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

Substitute presenter: Siyao Logan Peng (LMU Munich)

Existing datasets contain annotation errors

Motivation

- Categories exist, but they are fluid
- Not everything is plausible variation.
- Can we tease apart error from plausible human label variation?

Error vs. plausible **Human Label Variation**

Data Quality

Djamé.. @zehavoc · 20h

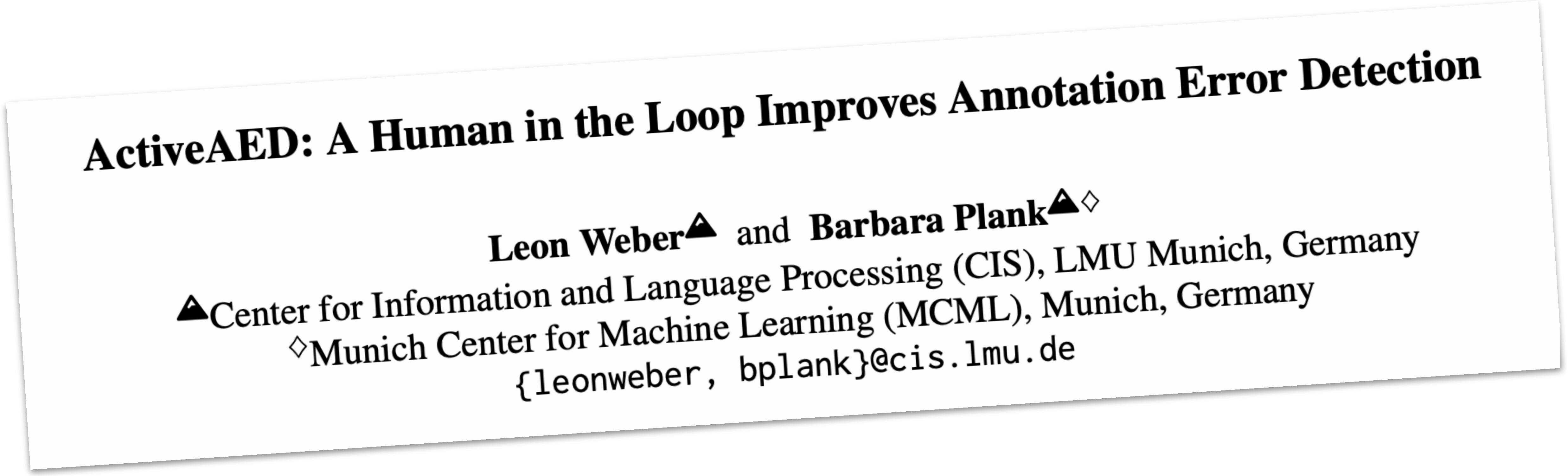
just found out this wonderful quote in an old paper where we described our efforts to parse the British National Corpus (100M words, back then it was huge, clusters and all) work by @Wjrgo @jenfoster, Josef van Genaboth and I
web.stanford.edu/group/cslipubl...

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)

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
- Prior work: we proposed to combine AED with human in the loop for classification tasks: **Active AED**
  - Datamaps
  - Active Learning


ActiveAED: A Human in the Loop Improves Annotation Error Detection

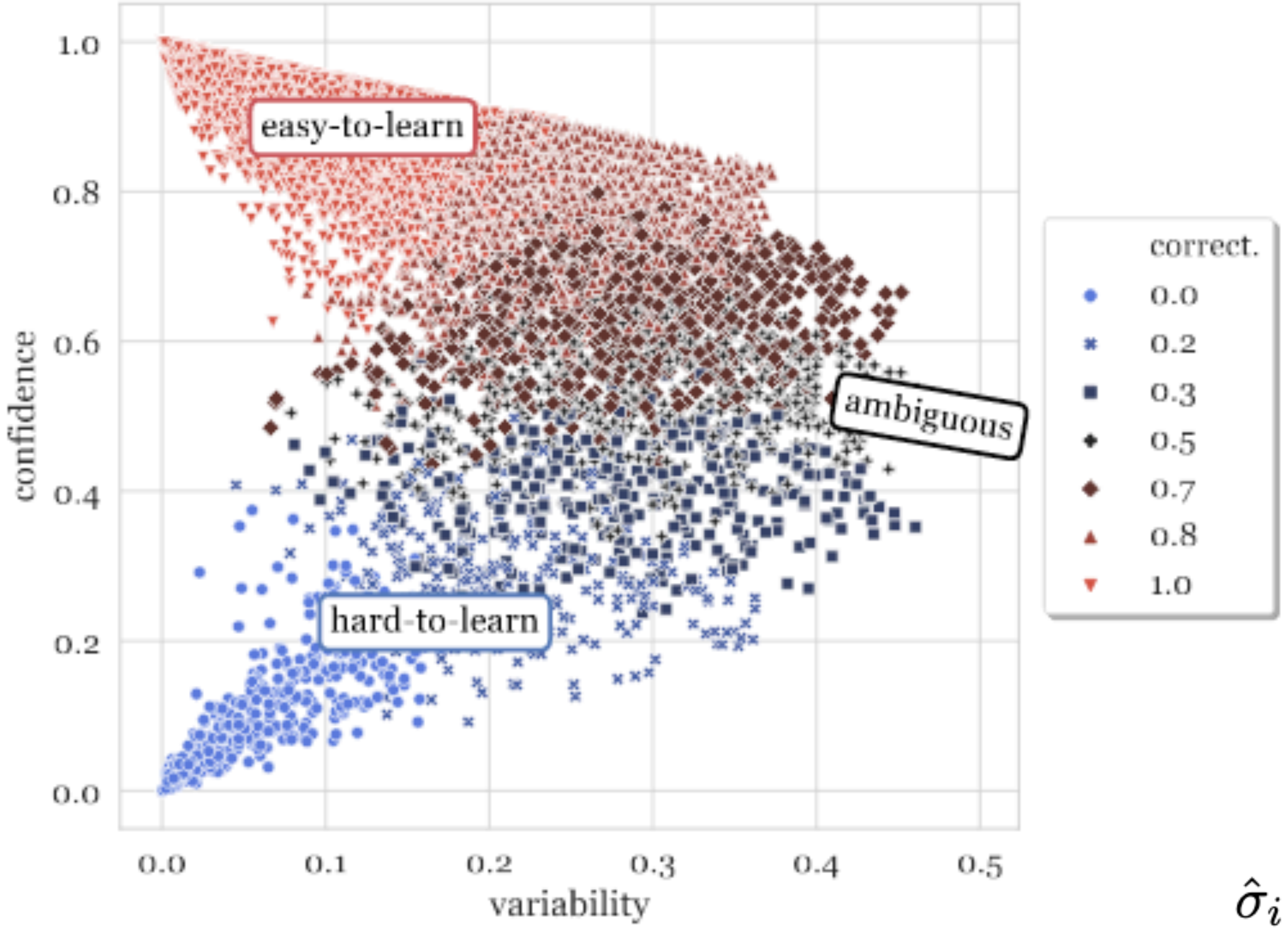
Leon Weber and Barbara Plank

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)

We adapt AED methods from earlier classification tasks (Swayamdipta et al., 2020)

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



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**.

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

(Swayamdipta et al, 2020)

Following earlier work on ActiveAED
Weber & Plank 2023

Our solution: ActiveAED

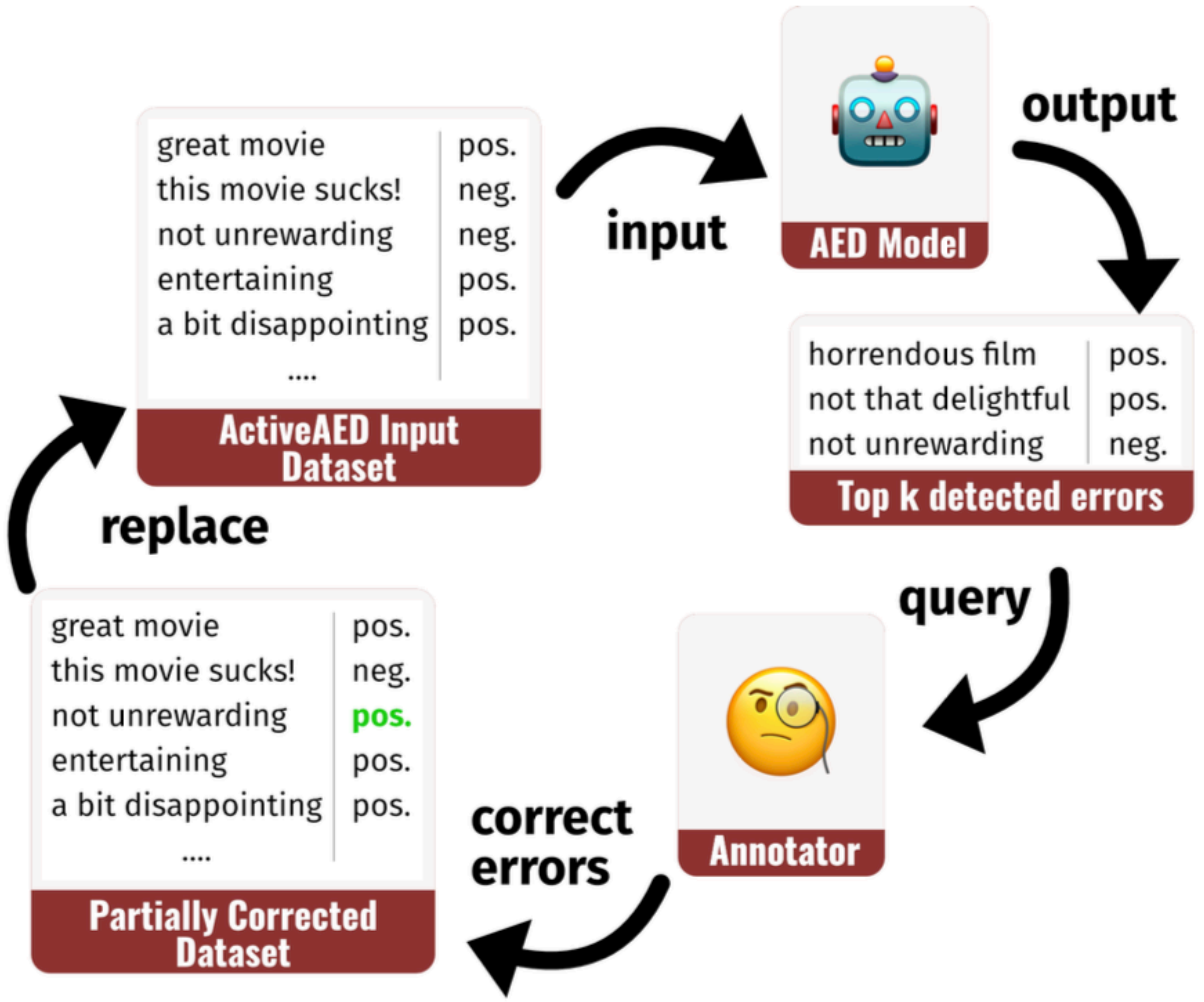
- **ActiveAED: Involve human annotator in pipeline, by repeatedly querying for error corrections**

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

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

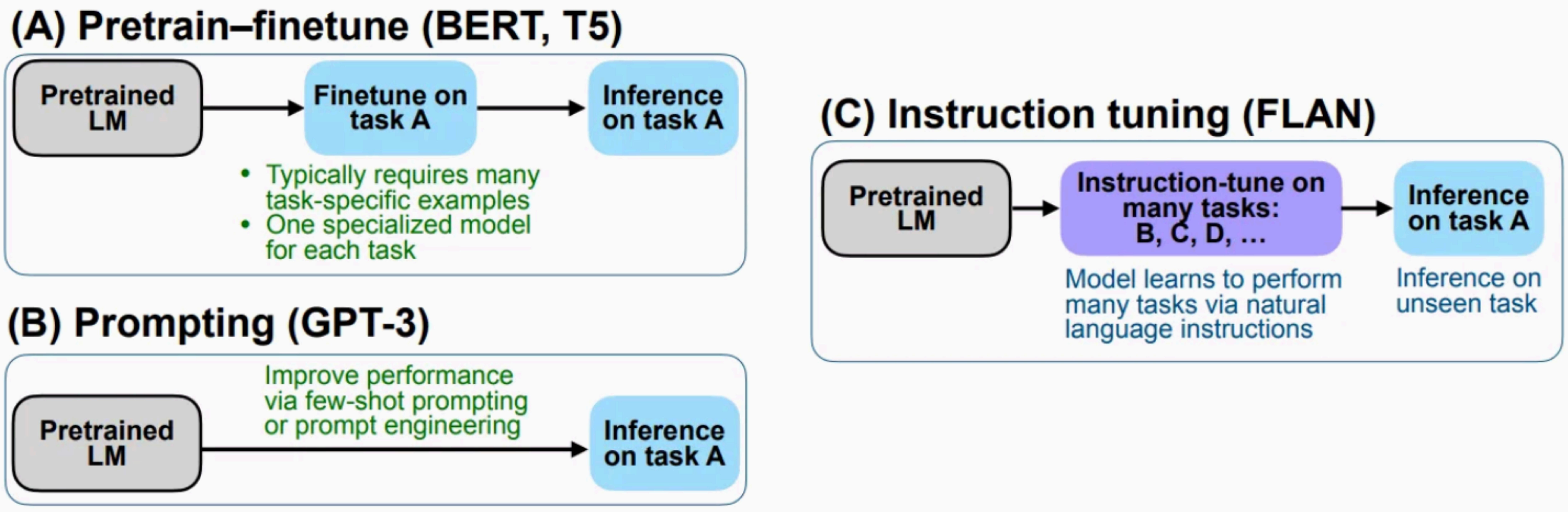
- **Our novel ensembling scheme merges training-dynamics-based and cross-validation-based AED for improved results**

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



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

From Pretrain-finetune to Instruction Tuning



Pretrain-finetuning vs prompting vs instruction tuning (Wei et al., 2022).

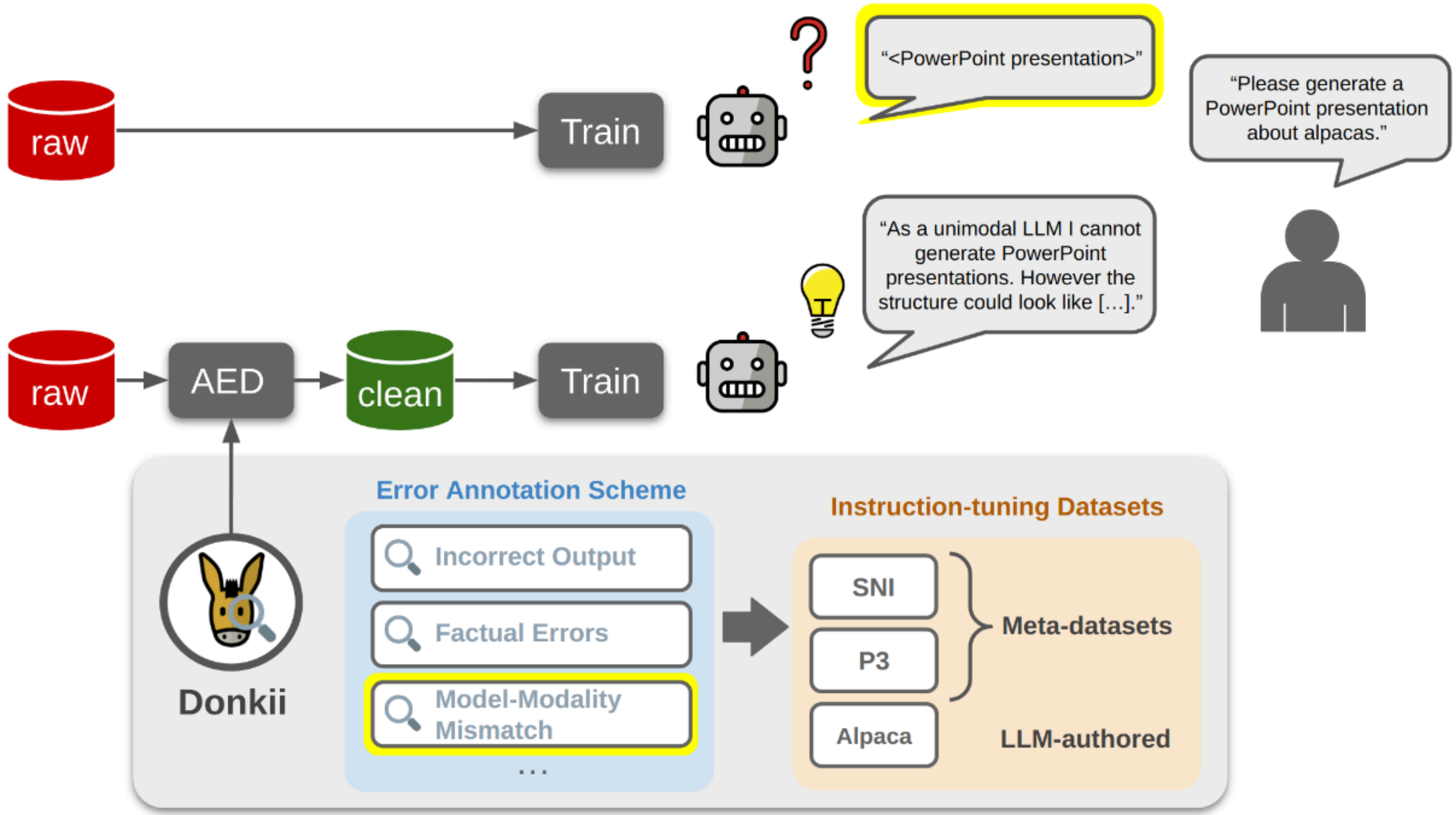
Image credit: https://jasonwei20.github.io/files/FLAN%20talk%20external.pdf

Instruction Tuning & AED

- Finetuning Datasets store input-output pairs in form of instructions.
- Qs: What kind of errors are there? How can we best detect them?

(Self-)Instruction-Tuning Datasets Contain Many Kinds of Tasks

Donkii: Detecting Errors in InstT Datasets

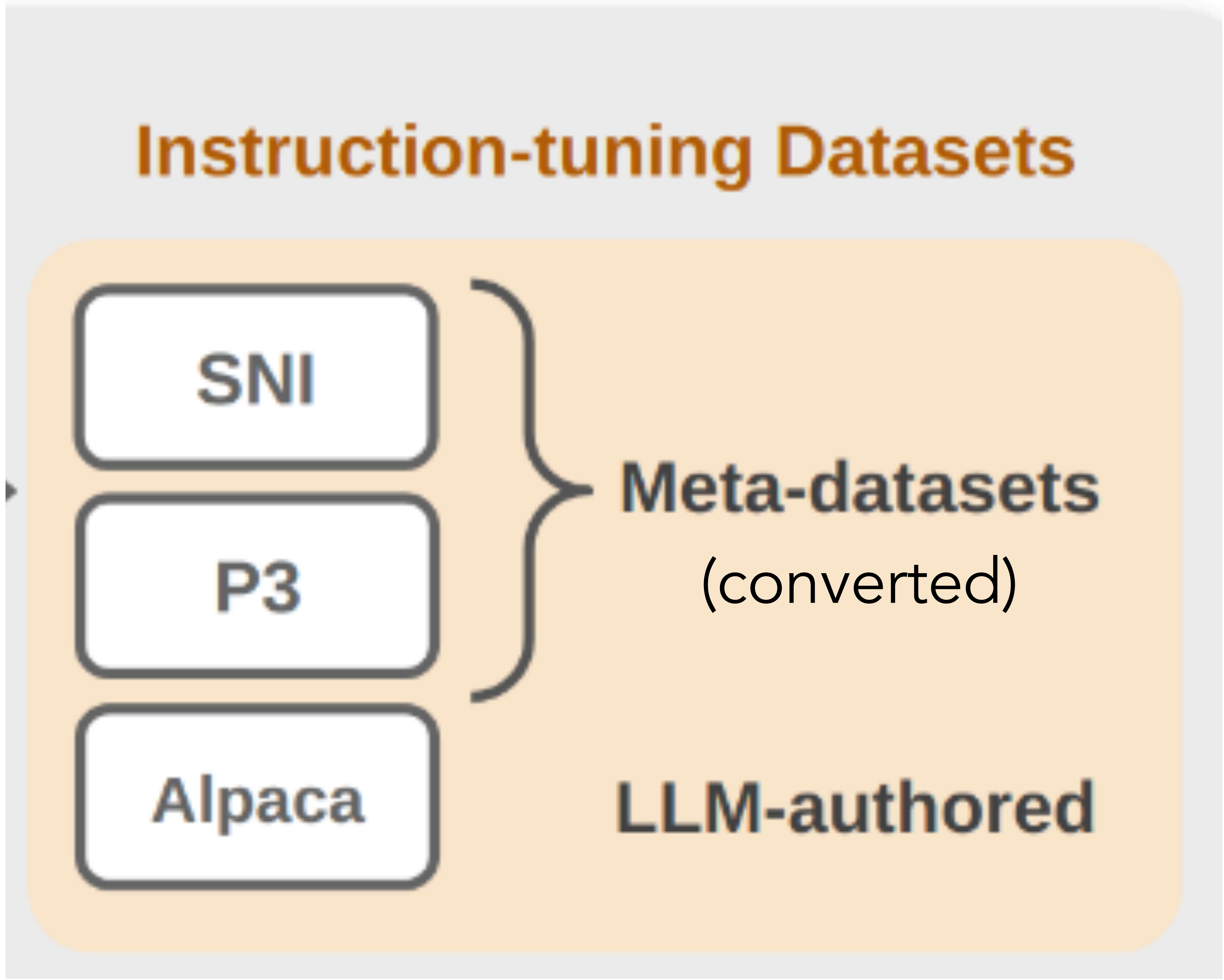


Three kinds of InsT Datasets

SNI: Supernatural Instructions

P3: Public Pool of Prompts

Alpaca: LLM self-instructed

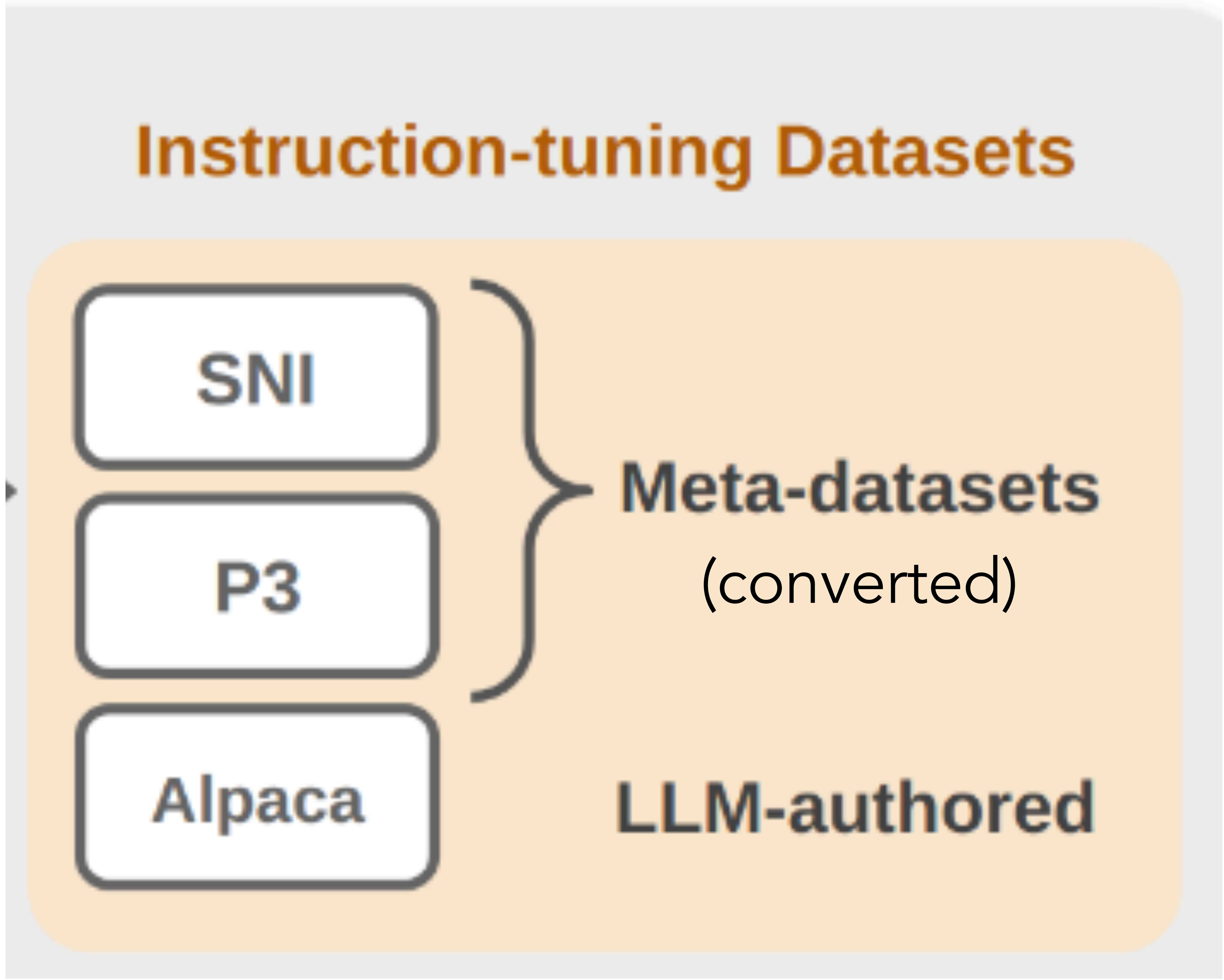


Three kinds of InsT Datasets

SNI: Supernatural Instructions

P3: Public Pool of Prompts

Alpaca: LLM self-instructed



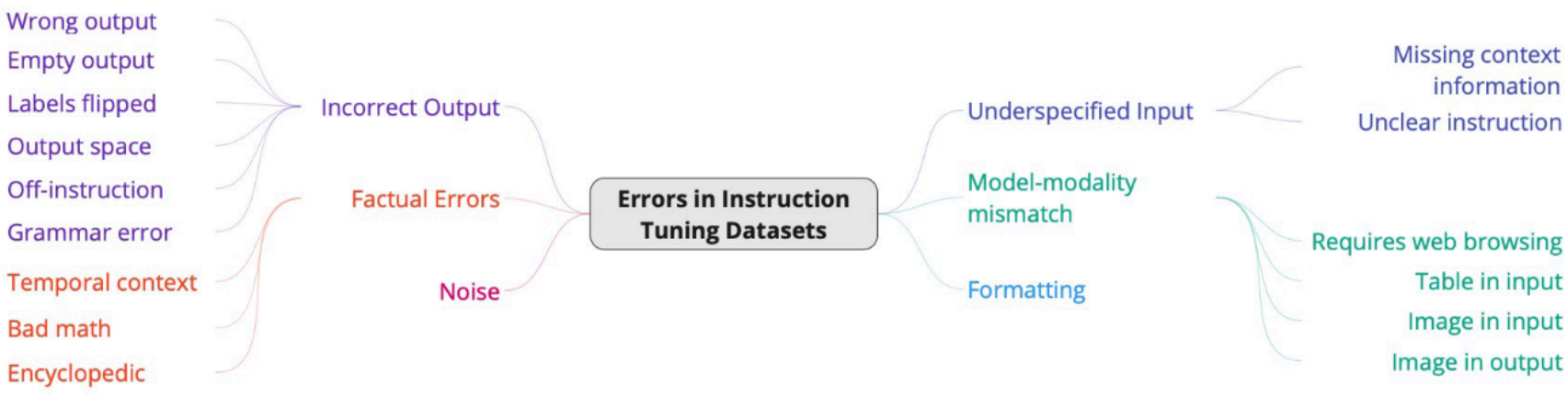
Error Sources:

collect from Github issues

synthetic noise (truncate, flip, gen)

collect from AlpacaDataCleaned

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;

Overview of Donkii Datasets

- Three data sections:
  - $\mathbf{X}_*$ (clean), $\mathbf{X}_{\mathbf{err}}$ (errors), $\mathbf{X}_{\mathbf{unk}}$ (rest, unknown error rate)
- Overview of resulting datasets:

	Source data	$\lvert\mathcal{X}_{\text{unk}}\rvert$	$\lvert\mathcal{X}^*\rvert$	$\lvert\mathcal{X}_{\text{err}}\rvert$	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{T}_{\text{err}}\rvert$	$\bar{L}_{\text{inp}}$	$\bar{L}_{\text{out}}$	Err	Prov
P3	Sanh et al. (2022)	399,472	12,237	12,237	417	20	118	9	Syn.	Meta
SNI	Wang et al. (2022b)	101,783	1,088	585	1,613	17	165	6	Nat.	Meta
ADC	Taori et al. (2023) (Ruebsamen and Contributors, 2023)	48,425	173	146	-	-	15	44	Nat	LLM

Table 1: Statistics for the three Donkii datasets. $\lvert\mathcal{T}\rvert$ denotes the total number of tasks, and $\lvert\mathcal{T}_{\text{err}}\rvert$ the number of tasks with at least one instance with an error. Note, that ADC does not provide a grouping of instances into tasks. $\bar{L}_{\text{inp}}/\bar{L}_{\text{out}}$ denotes the average input/output length in white-space-delimited tokens. 'Err' is the type of error (synthetic or naturally ocurring) and 'Prov' the provenance (meta-dataset vs LLM-authored). 'Lic' is the license under which the authors published their data.

How well does AED do on Instruction Tuning Data?

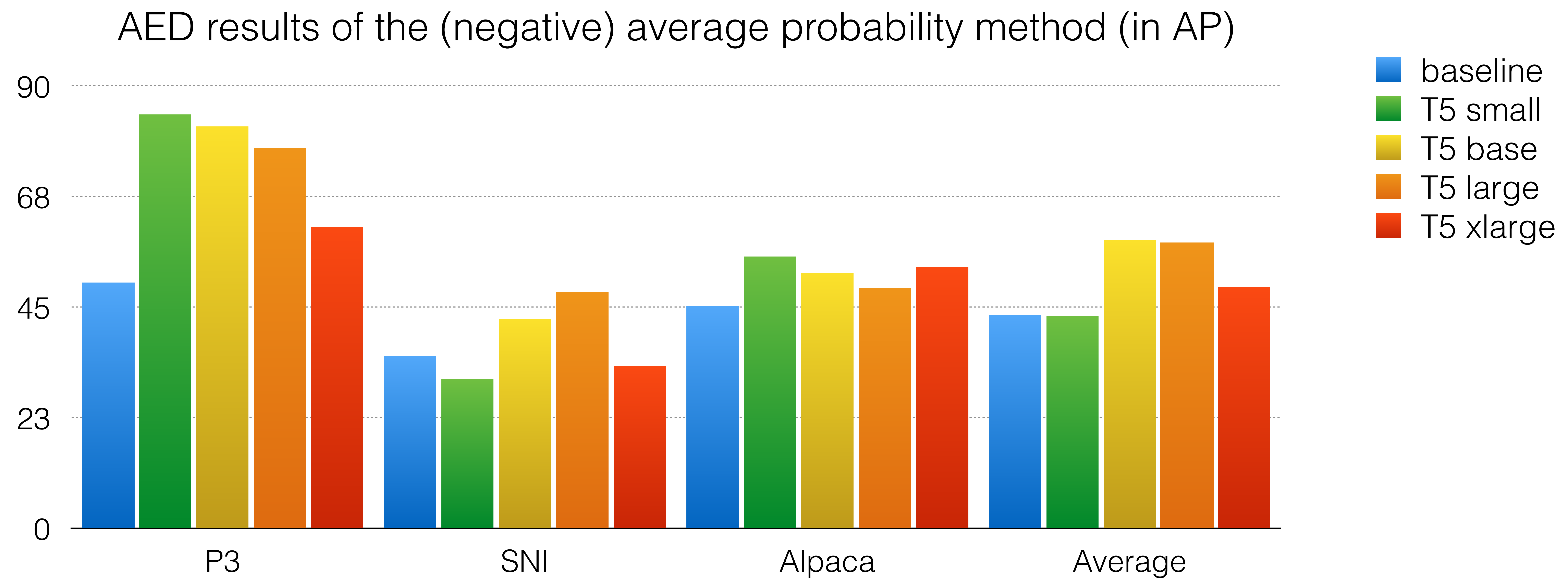
- Follow Klie et al. (2022) and use a ranking (scoring) approach
- Score for each instance (higher score, more likely an error)
- Model and score: T5 models (four sizes, three seeds) and training dynamics (with four different metrics) calculated over E epochs (e.g. avg probabilty, PPL, min prob, AUM) - e.g. **average probability**:

$$P_{\mu} = -\frac{1}{E}\sum_{e=1}^{E}\frac{1}{L}\sum_{l=1}^{L}p_{e,l},$$

- Evaluation metric: AP (average precision)

Results

- Baseline: Proportion of errors estimated from $\mathbf{X}_*$ (clean) and $\mathbf{X}_{\mathbf{err}}$ (errors)
- On average, average prob (Pμ) performed the best (Figure below)
- Perplexity second (see Table 4 in paper for details)



Results per category

The results differ strongly across error categories and dataset.

P3	out (9777)	inp (2460)	-	-	-
rand	50.0	50.0	-	-	-
P_μ	$89.4_{0.9}$	$68.0_{0.1}$	-	-	-
ADC	out (13)	inp (13)	noi (77)	fac (14)	mul (29)
rand	37.0	48.0	48.4	29.8	50.9
P_μ	$62.6_{0.8}$	$72.2_{0.2}$	$49.8_{0.4}$	$55.7_{0.8}$	$61.5_{0.5}$
SNI	out	form (64)	noi (2)	-	mul (3)
rand	38.2	50.0	3.0	-	2.3
P_μ	$51.7_{1.7}$	$51.9_{0.9}$	$30.6_{8.6}$		$14.9_{3.9}$

Table 5: Results per error category. All scores are AP (higher is better) in percent of P_μ using the best performing model size for the dataset. The category names are abbreviated: out: incorrect output, inp: underspecified input, noi: noise, fac: factual error, mul: multi-modality, form: formatting. The number in brackets gives the number of instances per category.

Results

- P3: Synthetically introduced errors are easier to detect
- We recommend to start with a ‘base' sized model for a new InstT dataset

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

In this line of research …

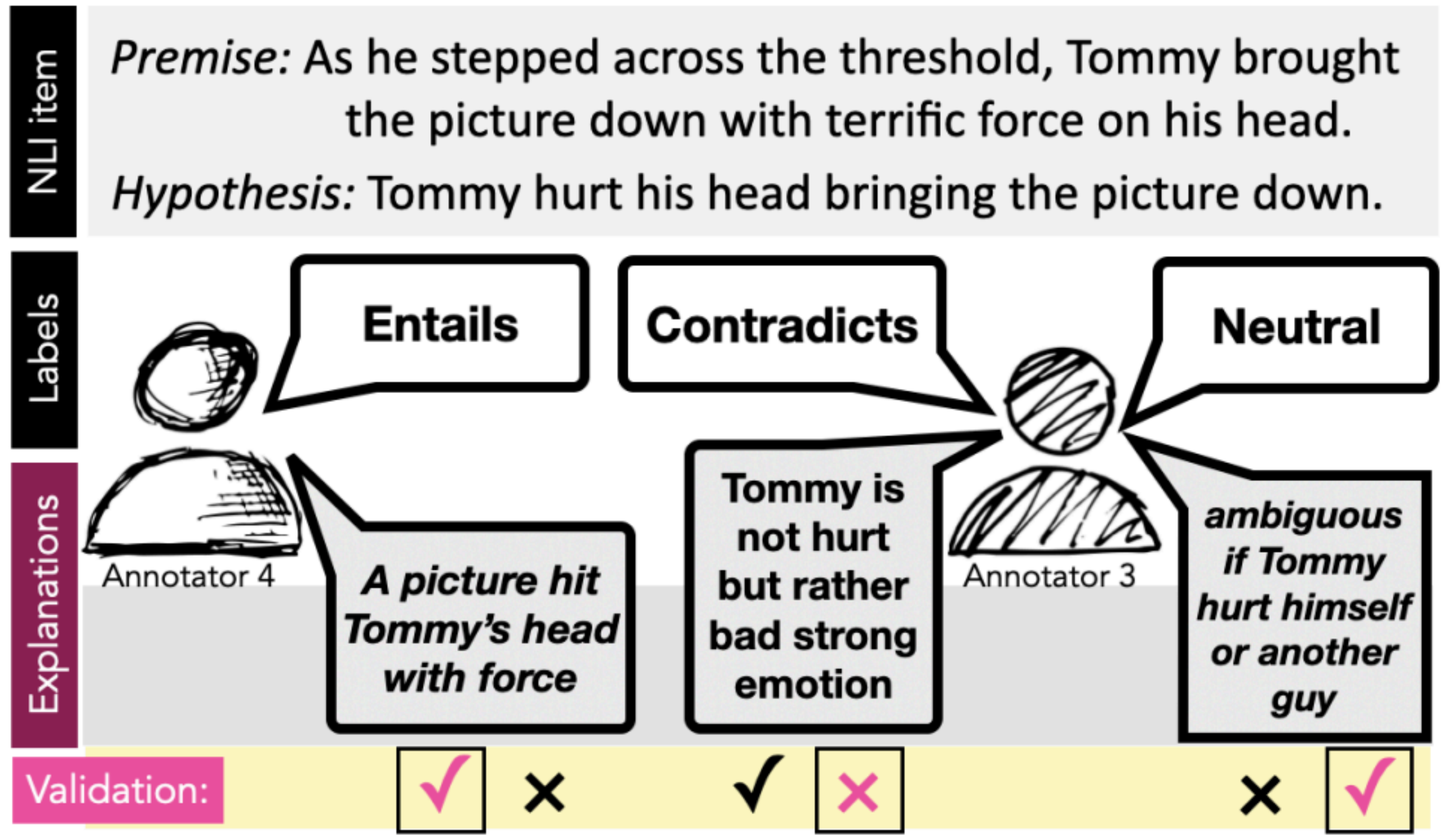


Figure 1: Variation or Error? We present a procedure and multi-label dataset, VARIERR, to tease apart annotation error from plausible human label variation. We leverage *ecologically valid explanations* and *validation* as two key mechanisms (boxed: self-validations; label "Contradicts" is an *error*); see §3-§4 for details.

Questions or Suggestions?

Paper

Github

Qs will be forwarded to Barbara Plank
b.plank@lmu.de