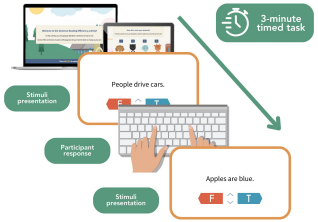
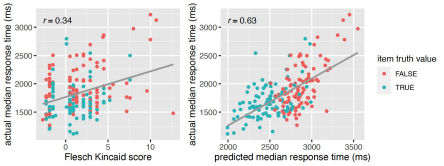
Generating and Evaluating Tests for K-12 Students with Language Model Simulations: A Case Study on Sentence Reading Efficiency

Eric Zelikman*, Wanjing Anya Ma*, Jasmine E. Tran, Diyi Yang, Jason D. Yeatman, Nick Haber

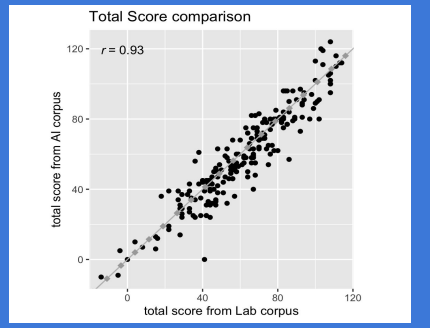
Developing expert-written educational tests is expensive and time-consuming, and determining the quality and difficulty of the test items requires collecting hundreds of student responses. Many tests require multiple administrations to monitor students' progress – ***parallel tests.*** In this study, we focus on tests of silent sentence reading efficiency, used to assess students' reading ability over time



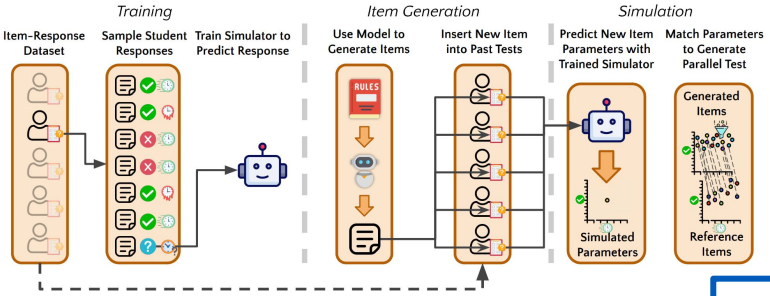
Item Response Simulator predicts response time better than readability metrics



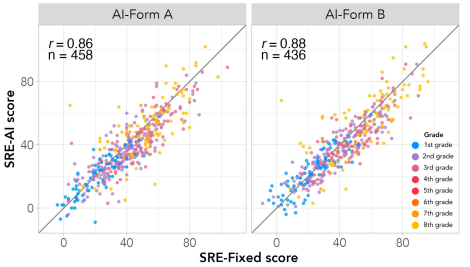
Our evaluation of a generated test with 234 students from grades 2 to 8 produces test scores highly correlated (r=0.93) to those of a standard test form written by human experts and evaluated across thousands of K-12 students.



We propose to fine-tune LLMs to simulate how previous students would have responded to unseen items. With these simulated responses, we can estimate each item's difficulty and ambiguity. We then solve unbalanced optimal transport for parallel test.



Follow-up validation study confirms the reliability of the generated AI forms across diverse student populations, suggesting the potential of more regular progress monitoring of reading ability in classrooms.



EMNLP Paper

Validation Paper

AI+Education Summit 2024
Website TBD: https://hai.stanford.edu/events/aieducation-summit

Advancing Human Learning with AI Technologies
2/1/2024 | Li Ka Shing Center, Stanford University