Evaluation Methods used for GenAI in Educational Contexts

Landscape Analysis and Evidence Synthesis Project Plan

Corresponding author: John Whitmer (jwhitmer@fas.org)
Last updated: 7/29/24

Running Meeting notes

Abstract

With the breakthrough of ChatGPT, large language models are having a tremendous impact on many fields, including education. A recent analysis by Lawrence Holt identified several EdTech product areas that Generative AI is being used for; this analysis tells us where products are being made, but it does not tell us how well these solutions work - and what evidence is being used to evaluate them.

Meanwhile, there is a deluge of open access, peer reviewed papers on Generative AI research that scholars and technologists are creating (initial hand-curated list with >100 papers here). Further, education measurement has well-established principles to evaluate assessments: validity, reliability, and fairness. This project seeks to bring these ideas together in a literature landscape analysis, and if that goes well, a synthesis of these approaches. This work will be conducted via a distributed group of researchers that are affiliated with the National Council of Measurement in Education AI Subcommittee.

Research Questions

This meta-analysis seeks to synthesize available literature about LLM model deployments in practice in order to answer the following questions:

1) In research studies, how often do researchers evaluate the validity, reliability, and fairness in studies applying Large Language Models? Do the use of these methods vary by the application area?
2) What methods are most frequently used to determine validity, reliability and fairness? Are there methods that are commonly used, and do these vary by application area?
3) Are these studies sufficiently similar that recommendations for baseline measures and metrics could be created? (future topic)

Application Areas & Committee Leads

We anticipate that the approaches to evidence may vary by the application area, and also seek parameters to guide and constrain this area of research. We have identified the following areas and members to lead initial scans in those areas.

~Proposed Evaluation Methods & Analysis Criteria

In initial meetings, we have identified the following as a draft list of evaluation methods and other criteria to use in the study analysis. These are anticipated to be refined over time.

Evaluation Constructs

Each of these methods should be analyzed by the application of whether it is used for

a. Validity - are the results appropriate to the desired purpose?
b. Reliability - does the system produce consistent results to the same question?
c. Fairness - are results tested to ensure that they are equally valid/fair for students (or other users) from different family backgrounds, social identities, and/or education levels?

Review Methods

1. Informal internal review - researcher conducts ad hoc analysis of outputs and results to see if they appear reasonable.
2. Human Subject Matter Expert Review - human experts review the results, usually through a rubric and experts not conducting the direct research
3. Automated (rule-based) evaluation methods - use non-AI automated approaches based on rules and decision systems to classify outputs; these could include semantic analysis, word counts/complexity measures, and other approaches.
4. Classification-based approaches (classify outputs) - results are classified using a machine learning or other method to classify results.
5. Automated/LLM-based evaluation systems (e.g. GPTScore; synthetic test data generation using LLMs - e.g. Ragas)
6. Datasets (external) included in evaluation - does the evaluation rely on external datasets and/or other baseline measures? If so, what are those?
7. Acceptance Criteria - what thresholds or measures are used to evaluate whether a model is successful?

Proposed Stages/Timeline

Item	Description	Deliverable	Timeline	Status

Review / Edit Current Zotero Group	Review literature in current group; de-duplicate entries, identify studies with empirical results. Find authors and metadata to use for literature review.	Short list of initial studies to categoriz e Metadata for longer research	1 week	Complete
Scope Project	Narrow in on key research questions, criteria for inclusion in literature, fields for analysis, scope and participation in project, Create a pre-registration for the study	Pre-regist ration for study	2 weeks	Complete
Initial Criteria Test	Conduct initial scan of research topics by area to test evaluation criteria and project scope	Initial dataset	6 weeks	CURRENT
Update Criteria	Revise criteria and update data collection approaches and participants	Updated materials	2 weeks	
Expand Literature Review	Conduct additional literature review to find studies.	Expande d list of studies for analysis	2 weeks	
Filter Results to Determine Sample for Analysis	Initial scan to ensure that studies meet criteria (e.g. are empirical with historical data or real students, include evaluation measures.	Final list of studies for analysis	1 week	
Classify/analyze found studies (some dual review	Review and classify studies	Study notes and classificat ion	8 weeks	

Preliminary Results	Summarize results		2 weeks	
Final Results	Write up full paper		4 weeks	

Open and Reproducible Science

All results for this project will be posted using an OSF Project, including the data files, the analysis scripts, and other resources. The results will be published as open pre-prints, although final articles may be published to journals

Googlesheet for literature analysis here

Search Keywords

The goal of this literature review is to identify papers (both pre-prints and peer-reviewed) that provide evidence of the quality of Generative AI outputs for education use cases. We are conducting this literature review to identify common measurement approaches being used and to synthesize some of the results from the field.

Keyword search terms
"Generative AI" or "GenAI" or ""Large Language Models" or "LLM" or "ChatGPT" AND "Education" or "K12 education" or "Higher education" or "education research" OR "Research" or "research study" or "evaluation" or "model quality" (because I think few publications won't be about research)

Date: > 2022 (e.g. advent of LLM/OpenAI)

Peer reviewed and not peer reviewed

Examples of the type of papers I think we'd like to see

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity* (arXiv:2302.04023). https://doi.org/10.48550/arXiv.2302.04023

Choi, J. H., Garrod, O., Atherton, P., Joyce-Gibbons, A., Mason-Sesay, M., & Björkegren, D. (2023). *Are LLMs Useful in the Poorest Schools? theTeacherAI in Sierra Leone* (arXiv:2310.02982). arXiv. https://doi.org/10.48550/arXiv.2310.02982

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial In a Large-Scale Online Course. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://edworkingpapers.com/ai21-483

Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2023). *How Useful are Educational Questions Generated by Large Language Models?* (arXiv:2304.06638). arXiv. http://arxiv.org/abs/2304.06638

Pardos, Z. A., & Bhandari, S. (2023). *Learning gain differences between ChatGPT and human tutor generated algebra hints* (arXiv:2302.06871). arXiv. https://doi.org/10.48550/arXiv.2302.06871

Wang, R., & Demszky, D. (2023). *Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction*. OSF. https://doi.org/10.35542/osf.io/5vrby

Databases to use for search
Edarxiv
ERIC (education resources information clearinghouse)
(whatever else you have access to!)