CollegeBoard

Predicting Item Difficulty for Pretest Items in Large-Scale Assessments

YoungKoung Kim | Psychometrics

College Board

December 10, 2025

Agenda

- Introduction : Motivation and Goals
- Two-Step Hybrid Approach
- Research Questions
- Method
- Results
- Summary and Discussion

How can difficulty prediction support item development?

⚡ Early check before pretesting

- Pretesting items is costly in time and resources
- TDs can get immediate feedback from the model during development

🧠 Feedback on linguistics features

- By combining LLM-predicted difficulty with interpretable linguistic features, the model can articulate *why* an item is predicted to be of a certain difficulty value
- Using the model as a companion tool, TDs can finetune item content iteratively until it aligns with the target difficulty

🧭 Enhanced pretest planning

- Predicted item difficulty could help inform the prioritization of items in the pretesting pipeline, which is especially important when there's a backlog of un-pretested items and limited slots

🔄 Continuous improvement

- Over time, TDs can develop a stronger understanding of the linguistic features that are most consistently associated with certain difficulty levels
- As more pretest results become available for training and validation, the model's predictive performance can also improve over time

What are the key goals for this research?

🎯 Achieve more precise difficulty estimates compared to human experts

- Since our TDs are already quite strong with nailing target difficulty buckets, we'd like the model to generate more fine-grained predictions so that we can align development efforts even more closely to specific item pool and test assembly needs.

🔍 Preserve interpretability for users

- Model should articulate which linguistic features (e.g., text complexity, stimulus length) are most predictive of item difficulty so that the model remains transparent and supports learning (versus just scoring).

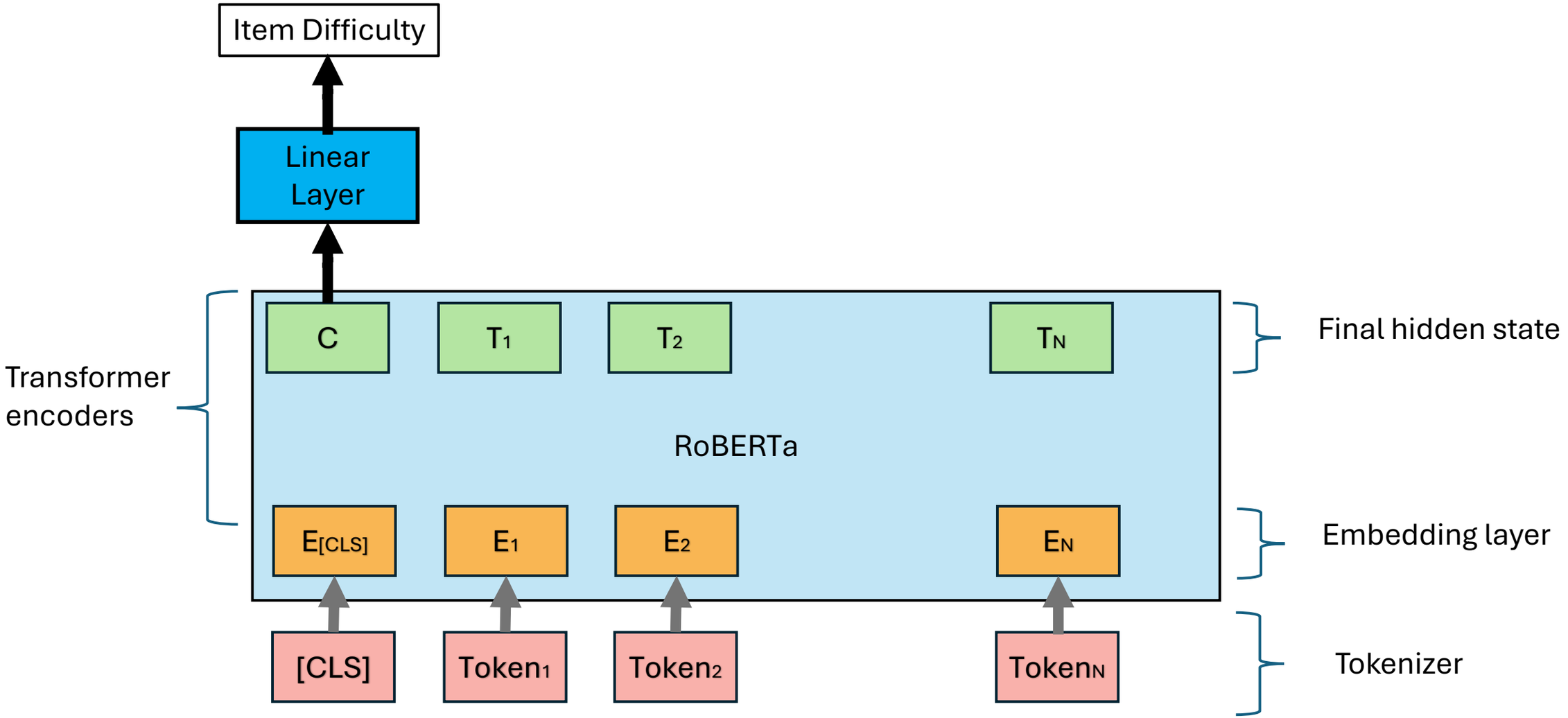
💡 Investigate the value of a hybrid approach that incorporates transformer-based models with a traditional feature-based approach

- Traditional ML approach uses linguistic features only. We want to see if there's any incremental value in using transformer-based models to enhance prediction accuracy while preserving interpretability.

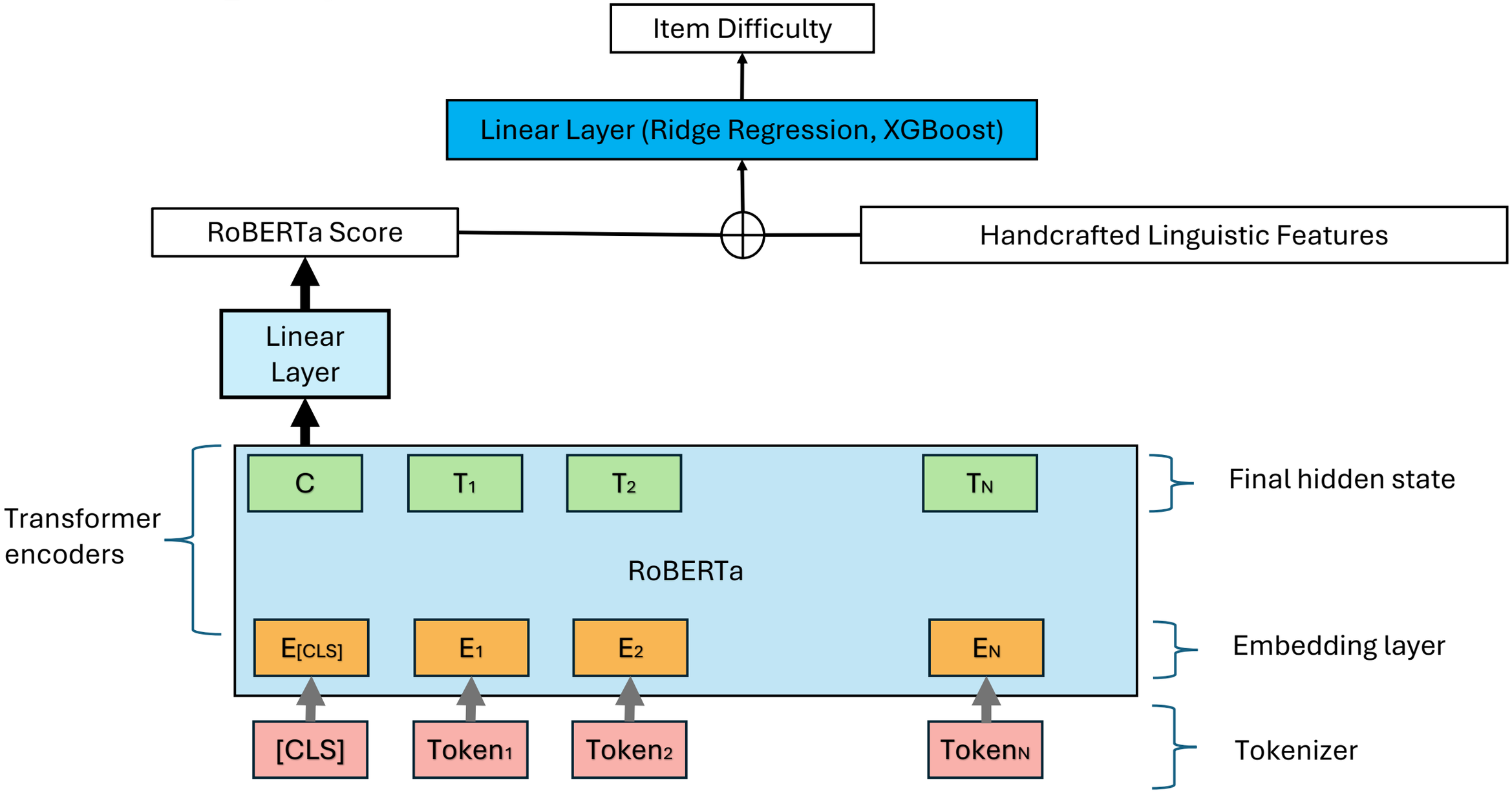
Modeling Item Difficulty Using Language Models

- Text based Modeling Approach
  - Feature-based approach
  - Transformer-based model approach
  - Hybrid approach
- Kim & Moses (2025) Two Step Hybrid Approach
  - An extension of Uto, Xie and Ueno (2020)
  - Step 1 (Transformer-Based Model Fine-Tuning) : Fine-tune encoder-only transformer models – BERT, RoBERTa, DeBERTa, ModernBERT – on item text to predict difficulty
  - Step 2 (Hybrid Predictive Modeling): Combine Step 1 predictions with linguistic features using machine learning models such as Ridge and XGBoost

RoBERTa for Item Difficulty Prediction



Two-Step Hybrid Model

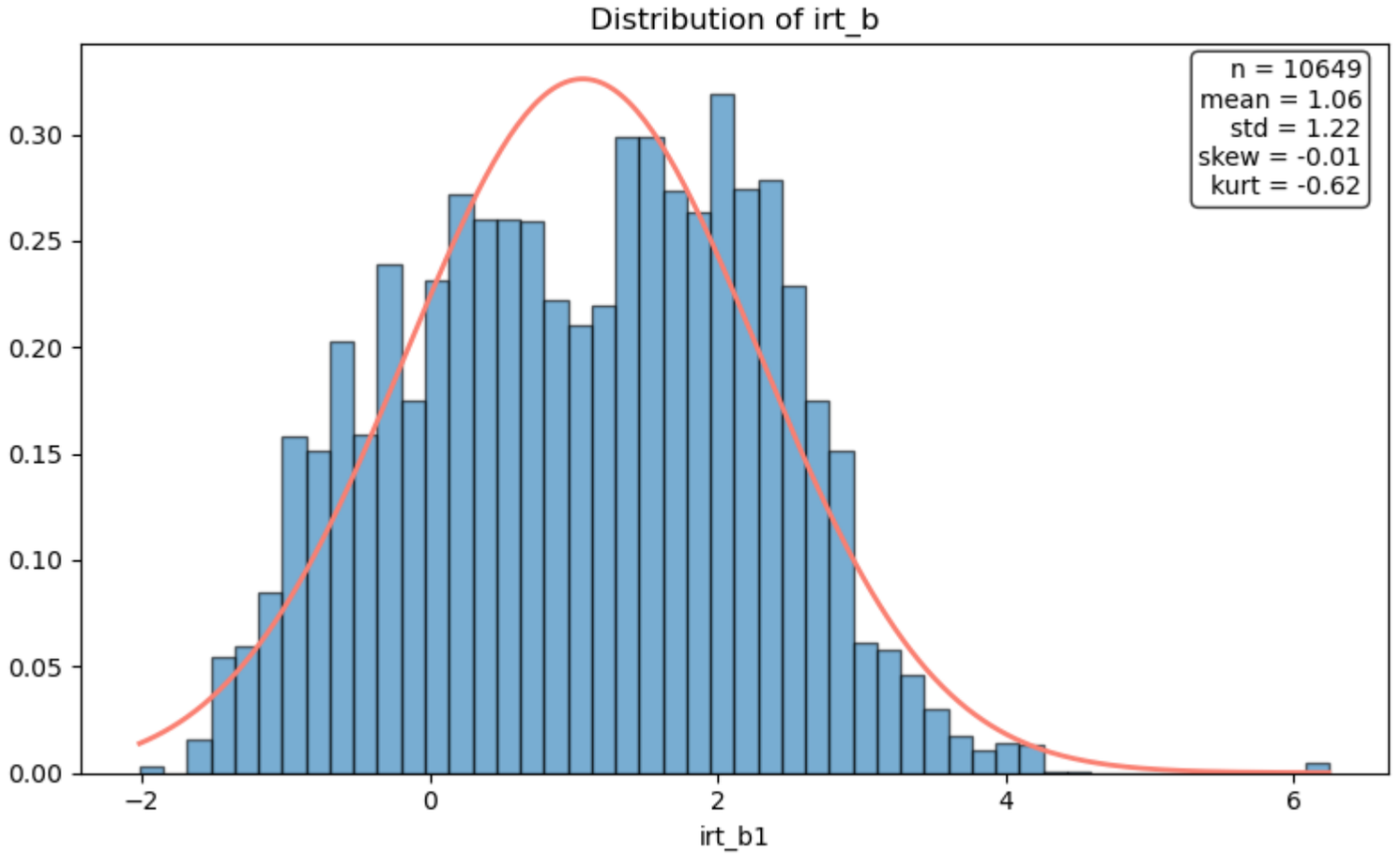


Research Questions

- Does a hybrid approach improve prediction accuracy compared with a linguistic-feature-only model?
  - Linguistic-feature-only model vs hybrid approach
- Which transformer model performs better for this task ?
  - RoBERTa
  - ModernBERT
- Which input configuration produces higher accuracy?
  - Stem + Stimulus
  - Stem + Stimulus + Options
  - Stem + Stimulus + Options + Skills
- Which ensemble model performs better for this task ?
  - Ridge Regression
  - XGBoost

Method : Data

- Data: Large-scale literacy assessment with multiple-choice items
- Target Item Difficulty
  - IRT b Parameter
- Train, Validation, Test Data
  - Train N = 8,232
  - Validation N = 1,164
  - Test N = 1,253



Method : Models

- Step 1 Model : Encoder-only transformer models
  - RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et. al, 2019)
  - ModernBERT: Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference (Warner, B. et. al, 2024)
- Step 2 Model :
  - Ridge Regression
  - XGBoost Regression

Linguistic Features

- Lexical Features
  - Word types, vocabulary richness, and word-level properties (e.g., counts of academic words, long words, specific parts of speech, pronouns)
- Syntactic Features
  - Grammatical structure and sentence composition indicators (e.g., counts of prepositions, auxiliary verbs, conjunctions, sentences)
- Discourse/Cohesion Features
  - Measures of semantic connections and flow between sentences (e.g., sentence similarity, causal connectives)
- Readability Features
  - Grade-level readability scores (e.g., Automated Readability Index, Coleman-Liau, Dale-Chall, Flesch-Kincaid, Gunning Fog)

Input Text

Stem + Stimulus	Stem + Stimulus + Options	Stem + Stimulus + Options + Skills
Which choice completes the text with the most logical and precise word or phrase? On the basis of extensive calculations and models, astronomers in the 1990s predicted that the collision of two neutron stars or a neutron star and a black hole could release a massive burst of gamma rays in an event called a kilonova. This _____ was confirmed with observations in 2017.	Question: Which choice completes the text with the most logical and precise word or phrase? Text: On the basis of extensive calculations and models, astronomers in the 1990s predicted that the collision of two neutron stars or a neutron star and a black hole could release a massive burst of gamma rays in an event called a kilonova. This _____ was confirmed with observations in 2017. Correct answer : theory Wrong answer1: evidence Wrong answer2: constant Wrong answer3: experiment	Question: Which choice completes the text with the most logical and precise word or phrase? Text: On the basis of extensive calculations and models, astronomers in the 1990s predicted that the collision of two neutron stars or a neutron star and a black hole could release a massive burst of gamma rays in an event called a kilonova. This _____ was confirmed with observations in 2017. Correct answer : theory Wrong answer1: evidence Wrong answer2: constant Wrong answer3: experiment Primary_content_classification: Craft and Structure Secondary_content_classification: Words in Context Tertiary_content_classification: Context-based Completion Context_classification: Science Subcontext_classification: Earth and Space Science

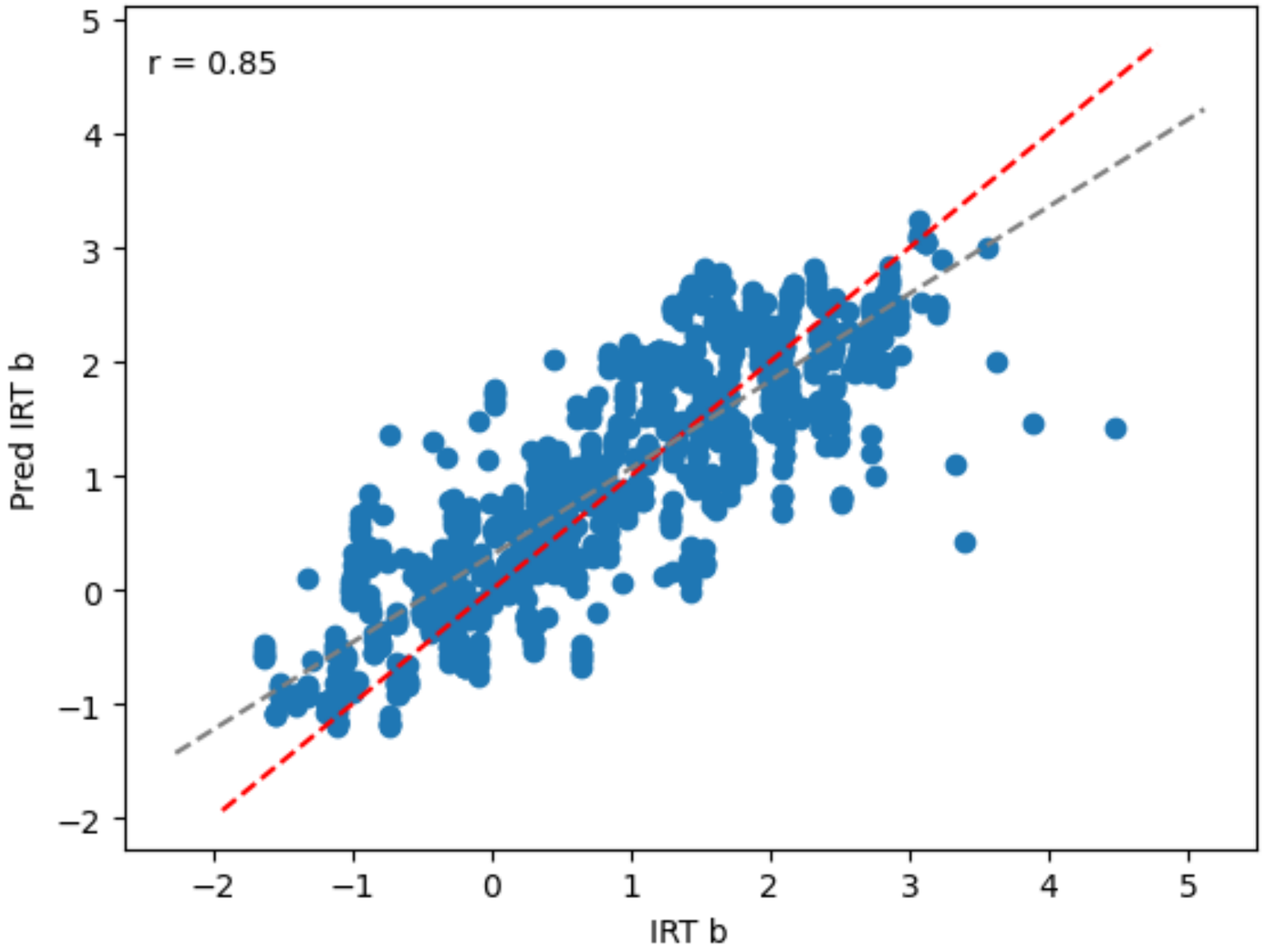
Model Evaluation

Input Text	Step 1 Model	Step 2 Model	Features	RMSE	R2	Corr Observed, Pred
Stem + Stimulus	N/A (Baseline)	Ridge Regression	Linguistic			
	N/A (Baseline)	XGBoost	Linguistic			
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score			
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score			
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score			
	ModernBERT	XGBoost	Linguistic + ModernBERT predicted score			
Stem + Stimulus + Options	N/A (Baseline)	Ridge Regression	Linguistic			
	N/A (Baseline)	XGBoost	Linguistic			
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score			
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score			
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score			
	ModernBERT	XGBoost	Linguistic + ModernBERT predicted score			
Stem + Stimulus + Options + Skills	N/A (Baseline)	Ridge Regression	Linguistic			
	N/A (Baseline)	XGBoost	Linguistic			
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score			
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score			
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score			
	ModernBERT	XGBoost	Linguistic + ModernBERT predicted score			

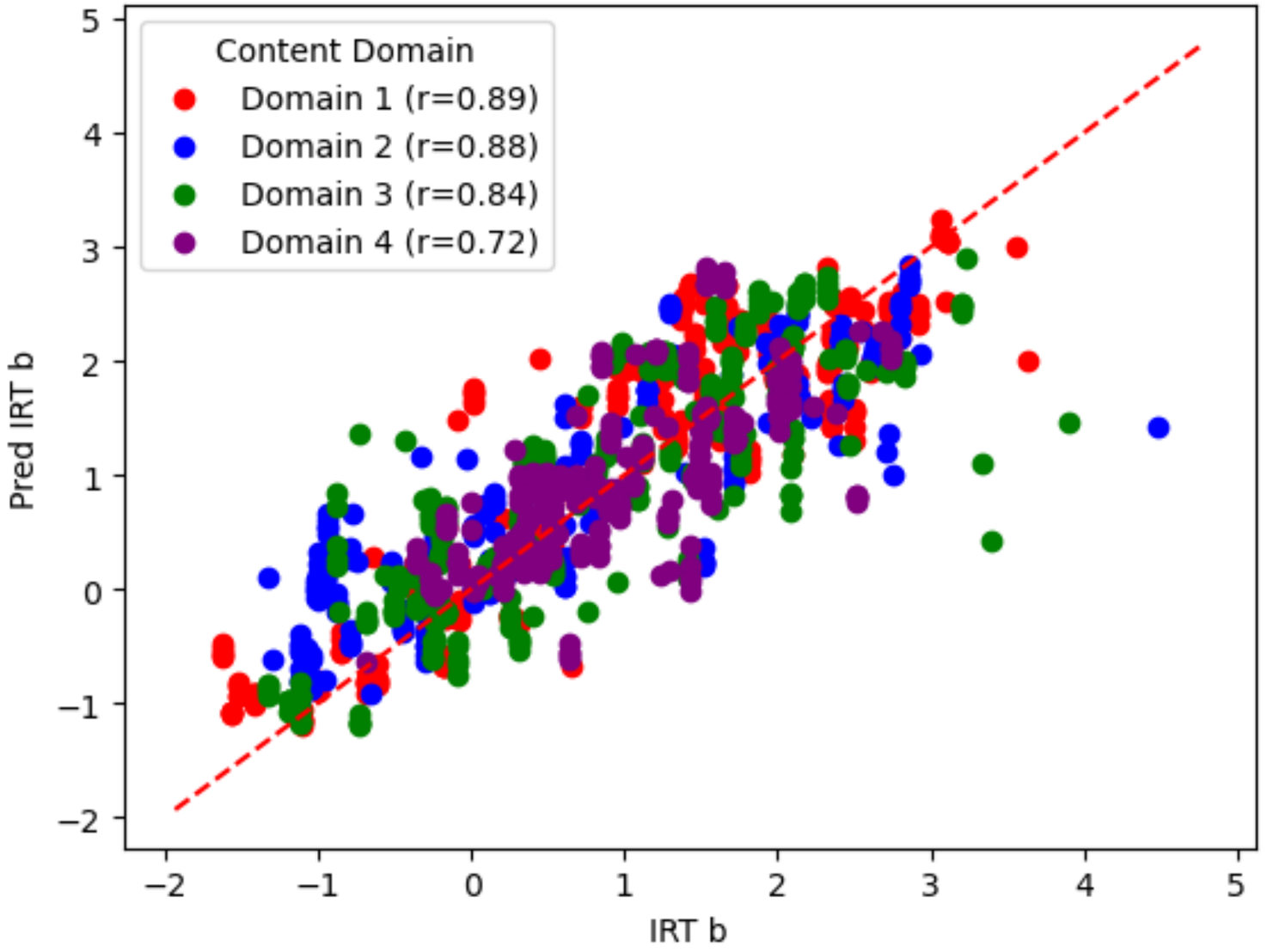
Results

Input Text	Step 1 Model	Step 2 Model	Features	RMSE	R2	Corr Observed, Pred
Stem + Stimulus	N/A (Baseline)	Ridge Regression	Linguistic	0.921	0.369	0.631
	N/A (Baseline)	XGBoost	Linguistic	0.897	0.402	0.674
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score	0.697	0.639	0.802
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score	0.695	0.640	0.803
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score	0.675	0.661	0.817
	ModernBERT	XGBoost	Linguistic + ModernBERT predicted score	0.676	0.659	0.816
Stem + Stimulus + Options	N/A (Baseline)	Ridge Regression	Linguistic	0.876	0.429	0.673
	N/A (Baseline)	XGBoost	Linguistic	0.876	0.429	0.687
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score	0.629	0.705	0.842
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score	0.636	0.699	0.840
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score	0.615	0.718	0.852
	ModernBERT	**XGBoost**	**Linguistic + ModernBERT predicted score**	**0.614**	**0.720**	**0.853**
Stem + Stimulus + Options + Skills	N/A (Baseline)	Ridge Regression	Linguistic	0.910	0.383	0.643
	N/A (Baseline)	XGBoost	Linguistic	0.910	0.384	0.646
	RoBERTa	Ridge Regression	Linguistic + RoBERTa predicted score	0.648	0.687	0.835
	RoBERTa	XGBoost	Linguistic + RoBERTa predicted score	0.653	0.682	0.833
	ModernBERT	Ridge Regression	Linguistic + ModernBERT predicted score	0.619	0.715	0.849
	ModernBERT	XGBoost	Linguistic + ModernBERT predicted score	0.620	0.714	0.849

Prediction Results

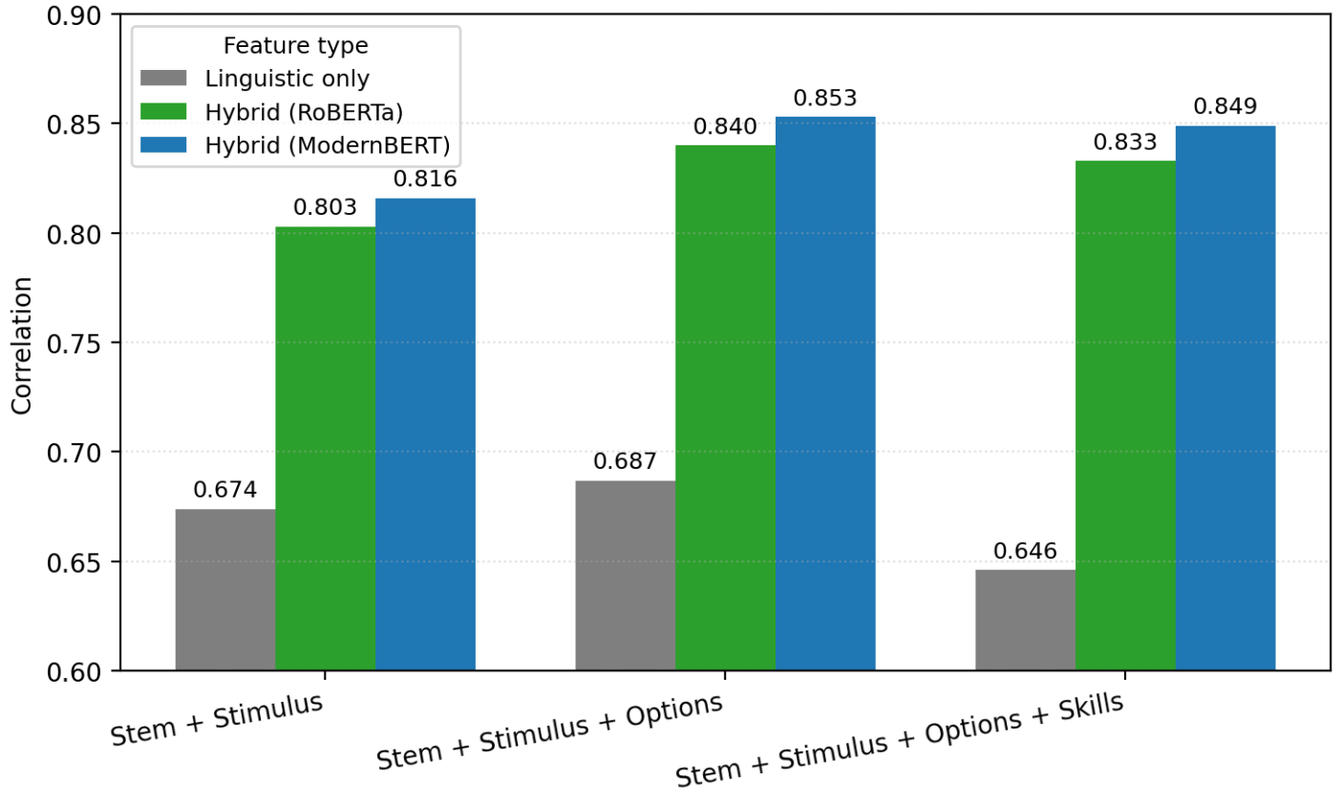


Prediction Results by Content Domain



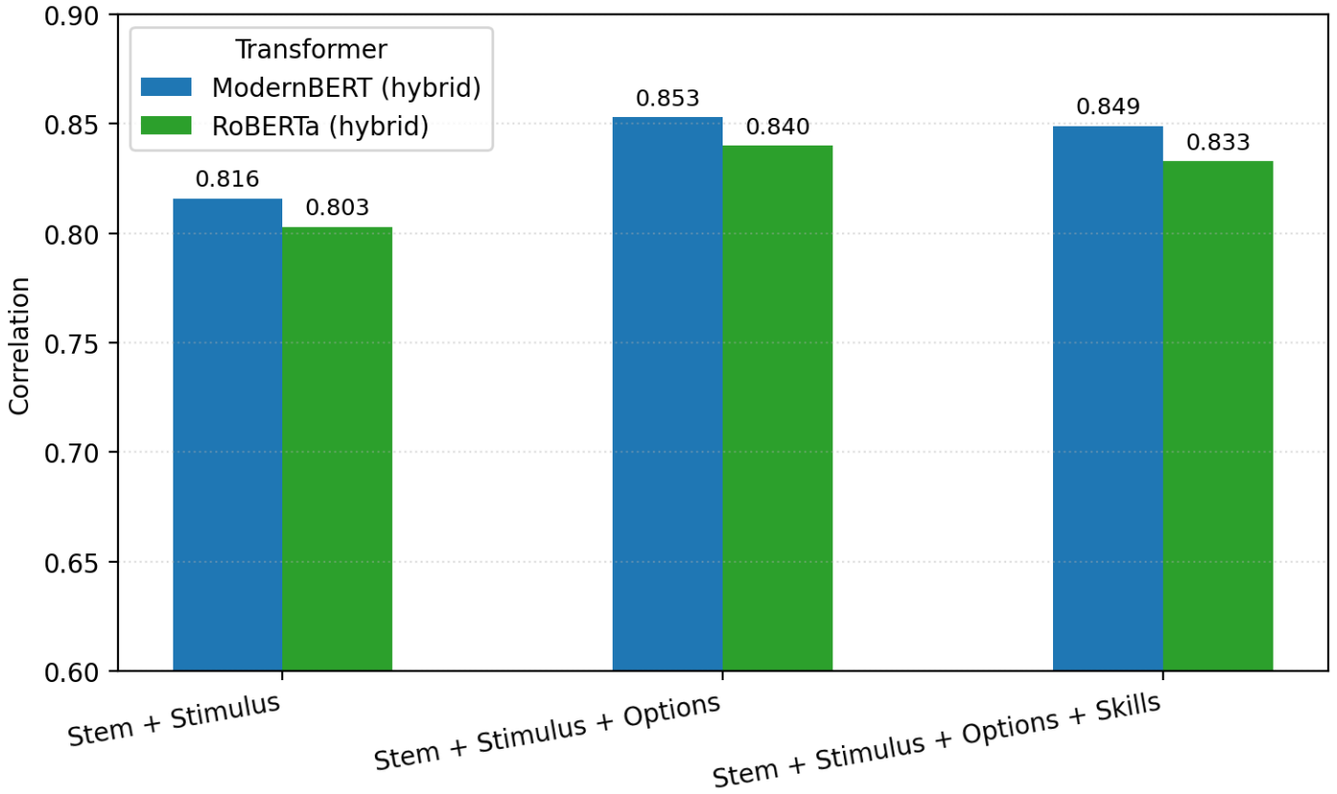
Q1 : Linguistic-feature-only model vs hybrid approach

- Does a hybrid approach improve prediction accuracy compared with a linguistic-feature-only model?
  - Yes, Two-step hybrid approach improved predictive accuracy



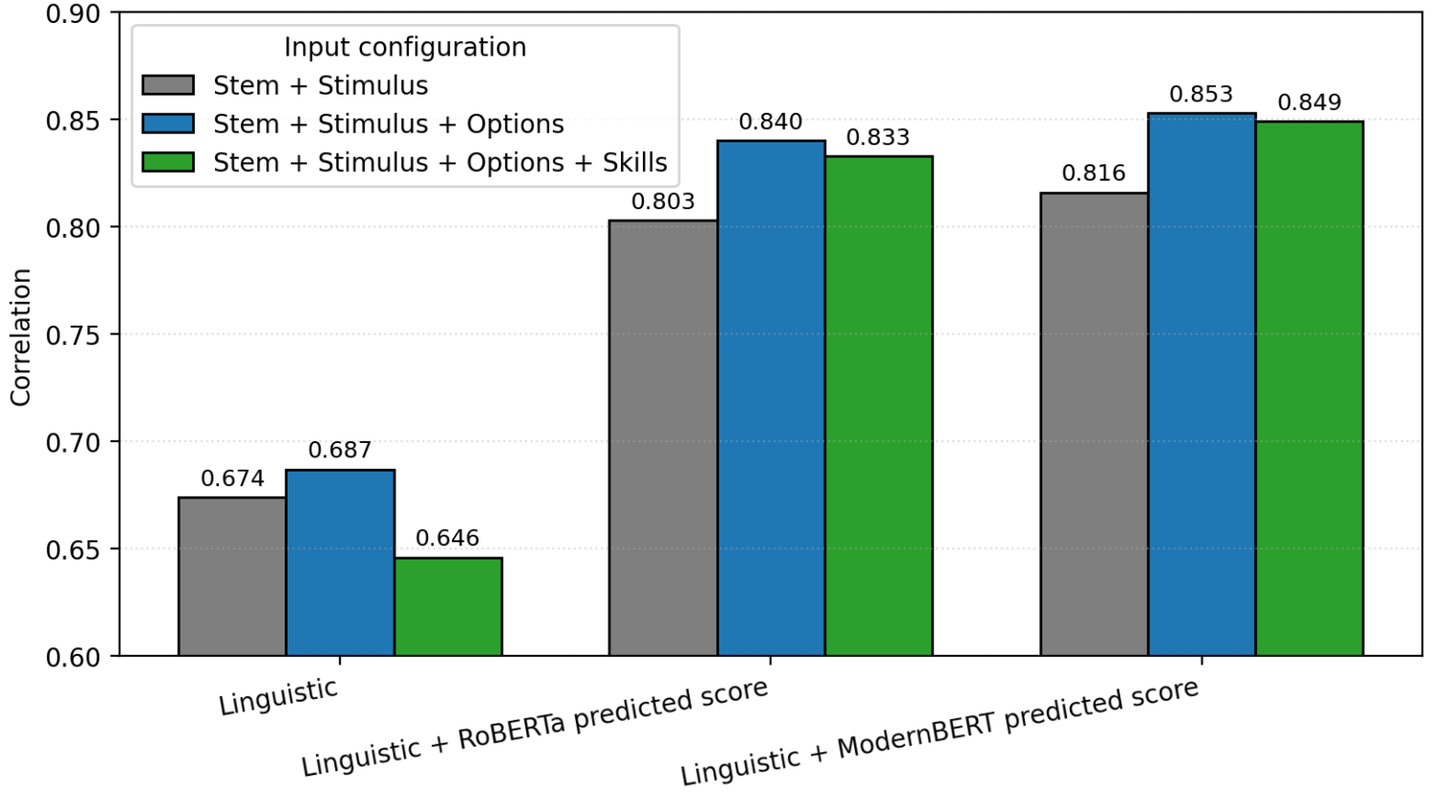
Q2 : RoBERTa vs. ModernBERT

- Which transformer model performs better for this task ?
  - ModernBERT slightly outperformed RoBERTa



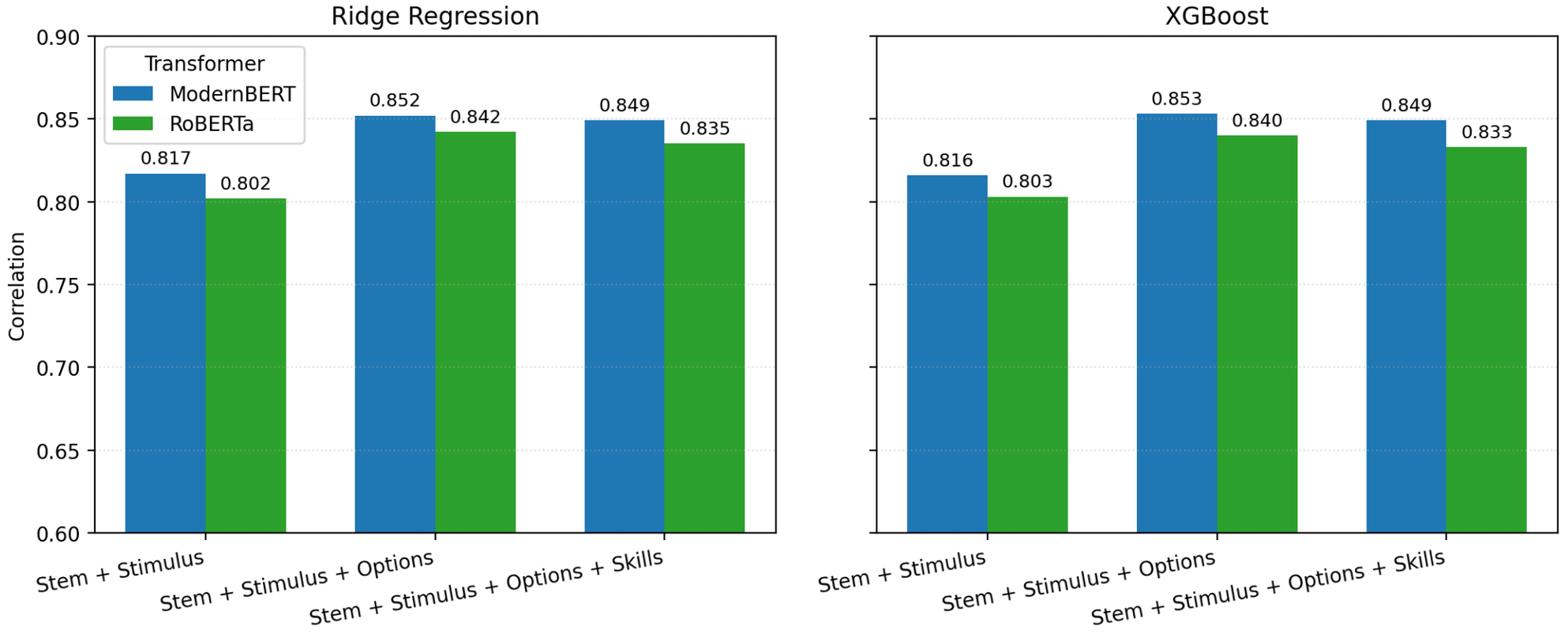
Q3 : Stem + Stimulus vs. Stem + Stimulus + Options

- Which input configuration produces higher accuracy?
  - Input text with stem, stimulus and options improved accuracy versus stem + stimulus only



Q4 : Ridge vs. XGBoost

- Which ensemble model performs better for this task ?
  - Both models had very similar accuracy



Summary and Discussion

- Summary
  - Two-step hybrid approach improved predictive accuracy
  - Transformer model comparison: ModernBERT slightly outperformed RoBERTa
  - Input text with stem, stimulus and options improved accuracy
  - Ensemble model comparison : XGBoost and Ridge Regression had very similar accuracy
  - Best model: two-step hybrid with ModernBERT and input text = stem + stimulus + options
- Limitations
  - Modest sample size
  - Only two transformer encoders evaluated
- Future work
  - Expand the study by collecting more training data and evaluating additional models on independent test sets, including DeBERTa, LLaMA, and GPT variants
  - Add features derived from generative AI models

Reference

- Kim, Y. K., & Moses, T. (2025, April). *Incorporating NLP features into transformer-based AES models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO, United States.
- Uto, M., Xie, Y., & Ueno, M. (2020, December). *Neural automated essay scoring incorporating handcrafted features*. In Proceedings of the 28th international conference on computational linguistics (pp. 6077-6088).

Thank You!

Questions?

ykim@collegeboard.org

CollegeBoard