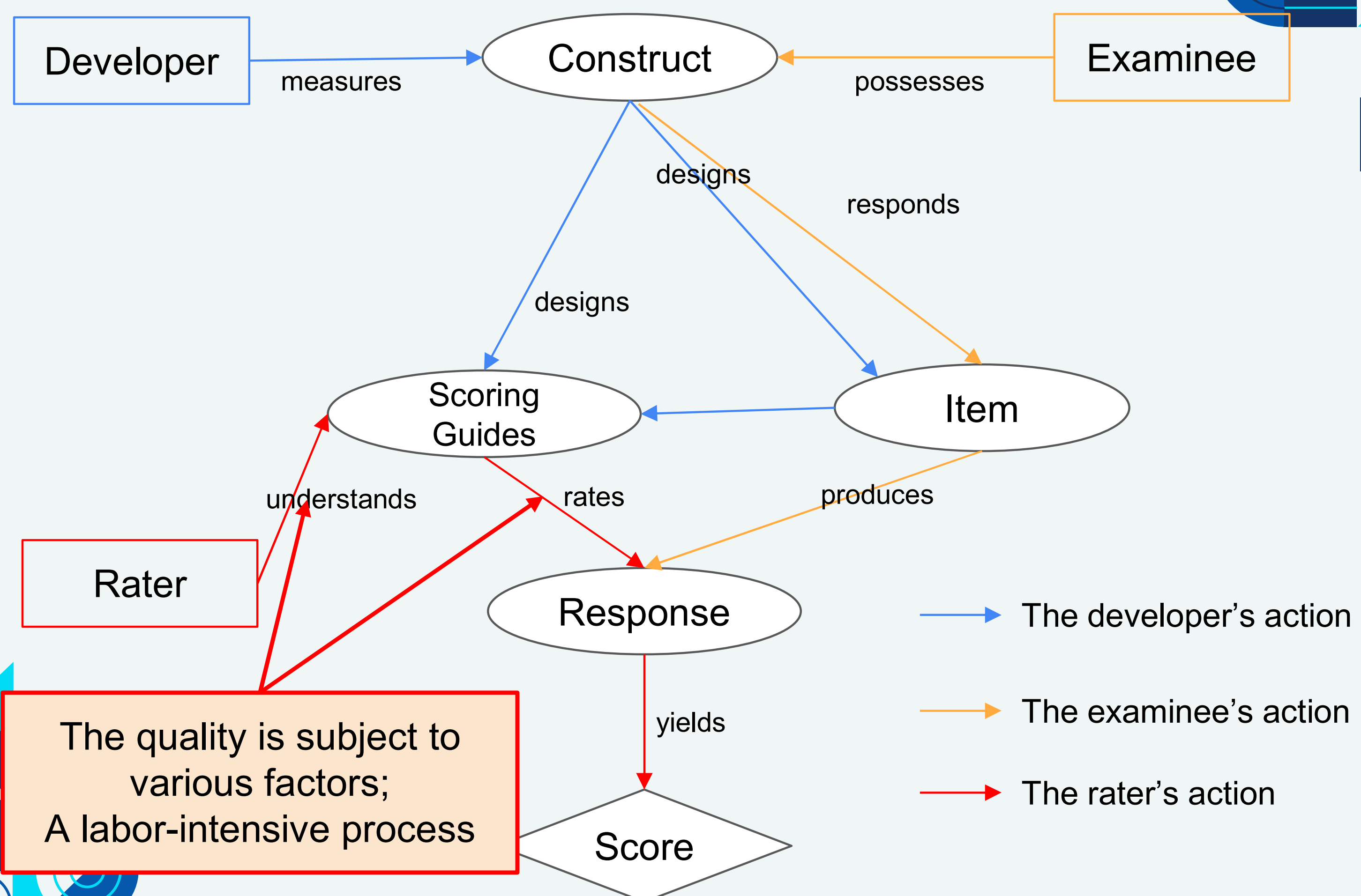
# Fine-tune GPT Models for Automatic Scoring Open-ended Response Items

Mingfeng Xue

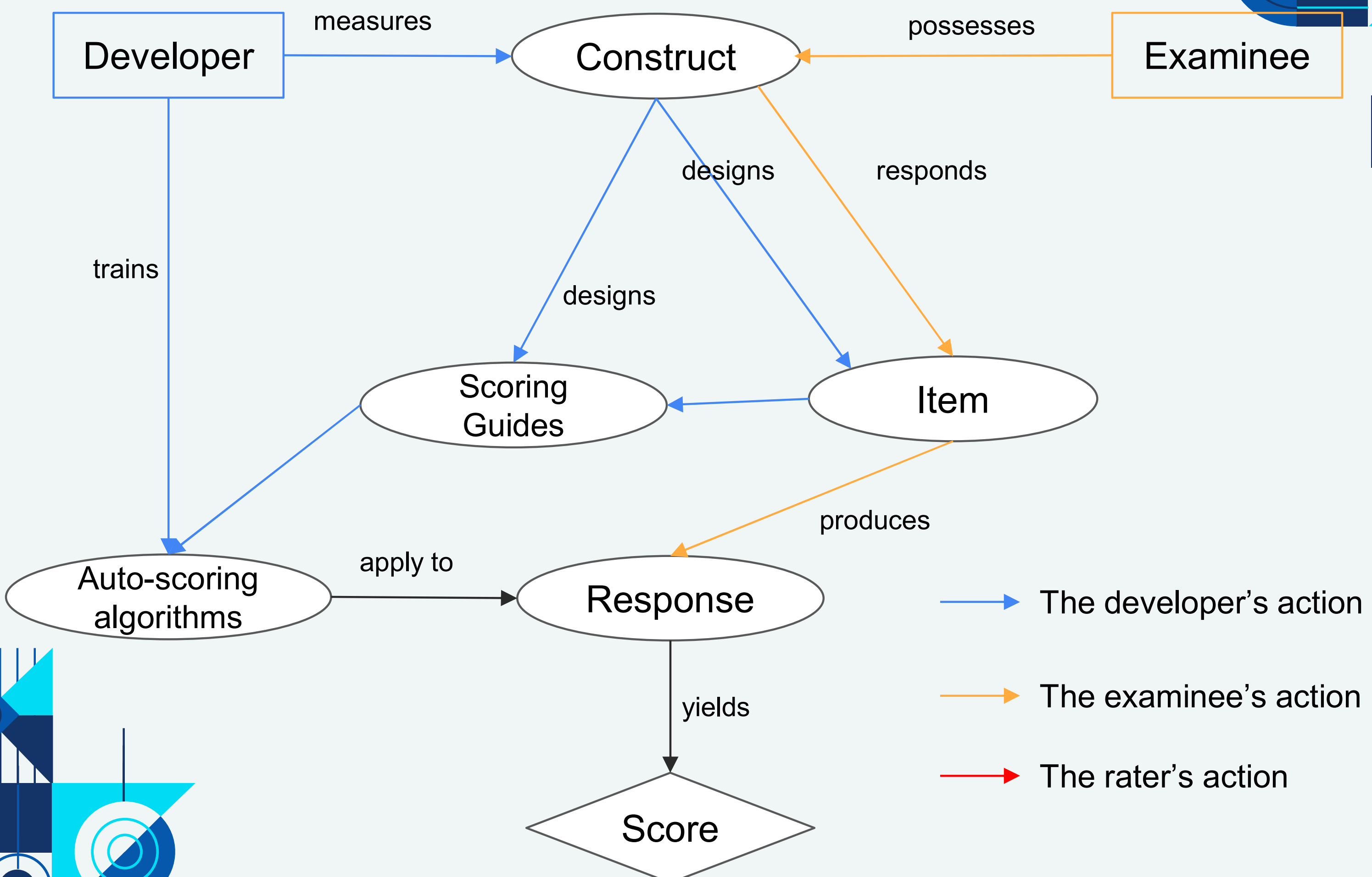
Bear Seminar

Nov. 7th, 2023

# Development and application open-ended items



# Development and application open-ended items



# Large language models (LLMs)

- Skip the feature engineering processes
- Thinking is directly correlated to language (e.g., think-aloud survey; Slobin, 1996)
- Open-ended responses are expressed in natural languages
- LLMs have proven to be effective in dealing various natural languages task (Bubeck, et al., 2023)
- Generative Pre-trained Transformer (GPT) is adopted because of its user-friendly API

# Fine-tune GPT

- ChatGPT outputs are inconsistent
- Fine-tuning is an approach to transfer learning in which the weights of a pre-trained model are trained on new data
- An application of the pretrain-finetune paradigm in LLMs
- Boost the performance of GPT in auto-scoring
- Make the auto-scoring more user-friendly

# Benefits of fine-tuning GPT in auto-scoring

- Consistency/ reliability
  - Outputs can be deterministic through proper settings
- Validity
  - Overcome the rater variability in manual ratings
  - Better align the scoring with test developers' intention in a border usages of the test
- Efficiency
  - Reduce cost
  - Increase scoring speed (especially important for some test scenarios, e.g., CAT)

# Research questions

- How consistent and accurate is ChatGPT in scoring?
- How accurate are fine-tuned GPT models in scoring under different conditions?
- What are the influence of autoscoring of fine-tuned GPT on latent trait estimates?
- How harsh are the fine-tuned GPT models in scoring in comparison to humans?

# Data

- # of students: 930 middle school students
- # of items: 7
- The construct measured: Pattern recognition
- 1/3 of the responses were doubly rated according to the scoring guides

| Item | Measurement goal | Maximum scores | # of responding students | Average response length | Standard deviation of response length |
|---|---|---|---|---|---|
| LZ2 | Compare two patterns at two places | 2 | 453 | 24.15 | 17.47 |
| LZ3 | Compare two pattern at two places | 2 | 452 | 20.81 | 12.66 |
| S9 | Describe one pattern among several | 3 | 434 | 29.32 | 27.98 |
| S12 | Describes two or more patterns among several | 2 | 470 | 25.82 | 26.18 |
| W13 | Describe the exact one pattern | 2 | 416 | 18.86 | 17.67 |
| W14 | Describes the exact two patterns | 2 | 440 | 18.65 | 17.49 |
| W15 | Describes the exact two patterns | 3 | 416 | 20.39 | 17.57 |

# Procedures

**Step 1**

Remove responses with less than three words

Split data into train and test sets at the ratio of 80:20

**Step 2**

Use oversampling techniques to generate train sets of various sample sizes: 10, 50, 100 per category, and all data

For each sample size, generate two sets with and without scoring guides

**Step 3**

Transform the train sets in to JSON file (system, user, assistant)

Fine-tune GPT model for each item, respectively, through Open AI's API on the train sets

**Step 4**

Generate prediction for the test sets

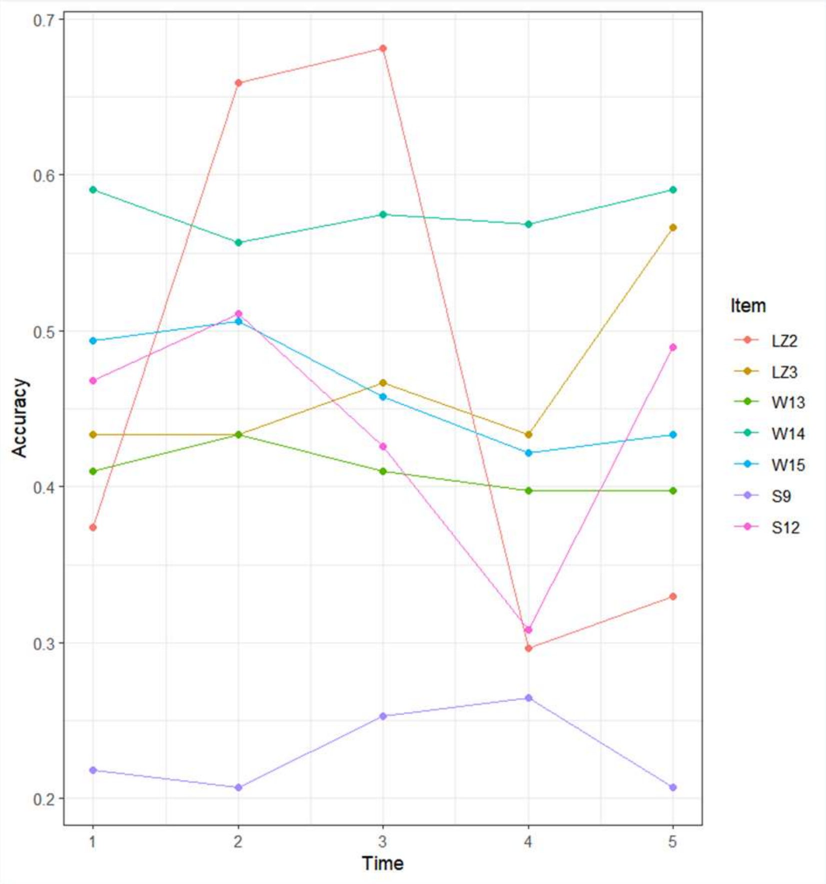
For ChatGPT, there is no training process, so I directly asked ChatGPT to produce scores according to the scoring guides five times

**Step 5**

Further analyses

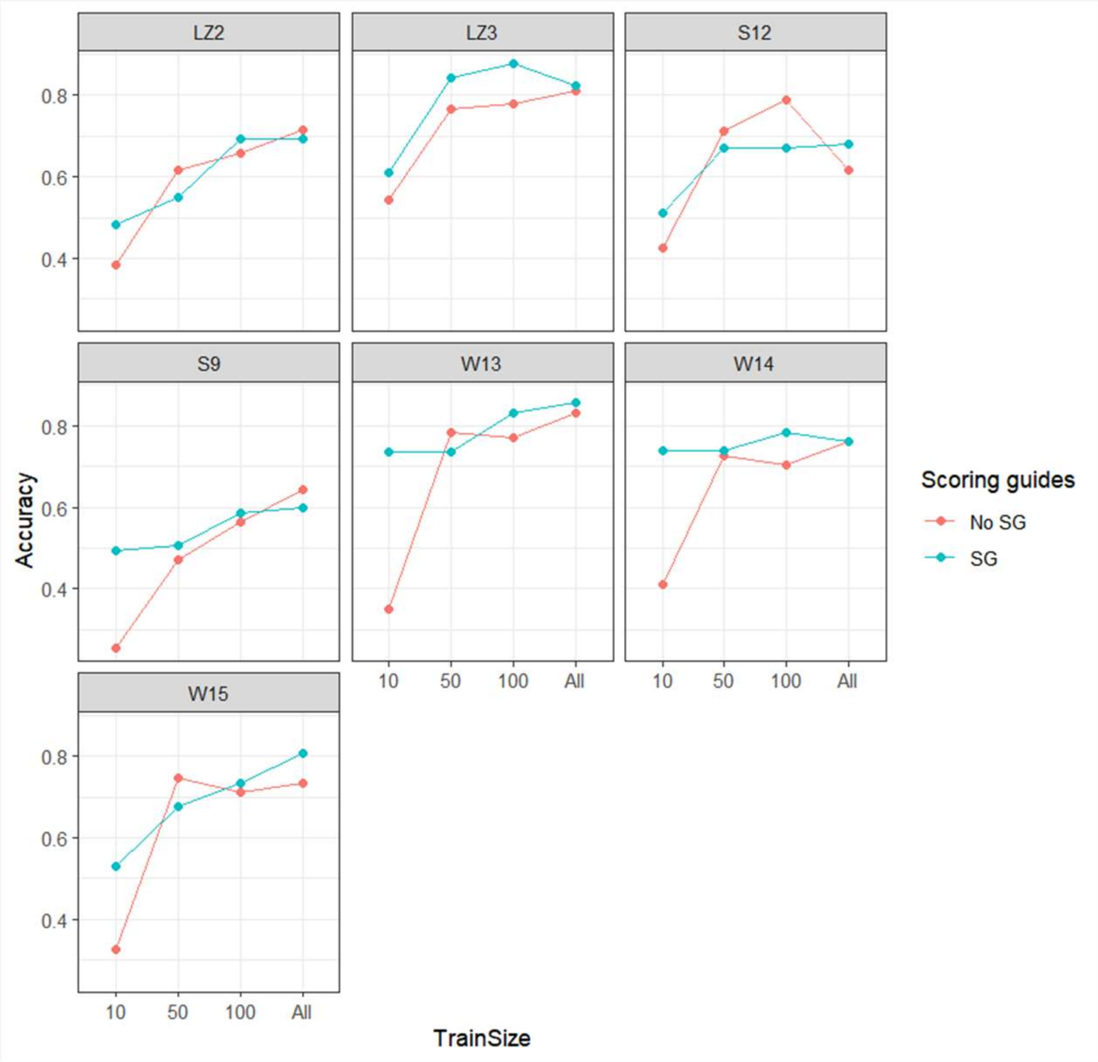
# Consistency of ChatGPT in scoring

| Item | Fleiss' Kappa |
|---|---|
| LZ2 | .578 |
| LZ3 | .766 |
| S9 | .681 |
| S12 | .349 |
| W13 | .914 |
| W14 | .722 |
| W15 | .636 |



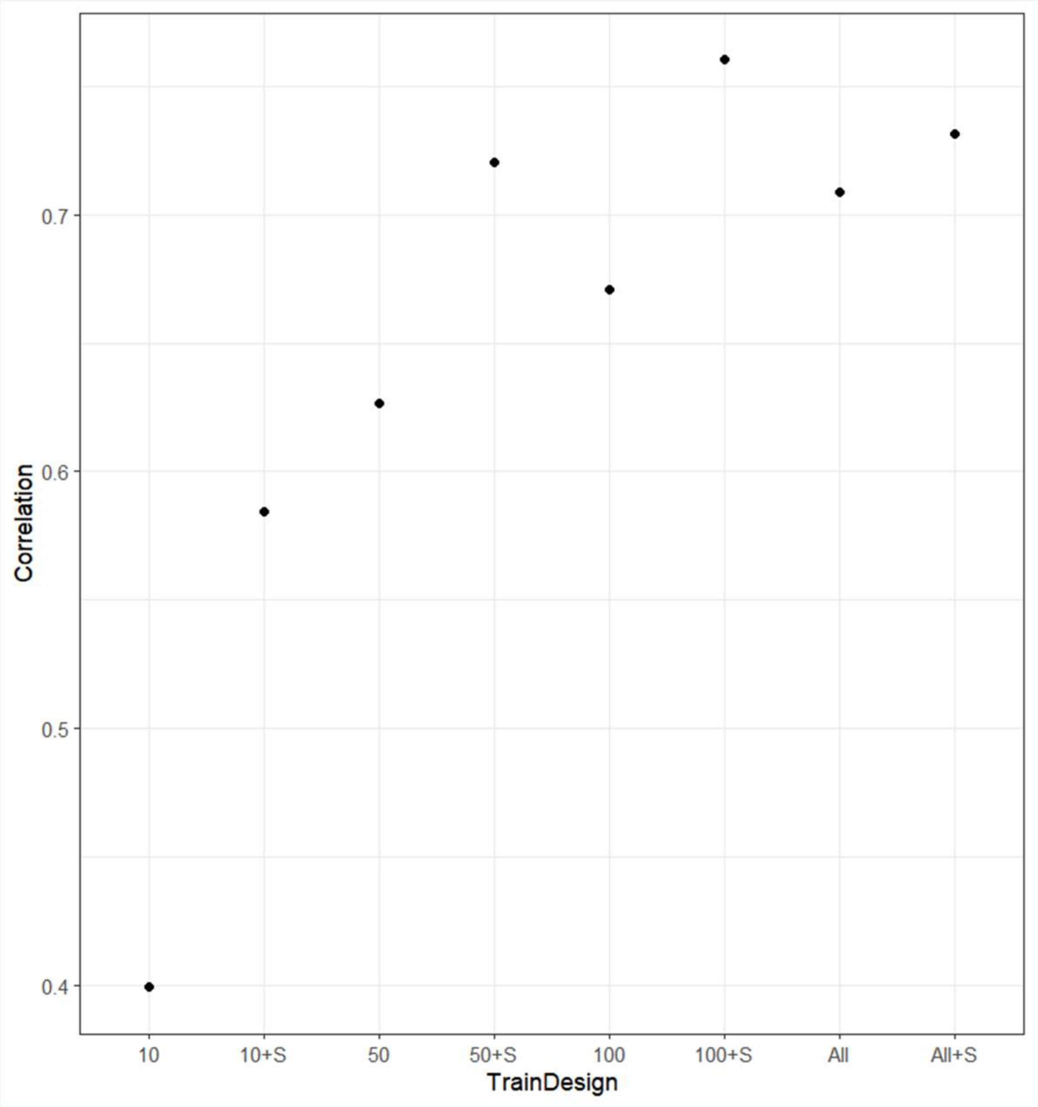
# Accuracy of autoscoring by fine-tuned GPT models

- As train sizes go up, the accuracy generally increases
- The inclusion of scoring guides increases the performance
- With 100 samples per category and scoring guides, the accuracy is the highest



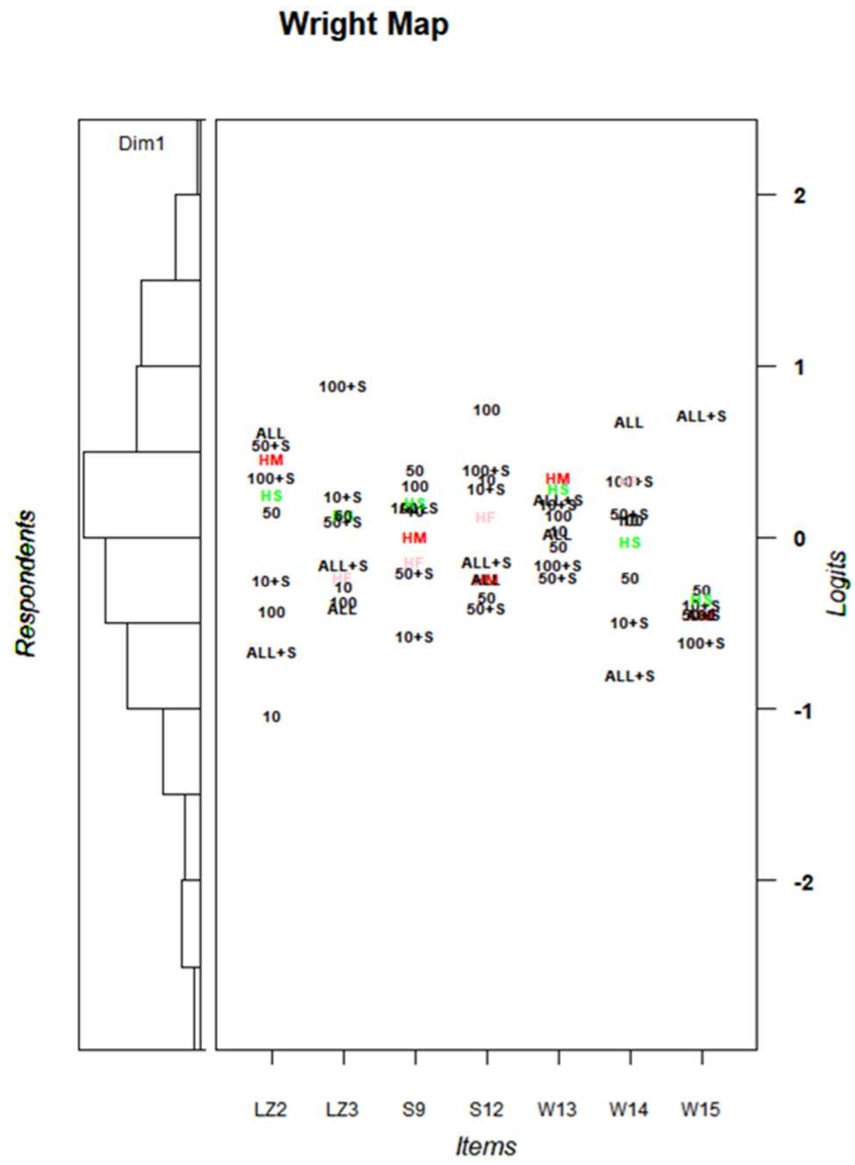
# The influence of autoscoring on latent trait estimates

- GPCM
- Cases with two responses and above
- Correlation of latent trait estimates between manual scoring and autoscoring



# Fine-tuned models as raters

- Many-facet models
- Rater x item design



# Future directions

- Pretrain-finetune ⟶ Pretrain-finetune-furtherFinetune
- Move onto GPT4
- Incorporate chain of thought or tree of thought into scoring
- Manual scoring –> GPT aided scoring –> GPT scoring

OpenAI DevDay, Opening Keynote

# Q & A

# Reference


Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Slobin, D. I. (1996). From “thought and language” to “thinking for speaking”.