PROMETRIC

Implementation of AI-human hybrid tools for responsible content development

Languages Canada Conference
February 29, 2024

Michael Holaday

Jennifer Flasko

Introductions

Michael Holaday

VP Revenue, Language Proficiency &
Prior Learning Assessments

Jennifer Flasko

Senior Content Development Lead

PROMETRIC

- Prior learning Assessments
- Students test for credit on what they've already learned
- 37 titles: business, humanities, math, physical sciences, social sciences technology
- Accepted by over 1900 institutions

dsst® | GET COLLEGE CREDIT.COM

PROMETRIC

PROMETRIC

finetune

A PROMETRIC COMPANY

Overview

- **Led by the AI experts at Finetune, we have been exploring how to leverage generative AI to enhance content development for our tests**.

- **In today's presentation, we will discuss:**
  - Introduction to Generative AI
  - Limitations and benefits
  - General vs customized models
  - Best practices for AI integration
  - Steps to AI implementation – based on Generate pilot experiences

Generative AI

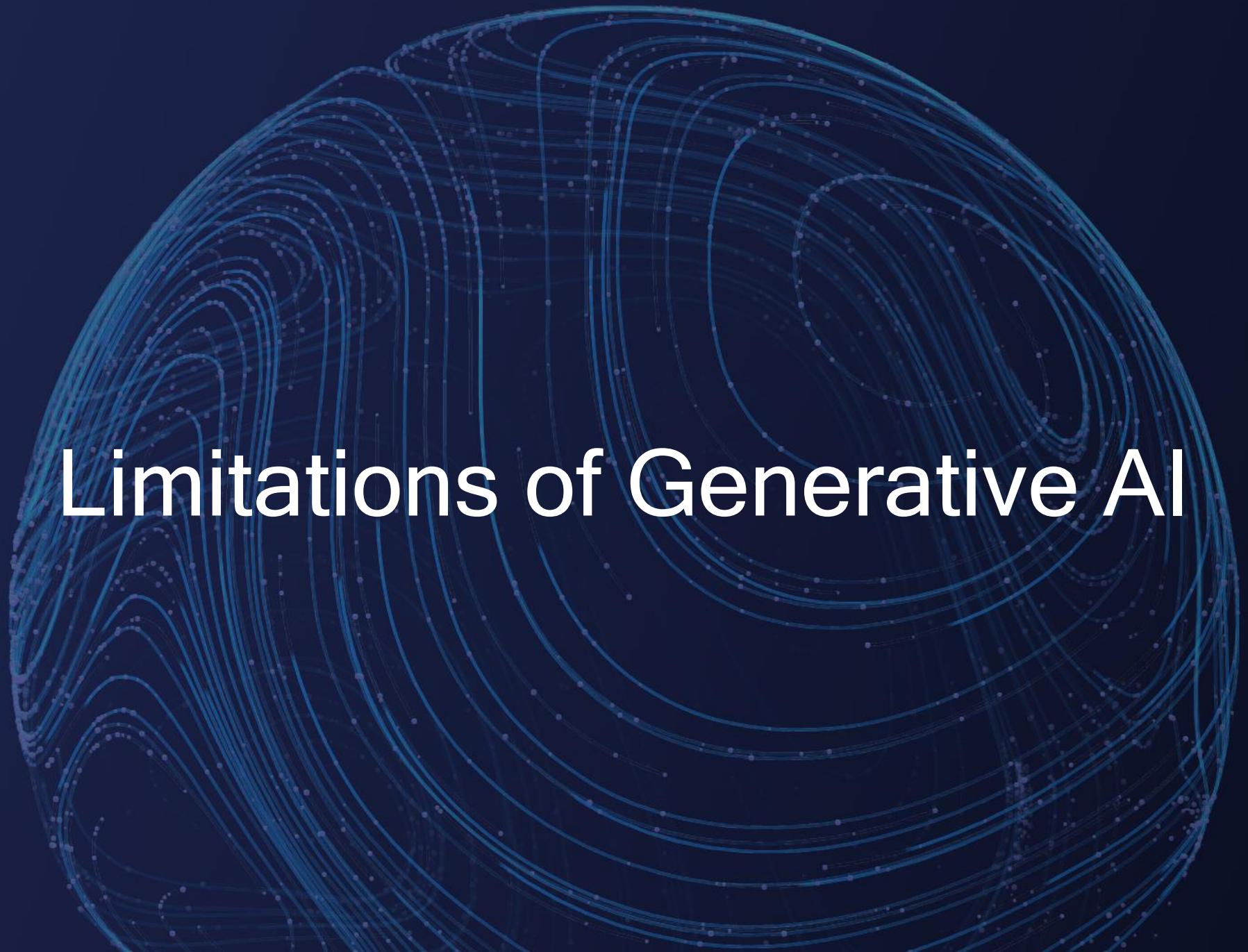
What is Generative AI, and how does it work?

- Generative AI = artificial intelligence that can *generate content*
  - e.g., images, music, or text
- Large language models (LLMs) = generative AI model for language tasks
  - e.g., answering queries, summarizing, translating, or generating text
- LLMS are trained on vast amounts of language data
  - e.g., millions of webpages, books, articles, etc.

What is Generative AI, and how does it work?

- Generative models are trained to predict the next word in a text
  - model uses previous words as context
  - assigns weights to previous words by their estimated importance
  - model predicts a word, then compares to actual word (from the training data)
  - model updates based on how close it was to being correct
  - process repeated over billions of words, across billions of texts

- Model learns the patterns, styles & structures of the language
  - Results in ability to produce original, human-like texts



Limitations of Generative AI

PROMETRIC

Bias

Availability Bias

- AI favors content more readily available in training data
  - Can result in reinforcing existing misinformation

Group Attribution Bias

- AI may attribute certain characteristics to groups (due to over-representation in training data)
  - Can result in perpetuating existing stereotypes and prejudices

Linguistic Bias

- AI favors prevalent linguistic styles, vocabularies, or cultural references
  - Can result in language use that is more relatable to certain groups
  - Regionally, culturally, or socio-economically exclusive language
    - e.g., idioms, colloquialisms, cultural or regional references, etc.

Inaccuracies

- Training data may include out-of-date, inaccurate; or conflicting information
- When predicting the next word, the model samples from a list of probable words
  - allows for greater creativity/diversity of output - but can result in inaccuracy
  - e.g., Google's Bard chatbot erroneously claimed the first images of an exoplanet were taken by the James Webb Telescope (when in fact, they were taken much earlier).

Hallucinations

- Information not present in the training data or grounded in any reality
- Not real concepts, people, places, or events – but stated as facts by the AI
- e.g., In 2023, a US attorney used ChatGPT to write a motion for a case; later discovered to be full of fabricated information and phony legal citations.

Untailored Results

- Freely available generative AI tools (e.g., ChatGPT) are general purpose
  - Not trained to perform any specific task - 'jack of all trades, master of none'
  - Fun to play around with, but difficult to achieve desired results

- For example, to develop test questions, the user must input all the requirements:
  - Format: multiple choice, fill-in-the-blank, short answer, etc.
  - Content: topic/theme, target language knowledge/skills, proficiency level, etc.

- Still, no guarantee that items will be appropriate for the test – and inputting test information introduces security concerns!

Security

- Risks of entering test information as prompts into publicly available AI tools:
  - sharing company's intellectual property with 3rd party
  - exposing confidential test information (e.g., test specs, example items, etc.)
  - compromising integrity of test --> damaging trust in test scores

- Generative AI tools may utilize user input for model training
  - model may reproduce content for other users
  - e.g., a content developer who enters test specifications into the tool may be teaching the model to reproduce test questions for other users (potentially test takers!)

General vs Customized Models

How does a Customized AI model differ from a General model?

- General purpose AI models:
  - trained on an extremely broad data set
  - not designed or trained for any specific task
    - used for emails, blogs, product reviews, resumes, etc.
  - often freely available to the public (security risks)
  - e.g., ChatGPT, Claude, BERT

- Users make queries (in the form of prompts) to get a response
  - prompts must be detailed & well-crafted – length, tone, context, ex's, etc.
  - trial and error; very time-consuming

- Some item authoring tools --> AI enhancements (e.g., chatboxes)
  - basically large language model (LLM) plug-ins
  - chatbox accessing a general LLM (same limitations)

How does a Customized AI model differ from a General model?

- Customized AI models:
  - trained on customized data
  - designed for a specific task
  - licensed and secure

- Customized AI item writing tools (e.g., Finetune's Generate)
  - interface designed as *AI-human-hybrid solution*
  - models developed specifically for each test

- Training data:
  - assessment and psychometric best practices
  - test blueprints, specifications, item writer guidelines, etc.
  - high-quality example items from the test

Benefits of Customized AI

Creativity

For the item writers

- Customized AI gives item writers a 'creative boost' - no more writer's block
  - Comes up with new and original ideas for items (*appropriate for the test*)
  - Supports writers in developing innovative new item types (scenarios for VR, live labs, simulations, etc.)

For the testing organization

- Customized AI can help enhance the diversity of items
  - Greater coverage of the construct (target knowledge, skills, abilities)
  - Less vulnerable to ‘over-practice’ by test takers
  - *Innovation begets innovation* – AI supports further advancements in assessment

Efficiency

For the item writers

- Customized AI gives item writers a 'jump start' – generates (editable) 'first draft'
  - Shifts time/effort from idea generation → editing & refining
- AI can generate high volume of items in ‘batches’
  - Select the best items, discard the rest & regenerate as needed

For the testing organization

- Higher volume of items = item bank health & test security
  - Sufficient replacement items in case of large-scale breach
  - Reduced exposure guards against cheating, harvesting & leaks

Consistency

For the item writers

- Customized AI generates items *aligned with test specifications*
  - Helps item writers meet assignment criteria
  - Supports new/returning writers & retain experienced writers

For the testing organization

- Customized AI tools can be shared among item writers
  - Achieve greater standardization of content
  - Support program continuity as contractual writers come and go

Reliability

For the item writers

- Customized AI tools can generate items from specific (pre-loaded) resources
  - Ensures all test questions are linked to learning materials (textbooks, curriculums)
  - Saves item writers' time searching through materials to find ideas for questions

For the testing organization

- Customized AI models can provide reference citations for items (validity & defensibility)
  - Helps writers ensure items are accurate and verifiable
- Customized AI models can be updated with new test resources
  - Knowledge is not fixed like general models

Security

For the item writer

- All the benefits of generative AI without the risks of general online tools
  - Avoid exposing confidential test information to 3rd parties or the public

For the testing organization

- Customized AI tools ensure security & ownership of test content
  - Access is restricted so test content remains secure
  - Generated content is owned by the testing organization

Best Practices for AI Implementation

1. Using a *Customized* AI Tool

Validity

Content is developed according to principles of assessment

Items are aligned with the objectives and requirements of the test

Security

Avoid security risks involved with using general AI tools

Confidential test information and content is secure

Protect the integrity of the test and meaning of test scores

2. Involving *Expert Humans* at Every Stage

AI model development

Test developers working with AI scientists & measurement experts to develop model

Item writers/subject matter experts testing & giving feedback on item quality

Content development

Collaborative AI-human hybrid item writing = human item writers *working with* AI

Items reviewed and edited by item writers and content development experts

Ensuring accuracy, fairness, and effectiveness of the content

3. Ensuring Ethical and Responsible Use of AI

Prometric Launching Digital AI Badge

Deployed internally – launching March 3

Educating Test Developers & Item Writers

Important for Prometric as a testing organization to ensure the responsible use of AI

Developed an AI badging program for our staff – will soon be available to others

Course outlines the benefits and limitations of AI, and best practises for implementation

AI Implementation

Planning

Objectives

- Assess goals for AI implementation – realistic, appropriate?
  - Consider needs of the test & capabilities of AI

Timeline

- Assess expected timeline – realistic, achievable?
  - Investment into custom AI model → payoff in item quality

Team

- Involve expert humans at every stage
  - AI scientists, measurement experts, content development specialists, item writers & reviewers

Customization

Test Information

- Blueprints or specifications
- Item writer guidelines or style guides
- Item taxonomies, difficulty targets, etc.

Example Items

- High-quality test items – quality over quantity
- Representative of test specifications

Challenges & Opportunities

- Identifying difficult-to-develop item types
- New item types

Piloting

Participants

- Trial with small group of *experienced* item writers/reviewers
- Supervised by Content Development Specialists

Process

- Trusted item writers 'test' the tool
  - Evaluate and edit the AI-generated items
  - Edits from trusted item writers help refine the model
- Content Dev Specialists review AI-assisted items for quality
- Feedback on item quality to AI developers
- Iterative process: model may need updates before deployment
- After deployment: AI model learns & gets better with use

Implementation

Company Education

- Educating test development team on ethical & responsible AI use

Item Writer Training

- Foundational item writer training remains essential
- Practical training on the AI tool *& ethical, responsible use*

Ongoing Support

- Learning how to use AI tools requires time & support
  - Some item writers adapt more easily than others

Expert Item Review

- Expert human review & editing of items is critical
  - Ensure accuracy, fairness, & appropriateness of test content

Summary

- What is generative AI, and what is its potential for enhancing the content development process?
- Limitations of generative AI and the differences between general and customized models
- Benefits of customized AI and the synergy between AI and human item writers and test developers
- Best practices for AI implementation:
  - Customized AI models for validity & security
  - Ethical and responsible AI integration
  - Continued reliance on human expertise in the content development process

PROMETRIC

Thank you

Contact

Michael Holaday michael.holaday@prometric.com
Jennifer Flasko jennifer.flasko@prometric.com

PROMETRIC