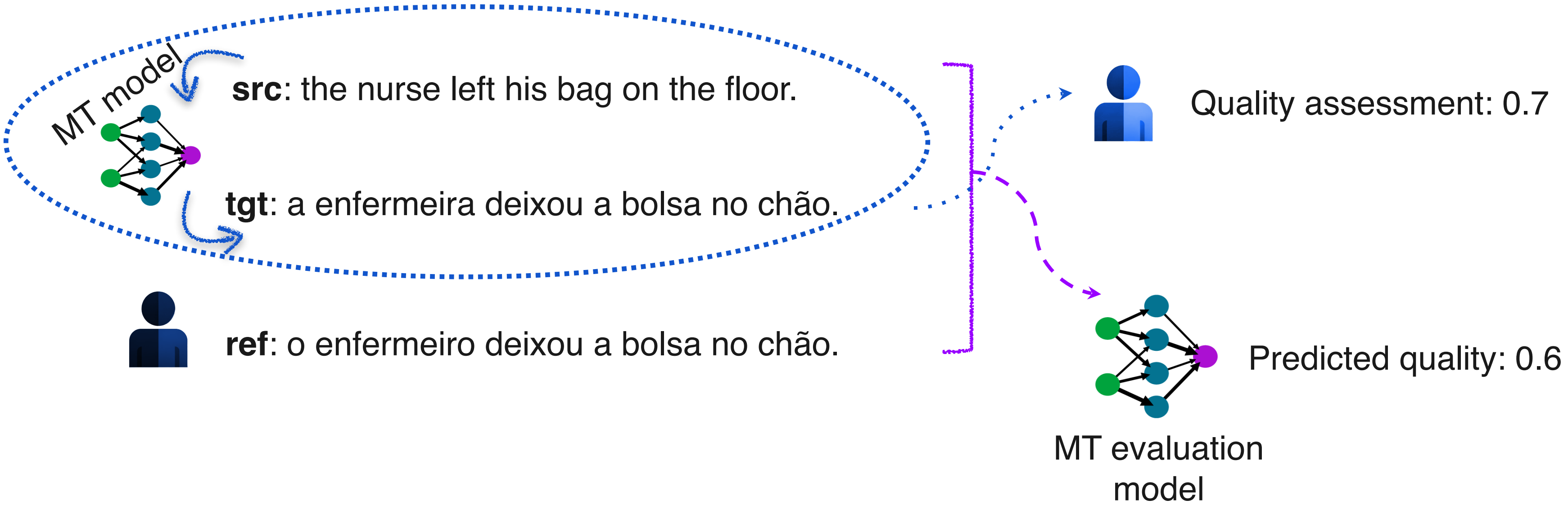
… other linguistic aspects

- Style preference
- Formality
- Example difficulty
- Syntactic complexity

❋(Zerva and Martins, 2023)

Conformalising MT

Conformalising MT



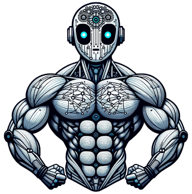
What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão.

❋(Kuhn et al. , 2023; Ye et al., 2024)

What about generation?

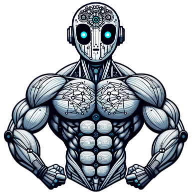
the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão.



❋(Kuhn et al. , 2023; Ye et al., 2024)

What about generation?

the nurse left his bag on the floor. ⇨ a enfermeira deixou a bolsa no chão.

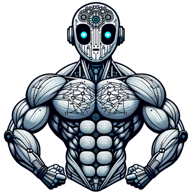


sample

❋(Kuhn et al. , 2023; Ye et al., 2024)

What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão.



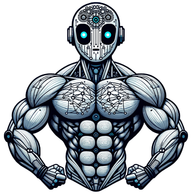
sample

a enfermeira deixou a bolsa no chão
a enfermeira deixou a bolsa no chão
a enfermeira deixou a sua bolsa no chão
o enfermeiro deixou a bolsa no chão
a enfermeira deixou a mochila dele no chão

❋(Kuhn et al. , 2023; Ye et al., 2024)

What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão.



sample

a enfermeira deixou a bolsa no chão
a enfermeira deixou a bolsa no chão
a enfermeira deixou a sua bolsa no chão
o enfermeiro deixou a bolsa no chão
a enfermeira deixou a mochila dele no chão

Sentence level uncertainty

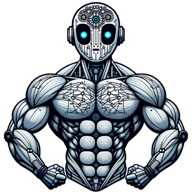
Access to output probabilities?
- ➡ Entropy-based uncertainty

No access to output probabilities?
- ➡ Deviation of output tokens
- ➡ Ask the model!

❋(Kuhn et al. , 2023; Ye et al., 2024)

What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão.



➡ sample ➡

a enfermeira deixou a bolsa no chão
a enfermeira deixou a bolsa no chão
a enfermeira deixou a sua bolsa no chão
o enfermeiro deixou a bolsa no chão
a enfermeira deixou a mochila dele no chão

Sentence level uncertainty

Access to output probabilities?

- ➡ Entropy-based uncertainty

No access to output probabilities?

- ➡ Deviation of output tokens
- ➡ Ask the model!

Sentence level conformal prediction

- ➡ As a sentence classification task
  - ○ Treat each sample as a label
- ➡ Use one of the uncertainty estimates as non-conformity

❋(Kuhn et al. , 2023; Ye et al., 2024)

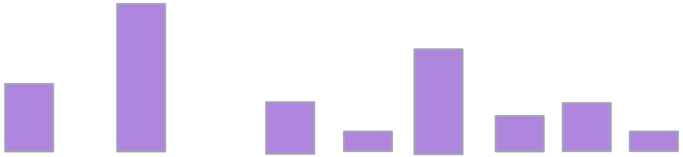
What about generation?

the nurse left his bag on the floor. ➪

❋(Ulmer et al., 2024)

What about generation?

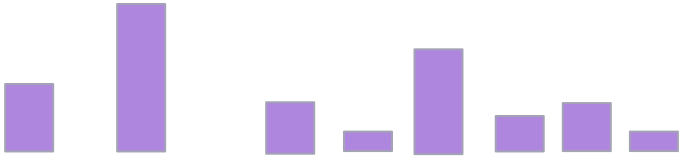
the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão .



❋(Ulmer et al., 2024)

What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão .



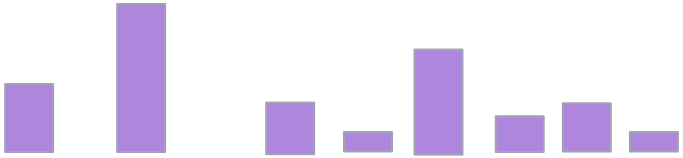
Word level uncertainty

- ➡ Output probabilities
- ➡ Entropy-based methods
- ➡ Sampling + semantic entropy

❋(Ulmer et al., 2024)

What about generation?

the nurse left his bag on the floor. ➪ a enfermeira deixou a bolsa no chão .



Word level uncertainty

- ➡ Output probabilities
- ➡ Entropy-based methods
- ➡ Sampling + semantic entropy

Word level conformal prediction

✗ exchangeability assumption

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

$$\mathbb{P}\Big(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\Big) \geq 1 - \alpha$$

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha$$

non ex.

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

$$\mathbb{P}\big(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\big) \geq 1 - \alpha$$

non ex.

$$\mathbb{P}\big(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\big) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \epsilon_i$$

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

coverage gap

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha$$

non ex.

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \epsilon_i$$

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

coverage gap

$$\mathbb{P}\big(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\big) \geq 1 - \alpha \quad \xrightarrow{\text{non ex.}} \quad \mathbb{P}\big(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\big) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \epsilon_i$$

We want this to be small!

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

coverage gap

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha$$

non ex.

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \epsilon_i$$

$$\epsilon_i = d_{TV}\left((x_i, y_i), (x_{test}, y_{test})\right)$$

We want this to be small!

… not that easy to compute

❋(Ulmer et al., 2024)

Conformalised Generation

Non-exchangeable CP bound (Barber et al., 2023)

coverage gap

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha$$

non ex.

$$\mathbb{P}\left(Y_{\text{test}} \in C_{\hat{q}}(X_{\text{test}})\right) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \epsilon_i$$

$$\epsilon_i = d_{TV}\left((x_i, y_i), (x_{test}, y_{test})\right)$$

We want this to be small!

… not that easy to compute

meaningful weights ⇒ small coverage gap

❋(Ulmer et al., 2024)

Conformalised Generation

Our solution:

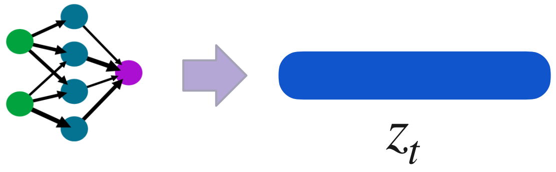
- Use the hidden representation of our LM
- Select a calibration set at every step of generation
- kNN to dynamically select the calibration set from a datastore
- distance metric to compute the weights

❋(Ulmer et al., 2024)

Conformalised Generation

Our solution:

- Use the hidden representation of our LM
- Select a calibration set at every step of generation
- kNN to dynamically select the calibration set from a datastore
- distance metric to compute the weights

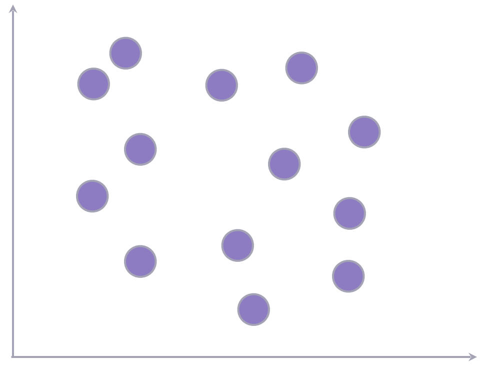


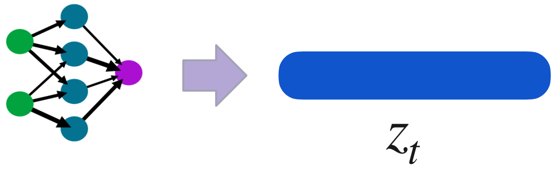
❋(Ulmer et al., 2024)

Conformalised Generation

Our solution:

- Use the hidden representation of our LM
- Select a calibration set at every step of generation
- kNN to dynamically select the calibration set from a datastore
- distance metric to compute the weights



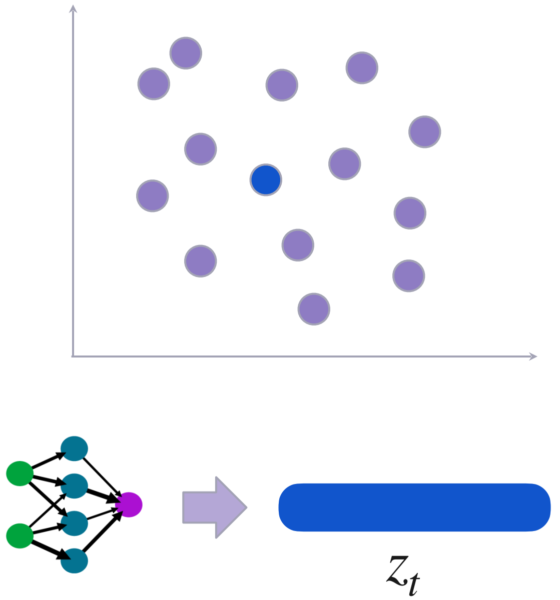


❋(Ulmer et al., 2024)

Conformalised Generation

Our solution:

- Use the hidden representation of our LM
- Select a calibration set at every step of generation
- kNN to dynamically select the calibration set from a datastore
- distance metric to compute the weights

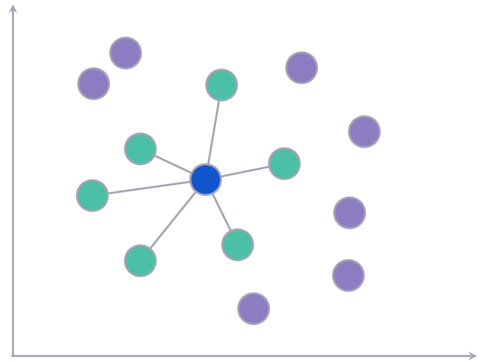


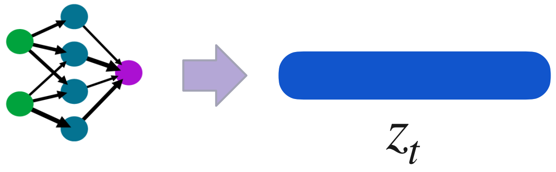
❋(Ulmer et al., 2024)

Conformalised Generation

Our solution:

- Use the hidden representation of our LM
- Select a calibration set at every step of generation
- kNN to dynamically select the calibration set from a datastore
- distance metric to compute the weights



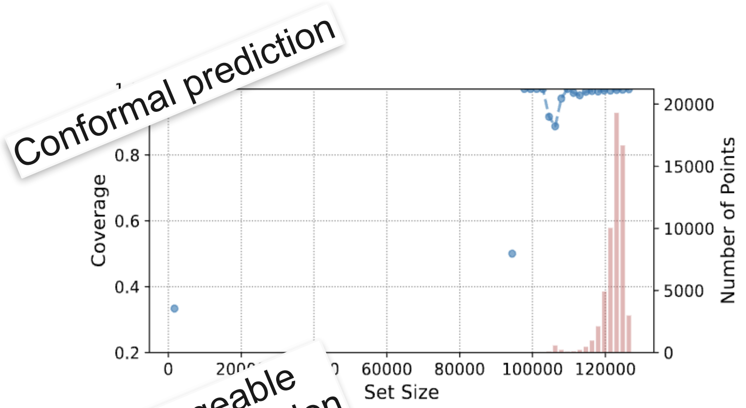


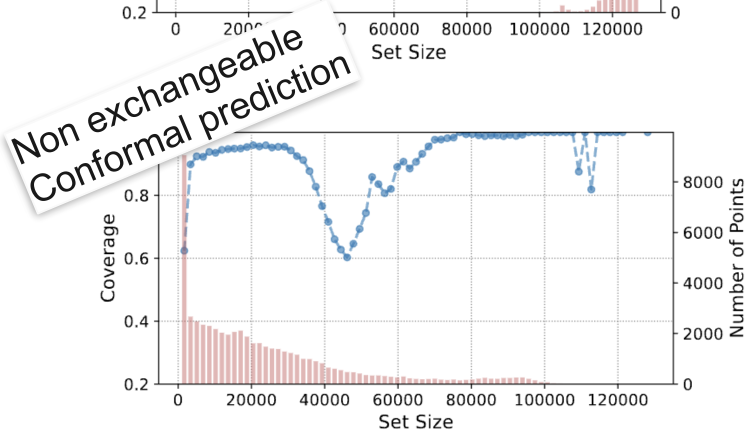
❋(Ulmer et al., 2024)

Conformalising Machine Translation

❋(Ulmer et al., 2024)

Conformalising Machine Translation

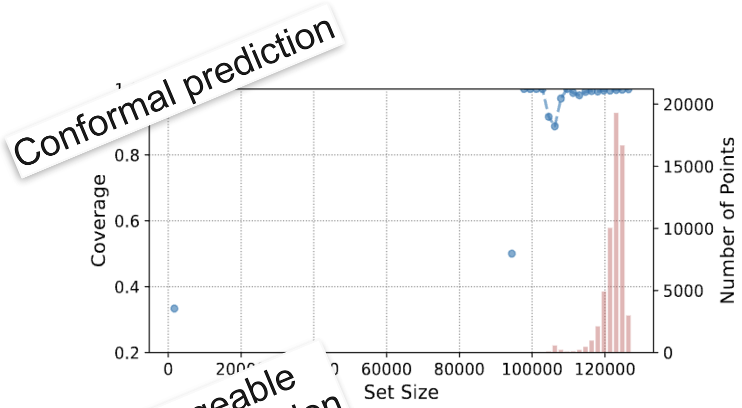


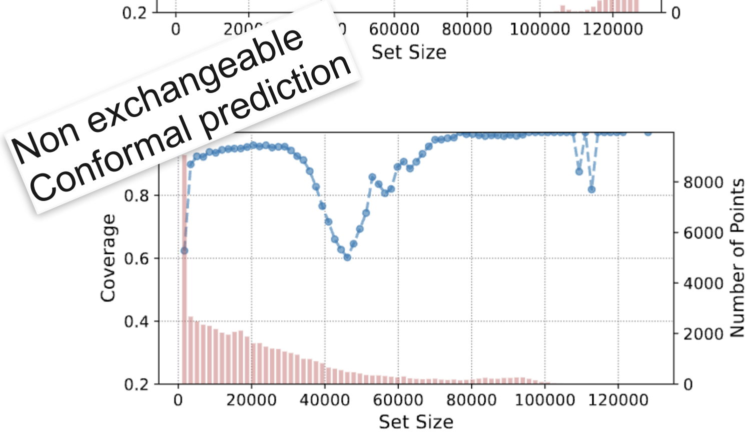


❋(Ulmer et al., 2024)

Conformalising Machine Translation

✓ Tighter confidence intervals
✓ Better “worst-case” coverage



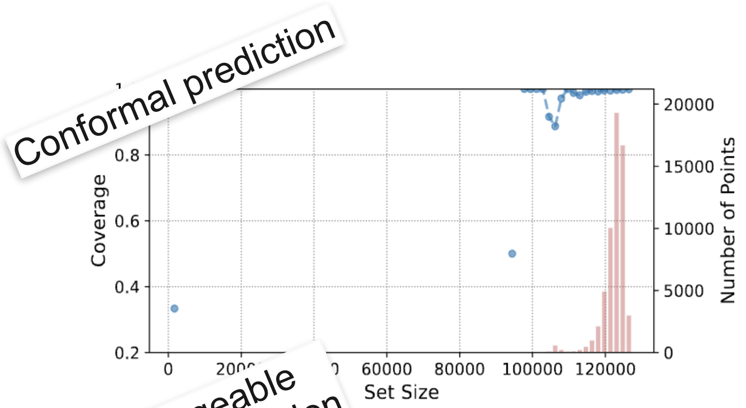


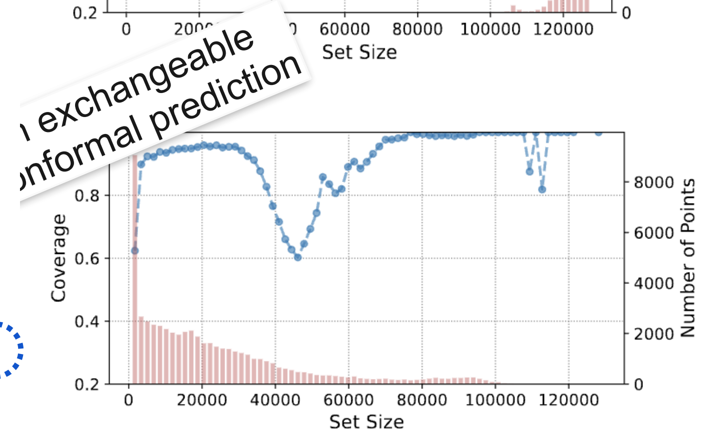
❋(Ulmer et al., 2024)

Conformalising Machine Translation

- ✓ Tighter confidence intervals
- ✓ Better "worst-case" coverage

- ✓ Comparable or even better performance to nucleus and top-k sampling





	En-De			En-Ja		
	BLEU	COMET	ChrF	BLEU	COMET	ChrF
Nucleus	27.63	0.89	54.8	10.61	0.59	36.52
Top-k	27.63	0.89	54.79	10.61	0.59	36.52
Conformal	27.63	0.89	54.8	10.61	0.59	36.52
Non-Ex Conformal	27.65	0.9	54.82	10.74	0.59	36.61

M2M100 - WMT 2022

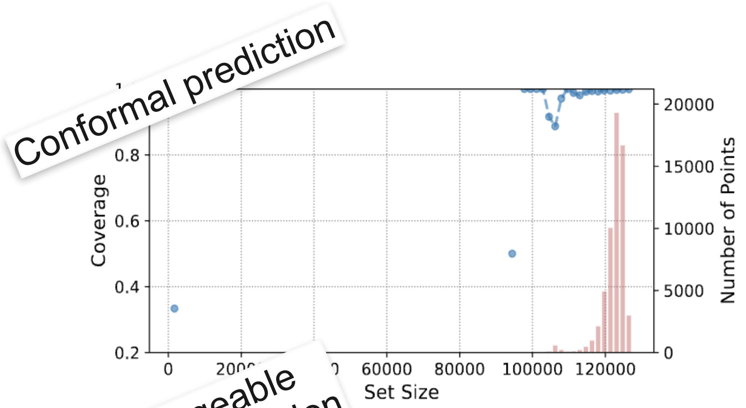
❋(Ulmer et al., 2024)

Conformalising Machine Translation

- ✓ Tighter confidence intervals
- ✓ Better "worst-case" coverage

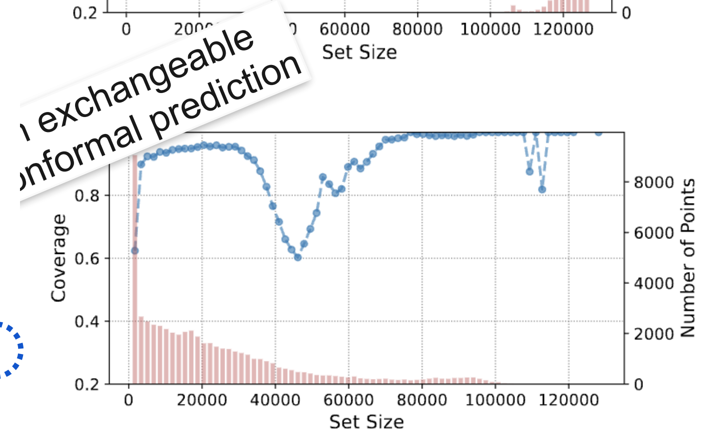
- ✓ Comparable or even better performance to nucleus and top-k sampling

- ✓ Robust to noise injection!



	En-De			En-Ja		
	BLEU	COMET	ChrF	BLEU	COMET	ChrF
Nucleus	27.63	0.89	54.8	10.61	0.59	36.52
Top-k	27.63	0.89	54.79	10.61	0.59	36.52
Conformal	27.63	0.89	54.8	10.61	0.59	36.52
Non-Ex Conformal	27.65	0.9	54.82	10.74	0.59	36.61

M2M100 - WMT 2022



❋(Ulmer et al., 2024)

Beyond coverage

We can calibrate for any loss function

❋(Farinhas et al., 2024)

Beyond coverage

We can calibrate for any loss function

monotone
bounded

❋(Farinhas et al., 2024)

Beyond coverage

We can calibrate for any loss function

monotone
bounded

- False negative rate
- Token-level F1 score
- λ-insensitive absolute loss

❋(Farinhas et al., 2024)

Beyond coverage

We can calibrate for any loss function

monotone
bounded

- ✤ False negative rate
- ✤ Token-level F1 score
- ✤ λ-insensitive absolute loss

Robust method

- ✓ Distribution shifts
- ✓ Changepoints

❋(Farinhas et al., 2024)

Beyond coverage

We can calibrate for any loss function

monotone

bounded

- ✤ False negative rate
- ✤ Token-level F1 score
- ✤ λ-insensitive absolute loss

Robust method

- ✓ Distribution shifts
- ✓ Changepoints

width **adapted** to the distribution shifts while maintaining performance for the controlled value

❋(Farinhas et al., 2024)

Beyond coverage

We can calibrate for any loss function

⋮

monotone
bounded

- ✤ False negative rate
- ✤ Token-level F1 score
- ✤ λ-insensitive absolute loss

Robust method

- ✓ Distribution shifts
- ✓ Changepoints

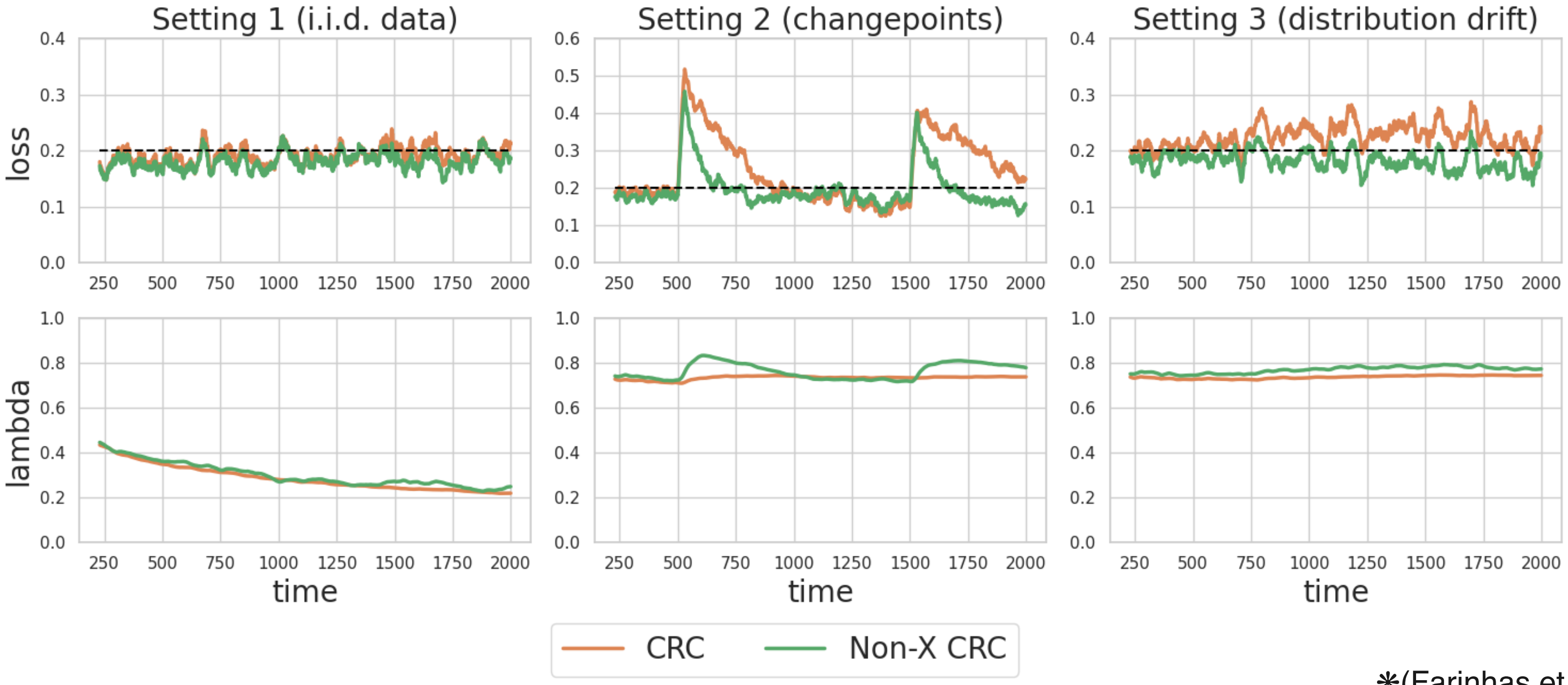
Efficient method

- ✓ Tighter prediction sets

width **adapted** to the distribution shifts while maintaining performance for the controlled value

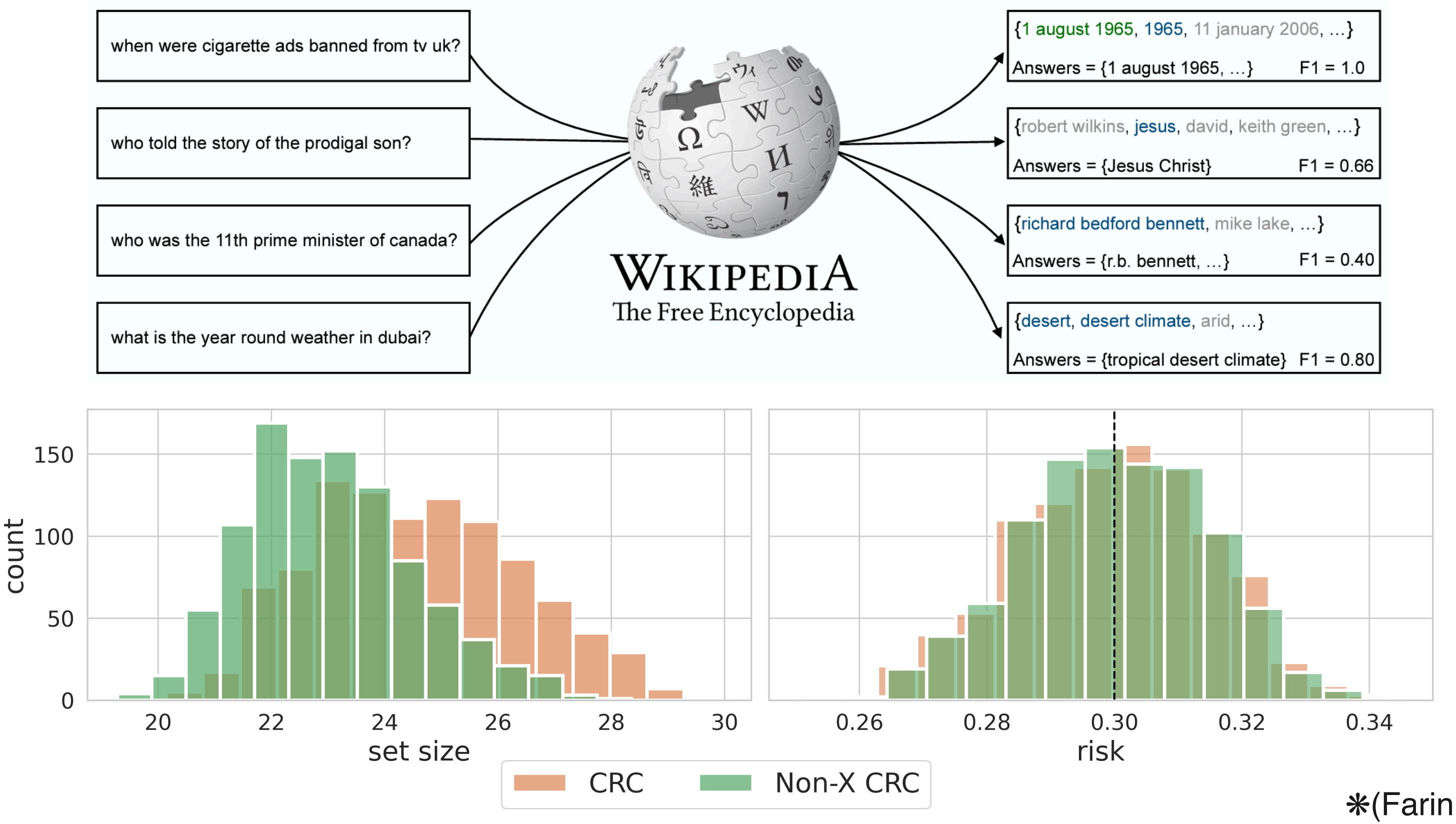
❋(Farinhas et al., 2024)

Simulated time-series data



❋(Farinhas et al., 2024)

Open QA

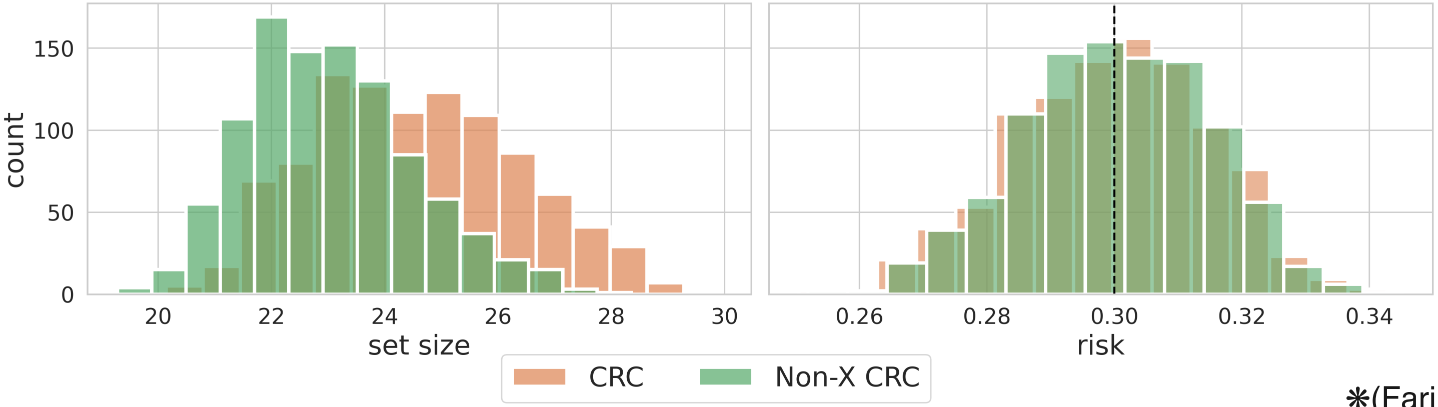


Token level
F1-score

*(Farinhas et al., 2024)

Open QA

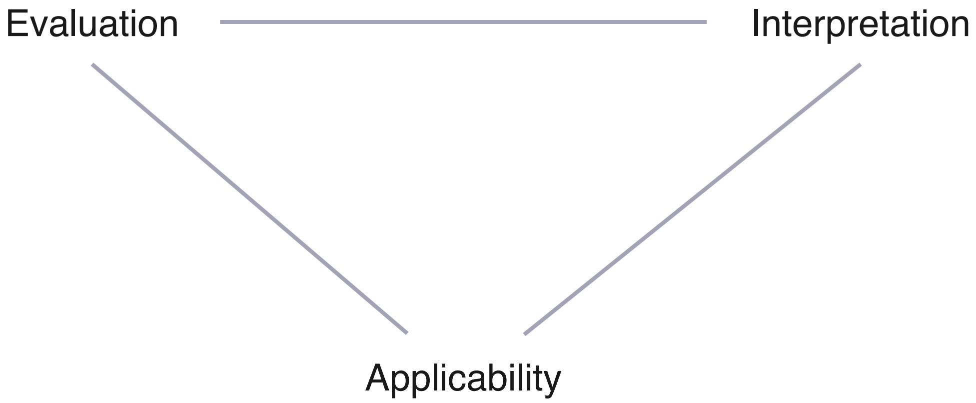
Token level F1-score



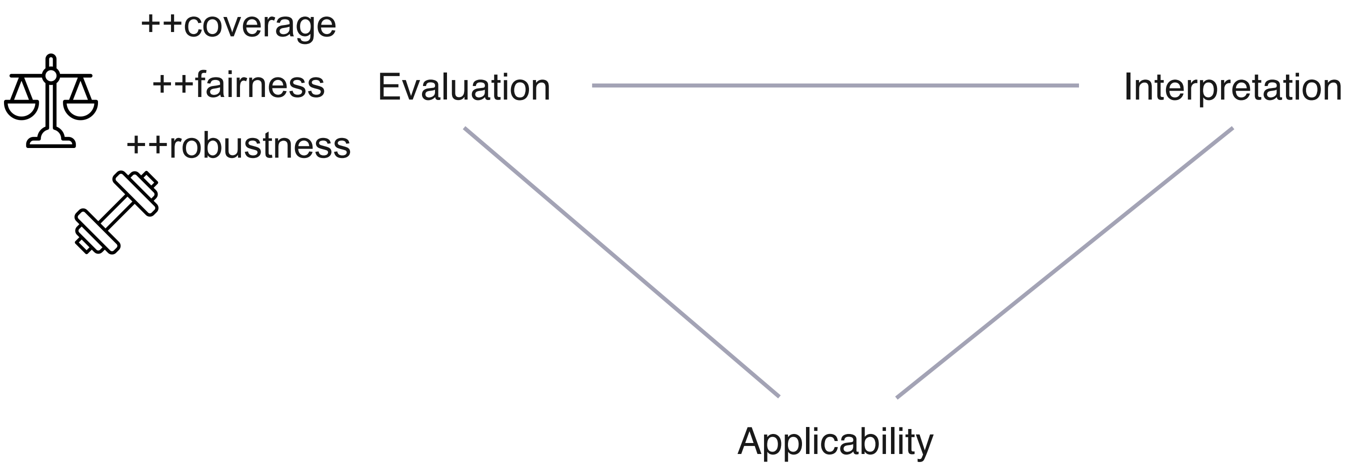
❋(Farinhas et al., 2024)

Conformal prediction

Conformal prediction



Conformal prediction



Conformal prediction

++coverage

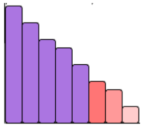
++fairness

++robustness

Evaluation

Interpretation

prediction sets



Allow for further analysis and interpretation

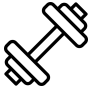
Applicability

Conformal prediction

++coverage

++fairness

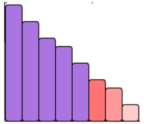
++robustness



Evaluation

Interpretation

prediction sets



Allow for further analysis and interpretation

Applicability

non-parametric

flexible calibration target

Conformal prediction

++coverage

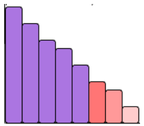
++fairness

++robustness

Evaluation

Interpretation

prediction sets



Allow for further analysis and interpretation

Applicability

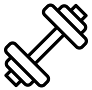
non-parametric

flexible calibration target

Efficiency?

Better calibration weights?

Conformal prediction



++coverage

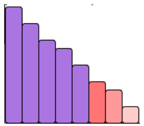
++fairness

++robustness

Evaluation

Interpretation

prediction sets



Allow for further analysis and interpretation

Applicability

Efficiency?

Better calibration weights?

non-parametric

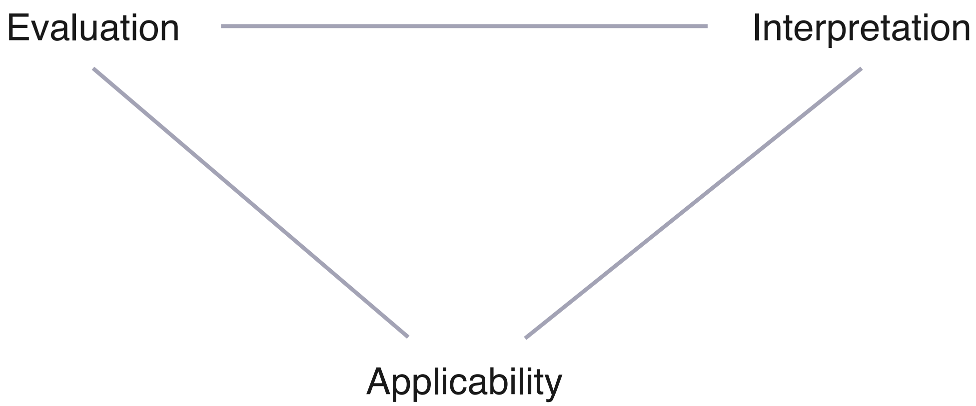
flexible calibration target

Different losses?

Interpretation of output?

Overall

Towards a more accessible version of uncertainty



Thank you!

References

1. Angelopoulos, A.N. and Bates, S., 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*
2. Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H.S., Fernández, R., Plank, B., Sennrich, R., Zerva, C. and Aziz, W., 2023. Uncertainty in natural language generation: From theory to applications. arXiv preprint arXiv:2307.15703.
3. Baan, J., Aziz, W., Plank, B. and Fernández, R., 2022, December. Stop Measuring Calibration When Humans Disagree. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 1892-1915).
4. Barber, R.F., Candes, E.J., Ramdas, A. and Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*, *51*(2), pp.816-845.
5. Farinhas, A., Zerva, C., Ulmer, D. and Martins, A.F., 2023. Non-exchangeable conformal risk control. arXiv preprint arXiv:2310.01262.
6. Giulianelli, M., Baan, J., Aziz, W., Fernández, R. and Plank, B., 2023, December. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 14349-14371).
7. Glushkova, T., Zerva, C., Rei, R. and Martins, A.F., 2021, November. Uncertainty-Aware Machine Translation Evaluation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3920-3938).
8. Kuhn, L., Gal, Y. and Farquhar, S., 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
9. Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. Advances in neural information processing systems, 31.
10. Ulmer, D., Zerva, C. and Martins, A.F., 2024. Non-Exchangeable Conformal Language Generation with Nearest Neighbors. *arXiv preprint arXiv:2402.00707*.
11. Xiao, Y., & Wang, W. Y. (2021, April). On Hallucination and Predictive Uncertainty in Conditional Language Generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 2734-2744).
12. Ye, F., Yang, M., Pang, J., Wang, L., Wong, D.F., Yilmaz, E., Shi, S. and Tu, Z., 2024. Benchmarking LLMs via Uncertainty Quantification. *arXiv preprint arXiv:2401.12794*.
13. Zerva, C. and Martins, A.F., 2023. Conformalizing machine translation evaluation. arXiv preprint arXiv:2306.06221.
14. Zerva, C., Glushkova, T., Rei, R. and Martins, A.F., 2022, December. Disentangling Uncertainty in Machine Translation Evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8622-8641).
15. Zerva, C. and Martins, A.F., 2023. Conformalizing machine translation evaluation. *arXiv preprint arXiv:2306.06221*.