Оценка неопределенности в NLP: от классификации к большим языковым моделям

Артём Важенцев

Научный сотрудник AIRI

Related Material

- **Tutorials:**
  - Uncertainty Estimation for Natural Language Processing. Adam Fisch, Robin Jia, Tal Schuster. COLING-2022.
  - Practical Uncertainty Estimation and Out-of-Distribution Robustness in Deep Learning. Dustin Tran, Balaji Lakshminarayanan, Jasper Snoek. NeurIPS-2020.
  - Uncertainty Quantification for Large Language Models. Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. ACL 2025

- **Workshops:**
  - UncertaiNLP @ EACL-2024 & EMNLP-2025
  - QUESTION @ ICLR-2025

- **Benchmark:** Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. Vashurin et al. @ TACL-2025

01

Background on Uncertainty Quantification

Why we need to estimate uncertainty of model predictions?

Consider we have a trained neural network model for **binary classification**

$$P(y = 1|x) = 0.9$$
$$y_{true} = 1$$

$$P(y = 1|x) = 0.2$$
$$y_{true} = 0$$

AIRI

Why we need to estimate uncertainty of model predictions?

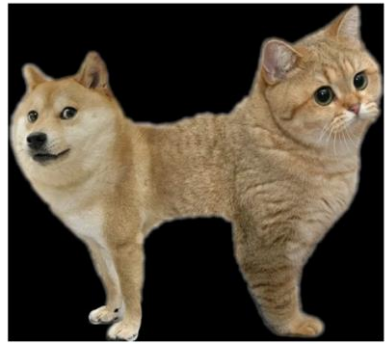
Consider we have a trained neural network model for **binary classification**

$P(y = 1|x) = 0.9$

$y_{true} = 1$

$P(y = 1|x) = 0.2$

$y_{true} = 0$



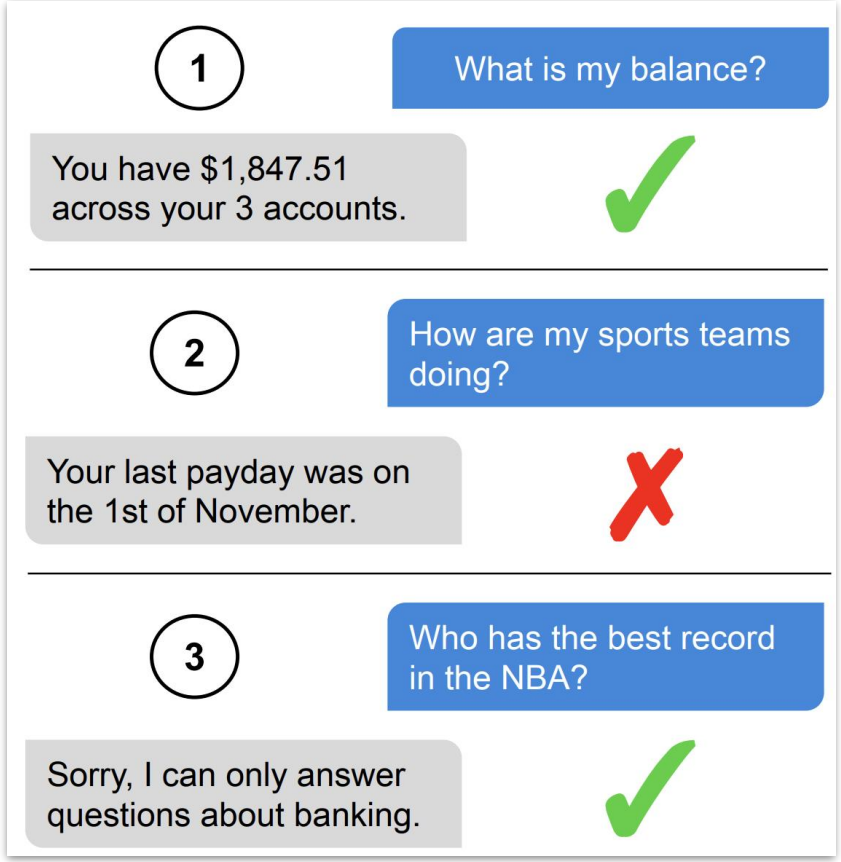
$\boldsymbol{P(y = 1|x) = 0.8}$

$y_{true} = ???$

AIRI

Background

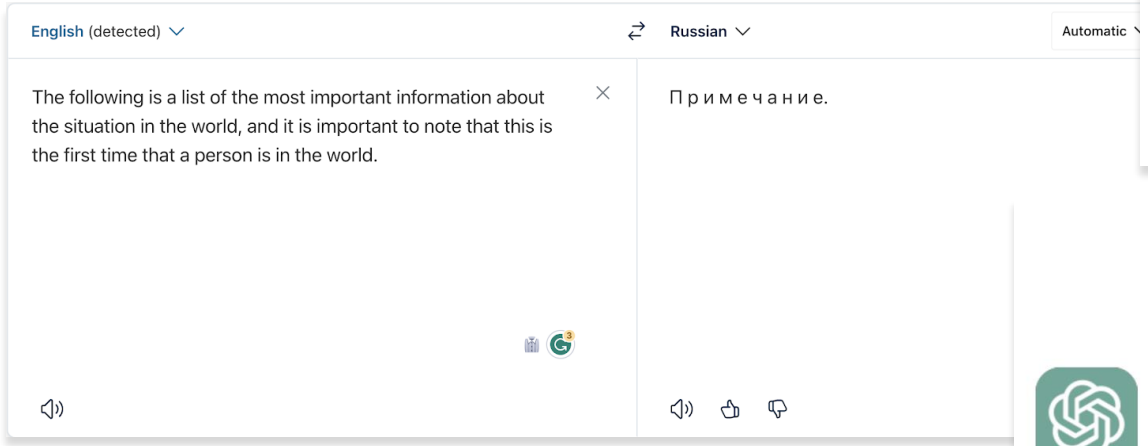
- Uncertainty Quantification (UQ) is a subfield of machine learning that seeks to model the degree to which model predictions can be trusted.
- Ideally, the instances with the **most uncertain predictions should correspond to errors**.

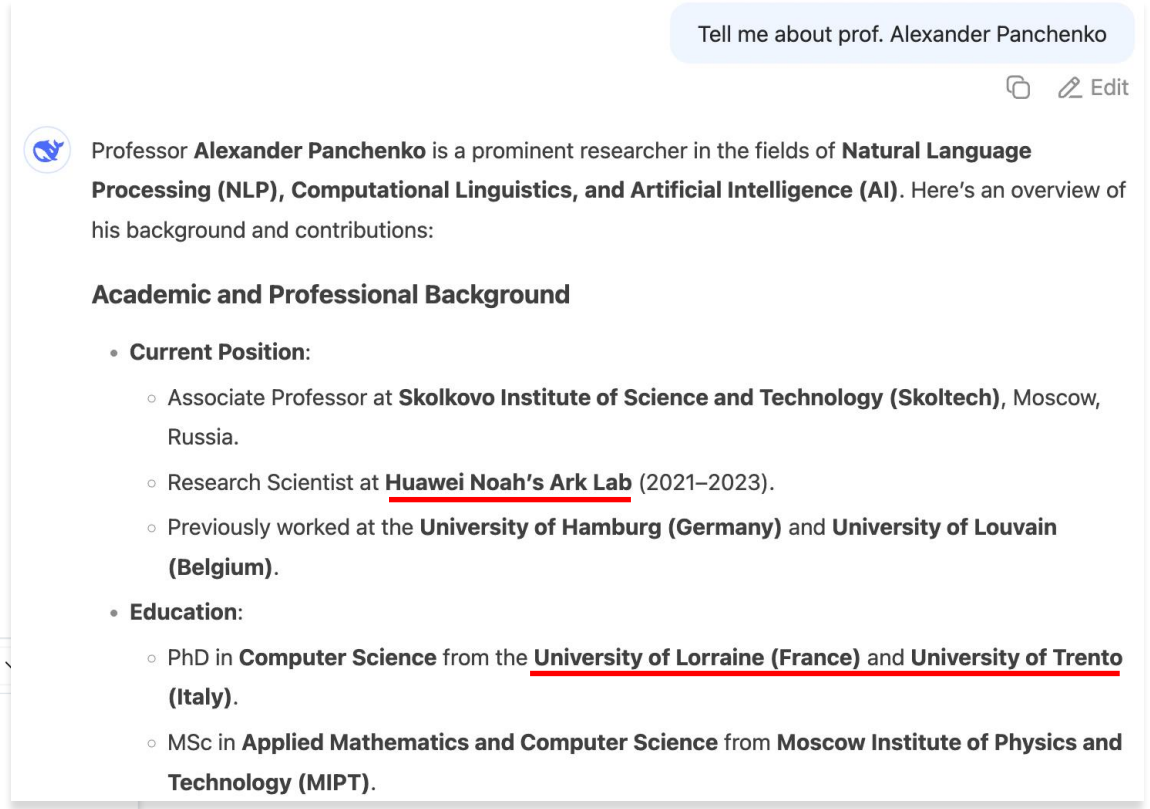


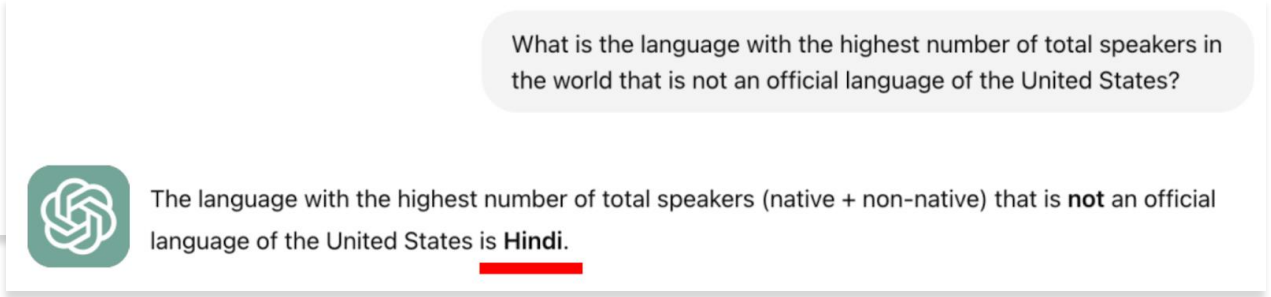
Larson et al. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. EMNLP-IJCNLP 2019.

LLMs Hallucinate

- LLMs deliver impressive performance, yet still generate **hallucinations and incorrect facts.**
- Such errors hinder their use in safety-critical applications.

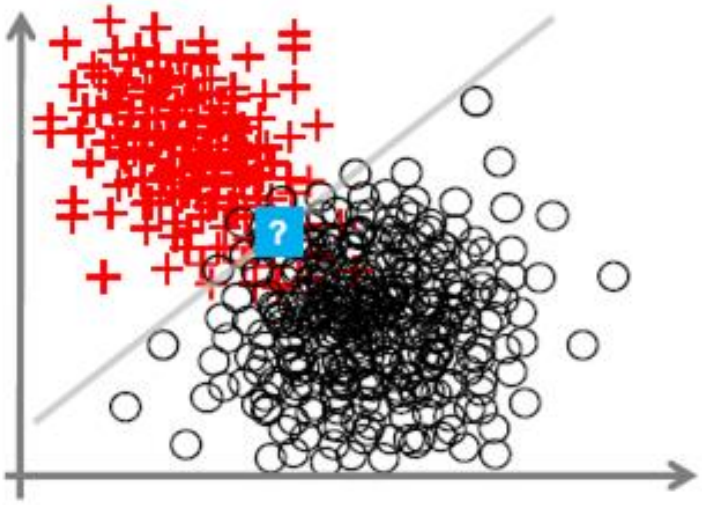




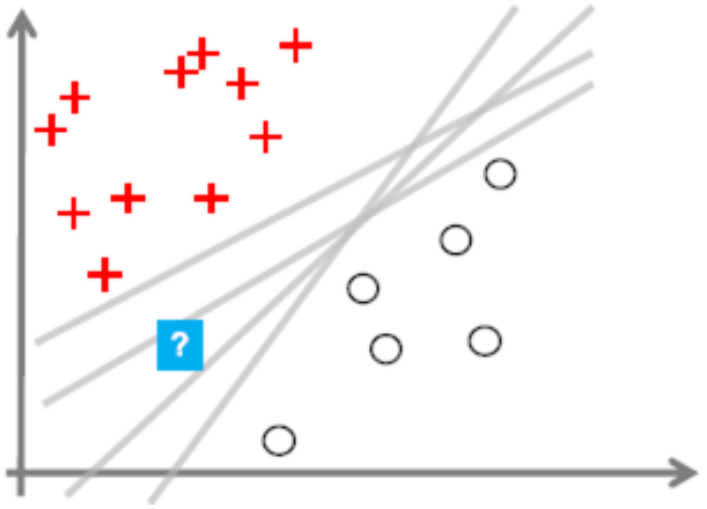


Notion of Uncertainty

Uncertainty quantification in deep learning focuses on analyzing and quantifying uncertainty to improve the reliability of model predictions.



Aleatoric uncertainty: occurs from ambiguity, randomness, and noise in data.



Epistemic uncertainty: pertains to a lack of knowledge about model parameters

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. 2011.

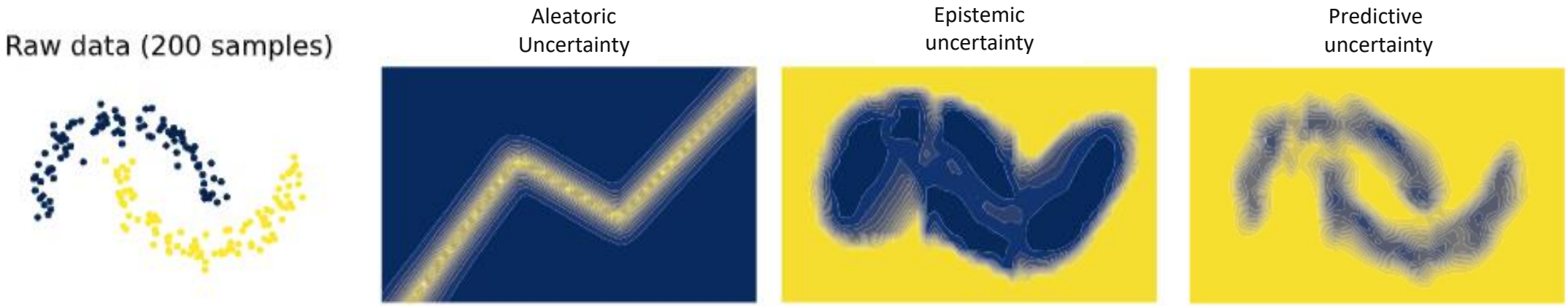
What is Uncertainty?

- There is no unified way for specifying uncertainty scores. They can be measured in various ways: **distances, probabilities, entropy, error**, etc.
- Information theory / Bayesian statistics provides a principled way of measuring uncertainty. It is an **entropy of a probability distribution**.

Two Sources of Uncertainty

$$U_{pred} :\triangleq U_{epistemic} + U_{aleatoric}$$

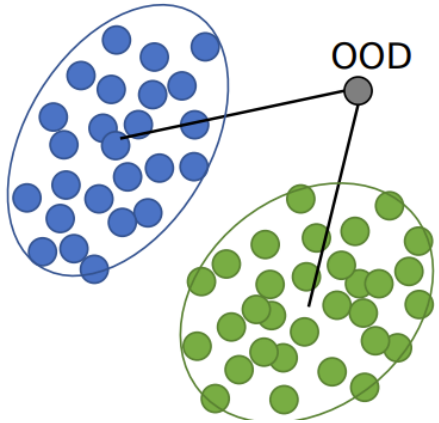
$$H(Y|x, D) = I(Y, W|x, D) + E_{w\sim p(W|D)}[H(Y|x, w)]$$



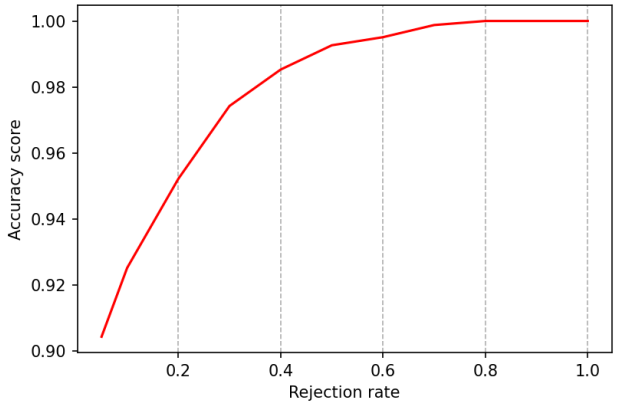
Applications of Uncertainty

Uncertainty quantification methods play a crucial role in various practical applications:

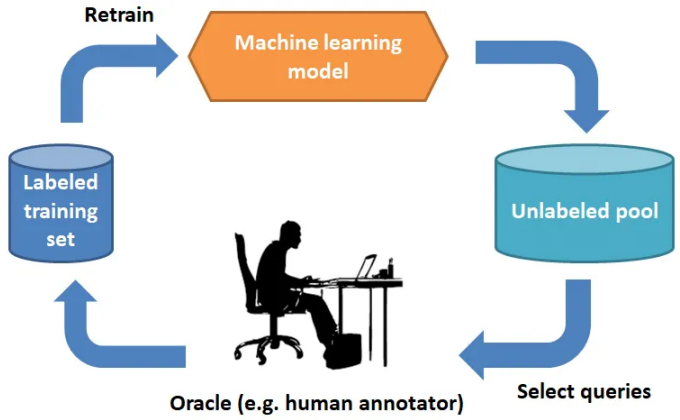
- Out-of-distribution (OOD) detection



- Selective prediction



- Active learning



02

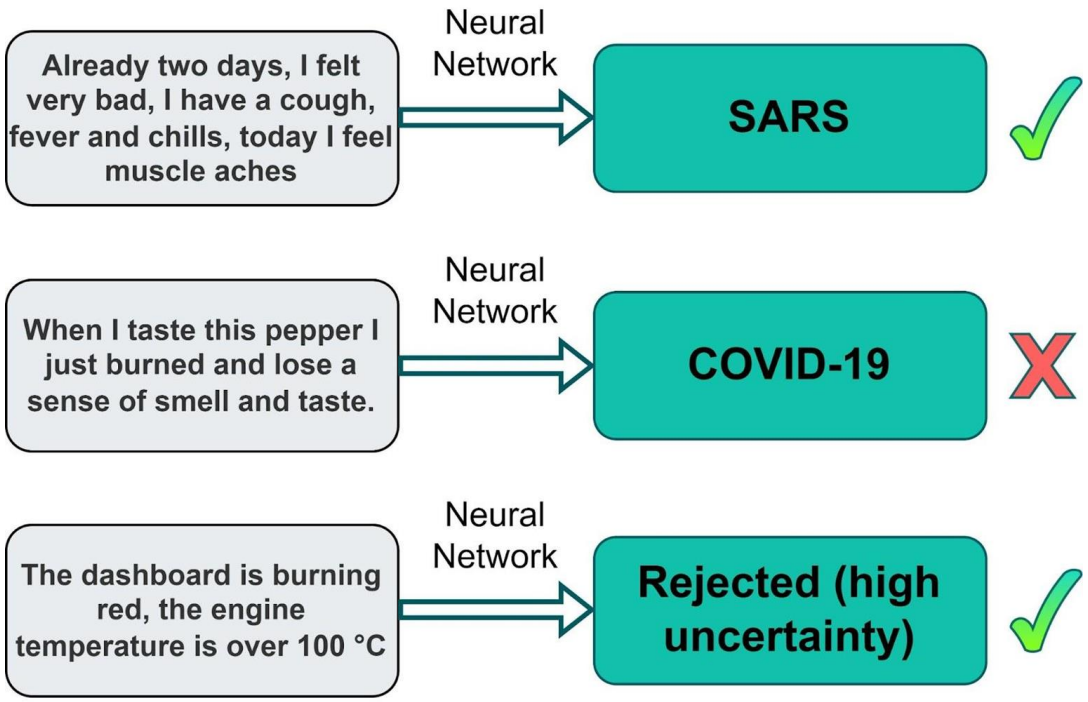
UQ for Text Classification Models

Problem Statement

Selective classification aims not only to make the prediction for a given instance but also to estimate the model's uncertainty associated with that prediction.

Applications:

- → hate speech detection in social networks
- → medical diagnostics



Softmax Response

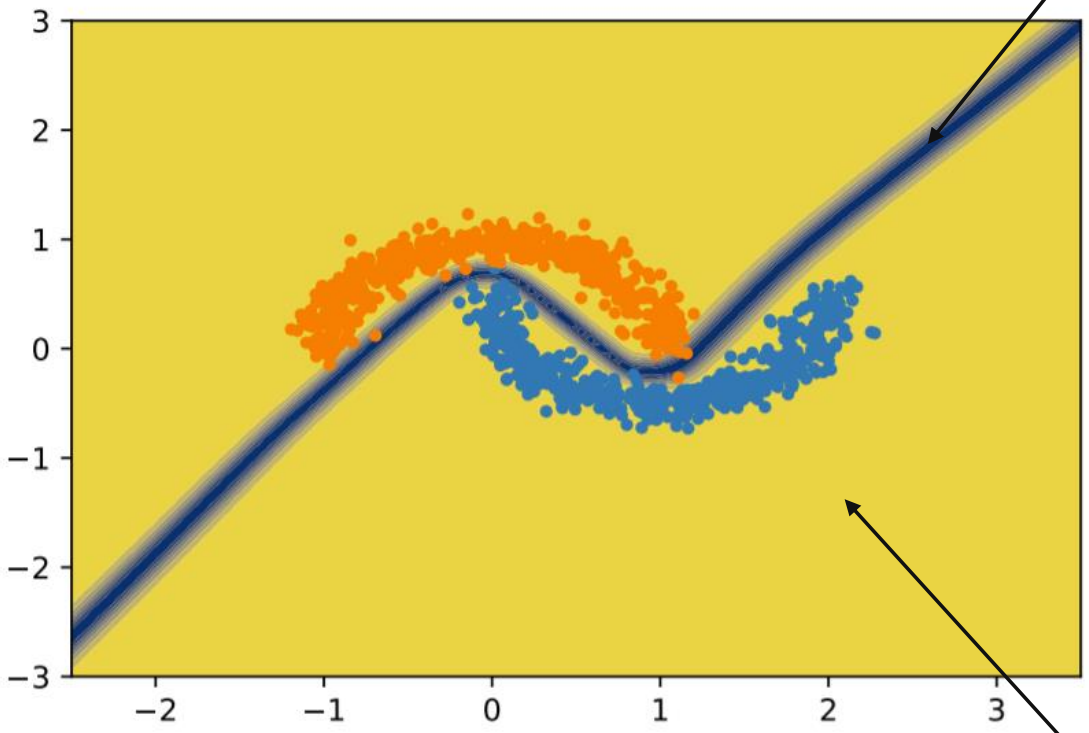
Softmax response (SR) is a trivial baseline for UE a trained model that uses the probabilities generated via the output softmax layer of the neural network. The smaller this maximum probability is, the more uncertain model is:

$$u_{SR}(x) = 1 - \max_{c \in C} p(y = c \mid x),$$

where $p(y = c \mid x)$ - probability of sample $\boldsymbol{x}$ belong to class $y = c \in C$.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. NeurIPS 2017.

Why Simple Softmax Probabilities are Bad for UQ?



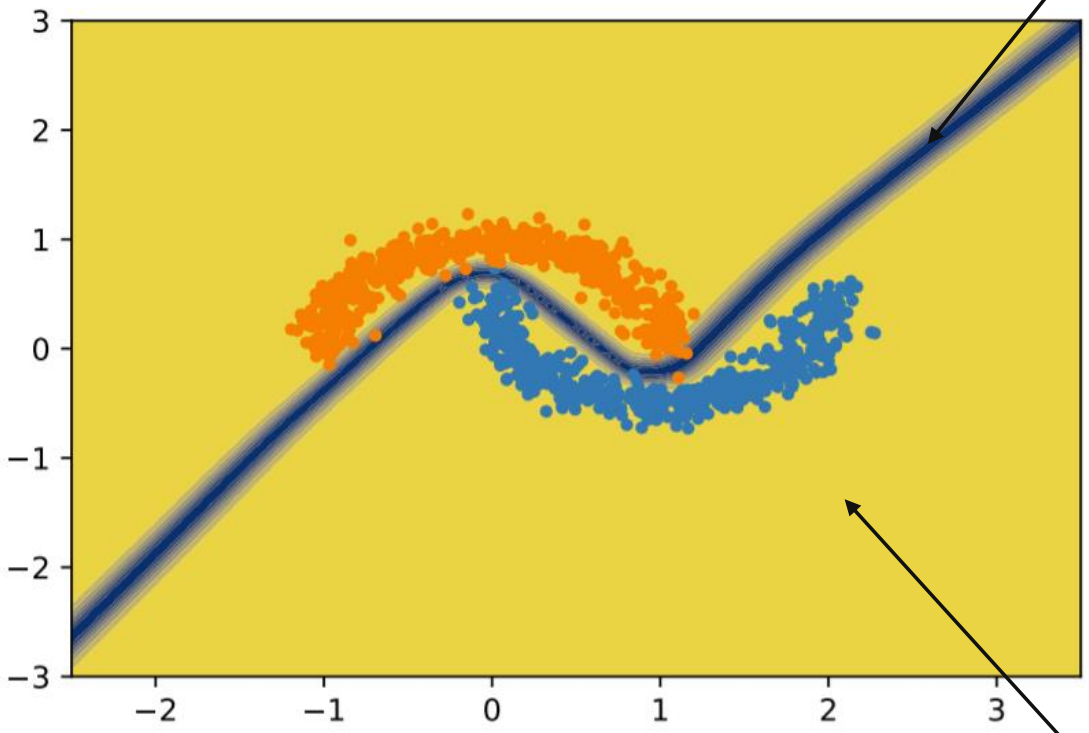
- High uncertainty

What we get usually from softmax

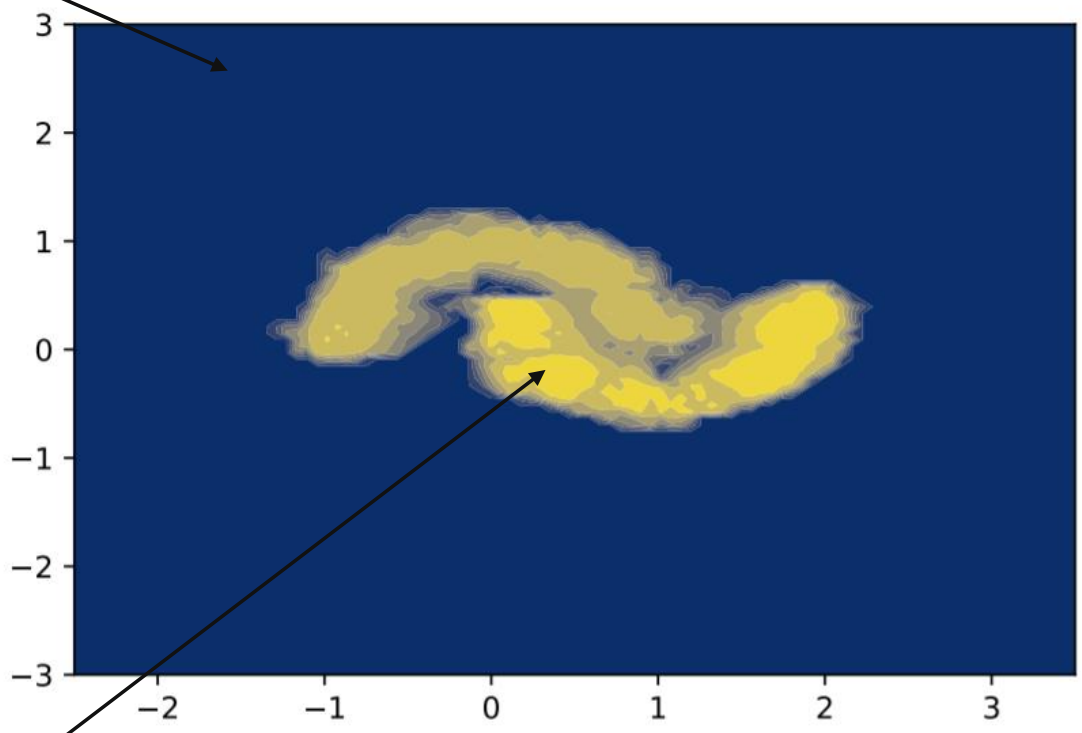
- Low uncertainty

Why Simple Softmax Probabilities are Bad for UQ?

- High uncertainty



What we get usually from softmax



What we would like to have

- Low uncertainty

AIRI

Deep Ensemble

Consider we have conducted T independent models. We can use the following ways to quantify uncertainty with the standard **Deep Ensemble**:

Overhead in:
- **memory footprint**
- **inference time**
- **training time**

Sampled maximum probability (SMP)

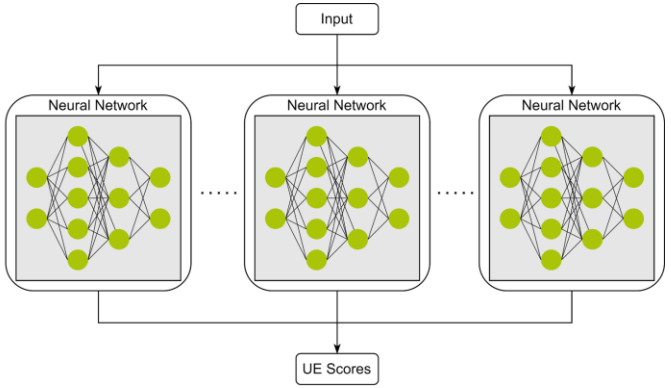
$$u_{\text{SMP}} = 1 - \max_{c \in C} \frac{1}{T} \sum_{t=1}^{T} p_t^c,$$

where p_t^c is the probability of the class c for the t-th stochastic forward pass.

Probability variance (PV)

$$u_{\text{PV}} = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{T} \sum_{t=1}^{T} \left(p_t^c - \overline{p^c} \right)^2 \right),$$

where $\overline{p^c} = \frac{1}{T} \sum_t p_t^c$ the probability for a class c averaged across T stochastic forward passes.



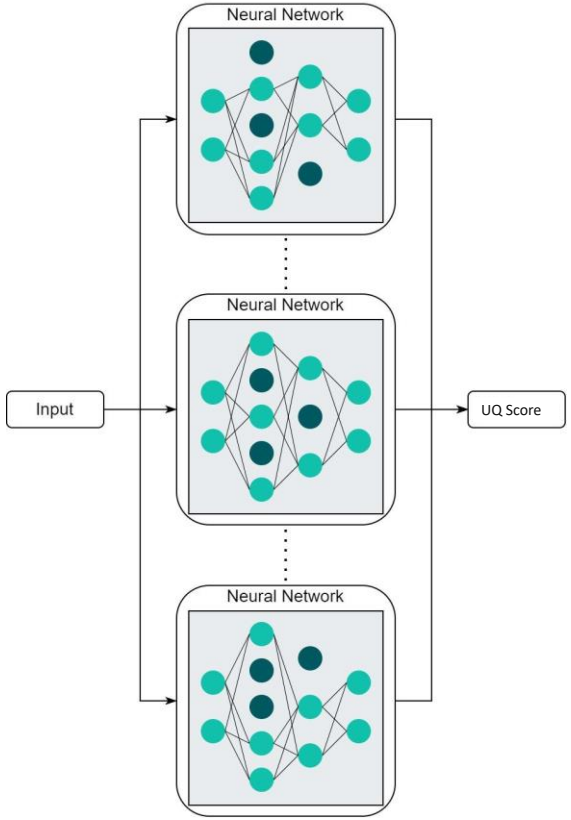
Bayesian active learning by disagreement (BALD)

$$u_{\text{BALD}} = -\sum_{c=1}^{C} \overline{p^c} \log \overline{p^c} + \frac{1}{T} \sum_{c,t} p_t^c \log p_t^c$$

Lakshminarayanan et al. Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS 2017.

Monte Carlo Dropout

Consider we have conducted T stochastic forward passes. We use the following ways to quantify uncertainty with the standard **MC dropout**:



Sampled maximum probability (SMP)

$$u_{\text{SMP}} = 1 - \max_{c \in C} \frac{1}{T} \sum_{t=1}^{T} p_t^c,$$

where p_t^c is the probability of the class c for the t-th stochastic forward pass.

Probability variance (PV)

$$u_{\text{PV}} = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{T} \sum_{t=1}^{T} \left(p_t^c - \overline{p^c} \right)^2 \right),$$

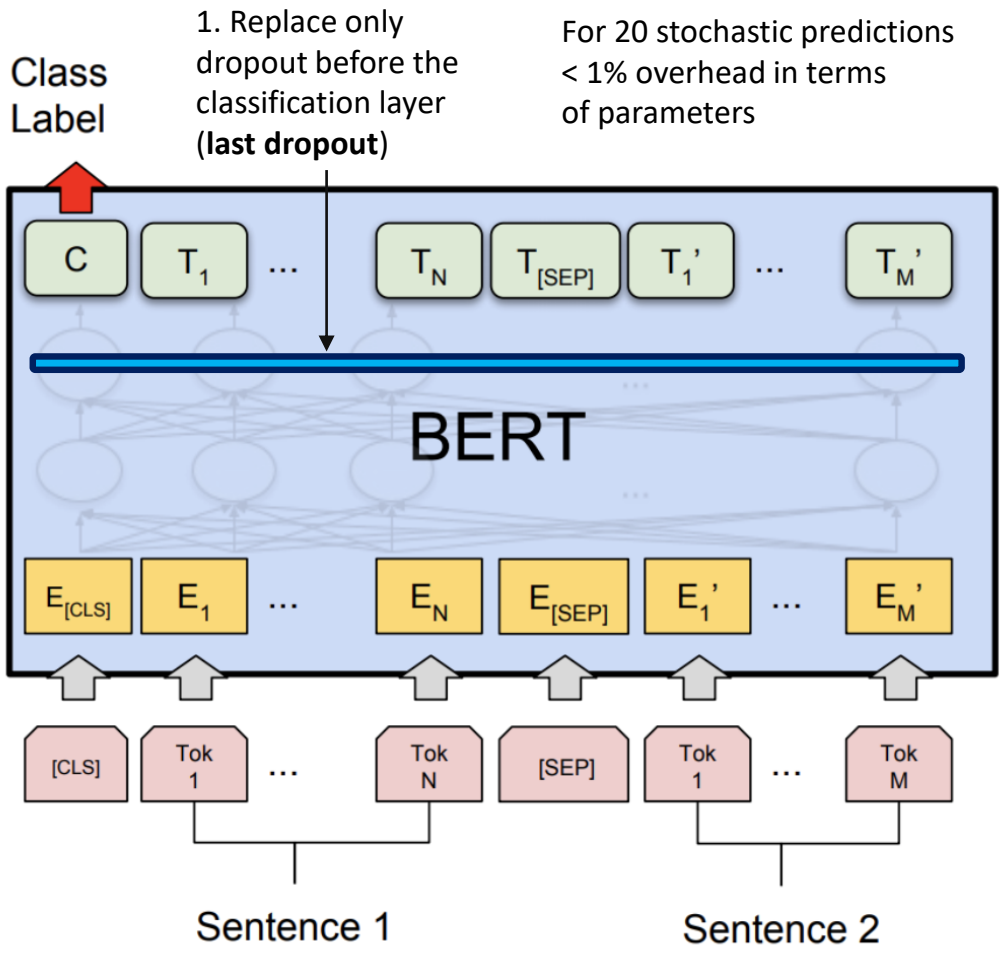
where $\overline{p^c} = \frac{1}{T} \sum_t p_t^c$ the probability for a class c averaged across T stochastic forward passes.

Bayesian active learning by disagreement (BALD)

$$u_{\text{BALD}} = -\sum_{c=1}^{C} \overline{p^c} \log \overline{p^c} + \frac{1}{T} \sum_{c,t} p_t^c \log p_t^c$$

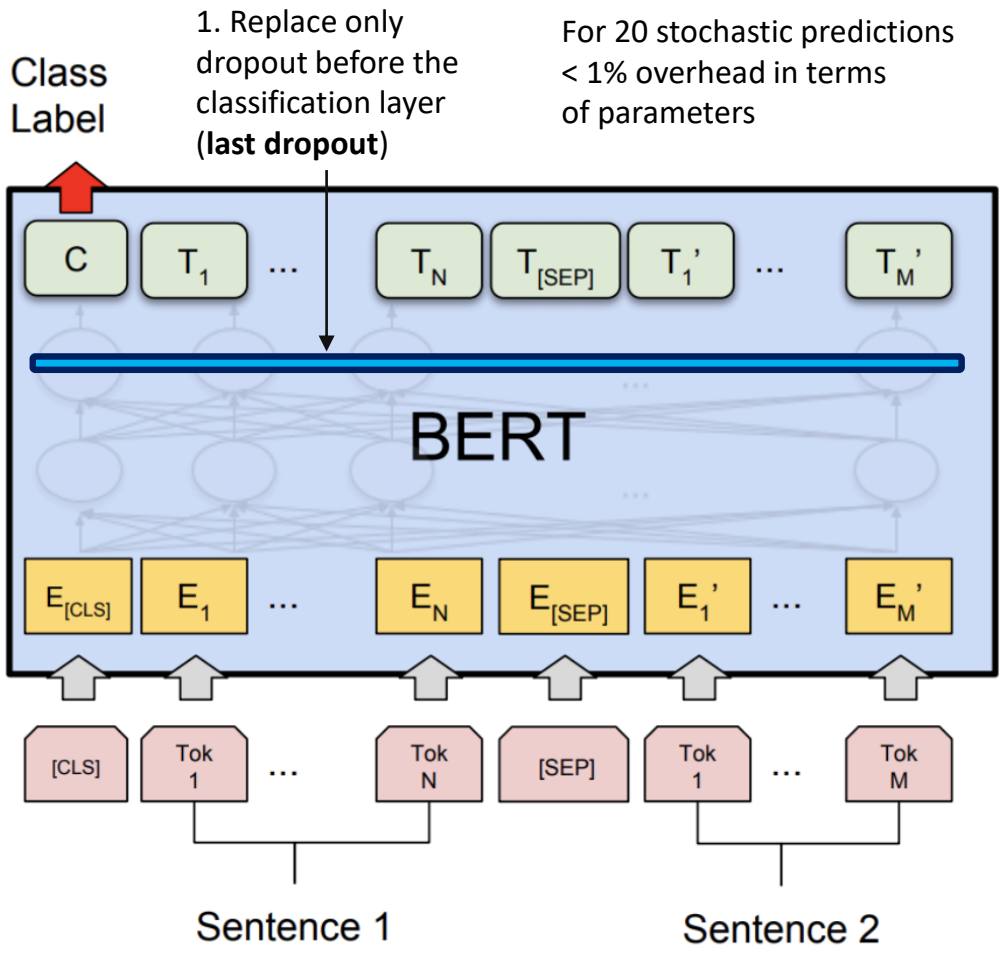
Gal et al. Deep bayesian active learning with image data. ICML 2017.

MC Dropout Options in Transformers



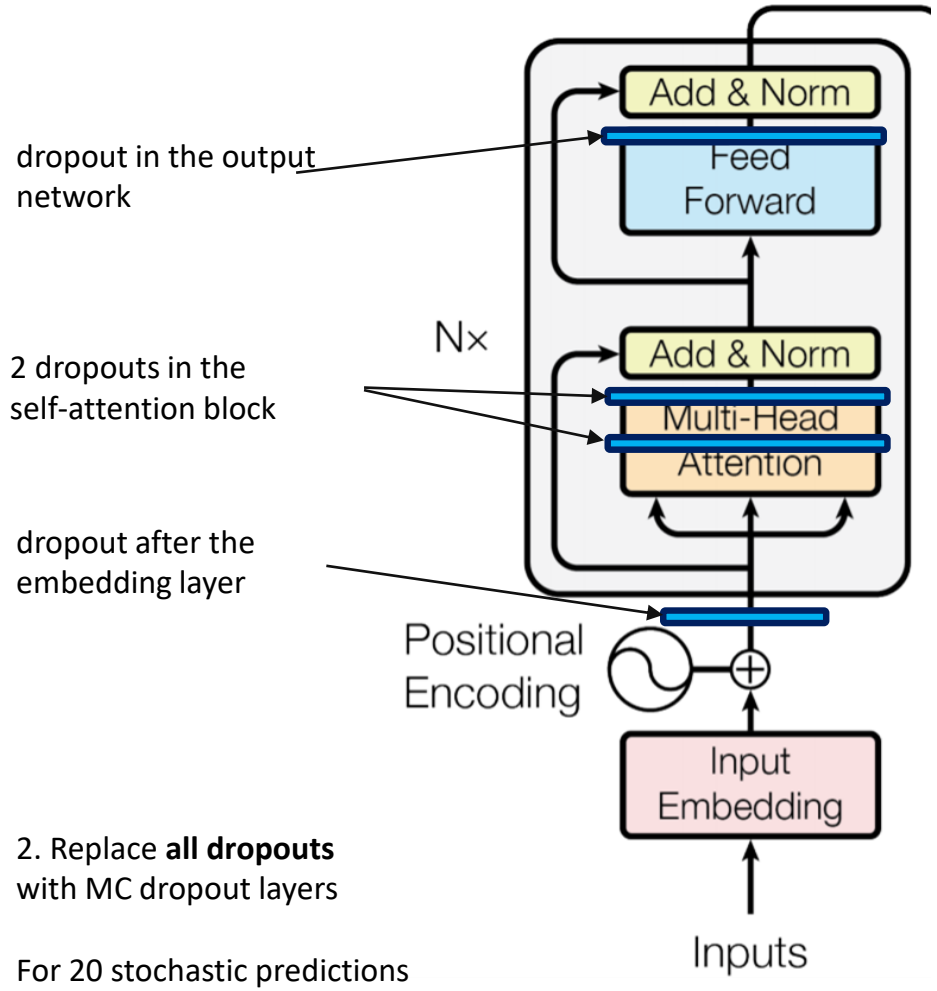
Shelmanov et al. How certain is your Transformer? EACL 2021.

MC Dropout Options in Transformers



1. Replace only dropout before the classification layer (**last dropout**)

For 20 stochastic predictions
< 1% overhead in terms of parameters



2. Replace **all dropouts** with MC dropout layers

For 20 stochastic predictions
~ 1900% overhead

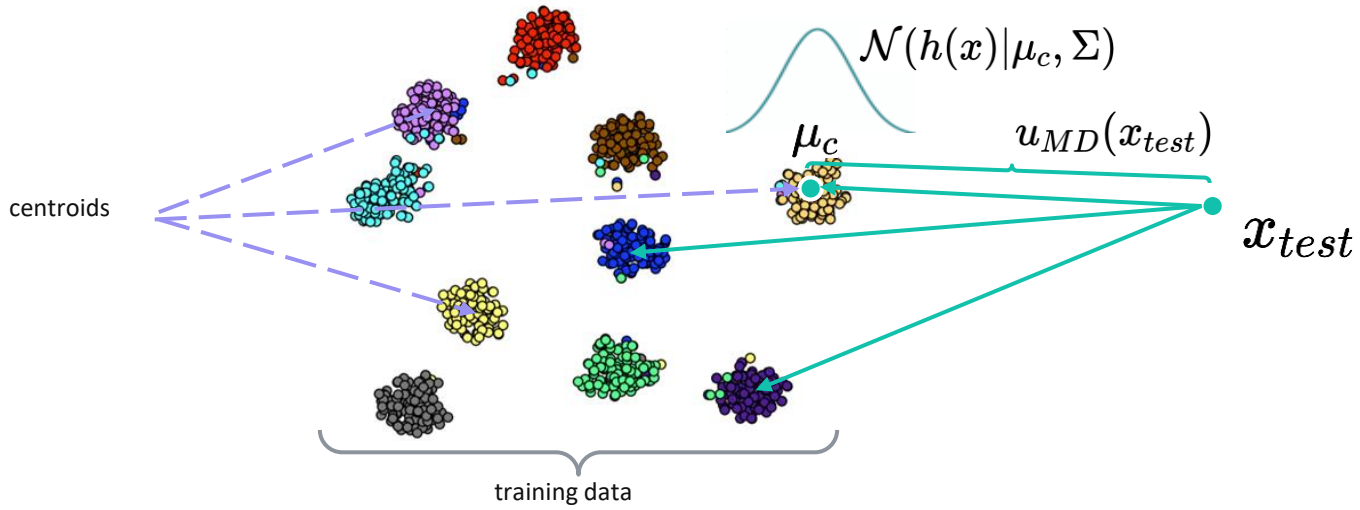
Shelmanov et al. How certain is your Transformer? EACL 2021.

Mahalanobis Distance

Mahalanobis distance (MD) is proportional to the negative log-likelihood of a multivariate normal distribution, up to an additive constant:

$$u_{MD} = \min_{c \in C} (h_i - \mu_c)^T \Sigma^{-1} (h_i - \mu_c),$$

where h_i is a hidden representation of a i-th instance, μ_c is a centroid of a class c, and Σ is a covariance matrix for hidden representations of training instances.



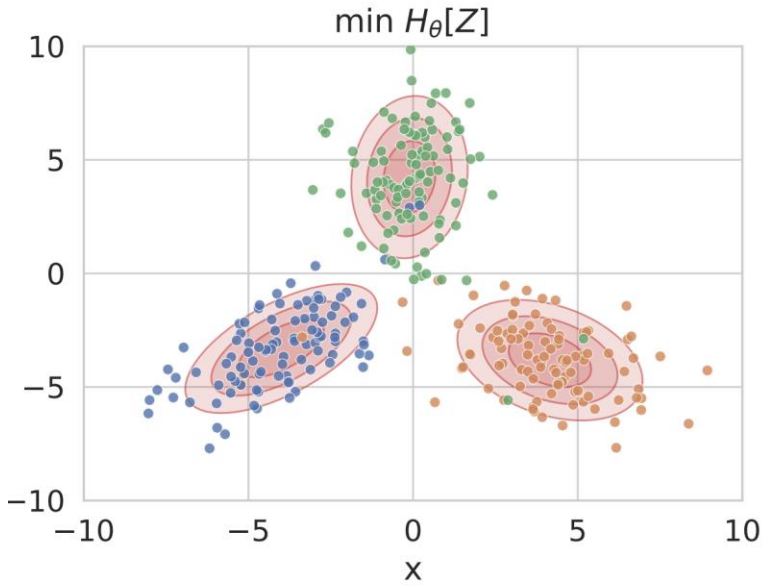
Lee et al. A simple unified framework for detecting outof-distribution samples and adversarial attacks. NeurIPS 2018.

Deep Deterministic Uncertainty

$$\tilde{U}_{\mathrm{E}}^{\mathrm{DDU}}(\mathbf{x}) = \sum_{c \in C} p(h(\mathbf{x}) \mid y = c) p(y = c)$$

$$p(h(\mathbf{x}) \mid y = c) \sim \mathcal{N}(h(\mathbf{x}) \mid \mu_c, \Sigma_c)$$

$$p(y = c) = \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \mathbf{1}[y_i = c]}{|\mathcal{D}|}$$



GMM with 3 components fitted to a synthetic dataset with 3 different classes

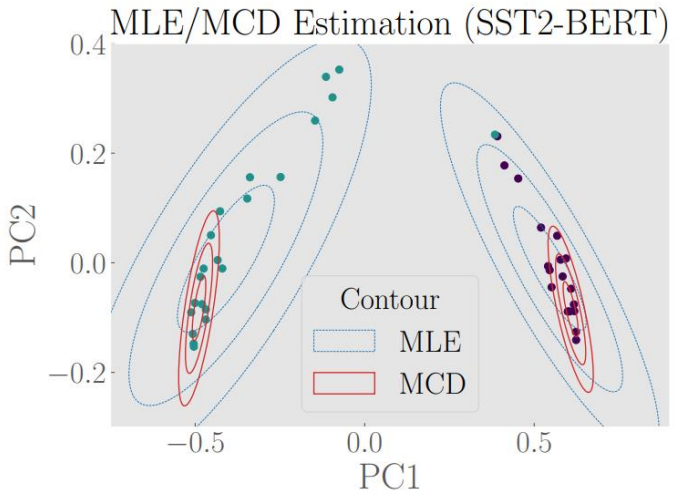
Mukhoti et al. Deep deterministic uncertainty: A new simple baseline CVPR 2023.

Robust Density Estimation

Idea: Removing outliers from the training dataset for parameter estimation in MD.

Method:

- → Do not share the covariance matrix between classes
- → Use Minimum Covariance Determinant (MCD) to find a subset of instances that minimizes the determinant of Σ for each individual class
- → PCA with an RBF kernel.



This results in a robust covariance estimation consisting of centered data points rather than outliers.

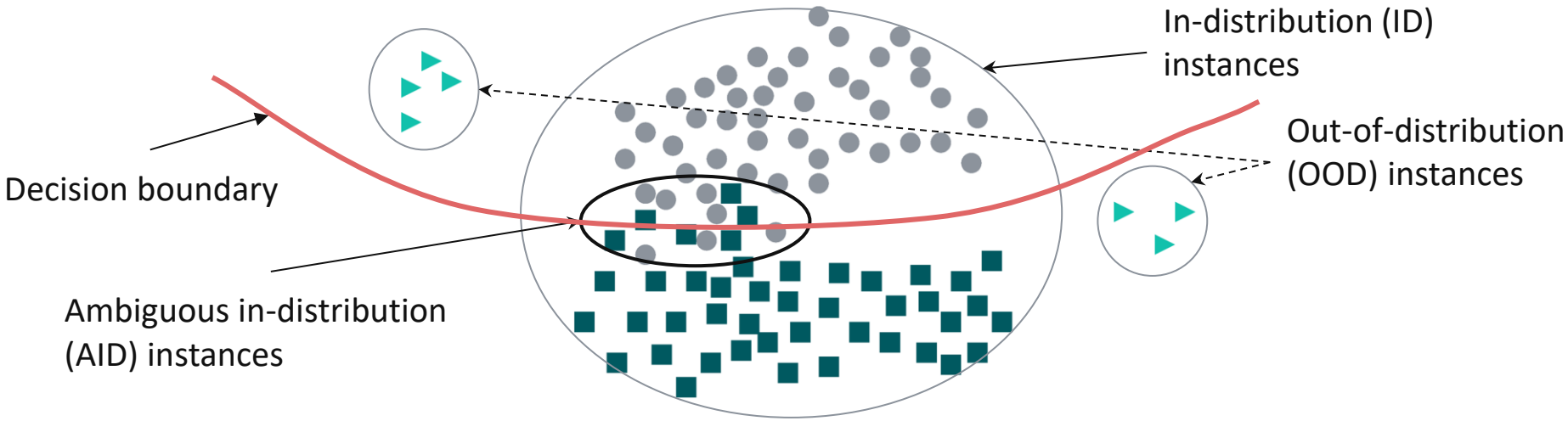
Yoo et al. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. Findings of ACL 2022.

Motivation

Goal: build a reliable selective classification methods for ambiguous text classification tasks.

Classification mistakes usually arise from two sources:

- → OOD areas – can be detected with epistemic UQ methods
- → Ambiguous in-distribution (AID) areas – can be detected by aleatoric UQ



Motivation

Following the Bayesian approach, the **total** uncertainty of a model prediction of an instance $\boldsymbol{x}$ for the given training dataset is computed as follows:

$$U_{\mathrm{T}}(\mathbf{x}) = U_{\mathrm{A}}(\mathbf{x}) + U_{\mathrm{E}}(\mathbf{x}),$$

where $U_{\mathrm{A}}(\mathbf{x})$ an aleatoric uncertainty and is $U_{\mathrm{E}}(\mathbf{x})$ stemic uncertainty.

Methods for quantifying <u>epistemic</u> uncertainty:

- → Mahalanobis Distance (MD)
- → Robust Density Estimation (RDE)
- → Deep Deterministic Uncertainty (DDU)

Methods for quantifying <u>aleatoric</u> uncertainty:

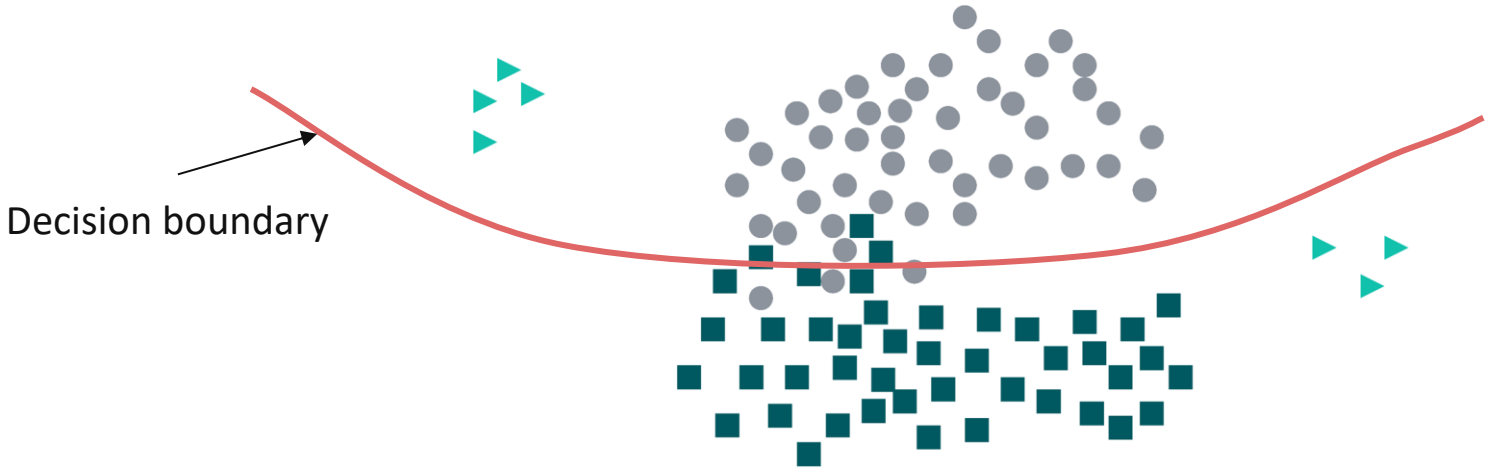
- → Softmax response (SR)
- → Entropy

AIRI

Hybrid Uncertainty Quantification (HUQ)

Input: Validation dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, hyper-parameters $\delta_{\min}, \delta_{\max}, \alpha$, input point $\mathbf{x}$, ranking function $R(u, \mathcal{D})$

Output: Uncertainty estimates $U_{\mathrm{HUQ}}(\mathbf{x})$

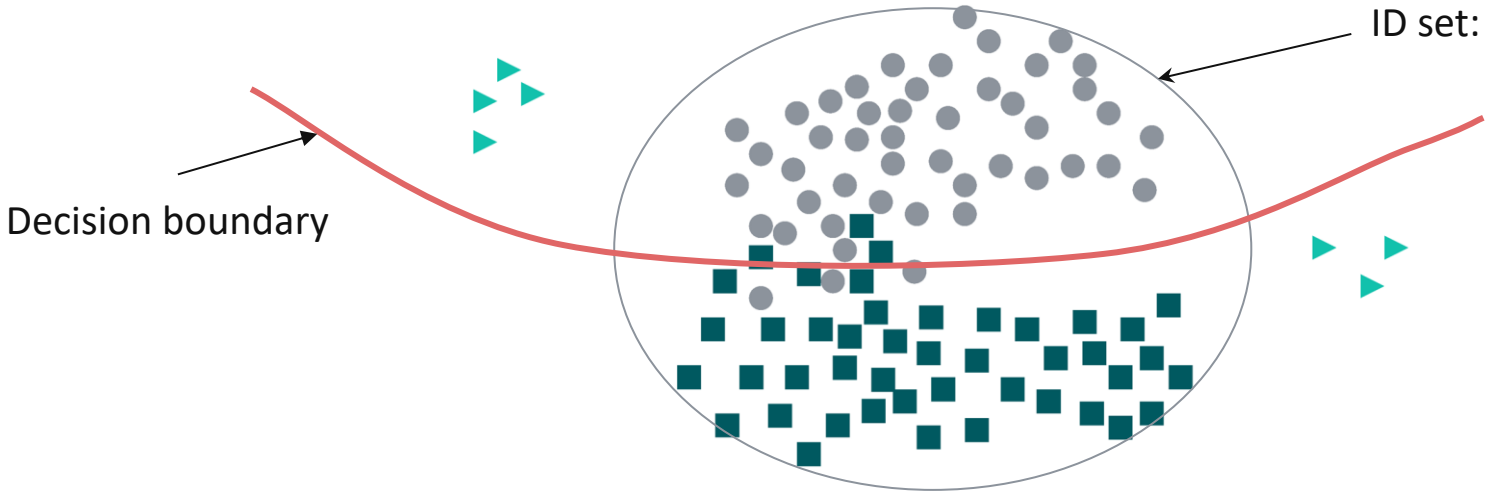


Vazhentsev et al. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. Findings of ACL 2023.

Hybrid Uncertainty Quantification (HUQ)

Input: Validation dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$-parameters , i$\delta_{\min}, \delta_{\max}, \alpha$oint , ranking functi$\mathbf{x}$n $R(u, \mathcal{D})$

Output: Uncertainty estimates $U_{\text{HUQ}}(\mathbf{x})$



ID set: $\mathcal{D}_{\text{ID}} = \{(\mathbf{x}_i, y_i) : (\mathbf{x}_i, y_i) \in \mathcal{D}, U_{\text{E}}(\mathbf{x}_i) \leq \delta_{\min}\}$

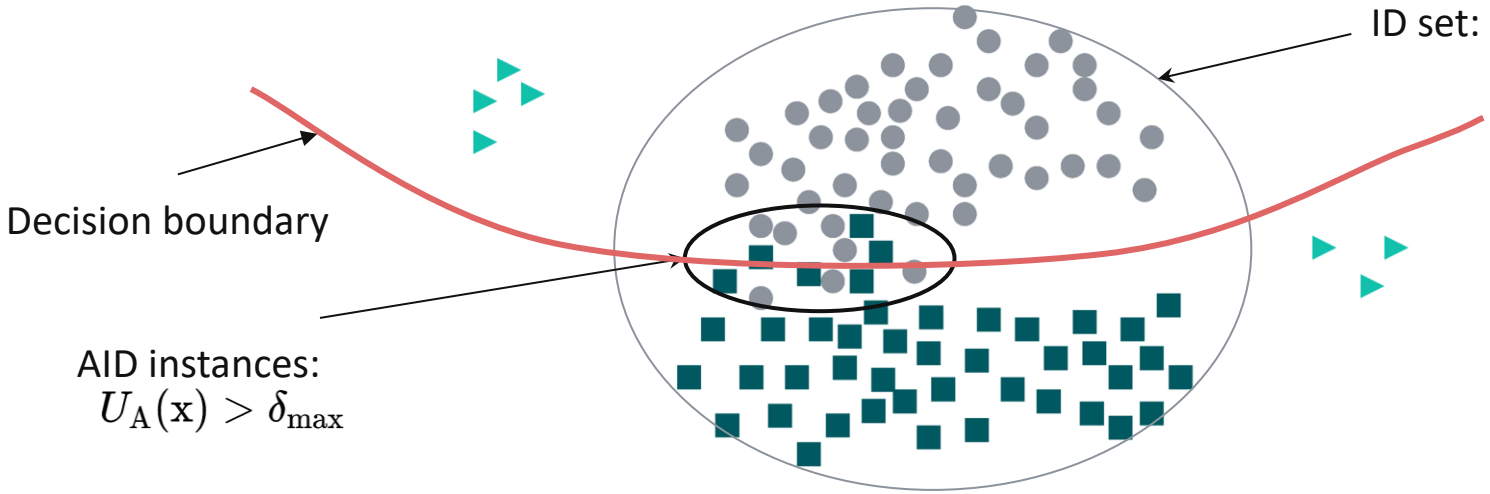
Vazhentsev et al. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. Findings of ACL 2023.

Hybrid Uncertainty Quantification (HUQ)

Input: Validation dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, parameters $\delta_{\min}, \delta_{\max}, \alpha$, point $\mathbf{x}$, ranking function $R(u, \mathcal{D})$

Output: Uncertainty estimates $U_{\text{HUQ}}(\mathbf{x})$

1. If this point belongs to the AID area: $U_{\text{HUQ}}(\mathbf{x}) = R(U_{\text{A}}(\mathbf{x}), D)$

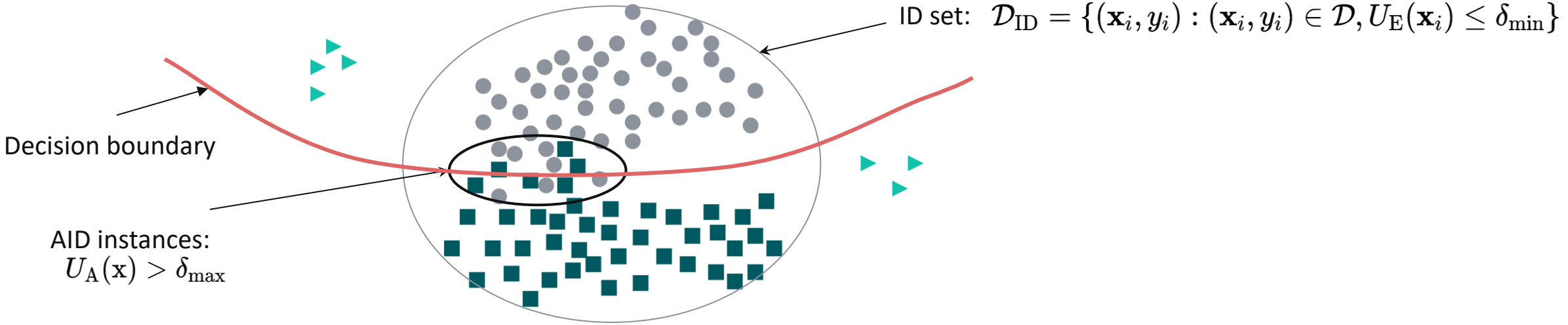


Hybrid Uncertainty Quantification (HUQ)

Input: Validation dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$-parameters , i$\delta_{\min}, \delta_{\max}, \alpha$oint , ranking functi$\mathbf{x}$n $R(u, \mathcal{D})$

Output: Uncertainty estimates $U_{\text{HUQ}}(\mathbf{x})$

1. If this point belongs to the AID area: $U_{\text{HUQ}}(\mathbf{x}) = R(U_{\text{A}}(\mathbf{x}), D)$
2. If this point belongs to the ID area, but not to AID: $U_{\text{HUQ}}(\mathbf{x}) = R(U_{\text{A}}(\mathbf{x}), D_{\text{ID}})$



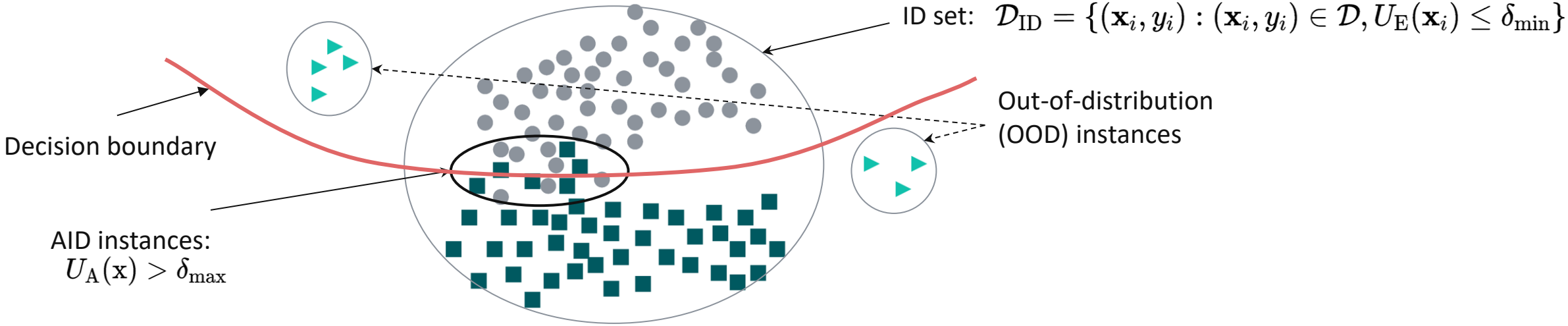
Vazhentsev et al. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. Findings of ACL 2023.

Hybrid Uncertainty Quantification (HUQ)

Input: Validation dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, parameters $\delta_{\min}, \delta_{\max}, \alpha$, point $\mathbf{x}$, ranking function $R(u, \mathcal{D})$

Output: Uncertainty estimates $U_{\text{HUQ}}(\mathbf{x})$

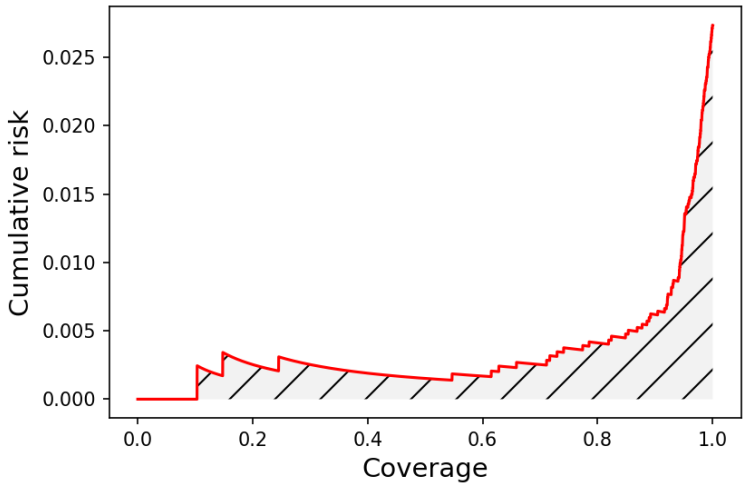
1. If this point belongs to the AID area: $U_{\text{HUQ}}(\mathbf{x}) = R(U_{\text{A}}(\mathbf{x}), D)$
2. If this point belongs to the ID area, but not to AID: $U_{\text{HUQ}}(\mathbf{x}) = R(U_{\text{A}}(\mathbf{x}), D_{\text{ID}})$
3. Otherwise: $U_{\text{HUQ}}(\mathbf{x}) = (1-\alpha)R(U_{\text{E}}(\mathbf{x}), D) + \alpha R(U_{\text{A}}(\mathbf{x}), D)$



Vazhentsev et al. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. Findings of ACL 2023.

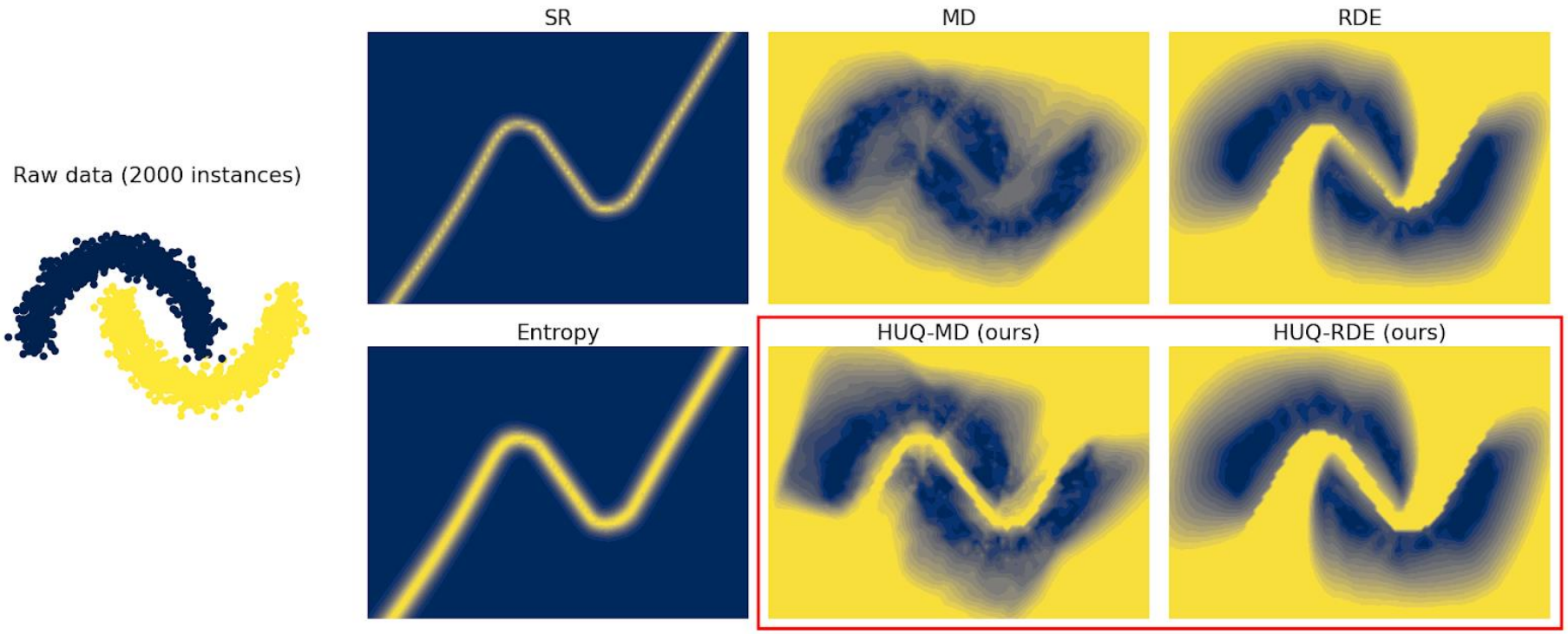
Experimental Setup

- **Models:** ELECTRA, BERT
- **Metrics:** AUC-RC↓ (area under the risk coverage curve)



- **Datasets:**
  - Paradetox, ToxiGen, Jigsaw, Twitter, ImplicitHate (Toxicity Detection)
  - SST-5, Amazon (Sentiment Analysis)
  - 20 News Groups

Toy Example



→ HUQ correctly identifies both regions with untrustworthy predictions: the area away from the training data distribution and the area around the model decision boundary.

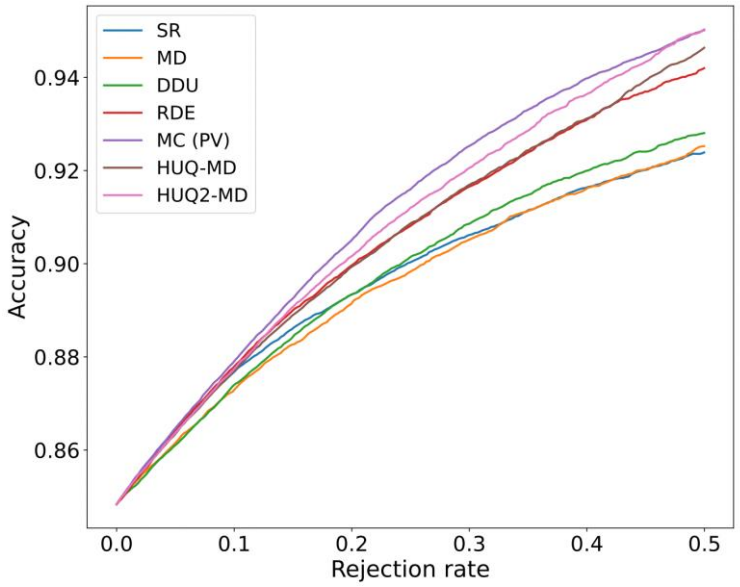
Results

Hybrid uncertainty quantification methods are usually **the best or the second best** after Ensemble. HUQ outperforms this baseline on Paradetox and SST-5.

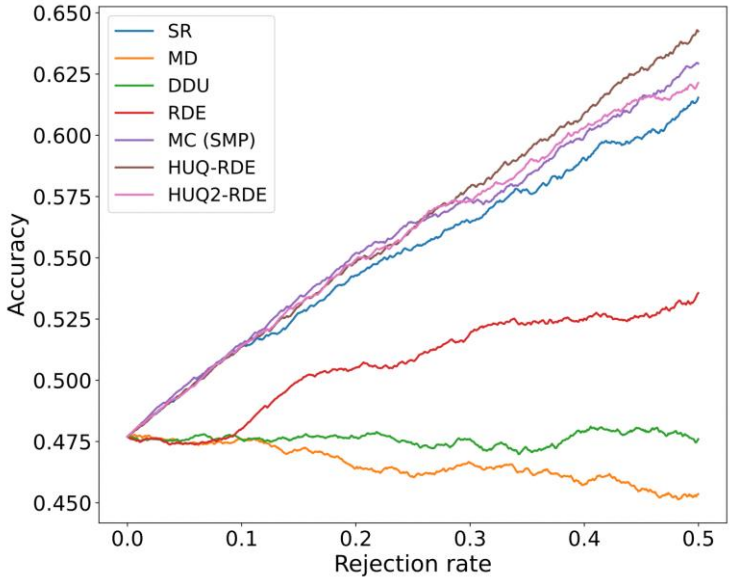
HUQ-DDU

Results: Medical Diagnostics Application

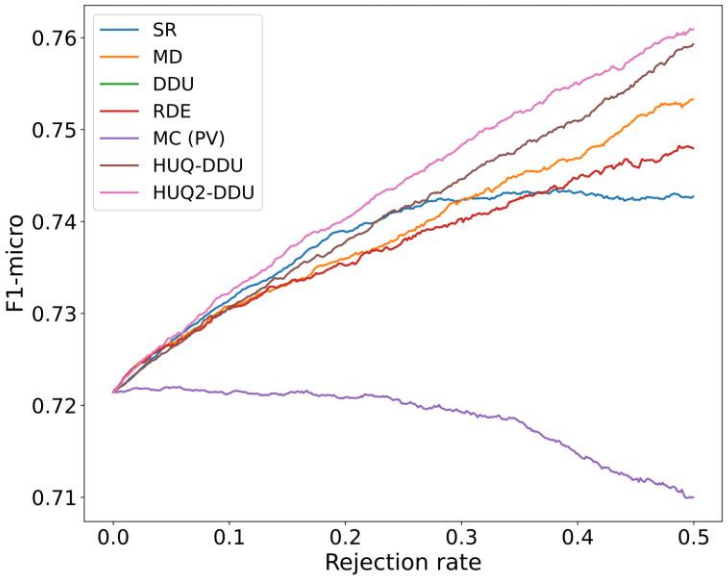
HUQ-2 and HUQ are consistently **the best or the second best** after MC dropout. While MC performs poorly on MIMIC-IV, HUQ-2 significantly outperforms all other methods



Mortality prediction



OV medical code prediction



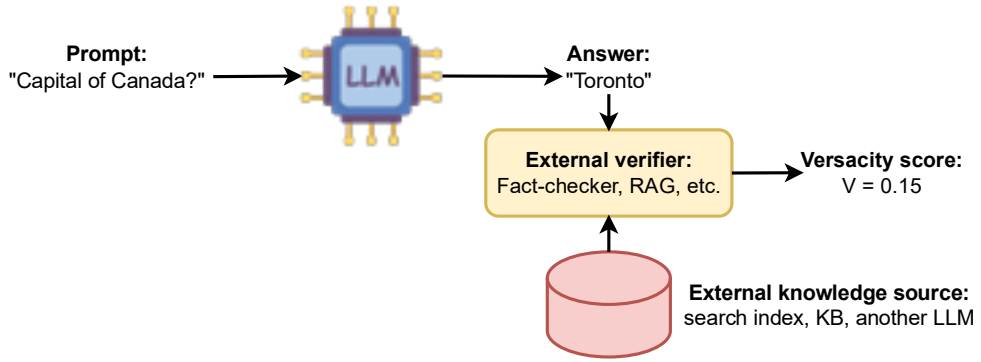
MIMIC-IV medical code prediction

03

UQ for Text Generation Models

Uncertainty as a Universal Hallucination Detector

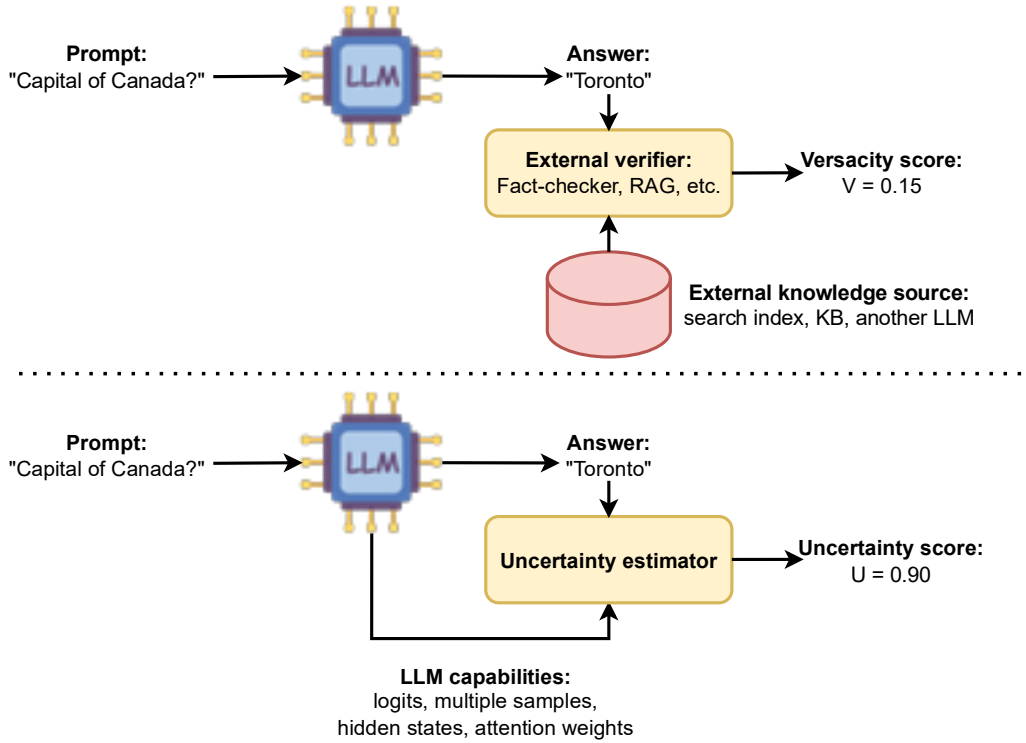
- Existing truthfulness assessment methods rely on external knowledge or large model ensembles, leading to **high computational costs** and limited applicability.



Uncertainty as a Universal Hallucination Detector

- Existing truthfulness assessment methods rely on external knowledge or large model ensembles, leading to **high computational costs** and limited applicability.

- **Uncertainty quantification** (UQ) offers a promising alternative, but it faces significant challenges in text generation.

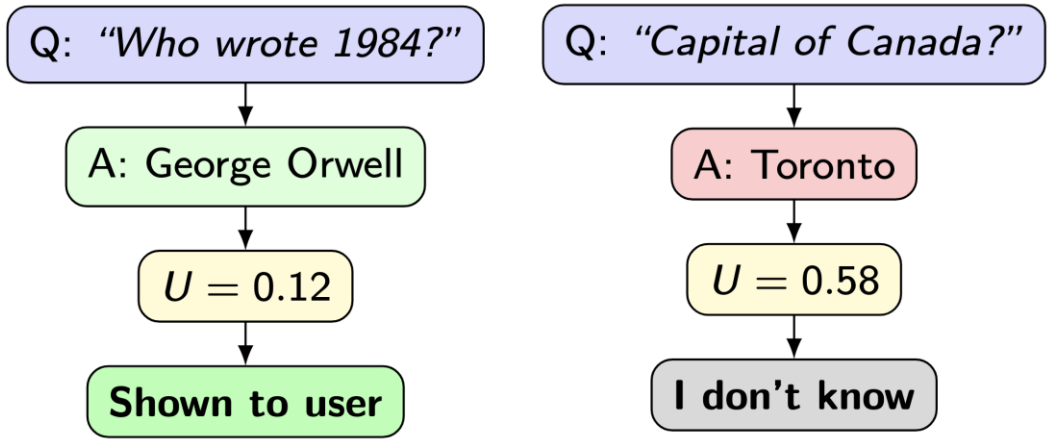


Reliability task: Selective generation

- **Goal:** Deliver answers only when the LLM is confident.
- **Selective rule:**

$$\text{output} = \begin{cases} \text{answer}, & \boldsymbol{U(x,y) < \tau} \\ \text{“I don’t know”}, & \boldsymbol{U(x,y) \geq \tau} \end{cases}$$

- By **rejecting** uncertain answers, we increase performance for remaining answers.



Reliability task: Claim-level hallucination detection

→ If $U(x, y, c) \geq \tau$, highlight claim $c \in \mathrm{C}(y)$ of LLM as potentially erroneous

Tell me a bio of Madonna.

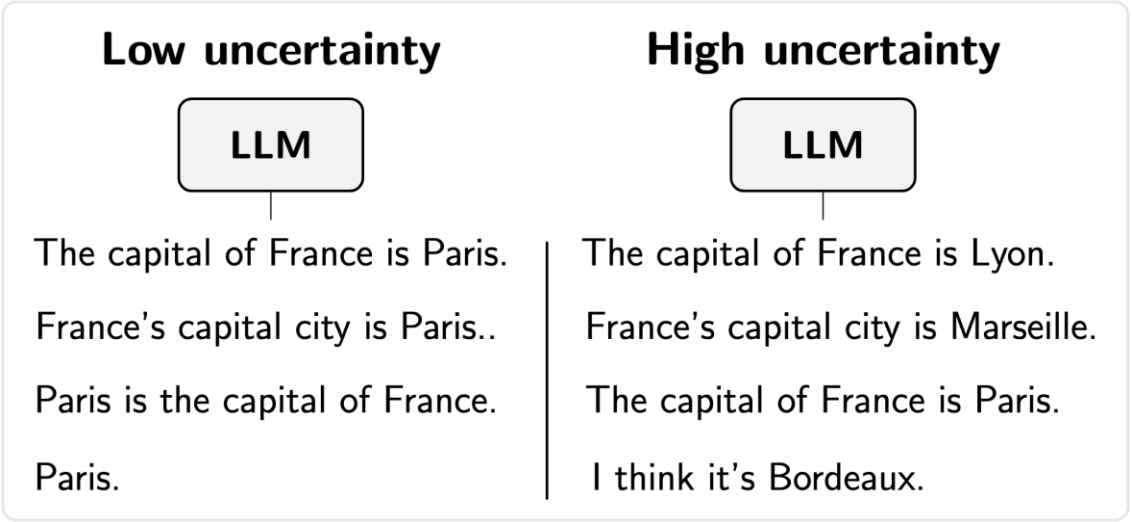
Madonna is an American singer, songwriter, actress, and business-woman. She was born on August 16, 1958, in Bay City, Michigan, and is best known for her provocative image and music. Madonna has sold over 300 million records worldwide, making her one of the best-selling music artists of all time. She has won numerous awards, including 20 Grammy Awards, a Golden Globe Award, and an Emmy Award. Madonna is also known for her philanthropic work and her activism on various social issues.

AIRI

Uncertainty Quantification Methods for LLMs

- **Unsupervised methods:** extract information from logits of LLM or multiple generations, ask LLM about its confidence.

  Weaknesses: limited effectiveness and computationally expensive.



Vashurin et al. Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. TACL 2025.

Information-based Methods

For a given:

$\boldsymbol{x}$ - input sequence (prompt)

$\boldsymbol{\theta}$ - model parameters

We can compute:

- Probability of the generated sequence: $P(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) = \prod_{l=1}^{L} P\left(y_l \mid \mathbf{y}_{<l}, \mathbf{x}, \boldsymbol{\theta}\right)$
- Maximum Sequence Probability (MSP): $U_{\text{MSP}}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) = 1 - P(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$
- Perplexity or Normalized Sequence Probability (NSP): $U_{\text{Perplexity}}\left(\mathbf{x}\right) = \exp\left\{-\frac{1}{L}\log P(\mathbf{y} \mid \mathbf{x})\right\}$

Fomicheva et al. Unsupervised quality estimation for neural machine translation. TACL 2020.

Sampling-based Methods

For a given:

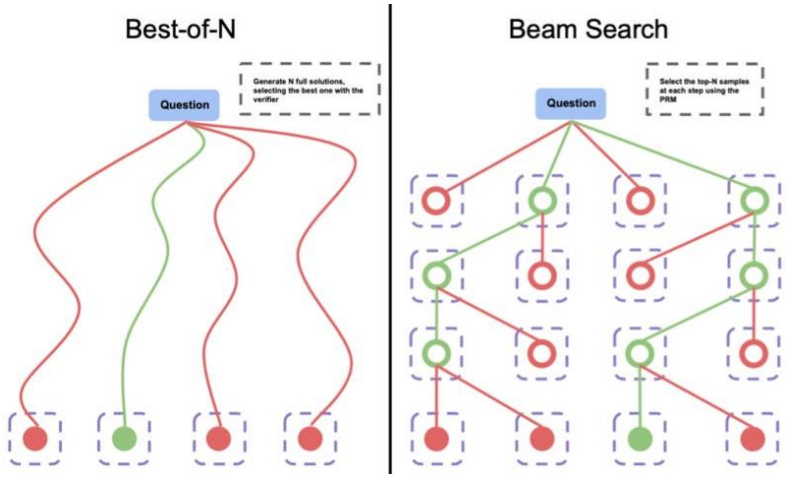
x - input sequence (prompt)

Θ - model parameters

We can generate:

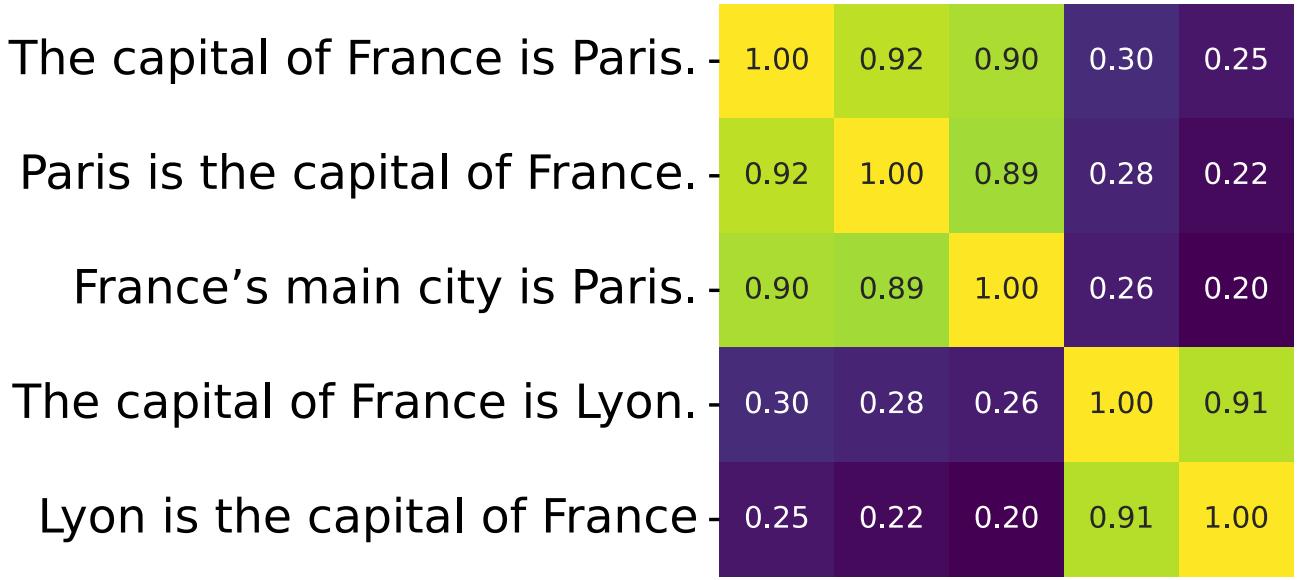
→ $y_1, y_2, \ldots, y_N$ - N sequences generated via sampling or beam search

Uncertainty score: quantifying consistency across multiple generations



Sampling-based Methods

- Construct a matrix S representing similarities between responses based on some semantic or lexical similarity measure, e.g. NLI entailment score or ROUGE



Lexical Similarity

- Lexical Similarity: compare samples via lexical metrics, e.g., ROUGE or BLUE
- Uncertainty is the **average lexical similarity** between the generated answers

$$\mathrm{U}_{\mathrm{LexSim}}(\mathbf{x}) = 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} s(\mathbf{y}^i, \mathbf{y}^j)$$

Fomicheva et al. Unsupervised quality estimation for neural machine translation. TACL 2020.

Graph-based Uncertainty Measures

- Sampled sequences are nodes, pairwise similarities are edges
- Then similarity matrix S becomes an **adjacency matrix of the graph**
- Degree matrix: $D_{ii} = \sum_{j=1}^{K} S_{ij}$ Normalized Graph Laplacian: $L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$
- Compute uncertainty by analyzing the graph connectivity:

1. Degree Matrix :

$$U_{Deg} = 1 - \text{trace}(D)/K^2$$

2. Sum of Eigenvalues of the Graph Laplacian:

$$U_{EigV} = \sum_{k=1}^{K} \max(0, 1 - \lambda_k)$$

Lin et al. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Model. TMLR 2024.

AIRI

Monte-Carlo Sequence Entropy

- Monte Carlo approximation of sequence entropy with N samples:

$$\mathrm{U}_{\mathrm{MCSE}}(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N} \log P\left(\mathbf{y}^i \mid \mathbf{x}\right)$$

- To ensure balanced contributions to the overall uncertainty from sequences of different lengths, we can employ a length-normalized version:

$$\mathrm{U}_{\mathrm{MCNSE}}(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N} \log \bar{P}\left(\mathbf{y}^i \mid \mathbf{x}\right)$$

Malinin & Gales. Uncertainty Estimation in Autoregressive Structured Prediction. ICLR 2021.

Semantic Entropy

- **Problem of MCSE**: semantic equivalence of different answers

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.5
Rome	0.4	0.4
London	0.1	0.1
Entropy	0.94	0.94

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.94	0.33

- **Idea:** Group the answers into clusters based on their meaning and calculate the entropy over semantic clusters:

$$\hat{P}(C_m \mid \mathbf{x}) = \sum_{\mathbf{y} \in C_m} P(\mathbf{y} \mid \mathbf{x})$$

$$\mathrm{U_{SE}} = -\frac{1}{M} \sum_{m=1}^{M} \log \hat{P}(C_i \mid \mathbf{x})$$

Kuhn et al. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR 2023.

AIRI

CoCoA: Bridging Confidence and Consistency

- A more flexible approach to confidence estimation can be achieved by combining various information-theoretic confidence measures with consistency analysis.

- CoCoA proposes a multiplicative form of this combination:

$$\mathrm{C}_{\mathrm{CoCoA}}(\mathbf{y}^*, \mathbf{x}) = \mathrm{C}_{\mathrm{inf}}(\mathbf{y}^*, \mathbf{x}) \cdot \mathrm{C}_{\mathrm{cons}}(\mathbf{y}^*, \mathbf{x})$$

- C_{inf} can be any information-theoretic confidence estimate, such as sequence probability, perplexity, mean token entropy etc., while C_{cons} is defined as:

$$\mathrm{C}_{\mathrm{cons}}(\mathbf{y}^*, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{y}^*, \mathbf{y}^i)$$

Vashurin et al. Uncertainty Quantification for LLMs through Minimum Bayes Risk: Bridging Confidence and Consistency. NeurIPS 2025.

Reflexive

- Black-box:

```
Provide your best guess and the probability that it is correct
(0.0 to 1.0) for the following question. For example:

Guess: <most likely guess>
Probability: <the probability between 0.0 and 1.0 that
your guess is correct>

Question: Who was the first president of the United States?
```

- Relies on the ability of the LLM to assess its own uncertainty

- White-box:

```
Question: Who was the first president of the United States?
Proposed Answer: George Washington was the first president.

Is the proposed answer:
  (A) True
  (B) False
The proposed answer is:
```

- Resulting confidence is based on the probability of the token encoding “True”:

$$U_{PTrue}(y) = 1 - P(\text{“}True\text{”} \mid \boldsymbol{x}, \boldsymbol{y})$$

Tian et al. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. EMNLP 2023.
Kadavath et al. Language Models (Mostly) Know What They Know. 2022

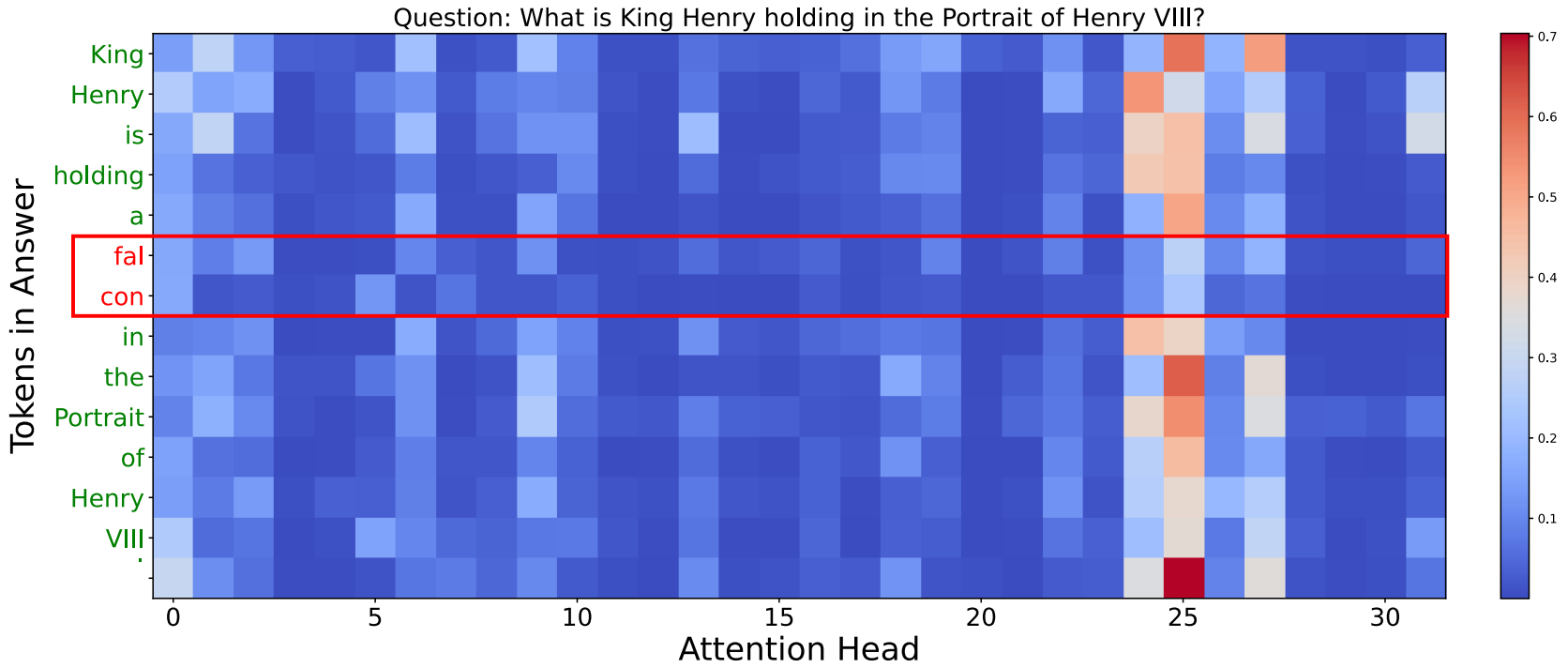
Identifying Hallucination-Associated Patterns in Attention Maps

Idea: identify patterns in attention maps that reveal hallucinations.

- *Question:* What is King Henry holding in the Portrait of Henry VII?
- *Correct Answer:* gloves and dagger.
- *LLM Answer (Llama-3.1 8b):* King Henry is holding a falcon in the Portrait of Henry VII.

Identifying Hallucination-Associated Patterns in Attention Maps

- Most attention heads show low weights
- **The 25th head:** high attention for correct tokens, low for the hallucinated token



Recurrent Attention-based Uncertainty Quantification: RAUQ

1. Select the most informative attention head per layer:

$$\mathbf{h}_l(\mathbf{y}) = \arg\max_{h=1\ldots H} \frac{1}{L-1} \sum_{i=2}^{L} a_{i,i-1}^{lh}$$

Recurrent Attention-based Uncertainty Quantification: RAUQ

1. Select the most informative attention head per layer:

$$\mathbf{h}_l(\mathbf{y}) = \arg\max_{h=1\ldots H} \frac{1}{L-1} \sum_{i=2}^{L} a_{i,i-1}^{lh}$$

2. Compute token-level layer-wise recurrent confidence score:

$$\mathbf{c}_l(y_i) = \begin{cases} P(y_i \mid \mathbf{x}), & \text{if } i = 1, \\ \alpha \cdot P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}) + (1-\alpha) \cdot a_{i,i-1}^{l\,\mathbf{h}_l} \cdot \mathbf{c}_l(y_{i-1}), & \text{if } i > 1. \end{cases}$$

Vazhentsev et al. Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs. 2025.

Recurrent Attention-based Uncertainty Quantification: RAUQ

1. Select the most informative attention head per layer:

$$\mathbf{h}_l(\mathbf{y}) = \arg\max_{h=1\ldots H} \frac{1}{L-1} \sum_{i=2}^{L} a_{i,i-1}^{lh}$$

2. Compute token-level layer-wise recurrent confidence score:

$$\mathbf{c}_l(y_i) = \begin{cases} P(y_i \mid \mathbf{x}), & \text{if } i = 1, \\ \alpha \cdot P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}) + (1-\alpha) \cdot a_{i,i-1}^{l\,\mathbf{h}_l} \cdot \mathbf{c}_l(y_{i-1}), & \text{if } i > 1. \end{cases}$$

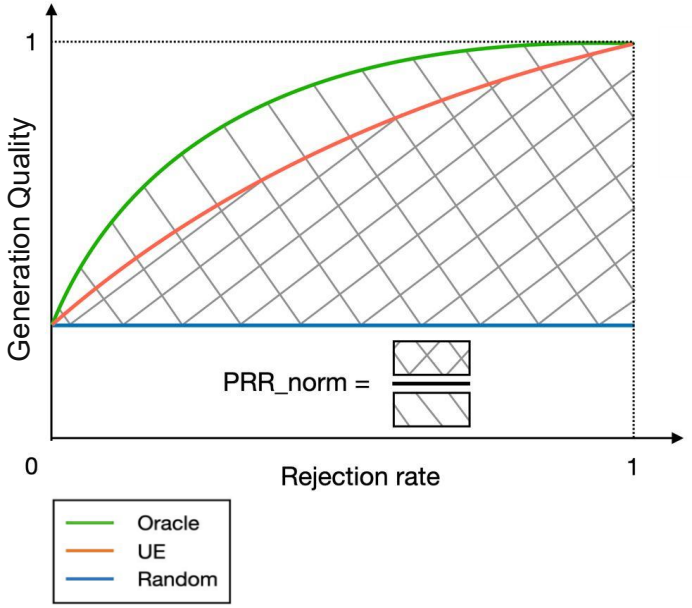
3. Aggregate the token-level layer-wise uncertainty scores to the final score:

$$\mathrm{U}_{\mathrm{RAUQ}}(\mathbf{y}) = \max_{l \in \mathcal{L}} \left[-\frac{1}{L} \sum_{i=1}^{L} \log \mathbf{c}_l(y_i) \right]$$

Vazhentsev et al. Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs. 2025.

Experimental Setup

- **Task:** sequence-level selective generation
- **Datasets:**
  - QA with short free-form answers: SciQ, CoQA, TriviaQA, MMLU
  - QA with long free-form answers: MedQUAD, TruthfulQA, GSM8k
  - Translation: WMT14 Fr-En, WMT19 De-En
  - Summarization: XSum, SamSum, CNN/DailyMail
- **LLMs:** Llama-3.1 8b, Gemma-2 9b, Qwen-2.5 7b, Falcon-3 10B
- **Metric:** PRR (50% max rejection)



Results

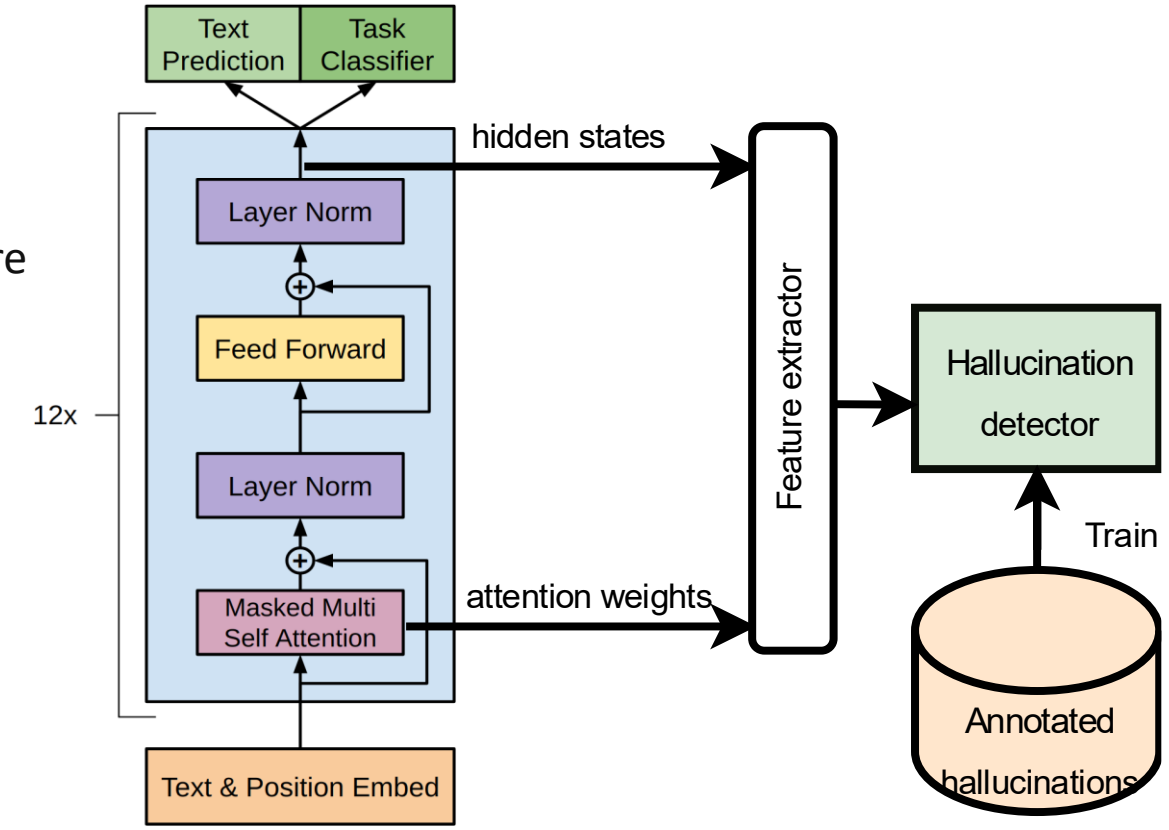
- RAUQ consistently outperforms prior methods with **minimal compute overhead (<1%)**
- **Best overall robustness** across models, tasks, and domains

UQ Method	Llama-3.1 8B			Qwen-2.5 7B			Gemma-2 9B			Falcon-3 10B			Mean
	QA	Summ	MT	QA	Summ	MT	QA	Summ	MT	QA	Summ	MT	
MSP	.347	.296	.397	.329	.151	.369	.361	.334	.381	.345	.177	.333	.318
Perplexity	.347	.419	.380	.343	**.254**	.406	.383	.375	.405	.356	.180	.439	.357
CCP	.285	.307	.340	.271	.186	.327	.329	.345	.320	.299	.128	.287	.285
Attention Score	.014	.126	.178	.038	.130	.142	.064	.103	.146	.054	**.192**	.089	.106
Focus	.320	.335	.361	.264	.186	.380	.416	.340	.385	.313	.139	.362	.317
Simple Focus	.342	.306	.415	.342	.136	.399	.396	.322	.422	.351	.095	.385	.326
DegMat NLI Score entail.	.306	.118	.239	.356	.154	.275	.337	.138	.259	.352	.132	.222	.241
Ecc. NLI Score entail.	.274	-.008	.284	.322	.002	.306	.298	.020	.290	.327	.038	.281	.203
EVL NLI Score entail.	.293	.114	.217	.349	.154	.245	.332	.133	.252	.351	.135	.206	.232
Lexical Similarity Rouge-L	.250	.131	.324	.334	.131	.327	.306	.161	.342	.285	.084	.275	.246
EigenScore	.232	.078	.285	.298	.061	.302	.267	.106	.226	.247	.051	.236	.199
LUQ	.287	.173	.214	.351	.196	.213	.344	.206	.259	.335	.121	.196	.241
Semantic Entropy	.254	.117	.315	.281	.092	.317	.291	.126	.337	.320	.133	.291	.240
SAR	.310	.170	.370	.351	.153	.393	.361	.235	.414	.334	.094	.337	.294
Semantic Density	.330	.153	.264	.352	.110	.291	.375	.167	.255	.358	.141	.280	.256
RAUQ	**.396**	**.428**	**.452**	**.358**	.213	**.438**	**.421**	**.392**	**.473**	**.392**	.181	**.465**	**.384**

Uncertainty Quantification Methods for LLMs

- **Supervised methods:** train a lightweight classifier on the information from the internal layers of LLMs to predict hallucinations.

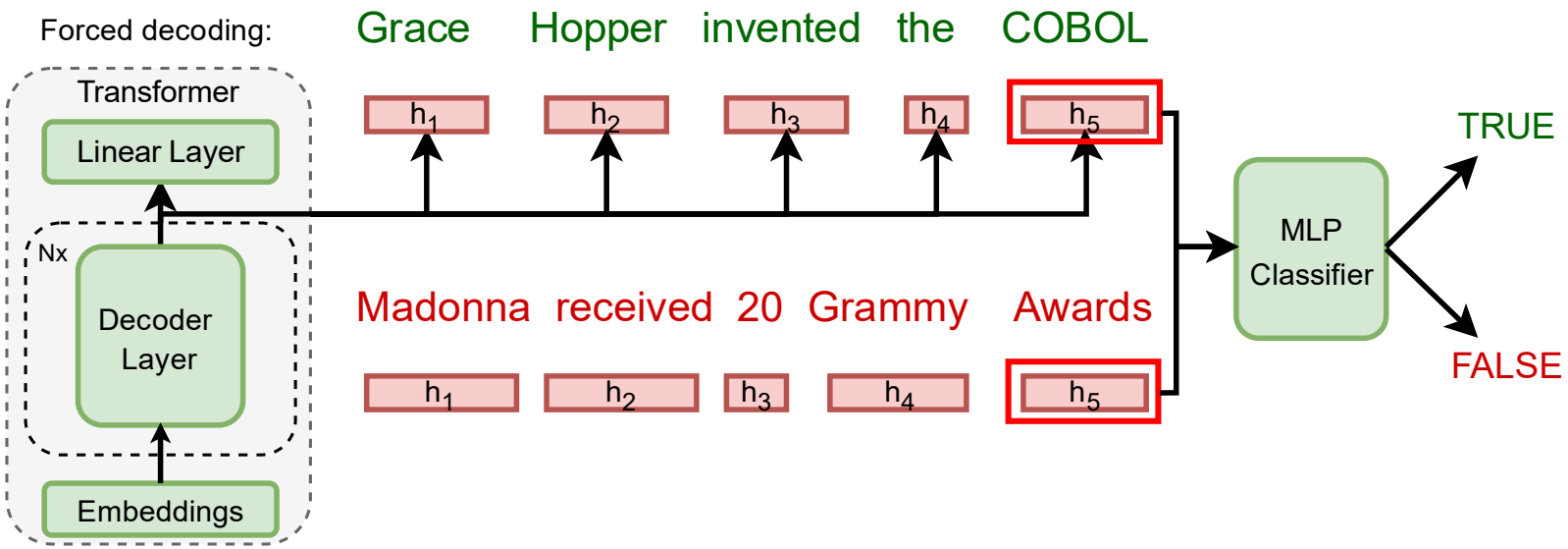
  Weaknesses: overfit to a particular domain and require annotated training data.



Azaria & Mitchell. The Internal State of an LLM Knows When It's Lying. EMNLP Findings 2023.

Statement accuracy prediction based on language model activations: SAPLMA

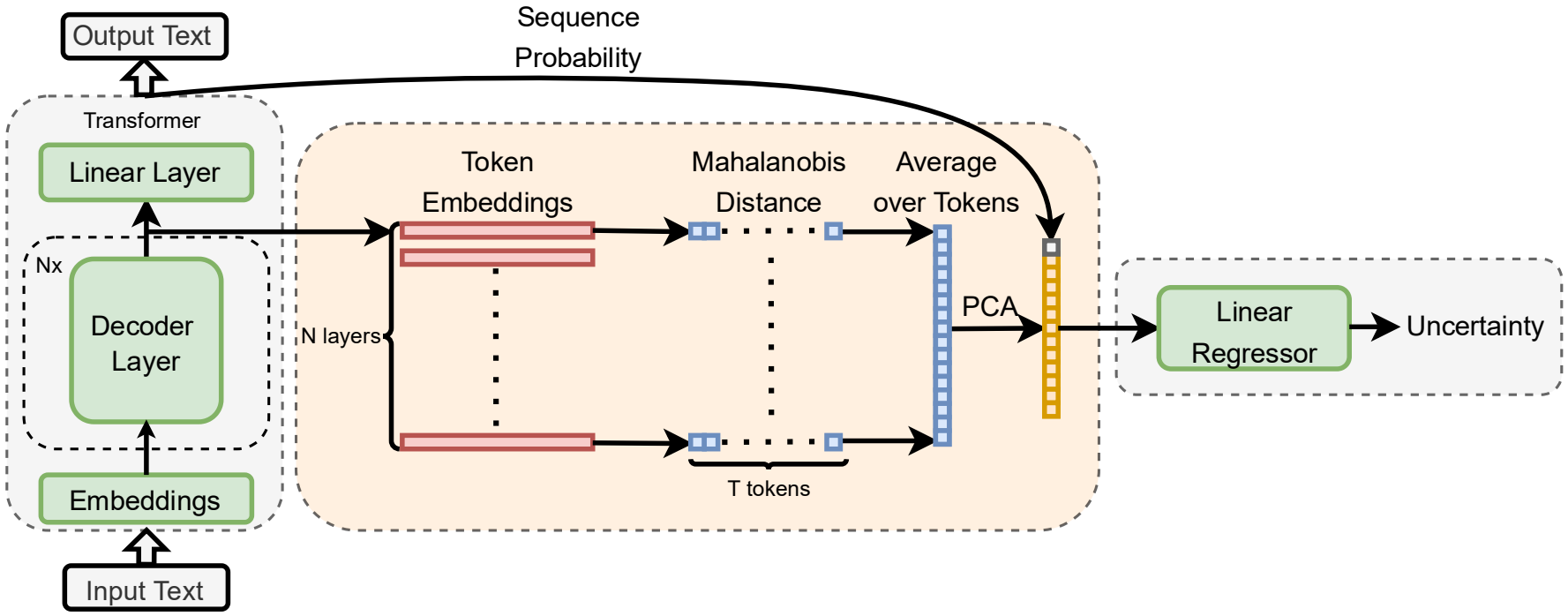
Idea: train on decoder layer activations to predict when LLM is uncertain.



Azaria & Mitchell. The Internal State of an LLM Knows When It's Lying. EMNLP Findings 2023.

Supervised average token-level relative Mahalanobis distance

Idea: aggregate token-level Mahalanobis distances to the cluster of “good” answers across all layers.

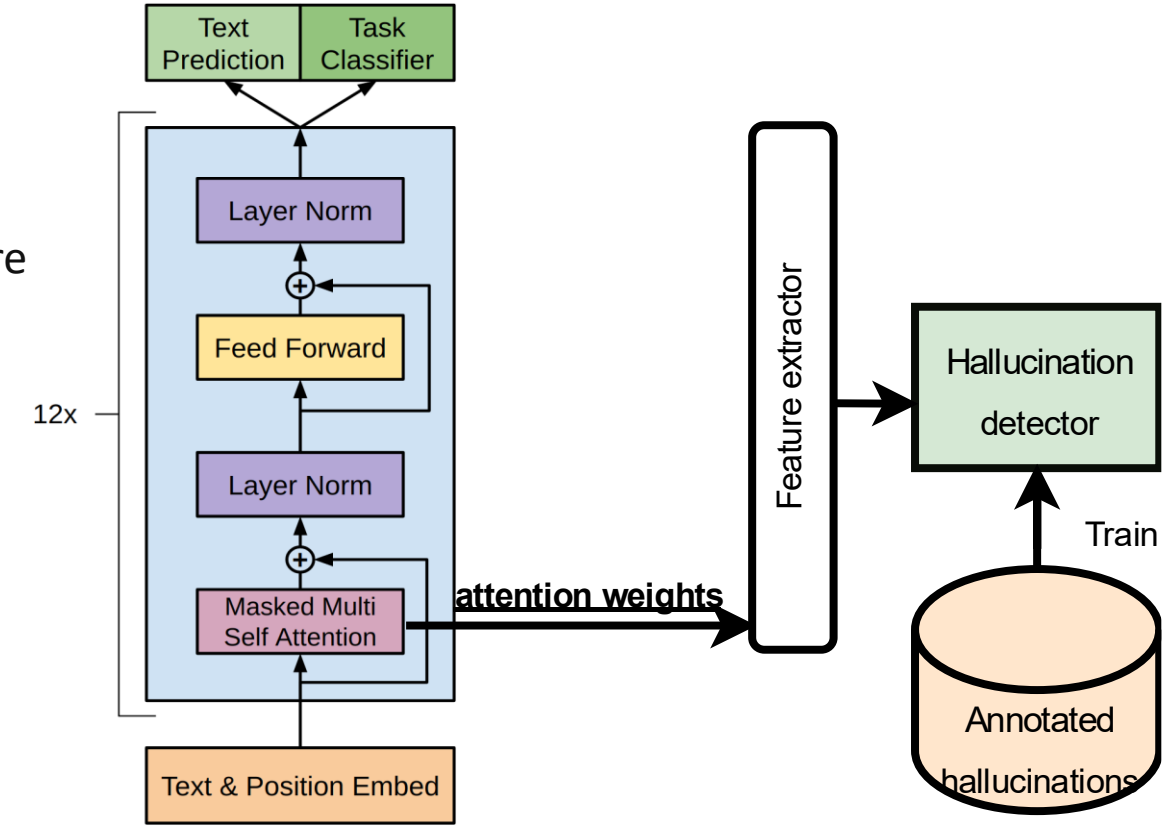


Uncertainty Quantification Methods for LLMs

- **Supervised methods:** train a lightweight classifier on the information from the internal layers of LLMs to predict hallucinations.

  Weaknesses: overfit to a particular domain and require annotated training data.

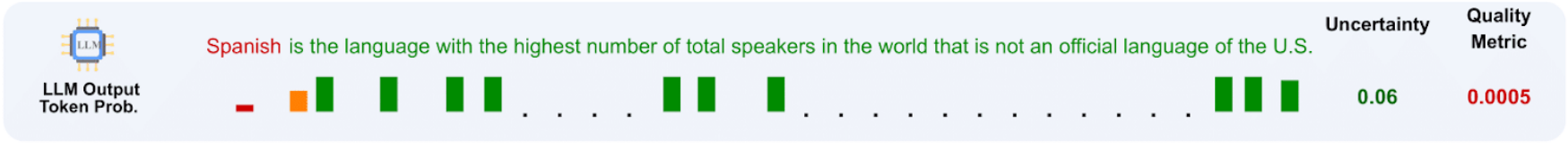
- **Attention-based** supervised methods emerge as the most effective approach.



Vazhentsev et al. Unconditional Truthfulness: Learning Unconditional Uncertainty of Large Language Models. EMNLP 2025.
Chuang et al. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. EMNLP 2024.

Conditional Dependency of Generation Steps

Problem: LLMs provide the conditional probability distribution, assuming all previous tokens are correct.



Conditional Dependency of Generation Steps

Problem: LLMs provide the conditional probability distribution, assuming all previous tokens are correct.

$$P(y_i \mid \mathbf{y}_{<i}, \mathbf{x})$$

We need the probability that does not depend on previously generated tokens:

$$P(y_i \mid \mathbf{x})$$

Toy simplification (1-step dependency): assume $P(y_i \mid \mathbf{y}_{<i}) \approx P(y_i \mid y_{i-1})$ and LLM generates only tokens that are true ("T") or false ("F").

$$\begin{aligned} P(y_i{=}\mathrm{T}) = {} & P(y_i{=}\mathrm{T} \mid y_{i-1}{=}\mathrm{T})\, P(y_{i-1}{=}\mathrm{T}) \\ & + P(y_i{=}\mathrm{T} \mid y_{i-1}{=}\mathrm{F})\big(1 - P(y_{i-1}{=}\mathrm{T})\big) \end{aligned}$$

AIRI

Trainable Attention-based Dependency (TAD)

Idea: attention implicitly encodes recurrent conditional dependency between generation steps, which we can learn.

$$C(y_i) = C(P(y_i \mid \mathbf{y}_{<i}, \mathbf{x}), \mathbf{C}_{<i}, Att_i)$$

Estimation of the unconditional confidence for the i-th token

Linear regression

LLM token probability

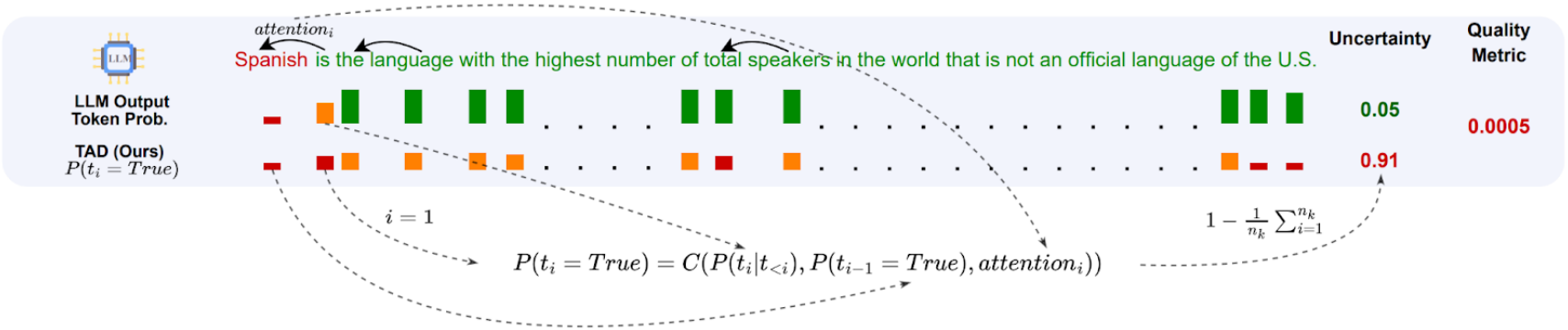
Predicted confidence of the previous tokens

Features derived from attention weights

Vazhentsev et al. Unconditional Truthfulness: Learning Unconditional Uncertainty of Large Language Models. EMNLP 2025.

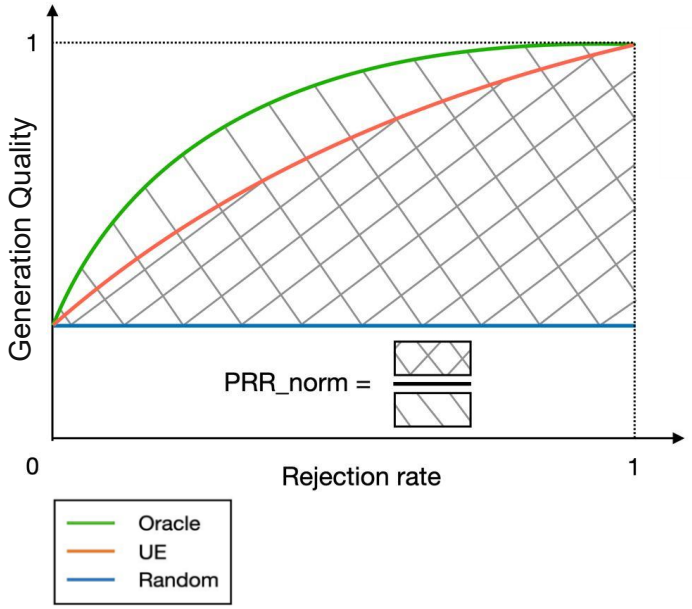
TAD: Inference Scheme

- **TAD** leverages the uncertainty from the previous step using a trainable model based on attention, resulting in a high overall uncertainty in the generated answer.



Experimental Setup

- **Models:** Llama-3.1 8b, Gemma-2 9b, Qwen-2.5 7b
- **Metrics:** Prediction Rejection Ratio (PPR) ↑
- **Datasets:**
  - QA with short free-form answers (SciQ, CoQA, TriviaQA, MMLU)
  - QA with long free-form answers (MedQUAD, TruthfulQA, GSM8k)
  - ATS (XSum, SamSum, CNN/DailyMail)
  - MT (WMT19 De-En)
- **UQ Baselines:**
  - Information-based methods (MSP, Perplexity, CCP)
  - Sampling-based methods (black-box methods, LexSim, Semantic Entropy, SAR)
  - Supervised methods (Factoscope, SAPLMA, Sheeps)



Results: In-Domain Performance

→ TAD **significantly outperforms** other supervised and unsupervised methods across various tasks and models.

UQ Method	XSum AlignScore	SamSum AlignScore	CNN AlignScore	WMT19 Comet	MedQUAD AlignScore	TruthfulQA AlignScore	CoQA AlignScore	SciQ AlignScore	TriviaQA AlignScore	MMLU Acc.	GSM8k Acc.	Mean PRR	Mean Rank
MSP	.077	.012	.339	.451	.030	-.088	.291	.551	.610	.654	.268	.291	12.91
Perplexity	.237	.250	.172	.466	.131	.274	.270	.385	.601	.400	.456	.331	10.45
Mean Token Entropy	.233	.280	.149	.475	.143	.356	.263	.342	.603	.225	.469	.322	10.55
CCP	.240	.025	.365	.388	.015	-.104	.215	.468	.596	.412	.281	.264	14.36
Simple Focus	.109	.116	.191	.496	.021	.093	.321	.536	.620	.550	.310	.306	11.55
Focus	.209	.144	.110	.452	.123	.189	.249	.462	.568	.037	.273	.256	14.55
Lexical Similarity Rouge-L	.122	.057	.122	.370	.075	.159	.297	.507	.531	.274	.511	.275	14.00
EigenScore	.077	-.010	.073	.374	.018	-.018	.281	.510	.500	.243	.537	.235	16.18
EVL NLI Score entail.	.139	.145	.068	.294	.122	.306	.329	.519	.571	.236	.372	.282	13.09
Ecc. NLI Score entail.	-.047	.032	-.015	.368	.107	.146	.294	.535	.543	.237	.386	.235	15.45
DegMat NLI Score entail.	.138	.145	.075	.332	.122	.300	.329	.540	.574	.235	.402	.290	12.36
Semantic Entropy	.016	.074	.106	.366	.073	.087	.265	.491	.536	.165	.380	.233	17.18
SAR	.128	.129	.107	.445	.088	.185	.318	.526	.585	.288	.459	.296	12.09
LUQ	.228	.170	.131	.265	.096	.322	.337	.449	.580	.321	.331	.294	12.18
Semantic Density	.080	.122	.213	.358	.095	.300	.386	.514	.603	.203	.381	.296	12.27
Factoscope	.185	-.032	.001	.069	.447	.137	.122	.345	.406	.844	-.101	.220	17.27
SAPLMA	.245	.326	.009	.345	.018	.321	.001	.374	.497	.418	.440	.272	14.09
MIND	.220	.133	.263	.365	.517	.314	.346	.496	.608	.883	**.738**	.444	7.36
Sheeps	.361	.313	.258	.487	.391	.476	.357	.487	.663	.883	.710	.490	4.73
LookBackLens	.436	.386	.369	.539	.497	.485	.352	**.600**	.585	.873	.627	.523	3.55
SATRMD	.338	.322	.254	.525	.362	.254	.315	.547	.623	.885	.566	.454	5.55
TAD	**.460**	**.416**	**.450**	**.553**	**.583**	**.500**	**.407**	.563	**.665**	**.893**	.701	**.563**	**1.27**

Results: Out-of-Domain Performance

- Supervised methods suffer a significant **performance drop** on out-of-domain data.
- **TAD** is the best-performing method on out-of-domain QA datasets.

UQ Method	CoQA AlignScore	SciQ AlignScore	TriviaQA AlignScore	MMLU Acc.	GSM8k Acc.	Mean PRR
MSP	.262	.459	.527	**.535**	.310	.419
SAR	.297	.439	.552	.275	.320	.377
Semantic Density	**.380**	.448	**.571**	.237	.197	.366
Factoscope	.016	.055	.161	.078	.049	.072
SAPLMA	-.030	.199	-.112	-.089	-.077	-.022
MIND	.044	.153	.237	.252	.230	.183
Sheeps	.092	.422	.295	.425	**.323**	.312
LookBackLens	.079	.365	.304	.422	.166	.267
SATRMD	.247	.349	.469	.205	.311	.316
TAD	.283	**.529**	.565	.512	.278	**.434**

AIRI

Conclusions

The key findings demonstrate the significant potential of UQ methods to enhance the model's predictions in every NLP task:

- Uncertainty quantification is a **crucial component** of ML-based systems.
- For practical purposes in classification tasks, consider **density-based UQ** methods like DDU, MD, RDE, etc.
- For ambiguous datasets, consider using **hybrid uncertainty quantification**, e.g. DDU + Entropy.
- For LLMs, **supervised methods** achieve state-of-the-art results for in-domain but experience a significant drop in performance when applied to out-of-domain.
- **Attention matrices** provide valuable information into the truthfulness of generations.
- Not all methods are applicable for claim-level UQ.

AIRI

airi.net

airi_research_institute

AIRI Institute

Telegram

AIRI