# AIME-Con 2-Hour Training Sessions: Oct 27 from 12:00–2:00pm

### 1. Getting Started with LLM Evaluation: A Primer for Psychometricians

***Presenter:*** Jodi Casabianca

***Brief description:*** This session introduces psychometricians and assessment scientists to evaluation methods for large language model (LLM) applications in education. Through lecture and hands-on activities, participants will explore evaluation pipelines and techniques—including error analysis, human review, and LLM-as-a-judge—demonstrating how psychometric principles can enhance rigor, validity, and interpretability in pipeline design.

***Presenter bios:*** Jodi Casabianca is Founder and Chief Scientist at BroadMetrics, a psychometrics and assessment consultancy that supports organizations, schools, and researchers to develop and evaluate their assessment systems. She previously worked as a senior measurement scientist at Educational Testing Service (ETS), where she contributed to psychometric efforts for various assessments, including TOEFL, GRE, Praxis, and supported K-12 clients. Much of her work has focused on the design, implementation, and evaluation of constructed response scoring systems in educational assessment, including AI scoring systems. Building on years of experience evaluating AI scoring models in high-stakes educational testing, Casabianca now focuses on applying psychometric methods to support the evaluation of LLMs in educational settings.

***Learning Objectives:*** This session will provide an overview of the different steps and analyses in the evaluation pipeline for LLM applications. Though we will start with a brief overview of LLMs to establish naming conventions and possible causes for errors, etc., participants should already have at least a general understanding of them.

By the end of this session, participants will be able to:

1. Summarize the foundational capabilities and limitations of LLMs to support accurate interpretation of evaluation results.

2. Deconstruct the key components of a task-specific LLM application, with attention to input/output structure, task prompts, and scoring logic.

3. Describe the core steps in an LLM evaluation pipeline, using education-focused use cases to illustrate various approaches.

4. Conduct basic trace-based error analysis and apply qualitative coding techniques to systematically identify and categorize common LLM errors.

5. Determine when it is appropriate to incorporate standardized tests or benchmark tasks into LLM evaluation, grounded in psychometric principles.

6. Identify and define key constructs relevant to LLM evaluation (e.g., factual accuracy, coherence, fairness, safety).

7. Explain the LLM-as-a-Judge method, including when it is appropriate to use and how to assess its reliability and alignment with human judgments.

8. Differentiate among types of expert human review, and design rating scales or rubrics for use in evaluation tasks.

9. Apply a principled, validity-driven design framework to develop a customized evaluation pipeline for a specific educational LLM use case, identifying appropriate sources of validity evidence.

***Software Requirements:*** N/A

---

**2. Creating Actionable Classroom Assessments: PLDs, Performance Tasks, and ChatGPT to the Rescue**

***Presenters:*** Bryan Drost & Char Shryock

***Brief description:*** Explore how Generative AI, used in conjunction with performance level descriptors, can transform classroom assessment design. This interactive session highlights AI's potential as a collaborative tool for creating authentic classroom assessments. Participants will engage in hands-on activities and receive tools to support research, analysis, and future classroom assessment development.

***Presenter bios:*** Bryan Drost, Ph.D. is a nationally recognized expert in curriculum, instruction, assessment, and technology integration with over two decades of experience in public education. He currently serves as an executive director for a school system in Northeast Ohio. In addition to his district leadership, Dr. Drost is the faculty lead at Ursuline College for the School Improvement Planning through Classroom Assessment and Data Analysis Certificate program. His instructional expertise spans both K–12 and higher education, and he is widely respected for his ability to bridge research and practice in meaningful ways. Dr. Drost serves as Co-Chair of the National Council on Measurement in Education (NCME) Classroom Assessment Committee and previously chaired the Standards and Test Use Committee. His work with NCME and the Ohio Department of Education has shaped state policy, professional learning frameworks, and classroom assessment practices nationally. He has contributed to the development of Ohio's model curricula, several large-scale assessment designs, and assessment literacy training across the country. Dr. Drost regularly speaks at conferences including ASCD, NCME, AMLE, and CoSN. His publications appear in Educational Leadership, Journal of Education, District Administration, Kappan, and more. As a committed advocate for instructional equity and innovation, Dr. Drost continues to support educators in designing assessment and instructional systems that improve outcomes for all learners.

Char Shryock is a nationally recognized expert in curriculum, instruction, and assessment, with over 35 years of experience as a teacher, technology specialist, Director of Curriculum, and

Superintendent. She has been selected to serve on multiple working committees with the Ohio Department of Education and Workforce, contributing to the development of state learning standards, model curriculum, and assessment systems based on her deep experience and practitioner insight. Char is a sought-after speaker and facilitator, presenting at national conferences including NCME, NSTA, ASCD, AMLE, and NCTE. Her recent publications in Kappan and IGI Global focus on the responsible use of Generative AI to improve instructional and assessment practices, reflecting her ongoing leadership at the intersection of innovation and evidence-based design. As a co-editor of ChatGPT: Navigating the Impact of Generative AI on Education Theory and Practice and co-author of multiple peer-reviewed articles, Char brings both practical insight and deep scholarly grounding to this session.

***Learning Objectives:***

1. Identify opportunities to integrate Generative AI into the classroom assessment design process.

2. Evaluate AI-generated classroom assessment ideas for alignment with learning goals and PLDs.

3. Refine classroom assessment tasks using AI and peer feedback, and evaluate benefits over traditional design methods.

4. Analyze performance level descriptors to guide classroom task development.

***Software Requirements:*** N/A

---

**3. Designing for Variability: AI-Driven Innovation to Facilitate Formative Assessment and Personalize Learning**

***Presenters:*** Samantha Goldman & Sean Smith

***Brief description:*** This session explores how AI supports teachers in conducting formative assessments and implementing personalized learning to meet the needs of all students. Participants will use hands-on tools and prompting strategies to generate adaptive supports and inform instruction, positioning AI as a thought partner in addressing learner variability and data-informed teaching.

***Presenter Bios:*** The presenters bring a combined depth of experience in special education, instructional design, and educational technology, with a strong emphasis on leveraging AI to support personalized learning and address learner variability. Their work spans K-12 and higher education, including leadership roles in federally funded research projects focused on AI-supported writing, virtual reality for social-emotional learning, digital tools for instructional planning and progress monitoring, and technology integration.

They have authored numerous peer-reviewed publications on topics such as Universal Design for Learning, AI integration in teacher preparation, and technology-supported writing strategies. Their recent work includes the development of an AI-driven feedback system designed to automatically score writing and provide personalized feedback on content, organization, and style to students with high incidence disabilities and their teachers. Their research has been recognized with national awards, and they frequently present at major conferences in special education and educational technology.

***Learning Objectives:***

1. Analyze how AI tools can generate formative feedback that informs instruction and supports effective educational practices, particularly for students with disabilities (SWDs).

2. Apply evidence-informed AI strategies to personalize instruction and reduce barriers in areas such as reading comprehension, executive function, and social communication, demonstrating measurable improvements in student engagement and performance.

3. Design and refine prompts that elicit actionable, student-specific feedback from AI systems, positioning prompt engineering as a core digital literacy skill in inclusive teaching and assessment.

***Software Requirements:*** N/A

---

**4. Deep Learning Model for Unstructured Data in Educational Assessment**

***Presenters:*** Mo Zhang, Akshay Badola, Andrew Hoang, Chen Li, & Hongwen Guo

***Brief description:*** This workshop presents transformer-based deep learning approaches for modeling unstructured process data in educational assessments, such as keystroke logs and click-stream data. Participants will gain hands-on experience applying these models, bypassing manual feature engineering, to enable more scalable, flexible, and robust analysis and applications.

***Presenter bios:*** The presenters are a multidisciplinary research team with extensive experience in psychometrics, natural language processing (NLP), computer science, statistics, and the application of artificial intelligence in educational assessment. Each team member has actively contributed to this line of research. The lead presenter, Dr. Mo Zhang, is a Senior Research Scientist at ETS. Dr. Zhang holds a Ph.D. in Educational Psychology. Her scientific work focuses on AI scoring and psychometric modeling of educational process data. She currently holds seven U.S. patents and has published widely in the field of educational measurement.

***Learning Objectives:***

This workshop introduces a generalizable methodology for modeling various types of unstructured data in educational assessments. Using click-stream data and keystroke logs as case

studies, we will demonstrate how adaptations of the transformer architecture can be applied to train language models that generate dense representations of log files commonly encountered in educational assessments. By the end of the workshop, attendees will have a good understanding of the approach and will be equipped to apply our provided code repositories to their own data.

***Software Requirements:*** Basic familiarity with Python programming is recommended. Hands-on activities and example scripts will be presented in Python.

---

**5. Introduction to AI Scoring in Python**

***Presenter:*** Christopher Ormerod

***Brief description:*** This workshop is an introduction to automated scoring using Python. We cover a range of methods from traditional frequency-based approaches to scoring using generative AI.

***Presenter bio:*** Dr. Christopher Ormerod has a PhD in Applied Mathematics and is currently a Principal Data Scientist at Cambium Assessment, where he leads a team that researches applications of AI to education and assessment. He has written several foundational papers on the applications of Large Language models to educational assessment and wrote the Python backend for Cambium Assessments automated scoring platform, formative feedback system, and speech scoring system. He was a member of one of the winning teams in the National Assessment of Educational Progress Automated Scoring Competition and currently cochairs the AIME group.

***Learning Objectives***

This hands-on workshop introduces participants to automated scoring using Python and machine learning techniques. Course materials include a code repository accessible through Google Colab notebooks.

By completing this workshop, participants will be able to:

*Master Core Concepts:* Grasp essential machine learning principles, particularly text classification methods and language models used in automated scoring applications

*Develop Technical Skills:* Install and effectively use Python libraries essential for machine learning implementation

*Navigate the complete machine learning workflow:* loading data, saving models, training algorithms, and calling text classifiers

*Apply Advanced Techniques:* Implement a range of text classification approaches, from traditional frequency-based methods (such as TF-IDF and n-grams) to modern fine-tuned language models.

***Software Requirements:*** Colab notebooks allow the participants to access and run code freely through a web interface. There are no installation or payment requirements.

---

**6. Designing and Evaluating Generative AI Simulations to Support Teacher Learning**

***Presenters:*** Beata Beigman Klebanov & Jaime Mikeska

***Brief description:*** In this training session, participants will learn about the design of a generative AI (GenAI) teaching simulation to support elementary teachers in learning how to elicit and attend to student thinking. They will try out the simulation, attempt to improve it, and learn about ways to evaluate the simulation responses.

***Presenter Bios:*** Drs. Mikeska and Beigman Klebanov co-lead a multi-year project on automating aspects of digital teaching simulations.

Dr. Mikeska is a Managing Principal Scientist in the ETS Research Institute. Her expertise is teacher education and development. She co-leads the ETS Teaching Pathways research program and has led several research projects on developing and using human-powered digital teaching simulations in teacher education.

Dr. Beigman Klebanov is a Principal Scientist in the ETS Research Institute. Her expertise is natural language processing in the context of educational applications. She has worked extensively on automated evaluation of language production, including essays in various genres, oral reading, and simulated classroom discussions.

***Learning Objectives:*** By participating in the training session, participants will:

1) Understand and value the importance of digital teaching simulations to support teachers' learning of key teaching competencies in math and science.

2) Learn about the design features of a GenAI teaching simulation and how it can be used to develop teachers' ability to elicit and attend to student thinking.

3) Understand through experience some of the main challenges in designing a teaching simulation with GenAI playing the student, including the tension between the elicitation construct and AI's 'eagerness-to-please' and the need to maintain a coherent 'knowledge boundary' for the student, including misconceptions.

***Software Requirements:*** Free software accessible via an ETS web app. No installation or payment requirements.

# AIME-Con 4-Hour Training Sessions: Oct 27 from 2:30–6:30pm

### 1. Introduction to AI-based Automated Item Generation and Automated Scoring

***Presenters:*** Duanli Yan & Alina Von Davier

***Brief description:*** This training session offers 1) an overview of the latest NLP techniques and large language models for automated scoring, alongside psychometric principles and practices for test development, 2) an overview of the design, development, evaluation, and quality control of automated scoring systems for practitioners on the applications of these systems.

***Presenter bios:*** Dr. Alina A. von Davier is the Chief of Assessment, Duolingo, where she leads the Duolingo English Test research and development area. She is also the Founder and CEO of EdAstra Tech, a service-oriented EdTech company. She is a researcher, innovator, and executive leader in the field of computational psychometrics, machine learning, and education.

Dr. Duanli Yan is an innovative researcher and developer in the field of modern psychometrics and educational measurement. She served as the director of data analysis and computational research in the research and development division in ETS, responsible for automated scoring engine upgrade evaluations. She has extensive experience in adaptive testing, psychometric research, test security, innovative technology research and implementations. Dr. Yan is also an adjunct professor at Fordham University and at Rutgers University, the State University of New Jersey.

They published extensively including many books and peer reviewed journals. They are co-editors for volume Computerized Multistage Testing: Theory and Applications (2014) which won 2016 AERA Division D Significant Contribution to Educational Measurement and Research Methodology award, Research for Practical Issues and Solutions in Computerized Multistage Testing (2024), co-authored Computerized Adaptive and Multistage Testing with R (2017). Dr. von Davier is a co-editor for Computational Psychometrics: New methodologies for a new generation of digital learning and assessment. Dr. Duanli Yan is a co-editor for Handbook of Automated Scoring: Theory into Practices, and Handbook of Research on Science Learning Progressions.

***Learning Objectives:***

This training offers a comprehensive overview to the many facets of automated item generation (AIG) and automated scoring (AS) in adaptive testing, drawing from real-world operational practices, published papers (Attali et al, 2022; von Davier et al., 2024) and the edited volume Handbook of Automated Scoring: Theory into Practice (Yan, Rupp, & Foltz, 2020). Participants will be guided through all operational aspects of AIG and AS, from design to implementation.

1. Demystifying the Black Box: The participants will gain an in-depth understanding of the various methods used for generating items and constructing automated scoring systems for

evaluations. We will discuss a human-in-the-loop approach to automation and the value of preserving human values in highly automated systems.

2. Implementing Systems in Operational Practice: Designing and implementing AIG and AS is crucial for educational assessments nowadays. However, transitioning them into operational systems and deploying them requires a more complex process, involving different implementation models and associated procedures. Participants will learn preprocessing textual data, filtering unsoarable essays and diverting them to hand-scoring, model building, score assignment, and reporting.

3. Evaluating and Maintaining Systems Over Time: The participants will learn various approaches to assessing the performance of AIG and AS systems, including comparisons to human test development and scoring using evaluation metrics, as well as managing system changes in operational practices. See Analytics for Quality Assurance in Assessment (von Davier, Liao, et al., 2022) for an example of such a system.

4. Exploring Open Issues, Future Directions, and Engaging in General Discussion: The participants will learn applications of AIG and AS, discuss potential challenges in implementation and the requirements for advancing the field with AI, NLP, psychometrics, ethics, and human values in strengthening the operational use of AIG and AS.

***Software Requirements:*** Laptop with ChatGPT

---

**2. Fundamentals of Generative AI for Item Development**

***Presenters:*** Henry Makinde, Hope Adegoke, & Mubarak Mojoyinola

***Brief description:*** This interactive session introduces educators and measurement professionals to generative AI techniques, including prompt engineering, fine-tuning, retrieval-augmented generation (RAG), and agentic systems. Participants will develop practical skills for using large language models (LLMs) for automated item generation and gain hands-on experience implementing these techniques in Python using such libraries like TensorFlow and PyTorch as well as Hugging Face Transformers and vector databases–Python should be pre-installed.

***Presenter bios:***

*Lead Presenter: Henry Sanmi Makinde, Ph.D. Candidate*

Henry Makinde is a Ph.D. candidate in Educational Research Methodology with a minor in Computational Statistics at UNC Greensboro, specializing in AI-driven educational assessment innovation. During his internship with the National Commission on Certification of Physician Assistants (NCCPA), he successfully deployed LLMs for automated item generation, developing Chain-of-Thought frameworks and RAG systems. Henry presented his findings directly to the NCCPA Board of Directors and also at NCME Conference 2025. As a PhD candidate, Henry is

currently researching "Automatic Item Development using LLM." While serving as Intern Graduate Research Assistant with Guilford County Schools, he built data collection and reporting systems for over 70,000 students. He is currently working as an intern for the Physician Assistant Education Association (PAEA) where he performs daily psychometrics tasks, works on multiple projects, and serves on the standard settings committee. Henry has technical expertise in Python, R, PowerBI, and Hugging Face Transformers, with specialized experience in prompting strategies, RAG systems, and LLM fine-tuning for educational applications. He currently serves as President of the Educational Research Methodology Graduate Students Association and has extensive experience translating complex AI concepts into accessible knowledge through presentations at NCME, NCARE, IMPS and institutional workshops.

*Hope Oluwaseun Adegoke, M.Sc, Phd Student*

Hope Oluwaseun Adegoke is a Ph.D. student in the Educational Research Methodology program at the University of North Carolina at Greensboro. He holds a B.Sc. and M.Sc. in Statistics, as well as a second M.Sc. in Educational Research Methodology. With over five years of professional experience as a data manager and analyst, Hope brings a strong applied background in data science and educational measurement. His current research interest is aimed at leveraging AI to enhance assessment practices. His technical expertise spans supervised machine learning, large language model (LLM) fine-tuning, and Retrieval-Augmented Generation (RAG) systems. He has proficiency in R, Python, and other analytic software. Currently serving as a psychometric graduate assistant at the American Board of Pediatrics, where he contributes to multiple AI-driven initiatives aimed at advancing assessment practices.

*Mubarak Olumide Mojoyinola, M.Sc., Ph.D. Candidate*

Mubarak is a Ph.D. candidate in Educational Measurement and Statistics at the University of Iowa, specializing in the intersection of psychometrics and data science methods. He holds dual master's degrees in Statistics and Data Science, and a B.Sc. in Statistics.

His research focuses on applied statistics, psychometric modeling, and machine learning applications in educational assessment. His recent work leverages large language models (LLMs) for item difficulty prediction, developing innovative approaches to enhance test development processes for achievement and certification assessments. His work bridges traditional psychometric theory with cutting-edge AI technologies to advance the field of educational measurement.

***Learning Objectives:*** Participants will:

1. Understand core concepts of generative AI.

2. Develop hands-on skills in automated item generation using AI techniques.

3. Explore prompt engineering, fine-tuning, retrieval-augmented generation (RAG), and AI agent frameworks.

4. Learn to effectively integrate Python for implementing AI solutions in educational measurement specifically Item generation.

***Software Requirements:*** Participants are encouraged to bring laptops pre-installed with Python (Jupyter Notebook, Spider, Anaconda, etc.). All the software are open-source and free. However, we will be making some API calls that require us to pay for tokens, hence we will need to create an OpenAI account then fund our accounts with at least $10. No specific hardware requirements beyond standard laptops.

---

**3. Using Generative AI For Item Construction: State-Of-The-Art and Practical Lessons**

***Presenters:*** Sergio Araneda, Chris Foster, & Susan Weaver

***Brief description:*** This hands-on session explores state-of-the-art uses of generative AI for item construction. Participants will learn key principles, validation strategies, and practical techniques—including prompt engineering and scalable workflows—to effectively generate, manage, and validate large item pools using LLMs. Includes demos, discussions, and current research insights. Participants will use Scorpion, Caveon's platform for Test Delivery–no coding or software installation required.

***Presenter bios:*** Sergio Araneda is a research scientist at Caveon, Ph.D. in Psychometrics and Educational Measurement from University of Massachusetts Amherst, and Mathematical Civil Engineer from Universidad de Chile. He specializes in validity, item generation and test security, and has presented extensively and published on the integration of AI into assessment ecosystems. Some of his work includes a case study about the use of LLMs for item generation, a validity framework based on philosophy of technology, and an experiential approach to validation. Sergio also has deep experience in hands-on demonstration, item pool construction, and the application of philosophical frameworks to educational technology.

Susan Weaver is Director of Caveon's Secure Exam Development Services℠, where she leads efforts to integrate test security throughout the entire exam lifecycle—from program design and item development to data analysis and delivery. With over 25 years of experience in high-stakes testing, she brings deep expertise in tailoring assessment solutions to the unique needs of diverse testing programs. Susan has been a pioneer in applying prompt engineering techniques to AI-assisted item generation, and her work focuses on maintaining psychometric integrity and content security in rapidly evolving assessment contexts. She holds an M.S. in Administration from Central Michigan University and a B.B.A. from Baker University.

Christopher Foster is our Senior Data Scientist at Caveon and also a Ph.D. in Psychometrics and Educational Measurement from University of Massachusetts Amherst. As an Assessment Technology specialist with expertise in workflow automation, AI integration, and large-scale content delivery; He has led multiple projects focused on scaling AI-assisted item generation,

including the development of the Scorpion AI Assist, a tool to facilitate test construction using AI. His practical experience complements the theoretical and psychometric focus of the other presenters.

Together, the team brings a unique blend of psychometric, technological, and philosophical expertise to guide participants through a rich, structured, and practical learning experience.

***Learning Objectives:***

This training session is designed to help educational professionals, psychometricians, and AI practitioners:

1. Understand the theoretical foundations and current research behind the use of large language models (LLMs) in test item construction.

2. Identify and apply emerging best practices in AI-assisted item generation, including prompt engineering, content validation, and ethical considerations.

3. Gain hands-on experience using generative AI tools to construct items across domains and difficulty levels.

4. Learn how to manage and scale up item generation workflows, including the use of AI to build and maintain large item pools or virtual item banks.

5. Critically assess the role of LLMs in item development pipelines and learn frameworks for validating their use in test design.

By the end of the session, participants will have both conceptual and practical knowledge to responsibly and effectively use generative AI in item development.

***Software Requirements:*** No software or coding required. Attendees will need to create an account in Scorpion to access the projects for the various hands-on sections. This should be completed in advance of the session. There will be no fee associated with the creation of the account in Scorpion.

---

**4. Integrating Generative AI into R Workflows: From APIs to Shiny Apps**

***Presenter:*** Christopher Runyon

***Brief description:*** Learn to integrate large language models into R workflows through hands-on practice. Starting with architecture fundamentals, participants understand the why behind best practices before implementing various prompt engineering techniques, API interactions, multi-agentic systems, and LLM-powered Shiny applications. Moderate R fluency required; no LLM experience assumed.

***Presenter bio:*** Christopher Runyon is a Senior Measurement Scientist at NBME, where he integrates generative AI models across multiple assessment functions including content development, automated scoring, novel assessment methods, and internal process improvements. In the last two years he has taught similar AI-focused workshops several times: "Generative AI Fundamentals" (six sessions) and "Using GenAI to Help Score Performance Assessments" (three sessions), and given a dozen presentations on AI integration in medical education assessment through NBME's speaker's bureau. He serves as co-chair of the "AI in Medical Education" special interest group within the Directors of Clinical Skills Education. He also was part of a team that taught an NCME Training Session on “Using R Shiny” in 2021.

***Learning Objectives:***

By the end of this workshop, participants will be able to:

- Explain key LLM architectural features that inform effective integration practices

- Apply prompt engineering principles to achieve consistent, reliable outputs in R workflows

- Implement both single-use and conversational API interactions with LLMs from R

- Design and deploy simple multi-agentic systems for complex tasks

- Build interactive Shiny applications that leverage LLM capabilities for end-user functionality

***Software Requirements:***

- R (version 4.4 or higher recommended)

- R Studio

- Access to a GenAI model (online chatbot version)

- GenAI model API key (Both free and paid options will be identified in course prep materials)

---

**5. What you Need to Know to Unlock your Generative Potential**

***Presenters:*** John Behrens & Peter Foltz

***Brief description:*** This workshop is an overview of the “psychology” and behavior of large language (and multi-modal) models, addressing methodological issues and applications to measurement research & practice. Includes hands-on activities with web based tools and demonstration with supplemental resources for post-class learning. Some activities require a free .gmail account.

***Presenter bios:*** John Behrens is Professor of the Practice, and Director, of Technology and Digital Studies, and Concurrent Professor of the Practice of Computer Science & Engineering, at the University of Notre Dame. Over his 20 year industry career, he led the development of

globally deployed intelligent learning and assessment solutions impacting many tens of millions of learners and published widely on the impacts of technology and advanced data science (including AI) on assessment theory and practice. Over the last 2 years he taught four semesters of “Generative AI in the Wild”, presented more than 30 workshops on generative AI to business leaders, and has been featured in the popular press, including Forbes and The Wall Street Journal, for his insights on AI and its educational and societal impacts. This extensive experience teaching about generative AI, combined with a marked sense of humor, brings a compelling energy to the classroom.

Peter Foltz is a Research Professor at the University of Colorado, Boulder Institute of Cognitive Science and Executive Director of the NSF AI Institute for Student AI Teaming. His work covers machine learning and natural language processing for educational assessments, large-scale data analytics, cognitive skills in reading writing, speaking, team collaboration, and 21st Century skills learning, He has worked in academic and industry edtech settings over the past 30 years leading innovative research and development teams resulting in the transition from ideas to proofs of concept to products. He serves on a number of corporate and academic advisory boards around incorporating AI methods into education, both for large-scale assessment and for supporting learning.

***Learning Objectives:***

Participants will learn the following skills with demonstration, discussion, and hands on practice:

1. How large language and multi-modal models work and differ

2. Understanding the implications of LLM/LMMs as probabilistic language-based cognition simulators.

3. Using Evidence Centered Design (ECD) to drive evaluation clarity

4. How to assess model long run behavior using simulation and measurement models

5. How to apply these methods to assess complex learner performances

6. Common tools, repositories, and common generative research and application patterns relevant to measurement professionals.

***Software Requirements:*** GoogleAI Studio and ChainForge