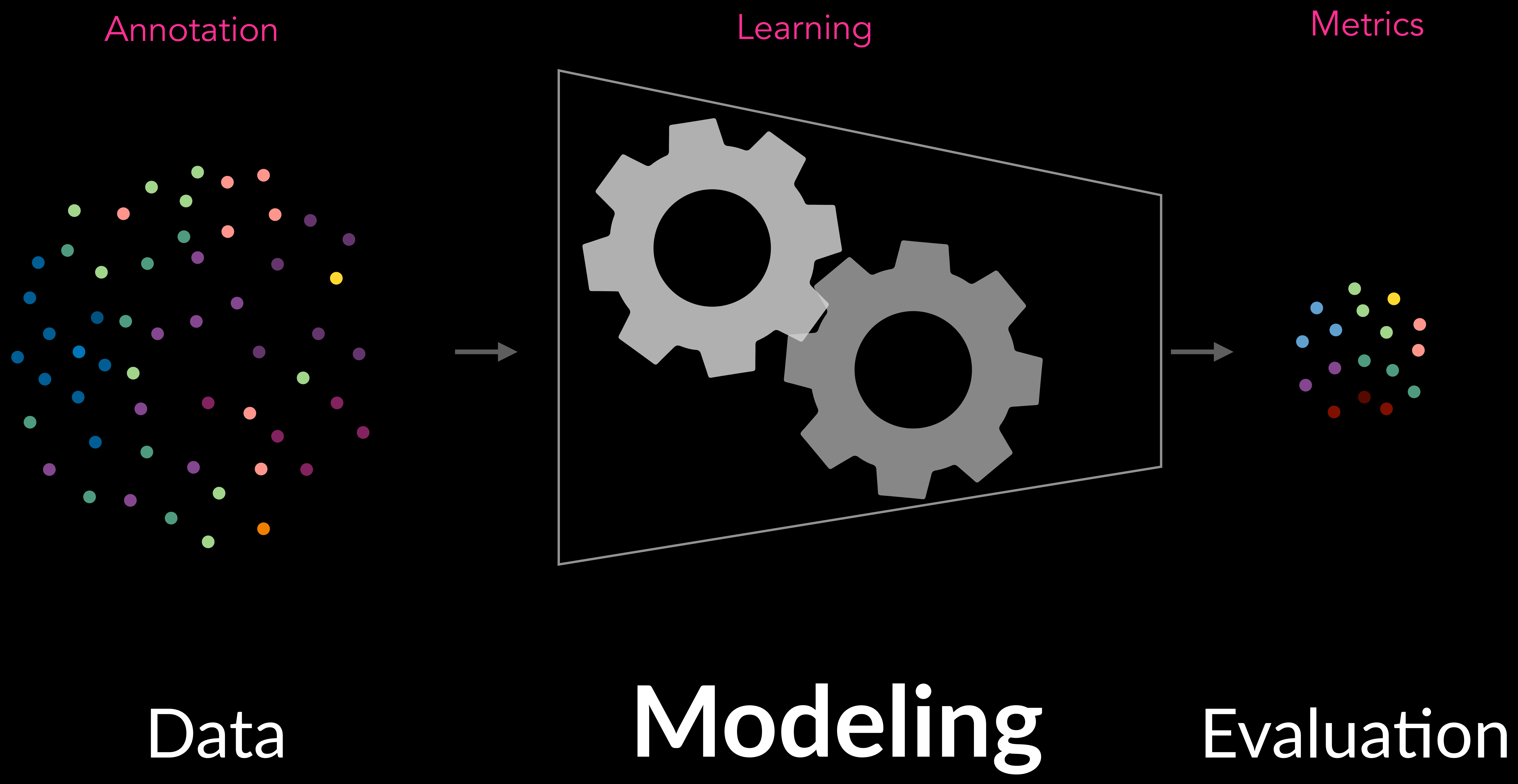
From Human Label Variation and Model Uncertainty to Error Detection (and Back)?

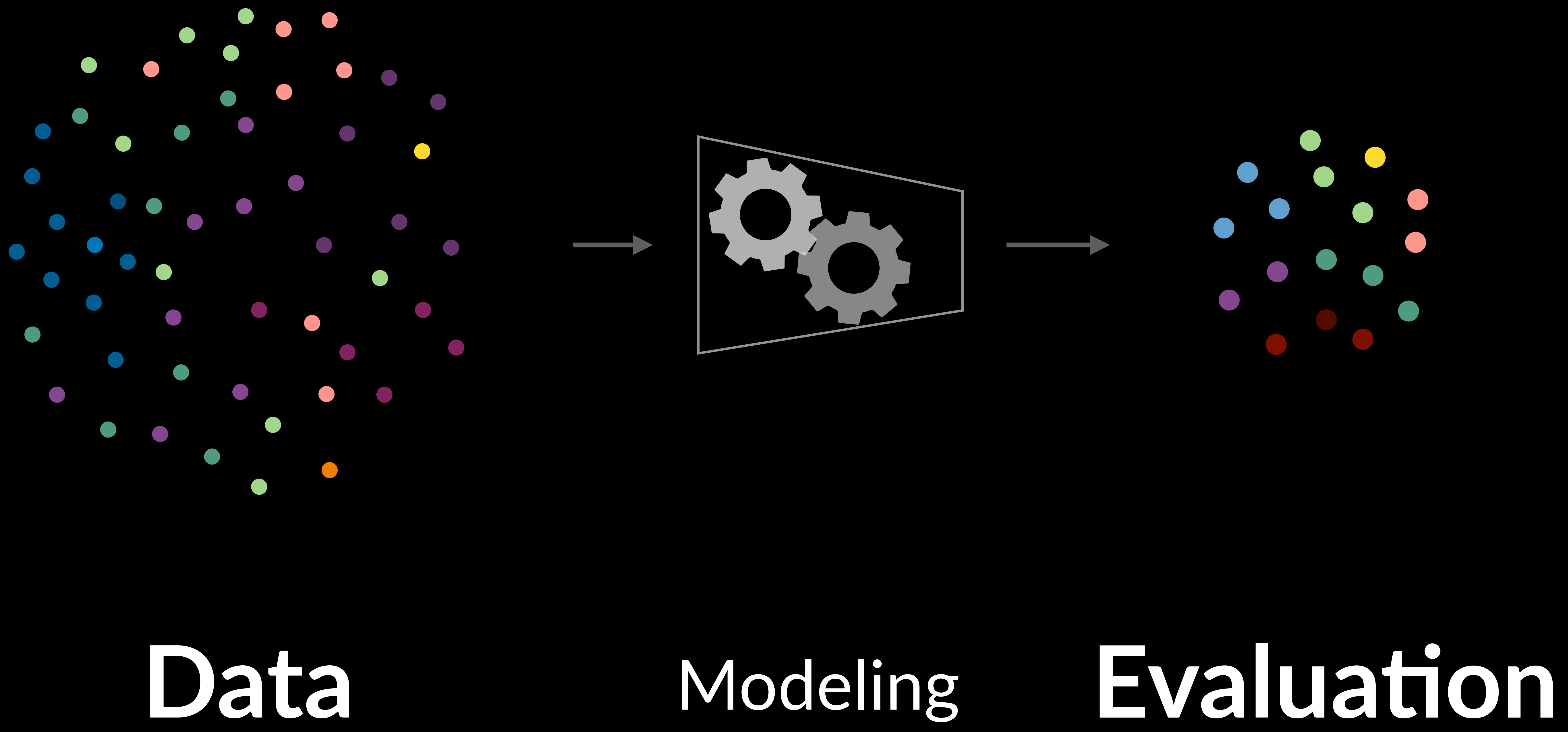
Prof. Dr. Barbara Plank
MaiNLP lab, CIS, LMU Munich
& IT University of Copenhagen

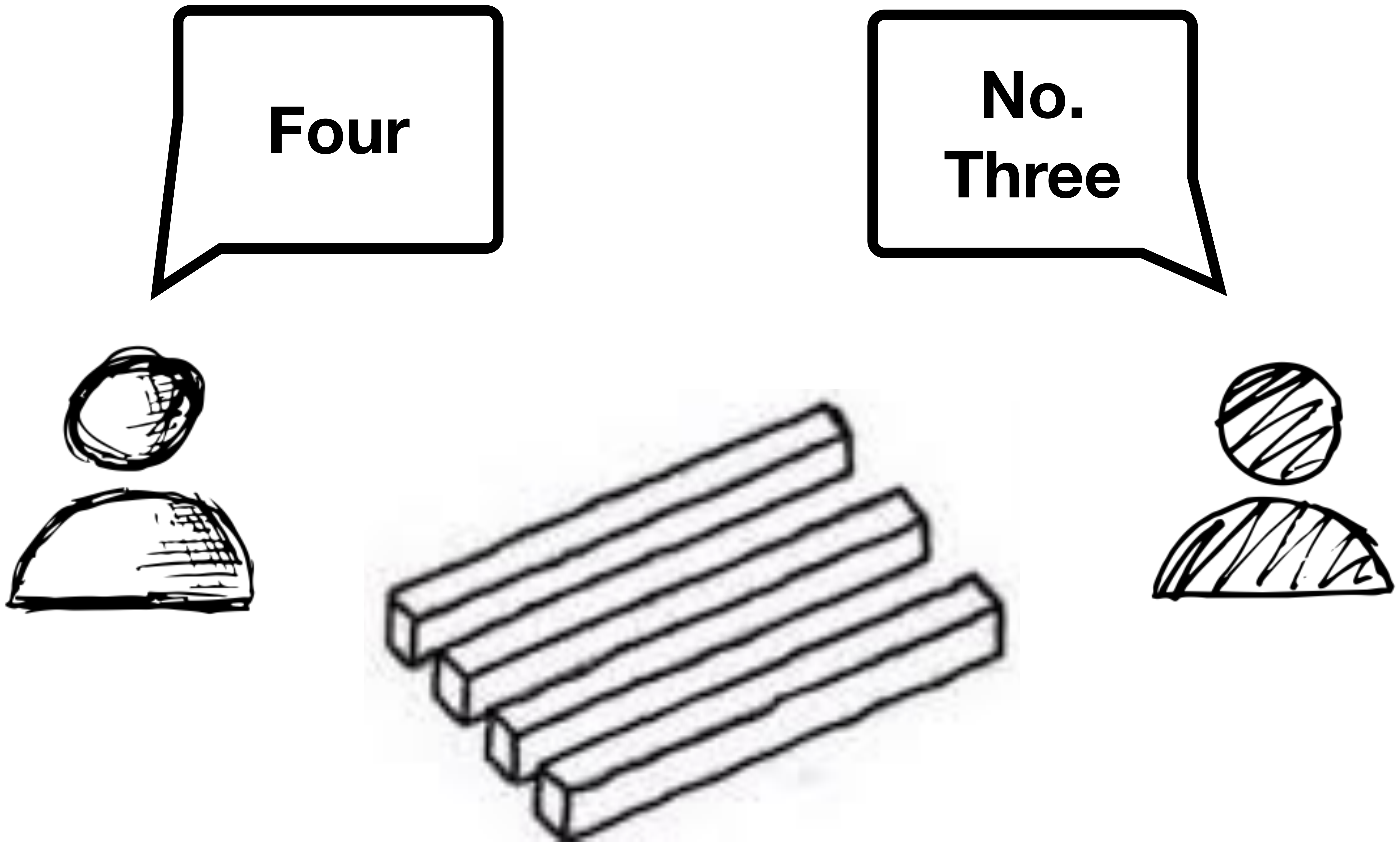
NLPerspectives workshop LREC-COLING 2024
May 21, 2024

A typical/traditional NLP/AI pipeline



Growing Importance of High-Quality Data and Evaluation




Four
No.
Three


The world is beyond digital - if we zoom out it is rich and diverse.

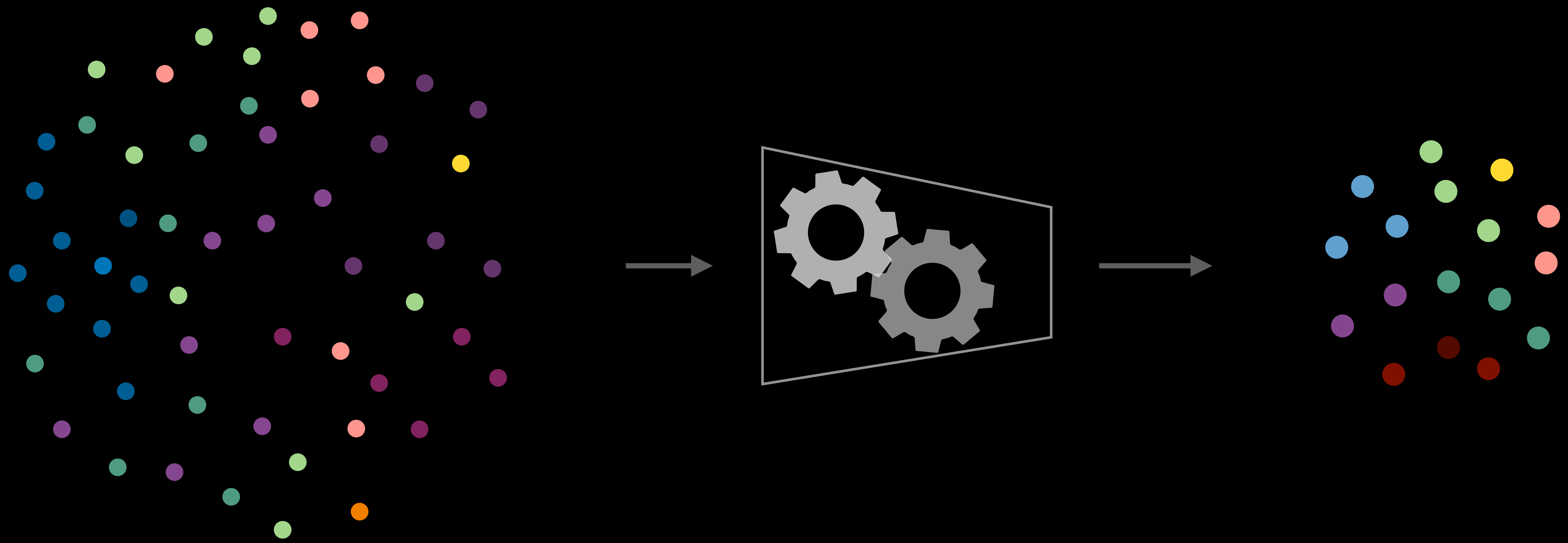
Src: Esher

Disagreement in human labeling is ubiquitous

— It impacts **all 3 stages** of the NLP pipeline

— It is one important form of **uncertainty**

Can we turn disagreement into advantage?



Data Modeling Evaluation

Growing Importance of
Data Quality > Data Quantity

The “it” in AI models is the dataset - talk by Thom Wolf 🤗

The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

I’ve been at OpenAI for almost a year now. In that time, I’ve trained a **lot** of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/

Evidence from a talk by Sara Hooker

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

- Recent work suggests smaller amounts of higher quality data remove the need for a larger model.
- This suggest larger models may just be compensating for problems in the data pipeline.

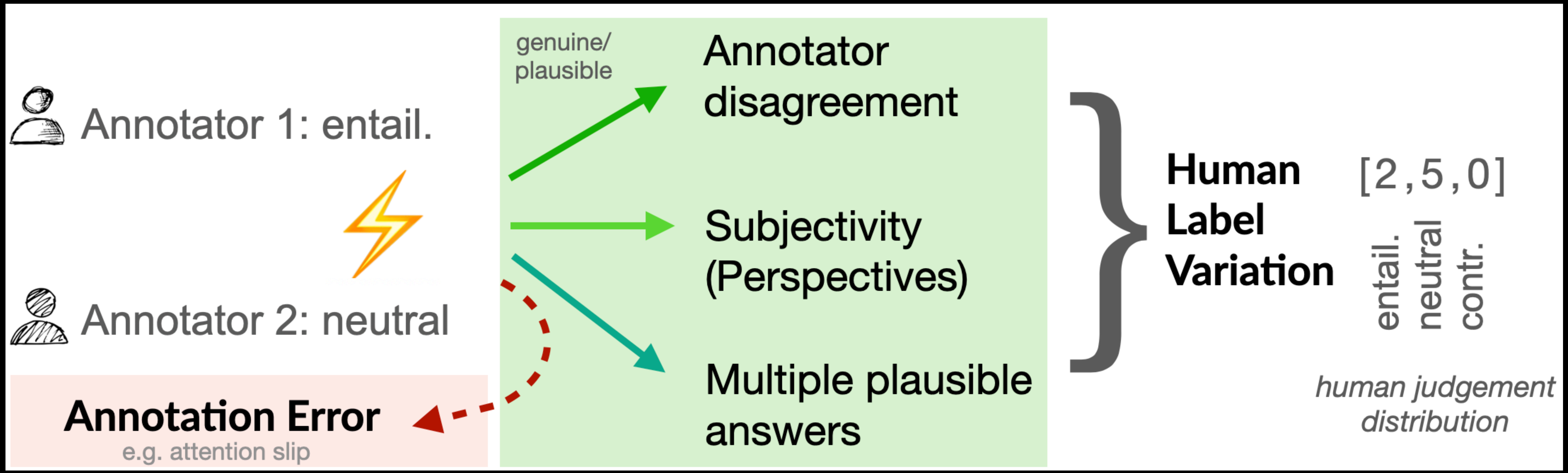
Hoffman et al. 2022 blog

Roadmap: Selected Case Studies

1. Humans and Uncertainty: What is *Human Label Variation?*
2. Models and Uncertainty: *Stop Measuring Calibration When Humans Disagree*
3. How to detect errors? *ActiveAED*
4. Plausible variation or error? *VariERR*

Disagreement or Variation?

- **Human Label Variation** (Plank, 2022 EMNLP)
  - plausible variation
  - to **reconcile different notions** in the literature (disagreement, perspectives, human uncertainty, hard cases)
  - preferred over **disagreement** as that implies two views cannot hold at the same time

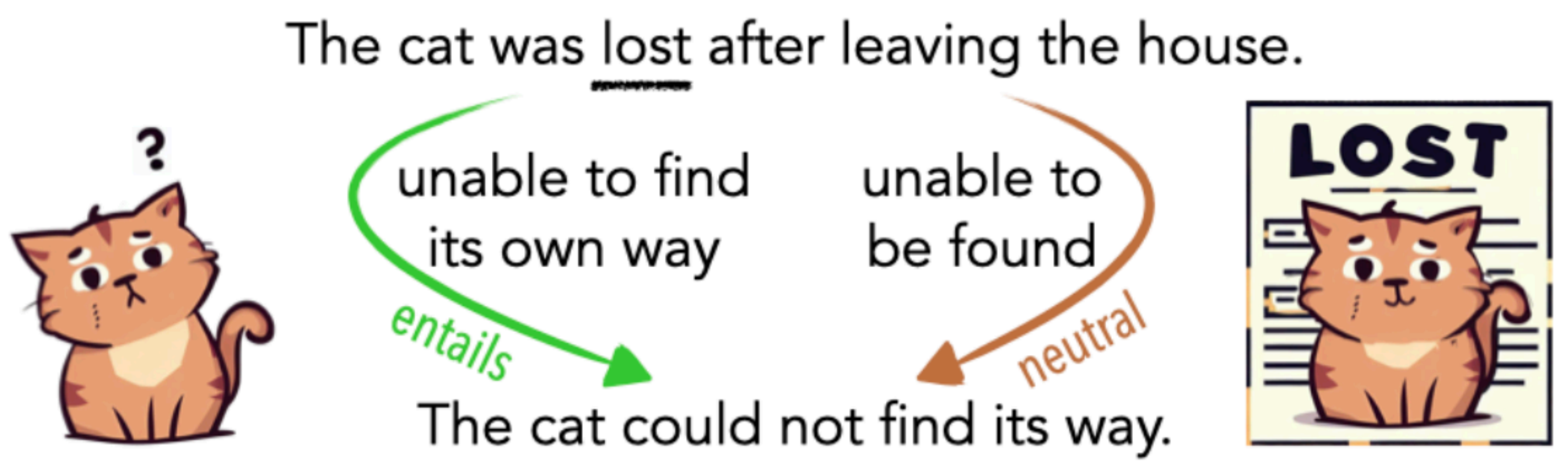


- In contrast to **errors**

Sources of human label variation

(Basile et al., 2021)

- **Stimulus characteristics** (ambiguity, task setup and difficulty)
- **Individual differences** (incl. cultural and socio-demographics): for example in hate speech or sentiment
- **Context and attention** (Intra-coder disagreement; attention slips play a non-negligible role as well; Beigman Klebanov et al., 2008)



Ambiguity (Example from Liu et al., 2023)

Examples

Lora Aroyo's NeurIPS 2023 keynote:

Is there a **SMILE** in this image?

YES **but** ...

Canada		
YES	NO	DNK
40%	40%	20%

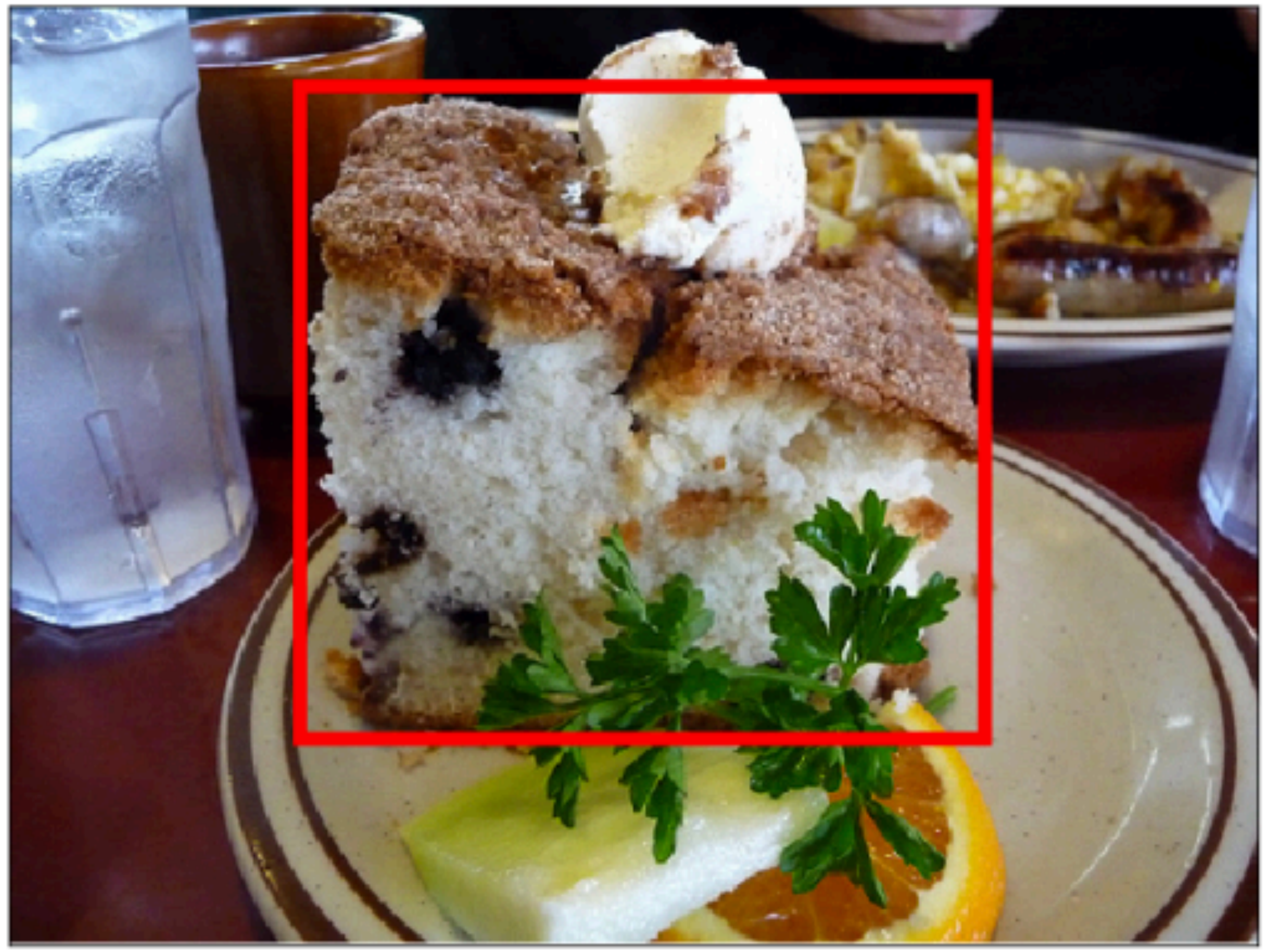
India		
YES	NO	DNK
70%	30%	0

USA		
YES	NO	DNK
50%	0	50%

adversarial example from the CATS4ML data challenge
https://github.com/google-research-datasets/cats4ml-dataset

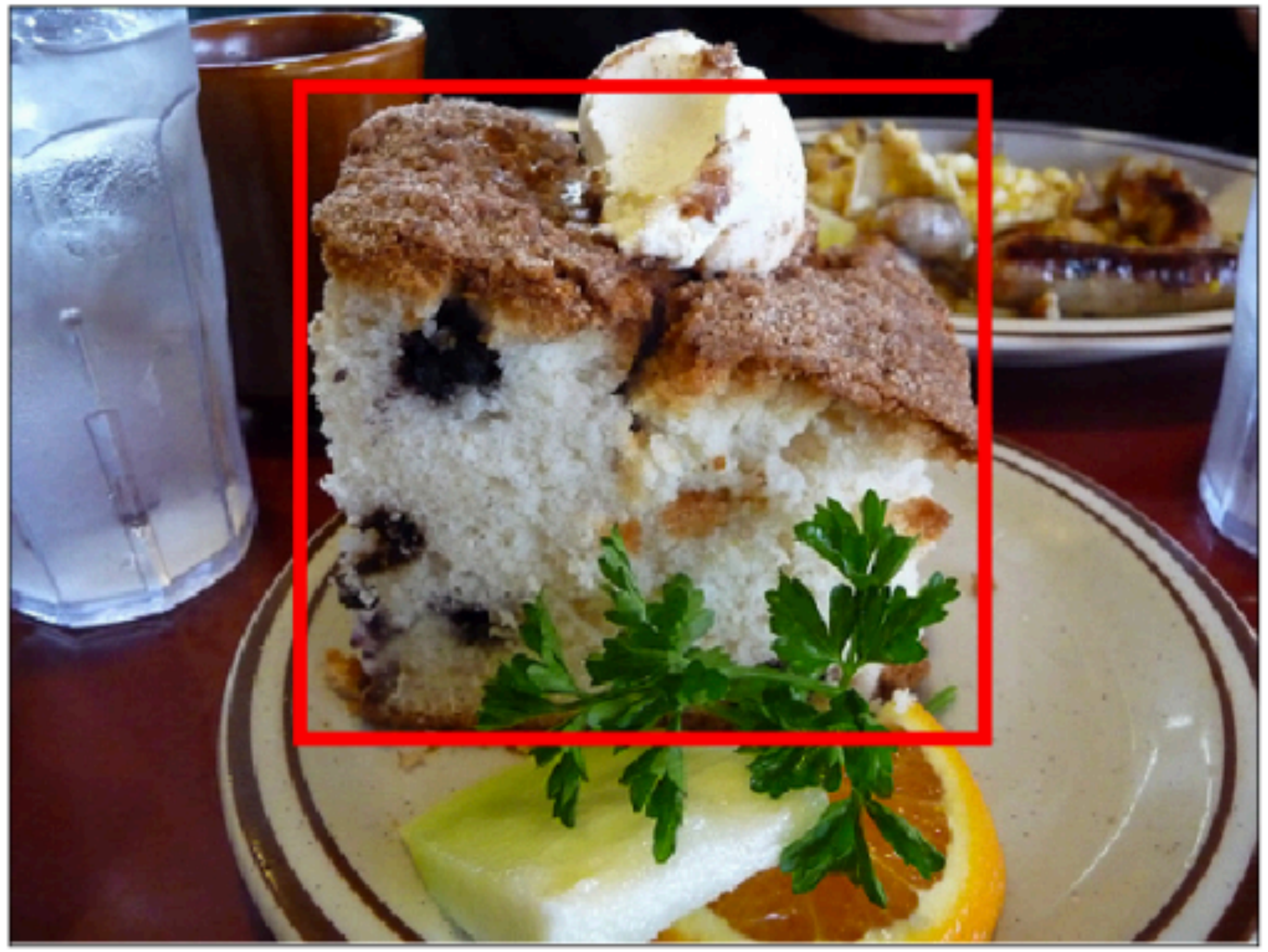
https://slideslive.com/39015341/the-many-faces-of-responsible-ai?ref=speaker-55217

Name the object



https://aclanthology.org/2020.lrec-1.710.pdf

Name the object



cake (53), food (19), bread (8), burger (6), dessert (6), snacks (3), muffin (3), pastry (3)

ManyNames dataset https://aclanthology.org/2020.lrec-1.710.pdf

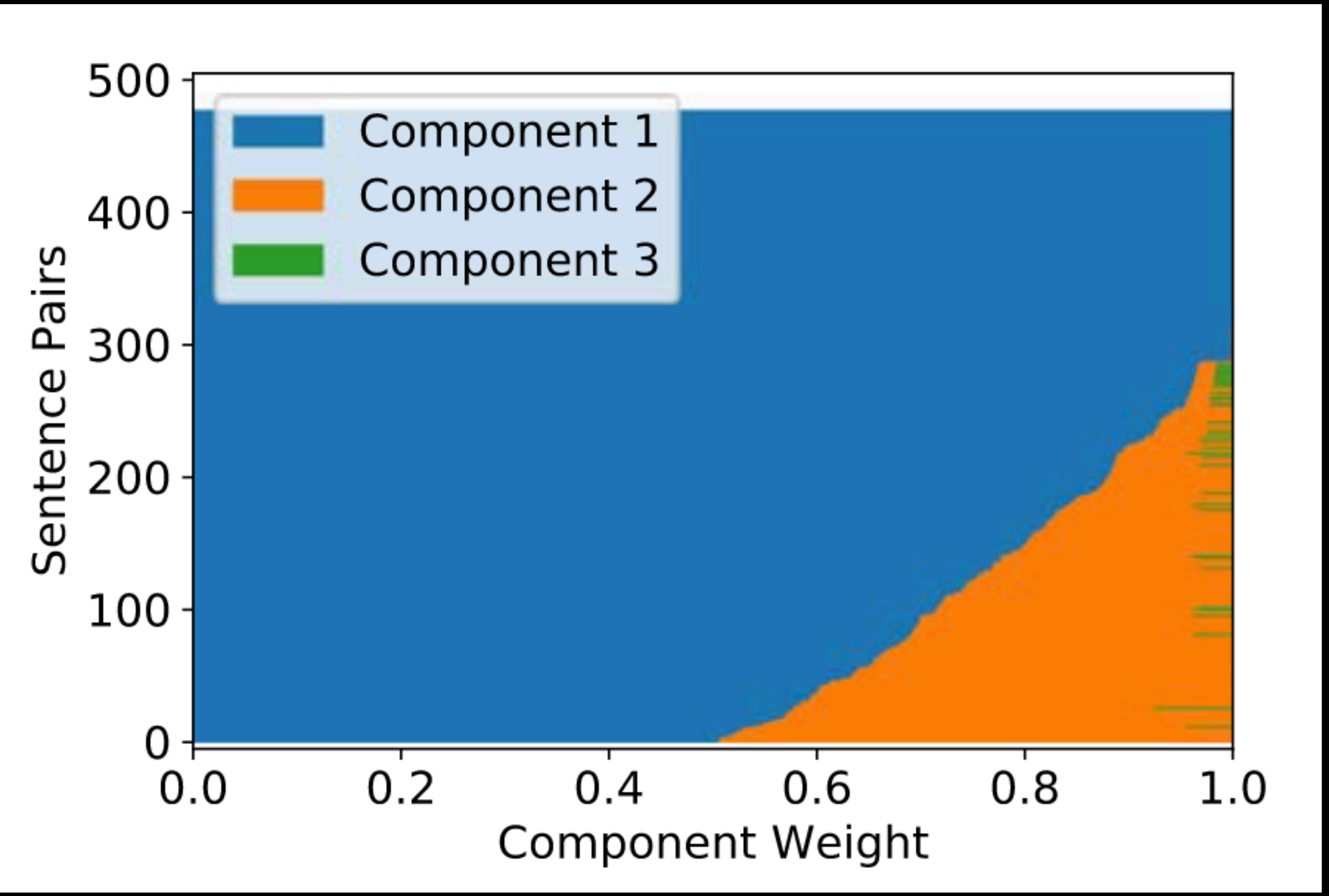
Natural Language Inference: Entailment? Neutral? Contradiction?

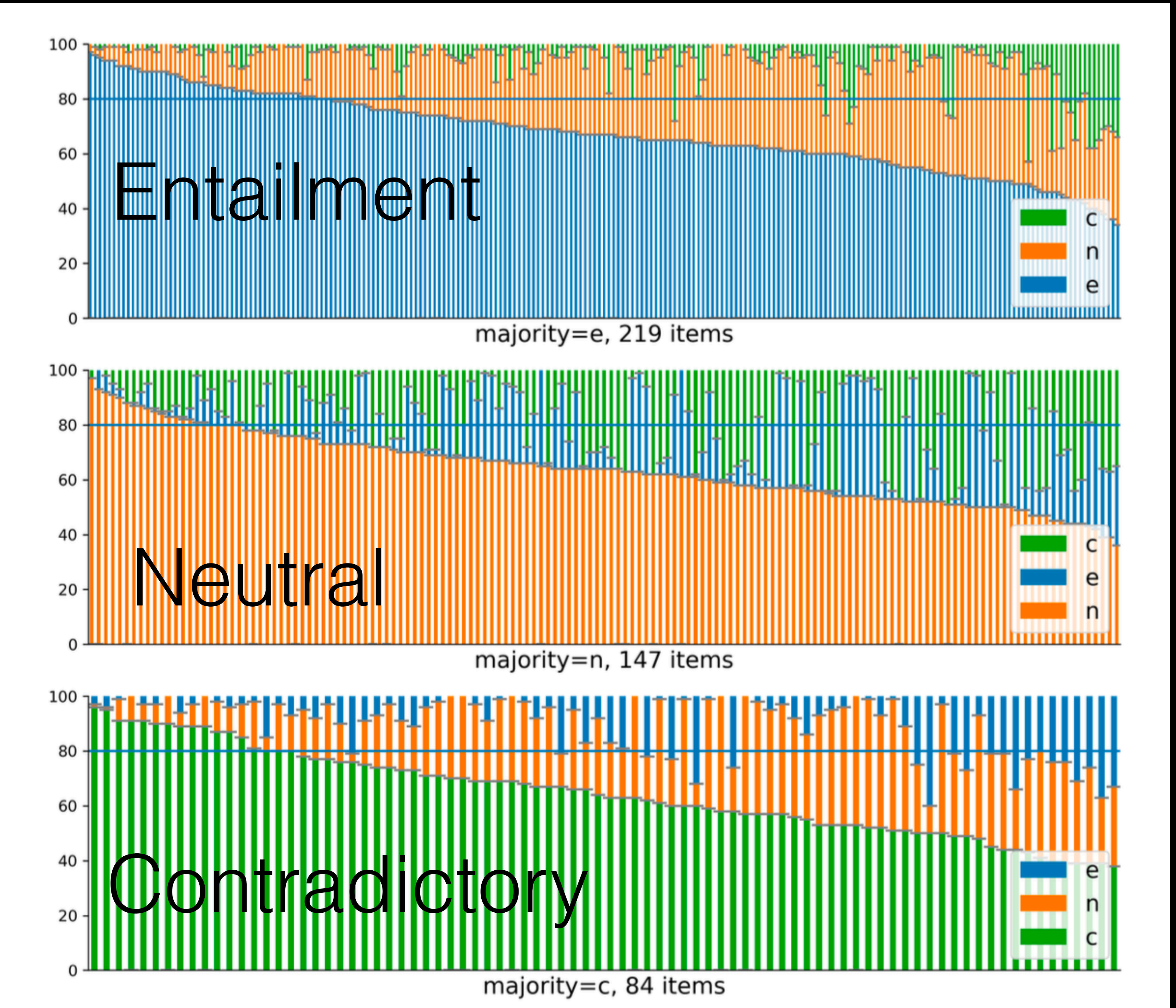
Context/Premise:	*Statement/Hypothesis:*	*[E, N, C]*
A boy in an orange shirt sells fruit from a street cart.	A boy is a street vendor.	[90, 10, 0]
A women wearing a red hat and black coat.	The women is asleep.	[0, 87, 13]
People walk amonst a traffic jam in a crowded city.	The cars are zooming past the people.	[3, 15, 82]
A women holding a child in a purple shirt.	The women is asleept at home.	[1, 53, 46]

ChaosNLI dataset by Nie, Zhou, Bansal (2020), Examples from Appendix of Baan et al., 2022

Natural Language Inference: How Frequent?

- "For 20% of the sentence pairs, there is a non-trivial second component (GMM; Pavlick & Kwiatkovski, 2019)





(Figure from Jiang & de Marneffe, 2022)

More NLP task examples (to name a few):

- **Toxic language detection:** Not all text is *equally toxic* for everyone (Sap et al., 2019). *Subjective language tasks* (Akhtar et al, 2021; Leonardelli et al., 2021; Ceras Curry et al., 2021)

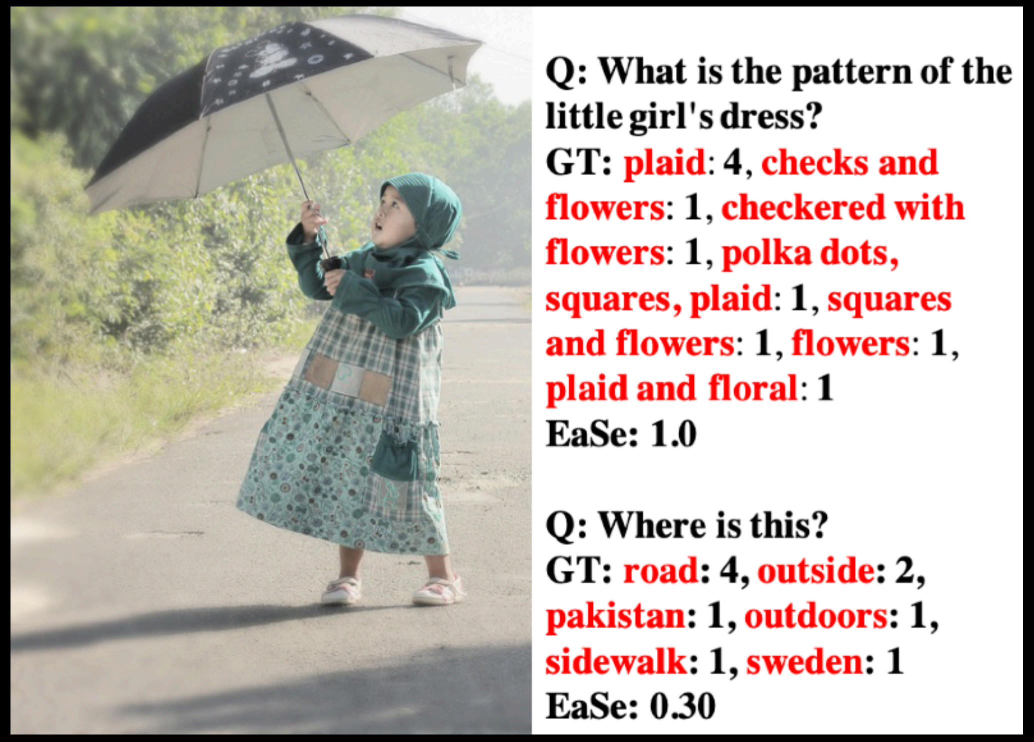
- **Understanding indirect answers** to polar questions (e.g. Damgaard et al., 2021)

Q: Hey. Everything ok?
A: I’m just mad at my agent

? Yes
? No
? Yes, subject to some condition
? Neither Yes nor no

- **Visual Question Answering** (Jolly et al., 2021)

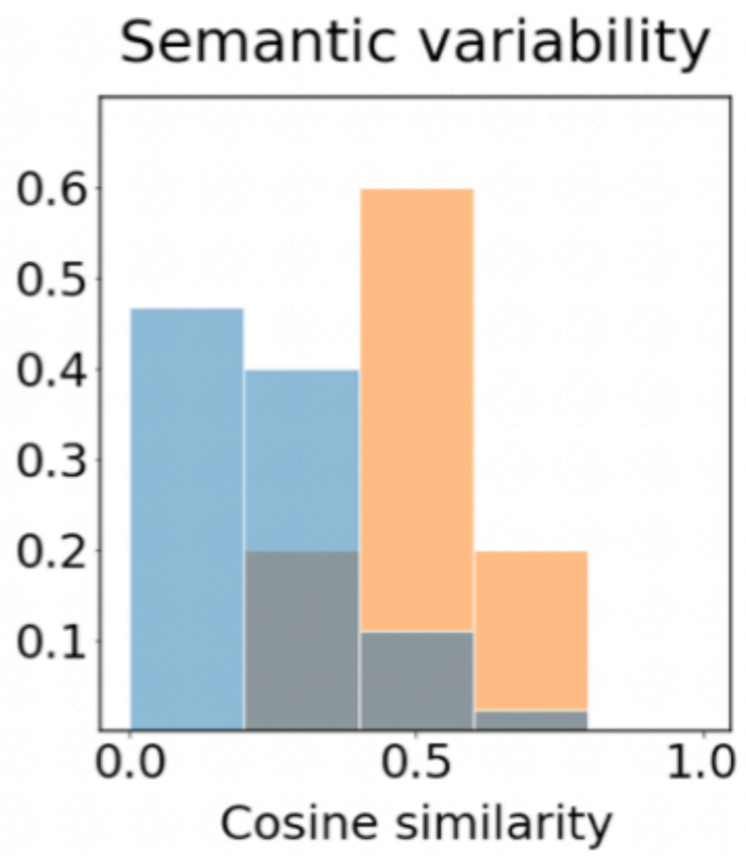
Q: Where is this?
GT: road (4), outside (2), outdoors (1), sidewalk (1), …



HLV not just labels: Natural Language Generation

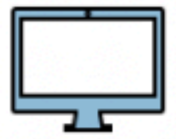
Dialogue context

It's very dark in here. Will you turn on the light?
Okay. But our baby has fallen asleep.
Then, turn on the lamp, please.
But where's the switch?



Humans

- Don't you know where the switch is?
- Switch is on the left side of the lamp.
- Just press the second switch on the board.
- Lamp is upon the study table and now you know where the switch is.
- I will light up the torch, so you can find the switch and turn on the lamp.



DialoGPT-medium, nucleus $p = 0.9$

- You don't have one.
- Where's the button?
- It’s on the top.
- Well, you'll want to turn it on.
- Turn it on.

(Guilianelli, Baan et al., 2023)

Is **Human Label Variation** So Bad? No.

It provides opportunities for more trustworthy, human-facing AI.

Many open questions

- How does Human Label Variation interact with Socially Aware NLP?
- How much Variation is there Within-Human Label Variation?
- How does Human Label Variation relate to Model Uncertainty?
- Is All Variation Plausible?
- How to tease apart Annotation Error from Plausible Human Variation?
- Do we Need More Labels or More Cases?

Human Label Variation: Growing interest

Toward a Perspectivist Turn in Ground Truthing for Predictive Computing

Valerio Basile[1], Federico Cabitza[2], Andrea Campagner[2], and Michael Fell[1]

Why Don't You Do It Right?
Analysing Annotators' Disagreement in Subjective Tasks

Marta Sandri, Dept. of Humanities, University of Pavia, Italy, sandri.marta97@gmail.com
Elisa Leonardelli, Fondazione Bruno Kessler, Trento, Italy, eleonardelli@fbk.eu
Sara Tonelli
Elisabetta Jezek

Understanding and Predicting Human Label Variation in Natural Language Inference through Explanations

Nan-Jiang Jiang[1] Chenhao Tan[2] Marie-Catherine de Marneffe[13]

DisaggregHate It Corpus: A Disaggregated Italian Dataset of Hate Speech

Marco Madeddu[1], Simona Frenda[1,2], Mirko Lai[1,2], Viviana Patti[1] and Valerio Basile[1]

More Labels or Cases?
Assessing Label Variation in Natural Language Inference

Cornelia Gruber[*1♠] Katharina Hechinger[*1♠] Matthias Aßenmacher[1,2♠]
Göran Kauermann[1♠] Barbara Plank[2,3♣]

Interpreting Predictive Probabilities:
Model Confidence or Human Label Variation?

Joris Baan, Raquel Fernández, Barbara Plank, Wilker Aziz

EPIC: Multi-Perspective Annotation of a Corpus of Irony

Simona Frenda[⋆⊙], Alessandro Pedrani[◇], Valerio Basile[⋆], Soda Marem Lo[⋆], Alessandra Teresa Cignarella[⋆⊙], Raffaella Panizzon[◇], Cristina Marco[◇], Bianca Scarlini[◇], Viviana Patti[⋆], Cristina Bosco[⋆], Davide Bernardi[◇]

ACTOR: Active Learning with Annotator-specific Classification Heads to Embrace Human Label Variation

Xinpeng Wang and Barbara Plank

Through the Lens of Split Vote: Exploring Disagreement, Difficulty and Calibration in Legal Case Outcome Classification

Shanshan Xu[1], Santosh T.Y.S.S[1], Oana Ichim[2]
Barbara Plank[3,4], Matthias Grabmair[1]

Annotator-Centric Active Learning for Subjective NLP Tasks

Michiel van der Meer, LIACS, Leiden University
Neele Falk, Institute for Natural Language Processing, University of Stuttgart

Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement

Gavin Abercrombie[1] and Verena Rieser[1,2] and Dirk Hovy[3]

When the Majority is Wrong: Modeling Annotator Disagreem… Subjective Tasks

Eve Fleisig[†] Rediet Abebe Dan Klein

Which Examples Should be Multiply Annotated?
Active Learning When Annotators May Disagree

Connor Baumler[*] Anna Sotnikova[*] Hal Daumé III

The Ecological Fallacy in Annotation:
Modelling Human Label Variation goes beyond Sociodemographics

Matthias Orlikowski[1], Paul Röttger[2], Philipp Cimiano[1], and Dirk Hovy[3]

Can Large Language Models Capture Dissenting Human Voices?

Noah Lee[*] Na Min An[*] James Thorne

Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation

Flor Miriam Plaza-del-Arco, Debora Nozza, Dirk Hovy

Roadmap: Selected Case Studies

1. Humans and Uncertainty: *The “Problem” of Human Label Variation*
2. Models and Uncertainty: *Stop Measuring Calibration When Humans Disagree*
3. How to detect errors? *ActiveAED*
4. Plausible variation or error? *VariERR*

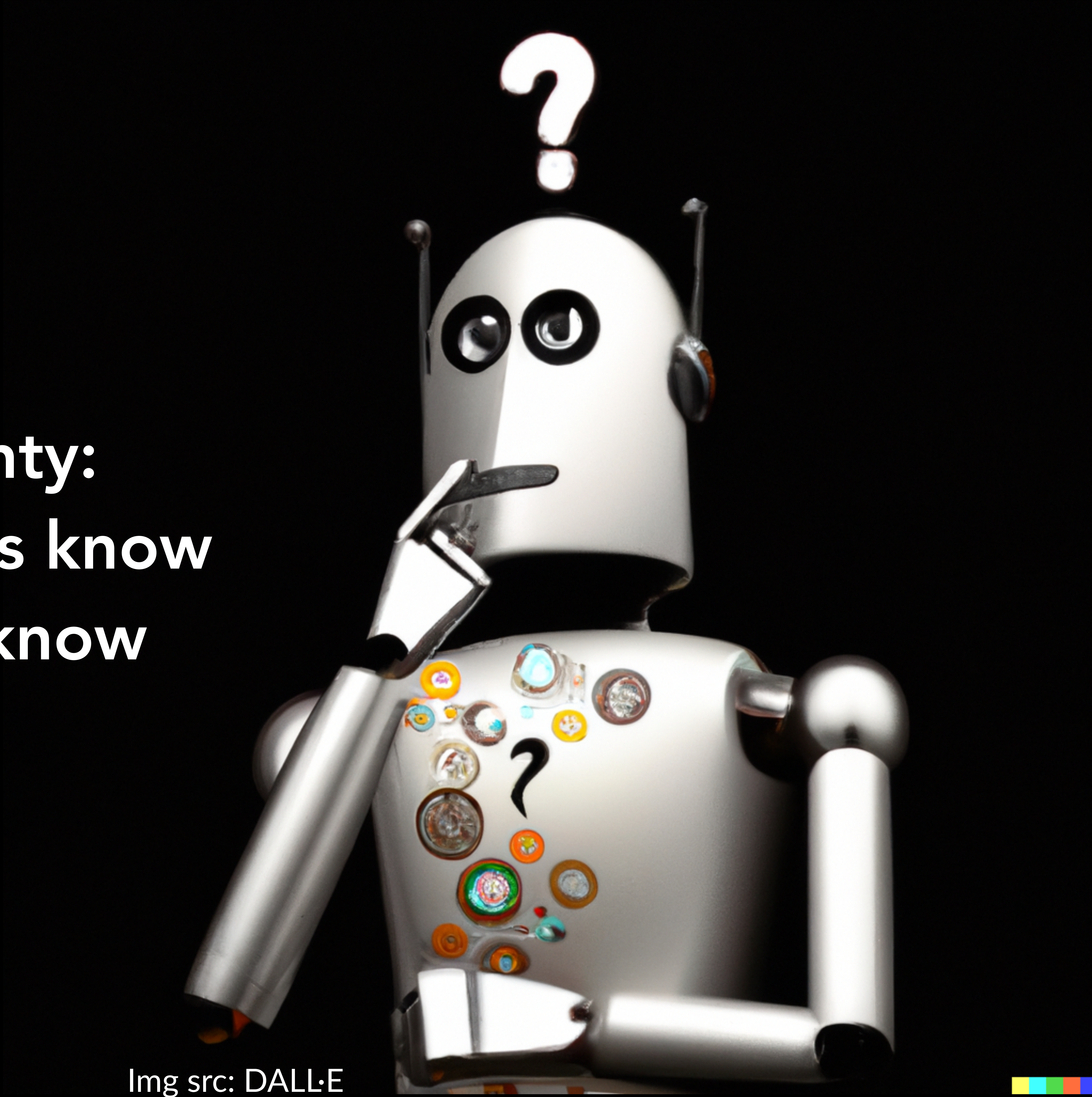
Stop Measuring Calibration When Humans Disagree

Joris Baan[1], **Wilker Aziz**[1], **Barbara Plank**[2,3,4], **Raquel Fernández**[1]
[1]University of Amsterdam, [2]IT University of Copenhagen, [3]MCML Munich, [4]LMU Munich
{j.s.baan,w.aziz,raquel.fernandez}@uva.nl, b.plank@lmu.de

(Baan, Aziz, Plank, Fernandez, 2022 EMNLP)

Uncertainty

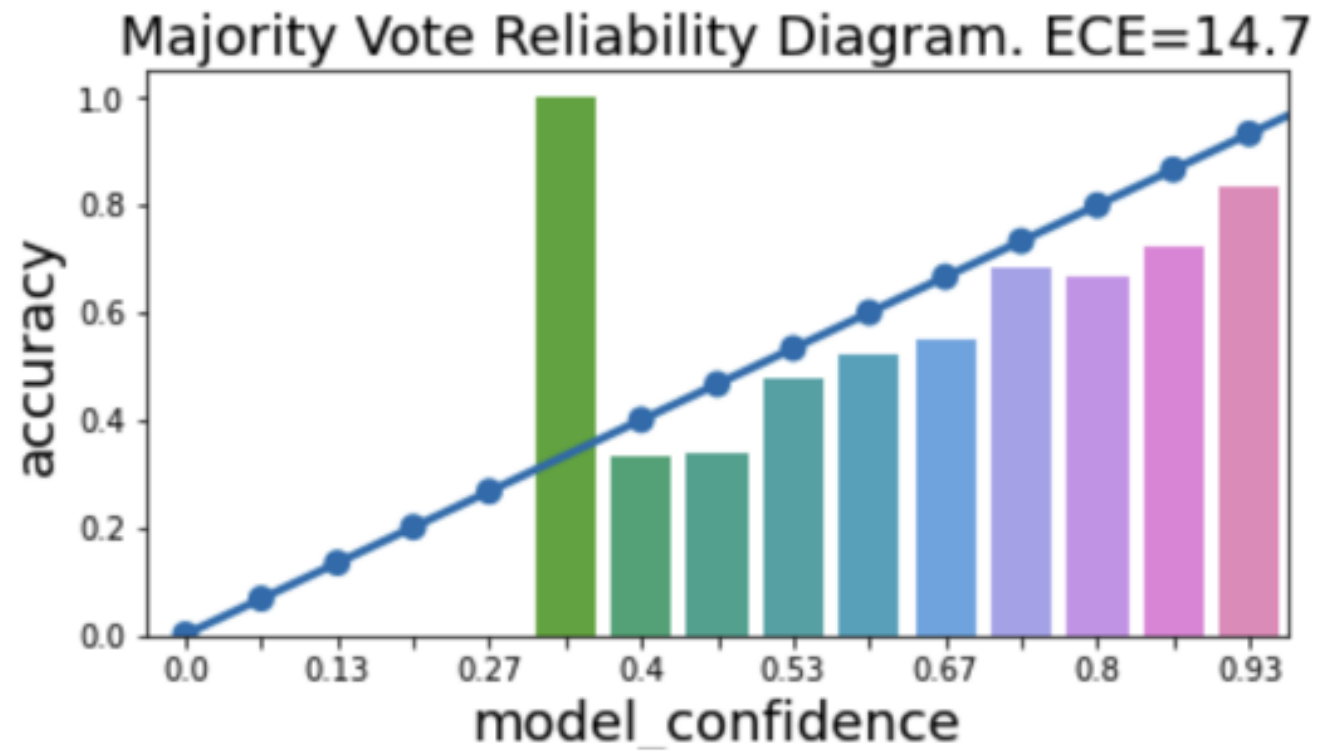
Model Uncertainty:
Models don’t always know
when they don’t know



Img src: DALL·E

More trustworthy models: Calibration & Model Uncertainty

- Calibration is a popular framework to evaluate whether a classifier knows when it does not know
- Reliability diagram to indicate how well calibrated a model is
  - ECE (expected calibration error)



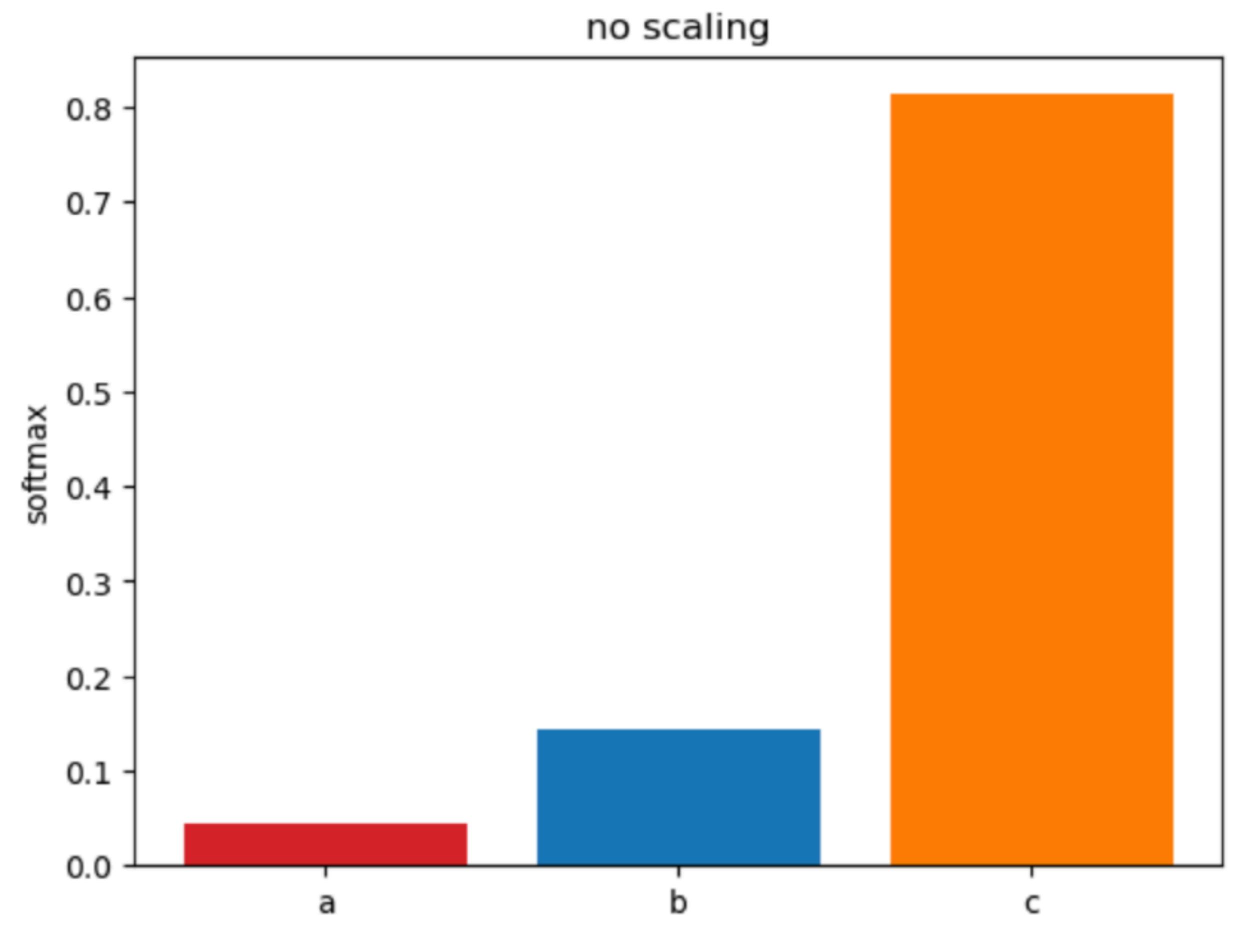
- What does calibration mean when there is no ground truth?
  - We examine calibration under the lens of HLV

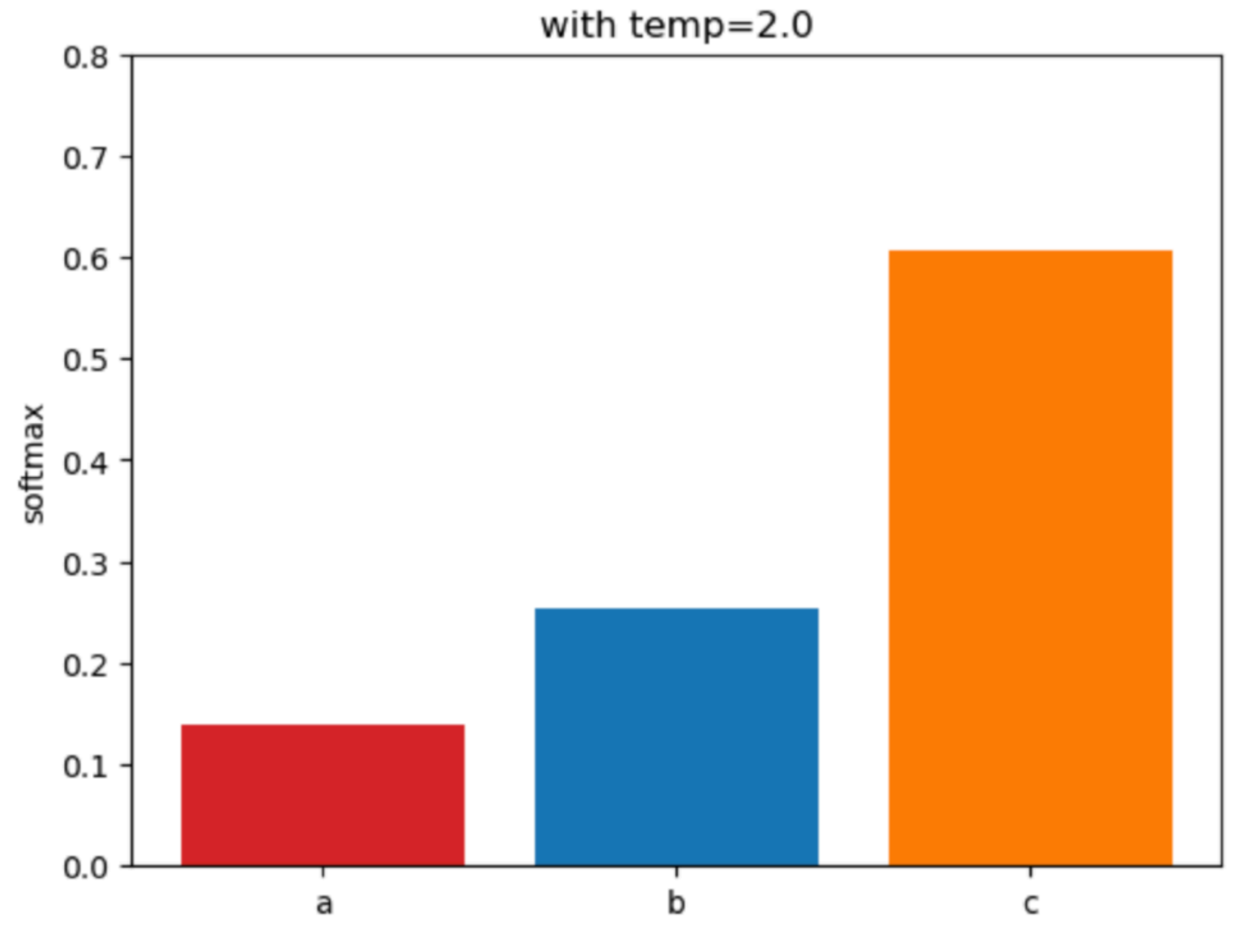
Calibration: Temperature Scaling

- Temperature Scaling is one way to do calibration. It is a post-processing technique to improve the calibration error. It works by dividing the logits by a scalar T:

$$softmax_T = \frac{e^{z/T}}{\sum_i e^{z_i/T}}$$

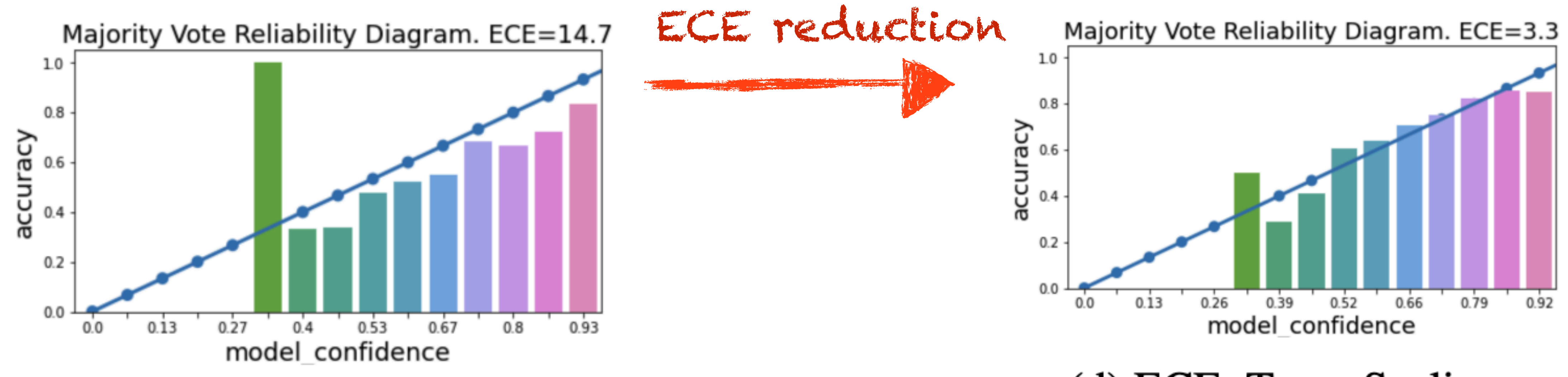
- If T>1.0, it makes the model less confident about its predictions:





Calibration to majority? BAD.

- Temperature Scaling can help improve ECE:



(c) ECE: Vanilla

(d) ECE: Temp Scaling

- However, we observe that **despite low ECE**, an oracle is still miscalibrated:

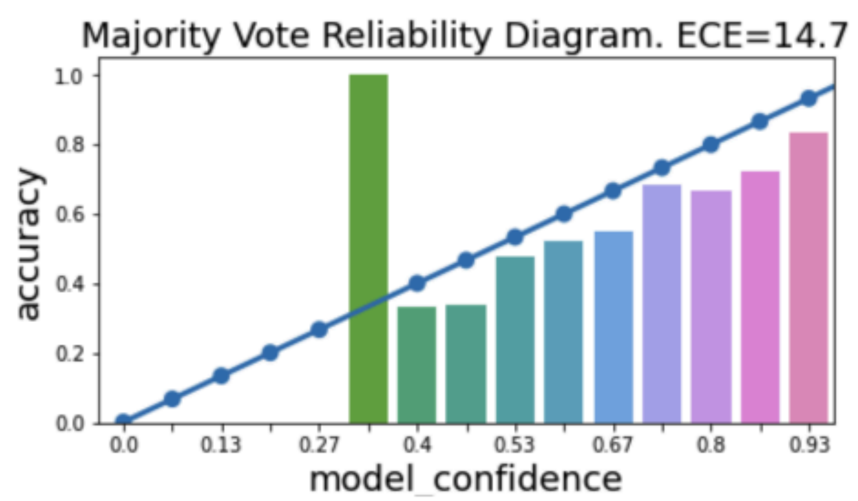
	RoBERTa	RoBERTa-TS	Oracle
Acc ↑	0.74±0.01	0.74±0.01	1.00
ECE ↓	0.14±0.01	0.03±0.01	0.25

- What can we do? Measure **Human Calibration Error (DistCE)**:
  - Total variation distance between predictive distribution and human judgement distribution (range: 0…1)

$$\mathrm{DistCE}(x) = \mathrm{TVD}(\mathbf{f}(x), \bar{\boldsymbol{\pi}}(x))$$

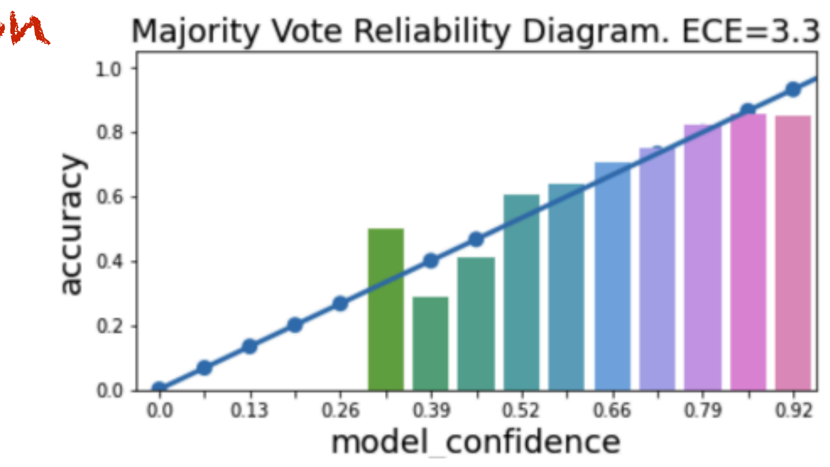
Calibration in Light of HLV

- DistCE = instance-level analysis, enables a more fine-grained view on model calibration (Baan et al., 2022). Recall:



(c) ECE: Vanilla

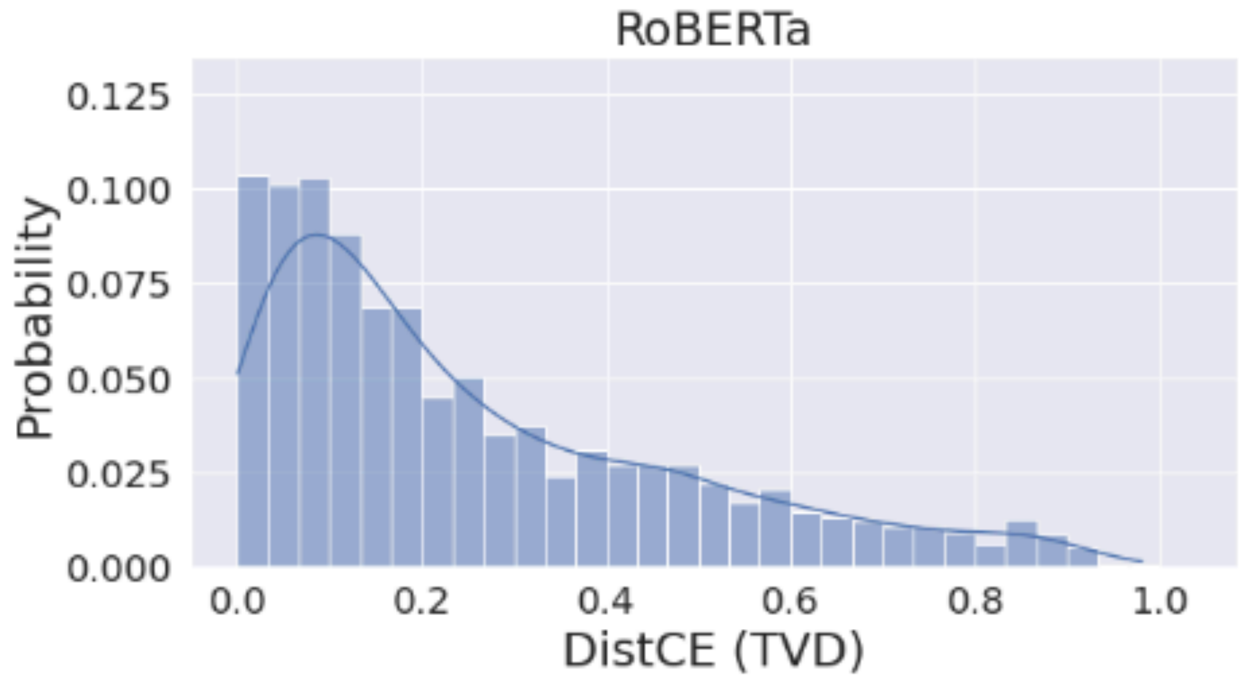
ECE reduction



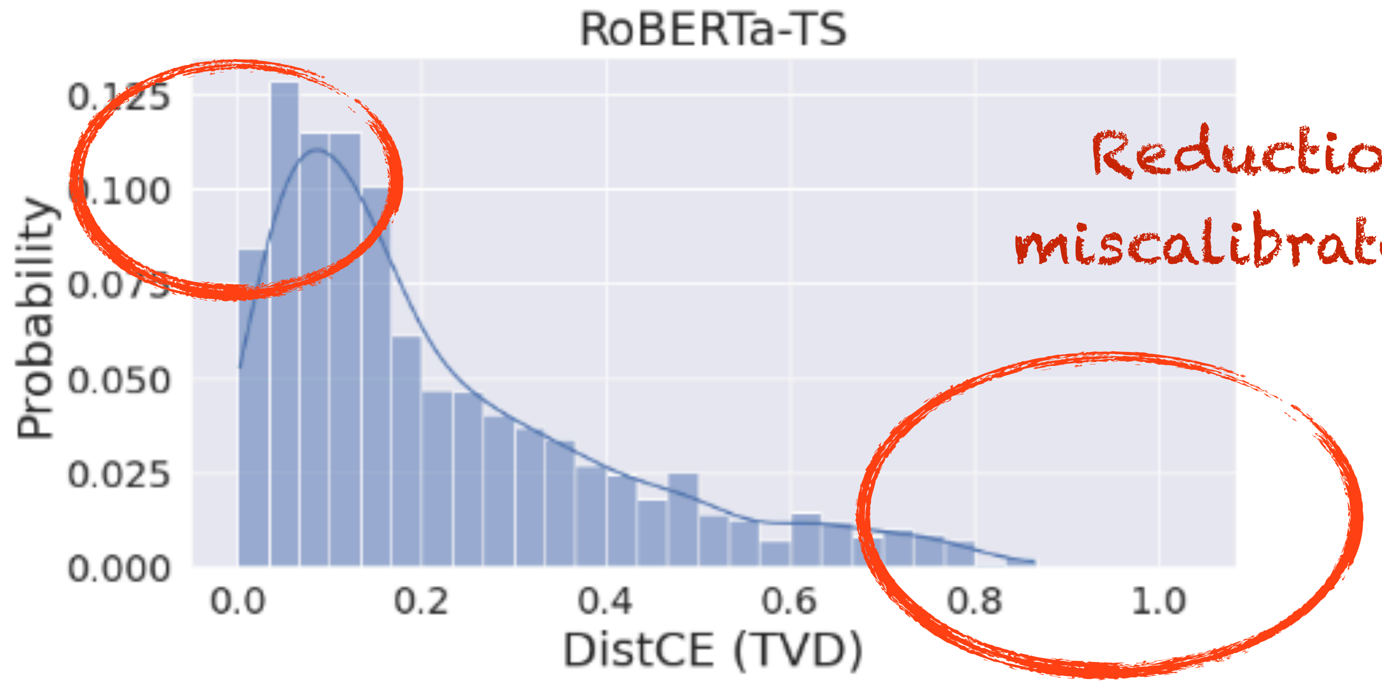
(d) ECE: Temp Scaling

BUT also fewer perfectly calibrated instances!

- DistCE ⬇ (0: perfectly calibrated to human judgement)



(a) DistCE: Vanilla



(b) DistCE: Temp Scaling

Reduction of most miscalibrated instances

Take-home message

(Baan et al., 2022 EMNLP)

- We showed that calibration to human majority is flawed
- We suggested to look at calibration in light of HLV
  - Proposed several measures (more in the paper), incl. **Human calibration error** (DistCE), that provide us **instance-level** insights
  - More nuanced insights into model **uncertainty**
- **Limitation**: requires data with human label variation

Roadmap: Selected Case Studies

1. Humans and Uncertainty: *The "Problem" of Human Label Variation*
2. Models and Uncertainty: *Stop Measuring Calibration When Humans Disagree*
3. How to detect errors? *ActiveAED*
4. Plausible variation or error? *VariERR*

Existing datasets contain annotation errors

Data Quality

Djamé.. @zehavoc · 20h

just found out this wonderful quote in an old paper where we described our efforts to parse the British National Corpus (100M words, back then it was huge, clusters and all) work by @Wjrgo @jenfoster, Josef van Genaboth and I
web.stanford.edu/group/cslipubl...

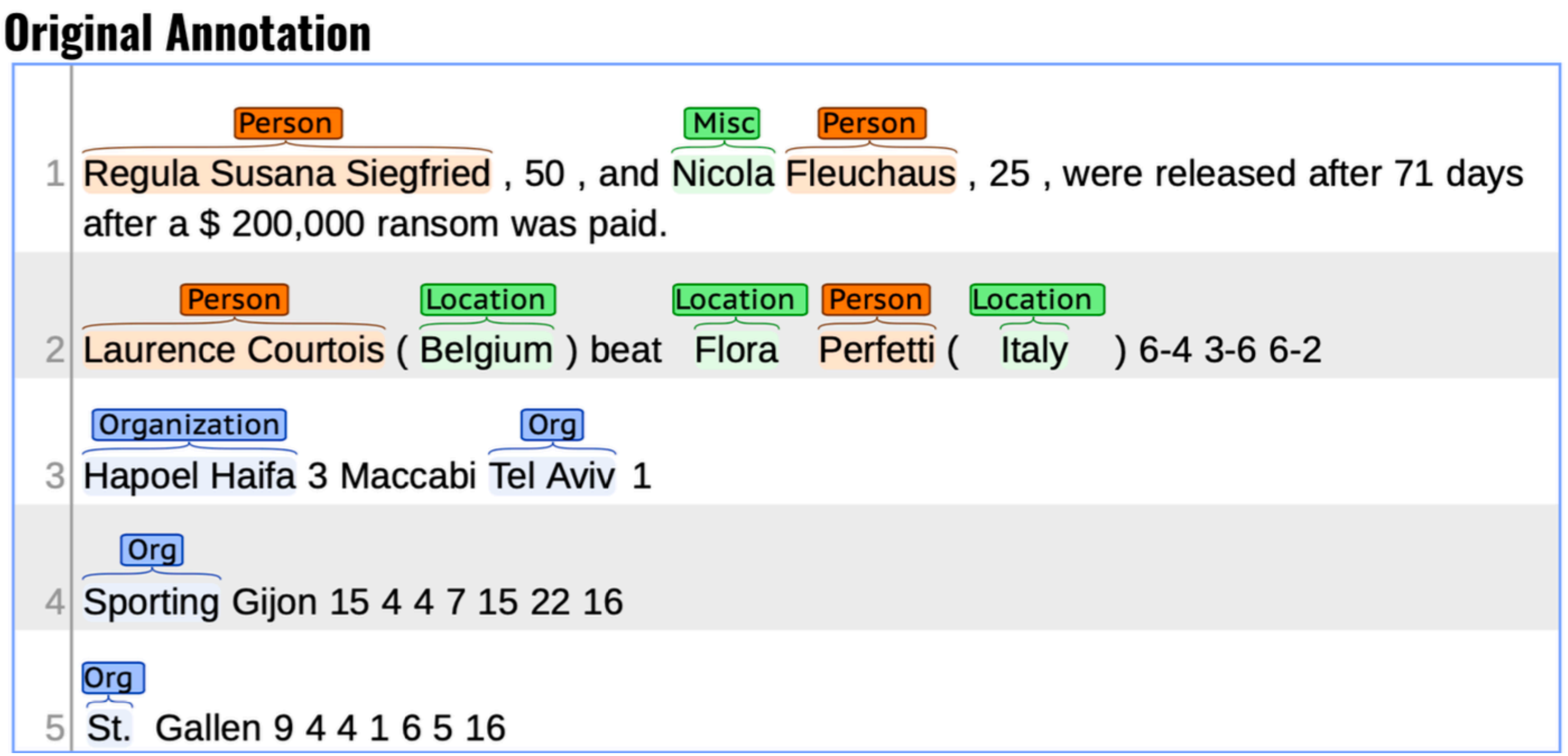
Still applies today imho

> *"Cleaning is a low-level, unglamorous task, yet crucial: The better it is done, the better the outcomes. All further layers of linguistic processing depend on the cleaniness of the data."*
> (Kilgarriff, 2007, p.149)

Example: Sentiment (Imdb)

Review	Original Label
SPOILERS AHEAD It is really unfortunate that a movie so well produced turns out to be such a disappointment. [...]	Positive
Lois Weber's film "Hypocrites" was and still kind of is a very bold and daring film. I enjoyed it and was very impressed by the filming and story of it. [...]	Negative

Example: NER (CoNLL 2003)



What to do about it?

Annotation Error Detection (AED)

- A long-standing task (e.g. Dickinson & Meurers, 2003); recently surveyed comprehensively by Klie, Webber, Gurevych (2022)
- Typical AED methods are post-hoc processing
- We propose to combine AED with human in the loop: **Active AED**

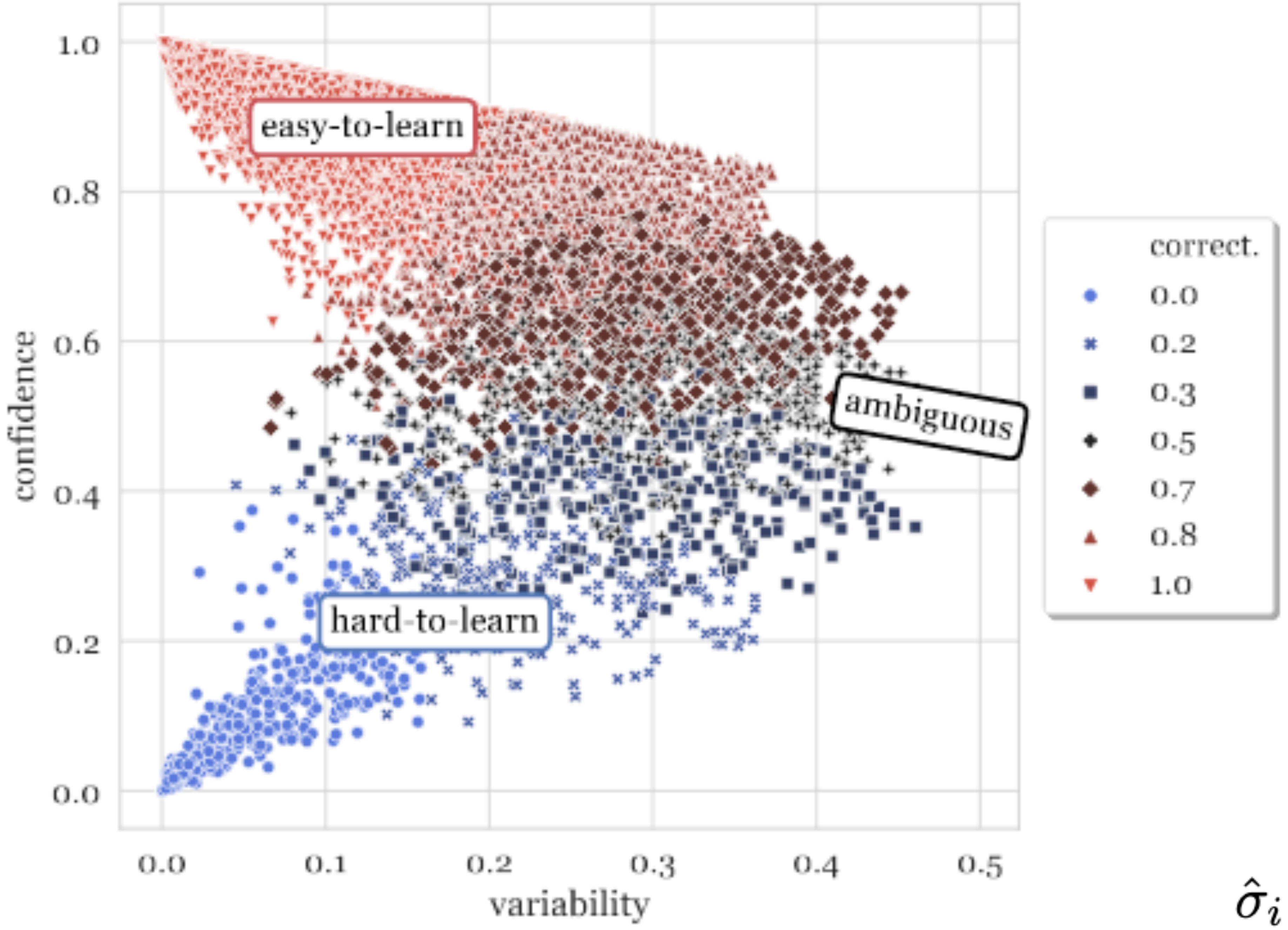
ActiveAED: A Human in the Loop Improves Annotation Error Detection

Leon Weber▲ and **Barbara Plank**▲◇
▲Center for Information and Language Processing (CIS), LMU Munich, Germany
◇Munich Center for Machine Learning (MCML), Munich, Germany
{leonweber, bplank}@cis.lmu.de

(Weber & Plank, 2023 ACL Findings)

Dataset cartography: Training dynamics

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\boldsymbol{\theta}^{(e)}}(y_i^* \mid \boldsymbol{x}_i)$$



$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\boldsymbol{\theta}^{(e)}}(y_i^* \mid \boldsymbol{x}_i) - \hat{\mu}_i\right)^2}{E}}$$

Data map for SNLI train set, based on a ROBERTA-large classifier. The x-axis shows **variability** and y-axis, the **confidence**; the colors/shapes indicate **correctness**.

(Swayamdipta et al, 2020)

Our solution: ActiveAED

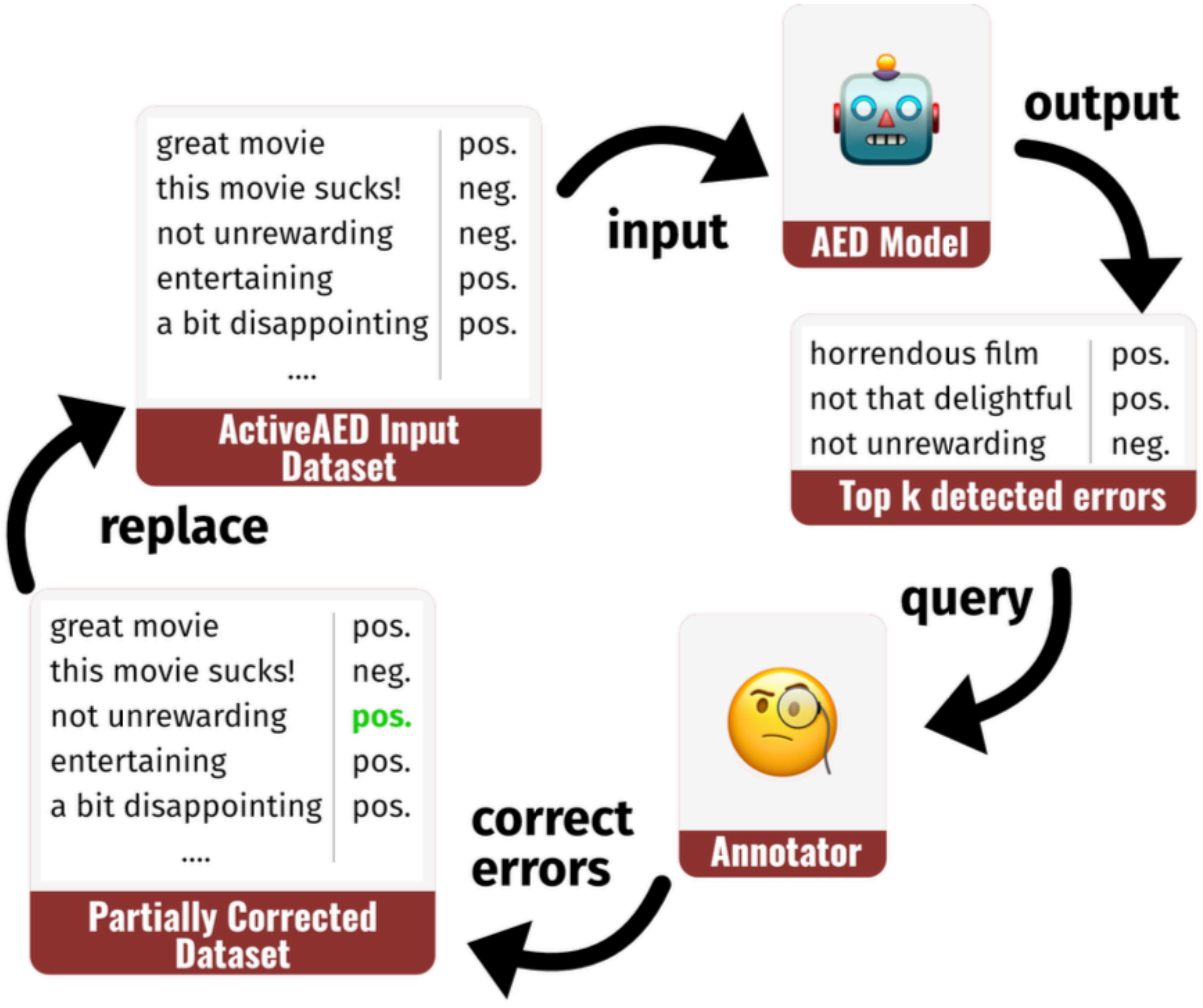
- ActiveAED: Involve human annotator in pipeline, by repeatedly querying for error corrections

- Can be used with any scoring-based method. We use Area-Under-the-Margin (Pleiss et al. 2020)

$$s_i = \frac{1}{E}\sum_{e=1}^{E} \max_{y' \neq y_i} p_{\theta_e}(y'|x_i) - p_{\theta_e}(y_i|x_i)$$

- Our novel ensembling scheme merges training-dynamics-based and cross-validation-based AED for improved results

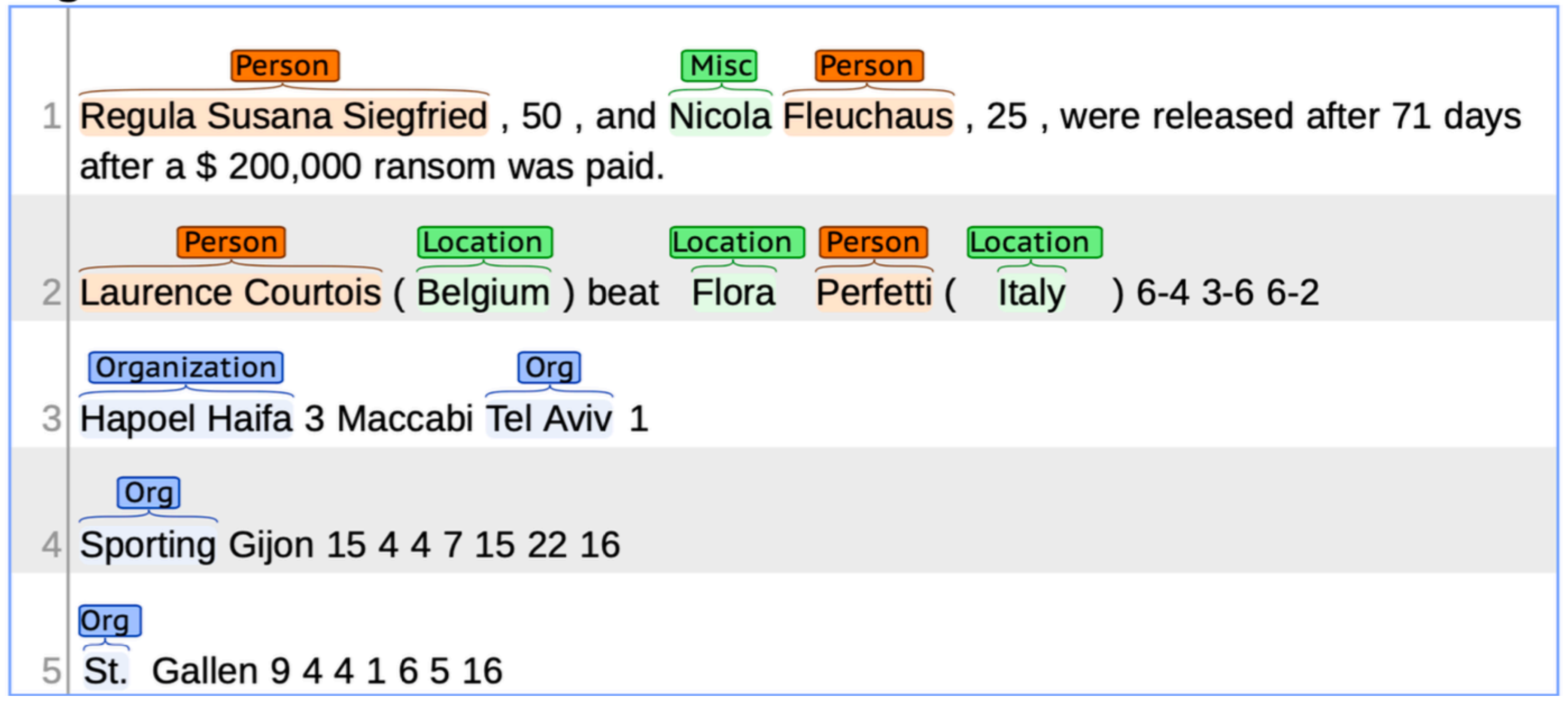
$$\begin{aligned} s_i^{train} &= \tfrac{1}{E-1} \textstyle\sum_{c \in train_i} s_{c,i} \\ s_i &= \tfrac{1}{2}(s_i^{train} + s_i^{test}) \end{aligned}$$



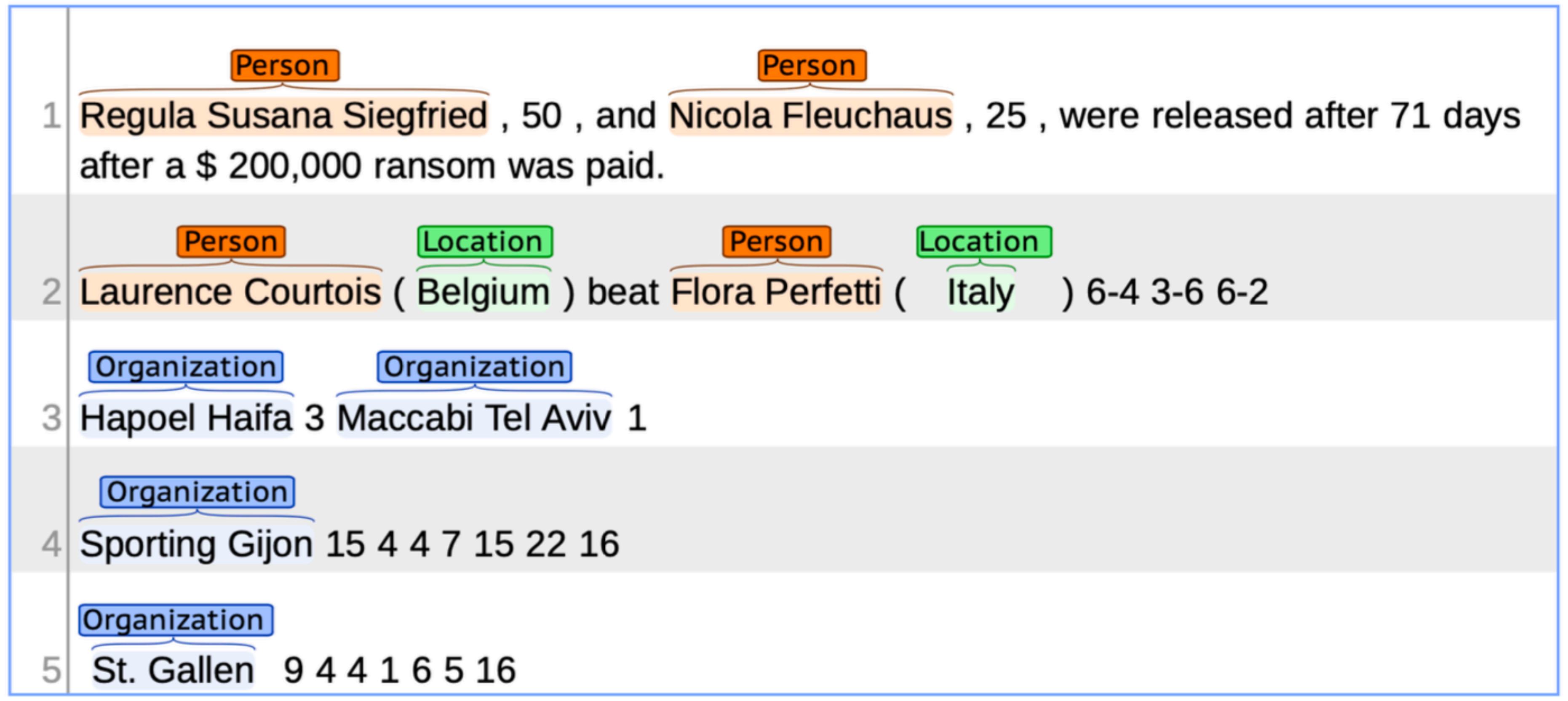
Main results

	ATIS	SI-Flights	IMDb	SST	GUM	CONLL-2003	SI-Companies	SI-Forex
CU	91.7±1.4	80.9±0.5	31.6±1.3	42.7±1.0	98.8±0.1	25.2±0.6	96.1±0.2	84.2±2.0
DM	97.2±0.2	79.2±2.4	30.1±3.0	47.1±1.0	99.3±0.1	30.2±0.7	97.5±0.2	80.6±0.9
AUM (p)	98.0±0.1	78.9±2.3	30.1±3.0	47.1±1.0	99.0±0.1	30.2±0.7	97.3±0.3	81.1±0.9
AUM (l)	97.3±0.4	72.6±0.3	27.5±2.5	39.6±1.3	**99.5±0.1**	29.3±0.2	97.2±0.2	66.6±1.5
ActiveAED	**98.6±0.1**	**86.6±0.5**	**36.6±0.1**	**53.0±0.2**	98.5±0.0	**33.3±0.2**	**99.3±0.0**	**89.7±0.6**
w/o active	98.7±0.1	80.3±0.6	36.0±0.4	52.9±0.4	98.4±0.0	31.7±0.4	97.9±0.1	85.5±0.6

Original Annotation



Corrected Annotation



Conclusion

(Weber & Plank, 2023)

ActiveAED: 🧐 A human in the loop improves Annotation Error Detection.

So far studied on AED were limited to (discriminative) classification tasks

DONKII: Characterizing and Detecting Errors in Instruction-Tuning Datasets

Leon Weber-Genzel and **Robert Litschko** and **Ekaterina Artemova**[*]
and **Barbara Plank**

MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

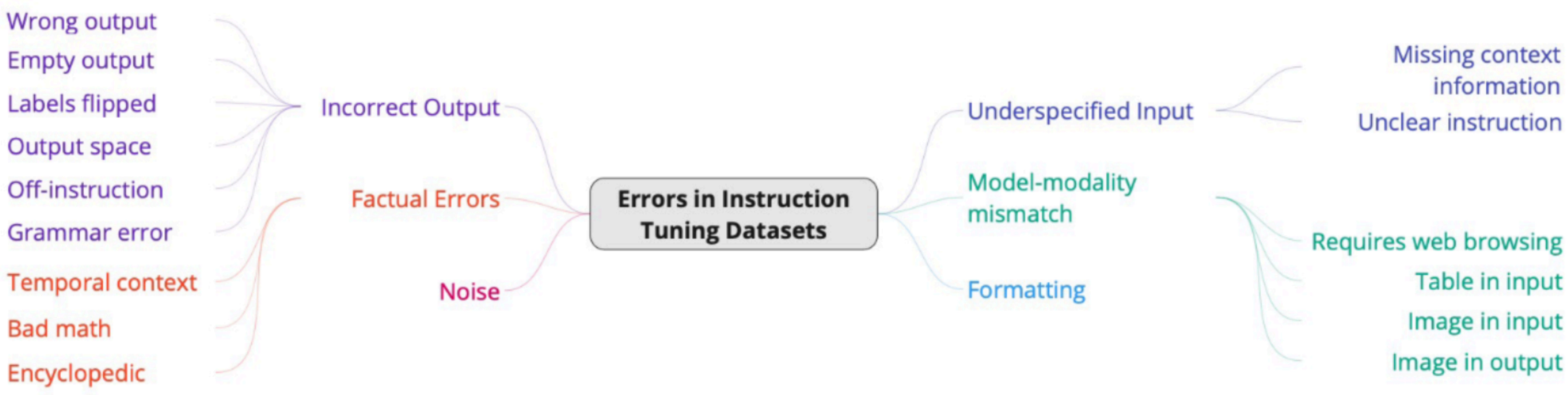
Munich Center for Machine Learning (MCML), Munich, Germany

Department of Computer Science, IT University of Copenhagen, Denmark

{leonweber, robert.litschko, b.plank}@lmu.de

(Weber et al., 2024 LAW@EACL 2024)

Taxonomy of Error Types



Examples

Error	Source	Instruction	Input	Output (shortened)
Requires web browsing	Alpaca	Summarize the New York Times article about the rise of AI.	`www.nytimes.com/2020/06/16/opinion/ai-growth.html`	The New York Times article titled "The Rise of Artificial Intelligence" looks at the growing impact of AI on the world today. <...>
Bad math	Alpaca	Calculate the mean of the numbers 3, 4, 6, 10, 11		The mean of the numbers 3, 4, 6, 10, 11 is 6.4. This is calculated by adding the numbers together and then dividing by the number of values (5). (3 + 4 + 6 + 10 + 11 = 34; 34 / 5 = 6.4).
Unclear instruction	Alpaca	Sort the following 10 fruits into three categories.	Bananas, Apples, Pineapples, Plums, Cherries, Peaches, Grapes, Watermelons, Mangoes, Oranges	Category 1: Bananas, Apples, Pineapples, Plums, Cherries;

🤔 What's an error vs plausible variation?

Roadmap: Selected Case Studies

1. Humans and Uncertainty: *The "Problem" of Human Label Variation*
2. Models and Uncertainty: *Stop Measuring Calibration When Humans Disagree*
3. How to detect errors? *ActiveAED*
4. Plausible variation or error? *VariERR*

Motivation

- While **Human Label Variation** exists, so do **errors**.
- Annotators are inevitably prone to make errors.
- We lack both a theory and operationalizable procedures to answer the RQ:
    - **Can we tease apart error from plausible human label variation?**

Error vs. plausible **Human Label Variation**

VariErr NLI: Separating Annotation Error from Human Label Variation

Leon Weber-Genzel[▲,🤖]* **Siyao Peng**[▲,🤖]* **Marie-Catherine de Marneffe**[✒] **Barbara Plank**[▲,🤖]

▲ MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

🤖 Munich Center for Machine Learning (MCML), Munich, Germany

✒ FNRS, UCLouvain, Belgium

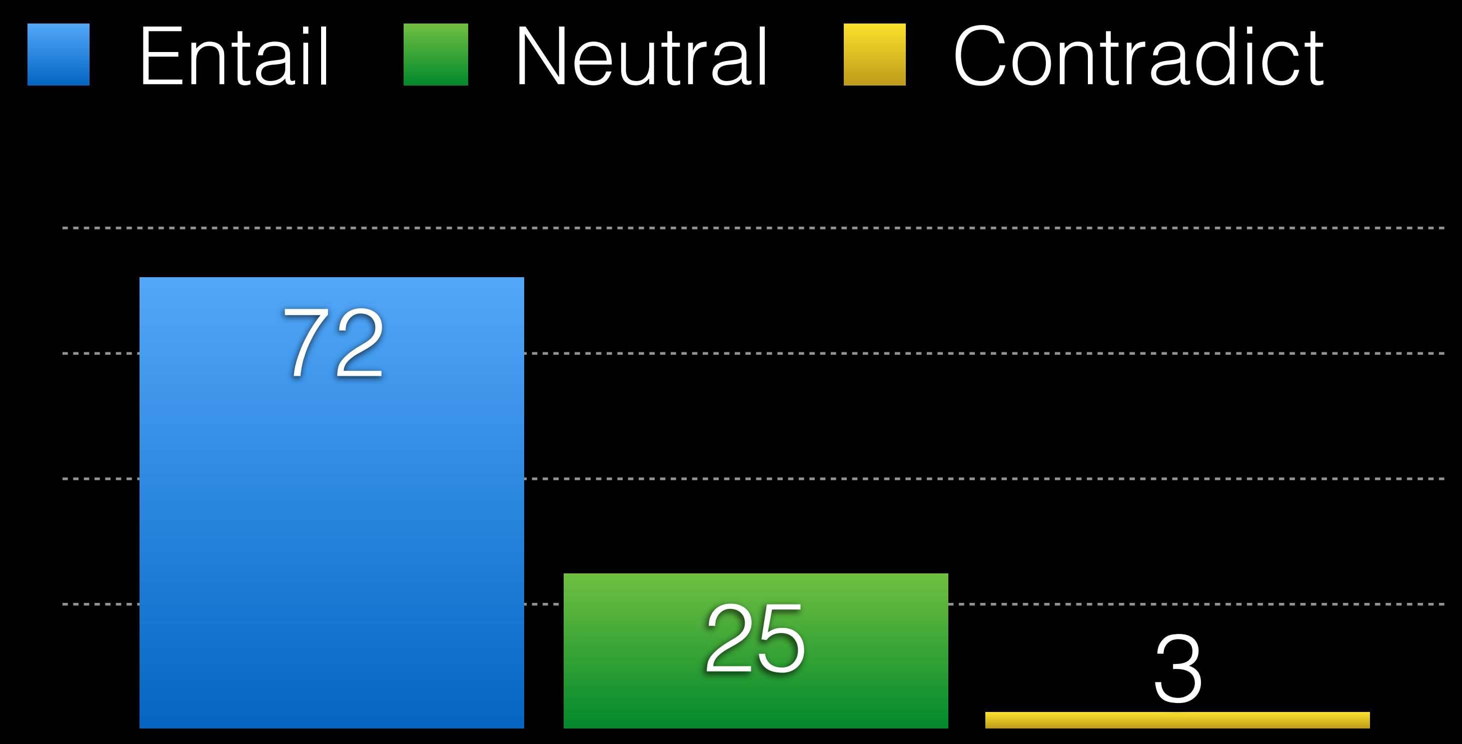
{leonweber,siyaopeng,bplank}@cis.lmu.de marie-catherine.demarneffe@uclouvain.be

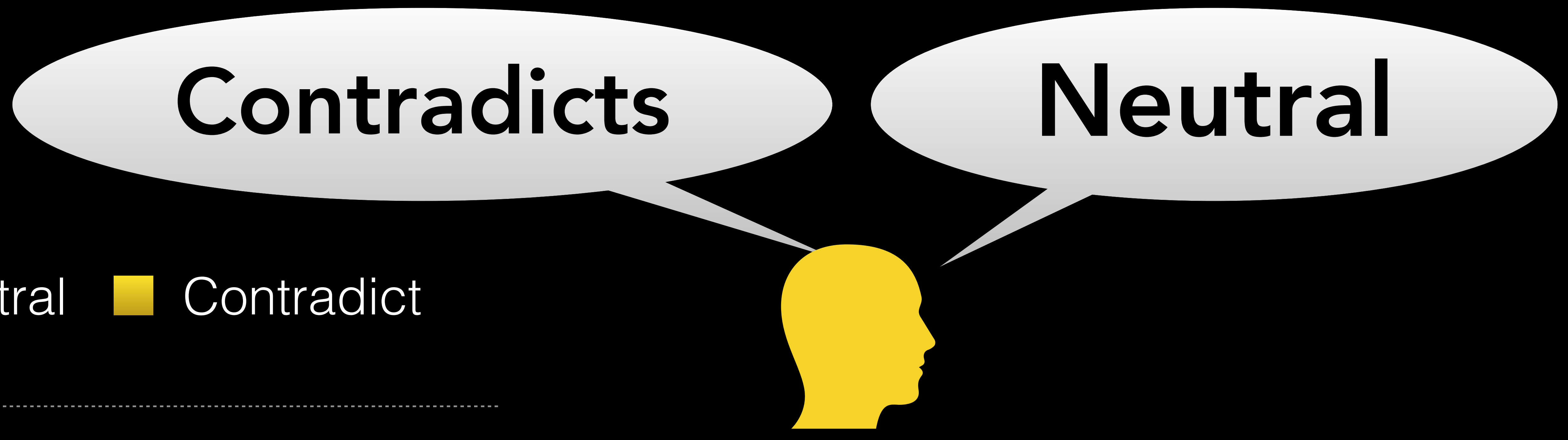
(Weber-Genzel, Peng et al., 2024 To Appear at ACL)

Natural Language Inference

Premise: As he stepped across the threshold, Tommy brought the picture down with terrific force on his head.

Hypothesis: Tommy hurt his head bringing the picture down.





From ChaosNLI (100 labels per instance)
by Nie, Zhou, Bansal (2020)

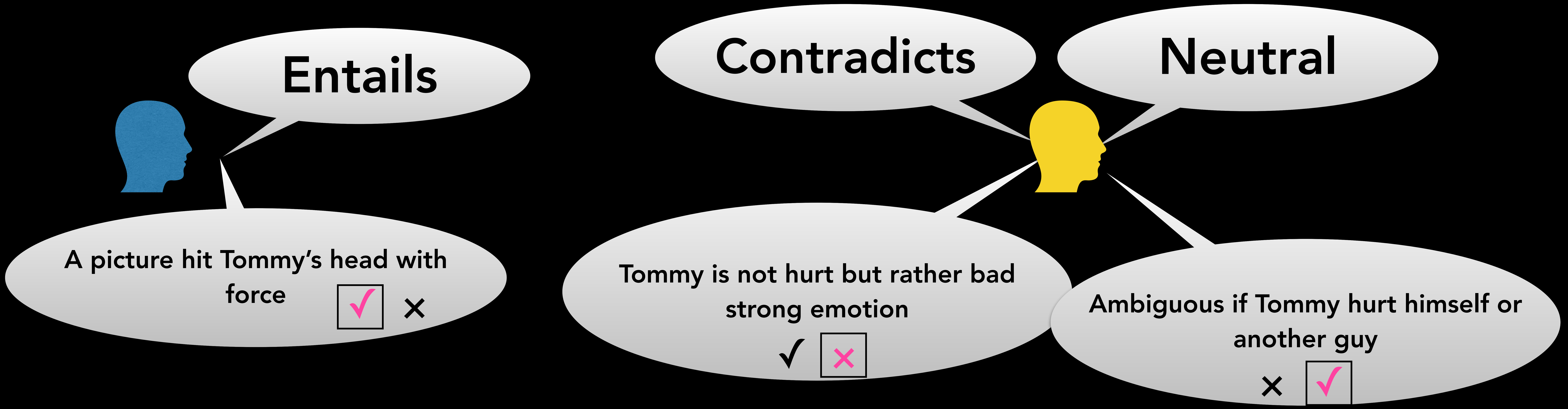
We propose a two step-procedure: 1) Explanations

Premise: As he stepped across the threshold, Tommy brought the picture down with terrific force on his head.

Hypothesis: Tommy hurt his head bringing the picture down.

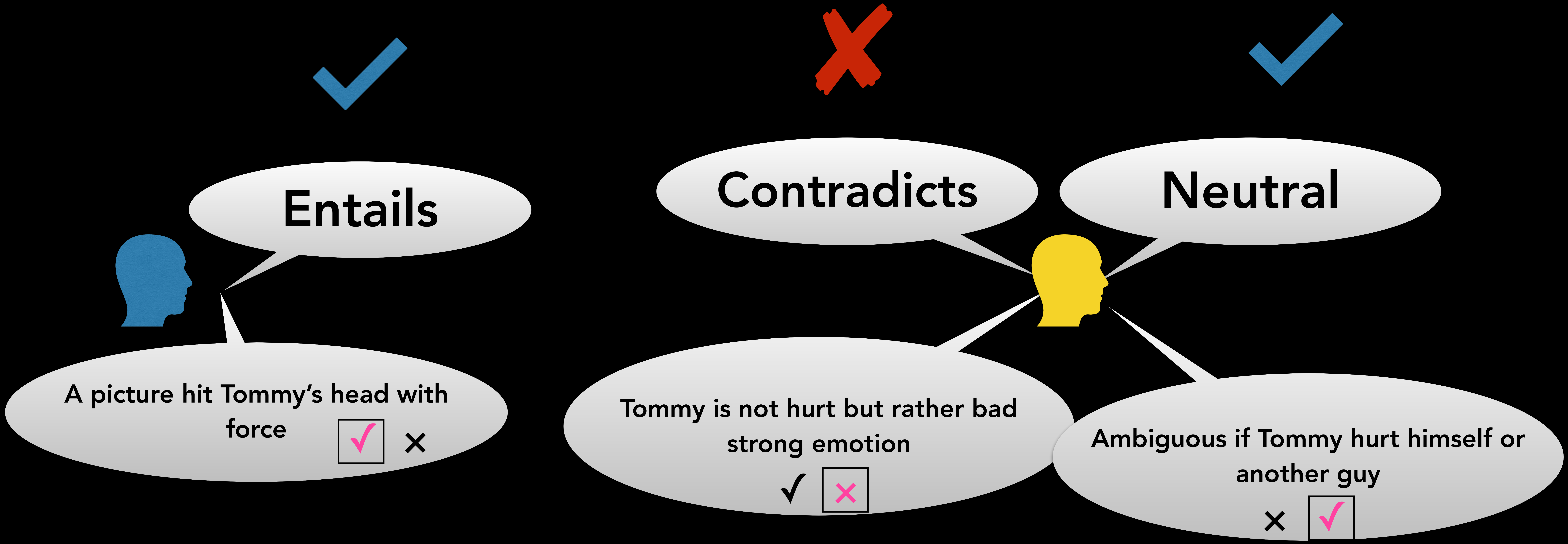
- **Ecologically valid explanations** inspired by (Jiang et al., 2023)

We propose a two step-procedure: 2) Validations



- Another kind of validation: see your own and peer's label-explanation pairs

ValiErr: Defining Errors



- **Self-validated**: any self-validated label-explanation pair is plausible, otherwise it is an **error**
- **Peer-validated**: A label-explanation pair is peer-validated if >=2 annotators approved it

Example from VariErr NLI:

Premise: `Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.`
Hypothesis: `Newspaper preprints can cost as much as $5.`
Label-explanation pairs: `Before:{E:1,N:2,C:1}  Self-validated:{N:2}  Peer-validated:{N:2,C:1}`
Label: `[N]  Errors:[ E, C ]`

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	4	`5 dollars for a piece of newspaper.`	✗	✗	✗	✗
N	1	`The context only mentions how low the price may be, not how high it may be.`	✓	✓	✓	✓
	3	`The maximum cost of newspaper preprints is not given in the context.`	✓	✓	✓	✓
C	2	`The context says 5 or 6 cents, not $5.`	✗	✗	✓	✓

(a) *id:* 72870c

Table 1: Sample annotations from VariErrNLI corpus. L: Label, A: Annotator; E: Entailment, N: Neutral, C: Contradiction; ✓: ‘yes’; ✗: ‘no’; ?: ‘idk’; magenta: self-judgments, black: peer-judgments, `Err`: label error.

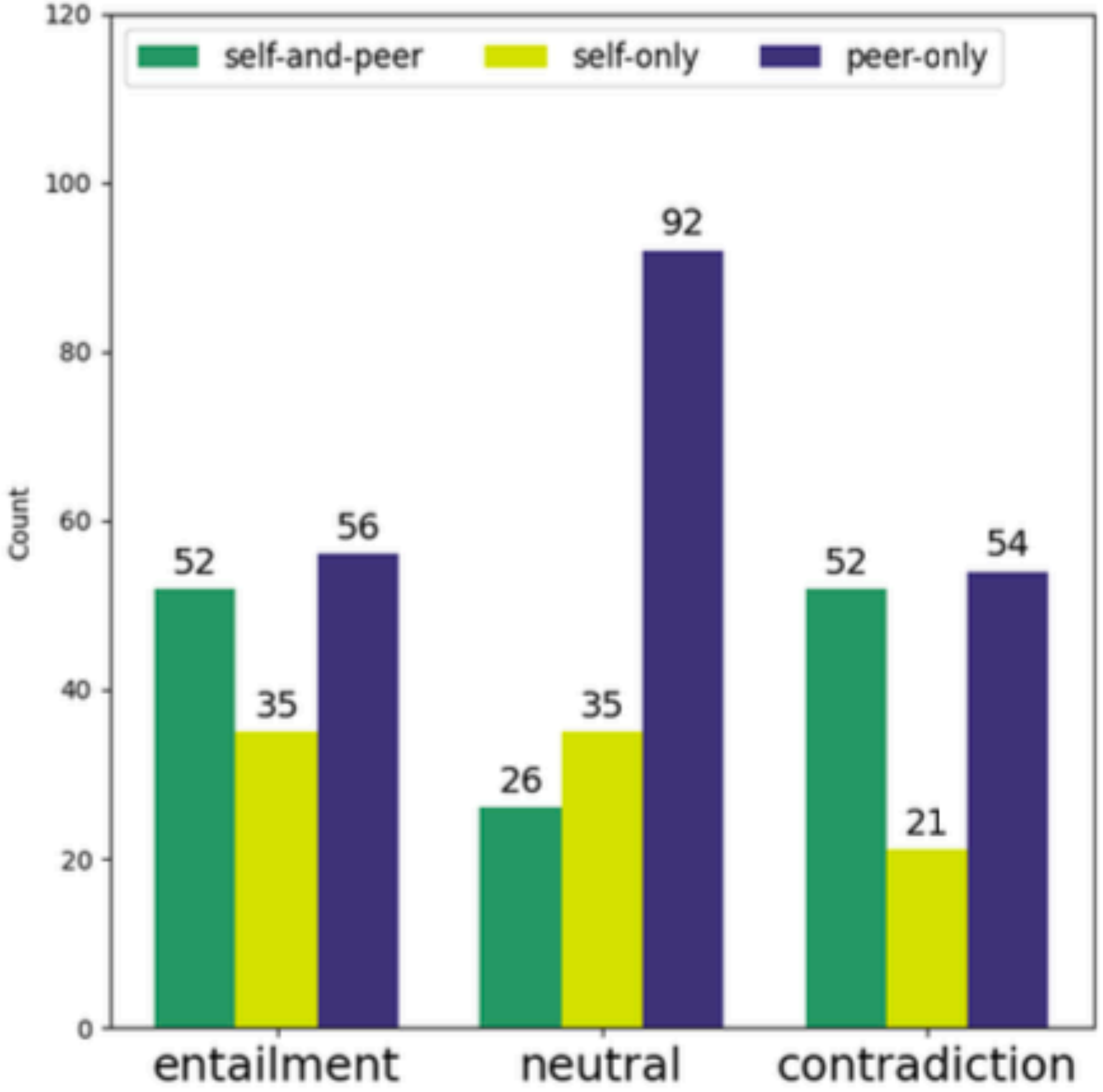
VariErr dataset

- VariErr NLI: re-annotated 500 NLI items from scratch, 1,933 label-explanation pairs
- 88.57% (1,712/1,933) are self-validated, 82.82% are peer-validated (1,601/1,933)
- Overall, 37% of items had self-identified errors (188/500)

Validation	FreqType	E	N	C	$\sum$	IAA
before validation	*repeated*	554	977	402	1,933	0.35
	aggregated	263	403	212	878	
self-validated	*repeated*	467	916	329	1,712	0.50
	aggregated	210	380	159	749	
peer-validated	*repeated*	446	859	296	1,601	0.69
	aggregated	177	335	130	642	

Statistics on VariErr

- Number of label-explanation pairs that were rejected in phase 2
- Most Entailment and Contradiction annotations are rejected by both self- and peer-validations

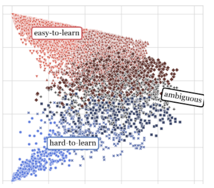


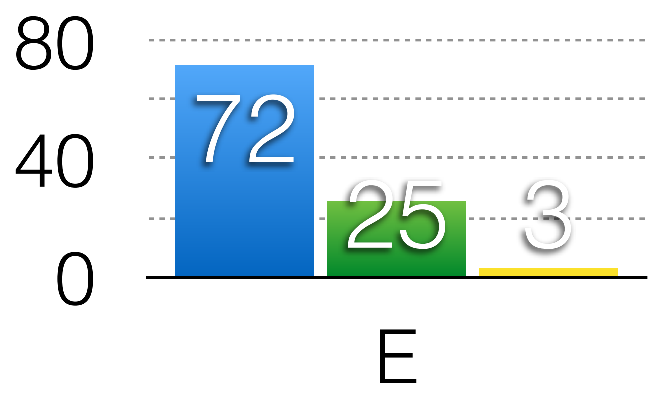
How good is Annotation Error Detection on VariErr?

- We model AED as a ranking task
- scorer to rank the list of labels with errors high
- from 500 items, give list of 878 item-label pairs to scorer
- compare ranked lists to self-flagged errors
- Metrics: Average Precision (AP),
  P/R/F1 of top 100 ranked labels P@100, R@100
- Five AED models: two variants of datamaps, metadata archaeology, two GPTs* (GPT-3.5 and GPT-4)

```
System:
You are an expert linguistic annotator.

User:
We have collected annotations for an NLI
    instance together with reasons for the
    labels. Your task is to judge whether the
    reasons make sense for the label. Provide
    the probability (0.0 - 1.0) that the
    reason makes sense for the label. Give
    ONLY the reason and the probability, no
    other words or explanation. For example:

Reason: <The verbatim copy of the reason>
Probability: <the probability between 0.0 and
    1.0 that the reason makes sense for the
    label, without any extra commentary
    whatsoever; just the probability!>.

Context: {CONTEXT}
Statement: {STATEMENT}

Reason for label {LABEL}: {REASON_1}
Reason for label {LABEL}: {REASON_2}
[...]
Reason for label {LABEL}: {REASON_n}

Reason {REASON_1}
Probability:
```

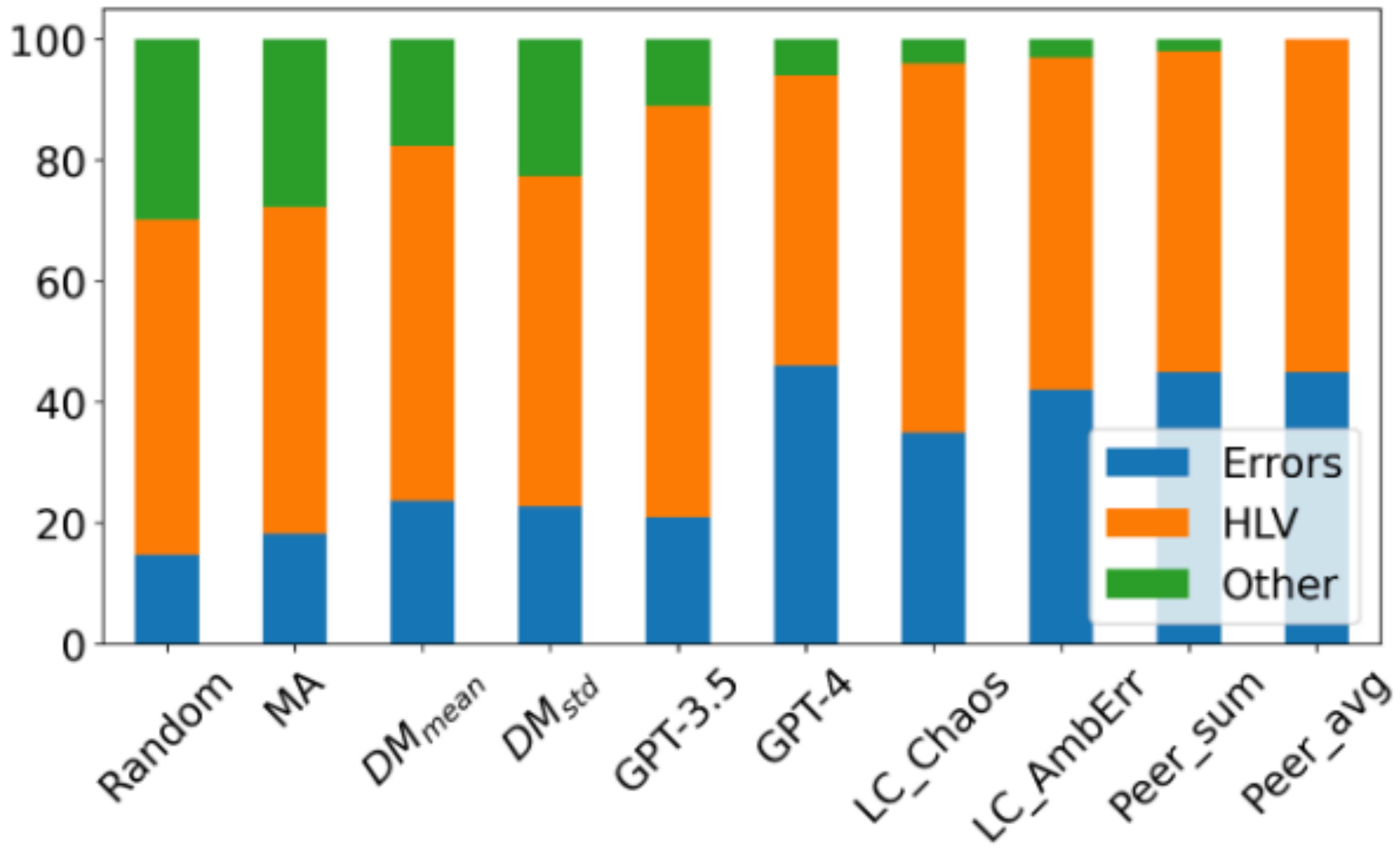
How good is Annotation Error Detection on VariErr?

- Random baseline AP **14.7**
- Data Maps: 22 AP 
- GPTs: GPT-3.5 **17.6**, GPT-4 **31.3**
- Human label count heuristics:
  - 32.5 (ChaosNLI 100 voters)
  - **40.8** (4 voters)



- Human heuristics outperforms GPTs, best with explanations:
- Peer heuristics from VariErr:
  - **46.5** (sum peer-validations)

- Human validation is a strong means to detect errors in data with high HLV
- Heuristics from VariErr performs (4 > 100)
- Analysis: What instances were selected?



Complementarity

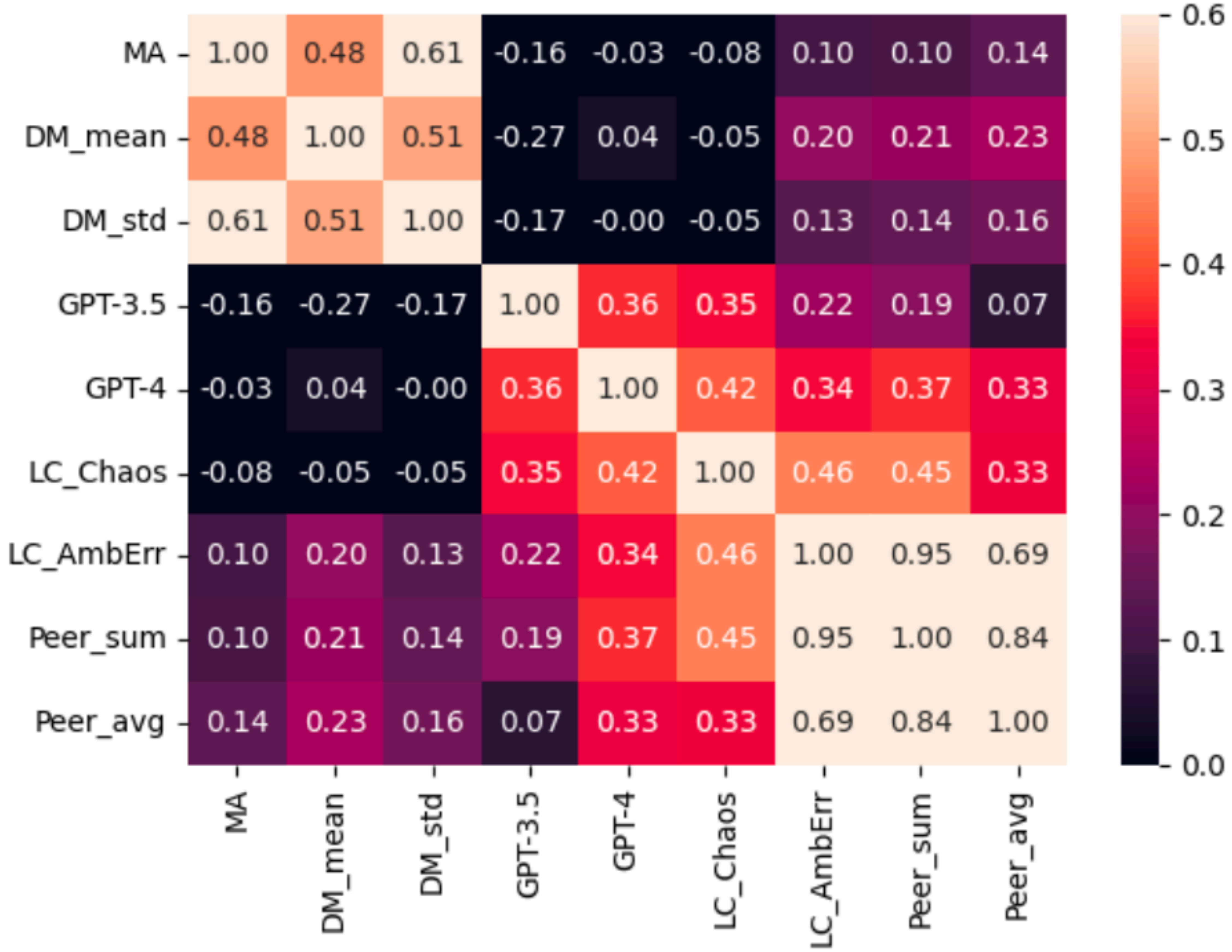


Figure 3: Correlations among scorer predictions.

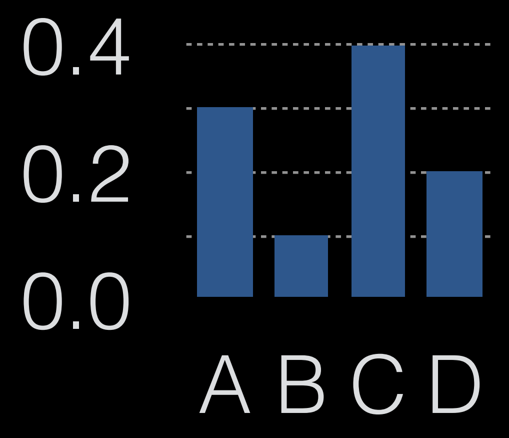
- GPT-4 correlation to LC_chaos is 0.42

Human Label Variation
- many exciting connections -

- high-stake human decision support (e.g. law)
- learning from less but higher quality data
- active learning
- model uncertainty
- statistics and data-generation process
- LLMs that react as humans do
- human values and LLM alignment

Take-home message

✓ **Human label variation** is signal (annotation errors though do exist)



✓ Let's **embrace** it in all stags of the AI pipelines - to not continue to model only the "mode"

✓ HLV will help us develop **trustworthy human-facing AI**

From Human Label Variation and Model Uncertainty to Error Detection (and Back)?

IT UNIVERSITY OF COPENHAGEN

Thanks to my research team, collaborators and funders:

DANMARKS FRIE FORSKNINGSFOND

Munich Center for Machine Learning

http://mainlp.github.io