CS 293/EDUC 473

Validation

Announcements & reminders

- Thanks for being so engaged! Keep up the great work!
- Pushed A2 and A3 deadlines by a week; final paper deadline remains the same
- Will try to get you feedback on A1 by Wednesday morning

Today’s class

- Debrief from guest visits
- Validation-related activity
- Tips
- Paper discussion led by Andrew, Avi and Zane

How did you feel after Dan Meyer's lecture?

What were some "aha" moments for you--within and across the three visits?

Nobody has responded yet.

Hang tight! Responses are coming in.

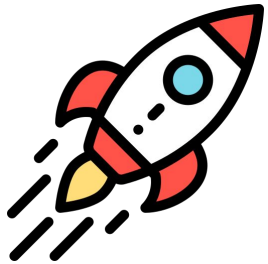
Where are we?

Identify Problem

Data Exploration

Algorithm Development & Validation

Tool Development



Deployment

Overarching Themes:

Bias & Fairness

Working closely with teachers

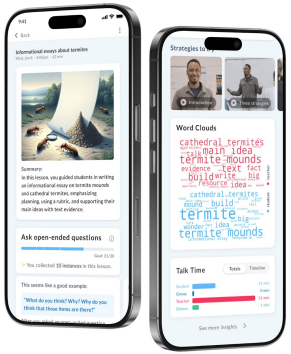
Evals are all you need.

Modeling decisions, tool development, iteration follow from your evaluation pipeline.

Three use cases = three teams

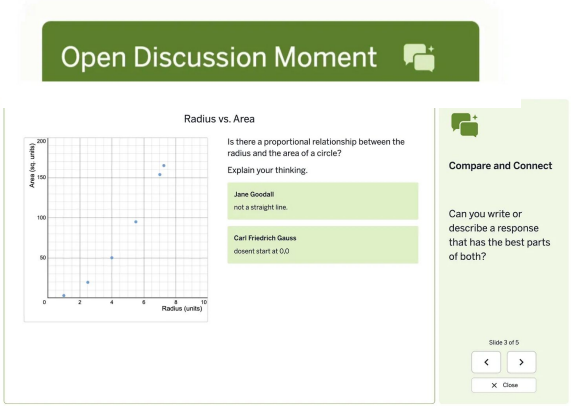
TeachFX

TALK MOVES

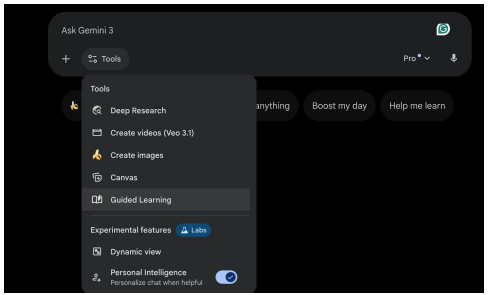


Teacher feedback tool

Amplify Desmos Math



Discussion facilitation tool for teachers

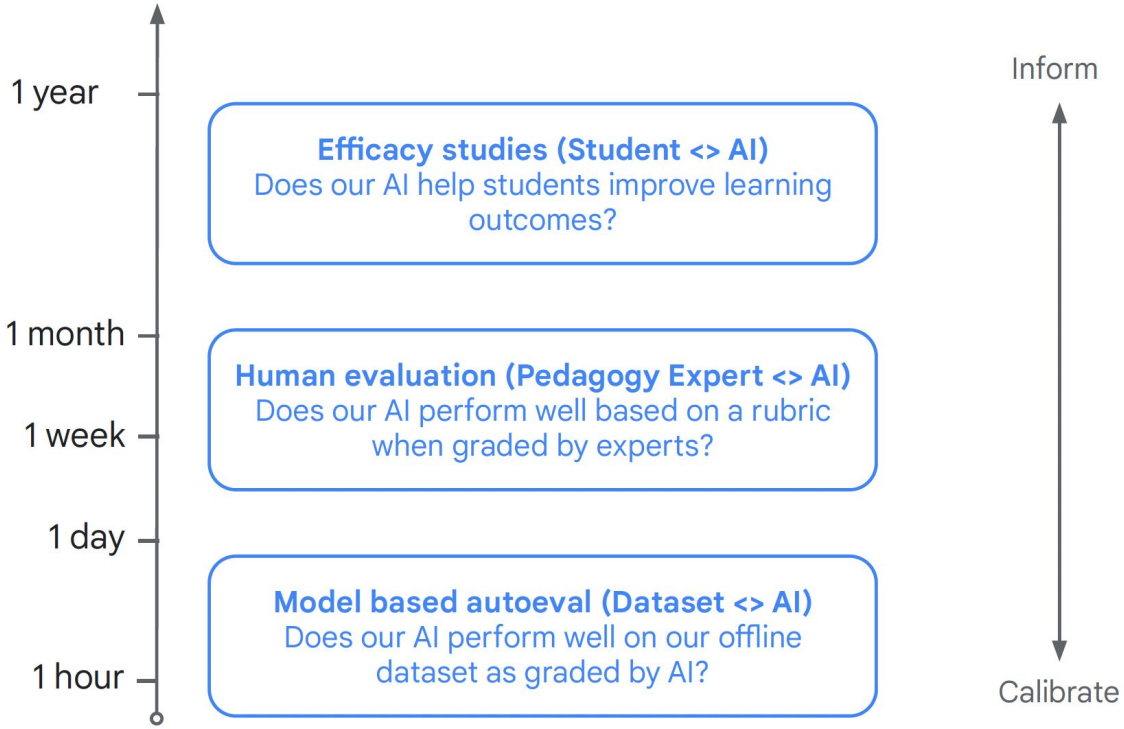


AI chatbot for students

On whiteboard, based on <u>tool developers'</u> POV:

1. *Our tool's input is…*
2. *Our tool's output is…*
3. *The output will be used to…*
4. *We know it is **valid** if…*
5. *We know it is **useful** if…*
6. *We know it is **fair** if…*

From LearnLM:



From a user's point of view:

1. *We know it is **valid** if…*
2. *We know it is **useful** if…*
3. *We know it is **fair** if…*

your birthday 🎂 :

Jan-March

April-June

July-Sep

Oct-Dec

In your project groups, consider:

1. *My construct is…*

2. *My model output will be used to…*

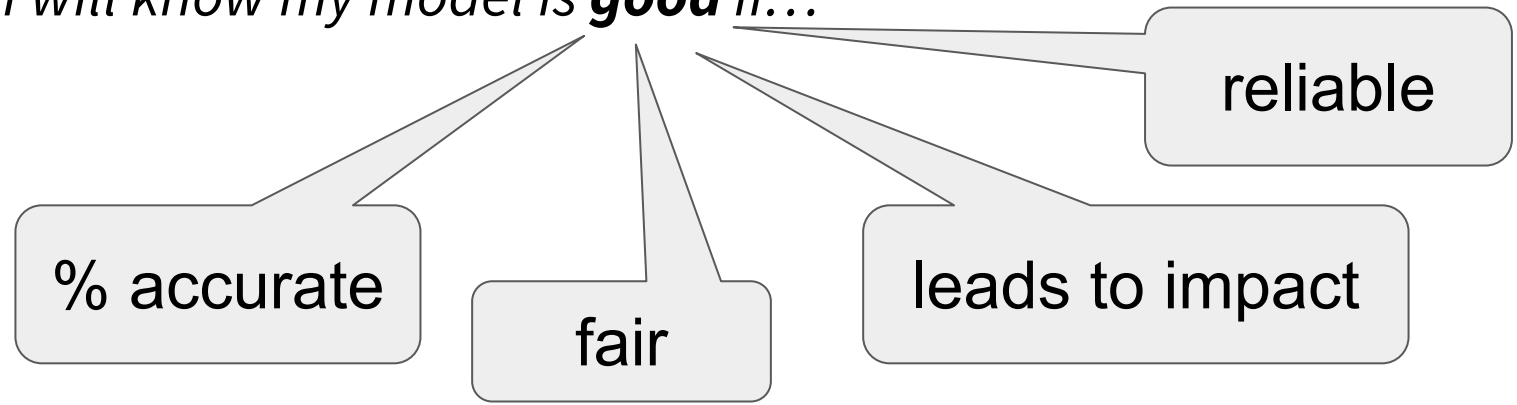
3. *I will know my model is good if…*

In your project groups, write on a whiteboard:

1. *My construct is…*

2. *My model output will be used to…*

3. *I will know my model is **good** if…*



💡 Quick tips

- Make sure your final test set is **truly held out**
  - i.e. not included in the training data; not even using it to improve prompts
- If IRR is low:
  - redefine construct(s) – can they be more specific without overfitting/reducing coverage?
  - continue rater calibration
  - can you get at a similar construct with a different kind of task?
- If model performance is low:
  - try different data pre-processing methods
    - consider the unitization and prevalence problems!
  - iterate on prompts (if using LLMs)
  - if labeled dataset is large enough, try SFT with smaller LMs (RoBERTa, BERT)
  - rethink construct

Dataset specific tips

- NCTE: splitting long utterances into sentences
- Coteach: repetitive queries; decide unit of analysis (thread vs query)
- Mathfish: split long multipart problems
- Persuade: break down essays into component parts

Mostly all:

- Use a context window!
  - context context context context [TARGET] *the segment I really care about* [/TARGET] context context context context
  - build a segmenter model first! (BIO tagging or generate text with tags)
- Mask out variable content
  - Numbers in NCTE/mathfish, “landform” and “Mars” in Persuade, student/teacher monikers in NCTE

Edu-ConvoKit

Dataset

Speaker	Text
Tutor	Alice, what units would we use to measure speed?
Student	Miles per hour.
…	…

e.g., tutoring or classroom conversation

A Pre-Processing

Speaker	Text
Tutor	[STUDENT] what units would we use to measure speed?
Student	Miles per hour.
…	…

e.g., anonymization, grouping utterances

B Annotation

Speaker	Text	Feature
Tutor	[STUDENT], what units would we use to measure speed?	
Student	Miles per hour.	
…	…	

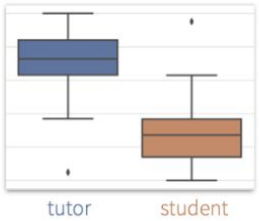
e.g., talk time, student reasoning, teacher talk moves

C Analysis

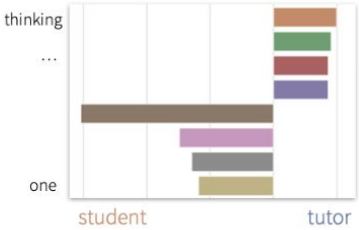
Qualitative e.g., student reasoning examples

> Student: Because, um, because, let's say at this point you're not traveling at fifteen miles an hour […]

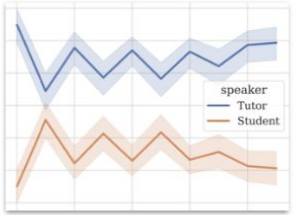
Quantitative e.g., talk time



Lexical e.g., log odds



Temporal e.g., talk over time



GPT e.g., summary of conversation

> 🤖: The tutor asked the student to determine which is larger, one half or two thirds, and by how much. The students used different colored rods to represent fractions […]

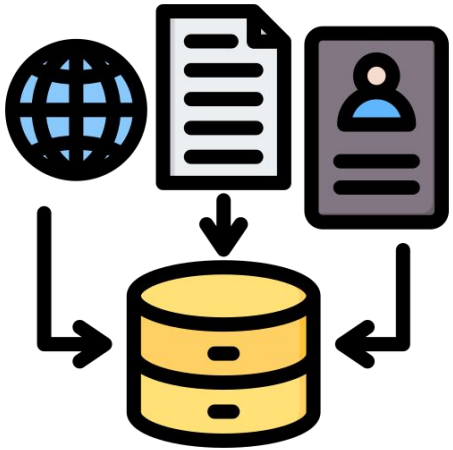
Reading discussion by Andrew, Avi and Zane

Extra slides

How do we define measurement?

Data

Structured (e.g. likert scale responses) / Unstructured (e.g. language)

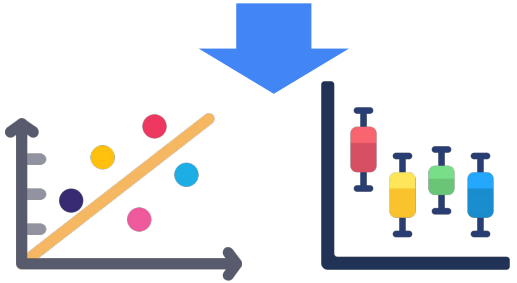


Score / Label

Measuring a target construct

Binary: Is this utterance on task?
Continuous: To what extent does the student feel empowered in this classroom?
Categorical: What is the topic of this lesson?

You can use these in quantitative analyses:



Most aspects of a quantitative research project / intervention / tool require measurement

- Identifying & analyzing teaching practices
- Evaluating fairness & bias
- Identifying need for intervention
- Understanding teachers' and students' perceptions of a tool
- Measuring outcomes
- …

None of these are trivial to measure

Most aspects of a quantitative research project / intervention / tool require measurement

- Identifying & analyzing teaching practices
- Evaluating fairness & bias
- Identifying need for intervention
- Understanding teachers' and students' perceptions of a tool
- Measuring outcomes (e.g. student learning)
- …

Ex. Challenges and Issues

Subjectivity and context-dependence

Sparse data & oversimplification of demographic categories

High stakes & prone to bias

Low response rate & self-reporting bias

Choice of outcomes are often the most controversial

Do you need to identify the measurement target?

E.g., type of classroom practice, dimension for user attitude (e.g. difficulty).

No: Skip to next step

Yes: Pick the target at the intersection of **promise & feasibility**

- Lit review
- Talk to people
- Look at data

Example brainstorming spreadsheet (list of discourse practices relevant to math ed)

Category	Feature Associated with Better Learning Outcomes	What to Measure	Examples
General pedagogical	student participation	number of words uttered by students/minute; and/or length of student utterances	
General pedagogical	test-orientedness	measure the number of references to standardized testing	MCAS, DCCAS
General pedagogical	instructional time	amount of instructional time vs off task time	procedural talk vs instructional talk; noise in the classroom
General pedagogical	wait time after questions	amount of wait time after questions	
General pedagogical	check-in on students	number of times teacher asks questions that check in to see if students are following along	“make sense?” “Any questions”? Thumbs up? Hold your boards up; “student .. looks puzzled”
Math specific	use of math terms	density of math terms from teachers & students; measure to what extent teachers press students to use such terms	angle, fraction
Math specific	use of sloppy (math) terms	density of sloppy terms from teachers and from students	borrowing, top and bottom, cancelling
Math specific	use of proofs, mathematical reasoning/explanation	presence of proofs, mathematical reasoning/explanation in teacher & student talk	
Math specific	degree of direct instruction & focus on memorization	estimate the degree to which the teacher is doing direct instruction	teacher talk with short student answers interspersed; words like remember, recall; first thing you do when you...what do you do next..we’re also going to have to do what?
Math specific	cognitive demand of (math) questions	estimate the degree of cognitive demand of questions (teachers & students)	why, explain, what does that mean, different, difference, compare, what’s missing, how do these relate
Math specific	teachers' evaluation of student contributions	see whether and how the teacher remediates students' misunderstandings;	correction, reformulation, repetition, praise
Math specific	uptake	degree to which teacher uses students' mathematical contributions in subsequent instruction; students' uptake of other students' ideas (with minimal teacher orchestration)	
Classroom climate	positive references to students	degree to which teacher uses student names in a positive way	positive: “Geoffrey’s idea” or “Marie, tell us what you are thinking” “I think Nonie solved the problem in the same way”; negative: Geoffrey!
Classroom climate	affirmation of knowledge and skill	degree to which teacher encourages students	“You totally understand this, you just need to tweak what you’re saying a little bit”
Classroom climate	broad regard	degree to which the teacher shows interest in the students' lives	asking non-academic questions

Does an NLP measure exist already for what you want to do?

Yes: Skip to validation (on your domain)

No: **Develop a measure** (in most cases) following the standard paradigm → next slide

Standard NLP measure development workflow

1. **Create high quality validation set**
   a. With sufficient # of examples to capture relevant variation (rule of thumb: at least 1k examples for a relatively straightforward measure, 2k for more subjective ones)
   b. When possible, create a held-out test set too (that you only evaluate on at the very end)
2. **Iteratively develop & validate model**
   a. Supervised paradigm: label training data → train classification/regression model
   b. Unsupervised/self-supervised paradigm: leverage unlabeled data

Standard NLP measure development workflow

1. **Create high quality validation set**
    a. With sufficient # of examples to capture relevant variation (rule of thumb: at least 1k examples for a relatively straightforward measure, 2k for more subjective ones)
    b. When possible, create a held-out test set too (that you only evaluate on at the very end)

what if creating a validation set is not at all trivial because the construct is highly subjective?

What to do if your **interrater agreement is fair to moderate**?

Even when working with domain experts & doing several rounds of rater training and discussion

First ask: why is agreement low?

Potential cause	Potential solutions
Poorly defined construct	• Improve definition & coding scheme! • Revise input level of granularity (e.g. sentence vs utterance vs segment)
Context-dependence of construct	• (When possible) Add more context • (When appropriate) Pre-define context
Intersubjectivity (diff. people might perceive or react to the same thing differently)	This may be important variation that's worth preserving

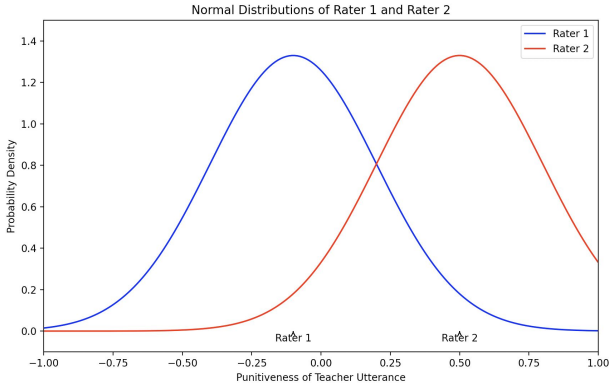
How to handle **inherently subjective** constructs?

During annotation

- Have multiple annotators (the more the better) for each example.

Processing Annotations

- Z-score judgments before aggregating
- Create different subsets of the data (subjective subset; less subjective subset) for evaluation



How to handle **inherently subjective** constructs?

During annotation

- Have multiple annotators (the more the better) for each example.

Processing Annotations

- Z-score judgments before aggregating
- Create different subsets of the data (subjective subset; less subjective subset) for evaluation

Modeling

- Incorporate confidence into the measurement
  - e.g. build a model that predicts rater agreement as a proxy for confidence
- Train representations separately for each rater and then combine them into a shared representation (Davani et al., 2022)

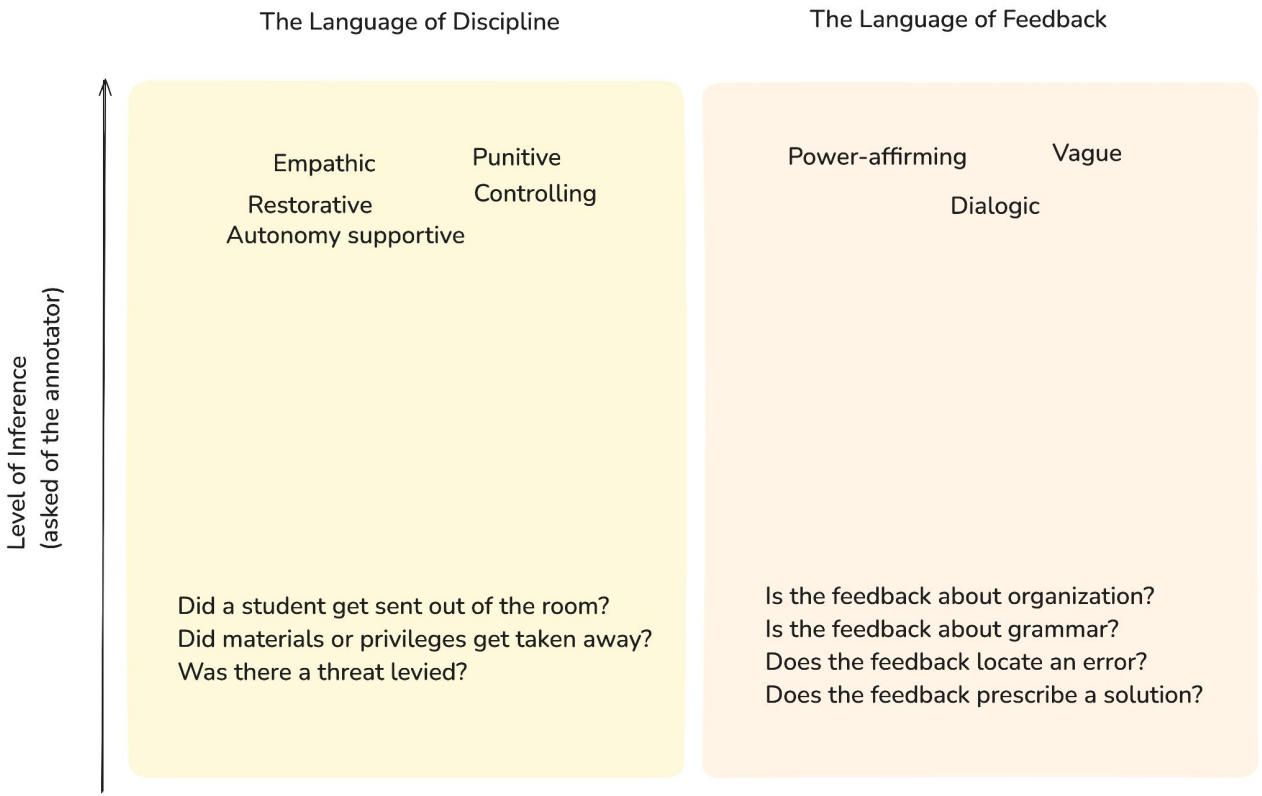
Validation

- Check if results are robust to variations in data or modeling decisions
  - e.g. leave-out validation
- Don't rely too heavily on your "ground truth" values
  - Correlate your measure with other relevant variables (e.g. overall instruction quality) to understand if it relates to positive or negative outcomes
  - Estimate the impact of applying your measure to address a specific issue

Application

- Don't use for making high stakes decisions!

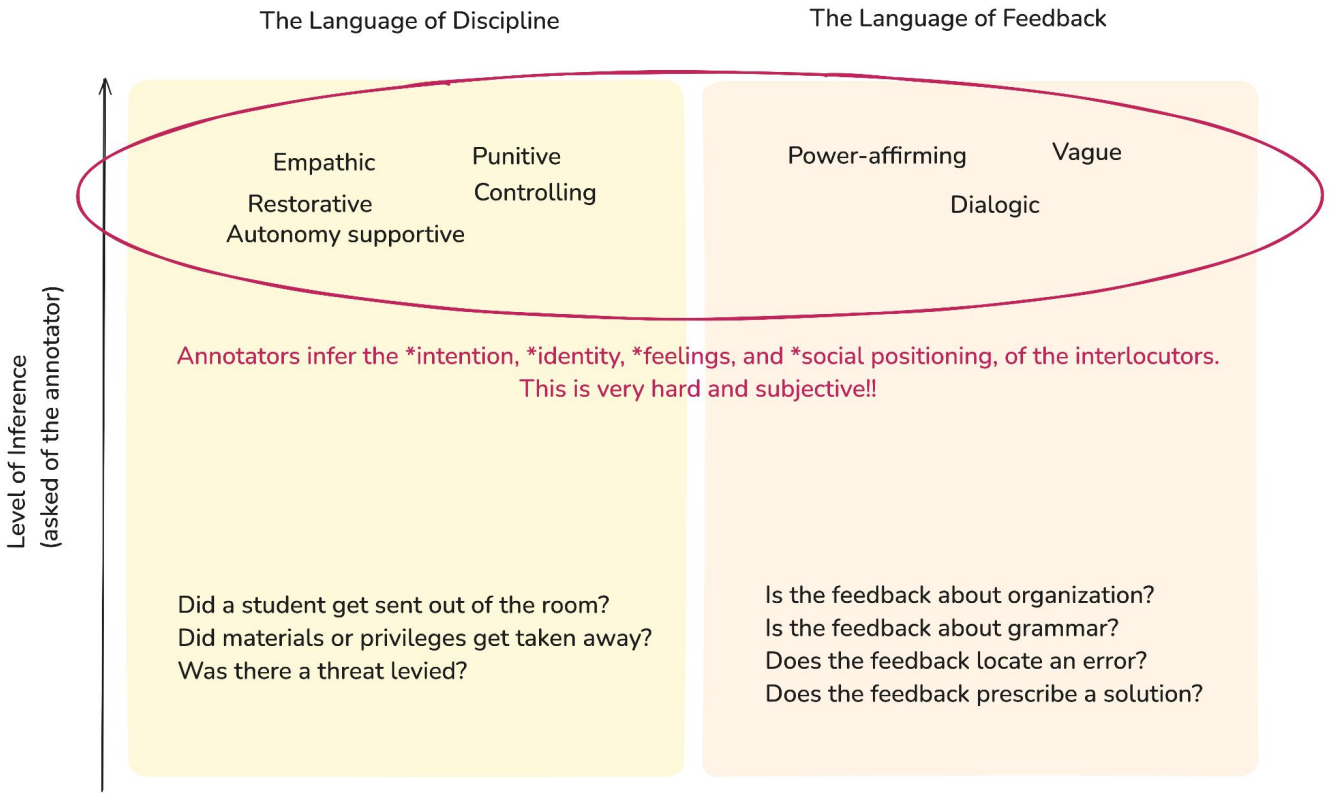
A few examples from lab work:



*all of these words are from the literature!

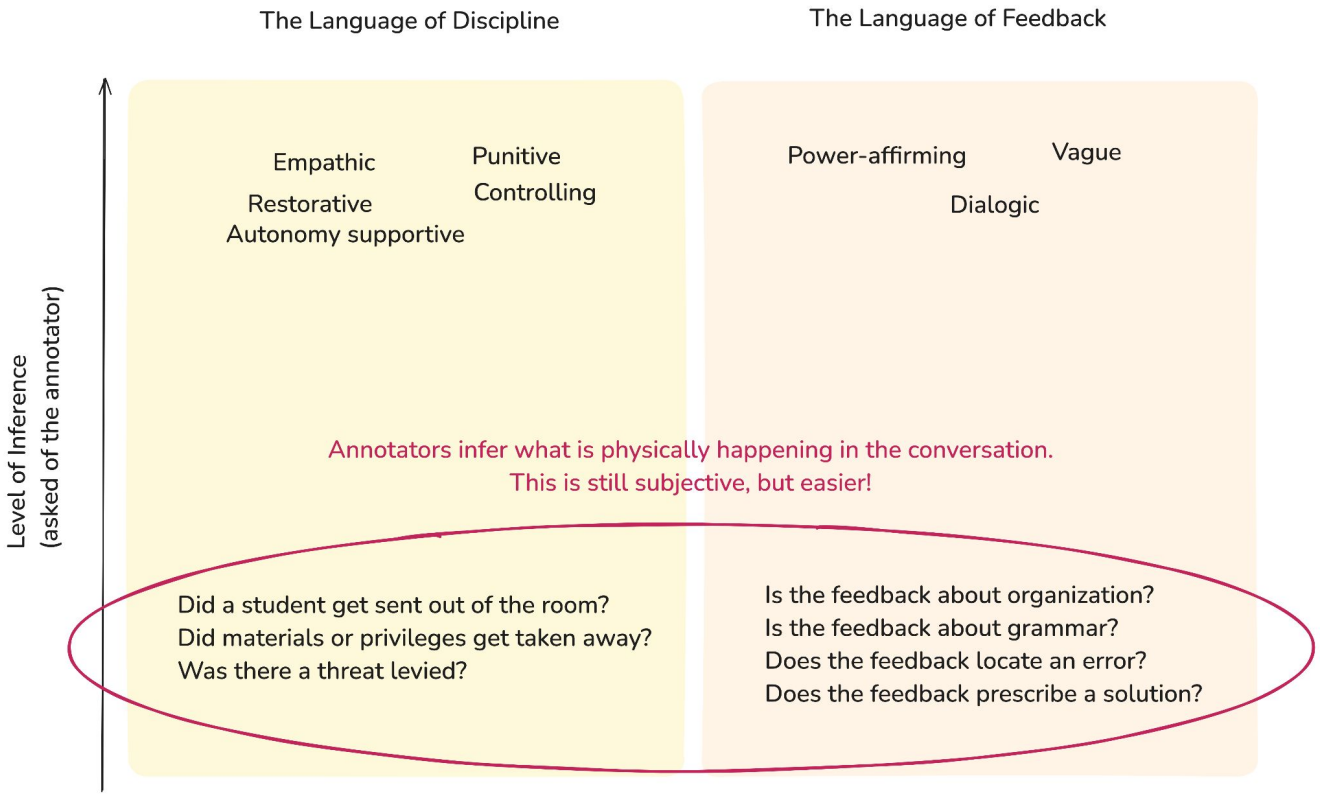
domain problems != technical problems

A few examples from lab work:



*IRR tends to be low for higher inference constructs

A few examples from lab work:



Do we expect LLMs to have the context to make this judgement based on its training?
Often for conversational data… we think no.

*is the construct so literal that we trust an LLM to infer instead?

Supervised modeling: LLMs or smaller models?

Smaller models (RoBERTa, BERT, etc.)	LLMs
Resources: https://simpletransformers.ai/; https://huggingface.co/docs/transformers/index	GPT, Claude, Gemini, DeepSeek
Pros: ● Downloadable → more transparency & control ● Needs little compute; **greener** choice! ● Can achieve similar performance to LLMs when sufficient labeled data is available	Pros: ● Very good at few shot learning ● Can be tuned with instructions
Cons: ● Require more training data ● Can't be tuned with instructions or via interacting with the model	Cons: ● Most cannot be downloaded ● Some models can't be finetuned

Supery

or both?

model distillation: use LLMs to generate data for finetuning smaller models

Smaller	
Resource https://hu	
Pros: ● Downloadable ● Needs little cor ● Can achieve sir sufficient label	few shot learning with instructions
Cons: ● Require more t ● Can't be tuned interacting with	be downloaded s can't be finetuned