AI-Assisted HQIM Review in South Carolina

A White Paper presented to SCDOE April 2025

Examining the efficacy of rubric-based quality evaluation in the textbook adoption process

teg

Executive Summary

AI-Assisted High-Quality Instructional Material Review in South Carolina

In 2024, the South Carolina Department of Education (SCDOE) partnered with Trinity Education Group (TEG) to pilot the use of EVA (Evaluation Assistant)—an AI-powered review system—to support the state's adoption of high-quality mathematics instructional materials.

Key Findings

- **High Correlation with Teacher Reviews**: EVA's review outcomes aligned strongly with SCDOE's teacher team scores (r = .81), demonstrating that AI can support or replicate expert educator judgment.
- **Full-Text, Standards-Aligned Evaluation**: EVA ingested and analyzed complete textbooks across 28 materials, evaluating each against South Carolina's custom HQIM rubric, based on EdReports and state standards.
- **Data-Rich, Granular Output:** EVA produced highly detailed, standards-aligned feedback that outperformed traditional correlation documents, especially in Gateway 1 (standards alignment), where it often identified evidence human reviewers missed.
- **Efficiency and Continuity**: Unlike teacher teams who must stop reviewing if a material fails a gateway, EVA can complete full reviews regardless—supporting iterative improvement and informed publisher feedback.

Recommendations for State Adoption programs and State Boards of Education

- **Enhance Transparency & Consistency**: Use EVA to standardize the HQIM review process and minimize variability across reviewers and review cycles.
- **Streamline Resource Allocation**: Incorporate EVA to reduce reliance on large teacher review teams, saving time and budget while maintaining rigor.
- **Leverage EVA alongside expert teachers**: Use EVA in pre-screening or reevaluation phases to prioritize reviewer time and accelerate decision-making.

EVA represents a scalable, efficient, and consistent "assistant" for states seeking to implement HQIM evaluation into their instructional materials adoption processes.

TABLE OF CONTENTS

Introduction

Background

The growing evidence in support of high quality instructional materials (HQIM) has led to an increase in state policies and practices supporting HQIM adoptions and stronger alignment of materials to rigorous state standards. In 2024, ExcelinEd released a report indicating that 33 states within the United States had established some form of guidance to assist school districts in the adoption of High Quality Instructional Materials (HQIM). The nature of this guidance, however, was found to vary considerably from state to state.

In some states, the guidance took the form of a recommended list of instructional materials, while in other states, the guidance was more prescriptive, mandating that districts select their instructional materials from a state-approved list.[1] The nature and quality of guidance around HQIM implementation was found to vary considerably from state to state. As states and local districts try to implement HQIM approaches, there are many challenges to consider, including teacher training, district evaluation of materials, and state adoption process changes to ensure HQIM options are available to every district.

Problem Statement

It is this last challenge that South Carolina found itself wrestling with in Spring 2024. When the time came to start the mathematics materials adoption process, the Chief Academics Officer (CAO) and the Instructional Materials and Standards team recognized an opportunity to improve the adoption process by creating its own HQIM rubric to align to new Mathematics standards. Textbook reviews are time consuming regardless of whether it is an HQIM review or not, and the complexity of the new HQIM process on top of the normal review process meant that there was a steep learning curve, coupled with a time crunch to complete the adoption within the legislatively defined timeline. All of these factors, an inherently time consuming process, new layers of complexity around HQIM, and the need for reviewers to upskill, created a situation where technological innovation was necessary.

Several South Carolina DOE staff attended a Council of Chief State School Officers (CCSSO) conference at which TEG was demonstrating a new way to support instructional resource alignment utilizing generative AI with CASE-compliant standards libraries. Inquiry into the capabilities of TEG's tool–known as the Evaluation Assistant, or EVA–started a conversation between SCDOE and TEG about how to innovate to support SCDOE's new HQIM adoption initiative.

---

[1] CurriculumHQ, 2024

The partnership with TEG offered SCDOE clear advantages in delivering on their new HQIM adoption process. The SCDOE was able to use EVA to quickly and efficiently analyze and review a selection of mathematics textbooks, a boon for them in their new HQIM review process. It further provided a wealth of digital data from which to better inform publishers of how to improve or enhance their HQIM materials to best meet SCDOE's needs. The partnership also provided TEG with valuable insights and data into their tool, as well as its performance, efficiency, and efficacy within the adoption context.

Purpose of White Paper

The purpose of this paper is to explore the innovative use of AI, specifically TEG's EVA, to review instructional materials against both state standards of learning as well as new, high-quality instructional material rubrics, such as that from EdReports.[2] This paper will seek to provide a comprehensive understanding of the process that was used to complete the reviews, take a deep dive into the results and insights gleaned from the evaluation, and where possible compare EVA results to the existing teacher-team reviews.

Furthermore, the paper will shed light on the challenges encountered during the development and implementation of EVA for textbook HQIM reviews, as well as the steps taken to overcome those challenges. There will also be a discussion of the measures implemented to ensure that the intellectual property rights of all stakeholders was safeguarded. Finally, the paper will conclude with a discussion of the future of EVA and the potential implication for the future.

Methodology

Pilot Creation Process

- Based on inquiry from SCDOE leadership, TEG proposed and constructed a pilot study of an EVA-powered review of textbook materials alongside the state's existing adoption review process, which consisted of identifying skilled, experienced educator subject matter experts organized into review teams, trained to evaluate texts for HQIM value, and given copies of the publisher-submitted materials.

[2] EdReports, n.d.

- TEG attended the SCDOE training for teacher review teams. Staff conducted interviews with SCDOE subject matter experts, teacher review teams, and publishers in order to document the training that teacher teams received.
- TEG technical staff then translated the training into a custom prompt architecture that was intended to produce a similar level and depth of instructional material review in conformance with the guidance given to the teacher team reviewers.
- SCDOE identified four basal textbook publishers whose materials would be evaluated by teacher review teams, and would also be evaluated using TEG's EVA AI-powered system.
- TEG then provided the following technical services:
    - Ingested SCDOE HQIM rubric, training and reviewer guidance from SCDOE consultants
    - Ingested SC state standards (in CASE format)
    - Ingested texts from publishers
    - Parsed all content using TEG's content parsing platform
    - Evaluated content using TEG's "context-aware reasoning" prompt engineering through a tuned and secure instance of OpenAI 4o
    - Applied recursive and self-reflective machine learning to results to confirm accuracy, reduce bias and eliminate hallucinations
    - Structured and formatted evidence to support HQIM and standards alignment
    - Provided a single administrative dashboard for the detailed and itemized review of full-text evaluations.
- SCDOE then reviewed the EVA output and shared feedback with publishers as needed.
- SCDOE, as part of the pilot project, asked TEG to compile a comparative analysis of the results of teacher review teams against the EVA system.

SCDOE's High-Quality Instructional Materials Rubric

Background

SCDOE, understanding the importance of using HQIM for improving student achievement, contracted EdReports to build a rigorous, state-approved HQIM rubric specific to South Carolina's requirements. They intentionally aligned the rubric with the South Carolina College and Career Ready Standards (SC-CCR) and the "Profile of the South Carolina Graduate" which describes the characteristics and competencies that a successful high school graduate should display.[3]

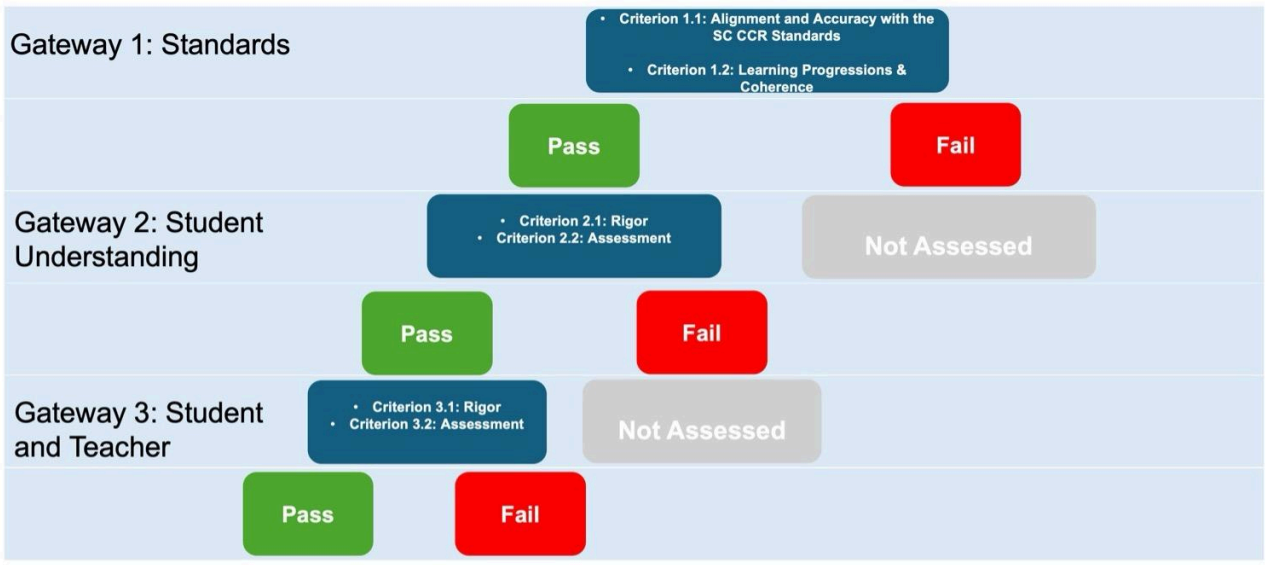
The alignment of standards and intended student outcomes ensured that the evaluation process would thoroughly assess the material's ability to support the state's educational goals and priorities. That process and intentionality led to the creation of a comprehensive, three gateway rubric that allows educators and administrators to evaluate the materials and select those materials that are most likely to lead to student success in the classroom and throughout their lives.

Design

The SCDOE rubric utilizes a similar approach to the EdReports "gateway" structure to evaluate materials. The rubric consists of three sequential gateways that are subdivided into indicators that check the material for alignment to standards and effectiveness in teaching and learning.

For a material to be approved it must be rated "exemplifies expectations" for each gateway, in order. If a material does not pass Gateway 1, then it cannot proceed to Gateway 2, and if it does not pass Gateways 1 and 2, then it cannot proceed to Gateway 3. If the material makes it to Gateway 3, and is rated "exemplifies expectations" then it is put on the approved materials list. Using a sequential system is beneficial in two ways: first, it guarantees that materials meet core expectations and standards alignment before proceeding any further, second, it saves valuable state resources and teacher review team time by stopping a review as soon as a material does not exemplify expectations within that gateway.

[3] South Carolina Department of Education & EdReports, 2024



If a material passes all three gateways it is put on the adopted materials list that districts use to select their materials.

SCDOE Structure

The SCDOE HQIM rubric is a multi-layered evaluation system made up of three gateways that serve as foundational pillars for the review. Each gateway is further divided into two criteria for a total of six criteria. Each gateway's criteria define and articulate the expectations and requirements of each gateway, and are the measured components of the gateway. Finally, each criterion is divided into multiple indicators, adding precision and granularity to the process, that are scored to determine if the material passes the criteria. It is at the indicator level that concrete evidence is gathered for the evaluation.

The table below shows the breakdown of each gateway, its criteria, and their indicators. The scores necessary to pass each criteria and gateway are also listed in the table.

	Gateway 1		Gateway 2		Gateway 3	
Criteria	Indicator 1.1	Indicator 1.2	Indicator 2.1	Indicator 2.2	Indicator 3.1	Indicator 3.2
Number of Indicators	2	3	6	4	9	6
Indicator Passing Score (Points)	8	>4	>10	>6	>14	>10
Gateway Passing Score	>12		>16		>24	

AI Solution Design by TEG

Background

At a fundamental level, EVA uses specific evaluation criteria, such as a rubric, and a set of academic standards to review an instructional material. The process allows for a quick and efficient review of an entire textbook, and produces a robust set of results. Those results show whether the content is aligned to the evaluation criteria/indicators, and the state standards. The end result is both comprehensive, giving a big picture view of the material, and granular, revealing the fine details about the alignment.

The results are then coded into a user-friendly dashboard that allows the end user to interact with, review, and use in whatever way that suits their needs and objectives. Most textbook companies provide a correlation document to help reviewers find where a standard or piece of content exists within a textbook, but EVA is able to do that independently of the correlation document. In fact it can even expand the scope of the correlation to find even more examples than would be available in a standard correlation document. This can be particularly useful when trying to determine if a specific standard passes and the evidence in the publisher's document is limited.

Data Collection

The initial Pilot Project Scope outlined that EVA was to focus on materials in grades 3, 8, and Algebra 1. The SCDOE would provide TEG with the publisher-submitted materials, the State Standards and the SCDOE rubric for the evaluation. During the course of the pilot, the grade level and number of evaluations changed as SCDOE's adoption process developed and as TEG's technical data ingestion model evolved. The initial set of reviews performed by EVA were on nine

textbooks from four different publishers covering six grade levels. The textbooks also came from a mix of proprietary and open sourced publishers.

The table below shows the mixture of grades and publisher status. Please note that the material titles have been anonymized.

Grade	Licensed 1	Licensed 2	CC-attribution 1	CC-attribution 2	Total
Grade 3		1			1
Grade 4	1				1
Grade 5			1		1
Grade 6	1				1
Grade 8		1	1	1	3
Algebra 1		1		1	2
Total	2	3	2	2	9

Given how quickly and efficiently EVA performed in the initial review stage, SCDOE requested that TEG use EVA to re-review materials that its teacher review teams did not pass. This set of reviews was slightly different from the original because in some instances SCDOE only wanted one gateway reviewed and in others they wanted the entire evaluation. This gave TEG the opportunity to review an additional 19 materials, either partially or completely.

The breakdown of the additional 19 materials is below.

Grade	Licensed 1	Licensed 2	Licensed 3	Licensed 4	CC-attribution 1	Total
Kindergarten	1					1
Grade 1	1					1
Grade 2	1					1

Grade	Licensed 1	Licensed 2	Licensed 3	Licensed 4	CC-attribution 1	Total
Grade 3	1				1	2
Grade 4	1					1
Grade 5	1					1
Grade 6	1			1		2
Grade 7	1			1		2
Grade 8	1			1		2
Algebra 1		1				1
Geometry	1	1	1			3
Algebra 2	1	1				2
Total	11	3	1	3	1	19

The combination of the original and additional materials gave a total of 28 materials to review. In the next section, using those 28 materials, this paper will explore EVA's performance relative to the teacher team reviews.

Results Analysis

Of the 28 materials that SCDOE requested TEG to review using EVA, 22 had both a teacher team review and an EVA review that could be used for comparison. This provided a substantial set of data to analyze the comparative results generated by EVA and teacher team reviewers.

The table below shows the data that was collected. The textbooks have been labeled as textbook 1, textbook 2, etc. Beside the textbook label is either a G1, G2, or G3, representing the gateway that the score represents. If a score is blank then there is no score for that particular gateway.

The table contains 22 materials with three gateways each, for a total of 66 items. In the teacher team review column there are 28 values and 39 null values and in the EVA column there are 58 values and nine null values. The null values in the teacher team review column stem from structural issues within the rubric, which will be explained more in the next section. The null values in the EVA column are present because TEG was only asked to review certain gateways for that specific text. One interesting takeaway from the discrepancy in null values between teacher team reviews and EVA is that teacher team reviews are required to stop if a gateway fails. It would be a waste of time for them to continue. But EVA is able to quickly work through the other gateways with minimal impact on time or resources, and thereby can provide valuable guidance and improvement data for publishers.

Textbook	Teacher team review score	EVA score	Difference in score	Textbook	Teacher team review score	EVA score	Difference in score
Textbook 1 G1	10	14	4	Textbook 12 G1	12	12	0
Textbook 1 G2		17		Textbook 12 G2	18	12	-6
Textbook 1 G3		24		Textbook 12 G3	24	21.275	-2.725
Textbook 2 G1	10	14	4	Textbook 13 G1	10	12	2
Textbook 2 G2		17		Textbook 13 G2		16	
Textbook 2 G3		22		Textbook 13 G3		20.5	
Textbook 3 G1	14	14	0	Textbook 14 G1	8	13	5
Textbook 3 G2	19	20	1	Textbook 14 G2		18	
Textbook 3 G3	24.45	30	5.55	Textbook 14 G3		26	

Textbook	Teacher team review score	EVA score	Difference in score	Textbook	Teacher team review score	EVA score	Difference in score
Textbook 4 G1	10	14	4	Textbook 15 G1	10	8	-2
Textbook 4 G2		17		Textbook 15 G2		14	
Textbook 4 G3		26.56		Textbook 15 G3		20	
Textbook 5 G1	14	14	0	Textbook 16 G1	10	13	3
Textbook 5 G2	18	18	0	Textbook 16 G2		18	
Textbook 5 G3	30	25.36	-4.64	Textbook 16 G3		25.42	
Textbook 6 G1	10	9	-1	Textbook 17 G1	10	12	2
Textbook 6 G2		16		Textbook 17 G2		16	
Textbook 6 G3		22		Textbook 17 G3		25	

Textbook	Teacher team review score	EVA score	Difference in score	Textbook	Teacher team review score	EVA score	Difference in score
Textbook 7 G1	10	9	-1	Textbook 18 G1	6	14	8
Textbook 7 G2		17		Textbook 18 G2		16	
Textbook 7 G3		24		Textbook 18 G3		23.5	
Textbook 8 G1	10	14	4	Textbook 19 G1	6	13	7
Textbook 8 G2				Textbook 19 G2		15	
Textbook 8 G3				Textbook 19 G3		24	
Textbook 9 G1	10	14	4	Textbook 20 G1	6	14	8
Textbook 9 G2				Textbook 20 G2		17	
Textbook 9 G3				Textbook 20 G3		23	

Textbook	Teacher team review score	EVA score	Difference in score	Textbook	Teacher team review score	EVA score	Difference in score
Textbook 10 G1	10	14	4	Textbook 21 G1	10	11	1
Textbook 10 G2				Textbook 21 G2		16	
Textbook 10 G3				Textbook 21 G3		22	
Textbook 11 G1	4	9	5	Textbook 22 G1	10	14	4
Textbook 11 G2				Textbook 22 G2		16	
Textbook 11 G3				Textbook 22 G3		24	

Broadly speaking, EVA and teacher team reviews were aligned with one another, which is borne out by the strong positive correlation (r = .81) between the two groups' scores across all textbooks. On average, the difference between the EVA scores and teacher team review scores, where a score exists for both, is 2.08 points higher from EVA.

However, the high correlation score masks one important difference in the evaluative process between teacher team reviews and EVA reviews: EVA's ability to take in large amounts of data, quickly search that data, and make connections between the data and the standards leads to a more thorough correlation and evidence collection for certain indicators, especially in Gateway 1.

To detail this important difference, Gateway 1 is the only gateway where there is a guaranteed score from both teacher team reviews and EVA (due to the fact that teacher team reviews did not proceed any further than Gateway 1 if they did not "exemplify expectations," while EVA continued the review). Looking solely at those Gateway scores, the correlation between teacher team reviews and EVA on Gateway 1 is still positive, but quite a bit weaker (r = .17), and this is directly attributed to EVA's ability to evaluate the full text for alignment to each standard. The average difference between EVA and teacher team review scores is three points higher for EVA.

When reviewing the data from teacher team reviews one obvious difference becomes apparent quickly. Gateway 1 focuses entirely on the alignment of the South Carolina College and Career Ready standards. Teacher team reviews rely on correlation documents to find relevant information in the textbook. If they do not find it, or it does not satisfy what they believe to be strong alignment, they can score the material lower. The reliance on correlation documents is a limitation of teacher team reviews: textbooks are lengthy and complex with unique organizational structures, and reviewers often lack the time to search for more evidence if the correlation document they have is wrong or does not provide enough information.

EVA, on the other hand, does not have this same limitation. EVA ingests the textbook in its entirety, and is therefore able to quickly and efficiently search for further evidence, outside of what is given for the review via a correlation document. This means that EVA is often able to find evidence of alignment to standards whenever a teacher team review cannot, and that in turn leads to higher scores overall for Gateway 1 in the Pilot.

One unfortunate outcome of a less extensive teacher team review of standards alignment is that publishers may dispute the results of a review, and may have more time and resources than the state's teacher review team to identify and support the validity of their correlations. EVA has the potential to provide a wealth of rich, page-specific detail with which to back-up its correlation findings in Gateway 1, making it much less likely that publishers may dispute the results.

Discussion

The collaboration between SCDOE and TEG showed the viability of using AI to review textbooks for HQIM adoption and the overall high correlation of EVA results to traditional teacher team reviews. This opens up many possibilities for new technology to support or enhance statewide textbook adoptions models, even when the adoption seeks to evaluate materials for conformance with complex rubrics like EdReports' HQIM or other quality rubrics.

Opportunities

- Create a more efficient review model: The data overwhelmingly suggests that EVA provided high-quality reviews, and in cases where books failed the teacher team review review but passed the EVA review, it provided insights and analysis beyond what a teacher team review could do within the time and resource constraints they have to work within.

  The ability to continue a review even when gateways fail creates efficiencies and cost savings in multiple ways. Firstly, it limits the number of reviewers needed for any given review as well as the amount of time that state subject matter experts have to devote to the process. Secondly, EVA is able to complete an entire review even if a gateway fails. For teacher team reviews this would be costly in both time and resources, but it is beneficial because it helps with the reevaluation process. Textbook publishers that want to resubmit their materials can identify issues beforehand and make any necessary corrections, leading to significant time savings and avoiding legal entanglements later in the process.

- Mitigate Risk through consistency and objectivity in the review process: The adoption process for publishers is a high-stakes, competitive endeavor. Ensuring a consistent, fair and objective review process for all submissions was critical for SCDOE. The use of EVA in the review process reduced the risk of teacher team review bias–positively or negatively–and the appearance of bias. By using EVA, independently or in collaboration with teacher team review, the risk for bias in quality-based rubrics is reduced. Studies show how systematic review processes that depend on quantitative factors are favored over qualitative reviews, partially to remove bias.[4] A separate paper on the textbook selection process in the United States found that one factor that could impact textbook adoptions was bias coming from external groups such as publishers and other groups.[5]

[4] Fey & Matthes, 2018
[5] Watt, 2009

Further research on the effectiveness of group decision making says that one liability is social pressure to conform to the ideas of the committee.[6]

The EVA model works on a "just the facts approach" that does not require teacher team review judgement and/or incomplete data. An AI-assisted approach with definable, *defensible* prompting strategies brings objectivity and consistency to the review process, and ensures outputs are based entirely on data.

Fully or partially automating the review process improves accuracy and efficiency without sacrificing fairness and consistency. A possible use for EVA and other AI-enhanced technology, could be to perform an initial review and then have a teacher team review those results, and sign off on the full review. Such a model would control risks inherent to AI and also control risks that come from using teacher team reviewers.

- Provide equity by helping to facilitate an HQIM process with fidelity: The main reason for using HQIM is to raise the quality of materials for students and ensure that every student is getting access to the same high standard of learning materials.[7] As discussed earlier, EVA can evaluate materials quickly, cheaply, efficiently, and consistently. Using EVA ensures that any cost-based arguments against HQIM (such as "It's too time-consuming" or "it's not worth the cost to rewrite textbooks", etc.) can be overcome.

  A benefit not to be overlooked is that the use of AI also ensures that the same rigorous process is being run for every single material, no matter the subject, grade level, or reviewer ability. This consistency ensures materials compete on the same level, and that the only metric that they are being judged on is quality and alignment to the state standards. States, districts, families, and all other stakeholders can be confident that an EVA review was done in a manner that was consistent and fair, and the best materials, based on the rubric, rise to the top.

Challenges and Limitations

The data and analysis in this paper shows that the benefits of EVA are compelling, but as with all new technology limitations and concerns exist. One of the unexpected positive benefits of structuring the Pilot Project with SCDOE in a flexible way is that, while issues were identified, many were able to be addressed in the iterative technical approach TEG took. The issues surfaced helped to shape the prompting, even the process of ingesting materials, in order to make EVA as consistent and effective as it became.

[6] Bates, 2014
[7] The New Teacher Project, 2018

The main weaknesses the team found were: Misaligned content, hallucinations in the review output, and bad examples provided as evidence of indicators and criteria

- Misaligned Content: Early runs of EVA on full-text materials found some misaligned review items. These misalignments occurred within areas of the standards that were vague or were very close to other standards. These misalignments also occurred with mathematical practice standards when the math practices were not explicitly called out within the publisher text.

  TEG's technical team reviewed these misses and adjusted the prompt workflow to remedy them. One of the improvements was to apply extra context when running reviews. Instead of using just the standards, the team sought out and found supporting documents from SCDOE such as instructional focus documents and standards documents that gave more detail about individual standards. Using extra context helped ensure that the standards matching was statistically valid and no further misalignments were present.

- Hallucinations: Another issue the team discovered early on was hallucinations in the review output. The initial cause of the hallucinations was that TEG's tech had not accounted for the sheer size and complexity of full-text publisher materials submissions. Often, these submissions were in excess of 1,500 pages per grade level. The hallucinations were interesting artifacts because they were indeed false statements, but not created out of whole cloth. Instead, the LLM created "examples" by summarizing the content and presenting that as evidence, even when the "example" needed to be manufactured by the LLM.

  Once this issue was discovered, TEG's tech team took several steps to ensure that it did not continue to occur. While some of these steps remain proprietary to EVA's algorithms and code, one step that illustrates the team's problem-solving was to tell EVA to use only the evidence on the page, the same as teacher reviewers. While this limited the creativity of EVA, it also led the team to discover that EVA was able to make inferences from the content when correctly prompted and provided with linguistic guardrails.

- Bad Examples: Finally, EVA occasionally chose bad examples as evidence. The bad examples were similar to the misalignments that the team found early on. The model misunderstood the standard and/or the content and found examples that matched the misunderstanding.

  This issue was remedied in large part by fixing the previous two issues discussed. Adding context and forcing EVA to explicitly use the content within the materials eliminated a large swath of the bad examples.

- Praxis and Perimetric Knowledge: After the reviews were complete the team discovered that the model struggled with Gateway 3, which is focused on supports for teachers and

students. This gateway requires the most external knowledge of the three. For example, it measures a textbook's recall of past grade level content and its foreshadowing of future content. Teacher reviewers, who are almost all experts in the grade level content they are reviewing, know this implicitly, and that makes it easy for them to score this section. The version of EVA that TEG used for this review cycle had no capacity to account for that perimetric, or external, knowledge and the scores and evidence in that section were not as strongly correlated to teacher team reviews as Gateways 1 and 2. TEG's tech team believes that this limitation can be addressed in future adoption processes by ingesting the full grade range for specific publishers.

Intellectual Rights Protections

An important part of obtaining the cooperation and participation of publishers in this Pilot Project was TEG's ability to respect and technically protect the intellectual property of all stakeholders. TEG developed a unique, secure methodology for implementing EVA during the Pilot Project. This methodology was designed to maximize EVA's ability to access the data in the material while simultaneously ensuring that no copyrighted material was used in an unauthorized manner.

Because EVA is not a RAG[8] model of AI, it can prohibit EVA's LLM from using any of the materials for training purposes. Once a particular publisher evaluation was completed, original source materials and data were expunged from EVA and destroyed. Further, TEG utilized a secure Azure server exclusive to this project, so no data was provided back to OpenAI or any other LLM. These steps allowed TEG to process and review materials, use fair use quotes to validate those reviews, but not infringe on the intellectual property rights of any publisher or organization.

[8] A "RAG" model is an AI process known as Retrieval-Augmented Generation (RAG), which optimizes output of a large language model by referencing a knowledge base (such as a textbook) outside of the LLM's training data before generating a response.

Conclusion

Summary of Findings

This paper explores AI, specifically TEG's EVA AI, as a review tool for textbooks and instructional materials. It uses the pilot project between TEG and SCDOE to delve into the capabilities and outcomes of EVA's use in the review process. The key findings include the following:

- Overall Alignment: While there were differences between EVA and teacher team reviews, the difference was not large. EVA scores were on average around 2 points higher than teacher team reviews. The strong positive correlation between EVA and teacher team review scores (r = .81) suggests that EVA was generally inline with the work that its teacher team review counterparts produce.

- Gateway 1 Differences: Even though EVA and teacher team review scores were generally in line with each other, differences were noticeable when viewing individual components of the rubric. EVA outscored teacher team reviews on Gateway 1 of the rubric by an average of three points. The difference is also apparent in the lower correlation (r = .16) between EVA and teacher team review on Gateway 1.

  The difference is due to advantages inherent in using AI. Specifically, EVA is able to ingest and reference the entire material for the evaluations of standards. This ability gives it an advantage over a teacher team review reviewer who is reliant upon correlation documents to find evidence. Because EVA can scour the entire material, it can find evidence that teacher team review reviewers miss because of their limitations in time and resources.

- Efficiency and Cost Savings: EVA presents an opportunity for a more efficient and cheaper review process. By using EVA the upfront work, including preparing materials and finding and hiring reviewers, can be reduced dramatically. Those gains in efficiency should translate directly into cost savings.

  Another cost reduction could come from the need for fewer reviewers or time to complete a review. There are multiple ways EVA usage could lead to cost reductions. EVA could replace a reviewer or team of reviewers, which would reduce labor costs. EVA could also act as an aid that streamlines parts or all of the review, which saves time, and in turn saves money.

- Consistency and Objectivity: The use of EVA can mitigate the risk of bias that teacher team reviews bring to a review process. That reduction leads to fairer and more consistent reviews that push quality to the forefront. That shift also functions as a risk

mitigation for states because publishers cannot claim their scores were impacted by biased individual reviewers.

- Equity and HQIM Fidelity: Using EVA for a review is a great way to guarantee that the process is consistent, and that materials are judged only on their quality and alignment to standards. The same process will occur every time, no matter the grade or subject. Ultimately this will build confidence in the process among all stakeholders, as they understand that no matter the outcome the process to reach it was the same for every reviewed item.

Research and References

AI Research

EVA's technical development was informed by many sources. The following list highlights some of the key resources that – while not explicitly referenced in the paper – represent the research contribution that TEG's technical team provided.

ANU College of Engineering and Computer Science. (n.d.). *InfluenceMap*. InfluenceMap. Retrieved April 22, 2025, from https://influencemap.cmlab.dev/

Chiang, C.-H., & Lee, H. (2023, May 3). Can Large Language Models Be an Alternative to Human Evaluation? *Arxiv.org*. https://arxiv.org/pdf/2305.01937

Eitan, A. T., Smolyansky, E., Harpaz, I. K., & Perets, S. (n.d.). *Find and explore academic papers*. Connected Papers. Retrieved April 22, 2025, from https://www.connectedpapers.com/about

Knoth, P. (2013, July 8-12). From open access metadata to open access content: two principles for increased visibility of open access content. *Open Repositories*, *2013*. https://oro.open.ac.uk/37824/

References

Bates, S. B. (2014, September). Committee effectiveness in higher education: The strengths and weaknesses of group decision making. *Research in Higher Education Journal*, *25*. https://files.eric.ed.gov/fulltext/EJ1055342.pdf

CurriculumHQ. (2024, October 31). *Trends in State Action*. CurriculumHQ. Retrieved April 18, 2025, from https://curriculumhq.org/trends-in-state-action/

EdReports. (n.d.). *Explore Reports*. EdReports. Retrieved April 21, 2025, from https://edreports.org/reports/math

Fey, C.-C., & Matthes, E. (2018). Textbook Quality Criteria and Evaluation. In *The Palgrave Handbook of Textbook Studies* (pp. 157-167). Palgrave Macmillan US.

The New Teacher Project. (2018, September 21). *The Opportunity Myth* [Online Article]. TNTP.org. https://tntp.org/wp-content/uploads/2023/02/TNTP_The-Opportunity-Myth_Web.pdf

South Carolina Department of Education & EdReports. (2024, November 13). *The SC Review Tool & List Development*. Statewide Webinars for the Textbook Adoption Process. Retrieved April 18, 2025, from https://docs.google.com/presentation/d/1BfRu65MSSKnMT2OFprGJpco82Ay5ambgfXpk0Vq4YTs/edit#slide=id.g12f1afe3aaa_0_0

Watt, M. G. (2009). Research on the Textbook Selection Process in the United States of America. *IARTEM e-Journal*, *2*(1). https://files.eric.ed.gov/fulltext/ED506523.pdf