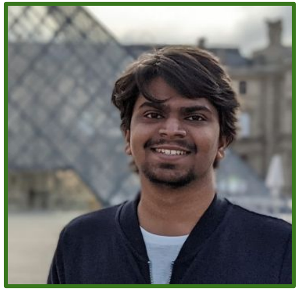
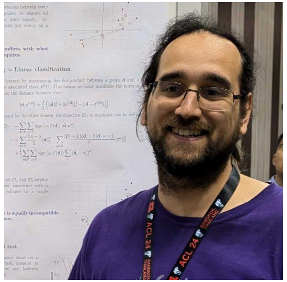
UNIVERSITÉ DE LORRAINE

UNIVERSITY OF HELSINKI

Data complexity is not Data uncertainty



Aman Sinha[1]



Timothee Mickus[2]

Raul Vazquez[2]

[1]Université de Lorraine, Nancy, France
[2]University of Helsinki, Finland

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

- My thesis topic broadly related to **Medical language understanding**.

Bonjour! Hello! I'm Aman Sinha

- I'm a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

- My thesis topic broadly related to **Medical language understanding**.
  - medical data sources, multi-modality, and structures.

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

- My thesis topic broadly related to **Medical language understanding**.
  - medical data sources, multi-modality, and structures.
  - interpretability of medical models

Bonjour! Hello! I'm Aman Sinha

- I'm a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

- My thesis topic broadly related to **Medical language understanding**.
  - medical data sources, multi-modality, and structures.
  - interpretability of medical models
  - uncertainty & reliability of medical models.

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are:

Marianne Clausel (IECL)

Mathieu Constant (ATILF)

Xavier Coubez (ICANS)

- My thesis topic broadly related to **Medical language understanding**.
    - medical data sources, multi-modality, and structures.
    - interpretability of medical models
    - uncertainty & reliability of medical models.

Bonjour! Hello! I'm Aman Sinha

- I'm a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are: Marianne Clausel (IECL), Mathieu Constant (ATILF), and Xavier Coubez (ICANS)
- My thesis topic broadly related to **Medical language understanding**.

Today, I will be talking about the interaction of data uncertainty and data complexity.

Bonjour! Hello! I’m Aman Sinha

- I’m a final year PhD Student at Université de Lorraine, Nancy and ICANS Strasbourg
- My thesis supervisors are: Marianne Clausel (IECL), Mathieu Constant (ATILF), and Xavier Coubez (ICANS)
- My thesis topic broadly related to **Medical language understanding**.

Today, I will be talking about the interaction of data uncertainty and data complexity.

Your Model is Overconfident, and Other Lies We Tell Ourselves

Timothee Mickus **Aman Sinha** **Raúl Vázquez**
University of Helsinki Université de Lorraine ICANS Strasbourg
firstname.lastname@{[1]helsinki.fi,[2]univ-lorraine.fr}

To appear at ACL Main 2025, Vienna, this July

Context

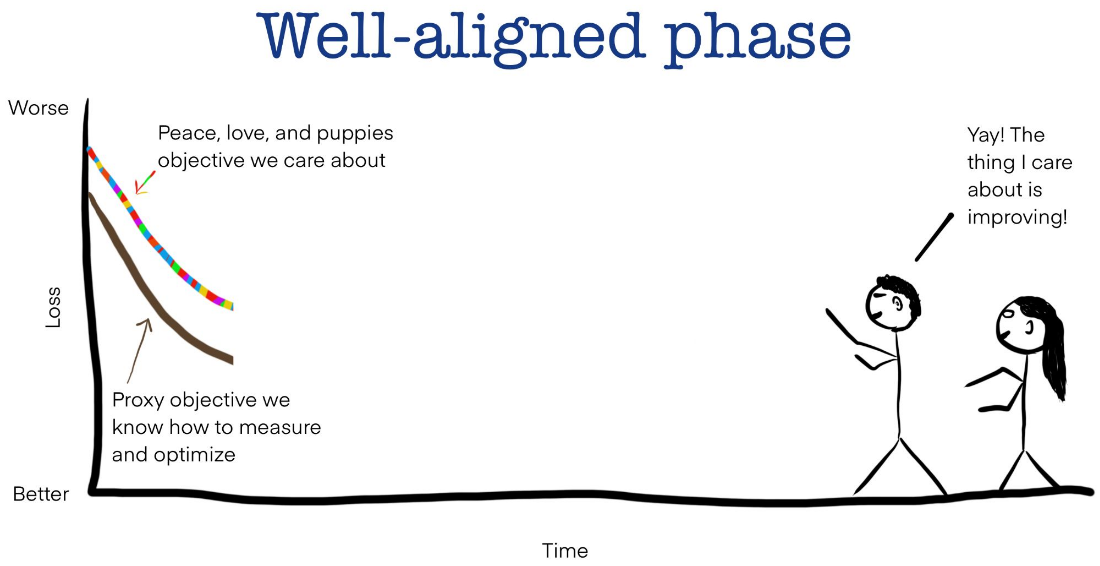
Motivation

Context

Source : https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

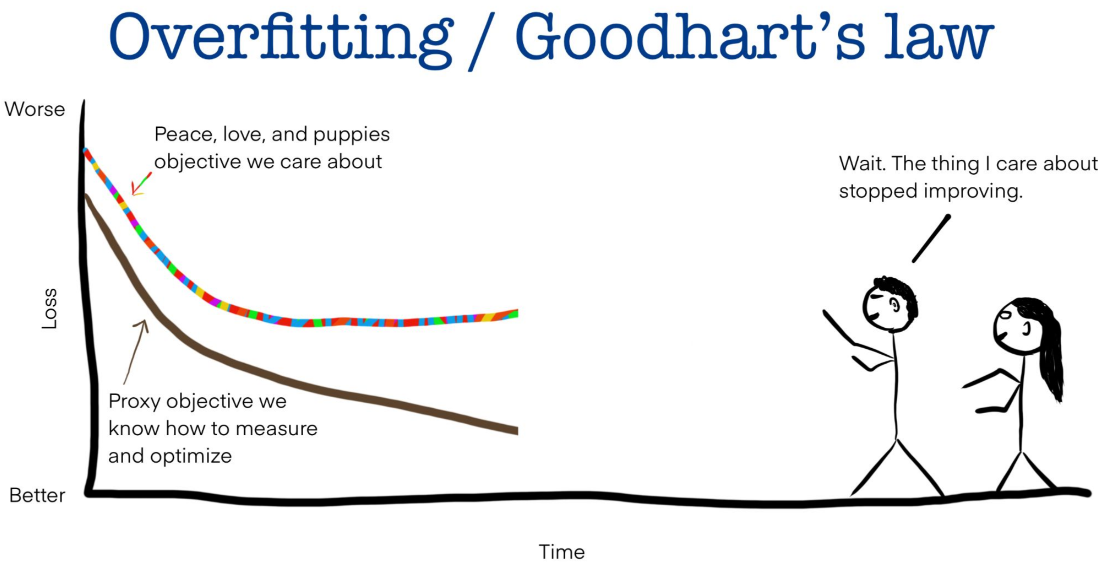
Context

Source : https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html



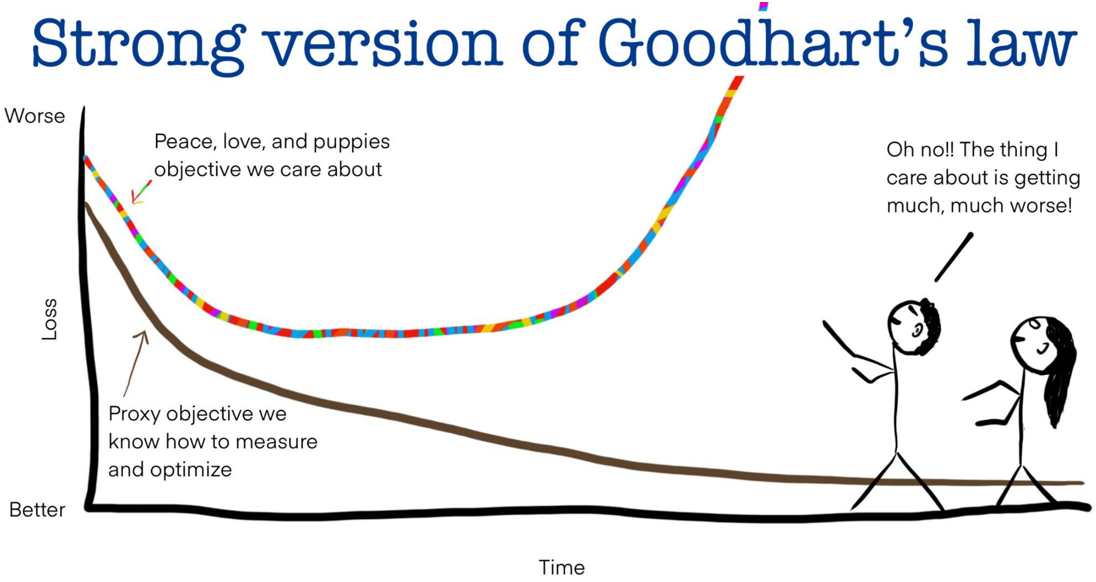
Source : https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

Context



Source : https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

Context



"When a measure becomes a target, if it is effectively optimized, then the thing it is designed to measure will grow worse."

Source : https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html

Motivation - I

- We expect any model to be good in terms of *performance*, *interpretation*, and *calibration*, as a wholesome behavior.

Motivation - I

- We expect any model to be good in terms of ***performance***, ***interpretation***, and ***calibration***, as a wholesome behavior.

From Greek mythology, Cerberus, often referred to as the hound of Hades

Wiki : https://en.wikipedia.org/wiki/Cerberus

Motivation - I

- We expect any model to be good in terms of ***performance***, ***interpretation***, and ***calibration***, as a wholesome behavior.

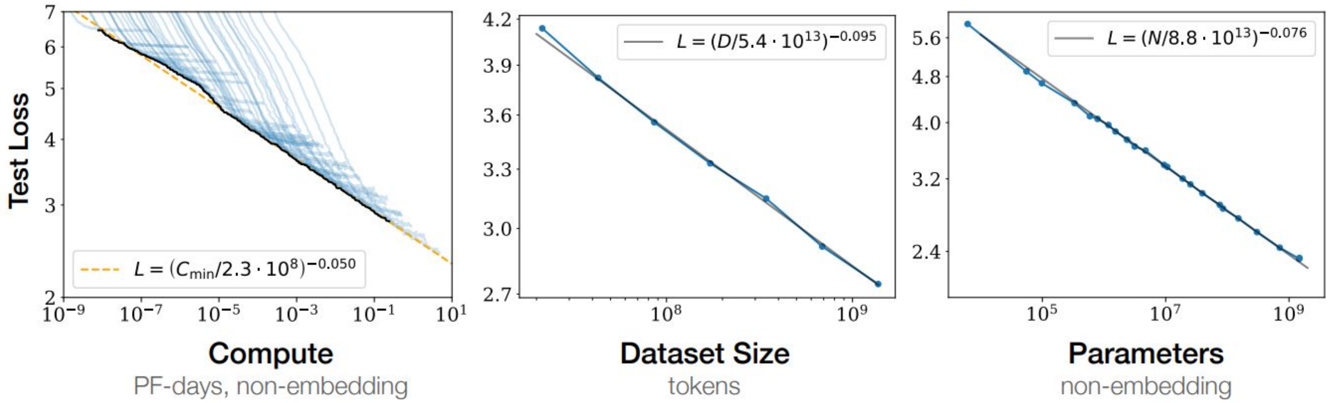


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al., 2020 : https://arxiv.org/pdf/2001.08361

Motivation - I

- We expect any model to be good in terms of *performance*, ***interpretation***, and *calibration*, as a wholesome behavior.

Decision Tree (if/else) → **LLMs**

```python
def recommend_activity_marseille(interest, preference):
    if interest == "outdoors":
        if preference == "land":
            return "Calanques Hiking"
        if preference == "sea":
            return "Boat Tour"

    if interest == "culture":
        if preference == "modern":
            return "MuCEM Museum"
        if preference == "historic":
            return "Old Port Walk"
```

Source : https://huggingface.co/mlabonne/BigLlama-3.1-1T-Instruct

Motivation - I

- We expect any model to be good in terms of ***performance***, ***interpretation***, and ***calibration,*** as a wholesome behavior.

What does being calibrated imply?

[Guo et al 2017] : https://arxiv.org/abs/1706.04599

💡 What does being *calibrated* imply?

- If a model is calibrated (aka **uncertainty-aware** or **reliable**) ,

P (model is correct | **confidence** is α) = α

It means, α-fraction of the predicted classes with α **confidence** should be correct.

Source : https://sites.google.com/view/uncertainty-nlp

Motivation - I

▸ We expect any model to be good in terms of *performance*, *interpretation*, and ***calibration,*** as a wholesome behavior.

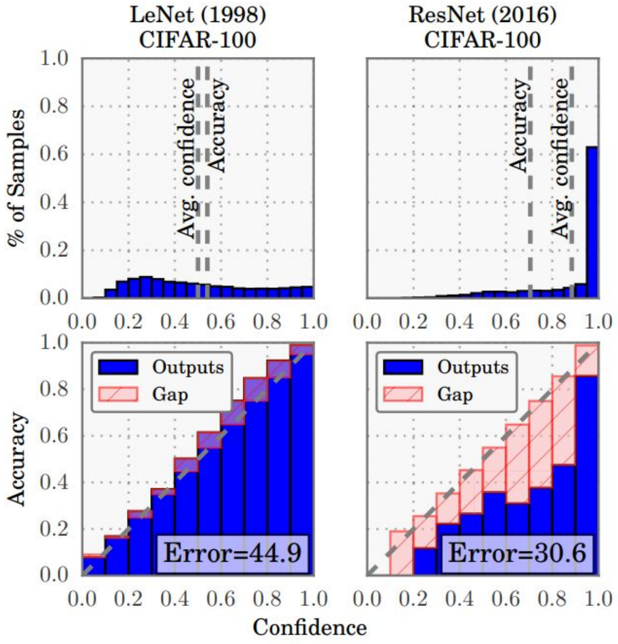


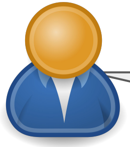
Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

Modern neural networks exhibit a strange phenomenon: **probabilistic error and miscalibration worsen** even as classification error is reduced [Guo et al 2017].

[Guo et al 2017] : https://arxiv.org/abs/1706.04599

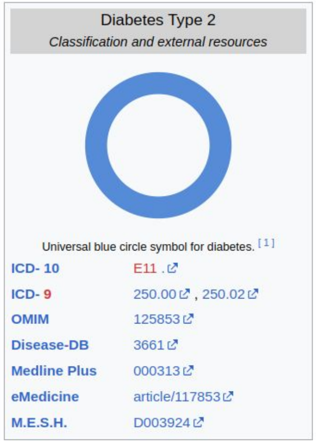
Motivation - I

- We expect any model to be good in terms of *performance*, *interpretation*, and ***calibration***, as a wholesome behavior.



आईसीडी १० E११ का मुख्य कारण क्या है ?
(Quelle est la cause principale du code CIM-10 E11 ?)

आईसीडी १० ई११ का मुख्य कारण पर्यावरणीय क्षरण है।
(La cause principale du code CIM-10 E11 est la dégradation de l'environnement.)



Diabetes Type 2
Classification and external resources

Universal blue circle symbol for diabetes. [1]

ICD-10	E11 .
ICD-9	250.00 , 250.02
OMIM	125853
Disease-DB	3661
Medline Plus	000313
eMedicine	article/117853
M.E.S.H.	D003924

Source : Mu-SHROOM Shared Task 2025 🍄 (Vázquez et al. 2025)

Motivation - II

- What makes an example unclear?

Motivation - II

- What makes an example unclear?

Uncertainty quantification → $P(y^*|\, x, \mathcal{D}) = \int \; P(y^*|x, \Theta) \;\; P(\Theta|\, \mathcal{D}) \quad d\Theta$

Motivation - II

- What makes an example unclear?

Uncertainty quantification ⇒ $$P(y^*|\, x, D) = \int \underbrace{P(y^*|x, \Theta)}_{\textit{data}}\ \underbrace{P(\Theta|\, D)}_{\textit{model}}\ d\Theta$$

Motivation - II

- What makes an example unclear?

Uncertainty quantification ⇒ $P(y^*|\, x, \mathcal{D}) = \int \underbrace{P(y^*|x, \Theta)}_{\textit{data}} \; \underbrace{P(\Theta|\, \mathcal{D})}_{\textit{model}} \; d\Theta$ ⇒ **intractable**

Motivation - II

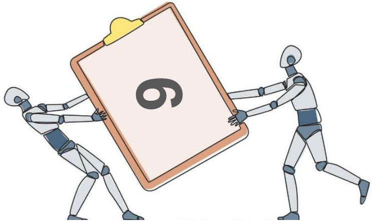
- What makes an example unclear?

Uncertainty quantification

$$P(y^*|\, x, D) = \int \underbrace{P(y^*|x, \Theta)}_{\textit{data}} \; \underbrace{P(\Theta|\, D)}_{\textit{model}} \; d\Theta$$

intractable

"inherent ambiguity or noise of the language"



Source : Google Images

Motivation - II

- When is an example unclear?

Motivation - II

- When is an example unclear?

> to amongst others attention slips. Crucially, HLV assumes humans usually provide their best judgements, and variation emerges due to, e.g., ambiguity of the instance, uncertainty of the annotator, genuine disagreement, or simply the fact that multiple options are correct. Aggregation obfuscates this real-world complexity.

Plank et al., 2022

Motivation - II

- When is an example unclear?

to amongst others attention slips. Crucially, HLV assumes humans usually provide their best judgements, and variation emerges due to, e.g., ambiguity of the instance, uncertainty of the annotator, genuine disagreement, or simply the fact that multiple options are correct. Aggregation obfuscates this real-world complexity.

Plank et al., 2022

- **Aleatoric Uncertainty** It is also known as data uncertainty, which refers to the uncertainty inherent in data due to its randomness or noise. This type of uncertainty is irreducible, meaning it cannot be eliminated through model improvements or tuning. It can arise from a variety of sources, such as noisy observations, overlapping classes, ground truth errors, inherent randomness, or other factors that are not entirely predictable.

Hu et al., 2023

Motivation - II

- When is an example unclear?

to amongst others attention slips. Crucially, HLV assumes humans usually provide their best judgements, and variation emerges due to, e.g., ambiguity of the instance, uncertainty of the annotator, genuine disagreement, or simply the fact that multiple options are correct. Aggregation obfuscates this real-world complexity.

Plank et al., 2022

- **Aleatoric Uncertainty** It is also known as data uncertainty, which refers to the uncertainty inherent in data due to its randomness or noise. This type of uncertainty is irreducible, meaning it cannot be eliminated through model improvements or tuning. It can arise from a variety of sources, such as noisy observations, overlapping classes, ground truth errors, inherent randomness, or other factors that are not entirely predictable.

Hu et al., 2023

Easy examples: (Low $\mathrm{PD}_{\mathrm{Val.}}$, Low $\mathrm{PD}_{\mathrm{Train}}$). Such examples are often visually typical members of their class and the predicted label nearly always matches the ground truth.

Looks like a different class: (Low $\mathrm{PD}_{\mathrm{Val.}}$, High $\mathrm{PD}_{\mathrm{Train}}$). In the validation set, there is a clear (and nearly always incorrect) classification for such an input, but it is difficult to connect such inputs to other examples of their ground truth class during training. Mislabeled examples are of this kind, as are visually confusing images which at first appear to show something else.

Ambiguous unless the label is given: (High $\mathrm{PD}_{\mathrm{Val.}}$, Low $\mathrm{PD}_{\mathrm{Train}}$). These examples are difficult to connect to their predicted class in the validation split but easy to connect to their ground truth class during training. These points may, for example, visually resemble both their own class and another class. They are likely to be misclassified.

Ambiguous: (High $\mathrm{PD}_{\mathrm{Val.}}$, High $\mathrm{PD}_{\mathrm{Train}}$). These examples may be corrupted or show an example of a rare sub-class. Predictions for these inputs can depend strongly on the random seed used for training and initialization.

Baldock et al., 2021

Complexity, uncertainty, human variation

Complexity, uncertainty, human variation

- Same root causes

Complexity, uncertainty, human variation

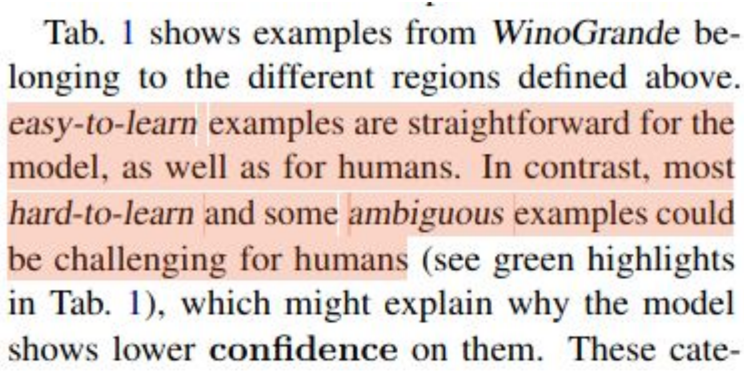
- Same root causes

Tab. 1 shows examples from *WinoGrande* belonging to the different regions defined above. *easy-to-learn* examples are straightforward for the model, as well as for humans. In contrast, most *hard-to-learn* and some *ambiguous* examples could be challenging for humans (see green highlights in Tab. 1), which might explain why the model shows lower **confidence** on them. These cate-

Swayamdipta et al., 2020

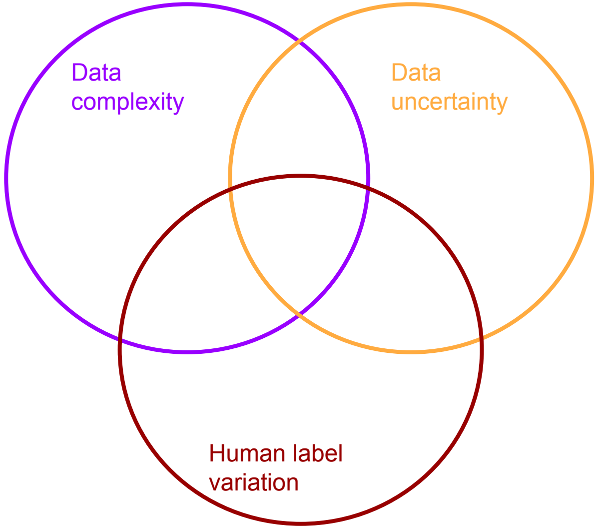
Complexity, uncertainty, human variation

- Same root causes



Tab. 1 shows examples from *WinoGrande* belonging to the different regions defined above. *easy-to-learn* examples are straightforward for the model, as well as for humans. In contrast, most *hard-to-learn* and some *ambiguous* examples could be challenging for humans (see green highlights in Tab. 1), which might explain why the model shows lower **confidence** on them. These cate-

Swayamdipta et al., 2020

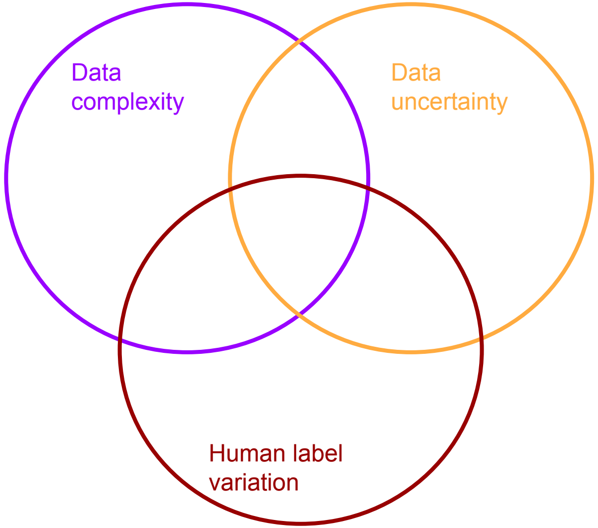


Complexity, uncertainty, human variation

- Same root causes

Tab. 1 shows examples from *WinoGrande* belonging to the different regions defined above. *easy-to-learn* examples are straightforward for the model, as well as for humans. In contrast, most *hard-to-learn* and some *ambiguous* examples could be challenging for humans (see green highlights in Tab. 1), which might explain why the model shows lower **confidence** on them. These cate-

Swayamdipta et al., 2020



Are these truly similar concepts?

Experimental Protocol

Dataset : ChaosNLI

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** **E E E N N**	**Entailment** **$E^{(50)}$ $N^{(50)}$**	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** **E E N N E**	**Neutral** **$N^{(93)}$ $E^{(7)}$**	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** **C N N C C**	**Contradiction** **$C^{(56)}$ $N^{(44)}$**	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** **N E N N E**	**Entailment** **$E^{(94)}$ $N^{(5)}$ $C^{(1)}$**	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Source : [What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data (ChaosNLI)?]

Dataset : ChaosNLI

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** **E E E N N**	**Entailment** $\mathbf{E^{(50)}\ N^{(50)}}$	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** **E E N N E**	**Neutral** $\mathbf{N^{(93)}\ E^{(7)}}$	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** **C N N C C**	**Contradiction** $\mathbf{C^{(56)}\ N^{(44)}}$	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** **N E N N E**	**Entailment** $\mathbf{E^{(94)}\ N^{(5)}\ C^{(1)}}$	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Re-annotation of SNLI, MNLI, αNLI

Source : [What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data (ChaosNLI)?]

Dataset : ChaosNLI

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** E E E N N	**Entailment** $E^{(50)}$ $N^{(50)}$	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** E E N N E	**Neutral** $N^{(93)}$ $E^{(7)}$	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** C N N C C	**Contradiction** $C^{(56)}$ $N^{(44)}$	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** N E N N E	**Entailment** $E^{(94)}$ $N^{(5)}$ $C^{(1)}$	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Re-annotation of SNLI, MNLI, αNLI

- 100 annotators per datapoint

Source : [What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data (ChaosNLI)?]

Dataset : ChaosNLI

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** **E E E N N**	**Entailment** **E**$^{(50)}$ **N**$^{(50)}$	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** **E E N N E**	**Neutral** **N**$^{(93)}$ **E**$^{(7)}$	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** **C N N C C**	**Contradiction** **C**$^{(56)}$ **N**$^{(44)}$	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** **N E N N E**	**Entailment** **E**$^{(94)}$ **N**$^{(5)}$ **C**$^{(1)}$	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Re-annotation of SNLI, MNLI, αNLI

- 100 annotators per datapoint
- Provides distribution over labels

Source : [What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data (ChaosNLI)?]

Dataset : ChaosNLI

Context	Hypothesis	Old Labels	New Labels	Source	Type
		majority and individual labels			
With the sun rising, a person is gliding with a huge parachute attached to them.	The person is falling to safety with the parachute	**Entailment** E E E N N	**Entailment** $E^{(50)}$ $N^{(50)}$	SNLI	Low agreements
A woman in a tan top and jeans is sitting on a bench wearing headphones.	A woman is listening to music.	**Entailment** E E N N E	**Neutral** $N^{(93)}$ $E^{(7)}$	SNLI	Majority changed
A group of guys went out for a drink after work, and sitting at the bar was a real a 6 foot blonde with a fabulous face and figure to match.	The men didn't appreciate the figure of the blonde woman sitting at the bar.	**Contradiction** C N N C C	**Contradiction** $C^{(56)}$ $N^{(44)}$	MNLI	Low agreements
In the other sight he saw Adrin's hands cocking back a pair of dragon-hammered pistols.	He had spotted Adrin preparing to fire his pistols.	**Neutral** N E N N E	**Entailment** $E^{(94)}$ $N^{(5)}$ $C^{(1)}$	MNLI	Majority changed

Table 3: Examples from ChaosNLI-S and ChaosNLI-M development set. 'Old Labels' is the 5 label annotations from original dataset. 'New Labels' refers to the newly collected 100 label annotations. Superscript indicates the frequency of the label.

Re-annotation of SNLI, MNLI, αNLI

- 100 annotators per datapoint
- Provides distribution over labels

We're considering SNLI

Source : [What Can We Learn from Collective HumAn OpinionS on Natural Language Inference Data (ChaosNLI)?]

Models

- Each time we consider a **pool** models of trained on NLI

Models

- Each time we consider a **pool** models of trained on NLI

<1B models

	N_{param}	θ_{i}		N_{param}	θ_{i}
BU_L-2_H-128_A-2	4385920	m_1	BU_L-2_H-512_A-8	22458880	m_{13}
BU_L-4_H-128_A-2	4782464	m_2	BU_L-4_H-512_A-8	28763648	m_{14}
BU_L-6_H-128_A-2	5179008	m_3	BU_L-6_H-512_A-8	35068416	m_{15}
BU_L-8_H-128_A-2	5575552	m_4	BU_L-2_H-768_A-12	38603520	m_{16}
BU_L-10_H-128_A-2	5972096	m_5	BU_L-8_H-512_A-8	41373184	m_{17}
BU_L-12_H-128_A-2	6368640	m_6	BU_L-10_H-512_A-8	47677952	m_{18}
BU_L-2_H-256_A-4	9591040	m_7	BU_L-4_H-768_A-12	52779264	m_{19}
BU_L-4_H-256_A-4	11170560	m_8	BU_L-12_H-512_A-8	53982720	m_{20}
BU_L-6_H-256_A-4	12750080	m_9	BU_L-6_H-768_A-12	66955008	m_{21}
BU_L-8_H-256_A-4	14329600	m_{10}	BU_L-8_H-768_A-12	81130752	m_{22}
BU_L-10_H-256_A-4	15909120	m_{11}	BU_L-10_H-768_A-12	95306496	m_{23}
BU_L-12_H-256_A-4	17488640	m_{12}	BU_L-12_H-768_A-12	109482240	m_{24}

Models

- Each time we consider a **pool** models of trained on NLI

<1B models

	N_{param}	θ_i		N_{param}	θ_i
BU_L-2_H-128_A-2	4385920	m_1	BU_L-2_H-512_A-8	22458880	m_{13}
BU_L-4_H-128_A-2	4782464	m_2	BU_L-4_H-512_A-8	28763648	m_{14}
BU_L-6_H-128_A-2	5179008	m_3	BU_L-6_H-512_A-8	35068416	m_{15}
BU_L-8_H-128_A-2	5575552	m_4	BU_L-2_H-768_A-12	38603520	m_{16}
BU_L-10_H-128_A-2	5972096	m_5	BU_L-8_H-512_A-8	41373184	m_{17}
BU_L-12_H-128_A-2	6368640	m_6	BU_L-10_H-512_A-8	47677952	m_{18}
BU_L-2_H-256_A-4	9591040	m_7	BU_L-4_H-768_A-12	52779264	m_{19}
BU_L-4_H-256_A-4	11170560	m_8	BU_L-12_H-512_A-8	53982720	m_{20}
BU_L-6_H-256_A-4	12750080	m_9	BU_L-6_H-768_A-12	66955008	m_{21}
BU_L-8_H-256_A-4	14329600	m_{10}	BU_L-8_H-768_A-12	81130752	m_{22}
BU_L-10_H-256_A-4	15909120	m_{11}	BU_L-10_H-768_A-12	95306496	m_{23}
BU_L-12_H-256_A-4	17488640	m_{12}	BU_L-12_H-768_A-12	109482240	m_{24}

1B models

Llama 3.2

OLMo

BLOOM

Pythia

Models

- Each time we consider a **pool** models of trained on NLI

<1B models

	N_{param}	θ_i		N_{param}	θ_i
BU_L-2_H-128_A-2	4385920	m_1	BU_L-2_H-512_A-8	22458880	m_{13}
BU_L-4_H-128_A-2	4782464	m_2	BU_L-4_H-512_A-8	28763648	m_{14}
BU_L-6_H-128_A-2	5179008	m_3	BU_L-6_H-512_A-8	35068416	m_{15}
BU_L-8_H-128_A-2	5575552	m_4	BU_L-2_H-768_A-12	38603520	m_{16}
BU_L-10_H-128_A-2	5972096	m_5	BU_L-8_H-512_A-8	41373184	m_{17}
BU_L-12_H-128_A-2	6368640	m_6	BU_L-10_H-512_A-8	47677952	m_{18}
BU_L-2_H-256_A-4	9591040	m_7	BU_L-4_H-768_A-12	52779264	m_{19}
BU_L-4_H-256_A-4	11170560	m_8	BU_L-12_H-512_A-8	53982720	m_{20}
BU_L-6_H-256_A-4	12750080	m_9	BU_L-6_H-768_A-12	66955008	m_{21}
BU_L-8_H-256_A-4	14329600	m_{10}	BU_L-8_H-768_A-12	81130752	m_{22}
BU_L-10_H-256_A-4	15909120	m_{11}	BU_L-10_H-768_A-12	95306496	m_{23}
BU_L-12_H-256_A-4	17488640	m_{12}	BU_L-12_H-768_A-12	109482240	m_{24}

1B models

Llama 3.2

OLMo

BLOOM

Pythia

Guarantees that we have pool of heterogeneous models: we can average out model quirks

Models

- Each time we consider a **pool** models of trained on NLI

<1B models

	N_{param}	θ_i		N_{param}	θ_i
BU_L-2_H-128_A-2	4385920	m_1	BU_L-2_H-512_A-8	22458880	m_{13}
BU_L-4_H-128_A-2	4782464	m_2	BU_L-4_H-512_A-8	28763648	m_{14}
BU_L-6_H-128_A-2	5179008	m_3	BU_L-6_H-512_A-8	35068416	m_{15}
BU_L-8_H-128_A-2	5575552	m_4	BU_L-2_H-768_A-12	38603520	m_{16}
BU_L-10_H-128_A-2	5972096	m_5	BU_L-8_H-512_A-8	41373184	m_{17}
BU_L-12_H-128_A-2	6368640	m_6	BU_L-10_H-512_A-8	47677952	m_{18}
BU_L-2_H-256_A-4	9591040	m_7	BU_L-4_H-768_A-12	52779264	m_{19}
BU_L-4_H-256_A-4	11170560	m_8	BU_L-12_H-512_A-8	53982720	m_{20}
BU_L-6_H-256_A-4	12750080	m_9	BU_L-6_H-768_A-12	66955008	m_{21}
BU_L-8_H-256_A-4	14329600	m_{10}	BU_L-8_H-768_A-12	81130752	m_{22}
BU_L-10_H-256_A-4	15909120	m_{11}	BU_L-10_H-768_A-12	95306496	m_{23}
BU_L-12_H-256_A-4	17488640	m_{12}	BU_L-12_H-768_A-12	109482240	m_{24}

1B models

Llama 3.2

OLMo

BLOOM

Pythia

Guarantees that we have pool of heterogeneous models: we can average out model quirks

(let's keep it simple, and **we will talk about 1B models**)

Metrics

- Contrasting values of metrics for human label variation / data difficulty / data uncertainty.

Human label variation	Data complexity	Data uncertainty
HUMAN-BASED	**MODEL-BASED**	
	With reference	*Without reference*
Entropy [Nie et al., 2020]	Early-exit [Baldock et al., 2021]	CP set size [Vovk et al., 2005]
Dissensus	Confidence [Swayamdipta et al., 2020]	Model entropy
	Early acquisition	Dissensus across model decisions
	Avg. acc. across models	Entropy across model decisions
	Avg. acc. across training	

Metrics

- Contrasting values of metrics for human label variation / data difficulty / data uncertainty.

Human label variation	Data complexity	Data uncertainty
HUMAN-BASED	**MODEL-BASED**	
	With reference	*Without reference*
Entropy [Nie et al., 2020]	**Early-exit** [Baldock et al., 2021]	**CP set size** [Vovk et al., 2005]
Dissensus	Confidence [Swayamdipta et al., 2020]	Model entropy
	Early acquisition	Dissensus across model decisions
	Avg. acc. across models	Entropy across model decisions
	Avg. acc. across training	

💡 **Entropy** [Nie et al., 2020]

Diversity in the label distribution, better accounting for both dominant and minority labels.

$$H_{ent} = -\Sigma_{yi \in Y} Pr_H(y_i |x) \log Pr_H(y_i |x$$

NB: Human-based

💡 **Entropy** [Nie et al., 2020]

Diversity in the label distribution, better accounting for both dominant and minority labels.

$$H_{ent} = -\Sigma_{y_i \in Y} Pr_H(y_i|x) \log Pr_H(y_i|x$$

example:

$\overbrace{\boxed{p}\ \boxed{n}\ \boxed{p}\ \boxed{p}\ \boxed{n}\ \boxed{n}\ \boxed{n}\ \boxed{n}\ \boxed{n}\ \boxed{p}}^{\text{10 Annotators}}$ {p: 4, n:6} | $H_{ent} = 0.9709$

NB: Human-based

Diversity in the label distribution, better accounting for both dominant and minority labels.

$$H_{ent} = -\Sigma_{yi \in Y} Pr_H(y_i\,|x) \log Pr_H(y_i\,|x$$

example:

$\overbrace{\text{p n p p n n n n n p}}^{\text{10 Annotators}}$ {p: 4, n:6} | $H_{ent} = 0.9709$

p n n p n n n n n p {p: 5, n:5} | $H_{ent} = 1.0000$

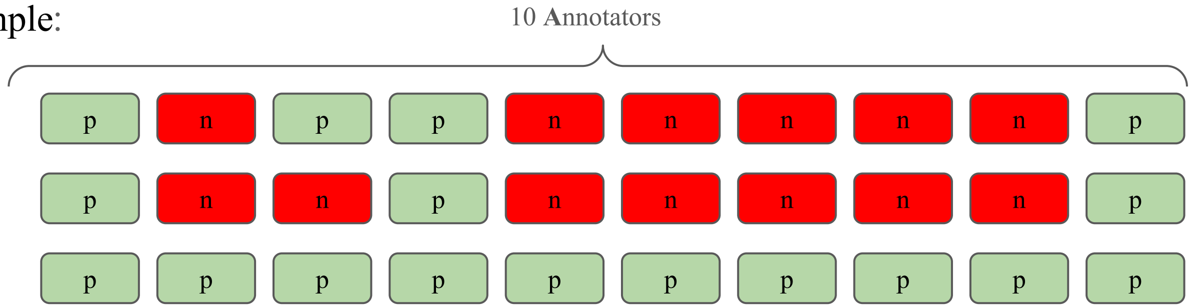
NB: Human-based

💡 **Entropy** [Nie et al., 2020]

Diversity in the label distribution, better accounting for both dominant and minority labels.

$$H_{ent} = -\Sigma_{yi\in Y} Pr_H(y_i\,|x) \log Pr_H(y_i\,|x$$

example:



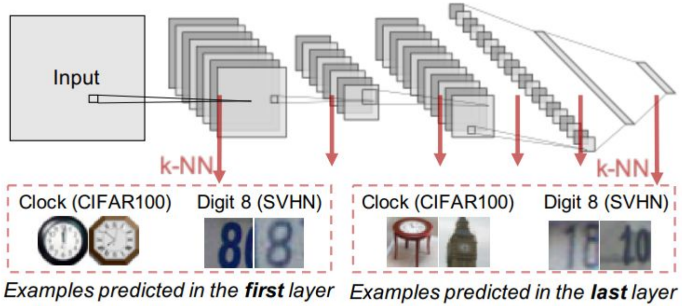
$\{p: 4, n:6\}$ | $H_{ent} = 0.9709$

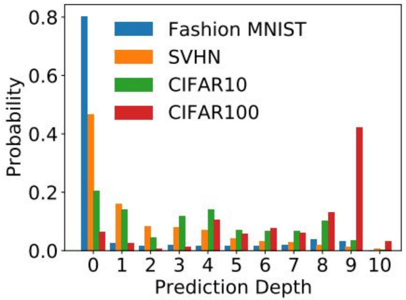
$\{p: 5, n:5\}$ | $H_{ent} = 1.0000$

$\{p: 10, n:0\}$ | $H_{ent} = 0$

NB: Human-based

💡 **Early-exit** [Baldock et al., 2021]

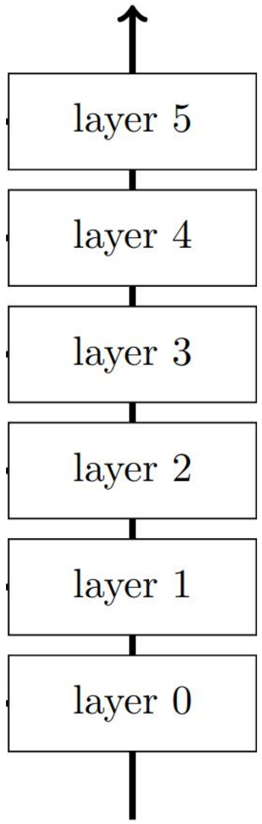




> *"Deep models use fewer layers to (effectively) determine the prediction for easy examples and more layers for hard examples"*

💡 **Early-exit** [Baldock et al., 2021]

Given a deep learning model with layers of the same shape



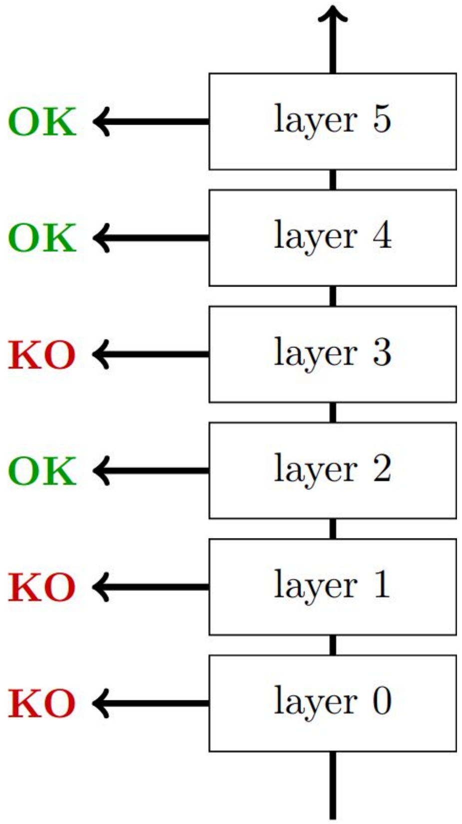
NB: Model-based, uses a reference

💡 **Early-exit** [Baldock et al., 2021]

Given a deep learning model with layers of the same shape

- get the predictions at every layer

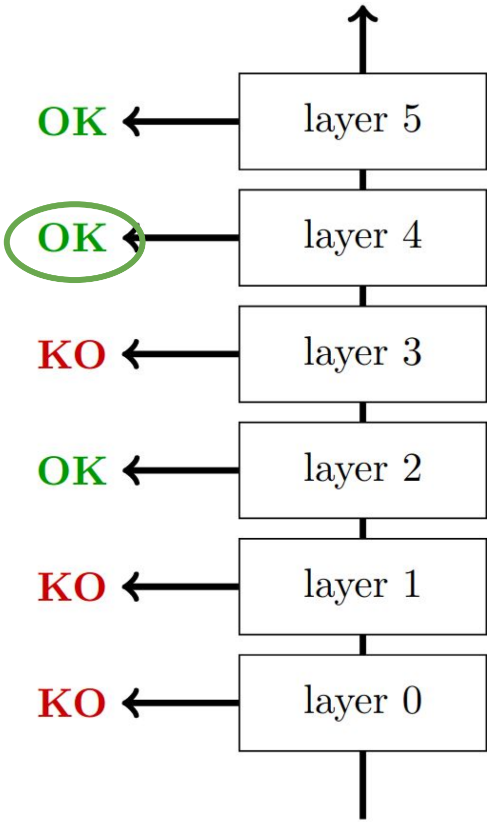
NB: Model-based, uses a reference



Early-exit [Baldock et al., 2021]

Given a deep learning model with layers of the same shape

- get the predictions at every layer
- select the layer where you start getting consistently correct predictions



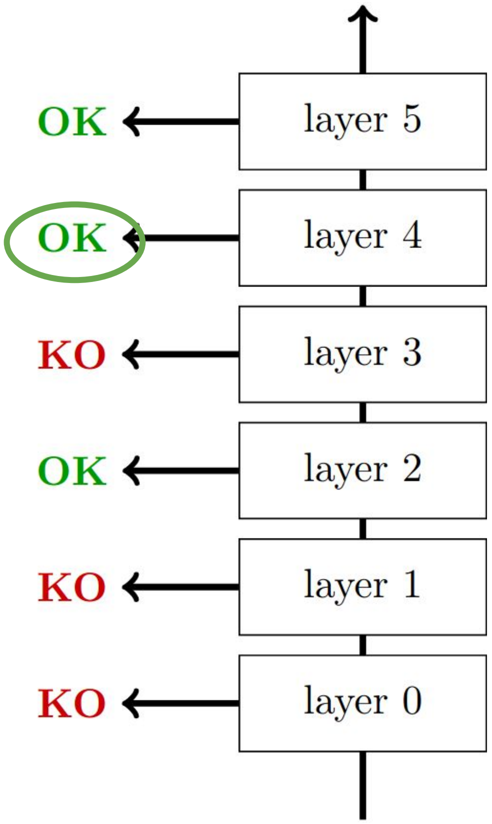
NB: Model-based, uses a reference

💡 **Early-exit** [Baldock et al., 2021]

Given a deep learning model with layers of the same shape

- get the predictions at every layer
- select the layer where you start getting consistently correct predictions

Examples for which this layer is lower are *easier*



NB: Model-based, uses a reference

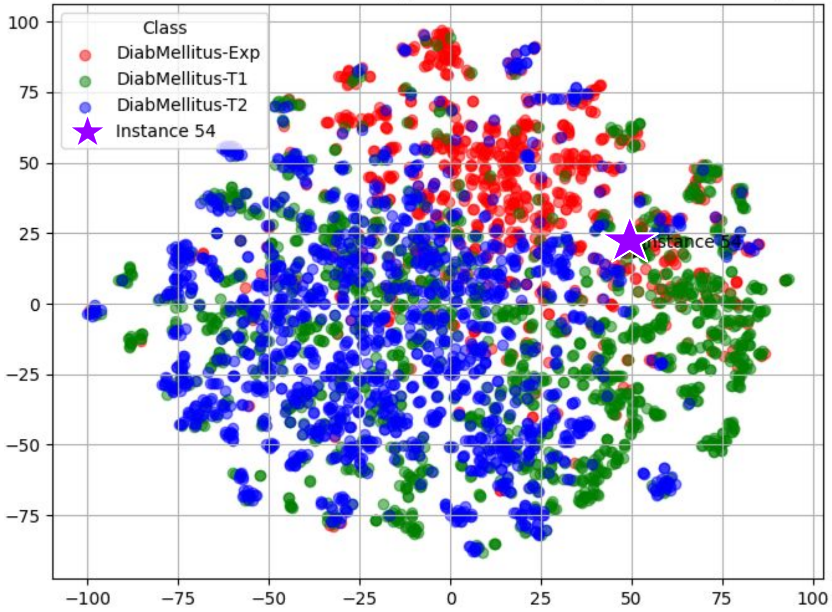
💡 **CP set size** [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

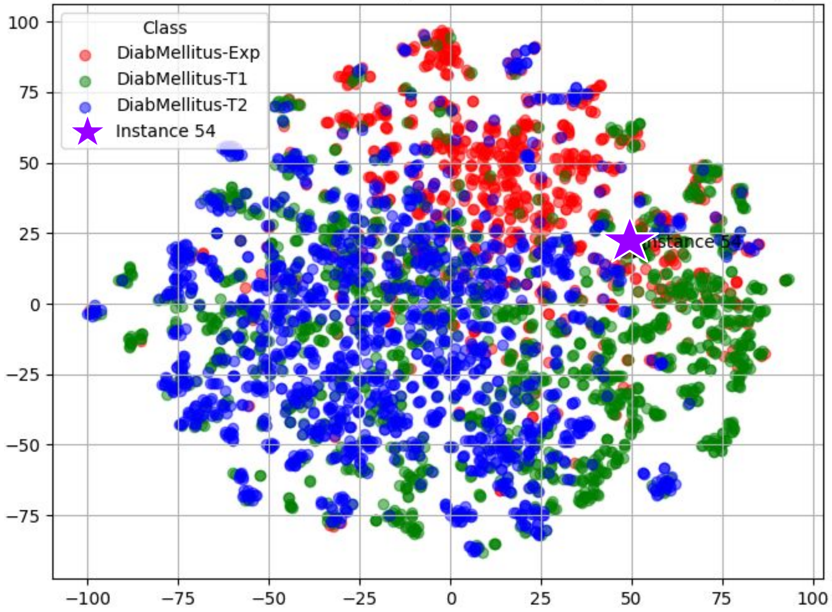
example:



💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:

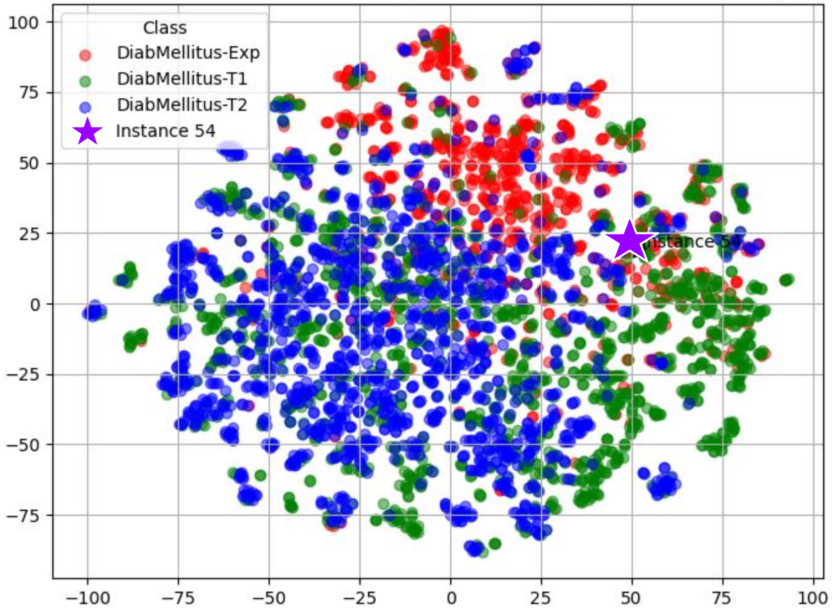


Is that a Diab-Exp? Is that a Diab-T1?
Is my classifier confident enough to decide on a single answer?
How ok am I with my classifier being wrong?

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:



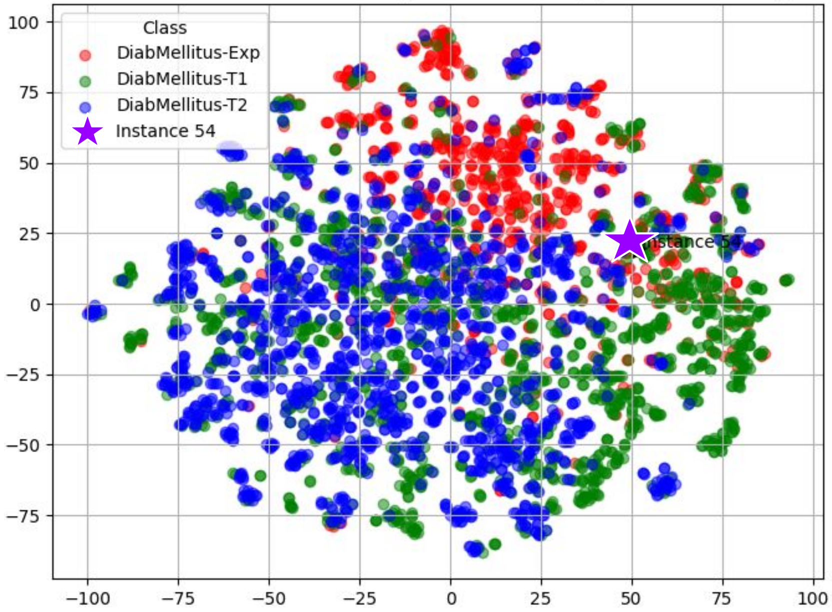
In practice:

$$\hat{P}(Y|\bigstar) \rightarrow \underset{0.45}{\mathrm{Diab}_{\mathrm{Exp}}},\ \underset{0.35}{\mathrm{Diab}_{\mathrm{T1}}},\ \underset{0.20}{\mathrm{Diab}_{\mathrm{T2}}}$$

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:



In practice:

$$\hat{P}(Y|\star) \rightarrow \mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}, \mathrm{Diab}_{\mathrm{T2}}$$

0.45 0.35 0.20

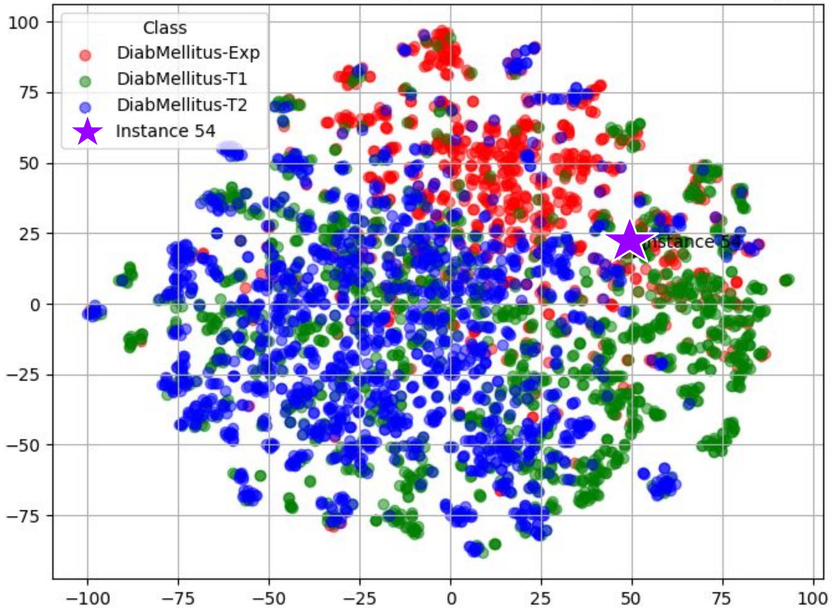
$$\downarrow$$

$$U(\{\mathrm{Diab}_{\mathrm{Exp}}\}, \hat{P}, u_{f1}) = 0.5$$

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:



In practice:

$$\hat{P}(Y|\bigstar) \rightarrow \mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}, \mathrm{Diab}_{\mathrm{T2}}$$

0.45 0.35 0.20

$$\downarrow$$

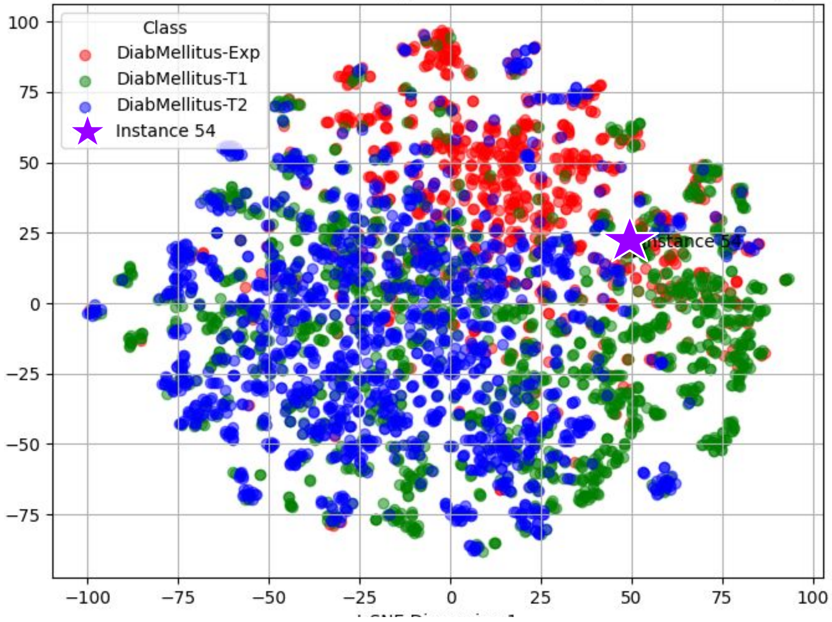
$$U(\{\mathrm{Diab}_{\mathrm{Exp}}\}, \hat{P}, u_{f1}) = 0.5$$

$$U(\{\mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}\}, \hat{P}, u_{f1}) = 0.6$$

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:



In practice:

$$\hat{P}(Y|\bigstar) \to \underset{0.45}{\mathrm{Diab}_{\mathrm{Exp}}}, \underset{0.35}{\mathrm{Diab}_{\mathrm{T1}}}, \underset{0.20}{\mathrm{Diab}_{\mathrm{T2}}}$$

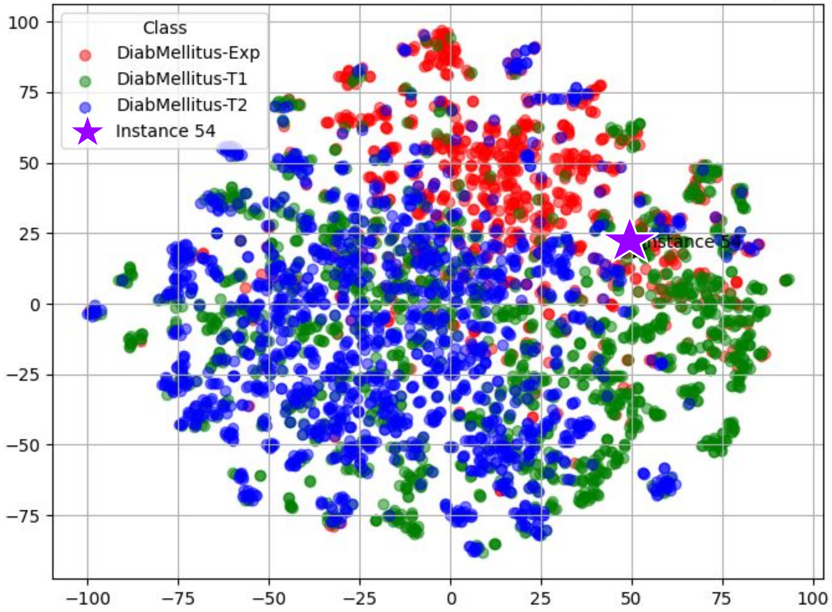
$$\downarrow$$

$$U(\{\mathrm{Diab}_{\mathrm{Exp}}\}, \hat{P}, u_{f1}) = 0.5$$
$$U(\{\mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}\}, \hat{P}, u_{f1}) = 0.6$$
$$U(\{\mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}, \mathrm{Diab}_{\mathrm{T2}}\}, \hat{P}, u_{f1}) = 0.465$$

💡 CP set size [Vovk et al, 2005]

Conformal prediction considers the set of labels one must select to statistically guarantee that they contain the right answer (with a given risk)

example:



In practice:

$$\hat{P}(Y|\bigstar) \rightarrow \mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}, \mathrm{Diab}_{\mathrm{T2}}$$

0.45 0.35 0.20

$$\downarrow$$

$$U(\{\mathrm{Diab}_{\mathrm{Exp}}\}, \hat{P}, u_{f1}) = 0.5$$

$$\boxed{U(\{\mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}\}, \hat{P}, u_{f1}) = 0.6}$$

$$\boxed{U(\{\mathrm{Diab}_{\mathrm{Exp}}, \mathrm{Diab}_{\mathrm{T1}}, \mathrm{Diab}_{\mathrm{T2}}\}, \hat{P}, u_{f1}) = 0.465}$$

In our case:

1. the number of labels we need to select is an indicator of uncertainty

NB: Model-based, no reference needed

Results

Human-based vs. model-based indicators

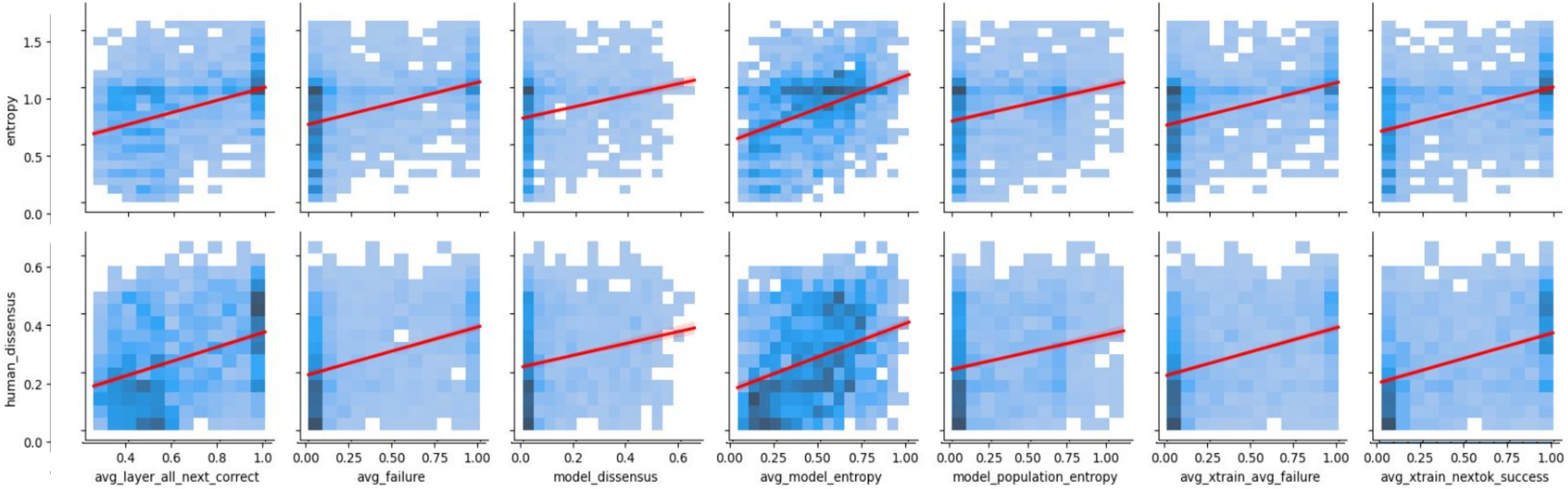
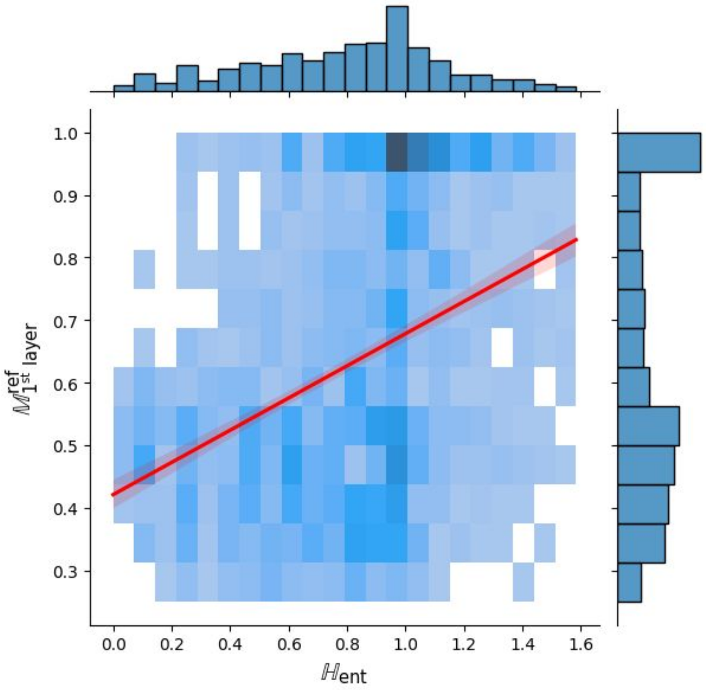
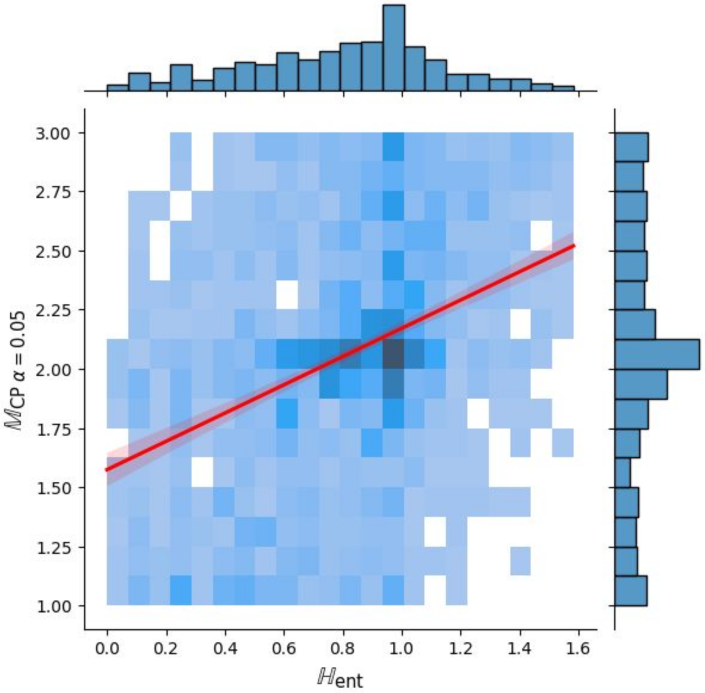


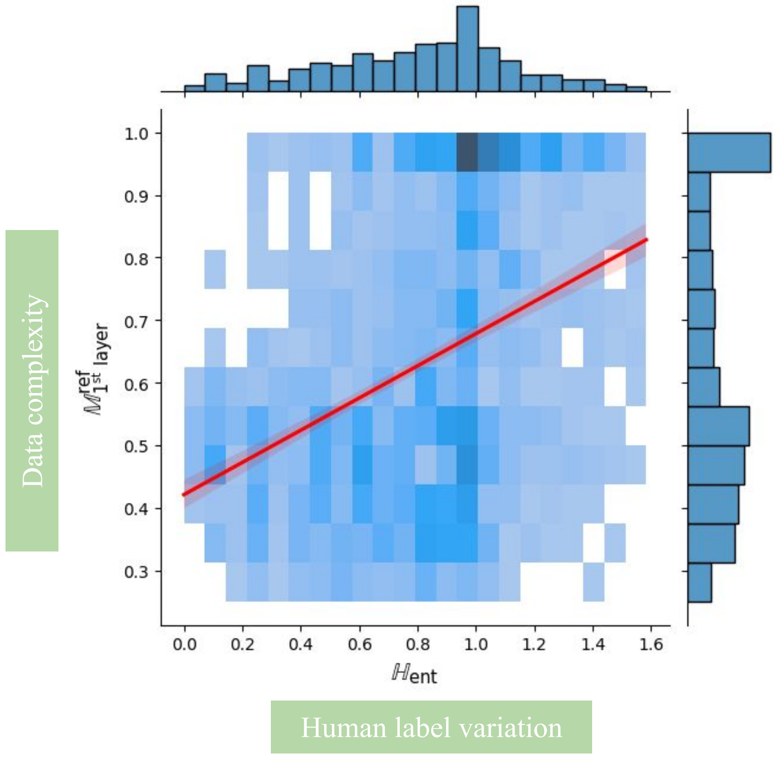
Figure : Interaction between human-based and model-based indicators

Human-based vs. model-based indicators

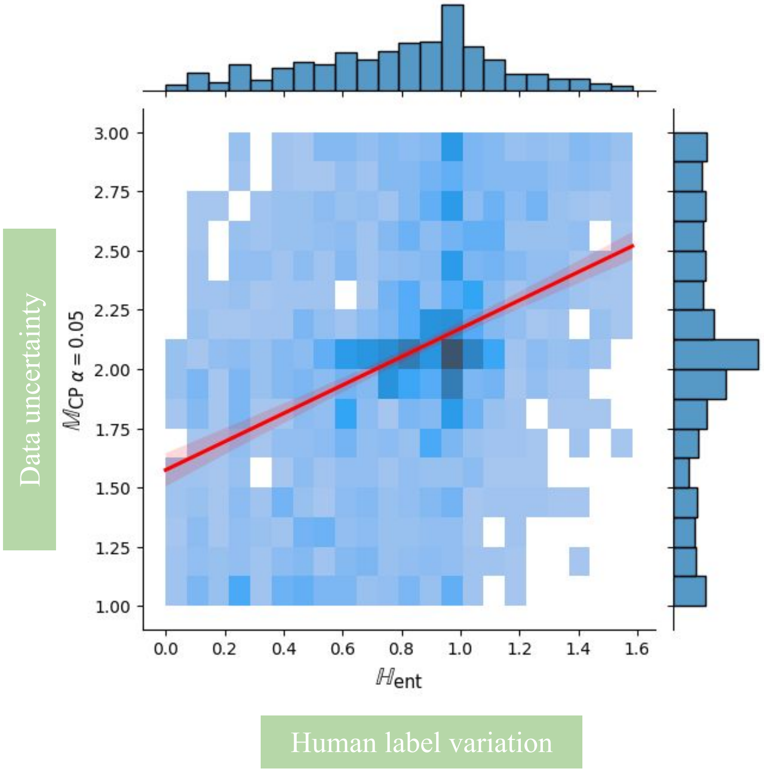




Human-based vs. model-based indicators



Residual (R2 Variance) = 0.1313



Residual (R2 Variance) = 0.0766

there is actually some order to this chaos

		Human label variation	
		$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
Data uncertainty	$\mathbb{M}_{\text{dis}}$	0.1947	0.1772
	$\mathbb{M}_{\text{ent}}$	0.2183	0.1970
	$\mathbb{M}_{\text{avg ent}}$	0.2811	0.2398
	$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.2767	0.2315
	$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.2819	0.2393
	$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.2482	0.2157
Data complexity	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.3497	0.3330
	$\mathbb{M}^{\text{ref}}_{1^{\text{st}} \text{ layer}}$	0.3624	0.3387
	$\mathbb{M}^{\text{ref}}_{1^{\text{st}} \text{ ckpt}}$	0.3682	0.3443
	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.3477	0.3274
	$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.3670	0.3428

there is actually some order to this chaos

- In practice, we observe **(low)** correlations throughout

		Human label variation	
		$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
Data uncertainty	$\mathbb{M}_{\text{dis}}$	0.1947	0.1772
	$\mathbb{M}_{\text{ent}}$	0.2183	0.1970
	$\mathbb{M}_{\text{avg ent}}$	0.2811	0.2398
	$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.2767	0.2315
	$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.2819	0.2393
	$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.2482	0.2157
Data complexity	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.3497	0.3330
	$\mathbb{M}^{\text{ref}}_{\text{1}^{\text{st}}\text{ layer}}$	0.3624	0.3387
	$\mathbb{M}^{\text{ref}}_{\text{1}^{\text{st}}\text{ ckpt}}$	0.3682	0.3443
	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.3477	0.3274
	$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.3670	0.3428

there is actually some order to this chaos

- In practice, we observe **(low)** correlations throughout
- whether an indicator **uses a reference** drives the correlation up

		Human label variation	
		$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
Data uncertainty	$\mathbb{M}_{\text{dis}}$	0.1947	0.1772
	$\mathbb{M}_{\text{ent}}$	0.2183	0.1970
	$\mathbb{M}_{\text{avg ent}}$	0.2811	0.2398
	$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.2767	0.2315
	$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.2819	0.2393
	$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.2482	0.2157
Data complexity	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.3497	0.3330
	$\mathbb{M}^{\text{ref}}_{\text{1st layer}}$	0.3624	0.3387
	$\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$	0.3682	0.3443
	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.3477	0.3274
	$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.3670	0.3428

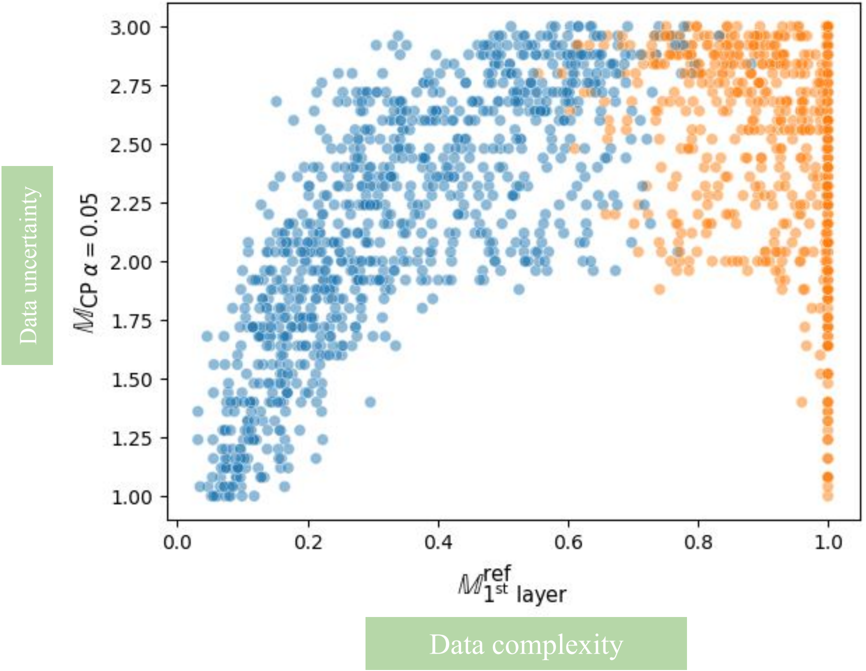
there is actually some order to this chaos

- In practice, we observe (low) correlations throughout
- whether an indicator **uses a reference** drives the correlation up
- correlations are much higher when comparing two metrics within a group of indicators

		Human label variation	
		$\mathbb{H}_{\text{ent}}$	$\mathbb{H}_{\text{dis}}$
Data uncertainty	$\mathbb{M}_{\text{dis}}$	0.1947	0.1772
	$\mathbb{M}_{\text{ent}}$	0.2183	0.1970
	$\mathbb{M}_{\text{avg ent}}$	0.2811	0.2398
	$\mathbb{M}_{\text{CP }\alpha=0.05}$	0.2767	0.2315
	$\mathbb{M}_{\text{CP }\alpha=0.1}$	0.2819	0.2393
	$\mathbb{M}_{\text{CP }\alpha=0.2}$	0.2482	0.2157
Data complexity	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.3497	0.3330
	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ layer}}$	0.3624	0.3387
	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ ckpt}}$	0.3682	0.3443
	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.3477	0.3274
	$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt }p}$	0.3670	0.3428

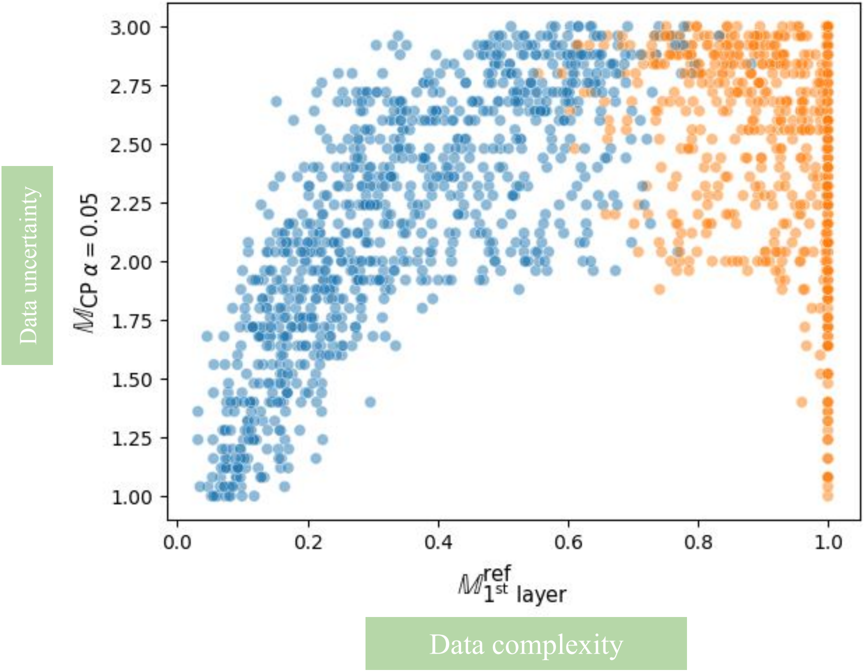
Reference-free vs. reference-dependent indicators

- U-shaped curve



Reference-free vs. reference-dependent indicators

- U-shaped curve
- orange: models tend to fail; blue: they tend to succeed



More generally

	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{fail}}$	$\mathrm{M}^{\mathrm{ref}}_{1^{\mathrm{st}}\ \mathrm{layer}}$	$\mathrm{M}^{\mathrm{ref}}_{1^{\mathrm{st}}\ \mathrm{ckpt}}$	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{avg\ ckpt}}$	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{avg\ ckpt}\ p}$
$\mathrm{M}_{\mathrm{dis}}$	−0.7761	−0.7593	−0.7737	−0.7136	−0.6838
$\mathrm{M}_{\mathrm{ent}}$	−0.7131	−0.7075	−0.7174	−0.6539	−0.6140
$\mathrm{M}_{\mathrm{avg\ ent}}$	−0.5615	−0.5264	−0.5283	−0.5303	−0.5111
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.05}$	−0.3670	−0.3633	−0.3515	−0.3389	−0.3037
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.1}$	−0.4761	−0.4565	−0.4575	−0.4453	−0.4182
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.2}$	−0.6427	−0.5836	−0.5967	−0.6156	−0.6116

when models tend to fail

	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{fail}}$	$\mathrm{M}^{\mathrm{ref}}_{1^{\mathrm{st}}\ \mathrm{layer}}$	$\mathrm{M}^{\mathrm{ref}}_{1^{\mathrm{st}}\ \mathrm{ckpt}}$	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{avg\ ckpt}}$	$\mathrm{M}^{\mathrm{ref}}_{\mathrm{avg\ ckpt}\ p}$
$\mathrm{M}_{\mathrm{dis}}$	0.9536	0.8928	0.9159	0.9003	0.8934
$\mathrm{M}_{\mathrm{ent}}$	0.9464	0.8891	0.9107	0.8980	0.8962
$\mathrm{M}_{\mathrm{avg\ ent}}$	0.8803	0.8955	0.9116	0.8694	0.9315
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.05}$	0.7748	0.7939	0.7971	0.7759	0.8546
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.1}$	0.8546	0.8601	0.8816	0.8491	0.9103
$\mathrm{M}_{\mathrm{CP}\ \alpha=0.2}$	0.8996	0.8982	0.9320	0.8799	0.9166

when models don’t

More generally

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ layer}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	-0.7761	-0.7593	-0.7737	-0.7136	-0.6838
$\mathbb{M}_{\text{ent}}$	-0.7131	-0.7075	-0.7174	-0.6539	-0.6140
$\mathbb{M}_{\text{avg ent}}$	-0.5615	-0.5264	-0.5283	-0.5303	-0.5111
$\mathbb{M}_{\text{CP } \alpha=0.05}$	-0.3670	-0.3633	-0.3515	-0.3389	-0.3037
$\mathbb{M}_{\text{CP } \alpha=0.1}$	-0.4761	-0.4565	-0.4575	-0.4453	-0.4182
$\mathbb{M}_{\text{CP } \alpha=0.2}$	-0.6427	-0.5836	-0.5967	-0.6156	-0.6116

when models tend to fail

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ layer}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\text{ ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$
$\mathbb{M}_{\text{dis}}$	0.9536	0.8928	0.9159	0.9003	0.8934
$\mathbb{M}_{\text{ent}}$	0.9464	0.8891	0.9107	0.8980	0.8962
$\mathbb{M}_{\text{avg ent}}$	0.8803	0.8955	0.9116	0.8694	0.9315
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.7748	0.7939	0.7971	0.7759	0.8546
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.8546	0.8601	0.8816	0.8491	0.9103
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.8996	0.8982	0.9320	0.8799	0.9166

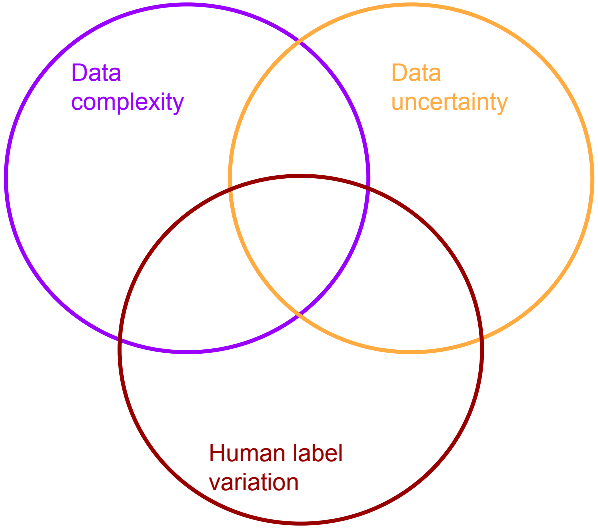
when models don’t

Reference-free model-based indicators systematically conflate model successes and model failures

Takeaways

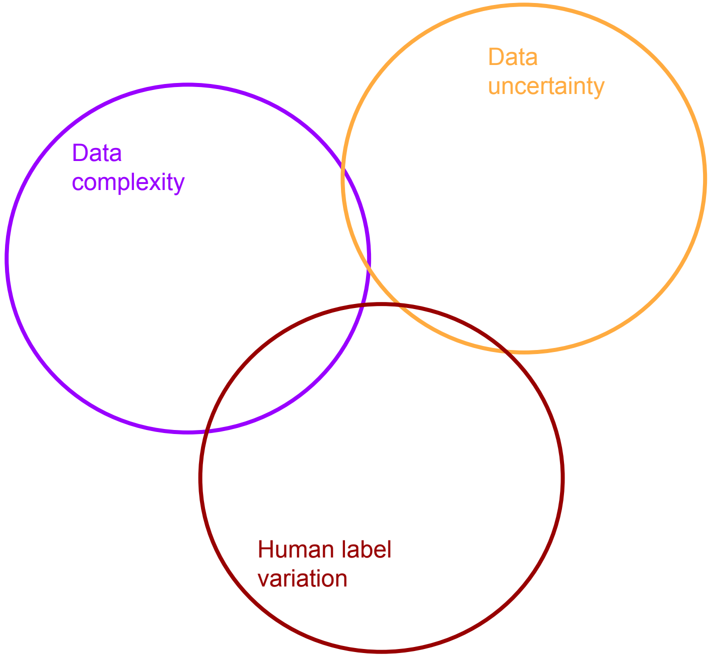
So what?

- Model-based indicators **align poorly** with human-based indicators!



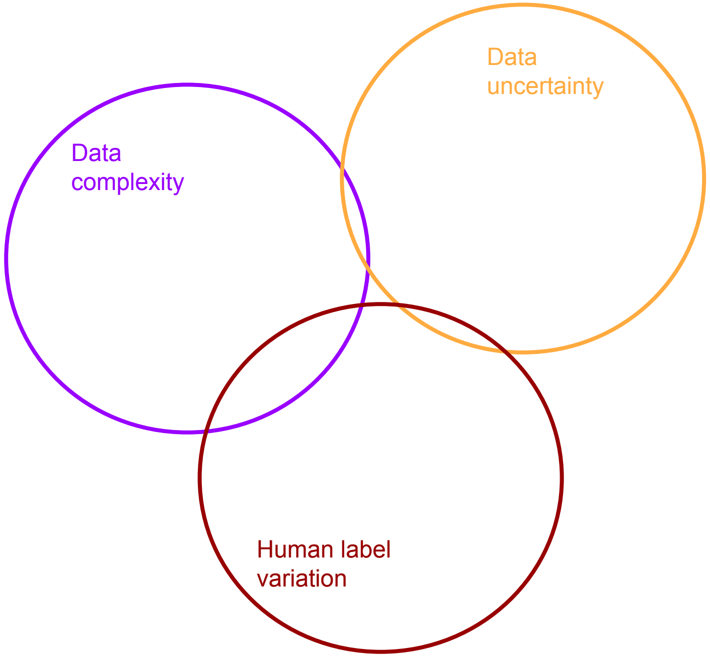
So what?

- Model-based indicators **align poorly** with human-based indicators!
- If ***what was easy/difficult for humans was also easy/difficult for models***, why does **data complexity** more aligned with **human disagreement** than **data uncertainty**.



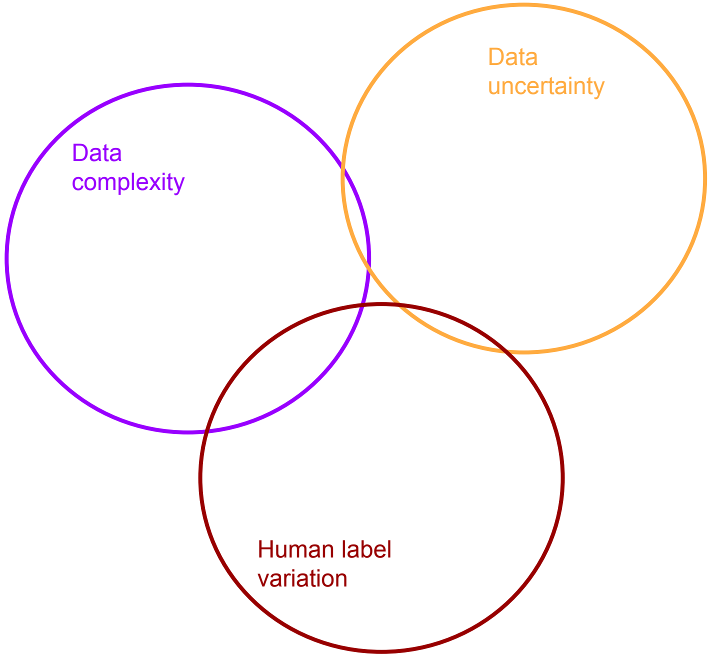
So what?

- Model-based indicators **align poorly** with human-based indicators!
- If ***what was easy/difficult for humans was also easy/difficult for models***, why does **data complexity** more aligned with **human disagreement** than **data uncertainty**.
- We need to **rethink practices such as active learning** (where difficulty of samples is taken in account during training)



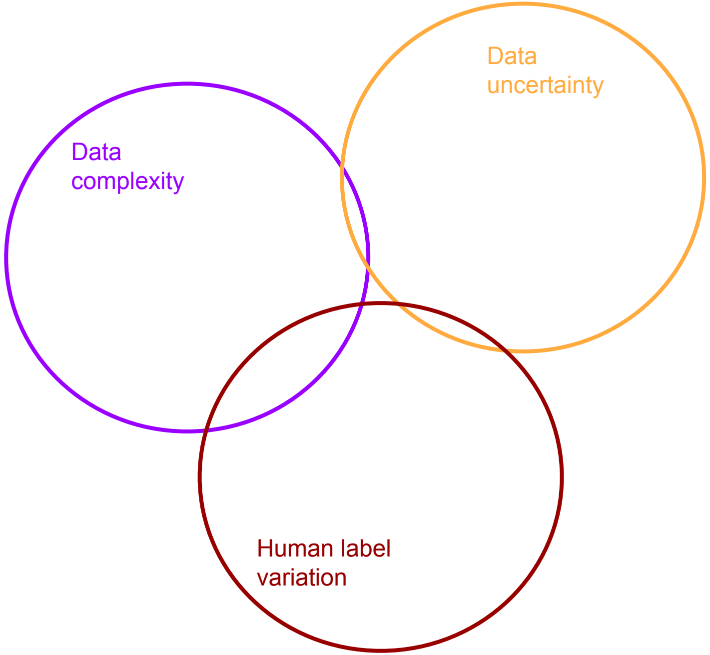
So what?

- Model-based indicators **align poorly** with human-based indicators!
- If ***what was easy/difficult for humans was also easy/difficult for models***, why does **data complexity** more aligned with **human disagreement** than **data uncertainty**.
- We need to **rethink practices such as active learning** (where difficulty of samples is taken in account during training)
- not all framework are created equal (out of the box **CP has flaws**)



So what?

- Model-based indicators **align poorly** with human-based indicators!
- If ***what was easy/difficult for humans was also easy/difficult for models***, why does **data complexity** more aligned with **human disagreement** than **data uncertainty**.
- We need to **rethink practices such as active learning** (where difficulty of samples is taken in account during training)
- not all framework are created equal (out of the box **CP has flaws**)
- Linguistic ambiguity remains distinct for model-based indicators



Appendix - I

- **Low residual** : Non-linear relationship between human and model based indicators.

	$\mathbb{H}_{\text{dis}}$	$\mathbb{H}_{\text{ent}}$
$\mathbb{M}_{\text{dis}}$	0.0314	0.0379
$\mathbb{M}_{\text{avg ent}}$	0.0575	0.0790
$\mathbb{M}_{\text{ent}}$	0.0388	0.0477
$\mathbb{M}_{\text{CP } \alpha=0.05}$	0.0536	0.0766
$\mathbb{M}_{\text{CP } \alpha=0.1}$	0.0573	0.0795
$\mathbb{M}_{\text{CP } \alpha=0.2}$	0.0465	0.0616
$\mathbb{M}^{\text{ref}}_{\text{fail}}$	0.1109	0.1223
$\mathbb{M}^{\text{ref}}_{1^{\text{st}} \text{ layer}}$	0.1147	0.1313
$\mathbb{M}^{\text{ref}}_{1^{\text{st}} \text{ ckpt}}$	0.1186	0.1356
$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	0.1072	0.1209
$\mathbb{M}^{+\text{ref}}_{\text{avg ckpt } p}$	0.1175	0.1347

Table : Proportion of explained variance (R2) of linear regressions predicting a model-based indicator from a human-based indicator.

Appendix - II

	$\mathbb{M}^{\text{ref}}_{\text{fail}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{layer}}$	$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$	$\mathbb{M}^{\text{ref}}_{\text{avg ckpt}\ p}$
$\mathbb{M}_{\text{dis}}$	0.5154	0.4966	0.5029	0.5035	0.5048
$\mathbb{M}_{\text{ent}}$	0.5168	0.4984	0.5047	0.5073	0.5127
$\mathbb{M}_{\text{avg ent}}$	0.5292	0.5419	0.5468	0.5292	0.5560
$\mathbb{M}_{\text{CP}\ \alpha=0.05}$	0.4860	0.4958	0.4972	0.4916	0.5286
$\mathbb{M}_{\text{CP}\ \alpha=0.1}$	0.5216	0.5290	0.5353	0.5241	0.5527
$\mathbb{M}_{\text{CP}\ \alpha=0.2}$	0.5232	0.5338	0.5437	0.5188	0.5338

Table : Spearman correlation between reference dependent and reference-free indicators

Thank you for your attention!

Any questions?

Interested in my work ?

Want to read our full paper ?

I am looking for postdoc opportunities in NLP, healthcare, LLM Evaluations