The First Workshop on Large Foundation Models for Educational Assessment

Sheng Li SHENGLI@VIRGINIA.EDU
Dongliang Guo DONGLIANG.GUO@VIRGINIA.EDU
Daiqing Qi DAIQING.QI@VIRGINIA.EDU
School of Data Science, University of Virginia

Zhongmin Cui ZHONGMIN.CUI@CFAINSTITUTE.ORG
CFA Institute

Jiasen Lu JIASEN_LU@APPLE.COM
Apple Inc.

Deborah Harris DEBORAH-HARRIS@UIOWA.EDU
College of Education, University of Iowa

Shumin Jing SHUMIN317@GMAIL.COM
Smarter Balanced

Preface

Advanced generative artificial intelligence (AI) techniques, such as large language models and large multimodal models, are transforming many aspects of educational assessment. The integration of AI into education has the potential to revolutionize not only the test development and evaluation but also the way students can learn. Over the past years, some successful adoptions of machine learning in this area have been using natural language processing for automated scoring, or applying collaborative filtering to predict student responses. Rapid advances in large foundation models (e.g., ChatGPT, GPT-4, Llama, Gemini, DeepSeek) demonstrate the potential of intelligent assessment with data-driven AI systems. These models could potentially benefit test construct identification, automatic item generation, multimodal item design, automated scoring, test security, and assessment administration. Meanwhile, new research challenges arise at the intersection of AI and educational assessments. For example, the explainability and accountability of current large foundations models are still inadequate to convince stakeholders in the educational ecosystem, which limits the adoption of AI techniques in large-scale assessments. In addition, it is still unclear whether large foundation models are capable of assisting complex assessment tasks that involve creative thinking or high-order reasoning. Tackling these research challenges would require collaborative efforts by researchers and practitioners in both AI and educational assessment.

This one-day workshop provided a forum for researchers in AI and educational assessment to review and discuss the recent advances in the application of large foundation models

for educational assessment. The workshop received 23 submissions. Each submission was reviewed by at least three domain experts in AI and educational assessment. Finally, nine papers were accepted for oral presentations, and eight papers were accepted for poster presentations. These papers cover the latest research efforts on applying large foundation models to various assessment tasks, such as automated feedback generation for open-ended questions, generating reading assessment passages using large language models (LLMs), and evaluating LLM robustness with incorrect multiple-choice options.

The workshop features five keynote speakers who are renowned experts in AI and educational assessment, from academia and industry. Professor Diyi Yang from the Stanford University shared her recent works on social skill training using LLMs. Professor Vered Shwartz from the University of British Columbia discussed several factors that need to be considered when deciding whether and how to utilize LLMs for grading students' exams. Professor Hong Jiao from the University of Maryland shared some successful use cases of AI in test development, such as automated scoring, cheating detection, and process data analysis. She also discussed additional possibilities and opportunities that AI can bring to enhance educational assessment practices, such as generative AI for item generation and item parameter prediction modeling. Dr. Susan Lottridge from the Cambium Assessment discussed the key elements and steps her team took when designing and building a writing feedback tool called "Write On with Cambi!". This tool walks students through reviewing their essay using structured feedback. Dr. James Sharpnack from Duolingo discussed the construction of the Duolingo English Test (DET), a high-stakes, large-scale, fully online English language proficiency test built using human-in-the-loop (HIL) AI.

As a final note, the editors of this workshop proceeding would like to thank the reviewers who contributed to the review process and provided valuable feedback to authors. Without their help this proceeding would not have been possible.