Leveraging Calibrated Uncertainty Estimates for Biomedical Applications

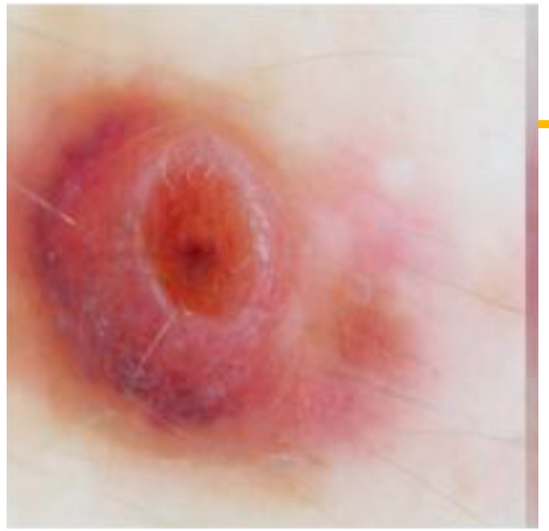
Florian Buettner

MOTIVATION - AI SYSTEMS IN MEDICINE

Decisions made downstream of an AI prediction need to know how much to trust it.

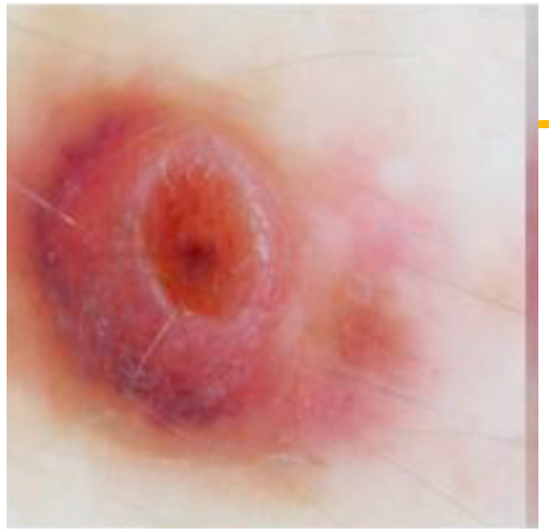
Precision oncology	A confident misclassification of a malignant lesion can delay or misdirect treatment.
Clinical decision support	LLMs answer patient- or clinician-facing questions where a hallucinated wrong answer is worse than an abstention.

CAN YOU TRUST YOUR MODEL'S PREDICTIONS?



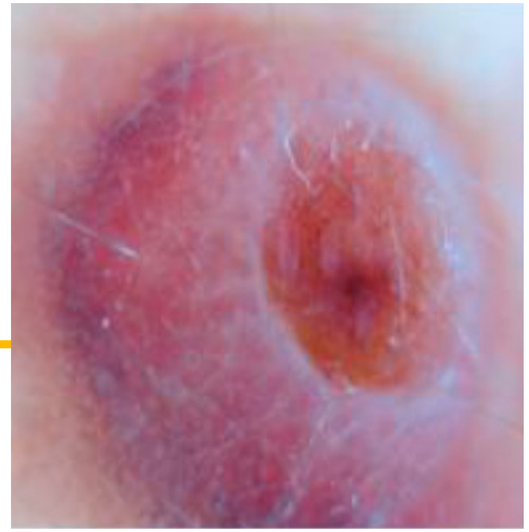
Nevus (benign) **99.8%**

CAN YOU TRUST YOUR MODEL'S PREDICTIONS?



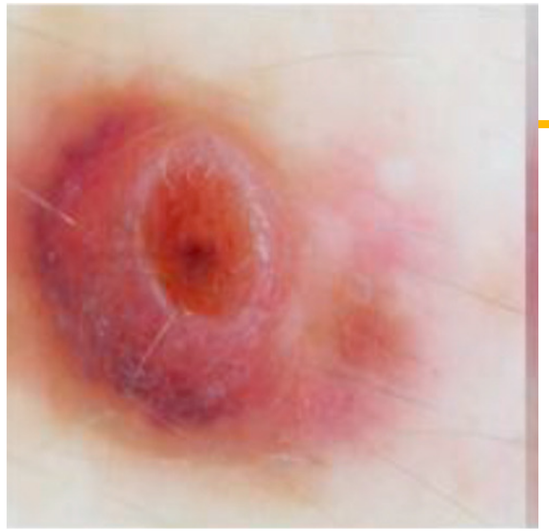
Nevus (benign) **99.8%**

Melanoma **93%**



Maron et al., EJC, 2021
Hekler, Brinker, Buettner, AAAI, 2023

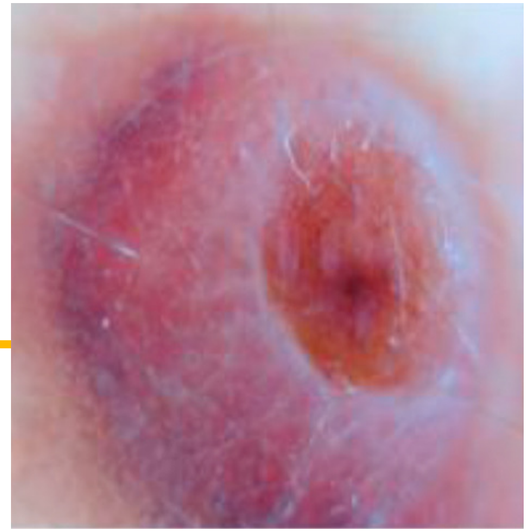
CAN YOU TRUST YOUR MODEL'S PREDICTIONS?



Nevus (benign) **99.8%**

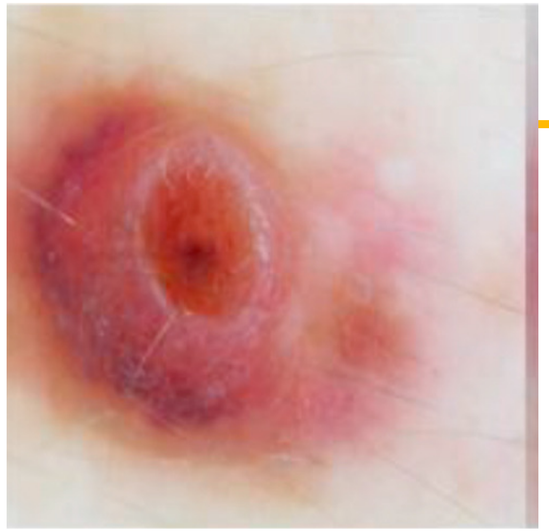
Melanoma **93%**

53%



Maron et al., EJC, 2021
Hekler, Brinker, Buettner, AAAI, 2023

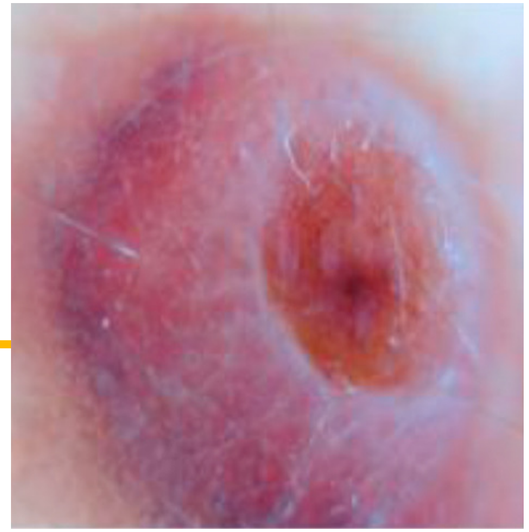
CAN YOU TRUST YOUR MODEL'S PREDICTIONS?



Nevus (benign) **99.8%**

Melanoma **93%**

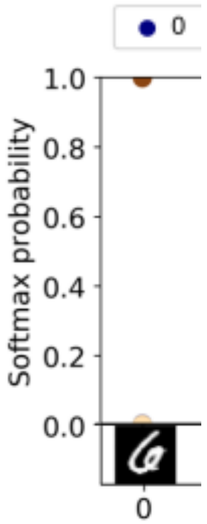
53%



Calibrated and reliable uncertainty estimates at the level individual predictions are key for trustworthiness

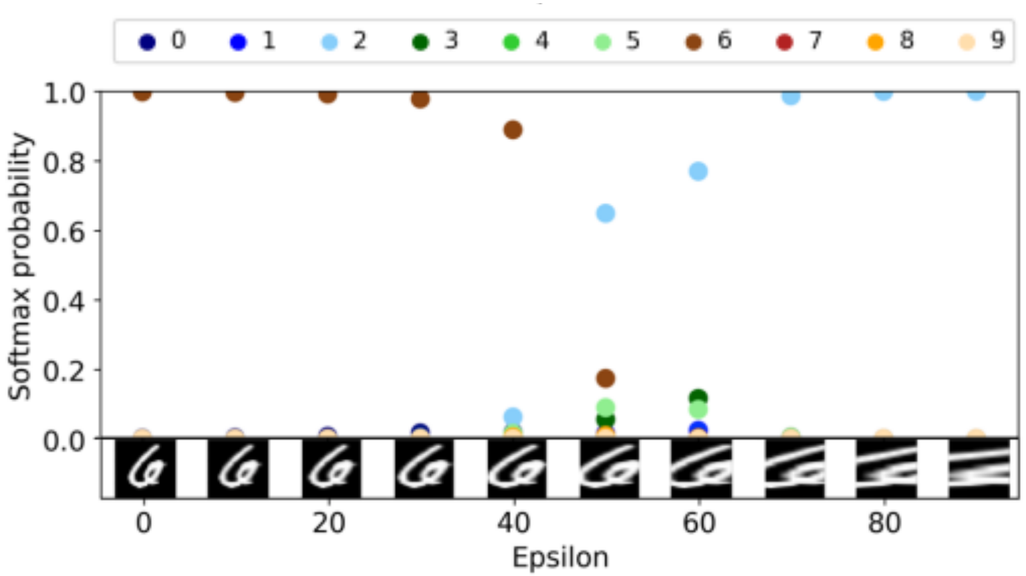
Maron et al., EJC, 2021
Hekler, Brinker, Buettner, AAAI, 2023

TRUSTWORTHINESS UNDER DATA DRIFT



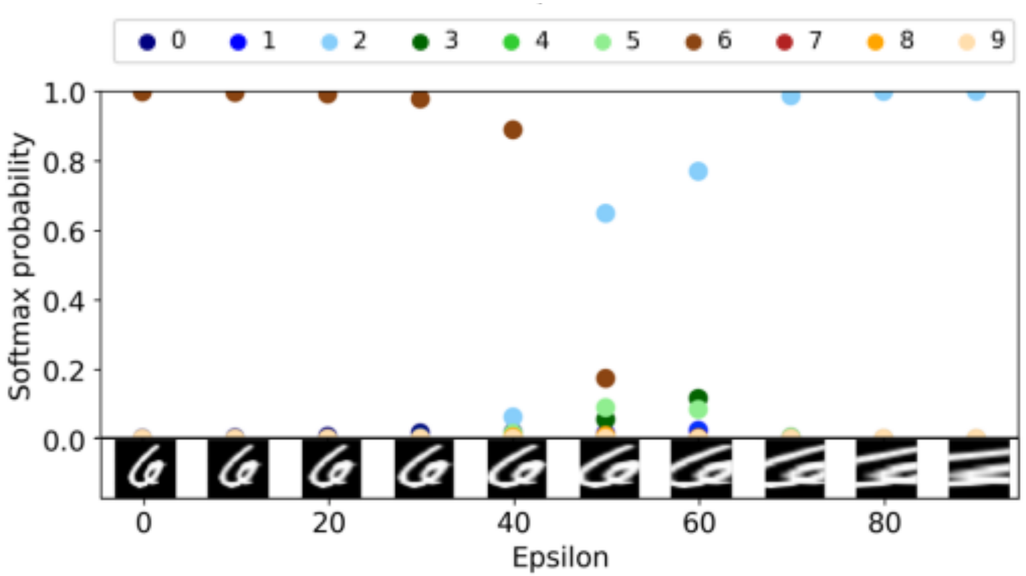
Ovadia et al., NeurIPS 2019
Tomani & Buettner, AAAI 2021
Tomani,…,Buettner, CVPR 2021

TRUSTWORTHINESS UNDER DATA DRIFT

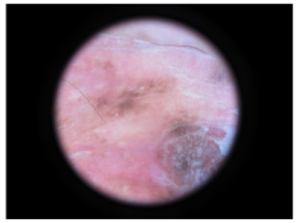


Ovadia et al., NeurIPS 2019
Tomani & Buettner, AAAI 2021
Tomani,…,Buettner, CVPR 2021

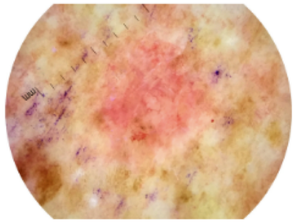
TRUSTWORTHINESS UNDER DATA DRIFT



Barcelona



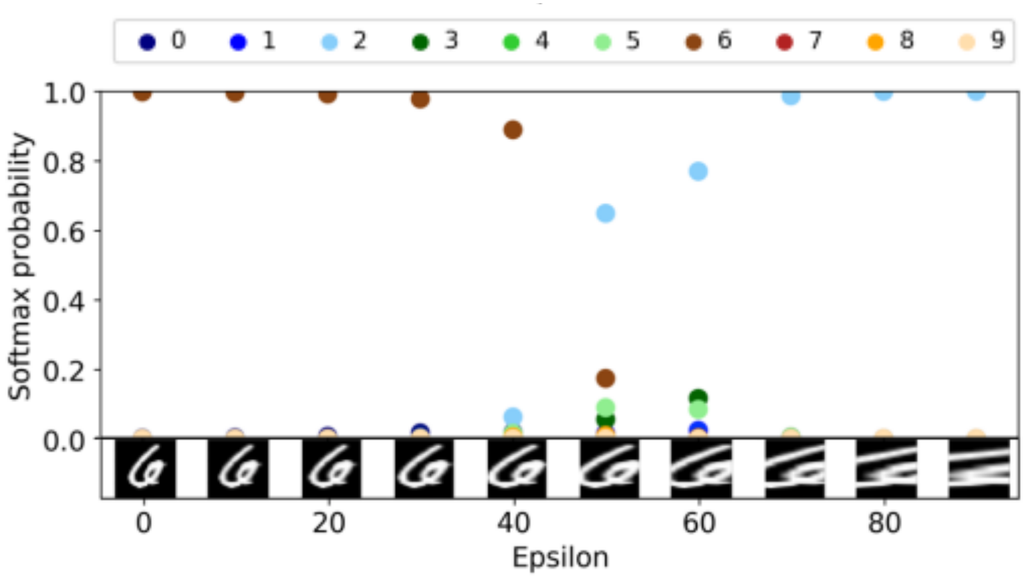
Sidney

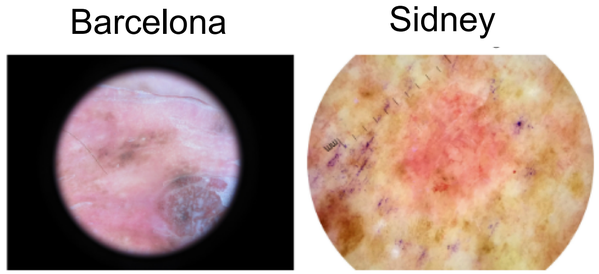


Ground truth	Melanoma	Melanoma
Label	Melanoma	Nevus
Confidence	99 %	99 %

Ovadia et al., NeurIPS 2019
Tomani & Buettner, AAAI 2021
Tomani,…,Buettner, CVPR 2021

TRUSTWORTHINESS UNDER DATA DRIFT





Ground truth	Melanoma	Melanoma
Label	Melanoma	Nevus
Confidence	99 %	99 %

Deep neural networks

- Model is not calibrated and unreliable
- Makes wrong predictions with high confidence, under data shift
- Predictions are not actionable: wrong predictions cannot be distinguished from correct ones

Ovadia et al., NeurIPS 2019
Tomani & Buettner, AAAI 2021
Tomani,…,Buettner, CVPR 2021

WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

- Previously: focus on ResNets and ViT, no web-scale pre-training, litte focus on training recipe

Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

- Previously: focus on ResNets and ViT, no web-scale pre-training, litte focus on training recipe
  - Does not reflect good practice of applied researchers

Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

- Previously: focus on ResNets and ViT, no web-scale pre-training, litte focus on training recipe
  - Does not reflect good practice of applied researchers
- Open question: What effect does the interplay between architecture and training recipe have on calibration?

Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

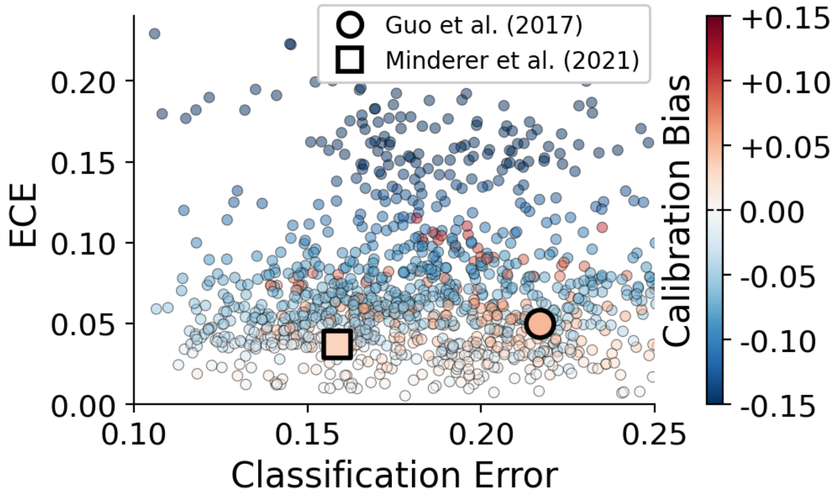
WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

- Previously: focus on ResNets and ViT, no web-scale pre-training, litte focus on training recipe
  - Does not reflect good practice of applied researchers
- Open question: What effect does the interplay between architecture and training recipe have on calibration?
- Large-scale re-evaluation of calibration behaviour of more than 1,000 vision models

Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

WHAT ABOUT MODERN ARCHITECTURES AND TRAINING RECIPES?

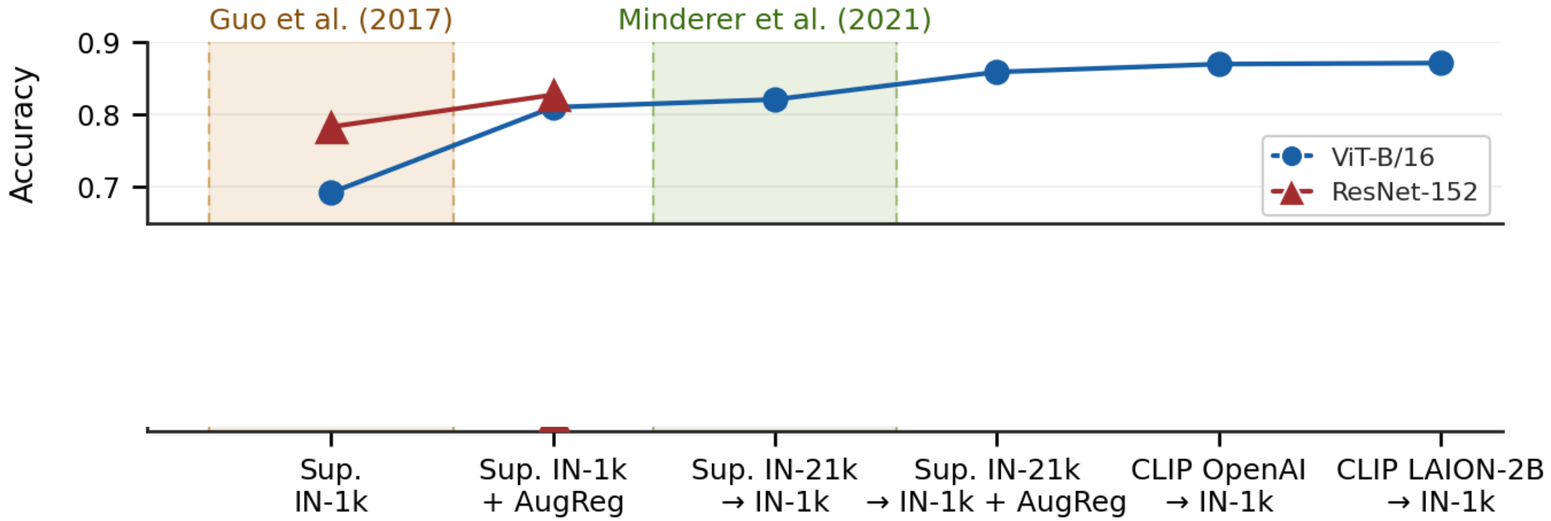
- Previously: focus on ResNets and ViT, no web-scale pre-training, litte focus on training recipe
  - Does not reflect good practice of applied researchers
- Open question: What effect does the interplay between architecture and training recipe have on calibration?
- Large-scale re-evaluation of calibration behaviour of more than 1,000 vision models



Guo et al., ICML 2017
Minder et al. NeurIPS 2022
Hekler, …, Buettner, arxiv 2025

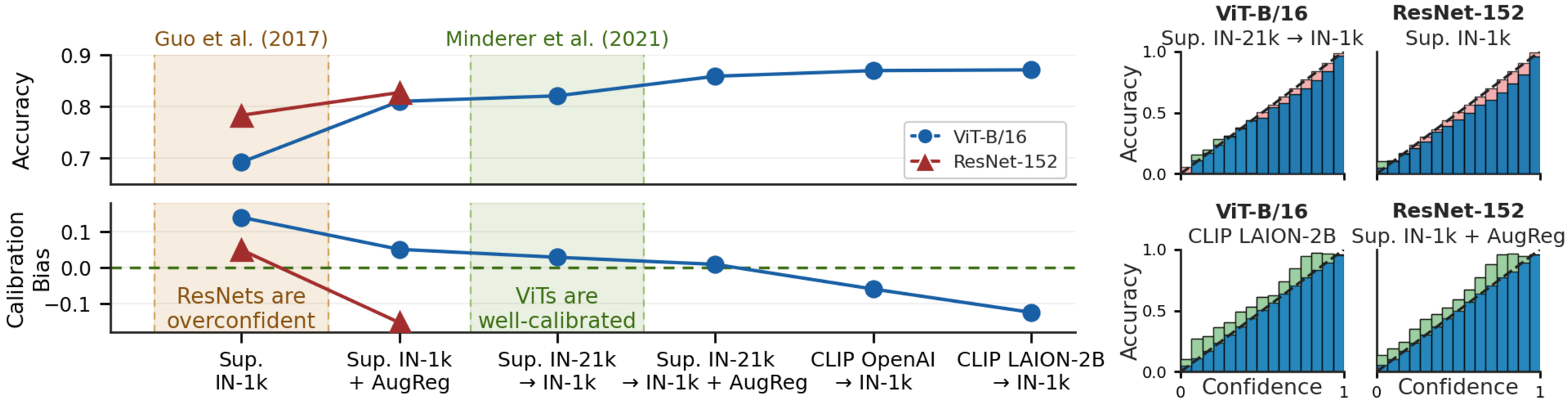
IT'S THE TRAINING RECIPE, NOT THE ARCHITECTURE

- Aggressive augmentations and large-scale pretraining leads to increasing underconfidence



IT'S THE TRAINING RECIPE, NOT THE ARCHITECTURE

- Aggressive augmentations and large-scale pretraining leads to increasing underconfidence

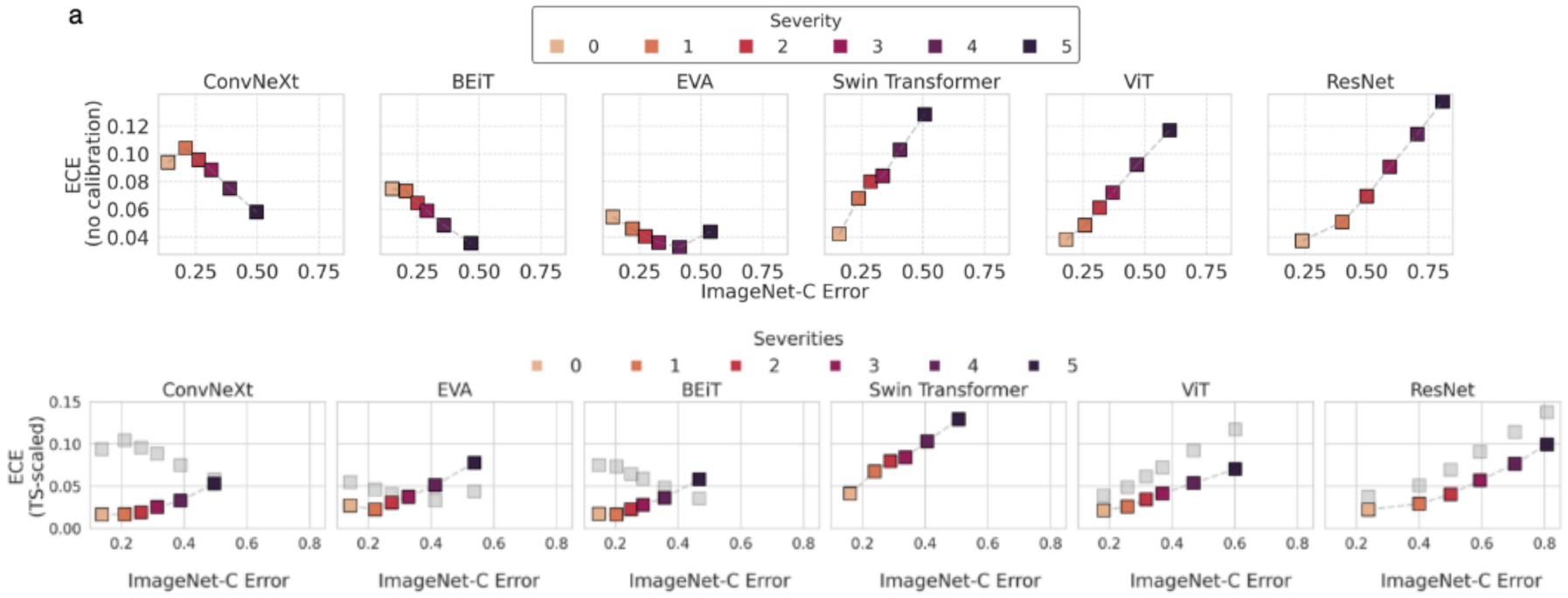


MIND THE DISTRIBUTION SHIFT WHEN RECALIBRATING

- Post-hoc recalibration of under confident models can exacerbate overconfidence under distribution shift

MIND THE DISTRIBUTION SHIFT WHEN RECALIBRATING

- Post-hoc recalibration of under confident models can exacerbate overconfidence under distribution shift



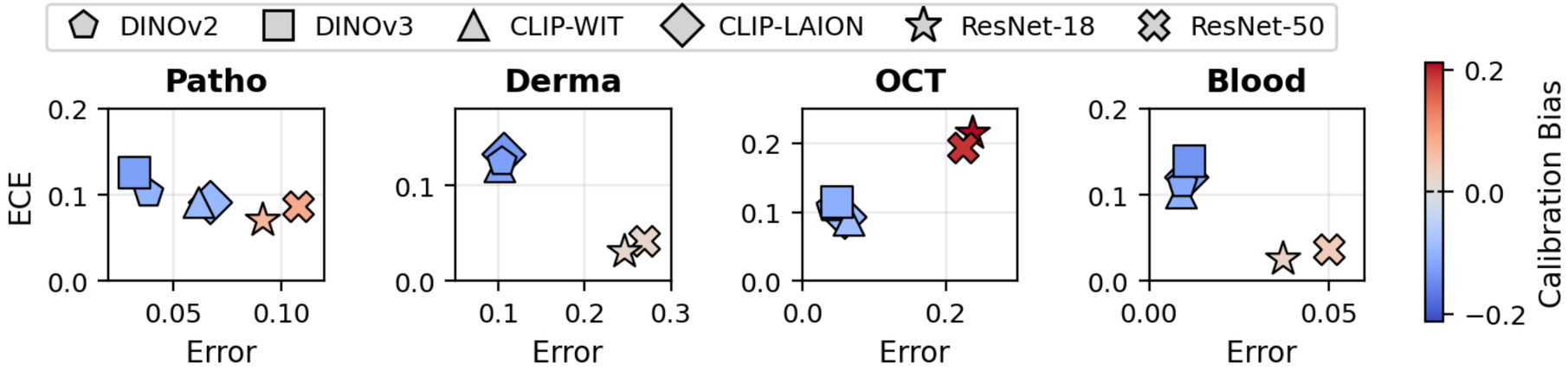
Temperature scaling can help - or harm

WHAT ABOUT MEDICAL APPLICATIONS?

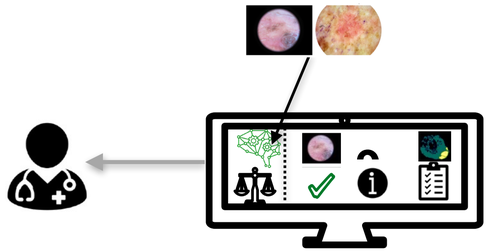
- Assess 4 common medical use-cases
  - Survival prediction based on pathology images
  - Skin lesion classification
  - OCT images of the retina
  - Cancer classification based on blood smears

WHAT ABOUT MEDICAL APPLICATIONS?

- Assess 4 common medical use-cases
  - Survival prediction based on pathology images
  - Skin lesion classification
  - OCT images of the retina
  - Cancer classification based on blood smears



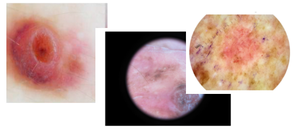
FROM DISCOVERY TO TRANSLATION: TRUSTWORTHY AI TOOLS FOR PERSONALISED ONCOLOGY



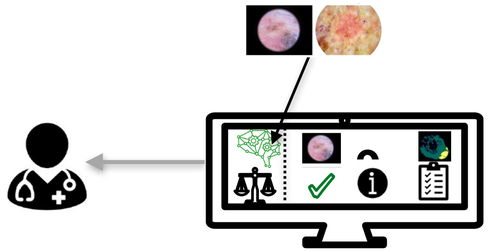
Diagnosis

Audit - Improve - Monitor model trustworthiness

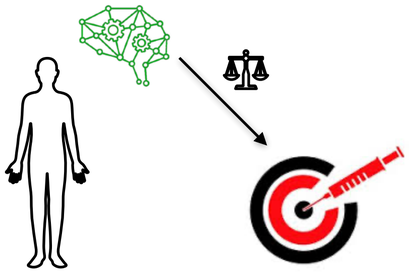
Diagnosis of skin lesions



FROM DISCOVERY TO TRANSLATION: TRUSTWORTHY AI TOOLS FOR PERSONALISED ONCOLOGY



Diagnosis

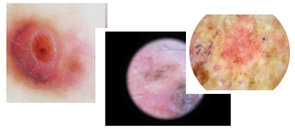


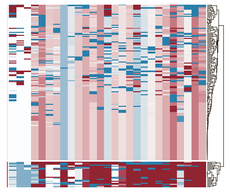
Stratification

Audit - Improve - Monitor model trustworthiness

Diagnosis of skin lesions

Stratification and modelling therapy outcome of blood cancer patients





GOETHE UNIVERSITÄT FRANKFURT AM MAIN

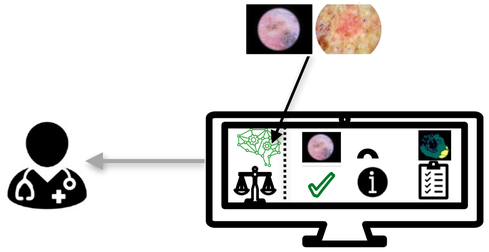
erc European Research Council

MLO MACHINE LEARNING IN ONCOLOGY

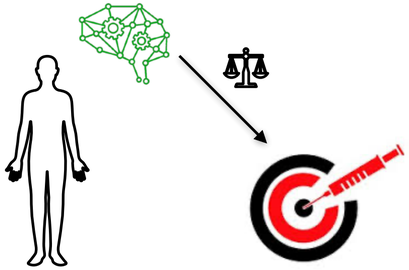
uct Universitäre Centren für Tumorerkrankungen Rhein-Main

dkfz. German Cancer Consortium Partner site Frankfurt|Mainz

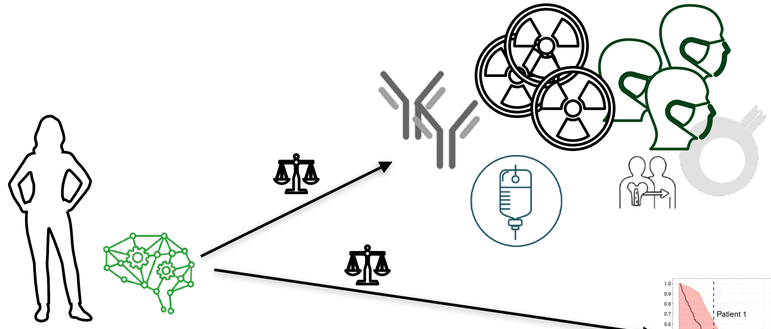
FROM DISCOVERY TO TRANSLATION: TRUSTWORTHY AI TOOLS FOR PERSONALISED ONCOLOGY



Diagnosis



Stratification



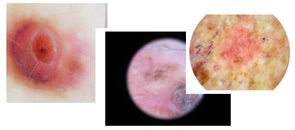
Outcome prediction and therapy recommendation

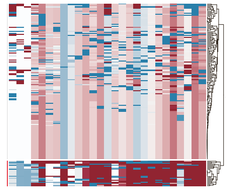
Audit - Improve - Monitor model trustworthiness

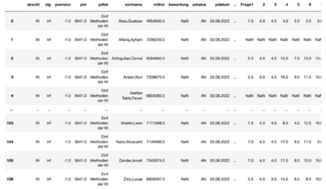
Diagnosis of skin lesions

Stratification and modelling therapy outcome of blood cancer patients

Therapy recommendation for metastatic breast cancer patients







GOETHE UNIVERSITÄT FRANKFURT AM MAIN

erc European Research Council

MLO MACHINE LEARNING IN ONCOLOGY

uct Universitäre Centren für Tumorerkrankungen Rhein-Main

dkfz. German Cancer Consortium Partner site Frankfurt|Mainz

DESIGNING MODEL AUDITOR AS AI AGENT

How can we identify and mitigate clinically relevant failure modes?

DESIGNING MODEL AUDITOR AS AI AGENT

How can we identify and mitigate clinically relevant failure modes?

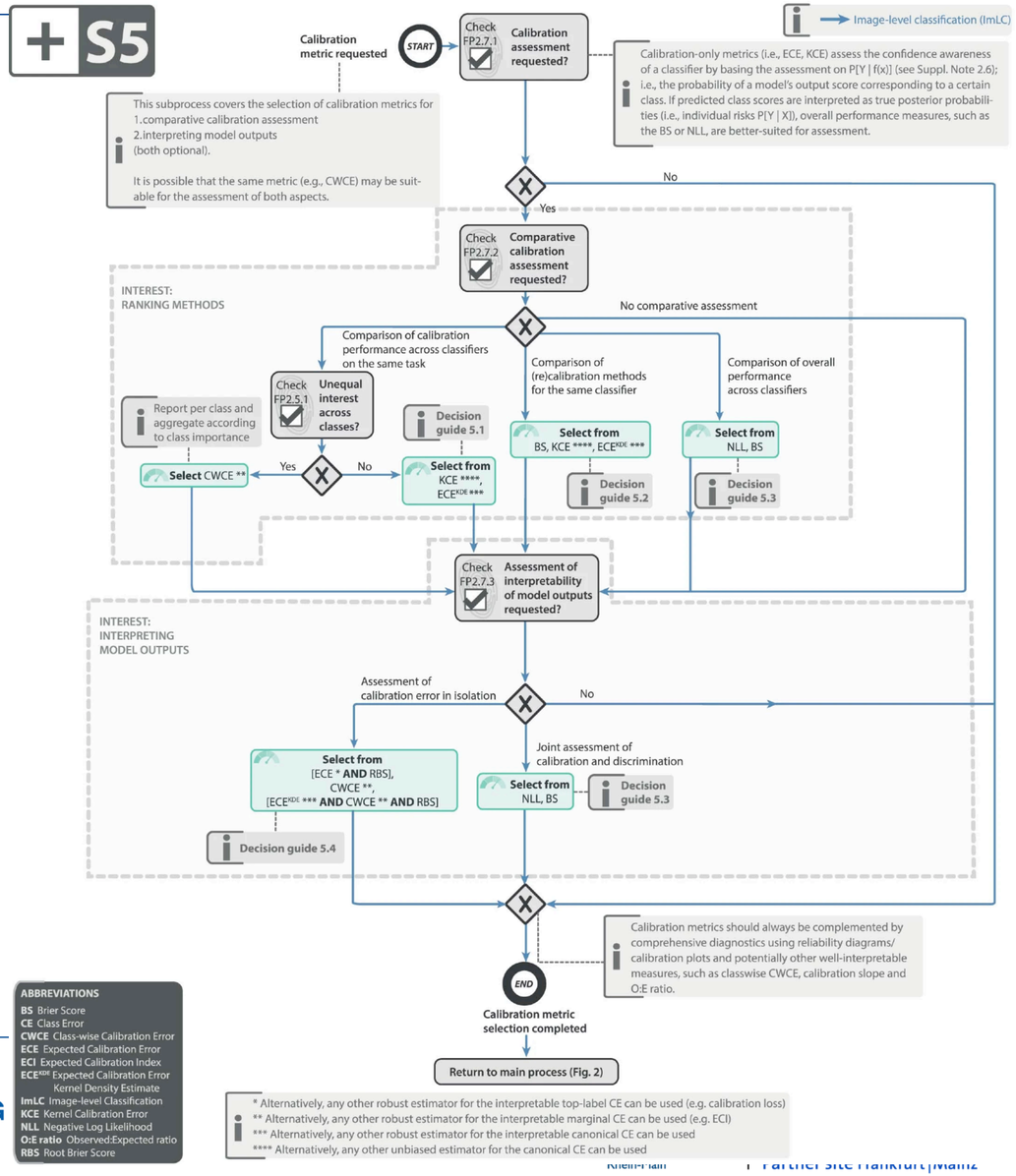
- What metrics are important?
- What domain shifts are important?
- How can we mitigate them?
- Challenge: Specific to domain and deployment environment

Gruber & Buettner, NeurIPS 2022
Gruber, …,Buettner/Blaschko, AISTATS 2024
L. Maier-Hein, …, Buettner, …, Jäger, Nat Meth 2024

DESIGNING MODEL AUDITOR AS AI AGENT

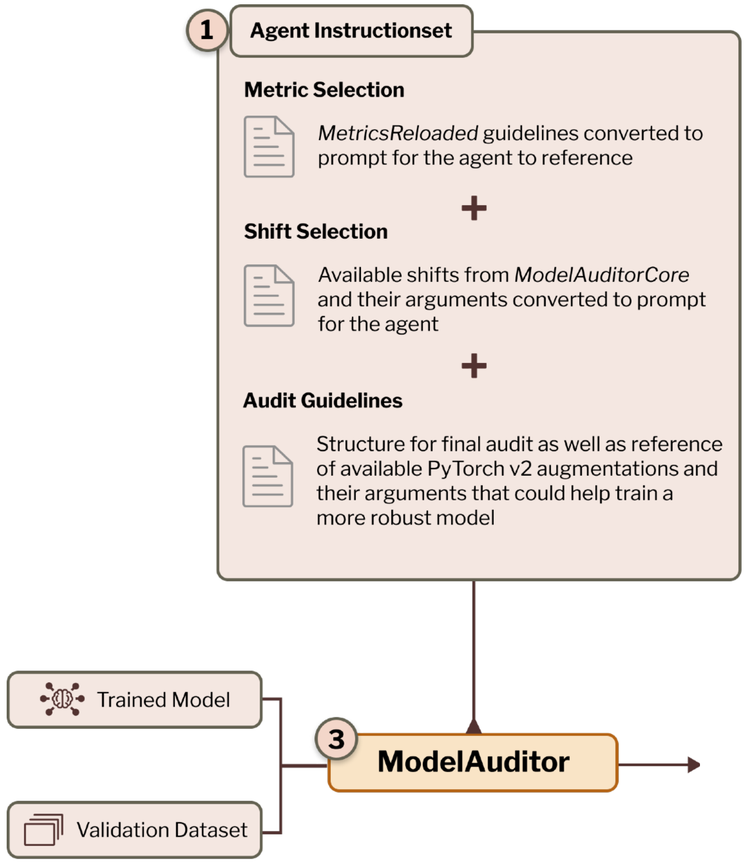
How can we identify and mitigate clinically relevant failure modes?

- What metrics are important?
- What domain shifts are important?
- How can we mitigate them?
- Challenge: Specific to domain and deployment environment

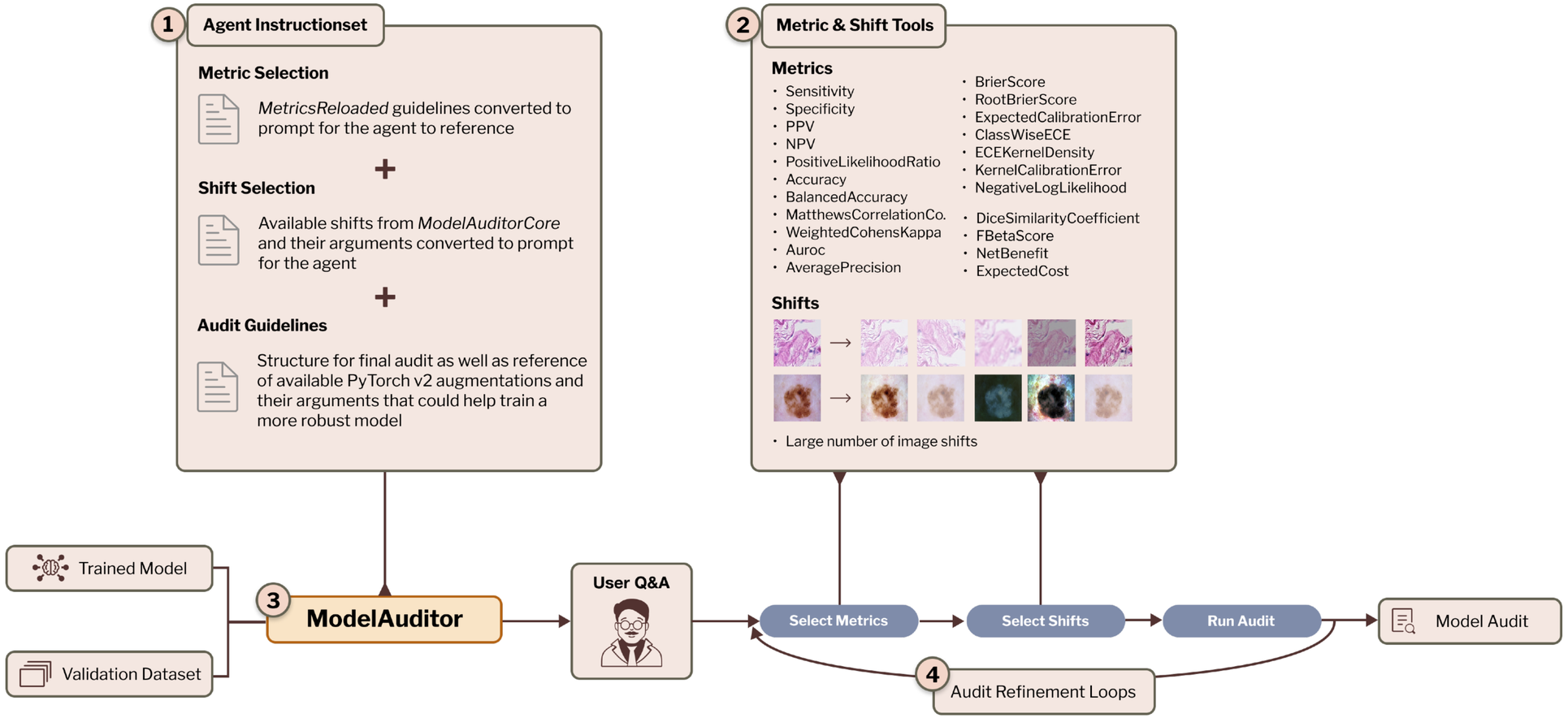


Gruber & Buettner, NeurIPS 2022
Gruber, …,Buettner/Blaschko, AISTATS 2024
L. Maier-Hein, …, Buettner, …, Jäger, Nat Meth 2024

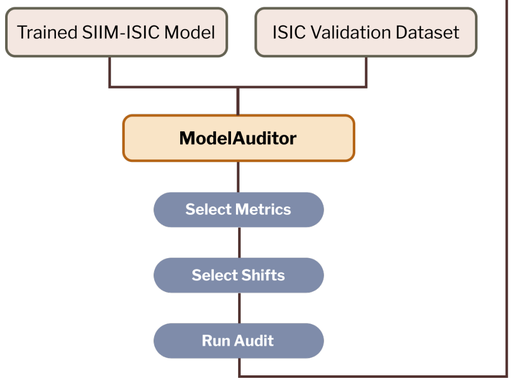
MODEL AUDITOR



MODEL AUDITOR



AUDIT OF SIIM-ISIC MODEL REVEALS ACTIONABLE FAILURE MODES



Analysis of Skin Lesion Classification Model Performance

Key Failure Cases and Strengths

Critical Failure Points:

1. Brightness and Contrast Sensitivity: The **model's performance varies significantly with lighting changes**. Sensitivity drops to 0.47 at normal brightness compared to 0.59 at lower brightness, **indicating inconsistent detection of malignant lesions under different lighting conditions**.

2. Overall Low Sensitivity: Across most transformations, sensitivity ranges from 0.47-0.71, meaning the model misses 30-45% of malignant lesions. This is particularly concerning since false negatives (missed malignancies) are more critical in this medical context.

3. Poor Calibration: Expected calibration error ranges from 0.08-0.11 across all transformations, indicating the model's probability estimates aren't reliable for clinical decision-making.

Notable Strengths

1. Good Discriminative Ability: The model maintains high AUROC values (0.92-0.95) across all transformations, showing strong underlying discriminative power.

2. Zoom Performance: **The model performs best with zoom-in at 0.8 severity (sensitivity 0.71), suggesting that focusing more closely on lesions improves malignancy detection.**

3. Orientation Robustness: Performance is consistent across rotations and flips, with vertical flip showing particularly good sensitivity (0.687).

Real World Impact

These findings have serious clinical implications:

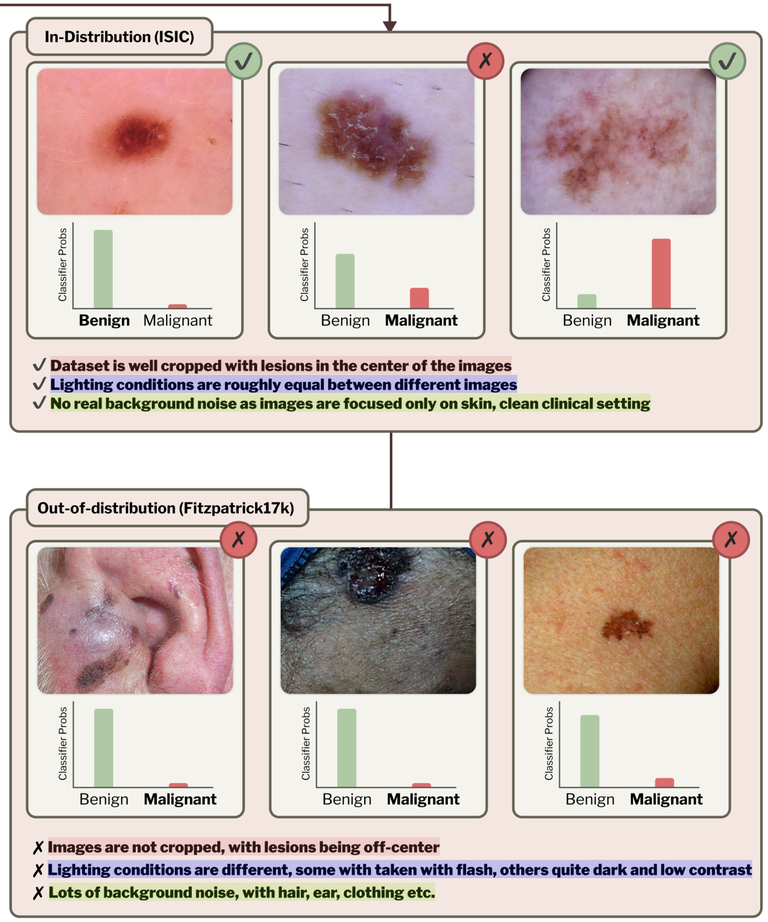
1. Missed Diagnoses: With sensitivity values mostly between 0.55-0.70, the model could miss a significant portion of malignant lesions, potentially leading to delayed treatment and worse patient outcomes.

2. Inconsistent Performance: **The model's sensitivity to lighting conditions means its performance will vary significantly depending on clinical photography settings, camera equipment, and ambient lighting.**

3. Unreliable Confidence Scores: The high calibration error means clinicians cannot reliably interpret the model's confidence, which could lead to inappropriate clinical decisions.

The ModelAuditor then recommended brightness, contrast and saturation augmentations to improve the shortfalls of the model. It further recommended optimization of the threshold for sensitivity as well as temperature scaling for post-hoc calibration.

Manual Verification



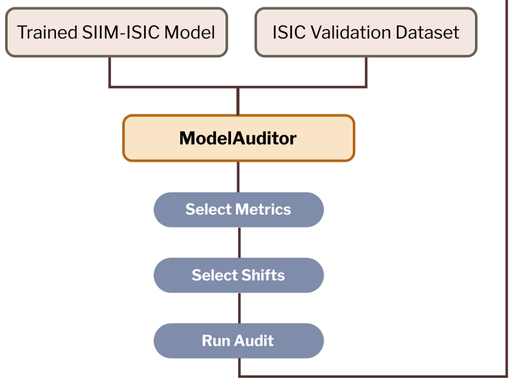
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

MLO MACHINE LEARNING IN ONCOLOGY

uct Universitäre Centren für Tumorerkrankungen Rhein-Main

dkfz. German Cancer Consortium Partner site Frankfurt|Mainz

AUDIT OF SIIM-ISIC MODEL REVEALS ACTIONABLE FAILURE MODES



Analysis of Skin Lesion Classification Model Performance

Key Failure Cases and Strengths

Critical Failure Points:

1. Brightness and Contrast Sensitivity: The **model's performance varies significantly with lighting changes**. Sensitivity drops to 0.47 at normal brightness compared to 0.59 at lower brightness, **indicating inconsistent detection of malignant lesions under different lighting conditions**.

2. Overall Low Sensitivity: Across most transformations, sensitivity ranges from 0.47-0.71, meaning the model misses 30-45% of malignant lesions. This is particularly concerning since false negatives (missed malignancies) are more critical in this medical context.

3. Poor Calibration: Expected calibration error ranges from 0.08-0.11 across all transformations, indicating the model's probability estimates aren't reliable for clinical decision-making.

Notable Strengths

1. Good Discriminative Ability: The model maintains high AUROC values (0.92-0.95) across all transformations, showing strong underlying discriminative power.

2. Zoom Performance: **The model performs best with zoom-in at 0.8 severity (sensitivity 0.71), suggesting that focusing more closely on lesions improves malignancy detection.**

3. Orientation Robustness: Performance is consistent across rotations and flips, with vertical flip showing particularly good sensitivity (0.687).

Real World Impact

These findings have serious clinical implications:

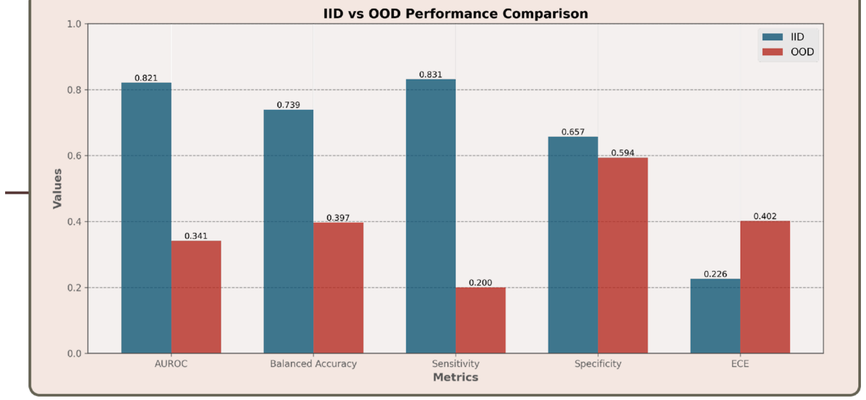
1. Missed Diagnoses: With sensitivity values mostly between 0.55-0.70, the model could miss a significant portion of malignant lesions, potentially leading to delayed treatment and worse patient outcomes.

2. Inconsistent Performance: **The model's sensitivity to lighting conditions means its performance will vary significantly depending on clinical photography settings, camera equipment, and ambient lighting.**

3. Unreliable Confidence Scores: The high calibration error means clinicians cannot reliably interpret the model's confidence, which could lead to inappropriate clinical decisions.

The ModelAuditor then recommended brightness, contrast and saturation augmentations to improve the shortfalls of the model. It further recommended optimization of the threshold for sensitivity as well as temperature scaling for post-hoc calibration.

3 **Results Out-of-distribution (Practitioner selected)**



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

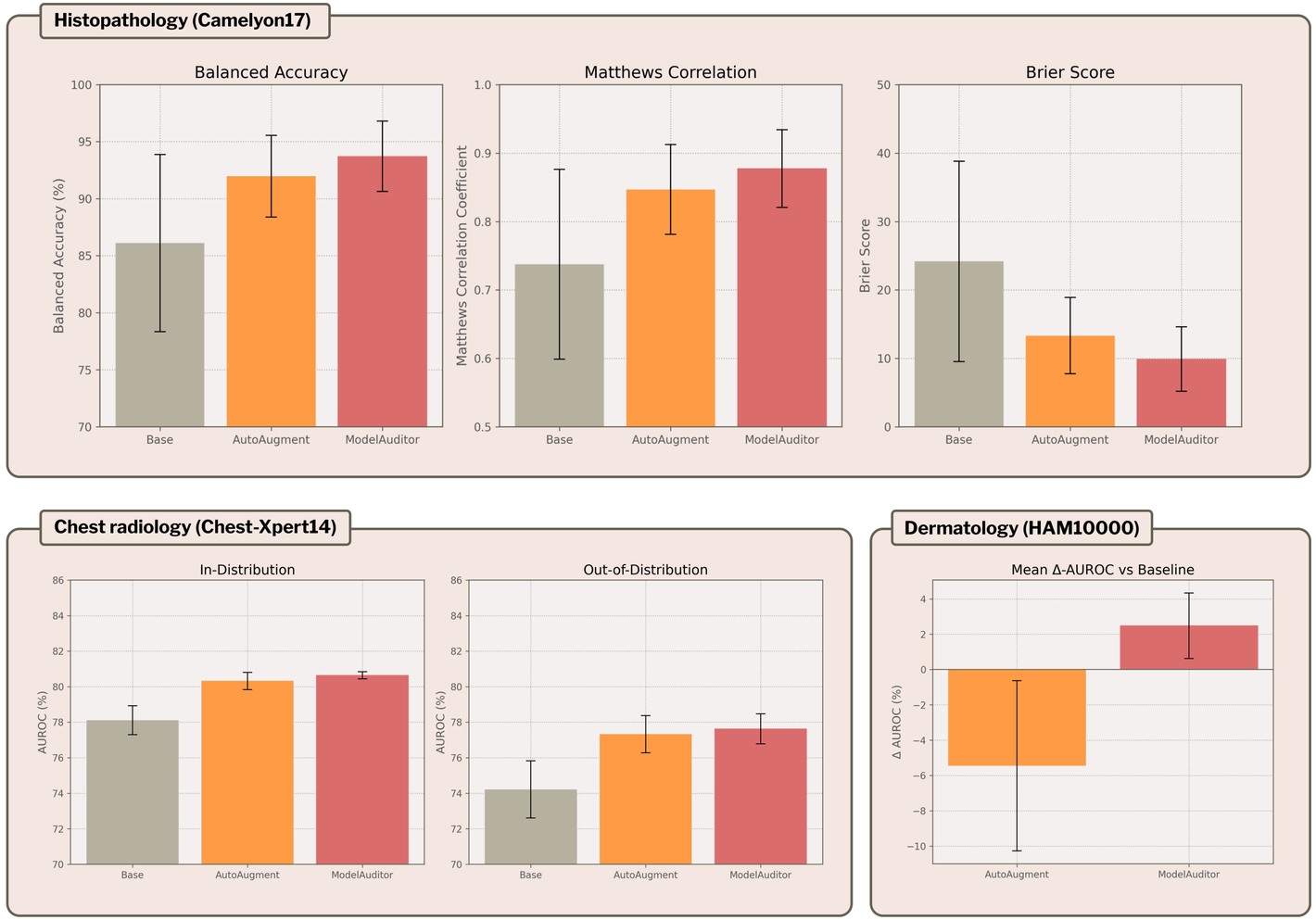
MACHINE LEARNING IN ONCOLOGY

uct Universitäre Centren für Tumorerkrankungen Rhein-Main

dkfz. German Cancer Consortium Partner site Frankfurt|Mainz

MODEL AUDITOR RECOVERS LOST PERFORMANCE UNDER DISTRIBUTION SHIFT

- Perform model audit on different domains
  - Histopathology
  - Chest Radiography
  - Dermatology
- Retrain models using targeted augmentations suggested by Model Auditor
- Compare to AutoAugment as strong baseline



UNCERTAINTY IN GERNERATIVE MODELS

Gruber & Buettner, ICML 2024

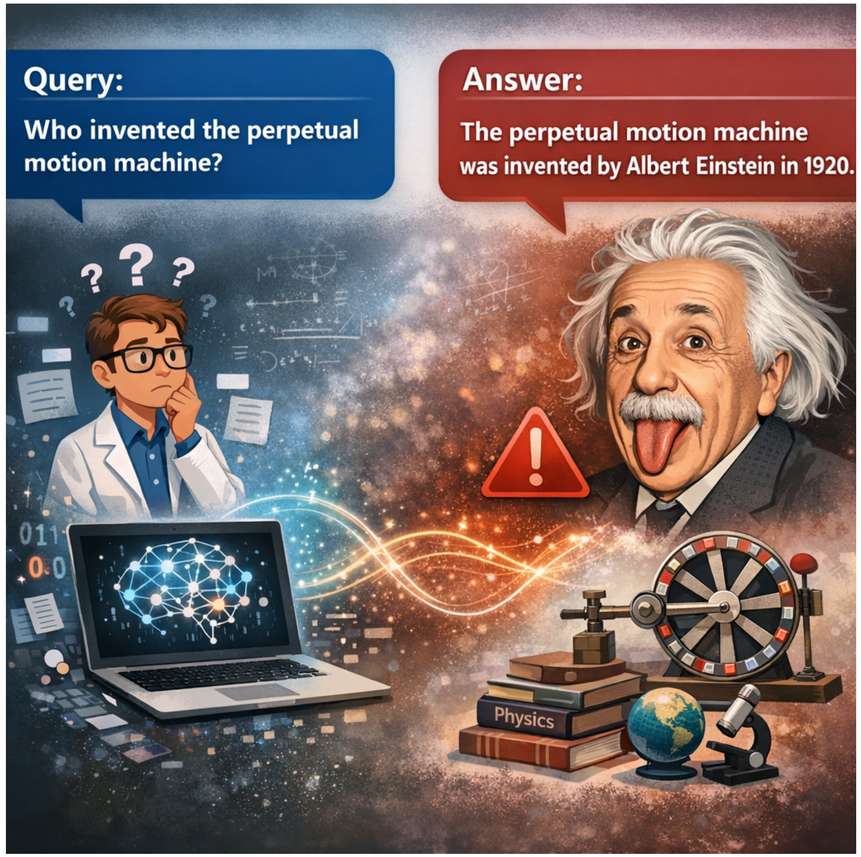
UNCERTAINTY IN GERNERATIVE MODELS

- Uncertainty in classification setting
  - (Calibrated) confidence scores

Gruber & Buettner, ICML 2024

UNCERTAINTY IN GERNERATIVE MODELS

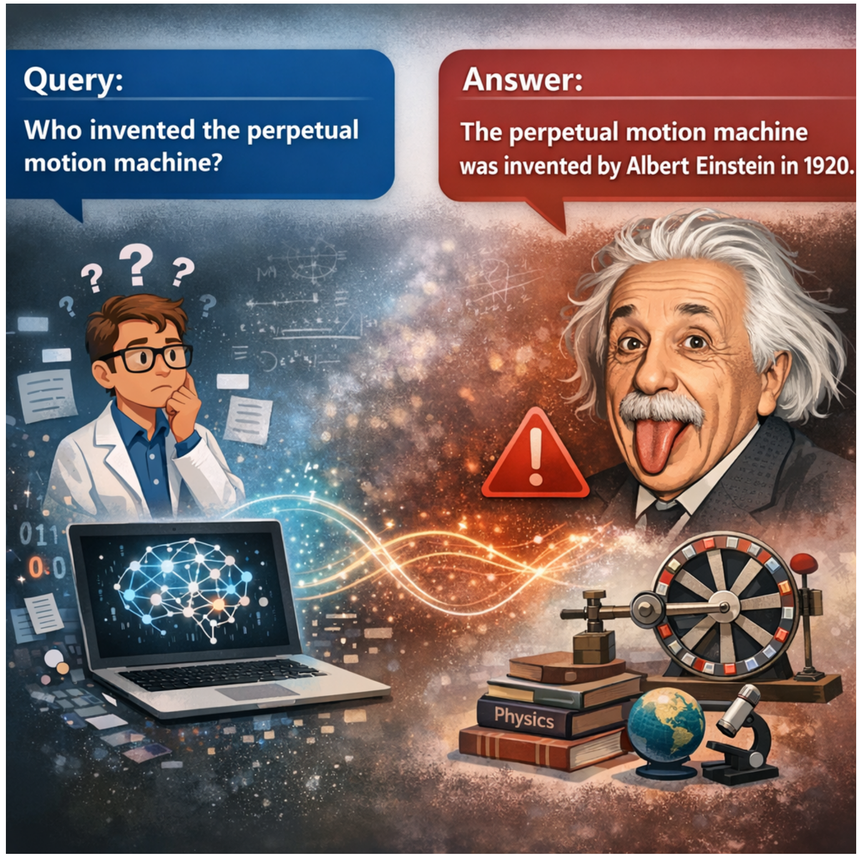
- Uncertainty in classification setting
  - (Calibrated) confidence scores
- How can we quantify uncertainty in generated input?



Gruber & Buettner, ICML 2024

UNCERTAINTY IN GERNERATIVE MODELS

- Uncertainty in classification setting
  - (Calibrated) confidence scores
- How can we quantify uncertainty in generated input?
- Can we use the loss function directly to estimate sample-specific uncertainty?



Gruber & Buettner, ICML 2024

PREDICTIVE UNCERTAINTY IN LLMS

Gruber & Buettner, ICML 2024

PREDICTIVE UNCERTAINTY IN LLMS

- LLMs predict probability distributions of tokens

Gruber & Buettner, ICML 2024

PREDICTIVE UNCERTAINTY IN LLMS

- LLMs predict probability distributions of tokens
- But: we often only have access to outputs

Gruber & Buettner, ICML 2024

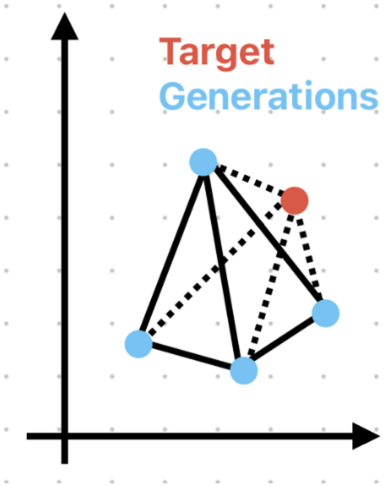
PREDICTIVE UNCERTAINTY IN LLMS

- LLMs predict probability distributions of tokens
- But: we often only have access to outputs
- Is there a score that can be computed on finite outputs only?

Gruber & Buettner, ICML 2024

PREDICTIVE UNCERTAINTY IN LLMS

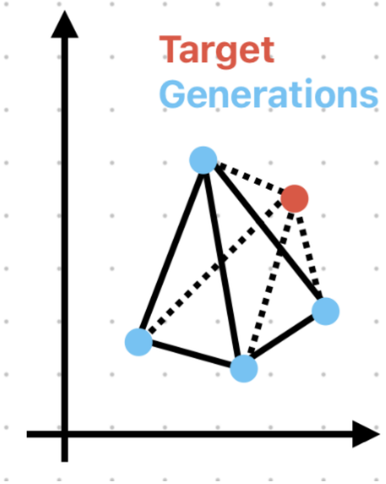
- LLMs predict probability distributions of tokens
- But: we often only have access to outputs
- Is there a score that can be computed on finite outputs only?
- Enter kernel score: $S_k(P, y) = \| P \|_k^2 - 2\, \mathbb{E}_{\hat{Y} \sim P} \left[k\left(\hat{Y}, y\right) \right]$

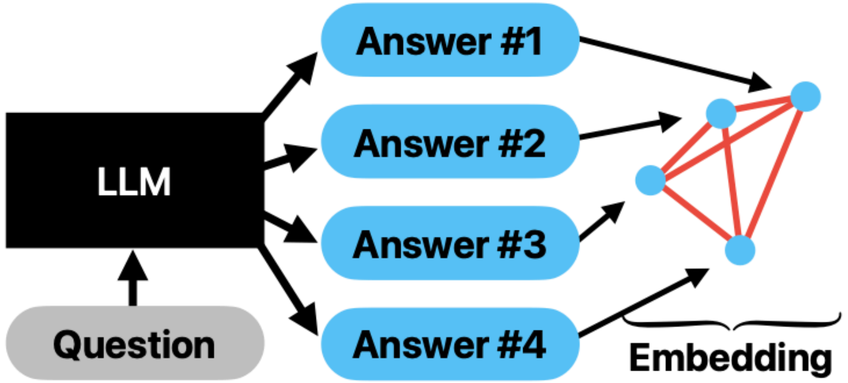


Gruber & Buettner, ICML 2024

PREDICTIVE UNCERTAINTY IN LLMS

- LLMs predict probability distributions of tokens
- But: we often only have access to outputs
- Is there a score that can be computed on finite outputs only?
- Enter kernel score: $S_k(P, y) = \| P \|_k^2 - 2\, \mathbb{E}_{\hat{Y} \sim P} \left[k\left(\hat{Y}, y\right) \right]$

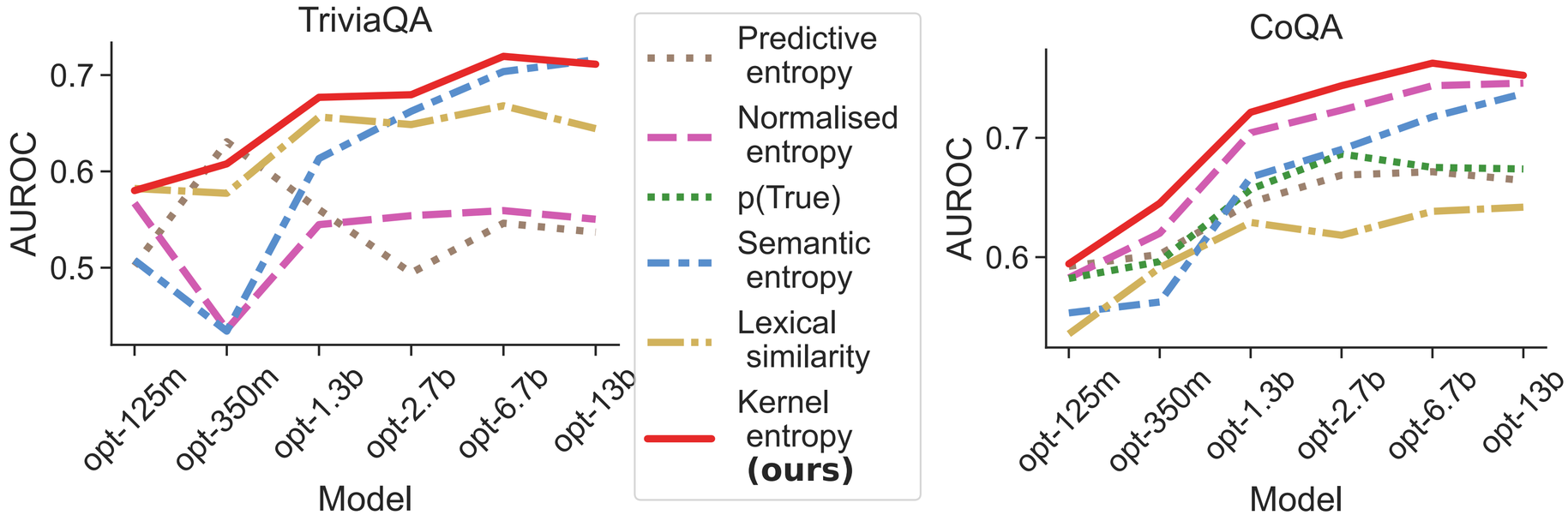




Gruber & Buettner, ICML 2024

APPLICATION LLM

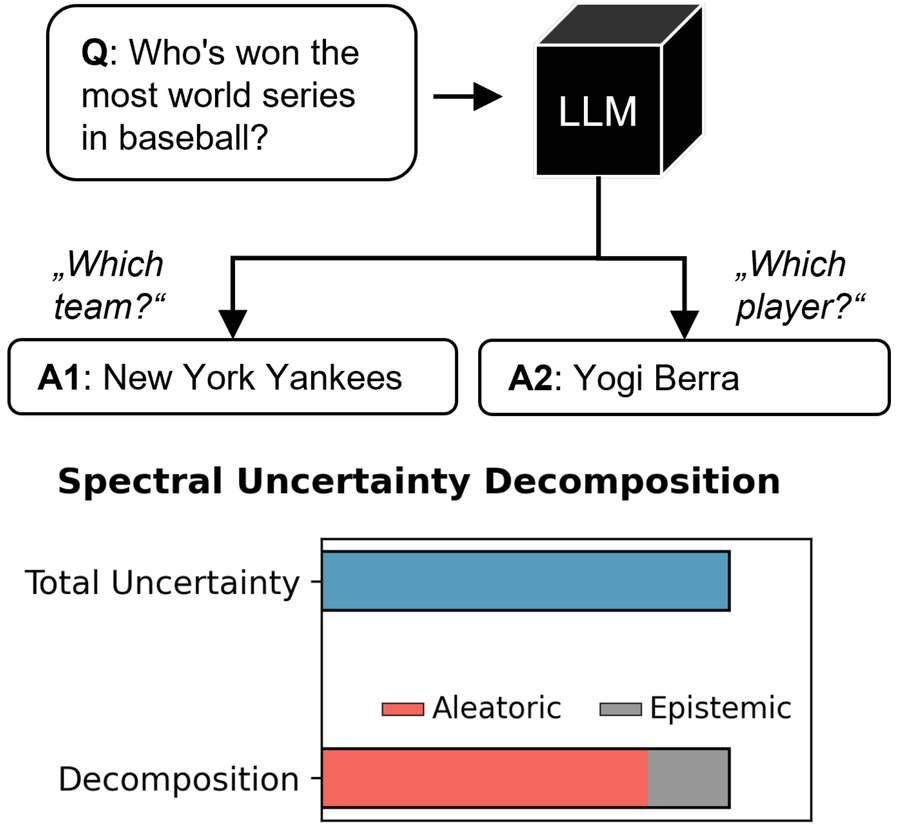
- Evaluate Kernel Entropy as uncertainty measure in Q&A tasks
  - Threshold kernel entropy to compute AUC for correct answer



UNCERTAINTY BEYOND HALLUCINATIONS

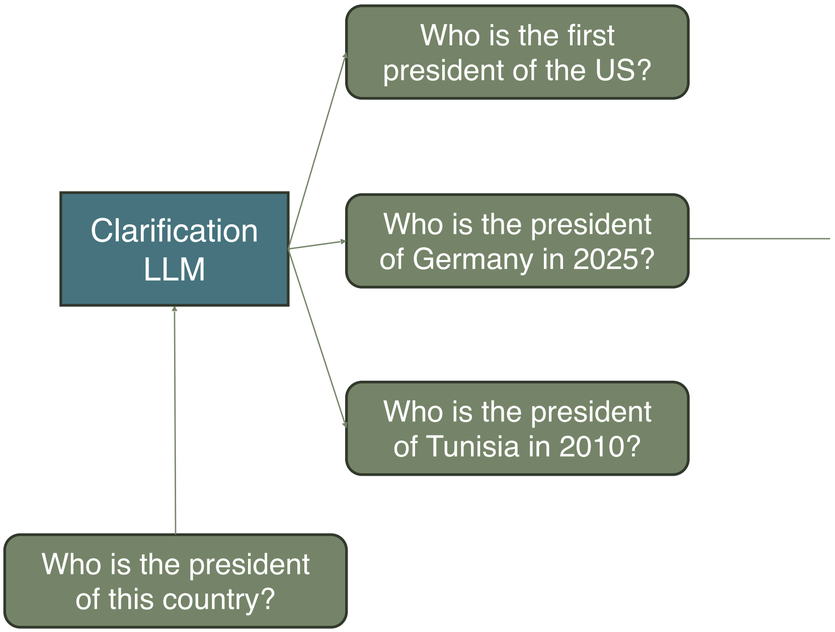
Walha, Gruber…,Buettner, AAAI 2026

UNCERTAINTY BEYOND HALLUCINATIONS



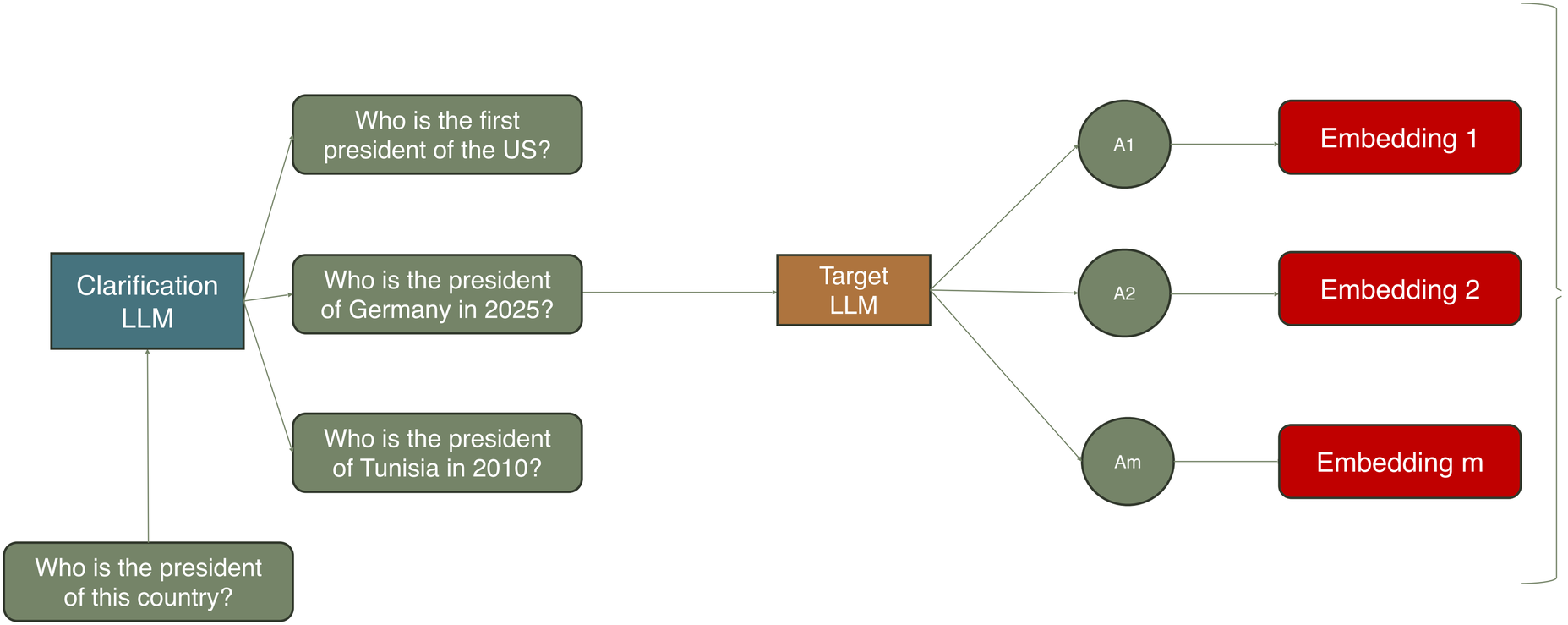
Walha, Gruber…,Buettner, AAAI 2026

SPECTRAL UNCERTAINTY



Generate n clarifications of the original question using an auxiliary model (outer samples)

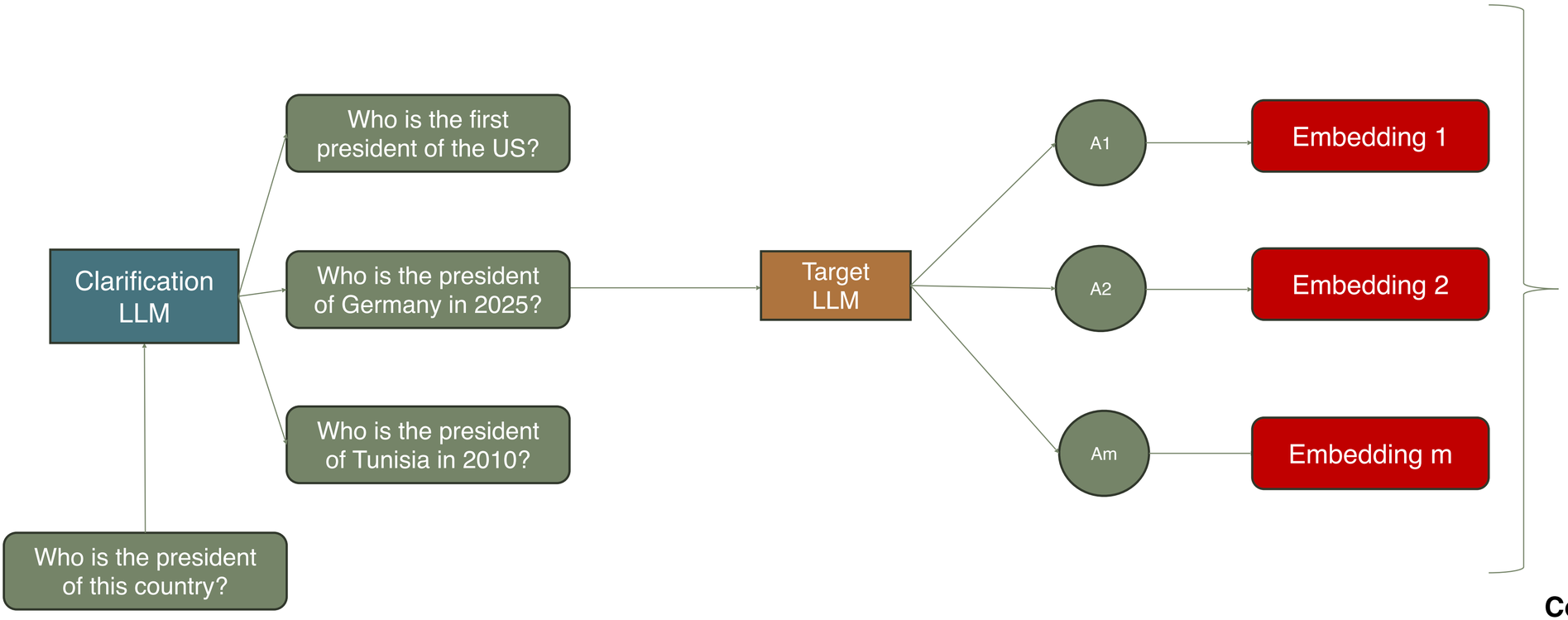
SPECTRAL UNCERTAINTY



Generate n clarifications of the original question using an auxiliary model (outer samples)

Generate m answers for each clarification using the target model (inner samples) and compute their embeddings

SPECTRAL UNCERTAINTY



$$SpecE\left(Clarification_1\right) = -Tr[K_1 \log K_1]$$

$$SpecE_{total} = -Tr[K_{total} \log K_{total}]$$

Generate n clarifications of the original question using an auxiliary model (outer samples)

Generate m answers for each clarification using the target model (inner samples) and compute their embeddings

Compute Kernel Matrices using the embeddings:

- $K_i \in R^{m \times m}$ **for every clarification i**
- $K_{total} \in R^{nm \times nm}$ **for all generated answer embeddings**

Then compute the VNEs of all matrices.

WHY IT WORKS

$$H_{VN}\left(\mathbb{P}_Y\right) = \mathbb{B}_{-H_{VN}}\left(\mathbb{P}_{Y|C}\right) + \mathbb{E}_C\left[H_{VN}\left(\mathbb{P}_{Y|C}\right)\right]$$

$SpecE_{total}$ $\qquad\qquad$ $\dfrac{\sum_{i=1}^{n} SpecE(Clarification_i)}{n}$

Total uncertainty **Aleatoric uncertainty** **Epistemic uncertainty**

$$-\sum_{i=1}^{nm} \hat{\lambda}_i^{\text{out}} \log \hat{\lambda}_i^{\text{out}} \qquad -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{\lambda}_{ij} \log \hat{\lambda}_{ij}$$

BETTER AMBIGUITY DETECTION AND TOTAL UNCERTAINTY

Ambiguity detection

BETTER AMBIGUITY DETECTION AND TOTAL UNCERTAINTY

Ambiguity detection

Uncertainty Method	Phi-4 14B		LLaMA 4 Maverick	
	AUROC (%)	AUPR (%)	AUROC (%)	AUPR (%)
AmbigQA				
Semantic Entropy	53.29	51.85	46.14	49.36
Kernel Language Entropy	49.88	48.11	45.59	48.84
Predictive Kernel Entropy	48.37	48.94	45.10	49.12
Input Clarification Ensembling (aleatoric)	63.46	62.23	59.51	60.12
Spectral Uncertainty (aleatoric)	**69.15**	**67.98**	**60.39**	**60.48**
AmbigInst				
Semantic Entropy	60.58	69.18	55.88	64.37
Kernel Language Entropy	60.60	69.35	55.80	61.83
Predictive Kernel Entropy	75.93	79.90	66.83	71.31
Input Clarification Ensembling (aleatoric)	71.70	80.62	69.66	79.04
Spectral Uncertainty (aleatoric)	**86.37**	**90.10**	**85.95**	**89.46**

BETTER AMBIGUITY DETECTION AND TOTAL UNCERTAINTY

Ambiguity detection

Uncertainty Method	Phi-4 14B		LLaMA 4 Maverick	
	AUROC (%)	AUPR (%)	AUROC (%)	AUPR (%)
AmbigQA				
Semantic Entropy	53.29	51.85	46.14	49.36
Kernel Language Entropy	49.88	48.11	45.59	48.84
Predictive Kernel Entropy	48.37	48.94	45.10	49.12
Input Clarification Ensembling (aleatoric)	63.46	62.23	59.51	60.12
Spectral Uncertainty (aleatoric)	**69.15**	**67.98**	**60.39**	**60.48**
AmbigInst				
Semantic Entropy	60.58	69.18	55.88	64.37
Kernel Language Entropy	60.60	69.35	55.80	61.83
Predictive Kernel Entropy	75.93	79.90	66.83	71.31
Input Clarification Ensembling (aleatoric)	71.70	80.62	69.66	79.04
Spectral Uncertainty (aleatoric)	**86.37**	**90.10**	**85.95**	**89.46**

Total uncertainty

Uncertainty Method	Phi-4 14B		LLaMA 4 Maverick	
	AUROC (%)	AUPR (%)	AUROC (%)	AUPR (%)
TriviaQA				
Semantic Entropy	84.70	71.10	71.64	43.75
Kernel Language Entropy	86.20	76.64	71.95	45.72
Predictive Kernel Entropy	85.88	74.66	73.85	45.86
Input Clarification Ensembling (total)	89.45	74.54	82.76	55.95
Spectral Uncertainty (total)	**91.92**	**80.79**	**84.82**	**60.84**
Natural Questions (NQ)				
Semantic Entropy	76.24	74.36	70.24	52.35
Kernel Language Entropy	**82.77**	81.84	71.60	60.79
Predictive Kernel Entropy	77.67	76.57	70.56	58.73
Input Clarification Ensembling (total)	81.91	81.04	74.93	60.72
Spectral Uncertainty (total)	81.63	**81.98**	**75.02**	**62.87**

BUT HOW DO I CALIBRATE IT?

BUT HOW DO I CALIBRATE IT?

STAY TUNED…..

SUMMARY

SUMMARY

- Formalise and measure calibration
  - MetricsReloaded guides practitioners to the right calibration metric for their tasks

SUMMARY

- Formalise and measure calibration
  - MetricsReloaded guides practitioners to the right calibration metric for their tasks
- Audit → Improve → Monitor
  - ModelAuditor identifies clinically relevant failure modes and recommends targeted fixes

SUMMARY

- Formalise and measure calibration
  - MetricsReloaded guides practitioners to the right calibration metric for their tasks
- Audit → Improve → Monitor
  - ModelAuditor identifies clinically relevant failure modes and recommends targeted fixes
- Principled and practical uncertainty quantification for LLMs via spectral entropy

ACKNOWLEDGEMENTS

MLO Lab
Achim Hekler
Lukas Kuhn
Nassim Walha
Sebastian G. Gruber
Azza Jenane
Giuseppe Serra
Thomas Decker
Dustin Eisenhardt
Alex Koebler
Hendrik Mertens
Arber Qoku
Kevin De Azevedo
Sareh AmeriFar
A. Yavuz Çakır
Tyra Stickel
Yihao Liu
…

LMU Munich
Eyke Hüllermeier
Alireza Javanmardi

DKFZ
Lena Maier-Hein
Paul Jäger
Annika Reinke

MERCK
SIEMENS
Ingenuity for life
DFG
Deutsche Forschungsgemeinschaft
Federal Ministry for Economic Affairs and Climate Action
erc
European Research Council

mlo-lab.github.io

Florian Buettner
florian.buettner@dkfz.de