4/8 — final mtg/thesis edits

3/4

* Thetas do not correlate — noise???
* JudgeBench
* MT, NER, etc — see if models can correctly judge just 1/0 pairs?
* Do something to demonstrate my hypothesis??
  + Try comparison test?????
* Any format but meet/exceed page
* Second reader emailed!!!

2/11

Initial results:

* Thetas do not correlate that well (r2 around 0.52) with available benchmarks?
* ICC curves look a little weird?

Questions:

* (Tanvi’s suggestion) few/zero/multishot prompting to expand # of models ?
* Expand from only essay grading?

Moving forward:

* JudgeBench LLM benchmark – maybe run that?
* DRAFT PAPER DUE SOON (2/20)

12/09

Questions/Issues

* Model inference with rubric per-category scoring format:
  + Run inference for each category?
* Robustness of only having like 15 models (students)? (similar to Tanvi’s final proj)
* Missing data for rubric item sometimes…?
* “Correctness”-- would it help for item param estimation to treat each category as its own question? Or vector similarity ..? tldr ask about per rubric scoring correctness

Question: Seq2Seq or Causal LM or Sequence Classification

(Classification task but heavily reliant on rubric prompt?)

* To fine tune or not?
* ^^ Could also compare?

A: try Seq2Seq classification and see?

Look into Adroit! — fill out application page

Get the zero-shot working first; then explore fine-tuning if needed

Focus on finishing the pipeline of operations

First: just with this dataset, what is existing performance with IRT

NEXT TIME: how was item response theory used exactly in these papers????

* IRT in non binary, writing — CREATE SLIDES !!! or latex :)

Task— narrowing the scope, what kind of writing???

* Argumentative,
* Asap, persuade 2.0 — what is available — exploratory

LLM Selection and Prompt Engineering

* Which LLMs are best suited for the writing domain and the specific writing features you aim to evaluate?
  + strengths/limitations
  + How to prompt
  + How will you handle potential biases in LLM judgments, such as position bias or sensitivity to prompt wording, to ensure a fair comparison?●

IRT Model Implementation and Analysis:

* How will you interpret the item difficulty and discrimination parameters obtained from the IRT analysis for both human and LLM judgments?■
* How will you compare the IRT-derived parameters between human and LLM judgments to assess the consistency and accuracy of LLMs as writing judges?