“All that glitters”: Techniques for Evaluating Validity, Bias, Fairness, and Helpfulness with Unreliable Labels

Case Data: Ratings of Teaching Quality

Michael Hardy, Stanford University

Humans are Unreliable Annotators

What to do about pyrite in our “gold” labels?

- **Human annotations** always some amount of **error**,
- **Disentangling** individual human rater biases and other **sources of variation** (**Inputs**, **Criteria**, and **Raters**) can improve model and dataset evaluation

Case Study: Rating Teaching Quality

Classroom Observations and Teacher Support

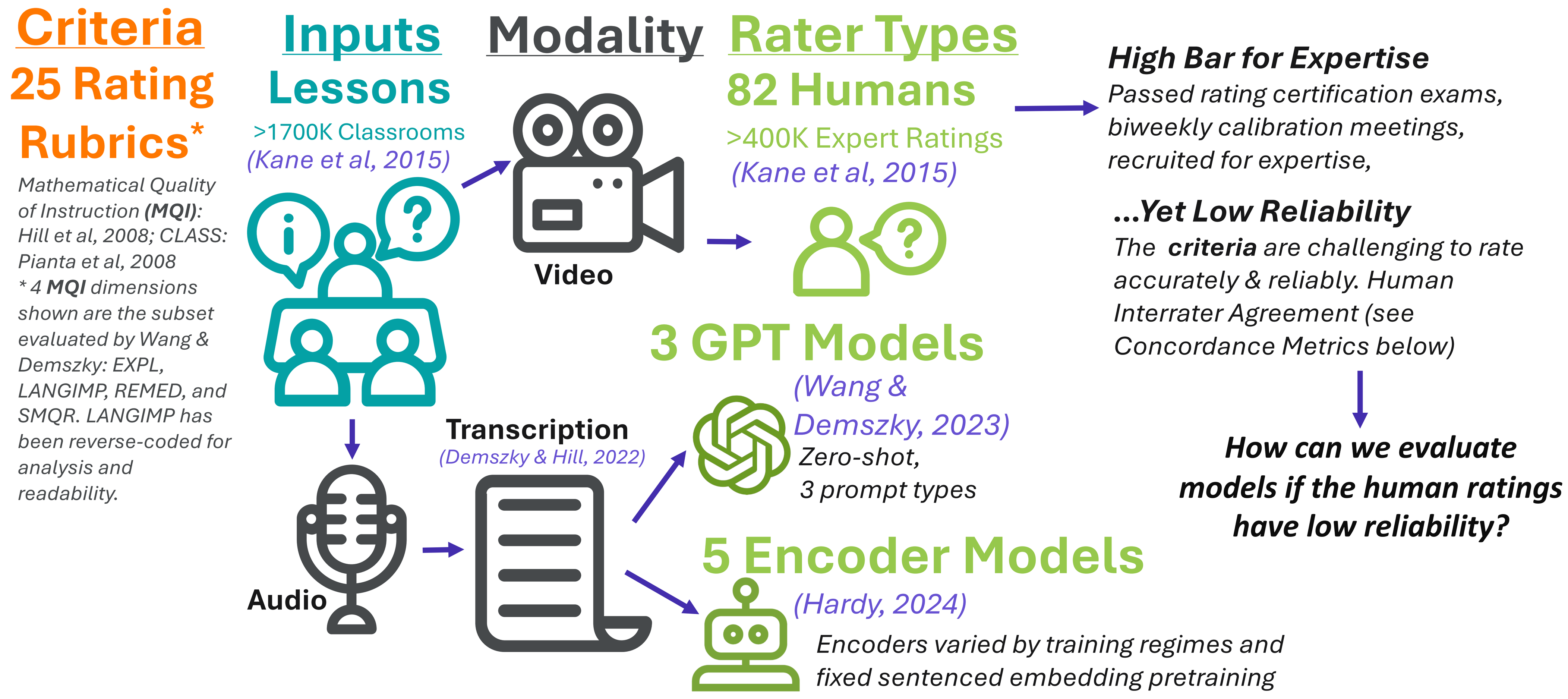
- **Universal**, **time consuming**, **high-stakes,** in person
- Paradoxically **developmental** & evaluative
- Key step in coaching teachers: **diagnosing needs**
- **Low reliability** even with **expert human raters,** due to **complexity** & many sources of **variation** (**teachers**, **lessons**, **students**, **raters**, **rubrics**, etc.)

Can automated ratings of classroom instruction improve human rating quality?

Evaluating with Low Quality Labels

More robust methods for measuring **validity**, **bias**, **fairness**, & **helpfulness** can **improve model development, data** and **evaluation** in the absence of “gold” labels.

Data & Design: Human and LLM Raters



Evaluation Techniques

Application: Aspects of Instructional Quality

Concordance Metrics

Typical Eval Methods

1. **SOTA** (even *“super-human”*) Encoder Model **Performance!**
2. **Low** “gold/ground truth” human **Reliabilities?**

*So, are SOTA results **good!?***

Methods Details

Interrater Reliability (IRR)
- **C’ κ**: Cohen’s κ
- **QWK**: Quadratic Weighted κ
- **%Agr**: % exact agreement
- **Agr±1**: % agreement w/in 1 rating category

Annotation Group Correlations
- **ICC**: Intraclass correlation coef.
- **AICC**: Adjusted ICC

Pointwise Correlations
- ρ : Linear/Pearson’s ρ
- r_s: Rank/Spearman’s r_s

Teacher Explanations

Metric	Human	Encoder	GPT
C’s κ	0.23	**0.27**	0.05
QWK	0.27	**0.44**	0.04
%Agr	0.7	**0.71**	0.32
Agr±1	**0.98**	0.97	0.86
ICC	0.15	**0.16**	0.16
AICC	0.52	**0.54**	0.53
ρ	0.27	**0.45**	0.07
r_s	0.26	**0.43**	0.07
$E\rho^2$	**0.15**	**0.15**	0.08
Φ	**0.12**	**0.14**	0.08
ϱ_{hm} (95%CI)		**0.85** (0.7, 1.0)	-0.18 **(-0.7, 0.3)**

Precision of Language

Metric	Human	Encoder	GPT
C’s κ	**0.25**	0.2	0.0
QWK	0.29	**0.34**	0.0
%Agr	**0.8**	**0.8**	0.31
Agr±1	**0.99**	0.98	0.98
ICC	**0.12**	**0.12**	**0.12**
AICC	0.45	**0.46**	0.45
ρ	0.29	**0.34**	0.01
r_s	0.28	**0.3**	0.0
$E\rho^2$	0.09	**0.15**	0.08
Φ	0.08	**0.14**	0.08
ϱ_{hm} (95%CI)		**0.85** (0.7, 1.0)	-0.18 **(-0.7, 0.3)**

Remediating Student Errors

Metric	Human	Encoder	GPT
C’s κ	0.26	**0.28**	0.0
QWK	0.32	**0.41**	0.0
%Agr	0.66	**0.68**	0.15
Agr±1	0.96	**0.97**	0.58
ICC	0.14	**0.15**	0.13
AICC	0.49	**0.52**	0.48
ρ	0.32	**0.41**	-0.01
r_s	0.32	**0.4**	0.0
$E\rho^2$	**0.13**	0.10	0.05
Φ	**0.11**	0.09	0.04
ϱ_{hm} (95%CI)		**0.95** (0.7,1.0†)	0.06 **(-0.5, 0.7)**

Student Questions and Reasoning

Metric	Human	Encoder	GPT
C’s κ	0.24	**0.26**	0.04
QWK	0.3	**0.36**	0.08
%Agr	**0.76**	**0.76**	0.39
Agr±1	0.98	**0.99**	0.91
ICC	0.18	**0.2**	0.2
AICC	0.57	**0.59**	**0.59**
ρ	0.3	**0.36**	0.14
r_s	0.29	**0.34**	0.12
$E\rho^2$	**0.14**	0.09	**0.0**
Φ	**0.13**	0.09	**0.0**
ϱ_{hm} (95%CI)		**1.0†** (1.0†, 1.0†)	**0.0** (0.0, 0.0)

Confidence & Validity

Generalizability of Ratings

1. Some **SOTA** correlations are **spurious**
2. “Super-human” **reliability doubtful**

*Eval metric choice can **outweigh models***

Methods Details

ϱ_{hm} = 0 suggests model-human correlations were not based on same construct.

Generalizability Studies to Deconstruct Sources of Variation

Generalizability Study estimating annotation quality across sources of variance where $E\rho^2$ and Φ are estimated using a rater by observation-within-individual teacher design, $R \times (O{:}I)$. Φ is a measure of “**dependability**”: in this case, it represents the extent to which the numeric **ratings assigned would persist under changing sources of variation** (e.g., same teacher, different lesson/day)

Disattenuation to Detect Spurious Correlations

Correct some impact of **measurement error**: if latent teacher construct is roughly the same across multiple lessons, use model-human correlation on complementary lessons, scaled by dependability:

$$\varrho_{hm} = \frac{\mathrm{Corr}[\mathrm{Score}(i, lesson = \mathfrak{L}, r_{human}), \mathrm{Score}(i, lesson \neq \mathfrak{L}, r_{model})]}{\sqrt{\Phi_h \Phi_m}}$$

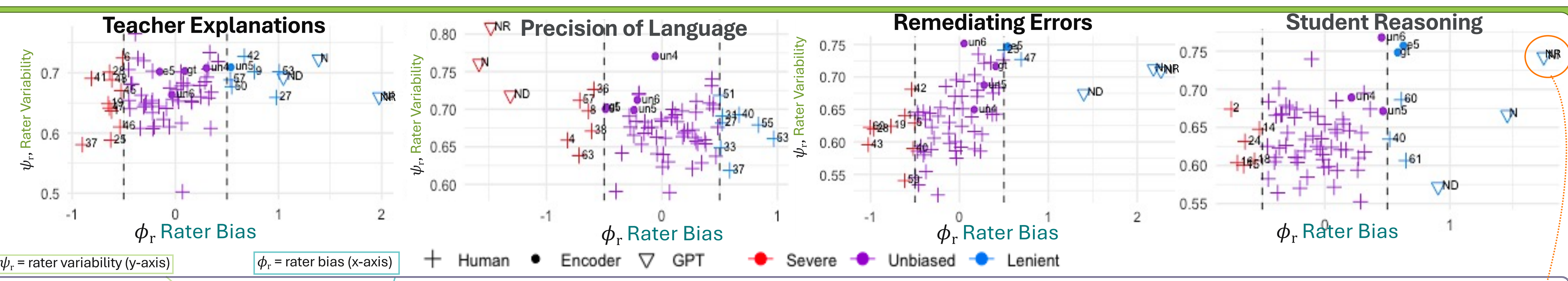
† Reported disattenuated correlations of 1.0 do not mean perfect correlation: it generally means that *measurement error is not randomly distributed.*

Disentangling Rater Biases

Rater Effect Models for De-biasing

1. Accounting for individual rater biases can help **estimate “gold” labels** in training or eval

Annotator ids** with can be used to estimate **individual rater biases and behaviors



Methods Details

Disentangling rater effects with multidimensional hierarchical rater models (MHRM) having a rater response Signal Detection Theory (SDT) first stage and a generalized partial credit Item Response Theory (IRT) second stage which estimates the latent “gold” label, ξ_{tij}

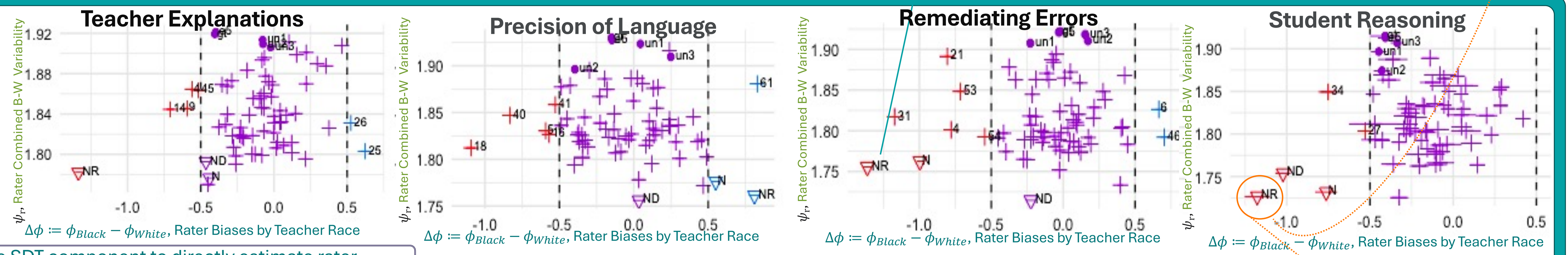
$$\text{MHRM}\begin{cases}\theta_i \sim \mathcal{N}(\mu, \sigma^2), \text{ distribution of latent "true" scores} \\ \xi_{tij} \sim P[\xi_{ij} = \xi \mid \theta'_i, \alpha_j, \gamma_{j\xi}] = \frac{\exp\{(k-1)\alpha_j\theta'_i - \sum_{k=1}^{k}\gamma_{jk}\}}{\sum_{h=1}^{K_j}\exp\{(k-1)\alpha_j\theta'_i - \sum_{k=1}^{h}\gamma_{jk}\}} \text{ (IRT model)} \\ X_{tijr} \sim P[X_{ijr} = k \mid \xi_{ij} = \xi] \propto \exp\left\{-\frac{1}{2\psi_r^2}[k - (\xi + \phi_r)]^2\right\} \text{ (SDT model)}\end{cases}$$

Rater Effect Plots: Each point is an individual rater: a “+” marker is a single human rater; “•” and “▽” are specific encoder and GPT models, respectively. X-axis is rater bias (ϕ above and $\Delta\phi \coloneqq \phi_{Black} - \phi_{White}$ for fairness plots below). **Right is more lenient, left more severe**. Color (via x-axis) are bias categories. Y-axis is rater variability (lower is more consistent. Horizontal lines 95% CI for bias via MCMC Bayes Estimation

Fairness across Racial Lines

Independence of Teacher Race

***GPT** models mostly show **negative bias against Black teachers** relative to White teachers*



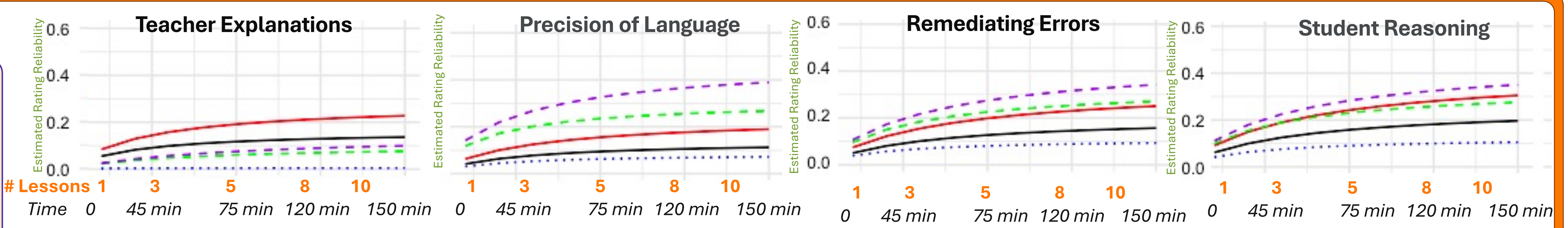
Methods Details The MHRM is extended to include teacher race covariate in the SDT component to directly estimate rater variability based on teacher behaviors. The errant categorical centrality tendencies of GPT “reasoning” models are measurably reduced for Black teachers: could they be receiving more *accurate* ratings from the models?

Helpfulness of Ratings

Human-in-Loop Decision Studies

1. **Encoder Models improve human label quality at least as much as another human f**or most items.
2. **GPT models did not positively impact** human label reliability for any of these items.

Humans most in need of label support may be most susceptible to confident GPT misguidance



Methods Details Decision Studies manipulate variance components from G-studies (e.g., how many raters, how long, model vs human, same vs different lesson). Based on low reliabilities, only human-in-the-loop (HIL) scenarios for models are displayed. All estimations used more conservative absolute error calculations. The x-axis above shows number of classroom observations and human time for each 15 min observation, for various changes to how classroom observations are conducted.

Changes to Classroom Observations

- **Baseline Human** observes various 15 mins of segments classes of same teacher
- **Additional Human** observes 15 min of *different* classes of same teacher
- **Single Encoder** observes entire class w/ 15 min HIL
- **Encoder ensemble** observes entire class w/ 15 min, HIL
- **GPT ensemble** observes entire class w/ 15 min, HIL