Enhancing Item Parameter Prediction with Transfer Learning

Mingfeng Xue[1,2] and He Ren[3]

[1]School of Education, UC Berkeley
[2]School of Education, UNC Greensboro
[3]College of Education, University of Washington

IMPS 2025, Minneapolis, Minnesota

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION

Introduction

- Traditional item calibration requires field testing
  - Labor-intensive
  - Costly
  - Inefficient (asynchrony between item writing and revision)
- Item parameter prediction from text (deep learning with language models)
  - Most items are expressed in text
  - Transformer-based language models have shown impressive performance in natural language processing (NLP) tasks

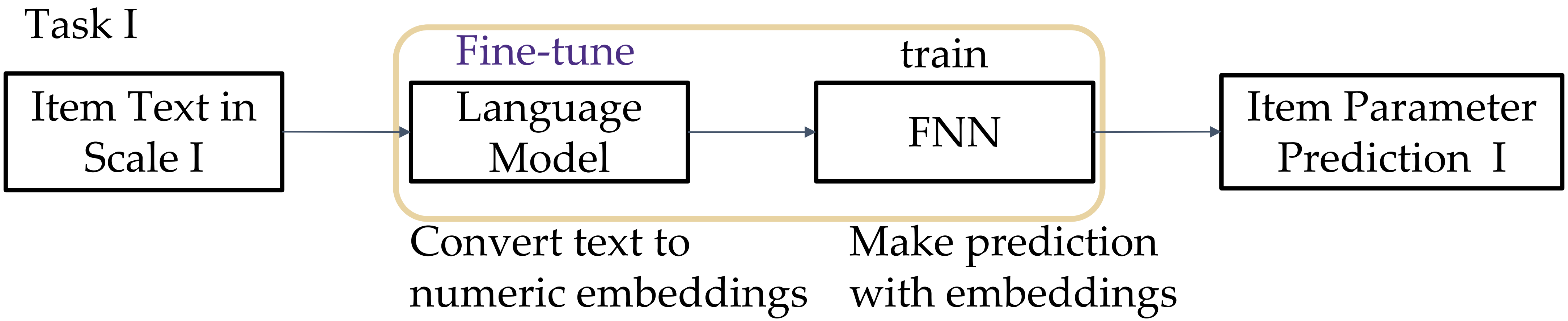
Motivation

> Small sample issue: A psychological scale can consist of three to hundreds of items (too small for deep learning)

> Transfer learning: a machine learning technique where a model trained on one task (denoted as base model) is transferred to help solve a different, but related task

- A good base model is critical

A Diagram for Transfer Learning in Our Case


Task I

Fine-tune

train

Item Text in Scale I

Language Model

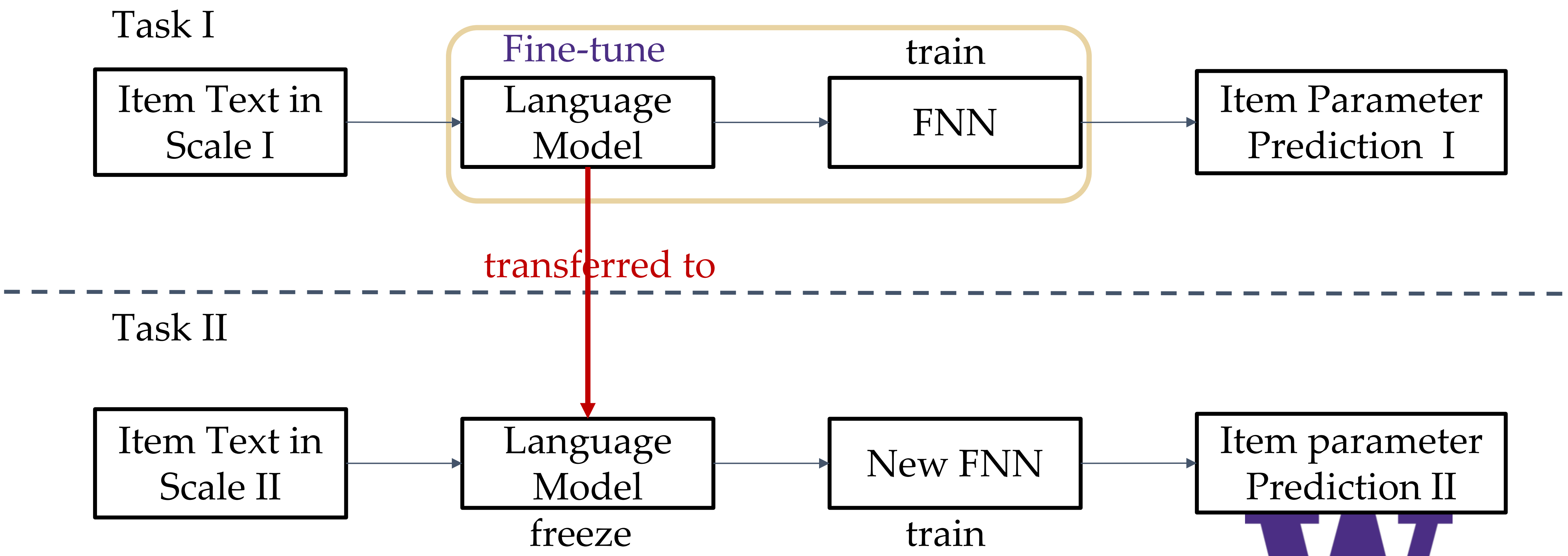
FNN

Item Parameter Prediction I

Convert text to numeric embeddings

Make prediction with embeddings

A Diagram for Transfer Learning in Our Case



Goals

> Evaluate the performance of transfer learning for item parameter prediction

- The influence of fine-tuning in transfer learning with language models
- The performance of two kinds of transfer learning:
  - The same prediction goals but different data

    i.e., the base language model is trained on predicting difficulty of one scale and transferred to predict difficulty of another scale
  - Different prediction goals and different data

    i.e., the base language model is trained on predicting item-pair similarity and transferred to predict difficulty of another scale

Data

> International Personality Item Pool (IPIP): 1142 respondents

> Apply GPCMs to calibrate item parameters

> Revised NEO Personality Inventory (NEO PI-R)

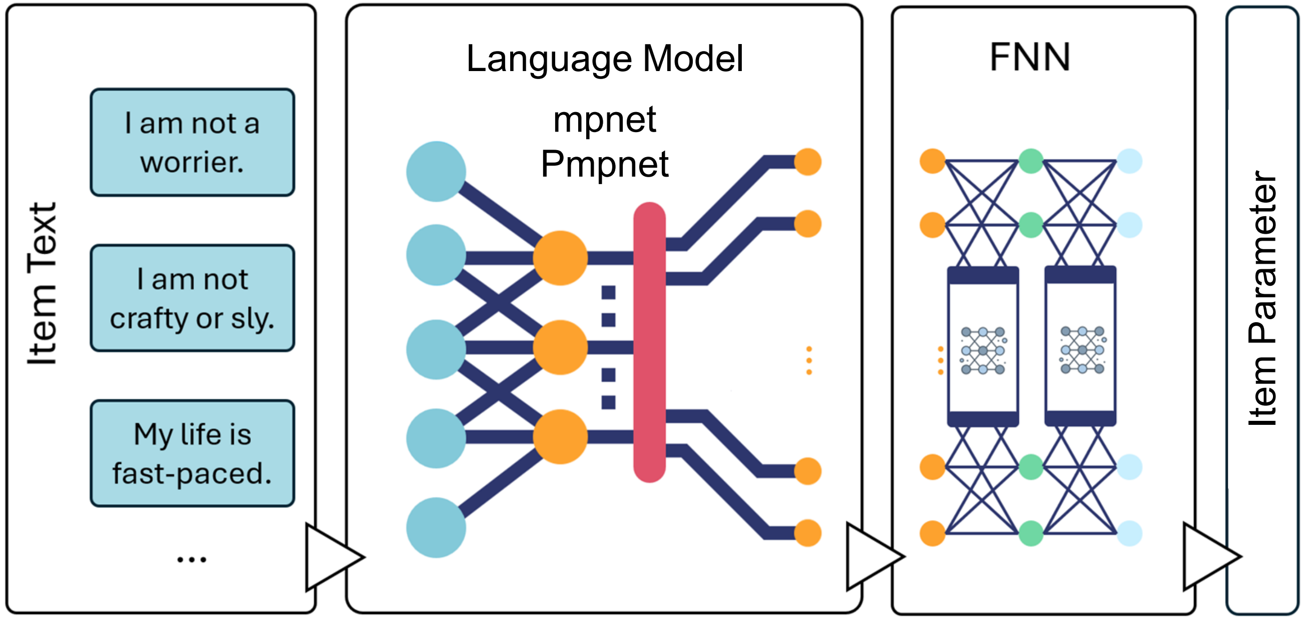
- Big-five personality: Neuroticism, Extraversion, Openness to experience, Agreeableness, Conscientiousness
- 240 items on a five-point Likert scale
- Intercept: M=-0.41, SD=1.05; Slope: M=0.67, SD=0.27
- Example. Agreeableness: I couldn’t deceive anyone even if I wanted to

Data

> The Sixteen Personality Factor Questionnaire (16PF)

- 156 items on a five-point Likert scale
- Intercept: M=-0.97, SD=2.22; Slope: M=0.76, SD=0.35
- 16 personality factors, e.g., warmth, anxiety
- Example. Anxiety: I am afraid that I will do the wrong thing

Model Structure

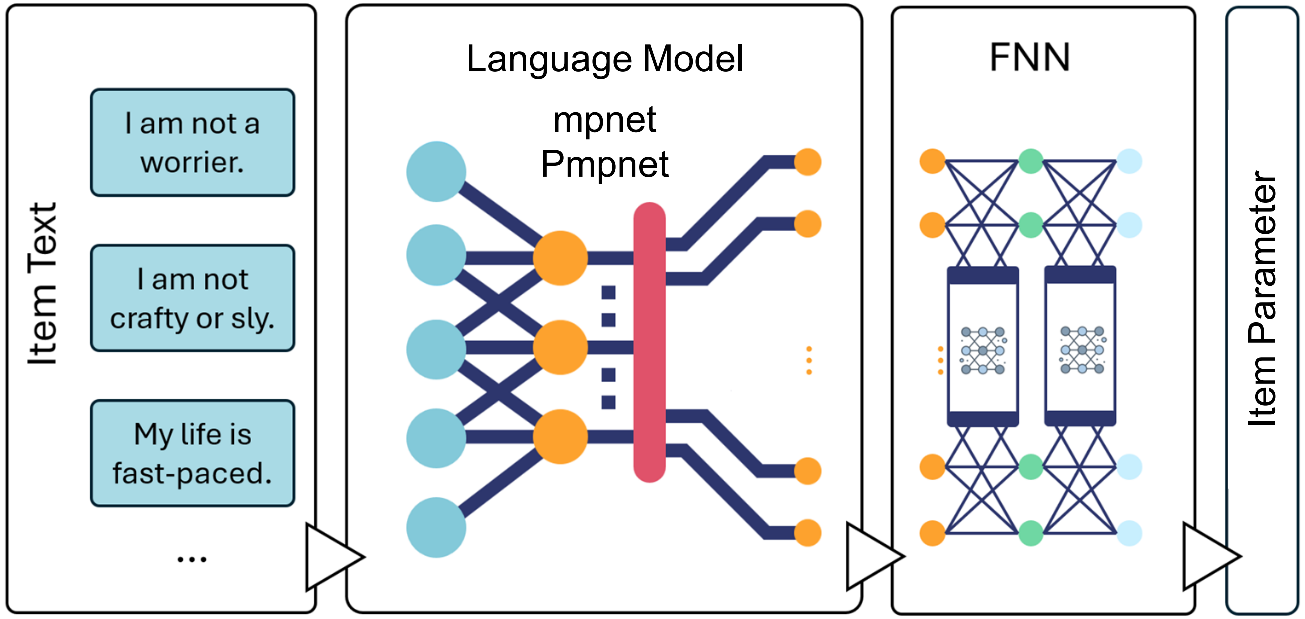


> Language model

- mpnet: a sentence BERT model (all-mpnet-base-v2)
  - Can be directly used or fine-tuned on our data
- Pmpnet: fine tuned by Wulff & Mata (2025) on predicting item pair similarity in 200,000 items in the IPIP (i.e., a base model with a different goal)

Model Structure

> FNN:
  - Three layers
  - Relu activation function
  - Dropout rate of 0.3
  - 30 epochs
  - 5-fold cross validation



Metrics

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R squared (R2)
- Correlation

Results: Difficulty Prediction

> The role of fine-tuning

	RMSE	**MAE**	**R2**	**Correlation**
NEO PI-R				
mpnet + **FNN**	1.012	0.811	0.071	0.413
mpnet + **FNN**	**0.974**	**0.762**	**0.140**	**0.431**
16PF				
mpnet + **FNN**	1.981	1.581	0.140	0.454
mpnet + **FNN**	**1.927**	**1.554**	**0.185**	**0.457**

> Fine-tuning increases item difficulty prediction for NEO PI-R by 100% and 16PF by 32% using R2 as the metric

Red: Fine tuned
Purple: Train

Results: Difficulty Prediction

> The role of transfer learning with base model trained on item difficulty

	RMSE	**MAE**	**R2**	**Correlation**
16PF				
mpnet + **FNN**	1.981	1.581	0.140	0.454
mpnet + **FNN**	1.927	1.554	0.185	0.457
mpnet_neo + **FNN**	**1.877**	**1.553**	**0.227**	**0.551**

> It further increases item difficulty prediction for 16 PF by 30% using R2 as the metric

Red: Fine-tuned
Purple: Train
Green: Trained on another task

Results: Difficulty Prediction

> The role of transfer learning with base model trained on item pair similarity

	RMSE	MAE	R2	Correlation
16PF				
mpnet + FNN	1.981	1.581	0.140	0.454
mpnet + FNN	1.927	1.554	0.185	0.457
mpnet_neo + FNN	1.877	1.553	0.227	0.551
Pmpnet + FNN	**1.496**	**1.182**	**0.509**	**0.715**

> It further increases item difficulty prediction for 16PF by 200% using R2 as the metric

Red: Fine-tuned
Purple: Train
Green: Trained on another task

Results: Discrimination Prediction

Red: Fine-tuned
Purple: Train
Green: Trained on another task

	RMSE	**MAE**	**R2**	**Correlation**
NEO PI-R				
mpnet + FNN	0.412	0.313	0.494	0.717
mpnet + FNN	0.350	0.254	0.635	0.798
Pmpnet + FNN	0.329	0.249	0.677	0.826
16PF				
mpnet + FNN	0.690	0.552	0.292	0.587
mpnet + FNN	0.681	0.515	0.310	0.598
mpnet_neo + FNN	0.664	0.502	0.344	0.631
Pmpnet + FNN	**0.518**	**0.381**	**0.601**	**0.785**

> Discrimination prediction have higher accuracy

> Similar patterns about fine-tuning and transfer learning are found

Conclusions

- Fine-tuning help increase the accuracy of item parameter prediction
- Transfer learning can further improve item parameter prediction even when the item sample size is limited (e.g., 16PF) or the base model is trained for a different purpose (e.g., Pmpnet)
- Item discrimination is easier to be predicted than item difficulty

Thanks!

mingfengxue@berkeley.edu
heren@uw.edu

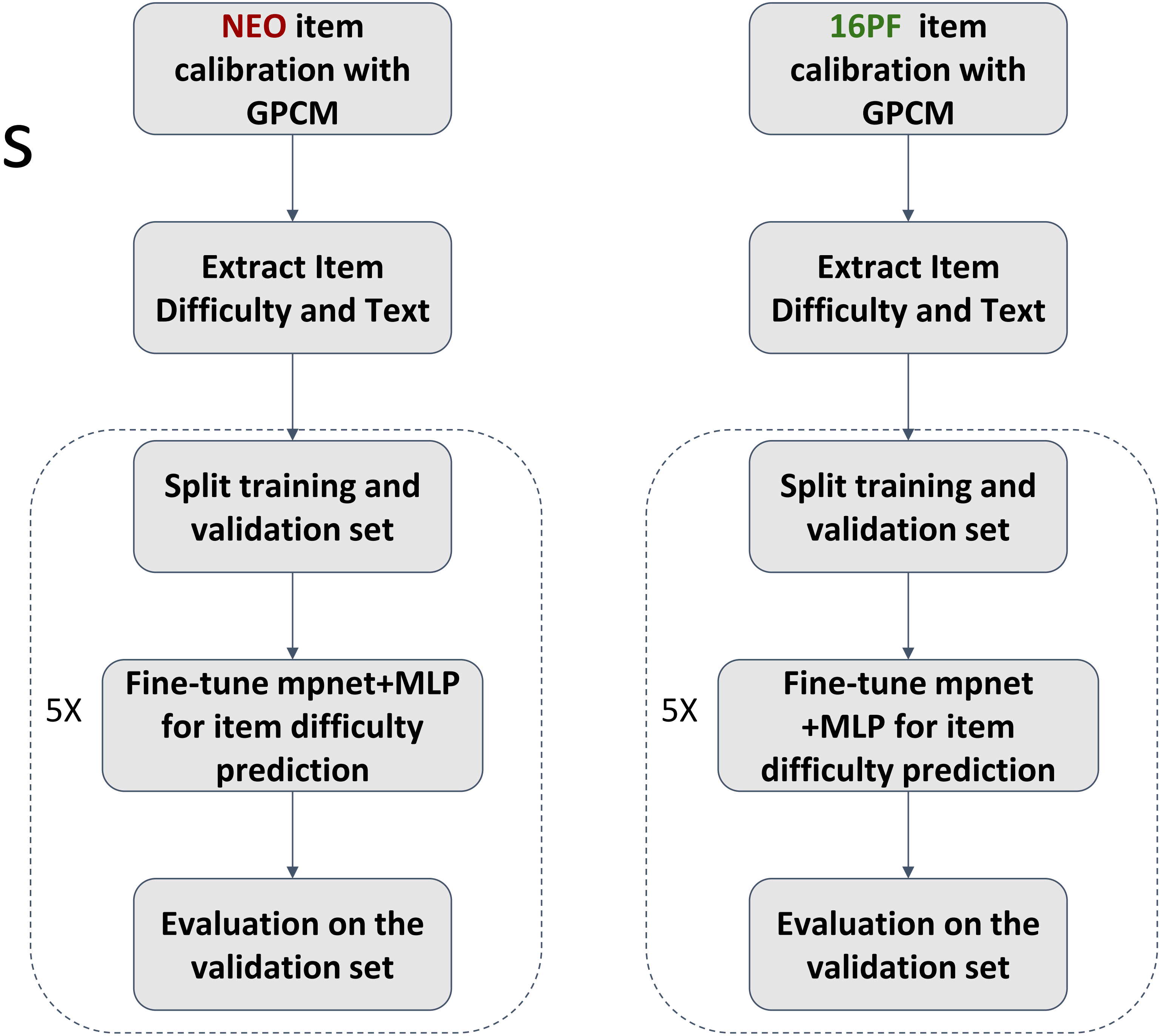
Personal Homepage

Mingfeng Xue and He Ren

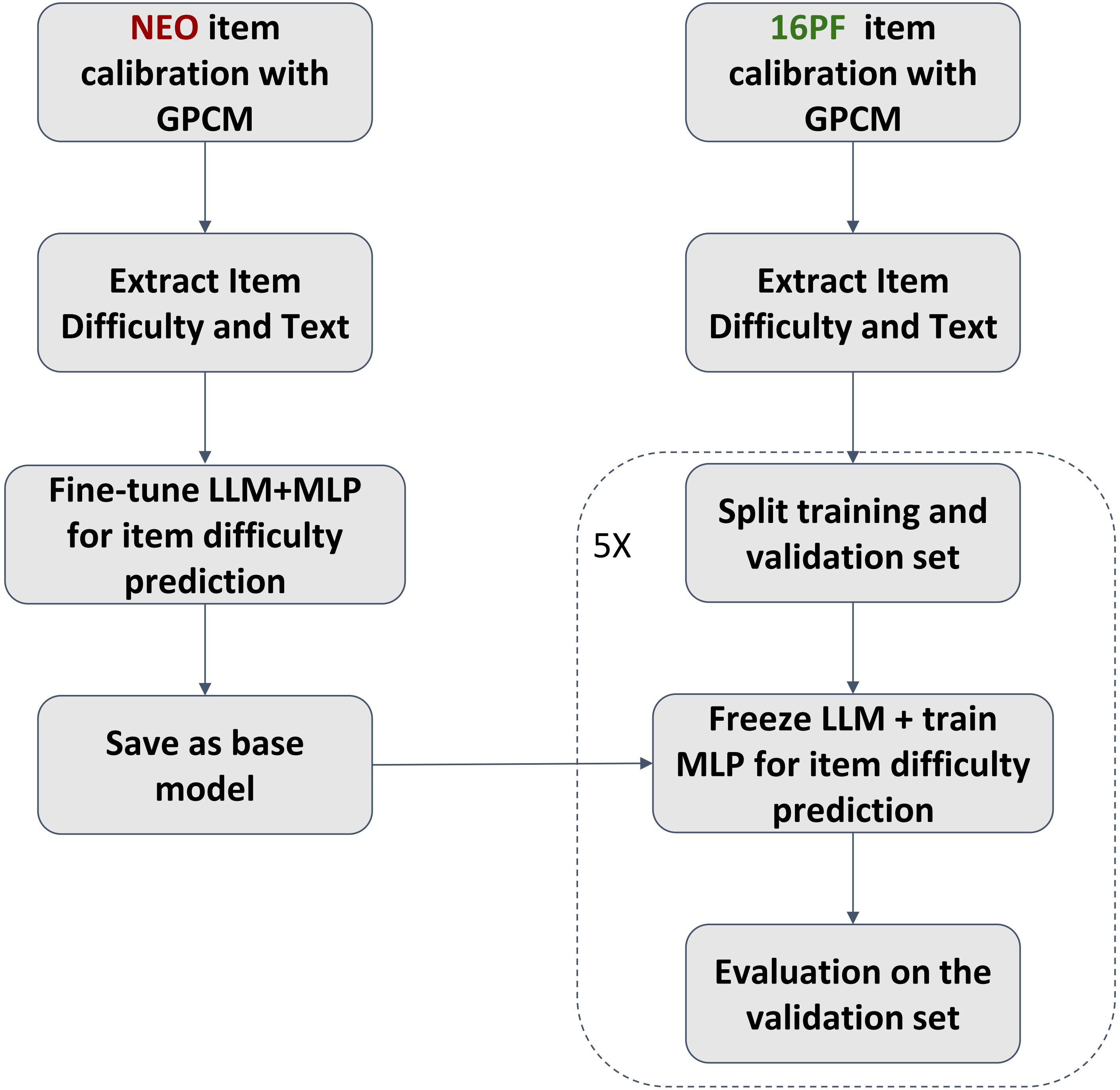
References

Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, *9*(5), 944-954. https://doi.org/10.1038/s41562-024-02089-y

Process


NEO item calibration with GPCM
Extract Item Difficulty and Text
Split training and validation set
5X
Fine-tune mpnet+MLP for item difficulty prediction
Evaluation on the validation set
16PF item calibration with GPCM
Extract Item Difficulty and Text
Split training and validation set
5X
Fine-tune mpnet +MLP for item difficulty prediction
Evaluation on the validation set

Process


NEO item calibration with GPCM

Extract Item Difficulty and Text

Fine-tune LLM+MLP for item difficulty prediction

Save as base model

16PF item calibration with GPCM

Extract Item Difficulty and Text

5X

Split training and validation set

Freeze LLM + train MLP for item difficulty prediction

Evaluation on the validation set